

한국정보과학회 학술심포지움 논문

UCWIT 2023

2023 디지털 정보 통신 기술 포럼

Proceedings of UCWIT 2023

Department of Information Convergence Engineering

Pusan National University, Nov. 24-25, 2023

제 16권 제2호

일시 : 2023. 11. 24(금) ~ 2023.11.25(토)

장소 : 부산대학교

주최 : 한국정보과학회

(사)한국과총 부산울산지역연합회

주관 : 한국정보과학회 영남지부

부산대학교 Digital-X AIoT 연구센터

협찬 : MTDATA, 대신정보통신, (주)우리아이티

대회장 인사말

어느덧 겨울이 오고 2023년도를 마무리하는 이 시점에 올 한 해에도 많은 성과가 있으셨기를 바랍니다. 명실상부 IT 분야 전국 최대 규모의 교육, 연구 및 산학협력 인프라를 보유하고 있어 국내외로 우수한 IT 전문인력을 배출하고 여러 연구성과를 자랑하는 부산대학교에서 2023년 한국정보과학회 영남지부 학술대회를 개최하게 되어 매우 기쁘게 생각합니다. 본 행사에 참석해 주신 많은 회원 여러분과 본 행사를 위해 준비해 주신 관계자 분들에게 깊은 감사의 말씀드립니다.

한국정보과학회는 컴퓨터와 소프트웨어 및 정보통신 분야에서 우리나라 최고의 학회로 발전을 거듭하고 있으며 국내외 IT 산업에 크게 기여하고 있습니다. 본 학술대회 행사는 한국정보과학회와 사단법인 학국과총 부산울산지역연합회가 공동 주최하고 한국정보과학회 영남지부, 부산대학교 Digital-X AIoT 연구센터가 공동 주관하고 있습니다. 한국정보과학회는 기술 분야별 소사이어티 활성화와 지부 운영을 적극 지원하고 있으며 영남지부의 본 행사와 같은 활발한 연구 활동은 한국정보과학회 발전을 비롯하여 국가 IT 발전에 지대한 공헌을 할 것이라 기대됩니다.

한국정보과학회 영남지부는 "유비쿼터스 컴퓨팅과 웹 정보기술 (Ubiquitous Computing and Web Information Technology, UCWIT)"이란 명칭의 학술대회를 지금까지 운영해 오고 있습니다. 본 행사는 유비쿼터스와 웹 정보기술을 중심으로 인공지능과 빅데이터를 비롯한 정보 통신 및 전통 산업과의 IT 융합 기술을 주제로 분야별 전문가들이 모여 최신 주제를 토론하는 학술대회입니다. 특히나, 요즘 인공지능, 빅데이터, 블록체인, 메타버스, 6G와 같은 4차 산업혁명을 주도하는 요소 기술들이 급속도로 발전하는 가운데 본 행사를 통해 해당분야의 다양한 연구 성과들을 공유하며 토론할 수 있는 좋은 기회가 될 것이라 기대됩니다.

끝으로, 그동안 본 학술대회 개최를 위하여 애써온 조직위원장, 학술위원장, 집행위원장 및 위원님들의 노고에 심심한 감사의 마음을 전합니다. 또한 본 대회 개최를 위해 장소를 제공하고 여러 가지 지원을 해주신 부산대학교 관계자 여러분과 후원기관 여러분에게 고마움을 전합니다. 이번 학술회의에 참석하신 참석자 여러분의 관심과 협력에 깊은 감사의 말씀을 드리며 항상 건강하시길 바랍니다.

대회장 한국정보과학회 회장 **이 원 준**

조직/학술 위원장 인사말

열역학 제2법칙에 따르면 닫힌계에서 엔트로피는 점점 증가한다고 합니다. 우리가 사는 세상도 이 법칙을 따르는 것 같습니다. 새로운 질병이 나타나기도 하고, 새로운 전쟁이 일어나기도 합니다. 세상이 점점 어지러워질 때 등장하는 구세주는 이젠 당연히 컴퓨터와 IT 기술입니다. 이러한 상황에서 개최되는 한국정보과학회 학술심포지움 유비쿼터스 컴퓨팅과 웹 정보기술 (Ubiquitous Computing and Web Information Technology, UCWIT)는 그 의미가 남다르다 하겠습니다. 한국정보과학회 영남지부 주관의 UCWIT 2023 행사는 산업체, 연구소 및 학계에서 활동하시는 지역 IT 전문가들을 위한 교류의 장입니다.

우리 지역에서 국제적 경쟁력을 갖춘 지역의 주력산업인 기계, 자동차, 조선, 에너지, 철강, 물류 등 전통산업은 많은 변화를 요구하고 있으며, 그 변화의 중심에는 제4차 산업혁명이라고 일컬어지는 ICT와의 융합이 차지하고 있습니다. 이번 학술심포지움을 통하여 지역의 산업이 더욱 업그레이드되어 경쟁력을 강화하는 기회가 될 수 있기를 기대합니다.

조직위원회에서는 제4차산업혁명의 핵심 이슈들을 중심으로 금년 행사를 준비하였습니다. 본 행사는 논문발표와 특별강연으로 구성됩니다. 논문발표는 총 3개의 세션으로 이루어집니다. 산학연 각계에서 제출된 논문은 메타버스/콘텐츠, 인공지능/딥러닝, IOT/클라우드/빅 데이터 및 학생들의 참신한 아이디어를 구현한 프로그램들에 대한 다양하고 흥미로운 주제를 다루고 있습니다.

본 행사가 활발한 교류의 장, 그리고 격의 없는 토론의 장이 될 수 있도록 여러분의 많은 참여를 바랍니다. 그리하여 본 행사가 지역 IT 분야의 학문적 발전과 관련 산업의 성장에 이바지할 것을 기대합니다.

행사 준비를 위해 수고하신 이원준 한국정보과학회 회장님을 비롯하여, 영남지부 운영위원님들, 대회 조직위원님들과 프로그램위원님들께 감사드립니다. 아울러 물심양면으로 지원해주신 MTDATA, 대신정보통신, ㈜우리아이티 및 행사 장소를 기꺼이 제공해 주신 부산대학교 차정인 총장님께 감사드리며 학문적 성과의 공유를 위해 기꺼이 논문을 투고해주신 저자님들께 진심으로 감사합니다.

조직위원장 **김 동 균** (경북대, 한국정보과학회 영남지부 지 부 장)

학술위원장 **유 영 환** (부산대), 집행위원장 **김 정 구, 김 종 덕** (부산대)

대회 조직

대회장	한국정보과학회장 이원준 (고려대학교)
조직위원장	정보과학회 영남지부장 김동균 (경북대학교)
학술위원장	유영환 (부산대학교)
학술위원	최윤호, 송길태, 박진선, 김원석, 김태운 (부산대학교)
집행위원장	김정구, 김종덕 (부산대학교)
고문	강순주 (경북대) 김석규 (안동대) 김순철 (대구대) 김영학 (금오공대) 변정용 (동국대) 송하주 (부경대) 우용태 (창원대) 정민수 (경남대) 정중수 (안동대)

행사 일정

11/24(금)

일시	행사 내용	비고
10:30 ~ 11:00 ('30)	학회 등록	국제회의장
11:00 ~ 12:00 ('60)	Poster 발표	국제회의장
12:00 ~ 13:30 ('90)	점심 식사	애슐리 퀸즈
13:30 ~ 14:00 ('30)	개회식 개회사 : 김동균 한국정보과학회 영남지부장 축사 : 진성호 한국과총 부산울산지역연합회장 환영사 : 백윤주 부산대학교 정보의생명공학대학장	국제회의장
14:10 ~ 15:40 ('90)	Oral 발표	
15:40 ~ 16:00 ('20)	휴식	
16:00 ~ 16:20 ('20)	Explainability of Deep Neural Networks with Comparative Neuron Activations and Gradients 경북대 - 남우정 교수님	
16:20 ~ 16:40 ('20)	Code and Performance Portability on Heterogeneous Computing 경북대 - 남덕윤 교수님	
16:40 ~ 17:00 ('20)	AI 시대 도래에 따른 차세대 컴퓨팅 하드웨어 전망 경북대 - 김정근 교수님	
17:00 ~ 17:20 ('20)	환경 변화에 강인한 멀티 모달 3D 시각 인지 기술 부산대 - 박진선 교수님	
17:20 ~ 17:40 ('20)	Autonomous intelligence driven by visual correspondence 부산대 - 전상률 교수님	
17:40 ~ 18:00 ('20)	추천 시스템의 개념과 딥러닝 기반 방법 부산대 - 조준수 교수님	
18:00 ~	폐회사	

11/25(토)

일시	행사 내용	비고
10:00 ~ 12:00 ('120)	한국정보과학회 영남지부 총회	국제회의장

Oral 발표 정보

일시	행사장		
	Room #1 (기계관 국제회의장)	Room #2 (기계관 101호)	Room #3 (기계관 102호)
14:10 ~ 15:40 ('90)	Metaverse, Contents & Network Technology	Recent Advances in Deep Learning	Application of Artificial Intelligence

Session #1 : Metaverse, Contents & Network Technology

좌장 : 김 종 덕, 유 영 환 (부산대학교)

번호	제목 저자 및 소속
1-1	Infeasible 상황에서도 견고한 IoT 네트워크를 위한 전력 할당 방식 응우옌 쉰안 등 (부산대학교)
1-2	6TiSCH 네트워크에서 패킷 큐 활용도 기반 향상된 'Adapting to Traffic' 기법 박정배 (부산대학교)
1-3	공장용 LED를 활용한 Li-Fi 기반 통신 시스템 구현 김준영 (포항공과대학교)
1-4	UAV를 이용한 Wi-Fi 음영 지역 탐지 및 해결 제안 시스템 김동주 (부산대학교)
1-5	크롬 확장기능을 사용한 브라우저 기반의 코드 작성기 이선호 (부경대학교)
1-6	지식그래프를 활용한 버티포트 리소스 관리 방안에 대한 연구 고동연, 김영인, 김유진, 김재렬, 송나흔 (경북대학교)

Session #2 : Recent Advances in Deep Learning

좌장 : 김 태 운 (부산대학교)

번호	제목 저자 및 소속
2-1	스마트 팩토리를 위한 PINN 김재우 (창원대학교)
2-2	LSTM과 XGBoost를 이용한 전력 예측에서의 ANN 모델 성능 비교 - 다중 변수 추출 시 데 이터분석의 중요성 김다영, 석혜원, 송득모, 정설영 (경북대학교)
2-3	MmWave 레이더 기반 일상 행동 인식을 위한 경량화 딥러닝 시스템의 설계 및 구현 신기훈, 임재봉, 김태구, 조용훈, 정혜선, 신채림 (부산대학교)
2-4	NR-V2X Mode2에서 강화 학습 기반 적응형 Modulation Order 선택 기법 조의리 (경북대학교)
2-5	SSD의 쓰기 증폭 개선을 위한 머신러닝 기반 Hot/Cold 페이지 분류 기법 윤성준 (부산대학교)
2-6	WaveTR: 분산 시스템을 위한 메시지 스트림 프레임워크 김민섭, 손자영, 이우령, 홍은솔 (경북대학교)

Session #3 : Applications of Artificial Intelligence

좌장 : 박 진 선 (부산대학교)

번호	제목 저자 및 소속
3-1	차량 주행 상황에서 촬영된 이미지의 분포 정보를 활용한 안개 분류 기법 임도현 (경북대학교)
3-2	더 나은 비전-언어 모델링을 위한 속성-어댑터 조용현 (부산대학교)
3-3	멀티모달 기반 IP 비디오월 제어 시스템 윤새봄, 최문석, 우성현, 이정민, 문서정, 정설영 (경북대학교)
3-4	Bert 기반 오픈 도메인 뉴스 이벤트 추출 방법 황해루이 (아주대학교)
3-5	NSGA-II를 활용한 인테리어 상품 최적 조합 추천 프레임워크 지다인, 김서윤, 김현민, 이정미, 장진영, 조유담 (서울여자대학교, (주)에이아이닷엠)
3-6	엣지컴퓨팅 기반의 영상분석기술을 활용한 응급상황 이벤트 기술개발 백정훈 (경북대학교)

포스터 세션

좌장 : 김 종 덕, 유 영 환, 김 태 운, 박 진 선 (부산대학교)

번호	제목 저자 및 소속
4-1	객체 위치 추정을 통한 효율적인 QR 코드 출입증 인식 오중현 (경북대학교)
4-2	거대 언어 모델의 효율적인 학습 보안을 위한 LangChain 활용 연구 곽건영, 손홍석, 심규민, 채정민, 정설영 (경북대학교)
4-3	기계학습을 이용한 IoT 기반 재난 시스템 Kubernetes Custom Autoscaler 설계 박준희 (경북대학교)
4-4	GPT-4 텍스트 증강 데이터 기반 소규모 챗봇에서 BERT와 Sentence-BERT의 성능 비교 은지수, 김민국, 김수민, 서상훈 (경북대학교)
4-5	대규모 언어 모델을 활용한 강의평가 질의응답 서비스 신진철 (경북대학교)
4-6	이상행동 감지 모델 시스템 김승준, 우승택, 박한솔, 김세아, 김민준, 정설영 (경북대학교)
4-7	음성인식 및 감정분석 기술 기반 감정일기 어플리케이션 설계 및 구현 김태환, 이휘찬, 오동휘, 한의진 (경북대학교)
4-8	SentenceBERT와 ChatGPT를 활용한 갱년기 상담 챗봇 서비스 고도화 이은주, 서동혁, 이채현, 최혜성, 오유나 (경북대학교)
4-9	수상레저기구 위치정보 제공을 위한 프로토콜 최연욱 (경북대학교)
4-10	SQL 최적화를 통한 요금계산시스템의 트랜잭션 시간 단축 최윤정 (경북대학교)
4-11	쿠버네티스 기반 분산형 교내 IoT 모니터링 시스템 이현, 김은지, 장서윤, 이창홍 (부산대학교)
4-12	경량화 플랫폼 Jetson TX2를 활용한 YOLOv5 및 DeepSORT 기반 버스 내부 모니터링 시스템 구현 김민욱, 황정호, 황현정, 김윤섭 (부산대학교)
4-13	패치 기반 생성형 모델을 적용한 메타버스 이미지 리타겟팅 연구 김찬호, 김은지, 김건아, 문채원 (경북대학교)
4-14	웹서비스에서 자바스크립트의 보안 강화를 위한 연구 동향 분석 이인규 (경북대학교)
4-15	인공지능 학습을 위한 데이터 증강 및 전처리 과정에 대한 연구 황효성, 김규회, 채준혁 (경북대학교, (주)퀀텀에이아이)
4-16	빅데이터 기반 온라인 이미지 편집 솔루션 박준서, 가파로브 자흔기르, 정지후, 이상준 (경북대학교)
4-17	딥러닝을 활용한 맞춤형 신발 소재별 관리 시스템 윤해진 (경북대학교, (주)스마트룸뱅크)
4-18	차량 내부 공기 질 실시간 분석에 따른 국부공기청정정화 시스템 효율 증대 배상우 (경북대학교)
4-19	웹 기반 악취방지시설 조합 예측 서비스 김다진, 박지예, 이아현, 정태훈 (경북대학교)
4-20	고회수율 RO 시스템 제어를 위한 인공지능 모델 설계 전민재, 이상현, 배진성, 배민석 (경북대학교)

4-21	재활의학적 통증 환자의 통증 기록관리 어플리케이션 설계 엄소민 (경북대학교)
4-22	음성인식 및 감정분석 기술 기반 감정일기 어플리케이션 설계 및 구현 김태환, 오동휘, 이휘찬, 한의진 (경북대학교)
4-23	쿠버네티스와 아파치 카프카를 이용한 마이크로서비스 아키텍처 구현에 대한 연구 이상희, 김현우, 신성현, 정현정, 조동필 (경북대학교)
4-24	데이터 증강 학습 이용한 딥러닝 기반 실시간 전동 킥보드 검지 시스템 김도현 (경북대학교)
4-25	연합 학습에서의 이질성 분류 및 그 영향에 관한 최신 연구 동향 맹키스투 테스파후넝 민우엘렛 (부산대학교)
4-26	생성형 AI를 활용하는 IT 시스템의 빠른 구현을 위한 레퍼런스 아키텍처 제안과 적용 가능성 연구 송혜경, 권다운, 제유나, 하재현 (경북대학교)
4-27	긴급신고 상황에 교환기에 종속된 CTI시스템에 대한 향후 확장방향에 대한 연구 추현우 (경북대학교)
4-28	사용자 위치 및 관심사 기반의 오프라인 번개 모임 앱 고도화 손석현 (경북대학교)
4-29	강철영상의 미세조직 분할을 위한 UNet3+ 개선 비샬 (금오공과대학)

Infeasible 상황에서도 견고한 IoT 네트워크를 위한 전력 할당 방식

응우옌 쉘안 동^o 황원주

부산대학교

tung.nguyenxuan@pusan.ac.kr, wjhwang@pusan.ac.kr

Power allocation scheme for robust IoT networks under infeasible circumstances

Nguyen Xuan Tung^o Won Joo Hwang

Pusan National University

요약

Considering the infeasible circumstances in Internet-of-Things (IoT) networks, where the system cannot satisfy all IoT devices (IDs) their required data throughput, namely infeasible circumstances. Our work presents a novel framework to handle this issue. We propose an iterative algorithm that gradually increases the number of satisfied IDs and enhances the total network throughput. Therein, the power allocation can be obtained via a designed fixed-point algorithm, which can serve more IDs (up to 50%) than applying an equal power allocation method.

1. Introduction

The Internet of Things (IoT) has revolutionized our interaction with the world with numerous applications [1]. However, managing IoT networks poses several challenges [2]. Although previous works have made significant improvements to system performance, limitations remain, particularly in satisfying the Quality of Service (QoS) to users. Optimizing the system solely by the power budget, [3], leads to disparities in service quality, while some IDs experience an unnecessarily high data throughput, others suffer from very poor service. When all users request minimum data throughput at the same time, the system may have no solution, namely infeasible circumstances.

Nevertheless, the study of this problem in IoT networks is still limited. Especially, in dense IoT networks [4] with intense interference. Therefore, it is crucial to identify and provide services to as many IoT devices as the system can satisfy. Being inspired, this work studies an IoT system where multiple APs communicate with IoT devices in the downlink way. Our objective is to handle the infeasible problems and maximize the number of IoT devices satisfying the required data throughput. We propose a dual-objective optimization problem that optimizes both the number of satisfied IoT devices and the total data throughput. Extensive simulations demonstrate a significant improvement in comparison to an equal power allocation method.

2. System Models and Problem Formulation

We consider a downlink IoT system with K APs and N IDs as in Fig.1. An AP simultaneously serves multiple IDs.

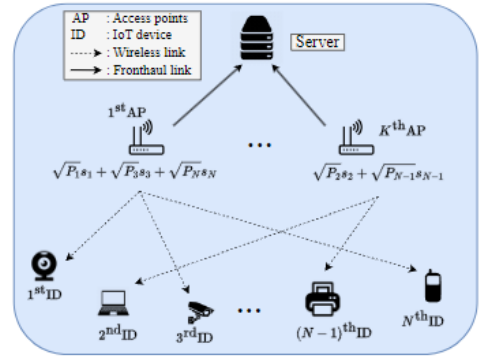


Fig.1: A downlink IoT network with multiple APs and IDs.

The n -th ID receives the desired signal s_n . The received signal at the n -th ID is $y_n = \mu_{k,n} h_{k,n} \sqrt{P_n} s_n + \sum_{n' \neq n, n' \in \mathcal{N}} \sum_{k \in \mathcal{K}} \mu_{k,n'} h_{k,n'} \sqrt{P_{n'}} s_{n'} + n_n$. Therein, $h_{k,n}$ is the channel between the n -th ID and the k -th AP, n_n is the additive white Gaussian noise (AWGN) at the n -th ID, which is distributed as $\mathcal{CN}(0, \sigma^2)$. Here, $\mu_{k,n}$ indicates the ID-AP association, which is determined by the strongest large-scale fading channel. The data throughput of the n -th ID is computed as

$$R_n(\mathbf{p}) = \log_2 \left(1 + \frac{|\mu_n^T h_n|^2 P_n}{\sum_{n' \neq n, n' \in \mathcal{N}} |\mu_{n'}^T h_n|^2 P_{n'} + \sigma^2} \right). \quad (1)$$

Where $\mu_n = [\mu_{1,n}, \dots, \mu_{K,n}]^T \in \mathbb{B}^K$, $\Psi = [\mu_1, \dots, \mu_N] \in \mathbb{B}^{K \times N}$, $\mathbf{h}_n = [h_{1,n}, \dots, h_{K,n}]$, and $\mathbf{p} = [P_1, \dots, P_N]^T \in \mathbb{R}^N$. To maximize the number of satisfied IDs and the total data throughput, we consider the following dual-objective problem:

$$\begin{aligned} & \underset{\mathbf{p}}{\text{maximize}} && (|\mathcal{Q}|, R_{tot}(\mathbf{p})) \\ & \text{Subject to} && \sum_{n \in \mathcal{N}} \mu_{k,n} P_n \leq P^{max}, \forall k \in \mathcal{K} \quad (2) \\ & && R_n(\mathbf{p}) \geq R_n^{thr}, \forall n \in \mathcal{Q}(\mathbf{p}) \end{aligned}$$

where $\mathcal{Q}(\mathbf{p}) = \{n | n \in \mathcal{N}, R_n(\mathbf{p}) \geq R_n^{thr}\}$.

3. Proposed Solution

Note that if the system optimizes the sum of data throughput without meeting all the IDs' QoS requirements, the power budget would be allocated to the IDs with good channel conditions only. To share more power with poor channel conditions IDs, we set the data throughput requirement of satisfied IDs equal to their requested levels. Thus, less power budget is required from these IDs to satisfy the request and the remaining power budget will be shared with other IDs. We repeat two main steps: (1) updating the satisfied IDs by reallocating the power budget; (2) solving the following data throughput maximization problem with fixed data throughput constraints as

$$\begin{aligned} & \underset{\mathbf{p}}{\text{maximize}} && R_{tot}(\mathbf{p}) \\ & \text{Subject to} && \sum_{n \in \mathcal{N}} \mu_{k,n} P_n \leq P^{max}, \forall k \in \mathcal{K} \quad (3) \\ & && R_n(\mathbf{p}) = R_n^{thr}, \forall n \in \mathcal{Q}^*(\mathbf{p}^{*,(t-1)}) \end{aligned}$$

where t is the iteration index. The above problem can be solved by applying the log approximation transformation $\log_2(1+z) \geq \alpha(z') \log_2(z) + \beta(z')$ with non-negative z and z' , and $\alpha(z') = z'/(1+z')$ and $\beta(z') = \log_2(1+z') - z'/(1+z') \log_2(z')$. The partial Lagrangian function on the power allocation of unsatisfied IDs and adjusting the power of satisfied IDs accordingly. We repeat solving (3) and update the satisfied set as in Algorithm 1 until convergence.

4. Performance Evaluation

In this section, the performance of our proposed algorithm is evaluated. For better illustration, we evaluate the system performance with different AP association methods including the closest AP association and the coalition game method (CG-APA). Meanwhile, the proposed power allocation (DIF-PA) is compared to the equal power allocation. Fig.2 illustrates the detailed data throughput of IDs when the system deploys different power allocation solutions and AP association methods. The network has 8 IDs with the minimum data throughput requirement of 0.5 (bits/s/Hz). When the system applies Equal-PA and the proposed CG-APA, only 4 IDs can meet their requirements.

The figure also shows that the proposed power allocation can provide higher total data throughput in comparison to the equal method.

Algorithm 1: Alternative algorithm to solve problem (2).

- 1: **Input:** The system parameters N, K, σ, P^{max} , and a tolerant ϵ .
- 2: **Initialize:** $\mathbf{p}^{*,(0)}$ and calculate $\mathcal{Q}^*(\mathbf{p}^{*,(0)})$.
- 3: **Repeat**
- 4: $t = t + 1$.
- 5: Solve (3) to obtain $\mathbf{p}^{*,(t)}$.
- 6: Update $\mathcal{Q}^*(\mathbf{p}^{*,(t)}) = \{n | R_n(\mathbf{p}^{*,(t)}) \geq R_n^{thr}, \forall n\}$.
- 7: **Until** $(|\mathcal{Q}^{*,(t)}| == |\mathcal{Q}^{*,(t-1)}|) \& (R_{tot}(\mathbf{p}^{*,(t)}) - R_{tot}(\mathbf{p}^{*,(t-1)}) \leq \epsilon)$.
- 8: **Output:** $\mathbf{p}^{*,fin}$.

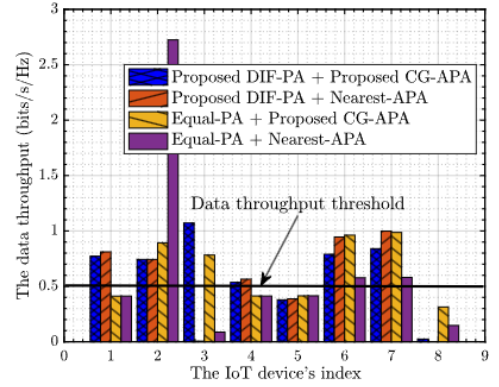


Fig. 2: The data throughput of IDs with different solutions.

4. Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00260098) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

References

- [1] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," IEEE Internet of Things J., vol. 4, no. 5, pp. 1125–1142, Mar. 2017.
- [2] C. A. Ouedraogo, S. Medjiah, C. Chassot, K. Drira, and J. Aguilar, "Adaptive performance analysis in IoT platforms," IEEE Trans. Netw. Service Manage., vol. 19, no. 4, pp. 4764–4778, Jul. 2022.
- [3] T. Park, G. Lee, W. Saad, and M. Bennis, "Sum rate and reliability analysis for Power-Domain Nonorthogonal Multiple Access (PD-NOMA)," IEEE Internet of Things J., vol. 8, no. 12, pp. 10 160–10 169, Jan. 2021.
- [4] X. Liu, Z. Qin, Y. Gao, and J. A. McCann, "Resource allocation in wireless powered IoT networks," IEEE Internet of Things J., vol. 6, no. 3, pp. 4935–4945, Jan. 2019.

6TiSCH 네트워크에서 패킷 큐 활용도 기반 향상된 'Adapting to Traffic' 기법

박정배^o 정상화

부산대학교 정보융합공학과

wdqo94@gmail.com, shchung@pusan.ac.kr

Enhanced 'Adapting to Traffic' Technique Based on Packet Queue Utilization in 6TiSCH Networks

Jeongbae Park^o Sang-Hwa Chung

Department of Information Convergence Engineering, Pusan National University

요 약

IPv6 over the Time-Slotted Channel Hopping(6TiSCH)는 산업 환경에서 안정적이고 신뢰성 있는 무선 통신을 제공하는 프로토콜 스택이다. 6TiSCH의 셀 스케줄링 기법인 Minimal Scheduling Function(MSF)의 동적 셀 할당 메커니즘인 'Adapting to Traffic'은 패킷 큐 활용도와 무관하게 작동하므로, 트래픽 변화로 인해 발생하는 큐 오버플로우로 패킷 손실이 발생할 수 있다. 본 논문에서는 패킷 큐 활용도를 기반으로 셀 추가 및 삭제 판단 주기를 동적으로 조정하여 패킷 큐 오버플로우를 줄이고 유휴 청취를 감소시킨다. 6TiSCH 시뮬레이터를 활용한 실험을 통해 기존 기법에 대비 패킷 전송률, 처리량, 네트워크 에너지 소모량에서 성능 향상을 확인하였다.

1. 서 론

현대 산업 환경에서 무선 네트워크의 중요성이 급증하고 있다. 무선 네트워크를 통해 설비와 자원을 실시간으로 모니터링하고 제어함으로써 공정 최적화, 에너지 효율 향상, 장비 상태 예측, 유지보수 등이 가능하다.

그러나 산업 환경에서의 무선 통신은 철제 구조물로 인한 다중 경로 페이딩 및 무선 신호 간섭으로 인해 통신 신뢰성이 현저하게 감소한다. 특히 산업 분야에서는 실시간 데이터 수집 및 제어 시스템의 안전성이 중요하므로 안정적이고 예측 가능한 무선 통신이 필수적이다.

IPv6 over the Time-Slotted Channel Hopping(6TiSCH)[1]는 산업 환경에서 안정적이고 신뢰성 있는 무선 통신을 제공하기 위한 프로토콜 스택 중 하나다. 이 프로토콜 스택은 IEEE 802.15.4e 표준에 기반하며, 산업 환경에서도 효과적으로 동작한다. 6TiSCH의 셀 스케줄링 기법인 Minimal Scheduling Function(MSF)[2]의 'Adapting to Traffic'은 트래픽 양에 따라 셀을 동적으로 추가 또는 삭제하는 기법이다. 그러나 이 기법은 셀 활용도만을 고려하기 때문에 트래픽 변동이 빈번한 네트워크에서 패킷 큐 오버플로우로 인한 패킷 손실이 발생할 수 있다. 따라서 본 논문에서는 패킷 큐 활용도를 고려한 개선된 'Adapting to Traffic' 기법을 소개한다.

2. 배경지식

TSCH는 6TiSCH의 중요한 메커니즘 중 하나로, 무선 센서 네트워크에서 안정적이고 신뢰성 있는 통신을 지원하기 위한 Media Access Control(MAC) 기술이다. TSCH는 채널 호핑과 시분할 통신을 결합하여 외부 간섭을 최소화하고, Absolute Slot Number(ASN) 기준으로 네트워크에 참여한 모든 장치에 시각 동기화된다. 각 장치 쌍은 특정한 타임 슬롯과 주파수 채널을 사용하여 통신하며, 이를 통해 충돌을 최소화하고 예측 가능하며 신뢰성 있는 통신이 가능하다. 일정 개수의 타임 슬롯이 모여 슬롯프레임을 구성하고, 슬롯프레임이 무한히 반복되는 구조를 가지고 있다.

특정한 타임 슬롯과 주파수 채널 조합은 셀이며, 셀 할당 방법을 셀 스케줄링이라고 한다. MSF는 6TiSCH에서 사용되는 셀 스케줄링 기법이다. MSF는 여러 종류의 셀을 사용하여 네트워크 노드 간 통신을 관리한다. 셀에는 Minimal cell, Autonomous cells, 그리고 Negotiated cells가 있다. 그림 1은 6TiSCH 네트워크에 참여한 장치의 슬롯프레임을 보여주며, MSF를 사용하여 할당된 셀을 확인할 수 있다.

6TiSCH Operational sublayer(6top)[3]는 6TiSCH의 링크 레이어 상단에 위치하며, 셀 할당, 삭제, 재배치 등을 위해 6top Protocol(6P)을 사용한다.

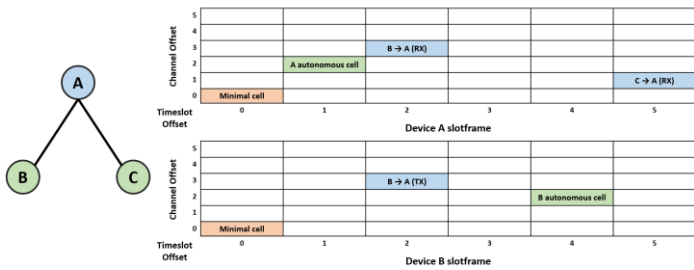


그림 1 MSF을 활용해 셀이 할당된 슬롯프레임

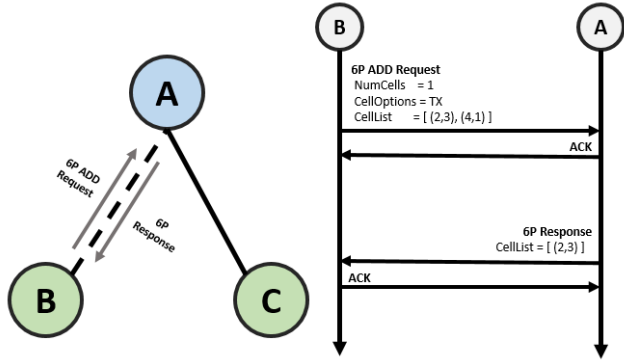


그림 2 6P 트랜잭션

6P 트랜잭션을 통해 셀을 추가하려는 노드는 상대 노드에게 후보 셀 리스트와 추가할 셀의 수를 보낸다. 상대 노드는 이 정보를 기반으로 가능한 셀을 선택하고 응답한다. 그림 2는 통신 노드 A와 B의 6P 트랜잭션의 예시를 보여준다.

'Adapting to Traffic' 기법은 MSF에서 제공하는 기능으로 트래픽 양에 따라 셀을 동적으로 추가 또는 삭제하는 기법이다. 부모 노드와 Negotiated cell에서 NumCellsElapsed의 값이 1씩 증가하며, 해당 셀에서 패킷을 송신 또는 수신 했을 경우 NumCellsUsed의 값이 증가한다. NumCellsElapsed의 값이 MAX_NUMCELLS에 도달했을 경우 NumCellsUsed가 LIM_NUMCELLSUSED_HIGH보다 크다면 6P 트랜잭션을 통해 셀을 추가하고 LIM_NUMCELLSUSED_LOW 보다 낮을 경우 셀 삭제를 위한 6P 트랜잭션을 시작한다.

3. 관련 논문

트래픽 변동에 취약한 MSF의 단점을 보완하기 위해 몇 가지 개선된 기술이 제안되었다. A-MSF[4]는 노드가 셀 활용률을 50% 달성하기 위해 여러 개의 셀을 동적으로 추가하거나 제거한다. IMSF[5]는 큐 활용도를 고려하여 미래의 트래픽 수요에 맞게 셀을 추가한다. EMSF[6]는 MSF의 임계치 기반 메커니즘 대신 푸아송 알고리즘을 활용하여 예측 접근 방식으로 셀을 추가하거나 삭제한다.

트래픽 수요에 따라 적절한 개수의 셀을 한번의 6P 교섭으로 할당할 경우, 6P 트랜잭션에서 발생하는 오버헤드를 줄일 수 있다는 장점이 있다. 하지만 MAX_NUMCELLS의 개수는 고정되어 있기 때문에 6P 교섭 전까지 패킷 큐 오버플로우가 발생할 수 있다. 특히, 슬롯프레임의 길이가

길고 슬롯 프레임에 할당된 셀이 적다면 셀을 추가하는데 많은 시간이 걸릴 수 있다.

4. 제안 기법

기존의 방식에서는 MAX_NUMCELLS의 크기가 고정되어 있어 패킷 큐의 활용도와는 무관하게 셀을 추가하거나 삭제하는 판단 주기가 결정되었다. 그러나 제안 기법에서는 패킷 큐의 활용도에 따라 셀 추가 또는 삭제의 판단 주기를 동적으로 부여한다. 패킷 큐의 활용도가 높을수록 셀 추가 판단 주기를 줄이며, 셀 추가 시 패킷 큐 오버플로우를 방지할 수 있다. 패킷 큐의 활용도가 낮을수록 셀 삭제 판단 주기를 줄여, 셀 삭제를 통한 유휴 청취 감소로 통신 장치의 에너지 소모를 줄일 수 있다. 패킷 큐 활용도 기반 'Adapting to Traffic' 기법은 각 Negotiated cell에서 다음과 같은 동작을 수행한다.

1. NumCellsElapsedAdd, NumCellsElapsedDel을 1씩 증가시킨다.
2. 해당 셀에서 패킷을 전송했다면, NumCellsUsedAdd와 NumCellsUsedDel을 1씩 증가시킨다.
3. NumCellsElapsedAdd가 MAX_NUMCELLS_ADD보다 크거나 같고, 셀 활용도가 LIM_CELLUTILIZATION_HIGH 이상이라면, 셀 추가를 위한 6P 트랜잭션이 시작되며, 6P 트랜잭션을 하지 않았더라도 매번 NumCellsElapsedAdd와 NumCellsUsedAdd를 초기화한다.
4. NumCellsElapsedDel이 MAX_NUMCELLS_DEL보다 크거나 같고, 셀 활용도가 LIM_CELLUTILIZATION_LOW 이하라면, 셀 삭제를 위한 6P 트랜잭션이 시작되며, 6P 트랜잭션을 하지 않았더라도 매번 NumCellsElapsedDel과 NumCellsUsedDel을 초기화한다.

MAX_NUMCELLS_ADD는 패킷 큐 활용도가 높을수록 작아져야 하므로 수식 (1)과 같다. 수식 (1)의 α 와 β 는 네트워크 환경에 맞게 조절할 수 있다. 예를 들어 버스트 트래픽이 자주 발생하고 패킷 유실이 치명적인 서비스의 경우 패킷 큐에 따른 영향을 많이 줄 수 있도록 높은 α 를 설정할 수 있다. MAX_NUMCELLS_DEL은 패킷 큐 활용도가 높을수록 커져야 하므로 수식 (2)와 같으며, 기존 기법에서는 MAX_NUMCELLS이 일정하게 유지되기 때문에 셀 추가 및 삭제 기준을 셀의 개수로 나타낸 LIM_NUMCELLSUSED_HIGH와 LIM_NUMCELLSUSED_LOW로 설정되었다. 그러나, 제안된 기법에서는 MAX_NUMCELLS_ADD와 MAX_NUMCELLS_DEL이 패킷 큐 활용도에 따라 동적으로 조절되므로, 셀 추가 및 삭제를 판단하기 위한 기준으로 LIM_CELLUTILIZATION_HIGH와 LIM_CELLUTILIZATION_LOW를 수식 (3)과 수식 (4)와 같이 정의한다.

$$\begin{aligned} \text{MAX_NUMCELLS_ADD} &= \text{MAX_NUMCELLS} \\ &\times \left(1 - \left(\alpha \times \frac{\text{The number of packet}}{\text{Packet queue length}} + \beta \right) \right) \end{aligned} \quad (1)$$

$$\begin{aligned} \text{MAX_NUMCELLS_DEL} &= \text{MAX_NUMCELLS} \\ &\times \left(\gamma \times \frac{\text{The number of packet}}{\text{Packet queue length}} + \delta \right) \end{aligned} \quad (2)$$

$$\text{LIM_CELLUTILIZATION_LOW} = \frac{\text{LIM_NUMCELLSUSED_LOW}}{\text{MAX_NUMCELLS}} \quad (3)$$

$$\text{LIM_CELLUTILIZATION_HIGH} = \frac{\text{LIM_NUMCELLSUSED_HIGH}}{\text{MAX_NUMCELLS}} \quad (4)$$

5. 성능 평가

시뮬레이션은 6TiSCH 시뮬레이터[7]를 활용했으며 실험은 총 1200 슬롯프레임동안 진행되었다. 한 쌍의 부모와 자식 노드를 생성 후 버스트 트래픽이 발생하는 구간을 제외하고 자식 노드는 10초에 패킷 1개를 생성한다. 400 슬롯프레임에서 700 슬롯프레임까지 버스트 트래픽을 발생시켜 기존의 6TiSCH 네트워크와 제안된 기법이 적용된 6TiSCH 네트워크의 성능을 확인하였다. 버스트 트래픽 구간에서 패킷 생성 주기는 1초, 0.8초, 0.6초, 0.4초, 0.2초이다. 각 패킷 생성 주기마다 실험이 시행됐으며 모든 실험 결과는 5회 실행된 평균이다. 실험에서 패킷 큐 사이즈는 10으로 설정했으며, 수식 (1)의 α 는 9, β 는 0으로 수식 (2)의 γ 와 δ 는 0.5로 설정하였다.

그림 3과 그림 4은 시간에 따른 각 슬롯 프레임의 패킷 큐 활용도를 나타내며, 버스트 트래픽 구간 동안 0.2초 간격으로 패킷이 생성되었다. 기존 6TiSCH 네트워크에서는 셀 추가가 제안 기법에 비해 상대적으로 더딘 것을 확인할 수 있다. 반면에 제안된 기법을 적용한 6TiSCH 네트워크에서는 패킷 큐 활용률이 증가함에 따라 네트워크가 신속하게 셀을 추가하여 패킷 큐의 활용률을 낮추는 결과를 확인할 수 있다.

5.1 패킷 전송률

그림 5은 버스트 트래픽 구간에서 패킷 생성 주기에 따른 기존 기법과 제안된 기법의 패킷 전송률을 나타낸다. 패킷 생성 주기가 더 짧아질수록, 두 기법 모두 패킷 큐 오버플로우로 인해 패킷 전송률이 낮아지는 것을 확인할 수 있다. 제안된 기법의 경우 패킷 큐 오버플로우로 인해 전송률이 100%에서 90.78%로 감소하였으며, 반면에 기존 기법의 경우 100%에서 60.01%로 패킷 전송률이 낮아진 것을 확인할 수 있다.

5.2 처리량

그림 6는 버스트 트래픽 구간에서 패킷 생성 주기에 따른 기존 기법과 제안 기법의 처리량을 나타낸다. 패킷 전송 주기가 1초당 한 개의 패킷인 경우, 기존 기법은 374개의 패킷을 처리하며, 제안 기법은 372개의 패킷을 처리한다. 그러나 패킷 전송 주기가 0.2초인 경우, 기존 기법은 947개의 패킷을 처리하며, 제안 기법은 1435개의 패킷을 처리하므로 처리량의 차이가 발생하는 것을 확인할 수 있다.

5.3 네트워크 에너지 소모량

그림 7는 전체 네트워크의 에너지 소모량을 표시하며, 단위는 mA이다. 버스트 트래픽의 패킷 전송 주기가 0.2초에서는 제안된 기법의 에너지 소모가 더 많은 결과가 나타났다. 이는 기존 기법에서 많은 패킷 손실이 발생했기 때문이다. 패킷 손실이 발생할 때, 전송을 하지 않기 때문에 실험 전체의 패킷 전송 횟수가 줄어든다.

각 패킷 전송 시 송신자는 패킷을 전송하고 ACK 수신에 에너지를 소모하며, 수신자는 패킷을 수신하고 ACK를 전송하는 과정에서 에너지를 소모한다. 그러나 패킷을 전송하지 않았을 때, 수신자는 패킷을 기다리는 유휴 청취에만 에너지를 사용한다. 따라서 많은 패킷 손실이 발생한 0.2초에서는 제안된 기법의 에너지 소모가 더 많다. 전송 주기가 1초에서 0.4초까지는 기존 기법의 에너지 소모량이 더 많다. 이는 제안 기법에서는 패킷 큐의 활용도가 낮을 때 셀 삭제 판단 주기가 짧아 불필요한 셀을 삭제하여 유휴 상태의 청취를 줄였기 때문이다.

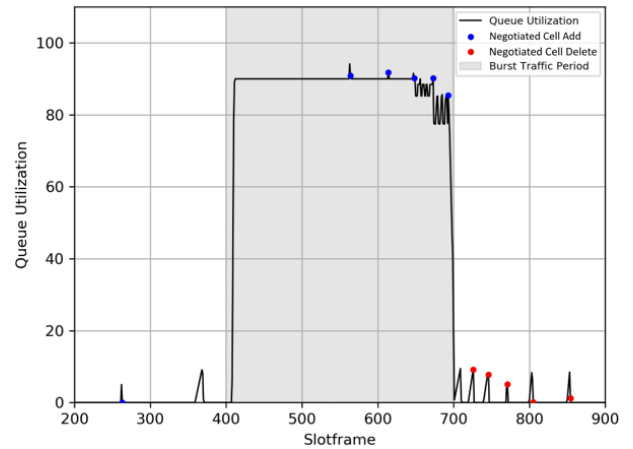


그림 3 시간에 따른 패킷 큐 활용도 (기존 기법)

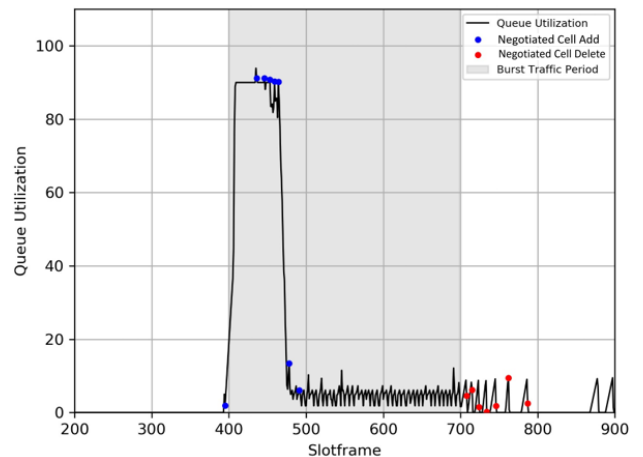


그림 4 시간에 따른 패킷 큐 활용도 (제안 기법)

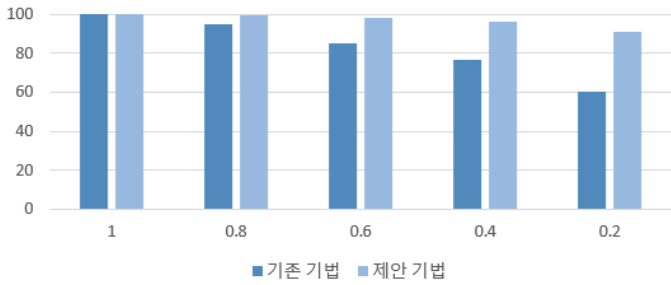


그림 5 패킷 전송률

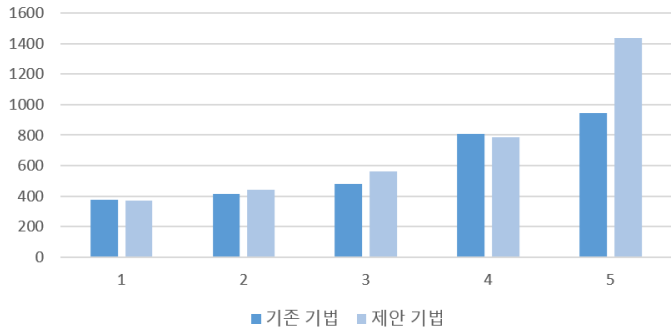


그림 6 처리량

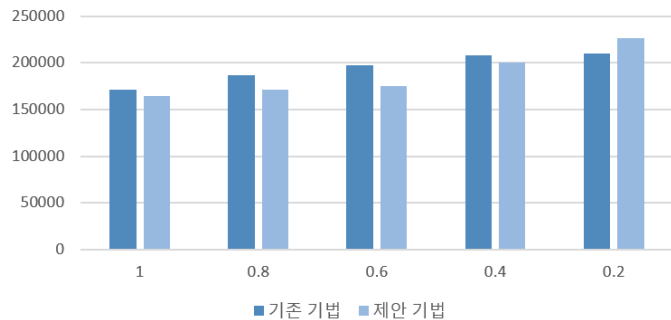


그림 7 네트워크 에너지 소모량

6. 결론 및 향후 연구

본 논문에서는 패킷 큐 활용도 기반 'Adapting to Traffic' 기법을 제안했으며, 버스트 트래픽 구간의 패킷 생성 주기를 변경하며 기존 기법과 비교했다. 패킷 전송률은 최대 30.77%의 차이를 나타내며, 처리량 역시 488개의 차이를 보였다. 이를 통해 제안된 기법을 적용함으로써 MSF의 성능이 향상되었음을 확인할 수 있다. 네트워크 에너지 소모량의 경우 패킷 생성 주기가 1초에서 0.4초까지는 기존 기법이 더 많은 에너지를 소모했으나, 0.2초에서는 처리량 차이로 인해 제안 기법이 더 많은 에너지를 소모하는 것을 확인할 수 있었다.

향후 연구에서는 셀 추가 판단 주기의 단축으로 인해 발생하는 6P 트랜잭션의 오버헤드를 확인하고, 트래픽 대비 과도한 셀 할당 여부를 확인할 것이다. 뿐만 아니라 셀 삭제 판단 주기 감소로 인한 네트워크 에너지 소모량 변화량에 대해 면밀히 확인할 예정이다. 이후 트래픽 변화에 대응이 어려운 매우 긴 슬롯프레임의 가진 6TiSCH 네트워크 적용해 A-MSF, IMSF 또는 EMSF와 같이 개선된 MSF 기법과 제안 기법의 성능 비교 할 것이다.

7. 사사

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2023-RS-2023-00260098)

참고 문헌

- [1] P. Thubert, "An architecture for IPv6 over the time-slotted channel hopping mode of IEEE 802.15.4 (6TiSCH)," in RFC 9030, 2021.
- [2] T. Chang, M. Vucinic, X. Vilajosana, S. Duquennoy, and D. Dujovne, "6TiSCH minimal scheduling function (MSF)," RFC 9033, DOI 10.17487/RFC9033., 2021.
- [3] Q. Wang, X. Vilajosana and T. Watteyne, "6top Protocol (6P)", Internet Engineering Task Force Std., RFC8480, August 2018
- [4] T. Chang, M. Vučinić, X. V. Guillén, D. Dujovne, and T. Watteyne, "6TiSCH minimal scheduling function: Performance evaluation," Internet Technology Letters, vol. 3, no. 4, p. e170, 2020.
- [5] K. N. Tapadar, M. Khatua, and V. Tamarapalli, "IMSF: Improved minimal scheduling function for link scheduling in 6TiSCH networks," in 23rd International Conference on Distributed Computing and Networking, pp. 124–127, 2022.
- [6] T. Hamza and G. Kaddoum, "Enhanced minimal scheduling function for IEEE 802.15.4e TSCH networks," in 2019 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6, IEEE, 2019.
- [7] E. Municio, G. Daneels, M. Vucinic, S. Latre, J. Famaey, Y. Tanaka, K. Brun, K. Muraoka, X. Vilajosana and T. Watteyne, "Simulating 6TiSCH Networks", Wiley Transactions on Emerging Telecommunications (ETT), vol. 30, no. 3. 2019

공장용 LED를 활용한 Li-Fi 기반 통신 시스템 구현

김준영*, 김경준

포항공과대학교

kjy2254@postech.ac.kr

Implementation of a Li-Fi based communication system using factory LED

Jun-yeong Kim*, Kyungjun Kim

Pohang Univ. of Science and Tech.(POSTECH)

요약

지능형 IoT(Artificial Intelligence IoT, AIoT) 기술이 발전함에 따라 스마트 제조업, 생활 및 상업용 서비스 등 다양한 환경에서 지능형 로봇의 활용이 늘어나고 있다. 이러한 지능형 로봇들의 제어와 이들이 생산해 내는 대용량의 데이터는 기존 네트워크 환경에서는 수용하기 힘든 저지연, 초고속의 환경을 필요로 한다. LiFi(Light Fidelity)는 빛을 이용한 통신 방법으로 LED의 빠른 깜빡임을 이용해 데이터를 전송함으로써, 주파수의 혼선이 없고, 최대 10Gbps의 빠른 속도를 구현할 수 있는 장점이 있어 차세대 통신으로 주목받고 있다. 하지만, 송수신기 간의 장애물이 있거나 먼 거리에서는 빛의 변화 감지에 어려움이 있어 실제 현장에 적용하기에는 어려움이 있다. 본 논문에서는 이러한 문제 해결에 도움이 되고자 공장용 LED와 같이 고전력 광원을 사용한 먼 거리에서의 통신 시스템 구조를 제안하고, 이를 바탕으로 구현된 예시를 제공한다.

1. 서론

정보통신기술이 빠른 속도로 발전함에 따라 사물들은 서로 능동적으로 데이터를 주고받게 되었고, 그때 생성된 많은 양의 데이터를 활용해 의사결정을 하는 AIoT 기술이 주목받고 있다. 클라우드 기반의 AIoT 기술은 높은 컴퓨팅 환경을 제공하지만, 디바이스가 기하급수적으로 늘어남에 따라 클라우드의 부하 증가 및 네트워크 지연으로 인한 병목 현상으로 인해 시스템 전체 성능에 악영향을 미칠 수 있다[1].

이러한 문제를 해결하기 위한 방법으로 가시광통신(Visible Light Communications, VLC) 기술이 있으며, VLC는 기존의 무선 주파수와 달리 가시광을 사용해 통신을 한다. 가시광의 주파수 영역은 380THz~750THz로 전체 무선통신 주파수보다 1만 배 이상 넓은 영역을 가지고 있어[2], 포화된 네트워크 환경의 보완으로 많은 연구가 이뤄지고 있다. 하지만, VLC는 빛을 사용하기 때문에 송수신기 간에 장애물이 있거나 먼 거리의 경우 통신에 제약이 생길 수밖에 없다[3]. 따라서 스마트 팩토리과 같

은 넓고 장애물이 존재하는 현장에서 이를 적용하기에는 어려움이 있다. 이러한 상황에 도움이 되고자, 본 논문에서는 저출력 LED를 사용해 간단한 VLC 통신을 구현하고, 이를 공장용 고출력 LED에 적용해 거리별 성능 평가를 실시한다.

본 논문의 구성은 다음과 같다. 2장에서는 기본적인 VLC 통신에 대한 설명과 구현 및 고출력 LED로의 적용에 대해 설명하고, 3장에서는 구현된 시스템의 성능을 평가한다. 마지막으로 4장에서 결론 및 향후 계획에 대해 살펴본다.

2. Li-Fi 기반 통신 시스템

2.1 VLC 통신

컴퓨터에 존재하는 데이터는 모두 0과 1로 이진화 표현이 가능하다. VLC 통신의 송신부는 보내고자 하는 데이터를 0과 1로 이진화하여 이를 정의된 방법으로 LED를 제어한다. 본 논문에서는 0은 Off, 1은 On으로 표현하

는 On-Off Keying(OOK) 방식을 사용하였다. 수신부는 송신부가 제어한 LED의 광량을 조도 센서를 통해 확인하고, 임계값을 기준으로 0과 1로 데이터를 구분한다. 이러한 형태의 송수신부의 구현을 위해 아두이노를 사용해 LED와 조도 센서를 제어하였다. 구성요소의 자세한 사항은 표 1과 같으며, 전체 통신 흐름은 그림 1과 같다.

표 1. VLC 통신 테스트 구성요소

Component	Description	Role
Arduino	Arduino Uno	Microcontroller
LED	Neopixel Jewel	Transmitter
Photo Diode	BPW21R	Receiver

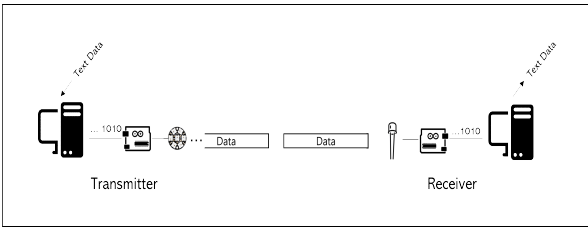


그림 1. 전체 통신 흐름

송신부의 아두이노는 시리얼 통신을 통해 컴퓨터로부터 텍스트 데이터를 입력받고, 해당 문자열을 이진화한다. 이진화된 문자는 최상위 비트(Most Significant Bit, MSB)부터 0은 Off, 1은 On으로 인코딩되어 LED를 일정 간격으로 제어해 데이터를 전송한다. 수신부는 일정 간격으로 조도 센서의 값을 읽으며, 임계값을 기준으로 0과 1로 디코딩한다. 디코딩된 문자는 시리얼 통신을 통해 수신부의 컴퓨터로 전송되어 송신부에서 작성한 텍스트를 확인할 수 있다.

두 기기 간 데이터 통신에는 동기화가 필요하다. 본 논문에서는 원활한 테스트를 위해 그림 2와 같이 비동기식 전송 방식을 사용하였다.

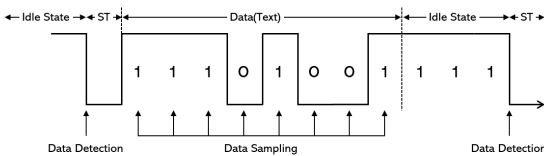


그림 2. 비동기 전송 방식

일반적으로 On(1) 상태에 있는 LED를 사용해 데이터를 보내기 때문에 Start bit(0)를 사용해 데이터의 전송을

알리고, 수신 측은 이를 감지 후 데이터를 일정 간격으로 읽음으로써 동기화를 맞추었다.

2.2 고휘력 LED를 사용한 데이터 전송

위 테스트에서 전송에 사용한 LED는 아두이노의 5V 전력을 사용한다. 따라서 LED가 낼 수 있는 출력이 제한되어 있어 송수신기 간 거리가 일정 이상 떨어질 시 조도 센서에 도달하는 빛의 양이 변화를 감지할 수 없을 수준으로 떨어진다. 이를 해결하기 위해 220V를 사용하는 고휘력 LED를 전원으로 사용하고, 이를 MCU(아두이노)를 통해 제어할 필요가 있다.

높은 전압의 기기를 제어하기 위해서는 릴레이를 사용할 필요가 있다. 릴레이는 유접점 릴레이와 무접점 릴레이가 존재하며, 본 논문에서는 응답시간이 빠르고 노이즈가 없는 무접점 반도체 릴레이를 사용하였다. 상세 연결 회로는 그림 3과 같다.

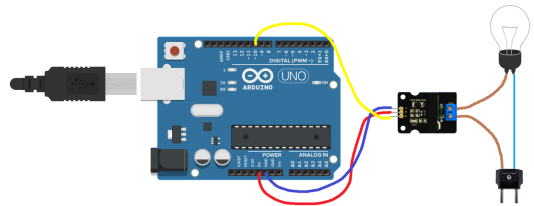


그림 3. 릴레이 연결 회로도

3. 성능 평가

그림 3의 회로를 사용해 2.1에서 진행한 통신을 적용해 거리별로 테스트를 진행하였다. 테스트를 진행한 환경은 낮 시간대의 외부 햇빛과 가정용 LED 전등이 존재하는 방안에서 진행하였으며, 테스트를 진행한 결과는 표 2와 그림 4와 같다.

표 2. 테스트 시 측정된 거리별 전압

Distance	Voltage		Average Voltage	
	Min	Max	On	Off
1M	977	1184	1175	1013
3M	909	1078	1068	932
5M	916	1021	996	945
10M	961	1036	1011	980

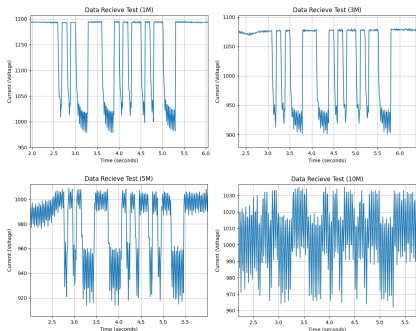


그림 4. 통신 테스트 시 측정된 전압 그래프

측정 결과, 동일한 신호를 전송하였을 때 거리가 멀어짐에 따라 On-Off 간의 절대적인 전위차는 줄어들었지만, 두 신호의 구분이 가능한 영역이 존재하는 것을 확인하였다. 따라서 이를 분석해 두 신호 간의 차이를 감지해 아날로그값을 디지털화하여 데이터를 복구하는 것이 가능하다.

4. 결론 및 향후 계획

본 논문에서는 AIoT 기술에 적용가능한 차세대 통신 방법으로 주목받고 있는 가시광통신과, 해당 기술에서 발생할 수 있는 통신 범위에 대한 문제를 해결하기 위해 고출력 LED를 적용한 방안에 대해 살펴보았다. 향후 제안한 내용을 바탕으로 외부 광원이 존재하는 환경에서 신호 구분이 가능한 가시광통신 알고리즘을 개발하고, 실제 스마트팩토리 등 현장에 적용가능한 형태로의 연구개발을 진행할 계획이다.

감사의 글

본 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2022R111A1A01066264, 차세대 AIoT 디바이스 연결성 보장을 위한 6G LiFi 네트워킹 기술 연구).

References

[1] M. Leba, S. Riurean, A. Ionica, “LiFi-the Path to a New Way of Communication.” *IEEE 12th Iberian*

Conf. on Information Systems and Technologies, Lisbon, Portugal, pp. 21-24, June 2017

- [2] 권근호, “A Study on the Li-Fi technological application prospect” *2016년도 한국통신학회 추계 종합학술발표회 논문집*, pp. 490-491 Nov. 2016
- [3] T. Komine and M. Nakagawa, “Fundamental analysis for visible-light communication system using LED lights”, *IEEE Transactions on Consumer Electronics*, Vol. 50, No. 1, pp. 100-107, 2004

UAV를 이용한 Wi-Fi 음영 지역 탐지 및

해결 제안 시스템

김 동 주^o, 김 동 혜, 박 동 한, 유 영 환

부산대학교 정보컴퓨터공학부

rlaehdwn9097@pusan.ac.kr, swhyny@pusan.ac.kr, vkfkdor00@pusan.ac.kr, ymomo@pusan.ac.kr

Detecting Wi-Fi Coverage Holes and

Solution Suggestion System Using UAV

Dong-hye Kim, Dong-han Park, Dong-ju Kim^o, Young-hwan Yoo
School of Computer Science and Engineering, Pusan National University

요 약

5G 무선 네트워크는 빠른 속도와 높은 전송량을 제공하기 위해 높은 주파수 대역을 사용하지만, 이러한 전파적 특성 때문에 음영지역(Coverage Holes)이 발생한다. 다가오는 6G 무선 네트워크는 테라헤르츠(THz) 대역의 주파수를 이용하며 이에 따라 5G보다 더 많은 음영지역이 생길 것으로 예측된다. 본 논문은 6G 서비스 시나리오 중 하나인 초공간 시나리오에 기반하여 UAV(Unmanned-Aerial Vehicles)를 활용한 음영지역 탐색 및 해결 방안을 제안한다. 음영 지역을 해결하기 위해 UAV와 임베디드 보드를 이용하여 6G의 테라헤르츠 대신 부산대학교 교내의 PNU-WiFi 전송 속도를 측정하고 와이파이 확장기의 설치 위치를 특정할 수 있는 알고리즘을 제안한다. 설계된 시스템은 드론을 이용하므로 음영 지역을 파악을 위한 무선 네트워크 신호 세기 측정 시 공간의 제약을 받지 않아 효율적이다. 기계학습을 이용하여 와이파이 확장기 설치 위치 특정을 할 수 있으며 특정된 위치에 확장기 설치를 통해 음영 지역에서의 와이파이 속도가 향상됨을 확인했다. 이 시스템을 구현함으로써 6G 초공간 시나리오의 UAV를 활용한 효율적인 무선 네트워크 음영 지역 해소를 기대할 수 있다.

1. 서 론

5G 무선 네트워크는 이전 세대의 통신 기술인 4G LTE를 대체하고 발전시킨 기술로, 현재 활발하게 개발되고 있는 중요한 통신 기술이다. 4G LTE가 1.8/2.1/2.6GHz의 낮은 주파수 대역을 사용하는 것과 달리 5G는 빠른 속도와 높은 전송량을 제공하기 위해 3.5GHz~6GHz (Sub-6G) 주파수 대역을 포함하여 28GHz와 39GHz 등의 밀리미터파(mmWave)로 불리는 초고주파 대역까지 사용하고 있다. 하지만 초고주파의 경우 물리적 특성상 낮은 주파수에 비해 멀리까지 전파되지 못하고 장애물을 통과하는 투과력이 상대적으로 약한 특성이 있다. 이러한 특성으로 인해 4G LTE보다 더 많은 음영지역(Coverage Holes)을 생성한다. 음영지역은 안정적인 무선 성능을 위해 필요한 수신 신호 강도의 임계치 수준 미만이 지역을 의미한다[1]. 이러한 음영지역을 없애기 위해 새로운 네트워크 노드를 추가로 배치하는 것이 간단한 방법이지만 비용 측면에서 비효율적이다. 따라서 현재 5G 네트워크에서 발생할 수 있는 음영지역을 효율적으로 탐지하기 위한 User Modeling Adaptation and Personalization(UMAP) 알고리즘의 개발[2], 그리고 해결하기 위한 빔포밍, Massive MIMO와 같은 주파수 기술의 개발이 이루어지고 있다. 하지만 이러한 해결 방안들은 지상 인프라에 의지하고 다양한 환경 변화에 적용하기 힘들다는 단점이 있다.

다가오는 6G 무선 네트워크에서는 테라헤르츠(THz) 기반의 초광대역(Ultra Precision Positioning), 초고신뢰·저지연(Ultra High Reliability & Low Latency), 초공간(Ultra 3D coverage), 초연결(Ultra Massive Connectivity), 초절감(Ultra Low Energy)의 6가지 서비스 시나리오로 세분화하고 있다[3]. 그중 초공간 시나리오에서는 지상에서의 통신뿐만 아니라 Unmanned-Aerial

Vehicles(UAV)를 활용한 비 지상에서의 통신 기술도 제안되고 있다[4]. 5G보다 더 높은 주파수 대역을 사용하는 6G 특성 때문에 더 많은 음영 지역이 생길 것으로 예상된다. 6G에서 음영 지역의 탐색과 해결 방법으로 UAV를 이용한 음영지역을 제거하는 방식이 제안된다[4]. 이러한 방식은 지상 인프라에 의존하지 않아 공간적 제한이 적고 사용자의 이동성 특성에 따라 자율적으로 배치될 수 있다는 장점이 있다.

본 실험에서는 UAV를 이용한 음영지역의 탐지 및 해결을 제안한다. 그러나 실제로 6G 무선 네트워크가 현재 구현되어 있지 않으므로 교내에 존재하여 쉽게 측정이 가능한 네트워크 데이터인 PNU WiFi의 속도 데이터를 대신 이용하여 UAV를 이용한 음영지역의 탐지 및 해결을 구현한다. 따라서 본 논문에서는 UAV와 임베디드 보드를 이용해 특정 지역의 와이파이 데이터 전송량을 측정하고 음영지역을 분석 및 해결하는 기술을 제안한다.

논문 구성은 다음과 같다. 2장에서는 본 논문이 제시한 드론을 이용한 교내 Wi-Fi 음영 지역 확인 및 해결 시스템에 대해 자세히 설명한다. 3장에서는 제안한 방법의 결과를 분석한다. 마지막으로 4장에서는 결론과 향후 연구 방향을 간략히 서술한다.

2. 시스템 설계 및 구현

2.1. 데이터 수집 장치 및 실험 환경

데이터 수집장치는 Wi-Fi와 연결할 수 있는 모듈이 내장되어 있는 라즈베리파이 보드를 사용하였고, 데이터 시각화를 위한 위치 데이터를 측정하기 위해 GPS 모듈을 사용하였다. 그림 1은 시스템의 구성도로 수집장치를 이용해 데이터를 측정하고 이를 서버로 전송하여 저장한다. 이후 저장한 데이터를 이용하

※본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터사업의 연구결과로 수행되었음. (IITP-2023-RS-2023-00260098)

여 음영지역을 확인할 수 있게 시각적으로 나타내고 해결을 위한 결과를 웹에 나타낸다.

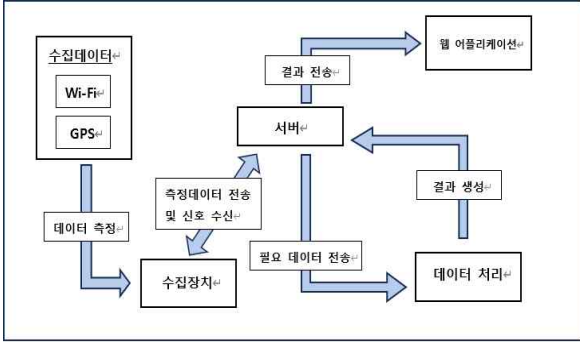


그림 1. 시스템 구성도

그림 2와 3은 실험 환경을 보여준다. NETGEAR AX1800 EAX20 와이파이 확장기를 실외용 배터리팩으로 전원을 연결하여 사용한다. DJI Mini 2 드론 위에 데이터 수집 장치를 넣은 케이스를 설치하고 내장된 프로그램을 실행하여 와이파이 다운로드 속도 데이터와 위도, 경도 데이터를 측정하여 저장한다.



그림 2. WiFi 확장기



그림 3. 데이터 수집 장치 및 드론

와이파이 다운로드 속도 데이터의 측정은 부산대학교 시월광장에서 측정하였으며 그림 4 처럼 총 16개의 지점에서 측정하였다. 이웃한 각 지점 간의 거리는 5m이다. 실외 환경이기에 유동인구가 많을수록 측정에 어려움이 있어 유동인구가 30분당 1~2명인 19시~23시 사이에 측정하였다.

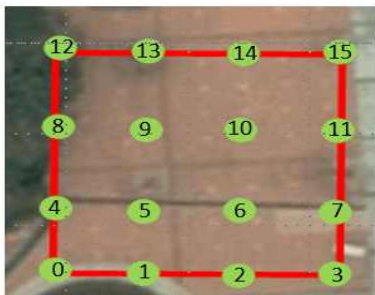


그림 4. 시월 광장 데이터 측정 지점

2.2. 회귀를 이용한 다운로드 속도 예측

와이파이 속도는 거리에 종속된다. AP(Access Point)라고 불리는 공유기, 확장기와 같은 무선 단말기와 가까울수록 높은 다운로드 속도를 가지게 된다. 이 사실을 바탕으로 확장기 설치 이후의 다운로드 속도 변화를 예측하고자 했다. 이를 위한 방법으로 확장기 설치 이후 각 지점의 속도를 학습하는 방식을 선택하였다. 이때 확장기와 현재 지점 간의 거리라는 독립 변

수와 속도라는 종속 변수가 있어 인과관계를 파악하고 속도 변화를 예측할 수 있는 회귀를 선택했다. 단순 선형 회귀 방식을 사용할 경우, 좋은 예측을 기대하기 어려웠다. 따라서 회귀의 방식은 차수가 많은 다항 회귀 방식을 사용했으며 MSE의 값이 가장 작게 나온 3차 함수를 선택하였다.

2.3. 확장기 위치 특정 알고리즘

2.3.1. 예측 데이터 생성

회귀를 이용하여 얻은 확장기와의 거리에 따른 속도를 확장기를 설치하지 않고 측정된 데이터와 비교한 후 더 큰 값을 적용하는 방식으로 예측 데이터를 생성한다.

이런 방식으로 1번 지점에서 16번 지점까지 확장기를 설치한 후 예측되는 Wi-Fi Download Speed 데이터를 생성한다. 이렇게 생성한 데이터를 이후 분류 과정을 통해서 나누고 분석한다.

2.3.2. k-Means Clustering

k-Means Clustering은 특정 임의의 지점을 선택해 군집 중심점을 설정하고, 해당 중심에 가장 가까운 포인트 들을 선택하는 군집화 기법이다. k-Means Clustering을 사용할 경우 데이터를 각 속도 군집 별로 분류 할 수 있는 알고리즘이다. 각 군집의 중심과 데이터들이 어느 군집에 속하는 지 파악이 가능하고 결과 해석이 간단하다.

최적의 클러스터 개수, 즉 k값을 찾기 위해 Elbow Method를 이용하였다. Elbow Method는 그래프가 꺾이는 부분의 k값을 최적의 값으로 판단하는 방법이다. 그러므로 그림 5의 결과에 따라 그래프가 가장 크게 꺾이는 부분인 3을 선택했다. 클러스터 개수를 3으로 정해 k-Means Clustering을 수행하여 Low speed, Middle speed, High speed의 3가지 군집으로 데이터를 분류했다.

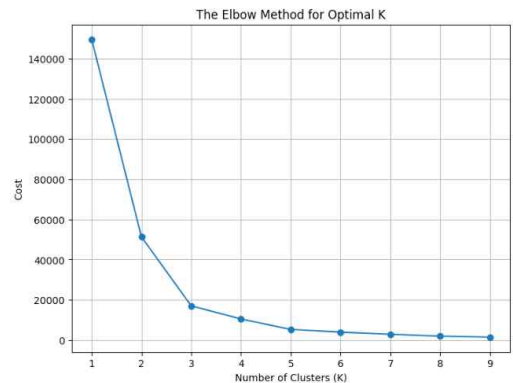


그림 5. Elbow Method 결과

2.3.3. 설치 위치 특정을 위한 기준 설정

분류된 데이터의 분석을 통해 음영지역을 가장 효율적으로 해결할 수 있는 확장기의 설치 위치 i 를 찾기 위한 방법을 제안한다. 단순히 속도가 가장 낮은 지점에 설치하는 결과가 아닌 속도가 낮은 지역을 전체적으로 해결할 수 있는 설치 지점을 찾는 것을 목표로 하였다. 이를 위해 분류 결과를 분석하여 3가지 기준을 만들었다.

확장기 설치 위치 i 에서 음영 지역 해결을 위한 3가지 기준은 표 1과 같다. x_i 는 낮은 다운로드 속도를 가지는 음영지역을 파악하기 위해 설정하였다. y_i 는 Download speed의 최소값이 클수록 확장기 설치 이전에 낮았던 속도가 개선되었음을 확인할 수 있는 지표로 보기 위해 설정하였다. z_i 는 전체적인 속도의 크기를 비교하기 위하여 설정하였다.

표 1. 설치 위치 특정을 위한 기준

기호	의미
x_i	i 에서 예상되는 Low speed 클러스터의 원소의 개수
y_i	i 에서 예상되는 Download speed 최소값의 크기
z_i	i 에서 예상되는 전체 Download speed의 평균

확장기 설치 위치 i 를 특정하기 위한 기준으로 Score 값 계산하는 수식을 정의하였다. Score 수식은 수식 1과 같다.

$$Score = \alpha x_i + \beta y_i + \gamma z_i \quad (1)$$

Score의 수식에 있는 각 가중치 α, β, γ 값은 다음과 같다. 각 가중치의 절대값들을 모두 더하여 1이 나오도록 설정했다. 실험에서 설정한 가중치로 α 는 -0.2, β 값은 0.5, γ 값은 0.3으로 지정하였다. 가장 높은 Score 값을 가지는 위치 i 가 확장기의 위치로 판단한다.

3. 결과

본 논문의 실험 환경인 부산대학교 시월 광장의 각 지점에서 거리에 따라 확장기 속도를 측정한 데이터와 선형 회귀를 결과는 그림 6에서 확인할 수 있다. 표 2에서는 확장기와의 거리에 따른 속도의 선형 회귀를 통한 예측 결과를 확인할 수 있다. 확장기와의 거리가 멀어질수록 속도가 줄어들어 21.21m에서는 18.09Mbps의 값을 가지고 가까울수록 속도가 커져 0m에서 31.73Mbps의 값을 가진다.

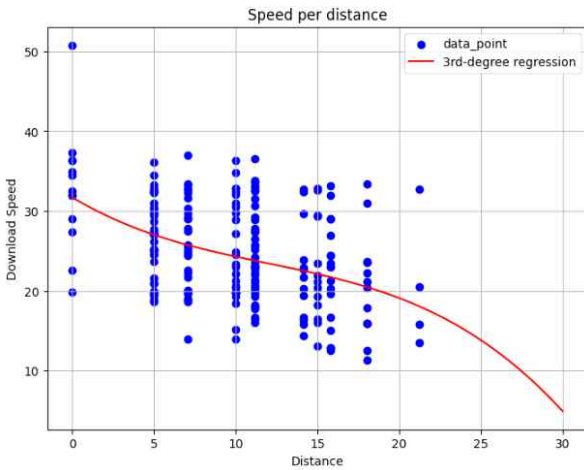


그림 6. 거리에 따른 확장기 세기 데이터 다중 회귀 결과

표 2. 확장기와의 거리에 따른 와이파이 예측 속도

확장기와의 거리(m)	예측 속도(Mbps)
0	31.73
5	27.15
7.07	25.81
10	24.3
11.18	23.78
14.14	22.47
15	22.20
15.81	21.76
18.03	20.58
21.21	18.09

선형 회귀로 얻은 확장기와의 거리에 따른 속도 예측 결과를

이용해 1~16번의 위치에 확장기를 설치 시 예측되는 각 지점의 속도와 기존 속도를 비교한다. AP와 확장기 중 속도가 더 큰 곳에 연결되어야 하므로 두 값 중 큰 값을 사용하여 16개의 예측 데이터를 생성한다. 생성된 데이터를 k -Means Clustering을 사용하여 Low speed, Middle Speed, High speed로 분류했다. 결과는 그림 7과 같다.

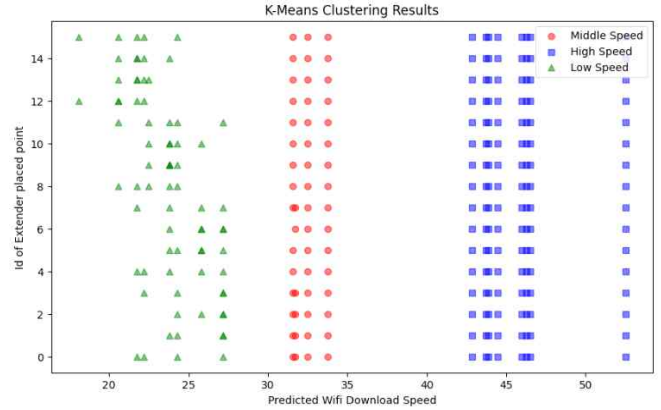


그림 7. k -Means Clustering 결과

최종적으로 클러스터링 결과 x, y, z 값을 구해 수식(1)에 대입하여 Score 값을 계산했다. Score의 값의 결과는 표3과 같다. 표3을 통해 음영지역을 가장 효과적으로 해결하기 위한 확장기의 위치를 가장 높은 Score를 나타낸 2번 지점을 설치 지점으로 지정하였다.

표 3. Score 계산 결과

i	x	y	z	Score
2	4	24.30	37.502500	22.600750
1	4	23.78	37.375625	22.302687
⋮	⋮	⋮	⋮	⋮
15	5	18.09	35.676875	18.748062
12	5	18.09	35.444375	18.678313

이를 확인하기 위해 확장기를 설치하기 전과 2번 지점에 설치한 후의 데이터 시각화 결과의 비교를 진행하였다. 그림 8은 확장기 설치 이전의 측정 결과를 시각화한 결과로 오른쪽 아래에서 왼쪽 아래 부분의 속도가 최소 5.95Mbps, 최대 15.5Mbps로 다른 지점에 비해 낮은 속도를 가지는 음영지역이 형성되어 있음을 볼 수 있다.



그림 8. 설치 이전의 Download Speed

그림 9의 경우 확장기 설치 이후의 결과로 확장기 설치 시의 속도와 기본 속도를 비교해 더 큰 값을 선택하게 하였다. 결과적으로 설치 이전에는 속도가 10Mbps 이하로 내려가는 음영지역이 존재하였지만 설치 이후에는 이전의 음영지역이라고 판단되는 곳의 속도가 최소 24.3Mbps에서 최대 31.73Mbps까지 상승한 것을 볼 수 있다.



그림 9. 설치 이후의 Download Speed

4. 결론 및 향후 연구

본 실험에서는 6G 네트워크의 음영지역을 해결하기 위해 제안되는 UAV를 이용한 음영지역 탐지 및 해결 방식을 교내 PNU WiFi의 속도 데이터로 구현했다. UAV와 임베디드 보드를 이용해 교내 PNU WiFi의 속도와 GPS를 측정하고 음영지역을 해결하기 위한 와이파이 확장기의 위치를 특정하는 시스템을 제안했다.

제안하는 시스템에는 GPS 모듈과 와이파이 모듈 등의 장치를 임베디드 보드에 활용하여 해당 위치에서의 와이파이 속도 데이터를 측정하고 시각화했다. 이를 통해 제안한 음영 지역 탐지 및 해결 제안 시스템에는 예상되는 와이파이 속도를 다중 선형 회귀를 이용해 예측했다. 예측된 결과를 k -Means Clustering을 이용해 3가지의 군집으로 분류하였다. 클러스터링 결과를 바탕으로 효과적인 확장기의 설치 위치를 예측하고 결과에서 이를 확인했다.

본 논문에서 제시된 UAV를 활용한 음영 지역 탐지 및 해결 방안은 와이파이를 이용하여 실험했다. 향후 연구에서는 6G 초공간 시나리오에서의 실현 가능성을 판단하기 위해 6G 기지국과 확장기를 활용하여 실험해보는 것이 필요하다. 위 실험에서 사용된 확장기는 드론의 출력보다 무거워 실험 환경에서 지상에 위치했다. 이 점을 고려하여 보다 경량화된 확장기를 드론에 장착해 이동식 확장기로서 실험해 보는 것이 필요하다.

참고 문헌

[1] A. Gómez-Andrades, R. Barco, and I. Serrano, "A method of assessment of LTE coverage holes," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, pp. 1-12, Dec. 2016.

[2] C. K. Anjinappa and I. Guvenc, "Coverage Hole Detection for mmWave Networks: An Unsupervised Learning Approach," in *IEEE Communications Letters*, vol. 25, no. 11, pp. 3580-3584, Nov. 2021.

[3] 고남석, 박노익, 김선미. "6G 모바일 코어 네트워크 기술 동향 및 연구 방향.", 한국전자통신연구원, 2021.

[4] Rizwana Ahmad, Moussa Ayyash, Haythem Bany Salameh, ReyadEl-Khazali and Hany Elgala, "Indoor Flying Networks for 6G: Concepts, Challenges, Enabling Technologies, and Opportunities," *IEEE Communications Magazine*, 22. Feb. 2023.

크롬 확장기능을 사용한 브라우저 기반의 코드 작성기

이선호[○], 신희용, 윤한진, 송하주

부경대학교 컴퓨터인공지능공학부

zzangsh2946@naver.com, boxone3840@gmail.com, dgw0620@naver.com, hajoosong@pknu.ac.kr

A Browser Based Code Editor using A Chrome Extension

Seonho Lee[○], Heeyong Sin, Hanjin Yun, Ha-Joo Song

School of Computer Engineering and Artificial Intelligence, Pukyong National University

요 약

4차산업 시대를 맞아 프로그래밍 능력에 대한 요구가 늘어나며 코딩 학습이 인기를 얻고 있다. 웹페이지, 블록그, 동영상 등의 다양한 코딩 학습 매체를 통해 소스 코드가 제공되고 있으나 텍스트 형식으로 제공되지 않는 경우가 흔하다. 또한 소스코드 작성, 컴파일, 실행과 테스트 환경을 구축하는 과정이 필요하다. 이에 본 논문은 OCR(광학문자인식)을 이용한 코드 추출 기능과 작성한 코드를 인공지능 챗봇을 통해 질문하는 기능을 탑재한 소스 코드 실행 기능의 크롬 확장 프로그램을 제안한다. 사용자는 동영상 또는 이미지로 제공되는 소스코드를 캡처하여 브라우저 상의 코드 작성기에서 작성하고 컴파일하여 결과를 확인할 수 있다. 따라서 개발환경의 구축이 필요없고 소스 코드가 제공되지 않는 환경에서도 초보자들이 간편하게 프로그래밍 과정을 경험해 볼 수 있다..

1. 서 론

통합개발환경이나 텍스트 편집기를 설치하고 설정을 구성하는 과정은 초보자에게는 어려울 수 있다. 이러한 번거로움을 해소하기 위해 온라인에서 바로 코딩을 시작할 수 있는 편리한 환경을 제공하고자 한다. 또한, 영상 강의는 많은 코딩 학습자들에게 인기가 있다. 그러나 영상을 시청하면서 따로 코딩 환경을 구축하고 결과를 확인하는 것은 번거로울 수 있기에 이 프로그램은 영상 재생 화면 옆에 즉시 코딩할 수 있는 창과 결과 확인 창을 제공하여, 학습자들이 즉각적으로 코드를 작성하고 실행 결과를 시각적으로 확인할 수 있도록 도와준다. 프로그래밍에서는 코드 작성 후 실행 결과를 확인하고, 오류를 찾아 수정하는 과정이 반복된다. 이 프로그램은 즉시 코드를 컴파일 후 실행하여 결과를 보여주고, 인공지능을 통해 오류의 원인과 해결책을 피드백 받을 수 있어 학습자들은 작성한 코드의 동작과 오류를 실시간으로 확인하고 해결할 수 있다. 이는 학습자들이 빠르게 문제점을 파악하고 개선하는 데 도움을 준다.

본 논문에서는 OCR(optical character recognition)기능과 AI 챗봇을 활용한 소스 코드 실행 크롬 확장 프로그램을 제안한다. 이를 통해 프로그래밍 초심자들이 간편하게 코딩 경험을 할 수 있도록 하였다.

2. 관련 연구

제안하는 확장 프로그램과 유사하게 소스 코드를 실행하고 결과를 확인할 수 있는 서비스로 'tutorialspoint', 'ideone', 'paiza.io' 등이 있다. 이러한 서비스들은 웹에서 사용자가 소스 코드를 적고 실행하여 실행에 대한 결과를 확인할 수 있는 서비스를 제공하며, 사용자의 소스 코드의 실행 결과를 확인할 수 있도록 하는 측면에서는 제안 서비스와 큰 차이가 없다. 반면 제안하는 서비스에서는 회원가입과 로그인 없이 서비스를 제공하며, 동영상의 코드를 문자 인식하여 제공함으로써 사용자에게 도움을 줄 수 있도록 하였다. 또한 코드의 오류나 궁금증

에 대하여 AI 챗봇에게 질문함으로써 문제를 해결할 수 있도록 한다. 이러한 차별성을 통해 사용자에게 보다 편리한 학습 환경과 학습 경험을 제공한다.

3. 브라우저 기반 소스 코드 작성기의 설계

그림 1은 제안 시스템의 구조를 보인 것으로 웹 서버는 스프링 부트를 사용해서 구축하였고, 웹 서버에서는 사용자의 소스 코드를 컴파일 후 실행한 결과를 반환, 문자 인식의 결과를 반환과 사용자의 질문에 대한 AI 챗봇의 답변을 반환하는 기능들을 REST API 방식으로 제공한다. 소스 코드를 컴파일 후 실행하기 위해서 웹 서버에 gcc 컴파일러, java 컴파일러와 python3 인터프리터를 설치한다. 사용자가 입력한 소스 코드와 입력값을 클라이언트로부터 전달받고 소스 코드와 입력값을 웹 서버에 저장하는데, 웹 서버는 다수의 클라이언트가 동시에 요청할 수 있기 때문에 동시성 문제를 해결하기 위해 사용자의 소스 코드, 입력값과 결과값을 random uuid를 사용한 파일명과 코드 언어에 맞는 확장자와 텍스트 확장자로 저장한다.

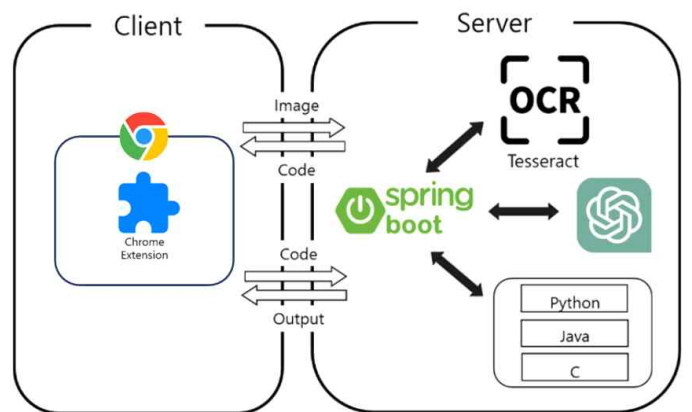


그림 1. 제안 시스템의 구조

C와 C++은 gcc 명령어를 사용해서 실행 파일을 생성하고, 실행 파일을 실행 후 exit code를 반환받아 성공했을 시 결과 파일을 읽어서 클라이언트에 반환한다.

python은 python3 명령어를 사용해서 인터프리터 방식으로 실행하고, 실행이 끝날 때까지 기다린 후 성공했을 시 결과 파일을 읽어서 브라우저에 반환한다.

Java는 java 명령어를 사용해서 컴파일 하지 않고 소스 코드를 바로 해석하는 방식으로 실행한다. javac 명령어로 바이트 코드 파일인 .class 파일을 만든 후 실행한다면, 파일명과 public class 명이 같아야 하는 규칙이 있다. 웹 서버는 동시성 문제를 해결하기 위해 모든 파일명을 random uuid로 저장하기 때문에 해당 규칙을 지킬 수 없다. 그래서 java 명령어를 사용해서 실행한다. 또한 실행이 끝날 때까지 기다린 후 성공했을 시 결과 파일을 읽어서 클라이언트에 반환한다.

문자 인식 기능을 위해 tesseract 라이브러리를 사용한다. 클라이언트는 문자가 포함된 이미지를 웹 서버로 전달하고, 웹 서버는 해당 이미지를 tesseract 라이브러리에서 제공하는 문자 인식 기능을 사용해서 결과를 클라이언트에 전달한다.

AI 챗봇을 사용한 질문에 대한 답변을 반환하기 위해 OpenAI에서 제공하는 ChatGPT API를 사용한다. 웹 서버에서는 클라이언트로부터 질문을 받으면 ChatGPT API를 호출하여 질문에 대한 답변을 받고, 클라이언트로 반환한다.

4. 크롬확장의 구현

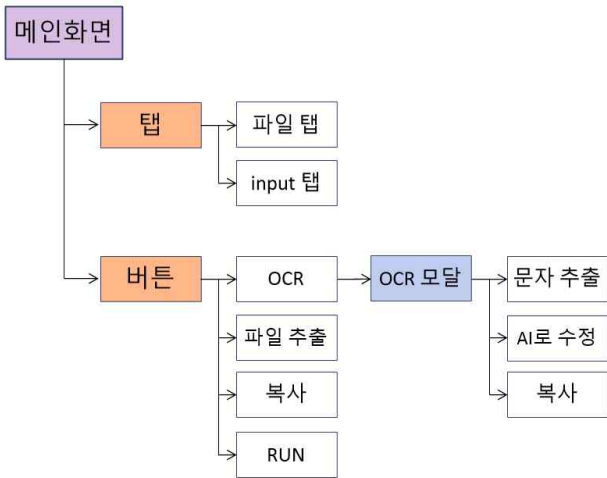


그림 2. 프로그램 메뉴 및 기능

제안하는 서비스의 클라이언트는 대중적인 사용을 위해 전 세계에서 가장 점유율이 높은 브라우저인 크롬 브라우저의 확장프로그램과 html, css, javascript를 사용하여 개발하였다. 그림 2는 메뉴 구성에 관한 그림이다. 화면은 크게 메인화면과 OCR 화면 두 가지로 구성되고 메인 화면에는 상단부터 TAB, BUTTON, RUN의 구성으로 나뉜다. BUTTON의 캡처 OCR 버튼을 통해 OCR 모달인 그림 4의 모습을 활성화할 수 있다.

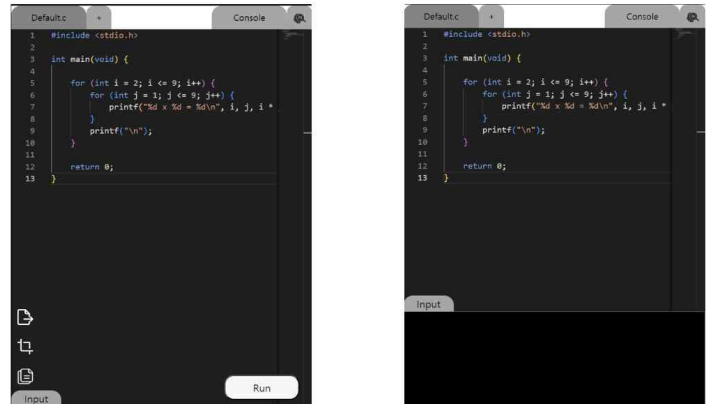


그림 3. 소스 코드 편집기

그림 3은 소스 코드 편집기로 제안하는 서비스의 클라이언트의 메인 화면이다. 탭은 크게 소스 코드 탭과 콘솔 탭으로 구분된다.

소스 코드 탭은 '+' 버튼을 이용하여 .c, .cpp, .py, .java 형식의 소스 코드 탭을 생성할 수 있다. 또한, 각 탭의 삭제기능을 추가하여 생성과 삭제가 자유롭게 구현하였다. 콘솔 탭은 작성한 코드와 Input 탭 상의 입력값을 바탕으로 서버에서 실행된 결과를 콘솔에 출력하여 확인할 수 있다. 메인화면에서의 버튼은 파일 추출, 캡처 OCR, 복사 버튼으로 구성된다. 파일 추출 버튼은 현재 활성화된 소스 코드 탭의 코드를 파일의 형태로 PC의 원하는 장소에 추출하여 저장할 수 있다. 캡처 OCR 버튼은 그림 4의 OCR 모달창을 띄워 캡처한 화면에 대한 OCR 기능을 사용할 수 있다. 복사 버튼은 현재 활성화된 소스 코드 탭의 코드의 전체를 복사한다. Run 버튼은 현재 활성화된 탭의 코드와 Input 탭의 input 값을 .c, .cpp, .java, .py와 같이 파일의 형태로 전달하여 컴파일을 실행한다. Input 탭은 통해 서버로 파일을 전송할 때, Input 탭에 적힌 Input 값을 input.txt 파일로 만들어 전송한다.

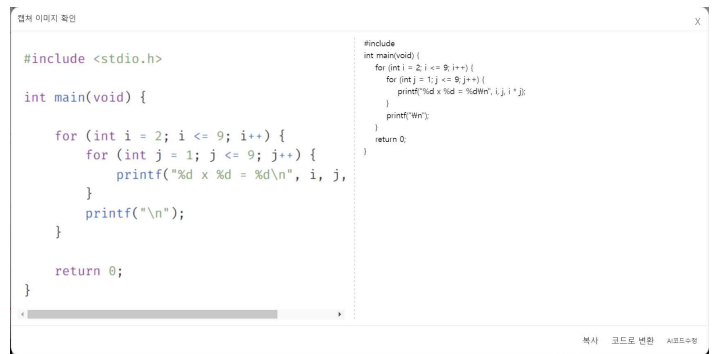


그림 4. 문자 인식 모달

그림 4의 문자 인식 모달창은 캡처 OCR 버튼을 통해 활성화 된다. 이 모달창은 복사된 화면창과 OCR 결과창으로 구분된다. 이 창들을 통해, 캡처를 통한 OCR 기능을 사용할 수 있다. 복사된 화면창은 ctrl + v를 통해 클립보드에 가장 마지막으로 저장된 파일이 이미지 파일이라면, 그 이미지파일을 띄어주는 창이다. OCR 결과창은 '코드로 변환' 버튼과 'AI코드 수정' 버튼을 통해 서버로 전송된 후, OCR 결과 추출된 텍스트를 출력하는 창이다. 코드로 변환 버튼은 클립보드의 가장 마지막으로 저장된

항목이 이미지이면, 이 파일을 .png 형식으로 서버에 전송하여 OCR 결과로 추출된 텍스트를 결과창에 출력한다. AI 코드수정 버튼은 현재 코드로 변환을 통해 출력된 결과를 OpenAI의 ChatGPT API를 사용하여 코드의 오류를 수정하여 추출된 결과를 결과창에 출력한다.

5. 결론

본 논문에서는 소스 코드를 작성 및 실행할 수 있고 OCR과 AI 챗봇 기능을 포함한 크롬 확장 서비스를 제안하였다. 사용자들은 별도의 IDE를 설치하여 환경 설정을 하지 않고, 확장 프로그램만 설치함으로써 간편하게 소스 코드를 실행하며 오류를 해결할 수 있다. 그러나, OCR 기능은 들여쓰기와 높은 인식률을 보이지 않는다는 점에서 문자 인식 기능의 개선이 필요하다. 본 논문에서 제안하는 크롬 확장 프로그램은 사용자 편의성과 개발 환경의 간소화를 목표로 하고 있으며, 성능 개선과 OCR 인식을 개선을 통해 보다 더 실용적인 서비스로 발전시킬 수 있을 것으로 기대된다.

참고문헌

[1] H.Y.Park(2023.Seq.1).Elementary Coding Boom for Nurturing Digital Talents... Smart Learning and Online Courses on the Rise.gpkorea. Available: <https://www.gpkorea.com/news/articleView.html?idxno=104666>

[2] G.H.Kang, J.H.Ko, Y.J.Kwon, N.Y.Kwon, S.J.Koh "A Study on Improvement of Korean OCR Accuracy Using Deep Learning" pp.693 - 695, 2018.

[3] G.C.Lee, J.S.Yoo "Development an Android based OCR Application for Hangul Food Menu" Proceedings of Journal of the Korea Institute of Information and Communication Engineering vol. 21, no. 5, pp. 951-959, 2017.

[4] D.W.Lee, Y.S.Jang, D.Y.Lee, S.B.Jeon, D.M.Seo "The Method of Number Detection in Analog Gauge using Tesseract-OCR" Proceedings of Korea Software Congress 2020(KSC2020) , pp. 228-230, 2020.

[5] Y.S.Son, H.L.Lee, S.H.Park, J.W.Shin "How to use Tesseract OCR technology to expand information usability" Proceedings of Korea Software Congress 2019(KSC2019) , pp. 1278-1280, 2019.

지식그래프를 활용한

버티포트 리소스 관리 방안에 대한 연구

고동연¹, 김영인¹, 김유진¹, 김재렬¹, 송나흔¹, 김정석², 정설영¹

¹ 경북대학교, ²SK 텔레콤

carb014@knu.ac.kr, hardworking@knu.ac.kr, ky107601@knu.ac.kr,
jimsil261@knu.ac.kr, etnosd01@knu.ac.kr, jeongseok.kim@sk.com,
snowflower@knu.ac.kr

Study on the Knowledge Graph-based Vertiport Resource Management Method

Dongyeon Go¹, Youngin Kim¹, Yujin KIM¹, Jaeyeol Kim¹, Naheun Song¹,

Jeongseok Kim², Seol-Young Jeong¹

¹ Kyungpook National University, ²SK Telecom

요 약

UAM (Urban Air Mobility) 분야는 도심 공간에서의 새로운 항공 이동 기술 및 서비스의 발전으로 주목받고 있다. 그러나 복수의 운항 사업자가 동일한 항로와 기착지를 공유함으로써 발생하는 혼잡과 안전 문제가 UAM 분야의 새로운 도전 요소로 부상하고 있다.

본 논문에서는 복수의 운항 사업자 간의 협력 강화를 위해 선결 문제인 버티포트 리소스 관리 문제를 해결하기 위한 새로운 시스템을 탐구했다. 복수의 운항 사업자 간의 비행 일정을 조율하고 비정상 상황에 대비하여 안전 및 효율성을 향상시키는 방안을 모색하기 위해 다양한 스케줄 데이터를 만들고 그 데이터를 바탕으로 한 시뮬레이션을 수행하는 시스템을 구성했다. 이 시스템은 버티포트 현재 상황 및 비정상 상황 판단을 위해 지식 그래프를 활용하였고 교통 관제 시스템의 신속한 결정과 유연한 대응이 가능한 체계를 구축했다. 이 실험을 통해 버티포트 리소스를 관리할 때 발생할 수 있는 문제를 정의하고 해당 문제를 해결하기 위한 효율적 방안을 마련하였다. 또한, 비정상 상황으로 인해 리소스 스케줄링이 필요한 경우 효율적으로 대응하는 방안을 제시했다. 본 논문에서 제안하는 버티포트 리소스 관리 시스템은 UAM 을 안전하고 지속 가능한 방향으로 진화시킬 것으로 기대한다.

1. 서 론

현재 대도시 지역들은 교통 체증이나 통근 지연 등으로 사회적 비용이 증가하고 있다. 하나의 해결책으로 Urban Air Mobility(UAM)이 각광받고 있다. UAM 은 친환경 electric Vertical Takeoff and Landing (eVTOL) 항공기를 기반으로 한 차세대 이동 수단으로, 도심 공간에서 교통 혼잡 문제를 해소하기 위한 목적을 가진다. UAM 은 승객이나 화물 운송 등으로 타 교통수단과 연계될 수 있고, 공공목적(긴급의료 등)과 관광사업을 위해 운용될 수 있다[1][2]. 하지만 UAM 이 상용화되기 위한 선결 과제들이 존재하고, 본 논문은 그 중 하나로 제한된 구역 내에서 항공기 교통량 증가를 대비한 효율적 교통 관리 시스템 개발에 집중했다.

UAM 교통 관리 시스템으로 항공기에서 사용하는 Air Traffic Management(ATM) system 을 그대로 사용하기에는 ATM 의 경우 도시 근처 또는 도시 내에서 상당한 거리의

공항 사이의 정기 비행을 관리하기 위해 설계되었고, 일반적인 비행 고도(예: 10,000 피트 이상)를 다루지만 UAM 의 경우, 요청에 따른 고밀도 단거리 비행을 관리하며, 도시 항공 공간과 매우 가까운 고도 (예: 10,000 피트 미만)에서 운항되는 항공기로 적용이 어렵다[3]. UAM 은 교통량이 많은 도시의 항공 공간에서 운항이 될 예정이므로, 현재 ATM 서비스가 수행할 수 있는 고도보다 더 낮은 고도 및 거리를 표준으로 운영을 해야 한다. UAM 과 유사한 환경의 교통 관리 시스템에 대한 통일된 인프라가 구축되지 않은 상황으로, 현재 실존하는 방식들인 UAS 나 공군이 사용하는 기술들을 채택해보려는 시도를 하는 중이다[3]. 그러나 이는 UAM 의 차량 크기 및 UAM 의 교통 밀도와 맞지 않는 문제가 발생할 수 있어 UAM 작업을 안전하게, 효과적으로 관리할 수 있는 ATM architecture 설계의 개선이 필요하다.

개선된 ATM 이 반드시 포함해야 하는 것은 버티포트 리소스 관리 방안이다. 버티포트의 수용량 변화정보, UAM 비행정보와 비행 계획과 관련된 정보를 종합하여 이해관계자들끼리 교류하고 UAM 항공기의 안전한 착륙을 보장하는 것이 중요하다[4]. 따라서 버티포트의 수용량 정보와 UAM 비행 관련 정보 리소스를 체계적으로 관리할 필요가 있다. 이를 위해 데이터에 쉽고 빠르게 접근하여 리소스를 관리할 수 있도록 지식그래프를 활용한 UAM Vertiport Resource Management 시스템을 설계하였다. 이를 통해 리소스를 효율적이고 지속적으로 관리하여 UAM 운행의 안전성과 효율성을 향상시키며 비정상 상황에도 유연하게 대비할 수 있는 기반을 마련하고자 한다.

따라서 본 논문에서는 버티포트의 상태 및 UAM 의 현재 운항 상태, 비정상 상황 등의 시나리오를 설정하고 시뮬레이션을 진행하여 flight plan 을 점검하고 비정상 상황에 더 효율적으로 대응하는 방법을 모색한다. 이를 통해 버티포트 리소스 관리 방안 및 일괄적으로 들어오는 flight plan 을 한번에 처리할 수 있게 한다.

2. 선행 연구

PSU(Provider of Services for UAM)는 운항 안전 정보 공유, 교통흐름 관리, 비행계획 승인, 항로 이탈 모니터링 서비스를 제공한다[1]. PSU 중심의 협력적 방법에 대한 한 연구에서, 시뮬레이션을 통해 비효율성, 연료 부족 비정상 상황, 항공기 간 거리 유지 실패 상황, 알고리즘 계산 시간을 분석하였다. PSU 가 제안하는 Strategic 방식은 계산 시간을 제외한 모든 시뮬레이션에서 가장 효율적으로 나타났다[5]. 따라서 안전한 운항을 위해 PSU 를 사용하는 것에 이어서 리소스 관리를 통해 효율을 극대화하는 방안을 탐구하였다.

3. 실험 및 시뮬레이션

3.1 시나리오 설계를 위한 가정

본 연구에서는 버티포트의 가용성, UAM 항공기 운항 상태, 실시간 수용량 변화 등의 정보를 UAM 교통관리 서비스 제공자에게 제공하는 버티포트 운영자[1]의 관점과 UAM 운항자, 버티포트 운영자, 운항 지원 정보 제공자 등 각 이해관계자에게 교통관리 서비스를 제공하는 UAM 교통관리 서비스 제공자[1]의 역할을 통합해 간단히 변수명 ‘Operator(이하 관제사)’로 가정하였다. 버티포트 리소스 관리의 복잡도를 낮추기 위하여 1 명의 관제사가 3 개의 버티포트를 관제하며 각 버티포트에는 4 개의 버티패드(Vertipad, UAM 1 개를 주차할 수 있는 공간)가 설치되어 있으므로 제한하였다. 버티포트 및 회랑의 위치와 형태 등은 승객 수요와 지형 정보를 고려하여 사전에 정의되어 있다는 가정을 적용하고, UAM 노선 상황도 이미 배정 되어있다고 가정한다[6]. UAM 은 250km 속도로 일정하게 운행한다고 가정하며, 각 항로들은 버티포트

사이의 직선 거리를 이용한다고 가정한다. 또한 주변의 비협력적인 비행체의 출현 등과 같은 모든 상황은 UAM 이 관조하고 있다고 가정한다.

3.2 지식 그래프

지식 그래프(Knowledge Graph)는 정보를 구조화하고 연결하는 방법 중 하나로, 그래프 형태로 정보를 표현하여 엔터티(개체)와 엔터티 간의 관계(Relationship)를 나타낸다. 각 노드는 어떤 유형의 엔터티를 나타내며, 간선은 엔터티 간의 관계를 나타낸다. 지식 그래프는 관계형 데이터베이스와는 달리 미리 정의된 스키마가 필요하지 않아 변경이 용이하며, 의미론적 관계를 저장할 수 있어 다양한 이해관계자가 전송하는 데이터를 분석하고 통합하는데 큰 역할을 한다. 본 연구에서는 효율적으로 UAM 의 비행계획을 관리하기 위해 UAM, Flightplan, Waypoints, Vertiport 클래스를 정의하고 이를 위한 속성들도 함께 정의했다. uam 은 uam 비행 정보가 들어가 있고 고유의 uam ID, 배터리 상태, 그리고 현재 상태와 위치를 받은 시간이 포함된다. 또한, 각 항공 운송 장치는 located_at 관계를 통해 현재 위치 정보를 Coordinate 클래스와 연결한다.

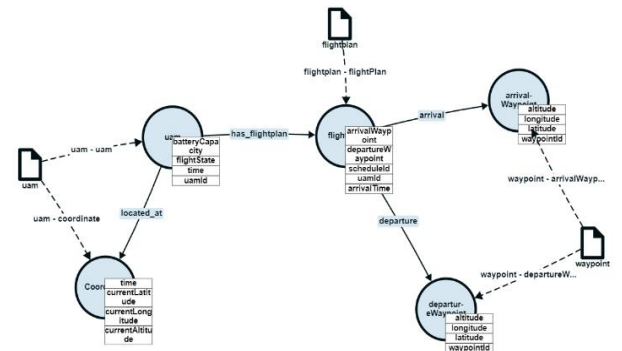


그림1. 비정상 상황 판단을 위한 지식그래프

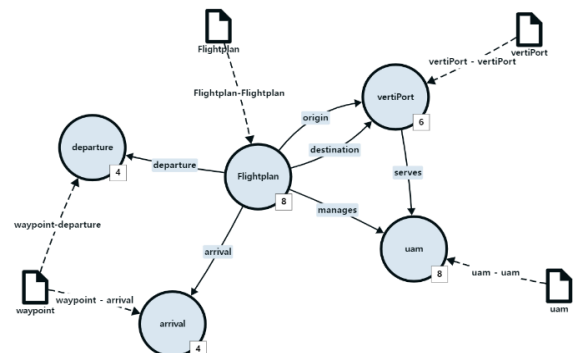


그림2. 현재 버티포트 상태 판단을 위한 지식그래프

Flightplan 은 UAM 의 일정을 나타내며, 각 일정에는 고유의 일정 ID, 출발 지점 ID, 도착 지점 ID, 출발 시간, 도착 시간,

그리고 해당 일정에 관련된 UAM 의 ID 가 포함된다. Schedule 클래스는 Vertiport 에게 출발지와 도착지를 각각 origin, destination 으로 mapping 해 전달하여 vertiport 가 vertipad 를 효율적으로 관리할 수 있도록 하고, UAM 이 비행을 잘 수행하고 있는지 manages mapping 으로 관리한다. arrivalWaypoint 는 도착 지점을 나타내며, 각 도착 지점에는 고유의 Waypoint ID 와 그에 해당하는 좌표(위도, 경도, 고도)가 포함된다. Coordinate 는 UAM 의 현재 위치를 나타내며, 이는 위도, 경도, 고도로 구성된다. departureWaypoint 는 출발 지점을 나타내며, 각 출발 지점에는 고유의 Waypoint ID 와 그에 해당하는 좌표(위도, 경도, 고도)가 포함된다.

Vertiport 클래스는 시간마다 Vertipad 별 상태를 나타낸다. Vertiport 는 UAM 에 현재 vertiport 상황을 serves mapping 으로 제공한다.

본 연구에서는 시뮬레이션을 위해 데이터를 생성했다. 데이터 형식은 다음과 같은 표로 정의되었다.

표 1. 데이터 형식

UAM	flightPlan	waypoint
recordId	flightPlanId	waypointId
uamId	uamId	latitude
time	departureWaypoint	longitude
latitude	arrivalWaypoint	altitude
longitude	departureTime	
altitude	arrivalTime	
batteryCapacity		
flightStatus		

UAM 은 1 초 간격으로 현재 위치를 보내며, 이동 경로는 출발지와 목적지 사이를 직선으로 가정한다. 이로 인해 배터리 용량과 시간이 기록된다. flightPlan 은 출발지, 목적지, 출발 및 도착 시간이 미리 정의되어 있으며, 해당되는 좌표는 waypoint 에 기록되어 있다.

3.3 시나리오 : 비정상 상황 발생 시

K-UAM 운용개념서 1.0 에 따르면 비정상 상황을 경미한 비정상 상황, 우발상황, 비정상 상황으로 구분하고 있다[1].

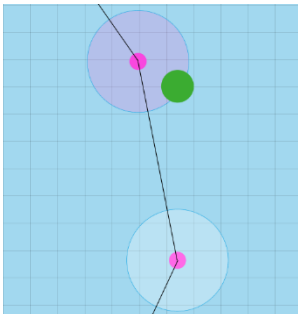


그림 3. Waypoint를 지나고 있는 UAM

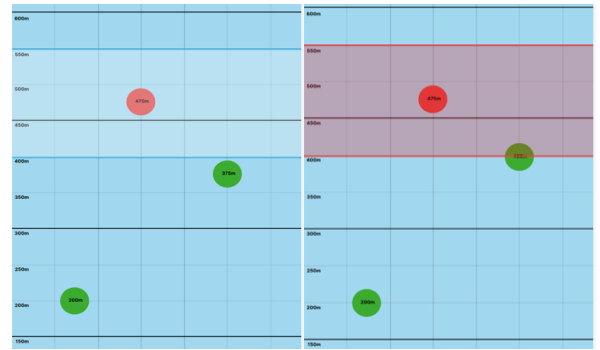


그림 4. 좌) 정상상황의 고도 우) 비정상 상황의 고도

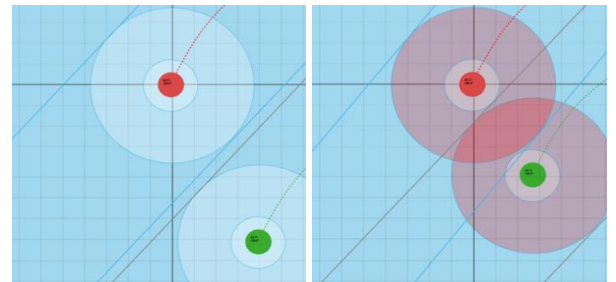


그림 5. 정상상황의 위도 및 경도 우) 비정상상황의 위도 및 경도

본 연구에서는 계획된 목적지에 도착할 수 없는 우발상황 및 비정상 상황이 아닌, 최초 계획된 목적지로 비행은 가능하나 일부 시스템 및 환경 등이 비정상인 경우인 비정상 상황만을 판단한다고 가정한다. 비정상 상황에는 지정된 항로를 벗어난 경우, 회랑 내 비협력적 비행체와 조우 등이 있다.

지정된 항로로의 이동을 판단하기 위해 항로마다 Waypoint 를 설정하여 Waypoint 에 일정 시간 내 적절한 오차범위로 도착하였는지를 확인한다. 그림 3 은 시나리오를 위해 설정한 Waypoint 와 UAM 이 Waypoint 를 지날 때의 상황을 나타낸다. Waypoint 에 도착하였다면 정상 상황으로 0 을, 도착하지 않았다면 비정상 상황으로 1 을 전송하도록 한다. Waypoint 는 항로내에서 5km 마다 배치한다고 가정한다.

회랑 내 비협력적 비행체 및 복수의 운항 사업자의 UAM 의 돌발 출현의 경우도 비정상 상황으로 판단하여, 관제 시스템에 1 을 전달하며 관제 시스템에 항로 내 접근한 비행체의 위도와 경도, 고도를 출력한다. 그림 4 와 그림 5는 정상적인 UAM 의 이동과 복수의 운항 사업자의 UAM 가 동일 항로에 접근할 경우 관제 시스템의 변화를 나타낸다.

4. 결론

본 논문에서는 지식그래프로 버티포트 리소스를 효율적으로 관리해 버티포트 사용량을 극대화했다. 또한, 비정상 상황 시나리오와 이에 대한 시뮬레이션을 통해 항로를 재설정하

는 동적 스케줄링 방안을 제시했다. 이를 통해 비정상 상황 발생 시에도 안전하게 flightPlan 업데이트가 가능함을 확인했다. 그러나 시뮬레이션에서는 관제사가 1명임을 가정했으므로 관제사가 다수인 경우, 지식그래프로 표현한 버티포트 리소스 관리 방안의 데이터 동기화 문제에 대해서 테스트할 필요가 있다. 나아가 다른 비정상 상황이 발생했을 때에 대한 해결 방안을 찾는 후속 연구도 필요하다.

Acknowledgement

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음”
(2021-0-01082)

5. Reference

- [1] Ministry of Land, Infrastructure and Transport, "K-UAM Concept of Operations 1.0", Sep, 2021.
- [2] P. KrishnaKumar, J. Witter, S. Paul, H. Cho, K. Hantu, and S. Chowdhury, "Fast Decision Support for Air Traffic Management at Urban Air Mobility Vertiports using Graph Learning," arXiv preprint arXiv:2302.05849, 2023.
- [3] S. Bharadwaj, S. Carr, N. Neogi and U. Topcu, "Decentralized Control Synthesis for Air Traffic Management in Urban Air Mobility," IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, VOL. 8, NO. 2, pp.598-608, 2021
- [4] A. Straubinger, "An overview of current research and developments in urban air mobility – Setting the scene for UAM introduction", Journal of Air Transport Management, Volume 87, August 2020, 101852
- [5] S. Bharadwaj, S. Carr, N. Neogi, and U. Topcu, "Decentralized Air Traffic Management for Advanced Air Mobility", IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, VOL. 8, NO. 2, pp.1-8, 2021
- [6] E. Yeonju, J. Daekeun, K. Hyounkyoung and Y. Chanhong, "Operational Scenarios Development and System/Subsystem Design for VIPP(Virtual Integrated oPeration Platform) for UAM", The Korean Society for Aeronautical and Space Sciences, Issue 11, 2022

스마트팩토리를 위한 PINN

김재우⁰¹ 최도진^{1*}
창원대학교 컴퓨터공학과
lgapt5@naver.com, dojinchoi@changwon.ac.kr

PINN for Smart Factory

Jaewoo Kim⁰¹ Dojin Choi¹
Department of Computer Engineering, Changwon National University
lgapt5@naver.com, dojinchoi@changwon.ac.kr

1. 서론

PINN은 Physics-Informed Neural Network의 약자로, 물리 정보 신경망을 의미한다[1]. 이는 물리학 정보를 통하여 인공지능 모델을 개발하는 기술이다[2]. 기존의 인공지능 모델은 스스로 loss를 줄이는 방향으로 학습을 진행하지만, PINN은 공식을 기반으로 학습하기 때문에 정확도가 높아 수치 해석 분야에서 현재 각광 받고 있다[3-7]. 본 논문에서는 PINN에 대한 개념과 활용될 수 있는 경우, 그리고 간단한 예시 및 주의할 점에 대하여 소개한다.

2. PINN 설명

PINN은 물리 정보 신경망을 뜻한다. 보다 정확하게 표현하면, PINN은 편미분 방정식의 해를 구해 주는 기술이다. PDE는 Partial Differential Equation의 약자로 부분 미분 방정식을 의미하는데, PINN은 손실 함수 값을 줄이기 위하여 신경망 훈련으로 PDE 솔루션을 근사화한다. 그 후, 데이터가 없는 신경망은 PDE를 훈련 손실에 내장하고, 사용자가 지정한 딥러닝을 통하여 초기 및 경계 조건을 적용하여 작동한다. 그림 1은 PINN을 간단히 나타낸 것이다.

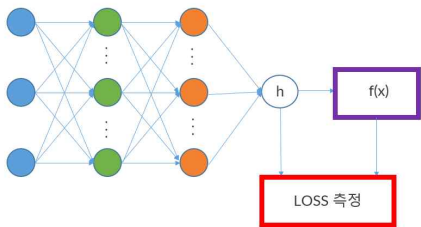


그림 1 PINN을 간단히 나타낸 그림

PINN은 물리 법칙을 모델에 통합할 수 있어, 모델의 예측 정확도를 더욱 높여준다는 장점이 있다. 또한, 데이터에 대한 의존도를 낮출 수 있어, 적은 데이터 양으로도 동작시킬 수 있다는 장점도 있다.

3. 활용 예시

PINN은 자연어 처리 등 미분이 되지 않는 경우에는 사용할 수 없다는 단점이 있다. 하지만 스마트 팩토리의 엣지 컴퓨팅에서 활용될 수 있을 것으로 기대된다. 예를 들어, 스마트 팩토리에서 각 기기별로 발생하는 데이터를 이용하여 고장 진단, 에너지 효율 최적화 등에 이용될 수 있을 것으로 보인다. 엣지에서 기록해야 할 만한 상황이 발생했을 때만 클라우드에 데이터를 전송하는 방식으로 설계하면, 클라우드에 큰 부하를 일으키지 않으면서도 적절한 성능을 낼 것으로 기대된다. 그림 2는 엣지 컴퓨팅에 사용될 수 있는 예시를 나타낸 것이다.

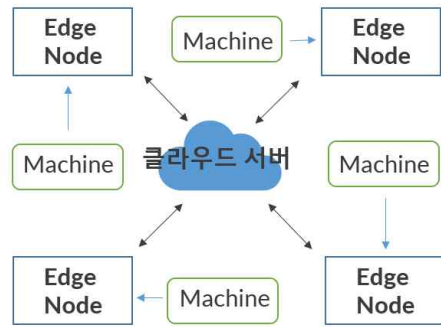


그림 2 활용 예시

4. 간단한 예시 및 주의할 점

실험 환경은 Google Colab을 사용하였다. 알고리즘을 구동시키기 위한 파이썬 라이브러리는 DeepXDE를 사용하였다. DeepXDE는 딥러닝을 사용하여 미분 방정식을 풀기 위하여 사용되는 라이브러리이다. 그림 3, 4, 5, 6에서 검정색 선은 물리 법칙을 나타낸 것이고, 빨간색 선은 학습에 의한 예측을 나타낸 것이다.

첫 번째 예시에서 학습된 방정식은 $\cos(t)$ 를 사용하였다. 신경망 구성은 입력을 하나의 노드로 받고, 16개, 8

* 교신저자 : dojinchoi@changwon.ac.kr

이 (성과)는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원과(No. RS-2022-00166906) 2023년도 창원대학교 강사-대학원생-교원 간 융합연구 지원 사업과 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.(2021RIS-003)

개, 4개, 2개로 노드를 이후에 구성하였으며, 마지막 출력 노드는 1개이다. 활성화 함수는 elu를 이용하였으며, optimizer는 adam, 초기 학습률은 0.01, epoch은 50으로 임의로 지정하였다. Time domain 설정 범위별 train loss와 test loss는 표 1과 같다. 모든 경우에서 loss가 상당히 낮은 편에 속하지만, time domain의 범위가 늘어날수록 loss가 증가하는 경향을 띄는 것을 볼 수 있었다.

표 1 time domain 범위별 loss

	0 ~ 2	0 ~ 5	0 ~ 10
train loss	0.85e-04	1.58e-02	3.34e-01
test loss	6.72e-04	1.46e-02	3.28e-01

prediction이 time domain이 0에서 2일 때, 0에서 5일 때는 상당히 정확하나, time domain이 0에서 10일 경우에는 prediction이 제대로 동작하지 않는 것을 볼 수 있었다. 그 이유는 time domain이 늘어날수록 데이터의 밀도가 줄어들어 prediction이 정확하지 않은 추세를 보이는 것으로 예상된다. 따라서 PINN에서도 데이터 양의 영향을 적게 받는다고는 하지만, 적절한 데이터가 있어야 좋은 성능을 보일 수 있는 것을 알 수 있었다. 그림 3은 time domain을 0에서 2로 설정하였을 경우에 나타난 차트이고, 그림 4는 0에서 5로 설정하였을 경우, 그림 5는 0에서 10으로 설정하였을 경우 나타난 차트다.

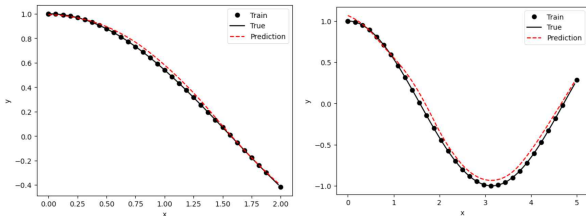


그림 3 time domain을 0에서 2로 설정하였을 경우 그림 4 time domain을 0에서 5로 설정하였을 경우

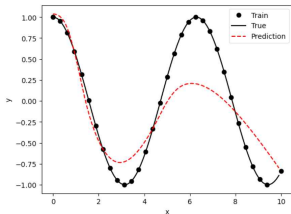


그림 5 time domain을 0에서 10으로 설정하였을 경우

두 번째 예시로 다차식의 샘플인 $x^{10} - x^9 + x^8$ 을 학습시켰을 때도 좋은 성능을 보였다. 신경망 구성은 입력을 하나의 노드로 받고, 16개, 8개, 4개, 2개로 노드를 이후에 구성하였으며, 마지막 출력 노드는 1개이다. 활성화

함수는 elu를 이용하였으며, optimizer는 adam, 초기 학습률은 0.01, epoch은 50으로 임의로 지정하였다. train loss와 test loss는 5.51e-03, 3.18e-03로 나타났다. 학습 후 나타난 차트는 그림 6과 같으며, time domain은 0에서 2로 설정하였다.

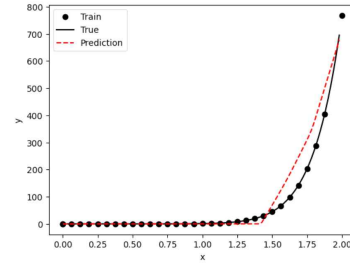


그림 6 다차식 학습 후 나타난 차트

하지만 제시한 두 예시들과 달리, 계산이 아주 복잡한 물리 공식은 코드로 변환하는 과정이 다소 어렵거나 학습이 잘 진행되지 않는 경우도 있었다.

5. 향후 연구

PINN은 강력한 알고리즘으로 현재 주목 받고 있다. 하지만 계산이 아주 복잡한 물리 공식은 코드로 변환하는 과정이 다소 어렵거나 학습이 잘 진행되지 않는 경우가 있어, 보다 효율적인 처리에 대한 연구를 진행할 예정이다. 또한 다양한 상황에서 PINN을 적용해보는 연구도 진행할 예정이다.

참 고 문 헌

- [1] M. Raissi, P. Perdikaris, G.E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, Volume 378, pp. 686-707, 2019.
- [2] Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, George Em Karniadakis, "Physics-informed neural networks (PINNs) for fluid mechanics: a review," *Acta Mechanica Sinica* 37, pp.1727-1738, 2021.
- [3] 최현조, 신명현, 박종호, 박진수, "물리 기반 인공지능 경망을 이용한 PIV용 합성 입자이미지 생성," *한국가시화정보학회지* Vol.21 No.1, pp.119-126, 2023.
- [4] 이창환, 예인수, 최호, "Digital Transformation을 위한 PINN(Physics-informed Neural Network)의 효용성 및 Feasibility Test," *대한기계학회 CAE 및 응용역학 부문 2022년도 춘계학술대회*, pp.69-70, 2022.
- [5] 이한별, 하중문, 이형진, 최원재, 오현석, "1차원 파동 방정식 풀이를 위한 물리지식기반 인공지능경망 개선," *대한기계학회 2022년 학술대회*, pp.1616-1618,

2022.

- [6] Seho Son, Hyunseung Lee, Dayeon Jeong, Kyung Ho Sun, Ki-Yong Oh, “A Physics-Informed Neural Network for Modeling Electromagnetism of an Electric Motor,” CAE 및 응용역학부문 2023년 춘계학술대회, pp.104-105. 2023.
- [7] 김승욱, 이승철, “물리기반 인공지능망을 활용한 리튬이온 전지 열화 예측,” 한국소음진동공학회 2020년도 추계 학술대회 초록집, pp.229, 2020.

LSTM과 XGBoost를 이용한 전력 예측에서의 ANN 모델 성능 비교 — 다중 변수 추출 시 데이터분석의 중요성

김다영^{0,1}, 석혜원¹, 송득모¹, 정설영¹, 김진해²

경북대학교¹, 위즈옥션²

rlaekdud@knu.ac.kr, hws2008@knu.ac.kr, deukmosong@naver.com,
snowflower@knu.ac.kr, linux33@daum.net

Comparison of ANN model performance in power prediction using LSTM and XGBoost — Importance of data analysis when extracting multiple variables

Dayeong Kim^{0,1}, Hyewon Seok¹, Deukmo Song¹, Seolyeong Jeong¹, Jinhae Kim²
Kyungpook National University¹, Wizauction²

요 약

전력 사용량 예측을 통한 실사용량 감소는 에너지 효율성과 지속 발전 가능성 면에서 중요하다. 이를 위해서는 정확한 전력량 예측이 동반되어야 한다. 본 논문에서는 여러 건물에서 수집된 전력 데이터 세트를 사용하여 주요한 특징을 추출해 학습하는 LSTM과 시계열 데이터 예측에 주로 사용되었던 XGBoost의 예측 성능을 비교 분석하고 건물의 에너지 효율을 위한 단기 전력 소비량을 예측하는 모델을 제안한다.

1. 서 론

현대 사회에서 에너지 효율성과 지속 가능성은 급속한 전력 수요 증가로 인해 더욱 중요해지고 있다. 건물 단위의 전력 사용량 예측은 전력 공급 및 관리의 핵심 요소 중 하나로, 전력 공급의 안정성과 비용 효율성을 향상시키는 데 결정적인 역할을 한다. 최근에는 딥러닝과 기계 학습 기술의 발전으로 건물 단위의 전력 사용량 예측에 관한 연구가 확장되고 있다. 이러한 연구는 전력 사용량을 감소 시키고자 하는 이들에게 도구로써 제공되어 에너지 효율성을 향상시키고 지속 가능한 에너지 사용을 촉진하는 데 기여할 수 있다.

본 연구에서는 예측 모델로 LSTM과 XGBoost를 활용하여 건물 단위의 전력 사용량 예측 성능을 비교하고자 한다. LSTM은 기존의 LSTM에 상관 관계 분석을 더한 모델로, 상관 계수를 통한 주요한 특징 변수를 추출해 성능을 향상시키고자 했다. XGBoost는 그라디언트 부스팅 기반의 앙상블 학습 모델로, 시계열 데이터를 통한 예측 면에서 강력한 성능을 보여 준다. 두 모델 중 어떤 모델의 성능이 전력 수요 예측에 있어서 더 뛰어난지 학습 데이터와 검증 데이터로 분리하고, 검증 데이터 예측을 통해 성능을 비교한다.

2. 데이터 셋 및 성능 지표

본 논문은 60개의 건물에서 1시간 단위로 수집된 전력 소비량 데이터 세트를 사용하였다. 데이터에는 건물 번호, 날짜 및 시간, 전력 사용량, 기온, 풍속, 습도, 강수량 등

총 10개의 변수가 포함되어 있다. 데이터 세트는 2020년 6월 1일부터 2020년 8월 24일까지의 2개월간의 전력 사용량 데이터를 종속 변수로 설정했고, 데이터 셋을 학습 데이터와 검증 데이터로 나누어 검증을 진행했다.

LSTM과 XGBoost의 정량적 성능 평가를 위해 성능 평가 지표로는 선행 연구들에서 일반적으로 사용되는 MAE와 MAPE를 사용하였다.[1]

MAE(Mean Absolute Error)는 예측값에서 실제 값의 차를 절댓값의 평균으로 한 값을 뜻한다. 직관적이고 평균화된 예측 오차를 제공해 모델 간의 비교에 적합하지만, 이상치에 둔감하다는 단점이 있어 이를 보완하고자 MAPE 값도 성능 지표로 사용한다.

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x|$$

수식 1 MAE 계산식

MAPE(Mean Absolute Percentage Error)는 예측값 대비 오차가 차지하는 정도를 나타내는 지표로, 실제 값과 예측값의 차이를 실제 값으로 나눈 것이다. 오차 평균의 크기가 작을 수록 좋은 모델이라 평가할 수 있다.

$$MAPE = \frac{\sum \frac{|A - F|}{A} \times 100}{N}$$

수식 2 MAPE 계산식

3. 데이터 전처리

데이터 셋이 시계열 데이터이기 때문에, 이를 회귀 문제로 변환하기 위해 시간 관련 변수(월, 주, 요일)를 추가했다. 그 외에도 전력 소비량의 시간대별 평균, 표준편차 등의 변수를 추가했다. 공휴일 여부가 전력 사용량에 큰 영향을 미친다고 판단해 이를 변수로 추가했다. 건물별 모델 생성 시 무의미한 태양광 발전 시설 변수를 삭제했다. 또한 온도와 습도라는 데이터가 있는 것을 이용하여 불쾌 지수라는 새로운 변수도 생성하였다. 불쾌 지수의 계산은 수식 3을 이용하였다.

$$\text{불쾌지수} = 1.8T - 0.55(1 - RH)(1.8 - 26) + 32$$

T: 기온(°C), RH: 상대습도(%)

수식 3 불쾌지수 계산식

4. 전력 수요 예측

4.1. LSTM 기반의 전력 수요 예측

4.1.1 건물 별 전력 사용량과 타 변수들의 상관 관계 분석

데이터 상관관계표는 데이터 셋 내의 변수 간의 관계 파악에 있어서 유의미한 정보를 주기 때문에 예측 모델의 데이터 전처리 과정에 있어서 중요한 역할을 한다.

따라서 건물 별 전력 사용량과 각 변수들의 상관 관계가 다를 것이라는 가설을 가지고 상관 계수를 계산해 시각화 하였다. 그림 1과 같이 결과를 확인할 수 있는데, 실제로 건물별로 전력 사용량과 변수들의 상관 계수가 다를 수 있었다. 예를 들면, 1번 건물의 경우에는 전력 사용량과 기온이 강한 양의 상관 관계를 가진다. 하지만 9번 건물의 경우에는 전력 사용량과 기온의 경우 상관 관계가 거의 없고, 일조와는 음의 상관 관계를 가지는 것을 볼 수 있다. 각 건물별로 전력 사용량과의 상관 관계가 0.3 이상인 변수를 이용해 학습 및 예측을 진행하였다.

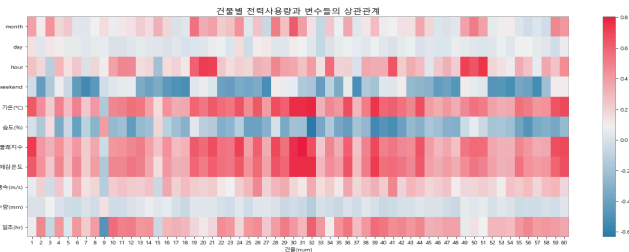


그림 1 건물별 전력사용량과 변수간의 상관 관계표

4.1.2. 하이퍼 파라미터 최적화

LSTM 모델을 이용한 단기 전력 예측 선행 연구의 모델을 참고해 하이퍼 파라미터를 결정하였다.[2][3] 최적화 기법으로 Adam을 사용했으며, 손실 함수로는 SMAPE를 사용했다. 활성 함수로는 tanh을 사용했다.

ReLU를 사용하지 않은 이유는 RNN이 CNN과는 다르게 이전 step의 값을 가져와 사용하기 때문이다. ReLU를 쓰게 되면 이전 값이 커짐에 따라 전체적인 출력이 발산하는 문제가 생길 수 있다. 따라서 과거의 값들을 재귀적으로 사용하는 RNN 모델에서는 이를 정규화 하는 것이 필요하다. 이를 위해 sigmoid보다 기울기의 역전파가 더 잘되는 tanh을 사용했다.

4.2. XGBoost 기반의 전력 수요 예측

4.2.1. 하이퍼 파라미터 최적화

60개의 건물들이 각자 다른 특성을 가짐에 따라 각각의 건물에 대해 XGBoost 모델을 사용하지만 개별 파라미터 값들을 다르게 하여 60개의 모델을 생성하였다. 다른 파라미터를 고정하지 않고, 전체 파라미터를 튜닝하고자 하면 시간이 오래 걸린다는 단점이 있다. 그래서 모델 내의 하이퍼 파라미터들은 sklearn의 gridsearchCV를 활용해 튜닝했다. 이후 표1과 같이 XGBoost의 early stopping 기능으로 n_estimators를 튜닝하고, 손실 함수로 사용한 weighted_mse의 alpha값을 튜닝했다.[4]

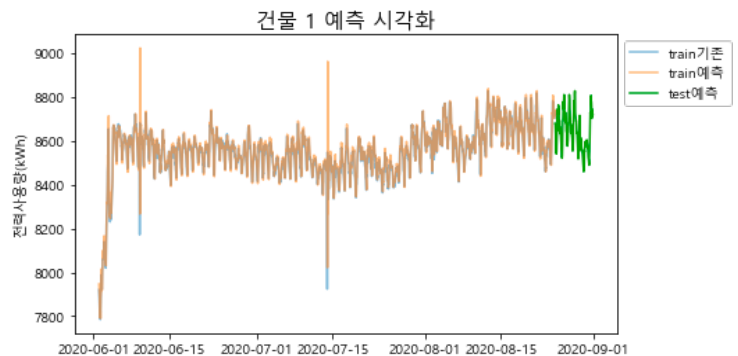
표 1 건물별 하이퍼 파라미터 튜닝 값

건물 번호	n_estimators	alpha
1	100	100
2	100	100
3	100	1
4	100	0
5	100	25

5. 전력 수요 예측 결과 및 평가

5.1. LSTM의 예측 결과

그림 2의 결과를 보았을 때, train set에 대해 기존의 데이터와 예측 값이 대체로 겹치는 양상을 보인다. 따라서 학습이 제대로 되었다는 것을 알 수 있다. 초록색으로 표시된 부분은 미래의 전력 수요를 예측한 것이다. 미래의 전력 수요를 예측한 값이 기존과 유사한 패턴으로 진행됨을 확인할 수 있다.



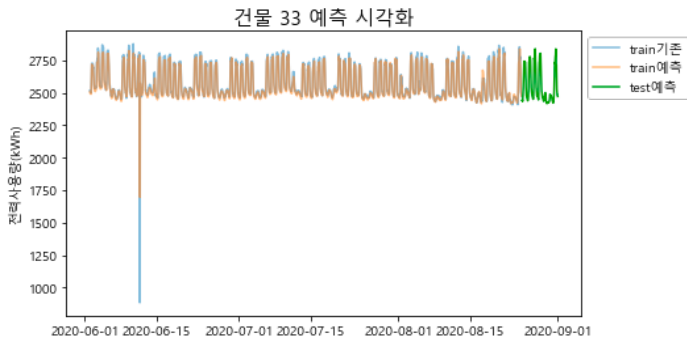


그림 2 LSTM 모델에서의 예측 결과 시각화

5.2. XGBoost의 예측 결과

아래 그림 3은 서로 다른 건물의 전력 수요 예측을 각각 나타낸 그림이다. 기존의 데이터는 파란색으로, 예측한 데이터는 초록색으로 표기했다. 예측 구간의 실제 전력 사용량은 주황색으로 표기했는데, 대부분 맞지만 일부 부분에서는 오차가 발생했다. 하지만 전체적인 패턴을 보았을 때, 모델이 예측한 전력 사용량이 이전의 실사용량과 비슷한 양상을 보이는 것을 확인할 수 있다.

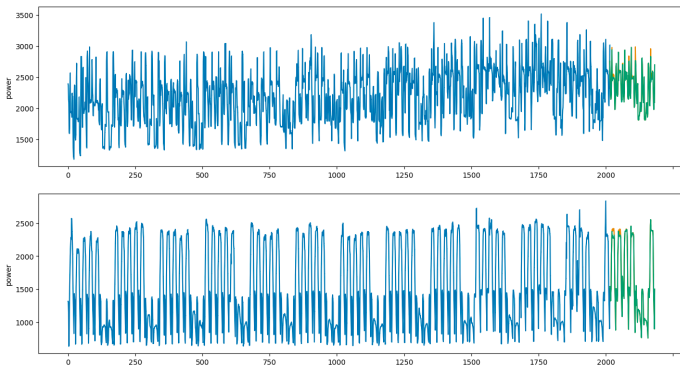


그림 3 XGBoost 모델에서의 예측 결과 시각화

5.3. 성능 평가

성능 평가 결과를 표기한 표 2에서 볼 수 있듯, LSTM 모델의 MAE 값은 102.3258, MAPE 값은 0.0279으로 XGBoost보다 MAE는 38.4799만큼, MAPE는 0.0111만큼 더 낮게 나온 것을 확인할 수 있었다. 이를 통해 LSTM 모델이 더 높은 예측 정확도를 보이는 것을 확인할 수 있었다.

표 2 두 모델의 성능 평가 결과

	LSTM	XGBoost
MAE	102.3258	140.8057
MAPE	0.0279	0.0390

6. 결 론

본 논문에서는 60개의 건물에서 1시간 단위로 수집된 약 2개월 간의 전력 사용량 데이터 세트를 사용해 LSTM과 XGBoost 모델의 예측 성능을 비교하는 연구를

진행하였다. 위의 성능 평가에서 알 수 있듯, LSTM과 XGBoost 모델 모두 좋은 성능을 보여주었다. 하지만 근소한 차이로 LSTM이 더 좋은 성능을 보여줬는데, 이는 XGBoost가 시계열 데이터 예측에 있어서 LSTM보다 높은 성능을 보여준다는 선행 연구들의 결과와는 상반되는 것을 볼 수 있다. 실험 결과를 미루어 보았을 때, LSTM 모델도 데이터 전처리 및 상관 관계표 도입을 통해 충분한 성능을 끌어낼 수 있다는 것을 보여준다. 또한, 전력 시계열 데이터 예측에 LSTM이 활용될 수 있음을 알 수 있었다.

하지만 모든 시계열 데이터에서가 아닌, 연구에 사용한 데이터 세트와 모델에 한해서만 그런 결과를 보여 준다는 점에서 'LSTM이 XGBoost보다 시계열 데이터 학습에 뛰어난 성능을 보인다'고 말할 수는 없다. 또, MAPE의 차이가 0.0111로 매우 작다는 점을 보아, 연구에서 사용한 LSTM과 XGBoost 모델의 성능이 근소한 차이를 보임을 알 수 있다. 이는 데이터 전처리 과정이나 학습 방식에 따라 극복할 수 있는 차이로, 데이터 전처리 과정과 상관 관계표 등의 데이터 분석의 중요성에 대해 알 수 있다.

추후 연구에서는 가족 구성원, 가구 면적 등의 상황 변수를 고려하고 산업용 건물에 대한 예측이 아닌, 거주지의 전력 사용량 단기 예측 모델을 만들면 공공의 이익으로 연결되어 그 활용도를 더욱 높일 수 있을 것으로 기대된다.

7. 참고 문헌

- [1] 최은선, SomAkhamixay Oui, 전유정, 김진실, 김재성, XGBoost와 NGBost를 이용한 전력데이터 예측 성능 비교에 관한 연구, 한국통신학회 학술대회논문집, 2020년도 한국통신학회 추계종합학술발표회 논문집, pp. 348-349, 2020.
- [2] 엄호용, 유대현, 백승목, 전력수요량 예측을 위한 LSTM 학습 방법에 따른 성능 비교 연구, 한국정밀공학회 2021년도 추계학술대회 논문집, 618-619, 2021
- [3] 김지은, 천관호, LSTM을 활용한 단기 전력수요 예측 기법, 2023년도 대한전기학회 하계학술대회 논문집, 1,450-1,451, 2023
- [4] 오재영, 함도현, 이용건, 김기백. XGBoost 기법을 이용한 단기 전력 수요 예측 및 하이퍼파라미터 변화에 따른 영향 분석, 전기학회논문지 제68권 제9호, pp. 1073-1078, 2019.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

MmWave 레이더 기반 일상 행동 인식을 위한 경량화 딥러닝 시스템의 설계 및 구현

신기훈[○], 임재봉 김태구, 조용훈, 정혜선, 신채림, *백윤주

부산대학교

(skh2929209, jaebonglim, tbg8577, kchoyh95, ahttjd55,

cofla0429)@pudan.ac.kr, *yunju@pusan.ac.kr

Design and Implementation Light-weight Deep Learning System for Human Action Recognition based on mmWave Radar

Shin Ki Hun[○], Lim Jae Bong, Kim Tae Gu, Cho Yong Hun, Jeong Hye Sun, Shin Chae

Rim, *Back Yun Ju

Pusan National Univ.

요 약

이 논문은 mmWave 레이더 데이터를 기반으로 한 일상 행동 인식(HAR)을 위한 경량 딥러닝 모델의 성능 향상을 지식 증류 기법을 사용한다. 이 연구는 지식 증류 적용 전후로 다양한 크기와 입력 길이를 가진 모델의 성능 변화를 평가하는 데 중점을 둔다. 본 연구는 mmWave에서 포인트 클라우드 데이터를 활용하는 경량 딥러닝 시스템의 설계 및 구현을 포함한다. 이 방법론은 향상된 표현력과 정확성을 위해 인간 관절 점을 사용하여 큰 크기의 teacher 모델을 학습시키는 것으로 시작한다. 이후 student 모델을 teacher 모델로부터 가지 치기와 지식 증류를 통해 학습한다. 학생 모델은 포인트 클라우드 데이터만을 사용하며, 효율적인 모델 구조에 최적화된 단일 모달 시스템을 대표한다. 모델의 크기와 입력 길이 측면에서 다양한 모델을 평가하며, 작은 모델과 짧은 입력 길이를 가진 모델이 성능이 낮아지는 경향을 보여 준다. 그러나 지식 증류 적용은 모든 모델에서 정확도를 크게 향상시켜, 최소 0.1%에서 7% 이상의 개선을 이루었다. 제안된 손실 함수는 특징 유사도 차이와 모델 예측 불일치를 결합하여 가장 좋은 성능 향상을 보여주었습니다.

1. 서 론

인간 행동 인식(Human Action Recognition, HAR)은 인간이 어떤 행동을 함에 있어 다양한 센서를 통해 이를 측정하고 나온 데이터를 이용하여 해당 행동을 인식하는 시스템이다. 이러한 HAR은 최근 몇 년간 스마트 홈, 스포츠 분석, 인간-컴퓨터 상호작용 등을 포함한 다양한 응용 분야에서 광범위한 응용과 함께 핵심적인 연구 분야로 부상하고 있으며 사용자의 환경에 따라 여러 가지 형태로 사용이 될 수 있다. 독거 노인 가정 이상 탐지는 가정 내 환경에서, 작업자 이상 상태 감지는 기업 내에서 사용될 수 있다.

전통적인 HAR 접근 방식은 크게 두가지로 구분할 수 있다. 접촉식 센서와 비접촉식 센서로 구분 가능하며 접촉식 센서는 웨어러블 센서[1], 비접촉식 센서는 다시 비전[2]과 RF[3]로 구분할 수 있다. 정교한 센서 기술과 웨어러블 장치의 보급은 정교한 시스템을 포함한 대

체 데이터 소스를 제공하며, 행동 인식을 위한 풍부하고 다차원의 데이터를 제공한다. 하지만 이 중 비전 기반의 데이터 측정 방식은 높은 계산 비용, 조명 및 가시적인 환경적 요인에 대한 민감성, 비디오 감시와 관련된 개인 정보 보호 문제와 같은 한계를 가지고 있다.

이에 30GHz에서 300GHz의 주파수 범위에서 작동하는 mmWave 기술은 유망한 대안을 제공한다. MmWave 센서는 움직임과 물체의 속도 및 각도를 포착하는데 사용되며 다양한 분야의 활용된다. 고주파수의 높은 대역폭을 가지는 mmWave는 가시적 환경에 덜 민감하고 비침습적이다. 이에 mmWave를 활용한 다양한 HAR 연구가 활발히 진행 중이다.

이와 더불어 딥러닝을 통해 mmWave의 고차원 데이터를 효율적으로 처리하여 HAR 분야를 놀라운 정확도를 가진 모델과 데이터에서 복잡한 패턴을 학습할 수 있는 능력을 제공한다. 하지만 전통적인 딥러닝 모델은 많은 계산 복잡성과 높은 자원 요구 사항이 존재한다.

또한 HAR을 위해 사람의 관절 포인트를 예측 후 행동을 인식하는 2 stage 방식[4], 추론을 위해 관절 포인트와 point cloud 데이터를 동시에 이용하는 멀티 모달 모델[5] 등 딥러닝의 크기를 고려하지 않은 연구가 대부분이다. 이러한 딥러닝 모델은 제한된 엣지 디바이스에 배치하여 동작하기엔 적합하지 않다. 높은 정확도를 유지하면서도 효율적인 경량 딥러닝 모델의 개발을 필요로 한다.

이 논문은 mmWave의 point cloud 데이터를 활용한 경량 딥러닝 시스템에 대해 제안한다. 우리는 사람의 관절 포인트를 활용한 충분한 크기의 딥러닝 모델을 사전 학습한 후 이를 이용하여 경량 딥러닝 모델을 학습한다. 이는 추론 시에 point cloud만 이용하는 싱글 모달의 시스템으로 효율적인 모델 구조를 가진다. Graph Convolution Network(GCN) 기반의 모델을 활용하여 가지치기, 지식 증류를 통해 모델의 크기를 효과적으로 줄여 엣지 디바이스에서 추론할 수 있는 시스템을 제공한다.

2. 본 론

본 논문은 point cloud를 이용하여 HAR을 하기 위해 상용 mmWave 디바이스를 통해 측정된 데이터를 사용하여 행동 식별 모델을 제안한다. 7가지의 행동에 대해 측정된 데이터와 GCN 기반 모델 학습 방법에 대해 소개한다.

2.1 데이터셋

본 연구에 사용된 데이터 세트는 오픈 데이터 세트를 활용하였다. DGHUA[5]는 동국대학교에서 측정된 데이터 세트로 평균 나이 23세로 이루어진 총 19명의 참가자로 측정되었다. 측정 디바이스는 TI사의 IWR1843BOOST와 Kinect를 사용하였고 각 디바이스를 통해 point cloud와 관절 포인트를 수집하였다. 측정 자세로는 총 7가지의 자세에 대해 존재한다. 걷기, 점프, 앉았다 일어나기, 양팔 벌리기, 오른팔과 오른발 올리기, 왼팔과 왼발 올리기, 앞으로 넘어지기로 구성되어 있다. 각 동작은 20초 동안 측정되었다. 레임으로 구성되어 있고 5~6회 반복 측정하였다. 한 프레임마다 총 25개의 포인트가 구성되어 있고 이에 (3, 400, 25)의 데이터 형태로 존재한다.

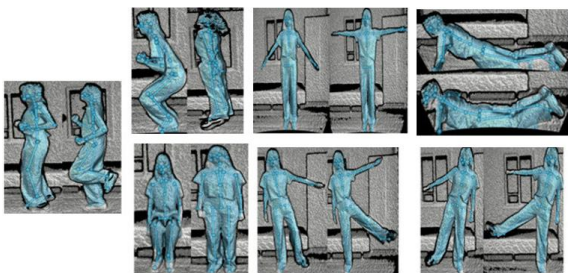


그림 1. 데이터 세트 동작 예시

2.1 딥러닝 모델 구현

제안하는 딥러닝 모델은 포인트간 관계를 효율적으로 학습하기 위해 GCN 기반의 HAR 모델이다. HAR 분야에서 포인트 간 관계성 파악에 유리해 좋은 성능을 보여주는 GCN은 관절과 point cloud 사이의 데이터 분석에도 마찬가지로 좋은 성능을 보여준다. 또한 각 포인트간 공간적, 시간적 관계성을 학습하기 위한 Spatial-Temporal GCN(ST-GCN) 구조를 이용하였다.

GCN 구조의 특성 상 feature extraction 시 많은 양의 메모리를 차지한다. 따라서 20초의 길이를 한 번에 학습에 이용할 시 연산에 필요한 메모리 양의 문제로 제대로 된 학습이 되지 않는다. 이는 이후 엣지 디바이스에서 추론에도 영향을 미친다. 따라서 10초, 5초, 2.5초로 데이터를 나누어 다양한 모델을 학습한다.

본 모델은 사전 학습된 큰 사이즈의 teacher 모델로 지식 증류를 이용하여 student 모델을 학습한다. Teacher 모델의 경우 표현력과 정확도가 높이기 위해 충분한 사이즈의 모델로 학습하였다. 이후 student 모델 학습은 가지 치기를 통해 모델의 사이즈를 감소시킨 후 teacher 모델로부터 지식 증류를 통해 성능 향상을 하였다. Teacher 모델의 경우 입력으로 관절 포인트를 이용한다. 이는 관절 포인트가 point cloud에 비해 표현의 정확도가 높다. 따라 보다 많은 정보를 teacher 모델에 학습시킨다. Student 모델은 point cloud만을 이용하여 학습하여 이후 추론 시에도 해당 데이터만을 이용하는 싱글 모달 시스템을 구현했다.

2.2 손실 함수

일반적인 지식 증류의 경우 두 모델 간 target loss간 차이만을 줄이기 위해 손실 함수를 설계한다. 하지만 이는 서로 다른 데이터로 학습된 teacher와 student 모델 사이의 지식 증류 시 성능 향상에 제한이 있다. 따라서 이를 해결하기 위해 조합된 손실 함수를 사용했다. 서로 다른 입력에 대해 추출된 feature가 유사하게 학습될 수 있도록 feature간 유사도 차이와 각 모델이 예측하는 결과 사이의 차이를 조합하여 모델의 손실 함수로 사용하였다.

3. 실험

본 논문에서 제안하는 HAR 모델을 학습하고 구성하기 위해 입력 데이터로 teacher 모델 학습 시엔 관절 포인트를 사용하였다. 20fps의 데이터를 전부 사용하지 않고 10초 단위로 크기로 윈도우를 만들어 학습에 사용하였다. 동일한 동작에 대해 슬라이딩 윈도우 적용 시 과적합 될 수 있어 앞과 뒤 데이터를 5초씩 떼어내는 것으로 10초를 구성하였다.

이후 student 모델에는 point cloud 데이터를 입력으로 사용하여 학습하였다. 입력 데이터 크기를 2.5초, 5초 그리고 10초로 총 3가지의 크기를 가지는 모델을 학습하였다. 또한 서로 다른 크기를 가지는 5개의 모델을

성능 비교를 위해 학습하였다. 평가 시에는 400 프레임의 데이터 모든 부분에 있어 성능 평가를 하기 위해 10 프레임씩 오버랩하여 입력 시간에 맞게 윈도우를 구성하여 사용하였다.

제안하는 손실 함수 평가를 위해 일반적[6]로 지식 증류시에 사용되는 방식인 target loss만을 사용했을 때와 feature 간 유사도 차이만을 사용했을 때 이 둘을 조합한 것과 지식 증류 평가를 위해 적용하지 않은 것까지 총 4개의 모델을 학습시켰다.

4. 결과

먼저 teacher 모델의 경우 크기응 약 380K개의 파라미터를 가지며 정확도는 98%로 학습되었다. 이와 동일한 사이즈를 가지고 입력으로 point cloud를 이용한 모델은 94%의 정확도를 보여주었다.

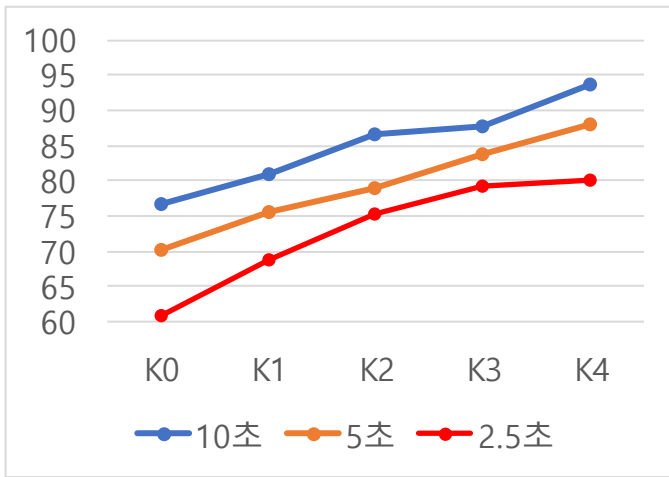


그림 2. 지식 증류 전 모델 정확도

4.1 지식 증류 전 모델 성능 평가

먼저 지식 증류를 하기 전 여러 가지 크기와 서로 다른 입력 길이를 가지는 모델을 학습하였다. 5개의 K급 크기에 대해 성능은 다음 그림 2와 같다. 모델의 크기가 작아질수록 입력 길이가 짧아 질수록 모델의 성능이 작아지는 것을 확인할 수 있다. 입력 길이가 제일 긴 10초에서 약 110K 크기의 모델의 성능 정확도가 약 93%로 제일 좋은 것을 확인할 수 있다.

4.2 지식 증류 후 모델 성능 평가

다음은 98%의 정확도를 가지는 teacher 모델을 이용하여 지식 증류를 하였을 때의 성능평가를 위해 실험하였다. 모든 모델 정확도 성능에서 작게는 0.1%에서 크게는 7% 넘게 성능 향상을 이끌어 내었다. 이는 모델의 크기가 줄어도 teacher 모델을 통해 성능을 향상할 수 있다는 것을 의미한다. 제일 좋은 성능 향상을 보여준 모델은 K3모델이고 이는 teacher 모델에 비해 약 1/16 배 줄었지만 성능은 약 12%p가 감소하였다.

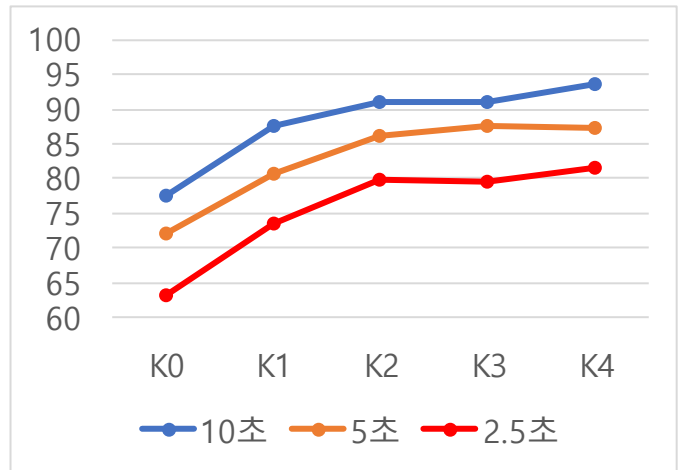


그림 3. 지식 증류 후 모델 정확도

4.3 손실 함수 평가

3가지의 손실 함수에 대해 K1, K2 크기에 대해 성능 평가를 나타내었다. Baseline 모델의 아무 기법도 적용하지 않은 모델의 정확도이다. Hidden layer는 feature 유사도 차이만을 지식 증류한 모델의 정확도이다. Probability는 모델의 예측한 target loss 차이만을 지식 증류한 모델의 정확도이다. 마지막으로 Combine은 모든 기법을 모두 적용한 모델의 정확도이다. 모두 적용하지 않았을 때 기법과 모델의 따라 오히려 성능 하락을 보여주는 경우가 존재하였다. 하지만 제안하는 손실 함수를 적용하였을 때는 가장 좋은 성능 향상을 보여주었다.

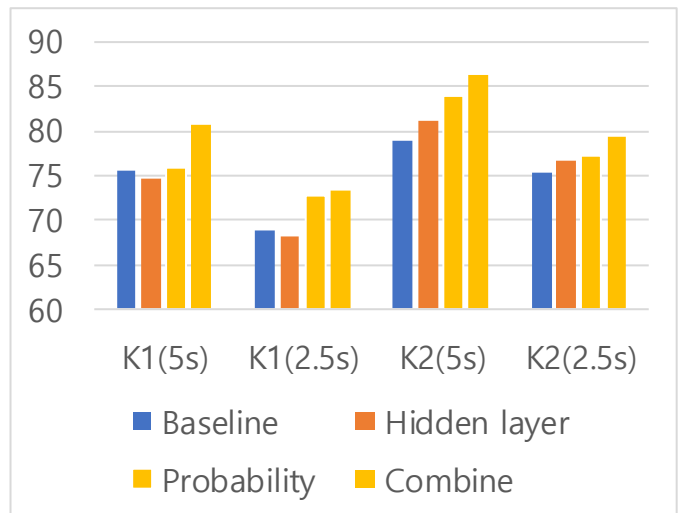


그림 4. 손실 함수 별 모델 정확도

4 결론

본 논문에서는 지식 증류 기법을 적용한 경량화 딥러닝 모델의 성능 향상에 대한 연구를 수행하였다. 다양한

크기와 입력 길이를 가진 모델을 대상으로 지식 증류 전후의 성능 변화를 평가하였다.

지식 증류를 적용하지 않은 상태에서는, 모델 크기가 작아지고 입력 길이가 짧아질수록 성능이 감소하는 경향을 확인하였다. 이에 지식 증류를 적용하여 모든 모델에서 정확도가 최소 0.1%에서 최대 7% 이상 향상되었다. 특히 K3 모델의 경우 크기는 약 1/16배 감소하였음에도 불구하고 성능은 12%p 감소하는데 그쳤다. 이는 지식 증류가 모델 크기에 비례하여 성능을 유지하거나 향상시킬 수 있음을 시사한다.

손실 함수에 대한 평가에서는, baseline 모델, feature 유사도, target loss 차이만을 지식 증류한 모델, 그리고 이 모든 기법을 적용한 모델에 대해 비교하였다. 제안된 손실 함수를 적용하였을 때 가장 좋은 성능 향상을 확인할 수 있었다.

향후 연구로는, 경량화된 모델을 엣지 디바이스에서 동작 가능하게 모델 구조 변화를 통해 메모리 사용 효율성을 증가하고, 이후 추론 시간 및 추론 정확도 향상 방안에 대해 연구한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2023-RS-2023-00260098)

이 연구는 2023년도 산업통상자원부 및 한국산업기술기술평가원(KEIT) 연구비 지원에 의한 연구임 (20015052)

참고 문헌

- [1] EK, Sannara, François Portet, and Philippe Lalanda. "Lightweight Transformers for Human Activity Recognition on Mobile Devices." arXiv preprint arXiv:2209.11750 (2022).
- [2] Chen, I-Chien, et al. "Multi-person pose estimation using thermal images." IEEE Access 8 (2020): 174964-174971.
- [3] Wang, Congming, Xiaohui Zhao, and Zan Li. "DCS-CTN: Subtle Gesture Recognition based on TD-CNN-Transformer via Millimeter Wave Radar." IEEE Internet of Things Journal (2023).
- [4] Lee, Gawon, and Jihie Kim. "Improving human activity recognition for sparse radar point clouds: A graph neural network model with pre-trained 3d human-joint coordinates." Applied Sciences 12.4 (2022): 2168.
- [5] Lee, Gawon, and Jihie Kim. "MTGEA: A Multimodal Two-Stream GNN Framework for Efficient Point Cloud and Skeleton Data Alignment." Sensors 23.5 (2023): 2787.
- [6] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

NR-V2X Mode2에서 강화 학습 기반

적응형 Modulation Order 선택 기법

조의리^o, 말릭 사드 무하마드, 박세리, 김동균
경북대학교 IT대학 컴퓨터학부

joeuri0402@knu.ac.kr, maliksaad@knu.ac.kr, psr0527@knu.ac.kr, dongkyun@knu.ac.kr

Reinforcement Learning based Adaptive Modulation Order Selection in NR-V2X Mode 2

Euri Jo^o, Malik Muhammad Saad, Seri Park, Dongkyun Kim
School of Computer Science and Engineering, Kyungpook National University

요 약

The fifth generation (5G) mobile communication network revolutionizes connectivity with its capabilities boasting ultra-high speeds, extensive connectivity, reliable communication, and remarkably low latency. Among its transformative features is NR-V2X communication, empowering vehicles to interact seamlessly with each other and their surroundings. The NR-V2X framework operates in two modes, with Mode 2 facilitating direct vehicle-to-vehicle interactions. This mode proves vital in scenarios where instantaneous communication is crucial, such as collision avoidance and traffic optimization. However, the efficient operation of vehicular networks necessitates a delicate balance between high transmission rates and the imperative of avoiding network overload. Addressing this challenge, this paper introduces a novel approach utilizing Reinforcement Learning (RL) as a congestion control mechanism. Specifically, the proposed RL scheme aims to optimize the modulation order, thereby enhancing the overall efficiency of the vehicular communication network. By leveraging RL, the system dynamically adapts to changing network conditions, striking a balance between the need for high data rates and the imperative of congestion management. This research contributes to the advancement of robust congestion control mechanisms in NR-V2X communication, ensuring the seamless integration of high-speed vehicular networks into the broader 5G landscape.

1. Introduction

The advent of the fifth generation (5G) mobile communication network marks a revolutionary stride forward, poised to transform the very fabric of our daily lives. 5G brings with it a suite of capabilities, including ultra-high speeds, expansive connectivity, reliable communication, and remarkably low latency, addressing the imperatives of our hyper-connected world. These features collectively promise not only faster internet for smartphones but also lay the groundwork for groundbreaking applications that will reshape industries and societies [1].

Among the myriad applications ushered in by 5G, one of the most transformative is New Radio Vehicle-to-Everything (NR-V2X) communication. At its core, NR-V2X empowers vehicles to communicate

seamlessly with each other and their surrounding environment, forging a fully interconnected transportation ecosystem. This innovation holds the key to elevating road safety, enabling the realization of autonomous driving, optimizing traffic management, and fostering the real-time exchange of critical information. As 5G networks unfold, NR-V2X stands as a testament to the potential of this technology to revolutionize the dynamics of transportation and pave the way for a smarter, more efficient future [2].

Within the NR-V2X framework, two distinct communication modes, Mode 1 and Mode 2, play pivotal roles in shaping the efficiency and effectiveness of vehicular networks. Mode 2, characterized by decentralized communication, facilitates direct interactions between vehicles. This proves particularly advantageous in scenarios where swift and immediate communication is paramount, such as in collision

avoidance and traffic optimization. However, the dynamic and decentralized nature of vehicular networks presents a challenge in effectively managing resources within NR-V2X Mode 2. Consequently, the need arises for robust congestion control mechanisms to strike a delicate balance between the demands of high transmission rates and the imperative of avoiding network overload [3].

In the remainder of this paper, related work is discussed in Section 2. Section 3 discusses the system model. Problem is defined in Section 4. Finally, the proposed scheme is discussed in Section 5 and the conclusion is drawn in Section 6.

2. Related Work

We have reviewed studies using MCS to reduce the size of messages in C-V2X and NR-V2X. Miguel Sepulcre et al. investigates the compression of V2X messages to alleviate communication channel load in vehicular networks using three compression algorithms [4]. The papers applying MCS in C-V2X are as follows. Andres Burbano-Abril et al. investigates the implementation of adapting the MCS within the C-V2X sidelink under various vehicular densities [5]. Ji-Cheng Yin et al. focus on optimizing the performance of LTE-V2V Mode 4 in C-V2X communications, specifically addressing the selection of MCS for improved transmission and reception efficiency. The study proposes vehicles to automatically choose a suitable MCS for communication [6]. Yu-Jen Ku et al. addresses the limitations of C-V2X Sidelink communication protocol. They investigate parameters, such as MCS and packet size [7].

The papers applying MCS in NR-V2X are as follows. Jin- Yan et al. addresses the challenge of determining an optimal MCS for 5G NR Sidelink in the context of V2X communication. This research aims to identify a globally optimal MCS value applicable to any scenario. [8] Shang-Huan Wu et al. focuses on improving the reliability of V2X communications by proposing a platoon-based groupcast approach using Deep Reinforcement Learning [9]. Aslihan Reyhanoglu, et al. focuses on enhancing the reliability of V2X communications by proposing a Machine Learning aided NR-V2X QoS prediction scheme [10]. Given the limited number of studies applying MCS to the transmission of CAM messages in NR-V2X broadcasting, we believe there is a need for further research on MCS selection in NR-V2X mode 2.

3. System Model

We consider a set of V vehicles, $V = \{1, 2, 3, \dots, M\}$ on the road. There are K resources denoted by $R_K = \{1, 2, 3, \dots, R_K\}$ in the network. Each resource is defined in the two-dimensional time and frequency domain, following the LTE standard, also depicted in Fig. 1. Every vehicle reserve resources from the available set in the network using semi-persistent scheduling for communication. The vehicle selects n resources for the corresponding Cooperative Awareness Message (CAM) transmission. The value of n (i.e., the number of resources) reserved for the CAM message depends on the size of data that the CAM message carries.

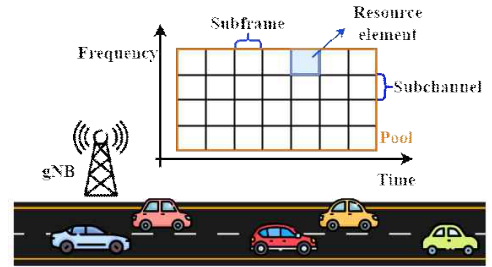


Fig. 1. System Model

3.1 CAM Structure

Fig. 2 illustrates the structure of Cooperative Awareness Messages (CAMs), which are messages exchanged within the Intelligent Transport System network. CAMs consist of mandatory and optional parts, with an average size of around 350 Bytes. However, the message size can vary (190 Bytes to 800 Bytes or potentially more) depending on the presence of optional containers. CAM messages include a Protocol Data

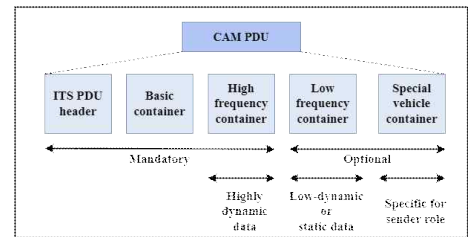


Fig. 2. CAM Packet Structure

Unit (PDU) header, a basic container, and a high-frequency container, all of which are mandatory. The header indicates the vehicle ID. The basic container contains data such as position and timestamp, while the high-frequency container includes dynamic data like heading, speed, acceleration, and curvature. Optional containers encompass the low-frequency container (vehicle role, trajectory, basic sensors) and the

special vehicle container (traffic conditions, road works, road accident) [11].

CAM message generation is aperiodic, occurring when the vehicle's position changes by more than 4m, the direction changes by $\pm 4^\circ$ or more, or the speed changes by 0.5m/s or more. Generation also happens if the time elapsed since the last CAM generation is equal to or greater than TGenCam and equal to or greater than TGenCam_Dcc. The CAM generation interval ranges from TGenCamMin = 100ms (10Hz) to TGenCamMax = 1,000 ms (1Hz).

The parameter TGenCam_Dcc is provided by the management entity, with a value range of TGenCamMin \leq TGenCam_Dcc \leq TGenCamMax. If provided with a value above TGenCamMax, TGenCam_Dcc is set to TGenCamMax. If the value is below TGenCamMin or not provided, TGenCam_Dcc is set to TGenCamMin. Note that the specified rules apply only to the mandatory part, excluding the optional part [12].

4. Problem Definition

Vehicles transmit CAM messages using resources reserved from previous transmissions. However, the variable size of CAM messages introduces situations where messages cannot be transmitted through the reserved resources if the CAM message size is substantial. In such instances, the resources already reserved and broadcasted to neighboring vehicles remain unused, causing resource wastage. Furthermore, transmitting larger-sized messages necessitates the reservation of new resources. As the number of available resources is limited, inefficient resource utilization can lead to channel congestion. When the channel becomes congested, collisions may occur between messages, resulting in improper delivery of the messages.

5. Proposed Scheme

We aim to enable communication between vehicles by utilizing reserved resources, with the reservation broadcasted to nearby vehicles. To achieve this, we adopt an approach involving modulation and coding schemes. Choosing a higher modulation order allows us to transmit a greater amount of information using the same resources. However, higher modulation orders may result in an increased bit error rate, representing the number of bit errors within a given time period caused by noise, interference, and distortion in the communication channel.

5.1 Deep Learning

To address the aforementioned issue, we propose a mechanism employing reinforcement learning to determine the optimal modulation order. Reinforcement learning is suitable for solving problems where finding an optimal value is challenging. The agent in our system is a vehicle facing a situation where the CAM (Cooperative Awareness Message) message to be sent is larger than the reserved resources.

State Each vehicle observes the state. The state includes the CAM message size and channel state information given as follows.

State
= {CAM message size, Channel State Information}

Action Based on the state observed by the vehicle i.e., agent, the action is performed. The action involves a set of modulation and coding schemes(MCS) given as follows.

Action = {2,4,6,8}

Each vehicle takes action to choose the optimal MCS.

Reward The reward is calculated based on the PDR given as below.

$$PDR = \frac{1}{N_j} \sum_{j=1}^{N_j} \eta_j \quad (1)$$

Here, N_j represents the number of vehicles around the agent vehicle. η is 1 if the vehicle has properly received the message; otherwise, it is 0. We define the Packet Delivery Rate (PDR), and our machine will be trained based on centralized training, with distributed execution in real-world situations.

6. Conclusion and Future Work

In this paper, we aim to enable communication between vehicles by utilizing reserved resources, with the reservation broadcasted to nearby vehicles. To transmit the CAM message that is larger than the reserved resources, we adopt an approach involving MCS. We propose a Reinforcement Learning scheme to determine the optimal modulation order for 5G NR V2X sidelink broadcast communication. The reward of Reinforcement Learning is calculated based on the PDR that we defined. We consider the number of vehicles around the agent vehicle, and whether the vehicle has properly received the message in this PDR calculation.

In future work, we intend to train our machine based on centralized training, with distributed execution in real-world situations. We will also investigate how to determine whether the surrounding vehicles have properly received the transmitted packet.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program(IITP-2023-RS-2022-00156389) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

REFERENCES

- [1] S. S. Husain, A. Kunz, A. Prasad, E. Pateromichelakis, and K. Samdanis, "Ultra-high reliable 5g v2x communications," *IEEE Communications Standards Magazine*, vol. 3, no. 2, pp. 46-52, 2019.
- [2] H. Bagheri, M. Noor-A-Rahim, Z. Liu, H. Lee, D. Pesch, K. Moessner, and P. Xiao, "5g nr-v2x: Toward connected and cooperative autonomous driving," *IEEE Communications Standards Magazine*, vol. 5, no. 1, pp. 48-54, 2021.
- [3] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. S. ahin, and A. Kousaridas, "A tutorial on 5g nr v2x communications," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1972-2026, 2021.
- [4] M. Sepulcre, J. Gozalvez, G. Thandavarayan, B. Coll-Perales, J. Schindler, and M. Rondinone, "On the potential of v2x message compression for vehicular networks," *IEEE Access*, vol. 8, pp. 214254- 214268, 2020.
- [5] A. Burbano-Abril, B. McCarthy, M. Lopez-Guerrero, V. Rangel, and A. O' Driscoll, "Mcs adaptation within the cellular v2x sidelink," in *2021 IEEE Conference on Standards for Communications and Networking (CSCN)*, pp. 111-117, 2021.
- [6] J.-C. Yin and S.-H. Hwang, "Variable mcs method for lte v2v mode4," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1368-1370, 2021.
- [7] Y.-J. Ku, B. Flowers, S. Thornton, S. Baidya, and S. Dey, "Adaptive c-v2x sidelink communications for vehicular applications beyond safety messages," in *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, pp. 1-6, 2022.
- [8] J. Yan and J. Harri, "Mcs analysis for 5g-nr v2x sidelink broadcast communication," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1347-1352, 2022.
- [9] S.-H. Wu, R.-H. Hwang, C.-Y. Wang, and C.-H. Chou, "Deep re- inforcement learning based resource allocation for 5g v2v groupcast communications," in *2023 International Conference on Computing, Networking and Communications (ICNC)*, pp. 1-6, 2023.
- [10] A. Reyhanoglu, E. Kar, F. E. Kumec, Y. S. C. Kara, S. Karaagac, B. Turan, and S. Coleri, "Machine learning aided nr-v2x quality of service predictions," in *2023 IEEE Vehicular Networking Conference (VNC)*, pp. 183-186, 2023.
- [11] CAR 2 CAR Communication Consortium, "Survey on ITS-G5 CAM statistics," *Tech. Rep. TR2052 V1.0.1*, December 2018.
- [12] ETSI, "Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service," *Tech. Rep. EN 302 637-2 V1.3.1*, Sep. 2014.

SSD의 쓰기 증폭 개선을 위한 머신러닝 기반 Hot/Cold 페이지 분류 기법

윤성준^o 안성용

부산대학교

yunekorea@pusan.ac.kr, sungyong.ahn@pusan.ac.kr

Machine-learning based Hot/Cold Page Classification for Reducing Write Amplification of SSDs

Sungjune Yune^o Sungyong Ahn

Pusan National University

요 약

SSD는 데이터의 덮어쓰기가 불가능하고 각 block에 대한 쓰기 횟수에 제한이 있는 단점으로 인해 새로운 데이터를 다른 곳에 쓰고 기존의 데이터를 무효화하는 방식으로 지우기를 수행하는 시점을 뒤로 미룬다.

정해진 임계점에 도달하면 Garbage Collection을 발동하여 무효화된 데이터를 정리하는 작업을 수행하는데 이 때 블록을 지우기 위해 블록 내의 데이터를 옮기는 과정에서 추가적인 쓰기(쓰기 증폭, Write Amplification)가 발생하여 SSD의 성능을 저하, 수명을 단축시킨다.

위 문제점을 해결하기 위해 무효화 될 시점이 비슷한 데이터를 예측하여 분리 저장해 유효 및 무효 데이터가 한 블록 내에 공존할 가능성을 줄이고자 하며, 이를 위해 머신 러닝 모델을 사용하여 무효화 시점을 예측하고 이를 SSD 시뮬레이터에 적용해 WA 개선의 결과를 확인하려고 한다.

1. 서 론

SSD는 빠른 입출력 속도로 현대 컴퓨터의 주요 저장장치로 쓰이지만 데이터의 덮어쓰기가 불가능하고 각 NAND block에 대한 지우기 횟수에 제한이 있어 새로운 데이터를 다른 곳에 쓰고 기존의 데이터를 무효화(invalidate) 하는 방식으로 지우기를 수행하는 시점을 뒤로 미룬다. 이렇게 무효 데이터의 수가 임계점에 도달하면 Garbage Collection(GC)을 발동하여 무효화된 데이터를 정리하고 빈 block을 확보하는 작업을 수행하는데 이 때 block을 지우기 위해 block 내의 유효 데이터를 옮기는 과정에서 쓰기 증폭(Write Amplification)이 발생하여 SSD의 성능을 저하시키고 수명을 단축시킨다.

이러한 문제점을 해결하기 위해 무효화될 시점이 비슷한 데이터를 예측하여 block별로 분리 저장하면 유효 및 무효 데이터가 한 block내에 공존할 가능성을 줄여 block을 지울 때 옮겨야 할 유효 데이터의 수가 적어질 수 있다.

기존의 연구 중에서도 각 block의 수명을 계산하여 class를 구분한 뒤 분리하여 저장하는 연구[1]가 있으며 본 논문에서는 종합적인 데이터의 무효화 시점을 예측하기 위해 머신 러닝 모델을 사용하고자 한다.

2절에서는 이번 연구의 배경을 설명한다. 3절에서는 입출력 트레이스를 이용해 ML 모델을 학습시키고 시뮬레이

터까지 적용하는 과정을 서술하고 4절에서 실험 결과에 대해 평가한 후 결론을 낸다.

2. 배경

낸드 플래시 메모리(NAND flash memory) 기반 SSD(Solid State Drive)는 빠른 I/O성능을 보여주지만 과거의 주요 저장장치였던 HDD에 비해 새로운 문제점을 가지고 있다. 낸드 플래시 메모리는 In-place update가 되지 않아 기존 데이터를 갱신하기 위해서는 지우기를 먼저 수행해야 한다. 한편, 낸드 플래시 메모리의 읽기 및 쓰기 단위는 수 KB의 Page인 반면 지우기의 단위는 수십 개의 Page가 모인 훨씬 큰 상위 단위인 Block으로 상이하며, Block별로 쓰고 지우는 횟수가 P/E Cycle(Program and Erase Cycle)로 제한되어 있다. 이러한 제약 사항을 보완하기 위해 SSD는 기존의 Page를 무효화(invalidation)하고 다른 곳에 새로 쓰는 방식으로 지우기를 뒤로 미룬다.

이렇게 invalid page가 많아지면 이를 제거하고 빈 공간을 확보하기 위한 Garbage Collection(GC)을 SSD에서 수행하는데 이 때 block안에 있는 유효(valid) page를 다른 안전한 곳으로 이주하면서 호스트의 요청보다 더 많은 쓰기를 수행하는 쓰기 증폭(Write Amplification, WA)가 발생하면서 다른 쓰기 작업의 성능을 저하시키고 P/E Cycle을 소모하여 SSD NAND의 수명을 감소시킨다.

이 GC 작업으로 발생하는 WA를 줄이기 위해서는 쓰여진 이후 무효화되기까지의 수명이 비슷한 page끼리 한 block 내에 모여 있는 것이 유리하다. GC를 위해 block을 지우는 시점에 대다수의 page가 무효 상태가 되어 이주할 page가 적거나 없을 것이기 때문이다.

기존의 연구 중 하나[1]는 사용자가 직접 쓴 block과 GC를 통해 옮겨진 block을 분리하고 각 block의 수명을 계산하여 class를 구분한 뒤 분리하여 저장한다. 본 논문에서는 이러한 수명 계산에 더해 종합적인 Hot/cold data의 지수인 hotness 예측을 위해 수명과 같은 page별 접근 정보를 저장하고 이를 LSTM[2] 모델에 적용하여 page의 Hotness를 예측한다.

3. 설계 및 구현

3.1 개요

이번 실험에서는 실제 서버에서 저장장치에 수행된 작업을 기록한 입출력 트레이스를 모델 학습과 시뮬레이션에 이용했다. 입출력 트레이스를 적절한 단위로 수정하고 ML 모델의 학습 및 분류에 필요한 hash table을 생성하여 입출력 트레이스 일부를 가져와 정보를 입력했다. 이 정보를 이용해 ML 모델을 학습시키고 학습된 ML 모델을 이용해 모든 입출력 트레이스에 대해 labeling을 진행했다. Labeling 완료된 입출력 트레이스는 그렇지 않은 입출력 트레이스와 함께 SSD의 동작을 모방한 시뮬레이터에 투입하여 WA 개선 여부를 확인했다.

3.2 입출력 트레이스 전처리

UMassTraceRepository[3]에 보관된 금융 기관에서 동작한 OLTP Application I/O를 이번 실험에 사용했다. 해당 trace에는 총 5,334,987개의 I/O가 기록되어 있으며, 그 중 쓰기는 4,099,354개(76.8%), 읽기는 1,235,633개(23.2%)로 구성되어 있고 각 I/O line은 [ASU, Logical Sector Number, I/O Size, Read/Write, Timestamp] 순으로 값이 기록되어 있다. 이를 SSDSIM에서 인식할 수 있는 형태로 변경해야 하며, page 단위로 Hotness label을 부여할 것이기에 각 I/O line을 page 단위로 분리해야 한다. 최종 I/O line의 형태는 [Timestamp, Logical Device Number, Logical Page Number, Offset, Page Write Size, Pages to Access, R/W Code]가 되며, 이를 인식하기 위한 SSDSIM의 수정은 3.4절에서 다룬다.

표 1. 기존 입출력 트레이스 예시

ASU	LSN	Size	R/W	Time
3	817767	3072	W	1.769472

표 2. 변경한 Page 단위 입출력 트레이스 예시

Time	LDN	LPN	Offset	I/O Size	Pages to Access	R/W code
1769472000	3	204441	3	1	3	0
1769472000	3	204442	0	4	2	0
1769472000	3	204443	0	1	1	0

ML 모델 학습을 위해 초기 100만개의 trace로 page별 접근 정보를 저장하는 hash table을 생성했다. 각 page의 정보는 접근 횟수, 접근 시간 간격, 접근 시간 표준편차, 소속된 쓰기 요청의 크기 평균, 가장 최근 데이터의 수명이 저장되며, 이 중 접근 횟수를 제외한 네 가지의 정보를 ML의 feature로 사용했다. 또한, 학습 이후에도 이 hash table을 업데이트하면서 차후 labeling에 지속적으로 사용했다.

ML 모델의 지도학습을 위해 Hash table의 값을 모두 가져와 K-means 분류를 통해 4개의 cluster(K=4)로 묶어 labeling하고 이 label을 지도학습에 사용했다.

3.3 모델 구성 및 학습, 예측

이번 실험에서는 이전 정보를 통해 다음 값을 예측할 수 있는 LSTM(Long-Short Term Memory)를 이용했으며 Label의 값은 3.2절에서 K-Means(K=4) cluster를 통해 지도학습을 수행했다. Single-layered LSTM은 4개의 cluster에 대해 학습할 때 과소적합이 발생하기 때문에 Stacked-LSTM 모델을 구성해 학습을 진행했다.

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 1, 4)]	0
lstm_2 (LSTM)	(None, 1, 64)	17664
lstm_3 (LSTM)	(None, 16)	5184
dense_1 (Dense)	(None, 4)	68

 Total params: 22916 (89.52 KB)
 Trainable params: 22916 (89.52 KB)
 Non-trainable params: 0 (0.00 Byte)

그림 1. Stacked-LSTM Summary

3.4 Offline labeling

학습이 완료된 이후에는 3.2절의 hash table을 다시 이용해 Page 단위로 변경된 I/O trace를 따라가면서 ① LPN의 정보를 가져와 기록하고, ② hash table을 업데이트 한다. 이렇게 I/O trace의 흐름에 대응하여 labeling을 위한 page 정보를 구성하고 학습된 Stacked-LSTM 모델을 이용해 각 page별 Hotness label을 예측한다. 예측된 label은 I/O trace에 append하여 다음과 같이 I/O Trace를 구성한다.

표 3. Labeling 완료된 입출력 트레이스 예시

Time	LDN	LPN	Off	I/O	P	RW	label
1769472000	3	204441	3	1	3	0	0
1769472000	3	204442	0	4	2	0	0
1769472000	3	204443	0	1	1	0	0

3.5 시뮬레이터 구성

Labeling의 결과를 적용하기 위해 실제 SSD의 I/O 및 GC 등과 같은 여러 기능을 모방한 SSDSIM[4]이라는 시뮬레이터를 사용했다. SSDSIM은 I/O Trace의 주소와 크기를 Sector 단위(512 Bytes)로 입력 받아 처리한다. 3.3절에서 구성한 ML 모델은 Page단위로 labeling하기 때문에 표 2처럼 page단위 I/O trace를 받아야 한다.

또한 Hotness labeling이 끝난 I/O trace를 받아 label에 따라 격리하여 저장해야 한다. 이를 위해 hotness를 저장하는 parameter를 SSD block 구조체에 저장하고 해당 parameter를 다루는 기능을 적절한 곳에 추가하는 등의 방식으로 시뮬레이터를 적절히 수정했다.

3.6 시뮬레이터 적용

수정이 완료된 시뮬레이터는 SSD의 NAND 하드웨어 구성 Parameter를 변경해가면서 labeling된 Trace와 그렇지 않은 Trace를 각각 투입하여 시뮬레이션 결과를 저장 및 비교했다.

표 4. SSD의 낸드 하드웨어 구성 parameter

Channel number	1
Chip number	2
Die number	2
Plane number	2
Block number	2048
Page number	64
Subpage number	4
Page capacity	2048
Subpage capacity	512

4. 실험 및 성능평가

4.1 평가 환경

이번 실험은 Ubuntu 20.04가 설치된 서버 컴퓨터에서 수행했다. Intel Xeon Gold 52180이 탑재되어 있고 128GB의 DDR4 DRAM이 장착되어 있다.

SSDSIM은 Label을 이용한 block별 격리 저장이 적용된 **WithML**과 그렇지 않은 **WithoutML** 두 가지로 구성했으며, OLTP Application I/O를 page 단위로 분리한 13,140,476 개의 trace를 labeling하여 실험에 적용했다.

4.2 평가 결과

4.2.1 모델 정확도

evaluation loss and metrics ##
[0.22040630877017975, 0.9468346238136292]

그림 2. 학습 완료 모델 평가 결과(loss, accuracy)

모델을 학습시키고 평가했을 때 loss값은 0.22, 정확도는 약 95%가 나왔다. 이 모델을 이용하여 page단위로 나뉜 입출력 트레이스에 대해 hotness labeling을 진행했다.

4.2.2 시뮬레이션 결과

표 4의 HW 구성 parameter에서 실험했을 때의 결과는 다음 표5, 6과 같다.

표 5. WithoutML의 결과

GC copy back count	396,210
Write flash count	8,315,907
Waste page count	18,213

표 6. WithML의 결과

GC copy back count	395,112
Write flash count	8,315,254
Waste page count	6377

SSD의 내부 작업을 수행하면서 불필요하게 소비된 page의 수인 waste page count는 18,213개에서 6,377개로 35%정도로 감소했다. SSDSIM에서 Waste page는 copy back등으로 기존의 페이지를 옮기거나 SSDSIM의 매핑 정책으로 인해 block 내 특정 page 주소에 쓰려고 하는 경우 목표 주소와 쓰여진 page 사이의 비어 있는 page 들을 바로 무효 page로 변경하는 경우에 발생한다. 이렇게 낭비된 무효 page는 결국 이후에 추가적인 GC를 발생 시키면서 성능을 저하시키고 P/E Cycle을 소모하여 수명을 감소시킬 수 있다. Hotness label에 따라 block별로 격리하여 저장하는 것이 매핑 정책에 긍정적 영향을 줬으며 Waste page의 수가 줄면서 이에 대해 미래에 발생할 수 있는 잠재적인 GC 발생 횟수 및 copy back의 수는 줄어들었다고 볼 수 있다.

GC copy back의 경우 WithML에서 약 1,100회 (0.28%) 감소했다. Copy back의 감소율은 기대치 이하인데, 이러한 결과가 발생하는데 대한 예상되는 이유는 K-means clustering을 위한 표본의 한계점이다. 초기 100만개의 page 입출력 트레이스를 이용할 때 page 접근 정보를 저장하는 hash table의 크기는 약 17만개인데 이를 이용해 clustering 및 모델 학습을 할 때의 page 정보의 분포와 시뮬레이터를 동작 시키면서 사용하는 이후 약 1200만개의 page 입출력 정보의 분포는 차이가 날 수 있다. 예를 들어 초기 100만개의 입출력 트레이스에서 확보한 최대

수명보다 더 긴 수명의 page 정보가 hotness 분류에 사용되는 경우 예상하지 못한 결과가 나올 수도 있다.

또한, SSDSIM에서는 SSD block의 상위 계층인 plane에서 하위 전체 page중 free page의 수에 따라 gc를 수행한다. WithML에서는 이에 더해 새로운 hotness label을 가진 page를 저장할 공간을 확보하기 위해 plane의 하위 block 중 label이 부여되지 않은 free block의 수가 일정 수 이하일 때 GC를 발동하도록 하는 기능을 추가했는데 이에 따라 기존의 SSDSIM의 조건에 따른 GC 발동에 더해 추가적인 GC가 발생하여 copy back이 소폭 증가했을 것으로 보인다.

5. 결론

본 논문에서는 page의 수명 계산에 더불어 수명을 포함한 page 접근의 여러 정보를 모아 hash table에 저장하고 이에 대해 LSTM 모델을 이용하여 종합적인 Hot/cold 지수인 hotness를 예측했다. 또한 SSDSIM이라는 간단한 시뮬레이터를 이용해 WA의 개선의 예상 결과를 확인해보았다. 결과는 그렇게 만족스럽지 않았으나 개선할 수 있는 부분이 많이 있으며, 앞으로 더 좋은 결과를 낼 가능성을 볼 수 있었다. 추후에는 hash table의 크기를 줄이고 ML 모델 경량화를 수행한 후 실시간 분류를 포함한 Gem5[5] 기반 SimpleSSD에도 실험을 해볼 예정이다.

Acknowledgment

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업(IITP-2023-RS-2023-00260098)의 연구결과로 수행되었으며 정부(과학기술정보통신부)의 재원으로 한국연구재단(No. RS-2023-00208046)의 지원을 받아 수행된 연구임.

6. 참고문헌

- [1] Qiuping Wang, Jinhong Li, and Patrick P. C. Lee, The Chinese University of Hong Kong; Tao Ouyang, Chao Shi, and Lilong Huang, Alibaba Group. Separating Data via Block Invalidation Time Inference for Write Amplification Reduction in Log-Structured Storage. In the Proceedings of the 20th USENIX Conference on File and Storage Technologies(FAST'22), pages 429-443, 2022.
- [2] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735-1780.
- [3] UMass Trace Repository. <https://traces.cs.umass.edu/index.php/Storage/Storage> Accessed on 2023-11-13
- [4] SSDSIM. <https://github.com/huaicheng/ssdsim>. Accessed on 2023-11-13

[5] Gem5. <https://www.gem5.org/> Accessed on 2023-11-13

[6] SimpleSSD. <https://docs.simplessd.org/en/v2.0.12/> Accessed on 2023-11-13

WaveTR: 분산 시스템을 위한 메시지 스트림 프레임워크

김민섭, 손자영, 이우령^o, 홍은솔
경북대학교

dmg04158@gmail.com, cospet2@gmail.com, vlfkdydtk@gmail.com, eunsolhh@gmail.com

WaveTR: Message stream framework for distributed system

Kim Minseob, Son Jayoung, Lee Woolyung^o, Hong Eunsol
Kyungpook National University

요약

높은 처리량의 분산 처리를 위한 컴퓨터 구조와 시스템이 연구 및 개발되었지만, 복잡성과 높은 비용은 여전히 해결되지 않은 과제다. 본 논문에서는 이러한 문제를 해결하면서 토픽에 따라 메시지를 분류하는 새로운 구조의 분산 메시징 서비스 WaveTR과 토픽 구독 알고리즘을 제안한다. WaveTR은 트랜지스터라는 이름의 노드를 통해 메시지를 분류하고 전달한다. 트랜지스터로 이루어진 네트워크는 데이터의 복제, 분산 처리, 계층화를 유리하게 하며 직관적이고 유연한 확장성을 제공한다. 트랜지스터는 프레임워크로 제공되므로 사용자에 따라 다양한 형태로 구현될 수 있다.

1. 서론

컴퓨터가 처리해야 하는 정보의 양과 처리되는 속도가 증가함에 따라 클라우드 환경을 이용한 여러 분산 처리 시스템이 개발되었고, 다양한 장치에서 발생하는 정보를 분류하고 수집하기 위한 메시징 서비스가 발전하였다. 이러한 서비스에는 대표적으로 Kafka와 RabbitMQ가 있다[1].

Kafka는 네트워크에서 메시지를 수신할 때 토픽에 따라 분류하고 그것을 디스크에 저장할 때는 여러 개의 노드로 이루어진 클러스터에 복제함으로써 비동기 메시지 서비스를 구현한다.

RabbitMQ는 AMQP(Advanced Message Queuing Protocol)라는 TCP 통신을 위한 저수준의 프로토콜을 구현한 비동기 메시징 서비스이다. Kafka의 토픽과 비슷한 개념의 큐를 통해 메시지를 분류하고 디스크에 저장하며, 클러스터를 지원함으로써 메시지를 여러 노드에 복제할 수 있다.

본 논문에서는 기존의 메시징 서비스와 유사하게 토픽에 따라 메시지를 분류하며 분산 처리와 확장성을 달성할 수 있는 새로운 메시징 서비스인 WaveTR과 분산 처리 환경에서 사용할 수 있는 토픽 구독 알고리즘을 제안한다.

WaveTR은 트랜지스터라는 이름의 노드로 네트워크를 구성한 분산 메시징 시스템이다. 트랜지스터는 토픽 구독 알고리즘에 의해 메시지에 포함된 토픽으로 메시지를 분류하고, 그 메시지를 구독하는 클라이언트나 트랜지스터로 전달하는 단순한 역할을 한다. 토픽과 트랜지스터로 구성된 네트워크는 토픽에 따라 메시지가 저장되는 위치를 분산 및 복제 할 수 있으며 네트워크의 계층화를 통해 시스템의 유연한 확장성을 달성한다.

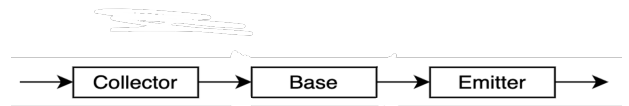
트랜지스터는 프레임워크로 제공되기 때문에 토픽을 구독하는 방법은 사용자의 구현에 따라 정적이거나 동적일 수 있으며, 메시지의 처리 방법도 외부 프로그램을 경유하지 않고 트랜지스터 내부

에서 이루어질 수 있다. 또한 디스크의 사용을 지원하면 더 강력한 비동기 메시지 큐를 제공할 수 있다.

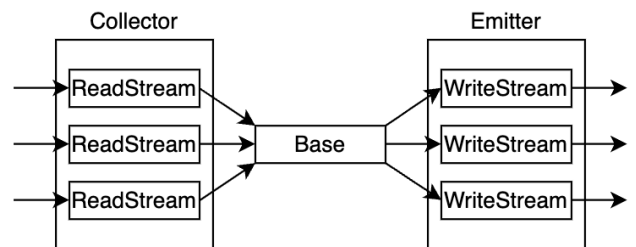
본 논문의 구성은 다음과 같다. 2장에서는 제안된 트랜지스터의 설계와 구조에 대하여 기술한다. 3장에서 토픽과 토픽 구독 알고리즘을, 4장에서 트랜지스터의 성능평가 및 테스트 결과를, 5장에서는 결론 및 향후 연구 방향을 기술한다.

2. 트랜지스터

트랜지스터는 WaveTR을 구성하는 독립적인 프로그램이다. 트랜지스터는 메시지의 입력, 토픽 구독 및 인덱싱, 출력을 담당하는 Collector, Base, Emitter로 구성된다. 트랜지스터를 지나는 데이터의 흐름은 <그림1>과 같다.



<그림 1> 트랜지스터 구조



<그림 2> ReadStream과 WriteStream

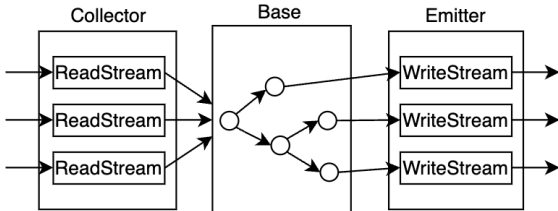
2.1. Collector와 Emitter

Collector는 트랜지스터의 메시지의 입력을, Emitter는 메시지의 출력을 담당한다. 이 둘은 연결된 노드마다 스트림을 가지고, 메시지는 스트림을 따라 트랜지스터를 이동한다.

<그림2>는 하나의 트랜지스터에서 스트림을 지나는 메시지의 흐름을 나타낸다. 스트림은 각각 독립적인 스레드를 가지며 ReadStream은 메시지 게시자로부터 수신한 메시지를 Base로 전달하는 기능을, WriteStream은 전달된 메시지를 메시지 구독자로 송신하는 기능을 한다. 각 WriteStream은 큐 형태의 자료구조로 송신할 메시지를 우선 메모리에 보관하여 간단한 비동기 메시지 큐를 구현하고, 이를 통해 메시지의 흐름을 조절하여 전체적인 지연 시간을 낮출 수 있다.

2.2. Base

Base는 토픽을 구독한 스트림으로 메시지를 전달할 수 있도록 내부적으로 인덱스를 가진다. Collector로부터 전달받은 메시지는 인덱스를 거쳐 Emitter로 이동하며, 각 WriteStream은 자신이 구독한 토픽에 맞는 메시지를 전달받는다.

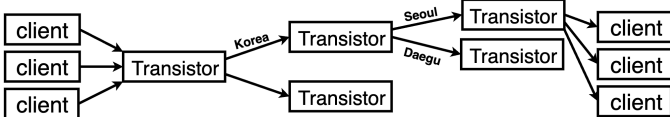


<그림 3> 트랜지스터 구조

<그림3>은 인덱스가 포함된 메시지 흐름을 나타낸다. Base의 인덱스는 다양한 형태로 구현될 수 있으며, 동적인 구독을 통한 인덱스 변화가 포함될 수 있다. 본 논문에서 구현한 Base는 인덱스 업데이트의 비동기 처리의 성능 저하를 줄이기 위해 인덱스와 그 복사본을 스왑하여 사용하는 롤링 업그레йд 방식을 사용하였다. 또한 스왑을 위해 Lock을 기다리는 동안에도 인덱스 업데이트를 처리할 수 있도록 하여 비동기 처리의 효율을 높였다[2].

2.3. 트랜지스터 결합

WaveTR은 스트림과 트랜지스터로 구성된 네트워크 시스템으로 처리가 필요한 데이터를 분산, 복제, 병렬 처리한다.

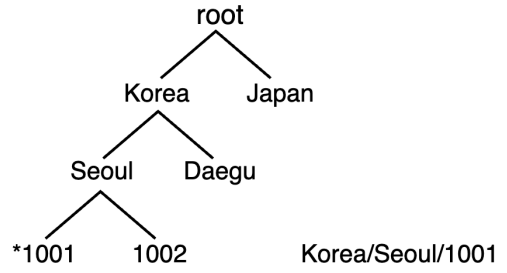


<그림 4> 트랜지스터 결합

트랜지스터는 결합을 통해 다계층을 형성할 수 있다. <그림4>는 이를 이용한 분산 시스템의 예시이다. 위 예시에서 트랜지스터는 토픽의 노드당 하나씩 존재하고, 각각의 트랜지스터는 그 노드에서 갈라지는 토픽만을 처리한다. 토픽의 복잡성과 트랜지스터의 분할은 성능에 큰 영향을 줄 수 있으므로 문제에 따라 적절한 토픽을 구성해야 한다[3].

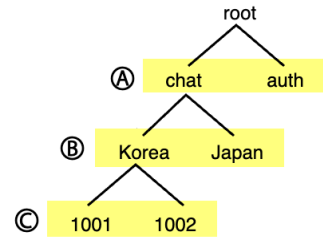
3. 인덱스

3.1. 토픽



<그림 5> 인덱스(좌)와 메시지 토픽(우)

구독자로부터 요청된 모든 토픽을 모으면 그 구조는 트리의 형태이다. 하나의 메시지는 루트에서 특정 노드까지의 경로를 토픽으로 가질 수 있고, Base의 인덱스는 트리 구조로 스트림의 구독을 관리한다. 현재 인덱스와 메시지의 토픽이 <그림5>와 같다면 "Korea/Seoul/1001"의 경로를 따라 트리를 이동하여 "*"로 표시된 노드의 스트림에 해당 메시지가 전달된다.

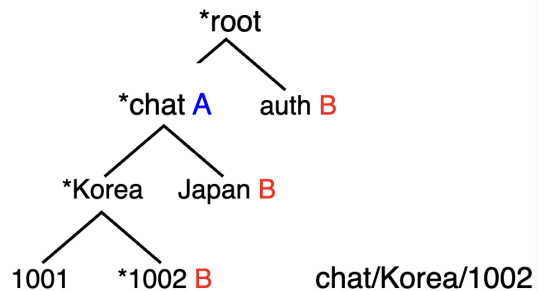


<그림 6> 레이어가 나뉜 토픽

트리 구조의 토픽은 메시지의 전달을 레이어로 추상화할 수 있다. <그림6>은 트리의 각 노드를 레이어로 나눈 예시이다. 레이어 A는 서비스 단위로 메시지를 구분하고 레이어 B는 지역 단위로 메시지를 구분한다. 각 레이어에서 분리된 메시지는 그 레이어를 구독하는 각각의 트랜지스터에 전달하여 분산 처리할 수 있다.

3.2. 토픽 구독 알고리즘

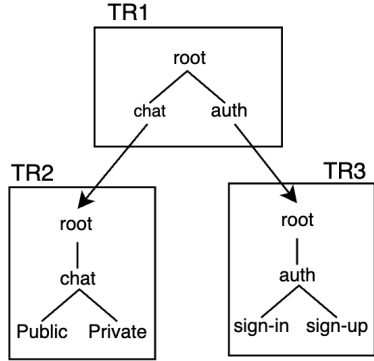
스트림은 하나 이상의 토픽을 구독하여 메시지를 전달받는다. 여기서 토픽은 인덱스 트리의 노드이다. 만약 어떤 스트림이 트리의 한 노드를 구독했다면 동일한 스트림은 해당 노드의 후손 노드를 구독할 수 없다.



<그림 7> 스트림의 구독

<그림7>은 두 스트림 A와 B가 토픽을 구독한 예시이다. 스트림 A는 "chat"을 구독했고 스트림 B는 "auth"와 "chat/japan", "chat/

korea/1002”를 구독했다. 스트림 A는 “chat”을 구독했으므로 그 노드의 후손인 “chat/korea”와 같은 토픽은 구독할 수 없다. 메시지는 메시지의 토픽에 포함되는 모든 노드에 전달된다. 예를 들어 위 상황에서 메시지의 토픽이 “chat/korea/1002”라면 *로 표시된 root, chat, korea, 1002 노드에 전달되므로 스트림 A와 B 모두 메시지를 전달받을 수 있다.



<그림 8> 멀티 레이어 토픽 인덱스

<그림8>은 트랜지스터의 레이어를 나누어 인덱스를 분리한 예시이다. TR1은 모든 종류의 메시지를 받되 채팅은 TR2, 인증은 TR3로 메시지를 전달하여 서비스를 분리한다. 구현에 따라, 통과한 토픽을 단순화시켜 TR2의 chat과 TR3의 auth를 제거할 수도 있다.

4. 성능 측정 및 평가

트랜지스터의 구현은 Go 언어, 네트워크 프로토콜에는 HTTP/2를 이용한 gRPC와 Protobuf를 사용하였다.

4.1. 단일 트랜지스터 성능 측정

트랜지스터(TR)에서 CPU를 예약하고, 실제 사용량에 도달할 수 있도록 메시지를 보내는 게시자(Pub)와 메시지를 받는 구독자(Sub)의 자원 양을 조절하여 트랜지스터의 성능을 측정하였다. core는 트랜지스터의 예약된 CPU의 코어, util은 실제 사용한 코어를 의미한다.

util	core	초당 전송량 (MS/s)	1코어 초당 전송량 (MS/s*core)
0.533	0.5	57,400	114,800
1.094	1.0	185,667	185,667
11.73	1.5	274,000	182,667
2.37	2.0	275,667	137,833
2.83	2.5	317,000	126,800
2.66	3.0	311,333	103,778
4.29	8.0	391,000	48,875

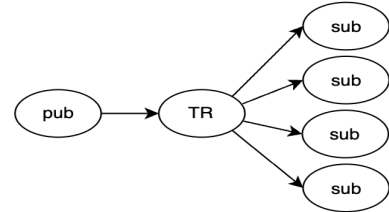
<표 1> 단위 시간당 메시지 수

<표1>는 자원에 따른 트랜지스터가 1초당 주고받을 수 있는 메시지 수를 나타낸다. 메시지의 크기는 30바이트로, WriteStream

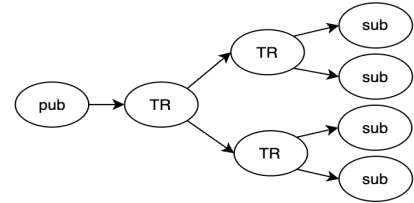
의 메시지 큐의 크기는 10으로 작은 값을 설정했다. Pub에게 메시지를 생성을 요청하고 10분 동안 관찰한 뒤 가장 높은 수치를 기록하였다.

4.2. 메시지 복제 성능

하나의 Pub으로부터 받은 메시지를 복제하여 다수의 Sub에게 전달하는 경우의 성능을 측정하였다.



<그림 9> 단일 계층 구조



<그림 10> 다계층 구조

성능은 <그림9>처럼 하나의 트랜지스터 사용하여 Sub에게 전달하는 경우와 <그림10>처럼 트랜지스터 3개를 두 계층으로 나누는 경우를 나누어 비교했다. 두 계층으로 나누었을 때는 각 트랜지스터가 동일한 CPU 자원을 갖도록 설정했다.

Sub	total core	1계층 초당 전송량 (MS/s)	2계층 초당 전송량 (MS/s)
2	3	162,750	59,043
	6	170,010	123,261
	9	184,530	183,134
4	3	238,300	115,022
	6	250,600	255,440
	9	259,812	358,444

<표 2> 메시지 복제 성능

4.3. 평가

<표1>에서 보여지는 단위 시간당 메시지 수는 1~1.5 코어의 자원에서 가장 높은 효율을 보여주었으며 그보다 높은 코어는 하나의 코어가 1초동안 처리하는 메시지 및 바이트 수가 점차 줄어들었다. 이 테스트 결과는 단일 프로그램에서의 메시지 전송은 성능에 따른 전송량에 한계가 존재하기에 높은 효율을 위해선 적절한 수의 프로그램으로 메시지를 나누어 처리해야함을 의미한다.

<표2>에 의하면 낮은 자원에서는 단일 계층에서의 메시지 전송량이 더 높지만 반대로 높은 자원에서는 두 계층으로 메시지를 분

산하여 처리하는 것이 더 높다. 트랜지스터의 수가 늘어나면 네트워크 내부에서 메시지의 추가 전송이 필요하여 성능이 감소하지만, 높은 자원이 필요한 시스템에서는 그보다 분산 처리로 얻는 성능 증가가 더 크다.

Sub	total core	1계층 1코어 초당 전송량 (MS/s)	2계층 1코어 초당 전송량 (MS/s)
2	3	54,250	19,681
	6	28,335	20,544
	9	20,503	20,348
4	3	79,433	38,341
	6	41,767	42,573
	9	28,868	39,827

<표 3> 코어당 초당 전송량

<표3>는 <표2>의 결과를 코어 수로 나누어 1코어당 초당 전송량을 계산한 결과다. 단일 계층은 코어가 증가함에 따라 하나의 코어가 처리하는 메시지 수가 점차 감소하는 반면, 두 계층으로 나눈 경우는 감소하지 않고 균등하다. 이는 다계층의 트랜지스터로 분산된 시스템이 자원 할당을 통한 성능 증가에 더 유리함을 나타낸다.

Sub	total core	1계층 1 Sub 초당 전송량 (MS/s)	2계층 1 Sub 초당 전송량 (MS/s)
1	3	91,691	
2		81,375	29,522
4		59,575	28,756
1	6	105,825	
2		85,005	61,631
4		62,650	63,860
1	9	111,488	
2		92,265	91,567
4		64,953	89,611

<표 4> Sub당 전송량

<표4>는 <표2>의 결과를 Sub 수로 나누어 하나의 Sub이 받는 초당 전송량을 계산한 결과다. 단일 계층의 경우 구독중인 Sub이 증가할 수록 효율이 감소하지만, 두 계층으로 나눈 경우는 비슷한 전송량을 보인다. 이는 다계층의 트랜지스터로 분산된 시스템은 많은 구독에 대해 더 효율적으로 대응할 수 있음을 나타낸다.

5. 결론 및 향후 연구 방향

본 논문에서는 토픽에 따라 메시지를 분류하는 WaveTR 구조와 토픽 구독 알고리즘을 제안하고, 이에 대한 성능을 분석하고 평가하였다. 트랜지스터는 네트워크를 구성하고 메시지를 분류 및 전달하는 독립적인 프로그램이다. 단일 계층의 트랜지스터로 구성된 시스템은 자원과 구독 수 증가에 따른 메시지 전송의 효율 감소를 보였지만, 다계층의 트랜지스터로 구성된 시스템은 높은 자원과 높은 구독 수에서 단일 계층보다 더 높은 성능을 보여주었다. 이는 다계층의 WaveTR 구조가 자원 할당을 통한 성능 증가에 유리하면서, 많은 구독에 대해 더 강건하게 대응할 수 있음을 보여준다. 또한 WaveTR은 노드를 결합하는 방식으로 시스템을 구성하기에 직관적인 설계, 경제적인 비용, 유연한 확장성이라는 장점을 가지며, 경우에 따라 노드를 서비스, 지역 단위로 레이어를 나누어 설계할 수 있다.

트랜지스터는 인덱싱과 탐색을 비롯한 구현에 성능 차이가 발생할 수 있고, 특정 도메인은 그에 맞는 구현이 필요할 수 있다. 따라서 트랜지스터의 구현을 위한 알고리즘 및 인덱싱 방법에 대한 후속 연구가 필요할 것으로 보인다.

사사(Acknowledgement)

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음" (2021-0-01082)

참고 문헌

- [1] Sharvari, T., & Sowmya Nag, K. (2019). A study on modern messaging systems-kafka, rabbitmq and nats streaming. CoRR abs/1912.03715.
- [2] Wolski, A., & Laiho, K. (2004, May). Rolling upgrades for continuous services. In International Service Availability Symposium (pp. 175-189). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [3] 남범준, 정현민, 권영우. (2023). 분산 메시징 시스템에서 Topic 과 Partition 관리의 필요성(pp. 673-674), 한국통신학회 학술대회논문집

차량 주행 상황에서 촬영된 이미지의 분포 정보를

활용한 안개 분류 기법

임도현* 고석주**
경북대학교 대학원 컴퓨터학부

Fog Classification Method Based on Distributional Characteristics of Images Captured in Vehicle Driving Scenarios

Lim Dohyeon^o Seok-Joo Koh

School of Computer Science and Engineering, Kyungpook National University

요약

안개와 같은 악천후 상황에서는 시야가 제한되어 운전 안전에 상당한 위험이 따르며, 차량간의 속도 차이가 증가하고 후행 거리가 감소함으로써 교통사고의 발생위험을 높인다. 본 연구에서는 차량 주행 환경에서 안개를 신속하게 감지하기 위한 중요성에 주목하여, 차량 주행 환경에서 촬영된 이미지를 기반으로 안개를 분류하는 알고리즘을 제안한다. Dark Channel Prior과 Sobel filter를 활용하여 이미지를 전처리하고, 이를 기반으로 안개를 분류하는 모델을 구축한다. 일반적인 CNN기반 분류 모델과는 달리, 해당 방법론은 이미지의 각 픽셀에 대해 기초적인 통계 및 분포정보를 추출하여 분류를 진행하였으며 실험결과, AI Hub에서 얻은 14,000장의 안개 이미지와 일반 주행 이미지를 기반으로 분류 정확도는 96.62%로 나타났다. 이를 토대로 차량 주행중 안개 상황을 빠르게 감지하여 운전자에게 경고를 주는 시스템의 개발이 가능함을 입증하였다.

Key words

Machine Learning, Image Classification, Fog Detection

1. 서론

도로 위에서 안개가 발생하면 운전자의 시거(視距)가 감소하게 되고, 이로 인해 차량 간의 속도 편차 증가와 차간 간격 감소로 교통사고 발생 위험이 증가 한다[1]. 또한, 안개 발생 시 교통사고 치사율(교통사고 100건당 사망자 수)은 8.8명으로 맑은 날 대비 5.5배 높은 것으로 집계되었으며, 특히 보행자 교통사고(차대 사람)의 경우 안개 발생 시의 치사율(25명)이 맑은 날(3명) 대비 8.3배 높은 것으로 확인되었다. [2]

안개 발생으로 인한 교통사고 예방을 위한 효과적인 방안은 운전자에게 안개 발생 현황을 공유하는 것이며, 이를 위하여 안개 관측을 수행하는 방안은 시정계의 시정거리 정보를 활용하는 방안, 인공위성의 영상신호로 안개 영역을 탐지하는 방안, CCTV 카메라의 영상 이미지로 안개를 감지하는 방안 등이 존재한다. 그러나 시정계 기반의 안개 관측은 인간활동이 비교적 적은 기상관측소에 시정계가 설치됨에 따라, 관측소 주변의 안개 상황이 도로의 안개 상황을 대표하지 못하는 사례가 빈번하고,

인공위성 기반 안개 관측은 다양한 장애요소와 실시간 대응에 난점이 있으므로 CCTV 기반의 안개 관측이 안개 상황을 보다 현실적으로 반영하는 것으로 알려져 있다. [3]

그러나 여전히 CCTV 기반의 안개 관측은 촬영 기기가 설치되지 않은 지역의 안개 발생 현황은 파악이 어려운 점이 있으므로, 본 연구에서는 차량 주행 환경에서 촬영된 이미지를 기반으로 주행 상황에서의 안개를 분류하는 알고리즘을 제안함으로써, 선행 운전자가 확인한 안개 발생 상황 정보를 후행 운전자에게 신속하게 전달하는 방법을 제시하고자 한다.

2. 관련 연구

2.1 안개 특성 분석

컴퓨터 비전과 컴퓨터 그래픽스 분야에서 안개 이미지 형성을 설명하는데 사용되는 모델은 아래와 같다.

$$I(x) = J(x)t(x) + A(1-t(x)),$$

이때 $I(x)$ 는 안개에 의해 왜곡된 실제 이미지, $J(x)$ 는

* 경북대학교 dhlim@knu.ac.kr, ** 경북대학교, sjkoh@knu.ac.kr

맑은 날씨의 이미지, $t(x)$ 는 대기투과율 A 는 대기 산란 광을 의미한다. 이를 해석하면, 물체에서 반사되어 들어오는 $J(x)$ 는 대기를 통과하면서 정보가 일부 손실되고 $J(x)t(x)$ 만큼의 정보가 남는다. 또한, 광원으로부터 형성된 A 가 정보와 섞이면서, 결과적으로 관찰자에게 $I(x)$ 가 전달된다. [4]

Dark Channel Prior는 이미지에서 픽셀값들을 비교하여 $J(x)$ 값을 추정하여 Dehazing task를 수행하는 방법론이며 $J(x)$ 값을 추정하는 모델은 아래와 같다.

$$J^{dark}(x) = \min_{c \in r, g, b} (\min_{y \in (\Omega(x))} (J^c(y)))$$

이때 $J^{dark}(x)$ 는 Dark Channel, $J^c(y)$ 는 픽셀 r, g, b 값 중 한 Channel, $\Omega(x)$ 은 x 를 중심으로 일정 범위 내의 픽셀들의 집합을 의미한다. 해당 가정을 기반으로 안개 발생 시의 $t(x)$ 를 연산이 가능하여 이미지에서 Dehazing task가 수행 가능할 뿐만 아니라, $t(x)$ 값을 통해 일반/안개 분류 task도 수행할 수 있다.

2.2 CNN 기반 분류 모델

합성곱 신경망(Convolutional Neural Network, CNN)은 이미지 인식 및 패턴 인식에 활용되는 모델로 합성곱 연산을 기반으로 특정 영역의 feature과 output이 상호작용하여 이미지 데이터를 분석 가능한 모델이다.

안개 분류 영역에 있어 CNN 기반의 모델을 통해 해무를 96%의 분류 정확도를 달성한 연구[5]가 진행되었으며, 안개와 유사한 패턴을 지닌 화염 및 연기 감지 분야에서도 딥러닝 기반의 합성곱 신경망을 활용하여 분류를 진행한 연구가 있다.[6]

합성곱 연산 및 풀링 연산을 통해 이미지의 이동에 대한 불변성을 가질 수 있으며, 이에 따라 객체나 패턴이 이미지 내에서 이동하여도 모델이 인식하는 특성이 있으며, 이로 인해 이미지에 대해 심층적인 분석 결과를 거둘 수 있다. 그러나 합성곱 신경망 모델의 경우 모델의 깊이가 깊어질수록 연산량이 증가하는 단점과 이미지의 어떤 특성이 실제로 어떤 방식으로 분류가 되어 영향을 미치는지 파악이 어려운 단점이 존재한다.

3. 제안하는 안개 분류 기법

본 연구에서 제안하고자 하는 방법론은 안개의 주요 특성을 고려하여 픽셀별로 이미지 전처리를 진행하고, 도출된 픽셀의 값의 기초 통계 및 분포 정보를 활용하여 안개를 분류하는 방법론이다.

기존의 CNN 분류 모델이 입력된 이미지의 패턴을 RGB 영역에서 패턴값을 기반으로 이미지를 분류한 것과 달리, 제안하고자 하는 방법론은 이미지 전처리 과정을 통해 픽셀별로 연산된 값의 기초 통계 및 분포 정보를 활용함으로써 도로 주행 상황에서의 노이즈 형태의 안개 정보를 분류하는 특성이 있다.

Dark Channel Prior과 Sobel filter를 활용한 이미지 특성 맵 도출 단계, 기초 통계 및 분포 정보에 기반한 데이터 특성 추출 단계, 추출된 데이터를 주요 변수로 활용하여 머신러닝을 진행하는 모델 생성 단계로 구성되어 있다.

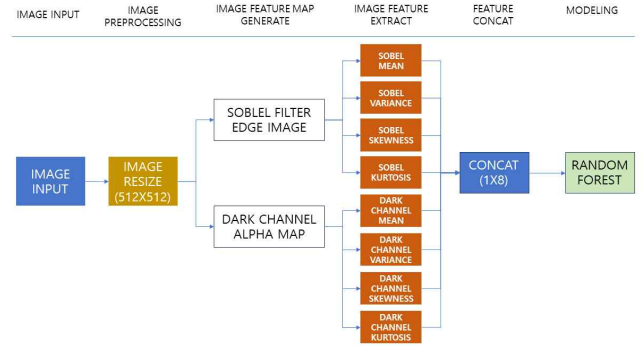


그림 1. 제안 분석 방법론 단계별 구조
Fig 1. Architecture of Proposed Analysis Methodology

3.1 이미지 특성 맵 도출 단계

데이터 맵 도출 단계에서는 다양한 크기로 입력된 이미지를 512X512의 균일한 이미지 형태로 보정 후 Dark Channel Prior의 Alpha map 이미지와 Sobel filter의 edge 이미지를 생성하여 이미지 특성 맵을 생성한다.

3.1.1 Dark Channel Prior

먼저 입력된 이미지는 512X512 크기로 보정되어 처리된다. RGB 채널로 분리한 후, 각 픽셀의 최솟값을 계산하여 최솟값 이미지를 생성한다. 이 최솟값 이미지에 64X64 크기의 커널을 적용한 침식 연산을 수행하여 Dark 이미지를 도출한다.

다음으로 Dark 이미지에서 상위 0.1%에 해당하는 픽셀을 선택하고, 각 채널(R, G, B)에서 최댓값을 찾아 배열 A에 저장한다. 이 배열 A의 최댓값으로 원본 이미지를 정규화하고, 정규화된 이미지의 각 채널에 대해 다시 최솟값 이미지를 생성한다. 이후에는 32X32 크기의 커널을 사용하여 두 번째 침식 연산을 수행한다. 이 과정을 통해 얻어지는 이미지는 분류모델에 적용 가능한 Alpha map으로 정의된다.

3.1.2 Sobel filter

Dark Channel Prior에서 적용된 형태와 같이 입력된 원시 이미지를 512x512 크기로 바로잡는다. 입력된 이미지를 그레이 스케일로 보정 후 수평 방향과 수직 방향으로 Sobel filter를 적용하여 각각의 절댓값을 계산한다. 이 두 방향에서 얻은 결과를 가중치를 더해 합친 Edge 이미지를 도출한다.

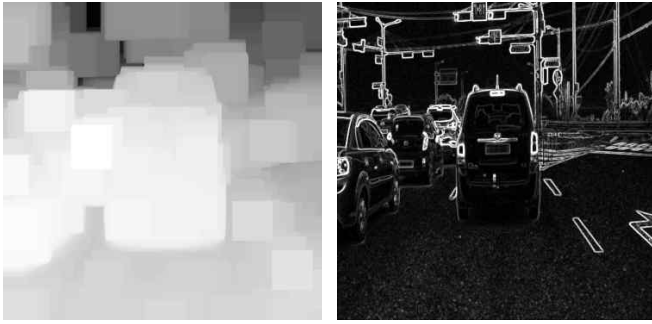


그림 2. 안개 이미지 특성 맵
Fig 2. Fog Image Feature Map

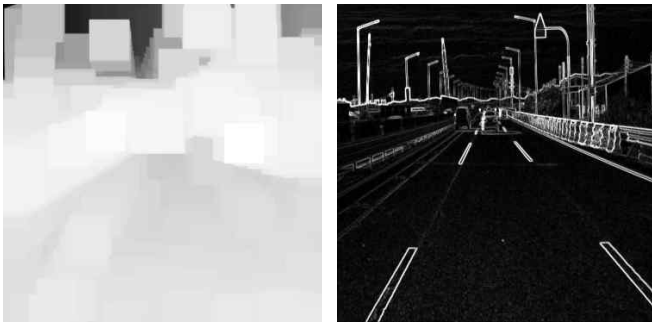


그림 3. 정상 이미지 특성 맵
Fig 3. Not Fog Image Feature Map

3.2 이미지 특성 정보 추출 단계

도출된 Alpha map 이미지와 Edge 이미지 내에서 이미지 특성 정보를 추출하고 이를 이용하여 분류를 진행한다.

각각의 이미지 특성 맵의 평균, 분산, 왜도, 첨도를 계산하여 이미지 특성 맵의 정보를 수치화한다. 이렇게 추출된 Alpha map 정보와 Edge 이미지의 수치적인 정보를 특성으로 사용하여 각 이미지를 특성 벡터로 표현한다.

이후 해당 변수를 하나의 테이블로 결합해 변수 데이터 세트로 정의하며 이와 같은 데이터를 Random forest 모델로 적용하여 분류를 진행한다.

4. 실험 결과

본 연구에서는 Dark Channel Prior과 Sobel filter를 활용한 안개 분류모델을 개발하기 위해 AI HUB 내 공개된 안개가 낀 상황에서 도로를 주행한 데이터셋을 활용하였다.

Randomforest 모델을 기반으로 14,000장의 안개 이미지와 일반 주행 이미지를 활용하여 분류를 진행하였으며, 홀드 아웃 기법을 적용하여 Train/Test 데이터로 나누어 모델을 평가하였다.



그림 4. 안개 이미지 데이터
Fig 4. Foggy Image Data

안개 탐지 모델의 성능을 측정하기 위해 정확도 (Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1 Score) 등의 주요 평가 지표를 사용하여 평가를 진행하였다.

표 1. 오차 행렬 결과
Table 1. Confusion Matrix Result

		Prediction	
		0	1
Reference	0	2050	72
	1	62	2016

* 0 : fog / 1: Not fog

오차 행렬값을 기준으로 Y축은 실제 값을 의미하고, X축은 예측 결과값을 의미한다. 전체 4,200장의 테스트 이미지 중 4,066장이 분류에 성공하였으며, 134장의 이미지가 분류에 실패하였다.

표 2. 정확도 평가 결과
Table 2. Accuracy Assesment Result

accuracy	precision	recall	F1-measure
96.62%	96.69%	97.06%	96.87%

전체 정확도는 약 96.62%, 모델 Precision은 약 96.69%, Recall 값은 약 97.06%, F1-measure 값은 약 96.87%로 높은 수치를 나타내었다.

5. 결론

본연구에서는 악천후 차량 운행 상황에서 촬영된 이미지를 Dark Channel Prior과 Sobel filter를 기반으로 전처리를 진행하고 Randomforest 모델을 기반으로 이진 분류를 수행하였다. 모델의 정확도 평가 결과 전체의 정확도는 약 96%를 도출하였다. 차량 주행 상황에서 안개에 대한 분류모델이 존재하지 않아 타 모델과 비교가 어려운 점이 있으나, 100시간의 주행을 통해 확보한 14,000장의 이미지를 활용하여 모델을 구축하였다는 점에서 분류모델의 신뢰성이 확보 가능할 것으로 기대된다. 또한, 구축된 모델을 기반으로 블랙박스 내 해당 모델을 적용한다

면 선행 운전자가 확인한 안개 상황 정보를 후행 운전자에게 전달이 가능할 것으로 기대된다.

그러나 활용된 데이터 세트가 안개의 발생으로 인한 시정거리 계산 정도가 명확하지 않은 한계점이 있으며, 이러한 한계는 차후 추가 연구를 통해 안개 발생으로 인한 촬영 데이터를 확보하고 이를 기반으로 추가 학습을 통해 정확도를 개선 가능할 것으로 보여진다.

향후 연구에서는 해당 오류를 개선하기 위하여 안개 발생 요인을 더 조사하고, 센서 데이터를 추가로 활용하여 이미지 데이터 세트와 센서 데이터셋을 결합하여 분석하는 방향으로 연구를 진행할 것이다.

관련 문헌

- [1] Michigan Department of Transportation, Providing Visibility and Visual Guidance to the Road User,1985
- [2] 도로교통공단,초겨울인 11월 안개 교통사고 주의, 2021,
https://www.koroad.or.kr/main/board/6/7476/board_view.do?&cp=25
- [3] 김경원, 파주안개감시측정소에서 CCTV를 통해 관측된 도로 안개의 강도 측정,2021,한국대기환경학회지 Vol.37, No(3), pp.338-401,2021
- [4] S. Yang, Q. Zhu, J. Wang, D. Wu, and Y. Xie, An improved single image haze removal algorithm based on dark channel prior and histogram specification, Proceedings of the 3rd International Conference on Multimedia Technology, pp. 279-292,2013
- [5] 김나경, 박수호, 정민지, 황도현, 앵호자리갈 운자야, 박미소, 김보량, 윤홍주 CCTV와 합성곱 신경망을 활용한 해무 탐지 기법 연구. 한국전자통신학회 논문지 V.15, No(6),pp.1081-1088,2020
- [6] 류진규, 곽동걸, 딥러닝 기반의 합성곱 신경망을 이용한 화염 및 연기 감지 알고리즘에 관한 연구,한국방재학회논문집 Vol.20 No.1,pp.223-232,2020

더 나은 비전-언어 모델링을 위한 속성-어댑터

조용현^o 박현수 박진선
부산대학교 정보융합공학과
부산대학교 정보컴퓨터공학과
{cm144dlqslek, jerome10, jspark}@pusan.ac.kr

Attribute-Adapter for Better Vision-Language Modeling

Yonghyeon Jo^o Hyeonsu Park, Jinsun Park
Department of Information Convergence Engineering, Pusan National University
School of Computer Science and Engineering, Pusan National University

요약

본 논문은 GPT를 활용하여 클래스에 해당하는 속성 정보를 생성하고, 이를 비전-언어 모델링에 활용할 수 있는 AT-Adapter를 제안한다. AT-Adapter는 클래스의 특징을 고려하여 속성을 분류하고, 이를 최종 로직에 통합함으로써 더욱 강건한 모델을 형성하였다. 실험 결과 16샷의 정량적 평가로 Food101, Oxford-Pets 데이터 셋에서 각각 78.94%, 89.70%의 높은 성능을 달성하였으며, 이는 기존 Tip-Adapter 모듈 대비 0.99%, 0.41%의 성능이 향상되었다.

1. 서론

비전과 언어는 주변 세계를 인식하고 환경과 다양한 상호 작용을 하기 위한 핵심적인 수단으로 자리 잡고 있다. 비전-언어 모델(VLM, Vision-Language Model)[1-6]은 비전과 언어의 모달리티를 결합하는 방법에 대한 다양한 접근을 제시하며 이에 대한 활용 방법이 대두되고 있다. 특히 CLIP(Contrastive Language-Image Pre-training)[1] 모델은 대규모 시각-언어 모델을 훈련하는 간단하면서도 효과적인 방법으로 드러났으며, 대규모 크기(400M)의 이미지와 텍스트를 활용하여 학습하는 새로운 접근법을 제공하였다. CLIP의 핵심 아이디어로 지도 학습을 통해 전이 가능한 비전 특성을 학습하는 것이 제안되었으며, 제로 샷 분류에서 좋은 성능을 보였다[1,4-8].

최근에는 CLIP에 다양한 방법론을 제안하는 연구가 진행되고 있다. 먼저 Gao Peng 등[9]은 Residual

Connection을 활용하여 사전 학습된 CLIP의 이미지 특징과 어댑터의 특징 정보를 결합하는 Clip-Adapter를 제안하였으며, Zhou Kaiyang 등[10]은 CLIP 모델의 텍스트 인코더에서 Learnable Continuous tokens를 파인튜닝(Fine tuning)함으로써 텍스트의 프롬프트를 최적화시키는 CoOp을 제안하였다. 마지막으로 Zhang Renrui 등[11]은 훈련 데이터에 대해 Key-Value의 캐시 가중치를 생성함으로써 역전파 없이 좋은 성능을 보이는 Tip-Adapter를 제안하였다. 하지만 이러한 방법들은 각 클래스에 대한 특징만을 활용하여 학습을 진행하기 때문에, 더 복잡하고 다양한 클래스를 분류해야 하는 퓨-샷 러닝[12,13]에서는 충분한 정보를 얻지 못하게 된다.

이를 해결하기 위해 본 논문에서는 GPT를 활용하여 클래스에 해당하는 속성 정보를 생성하고, 이를 활용할 수 있는 AT-Adapter를 제안한다. AT-Adapter는 클래스들의 속성 관점에서 특징을 분류하고 이를 logits에 결합함으로써 더욱 강건한 비전-언어 모델을 구축한다. 또한, 다

표 1. Food101[15] 데이터 셋의 GPT 속성 정보 출력값 예시

Category	Attribute
1. Cuisine	"Italian", "Japanese", "Mexican", "French", "American", "Indian", "others"
2. Types	"Beef", "Chicken", "Seafood", "Vegetarian", "Desserts", "others"
3. Method	"Grilled", "Fried", "Baked", "Raw", "Boiled", "others"
4. Menu	"Appetizers", "Main Courses", "Desserts", "Side Dishes", "Snacks", "others"
5. Flavor	"Spicy", "Mild", "Sweet", "Savory", "others"
6. Vegetable	"Vegetarian", "Non-Vegetarian"
7. Temperature	"Cold", "Hot", "Room Temperature", "Frozen", "others"

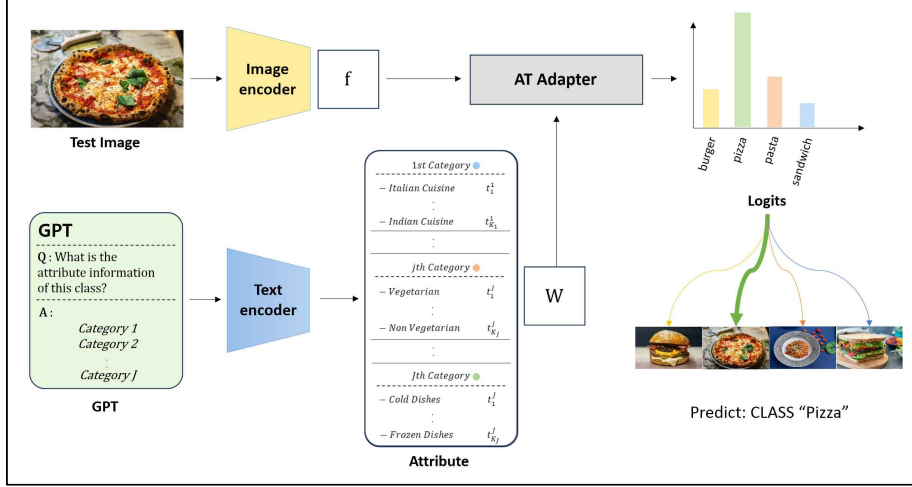


그림 1. AT-Adapter의 아키텍처

양한 모델에 적용 가능한 범용 모듈로 활용할 수 있어 모델의 특성과 상관없이 유연하게 적용될 수 있다.

2. 방법론

그림 1은 AT-Adapter의 아키텍처이다. GPT는 클래스의 속성 정보를 생성하여 훈련 시 필요한 속성 Key-Value의 캐시 가중치를 추출한다. AT-Adapter는 입력된 이미지를 속성 Key, Value 값을 사용하여 각 카테고리에 해당하는 속성으로 분류하도록 학습한다. 최종 logits을 출력할 때는 클래스 차원의 logits으로 변환하여 클래스를 예측하도록 한다. 본 장에서 2.1은 GPT를 활용하는 방법을 설명하며, 2.2는 AT-Adapter의 계산 과정과 Tip-Adapter[11]와 logits을 통합하는 과정에 대해 설명한다.

2.1. GPT 활용 방법

GPT[14]는 데이터 셋의 클래스 이름을 활용하여 식(2.1)에서 해당 클래스에 대한 속성 정보를 획득한다. 예시로 Food101[15] 데이터 셋의 경우, 속성 정보는 표1과 같다.

$$Category = GPT-3(Commands). \quad (2.1)$$

클래스 당 해당하는 카테고리는 Json 파일로 저장하여 훈련 시 속성 정보에 Key-Value의 캐시 가중치로 사용한다. 예시로 Food101[15] 데이터 셋의 “Pizza” 클래스에 해당하는 속성 정보는 “Italian Cuisine”, “Vegetarian Dishes”, “Baked Dishes”, “Main Courses”, “Mild Dishes”, “Vegetarian Dishes”, “Hot Dishes” 이다.

2.2 제안된 모듈

2.2.1 Tip-Adapter[11]

퓨-샷 러닝의 데이터 셋에서 C 는 클래스들의 개수, K 는 샷의 개수를 정의한다. 총 CK 개의 라벨링된 데이터 셋 $D \in R^{CK \times 3 \times H \times W}$ 이 주어지며, CLIP의 이미지 인코더를 통해 CK 개의 특징을 추출한다.

$$F = Clip_{img}(D), F \in R^{CK \times d} \quad (2.2)$$

텍스트의 경우 데이터 셋에 해당 프롬프트 템플릿을 적용하여 각 클래스 명을 기재한 후, Clip 텍스트 인코더를 통해 각 클래스에 속하는 특징을 추출한다.

$$W_c = Clip_{text}(Text_c), W_c \in R^{1 \times d} \quad (2.3)$$

$f_{test} \in R^{1 \times d}$ 는 이미지 인코더로 추출된 테스트 이미지의 특징이고, $L \in R^{C \times CK}$ 는 데이터 셋의 원-핫 인코딩된 라벨값이다. F 와 f_{test} 간의 유사성은 식 (2.4)의 affinity matrix를 통해 도출되며, 하이퍼파라미터 α 를 설정한 TAL(Tip-Adapter Logits)은 식 (2.5)와 같다.

$$A = \exp(-\gamma_1(1 - f_{test}F^T)) \quad (2.4)$$

$$TAL = \alpha AL^T + f_{test}W^T \quad (2.5)$$

2.2.2 AT-Adapter

2.1에서 GPT로 생성된 속성들은 텍스트 인코더를 통해 특징을 추출한다. 여기서 J 는 카테고리의 개수를 나타내며, K_j 는 j 번째 카테고리 해당하는 속성값들의 개수를 의미한다. 각 카테고리에 따른 행렬 가중치 W^j 는 식 (2.6)과 같이 표현된다.

$$W_{k_j}^j = Clip_{text}(t_{k_j}^j), j \in [1, 2, \dots, J], K \in [K_1, K_2, \dots, K_j] \quad (2.6)$$

$$W^j = Concat([W_1^j, W_2^j, \dots, W_{K_j}^j])$$

각 j 번째의 카테고리에 따른 W^j 와 f_{test} 간의 유사성은 식 (2.7)의 affinity matrix를 통해 도출된다.

$$B_j = \exp(-\gamma_2(1 - f_{test}W_j^T)), B_j \in R^{1 \times K_j} \quad (2.7)$$

B_j 의 값을 K 차원에서 각 속성에 해당하는 클래스 C 차원으로 변환하였다. 카테고리 j 에 따라 각 logits을 구한 후 이를 평균하여 ATAL(ATtribute-Adapter Logits)를 계산하였다. 여기서 $L_j \in R^{C \times K_j}$ 는 클래스에 따른 속성의 라벨값은 원-핫 벡터이다.

$$ATAL = \frac{1}{J} \sum_{j=1}^J B_j L_j^T, ATAL \in R^{1 \times C} \quad (2.8)$$

최종 logits은 하이퍼 파라미터 β 를 설정한 후 Tip Adapter Logits과 AT-Adapter Logits을 더해주었으며, argmax 값을 통해 테스트 이미지의 타겟을 예측하였다.

$$\begin{aligned} \text{Logits} &= TAL + \beta ATAL \\ &= \alpha AL^T + f_{test} W^T + \beta \left(\frac{1}{J} \sum_{j=1}^J B_j L_j^T \right) \end{aligned} \quad (2.9)$$

$$\text{Pred} = \text{argmax}(\text{Logits}) \quad (2.10)$$

3. 실험 결과

데이터 셋은 Food101[15], Oxford-Pets[16]로 실험을 진행하였다. Food101은 “a photo of a {classname}, a type of food”, Oxford-Pets의 경우, “a photo of a {classname}, a type of pet.”의 템플릿을 사용하였다. 하이퍼 파라미터 γ_1 는 1.89, γ_2 는 0.69, α 는 1, β 는 1.3으로 설정하였으며, Resnet50 백본으로 1, 2, 4, 8, 16 샷에 따라 실험을 진행하였다.

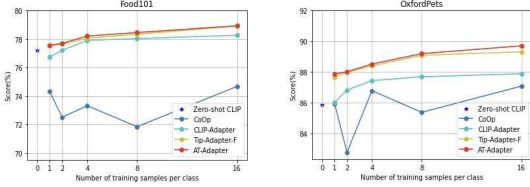


그림 2. Food101[15], Oxford-Pets[16] 데이터 셋의 성능 (%) 비교.

정량적 평가로 Tip-Adapter에 AT-Adapter를 적용하였을 때, 모든 샷에서 좋은 성능을 보였다. 특히, 16샷에서 Food101은 78.94%, Oxford Pets는 89.70%의 높은 성능을 달성하였으며, 이는 기존 Tip-Adapter 모델 대비 0.99%, 0.41% 성능 향상을 보였다.

표 2. Food101[15] 데이터 셋의 성능(%) 비교

Model	1	2	4	8	16
CoOp[4]	74.32	72.51	73.33	71.85	74.68
CLIP-Adapter[3]	76.75	77.22	77.92	78.04	78.28
Tip-Adapter-F[5]	77.55	77.65	78.06	78.35	78.93
AT-Adapter	77.57	77.71	78.22	78.47	78.94

표 3. Oxford-Pets[16] 데이터 셋의 성능(%) 비교

Model	1	2	4	8	16
CoOp[4]	85.92	82.74	86.77	85.36	87.07
CLIP-Adapter[3]	85.99	86.81	87.43	87.68	87.88
Tip-Adapter-F[5]	87.65	87.98	88.39	89.08	89.31
AT-Adapter	87.86	88.01	88.51	89.19	89.70

4. 결론 및 향후 연구

Chat GPT를 활용하여 속성 정보를 취득하는 방법에 대해 논의하였고, CLIP 모델에 제안한 AT-Adapter를 사용하여 Food101, Oxford-Pets 데이터 셋에 기존 모델 대비 높은 성능을 보였다. 하지만, Chat GPT에서 추출된 속성 정보가 불확실하거나 일부 클래스에 속성 정보가 편향되었을 경우 정확도가 크게 떨어졌으며, 또한 특정 데이터 셋에서는 속성 정보를 생성하지 못하였다. 향후에는 데이터 셋에 따른 속성 정보를 도출시키는 방법론과 속성 정보를 활용한 다양한 메서드를 개선하여 정확도를 높이는 연구를 수행할 것이다.

사사

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2023-RS-2023-00260098).

참고문헌

- [1]. Radford, Alec, et al. “Learning transferable visual models from natural language supervision.” *International conference on machine learning*. PMLR, 2021.
- [2]. Zhang, Pengchuan, et al. “Vinvl: Revisiting visual representations in vision-language models.” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
- [3]. Zhang, Yuhao, et al. “Contrastive learning of medical visual representations from paired images and text.” *Machine Learning for Healthcare Conference*. PMLR, 2022.
- [4]. Fürst, Andreas, et al. “Clobb: Modern hopfield networks with infoloob outperform clip.” *Advances in neural information processing systems* 35 (2022): 20450-20468.
- [5]. Yuan, Lu, et al. “Florence: A new foundation model for computer vision.” *arXiv preprint*.
- [6]. Jia, Chao, et al. “Scaling up visual and vision-language representation learning with noisy text supervision.” *International conference on machine learning*. PMLR, 2021.
- [7]. Li, Yangguang, et al. “Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm.” *arXiv preprint*

- arXiv:2110.05208*(2021).
- [8]. Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
 - [9]. Gao, Peng, et al. "Clip-adapter: Better vision-language models with feature adapters." *International Journal of Computer Vision*(2023): 1-15.
 - [10]. Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision*130.9 (2022): 2337-2348
 - [11]. Zhang, Renrui, et al. "Tip-adapter: Training-free clip-adapter for better vision-language modeling." *arXiv preprint arXiv:2111.03930*(2021).
 - [12]. Vedaldi, A., et al. "Convolutional architecture for fast feature embedding." *Cornell University*(2014).
 - [13]. Shen, Zhiqiang, et al. "Partial is better than all: revisiting fine-tuning strategy for few-shot learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 11. 2021.
 - [14]. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems*33 (2020): 1877-1901.
 - [15]. Bossard, Lukas, Matthieu Guillaumin, and Luc Van Gool. "Food-101-mining discriminative components with random forests." *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI* 13. Springer International Publishing, 2014.
 - [16]. Parkhi, Omkar M., et al. "Cats and dogs." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.

멀티모달 기반 IP 비디오월 제어 시스템

윤새봄⁰¹ 최문석¹ 우성현¹ 이정민¹ 문소정¹ 조영욱² 정설영¹

¹경북대학교

²주)하이퍼월

saebom.yoon@knu.ac.kr, cmschs0301@knu.ac.kr, wshdhkd@knu.ac.kr, djswpsk@knu.ac.kr

sjeong35@knu.ac.kr, c1122@naver.com, snowflower@knu.ac.kr

Multimodal-based IP Video Wall Controller System

Yoon Sae Bom⁰¹ Choi Mun Seok¹ Woo Seong Hyun¹ Lee Jeong Min¹ Moon So Jeong¹
Cho Young Wook² Jeong Seol Young¹

¹Kyung Pook National University

²Hyperwall

요약

우리가 세계를 경험하는 것은 멀티모달이다. 따라서 멀티모달 인터페이스를 활용한 제어 시스템은 유연한 사용자 경험을 제공할 수 있다. 현재 드론, 자율주행 모빌리티 등 스마트 시티 관제센터의 중요도와 사회안전망에 대한 높은 인식에 따라 관제센터 운영의 지능화 및 인간 친화적인 운영 방식이 전환이 필요한 상황이다. 본 논문에서는 AST(Audio Spectrogram Transformer) 기반 음성명령 인식 모델과 Whisper 기반 STT, Random Forest를 활용한 손동작 인식 모델을 결합한 멀티모달 기반 IP 비디오월 제어 시스템을 제안해 인간 친화적인 IP 비디오 월 컨트롤러 제어 시스템을 제안해 초 지능형 관제센터를 구현하는 초석을 마련한다.

1. 서론

우리를 둘러싼 세계 속엔 다양한 모달리티가 존재한다. 일반적으로 모달리티는 어떤 일이 발생하거나 경험되는 방식을 말한다. 대부분의 사람은 이것을 감각과 연관시킨다. 인공지능이 우리의 세계를 이해하기 위해서는 멀티 모달 메시지를 해석하고 추론할 수 있어야 한다. 그러기 위해서는 여러 감각에 대한 정보를 처리하고 연관시킬 수 있어야 한다[1].

본 논문에서는 음성 인식 및 제스처를 이용한 제어 시스템을 제안한다. 실시간 STT(Speech-To-Text)을 위해 자동음성인식을 위한 사전 훈련된 Whisper 모델과 ffmpeg프로그램을 합병하여 사용한다. 손동작 인식에 대해서는 인식 속도와 정확도를 모두 고려하여 머신러닝 모델을 선택한다. 최종적으로 음성과 손동작 모델들을 합병하여 멀티모달 기반 제어 시스템을 구축한다. 멀티모달기반 제어 시스템에서는 손동작과 음성을 인식하여 마우스와 키보드가 제어되고 이는 관제센터의 CCTV 모니터 시적인 입력으로 효율적으로 운영되게 도와주고 콘서트, 교육 등의 몰입형, 체험형 콘텐츠를 제작하는 데 활용도를 높일 수 있다.

2. 관련 연구

웹 환경에서 장애인의 디지털 기기 접근성 향상을 위한 멀티모달 인터페이스 연구[2]는 음성 명령식으로 Google API STT를 사용했다. 본 연구에 사용되는 모델은 음성 명령 인식에 특화된 모델로, 더 나은 성능을 보여줄 것으로 예상된다. 외부 환경에 강인한 딥러닝 기반 손 제스처 인식[3]은 GooLeNet을 활용한 손 제스처 인식 모델을 제안한다. 본 연구에서는 제스처 인식에 딥러닝보다 가벼운 머신러닝 방법론을 적용하여 더 빠른 인식 속도를 기대할 수 있다.

3. 시스템 상세 설계

3.1 실시간 음성명령 인식 시스템 구현

3.1.1 오디오 분류 모델 선정

음성 명령 인식을 위해 ‘speech_commands v0.02’ 데이터 셋을 사용한다. 오디오 분류를 위한 모델로 AST: Audio Spectrogram Transformer[4]에서 제안된 AST, facebook사의 Wav2Vec2-Base, openai사의 whisper-small을 사용한다. 성능은 데이터셋 안의 테스트 데이터셋 부분을 사용한다. 표1은 오디오 분류모델들의 Accuracy 성능이다. Wav-2Vec2과 AST모델은 앞서 언급한 데이터셋으로 미세 조정되어 배포된 모델이 이미 존재한다. 그래서 batch size를 줄여 성능 향상을 시도했지만, 기존 모델들보다 성능이 좋지 않았다. Whisper 모델은 미세 조정되어 배포된 모델이 존재하지 않는다. 따라서 이를 미세 조정하여 Accuracy 0.906을 달성했다. 하지만 기존에 배포된 AST 모델의 성능이 가장 좋아 이를 오디오 분류 모델로서 멀티모달 시스템에 적용한다.

Models	Wav2Vec2	AST	Whisper small
기존 미세조정된 모델	0.972	0.981	
직접 미세조정된 모델	0.902	0.889	0.906

<표1. 모델별 Accuracy 성능 비교>

3.1.2 AST-finetuned-speech-commands-v2 활용 실시간 음성명령 인식 모델

앞서 선정한 오디오 분류 모델과 실시간 음성 스트림을 위한 ffmpeg 프로그램을 합병하여 실시간 음성 명령 인식 모델을 구축한다. 합병 시 성능 최적화를 위한 매개변수 실험 결과는 표2와 같다. 특정 명령어의 확률인 prob_threshold는 0.8, 오디오

오 데이터를 처리할 때 사용되는 덩어리의 길이(초 단위)인 chunk_length_s는 1.5, 오디오 스트림에서 얼마나 자주 데이터를 읽을지(초 단위)를 결정하는 stream_chunk_s는 0.2로 최적화되었다. 그리고 음성 명령어와 실제 동작을 표3과 같이 대응시킨다.

정확도	prob_thres hold	chunk_len gth_s	stream_ch unk_s
95.55	0.8	2.0	0.25
96.66	0.8	1.5	0.25
74.44	0.8	2.5	0.25
81.10	0.8	2.0	0.3
94.44	0.8	1.5	0.3
65.55	0.8	2.5	0.3
94.44	0.8	2.0	0.2
97.76	0.8	1.5	0.2
64.44	0.8	2.5	0.2

<표2. 모델별 Accuracy 성능 비교>

음성 명령어	실제 동작
forward	중 인
backward	줌 아웃
two	더블 클릭
up	위
down	아래
right	오른쪽
left	왼쪽
follow	마우스 드래그
stop	멀티모달 시스템 종료

<표3. 음성 명령어 정의>

명령어를 chunk_length_s 단위로 stream_chunk_s마다 받아오기 때문에, 한 명령어를 말할때 여러번 인식되는 문제가 발생한다. 그래서 인식된 명령어를 리스트에 추가하고, 리스트의 맨 끝부분 데이터와 그 이전 데이터가 다를 때만 명령어로 인식을 받는 형태로 알고리즘을 구현했다.

3.2 실시간 손동작 인식 시스템 구현

3.2.1 손동작 인식 모델 학습 및 테스트

CPU 노트북 환경에서의 실행 속도를 고려하여 머신러닝 기법을 적용하였다. KNN, DecisionTree(이하 DT라고 한다.), PCA+DT, Randomforest(이하 RF라고 한다.)대해 학습 및 테스트 과정에서 측정한 정확도와 속도는 다음과 같다.

	KNN	DT	PCA +DT	RF
elaptime of train	13.721	39.034	10.148	19.904
accuracy (%)	99.87	97.37	97.75	100

<표 4. ML 모델 학습 시간과 정확도>

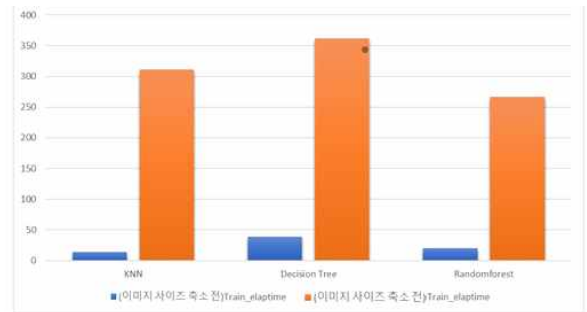
<표 4>를 참고하면, 머신러닝 모델 학습 시간은 PCA로 차원을 축소하고 DT로 학습을 시킨 PCA+DT가 학습 시간이 가장 적게

걸렸으며, 정확도는 RF와 KNN이 가장 높았다. <표 5>에서는 테스트 과정에서 KNN이 정확도 측면에서 가장 우수하였고 DT는 속도 측면에서 가장 좋게 측정되었다.

	KNN	DT	PCA +DT	RF
elaptime of test	2.586	0.0459	0.0966	0.1
accuracy (%)	100	67	67	95

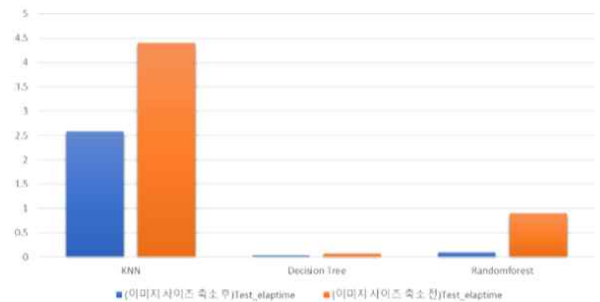
<표 5. ML 모델 테스트 시간과 정확도>

PCA를 적용하여 차원을 축소한 모델에서 정확도가 개선되지 않는다는 점을 통해, PCA 기법을 적용하는 대신 기존 이미지 크기를 (480, 290)에서 (145, 240)으로 줄여서 다양한 머신러닝 모델에 대해 학습 및 테스트를 진행하였다.

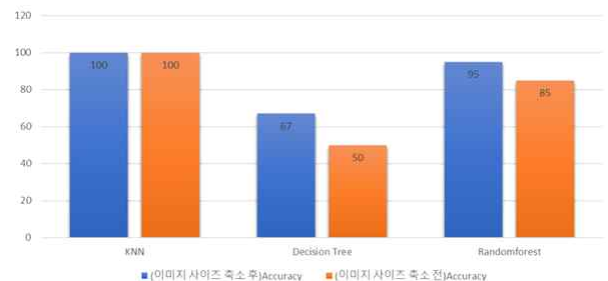


<그림 1. 이미지 축소에 따른 Train_elaptime 변화>

<그림 1>에서는 이미지 크기를 줄였을 때, 학습 시간이 현저하게 감소하는 것을 보여준다.



<그림 2. 이미지 축소에 따른 Test_elaptime 변화>



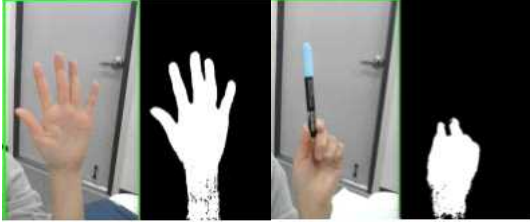
<그림 3. 이미지 축소에 따른 Accuracy 변화>

<그림 2>과 <그림 3>를 통해 이미지 크기를 줄였을 때, KNN과 Rf의 경우 테스트 시간이 크게 줄었다. 정확도 측면에서는 KNN은 정확도가 유지되었고 Rf의 경우 95에서 85로 줄었다. 속도와 정확도를 모두 고려하였을 때 KNN과 DT에 비해 Rf 모델이 멀티모달 시스템에서 가장 적합한 모델이라고 판단한다.

3.2.2 컴퓨터 비전을 통한 손 모양 이미지 처리

컴퓨터 비전을 사용하여 손 모양을 인식하는 과정은 다음 순서로 진행된다.

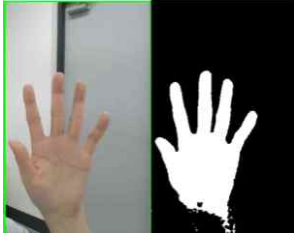
(1) 비디오 출력 화면에 손이 있는 영역과 아닌 영역으로 구분하여 나눈다.



<그림 4. 컬러 좌표계 적용>

(2) <그림 4>에서는 컬러 좌표계를 적용한다.

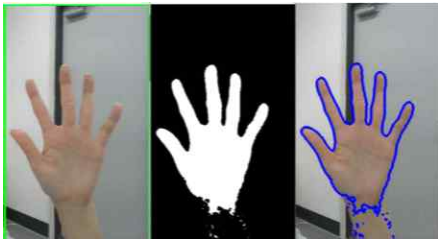
HSV는 색상(Hue), 채도(Saturation), 명도(Value)의 좌표를 써서 특정한 색을 지정한다. HSV를 통해 살생을 지정하여 살색 손만 감지된다.



<그림 5. 모폴로지 팽창, 침식 적용>

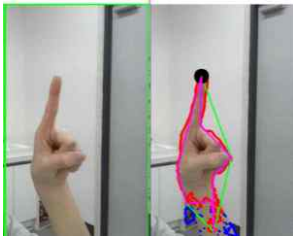
(3) <그림 5>에서는 모폴로지 팽창, 침식을 적용한다.

Morphology(형태학)는 이미지를 Segmentation 하여 단순화, 제거, 보정을 통해 형태를 파악하는 목적으로 사용된다. 커널의 모양을 타원형으로 하여 원본 이미지의 각 픽셀에 커널을 적용하여 팽창한다. 결과적으로 경계가 좀 더 부드러워지고 구멍이 메꾸어지는 효과가 있다. 이미지에 Dilation(팽창)을 적용 후, Erosion(침식)을 적용하면 이미지의 전체적인 윤곽을 파악하기에 적합하다.



<그림 6. 외곽선 검출 Contoures을 적용>

(4) <그림 6>에서는 외곽선 검출 Contoures를 적용한다. 이진 영상에서 흰색 객체의 외곽선을 표현한다.

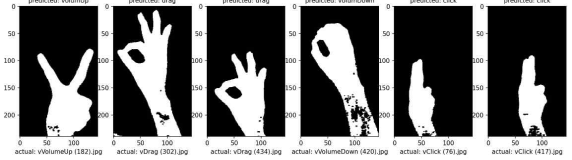


<그림 7. 손가락 끝점 좌표값 적용>

(5) <그림 7>에서는 손가락 끝점 좌표값 가져온다.

객체 궤적의 의미와 윈도우 상에서의 의미를 일치시켜야 한다. 이를 위한 매핑 과정에서 손가락 끝점에 대한 좌표값을 적용한

다.



<그림 8. 손모양 라벨>

(6) <그림 8>은 구별이 잘되는 다양한 손모양을 선정한다. 머신러닝 이미지 분류 모델에 손모양에 해당하는 동작을 나타내는 라벨과 함께 지도학습을 시킨다.

3.3 학습한 머신러닝 이미지 분류 모델에 손모양 테스트

손동작	키보드 및 마우스 제어
	<ul style="list-style-type: none"> ·해당 손동작에 음성으로 up을 입력하면, Volume 크기가 증가한다. ·해당 손동작에 음성으로 down을 입력하면, Volume 크기가 감소한다.
	<ul style="list-style-type: none"> ·해당 손동작에 음성으로 backward라고 입력하면, 화면이 Zoom out된다. ·해당 손동작에 음성으로 'forward'라고 입력하면, 화면이 Zoom in된다.
	<ul style="list-style-type: none"> ·해당 손동작에 음성으로 stop이라고 입력하면, 손동작 및 음성 인식 모드가 종료된다.
	<ul style="list-style-type: none"> ·해당 손동작을 취하면 마우스기 손 끝점을 따라 이동한다. ·해당 손동작에 음성으로 left를 입력하면, 마우스가 좌클릭 된다. ·해당 손동작에 음성으로 right를 입력하면, 마우스가 우클릭 된다. ·해당 손동작에 음성으로 더블을 입력하면, 마우스가 더블 클릭 된다.

<표 6. 손동작 및 음성 제어 기능 task>

<표 6>는 각각의 손동작과 음성에 대응되는 마우스와 키보드 동작에 대한 정의를 설명해 준다.

3.4 손동작과 음성 합성 시스템 구성

손동작 인식 모델과 음성 인식 모델 두 모델을 병합하는 시스템을 구현하는 과정에서 속도가 더 느려져, 비디오 캡처에서

손동작 흐름이 계속 끊기면서 출력되었다. 이 문제점을 해결하고자 두 모델을 각각의 프로세스로 실행시켰다. 음성 모델의 경우 인식한 음성 label 값을 voice_label.txt 파일에 시간과 함께 기록하여 저장하였다. 손 모양 인식 모델에서 손 모양을 인식하면 voice_label.txt 파일에서 실시간으로 업데이트된 음성 인식 정보를 읽어와서 손동작과 음성에 해당하는 마우스 및 키보드 제어 동작을 실행하게 하였다.

4. 시스템 구현 결과

팀원 A,B,C,D,E를 대상으로 손동작 및 음성을 접목한 마우스, 키보드 제어기능의 인식률을 측정한다. 각 기능 동작을 10번씩 시도했을 때, 시도한 10번 중 몇 번 제대로 동작했는지를 바탕으로 인식률을 구한다. 제스처 모델 중 KNN(정확도가 제일 우수, 속도가 느림), RF(정확도와 속도가 가장 적절하게 높음), DT(속도가 가장 좋음)는 정확도와 속도 측면에서 각각의 장단점을 가지고 있다. 이 세 모델에 대해 인식 속도와 인식 정확도를 모두 고려한 결과, 속도가 가장 좋았던 DT의 손동작과 음성 인식률은 <표 8>와 같다. 팀원들의 평가를 조사하였을 때, RL 모델에서 인식률이 더 좋았다는 의견이 우세했다. 그래서 <표 7>을 근거로 RL을 제스처 모델로 사용하기로 하였다. KNN의 경우에는 속도가 너무 느려, 비디오 캡처 화면에서 손동작 인식 화면이 계속 끊겨 실험을 중단했다. 음성 인식의 경우 파라미터 값으로 sampling_rate은 0.8, chunk_length_s은 1.5, stream_chunk_s은 0.2로 최적화시킨 실시간 음성명령 인식 모델을 멀티모달 시스템에 적용했다. 볼륨 제어부분에서는 손동작이 인식의 성능이 좋지 않아 이를 음성만으로 제어하는 것이 적합하다고 생각했다. 한편, 손동작과 음성을 병합하는 시스템을 구성하기 위해 서로 다른 프로세스에서 두 모델이 각각 동작하게 하였다. 음성 모델이 인식한 음성 결과를 파일에 저장하고, 손동작 모델과 파일을 공유하게 했다. 손동작 모델이 손 모양을 인식하고 공유 파일에서 실시간 음성 정보를 가져온다. 최종적으로 앞선 과정을 통해 멀티모달 시스템을 구축하였다.

인식률	마우스 제어				볼륨 제어		시작 및 종료		화면 조절	
	이동	좌클릭	우클릭	더블클릭	볼륨증가	볼륨감소	시작	종료	화면확대	화면축소
팀원 A	0.9	0.9	0.9	0.9	0.3	0.3	1	1	0.9	0.9
팀원 B	0.5	1	1	1	0.3	0.3	1	1	1	1
팀원 C	0.65	0.85	0.85	0.85	0.35	0.35	0.85	0.85	0.7	0.7
팀원 D	0.85	1	1	1	0.3	0.3	1	0.8	0.7	1
팀원 E	0.7	0.9	1	1	0.3	0.3	1	1	0.7	0.9

<표 8. Gesture모델: DT, Voice모델; finetuned AST + ffmpeg>

5. 결론

음성 명령 인식 부문에서 기존 연구의 STT를 사용하는 방법론보다 실시간 음성 명령 인식에 특화된 모델을 멀티모달 시스템에 적용했다. 한편, 제스처 인식 부분에서는 현재 많이 사용되고 있는 딥러닝 방식의 보편화된 멀티모달 적용 모델보다 더 가벼운 방식인 머신러닝 방식의 가벼운 멀티모달 시스템을 구현했다.

정확도와 속도 두 측면에서 모두 적합한 머신러닝 이미지 분류 모델을 선정하고 이미지 데이터 셋을 가공하였다 그리고 최적화된 하이퍼파라미터를 실험으로 찾아내어 정확도가 속도가 우수한 음성모델을 적용하였다. 두 모델이 하나의 파일을 공유하며 각각의 프로세스로 실행하였다. 이는 CPU 기반의 환경에서 머신러닝 모델을 적용함으로써 정확도와 속도 두 측면에서 성능을 향상할 수 있어 키보드 및 마우스 제어를 멀티모달 시스템 환경에 맞게 자연스럽게 실행시킬 수 있었다.

사사

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01082)

6. 참고문헌

[1] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607. 2019

[2] 박주현. "웹 환경에서 장애인의 디지털 기기 접근성 향상을 위한 멀티모달 인터페이스 연구." 국내박사학위논문. 2020

[3] 오동환, 이병희 and 김태영. (2018). 외부 환경에 강인한 딥러닝 기반 손 제스처 인식. 한국차세대컴퓨팅학회 논문지, 14(5), 31-39.

[4] Yuan Gong, Yu-An Chung and James Glass. AST: Audio Spectrogram Transformer. arXiv:2104.01778 (cs). 2021

인식률	마우스 제어				볼륨 제어		시작 및 종료		화면 조절	
	이동	좌클릭	우클릭	더블클릭	볼륨증가	볼륨감소	시작	종료	화면확대	화면축소
팀원 A	0.8	0.8	0.8	0.8	0.7	0.7	1	1	0.7	0.7
팀원 B	1	0.5	1	0.35	0.35	0.35	1	1	1	1
팀원 C	0.7	0.9	0.9	0.9	0.4	0.35	1	1	0.8	0.8
팀원 D	1	0.8	0.8	0.8	1	1	1	1	0.8	1
팀원 E	0.9	0.7	0.8	1	0.3	0.3	1	1	0.7	0.7

<표 7. Gesture모델: RL, Voice모델; finetuned AST + ffmpeg>

Bert 기반 오픈 도메인 뉴스 이벤트 추출 방법

황해루이, 정태선

아주대학교 대학원 인공지능학과

huanghairui523@ajou.ac.kr, tschung@ajou.ac.kr

Open-Domain News Event Extraction Method Based on Bert

Hairui Huang, Taesun Chung

Department of Artificial Intelligence, Ajou University Graduate School

Abstract

The open-domain event extraction task aims to identify and extract various types of event information from predefined text. The task is usually based on methods such as pre-training or neural topic modeling. However, there are a number of problems with existing methods. First, Existing pre-trained models suffer from insufficient feature vector extraction and excessively high embedding dimensions. Second, Existing methods are not rich enough in semantics and lack syntactic structural information, resulting in poor readability of results and insufficient extraction accuracy. Therefore, to address these issues, this paper first improves the open-domain event extraction method based on the neural topic model of BERT. Secondly, it dynamically integrates semantic and syntactic dependency information to obtain rich semantic and syntactic features, in order to further improve the model performance.

1. Introduction

With the development of Internet technology, information resources have become more abundant, and a greater amount of information appears in the form of unstructured text. Internet news is characterized by large amount of information, high timeliness, fast turnover and high credibility level, which makes it an important medium for obtaining information on events. The increasingly diverse types of news and the huge scale of news in the open news environment have created problems such as information overload and redundancy. It affects the news collection, judgment and processing. The text data to be processed for open-domain news event extraction is huge but the number of event types involved is very small, resulting in a very sparse amount of effective data to be processed.

The purpose of event extraction is to extract structured information from unstructured text, describing who, when, where, what, why, where, and how events occur in the real world. Existing event extraction tasks can be categorized into traditional and open domain event extraction. Traditional event extraction looks for words that belong to a specific

event pattern, and the extraction targets such as time, place, person, action. In open-domain event extraction tasks, events are usually treated as a set of related topic descriptions, constructed using classification or clustering, which extracts event information without specifying the event type^[1-3]. After research, it is found that most traditional event extraction corpora are based on knowledge, and the ACE2005 dataset has 33 event types, 599 documents, and 6000 labeled data, which means that there are less than 200 annotations for each event type on average, and at the same time, the distribution of event types in the corpus is very uneven. The event types in the open domain have noisy text and complex type structures, and it is cumbersome to annotate sufficient training data for each type of event type in real-world scenarios. Overall, open-domain event extraction is closer to real-life scenario information on social media. Therefore, exploring open-domain event extraction has important research implications.

The most important feature of open-domain event extraction is that it does not predetermine event types or patterns. It usually focuses on discovering new or unexpected events^[4-6], event text

generation^[7-8], and other general information extraction tasks^[9-11]. However, it also faces several challenges. First, the open domain involves a wide variety of domains and events, and accurate automatic event extraction requires understanding and representing different event types and domain knowledge. Second, open-domain event extraction needs to process a lot of text data but involves a small number of event types, resulting in very sparse effective data to be processed. Finally, there is a lot of noisy data, such as labeling errors and linguistic ambiguities, which need to be effectively filtered and corrected.

2. Related work

Presently, the research on technical methods for open-domain information extraction can be categorized into those based on clustering^[5], parsing^[12], Bayesian^[13] approaches, and the like.

2.1 Clustering-based approach

Social events are collections of various semantics, and related events tend to be cohesive. Therefore, density-based clustering algorithms can be used to detect and discover new events. Although the clustering method converges quickly and the algorithm still has the advantages of scalability and speed when dealing with large datasets, the method is not good at dealing with noise and outliers, and has difficulty converging on non-convex datasets.

2.2 Parsing-based approach

The results of syntactic analysis are often used to improve open domain event extraction methods. For example, verbs are useful for detecting event triggers, while nouns are useful for filtering event parameters. On the other hand, syntactic dependencies are useful for retrieving the same event roles and parameters in multiple sentences. The advantages of this method are rigor and precision, but the disadvantages are also very obvious, its cost is large.

2.3 Bayesian-based approach

Most open-domain event extraction Bayesian models assume that documents are joint distributions

of different types, slots, entities, and contextual features of events. This method has a reliable mathematical theory and robust classification efficiency, but still most of the methods do not pay enough attention to the semantic and dependent syntactic information of the sentence.

3. Proposed Methodology

3.1 Pre-training

Since the generated continuous contextual feature vectors need to be sent to the model as inputs, the previous pre-trained model cannot obtain a bi-directional feature representation that contains contextual information^[14], which can further affect the performance of the open-domain event extraction task. BERT has been proved to be useful for improving the accuracy of open-domain event extraction models in specific tasks after extensive research^[15]. BERT uses Transformer's encoder, which has the ability to utilize contextual information to obtain bidirectional representations of features, and adds word vectors, position vectors, and segmentation vectors, which fuses richer information and more accurately extracts the key trigger words and parameters, and can effectively improve the accuracy of open-domain event extraction models. Therefore, BERT is chosen as the pre-trained context-embedded language model to obtain the context representation.

3.2 Dimensionality reduction

The experiment revealed that utilizing a language model based on pre-trained contextual embeddings from BERT did not lead to significant improvements. This is attributed to the high dimensionality of the document embeddings, resulting in a sparse embedding space that requires dimensionality reduction. Although PCA and t-SNE are well-known dimensionality reduction methods, UMAP (Uniform Manifold Approximation and Projection) is a new technique proposed by Leland McInnes and others^[16], which has been demonstrated to preserve more local and global features in high-dimensional data. Although PCA and t-SNE are well-known dimensionality reduction methods. However, Umap is a new technique proposed by Leland McInnes et al^[17] that has been shown to preserve more local and

global features in high dimensional data. It constructs a weighted graph from high-dimensional data using edge strengths to represent the "proximity" of one point to another, and then projects this graph to lower dimensions, which has a number of advantages over t-SNE, the most obvious being that it has no computational constraints on the embedding dimensions and thus improves the speed, and better preserves the global structure of the data. Umap can be used in language models with different dimensional spaces, and it is based on reliable mathematical principles. Therefore, in this paper, we use the Umap method to reduce the dimensionality of document embedding. Figure 3-2 shows the overall model architecture.

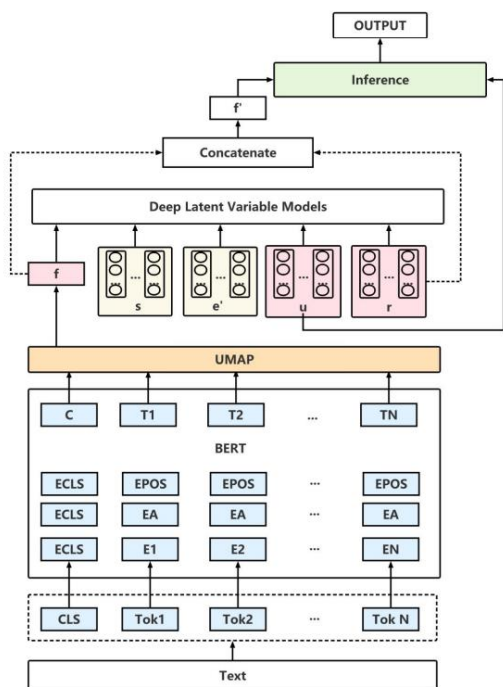


Figure3-1: Architecture diagram for open domain event extraction based on BERT neural topic modeling

3.3 Semantic understanding

The BERT model consists of a 13-layer Transformer encoder structure. However, during the training process, multi-layer Transformer encoding may suffer from gradient vanishing or exploding problem, which further loses semantic information. The Long Short-Term Memory (LSTM) model is proved to be an effective method to solve this problem. However, due to the problem that LSTM cannot capture the reverse sequence information, based on the semantic analysis, the Bi-LSTM(as

shown in Figure 3-2) model is selected, which combines the forward and backward LSTM, and can take into account the forward and backward information when processing sequence data, and is able to capture the long-distance dependency features.

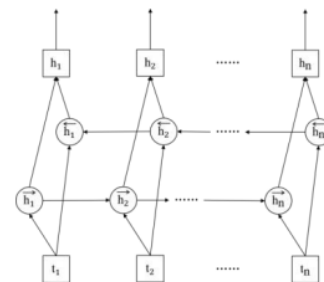


Figure 3-2: Schematic of BI-LSTM

4. Experiment

This experiment uses the GNBBusiness dataset. To validate the effectiveness of the proposed improved model in this paper, experiments will be conducted to compare it with the following models: Nguyen et al^[18] for the better performing model on the MUC4 dataset, von Luxburg et al^[19] proposed a clustering method that is very classical on the open domain event extraction task, and ODEE is the baseline model selected for this task, which was investigated on the open domain event extraction task. The results of the experiment are shown in Table 4-1.

	Schema Matching(%)		
	P	R	F ₁
Nguyen et al.(2015)	41.5	53.4	46.7
Clustering	41.2	50.6	45.4
ODEE	43.4	58.3	49.8
BERT-ODEE	42.1	56.6	48.3
BERT-Umap-ODEE	44.3	58.9	50.3
OurModel	44.9	60.1	51.3

Table 4-1: Overall performance of schema matching

It can be found from the experimental data that the model proposed in this paper has the best performance in terms of performance. In schema matching, the Precision (P), Recall (R), and F1 values of our approach are 44.9, 60.1, and 51.3, respectively. This represents an improvement of 1.5, 1.8, and 1.5 compared to the baseline model.

5. Conclusion

The open-domain event extraction task is very important for quickly understanding and summarizing a large amount of information. On the one hand, in order to solve the problems of insufficient extracted feature vectors, too high embedding dimensions and the same parameter type corresponding to multiple parameters, this paper proposes an improved method for open-domain event extraction based on the BERT neural topic model. On the other hand, in order to solve the problems of insufficient semantic richness and poor readability caused by the lack of syntactic structure information in the existing methods, this paper proposes an improved open-domain event extraction method that dynamically fuses semantics and dependent syntax. Based on the above two research methods, it can be verified through experiments that combining the neural topic model framework with BERT and incorporating rich semantic and dependent syntactic features have certain improvement effects, so the research on open-domain event extraction is of certain value and significance.

5. Reference

- [1]Yang Y, Wu Z, Yang Y, et al. A Survey of Information Extraction Based on Deep Learning[J]. *Applied Sciences*, 2022, 12(19): 9691.
- [2]Liu J, Min L, Huang X. An overview of event extraction and its applications[J]. *arXiv preprint arXiv:2111.03212*, 2021.
- [3]Li Q, Li J, Sheng J, et al. A Compact Survey on Event Extraction: Approaches and Applications[J]. *arXiv preprint arXiv:2107.02126*, 2021.
- [4]Veyseh A P B, Van Nguyen M, Min B, et al. Augmenting open-domain event detection with synthetic data from gpt-2[C]//*Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III* 21. Springer International Publishing, 2021: 644–660.
- [5]Peng H, Li J, Song Y, et al. Streaming social event detection and evolution discovery in heterogeneous information networks[J]. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2021, 15(5): 1–33.
- [6]Araki J, Mitamura T. Open-domain event detection using distant supervision[C]//*Proceedings of the 27th international conference on computational linguistics*. 2018: 878–891.
- [7]Fu Z, Bing L, Lam W. Open domain event text generation[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, 34(05): 7748–7755.
- [8]Martin L, Ammanabrolu P, Wang X, et al. Event representations for automated story generation with deep neural nets[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. 2018, 32(1).
- [9]Shen J, Zhang Y, Ji H, et al. Corpus-based open-domain event type induction[J]. *arXiv preprint arXiv:2109.03322*, 2021.
- [10]Naik A, Rosé C. Towards open domain event trigger identification using adversarial domain adaptation[J]. *arXiv preprint arXiv:2005.11355*, 2020.
- [11]Liu X, Huang H, Zhang Y. Open domain event extraction using neural latent variable models[J]. *arXiv preprint arXiv:1906.06947*, 2019.
- [12]Ritter A, Mausam, Etzioni O, et al. Open domain event extraction from twitter[C]//*Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012: 1104–1112
- [13]Wang R, Zhou D, He Y. Open event extraction from online text using a generative adversarial network[J]. *arXiv preprint arXiv:1908.09246*, 2019.
- [14]Yuan Q, Ren X, He W, et al. Open-schema event profiling for massive news corpora[C]//*Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018: 587–596.
- [15]Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure[J]. *arXiv preprint arXiv:2203.05794*, 2022.
- [16]McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. *arXiv preprint arXiv:1802.03426*, 2018.
- [17]McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. *arXiv preprint arXiv:1802.03426*, 2018.
- [18]Nguyen K H, Tannier X, Ferret O, et al. Generative event schema induction with entity disambiguation[C]//*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015: 188–197.
- [19]Von Luxburg U. A tutorial on spectral

clustering[J]. *Statistics and computing*, 2007, 17:
395-416.

NSGA-II를 활용한 인테리어 상품 최적 조합 추천 프레임워크

지다인^o(*), 김서윤(*), 김현민(*), 이정미(*), 장진영(*), 조유담(*), 최예림(*)(**)
서울여자대학교 데이터사이언스학과*, 주식회사 에이아이닷엠**
{dain,ji27, ksy9541, khm020401, jeongmi0711, jane050140, yudam,jo25}@gmail.com,
yerim.choi@swu.ac.kr

A Framework for Optimal combination recommendation of Interior Product Using NSGA-II

Dain Ji^o(*), Seoyun Kim(*), Hyunmin Kim(*), Jeongmi Lee(*), Jinyoung Jang(*),
Yudam Jo(*), Yerim Choi(*)(**)
Department of Data Science, Seoul Women's University*, ai.m Inc.**

요약

가구 추천 서비스는 현재 활발하게 제공되고 있다. 그러나 기존의 추천 서비스들은 단일 상품만을 사용자에게 추천하지만 가구는 다수의 상품을 함께 배치하는 것이 일반적이기에 사용자의 니즈를 충족시키기 어렵다. 본 연구에서는 사용자의 만족도에 영향을 미치는 구매 결정 요인인 예산과 리뷰를 고려하여 최적의 상품 조합을 도출하는 프레임워크를 제안한다. 실험에는 ‘오늘의 집’ 플랫폼에서 수집한 데이터를 사용하며, Vision Transformer과 Non-Dominated Sorting Genetic Algorithm II 을 활용하여 최적의 조합을 도출한다. 파레토 프론트(Pareto Front)로 확인한 결과 스타일 예측 및 조합 추천에서 우수한 성능을 보인다.

1. 서론

가구 추천 서비스는 현재 활발하게 제공되고 있다. 플랫폼 경제가 발전하고 오프라인 시장이 디지털로 확대됨에 따라 홈퍼니싱 시장도 활성화되고 있다[1]. 이러한 기술의 발달과 인테리어에 대한 관심이 증가하여 현재 모바일 앱 시장에는 다양한 종류의 인테리어 관련 어플리케이션 및 서비스가 등장하고 있다[2]. 예시로, 라이프스타일 플랫폼인 오늘의 집은 한 상품에 대해 유사한 상품을 추천해주는 서비스를 제공하고 있다.

하지만, 가구는 단일 상품만을 배치하는 것이 아니라 다수의 상품을 함께 배치하는 것이기 때문에 단일 상품만을 추천해주는 서비스는 사용자의 니즈를 만족시키지 못한다. 가구 배치는 상호 조화가 필요한 요소로, 함께 어우러지면 조화로운 효과를 낼 수 있지만, 그렇지 않은 경우도 존재한다[3]. 홍승현 외(2022)에 따르면 사용자들이 가장 기대하고 있는 서비스는 구매하고자 하는 가구 선택 시, 해당 가구와 어울리는 가구 또는 색 조합 등을 추천해주는 서비스로 나타났다[4].

또한, 1인 가구가 증가하고 있는 추세에서 1인 가구를 고려한 상품 판매가 중요해지고 있다. 인테리어 플랫폼 사용자들이 가장 선호하는 콘텐츠는 ‘중고거래’로, 필요한 가구를 저렴하게 사거나

편하게 팔 수 있다는 비용적 측면에서 장점이 존재한다[4]. 따라서 1인 가구를 위해 비용적 측면과 가구 조합을 고려한 추천 시스템이 필요하다.

인테리어 분야 외에도 다양한 분야에서 조합을 제시하는 연구가 진행되고 있다. 예를 들어, 패션 분야에서는 퍼스널 컬러를 무료로 분석할 수 있는 온라인 플랫폼을 제작하여 비용적 측면과 퍼스널 컬러를 고려한 색상의 의류를 추천하는 서비스를 제공하는 연구가 있다. 해당 선행 연구에서는 k-means 클러스터링을 사용하여 이미지에서 주요 색상을 찾아 퍼스널 컬러를 측정하고, 이에 맞는 색상의 의류를 추천한다[5]. e-스포츠 분야와 관련된 선행연구에서는 플레이어의 의상 조합에 도움을 주고자 협업 필터링에 기반한 게임 커스터마이징 아이템 조합 추천 방안을 제안한다. 해당 선행 연구에서는 아이템 기반 협업 필터링 방식을 사용하여 상품 간의 유사도를 구하고, 이에 맞는 의류 아이템 조합을 추천한다[6]. 이와 같은 연구들은 모두 조합을 고려한 연구를 진행 하였지만, 사용자가 원하는 예산과 만족도가 높은 상품을 고려하지 않는다는 한계점이 존재한다. 따라서, 본 연구는 기존 선행 연구의 한계를 극복하기 위한 대안으로 개인의 구매 결정 요인인 예산과 리뷰를 고려해 최적화된 상품 조합을 도출하는 프레임워크를 제안한다.

논문의 구성은 다음과 같다. 제2장에서는 프레임워크와 방법을 설명한다. 제3장에서는 실험을 위해 사용된 데이터를 제시하고 결과 분석을 진행한다. 끝으로 제4장에서는 본 연구의 요약과 한계점을 논의하며 논문을 마무리한다.

2. 방법

2.1 최적 조합 추천 프레임워크

본 연구에서 제안하는 프레임워크는 그림 1과 같다. 상품 데이터셋과 사용자의 입력 값(기존에 구매한 상품 URL, 구매하고자 하는 카테고리, 예산)을 활용하여 최적의 상품 조합을 제시한다. 카테고리 개수 및 예산 제한을 제약 조건으로 설정하고, 두 개의 목적함수(스타일 유사도 및 리뷰 평점 최대화)를 설정한 유전 알고리즘인 NSGA-II(Non-Dominated Sorting Genetic Algorithm II)를 활용한다. 이를 위해 상품 이미지를 라벨링하여 ViT(Vision Transformer)모델을 훈련시키고, 상품 크롤링 데이터셋에 적용하여 상품의 스타일을 예측한다.

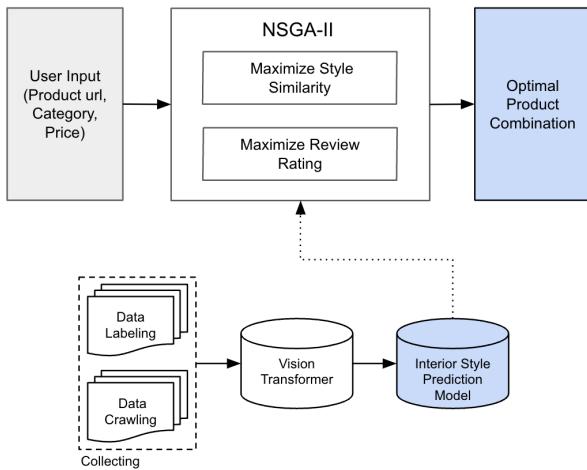


그림 1 인테리어 상품 최적 조합 프레임워크

2.2 최적 조합 도출

NSGA-II를 활용하여 최적의 조합을 제시한다. NSGA-II는 다목적 최적화 문제를 풀기 위한 다목적 유전 알고리즘의 대표적 방법이다[7]. NSGA-II는 일반적으로 목적치를 최소화 또는 최대화 시키는 최적화 알고리즘으로, 연산의 복잡도를 개선함과 동시에 우수해를 강조하고 공유 변수를 미리 선택할 필요가 없다는 장점을 지닌다[8]. 따라서 본 연구 목적에 적합한 NSGA-II를 이용하여 제약 조건을 충족하는 최적의 상품 조합을 찾기 위해 다목적 최적화를 진행한다.

NSGA-II의 실행 단계는 그림 2와 같다. 먼저, 스타일과 카테고리 설정 제한 및 제약조건을 고려하여 초기 모집단을 무작위로 생성하고, 생성된 개체들에 대해 목적함수를 계산하여 해당 개체의 적합도를 평가한다. 두 번째로, 개체들을 지배 등급에 따라 등급화하여 이를 기반으로 파레토 프론트를 형성한다. 세 번째로, 토너먼트 선택을 사용하여 부모 개체가 선택된 후 교차와 돌연변이를 통해 자식 개체가 생성된다. 마지막으로, 목적함수를

기반으로 적합도를 평가한 후 개체 병합 및 세대 교체가 이루어진다.

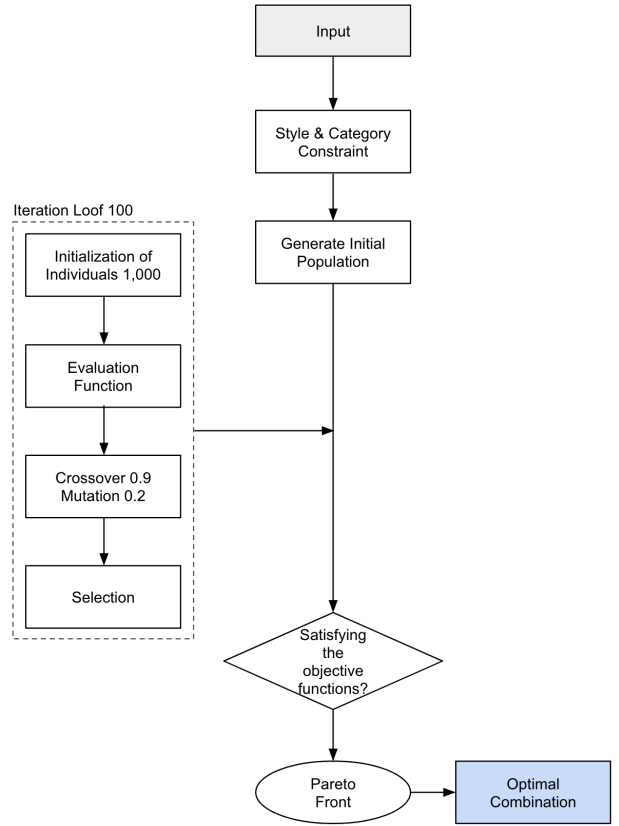


그림 2 NSGA-II 알고리즘 순서도

NSGA-II 설계를 위한 매개변수 설정은 표 1과 같다. NSGA-II 알고리즘에서 사용할 초기 개체 수를 1,000개로 설정하였고, 두 부모 개체에서 자식 개체를 생성하는 교차의 확률은 0.9, 개체의 유전자를 돌연변이시킬 확률은 0.2로 하였다. 세대를 100으로 설정하여 새로운 개체 집단을 형성하고 평가하는 프로세스를 진행한다.

표 1 NSGA-II 설계를 위한 매개변수 설정

	개체 수	교차	돌연변이	세대
NSGA-II	1,000	0.9	0.2	100

목적함수는 수식과 같다. 수식 (1)의 경우, 각 상품의 스타일 확률 벡터 값을 활용하여 스타일 유사도를 최대화한다. 스타일 확률 벡터 값은 예측된 스타일을 기반으로 추출한 값으로, 동일한 스타일인 경우에도 각 스타일에 대한 값의 차이가 존재한다. 이를 통해, 사용자가 보유하고 있는 상품과 유사한 상품을 찾고자 한다. 수식 (2)의 경우, 상품 리뷰 평점과 상품 리뷰 총 개수를 곱한 값을 로그 스케일로 변환하여 최대화한다. 이를 통해, 상품의 리뷰 평점과 리뷰 총 개수 간의 상관관계를 고려하면서, 사용자에게 높은 만족도를 제공하는 상품을 찾고자 한다.

$$\text{Maximize } \sum_{i=1}^N \sum_{j=1}^N \cos(S_i, S_j) \quad \dots\dots\dots (1)$$

1. N : 상품의 총 개수
2. S_i : 상품 i 의 스타일 확률 벡터
3. S_j : 상품 j 의 스타일 확률 벡터

$$\text{Maximize } \log(r_i \times m_i) \quad \dots\dots\dots (2)$$

1. r_i : 상품 i 의 리뷰 평점
2. m_i : 상품 i 의 리뷰 총 개수

제약조건은 다음과 같다. 다음의 두 가지 제약조건을 충족시키는 초기 모집단을 생성한다. 제약조건 1의 경우, 선택 가능한 상품 카테고리 수를 2개 이상부터 5개 이하로 제한한다. 제약조건 2의 경우, 조합된 상품의 가격 총합을 사용자가 지정한 예산 이하로 제한한다.

2.3 인테리어 스타일 예측

각 상품의 스타일을 예측하기 위해 ViT(Vision Transformer) 모델을 활용한다. ViT 는 자연어 처리에서 사용되는 트랜스포머를 이미지 분류 분야에 도입한 모델로 이미지 관련 태스크에서 뛰어난 성능을 보인다[9]. 본 연구에서는 Hugging Face[10]에서 제공하는 Auto-Train을 활용하여 사전 학습된 ViT 모델(google/vit-base-patch16-224)로 스타일 예측을 진행한다. 선행연구에 따라 가구는 사용 용도와 위치에 따라 종류가 굉장히 다양하므로[11], 채혜진(2019)에서 주로 언급된 5가지 카테고리 (Bed Frame, Chair, Lamp, Sofa, Table)로 제한하고[12], 서지우(2023)에서 제시된 선행연구의 기준을 참고하여 빈도수가 7 이상인 것을 기준으로 스타일을 Casual, Classic, Modern, Natural, Romantic 으로 총 5가지로 분류하였다[13]. ViT 모델을 학습시켜 총 5가지 스타일 중 한 가지 스타일로 예측한다.

3. 결과

3.1 실험 데이터

실험 데이터는 2023년 9월 30일 인기순을 기준으로 웹크롤링을 통해 수집한 오늘의 집 상품 데이터 셋이다. 카테고리, 브랜드, 상품명, 상품 번호, 리뷰 평균 별점, 리뷰 수, 정가, 판매가, 이미지 URL, 배송 타입, 스크랩 수, 위시 수, 무료 배송 여부, 판매 여부 총 15개의 항목을 수집하여 활용하였다. 5개의 카테고리(Bed Frame, Chair, Lamp, Sofa, Table)에 해당하는 상품 데이터를 수집해 총 39,960건의 상품 데이터를 수집하였다. 스타일 분류 모델을 학습하기 위해 상품명, 카테고리, 스타일, 판매가, 상품 URL, 이미지 URL, 색상 총 6개의 컬럼을 가진 5,000건의 라벨링 데이터 셋을 생성하였다.

3.2 스타일 예측 결과

인테리어 스타일 분류 모델을 통해 사용자가 원하는 예산과 스타일을 고려하여 5가지 카테고리 안에서 사용자의 요구에 맞는 조합을 보여줄 수 있도록 하였다. 이를 위해 ViT 모델을 훈련 및 학습을 진행하였고, 스타일 예측을 위해 학습한 ViT 모델의 성능은 표 2와 같다. ViT 모델은 5가지 카테고리에 대해 전체적으로 0.82

이상의 성능을 보였다. 이는 ViT 모델이 사용자의 니즈에 부합하는 맞춤형 상품을 제공하는 데에 효과적으로 활용될 수 있음을 보여준다.

표 2 ViT(Vision Transformer) 성능 결과

ViT	Bed Frame	Chair	Lamp	Sofa	Table
Test Accuracy	0.92	0.85	0.88	0.90	0.82
Macro F1 Score	0.92	0.83	0.80	0.89	0.78
Precision	0.92	0.88	0.88	0.90	0.87

3.3 조합 최적화 결과

조합 최적화를 평가하기 위한 방법으로 파레토 프론트(Pareto Front)를 사용하였다[7]. 모든 목적함수를 동시에 최대화하는 것은 현실적으로 어렵다. 여러 후보해 중에서 다른 해에 지배되지 않는 해를 ‘비지배해(Non-Dominated Solution)’라고 한다. 비지배해로 분류된 해들 중에서, 다른 해에 의해 더 나은 것으로 지배되지 않는 최적해를 ‘파레토 최적해(Pareto Optimal Solution)’라고 한다. 파레토 최적해는 목적함수와 제약조건에 의해 형성된 공간에서 생성되며, 이 공간을 ‘파레토 최적 프론트(Pareto Optimal Front)’라고 한다. 따라서, 파레토 최적 프론트 상의 해는 모든 목적함수에 대해 더 나은 해가 존재하지 않는 최종 선택된 해를 의미한다[7].

파레토 프론트(Pareto Front)를 이용하여 확인한 결과는 그림 3과 같다. 각각의 모든 점들은 생성된 모든 개체 집단을 나타내며, Non-Pareto Front Solutions 는 개체를 표현한 점이다. Pareto Front 는 목적함수와 제약조건에 의해 형성된 공간인 파레토 최적 프론트 상의 개체를 나타낸 점이다. 최종적으로 Selected Combination 이 최적의 조합으로 추천된 개체이다. 따라서, 최종 선택된 해가 파레토 최적 프론트에 위치하여 유의미한 결과를 도출하였다.

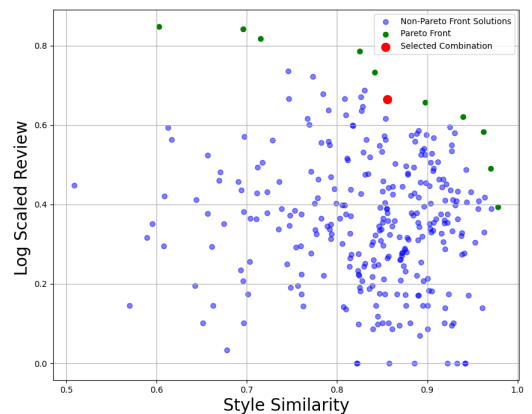





그림 3 인테리어 상품 조합의 파레토 최적해

실행한 인테리어 상품 조합 추천 결과는 표 3과 같다. 사용자가 기존에 가지고 있던 모던 스타일의 조명을 입력하고, 예산은 300,000원으로, 카테고리는 의자와 테이블로 지정하였다. 이에 대한 결과로 동일한 모던 스타일의 의자와 테이블이 추천 되었으며, 상품 리뷰 평점은 각각 4.55, 4.68의 높은 상품이었다. 또한, 이 추천된 상품들의 가격 총합은 174,000원으로, 설정한 예산 내에서의 조합을 제시하였다.

표3 인테리어 상품 조합 추천 결과

	Input(Lamp)	Output(Chair)	Output(Table)
Image			
Style	Modern	Modern	Modern
Price	65,000원	49,000원	125,000원
Review	4.5	4.55	4.68

4. 결론

본 연구에서는 NSGA-II를 활용한 가구 최적 조합을 도출하기 위해 라이프스타일 플랫폼인 ‘오늘의 집’에서 수집한 데이터를 활용하여 사용자가 선호하는 스타일과 상품 리뷰 만족도를 고려한 프레임워크를 제안한다. 구체적으로, 수집한 인테리어 상품 이미지 데이터를 기반으로 스타일 예측을 진행하였고, 제약조건이 있는 목적함수를 최대화하기 위해 NSGA-II를 활용하여 사용자의 선호를 고려한 최적의 조합을 제안한다. 파레토 프론트(Pareto Front) 그래프를 이용하여 목적함수를 고려한 조합을 시각적으로 표현한 결과, 제시한 최적의 조합이 파레토 최적 프론트 상에 위치하는 것을 확인하여 우수한 성능을 보이는 것을 확인할 수 있다. 연구에서 제시한 상품 조합 최적화 프레임워크는 향후 다양한 상품 조합에서 중요한 정보원으로 활용될 것으로 기대된다. 또한, 사용자의 정보를 반영하여 단일 상품이 아닌 조합 상품을 제시한다는 점에서 의의를 지닌다.

그러나, 해당 연구에는 몇 가지 한계점이 존재한다. 리뷰 개수가 적은 신상품의 특성상 리뷰 만족도 최대화 과정에서 고려도가 낮아 조합으로 제시되기 어렵다. 또한, 상품 카테고리를 5개로 설정하고 연구를 진행함으로써 카테고리의 다양성이 제한된다는 한계점이 존재한다. 향후 사용자 리뷰 데이터를 활용하여 사용자 유사도를 고려하고 상품 카테고리의 개수를 확장하여 연구를 진행한다면 더 유의미한 결과를 거둘 수 있을 것으로 기대된다.

감사의 글

본 연구는 서울여자대학교 SW중심대학추진사업단의 지원의

연구결과로 수행되었음 (2023)

참고 문헌

[1] 조혜진, and 이정교. “온라인 홈퍼니싱 플랫폼에서 공간콘텐츠의 전략적 요소에 관한 연구-오늘의 집 공간코디네이션 채널을 중심으로-.” 한국주거학회 32, no. 3 : 173-180. 2021.

[2] 조현경, and 서지은. “모바일 인테리어 어플리케이션의 기능성 평가.” 한국주거학회 32, no. 3 : 173-180. 2021.

[3] 홍상희, and 박병호. “디자이너의 경험을 토대로 한 거실용 가구의 디자인 디렉션 도출을 위한 제품형태와 이미지의 구조화.” 한국디자인문화학회 17, no. 3 : 775-785. 2021.

[4] 홍승현, and 권오정. “라이프 트렌드를 반영한 인테리어 플랫폼 콘텐츠 제안.” 한국주거학회 34, no. 2 : 299-303. 2022.

[5] 전석찬, 김건웅, 이우창, 서민성, and 조재혁. “퍼스널 컬러 진단을 통한 의상 컬러 조합 추천 서비스.” 한국정보기술학회, 653-657. 2023.

[6] 신정옥, and 임선영. “협업 필터링 기반 게임 커스터마이징 아이템조합 추천 방안 설계.” 한국정보처리학회 29, no. 2 : 659. 2022.

[7] 김솔지, 김태호, and 이흥철. "그리드 컴퓨팅을 위한 NSGA-II 기반 다목적 작업 스케줄링 모델." *한국컴퓨터정보학회논문지* 16, no. 7 : 13-23. 2011.

[8] 박순규, 이수복, and 이원철. "다목적 최적화를 위한 Goal-Pareto 기반의 NSGA-II 알고리즘." 한국통신학회논문지 32, no. 11 : 1079-1085. 2007.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint, arXiv:2010.11929, 2020.

[10] "HuggingFace." <https://huggingface.co/google/vit-base-patch16-224>.

[11] 정한결. "딥러닝 기반의 객체 탐지 기술을 활용한 인테리어 가구 인식 및 추천시스템과의 연동 방안." (석사학위, 한국과학기술원 : 경영공학부, 2022),31.

[12] 채혜진, and 오창섭. "1인 가구 주거 공간의 디자인 문화: 어플리케이션 <오늘의집> 의 '집들이' 게시물 중심으로." Archives of Design Research 32, no. 4 : 97-111. 2019.

[13] 서지우. "인구학적 특성에 따른 거실 공간실내 스타일 선호에 관한 연구 -20-30대 1인 가구를 대상으로-". (석사학위, 한양대학교 대학원, 2023),109.

엣지컴퓨팅 기반의 영상분석기술을 활용한

응급상황 이벤트 기술개발

백정훈^{01),2)}, 신용욱⁰⁰³⁾, 이채수⁰⁰⁴⁾, 이용주⁰⁰¹⁾

¹⁾경북대학교 일반대학원 컴퓨터학부, ²⁾(주)위니텍, ³⁾(주)테트라, ⁴⁾포인드(주)

bjh1974@naver.com, newshin@oktetra.com, ceo@4ind.co.kr, yongju@knu.ac.kr

Using Edge Computing-based Image Analysis Technology

Development of emergency event technology

Baek Jeong-hoon^{01),2)}, Sin Yonguk⁰⁰³⁾, Lee Chaesu⁰⁰⁴⁾, Lee Yongju⁰⁰¹⁾

¹⁾Kyungpook National University Graduate School, ²⁾WINITECH, ³⁾TETRA, ⁴⁾4IND

요 약

데이터 수집 및 실시간 계산에 중점을 둔 엣지 컴퓨팅을 활용하여, 경량화된 AI 영상분석기술 적용하여, 다양한 응급상황 이벤트를 감지하도록 한다. 이 논문에서는 이러한 엣지컴퓨팅 기반의 영상분석을 활용한 응급상황 감지기술을 개발하여, 각종 위험상황 대응 및 사고예방에 도움이 되도록 한다. 이러한 기술을 승강기라는 폐쇄 공간에서 일어날 수 있는 쓰러짐, 폭행, 화재등의 응급상황을 자동 감지하여, 선제적 대응 가능하게 하며, 국민의 생명과 재산을 지킬수 있도록 기술개발 적용한다.

1. 서 론

최근 다양한 분야에 엣지컴퓨팅을 활용한 기술을 도입하고 있다. 엣지단말장치를 통해 실시간 데이터 처리능력을 향상시켜 보다 빠른 판단을 하게 만든다. 이러한 엣지컴퓨팅 기술을 접목한 AI 영상분석기술을 적용하여, 응급상황 이벤트를 감지하는 기술을 개발한다. 4차산업혁명 기술이 적용된 응급상황 이벤트 기술을 활용한, 다양한 분야에 안전사고를 예방하게 한다.

이러한 기술적용을 통해, 각종 범죄와 사고를 예방할 수 있는 시스템이 개발돼 눈길을 끌고 있다.



[그림 1. 승강기내 사고에 대한 현황 및 문제점]

승강기에서 발생하는 각종 위험상황 대응 및 사고예방을 위해 감시,신고,초기대응,상황전파에서 긴급구조,사후관리까지 전 단계에 적용가능한 “승강기 스마트관제 플랫폼”에 기술을 적용하여 현재 시범운영중에 있다.

승강기 스마트관제 플랫폼은 AI, IoT, 센싱 등 4차 산업혁명 기술과 30억 건에 달하는 승강기 안전정보를 융합

해 승강기에서 발생하는 각종 위험상황을 예방할 수 있는 국민 안전을 위한 승강기 안정망이 조성되게 된다.

엣지컴퓨팅 기술을 적용한 IoT기반 지능형 안전장치 “Safety Hub”로 평상시 승강기 이용자에게 안전홍보영상, 디지털검사합격증 등의 승강기 안전정보는 물론, 지능형 AI 영상분석기술을 적용하여 “쓰러짐, 폭행, 화재” 등의 응급상황을 스스로 판단하여 신속하게 알려주는 방법을 제시한다.

2. 엣지컴퓨팅 기반 지능형 IoT 안전장치 개발

기존 범용프로세서(CPU)는 딥러닝의 높은 정확성은 대량의 데이터 처리,연산에 한계가 존재한다. 이를 극복하기 위해 인공지능 분야에 특화된 다양한 “연산유닛(Processing Unit)”이 등장한다.

인공지능(AD)시대 핵심인 신경망처리장치(NPU, Neural Network Process Unit)가 <그림 2>처럼 내장된 RK3588 칩셋 기반의 산업용 AI 임베디드 개발보드를 적용한다. 펜리스설계 및 KC인증등록한 EDGE PC를 개발한다.



[그림 2. AI 임베디드 개발보드 적용모델]



[그림 3. 지능형 IoT안전장치 구성내역]

[표 1. IoT안전장치 H/W, S/W 사양]

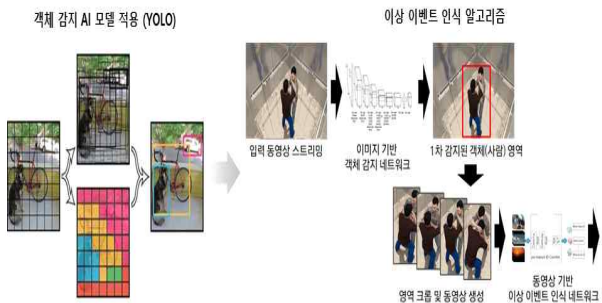
H/W	S/W	통신
-CPU: 8-core 64-bit -GPU: ARM Mali-G610 MP4 quad-core GPU -NPU: up to 6 TOPS -MEM: 8GB -저장공간: 64GB -모니터: 13인치 터치모니터 -Wireless: Support 2.4GHz, 5GHz -센서: 감지센서, 자이로센서 -카메라: IP카메라, WebCam -모델: LTE라우터모뎀	-O/S: 우분투 LTS 20.04 -S/W: 영상인식S/W, 음성인식S/W, 지능형 안전장치 S/W 센서감지 S/W -DBMS: MariaDB -WAS: Tomcat	기업 전용 LTE

엣지컴퓨팅 기술을 적용한 지능형 IoT안전장치 개발로, 승강기내에 설치되어, 카메라, 모션센서 및 가속센서등의 IoT 센서장치는 물론, “쓰러짐, 폭행, 화재” 등 비정상적인 응급 상황을 <그림 4>처럼 자동감지 및 선제적 상황대응 프로세서를 정립하여, 승강기내 발생하는 안전사고를 사전에 예방할수 있게 된다.



[그림 4. 응급상황 이벤트 검출 프로세스 아이디어]

이러한 응급상황 감지를 위해서는 <그림 5>처럼 다양한 상황에 맞는 영상학습 및 영상분석 경량화기술이 적용된다.

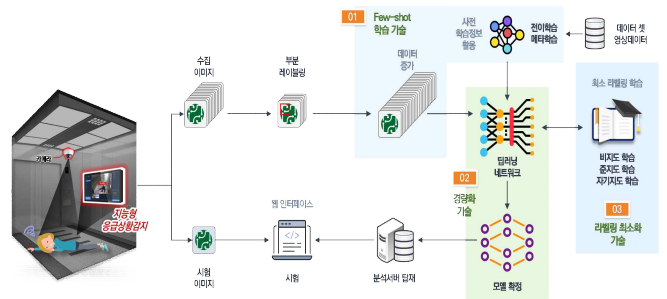


[그림 5. 객체감지 YOLO 모델 최적화 방안 도식화]

IoT안전장치 환경에 구동가능한 영상분석 자체 경량화 기술을 적용한다. <그림 6>과 같이 공공기관, 대학교, 연구소 등에서 공공 개방데이터 또는 연구 목적으로 배포한 영상, 이미지 사전데이터셋 구축하며, <그림 7>처럼 1FEW-SHOT 학습기술 적용 및 감지이벤트 검출이 가능한 딥러닝 기술, 효과적인 이벤트 검출을 위한 모델과 라벨링 최소화 기술을 적용한 엣지컴퓨팅 기반의 경량화 기술이 적용 된다.



[그림 6. 연구용 표준영상 사전데이터셋 활용]



[그림 7. 영상분석 경량화 기술적용]

이러한 영상분석 경량화 기술을 활용하여, 승강기내에 설치되는 엣지컴퓨팅 기반의 지능형 IoT안전장치를 통해 응급상황 이벤트를 <그림 8>과 같이 클라우드기반의 플랫폼 서비스를 통해 공단, 지자체, 관리주체등에게 응급상황을 웹브라우저상에 실시간으로 알려주는 통합관계 플랫폼 서비스를 제공하게 된다.



[그림 8. 응급상황 상황관계 플랫폼 서비스구성도]

1) few-shot learning : 한정된 샘플로부터 학습할 수 있는 머신러닝 방법론으로, 데이터가 부족한 상황에서 모델이 일반화하여 정확한 예측을 할 수 있게 해주는 방법론

3. 실증 및 검증결과

개발된 엣지컴퓨팅 기반의 IoT장치에 적용된 영상분석 경량화 기술을 활용하여, <그림 9>과 같이 대구교통공사 지하철 1호선, 2호선 및 지역 공공기관에 시범운영 설치되어 활용되고 있다. 2023년말 하반기에는 150대 추가 확대운영 설치되어, 전국 확대서비스를 위한 운영기반을 다지고 있다.

구분	기관명	대상 승강기
1	대구교통공사	1,2호선 지하철 역사내
2	부산시설공단	보행학교, 고속버스터미널 등
3	창원시청	여성회관 등
4	대구광역시교육청	도서관, 초·중·고 및 대학교
5	대구광역시 수성구청	복지관, 보건소, 평생학습센터 등
6	구미시청	봉곡도서관, 구미시설공단 등
7	인천교통공사	지하철 역사내
8	경주엑스포	엑스포 문화센터, 경주타워 등
9	대구광역시 달서구청	달구벌중합복지관, 달서아트센터 등
10	경원대학교	도서관, 기숙사 등

[그림 9. 시범운영서비스 설치현황]

<그림 10>처럼 KoLas 성능검증 기관인 ㄴ구미전자정보기술원(GERI)를 통해 성능시험을 통해, 쓰러짐 감지 95%, 싸움 감지 95%, 화재감지 90%이상의 정확도를 검증받았다.

검증항목	검정방법	단위	확인결과	비고
인공행동(쓰러짐)	인공행동	%	95	
인공행동(싸움)	인공행동	%	95	
인공행동(화재)	인공행동	%	90	
불법행위(쓰러짐)	인공행동	%	90	

[그림 10. 영상인식을 성능 검증결과]

향후, 지속적인 승강기 안전산업분야에 응급상황 자동 감지 플랫폼을 점차적으로 확대운영하도록 하여, 실시간 사고상황 알림,전파 등 승강기유지관리업체 디지털 업무환경 제공등으로, 사고·대응시간이 11분 이상 단축효과를 발휘하게 된다.

승강기에서 발생하는 각종 위험상황 대응 및 사고예방을 위해 감시·신고·초기대응·상황전파에서 긴급구조사후관리까지 수단단계에 적용되는 국민 안전을 위한 승강기 안전망이 조성된다.

ACKNOWLEDGMENT

본 과제(결과물)는 2022년도 디지털 공공서비스 혁신프로젝트 공모사업으로 지원을 받아 수행된 디지털 기반 승강기 통합관제 플랫폼 구축사업의 결과물입니다.

4. 결과 및 기대효과

최근 다양한 산업분야에도 엣지컴퓨팅 기술이 많이 활용 되어오고 있다. 특히, 엣지컴퓨팅 프로세서가 AI 기반 신속한 영상처리를 위해서 NPU 프로세스가 적용된 제품이 계속해서 출시되고 있다.

이러한 NPU 기반 엣지보드개발 및 영상처리 경량화기술을 적용한 AIoT 장치개발 및 프로세스 탑재로 지능형 응급상황 감지이벤트 기술이 개발되었다.

특정 산업분야인, 승강기에 도입하여, 최근 빈번하고 발생하고 있는 <그림 11>과 같이 승강기내 강력범죄 사전 예방 및 안전사고예방을 위해 응급상황 자동감지 플랫폼이 활용되어지고 있다.



"사람살려"...'똑똑한' 엘리베이터 폭행 자동 감지 / KBS 2023.07.26.

[그림 11. KBS 9시뉴스 응급상황 자동감지 플랫폼 소개]

참 고 문 헌

- [1] 지능형 사물인터넷 산업 동향, 정보통신기획평가원: <https://www.itfind.or.kr/publication/regular/weeklytrend/weekly/list.do>
- [2] Few shot learning 개념, 학습 방법, 프로세스 <https://www.thedatahunt.com/trend-insight/guide-for-few-shot-learning>
- [3] 승강기 통합관제플랫폼 구축 사업소개 (2022.10.4) <https://m.blog.naver.com/koelsa1671/222891627038>
- [4] 학술논문[한국정보기술학회 2023년도 하계종합학술대회 및 대학생논문경진대회] 지능형 영상인식 관계플랫폼을 활용한 승강기 응급구조절차 개선연구 (2023.06)
- [5] 딥러닝모델 경량화 기술 분석
- 지능형인프라기술연구단.기술보고서 2020.11.
- [6] "사람 살려"...'똑똑한' 엘리베이터 폭행 자동 감지
- KBS 뉴스 2023.07.26.

객체 위치 추정을 통한 효율적인 QR 코드 출입증 인식

오종현^{1,2} · 조원주² · 손민규² · 정순기¹

¹경북대학교 · ²씨앤피파트너스(주)

E-mail : spaceedu@knu.ac.kr / cnpartners@cnpartners.kr / sonmingyu93@cnpartners.kr / skjung@knu.ac.kr

Efficient QR code access recognition through object location estimation

Jong-Hyun Oh^{*1,2} · Won-Ju Cho² · Mingyu Son² · Soon Ki Jung¹

¹Kyungpook National University · ²CNP Partners Co.

요약

본 연구는 상단에 설치된 제1카메라와 전면에 위치한 제2카메라를 이용하여 출입 대상자 식별 프로세스의 효율성을 향상시키는 방법을 제안한다. 제1카메라는 객체의 위치를 추정하는 역할을 하며, 이 정보를 바탕으로 제2카메라는 설정된 범위 내의 이미지를 촬영하고 확대하여 QR코드를 인식한다. 이 시스템은 출입 대상자를 신속하게 인식할 수 있도록 설계되었으며, QR코드 인식을 통해 보안 및 출입 관리에 있어 높은 정확도와 효율성을 제공한다. 연구 결과는 특히 대규모 시설이나 빈번한 출입이 요구되는 환경에서의 적용 가능성을 시사한다. 이 시스템은 출입 관리 프로세스를 자동화하고, 인력 비용을 절감하며, 출입 절차의 신속함을 보장함으로써 전반적인 운영 효율성을 개선할 수 있을 것으로 기대된다

ABSTRACT

The research proposes a method to enhance the efficiency of the entry subject identification process using a top-mounted first camera and a front-facing second camera. The first camera plays a role in estimating the location of an object, and based on this information, the second camera captures and enlarges images within a set range to recognize QR codes. This system is designed to rapidly identify entry subjects, offering high accuracy and efficiency in security and access management through QR code recognition. The results of the study suggest potential applications, especially in large-scale facilities or environments requiring frequent access. By automating the access management process, reducing labor costs, and ensuring the swiftness of entry procedures, this system is expected to improve overall operational efficiency.

키워드

Object localization, QR-Code, Deep learning

1. 서론

현대 사회에서 QR 코드는 정보 공유 및 인증 방법으로 널리 활용되고 있다. 특히, 출입 관리 시스템에서 QR 코드의 중요성은 코로나19 팬데믹을 거치며 더욱 부각되고 있다. 이러한 출입관리 문제를 해결하기 위해 자동화된 무인 출입관리 시스템이 필요도 증가하고 있다.[1] 하지만, 기존의 QR 코드 인식 방식은 효율성과 정확성 면에서 한계를 보이고 있다. 이에 본 연구는 객체 위치 추정 기술을 통해 QR 코드 출입증 인식의 효율성과 정확도를 향상시키는 방법을 탐구한다.

기존 QR 코드 인식 시스템은 고정된 카메라 위치와 조명 조건 하에서만 최적의 성능을 발휘한다. 이로 인해 사용자가 QR 코드를 카메라에 정확히 맞추지 못할 경우 인식률이 급격히 떨어질 수 있다. 이러한 문제를 해결하기 위해, 본 연구는 객체 위치 추정 기술을 활용하여 QR 코드의 위치와 방향을 실시간으로 추정하고, 이를 기반

으로 더 빠르고 정확한 인식을 가능하게 하는 기술을 개발하고자 한다.

2. 객체 위치 추정 및 QR 검출

2.1 객체 위치 추정 기술을 이용하여 발 좌표 추정

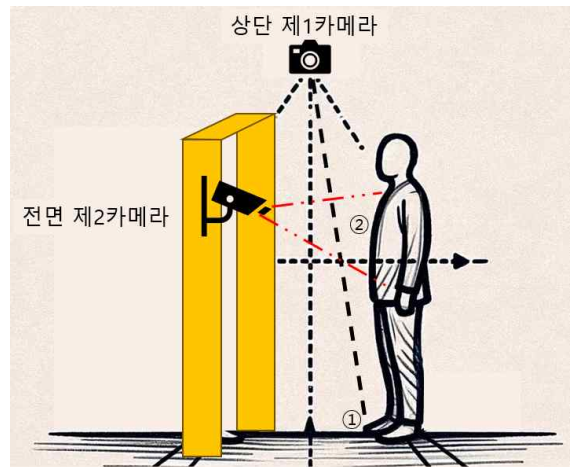


그림 1. 위치 추정 및 QR검출 컨셉

1) speaker

이 연구에서는 상단 카메라를 통해 촬영된 이동하는 사람의 이미지를 수집하고 라벨을 붙여 데이터셋을 만든 후, YOLO (You Only Look Once)[2] 모델을 사용하여 이 데이터셋을 이용해 모델을 훈련했다. 훈련된 모델은 각 사람에 대한 경계상자(Bounding Box)를 반환한다. 이후, 각 사람의 경계상자 정보를 기반으로 OpenPose, Skeleton-based 방식의 Pose Estimation 모델을 사용하여 신체 부위를 감지하고, 이 중에서 발(Foot) 부위를 식별한다. 이를 기반으로 객체 좌표를 추정 하게된다.[3] 그림 1의 컨셉과 같이 'Foot'는 객체가 서 있는 지점으로써 제 2카메라가 설정된 범위를 촬영하도록 명령하는 기준이 된다.

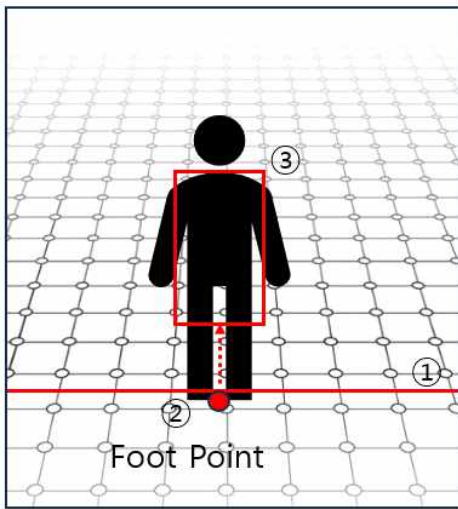


그림 2. Foot위치에 따른 ROI영역 예시

그림2는 제1카메라에 보여지는 화면으로 사람의 발이 제1카메라에 설정된 ①번 경계선을 넘을 경우 ②번 Foot Point를 잡게 되고 제2카메라는 ②번 기준에 따라서 임의로 정의된 위쪽 직사각형 ③번을 관심영역(ROI)로 추출 후 확대하여 QR코드를 인식하게 된다.

이때, 상단 제1카메라의 Foot FOV(Field of View)는 카메라 렌즈가 포착하는 바닥 영역의 범위를 나타낸다. 이는 카메라 렌즈의 시야각에 의해 형성된 원뿔 형태의 시각 영역으로, 카메라가 발의 이미지를 명확하게 캡처할 수 있는 최소 각도에서 정의된다. 이 각도는 카메라의 위치와 높이, 렌즈의 특성에 따라 달라지며, 발의 위치를 정확하게 탐지하고 추적하는 데 필수적인 요소이다. Foot FOV를 정의함으로써, 발의 위치, 움직임, 그리고 해당 공간 내의 다른 객체와의 상호작용을 이해할 수 있다.[3]

2.2 제1카메라 Foot Point에 대한 제2카메라 Calibration

제1 카메라와 제2 카메라의 Calibration은 카메라들 간의 상대적 위치와 방향을 결정하는 과정이다. 이를 위해, 두

카메라를 배치하고, 촬영된 이미지에서 패턴의 코너를 감지하고, 각 카메라의 내부 파라미터와 왜곡 계수를 계산한다. 내부 파라미터와 왜곡 계수는 아래와 같다.

- 내부 파라미터

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

- f_x 와 f_y 는 x 축과 y 축에 대한 초점거리
 - c_x, c_y 는 센서 상의 이미지가 형성되는 점
- 왜곡 계수

$$x_{distorted} = x(1 + k_1r^2 + k_2r^4 + k_3r^6)$$

$$y_{distorted} = y(1 + k_1r^2 + k_2r^4 + k_3r^6)$$

- (x, y) 는 왜곡되지 않은 픽셀 위치
- r 은 원점에서 픽셀까지의 거리

그 다음, 카메라들 사이의 상대적 위치(회전 및 이동 벡터)를 추정하고, 이미지를 정합하여 대응하는 점들이 같은 행이나 열에 나타나도록 한다. 본 연구에서 Calibration 과정을 OpenCV를 이용하여 파라미터를 추정하는 함수를 사용하였다.[4]

2.3 관심영역(ROI) 설정 및 QR검출 단계



그림 2. 이미지 범위에 따른 QR인식 성능 비교 실험그림

첫째, 관심 영역(ROI) 설정 단계에서는 객체 위치 추정 결과를 바탕으로 QR 코드가 존재할 가능성이 높은 특정 영역을 선택한다. 이 과정은 전체 이미지를 분석하는 것보다 연산 비용을 절감하는 데 도움이 된다.

둘째, 이미지 전처리 단계에서는 선정된 ROI에 대해 노이즈 제거, 명도 및 대비 조정과 같은 처리를 수행한다.

이러한 전처리 과정은 QR 코드 인식의 정확도를 높이는 데 중요한 역할을 한다.[5]

셋째, QR 코드 인식 단계에서는 전처리된 이미지 내에서 QR 코드를 인식한다. 일반적으로 사용되는 QR 코드 인식 알고리즘은 패턴 인식 기법을 통해 QR 코드의 위치, 크기, 그리고 방향을 정확하게 파악한다. 본 논문에서 QR코드를 검출하기 위해 Zxing[6] 알고리즘에서 제안하는 방법을 사용하였다.

마지막으로, 디코딩 단계에서는 인식된 QR 코드 내의 데이터를 추출한다. 이 데이터는 QR 코드에 암호화되어 있는 정보로, 디코딩을 통해 원래의 데이터 형태로 변환한다.

2.3 실험결과

	Full 이미지 QR추출	설정 이미지 QR추출
인식 성공 여부	실패	성공
처리시간(sec)	0.2024890	0.1555851

Full 이미지 QR추출시 QR코드 인식에 실패한 반면 설정 이미지 추출시에는 QR코드 인식 성공뿐만 아니라 처리시간도 23% 단축되었다.

3. 결론

본 논문에서는 딥러닝 기법을 이용하여 객체 추정 후 설정된 관심영역(ROI) 이미지에 QR코드 검출 시스템을 개발을 위한 환경을 구성하였다. 객체 탐지 후 전체 이미지에 대한 QR인식 보다는 관심영역(ROI) 이미지 설정에 따른 QR인식률과 인식속도의 효율성을 높아짐을 확인하였다. 향후 다양한 영상 데이터를 통한 실험으로 Calibration 과정 고도화 및 관심영역(ROI) 범위를 줄여가는 방법에 대한 추가적인 연구가 필요하다.

Acknowledgement

“본 논문은 과학기술정보통신부 및 정보통신기획평가원의 지능형화석신인재양성사업의 연구결과로 수행되었음”
(IITP-2022-RS-2022-00156389)

References

[1] 한국경호경비학회: 신현구, “방문객의 출입절차 자동화 시스템 도입을 통한 산업보안 수준 향상 방안”, 시큐리티 연구, 제63호, pp.165, 2020.

[2] YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, <https://arxiv.org/abs/2207.02696>

[3] 방송공학회: 김기철 외 6, “딥 러닝 및 칼만 필터를 이용한 객체 추적 방법”, 방송공학회논문지, 제24권, 제3호, pp. 501, 2019.

[3] 한국정보통신학회: 오종현, “상단 시야 인공지능 엡지 카메라를 이용한 출입 인가 시스템의 타인 따라가기 방지를 위한 객체 위치 좌표 추정 방법”, 춘계종합학술대회 논문집 제27권 1호, pp.319, 2023.

[4] 제어로봇시스템학회: 장예원, “OpenCV를 활용한 카메라 캘리브레이션과 마커 기반 증강현실”, 제38회 제어로봇시스템학회 학술대회, pp.477, 2023.

[5] 정보과학회논문지: 박동진, “기울기와 노이즈에 강인한 바코드 검출 시스템”, 시정보과학회논문지 제42권 제7호, pp.868, 2015.

[6] Zxing(Zebra Crossing) Project home, <http://github.com/zxing/zxing/>

거대 언어 모델의 효율적인 학습 보완을 위한

LangChain 활용 연구

곽건영⁰¹, 손홍석¹, 심규민¹, 채정민¹, 정설영¹, 김정근²

경북대학교 컴퓨터학부¹, Flash21²

krjs0907@gmail.com, bestof1226@gmail.com, gyu2781@gmail.com,

chaejm55@gmail.com, snowflower@knu.ac.kr, tsmall1000@naver.com

Study on the LangChain for efficiently complementary training of

Large language model

Gunyoung Kwak⁰¹, Hongseok Son¹, Gyumin Sim¹,

JungMin Chae¹, Seolyoung Jeong¹ and Jeonggeun Kim²

KyungPook National University¹, Flash21²

요 약

본 논문의 목적은 거대 언어 모델이 별도의 학습 과정을 거치지 않고도 특정 도메인 지식이나 학습 시점 이후의 정보를 적절하게 처리하는 것을 목표로 한다. 이를 위해 언어 모델에 RAG(Retrieval Augmented Generation), LLM Agent와 같은 연구 사항을 LangChain Framework를 활용하여 적용하였다. 이후 다양한 도메인 데이터에 대한 검증을 진행하였고 그 결과로 언어 모델이 기존에 대응하지 못했던 질문 사항에 높은 비율로 적절한 응답을 생성하는 것에 성공하였다. 이에 본 연구는 기존 언어 모델이 별도의 훈련 과정을 거치지 않고도 학습하지 않은 영역의 질의에 적절한 응답을 생성할 수 있음을 확인하였다는 점에서 의의가 있다.

1. 서 론

최근 거대 언어 모델의 부상으로 다양한 분야에서 이를 활용해 상용화하기 위한 시도가 증가하고 있다. 하지만 특정 도메인에 특화된 지식이나 학습 시점에 반영되지 않은 지식, 학습 시점 이후의 지식 등에는 적절하게 대응하지 못한다는 한계가 있었다.

그에 따라 특정 도메인의 지식을 학습시키기 위해 언어 모델을 미세조정(fine-tuning)하여 도메인의 지식을 학습시키는 방법이 있지만 이 같은 시도는 데이터셋의 준비, 고성능의 하드웨어, 시간 등을 필요로 한다는 단점이 있다. 또한 추가, 환율과 같이 유동적으로 변화하는 데이터에는 대응하지 못하며 정적인 데이터에만 대응할 수 있고 이를 해결하기 위해서는 학습 시점으로부터 갱신되는 정보를 지속해서 학습시켜 줘야 한다는 한계가 있다.

이에 본 논문은 LangChain Framework를 활용해 RAG(Retrieval Augmented Generation), LLM Agent와 같은 후속 연구 사항을 적용하여 별도의 학습 과정을 거치지 않고도 거대 언어 모델이 이와 같은 문제점에 유연하게 대처할 수 있도록 하는 것을 목표로 한다.

2. 관련 연구

2.1 RAG(Retrieval Augmented Generation)

해당 논문은 외부의 데이터베이스에서 정답을 도출하는 ODQA(Open-Domain Question Answering) 태스크와 대형코퍼스로 모델을 사전 학습시킨 LLM 연구를 결합해 기존의 Knowledge intensive Task의 성능을 향상한 연구이다.[1] 논문의 아이디어는 기존 ODQA 태스크의 Reader를 LLM에서 파생된 T5, GPT-3와 같은 Generator로 대체하는 방법론을 제시한 것이다. 이 연구를 기점으로 ODQA의 Task가 Span Prediction Task에서 Answer Generation Task로 전환되었고 다양한 생성 기반 태스크 수행이 가능해졌다. 또한 해당 연구의 Open-Domain QA Test와 Generation and Classification Test에서 ODQA의 DPR과 LLM의 T5와 비교했을 때 월등히 좋은 결과를 보여주었고 Retrieve된 지식이 없는 경우에도 11.8%의 정답률로 답변을 생성해 내는 결과를 보여주었다.

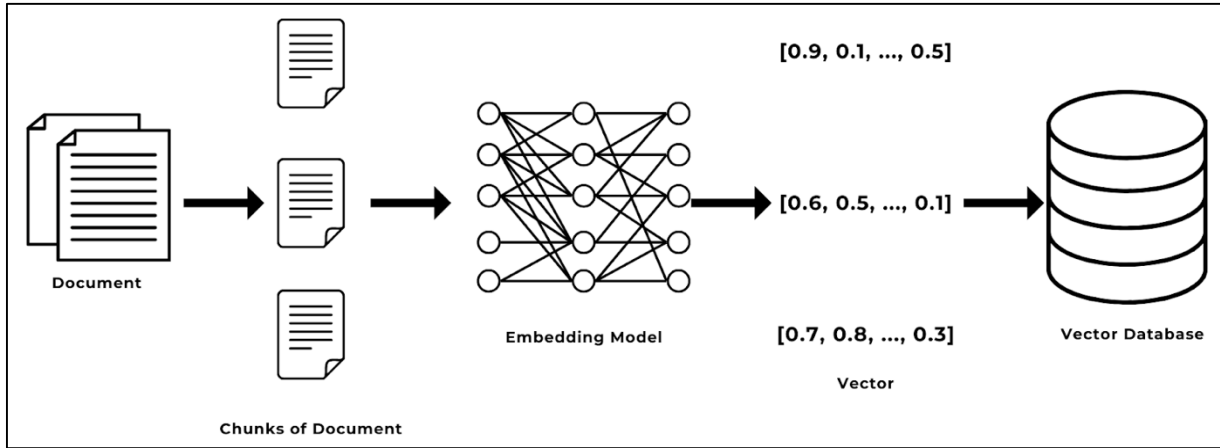


그림 1. 데이터가 임베딩되는 과정

2.2 Chain of Thought

해당 논문은 ‘Chain-of-Thought Prompting’을 사용하여 입력과 출력 사이의 중간 단계를 도입하여 모델의 산술 및 상식 추론 능력을 향상하는 방법을 제시하였다.[2] 제시한 방법을 통해 모델이 문제 해결을 위한 명확한 추론 단계를 거치게 되어 자연어 이해, 문제해결 및 상식적인 추론 능력이 향상되었다.

2.3 ReAct: Synergizing Reasoning and Acting in Language Models

ReAct는 ‘Chain-of-Thought’에서 제시한 방법에 Action을 추가하여 언어 모델이 환경을 인식하고 다음 행동을 결정하기 위해 지속해서 상호작용하는 방법을 제시한 연구이다.[3] 해당 기법으로 외부 지식을 활용해, 특정 작업을 수행하고 그 작업의 결과를 환경으로 다시 반영한다. 이렇게 제시한 방법은 유연한 방법으로 여러 언어 추론 및 의사결정 작업을 수행해 언어 모델의 성능을 한층 더 향상시켰다

2.4 Toolformer: Language Models Can Teach Themselves to Use Tools

Toolformer는 언어 모델의 성능 향상을 위한 방법론 중 하나로, 외부 도구를 활용하여 문장 생성에 참여할 수 있는 방법을 제시하였다.[4] Toolformer는 Self-supervised 방법으로 학습되어, 특정한 API만을 사용해 하나의 도메인에만 적용이 가능하거나, Zero shot이 불가능하다는 한계를 다양한 API 호출을 언제, 어떻게 할지를 학습하는 것으로 해결하였다.

3. 시스템 상세 설계

3.1 데이터가 임베딩되는 과정

데이터가 임베딩 될 때 그림1과 같은 과정을 거치며 상세 설명은 다음과 같다. 우선적으로 기존 데이터의 정보를 나누어서 청크라는 작은 단위로 변환 과정을 거친다. 그 이유로 청크는 추후 언어 모델에 문맥 정보로 제공되게 되는데, 언어 모델에는 토큰 한도 즉, 입력값에 대한 한계가 존재하기 때문에 이에 대처하기 위해 입력을 줄이는 방식이다. 그 후, 작게 나뉜 청크는 임베딩 모델을 통해 청크가 가지고 있는 데이터의 의미를 잘 반영할 수 있는 시맨틱 벡터로 변형된다. 변형된 벡터는 기존의 청크가 포함하고 있던 데이터와 함께 최종적으로 벡터 데이터베이스에 저장되게 된다.

일반적으로 청크가 클수록 보다 넓은 문맥에서의 답변이 가능하기에 청크 크기가 클수록 답변 능력이 좋아지지만, 언어 모델의 토큰 한도와 상충 관계에 있기 때문에 토큰 한도, 도메인의 특수성, 질문의 길이 등을 종합적으로 판단하여 적절한 크기의 청크 크기를 설정하는 것이 요구된다. 본 논문에서는 16K의 토큰 한도를 가진 “gpt-3.5-turbo-16k” 모델을 사용하여 토큰 한도 문제를 최소화하고자 하였으며 임베딩 모델은 “text-embedding-ada-002” 모델을 사용하였다. 또한, 청크 크기를 1000토큰, 청크 간 겹침 정도를 20토큰으로 설정하여 벡터 데이터베이스에 저장하였다.

위와 같이 구성된 데이터베이스에서 유동적으로 변하는 내용을 데이터베이스에 지속해서 반영하고 추후 일련의 과정을 거쳐 해당 정보를 언어 모델에 제공함으로써 정적인 데이터에만 대응할 수 있던 언어 모델이 유동적인 데이터에도 대응할 수 있게 하였다.

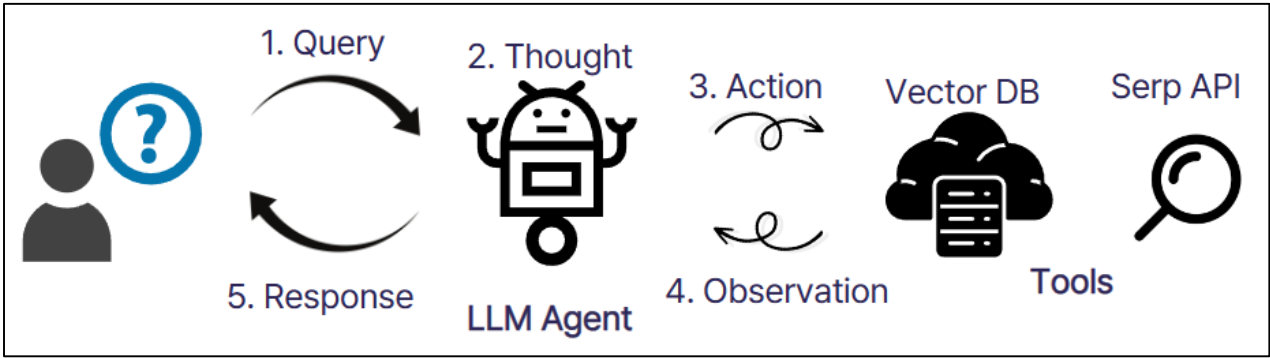


그림 2. 동작 구조

3.2 질문에 대한 응답을 생성하는 과정

그림2와 같은 과정을 거치며 작동 방식의 상세 설명은 다음과 같다. 언어 모델은 입력으로 들어온 질문(1. Query)에 대해 자신이 해야 할 일을 정의하는 과정(2. Thought)을 거친다. 그 후, 주어진 도구를 이용하여 정보를 얻는 행위를 하며(3. Action) 그에 따라 질문에 대한 관측값을 얻는다(4. Observation). 질문에 응답하기 위해 충분한 정보가 모일 때까지 2, 3, 4의 과정을 반복하며 충분한 정보가 모였다고 생각될 때, 해당 정보를 바탕으로 최종 응답을 생성한다(5. Response). 해당 과정에서 언어 모델에게 사용할 수 있는 도구로 벡터 데이터베이스, Serp API를 사용한 검색엔진을 주었다.

(i) 벡터 데이터베이스의 경우, 찾고자 하는 정보를 임베딩 모델을 통해 의미를 반영한 시맨틱 벡터로 변형한다. 그 후 해당 벡터와 벡터 데이터베이스 상에 존재하는 벡터들 간의 유사도를 측정 후 가장 유사한 3개의 청크를 뽑아내게 하였다. 이렇게 뽑힌 청크를 언어 모델에게 문맥 정보로 제공함으로써 특정 도메인 지식에 대한 질문에 학습 과정을 거치지 않고도 대응할 수 있게

하였다. 더불어 이 방식은 데이터베이스의 내용을 바꾸거나 다른 데이터베이스에 연결함으로써 새로운 도메인에 즉시 연결해 줄 수 있다는 장점이 있다.

(ii) 검색 엔진의 경우, 언어 모델이 찾고자 하는 정보의 내용을 웹상의 정보를 통해 찾아볼 수 있도록 하였다. 이를 통해 언어 모델이 학습되지 않은 정보나 데이터베이스에 존재하지 않는 정보에도 대응할 수 있을 것으로 기대된다.

4. 시스템 구현 결과

그림3에서 확인할 수 있듯이, 기존의 언어 모델이 대응하지 못했던 질문에 대해 벡터 데이터베이스, 검색 엔진을 이용하여 적절한 답변을 생성할 수 있음을 확인하였다.

위의 결과에서 더 나아가 언어 모델이 도메인에 특화된 지식, 학습되지 않은 지식에 대해 다양한 분야에서 잘 대응할 수 있는지 실험을 진행하였다. 보건의료정보원의 보건의료정보화 실태조사, 한국은행의 경제전망보고서, 과학기술정보통신부의 합성생물학 기술영향평가 보고서를 각각 시사, 경제, 생물 3종류로 분류하였고 이를

기존의 언어 모델
 Question = "2021년 11월 1일 비트코인의 가격은 얼마인가?"
 Answer: 죄송하지만, 현재의 날짜에 대한 정보는 제한된 범위 내에서만 제공됩니다. 비트코인의 가격은 시장 조건에 따라 변동되며, 실시간으로 확인해야 합니다. 비트코인의 가격을 확인하려면 실시간 가격 추적 웹 사이트나 암호화폐 거래소를 참조하시기 바랍니다.

Question = "현재 비트코인의 가격은 얼마인가?"
 Answer: 저는 현재의 비트코인 가격을 예측하거나 실시간으로 제공할 수 없습니다. 비트코인의 가격은 시장 변동성에 따라 지속적으로 변동하기 때문에, 실제 가격을 확인하기 위해서는 실시간 가격 정보를 제공하는 거래소 또는 금융 웹 사이트를 참고하시기 바랍니다.

RAG, Agent가 적용된 언어 모델
 Question = "데이터베이스에서 2021년 11월 1일 비트코인의 가격을 먼저 찾고, 검색을 통해 현재 한화 비트코인 가격을 찾아 두 값의 차이를 설명하라."

I need to find the historical price of Bitcoin on November 1, 2021, and the current price of Bitcoin in Korean won. Then I need to calculate the difference between the two prices and explain it in Korean.
 Action: 데이터베이스 조회
 Action Input: "데이터베이스에서 2021년 11월 1일 비트코인의 가격"
 Observation: '2021년 11월 1일 비트코인의 가격은 80,000,000.00 KRW입니다.'

Thought: I found the historical price of Bitcoin on November 1, 2021.
 Action: 검색 엔진 사용
 Action Input: "현재 한화 비트코인 가격"
 Observation: 46,204,980.75

Thought: Now I can calculate the difference between the two prices and explain it in Korean.
 Final Answer: 2021년 11월 1일 비트코인의 가격은 80,000,000.00 KRW이고, 현재 한화 비트코인 가격은 46,204,980.75 KRW입니다. 두 값의 차이는 33,795,019.25 KRW입니다.

그림 3. 실제 동작 예시

데이터베이스에 저장하였다. 그 후 각각의 데이터셋의 내용을 바탕으로 작성한 질문 중 기존 언어 모델이 적절한 응답을 생성하지 못한 10개의 질문을 RAG, Agent가 적용된 언어 모델에 재질문하여 실험을 진행하였다.

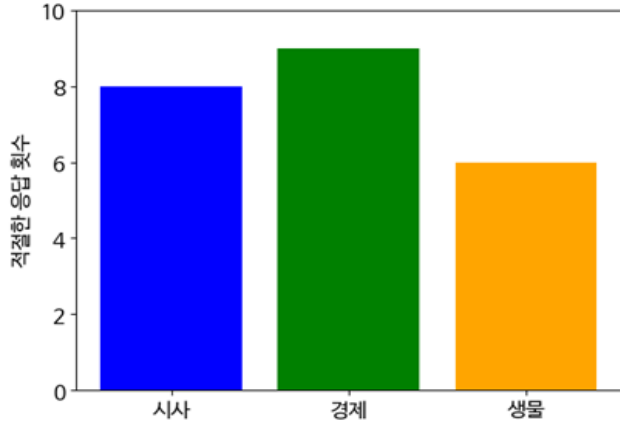


그림 4. 시사, 경제, 생물 분야에서의 적절한 응답 횟수

그에 따른 실험 결과는 그림4와 같다. 실험 결과를 관찰하였을 때, 전반적으로 높은 정확도로 응답을 생성하는 것을 확인할 수 있었다. 하지만, 상대적으로 도메인 지식이 어렵다고 생각될 수 있는 생물 분야 질의에서는 상대적으로 낮은 정확도를 보임을 확인할 수 있었다. 그에 대한 원인으로 일부 청크만을 보고 판단하여 응답을 생성하는 부분을 꼽아보았다. 도메인 지식이 어려울수록, 보다 많은 문맥 정보를 필요로 하지만 제한된 청크 크기, 개수로 인해 일부의 정보만으로 정답을 도출해 내기에 한계가 있었다고 예상된다.

5. 결론

본 논문에서는 Retrieval-Augmented-Generation, LLM Agent 연구 사항을 적용하여 언어 모델이 벡터 데이터베이스를 활용한 검색 증강, Serp API를 활용한 웹상의 정보 탐색을 통해 학습 과정을 거치지 않고도 학습하지 않은 정보를 처리할 수 있도록 하였다. 그에 따른 실험 결과로 언어 모델이 기존에 대응하지 못했던 질문 사항에 높은 비율로 적절한 응답을 생성하는 것을 확인하였다.

다만, 도메인 지식이 어려워질수록 답변 성능이 떨어짐을 확인하였고 주요 원인으로 언어 모델의 토큰 제한으로 인한 청크 크기, 개수의 제한에 따른 정보 부족을 원인으로 꼽았다. 해당 문제 사항이 존재하지만 최근 Ring Attention[5]과 같은 토큰 제한을 무한에 가깝게 늘리는 연구가 진행됨에 따라, 이와 같은 연구 사항이 적용된 언어 모델이 등장한다면 해당 단점을 보완할 수 있을 것으로 보인다.

사사문구

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

"이 연구는 과학기술정보통신부 및 정보통신산업진흥원의 "지역산업 SW인재양성 기반조성 사업" 지원으로 수행되었음."

참고문헌

- [1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Task", arXiv preprint arXiv:2005.11401, May 2020.
- [2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", arXiv preprint arXiv:2201.11903, Jan 2022.
- [3] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao, "ReAct : Synergizing Reasoning and Acting in Language Models", arXiv preprint arXiv:2210.03629, Oct 2022.
- [4] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, Thomas Scialom, "Toolformer: Language Models Can Teach Themselves to Use Tools", arXiv preprint arXiv:2302.04761, Feb 2023.
- [5] Hao Liu, Matei Zaharia, Pieter Abbeel, "Ring Attention with Blockwise Transformers for Near-Infinite Context, arXiv preprint arXiv:2310.01889, Oct 2023.

기계학습을 이용한 IoT 기반 재난 시스템 Kubernetes Custom Autoscaler 설계

박준희, 권영우

경북대학교 컴퓨터학부

devlos0322@gmail.com, ykwon@knu.ac.kr

Design of Kubernetes Custom Autoscaler for IoT-based Disaster Systems Using Machine Learning

JunHee Park

Department of Computer Science, Kyungpook National University

요약

이 논문은 IoT 기반 재난 대응 시스템의 비용 효율성과 안정성을 향상시키기 위해 기계 학습과 Kubernetes의 HPA 기능을 활용하는 새로운 Custom Autoscaler 설계를 제안한다. Custom Autoscaler는 재난 발생 전 시스템의 자원을 절약하고, 재난 시에는 예측 모델을 기반으로 한 신속한 scale-out을 통해 예상 리소스를 공급하여 인명과 재산 손실을 최소화하고자 한다. 본 연구는 기존 정적 리소스 할당 방식의 IoT 기반 재난 대응 시스템의 한계를 극복하고, 동적 자원 관리를 가능하게 하는 방법에 대해 제시한다. 이론적인 설계와 더불어 실제 재난 시뮬레이션 데이터를 사용하여 Custom Autoscaler의 효과를 검증할 프로토타입 구현을 통해 적용 가능성과 효과를 평가할 필요가 있다.

1. 서론

재난 상황 발생시, 신속하고 효과적으로 대응하는 것은 인명과 재산 손실을 최소화하는데 결정적인 역할을 한다. IoT(Internet of Things) 기반의 재난 대응 시스템은 이러한 상황에서 중요한 데이터를 수집하고 분석하여 실시간으로 정보를 모니터링하여 상황에 따라 적절히 대응할 수 있는 제반을 구축한다. 그러나 기존 시스템들은 재난 상황에 대응하기 위해 최대한의 정적 리소스를 제공하고 있다. 이에 본 연구는 재난 위험 이벤트 수신 시 기계 학습 기반 재난 상황에 따른 부하를 처리하기 위해 필요한 리소스량을 산정하고, 컨테이너 오케스트레이션 도구인 Kubernetes의 HPA(Horizontal Pod Autoscaler) 기능을 통해 자원 사용을 최적화하는 새로운 Autoscaler를 설계하고자 한다. 이를 통해 재난 발생 전 예측 모델을 바탕으로 평시에는 리소스를 절감하고, 이상 시에는 사전에 스케일링 함으로써 비용 효율성과 안정성을 동시에 달성하고자 한다. 본 논문은 이러한 시스템에 대한 설계를 제안한다.

2. 연구 배경 및 관련 연구

IoT 기반 시스템에서는 센서의 배터리 수명 또는 네트워크 비용, 데이터 처리비용 등 다양한 제약사항을

고려해야한다. 이러한 이유로 인해 관련 연구[1][2]와 같이 데이터 수집 주기를 상황에 따라 동적으로 조절하는 기술이 적용된다. 가변적인 수집 주기에 따라 시스템의 리소스를 최소화하여 비용을 절감하고 운영 시 안전성을 확보할 수 있는 방법은 클라우드 기반 환경에서 컨테이너 단위로 재난 시스템을 구성하는 서비스들을 배포하고, 컨테이너 오케스트레이션 도구인 Kubernetes를 통해 운영하는 것이다.

Kubernetes는 기본적으로 리소스의 사용량에 따라 컨테이너의 Replica를 조절하는 HPA를 내장하고 있다. HPA를 활성화시키면 Kube-controller-manager는 일정 주기(기본 주기: 15 초)에 따라 리소스의 상태를 확인하고 Replica 수를 조정한다. Replica 수를 조정하는 수식은 다음과 같이 정의된다.

$$R_{desired} = \text{ceil}[R_{current} * (\frac{M_{current}}{M_{desired}})]$$

$R_{desired}$ 는 목표로 하는 Replica의 개수이고, 이는 현재의 Replica 수인 $R_{current}$ 와 현재 리소스의 성능 값 $M_{current}$, 목표로 하는 성능 값 $M_{desired}$ 의 연산 결과를 올림 하여 정해진다.

이를 사용하면 센서의 데이터 트래픽에 따라 최적의 리소스를 사용할 수 있게 된다. 하지만 HPA 매트릭 풀링

주기와 컨테이너 로딩 지연 시간등으로 인해 Replica 수에 따라 생성된 Pod 가 트래픽을 처리하는데 지연이 생길 수 있다. 이 방식은 사전 연구 [3]에서 사후 대응 기법으로 분류되어 있으며, 점진적인 리소스 증가 시 대응이 가능하지만, 트래픽이 급격하게 변화할 경우 시스템 요청 처리에 취약하다는 한계가 있다.

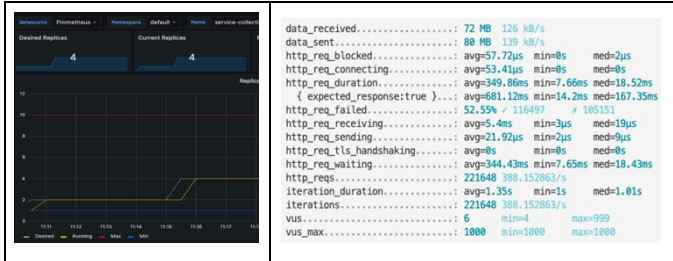


그림 1. (좌) HPA 를 통한 오토스케일링 예시, (우) 오토스케일링이 진행되는 동안 HTTP request failed 가 발생한 예시

그림 1 은 리소스 사용량을 통해 Autoscaling 이 수행되었지만, 그 과정에서 수신에 실패한 트래픽의 사례를 보여준다. 사전 연구[3][4]에 따르면 시스템 매트릭 기반 기계 학습을 통해 이전 데이터를 평가하고 향후 예측을 기반으로 미리 스케일링을 수행하는 사전 예방 기법을 제안하고 있다.

3. 제안 기법

본 논문에서는 사전 연구 [3][4]의 방식을 따라 기계 학습을 통해 HPA 를 수행하는 Custom Autoscaler 를 IoT 기반 재난 대응 시스템에 적용하면서, 재난 위험 이벤트를 통해 예측된 리소스를 즉각적으로 스케일링에 활용하는 방식에 대해 제안한다.

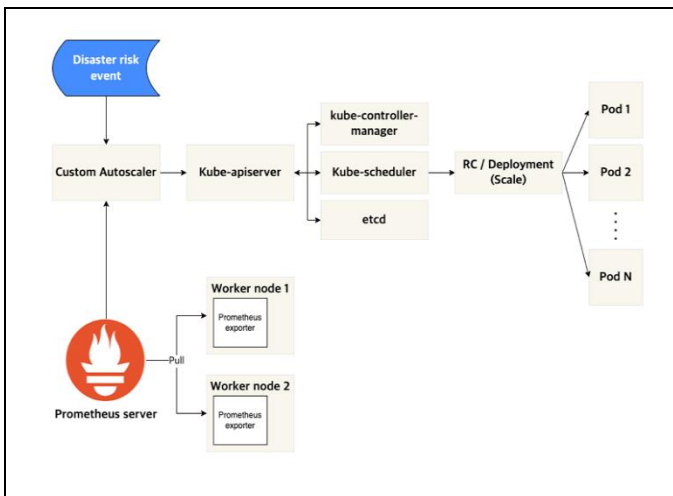


그림 2. Custom Autoscaler 시스템 아키텍처

그림 2 는 Kubernetes 와 매트릭을 수집하는 Prometheus server, HPA 를 포함한 kube-controller-

manager, 그리고 중간에서 스케일링을 증가하는 Custom Autoscaler 구조를 표현한 것이다. Prometheus server 에서 수집한 매트릭을 기반으로 Custom Autoscaler 내부에서 리소스 사용량을 예측한 후 kube-apiserver 를 통해서 리소스를 제어할 수 있게 된다.

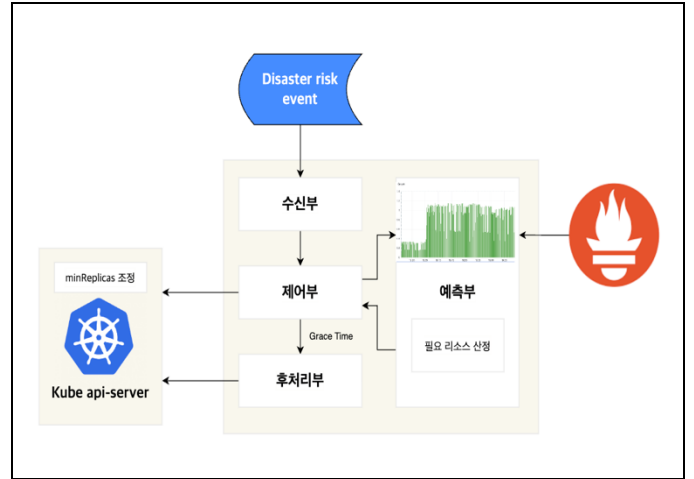


그림 3. Custom Autoscaler 의 구조

그림 3 과 같이 Custom Autoscaler 내부는 Prometheus 에서 수집한 매트릭을 기반으로 리소스를 예측하는 예측부, 외부에서 재난 위험 이벤트를 수신하는 수신부, 그리고 수신부와 예측부를 기반으로 HPA 를 제어하는 제어부, 그리고 재난 위험 종료 이벤트를 수신하여 리소스를 원래 상태로 되돌리는 후처리 부로 구성되어 있다. Custom autoscaler 가 재난 위험 이벤트를 수신하게 되면, 수집된 매트릭을 기반으로 필요한 리소스 수를 확보하기 위한 Replica 수를 산정한다. 산정된 Replica 수는 제어부를 통해 kube-apiserver 로 전달되어 HPA 의 minReplicas 옵션을 조정하게 되고, Scale-out 을 수행하게 되어 급격한 트래픽 변화에 대응할 수 있도록 한다. 이후 재난 위험 종료 이벤트를 수신하게 되면, Grace time 이후에 minReplicas 를 조정하고, Scale-in 을 수행하여 리소스를 평시 상태로 되돌리게 된다.

4. 결론 및 향후 연구

본 논문에서는 IoT 기반 재난 대응 시스템을 위한 Custom Autoscaler 를 설계하고 제안하였다. 이 Custom Autoscaler 는 재난 위험 이벤트를 감지 시, Kubernetes 의 HPA 기능과 기계학습 기반 리소스 예측 결과를 기반으로 스케일링을 수행하여 자원 활용을 최적화하는 방법을 제안한다. 재난 발생시 신속하고 효과적으로 대응할 수 있는 환경을 조성하여 재난 상황에서의 인명과 재산 손실을 최소화하며, 평시에는 리소스를 절약하여 운영 비용의 효율성을 높일 수 있다. 센서 데이터를 기반으로 한 실시간 예측 모델을 통해 정적 자원 배분 방식의 한계를 극복하고, 예측 가능한 동적 자원관리를 실현할 수 있다. 이러한 접근은 기존의

정적인 리소스 할당 방식을 개선하고, 재난 대응 시스템의 높은 가용성과 비용 효율성 사이의 균형을 찾는 데 기여할 수 있다.

본 연구는 향후 IoT 기반 재난 대응 시스템에 적용될 Custom Autoscaler 에 대한 제안을 증명하기 위한 프로토타입 구현이 필요하다. 재난 시뮬레이션 데이터를 통해 기본적인 기능을 검증하고, 이론으로 제시된 장점이 실제 환경에서 실현 가능성을 판단할 것이다. 또한 나아가 실제 재난 대응 환경에 Custom Autoscaler 를 적용하고, 그 효과를 평가하여 실제 작동환경에서의 문제점을 파악하고 개선할 수 있는 연구를 진행할 것이다.

사사

이 논문은 2023 년도 정부(과학기술정보통신부) 및 지자체(대구광역시)의 재원으로 (재)대구디지털혁신진흥원에서 주관하는 지역 디지털 혁신거점 조성지원사업의 지원을 받아 수행된 연구임(No. DBSD1-03)

참고 문헌

- [1] Mukesh Taneja, “A framework for traffic management in IoT networks”, IEEE, IC31, 2016
- [2] Simon Wong, James She, Kang Eun Jeon, “An Efficient Framework of Energy Status Reporting for BLE Beacon Network”, IEEE Internet of Things Journal, 2023.
- [3] Simon Shim, Ankit Dhokariay, Devangi Doshi, Sarvesh Upadhye, Varun Patwari, Ji-Yong Park, “Predictive Auto-scaler for Kubernetes Cloud”, IEEE SysCon, 2023.
- [4] Jonghwan Park, Jaegi Son, Dongmin Kim, “Resource Metric Refining Module for AIOps Learning Data in Kubernetes Microservice”, KSII, TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 17, NO. 6, 1545p, 2023.

GPT-4 텍스트 증강 데이터 기반 소규모 챗봇에서 BERT와 Sentence-BERT의 성능 비교

은지수⁰¹ 김민국¹ 김수민¹ 서상훈¹ 한정민² 정설영^{1*}

¹ 경북대학교 컴퓨터학부

² 림헬스케어(주)

*교신저자

eunjis0820@gmail.com, minkuk0118@gmail.com, sumin5494@gmail.com, spdlvka147@gmail.com,
hjm@rimhealthcare.co.kr snowflower@knu.ac.kr

Performance comparison of BERT and Sentence-BERT in small-scale chatbot based on GPT-4 text augmentation data

Jisu Eun⁰¹, Min-guk Kim¹, Sumin Kim¹, Sanghoon Seo¹, Jeong-min Han², Seol-Young Jeong^{1*}

¹Department of Computer Science, KyungPook National University

²RIM Healthcare Co., Ltd.

*Corresponding author

요약

성공적인 검색 모델 챗봇을 구현하기 위해서는 양질의 데이터셋이 필요하다. 그러나 항상 양질의 데이터셋을 보유할 수는 없다. 텍스트 데이터를 증강할 때, 우리는 기존에 널리 사용되고 있는 EDA를 대신하여 GPT-4를 이용한 텍스트 데이터 증강을 시험하였다. 또한, 검색 모델 챗봇에서 KoBERT를 이용한 다중분류 모델과 KR-SBERT를 이용한 코사인 유사도 계산 모델 중 어느 것이 더 적합한지 실험하였다. 그 결과 KR-SBERT가 KoBERT에 비하여 전체적으로 나은 성능을 보였다. GPT-4 프롬프트 엔지니어링을 통한 텍스트 증강은 KoBERT에서 약 35%의 향상을 보였고, KR-SBERT에서는 약 18%의 향상을 보였다. 이후 EDA와 GPT-4 증강을 직접적으로 비교하는 연구가 수행되길 기대한다.

1. 서론

성공적인 검색 모델 챗봇 (Retrieval-based Chatbot Model)을 구현하기 위해서는, 많은 어휘를 담고 있는 데이터셋이 필요하다. 다시 말해 다양한 어휘를 가지고 있는 많은 양의 데이터셋이 필요하다는 의미이다. 그러나 우리는 모든 상황에서 양질의 데이터셋을 확보할 수 없다. 데이터 증강(Data Augmentation) 기법은 데이터가 부족한 상황에서 인위적으로 데이터의 양을 증가시키는 방법이다. 전통적으로 이미지 데이터를 이용하는 연구에서 사용되었는데, 최근에는 텍스트 기반의 데이터 증강 기법도 연구되고 있다[1].

텍스트 기반의 데이터 증강 기법에는 여러 방법이 있으나, 많은 경우 영어가 아닌 데이터에는 적용이 어렵다. 또한 텍스트 데이터 증강 분야에서 널리 사용되고 있는 EDA(Easy Data Augmentation)는 BERT(Bidirectional Encoder Representations from Transformers)와 같이 사전 학습된 언어 모델에서 효과적이지 않다고 알려져 있다[2].

본 논문에서는 텍스트 데이터 증강 기법 중 하나로 OpenAI의 GPT-4 모델을 이용하여 텍스트를 증강하는

방법을 도입하고자 한다. GPT-4에게 적절한 질의를 함으로써 데이터셋의 양을 증가시킬 것이다. 증강된 데이터셋을 처리할 언어모델로 KoBERT[3]와 KR-SBERT[4]를 선정하였다. KoBERT와 KR-SBERT 중 어떤 모델이 검색 모델 챗봇에 적합할지 알아본다.

본 논문에서는 원본 데이터셋과 GPT-4로 증강한 데이터셋, 그리고 KoBERT를 이용한 다중분류기와 KR-SBERT를 이용한 코사인 유사도 계산기를 다룬다. 2종류의 데이터셋과 2종류의 언어 모델을 상호 조합하여 어떤 조합이 검색 모델 챗봇에 가장 효과적인 조합인지 제시하고자 한다. 성능의 비교는 정확도, 정밀도, 재현율, F1 점수를 이용해 비교한다[5].

2. 관련 연구

2.1. EDA

EDA는 텍스트 데이터를 증강하기 위한 방법 중 하나로, Wei et al.의 논문[2]에서 제시된 바 있다. EDA는 유의어 대체, 임의 삽입, 임의 교체, 임의 삭제 방법을 이용하여 텍스트를 증강하는 방법이다. 널리 사용되는 방법이지만 사전 학습된 모델에서는 EDA를 통한 데이터 증강으로

얻을 수 있는 성능 향상이 미미하다.

2.2. 프롬프트 엔지니어링 (Prompt Engineering)

프롬프트 엔지니어링은 GPT-4 와 같은 대규모 언어 모델과 더 효과적으로 대화하기 위하여 고안된 기술이다. 프롬프트는 대규모 언어 모델에 대한 상호작용과 출력을 사용자 지정할 수 있는 일종의 명령어이다[6]. 최근에는 정교한 프롬프트를 구성하여 데이터셋을 증강하려는 시도가 이루어지고 있다[7].

2.3. BERT 와 KoBERT

BERT 는 레이블이 없는 텍스트로부터 심층 양방향 학습을 한 사전학습된 모델이다[8]. 약 3300M 개의 영단어를 학습하였고, 별도의 모델 구조 변경 없이 출력층을 추가하는 것으로 미세조정 할 수 있다는 장점이 있다. KoBERT 는 BERT 가 기본적으로 지원하는 multilingual case 중 한국어의 성능을 개선하기 위해 제시된 모델이다. BERT 에 비하여 2.6% 한국어 처리 성능 향상을 기록한 것으로 알려져 있다. 본 논문에서는 KoBERT 를 질의응답 데이터셋에 대한 다중분류 문제를 해결하는 데 사용한다. 챗봇 구현 단계에서는 주어진 입력이 데이터셋에 존재하는 어떤 질문 집단에 속하는지 파악하고 해당 질문 집단에 대응되는 답변을 출력한다.

그러나 BERT 모델은 두 문장간 유사도를 구하는 문제에서는 비효율적이다[9].

2.4. SBERT 와 KR-SBERT

SBERT(Sentence-BERT)는 BERT 모델에서 NLI 및 STS 데이터셋으로 미세조정하여 문장 임베딩 성능을 개선한 모델이다. BERT 모델보다 빠르고 우수하게 의미론적으로 의미있는 (semantically meaningful) 문장 임베딩을 구할 수 있다[9]. KR-SBERT 는 KR-BERT[10] 모델에서 문장 임베딩 성능을 개선한 모델이다. SBERT 모델의 한국어 처리 성능을 증대하기 위해 제시되었다. SBERT 의 아이디어를 참조하여 구현한 것으로 알려져 있다[4]. 본 논문에서는 KR-SBERT 를 질의응답 데이터셋에 대하여, 문장의 임베딩 벡터 간 코사인 유사도를 계산하는 데 사용한다. 챗봇 구현 단계에서는 주어진 입력과 데이터셋을 비교하여 가장 코사인 유사도가 높은 질문을 찾고, 그에 대응되는 답변을 출력한다.

3. 실험 설계 및 평가

본 장에서는 트랜스포머 기반의 언어모델인 BERT 및 문장간 의미 유사도 측정 특화 모델인 SBERT 에서 한국어 처리 성능을 개선한 KoBERT 와 KR-SBERT 를 활용하여 성능 비교 실험을 진행한다. NLI 및 STS 데이터로 미세조정 후 배포된 KR-SBERT 와 달리 KoBERT 는 충분한 크기의 데이터셋으로 추가적인 미세조정 작업이 필요하다.

KoBERT 와 KR-SBERT 각 모델에 대하여 원본 데이터셋과 데이터 증강 후의 데이터셋(이하 증강 데이터셋) 두 가지를 적용해본다. 각 모델별로 데이터 증강 전과 증강 후의 성능을 측정한다. KoBERT 와 KR-SBERT 두 모델간 성능비교와, 각 모델의 데이터 증강 여부에 따른 성능비교를 수행하기 위한 지표로서 정확도, 정밀도, 재현율, F1 점수를 이용한다.

3.1. 가설 설정

본 논문에서 인용한 이론들을 바탕으로 두 가지의 가설을 수립하였다.

[가설 1] 같은 데이터셋으로 성능평가를 실행하면, 각 성능평가에 대해 KR-SBERT 가 KoBERT 보다 더 높은 값을 기록할 것이다.

[가설 2] 자연스럽게 다양한 문장을 생성하는 GPT-4 를 이용한 텍스트 증강은 KoBERT 와 KR-SBERT 에서 유의미한 성능 향상을 이룰 수 있을 것이다.

3.2. 실험 설계

본 실험을 수행하기 위해 챗봇의 질문으로 예상되는 문장 1085 개를 수동으로 작성하여 원본 데이터셋을 구축한 후, 프롬프트 엔지니어링을 활용하여 크기가 39284 인 증강 데이터셋을 구축하였다.

3.2.1. 데이터셋 구축 및 증강

실험에서 사용한 데이터셋은 림헬스케어(주)에서 보유하고 있는 65 개의 질문-응답 쌍을 기반으로 한다. 림헬스케어(주)는 자사의 “유리노”라는 제품에 대하여 자주 묻는 질문(FAQ) 65 개를 보유하고 있다. 65 개의 각 질문에 대해 유사문장을 수동으로 작성해서 총 1085 개의 원본 데이터셋을 구축하였다. 표 1 에서 그 중 5 개의 질문-응답 쌍을 확인할 수 있다.

그 후 원본 데이터셋과 GPT-4 를 활용하여 39284

표 1 림헬스케어(주)의 FAQ 데이터셋을 기반으로 구성된 원본 데이터셋 중 5 개를 추출하여 정리한 결과

	question	answer
82	아예 완전 치료할 수 있어?	유리노는 안경을 쓰면 눈이 좋아지는 것과 같은 원리로 ...
98	볼일 볼 때 착용해도 돼?	유리노는 질 내에 삽입하여 사용하는 기구로 ...
227	허가 받고 파는 제품이야 ?	유리노는 인체에 사용할 수 있는 3등급 의료기기로 ...
181	안 버리고 계속 써도 돼?	일회용 멸균 제품으로 재사용을 금지합니다. ...
626	유리노 제품에 특허가 부여되었나요?	유리노는 국내 특허는 물론 ...

크기의 증강 데이터셋을 구축하였다. 각 질문의 유사문장을 얻어내기 위해 문장의 말투, 어휘 수준, 질문이 나오게 된 맥락 등에 조건을 두는 few-shot learning 방법을 이용하여, GPT-4로부터 양질의 데이터를 얻고자 하였다.

GPT-4에게 전달한 프롬프트는 다음과 같다 :

GPT-4에게 전달한 프롬프트
<p>1) [FAQ 데이터셋에 있던 질문 목록 65개 중 하나]라는 문장을 뜻은 같고 다른 형태로 30개 적어줘.</p> <p>조건이 3가지 있어.</p> <ol style="list-style-type: none"> "유리노"라는 요실금 치료기구를 만드는 회사에게 하는 질문이야. 사람들이 챗봇에게 할법한 말투와 단어 수준으로 적어줘. [질문 의도]를 묻는 취지에서 벗어나지 말아줘.
<p>2) 위 내용을 전부 유지하되 조금 더 짧고 쉬운 말투로도 30개 적어줘.</p>
<p>3) 위 질문의 대답은 [해당하는 질문에 대하여 FAQ 데이터셋에 존재하는 답변]이야. 이 대답이 도출되기 위해 나올 법한 질문을 10개 적어줘.</p>

3)에서 생성한 질문들 각각에 대하여 다시 1), 2)를 수행한다.

위 방법으로 1개의 질문에 대하여

$$(1 + 10) * (30 + 30) = 660$$

총 660배 증강을 기대할 수 있으나 그 중 중복을 제거하여 39284 크기의 데이터셋이 되었다.

3.2.2. KoBERT

KoBERT는 원본 데이터셋을 훈련 집합:검증 집합:테스트 집합 6:2:2로 분리 후 미세조정 및 성능평가를 진행하였다. 클래스별 데이터 비율은 원본과 같게 했다.

가장 좋은 성능을 보이는 에포크(epoch) 값을 확인하기 위해 에포크 값당 훈련 정확도(train accuracy), 검증 정확도(validation accuracy), 손실함수(loss function) 결과를 구하였다, 그림 1은 원본 데이터셋을 사용하였을 때 훈련 중 최적의 에포크 값을 찾는 그래프이다. 배치 크기(batch size)는 32를 사용하였고, 다른 파라미터 값은 KoBERT 권장값을 준수하였다.

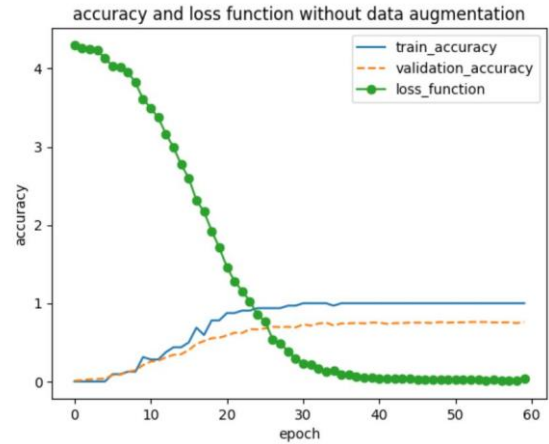


그림 1 증강 전 KoBERT의 정확도 및 손실함수 그래프

그림 1에서 성능 향상이 정체되며 손실함수 값이 더 줄어들이지 않거나 다시 커지는 기점에서 모델이 최대의 성능을 보이고, 그 이후로는 과대적합(overfitting)이 나타난다. 에포크 값이 30일 때를 전후로 정확도가 상승하지 않고, 35일 때를 전후로 손실함수 값이 정체되므로 최적의 에포크 값은 35이다.

증강 데이터셋을 사용하여 이 실험 과정을 반복하였다. 그림 2는 증강 데이터셋을 사용하였을 때 훈련 중 최적의 에포크 값을 찾는 그래프이다.

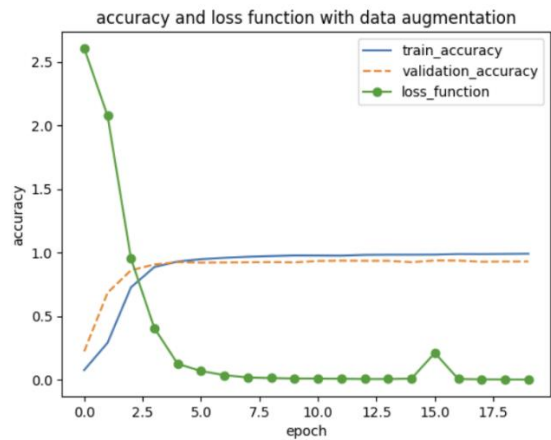


그림 2 증강 후 KoBERT의 정확도 및 손실함수 그래프

그림 2에서는 에포크 값이 8일 때를 전후로 정확도가 상승하지 않고, 15 이후로는 손실함수 값에 유의미한 변화가 없으므로 최적의 에포크 값은 14이다.

3.2.3. KR-SBERT

KR-SBERT는 미세조정이 끝난 모델이므로 따로 하이퍼파라미터 튜닝을 진행하지는 않고, 원본 데이터셋을 임베딩 집합: 테스트 집합=8:2로 분리해 KR-SBERT의 성능평가를 진행하였다. 그 후 증강 데이터셋에 대해서도 데이터셋을 임베딩 집합:테스트 집합=8:2로 분리해 성능평가를 진행하였다.

Google Colab의 T4 GPU 환경에서 KoBERT의 미세조정 및 KR-SBERT의 문장 임베딩 과정을 수행했다.

그 후 같은 환경에서 각 모델의 성능을 측정하였다.

3.3 성능 평가

표 2는 원본 데이터셋과 증강 데이터셋을 사용하였을 때 KoBERT와 KR-SBERT의 성능평가 결과를 기록한 표이다.

표 2 성능평가 결과

모델	정확도	정밀도	재현율	F1 점수
KoBERT_원본	0.7558	0.6595	0.6980	0.6782
KoBERT_증강	0.9217	0.9151	0.9271	0.9211
KR-SBERT_원본	0.7972	0.7914	0.7845	0.7879
KR-SBERT_증강	0.9263	0.9345	0.9305	0.9325

표 2에 따르면, 원본 데이터셋과 증강 데이터셋을 사용한 두 가지 경우 모두 KR-SBERT가 KoBERT보다 4가지 성능평가 지표에서 높은 값을 보인다. F1 점수 기준으로 비교하면 원본 데이터셋은 KR-SBERT가 약 11%p, 증강 데이터셋은 KR-SBERT가 약 1%p 높은 성능을 기록하였다. 증강 데이터셋을 이용했을 때 성능 차이는 아주 크다고 볼 수 없지만 [가설 1]이 참임을 확인할 수 있다.

GPT-4를 이용한 증강 이후 KoBERT와 KR-SBERT의 성능이 유의미하게 향상됨을 확인할 수 있었다. F1 점수를 기준으로 비교할 때 KoBERT는 약 35%의 향상을, KR-SBERT는 약 18%의 향상을 보였다. 따라서 [가설 2]가 참임을 확인할 수 있다.

4. 결론

본 논문에서는 첫째로 KoBERT와 KR-SBERT 중 어느 모델이 검색 모델 챗봇에 적합한지 알아보려고 하였다. 그 결과 전반적으로 KR-SBERT가 KoBERT에 대해 모든 성능평가 지표에서 앞서는 것을 확인할 수 있었다.

두번째로 부족한 텍스트 데이터셋을 증강시키기 위한 방법으로 GPT-4를 이용한 프롬프트 엔지니어링을 도입하였다. 원본 데이터셋과 증강 데이터셋을 KoBERT와 KR-SBERT 두 모델에서 비교하여 GPT-4를 이용한 텍스트 증강의 성능을 알아보았다. KoBERT와 KR-SBERT 모두 few-shot learning을 적용한 프롬프트 엔지니어링으로 눈에 띄는 성능 향상을 이루어냈다. 본 논문은 텍스트 데이터 증강 기법으로 GPT-4를 이용한 프롬프트 엔지니어링이 효과적일 수 있음을 제시하는 데 의미가 있다.

텍스트 증강이 불가능한 상황에서는 KR-SBERT가 확연히 유리하나, 텍스트 증강을 거친 후의 상황에서는 KoBERT와 KR-SBERT가 미미한 성능 차이를 보인다는 사실 또한 확인할 수 있었다.

추후 이 논문을 보완하여 EDA와 GPT-4 등의 프롬프트 엔지니어링을 이용한 텍스트 증강을 직접적으로 비교한 후속 연구가 수행되길 기대한다.

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01082)

참고문헌

[1] Bayer, Markus, Marc-André Kaufhold, and Christian Reuter. "A survey on data augmentation for text classification." *ACM Computing Surveys* 55.7 1-39. (2022).

[2] Wei, Jason, and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. p. 6382-6388. (2019.)

[3] SKTBrain, "KoBERT," GitHub repository, <https://github.com/SKTBrain/KoBERT>

[4] Park, et al. "KR-SBERT: A Pre-trained Korean-specific Sentence-BERT model", GitHub repository, <https://github.com/snunlp/KR-SBERT>, 2021

[5] Vujović, Ž. "Classification model evaluation metrics." *International Journal of Advanced Computer Science and Applications* 12.6 599-606. (2021)

[6] White, Jules, et al. "A prompt pattern catalog to enhance prompt engineering with chatgpt." *arXiv preprint arXiv:2302.11382* (2023).

[7] Younggyun Hahm et al.. "An Analysis of GPT-4 Prompting for Korean Aspect-based Sentiment Analysis." *한국정보과학회 학술발표논문집*, 2023, 760-762.

[8] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *Proceedings of NAACL-HLT*. p. 4171-4186. (2019)

[9] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* p. 3982-3992. (2019).

[10] Lee, Sangah, et al. "Kr-bert: A small-scale korean-specific language model." *Journal of KIISE* 47.7 682-692. (2020)

대규모 언어 모델을 활용한 강의평가 질의응답 서비스

신진철^{1,0}, 김근찬², 신성한², 조정래³, 정명훈⁴, 정설영²

¹경북대학교 원예과학과, ²경북대학교 컴퓨터학부, ³경북대학교 아동학부, ⁴구글코리아

wlscjf3877@knu.ac.kr, mouse4918@knu.ac.kr,

crossrunway01@gmail.com, wjdfoek3@knu.ac.kr,

jerryjg@google.com, snowflower@knu.ac.kr

Question-Answering Service Using Large-Scale Language Models for Course Evaluation

Jincheol Shin^{1,0}, Geunchan Kim², Sunghan Shin², Jeongrae Jo³, MyeongHoon Jeong⁴,
JeongSeol Young²

¹Dept. of Horticulture Science, Kyungpook-National University,

²Dept. of Computer Science, Kyungpook-National University,

³Dept. of Child Study, Kyungpook-National University,

⁴Google Korea

요약

본 연구는 LLM과 Vector DB를 활용한 대학 강의에 대한 질문 및 답변 서비스를 사용하였다. 사용자는 강의와 교수 관련 질문을 입력하고, 시스템은 이를 분석하여 가장 적합한 답변을 제공한다. 본 연구에서는 질문의 키워드 개수, Vector DB의 사용 유무, 문장의 개수에 따른 정확도를 측정하였다. 실험 결과, 3개의 키워드를 사용한 질문이 가장 높은 정확도를 보였으며, Vector DB를 활용하고 한 문장으로 구성된 질문이 가장 정확도가 높았다. 이러한 결과를 토대로, 대학 교육 환경에서 LLM 기술을 활용한 질문 및 답변 서비스는 사용자들의 정보 제공 및 지원을 개선하고 학습 경험을 향상시키는데 유용할 것으로 기대된다.

1. 서론

본 연구는 대학교 강의에 관한 질문과 답변을 제공하는 LLM(대규모 언어 모델) 서비스를 가장 효율적으로 활용할 수 있는 방안을 탐구하는 것을 목표로 한다. 대학생들과 교수님들을 대상으로 이용되는 이 서비스를 개발함으로써, 대학 교육 환경에서의 정보 제공 및 지원을 향상하고자 한다.

연구에서 주요한 관심사는 서비스 이용자들이 키워드 개수를 정할 때 LLM 서비스의 성능이 언제 가장 우수한지에 대한 측정이다. 이를 통해 최적의 서비스 이용 방법을 탐구하고 사용자들에게 가장 적합한 질문 형식을 제안하여 LLM을 활용할 때 최상의 결과를 얻을 수 있도록 한다.

또한 교육 분야에서 LLM 기술의 활용을 통해 학습 경험을 향상하고 학문적인 질문에 대한 답변을 효과적으로 제공하기 위한 새로운 방법을 탐구하며, 사용자들이 서비스를 최대한 활용할 수 있도록 돕는 것을 목표로 한다. 이러한 연구는 교육 환경을 더욱 풍부하게 만들어 학문적 성취를 촉진하고 지식 공유를 증진시킬 수 있을 것으로

기대된다.

2. 관련 연구 사례

최근 LLM을 이용한 자연어 처리 모델에 관해 다양한 연구가 진행 중이다. 평가자들로부터 자연어 모델을 벤치마크하여 LLM의 신경망의 깊이에 따른 성능을 측정하는 사례가 있었으며, 어떠한 구성의 LLM 모델이 가장 효율적인 결과를 내는지 도출한 사례가 있다 [1]. 한편 대화형 인공지능 챗봇의 경우 Context 및 Prompt의 설계가 매우 중요한데, AI 윤리 원칙에 따라 챗봇의 Prompt를 설계하여 올바른 답을 하도록 한 사례를 제시하기도 한다 [2]. 또한 우리의 연구와 유사하게 추천을 위한 챗봇의 추천 정확도를 향상하기 위해 다양한 촉진 전략을 포함한 Prompting 방법을 제시하며 [3], 시험 자동 채점에 LLM모델을 적용, 평가한 사례 [4]도 있었다.

3. 시스템 설계

3.1 시스템 구성

본 논문에서 제안하는 시스템은 LLM과 Vector DB, 그리고 RAG 시스템을 사용했다. LLM은 인간의 언어를 이해하고 생성하도록 훈련된 인공지능으로 구글의 chat-bison 모델을 사용했다. Vector DB는 데이터를 임베딩하여 저장해 놓을 수 있는 데이터베이스이다. 임베딩은 모든 유형의 데이터를 자산의 의미와 Context를 캡처하는 고차원 벡터로 인코딩하는 것을 의미한다. 이후 데이터베이스에 검색 시, 고차원 공간에서 인접한 데이터들을 찾을 수 있다. Vector DB는 구글의 PostgreSQL을 사용했다. RAG는 "Retrieval-Augmented Generation"의 약어로 정보 검색과 생성을 결합하여 텍스트 생성 프로세스를 개선하고 확장하는 방법론을 나타낸다. 요약하자면 RAG 시스템을 LLM 및 Vector DB를 결합하여 정보 검색과 생성을 향상하며, 사용자의 자연어 질문에 대한 효과적인 응답을 생성할 수 있게 시스템을 구성하였다.

3.2 데이터셋

우리 서비스는 학생들이 강의와 교수의 평가에 대해 질문하는 내용에 답변하기 위해 '에브리타임'이라는 어플에서 강의평들을 크롤링했다. 크롤링한 데이터들을 {key(강의명과 교수명) : value(해당 교수의 강의의 모든 강의평가 내용)} 형태로 저장하였고 모든 key와 그에 해당하는 value 값을 임베딩하여 데이터베이스에 저장하여 사용했다.

3.3 시스템 구조

사용자로부터 입력이 들어오면 그 입력을 LLM의 입력으로 넣어 과목명, 교수명과 질문 내용 총 3개의 키워드로 분리한다. 그림 1은 시스템 구조를 다이어그램으로 나타낸 것이다. 만약 키워드가 없는 경우엔 None을 반환하는데, 과목명, 교수명 키워드가 모두 None인 경우엔 Jail Breaking을 방지하기 위해 프로그램이 종료된다. LLM으로부터 분리된 키워드들은 데이터베이스에 질의할 때 사용된다. 과목명과 교수명은 DB에 저장되어 있는 tuple을 매칭하는 데 사용되고, 임베딩된 질문 내용과 L2 distance가 가장 짧은 data들을 LLM의 input token 제한 때문에 30개만 가져온다. 데이터베이스에서 가져온 Dataset은 다시 LLM의 Context에 입력되고, LLM에 지시할 다른 사항들도 Context에 같이 입력된다. Context가 입력된 LLM에 사용자의 입력을 다시 넣어 나오는 결과를 사용자에게 출력하는 방식으로 시스템이 설계되어 있다.

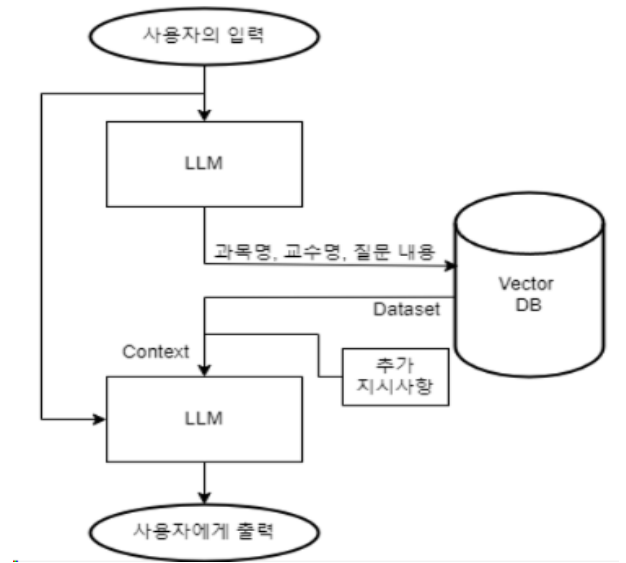


그림 1. 시스템 아키텍처

4. 시스템 평가

4.1 질문의 종류

본 논문에서 사용한 키워드는 '시험', '과제', '팀플', '출석'으로 총 4가지이다. 질문의 종류는 교수 이름만 포함된 질문을 '교수명'으로 지정하였고 과목 이름만 포함된 질문을 '과목명', 그리고 교수 이름과 과목 이름이 모두 포함된 질문을 '과목명+교수명'으로 지정하였다.

'교수명', '과목명', '교수명+과목명' 각각의 질문에서 키워드의 개수에 따라서 1단계부터 5단계까지 다섯 단계로 나누었다. 1단계는 키워드 개수가 0개인 것을 의미하며, 5단계는 키워드 개수가 4개인 것을 의미한다. 각 단계에서는 4개의 키워드로 만들 수 있는 모든 조합을 사용하여 질문을 만들었다. 예를 들어, 4단계는 키워드가 3개이므로 만들 수 있는 모든 조합의 경우는 (시험, 과제, 팀플), (시험, 과제, 출석), (시험, 팀플, 출석), (과제, 팀플, 출석)으로 총 4가지의 질문이 있는 것이다. 또한 각 단계에서 해당 질문의 문장 개수에 따라서 문장의 개수가 1개면 '한 문장'로 지정하였고 2개 이상이면 '여러 문장'으로 지정하였다. 예를 들어서 '교수명'이고 키워드가 '4개'인 '여러 문장'은 "OOO교수님 어때? 시험 어때? 과제 어때? 팀플 어때? 출석 어때?"로 정의한다.

4.2 결괏값 수집 방법

크게 Vector DB를 사용한 집단과 아닌 집단으로 나누었다. Vector DB를 사용한 집단은 3.3의 환경에서 각각의 질문을 Input 값으로 설정하여 얻은 결괏값을 수집하였다.

반면, Vector DB를 사용하지 않은 집단은 Google Cloud에서 제공하는 Generative AI Studio의 언어 모델의 프롬프트 모델을 사용하였다. 이 역시 각각의 질문을 Input 값으로 설정하여 얻은 결괏값을 수집하였다.

4.3 평가 방법

평가는 각각의 질문에 Optimal한 답변을 임의로 지정한 'Optimal output'와 Output을 각각 벡터화하여 두 문장 간의 Cosine 거리 차이를 통하여 정확도를 측정하였다.

4.4 평가 결과

실험결과를 통해 키워드의 개수에 따른 답변 완성도의 간략한 추이를 확인할 수 있었다. 아래의 그래프 중 주황색의 그래프는 한 문장으로만 질문을 구성했을 경우의 정확도 추이를 나타내며, 파란색의 그래프는 여러 개의 문장으로 질문을 구성하였을 경우의 정확도 추이를 나타낸다.

그림 2는 Vector DB를 사용하지 않은 경우의 결과이다. '교수명' 질문에서는 키워드가 2개이고 한 문장으로 이루어진 질문이 가장 정확도가 높았다. 반면, '과목명' 질문에서는 키워드가 4개이고 한 문장으로 이루어진 질문이 정확도가 가장 높았으며, '과목명+교수명' 질문에서는 키워드가 3개이고 한 문장으로 이루어진 질문이 정확도가 가장 높았다.

반면 그림 3은 Vector DB를 사용한 경우의 결과이다. Vector DB를 사용하지 않았을 때의 최저 정확도가 30%인데 비하여, 최저 정확도가 58%로 약 2배 정도 높았고 Vector DB를 사용하지 않았을 때의 정확도의 평균이 64%인데 비해, 평균이 72%로 더 높았다. 이 역시 마찬가지로 '과목명+교수명' 질문에서 키워드가 3개이고 한 문장으로 이루어진 질문이 가장 정확도가 높았다.

결과적으로 키워드의 개수가 3개인 경우, 정확도가 가장 높은 것을 확인할 수 있었다. 그중에서도 Vector DB를 사용하면서 한 문장으로 질문했을 때가 정확도가 더 높았다. 반면, 키워드의 개수가 2개이면서 Vector DB를 사용하지 않고 여러 개의 문장으로 이루어진 질문이 정확도가 가장 낮았다.

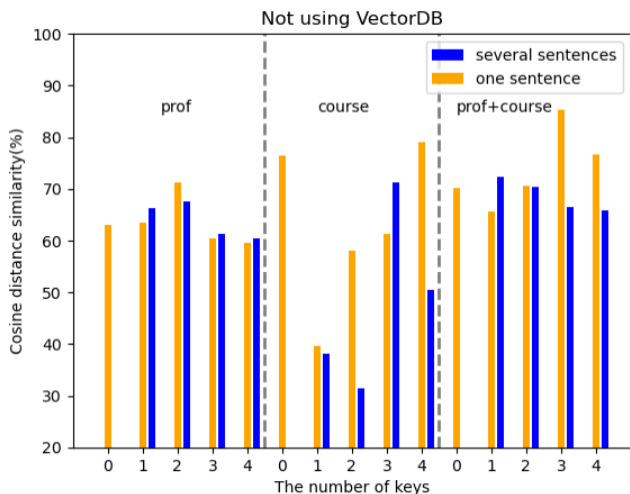


그림 2. Vector DB를 사용하지 않은 경우

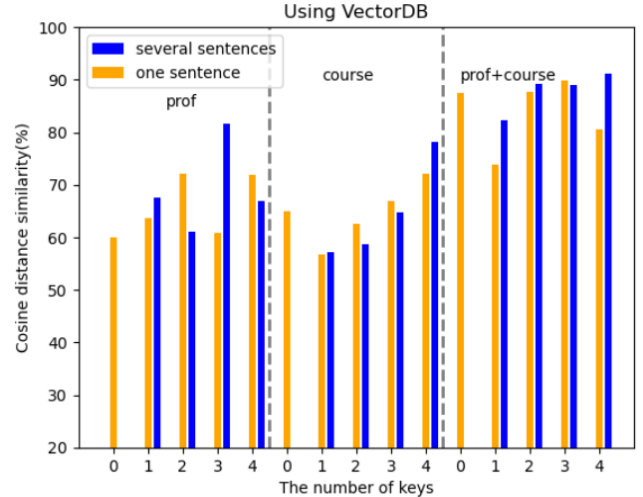


그림 3. Vector DB를 사용한 경우

5. 결과

본 연구는 대학 강의에 대한 질문 및 답변 서비스를 개발하고 향상하기 위해 LLM(대화형 언어 모델)과 Vector DB 기술을 활용하였다. 사용자는 강의 및 교수 관련 질문을 입력하면 시스템이 가장 적합한 답변을 제공한다. 연구는 다양한 질문 유형 및 서비스 설정을 실험하여 최적의 서비스 사용 방법을 찾았다. 결과적으로, 3개의 키워드를 사용한 질문이 가장 높은 정확도를 보였으며, Vector DB를 활용하고 한 문장으로 구성된 질문이 가장 효율적이었다. 이 연구를 통해 대학 교육 환경에서 LLM 및 Vector DB를 활용한 질문 및 답변 서비스가 사용자들의 정보 제공 및 학습 경험 향상에 유용할 것으로 기대된다.

6. 참고 문헌

[1] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, Yongbin Li, "Wider and Deeper LLM Networks are Fairer LLM Evaluators", arXiv preprint arXiv:2308.01862, 2023

[2] Junseong Bang, Byung-Tak Lee, Pangun Park, "Examination of Ethical Principles for LLM-Based Recommendations in Conversational AI", 2023 International Conference on Platform Technology and Service

[3] Hanjia Lyu, Song Jiang, Hanqing Zeng, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, Yinglong Xia, Jiebo Luo, "LLM-Rec: Personalized Recommendation via Prompting Large Language Models", 24 Jul 2023 arXiv:2307.15780 [cs.CL]

[4] Johannes Schneider, Bernd Schenk, Christina Niklaus, Michaelis Vlachos, "Towards LLM-based Autograding for Short Textual Answers", arXiv:2309.11508 [cs.CL] 9 Sep 2023

*"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

이상행동 감지 모델 시스템

김승준^(0,1), 우승택⁽¹⁾, 박한솔⁽¹⁾, 김세아⁽¹⁾, 김민준⁽¹⁾, 정설영⁽¹⁾, 김호영⁽²⁾

경북대학교 컴퓨터학부⁽¹⁾, (주)우경 정보 기술⁽²⁾

kimmokalover@gmail.com, dntmdxor99@naver.com, pjhcsols@naver.com,
seakim@knu.ac.kr, kmjj5275@naver.com, snowflower@knu.ac.kr,
hykim@wkitt.co.kr

System for Anomalous Behavior Detection

SeungJun Kim^(0,1), SeungTaek Woo⁽¹⁾, HanSol Park⁽¹⁾, SeAh Kim⁽¹⁾, MinJun Kim⁽¹⁾, SeolYoung Jeong⁽¹⁾, HoYoung Kim⁽²⁾

School of Computer Science and Engineering, Kyungpook National University⁽¹⁾,
Wookyung Information Technology⁽²⁾

요 약

최근 "묻지마 범죄"가 증가하고 있어 이러한 범죄를 줄이기 위한 노력이 확대되고 있다. 현재 이상행동 감지 시스템은 두 가지 주요 방법을 사용하고 있다. 직접 관제인에 의한 수동 감시 방식과 인공지능을 활용한 이상행동 감지 방식이다. 직접 관제인을 통한 감시는 인력 소모가 크다는 단점이 있다. 하지만 AI 기반의 이상행동 감지는 일반적으로 공간 및 시간적 특성을 구별하여 객체와 그들의 움직임 정보를 추출하고, 공간적 및 시간적 특성을 결합하여 객체의 신원과 행동을 분류하기 때문에 인력 소모나, 경제적으로 효과적이다. 하지만 대부분의 인공지능 모델은 현실의 비디오의 상태를 고려하지 않고, 비디오의 모든 프레임을 동일하다고 가정하고 있다. 그러나 Low-Delay 모드와 같이 I-Frame과 P-Frame이 있는 경우, P-Frame은 일반적으로 I-Frame에 비해 정보가 적을 수 있으므로 이러한 가정은 모델이 학습을 불안정하게 할 가능성이 있다. 또한 QP(Quantization Parameter)가 높을수록 비디오 프레임의 상세 정보가 크게 손실되어 모델이 움직임과 행동을 정확하게 파악하는 데 어려움을 겪을 수 있다. 따라서 본 논문에서는 비디오 부호화 기술을 고려하여 인공지능 모델을 사용한 웹 기반 감지 시스템을 제안한다. 구체적으로는 인공지능 모델을 사용하여 관리자에게 실시간으로 알림을 보내고 해당 내용을 다중 서버 구조를 통해 처리해, 사용자가 웹 기반 UI를 통해 이상행동을 확인할 수 있는 시스템 모델을 설계하고자 한다.

1. 서 론

범죄는 사회 안전과 안정에 심각한 영향을 미치는 사회 문제 중 하나로, 이를 예방하고 대응하기 위한 다양한 노력이 이루어지고 있다. 이러한 노력 중 하나는 현대 기술과 인공지능의 활용하여 안면 인식, 프레드폴 알고리즘 등이 있다. 본 논문에서는 비디오 객체 행동 인식 시스템을 통해 범죄와 연관된 문제를 다루고자 한다.

비디오 객체 행동 인식은 객체의 움직임과 행동을 이해하고 분류하는 기술로, 이를 통해 범죄 예방 및 수사에 도움을 줄 수 있다. 현재, 딥러닝과 비디오 부호화 기술, 그리고 다양한 웹 프레임워크의 발전으로 비디오 객체 행동 인식 모델의 정확도와 성능이 크게 향상되고, 사용자 친화적인 시스템들이 만들어지고 있다. 그러나 기존의 인공지능 모델들은 비디오의 상태를 고려하지 않고, 모든 프레임을 동일하게 다룬다. 하지만 대부분 현실의 비디오는 압축된 상태로, 현재 많이 사용되는 기술인 H.264/AVC(Advanced Video Coding)와 H.265/HEVC(High Efficiency Video Coding)로 압축되어 있어, Low-Delay와 Random-Access 모드 과정에서 프레임 정보 손실과 화질 변화가 발생하기 때문이다.[1][2] 이로 인해 객체의 움직임과 행동을 정확하게 파악하기 어려워진다. 그리고 기존의 관제 시스템은 실시간으로 동영상 데이터를 처리하는 특성 때문에 서버의 오버헤드가 크다.

본 논문에서는 이러한 관제 시스템의 특성과 비디오 부호화 기술의 특성을 고려하여, 인공지능 모델의 성능을 향상하고 효율적으로 영상 정보를 처리할 수 있는 자동 관제 시스템을 제안한다. 본 시스템은 기존의 인공지능 기반 행동 분석 모델을 새롭게 만드는 것이 아니라, 입력 데이터의 특성을 고려하여 모델의 성능을 향상하는데 중점을 두고 있다. 이는 범죄와 관련된 객체 행동 인식 분야에 활용될 수 있으며, 추후 범죄 예방 및 수사에 적극 활용될 수 있다. 그리고 실시간 범죄예방을 위한 실시간 정보 처리와 이를 통해 범죄와 관련된 비디오 객체 행동 인식 분야에서의 연구와 응용에 새로운 가능성을 제시한다.

2. 관련 연구

2.1 비디오 행동 인식

SlowFast는 비디오 행동 인식 딥러닝 모델이다.[4] 해당 모델은 공간적 구조를 파악하는 Slow Stream과 시간적 특징을 파악하는 Fast Stream을 분리하여 학습한다.

해당 모델의 Slow Stream에서는 모델에 입력되는 프레임의 시퀀스에서 τ 만큼의 프레임 간격을 두어 하나의 프레임을 선택해 일정한 개수의 프레임 스택을 Stream에 입력한다. τ 는 일반적으로 16으로 설정한다. 이는 30fps의 비디오에서 초당 약 2개의 프레임을 얻는 것과 같다. 따라서 Slow Stream에서는 시간적인 특징보다 공간적인 구조에 집중하고, 프레임의 객체가 무엇인지에 대한 정보를 학습한다.

Slow Stream과 반대로 Fast Stream에서는 시간적인 정보를 학습한다. 이때 Slow Stream보다 작은 프레임 간격을 두어 시간적인 특성을 파악한다. 이때 Fast Stream이 사용하는 간격은 τ/α 이다. 일반적으로 α 는 8로 설정한다. 따라서 τ 가 16이고, α 가 8이라면 Slow Stream보다 8배 많은 프레임 스택을 얻게 된다. 이는 30fps 비디오에서 초당 약 15개의 프레임을 얻는 것과 같다.

마지막으로 Fast Stream은 Slow Stream에 비해 적은 채널 용량을 가지고 있다. Slow Stream의 채널 용량이 C 일 때, Fast Stream의 채널 용량은 $C\beta$ 로 설정한다. 이때 β 는 일반적으로 1/8로 설정한다. Fast Stream이 적은 채널 용량을 가진다는 것은 공간적인 의미를 파악하는 능력이 Slow Stream에 비해 약하다는 것을 의미하며, 이는 Fast Stream이 시간적 특징에 집중하도록 만든다.

그림 1은 해당 모델의 개요이다. 그림 1에서 볼 수 있듯이 단계마다 Fast Stream의 정보를 Slow Stream으로 입력한다. 최종적으로 해당 모델은 Fast Stream과 Slow Stream의 정보를 종합하여 비디오의 객체가 어떤 행동을 하고 있는지 학습하게 된다.

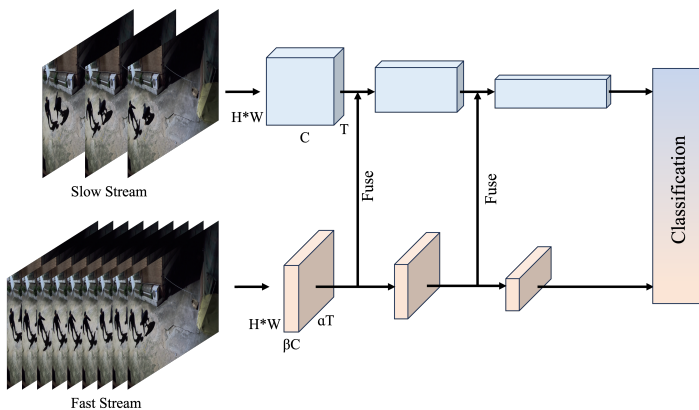


그림 1. SlowFast 모델 개요

2.2 분산 서버 구조

다중 서버 구조란 두 개 이상의 서버를 연계, 운영하는 기술로서 여러 개의 Transaction 요청이 들어왔을 때, 급격한 트래픽 변화에 대응하고 안정적으로 서비스를 제공하기 위한 기술이다. 분산 서버 처리를 위한 기술로는 대표적으로 로드밸런싱이 있는데, 이는 외부의 요청을 서버에 적절히 분산하여 처리하는 것을 의미한다.

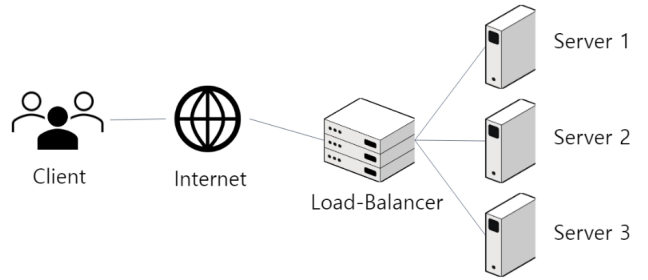


그림 2. 로드 밸런싱 개요

2.3 비디오 부호화 기술

2.3.1 Low-Delay 모드

Low-Delay 모드는 화상 통화, 스트리밍 방송, CCTV 등 다양한 분야에서 사용되는 비디오 부호화 모드이다. 해당 모드는 과거의 프레임을 참조하는 방식인 Inter 프레임 코딩을 사용하여 부호화하는 P-Frame과 프레임 내부의 정보만을 활용하는 방식인 Intra 프레임 코딩을 사용하여 부호화하는 I-Frame으로 이루어진다.

그림 3에서 볼 수 있듯이 하나의 GOP 내에는 하나의 I-Frame과 다수의 P-Frame이 존재한다. 이때 I-Frame을 부호화하는 양자화 파라미터인 QP는 일반적으로 가장 낮은 값을 가진다. 따라서 I-Frame은 가장 좋은 화질을 갖게 되지만 압축 효율이 감소한다. P-Frame을 부호화하는 QP는 일반적으로 I-Frame보다 높은 QP를 가지지만, 중간 혹은 끝부분에 I-Frame과 같은 QP를 가질 수도 있다. 결론적으로 P-Frame은 I-Frame에 비해 화질이 낮아지고, 세부 정보가 손실된다.

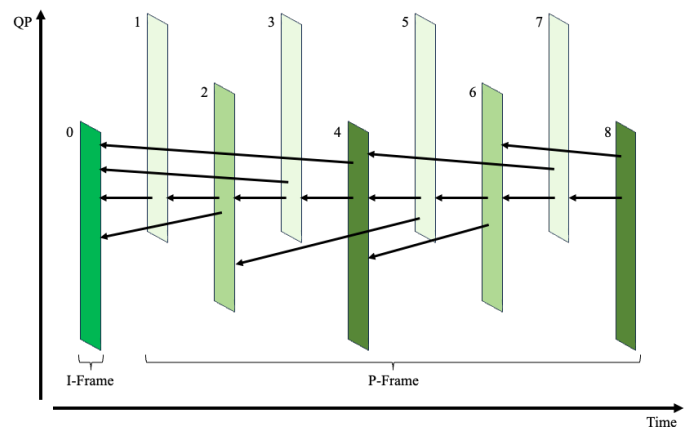


그림 3. Low-Delay 모드

2.3.2 Random-Access 모드

Random-Access 모드는 고화질 동영상 서비스를 제공하는 경우 등 다양한 분야에서 사용한다. 일부 지연을 허용하면서 우수한 화질을 제공하며, 임의접근을 지원한다. 해당 모드는 Low-Delay와 다르게 과거와 미래 프레

임까지 활용하는 B-frame을 추가로 사용한다. 그림 4에서 볼 수 있듯이 Random-Access 모드에서는 부호화 순서와 디스플레이 순서가 다를 수 있다. 또한, 임의접근을 지원하기 위해 주기적으로 I-Frame을 삽입한다. 이때 B-frame 과 P-frame은 I-frame보다 QP가 높거나 프레임 코딩 방식의 특징 때문에 화질이 낮고, 세부 정보가 손실될 수 있다.

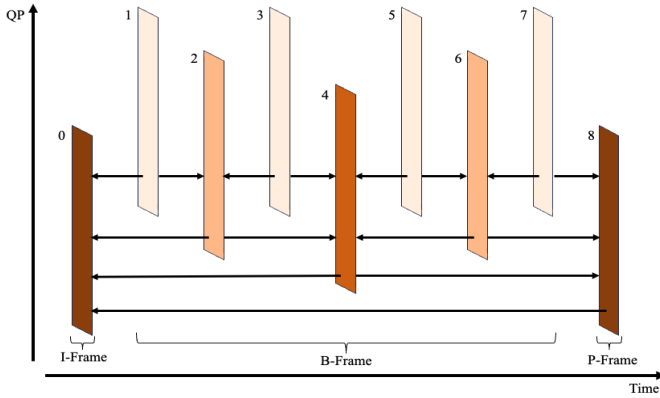


그림 4. Random-Access 모드

2.4 관제 시스템 아키텍처

그림 5처럼 기존의 관제 시스템은 일반적으로 중앙화된 아키텍처를 사용한다. 이는 영상 처리와 사용자 요청 처리를 하나의 중앙 서버에서 수행하는 방식을 의미한다. 이러한 아키텍처는 초기에 구현하기 쉽고 간단한 시스템을 구축하는 데 도움이 될 수 있으나, 여러 단점이 있다.

먼저 중앙 서버가 모든 작업을 처리하기 때문에 시스템의 확장성이 제한된다. 증가하는 트래픽 또는 요청에 대응하기 위해 중앙 서버를 계속 강화하거나 대규모 서버를 사용해야 할 수 있다. 따라서 비용이 증가하고, 과도한 자원 사용과 함께 단일 지점 장애의 위험이 증가한다. 또한 중앙 서버가 모든 요청을 처리하기 때문에 성능 병목 현상이 발생할 수 있으며, 서버 과부하로 인한 응답 지연 문제가 발생할 수 있다.

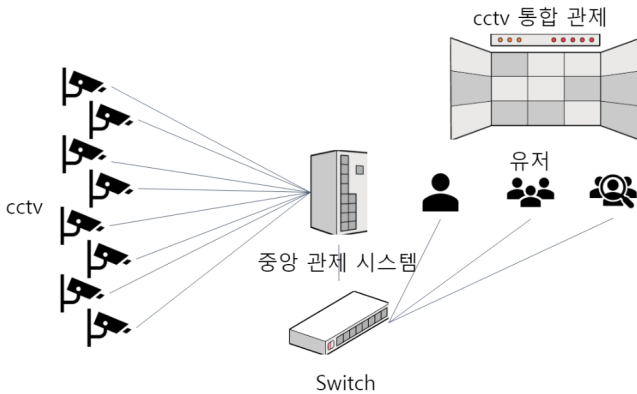


그림 5. 기존 관제 시스템 아키텍처

3. 시스템 설계

3.1 제안 인공지능 모델

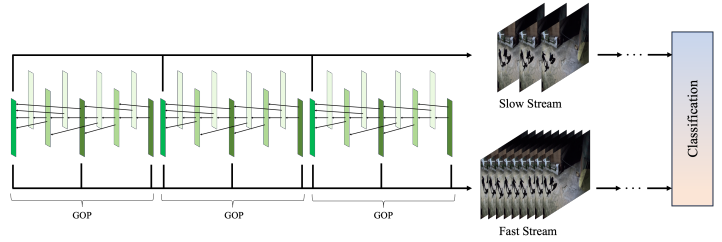


그림 6. 제안 인공지능 모델

앞선 기존 SlowFast 모델은 비디오 내부의 프레임 간의 특성을 고려하지 않고, 일정 구간마다의 프레임을 바탕으로 학습한다. 만약 Low-Delay로 부호화 복호화 과정을 거친 비디오를 학습할 때, Slow Stream에 입력되는 프레임이 모두 높은 QP를 가지는 P-Frame이라면 Slow Stream은 상대적으로 정보가 손실된 프레임만 선택하게 된다. 이러한 과정을 통해 학습된 모델은 최적화되지 못하고, 추론 과정에서 원하는 결과가 나오지 못할 가능성이 존재한다. 또한 Fast Stream에서도 높은 QP를 가지는 P-Frame 위주로 선택한다면, 프레임 내의 노이즈, 모션 블러로 인해 객체의 행동을 정확히 표현하지 못할 여지가 있다.

Random-Access 모드 또한 Low-Delay와 같이 프레임마다 특성이 다르고 정보의 양이 다르다. 해당 경우 또한 모델이 학습할 때, GOP 내에서 정보의 양이 상대적으로 적거나, 화질이 좋지 않은 프레임을 바탕으로 학습한다면, 비디오 내부의 정보를 완전히 활용하지 못한 학습 결과가 될 수 있다.

Low-Delay 모드와 Random-Access 모드에서 GOP 내에 정보 손실이 적은 프레임과 큰 프레임이 존재한다. 앞선 문제를 해결하기 위해 그림 6처럼 Slow Stream에 정보 손실이 적은 프레임인 I-Frame 혹은 QP가 낮은 프레임들만 넣게 된다면, 모델은 프레임 내의 객체를 명확히 인식할 수 있으며, 이를 통해 최적화된 결과에 도달할 수 있다. 또한 Fast Stream에도 정보 손실이 적은 프레임 위주로 학습한다면, 객체의 행동을 보다 잘 파악할 수 있고, 이를 통해 성능이 향상된다.

3.2 제안 시스템 아키텍처

기존의 아키텍처는 과도하게 중앙화 되어있어, 급격한 트래픽 변화에 대응하기 쉽지 않다. 또한 관제 시스템의 특성인 멀티미디어 전송은 일반적인 Transaction 처리에 비해 큰 비용을 요구하기 때문에 사용자에게 버퍼링과 같은 문제가 발생할 수 있다. 따라서 기존 중앙 집중적 서버의 단점을 극복한 다중 서버 기반 관제 시스템을 제안한다.

제안하는 시스템의 구조는 그림 7이다. CCTV를 통해 생성된 영상들은 비디오 제어 서버를 통해 기록되고 동시에 인공지능 서버를 통해 영상이 실시간으로 전달된다. 인공지능 서버에서는 실시간으로 영상 내에 이상행동 객체가 존재하는지 판단하고, 존재할 경우 해당 프레임과 메타 데이터를 클라우드 기반 파일 시스템 서버에 전달해 저장한다.

클라이언트 서버가 실시간 혹은 과거 영상 요청을 한

다면 로드밸런서를 통해 가장 오버헤드가 적은 서버의 주소를 알려주고, 클라이언트 서버는 해당 주소를 통해 파이프라인을 만들어서 전용 점대점 연결을 한다. 이때 사용되는 Scheduling 기법은 프로세스 수를 고려하는 Multi-Level Queue Scheduling을 사용한다.

파이프라인이 구축되면, 클라이언트가 연결을 종료할 때까지 계속해서 멀티미디어 데이터를 전송한다. 그리고 데이터가 저장된 클라우드에 접근하는 상황에도 클라우드 로드밸런서를 사용하여 I/O 연산량이 가장 적은 클라우드와 연결한다. 이를 통해 멀티미디어 전송 오버헤드를 최대한으로 줄임으로써 실시간성을 극대화하고, 사용자 친화적인 서비스를 제공할 수 있다.

[2] Sullivan, G. J., Ohm, J. R., Han, W. J., & Wiegand, T. (2012). Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12), 1649-1668.

[3] Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1933-1941).

[4] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slow fast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6202-6211).

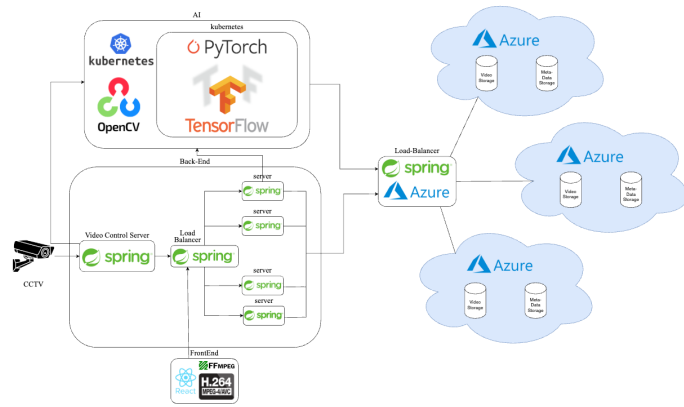


그림 7. 제안 시스템 아키텍처

4. 결론

기존의 관제 시스템은 중앙 집중적인 서버 구조로 과도한 트래픽 상황에서 실시간으로 동영상과 같은 멀티미디어를 전송할 때 문제가 발생할 수 있다. 또한 기존의 비디오 행동 인식 모델은 대부분 모델의 구조를 변경하거나, 더 깊고 넓은 네트워크를 구성하여 시스템의 성능을 향상하려 한다. 하지만 이러한 방법은 비디오의 모든 프레임이 압축되지 않고, 프레임마다 같은 정보의 양을 가지고 있음을 가정한다.

본 논문에서는 분산 서버 기반의 관제 시스템 모델과 함께 객체의 이상행동을 기존 인공지능 모델보다 더 정확하게 판별할 수 있는 시스템을 제안한다. 비디오 부호화 기술의 특성을 파악한 후, 이를 명시적으로 활용하는 모델을 같이 사용하여 미래에 발생할 수 있는 범죄를 효과적으로 예방할 수 있다.

5. Acknowledgement

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음” (2021-0-01082)

6. 참고 문헌

[1] Tamhankar, A., & Rao, K. R. (2003, July). An overview of H. 264/MPEG-4 Part 10. In *Proceedings EC-VIP-MC 2003. 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications (IEEE Cat. No. 03EX667)* (Vol. 1, pp. 1-51). IEEE.

(출판용 논문작성양식) * 저자정보 포함

음성인식 및 감정분석 기술 기반 감정일기 어플리케이션 설계 및 구현

김태환¹, 오동휘¹, 이휘찬¹, 한의진¹, 정현준², 정설영³

경북대학교, AtlasLabs

wksgo123@knu.ac.kr, hui1495@knu.ac.kr, kasterra@naver.com,
risingstaruj@gmail.com, jun.chung@atlaslabs.ai, snowflower@knu.ac.kr

Design And Implementation Of A Diary Application Using Voice Recognition and Sentiment Analysis Technology

Tae-hwan Kim¹, Dong-hwi Oh¹, Hwi-chan Lee¹, Ui-jin Han¹, Hyeon-jun Jeong²
and Seol-young Jeong³

Kyungpook National University, AtlasLabs

요약

본 논문에서는 STT(음성인식) 및 감정분석 기술을 활용한 감정일기 어플리케이션의 개발을 다룬다. 이 어플리케이션은 사용자가 일상적으로 느끼는 감정을 음성 인식을 통해 텍스트로 기록할 수 있고, 이를 감정 분석을 통해 감정 점수로 변환한 후, 한 달 동안의 감정 변화를 그래프로 시각화할 수 있다. 더 나아가서, 이 어플리케이션이 기분 기록지 작성을 돕는데 기여할 수 있으며, 따라서 사용자가 자신의 감정을 더 잘 이해하고 관리할 수 있도록 돕는 새로운 도구로서의 역할을 할 수 있음을 강조한다.

1. 서론

본 연구는 음성인식 기술(STT, Speech-To-Text)와 감정분석(Sentiment Analysis) 기술을 활용한 감정일기 어플리케이션의 설계 및 구현을 다룬다.

음성인식 기술(STT)은 사용자가 음성으로 감정을 표현하면 이를 텍스트로 변환하여 저장함으로써, 음성 메모를 효과적으로 기록할 수 있게 해준다. 감정분석 기술은 사용자가 저장한 텍스트를 분석하여 해당 텍스트의 감정 내용을 추출하고, 이를 정량화된 감정 점수로 표현한다. 음성 인식 기술과 감정분석 기술은 딥러닝과 빅데이터 기술을 기반으로 극적인 성능 향상을 이루었으며, 감정일기 어플리케이션의 개발은 이러한 기술의 상용화에 기여한다.

또한 현대 사회에서 스트레스와 감정 표현의 중요성이 커지고 있는 상황에서, 이 앱은 사용자가 일상적으로 느끼는 감정을 기록하고 시각적으로 분석할 수 있는 도구로 제공된다. 이를 통해 사용자는 감정 변화를 관찰하고 관리할 수 있으며, 개인의 감정 변화에 대한 통찰을 얻을 수 있다.

2. 관련 연구

이 논문에서는 기분 기록지(무드 차트)와 관련된 연구를

고려한다. 최종적으로 만들어진 어플리케이션이 기분 기록지를 작성하고 관리하는데 간접적으로 도움을 줄 수 있어야 함을 고려하였다. 기분 기록지는 사용자의 감정 상태를 일정 기간 동안 그래프로 시각화하여 보여주는 도구로, 사용자의 감정 변화를 파악하고 분석하는데 유용하다.[1]

여기에 더하여 AI 챗봇을 통해 기분기록지를 직접 작성하도록 하는 기존 어플리케이션[2] 또한 고려하였다. 본 어플리케이션은 사용자 접근성을 높이기 위해 기분 기록지를 직접 작성하도록 하는 방식이 아닌 보다 간접적인 방법을 취한다. 상담사 역할을 하는 챗봇을 배제하고, 개인이 기록하는 일기장의 방식을 채택하였다. 또한 일기장에 감정 분석 기술을 적용하여 보다 객관적인 시선으로 스스로 본인의 감정을 돌아볼 수 있게 했고, 텍스트로 작성하는 방식이 아닌 STT 기술을 활용하여 단순히 말하는 방식으로 일기를 기록하게 함으로써 접근성을 높였다.

3. 전체 시스템 구조 및 설계

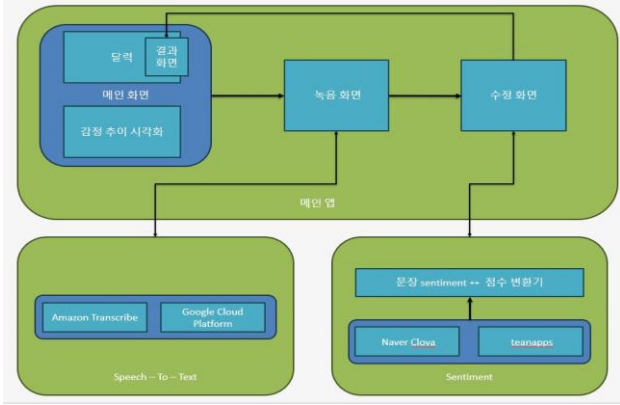


그림 1. 어플리케이션 흐름 및 구조도

Fig 1. Application Flow and Structure

그림 1에서는 본 어플리케이션의 작동 흐름을 간략하게 소개한다. 사용자가 앱을 실행하면, 녹음 버튼을 통해 음성으로 일기 형식으로 일일 감정을 표현한 음성 데이터를 받는다. 이어서 음성 인식 모듈에서 이진 음성 파일을 텍스트 형태로 변환한다. 변환된 텍스트는 감정 분석 모듈을 거쳐 임의의 감정 점수(Mood Score)로 변환된다.

3.1 음성 인식 모듈

음성 인식 모듈에서는 사용자의 음성이 입력되면 음성 신호를 문자로 변환하는 Speech To Text(STT)를 수행한다. 본 소프트웨어에서는 해당 기능을 구현하기 위해 시중의 STT 서비스 중 Google Cloud Speech API와 Amazon Transcribe를 사용한다. 둘은 퍼포먼스 및 비용 산정 문제 등 필요한 요구 조건에 따라 서버 단에서 교체될 수 있도록 했다.

3.2 감정 분석 모듈

음성 인식 모듈 과정에서 사용자의 음성 신호를 문자로 변환한 텍스트 파일을 감정 점수로 변환한다. 여기서 감정 점수를 다루는 Sentiment는 흔히 감정이라고 부르는 Emotion과는 다른 개념으로, 기쁨, 우울, 분노 등의 감정유형을 나타내기보다는 긍정,부정의 극성(Polarity)를 가진다.

$$Mood\ Score = 3 \times \frac{(W_{positive} - W_{negative})}{100}$$

그림 2. 감정 분석 결과를 기분 점수로 환산하는 수식
Fig 2. Formula for converting sentiment analysis results into mood scores

그림 2는 감정 분석 모듈이 Sentiment API로부터 받은 문장 별 Sentiment 분석 결과를 글 하나의 기분 점수로 환산하는 과정을 수식으로 나타낸 것이다. 동작 방식은 다음과 같다. 우선, 글을 문장으로 분할한다. 이후 각 문장에 긍정-중립-부정으로 분류하는 감정 분석을

수행한다. 이를 합산하여 전체 문단의 긍정-중립-부정 비율을 0~1 사이의 실수로 표현한다. 마지막으로 각 문장의 감정 분석 결과를 취합하여 -3 ~ +3 사이의 기분 점수로 변환한다. 긍정-부정-중립 중 가장 큰 감정에 가중치를 두고 긍정-부정의 비율을 고려하여 점수를 산정한다. 감정 분석 모듈에서는 시중의 자연어 처리 서비스 중 CLOVA Sentiment를 사용한다.

3.3 어플리케이션

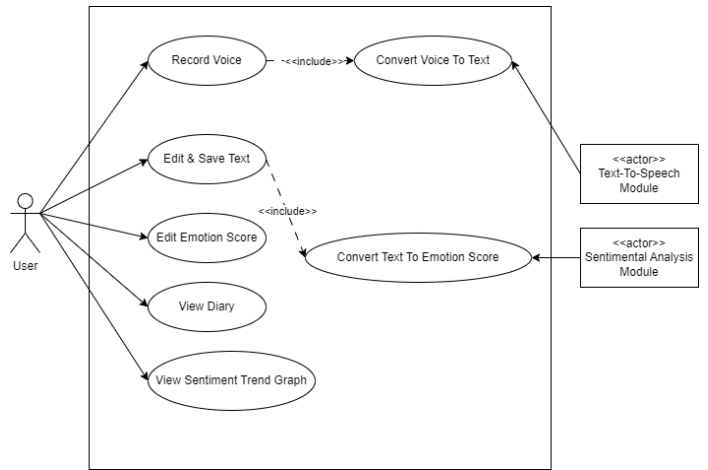


그림 3. 일기 어플리케이션 유즈케이스 다이어그램
Fig 3. Diary Application Use Case Diagram

그림 3은 본 어플리케이션에 필요한 기능을 유즈케이스 다이어그램으로 나타낸 것이다. 메인 어플리케이션의 주요 기능은 다음과 같다. 사용자의 음성 데이터를 스마트폰을 통해 녹음할 수 있다. 녹음된 음성 데이터는 음성 인식 모듈을 통해 텍스트로 변환된다. 텍스트는 다시 감정 분석 모듈을 통해 감정 점수로 변환된다. 사용자는 텍스트와 감정 점수를 일기 형태로 저장할 수 있다. 또한 사용자는 저장된 일기의 텍스트와 감정 점수를 자유롭게 수정할 수 있다. 어플리케이션은 하루의 일기와 감정점수를 사용자에게 보여준다. 한 달의 감정 변화 추이를 그래프 형식으로 시각화 할 수 있다. 메인 어플리케이션은 React Native를 이용하여 iOS/Android 크로스 플랫폼으로 개발하였다.

4. 시스템 구현 결과

다음과 같이 시스템을 구현하였다. 아래의 그림 4와 그림 5는 어플리케이션 화면을 캡처한 것이다.

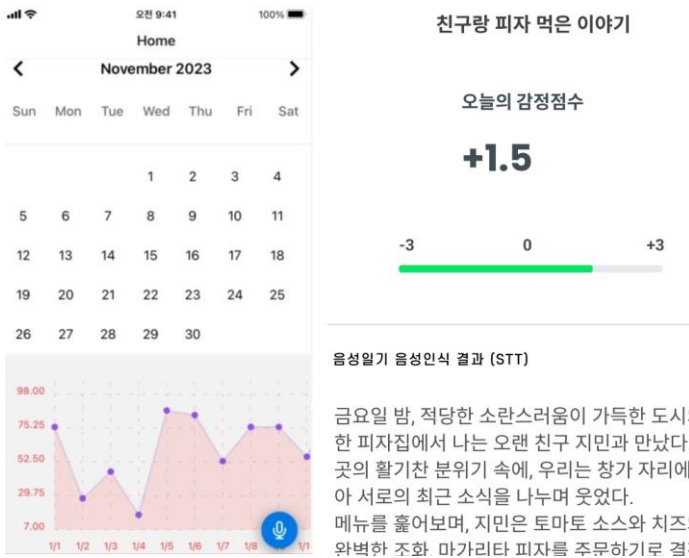


그림 4. 어플리케이션 메인 화면(a) 일일 결과 화면(b)
 Fig 4. Application Main Screen(a) Daily Results Screen(b)

그림 4(a)는 어플리케이션의 메인 화면이다. 크게 달력 부분과 해당하는 달의 기분 점수 추이를 나타낸다. 그림 4(b)는 달력에서 해당 날짜 버튼을 눌러 일일 결과 화면으로 진입한 것이다. 일일 결과 화면에는 당일 저장한 일기의 제목, 감정 점수, 내용이 기록된다.

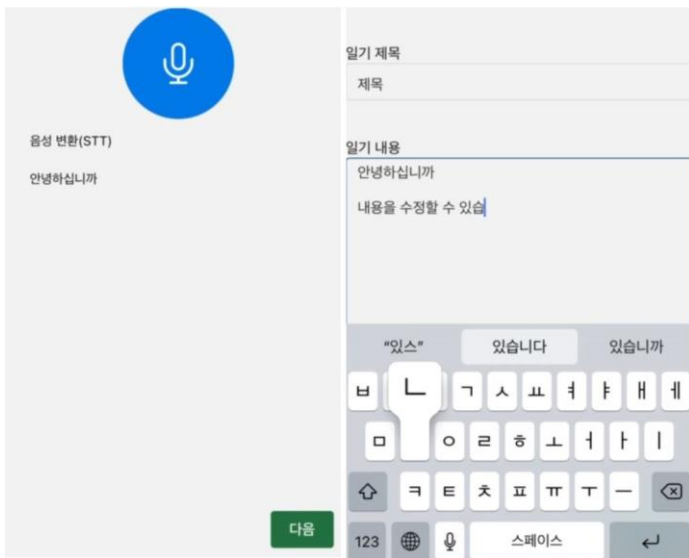


그림 5 녹음 화면(a) 수정 화면(b)
 Fig 5. Recording Screen(a) Editing Screen(b)

이은 그림 5(a)에서는 메인 화면의 녹음 버튼을 통해 진입할 수 있는 녹음 화면이다. 녹음 화면에서는 녹음 버튼을 눌러 사용자의 음성을 기록할 수 있고, 기록된 음성은 음성 인식 모듈을 통해 STT 처리되어 텍스트로 표시된다. 그림 5(b)는 녹음 화면에서 다음 버튼을 통해 진입하는 수정 화면이다. 수정 화면에서는 일기의 제목과

STT 처리된 텍스트를 수정할 수 있다. 수정 화면에서 다음 버튼을 누르면 감정 분석 모듈이 일기에 감정 점수를 부여한다. 작업이 종료되면 해당 기록이 저장된 그림 2(b)의 일일 결과 화면으로 진입한다.

5. 결론

본 연구에서는 Speech-To-Text와 Sentiment Analysis 기술을 활용한 감정일기 앱의 개발을 성공적으로 수행하였다. 사용자는 일상 감정을 기록하고 감정 점수를 통해 자신의 감정 상태를 시각적으로 파악할 수 있다. 이를 통해 개인의 감정 변화에 대한 보다 객관적인 자료를 얻을 수 있으며, 기분 기록지 작성에 있어 감정일기 앱을 활용할 수 있다. 이 연구는 감정 관리에 관한 새로운 사용자 경험과 자기 인식을 제공하는데 기여할 것으로 기대된다.

참고 문헌

- [1]한지연,김여진,조화연,이자경,and오유란."MoodChart Bot: AI 챗봇을 활용한 기분기록지 어플리케이션 설계 및 구현." 멀티미디어학회논문지 26.3 (2023): 503-508.
- [2]유희정, 최정은, 김정현, 최승원, 하태현, 하규섭, 차보석, 장재승. (2009). 환자용 기분기록지 개발 및 타당도 연구. 신경정신의학, 48(3), 174-181.

ACKNOWLEDGEMENT

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 중심대학사업의 연구결과로 수행되었음” (2021-0-01082)

- ¹ 경북대학교 컴퓨터학부
- ² (주)AtlasLab, VP of Engineering (교신저자)
- ³ 경북대학교 컴퓨터학부 교수 (교신저자)

SentenceBERT와 ChatGPT를 활용한 갱년기 상담 챗봇 서비스 고도화

이은주¹, 오유나², 서동혁¹, 이재현¹, 최혜성¹, 정설영(교신저자)^{1*}

¹경북대학교 컴퓨터학부, ²㈜리소프트

leeun5423@naver.com, call@resoft.kr, eutopia115@knu.ac.kr, good4170@gmail.com, bd2845@knu.ac.kr, snowflower@knu.ac.kr

Advancement of Menopause counseling ChatBot service Using SentenceBERT and ChatGPT

Eun-joo Lee¹, Yu-Na Oh², Dong-hyeok Seo¹, Chae-Hyun Lee¹, Hye-Seong Choi¹, Seol-Young Jeong^{1*}

¹School of Computer Science and Engineering, Kyungpook National University, ²Resoft Inc,

요약

본 논문에서는 갱년기 자가진단 셀프 케어 플랫폼 어플리케이션 <리프리티>의 Dialogflow기반 시나리오형 챗봇을 SentenceBERT와 ChatGPT를 활용하여 시기반 Retrieval+Generative 챗봇으로 고도화 하는 것을 제안한다. 고도화된 챗봇은 SentenceBERT와 ChatGPT를 활용하여 사용자의 질문에 가장 적절한 답변과 다음 질문을 예상하여 제공한다. SentenceBERT로 데이터셋에 존재하는 질문들의 다차원벡터를 계산해 저장해두고 사용자 질문의 다차원벡터와 코사인 유사도를 계산한다. 이후, 사용자 질문과 가장 유사한 질문을 데이터셋에서 선택해 답변을 제공한다. 코사인 유사도가 일정 값보다 작을 시 시스템은 사용자의 질문이 갱년기와 관련이 적다고 판단해 ChatGPT에 적절한 답변을 요청하여 사용자에게 답변을 제공한다. 추가적으로, 다음 질문을 ChatGPT를 통해 예측해 사용자에게 제시하여 대화의 흐름을 이어간다. 이와 같이 SentenceBERT와 ChatGPT를 활용하여 적은 데이터셋에서도 높은 정확도로 답변을 제공하고 자연스러운 대화의 흐름을 이어가는 챗봇 시스템을 제안한다.

1. 서론

갱년기 자가진단 셀프 케어 플랫폼 어플리케이션 <리프리티>에서는 사용자가 갱년기를 자가 진단하고 증상에 따른 헬스케어를 받을 수 있도록 챗봇 서비스를 지원한다. 해당 챗봇은 시나리오형 챗봇으로 사전에 개발자가 정해 놓은 질문과 답변을 선택하여 제한된 응답만을 제공한다. 이는 개발자가 원하는 방향으로 대화를 유도하여 대화를 이어나가는 방식이기 때문에, 상황에 맞는 유연한 답변을 제공하는 데 한계가 있어, 사용자별로 맞춤형 서비스를 제공한다는 목적을 달성하기에는 무리가 있다.

이를 해결하기위해 SentenceBERT와 ChatGPT를 활용하여 시기반 Retrieval+Generative 챗봇으로 고도화할 것이다. 이를 통해 적은 데이터셋만으로도 높은 정확도로 사용자의 다양한 질문에 자연스럽게 적절하게 답변하는 것을 기대한다.

2. 관련연구

2.1 Dialogflow

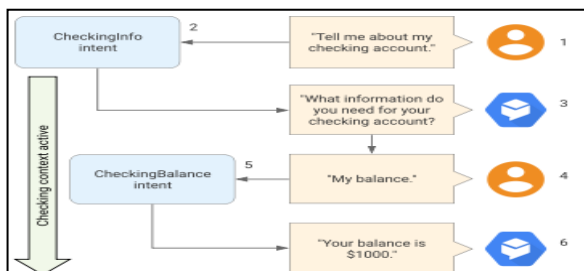


그림 1, Google Dialogflow

Dialogflow[1]는 GCP(Google Cloud Platform)에서 제공하는 NLP 플랫폼이다. 위 그림1과 같이 사전 질문과 그에 해당하는 답변을 지속적으로 이어지는 대화의 형태로 작성할 수 있다. 이렇게 작성된 질문과 답변들은 실제 사용자가 질문할 때 그 질문의 주제를 파악하여 사전에 작성했던 답변을 찾아 답변하는 모습을 볼 수 있다.

이처럼 Dialogflow를 이용하면, 직접 LLM(대형 언어 모델, Large Language Model)을 개발, 학습하는 과정을 거칠 필요 없이 제공되는 서비스만을 이용하여 쉽고 간편하게 챗봇을 제작할 수 있다.

하지만, 충분한 양의 질문과 답변의 flow가 작성되어 있지 않을 경우, 사용자가 질문을 하더라도 그에 해당하는 답변과 그 이후의 질문을 못해 정상적인 작동이 어렵다는 단점이 있다.

또한, 대화 도중 주제에 벗어난 질문을 할 때 답변을 유도하지 못하는 단점이 있다. 예를 들어 '날씨'와 관련된 대화를 할 때 질문자가 '기차'에 대해 질문하는 것과 같이 flow에서 벗어난 질문을 하면 올바른 답변을 하지 못하거나 작동하지 않는다. 따라서 Dialogflow를 이용해 챗봇을 개발할 경우 대량의 데이터셋을 통해 충분한 양의 flow를 구축할 필요가 있다.

2.2 SentenceBERT (SentenceBERT) with ChatGPT

SentenceBERT는 기존 BERT 모델에 문장(Sentence) 측면에서 향상을 가해 효과적으로 문장간 비교를 수행할 수 있는 Retrieval용으로 적합한 모델이다.[2]

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - GloVe	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	78.46	74.90	80.99	76.25	79.23	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	74.53	77.00	73.18	81.85	76.82	79.10	74.29	76.68

그림 2. BERT와 SentenceBERT간 문장 유사도(코사인 유사도 사용)

실질적인 ‘챗봇’ 구현에 필요한 질문간의 유사도를 찾기 위해서는 문장을 토큰화 하고, 이를 다차원 배열로 바꿔 벡터화 해야한다. 이렇게 생성된 벡터는 코사인 유사도를 이용하여 각 문장이 얼마나 유사한지를 비교할 수 있는데, 만 개의 문장들을 서로 분석하는데 BERT는 65시간이 걸렸지만, SentenceBERT는 5초(코사인 비교는 0.01)초로 획기적인 속도를 보여준다.[2] 또한, 그림 2에서도 볼 수 있듯이 기존 BERT모델에 비해 더 유사도가 높은 문장을 찾아낸 것을 확인할 수 있다. 이러한 문장 비교 방식과 성능 덕분에 적은 양의 질의응답 데이터셋을 가지고도 빠른 시간내에 질문을 비교하고 답변을 찾아내는데 적합하다. 따라서 해당 챗봇을 구현하기 위해 한국어로 파인튜닝 및 학습된 SentenceBERT 모델인 ‘ko-sroberta-multitask’ 모델을 사용하고자 한다. 하지만, SentenceBERT를 사용하더라도 데이터셋에 없는 질문을 할 때, 코사인 유사도가 낮아 적절한 대답을 제시하지 못한다는 단점이 있다. 이러한 단점을 생성형 AI인 ChatGPT를 이용함으로써 해결하고자 한다.

ChatGPT는 OpenAI사에서 LLM인 GPT-3.5를 기반으로 개발된 생성형(Generative) AI, 챗봇이다. LLM으로 학습된 데이터가 많은 ChatGPT 특성상, 수많은 질문을 하더라도 사용자의 요청 맥락에 맞게 맞춤형 답변을 생성하는 것을 볼 수 있다. [3]

또한, ChatGPT의 특성으로 기존 대화내역을 기억하고 이를 바탕으로 답변을 생성하기 때문에 일상적인 대화를 하더라도 Dialogflow처럼 대화를 이어나갈 수 있는 일종의 ‘흐름’을 형성할 수 있어 훨씬 더 자연스러운 대화를 만들 수 있다. 이를 통해, 사용자에게 실제 사람과 대화하는 듯한 자연스러움을 느끼게 할 수 있다.

하지만, ChatGPT는 특정 주제에 대한 전문적인 답변을 학습하지 못했기 때문에 전문적인 답변을 위해 전문적인 답변에 대한 데이터베이스를 구축하고, 위에서 제시한 SentenceBERT를 활용하여 질문에 대한 답변을 제시할 것이다. 이 때, SentenceBERT를 이용해 데이터베이스에서 답변을 찾지 못한 질문만을 ChatGPT에서 처리하는 시스템을 개발하고자 한다.

3. 시스템 구현

본 시스템에서는 갱년기 관련 데이터베이스를 활용한 SentenceBERT 모델과 자연스럽게 유연한 대화를 위한 ChatGPT를 사용한다. 소스코드 작성과 테스트는 Google Colaboratory를 통해 이루어졌다.

Q	A	Label
0	원래 침을 많이 자는 편이었는데, 요즘은 거의 못 자서 잠들기가 힘들다. why... 갱년기 불면증은 갱년기 여성들이 겪는 수면장애로, 수면의 질과 양이 감소하는 것이...	불면증
1	원인이 무엇인가요? 갱년기 불면증이 원인은 여성 호르몬인 에스트로겐의 분비 감소로 인한 수면의 질과 양...	불면증
2	수면에 억지 않고 잘 수 있는 방법이 있을까요? 갱년기 불면증 완화 방법으로는 처방 제를 유지, 천연 보충제 섭취, 건강한 식단,...	불면증
3	수면환경 개선은 어떻게 하나요? 수면환경 개선이라는 것이 어떤 것을 말씀하시는 건가요? 침대, 방, 조명 등을 말씀...	불면증
4	침대, 방, 조명 같은 거 맞아요. 밝히는 편만하게, 방은 시원하게, 조명은 어두운 것이 좋습니디. 또한, 스트레스...	불면증
5	수면환경 개선 방법 중에서 방을 시원하게 하는 방법은 무엇이지요? 커튼이나 블라인드로 창문을 가리기, 내부로 들어오는 열기의 대부분은 창문을 통해 들...	불면증

그림 3. 데이터셋 예시

Q	A	Label	emb=dlng
0	원래 침을 많이 자는 편이었는데, 요즘은 거의 못 자서 잠들기가 힘들다. why... 갱년기 불면증은 갱년기 여성들이 겪는 수면장애로, 수면의 질과 양이 감소하는 것이...	불면증	[0.42969008, -0.38425782, 0.54019003, -0.10943...]
1	원인이 무엇인가요? 갱년기 불면증이 원인은 여성 호르몬인 에스트로겐의 분비 감소로 인한 수면의 질과 양...	불면증	[0.08331859, -0.49912417, 0.22727801, -0.38341...]
2	수면에 억지 않고 잘 수 있는 방법이 있을까요? 갱년기 불면증 완화 방법으로는 처방 제를 유지, 천연 보충제 섭취, 건강한 식단,...	불면증	[0.19923586, -0.18425215, 0.542422, 0.3076777...]
3	수면환경 개선은 어떻게 하나요? 수면환경 개선이라는 것이 어떤 것을 말씀하시는 건가요? 침대, 방, 조명 등을 말씀...	불면증	[0.58547362, 0.14412792, 0.28519387, 0.1422882...]
4	침대, 방, 조명 같은 거 맞아요. 밝히는 편만하게, 방은 시원하게, 조명은 어두운 것이 좋습니디. 또한, 스트레스...	불면증	[0.25970787, -0.28704673, -0.61577587, -0.39591...]

그림 4. 질문의 임베딩 값이 함께 저장된 데이터셋

데이터셋은 기업에서 제공한 데이터베이스를 바탕으로 구성되었고, 추가적으로 크롤링한 데이터를 함께 포함하고 있다. 위의 그림3과 같이 갱년기 관련 질문, 그에 맞는 답과 분류 레이블로 이루어져 있다. 미리 크롤링한 갱년기 관련 데이터들을 활용하기 위해 SentenceBERT로 데이터셋에 있는 질문들의 임베딩 값을 구하여 데이터셋에 추가해 저장한다. 그림4에서 각 질문의 임베딩 값이 함께 저장된 것을 확인할 수 있다. 사용자가 질문을 던졌을 때 입력된 질문이 데이터셋에 저장된 질문과 유사할 경우 선행 학습된 답을 제시해주기 위해 SentenceBERT를 사용하여 입력된 질문의 임베딩 값을 구한다.

```
# calculate cosine similarity
def cos_sim(A, B):
    return dot(A, B)/(norm(A)*norm(B))
```

그림 5. 코사인 유사도 계산 함수

이러, 해당 질문의 임베딩 값과 데이터셋에 저장된 질문들의 임베딩 값의 코사인 유사도를 구한다. 그림 5와 같이 코드에서 정의한 cos_sim 함수를 통해 코사인 유사도가 가장 높은 질문을 가져오고 데이터셋에 저장된 이에 맞는 답을 제시하도록 한다. 하지만, 코사인 유사도가 가장 높은 질문을 가져온다고 하더라도 유사한 질문이 없어 어색한 답을 제시할 수 있다. 이를 방지하기 위해 그림 6과 같이 코사인 유사도가 0.64보다 높을 때만 유사하다고 판단하고 그에 해당하는 답을 제시하도록 하였다. 여기서 0.64라는 기준은 여러 번의 테스트를 통해 경험적으로 정한 값이다.

데이터베이스에 유사한 질문이 없을 경우(최대 코사인 유사도 값이 0.64 이하일 때) ChatGPT모델을 활용한다. ChatGPT API를 사용해 사용자의 질문을 ChatGPT로 보내고 제안된 답을 전달받아 사용자에게 제시하여 자연스러운 대화를 이끌어나간다.

```
if high_dist <= 0.64 :
    chat = openai.ChatCompletion.create(
        model="gpt-3.5-turbo", messages=texts
    )
    reply = chat.choices[0].message.content
    print(f"Repretty chatbot : {reply}")
    texts.append({"role": "assistant", "content" : reply})
else :
    print(f"Repretty chatbot : {answer}")
```

그림 6. 코사인 유사도에 따라 처리하는 방식을 구분

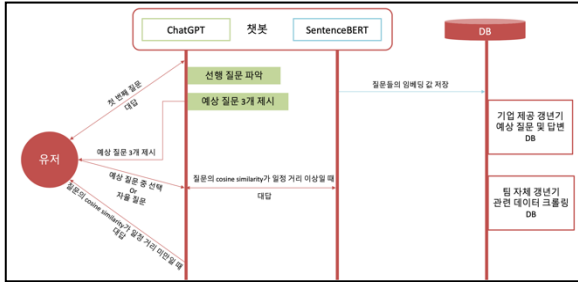


그림 7. 시스템 상세 설계도

그림7은 시스템 상세 설계도를 나타낸다. 대화 흐름을 간단하게 설명하자면 다음과 같다. 사용자가 선행 질문을 하고 리프리티 챗봇은 위에 설명한 두 개의 모델을 활용해 도출된 질문에 대한 답과 함께 사용자가 굳이 궁금해할 만한 질문 3개를 그림8에 나타난 코드를 통해 예측하고 이를 사용자에게 제시한다. 이를 통해 사용자의 입력 시간을 줄일 수 있다. 사용자는 보기에 제시된 질문들 중 선택할 수 있고, 제시된 질문이 아닌 그 외의 질문을 하고 싶을 땐 직접 입력해서 질문할 수 있다. 다음 질문을 주고 받을 때도 이 과정의 반복이다.

```
questions = [{"role": "user", "content": ""},
question = "" + text + ""에 후행될만한 질문을 갱년기와 관련하여 3개 예측해줘"
questions.append({"role": "user", "content": question,})
```

그림 8. ChatGPT에게 후행 질문 예측 요청

```
def isSimilar(question):
    theme = [{"role": "user", "content": ""},
question = "" + question + ""라는 질문은 갱년기와 관련이 있을까? 조금이라도 관련있으면 True, 그게 아니라면 False로만 대답해줘"
    # print(question)
    theme.append({"role": "user", "content": question,})

    # print(theme)
    chat = openai.ChatCompletion.create(
        model="gpt-3.5-turbo", messages=theme
    )
    reply = chat.choices[0].message.content
    # print(f"Repretry chatbot : {reply}")

    if(reply == "True"):
        return 1
    else:
        return 0
```

그림 9. isSimilar 함수

사용자의 선행 질문이 갱년기와 관련이 없고 다음 질문을 예측하기에 의미 없는 질문일 경우 예측 질문 3개를 제시하지 않는다. 이를 구분하기 위해 그림9와 같이 isSimilar 함수를 정의했다. 사용자의 질문은 ChatGPT로 넘겨져 갱년기와 관련이 있는 질문인지 판단된다. 이 때 ChatGPT에게 갱년기에 관한 상담을 해주는 상담가라는 역할을 부여하고 질문과 갱년기가 어느 정도 관련이 있을 땐 True, 아닐 시에는 False의 값을 리턴한다.

```
similar = isSimilar(text)
if similar == 1:
    chat = openai.ChatCompletion.create(
        model="gpt-3.5-turbo", messages=questions
    )
    reply = chat.choices[0].message.content
    print(f"예측 질문 : {reply}")
```

그림 10. True일 경우 ChatGPT에게 후행 질문 요청

False일 경우, 갱년기와 관련된 질문이 아니기 때문에 데이터셋에 저장된 질문과의 코사인 유사도를 비교할 필요가 없다. 따라서, ChatGPT가 사용자의 해당 질문을 받고 이에 맞는 답을 사용자에게 제시한다. 예를 들어, 사용자가 “안녕하세요”, “오늘 날씨는 어때요?”와 같은 일상적인 질문을 했을 땐 ChatGPT가 False의 값을 리턴하고 이어질 ChatGPT를 통해 답변만 제공할 뿐, 이어질 질문은 추천하지 않는다. True일

경우, 위에 설명한 두 가지 모델을 이용하여 적절한 답변을 제시해주고 그림10과 같이 다시 ChatGPT에게 후행될만한 질문을 예측하라고 한 후, 그 결과를 받아온다.

그렇게 사용자가 이어 할 법한 다음 질문 3가지를 질문하고 받아온다. 사용자는 3가지 질문 중 선택하거나, 직접 질문을 입력하여 대화를 이어간다.

추가로 ChatGPT API stream 기능을 이용하면 답변이 한글자씩 생성될때마다 결과값을 반환받을 수 있다. 이를 이용하여 답변이 전부 생성될때까지 기다리는 것 대신 한 글자씩 타이핑 하는 효과를 보여줘 사용자의 답답함을 해소할 수 있을거라 여겨진다.

4. 시스템 평가

최대 코사인 유사도와 갱년기와 관련성을 판별하는 isSimilar 함수의 반환값을 토대로 테스트 케이스를 아래와 같이 설정하여, 챗봇이 이상 없이 작동하는지 확인하기로 하였다,

1. 최대 코사인 유사도가 0.64 이상인 경우
2. 최대 코사인 유사도가 0.64 미만이지만, isSimilar 함수의 반환값이 True인 경우
3. 최대 코사인 유사도가 0.64 미만이고, isSimilar 함수의 반환값이 False인 경우

```
유저: 갱년기가 뭐야?
261
highest :
0.9072173
general chatbot : 갱년기는 여성의 생리기능이 노화되는 자연적인 과정으로서, 여성의 나이가 들면서 생리주기가 불규칙해지고, 갱년기의 정확한 원인은 아직 명확하게 밝혀지지 않았지만, 여성의 나이와 유전적인 요인이 중요한 역할을 합니다.
[{"role": "user", "content": ""}, {"role": "user", "content": "갱년기가 뭐야?"}]
예측 질문 : 1. 갱년기는 언제부터 시작되고 언제까지 지속되나요?
2. 갱년기 동안 어떤 신체적 변화가 일어나나요?
3. 갱년기가 정신적 건강에 어떤 영향을 미칠 수 있나요?
Repretry chatbot : 갱년기는 여성의 생리기능이 노화되는 자연적인 과정으로서, 여성의 나이가 들면서 생리주기가 불규칙해지고, 갱년기의 정확한 원인은 아직 명확하게 밝혀지지 않았지만, 여성의 나이와 유전적인 요인이 중요한 역할을 합니다.
```

그림 11. case 1 : 코사인 유사도 > 0.64

그림 11은 최대 코사인 유사도 값이 0.9072로, 데이터베이스에 저장된 문답을 가져오는 경우이다. 챗봇은 데이터베이스에 저장된 답을 가져와 사용자에게 제공한다. 또한 그 답을 본 사용자가 할 만한 다음 질문을 ChatGPT를 통해 예측하고 이를 사용자에게 추천하여, 사용자가 다음에 어떤 질문을 할지 생각할 필요 없이 사용자가 할 법한 질문을 제공하도록 했다.

```
유저: 잠이 안와
336
highest :
0.590363
general chatbot : 수면 문제를 해결하기 위해서는 잠을 못자서 피로감이 심해지는 경우, 잠을 자고 휴식하는 것이 중요합니다.
[{"role": "user", "content": ""}, {"role": "user", "content": "잠이 안와"}]
예측 질문 : 1. 갱년기는 잠에 영향을 주는 호르몬 중 하나일 수 있습니다. 갱년기에 어떠한 변화들이 수면 문제를 일으키는 잠이 안와를 유발하는 것일까요?
2. 갱년기 동안 잠이 안와는 경우가 많다고 들었는데, 어떤 이유로 잠이 안와야 할까요? 그리고 이를 개선하기 위해 일상생활에서 어떤 방법을 시도해볼 수 있을까요?
3. 갱년기가 정신적 건강에 어떤 영향을 미칠 수 있을까요? 그리고 이를 개선하기 위해 일상생활에서 어떤 방법을 시도해볼 수 있을까요?
Repretry chatbot : 잠이 안와라면 몇 가지 방법을 시도해볼 수 있습니다:
1. 편안한 환경 조성: 조용하고 어두운 방에서 잠자리를 마련하고, 편안한 온도, 적절한 습도를 유지합니다.
2. 활동 및 운동: 하루 동안 활동량을 늘리고, 규칙적인 운동을 통해 신체 활동을 늘리세요. 그러나, 운동은 수면 문제를 악화시킬 수 있으므로, 과도한 운동은 피하세요.
3. 스트레스 관리: 스트레스를 관리하기 위해 심호흡, 명상 또는 피트니스 클래스와 같은 스트레스 관리 프로그램을 시도하세요.
4. 수면 습관 유지: 항상 동일한 시간에 침대에 들어오도록 하여 내부 시계를 조절합니다. 수면 전에 휴대폰 사용을 피하고, 잠을 자기 전에 TV, 스마트폰 및 다른 전자 기기 사용을 피하세요.
5. 활동량 조절: 활동량을 줄이거나, 가벼운 운동이나 스트레칭을 시도하세요.
6. 활동 제한: 수면 시간을 줄이기 위해 침실에서 TV, 스마트폰 및 다른 전자 기기 사용을 피하세요.
7. 수면 전 리튬 조절: 수면 전에 가벼운 활동을 하는 것이 수면 문제를 악화시킬 수 있습니다.
8. 음식 및 음료 제한: 수면 전에 고지방 음식과 음료를 피하십시오. 카페인과 알코올은 수면 문제를 악화시킬 수 있습니다.
9. 수면 패턴 일시 변경: 잠을 못자고 있는 경우, 일찍 일어나는 것, 일찍 자는 것, 그리고 수면 패턴을 변경하세요.
10. 전문가와 상담: 만약 위의 방법들을 시도해도 잠이 안와한다면, 전문가에게 도움을 청하세요. 수면
```

그림 12. case 2 : 코사인 유사도 < 0.64, isSimilar = true

그림 12는 최대 코사인 유사도 값이 0.5902이며, 갱년기 증상과 관련성은 있으나, 데이터베이스에 관련 문답이 저장되어 있지 않아 ChatGPT를 통해 답변을 생성하고, 그 답변을 제공했다. 이 경우에는 데이터베이스에 답변을 제공할 수 있는 데이터셋이 존재하지 않기 때문에, 다음 질문을 추천하는 데 한계가 있다. 때

문에, “또 어떤 것이 궁금하세요?” 등의 다음 질문을 유발할 수 있는 질문을 추가적으로 제공해 사용자가 챗봇을 계속 이용할 수 있도록 했다.

```

[{"role": "user", "content": "안녕하세요?"}, {"role": "assistant", "content": "안녕하세요! 저는 Google Assistant입니다. 도움이 필요하시면 언제든지 말씀해주세요."}, {"role": "user", "content": "오늘은 어떤 날인가요?"}, {"role": "assistant", "content": "오늘은 2023년 10월 10일입니다. 화요일입니다. 오늘도 기분 좋게 하루를 보내시길 바랍니다."}, {"role": "user", "content": "오늘의 날씨를 알려주세요."}, {"role": "assistant", "content": "오늘의 서울 날씨: 맑음, 기온 15~22도. 미세먼지 농도 낮음. 쾌적한 날씨입니다."}, {"role": "user", "content": "건강에 좋은 음식이 뭐가 있을까요?"}, {"role": "assistant", "content": "건강에 좋은 음식으로는 과일, 채소, 견과류, 생선 등이 있습니다. 균형 잡힌 식단을 섭취하는 것이 중요합니다."}, {"role": "user", "content": "스트레스를 줄이는 방법이 있을까요?"}, {"role": "assistant", "content": "스트레스를 줄이는 방법으로는 규칙적인 운동, 충분한 수면, 명상 등이 있습니다. 스트레스 관리가 중요합니다."}, {"role": "user", "content": "오늘의 주요 뉴스는 무엇인가요?"}, {"role": "assistant", "content": "오늘의 주요 뉴스로는 AI 기술 발전, 기후 변화 대응 방안 등이 있습니다. 최신 뉴스를 확인하세요."}, {"role": "user", "content": "오늘의 주요 뉴스는 무엇인가요?"}, {"role": "assistant", "content": "오늘의 주요 뉴스로는 AI 기술 발전, 기후 변화 대응 방안 등이 있습니다. 최신 뉴스를 확인하세요."}, {"role": "user", "content": "오늘의 주요 뉴스는 무엇인가요?"}, {"role": "assistant", "content": "오늘의 주요 뉴스로는 AI 기술 발전, 기후 변화 대응 방안 등이 있습니다. 최신 뉴스를 확인하세요."}]]

```

그림 13. case 3 : 코사인 유사도 < 0.64, isSimilar = false

그림 12는 최대 코사인 유사도 값이 0.4345이며, 갱년기라는 주제와 관련성이 없다. 하지만, ChatGPT api를 이용하여 답변을 제공해 문제없이 챗봇과 사용자간 자연스러운 소통이 가능하도록 한다. 이 경우에는 사용자가 원래 목적에 맞게 이 챗봇을 사용하도록 “갱년기를 해결할 수 있는 방법에 대해서는 궁금하지 않으신가요?” 등의 질문을 이용해 사용자를 유도하는 방식을 이용했다.

5. 결론 및 제언

본 논문에서 제안한 갱년기 상담 챗봇 서비스는 병원에 내원하기에는 애매한 증상으로 인해 불편을 호소하는 갱년기 여성에게 삶의 질을 높일 수 있는 방법을 제안한다. 이전에 구현되어 있던 시나리오형 챗봇은 사용자가 처한 상황에 유연하게 대처하지 못해 맞춤형 서비스를 제공하기에는 제한이 있다.

이러한 문제점을 해결하기 위해, SentenceBERT와 ChatGPT를 이용하여 챗봇을 고도화하였다. 이 챗봇은 사용자의 질문을 SentenceBERT를 통해 구분하고, 임베딩 값에 따라 전문가 상담 데이터베이스에 기반한 Retrieval-based 챗봇이나, ChatGPT에 기반한 Generative-based 챗봇을 사용해 답변을 제공하는 구조로 이루어져 있다. 이를 통해, 다양한 질문에 따른 적절한 답변을 제공해 갱년기 여성이 건강한 삶을 유지하는 데에 기여할 것이다.

이 챗봇은 갱년기 여성의 효과적인 self-care를 도모하기 위해 개발되었다. 다만, 개발 진행 과정에서 추가적인 문제점을 찾을 수 있었기에, 다음과 같이 제언하고자 한다. 첫째, 데이터베이스의 데이터셋이 부족해 모든 질문에 전문적인 답변을 제공하지 못한 점이다. 데이터를 더욱 추가하고 서버 용량을 늘려 사용자에게 더욱 도움이 되는 답변을 마련해야 할 것이다. 둘째, 데이터베이스에 존재하지 않아 ChatGPT를 통해 답변을 얻은 질문이지만, 그 질문과 답변이 사용자에게 도움이 되는 경우이다. 위와 같은 질문을 데이터베이스에 즉시 추가하는 알고리즘을 도입해 ChatGPT에 대한 의존도를 줄이고, 이후 유사한 질문을 받았을 때 데이터베이스를 이용해 더욱 빠른 답변을 제공하도록 해야 할 것이다. 이처럼, 챗봇이 사용자에게 도움이 되는 질문

을 더욱 상세하고 빠르게 처리할 수 있다면, 사용자인 갱년기를 겪고 있는 사용자들이 자기주도적으로 삶의 질을 높이는 데 도움을 줄 것이라 기대한다.

6. 사사

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

"이 연구는 과학기술정보통신부 및 정보통신산업진흥원의 “지역산업 SW인재양성 기반조성 사업” 지원으로 수행되었음."

7. 참고문헌

[1] Sabharwal, Navin, et al. (2020). "Introduction to Google dialogflow." Cognitive virtual assistants using google dialogflow: develop complex cognitive bots using the google dialogflow platform : 13-54.

[2] Nils Reimers, Iryna Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 1-5, 학회 EMNLP 2019 (2019)

[3] 황정재. (2023). 알파고에서 챗GPT까지 AI 기술의 발전과 미래. Future Horizon. 55호. 34-37.

수상레저기구 위치정보 제공을 위한 프로토콜

최연욱⁰, 손희원, 권영우
(주)다이코기술서비스⁰¹⁾, 경북대학교 컴퓨터학부⁰²⁾, (주)위니텍, 경북대학교 컴퓨터학부
xnxn0403@knu.ac.kr, gmlnjs@winitech.com, ywkwon@knu.ac.kr

a protocol for providing location information for water leisure equipment

Choi Yeon-uk⁰, Son Hee-won, Kwon Young-woo
Daico Technology Service Co., Ltd.⁰¹⁾, Dept. of Computer Science Engineering, Kyungpook National University⁰²⁾,
Winitech Co., Ltd., Dept. of Computer Science Engineering, Kyungpook National University

요 약

최근 요트원정대, 도시어부 등 주요 방송매체의 활성화로 수상레저기구를 이용하는 이용자가 지속적으로 증가하고 있으며, 그에 따라 수상레저기구의 안전사고 또한 증가하고 있다. 수상레저기구는 위치발신장치를 의무적으로 비치하고 있는 선박(AIS), 어선(V-PASS)과 달리 위치발신장치를 비치할 의무가 없다. 이에, 수상레저기구의 사고가 발생할 경우 위치정보를 식별하는데 어려움이 있어 신속하게 구조하는 데 한계가 있다. 이 논문에서는 모바일 앱(바다내비앱)을 활용하여 모바일 사용자의 위치정보, 수상레저기구의 제원 정보를 해양경찰청 통합상황관리시스템으로 전송하여 모바일 앱 사용자들의 위치 및 수상레저기구의 제원 정보를 실시간으로 파악하여 관리할 수 있도록 모바일앱(바다내비앱)과 통합상황관리시스템을 연계하고 통합상황관리시스템에 수상레저기구를 표출함으로써 신속한 구조활동이 가능하게 한다.

1. 서 론

최근 요트원정대, 도시어부 등 주요 방송매체의 활성화로 수상레저기구를 이용하는 이용자가 지속적으로 증가하고 있다. 해양경찰청은 2010년 5,053대에서 2020년 31,503대로 최근 10년간 약 6배 이상 꾸준히 증가하고 있으며, 수상레저기구에 대한 체계적 관리와 안전성 확보를 위해 「수상레저기구의 등록 및 검사에 관한 법률」을 제정하고 있다[1].

수상레저기구 이용자가 지속적으로 증가하면서 수상레저기구의 안전사고 또한 증가하고 있으며, 한국해양교통안전공단(KOMSA)에 따르면 최근 5년간 월평균 수상레저기구 사고 선박 현황을 보면, 선박 사고 15,997척 가운데 수상레저기구 사고는 2,938척으로 18.4%를 차지했다[2].

모든 선박의 경우 「어선법」 제5조의2, 「선박안전법」 제30조에 따라 선박의 위치를 자동으로 발신하는 장치를 갖추고 이를 작동하여야 하며, 사고, 고장 등 긴급상황 발생 시 선박의 종류, 긴급구조요청, 선박 위치 확인 등이 가능해 신속한 구조활동이 가능하다.

하지만, 출항 신고 의무가 없는 이용객이 수상레저활동을 하다 사고가 발생할 경우, 수상레저활동에 활용되는 기구의 기본정보 및 위치 확인이 쉽지 않아 구조활동이 늦어지거나 구조활동에 어려움이 있어 대안 마련이 시급하다.

본 논문에서는 국민의 안전 및 사고예방을 위한 환경을 확보하고 신속하게 구조활동을 할 수 있도록 기존 시스템인 해양수산부 바다내비시스템과 해양경찰청 통합상황관리시스템을 연계하여 수상레저기구 위치표출 서비스를 제공할 수 있는 프로토콜을 제시하고자 한다.

2. 본 론

수상레저기구 위치표출 서비스는 기존 운영중인 해양수산부 바다내비(e-Navigation)시스템에서 수집한 수상레저기구 GPS 위치 정보, 수상레저기구 제원 정보를 TCP 소켓 통신을 활용하여 G/W연계서버로 정보를 수신하고, G/W연계서버를 통해 수집된 수상레저기구 GPS 위치 정보, 수상레저기구 제원 정보를 TCP 소켓 통신을 활용하여 실시간으로 통합상황관리시스템 GIS에 표출하는 것으로 다음 <그림 1>과 같이 구성하였다.



<그림 1> 수상레저기구 위치표출서비스 전체 개념
<Fig. 1> Overall concept of water leisure equipment location display service

수상레저기구 위치표출 서비스를 제공하기 위해 기존 해양수산부 바다내비(e-Navigation)시스템과 해양경찰청 통합상황관리시스템을 연계하여야 하며, 이를 위하여 수상레저기구 GPS 위치정보와 수상레저기구 제원 정보의 프로토콜 제시가 필요하다.

수상레저기구 GPS 위치정보와 수상레저기구 제원 정보의 프로토콜을 제시하기 위하여 (1)메시지 구조, (2)메시지 데이터 정의, (3)선종코드로 구성하였다.

(1) 메시지 구조

메시지 구조를 <표 1>과 같이 정의하였다.

<표 1> 메시지 구조(구분자 : ‘,’)

<Table 1> Message structure (separator: ‘,’)

구분	크기	설명	비고
시작문자	4	SENV	
시각	16	YYYYMMDDHHNNSSZZ	송신시각
메시지 타입	2	02 : 모바일 위치정보(수상레저 포함) 04 : 모바일 제원정보(정적정보) 91 : 시스템 수신 ACK	
데이터	가변	메시지 데이터 정의 참조	
끝문자	1	*	
CRC	2		
CRLF	2		

(2) 메시지 데이터 정의

메시지 데이터 정의는 (가)수상레저기구 GPS 위치정보, (나)수상레저기구 제원 정보, (다) 연결 G/W 연결상태 정보로 구성하였다.

(가) 수상레저기구 GPS 위치정보(수상레저 포함)

메시지타입		02	
구분	크기 (Byte)	설명	비고
선박 ID	가변	모바일 MRN - 수상레저기구일 경우 기구등록번호 (‘-’ 제외) (ex : 1307008)	
경도	가변	신호 발신위치의 경도(도 단위)	
위도	가변	신호 발신위치의 위도(도 단위)	
속도	가변	선박 속도	소수점 1자리까지
진행 방향	가변	선박 진행방향 (범위 : 0 ~ 359.9)	소수점 1자리까지
GPS 상태	1	GPS 장비상태 - 1 : 유효 - 2 : 유효하지 않음	
타겟 종류	가변	0 : N/A, 1 : 어선, 2 : 수상레저	

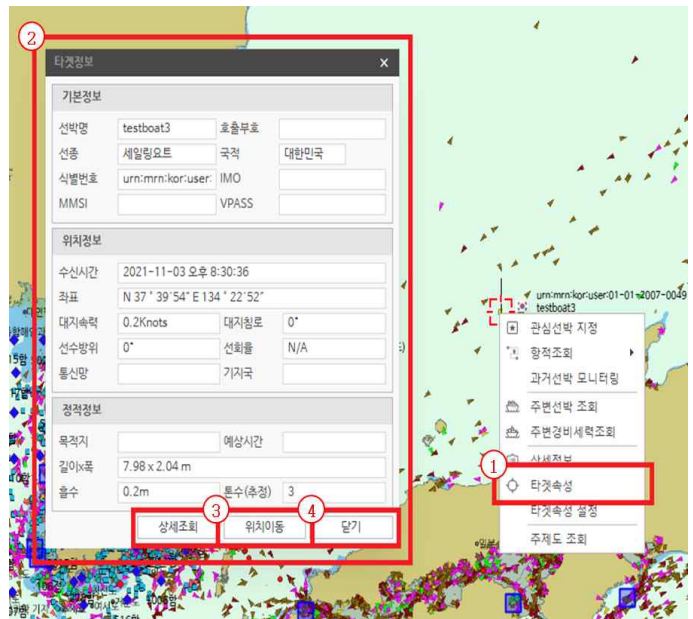
- 수신된 수상레저기구 이미지를 지도에 표출



(나) 수상레저기구 제원정보(정적정보)

메시지타입		04	
구분	크기 (Byte)	설명	비고
선박 ID	가변	모바일 MRN - 수상레저기구일 경우 기구등록번호 (‘-’ 제외) (ex : 1307008)	
관할권역 (등록지역)	가변	- 선박 관할 권역 코드 > 1-6 (1-6권역), 99(중앙) - 수상레저기구일 경우 등록지역	
선박명 (기구명)	가변	선박명(수상레저기구명)	
어선번호	가변		
선박번호	가변		
MMSI	가변	선박MMSI	
RFID	가변	선박 RFID	
IMO	가변	선박 IMO	
호출부호	가변	선박 호출부호(CallSign)	
기구등록 번호	가변	수상레저기구 등록 번호 - ex : 142070013 (‘-’ 제외)	
선종	6	- 선종코드 테이블 참조 - ex : A001	
길이	가변	선박 길이	
폭(너비)	가변	선박 너비	
깊이	가변	선박 깊이	
홀 수	가변	선박 홀 수	
툽 수	가변	선박 툽 수	
승선인원	가변	0 : N/A	

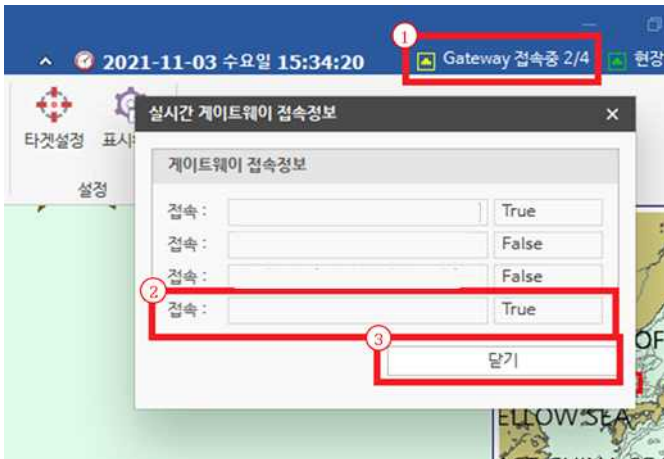
- 수신된 수상레저기구 정보를 조회



(다) 연결 G/W 연결상태 정보

메시지타입		02	
구분	크기 (Byte)	설명	비고
선박 ID	111111111	모바일 MRN - 수상레저기구일 경우 기구등록번호 (‘ - ‘ 제외) (ex : 1307008)	
경도	100	신고 발신위치의 경도(도 단위)	
위도	90	신고 발신위치의 위도(도 단위)	
속도	101.1	선박 속도	소수점 1자리까지
진행 방향	120.0	선박 진행방향 (범위 : 0 ~ 359.9)	소수점 1자리까지
GPS 상태	2	- 1 : 유효 - 2 : 유효하지 않음	
타겟 종류	0	0 : N/A, 1 : 어선, 2 : 수상레저	

- 수상레저기구의 실시간 연결 상태 정보를 확인



(3) 선종코드

선종코드는 「수상레저안전법」에 명시되어 있는 동력 수상레저기구 종류를 기준으로 하였다.

대분류	소분류	선종코드	비고
동력수상 레저기구 (G)	일반	G01	
	모터보트	G02	
	세일링요트	G03	
	수상오토바이	G04	
	스쿠터	G05	
	공기부양정	G06	

3. 결 론

코로나19 팬데믹 3년이 지나고 나서 첫 여름철을 맞아 수상레저 활동이 활발해진 가운데 수상레저기구 관련 사고가 잇따라 증가하는 등 주의가 요구되고 있다.

본 논문에서는 수상레저기구 사고 발생 시 신속하게 구조활동을 할 수 있도록 수상레저기구 위치와 제원정보를 해양경찰청 통합상황관리시스템에 표출할 수 있도록 수상레저기구 위치표출서비스 프로토콜을 제시하였다.

또한, 모바일 앱 기능 중 SOS 정보를 제공한다면 보다 더 신속·정확한 구조활동이 가능할 것이라고 판단된다.

끝으로 수상레저기구 사용자들이 출항 시 의무적으로 운영할 수 있도록 수상레저안전법을 개정하여 추진한다면 안정적인 수상레저활동이 가능할 것으로 판단된다.

Acknowledgement

이 연구는 과학기술정보통신부 및 정보통신산업진흥원의 “지역산업 SW인재양성 기반조성 사업” 지원으로 수행되었음.

참 고 문 헌

- [1] 세계뉴스, 2021년 수상레저 안전과 국민 편의를 위한 새로운 준비, <http://segyenews.com/news/newsview.php?ncode=10656178442> 20040
- [2] 노컷뉴스, 수상레저 성수기, 사고 하루 2건 꼴 발생...’5년간 인명피해 149명’, <https://www.nocutnews.co.kr/news/5953655>

SQL 최적화를 통한 요금계산시스템의 트랜잭션 시간 단축

최윤정^o 남덕윤*

경북대학교 컴퓨터학부; (주)위니텍

dahappy@knu.ac.kr, dynam@knu.ac.kr

Transaction Time Reduction for Billing Systems through SQL Optimization

Yunjeong Choi^o Dukyun Nam*

School of Computer Science and Engineering, Kyungpook National University;

Winitech Co., Ltd.

요약

본 논문은 요금계산의 속도 개선을 목표로 하며, 이를 위해 반복적으로 발생하는 SQL 쿼리 및 트랜잭션을 최적화하였다. 특히, 반복 SQL 쿼리에 대한 인덱스 스캔 방식을 개선하고, 트랜잭션 호출 수를 최적화하기 위해 시뮬레이션을 활용하여 검증하였다. 이러한 최적화를 실무에 적용하여 높은 속도 향상을 달성하였으며, 대용량 요금관리시스템에서도 더 나은 성능 향상을 기대할 수 있는 방법을 제시한다. 더불어, 추가적인 연구와 개발을 통해 더 나은 효율성을 달성하고, 다양한 비즈니스 시나리오에서의 적용 가능성을 탐구할 수 있을 것으로 기대한다.

1. 서론

상수도 요금관리시스템의 핵심 업무인 요금계산은 고객 만족도와 서비스 품질에 직접적으로 영향을 미치며 최근의 행정구역 개편으로 인한 요금이원화 정책, 데이터양의 증가, 지속적인 프로그램 사용으로 인해 요금계산 속도에 대한 대비가 필요하다.

현재 운영 중인 요금계산 프로그램은 최상위 정기분 요금계산 프로시저가 하위(내부) 프로시저들을 단계적으로 호출하여 요금계산을 수행하며, 사용자는 스레드(Thread) 방식을 통해 화면에서 요금계산 진행사항을 5초 간격으로 확인하고 있다.



그림 1. 요금계산 프로그램 개략도

본 논문은 현재 운영 중인 요금계산 프로시저 내에서 반복적으로 처리하는 SQL분석 및 시뮬레이션을 통해 속도 개선 방안을 제안한다.

2. 시뮬레이션 구현 및 검증

본 논문은 반복되는 SQL의 인덱스 스캔방식 개선 및 트랜잭션(Transaction) 호출 수 최적화에 대한 두 가지 방안을 소개하며, 이를 실제 요금계산과 유사한 시뮬레이션 프로시저를 통해 검증하였다.

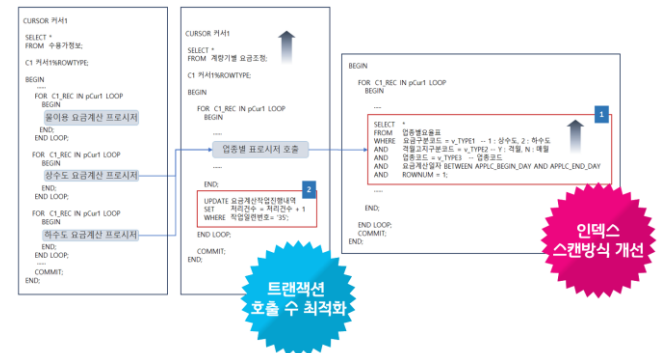


그림 2. 요금계산 프로시저 기본구조 및 개선사항



그림 3. 시뮬레이션 데이터베이스 구성

시물레이션 프로시저를 실행하기 위해서 설치계량기, 업종요금표, 작업진행내역 테이블을 생성하고, 각 테이블에 대한 임시 데이터를 구축하였다.

2.1 업종요금표 인덱스 스캔 방식 개선

업종요금표는 선분이력 형태의 기준정보 테이블로, 적용시작일자와 적용종료일자를 관리한다. 이 테이블은 상수도 및 하수도 업종별 사용료를 계량기별로 산출하는 용도로 사용된다. 기본키(PK: Primary Key)와 인덱스(INDEX)는 [요금구분코드 + 격월고지여부 + 업종코드 + 적용시작일자]로 구성되며, [적용종료일자]는 일반 속성으로 저장됩니다. 업종요금표의 데이터 특성으로 인해, 두 번째 부등호 조건인 적용종료일자는 스캔 범위를 줄이지 않아 별도 인덱스를 추가하지 않고 진행하였다.

표 1. 업종요금표 기본키 및 인덱스 구성

엔티티명 (테이블명)	식별자	기본키(PK)	인덱스(INDEX)
업종요금표 (ZT_INDUTYBY_TARIFF_TICKET)	요금구분코드	CHARGE_SE_CODE	PK_ZT_INDUTYBY_TARIFF_TICKET
	격월고지여부	BMNTH_NTIC_AT	
	업종코드	INDUTY_CODE	
	적용시작일자	APPLC_BEGIN_DAY	

업종요금표 테이블은 총 196건의 데이터를 포함하고 있으며, 기본키를 기준으로 1건만 조회할 때는 속도 차이가 거의 없다. 그러나 계량기별 업종 조회 프로시저를 통해 반복적으로 업종요금표를 조회하면 조회 건수에 따라 속도 차이가 발생한다. 5가지 인덱스 스캔방식을 시물레이션 프로시저로 구현하였으며, 상수도 및 하수도 업종요금표를 조회하는 SQL을 744,804번 반복적으로 실행하였고, 시물레이션은 총 10회를 거쳐서 수행하였다.

표 2. 업종요금표 인덱스 스캔방식 시물레이션 결과

NO	스캔 유형	구분	평균속도 (상수도+하수도)	비고
1	BETWEEN, INDEX_ASC, ROWNUM = 1 (기준)	평균작업시간(sec) 1건당 속도(µs)	27.800615 37.326081	프로시저 테스트 건수 : 10회 상수도업종 건수 : 372,402건 하수도업종 건수 : 372,402건 1µs = 1/1000000(sec) 27.800615 - 23.526313 = 4.274302(sec) (개선효과)
2	BETWEEN, INDEX_ASC, ROWNUM <= 1	평균작업시간(sec) 1건당 속도(µs)	28.715085 38.553882	
3	BETWEEN, INDEX_ASC, ROWNUM = 4	평균작업시간(sec) 1건당 속도(µs)	28.451197 38.199577	
4	BETWEEN, INDEX_DESC, ROWNUM <= 1	평균작업시간(sec) 1건당 속도(µs)	23.823001 31.985598	
5	<=, INDEX_DESC, ROWNUM <= 1 (선정)	평균작업시간(sec) 1건당 속도(µs)	23.526313 31.587254	

그림 4와 같이, 기존 방식은 업종요금표를 오름차순으로 스캔하여 과거시점부터 범위 검색(BETWEEN) 조건을 만족하는 첫 번째 값(ROW = 1)을 조회한다. 이 쿼리를 반복 실행하는 프로시저는 평균적으로 27.8초가 소요되었다.

```

/* 해당 고객번호, 계량기별 업종 요금 검색 */
SELECT *
FROM ZT_INDUTYBY_TARIFF_TICKET TK -- 업종요금표
WHERE CHARGE_SE_CODE = :i_CHARGE_SE_CODE
-- 요금구분코드 (1:상수도, 2:하수도)
AND BMNTH_NTIC_AT = :i_BMNTH_NTIC_AT
-- 격월고지여부 (Y: 격월고지, N: 매월고지)
AND INDUTY_CODE = :i_INDUTY_CODE -- 업종
AND :i_STDR_DAY BETWEEN APPLC_BEGIN_DAY AND APPLC_END_DAY
-- 기준일자(점이력 조회, 적용종료일자)
AND ROWNUM = 1;

```

그림 4. 기존 업종요금표 인덱스 스캔방식

그림 5와 같이, 시물레이션 결과에서는 가장 빠른 스캔 방식으로 기존 방식 대비 4.274초의 개선이 있었다. 1건당 속도 차이는 6마이크로초(us)로 미미하지만, SQL 반복 호출 횟수가 증가함에 따라 속도 차이가 커지는 것으로 확인되었다.

```

/* 해당 고객번호, 계량기별 업종 요금 검색 */
SELECT /*+ INDEX_DESC (TK PK ZT_INDUTYBY_TARIFF_TICKET) */
-- 내림차순 인덱스 힌트(hint) 처리
*
FROM ZT_INDUTYBY_TARIFF_TICKET TK -- 업종요금표
WHERE CHARGE_SE_CODE = :i_CHARGE_SE_CODE
-- 요금구분코드 (1:상수도, 2:하수도)
AND BMNTH_NTIC_AT = :i_BMNTH_NTIC_AT
-- 격월고지여부 (Y: 격월고지, N: 매월고지)
AND INDUTY_CODE = :i_INDUTY_CODE -- 업종
AND APPLC_BEGIN_DAY <= :i_STDR_DAY
-- 기준일자(점이력 조회, 적용시작일자 비교)
AND ROWNUM <= 1;

```

그림 5. 업종요금표 인덱스 스캔방식 개선(안)

그림 5의 쿼리를 실행계획(Explain Plan)으로 분석한 결과, 1) 업종요금표를 내림차순으로 인덱스 힌트(Hint)를 주면 적용시작일자를 최근시점부터 검색하는데 더 효과적이다. 2) 적용시작일자는 기본키 인덱스이며, 적용종료일자는 인덱스가 아닌 관계로 범위검색(BETWEEN) 보다는 적용시작일자를 이용한 부등호(<=) 방식이 더 효율적이다. 3) CURSOR_SHARING 파라미터의 설정 값에 따라 성능 편차가 발생할 수 있으므로, ROWNUM=1 대신 ROWNUM<=1을 사용하는 것이 좋다. ROWNUM<=1은 설정 값과 상관없이 COUNT STOPKEY가 동작하여 1건만 추출 후 수행을 종료한다 [1][2].

2.2 요금작업진행내역 트랜잭션 호출 수 최소화

현재 운영 중인 요금계산 프로그램은 최근 3개월 동안 평균적으로 37분 정도 소요되었다. 사용자는 작업진행사항(작업명, 시작일시, 종료일시, 경과시간, 처리건수)을 5초 간격으로 화면에서 처리건수가 증가되는 것을 확인하며, 이를 통해 요금계산이 중지되지 않고 정상적으로 작동하는 것을 시각적으로 확인할 수 있다. 그러나 1건씩 누적할 때마다 처리건수를 수정(UPDATE)하는

방식은 트랜잭션 호출 수에 따라 데이터베이스 성능 부하 측면에서는 비효율적이다[3].

NO	작업명	작업시작일자	작업종료일자	작업종료시간	처리건수
15	상수도 요금계산	2023/10/10 18:23:51	2023/10/10 18:23:51	00:01:00	248,330
16	하수도 요금계산	2023/10/10 18:23:51	2023/10/10 18:23:51	00:01:00	274,566


```

UPDATE ZT_OPERI_PROGRS_DTLS_2 -- 작업진행내역
SET   PROCESS_CO = PROCESS_CO + 1 -- 처리건수
WHERE OPERI_SN = :v_OPERI_SN; -- 작업일련번호
    
```

그림 6. 비효율적 처리건수 트랜잭션 방식(기존)

처리건수를 단일 건에서 여러 건으로 누적하고 수정(UPDATE)하는 방식을 통해 트랜잭션 호출 수를 줄이면 데이터베이스 성능 부하 및 처리 속도가 크게 개선된다. 전체 처리건수는 372,402건이며, 시뮬레이션은 총 3회를 수행하였다.

표 3. 작업진행내역 시뮬레이션 결과

NO	트랜잭션 발생 기준	트랜잭션 호출 수	평균수행속도 (sec)
1	처리건수를 1건 단위로 발생(기존)	372,402건	24.553
2	처리건수를 10건 단위로 발생	37,242건	3.047
3	처리건수를 100건 단위로 발생	3,726건	0.752
4	처리건수를 1,000건 단위로 발생(선정)	374건	0.518
5	처리건수를 10,000건 단위로 발생	39건	0.418

표 3을 살펴보면, 1번(기존 방식)과 4번(1,000건 단위로 트랜잭션 발생)을 비교했을 때 24초 정도의 속도 개선이 있었다. 3번부터 5번까지의 경우, 평균수행속도가 1초 미만으로 나타나므로 이 범위 내에서 최적화 작업을 진행한다면, 사용자에게는 실시간 처리건수를 제공하는 데에는 아무런 문제가 없을 것으로 보인다.

```

v_PROCESS_CO := v_PROCESS_CO + 1; -- 처리건수

/* 작업 처리건수 수정 */
IF v_PROCESS_CO = 1000 THEN

    UPDATE ZT_OPERI_PROGRS_DTLS_2 -- 작업진행내역
    SET   PROCESS_CO = PROCESS_CO + v_PROCESS_CO, -- 처리건수 = 처리건수 + 1,000
          TRNSC_CO = TRNSC_CO + 1 -- 트랜잭션 수
    WHERE OPERI_SN = v_OPERI_SN; -- 작업일련번호

    COMMIT;

    v_PROCESS_CO := 0;

END IF;
    
```

그림 7. 1,000건 단위로 트랜잭션 방식(예시)

하지만 처리건수를 누적하는 방식을 현재 운영 중인 요금계산 프로시저에 적용하는 데는 일부 문제점이 있다. 현재의 프로세스에서는 최상위인 정기분 요금계산 프로시저가 수용가정보(고객정보와 유사한 의미)를 커서(Cursor)를 통해 조회한 후, 상수도, 하수도 등과 같은 작업 프로시저들을 반복적으로 호출하여 해당 수용가의 계량기별로 요금계산을 수행하는 구조이다. 이에 따라 처리건수를 수정하는 트랜잭션 위치를 단위 작업 프로시저에서 최상위 정기분 요금계산 프로시저로 변경하여, 최상위 정기분 요금계산 프로시저와 단위 작업 프로시

저 사이에서 처리건수를 주고받아 1,000건 이상 처리될 때마다 처리건수를 수정하는 방식으로 구현해 보았다.

3. 실무 적용 결과

논문에서 제안한 반복적으로 발생하는 SQL에 대한 인덱스 스캔 방식 및 트랜잭션 호출 수 최적화를 통해 현재 운영 중인 요금관리시스템에 적용하여 시간 단축 효과를 보였다. 상수도 및 하수도 요금계산에서 이루어진 개선 작업은 직전 3개월 평균 대비 각각 17초와 8초의 속도 향상을 실현하였다. 또한, 물이용부담금 계산에 대해서는 처리건수 누적 횟수에 따라 시간 차이가 발생하지만, 10월에는 18초, 11월에는 11초의 성능 향상을 보였다.

표 4. 요금계산 단위작업별 결과

단위 작업명	구분 요금계산월	적용 전 3개월평균	적용 후	
			2023-10	2023-11
물이용 요금계산	인덱스 스캔 방식 개선	-	-	-
	트랜잭션 호출 수 최적화	-	1,200 건 단위로 호출	1,000 건 단위로 호출
	처리건수	287,940	283,322	290,473
	처리속도(sec)	63	45	52
상수도 요금계산	인덱스 스캔 방식 개선	-	적용	적용
	트랜잭션 호출 수 최적화	-	-	1,000 건 단위로 호출
	처리건수	277,701	274,566	279,461
	처리속도(sec)	87	80	70
하수도 요금계산	인덱스 스캔 방식 개선	-	적용	적용
	트랜잭션 호출 수 최적화	-	-	1,000 건 단위로 호출
	처리건수	244,488	248,300	241,871
	처리속도(sec)	84	80	76

트랜잭션 호출 수 최적화는 요금계산 전체에 확대 적용하여 더 나은 속도 향상을 기대할 수 있으므로, 지속적으로 개선 작업을 계획하고 있다.

4. 향후 기대효과

트랜잭션 호출 수 최적화 테스트에서 시뮬레이션 데이터를 10배(3,724,020건)로 확장하여 추가 실험을 수행한 결과, 4분 이상의 속도 개선 효과가 나타났다. 이로써 본 논문에서는 상수도, 전기, 가스 등과 같은 대용량 요금관리시스템에서 향후 성능 향상을 기대할 수 있는 방법으로 제시한다. 더불어 추가적인 연구와 개발을 통해 더 나은 효율성을 달성하며, 더 많은 비즈니스 시

나리오에서의 적용 가능성을 탐구할 수 있을 것으로 기대한다.

Acknowledgement

이 연구는 과학기술정보통신부 및 정보통신산업진흥원의 “지역산업 SW인재양성 기반조성 사업” 지원으로 수행되었음.

참 고 문 헌

- [1] 조시형, 오라클 성능 고도화 원리와 기법 II, (주)디비안, p56~58, p158~168, 2019.
- [2] 박성호, 오수영, SQL 튜닝의 시작: SQL의 이해와 효율적인 작성법, 엑셈, p347, 2013.
- [3] 조시형, 오라클 성능 고도화 원리와 해법 I, (주)비투엔, p392~403, 2017.

쿠버네티스 기반 분산형 교내 IoT 모니터링 시스템

이현⁰, 김은지, 장서윤, 이창홍, 김종덕*

부산대학교 정보컴퓨터공학부

*교신저자

hy00un@pusan.ac.kr⁰, gee4562@pusan.ac.kr, yrt7998@pusan.ac.kr, dlckdghd96@pusan.ac.kr,
kimjd@pusan.ac.kr*

Kubernetes-Based Distributed Campus IoT Monitoring System

요 약

교육 기관의 경우 다양한 활동으로 실시간 정보 수집과 처리가 필요하다. 이에 본 논문은 쿠버네티스를 이용한 높은 서비스 확장성과, 엣지 컴퓨팅을 이용한 대규모 실시간 IoT 데이터를 처리할 수 있는 엣지 환경의 프레임워크가 갖춰진 교내 모니터링 시스템을 구축한다. 교내 모니터링 시스템에서 관리되는 서비스인 IoT 센서 데이터를 수집하고 이를 사용자에게 보여주는 애플리케이션을 구현한다. 본 시스템은 이와 같은 서비스를 구축하고 실행하는데 필요한 환경을 제공한다. 쿠버네티스의 로드밸런싱 기능을 통해 네트워크 수준에서 트래픽 부하 분산 성능을 비교한다. 쿠버네티스의 HPA 기능을 통해 애플리케이션 수준에서 부하 분산 성능을 비교한다. 성능 평가를 바탕으로 교내 모니터링 시스템에 이를 적용한다.

1. 서 론

교육 기관의 경우, 광범위한 캠퍼스 내에서 다양한 활동이 이루어지고 있다. 강의실, 도서관, 연구실, 학생회관 등 교내 다양한 공간에서 학생들의 학습 활동, 연구 활동, 편의 시설 이용 등이 실시간으로 진행된다. 이러한 활동들을 효과적으로 지원하고 관리하기 위해, 실시간 정보 수집과 처리가 중요하다. 이를 위해 IoT 센서를 통해 다양한 데이터를 수집하고 이를 분석하는 IoT 시스템이 필요하다. [1]

이러한 대규모 실시간 모니터링 시스템을 클라우드 기반으로 구축하게 되면, 비용적인 부담과 데이터 처리 지연 등의 문제가 발생할 수 있다. 이러한 문제를 해결하기 위한 대안 중 하나로, 엣지 컴퓨팅 환경을 도입하는 방안을 고려할 수 있다. 엣지 컴퓨팅은 데이터가 생성되는 위치에 가까운 곳에서 데이터를 처리함으로써, 데이터 전송 지연을 최소화하고 효율성을 높일 수 있다.

엣지 컴퓨팅 환경에서는 각 노드의 자원이 한정적이므로, 이를 효과적으로 관리하고 분산하는 방법이 필요하다. 이에 쿠버네티스를 활용한 분산형 모니터링 시스템을 제안한다. 쿠버네티스는 클러스터 내의 노드 리소스를 효율적으로 관리하고 분산할 수 있어, 엣지 컴퓨팅 환경에서의 리소스 관리 문제를 해결할 수 있다.

따라서 본 논문에서 제안하는 시스템은, 엣지 컴퓨팅 환경을 구축하기 위해 라즈베리파이를 사용하고,

쿠버네티스의 경량화 버전인 k3s를 적용하여 분산형 모니터링 시스템을 구축하였다. 또한, 시스템의 성능을 평가하기 위해 부하 테스트 도구인 K6를 사용하였다. 이를 통해, 제안하는 시스템이 실시간 모니터링에 효과적이며, 대규모 IoT 센서 데이터 처리에 대한 문제를 해결할 수 있음을 보여준다.

2. 본 론

2.1. 관련 연구

2.1.1 엣지 컴퓨팅

데이터가 생성되는 지점 근처에서 처리하여 빠른 응답 처리를 할 수 있는 방법이다. 지연 시간이 발생하는 클라우드 환경이 아닌 엣지 컴퓨팅 환경을 통해 빠른 응답 처리가 가능하다. [2]

2.1.2 쿠버네티스 구성 요소

쿠버네티스는 컨테이너 오케스트레이션 플랫폼으로, 컨테이너화된 애플리케이션을 쉽게 배포하고 확장할 수 있다. Pod는 하나 이상의 컨테이너로 이루어진 기본 배포 단위이다. 노드는 물리/가상 환경을 의미하며 노드에 Pod가 배치된다. 클러스터란 여러 노드의 집합이다. 클러스터 내부에는 제어를 담당하는 마스터 노드, 실제 컨테이너가 배치되는 워커 노드로 이루어진다. [3]

2.1.3 k3s 기능

k3s는 쿠버네티스의 기본적인 기능을 제공하는 경량

화된 배포판이다. 자원이 제한 되어있는 엣지 환경에 적합하다. 쿠버네티스 기능 중 하나인 Service는 여러 Pod에 대한 진입 경로를 제공하며, 애플리케이션에서 Service의 IP와 Port를 통해 Pod에 접근할 수 있도록 한다. LoadBalancer는 외부로부터 들어오는 트래픽을 여러 Pod에 분산시키는 기능을 제공한다. 하나의 외부 IP를 사용해 path 경로에 따라 다양한 서비스를 연결하기 위해 Ingress를 사용한다. [4]

2.2. 시스템 설계

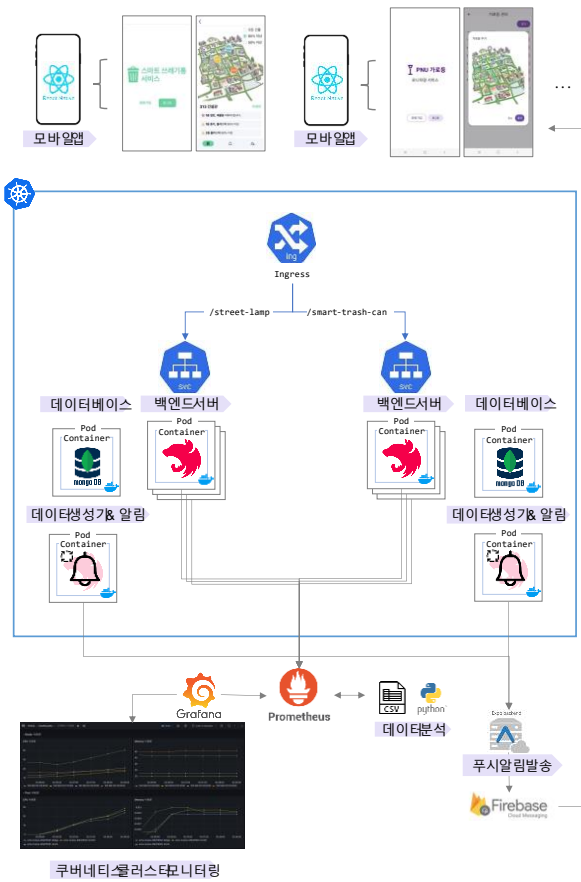


그림 1 교내 모니터링 시스템 구성도

그림1은 제안하는 교내 모니터링 시스템 구성도이다. 시스템에서 제공하는 서비스는 스마트 쓰레기통 서비스와 가로등 모니터링 서비스 2개를 모바일 애플리케이션 형태로 제공한다. 애플리케이션의 프론트엔드, 백엔드 기술로는 각각 React-Native, NestJS를 사용한다. 데이터를 저장하기 위해 MongoDB를 사용한다. 각 서비스는 데이터 생성기를 이용하여 대량의 센서 데이터를 관리하며, 가로등 및 쓰레기통에 대한 정보를 실시간으로 업데이트한다. 또한, 모바일 애플리케이션을 통해 실시간 알림 기능을 제공하여 사용자가 모니터링된 상태를 즉시 확인할 수 있다.

쿠버네티스 클러스터를 구축하기 위해 4개의 라즈베리파이를 사용했으며 마스터 노드 1개, 워커 노드 3개

로 구성된다. Ingress를 사용하여 1개의 외부 IP를 사용해 path 경로에 따라 서비스를 연결한다. 본 시스템의 안정적인 운영을 위해, 프로메테우스를 도입하여 시스템 상태를 지속적으로 모니터링한다. 또한 그라파나를 활용하여 클러스터 내부 자원 상태, 동작 여부를 실시간으로 확인하고 시각화한다. 시스템의 성능 평가를 위해 K6를 사용하여 부하 테스트를 실시한다. 이에 쿠버네티스가 효율적으로 부하를 분산하여, 각 Pod가 균등하게 자원을 사용하는지 확인한다. 해당 결과는 CSV 파일로 변환하여 데이터 분석이 가능하게 한다.

3. 성능 평가

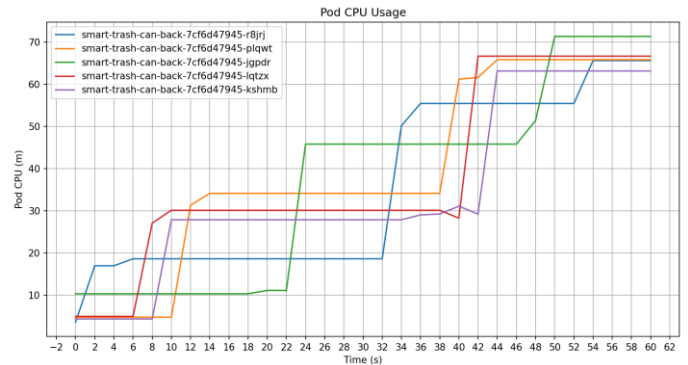


그림 2 로드밸런싱 CPU 변화량

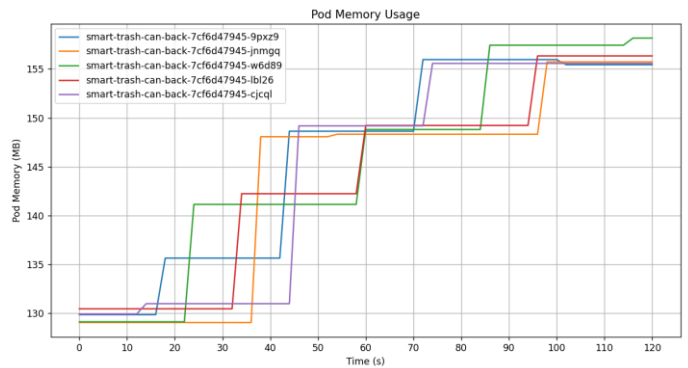


그림 3 로드밸런싱 Memory 변화량

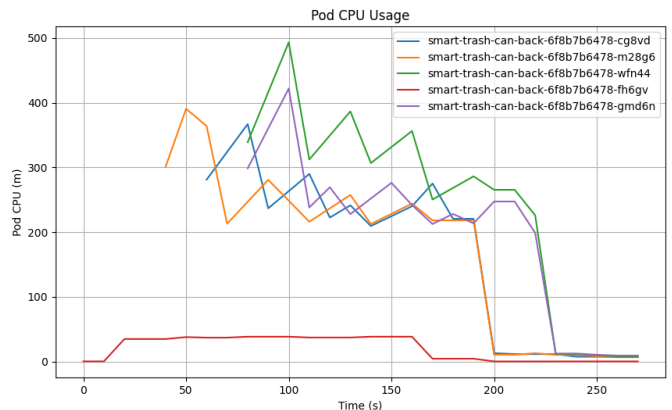


그림 4 오토스케일링 CPU 변화량

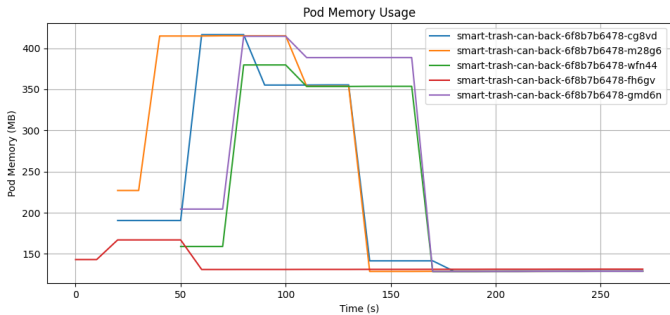


그림 5 오토스케일링 Memory 변화량

그림 2와 그림 3은 Pod 개수를 고정된 상태에서 Pod의 CPU, Memory 사용량 변화를 나타내는 그래프이다. 각 Pod에 CPU, Memory등이 일정하게 상승됨을 확인할 수 있다.

그림 4와 그림 5는 Pod 개수를 고정하지 않은 상태에서 Pod의 CPU, Memory 사용량 변화를 나타내는 그래프이다. 시간이 지남에 따라 Pod 개수가 1개에서 5개로 늘어나며, 부하테스트를 중지하고 일정 시간이 지난 후, 1개로 감소된다.

ACKNOWLEDGEMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음" (IITP-2023-RS-2023-00260098)

4. 결론

교육 기관에서 실시간 정보 수집과 처리가 필요하다는 문제를 인식하고, 이를 위해 IoT 시스템이 필요하다는 것을 서론에서 제시하였다. 그러나 기존의 클라우드 기반 시스템은 비용적인 부담과 데이터 처리 지연 등의 문제를 가지고 있다. 본 논문에서는 이러한 문제를 해결하기 위한 새로운 접근법인, 쿠버네티스를 이용한 높은 서비스 확장성과 대규모 실시간 IoT 데이터를 처리할 수 있는 엣지 환경의 프레임워크가 갖춰진 교내 모니터링 시스템을 구축한다. 성능 평가를 통해 각 Pod가 균등하게 CPU, Memory등의 자원을 사용하며, 이는 교내 모니터링 시스템이 트래픽이 증가되는 상황에서 유연하게 대응할 수 있음을 확인할 수 있다. 이후 연구에서는 쿠버네티스 로드밸런싱 알고리즘, 오토 스케일링의 알고리즘 분석을 실시한다.

5. 참고문헌

[1] 계보경, "데이터 기반 학교 운영을 위한 사물인터넷(IoT) 활용 환경 설계", 교육녹색환경연구, 20권, 3호, pp. 25-32, 2021

[2] 김귀훈, 홍용근, 표철식, "IoT와 SI를 위한 엣지 컴퓨팅 표준화 및 기술 동향", 한국통신학회지(정보화 통신)", 42권, 12호, pp. 49-56, 2017

[3] 이승환, 전성환, 정혜진, 나연목, "서비스의 연속성을 개선하기 위한 고가용성 쿠버네티스 클러스터 아키텍처", 한국정보과학회 학술발표논문집, pp. 1454-1456, 2021

[4] 김철원, 김근용, 김재인, 유학, 김성창, "마이크로서비스 기반 엣지 컴퓨팅 프레임워크 관리 기능 경량화를 위한 k3s 적용", 한국정보과학회 학술발표논문집, pp. 325-327, 2020

경량화 플랫폼 Jetson TX2를 활용한 YOLOv5 및 DeepSORT 기반 버스 내부 모니터링 시스템 구현ⁱ

김민욱^o, 황정호, 황현정, 김운섭, 김종덕*

부산대학교

*교신저자

{lcnwpe, hsk721, guswjd00, giet278, kimjd*}@pusan.ac.kr

Implementation of a Bus Interior Monitoring System based on YOLOv5 and DeepSORT utilizing the Lightweight Platform Jetson TX2

MinWook Kim^o, JeongHo Hwang, HyunJeong Hwang, YunSeob Kim, JongDeok Kim*

Pusan National University

*Corresponding author

요약

실시간으로 대중교통 내부를 모니터링 하는 것은 기존의 CCTV 시스템이 사후 대응에 그쳤던 한계를 극복할 수 있는 방안을 제시한다. 본 연구는 NVIDIA Jetson TX2를 활용하여 승객 수 측정과 안전 사고 감지의 작업을 실시간으로 수행하는 시스템을 제안하며, 특히 YOLO 모델의 크기 별 성능을 분석하여 에지 컴퓨팅 환경에서 최적의 성능을 구현하고자 한다. 이러한 시스템은 즉각적인 사고 대응은 물론 대중 교통의 서비스 품질을 개선하는 데 기여할 것으로 기대된다.

1. 서론

대중교통 시스템 중 버스는 많은 사람들이 일상적으로 의존하는 중요한 이동 수단이다. 그러나 버스 내부의 안전 관리는 많은 도전에 직면해 있으며, 특히 승객의 안전과 효율적인 운영을 담보하기 위한 실시간 모니터링 시스템의 필요성이 점점 더 부각되고 있다. 기존의 CCTV 시스템은 주로 사후 감시와 분석에 사용되어 왔지만, 이는 사고나 이상 상황 발생 시 즉각적인 대응을 할 수 있는 능력이 부족하다.

본 논문에서는 Jetson TX2를 이용한 실시간 버스 모니터링 시스템을 제안한다. NVIDIA Jetson은 고성능, 저전력 컴퓨팅을 제공하는 AI 컴퓨팅 플랫폼이다. 이 플랫폼은 GPU 기반 아키텍처를 사용하여 고급 영상 처리와 머신 러닝을 에지 단에서 처리할 수 있게 한다. 이를 통해 승객 수 측정과 넘어짐 감지를 실시간으로 수행함으로써 즉각적인 상황 인식과 대응이 가능하게 한다.

2. 관련 연구

2.1. YOLO (You Only Look Once)

YOLO는 실시간으로 다중 객체 탐지가 가능한 기술로 사물의 위치 정보 추론과 클래스 분류 과정을 한 번에 처리하는 1-stage-detection 방법을 사용한다. 이로 인해 기존 CNN 기반 추론 방식에 비해 속도 대비 높은

정확도의 검출을 제공하여 실시간 처리가 필요한 환경에서 강점을 가진다.[1]

2.2. DeepSORT (Simple Online and Realtime Tracking)

DeepSORT는 기존의 SORT(Simple Online and Realtime Tracking) 알고리즘을 개선한 딥러닝 기반의 객체 추적 알고리즘이다. SORT는 객체가 겹치는 폐색 현상, 프레임 밖으로 나갔다가 다시 들어오는 재진입, 노이즈 발생으로 인한 외관 변화가 발생할 때 잘못된 ID 스위칭이 발생하는데 DeepSORT는 이를 보완해 동일한 객체로 인식하고 추적 상태를 유지할 수 있도록 한다. 이러한 보완에서 탐지 및 추적되는 객체의 거리와 외형 정보를 결합하여 추적 정확도를 높이는 Matching Cascade 과정을 사용한다.[2]

2.3. MQTT (Message Queuing Telemetry Transport)

MQTT는 사물인터넷(IoT, Internet of Things) 환경에 적합하게 설계된 경량 메시지 프로토콜로 TCP/IP 상에서 브로커-클라이언트 아키텍처를 기반으로 동작한다. 이 프로토콜은 저대역폭, 높은 지연시간, 불안정한 네트워크를 가진 환경에서도 신뢰성 높은 메시지 전송을 가능하게 한다.

2.4. HLS (HTTP Live Streaming)

HLS 기술은 Apple Inc.에 의해 개발된 스트리밍 프로토콜로 실시간 및 온디맨드(On-Demand) 콘텐츠를 인터넷을 통해 전송하는 데 사용된다. HLS 는 비디오 파일을 여러 작은 세그먼트로 분할하여 HTTP 기반 파일로 제공한다. 분할된 파일은 일정 시간 동안의 비디오 데이터를 포함하며, 클라이언트는 이 조각들을 다운로드하여 연속적으로 재생함으로써 비디오 스트리밍이 가능하다.

2.5. crowdhuman_yolov5m 가중치

crowdhuman_yolov5m 는 crowdHuman 데이터셋에 YOLOv5m 을 이용하여 학습한 가중치이다. crowdHuman 데이터셋은 470,000 장에 존재하는 사람의 다양하고 혼잡한 상황에서 촬영된 이미지와 비디오로 구성되어 있고 각 이미지에 약 23 명의 사람이 존재한다. crowdhuman_yolov5m 에 사용된 train, validation 과 test 이미지 수는 표 1 과 같다.

표 1. train, validation, test 이미지 수

	train	val	test
# of images	15000	4370	5000

기존 yolov5m 모델에서는 사람 머리에 대한 가중치가 없는데 비해 crowdhuman_yolov5m 은 사람 머리에 대한 검출을 제공한다. 혼잡한 상황의 버스는 사람들이 겹쳐 있으면 사람 전체를 탐지하는 것이 어렵기 때문에 사람 머리에 대한 검출을 지원하는 모델을 사용하였다.

3. 승객 수 측정 및 안전사고 감지 시스템 구조

본 연구에서 구현된 부분은 두개의 주요 구성요소로 구분될 수 있다. 버스 내부에 설치된 Jetson 플랫폼을 이용한 영상 분석 기능과 분석 결과를 사용자에게 전달하는 통신 기능이다.

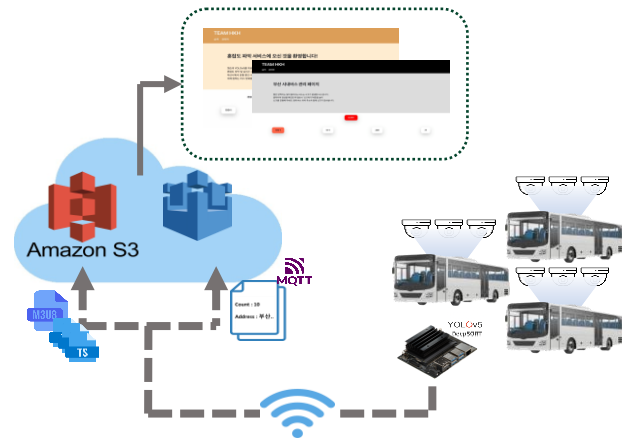


그림 1. 시스템 구성도

제안하는 시스템 구조는 그림 1 과 같다. 시스템의 첫 번째 부분은 Jetson 플랫폼을 이용하여 영상 데이터를 처리하는 데 중점을 둔다. 이 플랫폼은 버스 내부에 설치된 CCTV 로부터 RTSP(Real Time Streaming Protocol)을 이용하여 실시간 영상을 받아, YOLOv5 와 DeepSORT 알고리즘을 구동한다. crowdhuman_yolov5m 가중치를 기반으로 머리 및 사람을 탐지하여 승객의 승차 또는 하차 인원수를 측정하고, 승객의 넘어짐을 식별하는 기능을 수행한다.

시스템의 두 번째 부분은 분석된 정보의 전송에 관한 것이다. Jetson 플랫폼에서 처리된 승차 또는 하차 인원수와 사건 감지 데이터는 MQTT 프로토콜을 이용해 AWS IoT 로 전송된다. 각각의 버스와 영상 종류별로 토픽을 구분하여 승객의 승차 또는 하차 인원수를 발행하도록 설계하였다. 또한, 넘어짐 사건이 감지되면 사건 발생 후 약 30 초 동안 영상을 HLS 프로토콜을 통해 5 초 단위의 여러 개의 파일로 분할하여 AWS S3 버킷으로 업로드 되는 방식으로 설계하였다.

이러한 방법으로 버스 운행 시스템 관리자 측에서 다중 버스에 대해 실시간으로 승객의 승차 또는 하차 인원수, 사고 여부 및 사고 영상을 모니터링할 수 있도록 한다. 모든 데이터는 사용자와 관리자가 접근할 수 있는 웹 인터페이스나 애플리케이션을 통해 시각화 되어 제공될 수 있다.

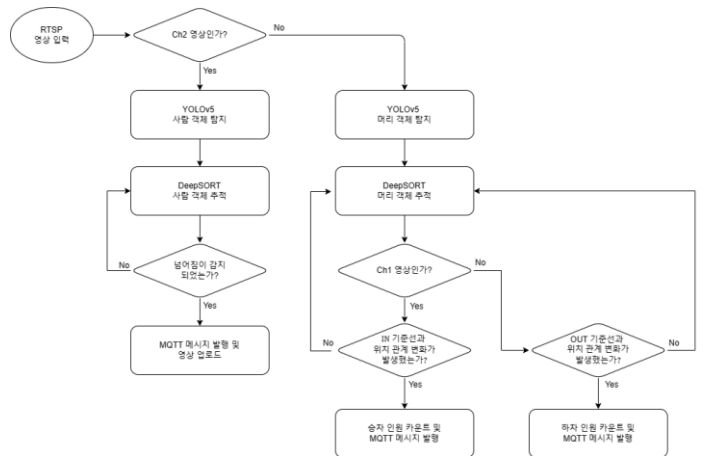


그림 2. 시스템 흐름도

그림 2 는 시스템의 동작 과정을 나타낸다. Jetson TX2 에서 RTSP 로 CCTV 영상을 입력 받고 어떤 구도의 CCTV 영상인지에 따라 분기한다.

Ch2 영상의 경우 사고 감지를 위해 사람 객체를 탐지 및 추적하고 현재와 이전의 바운딩 박스의 높이를 비교한다. 현재 바운딩 박스의 높이가 이전 것의 절반보다 작다면 넘어진 것으로 판단한다. 넘어진 것으로 판단되면 사고 발생으로 인식하여 `{bus#}/{accidentNum}` 토픽으로 MQTT 메시지를

발행한다. 이후, HLS 프로토콜을 이용하여 생성된 영상을 S3에 업로드 하는 과정을 거친다.

Ch1 과 Ch3 영상의 경우 승차 또는 하차 인원 수를 측정하기 위해 머리 객체를 탐지 및 추적하고 기준선과 위치 관계를 파악한다. 기준선과 객체의 위치 관계의 변화가 발생하면 승차 또는 하차 인원을 카운트하고 {bus#}/in 또는 {bus#}/out 토픽으로 MQTT 메시지를 발행한다.

4. 실험 및 평가

본 연구에서는 삼원 FA 에서 제공받은 버스 내부 CCTV 영상 데이터를 사용하였다. 해당 영상은 운행 중인 버스 내부를 여러 구도에서 촬영한 720×480 해상도와 6fps 영상 프레임을 가지는 3 개의 영상으로 구성된다. 입구와 출구 측의 2 개의 영상은 승객의 승차 또는 하차 인원 수를 측정하고 중앙의 1 개의 영상은 사고를 감지한다.

실험은 Jetson TX2를 이용하여 객체 탐지 및 추적을 통한 인원수 측정 성능 평가를 목적으로 하며, YOLOv5의 다양한 크기의 가중치 n, s, m, l, x와 crowdhuman_yolov5m을 사용하여 성능을 비교한다. 각 모델의 성능 평가는 인원수 측정 정확도, 처리 속도를 기준으로 하며, 에지 디바이스에서 실시간으로 영상을 처리하기에 가장 적합한 가중치를 탐색한다.

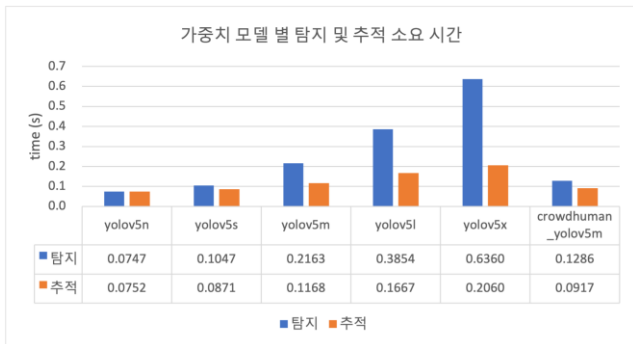


그림 3. 가중치 모델 별 탐지 및 추적 소요 시간

그림 3은 가중치 모델 별 한 프레임을 탐지 및 추적하는데 소요되는 시간을 비교한 결과이다. 일반적으로, 모델의 크기가 클수록 소요 시간이 증가하였다.

crowdhuman_yolov5m의 경우 n과 s버전보다 느리지만 m, l와 x에 비해 빠른 추론 속도를 나타낸다.

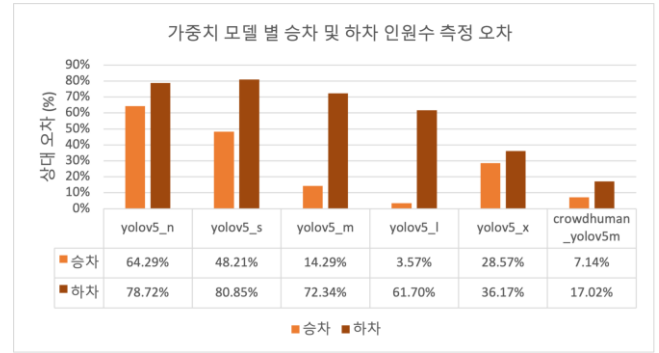


그림 4. 가중치 모델 별 승차 및 하차 인원수 측정 오차

그림 4는 가중치 모델 별 인원 수 측정 오차를 비교한 결과이다. 일반적으로 모델의 크기가 커질수록 오차가 감소하였지만 x의 경우 증가하였다. 특히 하차 영상의 경우 카메라 구도 문제로 인해 많은 인원이 겹치면서 상대적으로 큰 오차가 발생하였다.

crowdhuman_yolov5m모델은 동일 크기의 다른 모델에 비해 승차와 하차 장면 모두에서 비교적 낮은 오차를 보였다. 이는 crowdhuman_yolov5m모델이 머리 객체에 대한 검출을 지원하여 폐색 문제를 효과적으로 극복하였기 때문이다.

5. 결론

본 연구에서는 기존에 제공된 버스 내 CCTV 시스템에서 더 나아가 실시간 영상 분석을 위해 Jetson 플랫폼을 활용한 시스템을 제안하고 구현하였다. 모든 영상을 클라우드 서버로 전송하여 처리하는 방식에서 탈피하여 Jetson 에서 영상 처리를 통해 의미 있는 데이터만 전송하는 방식으로 네트워크와 서버 부하를 줄였다.

YOLOv5 모델이 지원하는 다양한 크기의 가중치 처리 속도와 승차 및 하차 인원수 측정의 정확도에 미치는 영향을 분석하여 에지 디바이스에서 최적의 가중치 모델을 탐색하고자 하였다. 표 2 와 표 3 을 종합하였을 때 n, s, m 과 l 모델을 사용하여 사람 객체에 대한 탐지 및 추적을 통해 승차 및 하차 인원수를 측정하였을 때 의미 있는 정확도로 인원을 측정하지 못하였다. 반면에 crowdhuman_yolov5m 을 사용하여 머리 객체를 탐지 및 추적하여 인원수를 측정하였을 때 추론 속도 대비 정확도가 가장 높았다. 에지 디바이스에 가장 적합한 가중치 모델을 찾기 위해 n 과 s 모델에 머리 객체를 학습시킨 모델을 사용하여 측정된 결과와 추가 비교가 요구된다.

향후 연구에서는 승객의 승차 또는 하차 동선을 분석하여 개선된 기준선을 적용하고 다양한 교통 환경에서의 적용 가능성을 탐구할 예정이다. 또한, HLS 의 Adaptive Bitrate Streaming 기술을 이용하여 네트워크 조건에 따라 동적으로 최적의 Bitrate 로 영상을 전송하는 방안도 고려할 수 있다.

이 시스템은 스마트 교통 솔루션의 범위를 확장하고, 승객 안전을 위한 신속한 대응에 있어서 중요한 역할을 할 것으로 기대된다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터사업의 연구결과로 수행되었음 (IITP-2023-RS-2023-00260098)

참 고 문 헌

- [1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, 2016 IEEE Conference on. (2016)
- [2] Nicolai Wojke, Alex Bewley, Dietrich Paulus , “SIMPLE ONLINE AND REALTIME TRACKING WITH A DEEP ASSOCIATION METRIC”, 2017 IEEE International Conference on. (2017): 3647.

ⁱ <https://github.com/pnucse-capstone/capstone-2023-1-26>

패치 기반 생성형 모델을 적용한 메타버스 이미지 리타겟팅 연구

김건아, 문채원, 김은지, 김찬호^o, 정설영*

경북대학교 컴퓨터학부

kga0416@knu.ac.kr m6011703@gmail.com now0104@knu.ac.kr coolho123@knu.ac.kr

snowflower@knu.ac.kr

Metaverse image retargeting study with a patch-based generative model

Geonah Kim, Chaewon Moon, Eunji Kim, Chanho Kim^o, Seolyoung Jeong*

School of Computer Science and Engineering, Kyungpook National University

요 약

메타버스 산업의 성장과 함께 메타버스 기술의 구현에 대한 중요성이 부각되고 있으나, 화면을 최적화하는 리타겟팅 기술에 대한 연구가 부족하다. 이 연구는 메타버스에서의 화면 리타겟팅 필요성을 언급하며, 패치 기반 이미지 생성 모델인 GPNN을 활용하여 메타버스 데이터의 리타겟팅을 수행하는 프레임워크를 실험적으로 적용했다. 이후 시각적 자연스러움을 평가하는 비교 분석을 실시하였으며, 그 결과 일관성 측면에서는 비교적 우수하나 완전성 측면에서는 개선이 필요함을 확인할 수 있었다. 실험결과를 통해 패치 기반 모델의 한계를 파악할 수 있었고, 향후 메타버스 이미지 기반의 리타겟팅을 수행하기 위해서는 객체에 대한 이해를 가진 모델 연구의 필요성을 파악하였다.

1. 서 론

메타버스 산업의 발전으로 메타버스의 활용 및 적용에 대한 연구가 활발히 이루어지고 있으나, 메타버스를 실현하는 기술 연구는 부족한 상태이다.[1] 특히, 메타버스는 3차원(3D) 공간을 기반으로 하여, 새로운 3D 데이터를 생성하는 데 비용과 시간이 많이 소요되는 한편, 기존의 3D 데이터를 활용하여 원하는 비율의 데이터를 생성할 수 있는 리타겟팅 기술에 대한 연구가 필요하다.

화면 리타겟팅은 메타버스 환경을 위한 중요한 기술 중 하나로, 다양한 화면 크기 및 비율에 맞게 콘텐츠를 최적화하는 것을 말한다. 컴퓨터 및 가상현실(VR) 기기와 같이 다양한 플랫폼에서 메타버스를 활용하는 사용자들에게 품질 높은 경험을 제공하기 위해 화면 리타겟팅 기술은 필수적이다.

본 연구는 메타버스 환경에서의 리타겟팅을 위해 패치 기반 생성형 모델(GPNN)을 실험적으로 활용한다. GPNN은 다른 이미지 생성 모델에 비해 시간이 적게 걸리고 자연 이미지를 생성하는 데 우수한 성능을 보이는 알고리즘이다. 따라서 GPNN을 활용하여 메타버스 데이터의 리타겟팅 수행 능력을 탐구하고자 한다.

이번 연구에서는 MPEG MIV 데이터를 이미지로 변환하여 활용하며, 입력 데이터에 대한 결과로 높이는 유지하고 너비를 2배로 조정하는 방식으로 진행한다. 이를 통해 메타버스 데이터의 리타겟팅을 실험적으로 보여준다.

2. 관련 연구

InGAN[2]은 한 장의 입력 이미지만으로 이미지의 내부 패치 분포를 학습하여 출력 이미지로 매칭하는 패치 기반 이미지 리타겟팅 모델이다. 모델 아키텍처는 여러 데이터 타입을 가진 이미지를 다양한 사이즈와 모양, 중형비를 가진 이미지로 리타겟팅하여 일관성 있고 완전한 결과를 도출한다. 그러나 입력 이미지 외의 추가적인 정보가 없기에 문맥적으로 의미가 맞지 않은 부자연스러운 결과를 도출한다는 한계가 존재한다.

패치 기반 모델인 GPNN[3]은 GAN과 유사한 생성 기능을 최근접 이웃 패치 기반 방법에 주입한 모델로, 빠른 생성 속도와 시각적으로 완성도 높은 이미지를 출력한다는 특징을 가진다. 모델의 구조는 멀티 스케일링 구조를 사용함으로써 이미지의 다양한 특징을 포착하고, 각 스케일마다 최근접 이웃 패치 방법을 통해 입력 이미지와 출력 이미지 간 내부 확률의 유사성을 강화한다. GPNN은 이러한 특성을 유지함과 동시에 파라미터 조절을 통해 리타겟팅, 이미지 합성, 이미지 편집 등의 응용에 적용할 수 있다. 본 연구에서는 GPNN의 리타겟팅 응용에 초점을 두고, 메타버스 콘텐츠에 적용한 결과를 살펴본다.

3. 실험 진행 및 결과

3.1. 실험 진행

GPNN은 패치 기반 생성형 모델들 중 높은 시각적 완성도와 빠른 생성 속도를 보유하고 있다. 이러한

특성을 고려하여, 본 연구에서는 GPNN 에 메타버스 콘텐츠를 기반으로 한 MIV 데이터 셋의 일부 프레임을 활용하여 리타겟팅 실험을 수행하였다(표 1). 입력 프레임의 해상도는 800 x N 으로 설정하였다. 입력 이미지의 높이를 유지하면서 너비를 2 배로 늘리는 리타겟팅을 수행하였다. 그림 1 은 본 연구에서 활용된 패치 기반 생성형 모델을 이용한 메타버스 콘텐츠 리타겟팅 프레임워크를 나타낸다.

실험 모델의 하이퍼 파라미터 값을 설정하기 위해 패치의 크기와 피라미드의 높이, 다운 스케일링 비율을 조절하여 총 4 가지의 경우로 실험을 진행하였다. Street 데이터를 활용한 실험 결과(그림 2), 모델에 입력하는 하이퍼 파라미터의 값에 따라 결과가 변하는 것을 확인하였다.

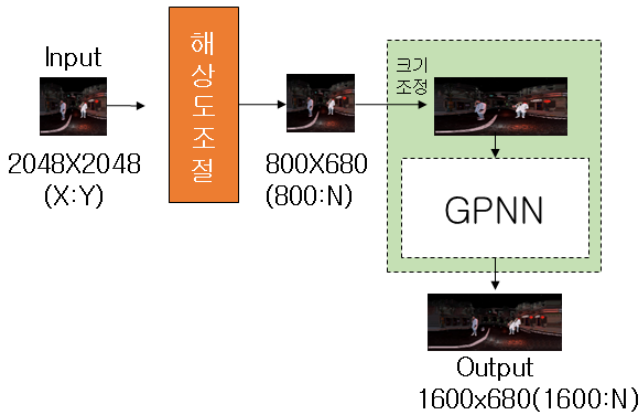


그림 1 메타버스 콘텐츠 리타겟팅 프레임워크

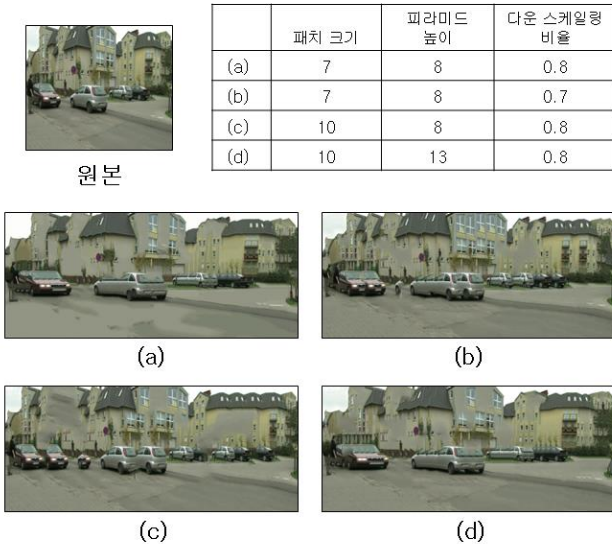


그림 2 하나의 데이터에 다양한 하이퍼 파라미터로 실험한 결과

그림 2 의 (b)와 (c)는 자동차 객체가 연속되지 않고 중복적으로 추가된 것을 볼 수 있으나, (a)와 (d)의 경우의 자동차 객체는 비교적 연속성을 가지고 생성되었다. 추가적으로 (a)의 배경에서 창문의 대칭성이 (d)에 비해 높은 것으로 나타났다. 이에, (a)는 일관성을 유지하며 생성된 것으로 확인할 수 있다. 따라서, 본 연구에서는 모델의 파라미터로, 패치 크기

7, 피라미드 높이 8, 다운 스케일링 비율 0.8 로 설정하여 진행하였다.

표 1 데이터 셋 명세

시퀀스 명	원본 해상도	GPNN 입력 해상도	리타겟팅 해상도
Cyberpunk	2048x2048	800x680	1600x680
Carpark	1920x1088	800x450	1600x450
Street	1920x1088	800x450	1600x450
Dancing	1920x1080	800x450	1600x450

3.2. 실험 결과

원본 이미지와 리타겟팅 이후 이미지를 비교하여 시각적 자연스러움을 평가한 결과, 여러 문제점을 확인할 수 있었다. 그림 3 과 그림 4 의 리타겟팅 결과를 보면 각각의 객체들의 크기와 비율이 비교적 잘 유지되었다. 이를 통해 일관성을 갖추었다고 판단할 수 있다. 그러나 완전성의 측면에서 살펴보면 이미지가 확장되면서 하나의 객체가 여러 개로 분할되어 객체의 형태가 완전하게 보존되지 않았다.

또한, 그림 4(b)의 왼쪽 하단을 보면, 넓은 부분에 직선으로 이어져 있는 화단 부분이 리타겟팅 이후 네 부분으로 끊어진 것을 볼 수 있다. 이는 GPNN 이 이미지를 생성할 때 전체 흐름을 반영하지 않고, 패치 사이즈를 바탕으로 초근접한 이웃 패치만을 학습하여 새로운 이미지를 생성함으로써 발생하는 문제로 추론된다.

결과적으로, 현재 모델은 메타버스 데이터에 대하여 일관성 측면에서는 어느 정도 성능을 보이지만, 완전성과 전체적인 흐름을 반영하는 능력에서 개선이 필요한 것으로 나타났다. 이는 패치 기반 생성형 모델의 구조적인 문제로 비롯되며, 모델을 유지하여 성능을 개선하는 것에는 어려움이 있음을 파악하였다.

따라서, 주변 문맥을 고려하여 더 넓은 영역의 정보를 학습할 수 있는 새로운 모델 아키텍처를 고안할 필요가 있다. 뿐만 아니라, 메타버스 이미지의 특성을 반영한 객체 감지 및 분할과 같은 컴퓨터 비전 기술을 함께 적용한다면 현재 모델의 한계점을 극복할 수 있을 것으로 기대된다.



그림 3 Dancing 리타겟팅



(a) GPNN 입력 이미지



(b) 리타겟팅 결과 이미지

그림 4 Carpark 리타겟팅

4. 결론

본 논문에서는 메타버스 데이터를 리타겟팅 하기 위해 패치 기반 이미지 생성형 모델인 GPNN 을 실험적으로 적용하였다. 그 결과 메타버스 객체들이 끊어지거나 중복적으로 생성되어 완전성이 보장되지 않는 문제가 있었다. 이는 패치 기반 모델이 근접 부분만을 학습하여 이미지를 생성하기 때문에 발생한 특성으로 파악된다. 특히, 사람과 같은 주요 객체가 중복적으로 생성된 경우 시각적으로 더 어색하게 인식될 수 있다.

결론적으로 대부분의 메타버스 콘텐츠는 복잡한 배경과 명확한 주요 객체를 가지고 있어서 기존의 패치 기반 생성 모델을 사용하는 것이 적합하지 않음을 알 수 있다. 따라서 주변 문맥을 고려하는 더 넓은 영역의 정보를 학습할 수 있는 새로운 모델 아키텍처의 개발을 고려하여야 한다. 또한, 메타버스 이미지에서 주요 객체의 중복 생성 및 이미지 완전성 부족 문제를 개선하기 위해서는 객체 감지 및 분할에 특화된 딥러닝 모델이나 컴퓨터 비전 기술을 통합하는 것이 유용할 것으로 판단된다. 이러한 발전된 모델들은 메타버스 데이터의 특성에 더 적합한 이미지 생성 및 리타겟팅을 가능케 할 것으로 기대되므로 이에 따른 후속 연구가 필요하다.

5. 참고 문헌

[1] SHIM, Mi-Na. Analysis of Research Trends on the Domestic Metaverse using Topic Modeling: Focusing on Research from 2007 to 2022, 디지털콘텐츠학회논문지 (J. DCS), 2457-2468. 2022. 23.12.

[2] SHOCHER, Assaf, et al. Ingan: Capturing and retargeting the " dna" of a natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 4492-4501. 2019.

[3] GRANOT, Niv, et al. Drop the gan: In defense of patches nearest neighbors as single image generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 13460-13469. 2022.

[4] SIMAKOV, Denis, et al. Summarizing visual data using bidirectional similarity. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE, p. 1-8. 2008.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음" (2021-0-01082)

웹서비스에서 자바스크립트의 보안 강화를 위한 연구 동향 분석

이인규, 성유성, 김동욱, 강인수, 신민, 김동균

경북대학교

skel137561@gmail.com, sndi123456789@gmail.com, ugi9297@naver.com, kin4496@naver.com,

smin_0820@naver.com, dongkyun@knu.ac.kr

Analysis of Research Trends for Security Enhancement in JavaScript in Web Services

Lee In Gyu, Seong Yu Seong, Kim Dong Wook, Kang In Su, Sin Min, Kim Dong Kyun
KyungPook National University

요약

자바스크립트는 웹 개발에서 가장 많이 사용되는 언어이다. 웹 서비스 및 어플리케이션 사용자의 증가에 따라 자바스크립트의 보안 문제는 중요한 이슈로 떠오르고 있다. 본 논문에서는 자바스크립트 웹 개발 및 서비스의 취약점 탐색 및 해결 방안과 관련하여 기존 연구 방법과 동향에 대해 살펴본다.

1. 서론

ICT 기술의 발전에 따라 웹 서비스는 웹 1.0부터 4.0을 거쳐 5.0까지 웹 기술 및 웹 서비스로 발전하였다. 이러한 웹 서비스의 개발을 위한 대표적인 기술은 HTML, CSS, 자바스크립트(Javascript)가 있다. 이중 자바스크립트는 인터프리터 방식의 스크립트 언어 특성으로, 직관적이고 높은 이식성을 갖고 있으며, 현재 웹 서비스 개발을 위해 가장 많이 사용하는 언어이다. 자바스크립트는 Node.js를 사용하여 서버 단과 클라이언트 단의 어플리케이션 개발에도 많이 사용되고 있다[1].

자바스크립트를 이용한 다양한 웹 서비스 개발의 중요성은 부각되고 있는 반면, 웹 서비스를 많이 사용함으로써 발생할 수 있는 취약점은 관심받지 못하고 있었다. 하지만, 보안 및 개인정보에 대한 중요도 및 민감도가 증가함에 따라, 안전한 웹 어플리케이션 및 서비스를 위한 보안 문제가 대두되고 있다.

본 논문에서는 자바스크립트의 취약점을 탐색 및 보안하기 위한 기존의 제안 기법과 연구 결과를 분석하고, 연구 동향 및 발전 방향에 대해 살펴본다.

2. 본론

먼저, [2]은 자바스크립트를 이용해 개발한 어플리케이션의 오염 및 취약점 탐지를 위한 방법 중 소스 코드를 통해 분석하는 정적 분석(static analysis)기법의 단점을 극복하기 위한 기법을 제안하였다. 제안 기법은 개발 단계에서 개발자와 보안 분석기가 상호작용하여 오염 및

취약점을 탐지할 수 있는 타입스크립트 컴파일러 API를 활용한 정적 분석 방법이다. 논문의 저자들은 오염 및 취약점 탐색을 위해 심볼(symbol)에 신뢰성 정보(Trust, Untrust, Null)를 명시적으로 추가 함으로써 생성되는 심볼의 초기 신뢰성을 확보하였다. 또한, 데이터 흐름의 분석을 위해 심볼의 데이터가 흘러가는 대상을 LHS로, 데이터가 들어오는 심볼을 RHS로 구분하여 데이터 흐름 분석을 통해 이동 경로에 대한 추적을 한다. 이때, 데이터 흐름 규칙 위반 여부와 지점을 탐지하여 취약점이 발생한 지점을 탐지한다.

제안 기법은 명시적 심볼 정보 입력과 데이터 흐름을 파악하여 정적 분석의 단점을 극복하였다. 제안한 새로운 정적 분석 방법을 이용하여 사용자 입력값의 유입 지점을 파악할 수 있어 개발자가 보안 분석에 있어 능동적으로 오염 및 취약 지점을 탐지할 수 있었다고 설명하였다.

[3]은 홈페이지에 접속하는 사용자(victim)의 PC에 은밀하게 악성프로그램 혹은 악성코드를 자동으로 감염시킬 수 있는 악성 URL에 탐지하는 시스템 구현 및 제안하였다. 제안한 시스템은 전체 소스 파일을 입력으로 하여 특정 웹 소스 파일의 자바스크립트 블록(block)을 정적 분석과 동적 분석을 이용하여 코딩된 URL 목록을 모두 추출한다. 정적 분석과 동적 분석으로 찾아낸 URL의 목록을 서로 빼면, 소스 코드에 은닉되어 있는 URL 목록을 찾을 수 있다.

한국인터넷진흥원(KISA) 등 신뢰성 높은 국내외 정보보호 기관으로부터 제공받거나 Zone-H[4]와 같이 사이버 침해사고에 대한 정보를 제공하는 홈페이지의 악성 URL

정보에 대한 DB를 구축한다. 이후, 은닉 URL과 유사도 측정 알고리즘(matching algorithm)을 사용하여 유사도를 점검한다. 0과 1 사이의 측정 결과를 이용하여 해당 URL을 3가지(악성, 의심, 정상)로 판단하여 악성 URL을 찾아내는 시스템을 제안 및 구현하였다.

박선녀 연구원의 연구 결과 논문인 [5]은 자바스크립트 엔진의 크래쉬가 발생하여 생성된 크래쉬 덤프 파일을 이용한 새로운 취약점 개발 도구를 설계 및 구현하였다. 본 논문에서 저자는 버그의 공격 가능성을 측정 및 분류하는 도구인 Exploitable과 AddressSanitizer의 정밀도 및 재현율에 대한 성능이 부족한 문제를 지적하였다. 이를 해결하기 위해 저자는 기계 학습을 이용한 새로운 자바스크립트 엔진의 보안 취약점 발견 도구를 설계 및 구현하였다.

제안한 시스템은 크래쉬 덤프 파일을 이용하여 8가지의 서로 다른 특성을 추출하였으며, 3가지의 벡터화 기법(LableEncoder, CountVectorizer, TfidfVectorizer)을 적용하여 벡터의 크기를 최적화하였다. 추출한 특성을 scikit-learn[6]의 기계 학습 모델을 활용하여 보안 취약점 분류를 위한 최적의 모델을 찾고자 하였다. 이를 위해 제안 시스템에서는 각 모델별 최적의 하이퍼파라미터를 그리드 서치 알고리즘을 이용하여 가장 높은 정확도의 파라미터를 선정하였다. 이후, 선정된 하이퍼파라미터를 학습한 최적의 모델을 기존의 분류 도구인 Exploitable과 AddressSanitizer의 정밀도 및 재현율을 비교 및 분석하였다. 표 1을 보면 실험 비교 결과, 자바스크립트 엔진 중 특히, SpiderMonkey에서 수집한 크래쉬 덤프파일을 이용한 실험 결과에서 약 95%의 정밀도를 보여 기존의 취약점 분류 도구 보다 약 40%의 향상률을 보였다.

타겟 시스템	자바스크립트 엔진	Chakra	V8	SpiderMonkey	오픈 엔진
CRScope	선택된 모델	RFC	RFC	DTC	RFC
	정확도	0.85±0.04	0.89±0.04	0.93±0.04	0.88±0.05
	정밀도	0.84±0.06	0.89±0.06	0.91±0.03	0.88±0.04
	재현율	0.87±0.03	0.90±0.02	0.92±0.07	0.86±0.07
	F1 점수	0.85±0.04	0.89±0.04	0.93±0.04	0.87±0.05
	AUC	0.85±0.04	0.89±0.04	0.93±0.03	0.88±0.05

표 1 CRScope 실험 결과

3. 결 론

본 논문에서는 웹 서비스 및 웹 개발에서 많이 사용되는 자바스크립트의 취약점 및 보안 관련 연구 동향에 대해 살펴보았다. 자바스크립트는 서버단과 클라이언트단 모두에서 많이 사용하는 웹 개발 언어로, 취약점 탐색 방법과 탐지한 취약점에 대해 해결하기 위한 방법에 대해 앞으로 많은 연구가 필요해 보인다. 또한, 웹 서비스 및 어플리케이션 사용자가 꾸준히 증가 함에 따라, 개인 정보와 악성 코드 등 추가적인 자바스크립트 사용으로 인해 발생할 수 있는 취약점 탐색 및 보안 관련 연구가 필요해 보인다.

ACKNOWLEDGEMENT

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음”(2021-0-01082)

참 고 문 헌

[1] J. Strimpel and M. Najim, Building Isomorphic JavaScript Apps, ISBN: 9781491932933, O'Reilly Media, Inc.,pp. 3-13, Sept. 2016

[2] 문태근; 김형식. 정적 오염 분석을 활용한 타입스크립트 코드의 보안 취약점 탐지. 정보보호학회논문지, 2021, 31.2: 263-277.

[3] 박휘량, et al. 웹 소스코드 파일에 숨겨진 악성 Javascript URL 에 대한 점검 체계. 한국컴퓨터정보학회 논문지, 2019, 24.5: 27-33.

[4] Zohn-H, <http://www.zone-h.org>.

[5] 박선녀. 기계 학습 기반의 자바스크립트 엔진 보안 취약점 분류 시스템. 2020.

[6] Buitinck, Lars, et al. “API design for machine learning software: experiences from the scikit-learn project.” arXiv preprint arXiv:1309.0238 (2013).

인공지능 학습을 위한 데이터 증강 및 전처리 과정에 대한 연구

황효성^{0,1}, 김규회¹, 채준혁¹, 정설영^{*,1}
여다영², 장휘영², 김재호³

¹경북대학교 컴퓨터학부, ²경북대학교 영어영문학과, ³퀀텀에이아이

hysee2@knu.ac.kr, K546h@knu.ac.kr, cjh4302@knu.ac.kr, snowflower@knu.ac.kr,
duekdud01@knu.ac.kr, ha5374@knu.ac.kr, quantumai@quantum-ai.ai

Study of Data Augmentation and Processing for Artificial Intelligence Learning

Hyo Seong Hwang^{0,1}, KyuHoi Kim¹, Junhyeok Chae¹, Seolyoung Jeong^{*,1},
DaYeong Yeo², Hwi Yeong Jang², Jaeho Kim³

¹Dept. of Computer Science, Kyungpook-National University

²Dept. of English Language and Literature, Kyungpook-National University

³Quantum AI

요약

제한된 상황에서 적은 데이터셋을 유의미하게 부풀리는 데이터 증강은 중요하다. 이 연구에서는 데이터를 효과적으로 증강하는 방법을 탐구한다. 두 가지 주요한 연구방법은 다음과 같다. 첫 번째 방법은 이미지 회전을 이용한 데이터 증강 방식으로, 다양한 라이브러리를 통해 구현하다. 두 번째 방법은 tensorflow를 활용한 데이터 증강 방식이다. 이 연구의 목표는 데이터 증강 기법을 최적화하여 이미지 인식 및 처리 성능을 향상시키고자 한다.

¹ 경북대학교 컴퓨터학부

² 경북대학교 영어영문학과

³ 퀀텀에이아이

1. 서론

인공지능의 성능 개선을 위한 연구가 끊임없이 진행되고 있다. 인공지능 모델을 훈련시키기 위해서는 풍부하고 다양한 데이터가 필요하다. 이러한 데이터를 확보하기 위해 데이터 증강과 전처리 기술이 널리 사용되고 있다. 데이터 증강(Data Augmentation)은 제한된 데이터셋을 가진 상황에서 인공지능 모델의 성능을 향상하기 위한 핵심 전략이다. 데이터 증강은 새로운 데이터를 수집하지 않고 기존의 데이터를 기하학적 변형을 통해 데이터의 수를 늘리는 방법이다[1]. 데이터 전처리(Data Preprocessing) 역시 인공지능 모델의 학습 과정에서 중요한 단계이다. 원시 데이터를 분석 및 처리에 적합한 형태로 변환하거나 정제하여 인공지능의 학습을 돕는다. 본 논문에서는 이미지 분류 작업을 위한 데이터 증강과 전처리 과정의 중요성을 강조하고, 다양한 방법들을 비교 분석할 예정이다.

2. 관련 연구

2.1 데이터 증강

데이터 증강(Data Augmentation)은 기존의 제한된 데이터셋을 확장하거나 다양화하는 방법 중 하나로, 학습 데이터의 양을 증가시켜 모델의 성능을 향상시키는 데 중요한 역할을 한다. 주로 이미지 데이터에서 사용되며, 원본 이미지에 여러 변형을 주어 새로운 이미지를 생성하는 것을 의미한다. 이미지 증강은 크기, 색상, 밝기, 대비, 회전 등으로 이미지를 수정하여 생성한다[2]. 기존의 데이터셋에 다양성을 추가하여 과적합(Overfitting)을 방지하고, 모델의 일반화 능력을 향상한다.

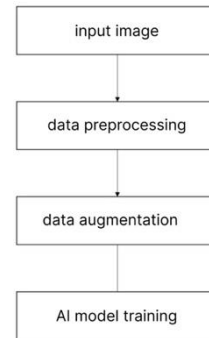
2.2 데이터 전처리

인공지능 모델의 성능 향상을 위해서는 유용한 데이터셋을 통해 학습하는 것이 중요하다. 데이터 전처리(Data Preprocessing)는 원시 데이터를 변환하거나 정제하는 과정을 의미하며 데이터 수집 과정에서 노이즈나 공정 환경 등의 문제로 데이터가 불완전할 수 있는데, 이들은 전처리 과정을 통해 처리함으로써 데이터셋의 품질을 높일 수 있다[3].

3. 시스템 설계

3.1 시스템 흐름도

해당 연구에서 제시하는 시스템의 흐름은 [그림 1] 과 같다.



[그림1] 시스템 흐름도

해당 시스템은 자동차 번호판 이미지를 input image로 사용한다.

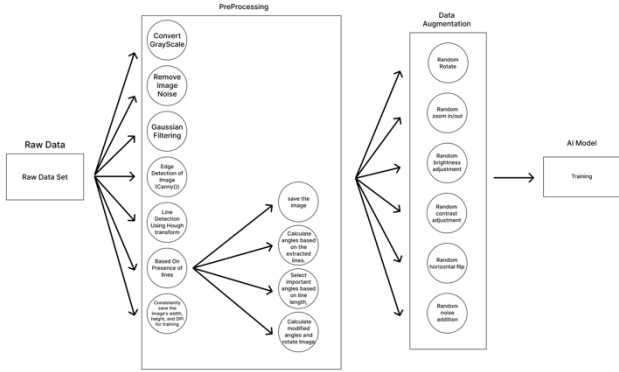
번호판 이미지는 다양한 교통 환경, 차량 유형, 그리고 기상 조건 하에 따라 다양한 각도, 조명 및 거리에서 촬영되었다. 이는 인공지능 학습에 필요한 데이터 셋으로 활용될 예정이다.

일관된 품질을 유지하기 위해 사이즈 조정, 규격 표준화 등을 통해 모델이 학습하는데 최적의 형태로 데이터 전처리를 수행한다. 하지만, 해당 데이터는 제한된 환경에서 촬영되었기 때문에 data set이 매우 적다는 한계를 지닌다. 이 시스템에서는 부족한 데이터셋 문제를 해결하기 위해 이후 전처리 과정을 거친 데이터를 이미지 회전, 색상 변환, 빛 조절 등의 기법을 활용하여 증강

한다.

3.2 시스템 아키텍처

시스템 아키텍처는 [그림2] 와 같다.



[그림2] 시스템 아키텍처

시스템은 번호판 데이터를 Raw Data Set으로 이용하여, 데이터가 전송되면 위 [그림2] 와 같은 구조로 전처리 작업 및 데이터 증강 작업을 수행한다. 데이터 전처리 과정은 7단계에 걸쳐 순차적으로 진행된다.

- 1) 이미지 GrayScale로 변환
- 2) 이미지 노이즈 제거
- 3) 가우시안 블러링 실행
- 4) 이미지의 엣지 검출
- 5) 확률적 허프변환에 의한 선분 검출
- 6) 선의 유무에 따라
 - 6-1) 추출된 선들 기반으로 각도 계산
 - 6-2) 각도 계산에 사용된 선의 길이가 선들의 중앙값보다 작은 각도 제외
 - 6-3) 남은 각도 분포기반으로 수정각도 계산
 - 6-4) 수정각도가 0 또는 90이 아닐 때 수정각도 만큼 이미지 회전
- 7) 학습을 위해 이미지 가로 길이, 세로 길이를 DPI에 일관되게 저장.

이후 데이터 증강을 실시한다. 데이터 증강은 아래 와 같이 실시한다.

- 1) 랜덤 회전, 밝기 및 대비 조절
- 2) 랜덤 노이즈 추가

- 3) 이미지 블러 처리 및 색조변환
- 4) 이미지 이동
- 5) 이미지 흑백 변환

이후 최종적으로 AI 모델에 데이터를 학습시킨다.

4. 시스템 구현

본 연구를 위해 데이터 전처리와 데이터 증강 시스템을 구축했다. 데이터 전처리에서는 이미지를 그레이스케일로 전환하고, fastNIMeansDenoising 과 가우시안 블러로 노이즈를 줄인 후, Canny 엣지 검출과 확률적 허프 변환을 통해 이미지를 보정하고 크기를 300x300으로 통일한다. 데이터 증강은 이미지에 회전, 밝기, 대비 조절과 노이즈 추가, 가우시안 블러 처리, 이동, 색조 변경, 흑백 변환 등을 적용하여 모델의 일반화 능력을 강화하고, 이를 인덱스를 붙여 저장함으로써 학습 데이터셋의 다양성을 확보한다.

4.1 데이터 전처리

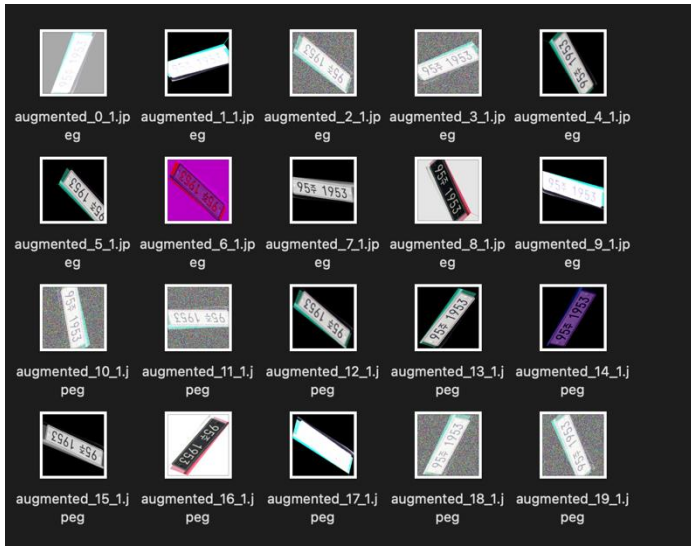
데이터 전처리에서 이미지를 모델 학습에 적합하게 변환한다. 초기 단계로 이미지를 그레이스케일로 바꾸어 색상 정보를 제거하며 픽셀의 밝기 값만을 이용해 계산 복잡도를 줄인다. 다음으로 노이즈 제거를 fastNIMeansDenoising 방법으로 진행하고, 세부 노이즈를 완화하기 위해 가우시안 블러링을 적용한다.

이후, Canny 방법으로 이미지의 엣지를 검출한다. 확률적 허프 변환으로 선분을 찾아, 해당 선분의 각도를 바탕으로 이미지 회전 각도를 조정한다. 각도의 중앙값을 기준으로 필터링한 후, 평균값으로 이미지를 회전한다. 마무리 단계에서는 학습 통일성을 위해 이미지 크기를 300x300으로 조정하여 저장한다.

4.2 데이터 증강 구현

데이터 증강은 기존의 데이터에 다양한 변형을 가해 모델의 일반화 능력을 향상시키는 방법이다. 구현된 기법으로는 회전(Spin)으로 이미지를 -360도에서 360도 사이로 랜덤하게 회전, 밝기 조절(Bright)로 이미지의 밝기를 -300에서 300 사이로 조절, 대비 조절(Contrast)로 0.0에서 5.0 사이의 범위로 이미지 대비를 조절하며, 노이즈

추가(Noise)를 통해 표준 편차 50의 정규 분포 노이즈를 이미지에 추가한다. 블러 처리(Blur)는 5x5 커널 크기로 이미지에 가우지안 블러를 적용, 이미지 이동(Shift)은 이미지를 x, y 축으로 0에서 100 사이로 이동, 색조 변경(ColorShift)은 RGB 채널의 색조를 -200에서 200 사이로 변경, 그리고 흑백 변환(Grayscale)을 통해 이미지를 흑백으로 변환한다. 이러한 데이터 증강 기법들은 주어진 이미지에 순차적으로 적용되고, 총 20번의 증강 과정을 거쳐 각각 저장되며, 저장될 때마다 원본 이미지의 파일 이름에 증강된 이미지의 인덱스가 포함되어 저장된다.



[그림3] 데이터 증강 구현 결과

5. 결 론

이 연구에서는 데이터 증강 및 데이터 전처리 과정을 통해 제한된 데이터 수집의 한계를 극복하고 인공지능 모델 학습에 도움이 되도록 충분한 양의 데이터를 생성하는 기법을 모색하였다.

구체적으로는 수집된 데이터에 빛 밝기조절, 부분 흐림, 회전 등의 노이즈를 추가하여 데이터를 증강하였다. 또한, 이미지 dpi, 사이즈 조정, 사이즈 표준화를 통해 입력 데이터의 일관성을 확보하여 모델이 다양한 해상도와 크기의 이미지를 처리할 수 있도록 도왔다.

연구에서 개발된 모델은 추후 실제 환경에서도 안정적이고 정확한 번호판 인식을 수행할 것으로 예상되고 이를 통해 높은 성능의 자동차 번호판 인식 시스템으로 활용될 수 있을 것으로 기대된

다.

참 고 문 헌

- [1] 김도희, 김명호 <객체 탐지 성능 향상을 위한 생성형 인공지능 기반 데이터 증강 기법 연구.한국컴퓨터정보학회 학술발표논문집>, 31(2), 51-54, 2023
- [2] 이주혁, 김미희 <대형 이미지 데이터셋 구축을 위한 이미지 이진화 기반 데이터 증강 기법>, 27(1), 59-64, 2023
- [3] 서호진, 김도현, 변재현 <제조품질 향상을 위한 데이터 전처리 프로세스>, 49(3), 248-257, 2023

빅데이터 기반 온라인 이미지 편집 솔루션

가파로브 자흔기르¹, 박준서¹, 정지후¹, 이상준¹, 최원서², 정설영¹
경북대학교 컴퓨터학부¹, 모바일앱개발협동조합²
pjun0650@knu.ac.kr

Online image editing solution based on Bigdata

Gaffarov Jakhongir¹, Park Junseo¹, Jeong Jihu¹, Lee Sangjun¹, Choi Wonseo², Jeong Seolyoung¹
Kyungpook National University School of Computer Science and Engineering¹
mobileappdev²

요 약

최근 이미지 콘텐츠 창작 활동이 급격히 늘어남에 따라 비전문가들도 많은 참여를 시도하고 있다. 그러나 현재 사용되는 많은 편집 프로그램들은 설치나 가입이 필요하고 많고 복잡한 기능들로 인해 사용에 어려움이 있다는 한계를 가지고 있다. 본 논문은 이 문제점을 해결하기 위해 개발한 빅데이터 활용 이미지 편집 서비스를 다루고 있다. 이 서비스는 url을 통해 접속하기만 하면 누구나 이용 가능하다. 그리고 빅데이터를 이용한 무료 이미지 검색 기능을 추가하여 콘텐츠 제작자들이 필요한 이미지를 쉽고 간편하게 찾을 수 있도록 해서 양질의 콘텐츠를 만들 수 있도록 돕는다.

1. 서 론

최근 들어 급격하게 미디어 콘텐츠에 관한 관심이 증가하고 관련 산업 역시 발달했다. 유튜브와 인스타그램과 같은 SNS를 중심으로 확산된 미디어 콘텐츠들 중에서 1인 미디어의 존재가 뚜렷하게 나타난다. 1인 미디어란 전문 제작자가 아닌 일반인들이 직접 기획부터 제작까지 모두 하는 콘텐츠를 말하는데 이들은 비전문가인 만큼 전문 도구와 편집 방식에 능숙하지 않다. 또한 시중에 나와있는 이미지 편집 프로그램은 별도의 설치, 가입, 결제가 필요한 경우가 많아 진입 장벽이 있다고 할 수 있다.

이러한 문제를 해결하기 위해 별도의 사전 절차가 필요 없는 웹 기반의 서비스를 개발하고자 한다. url을 통해 접속하면 별도의 절차 없이 필요한 편집 작업을 수행할 수 있는 서비스는 언제 어디서든지 활용이 가능하다는 점에서 기대한 효과를 볼 수 있다.

콘텐츠 제작 효과를 높이기 위해 빅데이터를 이용한 이미지 검색 기능을 제공하고자 한다. 빅데이터는 다양하고 방대한 데이터를 빠르게 처리하고 기존의 데이터 관리 방식이 아닌 새로운 방식으로 처리하여 새롭고 다양한 결과를 얻을 수 있다는 점에서 범 집행, 사회 서비스, 국토 안보, 질병 치료 등 다양한 분야에서 사용할 수 있다[1]. 콘텐츠 제작에 있어 다양한 이미지 소스 중 원하는 이미지를 찾는 시간을 줄이고 목적에 맞는 이미지를 빠르고 쉽게 찾을 수 있게 한다면 더욱 좋은 효과를 볼 수 있을 것이다.

2. 빅데이터 이미지 검색

빅데이터 기술은 미디어 콘텐츠 분야 전반에서 이미 많이 사용되고 있다. 글로벌 OTT 서비스 회사인 Netflix는 영상 콘텐츠 제작과 유통 전반에 걸쳐 사용자의 선호와 트렌드 분석에 빅데이터 분석을 이용하고 있고 유의미한 성과를 거두고 있다[2]. 본 서비스에서도 빅데이터를 이용한 이미지 검색으로 사용자에게 편의를 제공하고자 한다[2].

빅데이터를 활용한 이미지 검색 기능을 제공하기 위해 무료 이미지 검색 사이트인 Pixabay에서 제공하는 Pixabay API를 이용할 계획이다. 이 API는 검색어를 입력받아 검색어와 관련 있는 무료 이미지들을 데이터베이스에서 찾아 사용자에게 제공한다.

Pixabay는 각 이미지들을 해당하는 태그와 함께 데이터베이스에서 저장, 관리하는 방식으로 빅데이터를 운용한다. 이러한 빅데이터는 많은 양의 사진들을 효과적으로 관리하고 필요한 자료들을 적절하게 제공할 수 있다는 점에서 매우 효과적이다.

사용자는 원하는 사진에 맞는 키워드를 입력하여 얻은 이미지를 저장이나 구입 없이 편집에 즉각적으로 활용할 수 있다는 점에서 이점을 볼 수 있다.

사용자는 입력하는 검색어를 세부화하면서 필요한 이미지를 찾아나갈 수도 있다. 예시와 함께 설명하겠다.

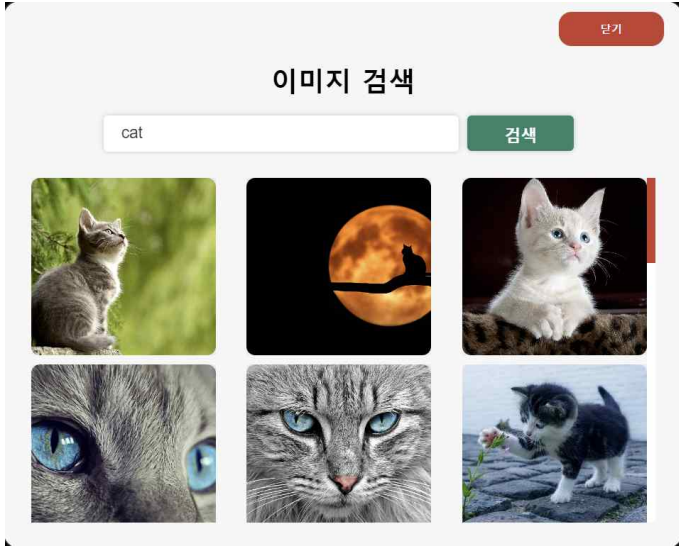


그림 1 키워드 검색

그림 1은 고양이 사진을 찾기 위해 'cat' 을 검색어로 입력한 화면이다. 보는 바와 같이 다양한 고양이의 이미지들이 결과로 나타난다. 그러나 사용자가 이 고양이 사진들 중 검은색 고양이만을 찾고 싶을 수도 있다.

그럴 때에는 아래 그림 2와 같이 'black+cat' 이라는 검색어를 입력하면 이에 맞춰 검은색 고양이 이미지만이 검색

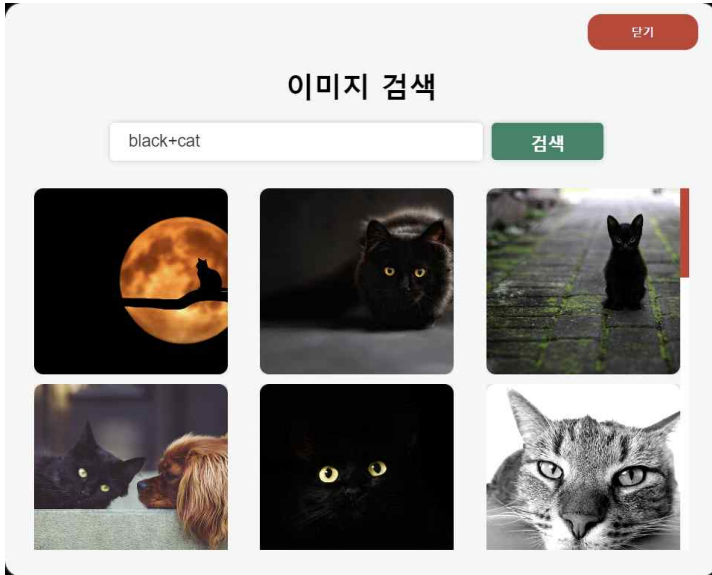


그림 2 키워드 상세 검색

이러한 예시와 같이 검색 키워드를 구체화하면서 원하는 이미지를 찾을 수 있는 것이 빅데이터 기반 이미지 검색의 장점이라고 볼 수 있다. 다량의 이미지를 찾기 위해 목록을 뒤지거나 할 필요 없이 검색어를 잘 설정하기만 하면 원하는 결과를 얻을 수 있기에 사용자의 편의에 큰 도움을 줄 수 있으리라 생각한다.

3. 시스템 설계

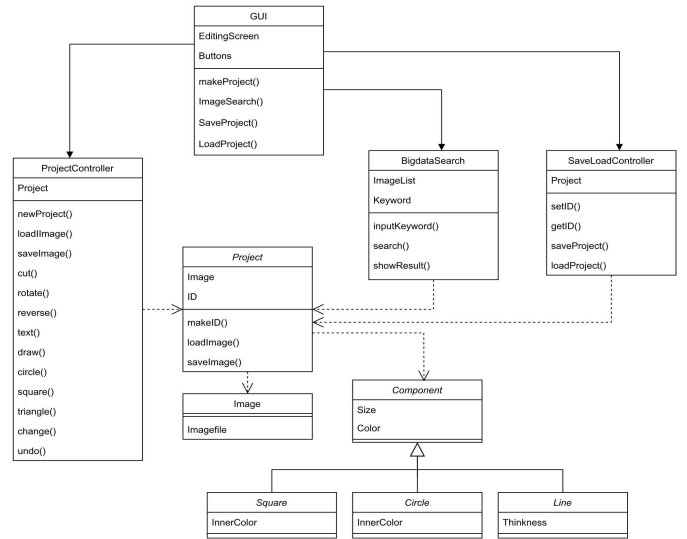


그림 3 클래스 다이어그램

그림 3은 시스템을 개발하기 앞서 시스템의 구조를 간략화한 다이어그램이다. GUI로 나타나는 버튼들과 편집창을 통해 사용자의 명령을 입력받으면 각 Controller 클래스들이 해당 명령을 수행한다. 기본적으로 사용자가 이미지 파일을 불러오면 해당 이미지는 Image 클래스로 저장되고 이 이미지 클래스를 포함한 Project 클래스에 각종 편집의 명령이 적용된다.

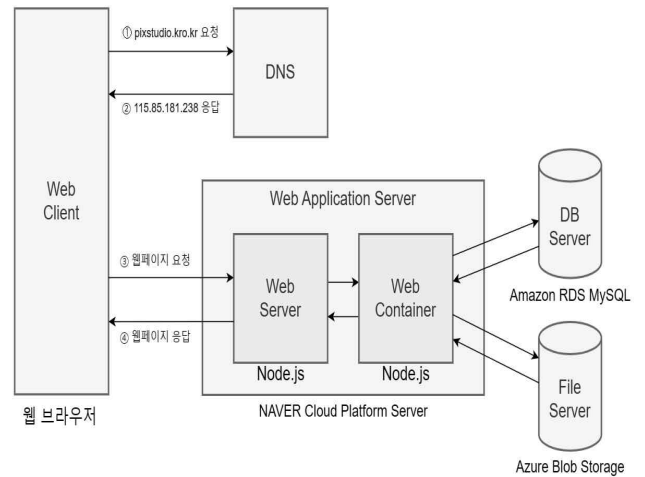


그림 4 백엔드 다이어그램

그림 4는 시스템의 데이터베이스 구조와 서버 구조를 담은 백엔드 다이어그램이다. Node.js 기반으로 구축된 웹 서버는 편집 웹페이지를 사용자에게 제공하고 마찬가지로 Node.js로 구축된 웹 컨테이너는 데이터베이스와 연동하여 프로젝트 정보를 주고받는데에 이용된다. 데이터베이스는 파일 서버와 DB서버를 구분하여 운용하고 있다. DB서버에는 ID, 파일서버에서의 저장경로, 저장시간이 저장되고 파일 서버에는 나머지 프로젝트 데이터들이 저

장된다.

3-1. 이미지 불러오기 및 저장

사용자는 디바이스에 저장된 이미지 파일을 불러와 편집창에서 편집 작업을 할 수 있다. 불러올 수 있는 이미지 파일의 형식은 jpg, jpeg, png로 제한한다. 편집이 완료된 이미지는 png 형식으로 디바이스에 저장된다.

3-2. 이미지 편집

사용자에게 제공되는 편집 기능으로는 이미지 자르기, 이미지 회전, 이미지 속성 변경이 있다. 이미지 속성 변경이란 필터를 적용하고 이미지의 밝기, 노이즈 정도, 색상을 조절하는 기능이다. 이 기능들을 선택할 수 있는 배너가 편집창 옆에 위치한다. 편집 메뉴를 선택하면 해당 기능을 이용 가능한 메뉴가 나타난다.

3-3. 컴포넌트 형식

이미지 편집 중에 추가된 도형, 텍스트 상자, 추가 이미지는 모두 컴포넌트로 사용된다. 컴포넌트는 이미지 파일과 별개로 위에 덧붙여진 대상으로 마우스를 이용해 이동하고 크기를 조절할 수 있다. 그리고 컴포넌트를 클릭하고 삭제 버튼을 누르면 해당 컴포넌트는 삭제된다.

3-3. 도형 추가

사용자는 이미지 위에 도형을 추가할 수 있다. 추가 가능한 도형으로는 원형, 사각형, 삼각형이 있고 추가한 도형 내부의 색과 테두리 색, 테두리 굵기를 조절할 수 있다. 추가된 도형은 컴포넌트 형식에 따라 조작 가능하다.

3-4. 텍스트 추가

사용자는 텍스트 상자를 추가할 수 있다. 텍스트 상자를 더블 클릭해 텍스트 상자 내의 텍스트 내용을 편집할 수 있다. 텍스트 상자를 선택하고 텍스트 메뉴의 버튼을 이용하여 텍스트의 스타일, 정렬 방향, 색상, 글씨 크기 역시 조절이 가능하다. 추가된 텍스트 상자는 컴포넌트 형식에 따라 조작 가능하다.

3-6. 편집 과정 임시 저장

사용자는 이미지를 편집하다 도중에 과정을 임시로 저장할 수 있다. 서비스 자체는 별도의 히스토리를 남기지 않고 웹페이지를 벗어나면 다시 복구가 되지 않는 시스템이기 때문에 이러한 기능을 추가했다. 사용자가 편집 임시 저장을 선택하면 편집창 위의 이미지와 추가된 컴

포넌트들이 데이터베이스에 임시로 저장되고 사용자에게 6자리 숫자로 이루어진 임의의 ID를 제공한다. 이후 다시 편집 과정을 불러와 이어서 편집을 하고 싶으면 사용하는 편집 과정 불러오기를 선택하고 ID를 입력하여 불러온 편집 과정을 이어 작업할 수 있다. 저장된 편집 과정은 24시간이 지나면 삭제된다.

4. 시스템 구현 결과

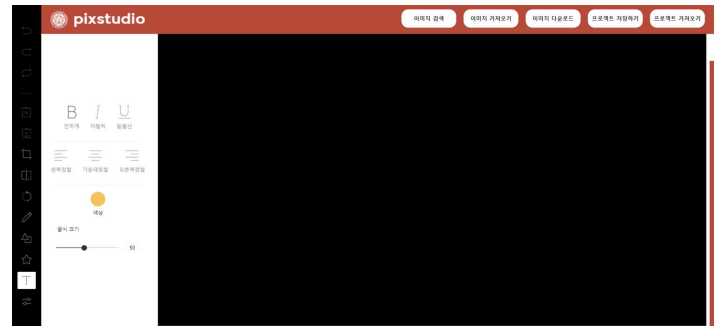


그림 5 초기 화면

시스템을 구현한 결과는 그림 5와 같은 웹페이지로 나타난다. 상단에는 이미지 검색, 이미지 가져오기, 이미지 다운로드, 프로젝트 저장하기, 프로젝트 가져오기 버튼이 있고 아래의 검은 화면은 불러온 이미지를 편집할 수 있는 편집창이다. 좌측에는 어떤 편집을 실행할지 선택하는 버튼들과 해당 편집에 사용되는 각종 버튼, 스크롤바 등이 위치하는 섹션이 있다.

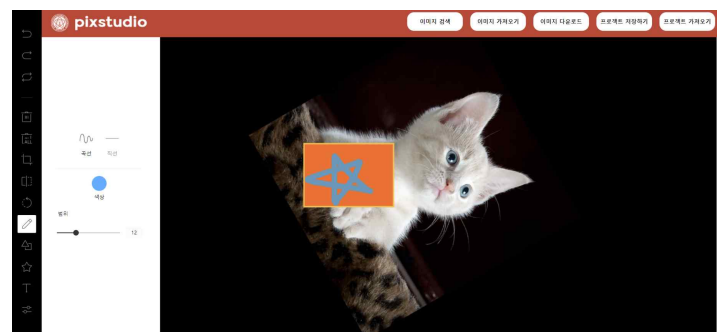


그림 6 편집 실행 화면

그림 6은 실제로 사진을 업로드하고 몇 가지 편집을 실행한 결과의 화면이다. 보이는 바와 같이 업로드된 사진을 회전하고 도형을 추가하고 선을 그릴 수 있는 것을 볼 수 있다.

5. 결론

본 논문에서는 빅데이터를 활용한 이미지 검색 기능을 가진 온라인 이미지 편집 솔루션 개발에 관한 내용을 기

술하였다. 온라인 기반의 서비스를 채택하여 사용자 편의를 도모하였고 무엇보다 빅데이터 기술을 이용해 사용자가 쉽고 정확하게 원하는 이미지를 찾을 수 있게 했다는 점에서 이 서비스가 많은 콘텐츠 제작자들에게 효과를 보여줄 수 있을 것이라 생각한다.

참 고 문 헌

[1] K. Michael and K. W. Miller, "Big Data: New Opportunities and New Challenges [Guest editors' introduction]," in *Computer*, vol. 46, no. 6, pp. 22-24, June 2013, doi: 10.1109/MC.2013.196.

[2] 김현영, 김재웅. (2019). 문화예술 콘텐츠 제작 및 유통에서의 빅데이터 활용 연구. *한국콘텐츠학회논문지*, 19(7), 384-392.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

딥러닝을 활용한 맞춤형 신발 소재별 관리 시스템

저자1 윤해진¹ 저자2 배상우², 김영진², 이창훈²
주식회사 스마트롬뱅이, 경북대학교 전자전기공학부
저자1 ivnavi@naver.com, 저자2 sbangrnd@gmail.com

Deep Learning-Based Customized Shoe Material Care System

저자1 Haejin Youn¹ 저자2 Sangwoo Bae², Youngjin Kim², Changhoon Lee²
SmartreumbangE Inc. and School of Electronic and Electrical Engineering, Kyungpook National University¹, SmartreumbangE Inc.²

요약

본 연구는 신발의 다양한 소재를 효율적으로 관리하기 위해 딥러닝을 활용하여 맞춤 관리를 제공하는 목적으로 수행되었다. 기존의 신발 관리 시스템은 신발 소재를 고려하지 않고 일괄적인 방식으로 관리되어 신발 손상을 초래하거나 효과적인 관리 효과를 얻기 어렵다. 본 연구에서는 신발 이미지를 활용하여 딥러닝 모델을 학습시키고, 이를 통해 신발 소재별로 맞춤 관리를 구현하고자 한다. 또한, 사용자가 적절한 사용 시간을 파악하기 어렵다는 문제를 해결하기 위해 수동으로 살균 시간을 선택하는 기존의 방식을 개선하고자 한다. 연구 결과를 통해 신발의 수명을 연장하고 효과적인 관리를 실현하는데 기여하고자 한다.

1. 서론

발은 가장 많은 땀샘이 있으며, 하루 250~500ml의 땀을 배출한다. 이로 인해, 신발은 늘 습한 상태이며 이 습한 환경은 세균을 증식시키고 발냄새, 발질환을 유발한다.[1] 이를 예방하기 위해서는 신발 관리가 중요하다. 전통적인 신발 관리 방법은 신발 세탁이다. 하지만 현재 신발 소재가 다양해지고 물세탁이 어려운 소재를 많이 사용하면서 대부분 신발 제조사에서는 신발 물세탁을 금하고 있다. 이러한 대안으로 최근 신발 내부를 건조, 살균해주어 위생적으로 관리해주는 신발 관리기가 많이 출시되고 있다. 본 논문에서는 신발 관리의 중요성과 기존 신발 관리기의 한계를 소개하고, 딥러닝을 활용한 신발 관리기의 제안 방법을 소개하여 신발 관리의 효율성을 개선하는 연구를 진행하고자 한다.

2. 문제 분석

신발은 다양한 소재로 제작되어 있기 때문에 각 소재의 특성에 맞춰 적절한 관리가 필요하다.[2] 기존의 신발 관리기는 신발 소재를 고려하지 않고 일괄적인 방식으로 관리되어 신발 손상을 초래하거나 효과적인 관리 효과를 얻기 어려웠다. 또한, 일괄적인 방식이 아닌 수동으로 관리 시간을 선택하는 방식 역시 사용자가 적절한 사용 시간을 파악하기 어려워 제대로 된 신발 관리가 어렵다.

3. 제안 기법

본 연구에서는 딥러닝을 활용하여 신발 소재별로 맞춤형 관리를 제공하는 신발 관리 시스템을 개발하고자 하

다. 신발 이미지를 활용하여 딥러닝 모델을 학습시키고, 이를 통해 신발 소재에 따라 최적의 관리 방법을 제공한다. 예를 들어, 가죽 소재의 신발은 특정한 방법으로 관리해야 하고, 메쉬 소재의 신발은 다른 관리 방법이 필요하다. 이러한 신발 소재별 맞춤 관리를 통해 신발의 수명을 연장하고 효과적인 관리를 실현할 수 있다.

또한, 사용자가 적절한 사용 시간을 파악하기 어렵다는 문제를 해결하기 위해 수동으로 살균 시간을 선택하는 기존의 방식을 개선하고자 한다. 딥러닝 모델을 활용하여 신발의 소재를 분석하고, 적절한 관리 시간을 자동으로 추천하는 기능을 구현한다. 이를 통해 사용자는 위생적이고 효율적인 신발 관리가 가능하다.

3.1 신발 소재별 신발 관리 방법 연구

1) 신발 소재별 종류 및 특징

신발 사진	종류	특징
	운동화 (통화)	- UPPER부와 밑창(솔)부의 경계가 뚜렷함 - 끈으로 묶는 방식이 많은편임
	러닝화	- 발목부가 낮음 - UPPER부가 메쉬로 되어있는 경우가 많음. - 끈이 없이 탄성으로 되어 있기도 함 - 비교적 가벼움.
	농구화	- 발목부가 높음 - 질은색이 비교적 많은편임 - UPPER부는 메쉬와 면이 조합되어 있음

표1. 신발 소재별 종류 및 특징

신발의 종류와 특징을 분석하여 관리 특성인 건조, 살균의 기준을 마련하였다. 본 연구에서는 신발의 종류와 특징을 표 1 과 같이 운동화, 러닝화, 농구화 3가지로 설정하고 향후 확대하고자 한다.

2) 신발 관리 방법 정립

① 건조 방법 정립

건조를 위해 신발에 포함된 물의 양을 표현해주는 함수율과 신발 건조를 위한 히터의 온도, 바람을 불어넣어 주는 팬의 풍량 및 풍속, 그리고 신발을 놓는 위치 등이 건조의 성능에 영향을 주는 주요 인자라 할 수 있다. 본 기술은 신발 구동장치를 신발 내부에 삽입한 후, 약 50℃의 열풍이 신발 내부를 순환하며, 신발 내부를 건조시킨다.

신발 침수 전, 후 무게를 비교한 수식 (1) 과 같은 시험 방법에 의한 건조율을 90% 이상을 목표로 한다.

$$\text{건조율(\%)} = \frac{(\text{침수 후 무게} - \text{가동 후 무게}) / (\text{침수 후 무게} - \text{최초 무게}) \times 100 (\%)}{\quad} \quad (1)$$

② 살균 방법 정립

일반적으로 신발 살균이라고 함은 UV 살균, 습가열에 의한 열풍대류방식 살균이 있다. 신발은 늘 습한 상태이며 UV의 경우 습한 환경에서는 살균 성능이 떨어진다.

본 연구에서는 신발 관리에 좀 더 적합한 열풍대류방식을 적용하여 신발을 1차 살균하고 2차로 UV로 한번 더 살균해준다. 열풍대류방식의 경우 40℃ 이상에서 살균이 일어난다. 하지만 60℃ 이하로 제한할 필요가 있는데 이유는 신발의 접착제들 대부분이 60℃ 이상에서는 화학반응을 일으켜 신발의 변형, 이염 등이 발생하기 때문이다.

3) 신발 관리 시험

① 건조시험

- 시험방법은 종류별 신발을 침수 후 건조성능 측정(각 5회)
- 침수량은 20~40g 으로 땀으로 인한 습한상태부터 외부 수분공급에 따른 침수상태까지 측정.



그림 1. 건조 시험 결과

- 5회 실험 결과의 평균값은 그림 1과 같이 러닝화 111.7%, 운동화 103%, 농구화 84.4% 건조율을 보였다.

② 살균시험

- 시험 방법은 시험참가자 3명이 각각의 종류별 신발을 일주일씩 착용하고 신발관리기 사용 전후의 ATP 측정을 통해서 오염도 개선정도를 비교 한다.
- ATP측정기의 세균에 대한 측정 편차가 크므로 1개의 요소에 3회씩 반복 측정하여 분도 상황을 고려하여 평가 한다.
- 결과는 그림 2 와 같이 러닝화 53%, 운동화 45%, 농구화 30% 줄어듦을 알 수 있었다.

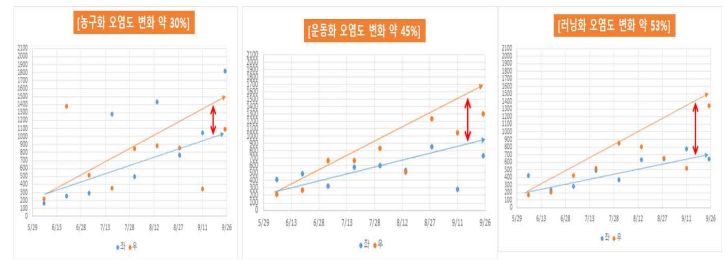


그림 2. 살균 시험 결과

③ 소재별 신발 관리 프로그램 도출

신발 종류별 관리시간(건조율90%, 살균 99%)

모드	기준습도 (%)	가동시간 (분)	부품별 작동시간			배터리 사용률
			FAN	HEATER	UV	
A U T O M O D E	60이하	30	FAN	10분	10분	25%
			HEATER	10분	10분	
			UV	10분	10분	
운 동 화	60이하	40	FAN	15분	15분	40%
			HEATER	15분	15분	
			UV	10분	10분	
농 구 화	60이하	60	FAN	30분	20분	55%
			HEATER	20분	20분	
			UV	10분	10분	

그림 3. 소재별 신발 관리 프로세스

시험 결과에 따라 러닝화, 운동화, 농구화의 신발 관리 시간은 (건조90% 이상, 살균 99%) 각각 30, 40, 60분으로 관리 프로그램 도출하였으며 프로세서는 그림 3 과 같다.

3.2 딥러닝을 활용한 신발 소재별 분류 프로그램 구성

1) 데이터 수집

연구를 위해 신발 소재에 대한 다양한 데이터를 수집해야하며, 이 데이터는 신발의 소재에 대한 정보와 해당 신발의 이미지로 구성될 수 있다. 데이터베이스나 인터넷에서 신발 소재와 관련된 정보를 수집하고, 직접 이미지를 촬영한다.[2]

2) ‘teachablemachine ‘ 활용한 학습

신발 소재별 이미지를 학습시켜 소재 분류를 테스트 하여 분류를 확인함.

3) App 개발

학습된 모델을 사용할 수 있도록 App 제작.

4. 향후 계획 및 결론

본 논문에서는 딥러닝을 활용한 맞춤형 신발 관리 시스템을 제안하였다. 신발 소재별로 최적의 관리 방법을 제공하여 신발의 수명을 연장하고 효과적인 관리를 실현할 수 있다. 향후 계획으로는 학습된 머신을 사용할 수 있도록 App를 제작할 계획이다. 또한 신발 소재 종류를 확대시키고 나아가 신발 사용 시간과 사람의 체질에 따라 적정한 사용 시간을 추천하는 기능을 추가할 계획이다. 이는 사용자 개인별 신발 관리를 통해 신발을 더욱 청결하고 올바르게 관리할 수 있도록 도움을 줄 수 있다.

Acknowledgement

이 연구는 창업성장기술개발사업 “모빌리티 탑승자의 헬스케어 및 편의 증진을 위한 발/신발관리 옵션장치 개발” 지원(S3283873)과 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성사업의 연구결과로 수행되었음(IITP-2023-RS-2022-00156389)

참 고 문 헌

- [1] 박수진, 신준식 and J.Kawasaki. (2003). 오존 표면 처리된 활성탄소의 암모니아 제거. 공업화학, 14(8), 1133-1137.
- [2] 이삼철 and 장용석. (2016). 오염된 신발의 효율적인 살균과 탈취를 위한 신발장의 설계. 한국산학기술학회논문지, 17(6), 199-206.
- [3] 이은민 and 김태훈. (2023). 구글 티처블머신을 활용한 정신과적 응급 대상자의병실 안전 모니터링 프로그램 개발. 문화기술의 융합, 9(1), 613-618.

차량 내부 공기 질 실시간 분석에 따른 국부공기청정정화 시스템 효율 증대

배상우¹, 윤해진², 김영진², 이창훈²
¹경북대학교 일반대학원 전자공학과, ²㈜스마트룸뱅크
¹sbangrnd@gmail.com, ²sbangsales1@naver.com

Increasing the efficiency of the local air cleaning and purification system according to the real-time analysis of air quality inside the vehicle

SangWoo Bae¹, HJ Youn², Y J Kim², C H Lee²
¹KYUNGPOOK NATIONAL UNIVERSITY
SCHOOL OF ELECTRONIC AND ELECTRICAL ENGINEERING, ²SmartreumbangE Inc.

요 약

논문은 교통수단 내부에서 전체 공기청정정화 시스템 대신 탑승객 별 개별 시스템을 적용하여 호흡기질환 예방 및 졸음운전 환경을 개선할 수 있는 시스템을 고안한 것이다. 개별적 적용을 통해 차량 전체가 아닌 탑승객 호흡기 주변을 국부적으로 집중관리 함으로써 기존의 미세먼지 제거 수준의 공기청정정화 시스템을 부유세균, 바이러스 제거 수준으로 업그레이드하고, 탑승객간의 오염전파를 방지한다. 나아가 실험과정에서 실시간 검출되는 오염원 데이터와 시스템 가동에 따른 개선된 데이터를 학습시켜 실시간 분석에 따른 교통수단 내부 효율적인 위생환경 및 졸음환경 개선을 제공하고자 한다

1. 서 론

기후, 팬데믹에 따른 차량용 공기청정기 수요 급증하고 있다. 미세먼지 이슈와 세계적 팬데믹으로 인해 공기청정기는 가정을 넘어 자동차까지 영역을 넓혔다. 팬데믹은 미세먼지에 집중한 휴대용 공기청정기에서 각종 부유세균 살균이 가능한 차량 융합 시스템으로 전환되고 있으며, 휴대용공기청정기에서 차량융합 공기청정-정화 시스템으로 고도화 되고 있다.

특히, 기존 시스템의 한계로 신개념 시스템 필요한데, 공조장치 설치형의 경우 차량 전체를 커버하기에 용량이 부족, 탑승객이 없을 시에만 작동가능한 단점이 있으며, 휴대용 제품의 경우 대부분 미세먼지에 집중, 부유 세균, 바이러스 살균 성능이 떨어진다. 그리고 가장 중요한 탑승객의 날 숨(비말)으로 인한 오염원 제거에 미비하며, 이산화탄소 농도 상승에 취약함을 보이고 있어 신개념 기술이 필요한 시점이다.

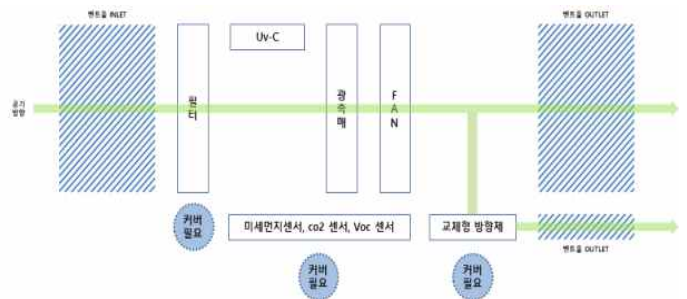
이에, 기존 필터 기반에 미세먼지 제거에 광촉매 기술을 더하여 부유세균을 제거하며, 이를 탑승객 개별적으로 호흡기 주변에 국부적으로 집중하여 효율적인 오염원 제거 및 오염원 방지를 구현하고자 한다. 아울러, 실험과정에서 획득한 오염원 검출, 제거 데이터를 학습시켜 실시간

간 교통수단 내에서 위생상태를 유지하고 졸음환경을 개선할 수 있는 시스템을 고안하고자 한다.

2. 본 론

1) 공기청정모듈 제작

' 호흡기 주변 공기청정, 정화 ' 를 주기능으로 ' 졸음 방지, 개선 기능 ' 을 위해 미세먼지+부유세균+휘발성유기물질 제거 가능한 모듈을 제작 미세먼지와 부유세균을



<그림 1. 국부공기청정정화장치 구성>

제거를 위하여 미세먼지+부유세균+휘발성유기물질 제거 가능한 모듈을 제작 미세먼지와 부유세균을 제거를 위하여 모듈을 위와 같이 구성하고, 호흡기 주변에 국부적인

공기정화존 형성을 위해 지속적인 깨끗한 공기 공급관로를 설계하였다.



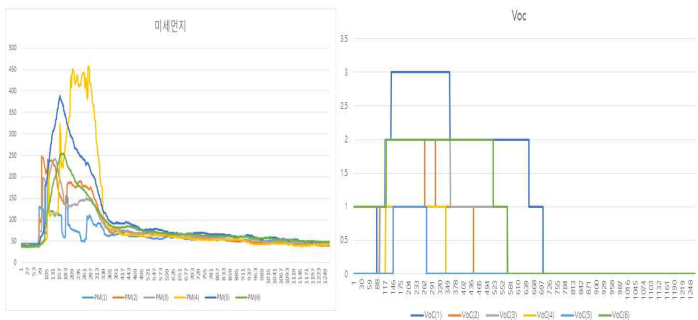
<그림 2. 국부공기청정구역 형성을 위한 관로설계>

2) 성능확인 및 개선을 위한 테스트베드 제작
실차와 유사한 환경에서 각종 시험을 진행할 수 있는 테스트베드를 제작하여 오염원 검출하고, 개선된 데이터를 분석하여 시스템의 세부요소인 볼륨량, 광촉매강도 등을 조절한다.



<그림 3. 성능확인을 위한 테스트베드>

3) 실험데이터 AI학습을 통한 실시간 교통수단 실내 청결유지 및 줄음환경 개선



<그림 4. 오염원 검출 및 개선 데이터>

오염원 검출과 시스템가동에 따른 오염개선 데이터를 확보 후 AI로 학습시킨 후 '실시간 공기모니터링'을 통한 효율적인 기기가동을 구현하고자 함.

4) 분석데이터 국부공기정화장치 시스템에 적용

① 센서 데이터 분석 적용 :

센서 및 AI를 사용하여 차량 내부의 환경 센서 데이터를 실시간으로 분석 후 VOC(차량 내 휘발성 유기물질), 미세먼지, 이산화탄소, 온도, 습도 등과 같은 데이터를 모니터링하여 호흡기 질환 및 졸음운전 위험 요인을 탐지 후 국부공기청정정화 시스템에 전달.

② 국부공기청정정화시스템 효과 : 호흡기 주변을 빠르고, 확실하게 필터, 광촉매 살균을 통해 VOC, 부유세균 제거를 통한 호흡기질환 예방효과 제공, 이산화탄소 농도에 따른 아로마테라피 분사를 통해 줄음환경 개선.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30			
먼지센서	상									중									하														
Voc센서	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하
공기질 표시	나쁨			나쁨			보통			나쁨			보통			좋음																	
송풍강도	(강)			(강)			(약)			(강)			(약)			(약)																	
CO2센서	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하	상	중	하
방향모터 구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	구동	

<그림 5. 오염도 검출 데이터에 따른 환경구분 및 가동 시나리오>

5) 성능평가

미세먼지제거효과와 국부적인 살균력을 확인을 위해 공인인증시험을 실시하였고 만족한 수치를 얻었다.

① 미세먼지 제거능력 평가시험(CADR)

: 시험환경 챔버크기 8.0 m³(±0.5) 환경에서 20분간 미세먼지(0.3μm) 제거능력을 시험

- 시험결과 : 0.1m³/min

- 발행처 : 한국건설생활환경시험연구원(KCL)

② 부유세균저감을 평가시험

: 폐렴균을 8.0 m³챔버로 분사한 후 제품가동 유/무의 부유세균농도를 대조하여 감소율을 평가 함.

- 시험결과 : 99.9 % 살균

- 발행처 : 한국건설생활환경시험연구원(KCL)

1) 공기 정화 장치

시험항목	단위	시험방법	시험결과	비고	시험장소
형상효율	m³/min	(1)	0.10	(21 ± 1) °C, (45 ± 5) % R.H.	A
정화면적	m²	(1)	0.8		

※ 작동모드 : S/W ON
 ※ 시험용 필터의 크기 : (8.0 ± 0.5) m
 ※ 시험장소
 A : 인천광역시 남동구 당포로 25

시험 항목	시험방법	시험 결과			시험환경
		가동 전 농도 (CFU/m³)	가동 후 농도 (CFU/m³)	감소율 (%)	
부유미생물 저감 시험 (폐렴균)	공기 정화 장치 의뢰자 제시	1.6 × 10 ⁴	1.3 × 10	99.9	(23.0 ± 0.5) °C (50.5 ± 1.0) % R.H.

<그림 6. 성능확보 성적서>

3. 결 론

본 논문은 교통수단 탑승객에 개별 공기청정정화시스템을 적용하여 호흡기질환을 예방, 전파방지, 줄음환경 개선 등 위생성과 안전성을 높이는 제안이다. 더욱이 오염원 검출 및 개선에 대한 데이터를 AI학습시켜 '실시간 공기질 모니터링에 따른 자동 제어시스템'으로 발전시키는 단계라 할 수 있다. 이를 통해 에너지낭비를 줄일 수 있고, 센싱비용 등을 줄일 수 있으며, 확보한 데이터로 파생될 향후, 국부청정성능의 핵심인 광촉매에 담지, 파장조절 등으로 성능을 강화하여 개별위생도 증가와, 각종 질환 데이터 학습을 통한 줄음개선 외 한방테라피 등으로 발전시키고자 한다.

Acknowledgement

본 연구내용은 2022~23 퍼스널케어융합얼라이언스지원사업, 현대자동차그룹 오픈이노베이션 'Zero1ne'의 연구결과와 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성사업의 연구결과로 수행되었음 (IITP-2023-RS-2022-00156389)

웹 기반 악취방지시설 조합 예측 서비스

김다진⁰¹, 박지예¹, 이아현¹, 정태훈¹, 김세원², 정설영¹

¹경북대학교 컴퓨터학부, ²(주)태성환경연구소

rew0316@gmail.com, sw091313@gmail.com, ahn11297@gmail.com, taehoon1127@gmail.com,
ksw@ts-ei.com, snowflower@knu.ac.kr

A Web-based Prediction Service for Combining Odor Control Facilities

Dajin Kim⁰¹, Jiye Park¹, Ahyeon Lee¹, Taehoon Jeong¹, Sewon Kim², Seol-Young Jeong¹

¹School of Computer Science and Engineering, Kyungpook National University

²Taesung Environmental Research Institute Co.

요 약

악취배출사업장에는 악취배출시설에서 배출되는 악취가 법률상의 배출허용 기준 이하로 배출될 수 있도록 악취방지시설을 설치해야 한다. 그러나 악취방지시설 기본설계 과정에서 많은 시간과 인력이 소모되므로 자동화가 필요하다. 또, 기존의 악취방지시설 예측 프로그램은 가용성과 시인성이 낮고 결과물 관리가 힘들다. 따라서 본 연구는 악취방지시설 예측에 필요한 기본 정보를 입력하면 그에 따른 악취방지시설 조합을 제시하고 각 조합의 단가와 제거 효율을 보여주는 데이터베이스 기반 웹 서비스를 개발하고자 한다.

1. 서 론

악취배출사업장은 악취배출시설에서 배출되는 악취가 법률상의 배출허용 기준 이하로 배출될 수 있도록 악취방지시설을 설치해야 한다[1]. 악취는 다른 환경적 요인과 다양한 악취기여물질이 복합적으로 나타나는 경우가 많아 전처리시설, 방지시설, 후처리시설 및 탄소흡착시설의 구성이 필요하다. 그에 따른 악취방지시설 구성, 설치 예산 산정, 각 시설별 효율 계산 등에 많은 검토가 필요하여 많은 시간과 인력이 소모된다[2]. 따라서 각 시설의 사용조건과 제거 효율을 기반으로 악취방지시설 구성을 결정하고 예산을 산정하는 과정을 자동화할 필요가 있다.

악취 방지시설 구성과 예산을 산정하는 기존의 프로그램이 존재하나, 이는 외부 현장에서 사용하기 어려운 설치형 프로토타입 프로그램으로 가용성과 시인성이 좋지 않고, 입력과 결과가 저장되지 않아 과거의 예측 기록을 이용하기 어렵다. 본 논문은 이를 보완하여 시·공간의 제약을 받지 않는 웹을 기반으로, 입력과 결과를 저장할 수 있는 서비스를 제공하는 것을 목표로 한다. 또한, 입력과 결과에 더해, 프로그램 내부 데이터를 데이터베이스로 관리하여 유지관리 비용을 절감할 것이다.

따라서 본 논문은 악취방지시설 예측에 필요한 기본 정보를 입력하면 그에 따른 악취방지시설 조합을 제시하고 각 조합의 예산을 산정하는 데이터베이스 기반 웹 서비스를 개발하고자 한다.

2. 관련 연구

2.1 유사 시스템 개발 사례

연구에 앞서, 본 논문에서 제안하는 프로그램과 유사한 시스템 개발 사례를 조사하였다. 아래의 두 연구는 웹 기반에, 기본 정보를 바탕으로 기본설계를 자동화한다는 점에서 본 논문과 유사성이 있다.

교량 구조물 설계를 위한 자동화 시스템[3]은 전체적인 설계 및 유지관리 등을 자동화하는 웹 서비스를 제공한다. 설계과정을 모듈화, 데이터베이스화하고 통합 시스템을 네트워크상에 구축함으로써, 소모되는 인력과 시간을 줄이고 시·공간의 제약성을 없앴으며, 유지관리 비용을 절감하였다. 또한, ASP(Application Service Provider) 시스템을 통하여 확장성이 높은 시스템을 제공한다.

로봇 기초 설계를 위한 웹/모바일 기반 대화형 서비스[4]는 기초 설계과정 중 구동기 및 제어기 선정을 자동화하는 프로그램이다. 기존의 자동화 시스템에서 사용하던 시뮬레이션 방식이 상호 의존적인 경향을 보이는 문제를 개선한 결과를 제공한다. 또한, 기본적인 설계 정보만을 이용해 구동기를 선택할 수 있도록 하며, 시공간적 제한을 받지 않도록 웹과 모바일 기반으로 개발하였다.

2.2 기존 악취방지시설 예측 프로그램

기존의 악취방지시설 예측 프로그램은 입력받은 22종의 지정악취물질[5] 농도와 방해인자 값, 추가설비 여부, 목표희석배수에 따른 악취방지시설 조합을 예측해 제시하는 프로그램이다. 이는 설치형 프로토타입 프로그램으로, 외부 현장에서 사용하기에는 적절하지 않다. 또, 예측한 결과를 저장할 수 없어 이용 및 관리에 어려움이 있고, 각 물질이나 방지시설별 기본값이 프로그램 내부적으로 포함되어 있어 해당 값의 관리가 힘들다는 아쉬움이 있다.

3. 시스템 상세 설계

3.1 시스템 기능

본 연구는 악취방지시설 예측에 필요한 인자인 악취기여물질, 방해인자, 추가설비, 목표회석배수와 악취방지시설의 사용조건을 바탕으로 악취방지시설 조합을 제시하여 비교·선정할 수 있는 웹 서비스를 개발하는 것을 목적으로 한다. 이후 사용자가 특정 조합을 선택하면 해당 시설 조합을 사용했을 때의 악취기여물질과 방해 인자의 기대 농도, 예측 이론적 회석배수, 시설별 단가를 제공한다. 또한 설정 페이지에서 사용자는 입력 인자의 최소감지농도, 수용성 여부, 악취방지시설의 제거 효율, 단가를 설정할 수 있다. 그림 1은 악취방지시설 예측 프로그램의 기능적 요구사항을 사용사례 다이어그램으로 나타낸 것이다.

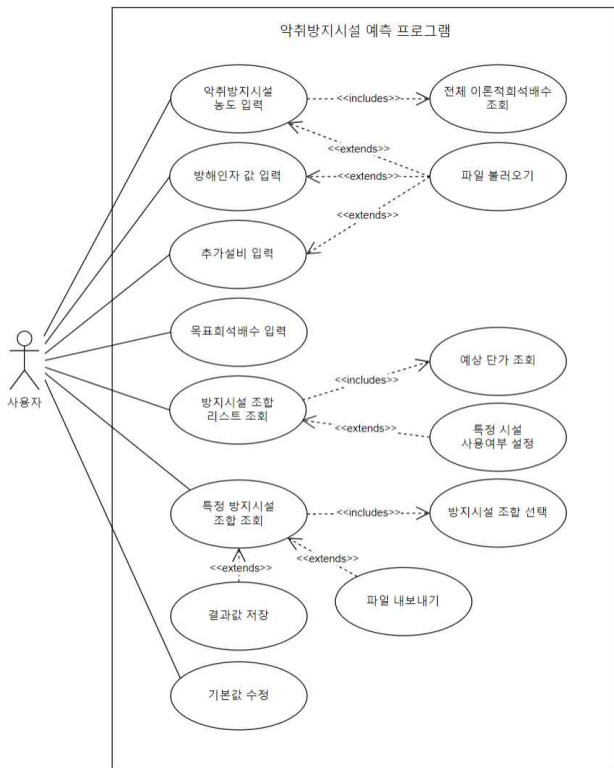


그림 1. 사용사례 다이어그램

3.1.1 입력

악취기여물질의 농도와 방해인자의 값 및 목표값, 그리고 추가설비 사용 여부, 풍량, 목표회석배수를 입력받는다. 입력창 혹은 csv, tsv 형식의 파일을 통해 입력받거나 데이터베이스에 저장된 정보를 불러올 수 있다. 입력창에서 악취기여물질 섹션에는 목표농도인 최소감지농도와 수용성 여부를 표시한다. 표 1은 입력인자의 분류와 예시를 보여준다.

표 1. 입력인자 분류와 예시

분류	예시
악취기여물질	암모니아, 황화수소, 톨루엔, 아세트알데하이드, 이산화탄소 등
방해인자	먼지, COx, 산가스, 오존, 온도, 유기실리콘화합물 등
추가설비	폐수처리장 시설, 에너지 활용시설 등

이를 바탕으로 악취기여물질 각각의 이론회석배수, 이론회석배수의 합, 기여율을 계산한다. 이론회석배수는 (농도)/(최소감지농도)로 계산되며, 농도가 최소 감지 농도 이하가 되도록 하는 회석배수를 의미한다. 악취기여물질과 방해인자, 추가설비, 목표 회석배수는 악취방지시설의 사용 가능 여부와 악취방지시설 조합을 결정하는 데 사용된다. 풍량은 시설 단가를 결정하는 데 사용되며, 단가는 기본 단가에 풍량에 따라 선형적으로 비용이 추가된다.

3.1.2 악취방지시설 예측 조합

입력된 값과 악취방지시설의 사용조건을 바탕으로 회석배수 합이 목표회석배수보다 작도록, 입력된 방해인자 값이 목표값보다 작게 하는 악취방지시설 조합 리스트를 찾아 제공한다. 배제할 악취방지시설을 선택하면 조합 리스트에 실시간으로 반영된다. 표 2는 악취방지시설 구성요소와 그 예시를 보여준다.

표 2. 악취방지시설 구성요소와 예시

구성요소	예시
전처리시설	cyclone, bag filter, packed tower scrubber - AOP 등
방지시설	packed tower scrubber - NaOHAOP, bio filter, high temperature plasma 등
후처리시설	activate carbon tower, SCR, SNCR 등
탄소흡착시설	amine absorption & desorption tower

사용조건에 경우 전처리 시설과 후처리 시설, 탄소흡착시설은 방해인자와 추가설비에 의존하고, 방지시설은 악취기여물질과 방해인자, 추가설비에 의존한다. 사용조건은 1차 조건과 2차 조건으로 나뉜다. 1차 조건은 추가설비 필요 유무, 방해인자 값의 범위 등 사용 가능 여부에 대한 조건이고, 2차 조건은 제거 효율, 우선순위 등 효율에 대한 조건이다. 1차 조건의 예로 전처리시설의 BagFilter는 먼지 농도가 100ppm 이하일 때에는 사용할 수 없고, PackedTowerScrubberAOP는 추가설비 항목의 폐수처리장이 사용 가능할 때만 사용할 수 있다. 2차 조건의 예는 전처리시설의 경우 먼지 제거설비, 유분 제거

설비, 수분 및 온도 제거설비 순서로 우선순위가 높게 부여되고, 먼지 제거설비 내에서 Packed tower Scrubber - AOP는 우선순위가 낮다. 사용조건은 시스템의 확장성을 위해 데이터베이스에 분류하여 저장한다.

3.1.3 악취방지시설 예측 조합 선택

특정 악취방지시설 조합을 선택하면 해당 조합에 대한 시설별 이미지와 적용했을 때의 결과를 보여준다. 결과에는 악취기여물질의 예측 농도, 방해인자의 예측값, 예측 희석배수, 단가가 해당한다. 결과 내용은 입력했던 데이터와 함께 데이터베이스에 별칭을 통해 저장하거나 csv, tsv 형식의 파일을 통해 내보낼 수 있다.

3.1.4 기본값 수정

설정 페이지에서 사용자는 악취기여물질의 최소감지값과 수용성 여부, 악취방지시설의 제거 효율과 우선순위, 단가를 수정할 수 있다.

3.2 시스템 구조

본 연구에서 개발하는 시스템은 크게 Human Computer Interaction Layer, Problem Domain Layer, Data Management Layer의 세 계층으로 나뉜다. 그림 2는 시스템의 구조를 패키지 다이어그램으로 나타낸 것이다.

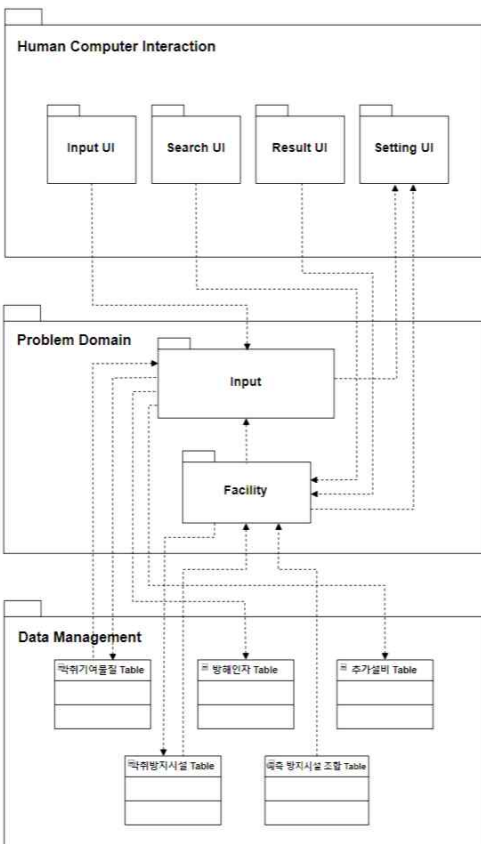


그림 2. 패키지 다이어그램

Human Computer Interaction Layer는 사용자와 직접적으로 상호작용하는 UI들로 이루어진다. Problem Domain Layer는 입력인자에 해당하는 Input 패키지와 악취방지시설 구성요소에 해당하는 Facility 패키지가 내부 알고리즘을 구성한다. Data Management Layer는 각 데이터베이스 구조에 해당하여 Problem Domain Layer와 서로 연관되며, 각 계층 내 패키지들은 데이터의 흐름에 따라 그림 2와 같은 의존 관계를 이룬다.

4. 시스템 구현 결과

시스템은 입력인자, 예측설비, 결과, 설정의 네 개의 탭으로 구성된다. 각 탭에 대해 사용자의 입장에서 사용 경험을 설명하고자 한다.



그림 3. 입력인자 탭

그림 3은 첫 번째 탭인 입력인자 탭이다. 입력인자 탭에서 사용자는 악취방지시설 예측에 필요한 악취기여물질, 방해인자, 추가설비를 입력한다. 입력은 입력창 혹은 파일을 통해 입력하거나 검색을 통해 데이터베이스에 저장된 정보를 불러올 수 있다. 이론적 희석배수 조회 버튼을 누르면 악취기여물질 농도에 대한 이론적 희석배수를 계산하여 출력한다. 사용자는 희석배수를 확인한 후 목표희석배수를 설정할 수 있다. 목표희석배수를 입력하고 방지시설 예측 버튼을 누르면 다음 페이지인 예측설비 탭으로 이동한다.



그림 4. 예측설비 탭

그림 4는 두 번째 탭인 예측설비 탭이다. 예측설비 탭에서 입력한 인자에 따른 사용 가능한 전처리시설과 방지시설, 후처리시설, 탄소흡착시설 리스트를 볼 수 있다. 해당 리스트를 바탕으로 목표회석배수에 부합하도록 조합된 악취방지시설 리스트를 그 아래에서 볼 수 있다. 사용자는 단가와 이론적 회석배수 결과를 통해 각 조합을 비교할 수 있다. 조합을 선택하면 구성 시설에 대한 간단한 정보를 리스트 오른쪽에서 볼 수 있다. 결과 조회 버튼을 클릭하면 다음 페이지인 결과 탭으로 이동한다.

세 번째 결과 탭에서는 선택한 악취방지시설 조합을 적용했을 때의 상세 결과를 볼 수 있다. 해당 방지시설 조합을 적용한 후의 악취기여물질 예측 농도와 방해인자의 예측값, 예측 회석배수, 단가가 이에 포함된다. DB에 저장하기 버튼을 누르면 입력과 결과를 데이터베이스에 저장할 수 있다. 파일로 내보내기 버튼을 누르면 파일로 저장할 수 있다.

마지막인 설정 탭에서는 입력인자 탭에 나타나는 기본값인 악취기여물질의 최소 감지 농도, 각 시설의 단가와 제거효율을 변경할 수 있다. 입력인자 탭의 악취기여물질, 방해인자, 추가설비와 예측설비 탭의 전처리시설, 방지시설, 후처리시설, 탄소흡착시설의 리스트 정보는 시스템의 확장성을 위해 데이터베이스에 저장된다. 웹 페이지에서 이 데이터베이스의 정보를 읽어 표시하기 때문에 데이터베이스의 값을 수정하면 웹 페이지에서도 수정된 정보가 반영되어 유지관리의 편의를 증대시킬 수 있다. 이때 데이터베이스에 직접 접근하여 값을 수정하려면 관련 지식이 필요하므로, 이를 보완하기 위해 설정 탭을 통하여 웹 기반 데이터베이스 관리 툴을 사용할 수 있도록 하였다.

5. 결론

본 논문에서는 악취방지시설 설계에 필요한 기본 데이터를 바탕으로 악취방지시설 조합을 예측하고 각 조합의 단가와 제거효율을 보여주는 웹 기반의 프로그램을 제안하였다. 제안한 프로그램은 자동화를 통해 신속한 기본 설계가 가능하며 효율성을 높일 수 있다. 웹 기반의 프로그램으로 구현하여 시간과 공간의 제한 없이 사용 가능한 환경을 제공할 수 있다. 또, 내부 데이터를 데이터베이스화하여 수정과 관리에 용이하며, 결과물을 파일과 데이터베이스에 저장하여 보관함으로써 정보의 연속성과 사용자 편의를 증대시킬 수 있다.

본 논문은 사용자가 직접 또는 파일을 통하여 기본 데이터를 입력하는데, 이때 악취배출사업장 데이터베이스와 연동하여 방지시설 예측 결과물을 관리하는 기능은 제공하지 않고 있다. 추후 연구를 통하여 악취배출사업장과 사업장 내 배출구별 데이터를 이용해 악취방지시설을 관리하는 기능을 추가할 예정이다.

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음” (2021-0-01082)

참 고 문 헌

- [1] 악취방지법(2004. 2. 9.) 제8조(2023. 3. 28. 법률 제 19310호로 개정된 것)
- [2] 이병관, 복합 탈취시스템을 이용한 악취물질의 제거, 국내박사학위논문 호서대학교 대학원, 2021.
- [3] 한영환, 구조물 설계를 위한 자동화 시스템 개발, 한국정보기술학회논문지, 7, 6, 189-197, 2009.
- [4] 김경달, 이민철, 자동화 및 로봇 기초 설계를 위한 웹 /모바일 기반의 대화형 서비스 개발. 한국정밀공학회 2013년도 춘계학술대회 논문집, 495-496, 2013.
- [5] 악취방지법 시행규칙(2005. 2. 7) 제2조(2023. 9. 27. 환경부령 제1055호로 개정된 것)

고회수율 RO 시스템 제어를 위한 인공지능 모델 설계

전민재¹, 이상현¹, 배진성¹, 배민석¹, 정설영¹, 임환규²

경북대학교 컴퓨터학부¹, (주)아쿠아웍스²

ssoui2@knu.ac.kr, qhemf0410@naver.com, bae1776@gmail.com, otuera1123@gmail.com,

snowflower@knu.ac.kr, okhawinkyu@gmail.com

Design of optimal artificial intelligence model for high-recovery RO system control

Min-Jae Jeon¹, Sang-Hyun Lee¹, Jin-Sung Bae¹, Min-Seok Bae¹, Seol-Young Jeong¹, Hwanky Lim²

Kyungpook National University School of Computer Science and Engineering¹, Aquaworks²

요약

수자원을 효율적으로 사용하기 위해서는 하수처리나 해수담수화 등 사용할 수 없는 물을 재사용 해야 한다. 하수처리 시스템 중 하나인 RO시스템을 효율적으로 작동하기 위해서는 적절한 시기에 RO필터의 약품 세척, 막 교체를 해야 한다. 기존에는 운영자의 경험에만 교체 시기를 의존하였지만, 공정의 정확성과 효율을 높이기 위해 여러 인공지능 모델을 비교하여 최적의 모델을 찾아 보았다.

1. 서론

하수처리 방법 중 하나인 RO시스템은 역삼투막 방식으로 멤브레인 필터를 거쳐 순수한 물을 얻는 시스템이다. 그 중 막 여과 공정은 RO필터를 이용하여 기상 또는 액상의 혼합물을 막 양면의 농도 차, 압력 차, 전위 차 등에 의하여 하수를 분리, 농축, 정제한다. 따라서 공정을 진행하다 보면 RO필터의 성능이 떨어지게 되고, 약품세척이나 RO필터 교체를 필요로 하게 된다. 기존에는 교체 시기를 시스템 운영자의 경험에만 의존하였지만, 수질 데이터베이스를 통한 예측으로 세척, 교체 시점을 최대로 증가시킴으로써 정확도를 높이고, 공정의 운영비를 절감하는 것이 본 연구의 목적이다. 세척과 교체 시기의 기준은 RO필터의 압력 값으로 하수처리 시설에서 얻은 수질 데이터를 활용해 인공지능 알고리즘으로 필터의 압력을 예측 하였으며, 알고리즘 별로 성능을 비교해 보았다.

2. 이론 배경

2.1 LSTM(Long short term memory)

LSTM 은 순환신경망(RNN)의 한 종류로 시계열 데이터와 같은 순차적 데이터를 예측하는데 사용된다. RNN은 학습을 계속할수록 매개변수인 가중치가 업데이트 되지 않는 기울기 소실 문제가 발생한다. LSTM은 RNN의 이런 문제를 해결하기 위해 나온 모델이다. LSTM은 [그림1]과 같이 3개의 게이트로 구성되어 있다.

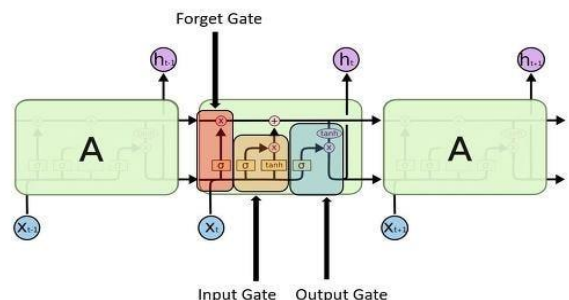


그림 1. LSTM 구조

망각 게이트는 이전 상태의 정보의 기억 여부를 결정한다. 입력 게이트는 현재 정보의 중요도를 판단하여 새로운 정보 중 어떤 정보를 저장할지 결정한다. 그리고 출력 게이트는 망각 게이트와 입력 게이트에 의해 변경된 cell state로부터 어떤 값을 output으로 도출할지 결정하는 게이트이다. LSTM은 이 3가지 게이트를 사용하여 RNN에 비해 비교적 먼 거리의 정보를 효과적으로 전달할 수 있다[1].

2.2 ARIMA(Autoregressive Integrated Moving Average)

ARIMA는 자기자신을 종속변수로 하고 이전 시점의 시계열을 독립변수로 갖는 AR모델과 자기자신을 종속변수로, 해당 시점과 과거의 distribution error들을 독립변수로 갖는 MA모델을 합친 모델인 ARMA모델에 추세까지 고려한 모델이다. ARMA 모델은 시계열 데이터가 정상성을 만족한다는 가정하에 적용이 가능하지만 ARIMA 모델은 차분(Integrated)이라는 개념을 통해 정상성을 가지지 않는 상황에서도 예측이 가능하다. ARIMA 모델의 표시는 ARIMA(p,d,q)로 표시되며, d차 차분한 데이터에 AR(p)모형과 MA(q)모형을 합친 모델을 의미한다.

ARIMA 모델은 자기회귀(AutoRegressive, AR), 차분(Integrated, I), 이동평균(Moving Average, MA)이라는 세가지 구성 요소가 합쳐져 시계열 데이터의 경향과 패턴을 파악하고, 미래 값을 예측하는데 효과적이다[2].

2.3 Prophet

Prophet은 시계열 데이터를 예측하기 위한 개방형 소스 머신러닝 알고리즘 중 하나이며, Facebook에서 개발한 라이브러리다. Prophet은 데이터에 주기적인 패턴과 휴일 효과를 자동으로 모델링 할 수 있는 기능을 제공하고, 이상치(Outliers)를 감지하고 처리하는 기능을 내장하고 있어, 이상치가 예측에 미치는 영향을 줄일 수 있다. Python 및 R에서 사용할 수 있으며, 간단한 API를 제공하여 비전문가도 쉽게 모델을 구축하고 예측을 수행할 수 있다. Prophet은 Trend, Seasonality, Holiday의 3가지 구성요소로 이루어져 있으며 다음의 관계식을 따른다.

$$y(t) = g(t)+s(t)+h(t)+E(t)$$

g(t)는 비주기적 변화를 반영하는 추세함수, s(t)는 주기적인 변화

(weekly, hourly--)를 반영, h(t)는 불규칙한 이벤트(특정 시기에 값이 감소/증가)를 반영, E(t)는 오차를 반영하는 관계식이다[3].

3. 데이터 수집 및 전처리

3.1 데이터 수집

수질데이터는 ㈜아쿠아웍스에서 수집한 수질 데이터를 바탕으로 분석을 진행하였다. 수집한 데이터에는 RO처리수 유량, RO순환수 유량, PF배출 유량, 유입 압력, 토출 압력이 30초 단위로 기록되어 있으며 본 연구에서는 그중 유입 압력 값만 추출하여 사용하였다.

3.2 데이터 전처리

데이터를 수집할 때 몇몇 오류로 인해 이상치(outlier)가 발생할 수 있다. 이런 이상치들로 인해 과잉적합(over-fitting)등의 문제가 발생하여 모델의 정확도를 떨어뜨릴 수 있다. 따라서 데이터의 이상치들을 제거해야 하며 IQR(Inter Quartile Range)의 1.5배를 벗어난 데이터를 이상치로 간주하였고[4], [그림2]의 raw data를 [그림3]과 같이 이상치를 제거하였다.

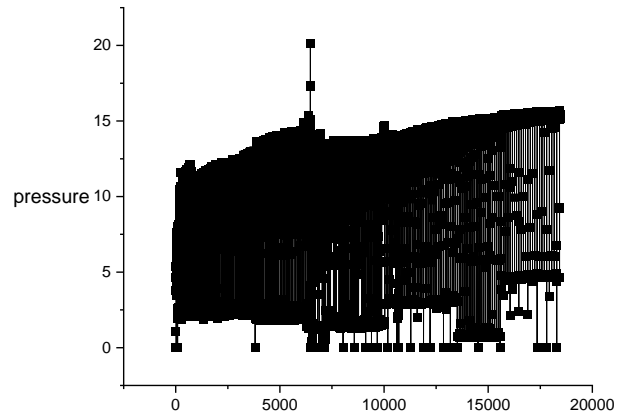


그림 2. Raw Data

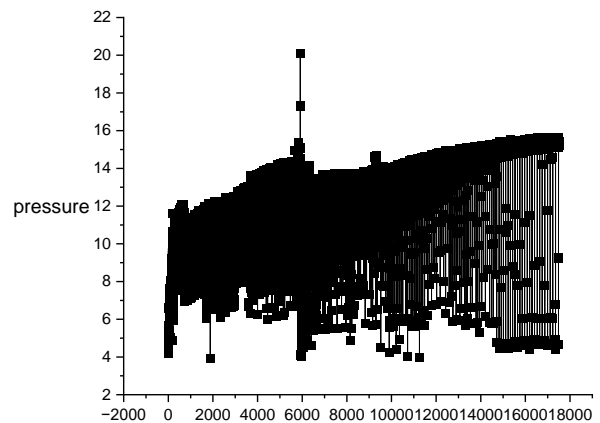


그림 3. 이상치 제거

4. 실험결과

수집한 수질 데이터로 앞서 소개한 LSTM, ARIMA, PROPHET을 사용하여 유입 압력 값을 예측하였다. 예측 모델의 성능 검증을 위해 수질 데이터의 80%를 훈련 집합으로 20%를 테스트 집합으로 구분하여 예측치와 비교하였다. 모델의 정확도를 평가하기 위해 평균 제곱 오차(RMSE)를 사용하였다.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

RMSE는 회귀(Regression)문제의 정확도를 측정하기 위해 사용되며 실제 값과 예측 값과의 차이를 나타낸다.

4.1 LSTM

LSTM의 은닉층 개수는 50개, 시퀀스 길이는 10, 활성화 함수는 relu, 옵티마이저는 adam을 사용하였다. 과거의 필터 압력 값들을 데이터로 넣으면 특정 시점의 압력 값을 출력하는 many-to-one을 사용하였다.

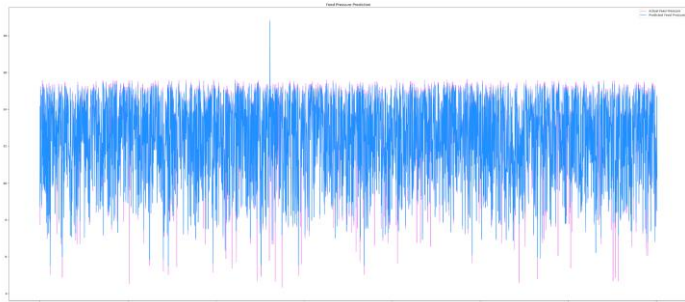


그림 4. LSTM 예측

[그림 4]처럼 예측결과가 나왔으며 LSTM 모델의 평균 제곱근 편차(RMSE)는 0.8이다.

4.2 ARIMA

ARIMA의 차수 (p,d,q)는 AIC(Akaike's information Criterion)로 추정[5]하여 AIC 값이 가장 낮은 차수(3,0,2)를 선택하였다[그림 5].

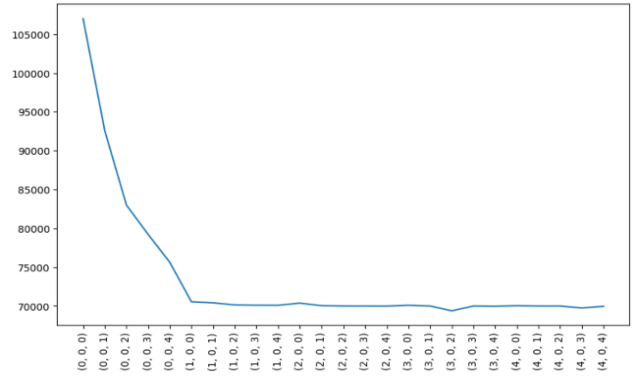


그림 5. ARIMA_AIC

다음은 차수 (p,d,q)가 (3,0,2)인 ARIMA model의 요약본이다.

SARIMAX Results						
Dep. Variable:	feed pressure	No. Observations:	16852			
Model:	ARIMA(3, 0, 2)	Log Likelihood:	-34683.362			
Date:	Fri, 03 Nov 2023	AIC:	69380.724			
Time:	01:49:00	BIC:	69434.850			
Sample:	0	HQIC:	69398.581			
			- 16852			
Covariance Type: opg						
	coef	std err	z	P> z	[0.025	0.975]
const	7.5519	3.901	1.936	0.053	-0.094	15.197
ar.L1	1.0343	0.020	52.953	0.000	0.996	1.073
ar.L2	0.6651	0.032	20.599	0.000	0.602	0.728
ar.L3	-0.6995	0.015	-47.983	0.000	-0.728	-0.671
ma.L1	-0.2598	0.020	-12.749	0.000	-0.300	-0.220
ma.L2	-0.7226	0.020	-36.488	0.000	-0.761	-0.684
sigma2	3.5887	0.024	148.251	0.000	3.541	3.636
Ljung-Box (L1) (Q):	50.70	Jarque-Bera (JB):	506612.47			
Prob(Q):	0.00	Prob(JB):	0.00			
Heteroskedasticity (H):	3.22	Skew:	-4.11			
Prob(H) (two-sided):	0.00	Kurtosis:	28.57			

그림 6. ARIMA 요약

다음은 훈련 데이터와 예측 데이터를 비교한 그래프이다.

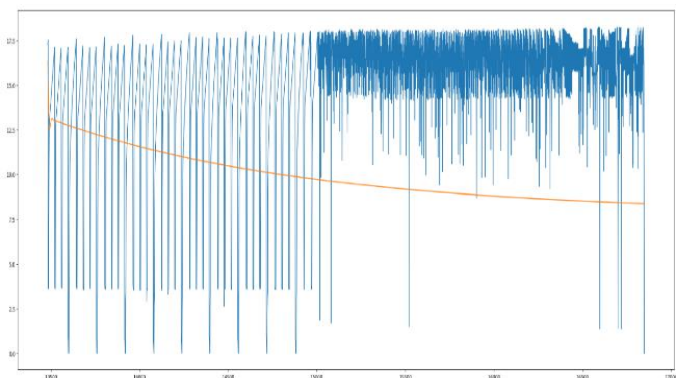


그림 7. ARIMA 예측

LSTM과 다르게 ARIMA는 데이터를 정확하게 예측하기보다는 추세를 예측한다. LSTM 보다 정확도가 떨어지며, ARIMA 모델의 평균 제곱근 편차(RMSE)는 6.56이다.

4.3 PROPEHT

$$y(t)=g(t)+s(t)+h(t)+\epsilon_i \quad [6]$$

g(t)는 비주기적 변화를 반영하는 추세함수, s(t)는 주기적인 변화를 반영, h(t)는 불규칙한 이벤트(특정 시기에 값이 감소/증가)를 반영, E(t)는 오차항이다. 주기적인 변화로 30초 간격의 계절성을 불규칙한 이벤트로 필터의 압력 값이 공정마다 바뀌는 시점을 반영하였다.

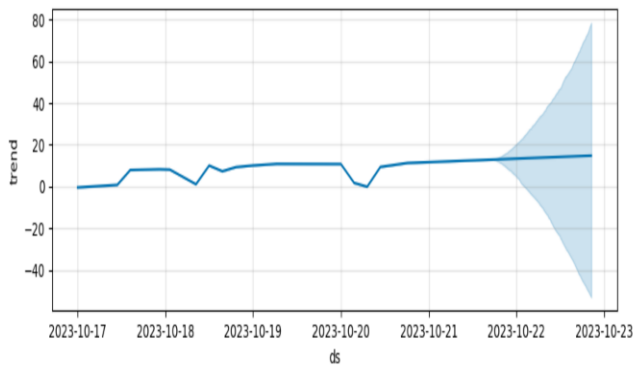


그림 8. PROPHET 요약

PROPHET 모델의 평균 제곱근 편차(RMSE)는 3.72이다.

4.4 결과비교

	RMSE(평균 제곱근 편차)
ARIMA	6.56
LSTM	0.8
PROPHET	3.72

표 1. 결과 비교

LSTM이 가장 뛰어난 성능을 보였고 ARIMA는 데이터에 대한 추정보다는 추세를 따르기에 실제로 사용하기에 적합하지 않았다. PROPHET은 LSTM보다는 성능이 떨어졌지만, 구현이 간단하고, 속도가 빨라 실제 시스템에 적용하기에 유연하다.

5. 결론

RO 시스템에서 멤브레인 필터를 효율적으로 사용하기 위해 수질 데이터를 사용하여 최적의 인공지능 모델을 찾아보았다. LSTM, ARIMA, PROPHET을 비교하였으며 그 중 LSTM이 가장 뛰어난 성능을 보였다. 인공지능 모델을 통해 필터의 압력 값을 예측하여 세척, 교체 시기를 알 수 있는 경우 운영자의 경험에만 의존한 이전과는 달리 보다 효율적이고 정확한 공정이 가능해진다.

References

- [1] 신동하, 최광호, 김창복, "RNN과 LSTM을 이용한 주가 예측을 향상을 위한 딥러닝 모델" Journal of KIIVol.15(2017): pp.916(2017)
- [2] 류병로, 한양수, "ARIMA 모형에 의한 하천 수질예측. Journal of Environmental Science International", 한국환경과학회지 Vol.7 Issue 4(1998): pp 433-440
- [3] 김준석, 강재환, 김성희, "윤주상 Prophet을 사용한 일변량의 시계열 예측" 한국정보통신학회 2020년도 춘계종합학술대회 논문집 제24권 제1호(2017): pp 329-331
- [4] "Oracle(n.d.) Oracle Fusion Cloud EPM Planning 작업. IQR" pp 351.
- [5] Prapanna Mondal¹, Labani Shit¹ and Saptarsi Goswami, "STUDY OF EFFECTIVENESS OF TIME SERIES MODELING (ARIMA) IN FORECASTING STOCK PRICES" International Journal of Computer Science, Engineering and Applications (IJCSA) (2014): pp 15-19
- [6] Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." The American Statistician 72.1 (2018): 37-45

※ 사사의 글

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

재활의학적 통증 환자의 통증 기록관리 어플리케이션 설계

엄소민, 이우진

(주)신라시스템, 경북대학교 컴퓨터학부

eumsomin@gmail.com, woojin@knu.ac.kr

Design of Pain Record Management Application for Pain Patients

Somin Eum, Woo Jin Lee

SILLA SYSTEM Co.Ltd, School of Computer Science and Engineering, Kyeongpook National University

요 약

현재, 통신기술의 발전과 사회의 노령화로 인해 만성 질환 환자 수가 급증하고 있으나 의사 인프라는 부족한 상황입니다. 본 연구는 의사-환자 소통 문제를 다루며, 짧은 진료 시간 동안 환자의 효과적인 정보 전달을 돕기 위한 정보통신 기반의 기록관리 어플리케이션을 개발하고자 합니다. 이 어플리케이션은 환자와 의사 간의 의사소통을 강화하고, 진료 효율성을 향상시켜 정확한 진단 가능성을 높이며, 의료 서비스 품질을 향상시킬 것으로 예상됩니다.

1. 서 론

현재 우리는 통신기술의 비약적 발전으로 컴퓨터 및 스마트기기들의 사용이 늘어났고, 이를 업무와 여가 활동등의 모든 부분에 사용하고 있다. 여기에 사회의 고령화와 교통사고를 비롯한 각종 사고등은 재활의학적 통증환자들이 늘어날 수밖에 없는 상황을 만들었다. 그렇다면 늘어나는 환자수에 비해 진료환경은 어떠한가.

2023년 OECD 보건 통계[1]에 따르면 대한민국은 최고수준의 의료 장비의 질과 수에 비해 임상 의사(한 의사 포함)는 인구 1,000명당 2.5명으로 OECD 국가 중 하위권에 속한다. 여기에 외래진료건수는 최상위에 위치해 있기 때문에 국내 병원의 의사 1인당 진찰건수는 OECD 국가 중 최고 수준일 수밖에 없으며, 그로 인해 평균 진료시간은 겨우 3분 이내일 정도로 매우 짧다.

현대 의료 서비스에서 의사와 환자 간의 원활한 의사소통은 기본적인면서도 핵심적인 요소이다. 그러나 짧은 진료 시간과 의료 인력 부족으로 인해 의사는 제한된 시간 동안 진료차트를 확인하고 환자의 증상 및 부작용을 청취하며 치료방법을 결정하고 약을 처방하고 검사를 지시하는 등의 일련의 많은 작업을 수행해야 하고, 환자들은 진료 시간에 대한 압박감으로 자신의 상태를 충분하게 그리고 구체적으로 설명하지 못하고 진료를 마치게 된다.

이런 문제의 궁극적인 해결책은 의료인원을 확충하는 것이겠지만, 그것은 법률적, 정책적인 사회적 협의가 필요하기에 본 논문은 환자의 입장에서 소통에 도움이

될 만한 해결책을 찾는 것을 목적으로 하였다.

이전부터 많은 곳에서 이런 해결책들에 대해 논의되어 왔으며, 특히 해외에서는 본 논문과 같은 취지의 어플리케이션들이 이미 개발되어 유, 무료로 서비스되고 있다.

본 논문은 이런 것들 참고, 학습하여 짧은 진료 시간을 더 효과적으로 활용하고 환자와 의사 간의 의사소통을 개선하기 위해 환자가 병원을 방문하기 전에 자신의 증상과 통증의 객관적인 기록을 관리하기 위한 통증 기록관리 어플리케이션을 설계하는 것을 목표로 한다.

2. 통증 기록관리 시스템

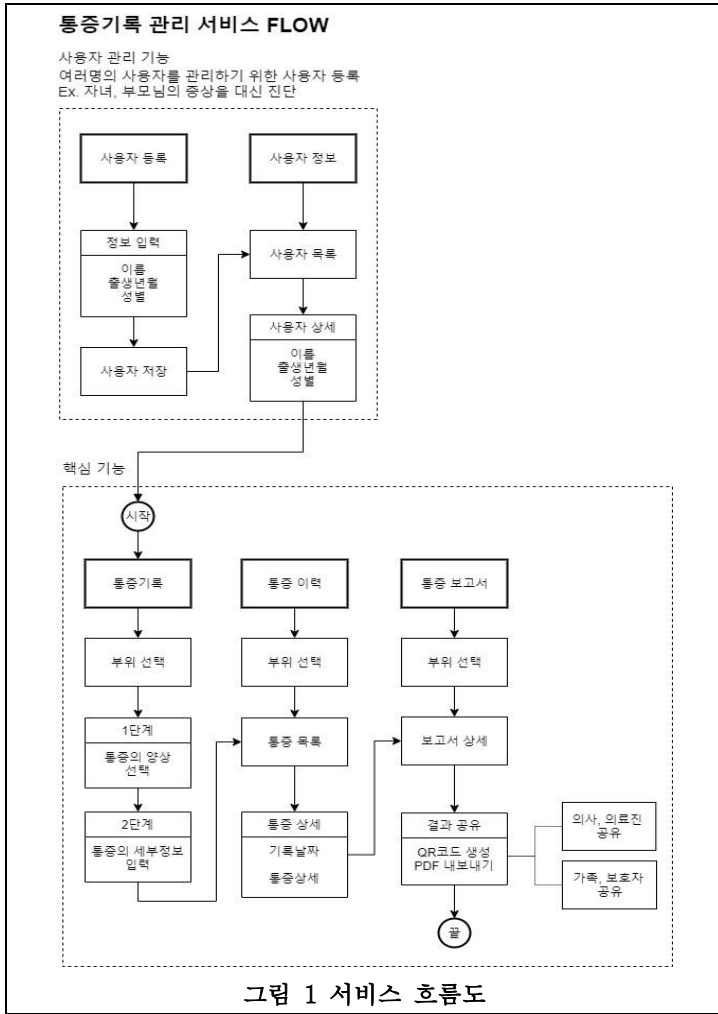
2.1 시스템 설계

통증 기록관리 어플리케이션을 시작하면 사용자의 등록 여부를 확인한다. 사용자 정보가 없을 경우에는 신규 사용자 등록 후에 어플리케이션을 사용할 수 있으며, 본인 외에도 어플리케이션의 사용이 어려운 가족들의 통증기록을 대신 관리할 수 있도록 사용자들 여러 명 등록할 수 있다.

통증이 발생하면 아픈 부위를 선택하여 증상을 기록할 수 있고, 통증 이력은 사용자별, 부위별, 날짜별로 저장된다. 기록된 정보는 통증 이력에서 달력 형태의 비로 제공되어 통증이 발생한 날짜를 쉽게 확인할 수 있다.

통증 기록이 7일이상 누적되는 경우 통증보고서가

생성되어 사용자별로 통증의 추이를 추적할 수 있도록 한다.



2.2 통증 기록

등록된 사용자의 최근 입력한 통증 기록이 있는지 확인하고 최근 입력한 데이터가 존재하는 경우에는 통증 부위별로 기록 개수와 마지막 기록일을 표시한다. 사용자의 새로운 통증을 기록할 수 있고, 기록된 통증의 증상이 지속되는 경우, 통증이 발생할 때마다 부위별로 통증을 기록할 수 있다. 통증 기록을 위한 문항은 근거기반 임상간호실무지침 (통증간호)[2]의 권고안 내용과 권고안 부록의 통증 초기사정 기록지와 통증 재사정 기록지를 참고하였다.

2.3 통증 이력

기록된 통증은 사용자별, 부위별, 날짜별로 저장되어 한달 단위로 통증의 기록을 한 눈에 볼 수 있다. 달력 UI에서 날짜 아래에 기분을 표현하는 아이콘과 동일한 색상 표시로 통증이 발생한 날짜의 기분을 시각적으로

표시한다. 달력에서 날짜를 선택했을 때 달력 하단 예목록으로 통증 기록을 출력한다. 목록에서 기분을 나타내는 아이콘과 통증의 정도를 알 수 있는 숫자를 표시하고, 상세 버튼을 눌러 상세정보를 확인할 수 있다.

2.4 통증 보고서

통증 보고서는 사용자별로 특정 부위의 통증 기록이 7일 이상 누적되어 기록된 경우 그 정보를 취합하여 생성한다. 보고서에는 기본적인 사용자의 정보 및 초기 통증 사정 기록의 통증 양상 정보를 출력하고, 통증의 경감 상태, 통증의 지속시간, 기분 상태 등의 추이 정보를 그래프로 제공한다. 보고서는 공유하기 링크를 생성하거나 PDF로 내보내기 기능을 사용하여 보호자나 의료진에 공유할 수 있도록 한다.

3. 시스템 설계 화면

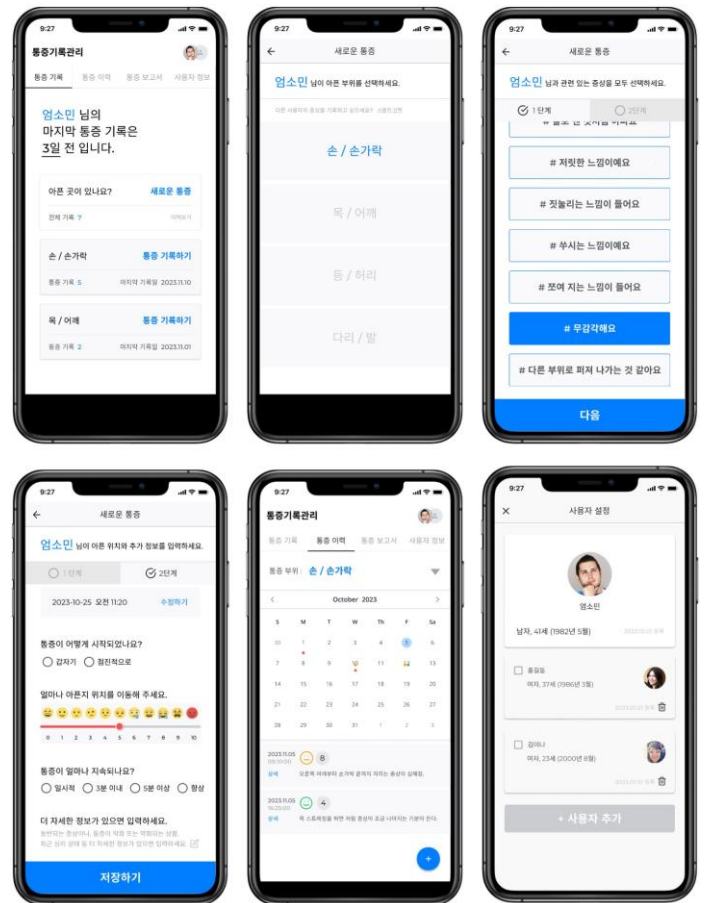


그림 2 앱 서비스 화면

4. 결 론

통증 기록관리 어플리케이션의 개발은 다음과 같은

기대효과를 가져올 것으로 예상된다. 먼저, 이 어플리케이션은 환자와 의사 간의 원활한 의사소통을 촉진하는 도구로 활용될 것이며, 이를 통해 환자는 어플리케이션을 통해 자신의 증상과 건강 상태를 사전에 기록하고 의사는 진료 이전에 해당 정보를 사전에 확인할 수 있어 의사와 환자 간의 의미 있는 대화와 정보 교환을 강화할 수 있을 것이다. 또한, 짧은 진료 시간을 효과적으로 활용하여 의료 현장의 효율성을 크게 향상시킬 것으로 기대된다. 의사는 환자에게 제공받은 보고서를 통해 사전에 환자의 정보를 확인하고 분석함으로써 진단에 필요한 추가적인 정보를 효과적으로 수집할 수 있다. 이는 의사가 정확한 진단을 내릴 수 있게 도와줄 것이다. 따라서 이 어플리케이션을 통해 정확한 진단이 가능해지고 환자들의 의료 서비스 품질이 향상될 것으로 기대된다.

5. 참고 문헌

- [1] OECD 보건통계(Health Statistics) 2023.
- [2] 은영, 근거기반 임상간호실무지침 (통증간호), 병원간호사회, 시리즈 11, 2018.
- [3] 이병혜, “의사-환자 간의 의사소통 장애요인 인식 차이와 소통만족도에 미치는 영향”, 커뮤니케이션학 연구, vol.19, no.1, pp. 35-54, 2011.
- [4] 김용, “환자 중심의 의료서비스 변화”, 과학기술학연구, vol.13, no.2, pp. 71-110, 2013.

(출판용 논문작성양식) * 저자정보 포함

음성인식 및 감정분석 기술 기반 감정일기 어플리케이션 설계 및 구현

김태환¹, 오동휘¹, 이휘찬¹, 한의진¹, 정현준², 정설영³

경북대학교, AtlasLabs

wksgo123@knu.ac.kr, hui1495@knu.ac.kr, kasterra@naver.com,
risingstaruj@gmail.com, jun.chung@atlaslabs.ai, snowflower@knu.ac.kr

Design And Implementation Of A Diary Application Using Voice Recognition and Sentiment Analysis Technology

Tae-hwan Kim¹, Dong-hwi Oh¹, Hwi-chan Lee¹, Ui-jin Han¹, Hyeon-jun Jeong²
and Seol-young Jeong³

Kyungpook National University, AtlasLabs

요약

본 논문에서는 STT(음성인식) 및 감정분석 기술을 활용한 감정일기 어플리케이션의 개발을 다룬다. 이 어플리케이션은 사용자가 일상적으로 느끼는 감정을 음성 인식을 통해 텍스트로 기록할 수 있고, 이를 감정 분석을 통해 감정 점수로 변환한 후, 한 달 동안의 감정 변화를 그래프로 시각화할 수 있다. 더 나아가서, 이 어플리케이션이 기분 기록지 작성을 돕는데 기여할 수 있으며, 따라서 사용자가 자신의 감정을 더 잘 이해하고 관리할 수 있도록 돕는 새로운 도구로서의 역할을 할 수 있음을 강조한다.

1. 서론

본 연구는 음성인식 기술(STT, Speech-To-Text)와 감정분석(Sentiment Analysis) 기술을 활용한 감정일기 어플리케이션의 설계 및 구현을 다룬다.

음성인식 기술(STT)은 사용자가 음성으로 감정을 표현하면 이를 텍스트로 변환하여 저장함으로써, 음성 메모를 효과적으로 기록할 수 있게 해준다. 감정분석 기술은 사용자가 저장한 텍스트를 분석하여 해당 텍스트의 감정 내용을 추출하고, 이를 정량화된 감정 점수로 표현한다. 음성 인식 기술과 감정분석 기술은 딥러닝과 빅데이터 기술을 기반으로 극적인 성능 향상을 이루었으며, 감정일기 어플리케이션의 개발은 이러한 기술의 상용화에 기여한다.

또한 현대 사회에서 스트레스와 감정 표현의 중요성이 커지고 있는 상황에서, 이 앱은 사용자가 일상적으로 느끼는 감정을 기록하고 시각적으로 분석할 수 있는 도구로 제공된다. 이를 통해 사용자는 감정 변화를 관찰하고 관리할 수 있으며, 개인의 감정 변화에 대한 통찰을 얻을 수 있다.

2. 관련 연구

이 논문에서는 기분 기록지(무드 차트)와 관련된 연구를

고려한다. 최종적으로 만들어진 어플리케이션이 기분 기록지를 작성하고 관리하는데 간접적으로 도움을 줄 수 있어야 함을 고려하였다. 기분 기록지는 사용자의 감정 상태를 일정 기간 동안 그래프로 시각화하여 보여주는 도구로, 사용자의 감정 변화를 파악하고 분석하는데 유용하다.[1]

여기에 더하여 AI 챗봇을 통해 기분기록지를 직접 작성하도록 하는 기존 어플리케이션[2] 또한 고려하였다. 본 어플리케이션은 사용자 접근성을 높이기 위해 기분 기록지를 직접 작성하도록 하는 방식이 아닌 보다 간접적인 방법을 취한다. 상담사 역할을 하는 챗봇을 배제하고, 개인이 기록하는 일기장의 방식을 채택하였다. 또한 일기장에 감정 분석 기술을 적용하여 보다 객관적인 시선으로 스스로 본인의 감정을 돌아볼 수 있게 했고, 텍스트로 작성하는 방식이 아닌 STT 기술을 활용하여 단순히 말하는 방식으로 일기를 기록하게 함으로써 접근성을 높였다.

3. 전체 시스템 구조 및 설계

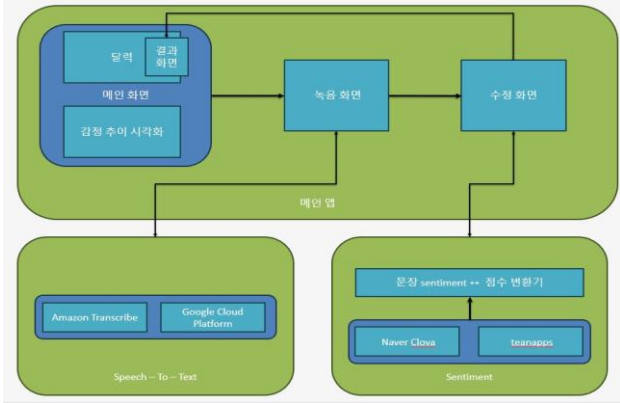


그림 1. 어플리케이션 흐름 및 구조도

Fig 1. Application Flow and Structure

그림 1에서는 본 어플리케이션의 작동 흐름을 간략하게 소개한다. 사용자가 앱을 실행하면, 녹음 버튼을 통해 음성으로 일기 형식으로 일일 감정을 표현한 음성 데이터를 받는다. 이어서 음성 인식 모듈에서 이진 음성 파일을 텍스트 형태로 변환한다. 변환된 텍스트는 감정 분석 모듈을 거쳐 임의의 감정 점수(Mood Score)로 변환된다.

3.1 음성 인식 모듈

음성 인식 모듈에서는 사용자의 음성이 입력되면 음성 신호를 문자로 변환하는 Speech To Text(STT)를 수행한다. 본 소프트웨어에서는 해당 기능을 구현하기 위해 시중의 STT 서비스 중 Google Cloud Speech API와 Amazon Transcribe를 사용한다. 둘은 퍼포먼스 및 비용 산정 문제 등 필요한 요구 조건에 따라 서버 단에서 교체될 수 있도록 했다.

3.2 감정 분석 모듈

음성 인식 모듈 과정에서 사용자의 음성 신호를 문자로 변환한 텍스트 파일을 감정 점수로 변환한다. 여기서 감정 점수를 다루는 Sentiment는 흔히 감정이라고 부르는 Emotion과는 다른 개념으로, 기쁨, 우울, 분노 등의 감정유형을 나타내기보다는 긍정,부정의 극성(Polarity)를 가진다.

$$Mood\ Score = 3 \times \frac{(W_{positive} - W_{negative})}{100}$$

그림 2. 감정 분석 결과를 기분 점수로 환산하는 수식
Fig 2. Formula for converting sentiment analysis results into mood scores

그림 2는 감정 분석 모듈이 Sentiment API로부터 받은 문장 별 Sentiment 분석 결과를 글 하나의 기분 점수로 환산하는 과정을 수식으로 나타낸 것이다. 동작 방식은 다음과 같다. 우선, 글을 문장으로 분할한다. 이후 각 문장에 긍정-중립-부정으로 분류하는 감정 분석을

수행한다. 이를 합산하여 전체 문단의 긍정-중립-부정 비율을 0~1 사이의 실수로 표현한다. 마지막으로 각 문장의 감정 분석 결과를 취합하여 -3 ~ +3 사이의 기분 점수로 변환한다. 긍정-부정-중립 중 가장 큰 감정에 가중치를 두고 긍정-부정의 비율을 고려하여 점수를 산정한다. 감정 분석 모듈에서는 시중의 자연어 처리 서비스 중 CLOVA Sentiment를 사용한다.

3.3 어플리케이션

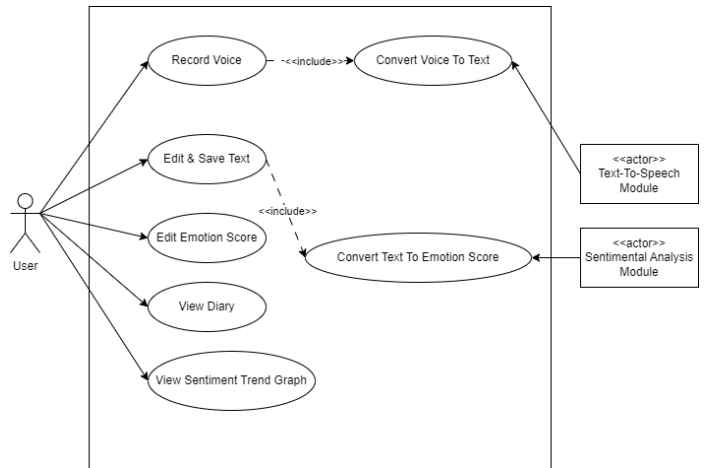


그림 3. 일기 어플리케이션 유즈케이스 다이어그램
Fig 3. Diary Application Use Case Diagram

그림 3은 본 어플리케이션에 필요한 기능을 유즈케이스 다이어그램으로 나타낸 것이다. 메인 어플리케이션의 주요 기능은 다음과 같다. 사용자의 음성 데이터를 스마트폰을 통해 녹음할 수 있다. 녹음된 음성 데이터는 음성 인식 모듈을 통해 텍스트로 변환된다. 텍스트는 다시 감정 분석 모듈을 통해 감정 점수로 변환된다. 사용자는 텍스트와 감정 점수를 일기 형태로 저장할 수 있다. 또한 사용자는 저장된 일기의 텍스트와 감정 점수를 자유롭게 수정할 수 있다. 어플리케이션은 하루의 일기와 감정점수를 사용자에게 보여준다. 한 달의 감정 변화 추이를 그래프 형식으로 시각화 할 수 있다. 메인 어플리케이션은 React Native를 이용하여 iOS/Android 크로스 플랫폼으로 개발하였다.

4. 시스템 구현 결과

다음과 같이 시스템을 구현하였다. 아래의 그림 4와 그림 5는 어플리케이션 화면을 캡처한 것이다.

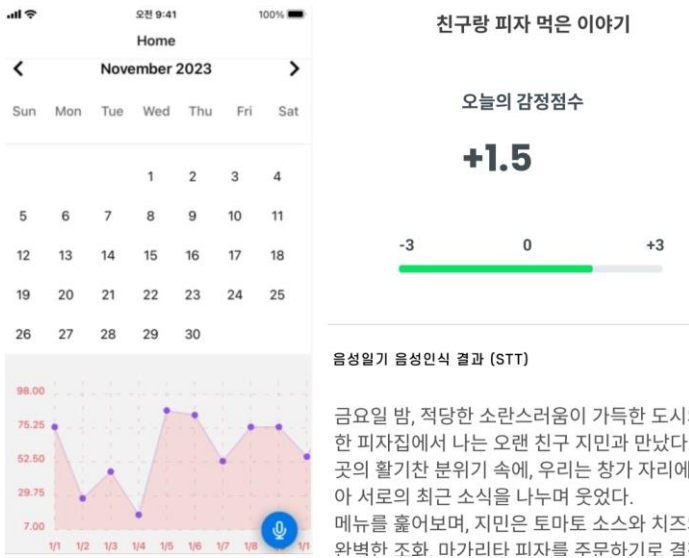


그림 4. 어플리케이션 메인 화면(a) 일일 결과 화면(b)

Fig 4. Application Main Screen(a) Daily Results Screen(b)

그림 4(a)는 어플리케이션의 메인 화면이다. 크게 달력 부분과 해당하는 달의 기분 점수 추이를 나타낸다. 그림 4(b)는 달력에서 해당 날짜 버튼을 눌러 일일 결과 화면으로 진입한 것이다. 일일 결과 화면에는 당일 저장한 일기의 제목, 감정 점수, 내용이 기록된다.

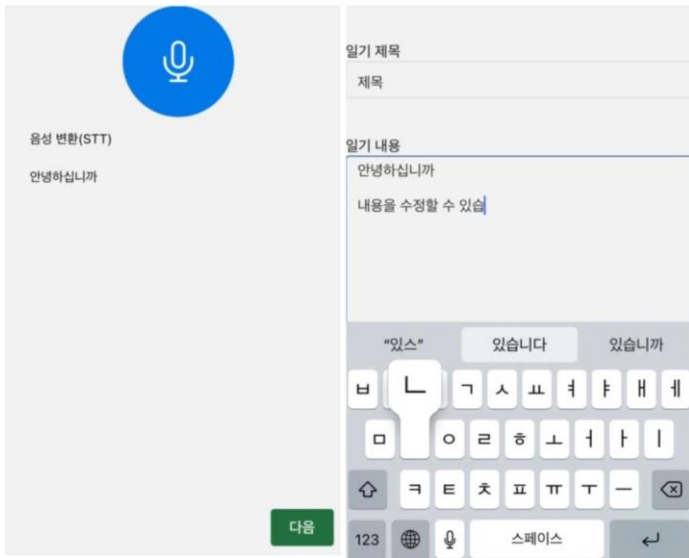


그림 5 녹음 화면(a) 수정 화면(b)

Fig 5. Recording Screen(a) Editing Screen(b)

이은 그림 5(a)에서는 메인 화면의 녹음 버튼을 통해 진입할 수 있는 녹음 화면이다. 녹음 화면에서는 녹음 버튼을 눌러 사용자의 음성을 기록할 수 있고, 기록된 음성은 음성 인식 모듈을 통해 STT 처리되어 텍스트로 표시된다. 그림 5(b)는 녹음 화면에서 다음 버튼을 통해 진입하는 수정 화면이다. 수정 화면에서는 일기의 제목과

STT 처리된 텍스트를 수정할 수 있다. 수정 화면에서 다음 버튼을 누르면 감정 분석 모듈이 일기에 감정 점수를 부여한다. 작업이 종료되면 해당 기록이 저장된 그림 2(b)의 일일 결과 화면으로 진입한다.

5. 결론

본 연구에서는 Speech-To-Text와 Sentiment Analysis 기술을 활용한 감정일기 앱의 개발을 성공적으로 수행하였다. 사용자는 일상 감정을 기록하고 감정 점수를 통해 자신의 감정 상태를 시각적으로 파악할 수 있다. 이를 통해 개인의 감정 변화에 대한 보다 객관적인 자료를 얻을 수 있으며, 기분 기록지 작성에 있어 감정일기 앱을 활용할 수 있다. 이 연구는 감정 관리에 관한 새로운 사용자 경험과 자기 인식을 제공하는데 기여할 것으로 기대된다.

참고 문헌

- [1]한지연,김여진,조화연,이자경,and오유란."MoodChart Bot: AI 챗봇을 활용한 기분기록지 어플리케이션 설계 및 구현." 멀티미디어학회논문지 26.3 (2023): 503-508.
- [2]유희정, 최정은, 김정현, 최승원, 하태현, 하규섭, 차보석, 장재승. (2009). 환자용 기분기록지 개발 및 타당도 연구. 신경정신의학, 48(3), 174-181.

ACKNOWLEDGEMENT

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 중심대학사업의 연구결과로 수행되었음” (2021-0-01082)

¹ 경북대학교 컴퓨터학부

² (주)AtlasLab, VP of Engineering (교신저자)

³ 경북대학교 컴퓨터학부 교수 (교신저자)

쿠버네티스와 아파치 카프카를 이용한 마이크로서비스 아키텍처 구현에 대한 연구

이상희^o, 김현우, 신성현, 정현정, 조동필, 정설영(교신저자)
경북대학교

이동욱(교신저자)
데이터스트림즈

lsh277604@gmail.com^o, hyunu.dev@gmail.com, sin6708k@gmail.com, wgwjh05169@gmail.com,
ehdvlf001@gmail.com, snowflower@knu.ac.kr, dwlee@datastreams.co.kr

Implementation of Microservices Architecture using Kubernetes and Apache Kafka

Sanghee Lee^o, Hyunwoo Kim, Seonghyeon Shin, , Hyeonjeong Jeong, Dongpil Jo
Seolyeong Jeong
Kyungpook National University

Dongwook Lee
DataStreams.Corp

요 약

이 논문은 마이크로서비스 아키텍처의 이해와 구현을 위한 방법을 제시한 쿠버네티스와 아파치 카프카를 사용하여 시스템을 설계하고 구현하였으며, 이를 통해 마이크로서비스 아키텍처의 효율적인 운용과 확장성, 빠른 배포, 서비스 간의 효율적인 통신을 실현할 수 있음을 보여준다. 또한, 마이크로서비스 아키텍처의 데이터 동시성 문제를 해결하기 위해 아파치 카프카를 도입하는 방법을 제시하였다. 이를 통해, 마이크로서비스 아키텍처를 제대로 이해하고 적절히 구현하지 않으면 발생할 수 있는 여러 문제를 예방하고, 마이크로서비스 아키텍처의 중요성을 강조하였다.

1. 서 론

마이크로서비스 아키텍처는 모놀리식 시스템을 작고 독립적인 서비스로 분할하는 아키텍처로, 2035년까지 현재 시장 규모 대비 6 배 이상의 성장이 예상될 정도로 최근 인프라 영역에서 가장 주목 받고 있는 기술이다.

기존의 모놀리식 아키텍처에서는 서비스 규모가 커짐에 따라 유지 보수가 어려우며, 확장성이 유연하지 않다. 또한 일부 장애가 서비스 전체로 번질 수 있다. 그 해결 방법으로 마이크로서비스 아키텍처를 채택할 수 있다. 그러나 이 아키텍처 채택 시, 별도로 배포된 마이크로서비스가 개별의 데이터베이스를 가지고 있어 이들 간의 데이터 동시성에 문제가 발생할 수 있다.

본 논문에서는 쿠버네티스(Kubernetes)를 활용하여 서비스 개별 배포 및 확장의 장점을 극대화하면서, 각 마이크로서비스 간의 데이터 실시간 동기화를 위해, 아파치 카프카(Apache Kafka)를 활용하여 마이크로서비스 아키텍처를 구현하였다.

2. 관련 연구

모놀리식 아키텍처는 단일한 대형 어플리케이션을 하나의 코드베이스로 구성하여, 모든 기능과 로직이 하나의 코드베이스 안에서 구현된다. 해당 구조는 구현 및 배포가 비교적 간단하고 네트워크 지연시간을 고려하지 않아 빠른 응답 시간을 보장한다. 그러나 이러한 구조는 어플리케이션의 크기가 커짐에 따라 구동시간이 늘어나고, 빌드 및 배포시간이 길어지게 된다. 수정사항이 생긴 경우에 다시 빌드하고 배포해야 하기 때문에 규모가 커지거나 유연성이 필요한 시스템에서는 적합하지 않다. 그에 반해 마이크로서비스 아키텍처는 다수의 독립적인 서비스로 구성되며 각 서비스는 서로 다른 기능을 수행한다. 이러한 구조는 서버 통신장애 등의 문제가 발생하더라도 그 영향이 해당 서비스로 제한되어 전체 시스템에 영향을 미치지 않아, 높은 가용성을 보장한다. 또한 새로 추가되거나 수정사항이 있는 마이크로 서비스만 빠르게 빌드 및 배포가 가능하다. [1]

하지만 각 서비스의 개발 환경이 분리되어 있어, 서비스 간 인터페이스가 변경될 경우 해당 변경사항이 다른 서비스에도 영향을 줄 수 있다. 그래서 마이크로서비스를 구현할 때에는 컨테이너 환경이 적합하다. 컨테이너 패키지에는 서비스를 개발하고 어플리케이션을 구동하기 위한 모든 정보와 요소들이 포함되어 있어, 마이크로서비스 구축이 용이하며, 개발 환경의 차이로 인한 오류를 방지할 수 있다.

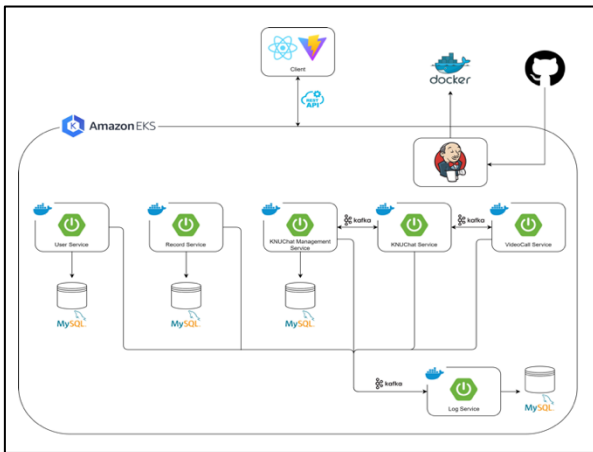
쿠버네티스는 이러한 컨테이너를 효과적으로 배포하고 관리하는 역할을 수행한다. 컨테이너 기반의 마이크로서비스 아키텍처를 구성하는데 쿠버네티스를 활용하면, 마이크로서비스 Deployment 의 Pod 수를 확장하고 클러스터의 총 용량을 확장하여 서비스 전체를 확장하는 등의 이점을 얻을 수 있다. 또한, 쿠버네티스를 이용하면 실행 중인 마이크로서비스를 모니터링하여 업데이트, 개선사항 등을 지속적으로 확인하고 관리할 수 있다. [2]

마이크로서비스 아키텍처의 복잡성은 데이터 동시성 관리에 영향을 미친다. 각 서비스는 자체의 데이터베이스를 가지고 있기 때문에, 서비스 간 데이터의 동시성을 유지해야 한다. 하나의 서비스에서 데이터가 수정되고 이 변경사항이 다른 서비스에 즉시 반영되어야 하는 경우, 독립적으로 운영되는 서비스 간의 데이터 동기화를 보장하는 것은 쉽지 않다. 이로 인해 데이터의 일관성이 손상될 수 있고, 이는 전체 시스템의 안정성에 영향을 미칠 수 있다. 이러한 문제를 해결하기 위해서는 별도의 데이터 관리 전략을 수립해야 한다. [3].

이를 해결하기 위한 데이터 관리 전략을 수립하기 위해 아파치 카프카를 선택했다. 아파치 카프카는 대량의 실시간 데이터 스트림 처리를 가능하게 하는 플랫폼으로, 마이크로서비스 아키텍처에서 데이터 동시성 문제를 해결하는 데 활용한다. 아파치 카프카는 데이터의 신뢰성과 일관성을 보장하고, 고가용성과 내결함성을 지원하며, 분산 환경에서의 확장성이 뛰어나다. 이를 통해 마이크로서비스 아키텍처의 데이터 동시성 문제를 해결하고 전체 시스템의 안정성과 성능을 향상시킬 수 있다. [4]

3. 시스템 상세 설계

3. 1. 시스템 구성



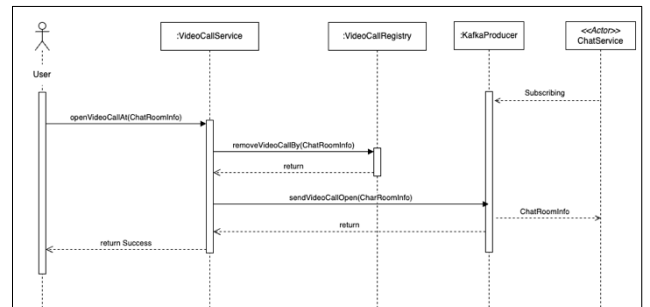
[그림 1] 시스템 구성도

본 플랫폼은 쿠버네티스 관리형 서비스인 Amazon Elastic Kubernetes Service (EKS)를 활용해 클라우드 환경에서 쿠버네티스 클러스터를 구축하였다. 사용자는 React application 을 통해 마이크로서비스로부터 서비스를 제공받는다.

마이크로서비스는 플랫폼이 제공하는 서비스를 단위로 나누어 각각 사용자 기능, 이력 기능, 채팅 정보 관리 기능, 소켓 기능, 영상통화 기능, 로그 데이터 관리 기능을 맡으며 필요한 경우 개별적인 DBMS 를 가진다. 이때 마이크로서비스 간 데이터 동기화는 pub/sub 모델을 가지는 아파치 카프카의 메시지 큐를 기반으로 이루어진다.

3. 2. 마이크로서비스 간 데이터 동기화

마이크로서비스는 다른 마이크로서비스의 이벤트 발생을 감지해야 하거나 데이터가 필요할 수 있다. 이를 위해 마이크로서비스마다 아파치 카프카 Producer 혹은 Consumer 를 가지고, 관련된 토픽에 메시지를 발행하고 이를 구독하여 데이터를 동기화한다.



[그림 2] 비디오채팅 시작 시퀀스 다이어그램

그림 2는 마이크로서비스 간 데이터 동기화를 설명하기 위한 예시로, 사용자가 영상통화를 시작할 때의 시퀀스 다이어그램이다. 본 플랫폼은 영상통화가 시작될 경우 채팅방에 관련 정보를 표시한다. 이를 위해 VideoCall 서비스로부터 KNUChat 서비스에 정보를 전달할 때, 아파치 카프카의 메시지 큐를 활용하여 두 서비스를 연동한다.

4. 시스템 구현 결과

이 장에서는, 로컬 환경에서 개발된 마이크로서비스가 Jenkins 를 활용하여 EKS, Elastic Container Registry (ECR) 및 Managed Streaming for Apache Kafka (MSK)를 통해 배포되는 과정을 상세히 서술하고자 한다.

4. 1. 로컬 환경에서의 마이크로서비스 개발

첫 번째 단계로, 로컬 환경에서 마이크로서비스를 개발하는 과정을 진행한다. 이 단계에서는 각각의 서비스 기능을 정의하고 구현하며, 개발된 서비스는 테스트를 통해 기능이 정상적으로 작동하는지 검증한다.

4. 2. Jenkins 를 활용한 CI/CD 파이프라인 구축



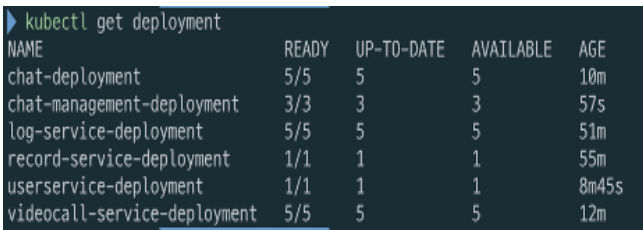
[그림 3] Jenkins 파이프라인 스크립트

마이크로서비스가 로컬 환경에서 개발 완료된 후, Jenkins 를 이용하여 지속적 통합(Continuous Integration)과 지속적 배포(Continuous Deployment)를 위한 파이프라인을 구축한다. 이 파이프라인은 코드 변경을 자동으로 감지하고, 해당 변경사항을 테스트한 후, 마지막으로 AWS 의 ECR 에 서비스를 배포한다.

4. 3. AWS ECR 을 이용한 마이크로서비스 배포

Jenkins 를 통해 실행된 CI/CD 파이프라인은 개발된 마이크로서비스를 AWS 의 ECR 에 배포한다. ECR 은 Docker 이미지를 저장하는 서비스로, 이를 활용하여 마이크로서비스의 컨테이너 이미지를 저장한다.

4. 4. AWS EKS 를 활용한 마이크로서비스 배포



[그림 4] EKS 에 배포된 Deployments

ECR 에 저장된 마이크로서비스의 이미지는, 이후 EKS 를 통해 Kubernetes 클러스터 위에 배포된다. EKS 는 관리형 Kubernetes 서비스로, 이를 활용하여 마이크로서비스의 확장성과 안정성을 보장한다.

4. 5. AWS MSK 를 활용한 마이크로서비스 간 통신

마지막으로, 배포된 마이크로서비스 간의 통신은 AWS 의 MSK 를 활용한다. MSK 는 Apache Kafka 를 관리하는 서비스로, 이를 활용하여 마이크로서비스 간의 메시지 기반 통신을 구현한다.

로컬 환경에서 개발된 마이크로서비스가 AWS 의 여러 서비스를 통해 배포되는 이러한 과정은 마이크로서비스 아키텍처의 효율적인 운용을 가능하게한다. 이 과정을 통해 마이크로서비스의 확장성을 높이며, 빠른 배포를 실현하고, 서비스 간의 효율적인 통신을 구현할 수 있다.

5. 결론

본 논문은 마이크로서비스 아키텍처에 대해 이해하고, 이를 효율적으로 구현하기 위해 쿠버네티스와 아파치 카프카를 활용한 시스템 설계 및 구현에 대해 다루었다.

3 장 시스템 설계 부분에서는 화상통화가 시작할 때를 예로 들어서 시퀀스 다이어그램을 그려서 아파치 카프카로 마이크로서비스 간의 데이터 동기화가 어떻게 진행되는지 설명하였다. 4 장 시스템 구현 부분에서는 AWS EKS 를 통해 컨테이너 오케스트레이션을 구현하고, AWS MSK 를 통해 마이크로서비스 간 데이터 동기화 문제를 해결하였다. 더불어 로컬 환경에서 개발한 마이크로서비스를 CI/CD 파이프라인을 통해 AWS ECR 에 빠르게 배포하는 과정을 기술하였다.

위 과정을 통해 마이크로서비스 아키텍처의 장점인 확장성과 빠른 배포를 유지하면서, 데이터 일관성 문제를 해결하여 시스템의 효율적인 운영을 가능하게 하였다. 따라서 본 논문은 쿠버네티스와 아파치 카프카를 이용하여 마이크로서비스 아키텍처를 구현하고자 하는 개발자들에게 유용한 정보를 제공할 것이다.

6. 참고 문헌

- [1] Amazon Web Services, <https://aws.amazon.com/ko/microservices/>
- [2] The Kubernetes Authors, 2023, <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes>
- [3] 고원영, 김진호, and 김영한, "웹 어플리케이션을 이용한 쿠버네티스 기반 마이크로서비스 아키텍처 실증 연구," in 한국통신학회 학술대회논문집, 2020, pp. 571-572.
- [4] 신정은, 권재환, and 김명호, "쿠버네티스 환경에서 웹 기반의 마이크로 서비스 배포 시스템," in 한국컴퓨터정보학회 학술발표논문집, 2020, pp. 45-48.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

데이터 증강 학습 이용한 딥러닝 기반 실시간 전동 킥보드 검지 시스템

김도현^{01‡}, 김민재^{1‡}, 도안탄히엔^{1‡}, 정지형^{1‡}, 김현우², 정설영(교신저자)

¹경북대학교 IT 대학 컴퓨터학부, ²사로리스

e-mail: goat4628@naver.com, ourtrip@knu.ac.kr, thanhchien260701@gmail.com,
wlgud3936@gmail.com, maker@saloris.world, snowflower@knu.ac.kr

Implementation of a Deep Learning based Realtime electric kickboard Detection System using a Data Augmentation

Do-Hyun Kim^{01‡}, Min-Jae Kim^{1‡}, Thanh-Hien Doan^{1‡}, Ji-Hyeong Jung^{1‡}, Hyun-Woo
Kim², Seol-Young Jeong(Corresponding author)

¹School of Computer Science and Engineering, Kyungpook National University

²SALORIS

‡These authors contributed equally to this work

요 약

공유 서비스의 등장으로 전동 킥보드의 수요가 증가함에 따라 전동 킥보드와 관련된 다양한 사건사고들이 늘어나고 있어 그에 대한 대비책을 필요로 한다. 지금까지 정부 차원에서 이러한 사고들을 줄이기 위한 다양한 시도가 있었지만 인력 및 비용 문제로 인해 원활히 이루어지지 않았다. 본 논문에서는 YOLOv8 알고리즘으로 학습시킨 인공지능 모델을 사용해 촬영한 이미지 및 영상에서 탑승자, 헬멧, 킥보드를 검지하여 헬멧 미착용자 및 2인 탑승자를 판별한다. 또한, 다양한 킥보드들의 색을 확인하여 공유 서비스 킥보드 회사를 알아낸다. 이를 위해 색에 따른 킥보드 검지의 정확도를 유의미한 수준까지 끌어올리기 위해 다양한 증강기법들을 비교 및 분석하여 가장 적합한 증강기법을 찾아낸다.

1. 서 론

개인형 이동장치(PM, Personal Mobility)인 전동 킥보드는 공유 서비스가 자리 잡은 후 지금까지 20개가 넘는 회사의 공유 킥보드 서비스가 이루어졌고 지금도 10개가 넘는 종류의 공유 전동 킥보드가 운영되고 있다. 이렇게 공유 서비스로서 활용도와 IT 친화적 시장 환경에 기반해 전동 킥보드 사용자 수는 급증하였다. 하지만 동시에 전동 킥보드의 위험성 또한 증가하였는데, 교통사고분석시스템 통계[1]에 따르면 2018년 사고 건수는 225건에 불과했으나, 매년 큰 폭으로 증가해 2020년 기준 897건, 2022년 기준 2386건으로 약 10배 이상이 증가하였다.

이를 해결하기 위해 2021년 개정 도로교통법에서 개인형 이동장치(PM)에 대한 범칙금을 강하게 책정하고 인명피해 사고에 대해서는 단속과 처벌을 강화하였다. 여러 노력에도 불구하고 실패한 이유들로 단속 인력 및 비용 문제, 이용자들의 낮은 안전의식 및 준법의식 등을 꼽을 수 있다. 그 중 비용 문제가 크게 자리 잡고 있는데 상당한 수의 전동 킥보드의 단속에 필요한

인력이 부족하고, 전동 킥보드를 관리 및 단속하는 비용 또한 만만치 않다는 것이다. 전동 킥보드는 다른 이동수단들에 비해 연구가 충분히 이루어지지 않았고 데이터의 부족으로 인해 적절한 모델이 구축하는 데까지 이어지지 못하였다.

본 논문에서는 경북대학교 내에서 공유 서비스가 이루어지고 있는 5개 회사(Gcooter, XingXing, Swing, Deer, Beam)의 킥보드를 대상으로 하여 인공지능 모델을 만들어 소개한다. YOLOv8 알고리즘을 활용해 헬멧, 사람, 킥보드를 탐지하여 PM 탑승 규정인 헬멧 착용 여부와 2인 이상 탑승 여부를 판별한다. 킥보드의 색상을 확인하여 그 킥보드가 어느 공유 서비스 회사의 킥보드인지도 판별한다. 부족한 데이터셋을 극복하기 위해 데이터 증강기법을 적극 활용하고 특히 구분 기준이 되는 색상과 관련된 증강기법을 분석하여 어떻게 효과적으로 정확도를 올릴 수 있는 지 알아볼 것이다.

본 논문의 구성은 다음과 같다. 2절에서는 관련 연구와 해당 딥러닝 모델의 원리에 대해 설명한다.

3절에서는 해당 전동 킥보드 탐지 시스템과 구조에 대해 설명하고 정확도를 높이기 위해 사용된 증강기법을 분석한다. 4절에서는 앞으로 해결해야 할 과제와 이 모델이 실제로 어떻게 활용될 수 있을지에 대해 언급하며 마무리한다.

2. 관련연구

2.1 딥러닝 기반 킥보드 탐지

기존의 영상기반 객체 탐지는 유사한 색상에 대한 오류에 취약하기 때문에 한계를 갖고 있다. 이를 해결하는 방법으로 이미지에서 관심 객체를 배경과 구분해 Bounding Box로 특정 객체를 딥러닝 CNN으로 학습하고 객체를 분류하는 기술이 많이 연구되고 있다[2]. 딥러닝 기반의 객체 탐지 모델은 One-Stage 모델과 Two-Stage 모델로 구분할 수 있다. 이 중 Two-Stage의 Detector는 정확도는 높지만 예측속도가 낮아 정확도는 떨어지지만 예측속도가 높은 One-Stage Detector인 YOLO 알고리즘[3]을 사용하기로 하였다.

2.2 학습 데이터 증강 기법

본 시스템에서는 정확도가 최소 80%를 넘어야 유효하다고 판단하였기 때문에 mAP[4]를 향상시킬 필요가 있었다[5].

일반적으로 데이터 학습에 필요한 데이터를 빅데이터 등을 통해 확보하지만 전동 킥보드 데이터를 충분히 확보하기 어렵고 이처럼 한정된 데이터로 학습할 시 과적합이 발생할 수 있다. 따라서 데이터 증강을 통해 이러한 과적합 현상을 방지하고자 하였다[6].

보통 이미지 데이터 증강에 사용하는 기법은 flip, exposure, crop 등의 기법이 사용된다. [7], [8]에서는 데이터 증강기법 중 한가지 기법만을 적용하여 데이터를 증강하였다.

이처럼 한 가지 기법만을 적용하여 데이터를 증강하면 일부 특징만 강조되는 문제가 있어 본 논문에서는 다양한 증강기법을 선정하여 킥보드 탐지에 효과적인 증강 데이터를 생성하였다.

하지만 색 공간을 변환하는 증강에는 여러 단점이 존재하고, 중요한 색상 정보가 삭제될 수도 있다[9]. 이에 극적인 변화 대신 증강기법 별로 30% 이내의 정도를 선택해 증강했다.

2.3 YOLOv8

기존의 Object Detector는 Two-Stage Detector로 Regional Proposal과 Classification이 순차적으로 이루어졌으나, YOLO 알고리즘은 One-Stage Detector로 두가지 과정이 동시에 이루어져 실시간 객체 탐지가 가능하다.

본 시스템에서는 이러한 YOLOv8 알고리즘을 이용해 빠른 처리속도로 전동 킥보드의 탐지가 가능했다.

[10]에 따르면 YOLOv8 모델이 YOLOv5모델보다 성능을 보이는 것을 알 수 있다. YOLOv8[11]은 신경망의 크기에 따라 n, s, m, l, x로 나누어지고 x로 갈수록 정확도는 높아지지만 속도는 느려진다는 단점이 있다.

3. 시스템 설계

3.1 실시간 킥보드 위반 금지 시스템

본 연구에서는 헬멧 미착용, 2인 탑승자 등 킥보드 위반 사항을 실시간으로 금지하는 시스템을 구현한다. 도로에서 설치된 카메라를 통해 킥보드를 촬영하고, YOLOv8를 이용해 위반 행위가 감지 되면, 해당 장소, 시간, 이미지를 데이터베이스에 저장한다. 데이터를 효과적으로 분류하기 위해 시스템은 킥보드 제조사도 분류하고, 제조사에 대한 정보도 데이터베이스에 저장한다.



Fig. 1. Electric Scooter Detection System Architecture

그림 1. 전동 킥보드 금지 시스템 구성도

3.2 킥보드 금지 모델 학습

가. 데이터 수집

킥보드 학습 이미지 데이터셋은 roboflow의 킥보드 데이터셋을 통해 수집했다. 또한 여러 제조사의 킥보드에 대한 약 20개의 영상을 직접 촬영한 후 영상에서 학습 이미지를 추출했다. 총 1004장의 원본 사진이 학습 데이터로 준비되었다.

나. 데이터 증강

기본 원본 데이터셋은 1004장을 바탕으로 학습 정확도를 높이기 위해서 여러 데이터 증강 기법을 실험하고 증강 기법들 간의 성능을 비교하고자 한다. 데이터셋의 train : valid : test의 비율은 6:2:2로 설정했다.

본 논문에서는 saturation, exposure, brightness 3개의 데이터 증강 기법을 사용한다. 적절한 수치를 선택하기 위해 5% 간격으로 최대 30%까지 늘려가며 증강한 후 mAP가 가장 높게 나타나는 값을 선택했다.

첫 번째 기법은 saturation 기법이다. Train set 604장의 데이터를 증강하였고 -20~20% 값으로 조절하였다.



Fig. 2. Kickboard image augmentation using saturation.
그림 2. Saturation을 이용한 킥보드 이미지 증강

두 번째 기법은 Exposure다. Train set 604장의 데이터를 증강하였고 -30~30% 값으로 조절하였다.



Fig. 3. Kickboard image augmentation using exposure.
그림 3. Exposure을 이용한 킥보드 이미지 증강

세 번째 기법은 Brightness이다. Train set 604장의 데이터를 증강하였고 -10~10% 값으로 조절하였다.



Fig. 4. Adjust kickboard image brightness.
그림 4. 킥보드 이미지 밝기 조절

다. 학습

본 논문에서는 경북대학교에서 공유 서비스가 이루어지고 있는 5개 회사(Gcooter, XingXing, Swing, Deer, Beam)의 킥보드를 대상으로 하여 딥러닝 모델을 설계한다. 색깔 기준으로 킥보드 제조사를 분류해서 5개 클래스(beam, gcooter, swing, yellow, other kickboard)로 나뉜다. 그 중 XingXing과 Deer의 두 회사 킥보드가 모두 노란색을 띠고 있어서 yellow 클래스에 해당된다. 위 5개 회사에 속하지 않는 개인 킥보드는 other kickboard 클래스로 구분한다.

학습 환경은 구글 Colab에서 제공하는 T4 GPU로 진행하였다.

라. 학습 모델 측정 결과

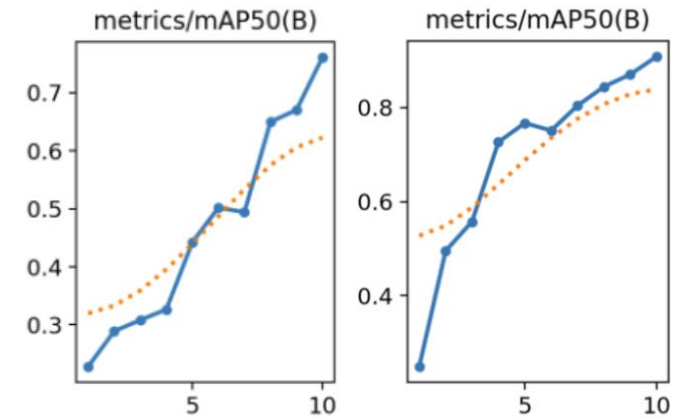
학습결과는 여러 증강기법들을 적용하는 데이터의 mAP(Mean Average Precision)을 비교하였다. 측정 결과는 다음 표1에 표시된다.

Table 1. Comparison of measured mAP among augmentation techniques

표1: 증강기법들의 측정 mAP 비교

	all	beam	gcooter	swing	yellow	Others
Original	0.761	0.876	0.707	0.699	0.778	0.499
Flip	0.857	0.883	0.924	0.734	0.938	0.741
Saturation	0.884	0.92	0.935	0.822	0.944	0.821
Exposure	0.892	0.942	0.951	0.813	0.961	0.82
Brightness	0.892	0.942	0.931	0.839	0.925	0.854
All-aug	0.907	0.98	0.962	0.83	0.931	0.865

그림 5(a)는 증강기법을 적용하지 않은 데이터셋의 mAP를 그림 5(b)는 세가지 증강기법을 적용한 mAP 그래프를 나타낸다. 전체적인 학습 모델 측정 결과는 표 1에서와 같이 Flip 기법에 비해 색에 영향을 주는 증강기법들이 mAP가 높게 나타났다. 또한 세가지 데이터 증강을 모두 적용한 모델의 mAP는 0.907로 이전의 0.761보다 0.146 증가하여 가장 큰 효과가 있음을 알 수 있다.



(a) no augmentation (b) all augmentations selected

Fig. 5. mAP measurement value

그림 5. mAP 측정값

마. 학습 모델 적용 결과

그림 6은 세 가지 증강기법을 적용하여 학습시킨 yolov8 모델로 테스트한 영상이다.

입력된 영상을 통해 킥보드가 검지되고 검지 당시의 날짜, 시간, Class 정보가 기록되었다.

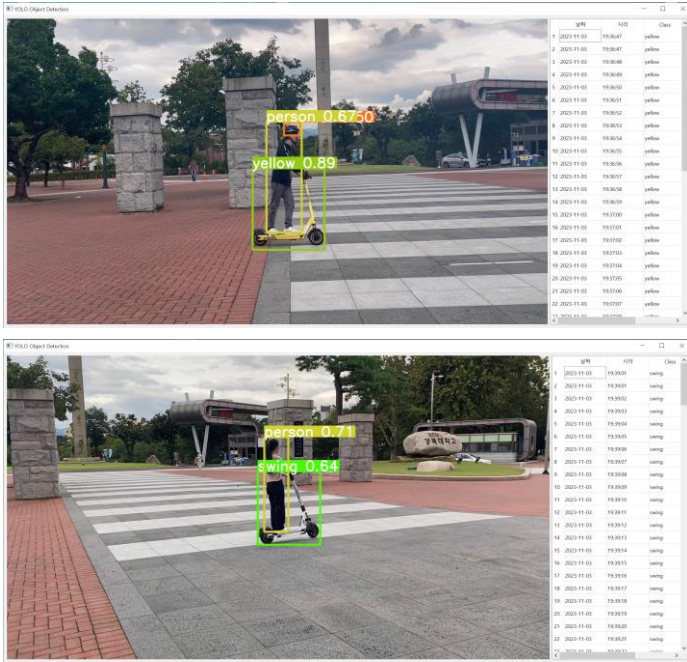


Fig. 6. Test Image for electric kickboard detection
그림 6. 전동 킥보드 검지 테스트 이미지

4. 결론

본 논문에서는 딥러닝에 사용된 YOLOv8 기술에 대해 알아보고 형태가 유사하지만 색상만 차이가 있는 공유 서비스 전동 킥보드들에 대해 색상으로 구분하는 방법에 대해 설명하였다. 데이터셋 부족과 같은 기존의 한계점을 극복하기 위해 여러 증강기법들 중 색상 및 밝기와 관련한 증강기법에 집중하여 소개하였다.

이러한 YOLOv8 기반의 객체 인식 모델과 데이터 증강 기법을 적용하여 일차적으로 단속하는 과정의 일부분을 자동화하여 인력과 단속 비용의 감소를 이끌어 낼 수 있고, 더 나아가 올바른 전동 킥보드 탑승법을 인지하는 공적인 효과를 이끌어 낼 수 있을 것이다. 이후에 사람들로 하여금 국민신문고에 신고하는 정책이 만들어졌을 때 활용되어 1차 거름망의 역할을 할 수 있을 것이다. 혹은, 길거리에 방치되고 있는 전동 킥보드들이나 전동 킥보드 전용 주차장에 있는 많은 킥보드들을 관리함에 있어서 비용 절감의 효과를 누릴 수 있을 것이다.

다만, 본 논문의 시스템은 주간 환경을 전제로 하여, 야간에서는 활용되지 못하는 점과 색이 비슷한 경우 구분에 어려움을 겪는 한계점을 가지고 있다.

향후에는 LIDAR와 같은 다양한 센서 데이터와 영상 데이터를 효과적으로 융합하여 킥보드 단속 판단과 킥보드 분류의 정확도를 더욱 향상시키기 위한 연구를 수행할 예정이다.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2021-0-01082)

"이 연구는 과학기술정보통신부 및 정보통신산업진흥원의 "지역산업 SW인재양성 기반조성 사업" 지원으로 수행되었음."

참고문헌

- [1] 김현중, "전동킥보드 공유서비스 안전실태조사", <https://www.kca.go.kr/smartconsumer/sub.do?menukey=7301&mode=view&no=1003197854>
- [2] Young-Jin Kim, Eun-Gyung Kim, "Image based Fire Detection using Convolutional Neural Network," Journal of the Korea Institute of Information and Communication Engineering, Vol.20, No.9, pp. 1649-1656, 2016.
- [3] J. Redmon, et. al., "You Only Look Once: Unified, Real-Time Object Detection," in Proc. CVPR, 2016. DOI: 10.48550/arXiv.1506.02640
- [4] Jonathan Hui, "mAP (mean Average Precision) for Object Detection", <https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173>
- [5] 경찰청, "무인교통단속장비 경찰규격서(경찰-6310-98-0001-차)", https://www.police.go.kr/user/bbs/BD_selectBbs.do?q_bbsCode=1001&q_bbscttSn=20220810140831836
- [6] Connor Shorten, Taghi M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning", Journal of Big Data, 6, Article number: 60 (2019), <https://journalofbigdata.springeropen.com/counter/pdf/10.1186/s40537-019-0197-0.pdf>
- [7] You-min Na, Dong-hwan Hyun, Do-hyun Park, Se-hyun Hwang, Soo-hong Lee, "AI Fire Detection & Notification System," Journal of the Korea Society of Computer and Information, Vol.25, No.12, pp63-71, 2020. DOI: 10.9708/jksci.2020.25.12.063
- [8] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V., "Autoaugment: Learning augmentation policies from data," arXiv preprint arXiv: 1805.09501., 2018.
- [9] Connor Shorten, Taghi M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning", Journal of Big Data, 6, Article number: 60, pp. 10 (2019), <https://journalofbigdata.springeropen.com/counter/pdf/10.1186/s40537-019-0197-0.pdf>
- [10] 이태희, 박천수, "YOLOv8을 이용한 실시간 화재 검출 방법", 반도체디스플레이기술학회지, 제22권, 제2호, 2023년 6월
- [11] Glenn Jocher, Mike Tune, <https://github.com/ultralytics/ultralytics>

연합 학습에서의 이질성 분류 및 그 영향에 관한 최신 연구 동향

멩기스투 테스파후넝 민우엘렛[○], 김태운^{*}

부산대학교

tesfahunegn9@pusan.ac.kr, taewoon@pusan.ac.kr

State-of-the-Art Research Trend on Taxonomy and Impact of Heterogeneity in Federated Learning

Mengistu Tesfahunegn Minwuyelet[○], Taewoon Kim^{*}

Pusan National University, Republic of Korea

Abstract

Federated Learning (FL) has emerged as a promising approach for collaborative, secure machine learning in distributive systems. However, the inherent heterogeneity in FL poses both challenges and opportunities. This paper aims to explore various forms of heterogeneity encountered in FL, including statistical, devices, network, communication, and model heterogeneity. We discussed their impact on FL performance and investigate protentional strategies to address these challenges. By highlighting the importance of understanding and managing heterogeneity, this paper provides insights into the design and implementation of an efficient and robust FL system.

1. Introduction

Federated Learning (FL) is a machine learning technique that allows multiple edge devices or nodes to collaboratively train a shared model while keeping the data decentralized. Instead of sending raw data to a central server for training as shown in Fig. 1. Inevitable heterogeneity in FL refers to the differences and diversity among participants or devices that collaborate to train a shared model without sharing their local data. FL faces the challenge of training on diverse data sets, devices, and networks that are beyond the control of the centralized FL server.

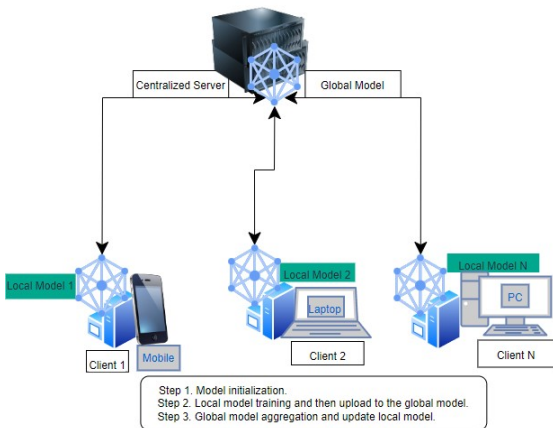


Figure 1. Overview of federated learning

This heterogeneity can cause the model to diverge, making the learning process ineffective. Heterogeneity in FL can stem from factors such as imbalanced data distribution, different hardware and network characteristics of client devices, unstable network connectivity, and limited device resources. In this paper, as shown in Fig. 2 we present a taxonomy of the heterogeneity in FL by classifying it into five types: statistical, device, architectural, model, network, and communication heterogeneity [1],[2].

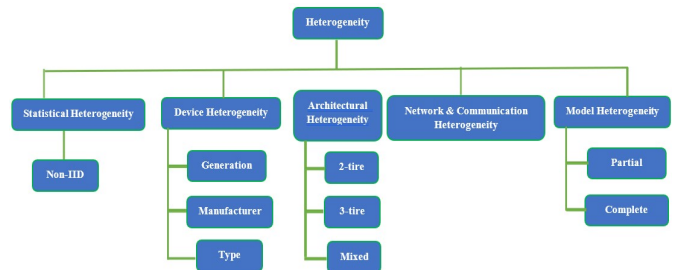


Figure 2. Proposed taxonomy of the heterogeneity in federated learning

2. Statistical Heterogeneity

Statistical heterogeneity in FL is a challenge that occurs when the data distributions of different clients are not independent and identically distributed or non-

IID for short. This can lead to bias in the global model or hinder convergence. Therefore, it is crucial for FL methods to tackle the issue of statistical heterogeneity and develop robust and efficient ways to aggregate local models or gradients from various clients [3]–[5]. In FL, the data heterogeneity between two clients, i and j , can be measured by comparing their respective local data distributions, P_i and P_j . When conducting a supervised task in FL, a client is chosen randomly, and its local data distribution, $P_i(x,y)$, is used to extract feature–label pairs from (x,y) [6]–[9].

From distribution perspective statistical heterogeneity is classified into the following classes. Label distribution skew refers to the difference in label $P_i(y)$ distributions between different clients, while the $P_i(x|y)$ situation remains the same. This can occur when the feature distribution is consistent across clients, but the label distribution may differ. Label preference skew implies that even when the feature distribution is consistent across clients, i.e., $P_i(x)=P_j(x)$, the label distribution may be different for different clients, i.e., $P_i(y|x) \neq P_j(y|x)$. Feature distribution skew occurs when the feature $P_i(x)$ distribution varies from customer to customer, but the distribution of $P_i(y|x)$ remains the same. Feature condition skew occurs when feature distributions may vary across clients i.e., $P_i(x|y) \neq P_j(x|y)$, even if $P_i(y)=P_j(y)$, mainly related to vertical FL which commonly performed in medical applications for example. Label noise skew represents the proportion of noisy labels contained, with varying degrees of label noise due to differences in expertise and input costs. Sample noise skew refers to the varying levels of sample noise in each client's private data, making communication between clients uncertain and complex. Quantity skew refers to the large difference in the quantity of different client data $P_i(x,y)$, resulting in different data amounts.

3. Device Heterogeneity

In a decentralized FL setting, client performance can vary due to differences in hardware, network connectivity, and power. Client devices in federated networks have varying computational capabilities, network connectivity (e.g., 3G, 4G, 5G and Wi-Fi), and power (i.e., battery level). This variability is due to differences in hardware (e.g., CPU and memory), network connectivity, and power. The training phase of FL can involve multiple devices from different generations, manufacturers, and types, resulting in a

network of heterogeneous devices with varying computational abilities, memory sizes, and battery capacities. As a result, the training period can vary significantly across clients, and it is not effective to treat all participants equally. To achieve optimal training results, FL needs to consider the heterogeneous hardware configurations.

This heterogeneity arises from the generation of a device as well, which significantly impacts its performance in FL tasks. For instance, newer devices have advanced hardware components, while older devices may have limited resources and outdated hardware. Manufacturers also influence device heterogeneity, with varying specifications, architectures, and optimization techniques. This can affect how devices handle tasks and interact with the system. Additionally, device types, such as smartphones, tablets, laptops, IoT devices, edge servers, or specialized hardware, also affect heterogeneity [10]–[14].

4. Architectural Heterogeneity

Architectural heterogeneity is the use of different types of hardware, software, or network components in a system. It can have advantages such as increased performance, scalability, reliability, and flexibility, but also challenges such as increased complexity, compatibility issues, and security risks. One way to classify architectural heterogeneity is based on the number of tiers or layers in a system such as 2–tier, 3–tier, and mixed architecture. A tier is a physical platform where a layer of software operates, serving as a logical component that performs a specific system function.

A two–tier system may consist of user devices and the central model aggregator, where a three–tier model may have a small number of local model aggregators placed between the user devices and the central aggregator. Also, there can be a mixed architecture that combines different layer architectures. A 2–tier architecture is simple and easy to implement, but it has limitations such as poor scalability, security, and maintainability. A 3–tier architecture is more complex and costly than a 2–tier architecture, but it has advantages such as modularity, flexibility, and maintainability. A mixed architecture is the most complex and challenging type of architecture, but it can also provide the most benefits in terms of performance, scalability, reliability, flexibility, and security.

5. Network and Communication Heterogeneity

FL is a decentralized machine learning approach that uses networked devices to train models while keeping data localized, addressing concerns about privacy and data ownership. This section focuses on network resource heterogeneity, which refers to the diversity and variation in network components and capabilities. With advancements in access network communication technology, communication resource allocation has become more diverse. In wireless communications, the resources of heterogeneous access networks include frequency, timing slot, space resource, and radio power. Communication cost can vary depending on the communication protocol, the network bandwidth, the network latency, or the network reliability. Communication cost can affect the performance, scalability, robustness, and efficiency of the global model, as well as the energy consumption and privacy of the clients [15].

6. Model Heterogeneity

FL requires each client to use a local model with the same architecture, aggregating network parameters into a global model. In IoT applications, clients may design unique local models due to individual requirements and hardware constraints. Model heterogeneity requires learning knowledge without sharing private data or model structure information. Transferring knowledge between heterogeneous clients is challenging, categorized into partial and complete heterogeneity. Partial heterogeneity occurs in real-life scenarios where some clients use the same model structure while others do not. In this paper, such a federated system is considered partial model heterogeneity. A FL model is trained for each isomorphic client subset. Common techniques like weighted averaging can aggregate intra-cluster models, but knowledge distillation is needed for inter-cluster models. Complete heterogeneity is a type of partial heterogeneity where structures of participant models differ in a FL framework. This results in unique models for each client, potentially leading to high learning costs and low communication efficiency [6].

7. Conclusion

In this paper, we have examined the various manifestation of heterogeneity in FL. We have

explored statistical, device, network, communication and model heterogeneity and discussed their impact on FL performance. By understanding and addressing these forms of heterogeneity, researchers and practitioners can enhance FL systems efficiency, fairness, and convergence. We conclude by emphasizing the need for further research in managing heterogeneity and promoting the development of robust and scalable solutions.

Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-RS-2023-00260098) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

- [1] Q. Yang et al., "Federated Machine Learning: Concept and Applications," Feb. 2019, [Online]. Available: <http://arxiv.org/abs/1902.04885>
- [2] M. Aledhari et al., "Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Access*, vol. 8, pp. 140699–140725, 2020.
- [3] J. Zhang et al., "Adaptive Federated Learning on Non-IID Data With Resource Constraint," *IEEE Transactions on Computers*, vol. 71, no. 7, pp. 1655–1667, 2022.
- [4] A. Ahmad et al., "Robust Federated Learning under Statistical Heterogeneity via Hessian-Weighted Aggregation," *Machine Learning*, vol. 112, no. 2, 2023.
- [5] Y. Li et al., "FedH2L: Federated Learning with Model and Statistical Heterogeneity," Jan. 2021, [Online]. Available: <http://arxiv.org/abs/2101.11296>
- [6] M. Ye et al., "Heterogeneous Federated Learning: State-of-the-art and Research Challenges," Jul. 2023, [Online]. Available: <http://arxiv.org/abs/2307.10616>
- [7] D. Gao et al., "A Survey on Heterogeneous Federated Learning," Oct. 2022, [Online]. Available: <http://arxiv.org/abs/2210.04505>
- [8] B. B. Sezer et al., "PPFchain: A Novel Framework Privacy-Preserving Blockchain-based Federated Learning Method for Sensor Networks," *Internet of Things*, vol. 22, Jul. 2023.

- [9] X. Ma et al., "A State-of-the-Art Survey on Solving Non-IID Data in Federated Learning," *Future Generation Computer Systems*, vol. 135, pp. 244–258, 2022.
- [10] S. Banabilah et al., "Federated Learning Review: Fundamentals, Enabling Technologies, and Future Applications," *Inf Process Manag*, vol. 59, no. 6, p. 103061, 2022.
- [11] A. Imteaj et al., "A Survey on Federated Learning for Resource-Constrained IoT Devices," *IEEE Internet Things J*, vol. 9, no. 1, pp. 1–24, 2022.
- [12] K. Zhang et al., "Challenges and Future Directions of Secure Federated Learning: A Survey," *Front Comput Sci*, vol. 16, no. 5, 2022.
- [13] T. Li et al., "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Process Mag*, vol. 37, no. 3, pp. 50–60, 2020.
- [14] J. Shin et al., "FedBalancer: Data and Pace Control for Efficient Federated Learning on Heterogeneous Clients," in *Proceedings of the Annual International Conference on Mobile Systems, Applications and Services*, 2022.
- [15] Q. Yu et al., "Research on Heterogeneous Communication Network for Power Distribution Automation," in *3rd International Conference on Material, Mechanical and Manufacturing Engineering*, 2015.

생성형 AI 를 활용하는 IT 시스템의 빠른 구현을 위한 레퍼런스 아키텍처 제안과 적용 가능성 연구

송혜경, 권다운, 제유나, 하재현, 최영락
경북대학교 컴퓨터학부

{sosschs9, momnpa333, jyn523, mini0950, ian.choi}@knu.ac.kr

Proposing Reference Architecture and Applicability for Rapid Implementation of IT Systems with Generative AI

Hyegyeng Song, Dawoon Kwon, Yuna Je, Jaehyeon Ha, Yeongrak Choi

요 약

생성형 AI 기술이 급속도로 발전함에 따라 일상 생활에서 AI 를 필요로 하는 분야에 생성형 AI 를 접목하여 빠르게 IT 시스템화를 하는 요구 사항이 증가하고 있다. 본 논문에서는 해커톤과 같이 IT 시스템으로의 빠른 프로토타이핑 및 구현을 필요로 하는 분야에서 생성형 AI 기술을 직접적으로 활용 가능한 아키텍처 제안과 적용 가능성을 살펴본다. 구체적으로는 생성형 AI 와 관련한 비즈니스와 기술을 수용하기 위해 요구사항과 계층적 구조 아키텍처에 반영하고, 이를 대중적인 오픈 소스 소프트웨어로 빠르게 구현 가능하도록 정리한다. 또한 해당 아키텍처에 대한 실제 구현을 위해 제안한 오픈 소스 소프트웨어를 기반으로 실제 기술 스택을 활용하여, API 를 통해 각 계층을 분리하고 프로토타이핑 구현을 통해 적용 가능성을 확인한다.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음"
(2021-0-01082)

1. 서 론

2018년 ELMo, BERT, GPT 와 같은 Pre-trained 언어 모델이 제안된 이후, GPT(Generative Pretrained Transformer) 시리즈에서는 최초 모델인 GPT-1 의 파라미터 수가 1.17 억개였음에도 이후 GPT-2 가 15 억, GPT-3 는 1.75 조에 달할 만큼 생성형 AI 기술 성장 속도가 매우 빠른 상황이다. [1] 또한 2022년 11월 OpenAI 에서 GPT 언어 모델에 대화형 인공지능 챗봇을 연동하여 출시한 ChatGPT 는 생성형 AI 기술에 자동 문서 생성, 이미지 생성과 같은 비즈니스 연결을 통해 기존 AI 서비스 대비 강력한 성능을 누구나 쉽게 사용할 수 있도록 지원하였다. 그로 인해 서비스를 시작한 지 5일 만에 이용자 100만 명을 돌파하고 2개월 만에 1억 명을 넘어서는 등의 폭발적인 관심을 유발하였으며, 뿐만 아니라 생성형 AI 기술을 기반으로 언어 모델을 직접 개발하거나 API 로 사용 가능한 인공지능 모델을 활용한 유의미한 사례가 도출되고 있다. [2] 일상 생활에서 AI 를 필요로 하는 다양한 분야에 기업, 조직, 개인이 접목하여 생성형 AI 기술의 실현 가능성을 살펴본다면, AI 를 보다 유익하게 활용하는 데 큰 도움을 줄 수 있을 것이다. 생성형 AI 를 활용한 서비스를 제공하려면 생성형 AI 와 관련한 비즈니스 요구사항을 명확히 이해하고 이에 맞게 IT

시스템을 빠르게 프로토타이핑 및 구현하는 것이 필수적이다. 이 때, 레퍼런스에 해당하는 아키텍처와 프로토타이핑을 위한 템플릿이 짧은 기간동안 구현 가능성을 검토하는 해커톤과 같은 행사에서 활용된다면 MVP(Minimum Viable Product)를 구현하고 제한된 기간 내에 비즈니스 아이디어의 실현 가능성을 확인하는 데 도움이 될 것이다. 본 논문에서는 생성형 AI 와 관련한 비즈니스 요구 사항 및 기술 스택을 유연하게 수용하는 계층적 구조 아키텍처를 제안하고, 대중적인 오픈 소스 소프트웨어를 기반으로 직접 프로토타입 구현을 통해 제안한 아키텍처에 대한 적용 가능성을 파악하고자 한다.

2. 아키텍처 제안

본 논문에서는 비즈니스의 명확한 요구사항 정의, 비즈니스에 대한 신속한 IT 시스템화를 위한 계층적 구조 설계, 대중적인 오픈 소스 소프트웨어를 활용을 통해 빠른 프로토타이핑 및 IT 시스템화를 하는 것을 목표로 한다. [3] 위에 제시한 3가지 명제를 기준으로 다음과 같은 아키텍처를 제안한다.

2.1. 비즈니스 요구사항

생성형 AI 기술을 활용하고자 하는 비즈니스 요구사항은 대상 비즈니스 도메인 및 분야에 따라 달라질 수 있기에, 본 논문에서는 국내에서 해커톤에 참여하는 주 대상이 쉽게 이해 가능한 한글 이야기에 대해 생성형 AI 기술을 활용하는 비즈니스를 선정하여 비즈니스 요구사항을 정리해보았다. 해당 정리한 내용을 다른 비즈니스 도메인 및 분야에 각 항목별로 기술한다면 생성형 AI 를 기반으로 하는 시스템에 대한 비즈니스 요구사항 정리에 도움이 될 것이다. 관련 자세한 비즈니스 요구사항 및 유스케이스 & 클래스 다이어그램은 [4]에서 확인할 수 있다.

- 시스템명: “생성형 AI 를 활용한 한국어 이야기 분석 및 퀄리티 추천 시스템”
- 생성 결과물: 생성형 AI 를 통해 평가/분석한 한글 이야기
- 생성형 AI 기술 적용 파트: 한국어 이야기 구조 및 감정 분석, 이야기 작성에 보조적인 역할
- 생성형 AI 에서 참고하는 데이터: 전통적인 한국어 소설 (구조적인 가이드 제공)
- 비즈니스 고려 사항: 회원 가입을 통한 무료/유료 서비스 구분

주요 기능명	상세 기능 항목	데이터
회원 관리	회원 가입, 로그인, 비밀번호 찾기, 회원 정보 수정	이메일 주소, 비밀번호, 닉네임, 회원 개인 정보 (이름, 생년월일 등)
스토리 (카드) 관리	카드 목록보기/추가/변경/삭제	이야기 단위 (전제, 설정, 인물, 개요)
카드에 대한 구조 분석	생성형 AI를 통해 카드에 대한 이야기 구조 분석 및 결과 확인	카드 구조 분석 결과 (유사도, 선별 근거, 추출 핵심 키워드)
카드에 대한 감정 분석	생성형 AI를 통해 카드에 대한 감정 분석 및 결과 확인	스토리 단계별 맥락 및 감정 변화

표 1 한국어 이야기 분석 및 퀄리티 추천 시스템을 예시로 한 비즈니스 요구사항 정리

2.2. 계층적 구조 아키텍처

2.2.1. 프론트엔드

이 레이어에서는 제공하고자 하는 정보를 시각화하기 위해 사용자에게 UI 를 제공하고, 사용자의 입력을 받는다. 입력에 따라 백엔드에 요청하고 그 결과를 사용자에게 시각화하여 제공한다. 사용자 경험을

향상시키고 상호작용을 효과적으로 지원하기 위해 직관성, 일관성, 효율성 등을 고려한 UI 디자인이 필요하다.

2.2.2. 백엔드

이 레이어에서는 계층적 구조에서 프론트엔드를 지원하고 제어하는 역할을 수행하며 데이터 관리, 비즈니스 로직 처리, 보안 및 성능 최적화를 위한 특정 기술과 도구를 활용한다. 적절한 웹 프레임워크를 선택하여 비즈니스 로직을 구현 및 관리하고, REST API 를 통해 다른 계층의 아키텍처와 데이터를 주고받는다. 또한 사용자 인증 및 인가를 담당하여 보안 조치를 취하고, 데이터베이스 관리 시스템을 활용하여 데이터를 효율적으로 관리한다. 이를 통해 해당 레이어에서는 전체 시스템이 원활하게 상호작용하고 안전하게 운영될 수 있도록 한다.

2.2.3. 브릿지 & 생성형 AI

이 레이어에서는 프론트엔드에서 전달받은 입력에 알맞은 답변을 가져오기 위해 LLM(언어 모델)기반의 생성형 AI 가 사용된다. 프롬프트 엔지니어링을 통해 답변 형식을 지정하더라도, Stochastic 한 인공지능의 특성으로 인해 원치 않는 답변 형식이 나올 수 있으며, 때로는 질문과 관련 없는 답변이 나올 수 있다.[5] 이러한 경우 파싱 작업을 통해 비즈니스 요구사항에서 원하는 형태로 변환하고, 만약 파싱 작업으로 해결할 수 없다면 다시 질문하여 올바른 답변을 얻어온다. 이후 구현한 REST API 를 프론트엔드에 직접 전달하는 것이 아닌, 시스템의 전반적인 상호작용을 관리하는 서비스 레이어(Service Layer)에 전달하기 위해 ‘브릿지’ 역할을 하는 가벼운 프레임워크(백엔드) 또한 필요하다.

2.2.4. 인프라

이 레이어에서는 시스템의 기반이 되는 하부 구조를 담당하며 여러 역할을 수행하여 버전 관리 및 전체 시스템의 안정성, 확장성, 성능을 보장한다. 주요 작업으로는 클라우드 배포를 통한 애플리케이션 운영, 로컬 테스트를 통한 코드 안정성 검증, 직접 서버 배포 및 운영체제, 네트워크, 보안 설정 담당, CI/CD 를 활용한 자동화된 배포 프로세스 구축, 스케일링 및 리소스 관리, 보안 및 인증 강화, 로그 및 모니터링 시스템을 활용한 감시와 이슈 추적 등이 있다. [6] 이러한 작업들을 통해 전반적인 시스템 운영을 안정적으로 지원한다.

2.3. 대중적인 오픈 소스 소프트웨어

각 레이어에 대한 대중적인 오픈 소스 소프트웨어를 선택하여 빠른 구현을 할 수 있다. 사용 가능한 오픈 소스 소프트웨어 및 본 논문에서 사용한 소프트웨어를 정리한 내용은 표 2 와 같다.

레이어명	사용 가능한 오픈소스 소프트웨어	본 논문에서 사용한 오픈소스 소프트웨어
프론트엔드	Angular, Nest, Next, React, Svelte, Vue	React
백엔드	Spring Boot, Django, Express.js, Flask, Ruby on Rails, ASP.NET, Laravel	Spring Boot
브릿지 & 생성형 AI	Express.js, FastAPI, Flask LangChain, Semantic Kernel (LLM - LLaMA 및 상용 API를 통한 LLM - Azure OpenAI, Bard, OpenAI 연동)	FastAPI LangChain (Azure OpenAI 연동)
인프라	OpenStack, Kubernetes Docker, Git, Jenkins	Git 사용 및 GitHub & 오픈 소스 인프라에 배포 가능한 솔루션을 상용 클라우드에 배포

표 2 각 레이어 별 오픈소스 소프트웨어 정리

3. 구현 결과

해커톤과 같은 행사에서 참고할만한 MVP 구현을 위해 총 4 명이 한 팀을 구성하고 각 아키텍처 계층 별로 역할 분담을 하여 시스템 구현 작업을 수행해 보았다. 첫 번째로 비즈니스 요구사항을 명확하게 정리함으로써 시스템의 핵심을 이해하고 목표를 공유하였다. 이후에는 UI 프로토타이핑을 진행하여 사용자 경험을 시각적으로 확인하고 이를 바탕으로 구현을 진행하였다. 특히, 각 아키텍처 계층을 분리하기 위해 REST API 를 개발 가능한 오픈 소스를 활용하여 API 를 통한 의존성 분리를 진행하였다.

사용자는 시스템을 이용하기 위해 회원가입 및 로그인을 진행한다. (그림 1) 회원가입 시 사용자 정보가 데이터베이스에 저장되며, 로그인 시 Spring Security 를 통한 사용자 인증 및 권한 부여가 가능하다.

사용자는 최대 3 개의 스토리 카드를 작성할 수 있고, 본인의 스토리 카드 내용을 데이터베이스에 저장하여 조회 및 수정이 가능하다. 스토리 카드 분석을 요청할 경우 Premise, Setting, Characters, Outline 을 생성형 AI 의 입력으로 답변을 파싱하여 JSON 형태로 받아온다. 선별한 5 개의 키워드를 바탕으로 유사한 문학을 3 개 선정하고 각 키워드에 대한 문학작품과의 유사도의 분석 결과를 막대 그래프로 시각화하여 나타낸다.

(그림 2)는 소나기(황순원)을 입력 예시로 하였으며, 죽은 시인의 사회(남궁창)와의 유사도를 보여주고 있다.

그림 1 회원가입 화면



그림 2 스토리 카드 분석 결과 화면

스토리 감정 분석을 요청할 경우, 스토리에 대한 요약본을 생성형 AI 의 입력으로 하고, 답변을 파싱하여 JSON 형태로 받아온다. 잘 알려진 소설 구조인 발단, 전개, 위기, 절정, 결말에 따른 단계 별 스토리 분석 결과와 전체적인 스토리 분석 결과를 제공한다. 단계 별 스토리 분석 결과는 단계 별 감정 분석 결과를 막대

그래프로 나타내고, 각 감정 별로 단계에 따른 점수 변화를 꺾은 선 그래프로 나타내 비교한다. (그림 3, 4) 전체적인 스토리 분석 결과는 감정 분석 결과를 막대 그래프로 나타내고, 그 판단 근거를 표로 나타낸다. (그림 5) 입력의 예시는 소나기(황순원)이다.

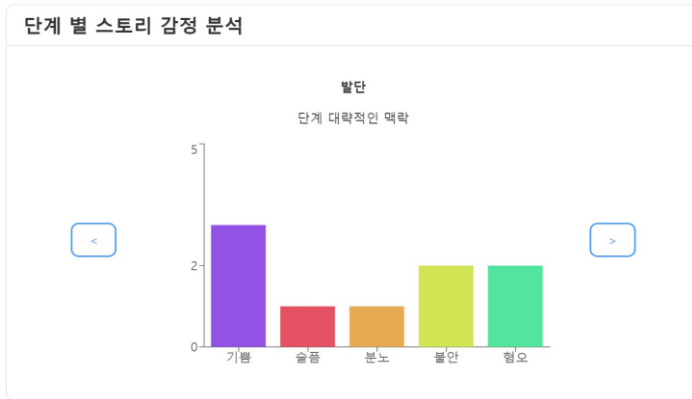


그림 3 스토리 단계 별 감정 분석 결과 화면

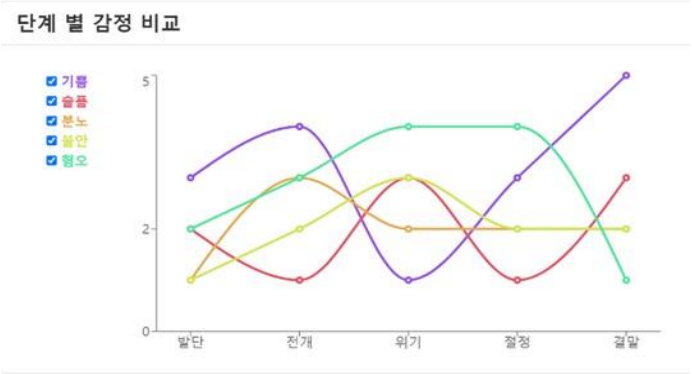


그림 4 스토리 단계 별 감정 비교 화면



그림 5 전반적인 스토리 분석 결과 화면

해당 결과를 바탕으로 정의한 비즈니스 요구 사항을 수용한 프로토타이핑 구현을 완료하였다. 해당 결과물은 GitHub 에 저장되어 있으며 [4], 누구나 fork 를 통해 템플릿으로 활용하여 다른 생성형 AI 비즈니스 요구 사항에 대해서도 빠른 요구사항 정리 및 프로토타이핑 구현이 가능할 것으로 보인다.

4. 결론

본 논문은 생성형 AI 를 활용하는 IT 시스템을 빠르게 구현하기 위한 레퍼런스 아키텍처를 제안하고 적용 가능성을 파악한다. 이를 확인하기 위해 빠른 프로토타이핑 및 IT 시스템화를 위한 필수 사항을 살펴보고 비즈니스의 요구사항을 명확히 정의하고, 계층적 구조 설계와 대중적인 오픈 소스 소프트웨어 활용을 통해 빠른 IT 시스템화를 달성하는 방법을 제시한다. 주요 결과로는 비즈니스 요구사항을 충족시키는 시스템을 구현하였으며, 해당 연구 결과를 활용해 다른 비즈니스에서 참조할 수 있는 레퍼런스로 제공한다.

현재 구현한 결과물은 해커톤과 같이 빠른 구현을 필요로 하는 분들이 쉽게 사용하고 기술을 빠르게 습득할 수 있도록 템플릿으로 제공하며, 생성형 AI 를 직접 활용하는 비즈니스에 초점을 맞추었다. 또한, 다양한 비즈니스를 직간접적으로 수용할 수 있는 형태와 웹이 아닌 플러그인 형태 등에 대한 레퍼런스 모델 또한 고민해보며 더 다양한 확장 가능성과 적용 시나리오를 탐구할 계획이다.

참고 문헌

[1] Min Zhang, Juntao Li, A commentary of GPT-3 in MIT Technology Review 2021, Fundamental Research, Vol 1, Issue 6, 2021, Pages 831-833.
 [2] 김태원. 챗 GPT 가 촉발한 생성형 AI 시대, 미래 모습과 대응 방안. FUTURE HORIZON,(55), 2-9. 2023.
 [3] V. Bilgram and F. Laarmann, "Accelerating Innovation With Generative AI: AI-Augmented Digital Prototyping and Innovation Methods," in IEEE Engineering Management Review, vol. 51, no. 2, pp. 18-25, 1 Secondquarter,june 2023,
 [4] <https://github.com/Team-CDP1-class>
 [5] Much to discuss in AI ethics. Nat. Mach. Intell. 4, 1055-1056. 2022
 [6] Hyun Kyoung Lee and Dong Kun Noh, "Design of the Efficient CI/CD (Continuous Integration and Continuous Deploy) Automation Model," Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 372-373, 2021.

긴급신고 상황에 교환기에 종속된 CTI시스템에 대한 향후 확장방향에 대한 연구

추현우 김동균

경북대학교

hwhchoo@gmail.com dongkyun@knu.ac.kr

Exploring Future Expansion Directions of Exchange-Dependent CIT System in Emergency Reporting Scenarios

Hyunwoo Choo, Dongkyun Kim

Kyungbuk University

요약

현대 사회에서 긴급상황에 신속하게 대응하기 위해 긴급신고 전화 시스템은 매우 중요한 역할을 수행하고 있습니다. 그러나 대규모 재난 발생 시, 긴급신고 전화 번호로의 동시 다수의 전화로 인해 호 폭주 상황이 발생할 수 있습니다. 이러한 상황에서 기존의 교환기 종속형 CTI시스템은 신고 전화를 적절히 처리하지 못하며, 일부 사용자들은 전화가 아예 연결되지 않는 문제에 직면할 수 있습니다. 본 연구는 클라우드 네이티브 기반의 CTI시스템을 소개하여 현재의 SIP기반 통신 인프라를 반영하고, 자원의 한계를 넘어선 CTI시스템이 대규모 재난상황에서 끊임 없는 긴급신고 전화 연결을 보장할 수 있는 방안에 대한 연구를 수행합니다.

1. 서론

현대 사회에서 긴급한 상황에 빠르게 대응하는 데에는 긴급신고 전화 시스템이 큰 역할을 하고 있습니다. 이러한 시스템은 긴급한 도움을 필요로 하는 시민들과 신속한 조치가 필요한 긴급 서비스 기관 간의 연결을 촉진하여 사람들의 안전과 생명을 보장하는 중요한 도구로 작용하고 있습니다.

그러나 대규모 재난이 발생하면 긴급상황에 대한 신고 전화 수가 급증하여 호 폭주 상황이 발생할 수 있습니다. 이는 기존의 교환기 종속형 시스템이 갖고 있는 물리적 한계에서 CIT시스템이 처리하기 어려운 과제 중 하나로, 일부 사용자는 전화가 아예 연결되지 않는 문제에 직면할 수 있습니다.

본 연구는 이러한 도전에 대응하여 클라우드 네이티브 기반의 CTI시스템을 소개하고자 합니다. 이 시스템은 현재의 SIP(Session Initiation Protocol) 기반 통신 인프라를 반영하여 자원의 한계를 넘어선 CTI 시스템이 대규모 재난 상황에서 끊임 없는 긴급신고 전화 연결을 보장할 수 있는 방법에 대한 연구를 수행할 것입니다.

이를 통해 본 논문은 대한민국의 긴급신고기관 중하나인 119소방 시스템의 예를 통해 CTI 시스템의 한계를 극복하고, 긴급한 상황에서 효율적이고 안정적인 서비스를 제공하기 위한 새로운 방향성을 모색하고자 합니다. 현대 사회의 긴급 대응 체계를 강화하고 안전성을 높이는 데에 이 연구가 기여할 것으로

기대됩니다.

2. 본문

대한민국 소방청이 매년 발행하는 소방통계연보에 따르면 119 접수대는 2021년 전국 합계 878대가 운영 중에 있습니다.

연도별	119종합상황실							소방서
	119접수대 현황				회선 현황			
	소 계	119 접수대	응급의료 상담 접수대	보조 접수대	평상 시 119전용회선수		시스템 장애시 비상접수 회선	119전용 회선수
				주	예 비			
2015	558	274	87	197	3,120	720	772	1,460
2016	598	302	88	208	3,030	720	320	1,670
2017	609	307	82	220	2,730	887	333	1,410
2018	689	316	80	293	3,240	810	499	1,611
2019	745	317	78	350	3,240	822	480	1,533
2020	850	338	65	447	3,360	702	457	1,571
2021	878	365	70	443	3,400	922	344	1,580

[그림 1] 연도별 119 접수대 현황

그림1에서 알 수 있듯이 2021년에 자연재난, 대형화재 등 119 신고 폭주에 대비한 보조 접수대의 운영대수가 평상시 접수대보다 21% 많은 것을 확인할 수 있다. 또한 2015년부터 증가하고 있는 접수대

현황을 나타내고 있는데, 119접수대가 33% 증가하는 동안 보조접수대는 동기간에 124% 증가한 것을 확인할 수 있습니다.

이와 같이 지금까지는 물리적 하드웨어를 증설하여 시민들에게 안정적이 서비스를 제공하고자 하였지만, SIP기반의 통신 시스템을 통해 접수를 한다면, 더 이상의 보조 접수대를 증설하지 않고도 호 폭주 상황에서 상대적으로 더 효과적인 대응이 가능할 수 있을 것입니다. 클라우드 네이티브화 된 신고접수시스템을 구축하게 된다면 스케일링 및 자원 관리의 이점을 통해 특정 지역이나 재난이 일어난 상황에 필요한 만큼의 자원을 동적으로 할당하여 대응할 수 있습니다. 또한, 클라우드 네이티브한 시스템의 이점 중 하나인 로드 밸런싱을 통해 트래픽을 분산 시키고 각 서버에 고르게 분배함으로써 호 폭주 상황에서의 효과적인 대응이 가능합니다.

4-2-2-2. 연도별 119신고 지령시스템 운영 현황(2012-2021)

연도별	합 계	출동관련 신고					유관 기관 이첩	안내 및 민원 전화	장난 전화	무응답	오접속	기 타	
		소 계	화 재	구조	구 급	대민 출동 및 기타							
2012	계	19,362,314	3,265,681	222,810	463,062	2,133,355	446,454	120,777	3,023,960	17,917	2,744,682	7,395,358	2,793,949
	유선	5978234	1,218,155	61,905	162,541	775,063	218,646	37,245	1,117,417	6,863	626,853	1,960,473	1,011,228
	무선	13384080	2,047,526	160,905	300,521	1,358,292	227,808	83,532	1,906,533	11,054	2,117,829	5,434,885	1,782,721
2013	계	13,788,198	3,372,929	206,857	433,113	2,183,678	549,281	127,368	3,510,436	7,198	1,758,802	3,779,428	1,232,037
	유선	4,758,240	1,190,299	51,847	132,055	734,678	271,719	39,297	1,305,252	3,800	558,732	1,150,015	510,845
	무선	9,029,958	2,182,630	155,010	301,058	1,449,000	277,562	88,071	2,205,184	3,398	1,200,070	2,629,413	721,192
2014	계	11,841,628	3,720,100	241,583	473,439	2,371,224	633,854	111,102	3,233,505	3,232	1,601,886	2,099,259	1,072,544
	유선	3,937,040	1,197,667	54,003	131,932	664,853	346,849	28,086	1,043,680	1,634	483,034	736,029	446,910
	무선	7,904,588	2,522,433	187,580	341,507	1,706,341	287,005	83,016	2,189,825	1,598	1,118,852	1,363,230	625,634
2015	계	10,415,466	3,766,869	243,360	476,865	2,462,140	584,504	76,747	3,114,211	2,267	1,231,053	1,176,110	1,048,209
	유선	2,991,529	1,139,660	51,416	115,197	636,413	336,634	15,032	835,236	845	313,711	388,098	298,947
	무선	7,423,937	2,627,209	191,944	361,668	1,825,727	247,870	61,715	2,278,975	1,422	917,342	788,012	749,262
2016	계	10,727,743	3,979,540	234,798	594,856	2,569,380	580,526	127,083	3,216,320	2,094	1,594,268	874,968	933,480
	유선	2,126,072	920,826	40,323	103,886	466,356	310,251	13,092	556,351	622	246,885	139,667	248,629
	무선	8,601,671	3,058,714	194,475	490,960	2,103,004	270,275	113,991	2,659,969	1,472	1,347,383	735,291	684,851
2017	합계	11,550,786	4,268,749	323,515	605,446	2,667,533	672,255	129,704	3,307,760	1,477	1,898,771	945,415	998,910
	유선	2,329,542	938,207	48,962	100,046	432,855	356,344	11,757	544,691	657	346,381	180,277	307,572
	무선	9,221,244	3,330,542	274,553	505,400	2,234,678	315,911	117,947	2,763,069	820	1,552,390	765,138	691,338
2018	합계	11,384,521	4,322,202	335,545	589,495	2,725,286	671,876	130,636	3,336,411	753	1,748,186	873,048	973,285
	유선	2,148,395	874,942	60,605	88,603	370,942	354,792	11,040	558,361	272	274,893	146,962	281,925
	무선	9,236,126	3,447,260	274,940	500,892	2,354,344	317,084	119,596	2,778,050	481	1,473,293	726,086	691,360
2019	합계	11,567,173	4,375,325	340,757	635,996	2,684,775	713,797	104,990	3,328,167	407	2,000,756	891,734	865,794
	유선	1,985,084	843,624	66,167	91,802	312,193	373,462	7,670	553,096	103	226,090	117,762	236,739
	무선	9,582,089	3,531,701	274,590	544,194	2,372,582	340,335	97,320	2,775,071	304	1,774,666	773,972	629,055
2020	합계	11,274,559	4,234,670	329,312	588,094	2,518,950	798,314	87,615	3,272,023	665	1,886,019	931,521	862,046
	유선	2,024,363	835,945	76,373	83,514	258,428	417,630	5,997	602,180	224	209,317	120,084	250,616
	무선	9,250,196	3,398,725	252,939	504,580	2,260,522	380,684	81,618	2,669,843	441	1,676,702	811,437	611,430
2021	합계	12,075,804	4,724,467	375,409	700,505	2,829,687	818,866	85,195	3,542,718	249	1,832,968	971,332	918,875
	유선	1,971,808	902,895	119,172	94,829	254,383	434,511	4,778	557,760	33	171,726	103,039	231,577
	무선	10,103,996	3,821,572	256,237	605,676	2,575,304	384,355	80,417	2,984,958	216	1,661,242	868,293	687,298

그림 2 연도별 119신고 지령시스템 운영 현황

연간 지령시스템으로 접수되는 신고 건수는 약 1200만 건 정도가 접수되고 있는데, 교환기를 탈피한 CTI시스템의 구축을 통해, 호 폭주 상황에서도 ARS시스템이 단순 통신 단절이 아닌, 신고자에게 최소한의 안내 서비스를 제공할 수 있다는 점을 강조하고자 한다. 교환기의 물리적 한계를 극복하고, CIT기술의 적극적인 활용을 통해 신고자들에게 신속하고 정확한 정보를 제공함으로

써 긴급 상황 대응의 효율성을 향상시킬 수 있습니다.

이러한 체계적인 인프라가 마련되어 있다면 정부가 최근 주장하고 있는 시시대에 걸맞는 시스템으로 발전할 수 있는 기반을 마련할 수 있을 것이다. AI 기술은 긴급상황에서 발생하는 다양한 상황을 접수자의 신속한 상황 판단을 지원하고 응답을 도와줄 수 있을 것입니다.

이러한 발전된 시스템을 통해, 국민들의 안전과 생명을 보다 신속하게 보호하는 긴급 신고 시스템의 혁신을 이루어 나가는 데에 본 연구가 도움이 될 것이라 기대합니다.

3. 결 론

본 연구는 119 신고접수 시스템의 연도별 현황을 통해 대한민국의 긴급 상황 대응 체계의 발전을 살펴보았습니다. 특히, 교환기 종속적인 시스템의 한계를 극복하고자 하는 노력으로써 보조 접수대의 증가를 다루었습니다.

교환기를 탈피한 CTI 시스템의 구축을 통해 호 폭주 상황에서의 효율적인 대응과 ARS 시스템의 높은 유연성을 강조하였습니다. 이러한 기술적인 발전은 단순 통신 단절이 아닌, 신고자에게 최소한의 안내 서비스를 제공하는 차원에서 긴급 상황 대응의 효율성을 높일 것으로 기대됩니다.

더 나아가, 이러한 체계적인 인프라는 정부가 최근 주장하고 있는 AI의 활용을 통한 신고접수 시스템 구축에도 중요한 기반이 될 것입니다. AI 기술의 도입은 긴급 상황에서의 판단력과 응답 능력을 향상시켜 국민의 안전을 보다 효과적으로 지킬 수 있는 가능성을 열어줍니다.

그러나, 신고접수를 받는 접수대와 접수자의 숫자는 물리적 한계가 있기에 시스템 도입과정에서 타당성을 잘 따져봐야 하겠습니다.

4. 참고문헌

[1] 2022 소방청 통계연보(배포용)
https://www.nfa.go.kr/nfa/releaseinformation/statisticinformation/main/;jsessionid=t-pgNSfw++sh+sBQhOk4h2-d.nfa22?boardId=bbs_0000000000000019&mode=view&cntId=44&category=&pageIdx=%20%20%20%20%20%20%20%20%20%20%20%20%201&searchCondition=&searchKeyword=

사용자 위치 및 관심사 기반의 오프라인 번개 모임 앱 고도화

윤재용*, 손석현, 김수진, 방수빈, 배주영, 여명동, **

Enhancing off-line meeting appbased on user location and interests

Yoon Jaeyong, Son seokhyeon, Kim Sujin, Bang Subin, Bae Juyoung, Yeo Myeongdong

요 약

본 논문은 위치 및 관심사 기반의 오프라인 번개 모임 앱의 사용자 알림 기능, 문자 기능, 시차 조정, 카테고리 필터링과 관리자 기능을 포함한 보안 기능의 향상과 베타 테스트를 통한 고도화 및 상용화를 목표로 한다. 어플리케이션의 서비스를 통해 오프라인 매칭 플랫폼으로 기능 할 수 있으며 본 연구는 서버 및 어플리케이션 재설계와 미비한 기능의 추가 구현을 제시한다.

I. 서 론

앱의 주요 서비스는 사용자의 위치와 관심사를 기반으로 단기 소규모 모임 생성 및 탐색에 도움을 주는 것이다. 사용자는 관심사와 모임의 위치, 내용, 시간, 인원, 마감 여부 등을 설정하여 ‘건수’라고 불리는 모임을 생성할 수 있다. 사용자마다 개별로 설정할 수 있는 관심사 카테고리나 현재 위치를 기반으로 모임을 탐색하고, 참여 신청과 승인을 통해 능동적으로 모임에 참여할 수 있다.

기존의 개발 결과물을 분석한 결과, 관리자 페이지와 API, 모임 참가 요청, 모임 초대, 모임 위치 필터링, 클릭 이벤트 등 서버와 앱의 UI/UX에 미구현된 부분이 다수 존재했다. 이런 부족한 구현 사항에 대한 수정 방안을 구상하고 보완하여 상용화가 가능할 정도로 고도화하는 것이 본 논문의 목표이다.

II. 본론

1. 시스템 개요

클라이언트의 경우 앱이 다양한 플랫폼에서 동작할 수 있도록 google의 Flutter 프레임워크를 사용했다. 서버의 경우 Spring Boot 프레임워크를 사용했으며 MariaDB로 데이터베이스를 구현했다. 또한 높은 생산성을 위해 Mybatis와 JPA를 함께 사용했다. 채팅과 알림 기능 구현은 FCM(Firebase Cloud Messaging)을 이용했다. 보안을 위해 Spring의 모듈 Spring Security와 OAuth2 인증 방식을 사용하여 보안 기능을 구현했다. 관리자 페이지의 경우 간단한 동적 웹 페이지를 구현하기 위해 Thymleaf를 사용했다. API 설계 및 문서화는 Swagger를 통해 진행하였다. 서버는 AWS Lightsail을 통해 배포된다.

2. DB 구조 및 다이어그램

기존의 경우 스프링 Security에서 제공하는 기본

*소속, Email, **소속, Email(교신전자표시), ...

※ 지원기관표기(사사표기)

적인 폼 로그인을 사용하였다. 현재 User 클래스에서 role 필드를 통해 관리하던 권한 설정을 제거했으며, 권한 및 보안에 관련된 설정은 UserPrincipal.record를 통해 별도로 관리한다. 이때 UserPrincipal은 Security 모듈의 userDetails와 OAuth2User를 상속받아 enum 타입의 사용자 권한을 가진다.

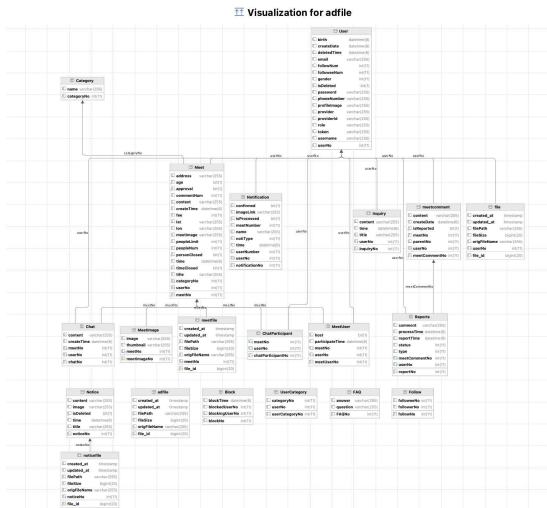


그림 1. 데이터베이스 다이어그램
Fig. 1. Database Diagram

3. 로그인 고도화

OAuth2.0 프로토콜을 기반으로 사용자 인증을 마치면, 핸들러에서 JWT 토큰을 생성해 클라이언트에게 넘겨준다. 기본 OAuth2.0 인증 처리 구현체를 사용하며 JWT 토큰은 Flutter의 SecureStorage와 웹의 localStorage에 저장된다. 사용자는 SNS 로그인과 일반 로그인 중 선택할 수 있으며, 로그인은 유효시간이 만료되거나 로그아웃 전까지 유지된다.

4. 관리자 페이지

관리자는 회원 관리, 콘텐츠, 공지사항, 문의사항, 광고, 데이터 통계에 대한 접근 권한을 가진다. 관리자는 회원이 생성한 모임 목록과 모임 내용, 팔로우 목록, 작성 댓글 등 상세 정보를 열람할 수 있으며 특정 회원을 차단 할 수 있는 권한을 가진다.

콘텐츠의 경우 관리자는 건수의 카테고리를 생성, 수정, 삭제가 가능하다. 또한 등록된 건수, 마감된 건수, 현재 진행 중인 건수에 대한 세부 사항을 열람할 수 있으며 건수를 삭제할 권한을 가진다. 관

리자는 건수 삭제와 같은 변경이 발생하며 유저에게 알림을 발송하며, Firebase의 FCM 서비스를 연동하여 선택적인 알림 발송이 가능하도록 구현한다.

공지사항, 문의사항 게시판의 경우 일반 사용자와 관리자가 모두 접근이 가능하다. 일반 사용자는 공지사항은 읽기 권한을 가지고, 문의사항의 경우 읽기와 쓰기 권한을 가진다. 이와 달리 관리자의 경우 두 가지 카테고리에 대하여 읽기, 쓰기, 삭제 권한을 가진다.

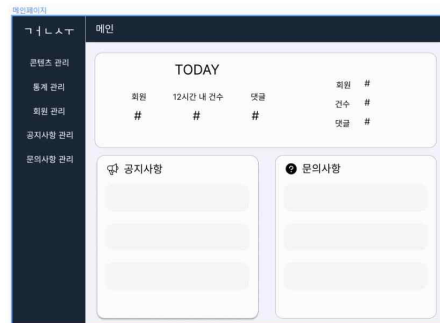


그림 2. 관리자 페이지 메인
Fig. 2. Admin page main

관리자는 앱 내 광고에 대한 모든 권한을 가지고 있어 광고 내용과 이미지를 등록, 수정, 삭제할 수 있다. 또한 관리자는 그림2처럼 관리자 페이지를 통해 각종 통계 자료에 접근할 수 있다. 통계 페이지에서 등록된 총건수, 새롭게 가입한 사용자의 수, 전월 대비 감소한 사용자 증가 감소 추이, 카테고리별 이용 내역 등의 통계 자료를 볼 수 있다.

III. 결론 및 향후 과제

본 논문에서는 사용자의 위치와 관심사를 기반으로 오프라인 매칭 서비스를 제공하는 어플리케이션의 설계와 구현을 중점적으로 살펴보았다. 이를 통해 관리자 권한 관리를 포함한 보안 사항과 미흡하게 구현된 기능의 수정 방안과 대안을 제시했다. 이 연구는 부족한 기능 재설계에 대한 방법을 제시한 사례로 사용가능 할 것으로 보인다.

향후 베타테스트를 진행하여 트래픽 발생량을 확인하고 대용량 트래픽 처리를 위한 Redis 및 Kafka 사용을 추가하는 것도 고려할 수 있을 것이다.

또한, 앱 사용자의 연령대, 성별에 따라서 관심사 키워드의 분포를 다중회귀분석을 활용하여 분석한다. 이를 관리자 페이지에서 확인함으로써 최근 사용하는 키워드의 추세를 파악할 수 있다. 이는 차후 어플리케이션 업데이트 방향성을 결정하는 것에 사용된다.

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음” (2021-0-01082)

참 고 문 헌

- [1] 회귀분석을 이용한 다중이용시설 실내오염물질 특성 연구 이세형 외 7명, 2013, 125-127

강철영상의 미세조직 분할을 위한 UNet3+ 개선

비샬^o 고재필

금오공과대학교 컴퓨터공학과

{bishalswain, nonezero}@kumoh.ac.kr

Refining UNet3+ for Segmentation of Microstructures in Steel Images

Swain Bishal Ranjan^o, Jaepil Ko

Dept. of Computer Engineering, Kumoh National Institute of Technology

Abstract

Segmentation of microstructures in steel images is crucial as it helps in assessing the quality and strength of the material. Traditional and existing deep learning models often struggle to capture both the textural details and structural patterns of the microstructures found in steel images. In this paper, we enhance the structural representation of steel images by modifying the UNet3+ design, including data processing, architectural modules, and a tailored loss function. Experimental results show that the modified model is effective in distinguishing microstructures not only within the same steel type and magnification, but also across different scales and steel types.

1. Introduction

The characterization of microstructures in steel is a critical aspect of materials science, playing a crucial role in determining the physical and mechanical properties of the material [1]. The microstructural composition of steel directly influences its strength, ductility, and overall performance in various applications [2]. Consequently, the precise identification and analysis of these microstructures are essential for ensuring the quality and reliability of steel materials.

Recently, deep learning approaches have been introduced to automate this process. Despite their advancements, existing deep learning models, such as standard convolutional neural networks, have struggled to accurately capture the complex textural details and diverse structural patterns found in steel microstructures [3]. One of the primary limitations of these models is their inability to handle the intricate variability inherent in steel microstructures.

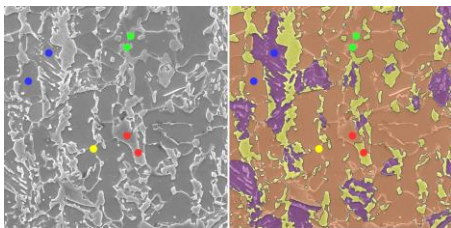


Figure 1. Illustration of the challenges in segmenting alloy steel microstructure: SEM (left) and its labeled image (right).

As shown in Figure 1, the challenges are multi-faceted: blue markings in the images illustrate regions that appear similar yet belong to different phases; green markings highlight areas with distinct textures and clear boundaries but of the same phase; red dots signify parts that, despite seeming similar, pertain to different phases; and yellow dots indicate that the presence of boundaries does not necessarily imply phase differences. Such complexities require a model to learn both structure and texture.

To address these challenges, we employ UNet3+ [4] architecture. The UNet3+ model is notable for its improved multi-scale feature extraction capability, essential for discerning subtle variations in the texture and structure of steel images. The full-scale skip connections of the model enable the integration of features across different layers, providing a richer and more detailed representation.

In this paper, we enhance the structural representation of steel images by modifying the UNet3+ design, including data processing, architectural modules, and a tailored loss function.

2. Refining UNet3+

2.1 Data Augmentation

We employ data augmentation to increase the diversity of training images, including horizontal and vertical flips, random rotations (0–90 degrees), scaling (1–2.5x), brightness adjustments, gamma value alterations, histogram

equalization, and sliding window operation. They simulate varying orientations, magnification levels, lighting conditions, and positional variances within steel microstructures.

2.2 Modifying UNet3+ for enhanced structural representation
To address specific challenges of microstructures in steel images, we incorporate – atrous convolutions [5], blur–pooling [6], and a broader kernel size of 7x7. Atrous convolution allows the model to expand the receptive field without losing resolution or increasing the number of parameters. This is beneficial for capturing the subtle yet critical variations in steel microstructures, where it is important to preserve details at different scales. Blur pooling layers were introduced in the model for reducing the spatial resolution of feature maps while preserving critical information, potentially enabling the model to identify larger–scale patterns and mitigate overfitting. The usage of a larger kernel size, such as 7x7, further aids in capturing broader contextual information, enabling the model to better understand and segment complex structural patterns that are common in steel images.

2.3 Combined Loss function

We use a combined loss function, which incorporates Focal [7], Jaccard [8], and Multi–Scale Structural Similarity (MS–SSIM) [9] losses specifically tailored to address specific challenges encountered in steel microstructure segmentation. Focal Loss tackles class imbalance, as it selectively emphasizes the learning of underrepresented phases or structures. Jaccard Loss is essential for precise boundary delineation between different phases in steel microstructures. MS–SSIM Loss concentrates on maintaining structural integrity and texture similarity. It ensures that the model captures the subtle textural nuances and structural patterns unique to each phase.

3. Experiments

3.1 Dataset and Experimental setup

The dataset comprised images of four different types of alloy steel, each characterized by varying tensile strengths and imaged at different magnifications:

- E–type Steel: Tensile strength of 580 MPa, magnifications of x2700, x3000, and x5000.
- A–type Steel: Tensile strength of 780 MPa, magnification of x5000.
- D3–type Steel: Tensile strength of 980 MPa, magnification of x5000.
- H2–type Steel: Tensile strength of 1108 MPa, magnification of x5000.

The initial phases of augmentation, training, and testing

used only the x2700 magnified E–type steel images. For scalability experiments, we extended our analysis to include images of x3000 and x5000 magnifications from E–type, as well as x5000 from A–type, D3–type, and H2–type steel.

We used a learning rate of 0.00001 and the model was trained for 30 epochs. We used the Rectified Linear Unit (ReLU) [10] activation function since it is particularly suitable for differentiating subtle differences in textures and structures, thanks to its efficiency in handling non–linear features without significant computational burden.

3.2 Results and Discussions

The results shown in Table 1 indicate that the modified UNet3+ model outperforms its predecessors – UNet, UNet++ [11] in all the tested categories. Notably, in high–magnification images (X5000 E–type, A–type, D3–type, and H2–type), the modified UNet3+ model demonstrates a significant improvement in both MPA and Dice scores, indicating superior segmentation capability.

TABLE I. Performance of UNet, UNet++ and UNet3+ across different magnifications and types of alloy steels.

Models	X3000 E–type	X5000 E–type	A–type	D3–type	H2–type
UNet	58.85 (0.5743)	28.82 (0.2768)	46.29 (0.4601)	39.9 (0.3842)	44.48 (0.4292)
UNet++	72.01 (0.7076)	80.49 (0.8028)	68.29 (0.6793)	54.73 (0.5361)	64.54 (0.6338)
Stock UNet3+	78.89 (0.7605)	81.04 (0.7932)	66.31 (0.6490)	73.28 (0.7272)	70.12 (0.6949)
modified UNet3+	79.30 (0.7809)	89.51 (0.8895)	74.37 (0.7293)	78.24 (0.7692)	76.13 (0.7471)

The modified UNet3+ model demonstrates an enhanced ability to discern and segment complex microstructures, addressing the initial challenges. The model effectively handles variations in scale, contrast, and the intricate nature of microstructures, which were previously problematic for standard models. Figure 2 shows qualitative segmentation results, where the modified UNet3+ model accurately identifies phases and boundaries that were ambiguously represented in the original images. The model's performance in distinguishing between similar–looking phases and correctly identifying phases with clear boundaries is particularly noteworthy. Moreover, the model adeptly navigates areas with deceptive similarities and correctly interprets boundary presence, which are crucial for accurate phase identification.

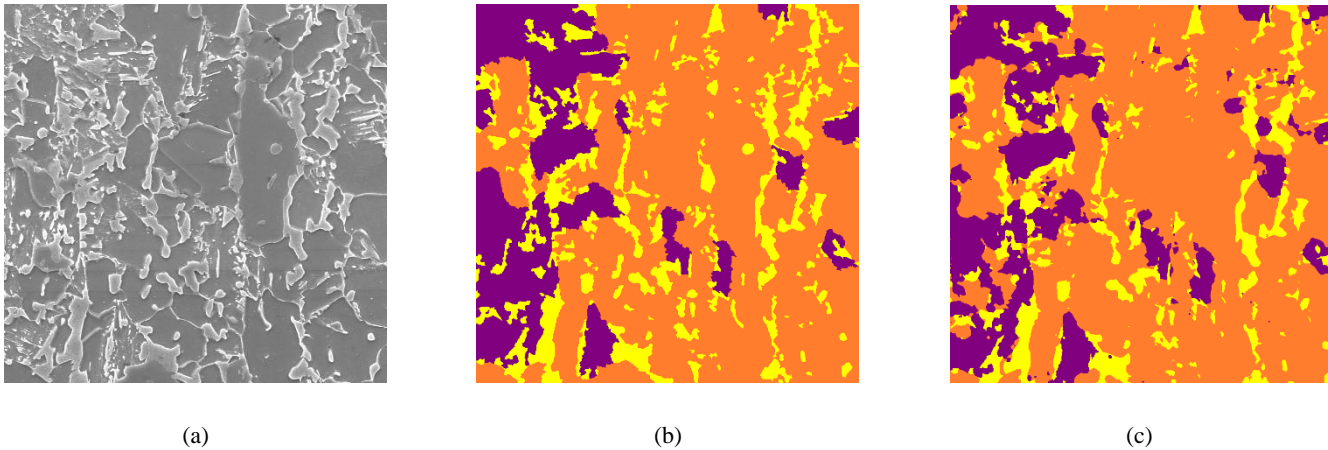


Figure 2. (a) is the input SEM image into the model (b) is the target label image (c) is the inferred image of the modified UNet3+ model. The label colors purple, orange and yellow correspond to bainite, ferrite and martensite respectively.

4. Conclusions

Based on the research, it can be stated that the use of UNet3+ architecture for microstructure segmentation in steel images results provides performance improvements over previous models. The adaptation of UNet3+ by incorporating atrous convolutions, blur-pooling and a broader 7x7 kernel size showed superior performance. The implementation of a combined loss function in our model facilitated the learning of both structural and textural aspects of various steel phases. The scalability experiments conducted across different magnifications and steel types (E-type, A-type, D3-type, and H2-type) have further validated the robustness and versatility of the UNet3+ model. These experiments underscore the model's capability to adapt to and accurately segment complex microstructures. There could be various other challenges including the impact of steel surface treatments, presence of microstructural defects, that could hinder the segmentation capability of the model. These challenges could serve as prime focuses for future work and related studies.

References

1. Bubryur Kim, N. Yuvaraj, Hee Won Park, K.R. Sri Preethaa, R. Arun Pandian, & Dong-Eun Lee Investigation of steel frame damage based on computer vision and deep learning. *Automation in Construction*, 132, 103941.2021
2. Bakhracheva, Yulia. Prediction of Fracture Toughness of Steels According to the Microstructure and Chemical Composition. *Materials Science Forum*, vol. 989, Trans Tech Publi, 2020, pp. 353–358.
3. Luengo, J.; Moreno, R.; Sevillano, I.; Charte, D.; Pel'aezVegas, A.; Fern'andez-Moreno, M.; Mesejo, P.; Herrera, F. A tutorial on the segmentation of metallographic images: Taxonomy, new metaldam dataset, deep learning-based ensemble model, experimental analysis and challenges. *Inf. Fusion* 2022, 78, 232–253.
4. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *arXiv* 2020, arXiv:2004.08790
5. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, & Alan L. Yuille. (2017). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.
6. Richard Zhang. (2019). Making Convolutional Networks Shift-Invariant Again.
7. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
8. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019; pp. 658–666.
9. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale Structural Similarity for Image Quality Assessment. In *Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, CA, USA, 9–12 November 2003; pp. 1398–1402. 2003
10. Abien Fred Agarap. (2019). Deep Learning using Rectified Linear Units (ReLU).
11. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. 2018.

Global IT Leader!

모든 비즈니스 영역을 통합하는 통찰력으로
고객의 니즈를 완벽히 분석한 최적의 서비스로
미래를 선도하는 최첨단 기술력으로

미래의 가치를 먼저 생각하는 기업

Total Solutions

- SI·NI 사업
- 보안솔루션

Smart Service

- Mobile 솔루션 사업
- 금융 솔루션 사업

Art Technologies

- 산업용 PDA 사업

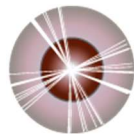
큰다 믿을득

대신정보통신주식회사 Daishin Information & Communications Co., Ltd.

서울특별시 금천구 가산디지털2로 169-16, 6층 (가산동, 하우스다가산퍼스타) Tel_02-2107-5000 Fax_02-2107-0515

www.dsic.co.kr

UCWIT 2023



한국정보과학회

KOREAN INSTITUTE OF INFORMATION SCIENTISTS AND ENGINEERS

06704 서울특별시 서초구 방배로 76, 401호 (방배동, 머리재빌딩)

전화 1588-2728 팩스 02-521-1352

<http://www.kiise.or.kr> kiise@kiise.or.kr