Portfolio for Dongin Jung

Dongin Jung, 정동인

dongin1009@gmail.com

0. Profile

- **Experience**
  - 크래프트테크놀로지스 (QRAFT Technologies)
    - LLM/NLP Researcher, AI Tech Lab, 2024.04 ~ 현재
    - LLM/NLP Researcher 인턴, AI Tech Lab, 2024.01 ~ 2024.03
      - 사내에서 수행중인 **금융** 관련 **LLM** 프로젝트의 성능 향상 및 오류 개선을 목표로 연구
      - 금융 도메인 특화 chat-LLM 연구: **금융 용어**와 한국어를 잘 이해할 수 있는 모델 학습 방법 연구
      - 번역/요약 자동 평가: 금융 문서의 번역 및 요약된 결과의 생성 품질을 오픈 **LLM 기반으로 평가** 및 **이유를 설명**할 수 있는 방법 개발

  - SK플래닛 (SK Planet)
    - AI Engineer 인턴, IoT 개발팀, 2023.04 ~ 2023.07
      - 팀 내 수행중인 **CV** 프로젝트에 최신 모델을 업무에 적용해 성능을 개선
      - Anomaly Detection: 여러 종류의 생산 부품의 **불량을 통합으로 탐지**할 수 있는 모델 도입
      - Object Detection & Tracking: **도로** 위의 여러 물체들을 **탐지하고 추적**할 수 있는 최신 모델 도입
      - Weather Removal: 비, 눈, 흐림 등 여러 날씨에서 촬영된 영상의 **노이즈를 제거해 맑은 상태로 복원**하는 최신 통합 모델 도입

0. Profile

- **Education**
  - **석사과정: 중앙대학교** (Chung-Ang University)
    - **AI학과**, Master of Science, 2021.03 ~ 2023.02
    - Supervisor: Yoon-Sik Cho, Data Science Lab @ CAU (full-time)
      - 연구 분야: 자연어 처리 및 이해(NLP&NLU), 그래프 신경망(GNN), 대조 학습(Contrastive Learning) 등
      - 학위 논문: 다중 가짜뉴스 탐지를 위한 데이터 관계 모델링 기법
    - Cumulative GPA: 4.34 / 4.5

  - **학사과정: 원광대학교** (Wonkwang University)
    - **컴퓨터소프트웨어공학과** (주전공), Bachelor of Science, 2017.03 ~ 2021.02
    - Supervisor: Jongmin Lee, Computing System Lab @ WKU (2020.02 ~ 2021.02, 학부연구생)
      - 연구 분야: 한국어 뉴스 수집 및 텍스트 마이닝, 언어 모델을 사용한 감정 분석
    - **디지털포렌식전공** (복수전공), Bachelor of Science, 2018.03 ~ 2021.02
      - 블록체인, 보안, 법학 관련 교과목 수강
    - Cumulative GPA: 3.91 / 4.5

1. Publications

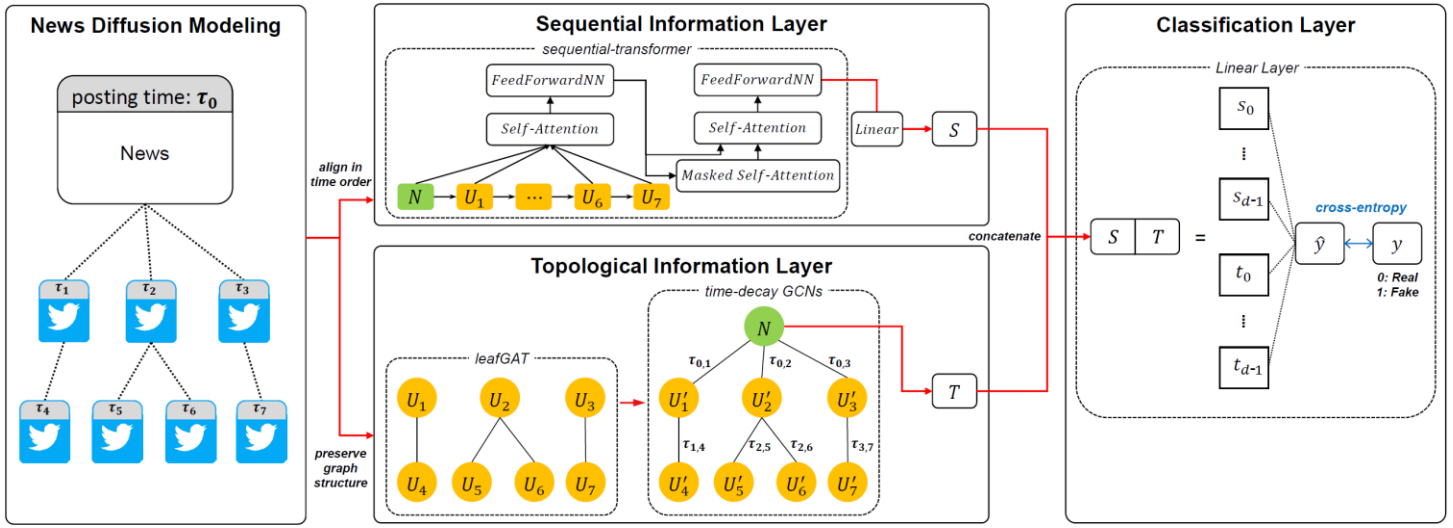
- 1-1. Topological and Sequential Neural Network Model for Detecting Fake News
  - IEEE Access
  - **Dongin Jung**, Eungyeop Kim, and Yoon-Sik Cho
- 1-2. Detecting Documents With Inconsistent Context
  - IEEE Access
  - **Dongin Jung**, Misuk Kim, and Yoon-Sik Cho
- 1-3. Automatic Conversation Turn-Taking Segmentation in Semantic Facet
  - International Conference on Electronics, Information, and Communication (ICEIC) 2023
  - **Dongin Jung**, and Yoon-Sik Cho
- 1-4. 이미지 검색을 위한 대조 학습 모델의 한국어 학습 방법
  - 대한전자공학회 2022년도 하계종합학술대회
  - **정동인**, 김응엽, 강성민, 조윤식
- 1-5. Video Retrieval with Tree-based Video Segmentation
  - International Conference on Database Systems for Advanced Applications (DASFAA 2023)
  - Seongmin Kang, **Dongin Jung**, and Yoon-Sik Cho
- 1-6. Exploiting Component Information with a Context-based Language Model for Effective Bug Triage *(in preparation)*
  - Dae-Sung Wang, **Dongin Jung**, Hyun-Taek Hong,Yoon-Sik Cho, and Chan-Gun Lee
- 1-7. Joint Contrastive and Supervised Learning in Human Activity Recognition *(in preparation)*
  - **Dongin Jung**, Jaehyeok An, and Yoon-Sik Cho

1-1. Topological and Sequential Neural Network Model for Detecting Fake News

- 기여: Fake News 전파의 topological과 sequential information을 결합
  - 핵심 기술: hierarchical GNN, supernode, time-decay edge score
  - 뉴스의 확산 구조를 효과적으로 모델링하기 위해 sequential, topological information을 활용해 가짜 뉴스 탐지
  - supernode로의 의미있는 전파를 위해 leafGAT, temporal-decay GCN으로 2단계의 계층적 GNN을 제안
  - 뉴스 전파 순서에 따라 Transformer를 통한 sequential feature 추출
  - 두 유형의 information을 결합해 기존 baseline 대비 가짜뉴스 탐지 성능 향상
  - https://github.com/dongin1009/TSNN-DFN

TABLE 2. **Fake news detection performance of TSNN and baselines. The evaluation results denoted as † are brought from the original paper because of cannot be reproducible.**

Model	*PolitiFact*		*GossipCop*	
	Accuarcy	F1-score	Accuarcy	F1-score
TSNN (ours)	**92.15**	**92.11**	**97.91**	**97.88**
UPSR† [28]	91.4	91.0	97.7	97.6
UPFD-SAGE [27]	79.75	79.71	97.45	97.43
UPFD-GAT [27]	81.27	81.25	97.38	97.35
UPFD-GCN [27]	82.78	82.71	97.51	97.48
Bi-GCN [26]	82.53	82.45	96.84	96.80
GCNFN [25]	84.81	84.78	95.48	95.42



1-1. Topological and Sequential Neural Network Model for Detecting Fake News

- 기여: Fake News 전파의 topological과 sequential information을 결합
  - (TABLE 3) topological information layer의 time-decay score를 조정하고
  - (TABLE 4) sequential information layer의 네트워크 구조를 변경해
  - 제안한 TSNN의 구조가 두 유형의 information을 효과적으로 추출함을 보임.

TABLE 3. A comparative analysis of the variants of the time-decay function in our time-decay GCNs module. It emphasizes the impact of changes in time unit and graph depth on the model's performance, with minute-based decay offering optimal performance.

Model	*PolitiFact*		*GossipCop*	
	Accuarcy	F1-score	Accuarcy	F1-score
minute-based	**92.15**	**92.11**	**97.91**	**97.88**
(*w/* Depth divide)	90.62	91.09	97.25	97.27
second-based	92.01	91.87	97.85	97.82
(*w/* Depth divide)	91.14	91.10	97.79	97.81
w/o time-decay	89.62	89.59	97.25	97.21
(*w/* Depth divide)	90.81	90.59	97.34	97.30

TABLE 4. Our ablation study demonstrates the impact of different model components on the overall performance of the TSNN model. Key findings show the superior performance of the 2-layered encoders-decoders transformer in capturing sequential information.

Model	*PolitiFact*		*GossipCop*	
	Accuarcy	F1-score	Accuarcy	F1-score
seq-transformer				
(2 enc-2 dec)	**92.15**	**92.11**	**97.91**	**97.88**
(3 enc-3 dec)	90.38	90.31	97.01	96.98
(4 enc-4 dec)	89.37	89.34	97.10	97.07
(4 encoders)	91.65	91.61	97.44	97.41
seq-LSTM	91.39	91.34	97.09	97.05
seq-GRU	91.14	91.10	97.05	97.01

1-2. Detecting Documents With Inconsistent Context

- 기여: 그래프 기반의 문서 내용 불일치 탐지
  - 핵심 기술: supernode, sentence embedding
  - 뉴스 본문 내용의 일관성을 학습해 불일치하는 문서를 탐지하는 모델 제안
  - 각 문장 또는 문단 단위로 분리한 후 sentence-transformer 임베딩 값을 노드 임베딩으로 사용
  - 문서를 대표할 수 있는 'supernode'를 생성해 분류
  - https://github.com/dongin1009/GraDID

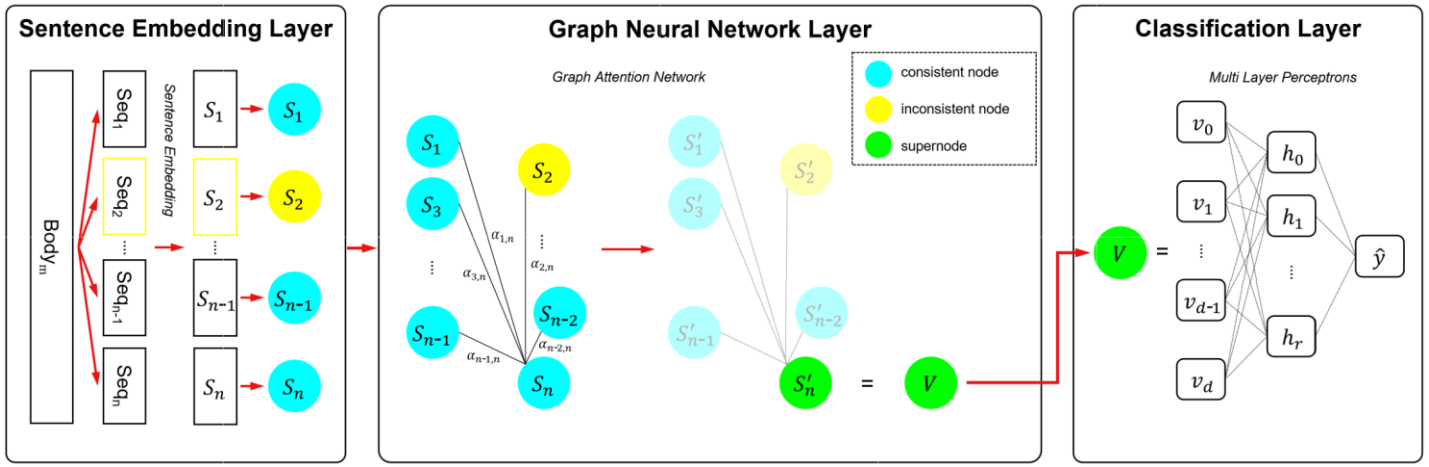


FIGURE 2. Overview of our GraDID model structure applied to inconsistent document. Sentence embedding layer encodes *n* sentences to *n* fixed-sized(*d* dim) vectors by sentence-transformer model. Graph neural network layer performs propagating information between nodes and makes supernode *V*. The edges have a attention score α between nodes. Classification layer classifies the supernode *V* through two-layered MLP.

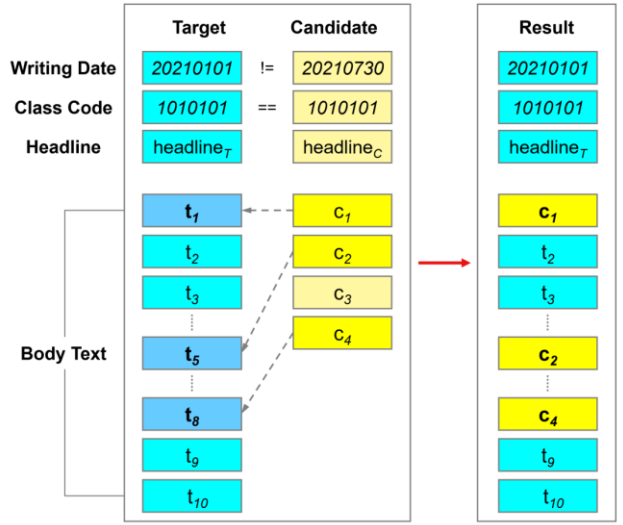


FIGURE 3. Our sentence substitution rules with some conditions. Left (blue) article is selected as target article, middle (yellow) article is chosen as candidate. These have different writing date with at least one common class code. We substitute 30% of the longer articles sentences with those from the shorter article. The substitution term is considered article's sentence length.

1-2. Detecting Documents With Inconsistent Context

- **기여: 그래프 기반의 문서 내용 불일치 탐지**
  - 영어 뉴스 데이터 NELA17과 한글 연합뉴스 데이터로 모델 평가
  - sentence-transformer 모델에 따른 성능 평가
  - 제목과 본문을 모두 사용하는 기존 모델 대비 본문의 일관성 탐지에 더 우수한 성능을 보임

TABLE 2. **GraDID with two variant sentence embedding models and two embedding-levels are compared to previous models on NELA17 dataset. We report the average scores from five runs with different random seeds. † denotes our reimplementation following the previous models with their settings, while other baselines are reported using their results.**

Model	NELA17 Dataset	
	Accuracy	AUROC
GraDID-paragraph		
w/ all-roberta-large-v1	**0.815** ± 0.003	**0.896** ± 0.003
w/ all-mpnet-base-v1	0.810 ± 0.005	0.890 ± 0.004
GraDID-sentence		
w/ all-roberta-large-v1	**0.774** ± 0.004	**0.853** ± 0.005
w/ all-mpnet-base-v1	0.767 ± 0.004	0.848 ± 0.004
POSHAN [19]	0.765	0.783
GHDE† [37]	0.759 ± 0.014	0.842 ± 0.017
MuSeM [18]	0.752	0.769
AHDE† [38]	0.757 ± 0.018	0.832 ± 0.019
BERT-Sent Pair	0.677	0.683
POSAt	0.648	0.669
LSTM	0.642	0.663
SVM	0.622	0.637

TABLE 3. **Proposed GraDID with two variant sentence embedding models in the YH-News dataset. Reported values are average values of five random seed results.**

Model	YH-News Dataset	
	Accuracy	AUROC
GraDID		
w/ sentence-KLUE-RoBERTa	**0.886** ± 0.001	**0.954** ± 0.001
w/ KoSentenceBERT	0.880 ± 0.001	0.949 ± 0.000
w/ LaBSE	0.850 ± 0.002	0.926 ± 0.002

1-3. Automatic Conversation Turn-Taking Segmentation in Semantic Facet

- **기여: Token 레벨의 대화 turn 분할**
  - 핵심 기술: token-level prediction, semantic segmentation
  - 대화의 turn 전환점을 token-level로 예측하는 task 제안
  - 기존 문장 단위 예측 또는 음성 신호(acoustic feature)를 사용한 것과 달리 텍스트의 의미적 측면에서 실시간 전환 예측
  - 구두점, 특수문자 제거 및 알파벳의 소문자화를 통해 실제 real-time voice speaking 상황을 가정
  - Recurrent Neural Networks 대비 GPT-2가 실제 전환점을 더 잘 구분함을 보임
  - https://github.com/dongin1009/semantic-turn-taking-prediction

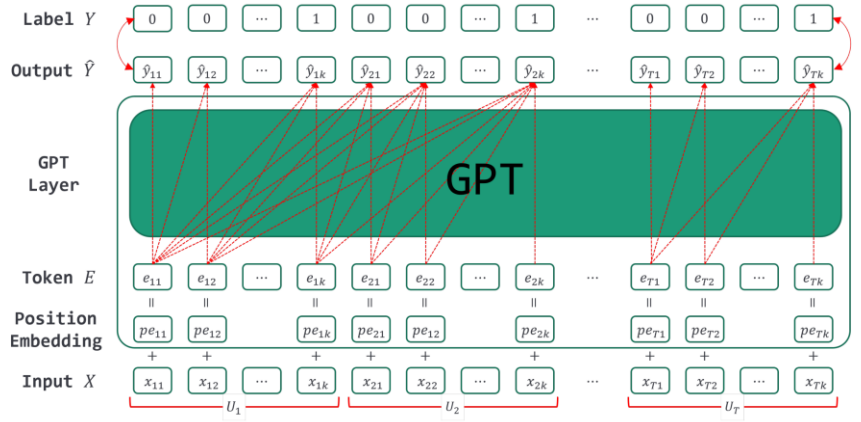


Fig. 2. A pipeline of token-level turn-taking segmentation. Input X is tokens from the tokenizer, and the model predicts in each token level whether the token is end-of-turn or not. In the current time point prediction, the future time point tokens are not used.

TABLE II
PERFORMANCE OF THE MODELS IN MULTIWOZ AND DAILYDIALOG.

Model	MultiWoZ				DailyDialog			
	lr	R	P	F1	lr	R	P	F1
GRU	1e-3	21.5	63.2	32.0	1e-4	28.7	58.1	38.4
LSTM	1e-3	33.9	65.7	44.7	1e-4	32.1	59.9	41.8
GPT-2	1e-5	**68.4**	**70.9**	**69.7**	1e-4	**60.7**	**62.9**	**61.8**

	actual conversation	GRU	LSTM	GPT-2
Multi WoZ	i need train reservations from nor wich to cam bridge // i have 133 trains matching your request is there a specific day and time you would like to travel // id like to leave on m onday and arrive by 1800 // there are 12 trains for the day and time you request would you like to book it now // before booking i would also like to know the travel time price and departure time please // there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these // no hold off on booking for now can you help me find an attraction called c in eworld cinema // yes it is a cinema located in the south part of town what information would you like on it // yes that was all I needed thank you very much // thank you for using our system	i need train reservations from nor wich to cam bridge i have 133 trains matching your request is there a specific day and time you would like to travel // [70.31%] id like to leave on m onday and arrive by 1800 // [71.35%] there are 12 trains for the day and time you request would you like to book it now // [72.77%] before booking i would also like to know the travel time price and departure time please // [74.53%] there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these // [69.79%] no hold off on booking for now // [71.62] can you help me find an attraction called c in eworld cinema yes it is a cinema located in the south part of town what information would you like on it // [71.96] yes that was all i needed thank you very much thank you for using our system	i need train reservations from nor wich to cam bridge // [89.71%] i have 133 trains matching your request is there a specific day and time you would like to travel // [93.70%] id like to leave on m onday and arrive by 1800 // [88.07%] there are 12 trains for the day and time you request would you like to book it now // [72.22%] before booking i would also like to know the travel time price and departure time please // [85.21%] there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these no hold off on booking for now can you help me find an attraction called c in eworld cinema yes it is a cinema located in the south part of town what information would you like on it // [91.95] yes that was all i needed thank you very much thank you for using our system	i need train reservations from nor wich to cam bridge // [54.12%] i have 133 trains matching your request is there a specific day and time you would like to travel // [73.99%] id like to leave on m onday and arrive by 1800 // [70.06%] there are 12 trains for the day and time you request would you like to book it // [57.14%] now // [87.52%] before booking i would also like to know the travel time price and departure time please // [88.65%] there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these // [78.43%] no hold off on booking for now can you help me find an attraction called c in eworld cinema // [71.26%] yes it is a cinema located in the south part of town what information would you like on it // [87.05%] yes that was all i needed thank you // [60.45%] very much // [65.44%] thank you for using our system
Daily Dialog	do you have maps of downtown area // yes here you are // how much is it // its free of charge // thanks so much //	do you have maps of downtown area yes here you are how much is it // [56.63%] its free // [57.91%] of charge // [58.73%] thanks so much // [52.60%]	do you have maps of downtown area // [72.55%] yes here you are // [68.82%] how much is it // [71.82%] its free of charge // [67.17%] thanks so much // [60.81%]	do you have maps of downtown area // [80.61%] yes here you are // [73.11%] how much is it // [52.79%] its free of charge // [64.57%] thanks // [58.60%] so much // [74.00%]

Fig. 3. Sample turn-taking segmentation output of models. Blue probabilities are denote turn ending probabilities, and reds denote incorrect predictions.

1-4. 이미지 검색을 위한 대조 학습 모델의 한국어 학습 방법

- **기여: 한국어 Language Model을 사용한 Ko-CLIP**
  - 핵심 기술: multimodal contrastive learning, prompt
  - Language-Image 쌍을 contrastive learning으로 학습한 CLIP의 Language 파트의 transformer를 한국어 모델로 대체
  - 데이터셋이 현저히 적은 한국어 텍스트-이미지 쌍을 효과적으로 학습하기 위해 language 모델에는 KLUE-RoBERTa-Large, visual 모델에 각각 Vision Transformer(ViT)와 ResNet을 사용
  - CIFAR의 한국어 텍스트에 Prompt Engineering 적용
    V1: '이것은 {text}이다.'
    V2: '이것은 {text}의 사진이다.'
  - https://github.com/dongin1009/Ko-CLIP

표 1. 한글 CIFAR 데이터셋에 제로-샷 분류 성능

	CIFAR-10(Korean)		CIFAR-100(Korean)	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
ViT-ver	88.75	99.49	34.26	60.29
ResNet-ver	59.37	95.75	21.08	44.85

1-5. Video Retrieval with Tree-based Video Segmentation

- 기여: 비디오 검색 작업에서 비디오를 프레임 변화에 집중해 표현
  - 핵심 기술: multimodal learning, video segmentation
  - 비디오-텍스트 검색에서 프레임들을 mean-pooling해 비디오를 표현하는 방식이 아닌 비디오의 장면 전환을 기점으로 sub-tree video를 만드는 방식을 제안
  - 텍스트 쿼리에 따른 비디오를 찾는 Recall 평가지표에서 기존 mean-pooling 대비 정확한 검색 결과를 보임

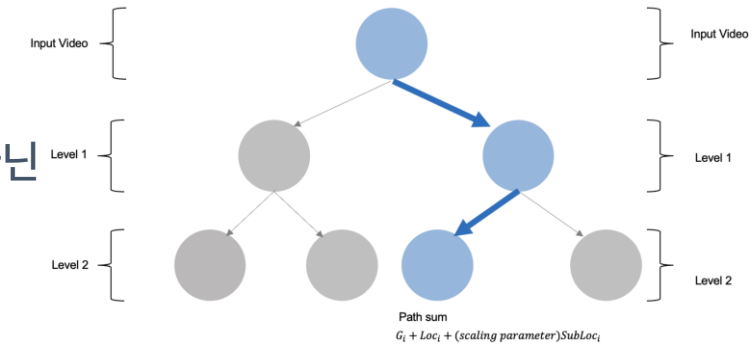


Fig. 3. Tree-based video segmentation and its global-local aggregation method

Table 1. Text-to-video retrieval performance on MSR-VTT [35] dataset with 7k-training data through data division of [25]. We compare with original CLIP4Clip [24] with two video aggregation methods. The meanP performs mean-pooling over images (parameter-free), and the seqTransf uses mean-pooling to aggregates frames obtained through transformer structure. For the performance metric, higher R@k is better, and lower MnR is better. The best performance is bold and the second best is underscored.

Methods	meanP				seqTransf			
	R@1↑	R@5↑	R@10↑	MnR↓	R@1↑	R@5↑	R@10↑	MnR↓
CLIP4Clip [24]	**43.6**	68.5	78.9	17.2	42.5	69.0	78.9	17.9
Ours (Div2)	**43.6**	69.4	79.3	16.5	42.9	**69.5**	80.1	**15.8**
Ours (Div4)	**43.6**	**69.8**	**79.7**	**15.5**	**43.7**	**69.5**	**80.4**	16.5
Test in Global(Div2)	43.5	68.4	78.7	17.7	42.2	68.7	78.5	17.8
Test in Global(Div4)	42.7	68.9	79.5	16.6	42.9	68.8	78.8	17.6

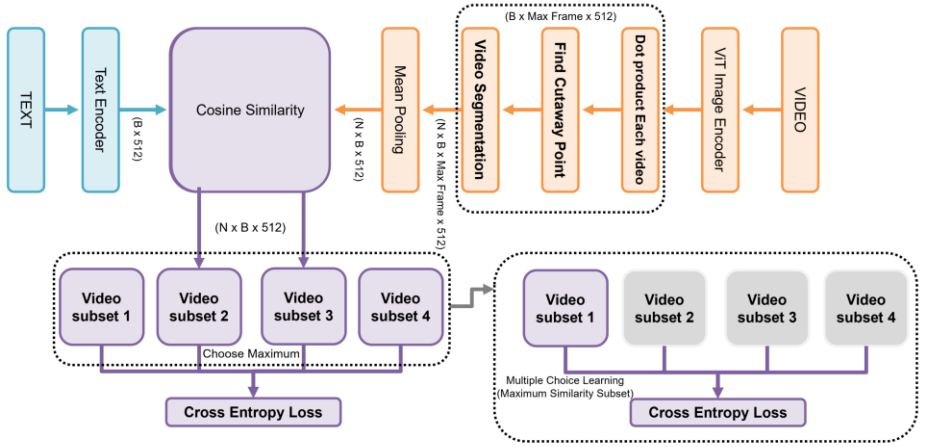


Fig. 1. The overall structure of proposed model based on CLIP4Clip [24]. Our methods are highlighted in a box with dotted lines, where other modules are borrowed from CLIP4Clip [24]. The content in parentheses is the dimension of the video. The output for final learning uses the concept of multiple choice learning to become a specific subset of video judged to have the most significant similarity to the text within the batch size.

1-6. Exploiting Component Information with a Context-based Language Model for Effective Bug Triage

- 기여: SW 이슈 리포트의 component feature를 사용해 버그 담당자 자동 예측
  - 핵심 기술: component embedding, Pretrained Language Model
  - 이슈 리포트의 textual feature(title+description)를 BERT를 사용해 developer(담당자)를 할당
  - BERT 임베딩에 component feature와 component embedding을 추가한 'Comp-BIAR' 모델 제안
  - 기존 word embedding based model 대비 developer assignment 성능 향상
  - component와 developer간의 상관관계 분석

Table 3: Summary of experimental results on previous and new issue datasets.

Source	Dataset	Method	R@1	R@5	R@10
Mani *et al.* [13]	*Core*	DBRNN-A	28.61	55.28	66.42
		DBRNN-A(GN)	15.37	31.84	40.91
		BIAR	34.37	61.11	71.28
		Comp-BIAR	**37.57**	**65.73**	**75.54**
	Firefox	DBRNN-A	28.46	57.42	68.90
		DBRNN-A(GN)	22.12	45.82	57.22
		BIAR	32.68	62.01	73.88
		Comp-BIAR	**35.88**	**67.20**	**77.87**
Wang *et al.* [20]	*Platform*	DBRNN-A(GN)	50.64	71.16	77.67
		BIAR	60.34	81.13	87.81
		Comp-BIAR	**61.25**	**85.16**	**92.45**
	Foundation	DBRNN-A(GN)	67.43	88.95	93.55
		BIAR	74.30	92.99	96.82
		Comp-BIAR	**83.94**	**96.53**	**99.02**
	Firefox New	DBRNN-A(GN)	31.93	60.89	72.83
		BIAR	54.24	75.76	82.95
		Comp-BIAR	**57.73**	**81.67**	**88.76**

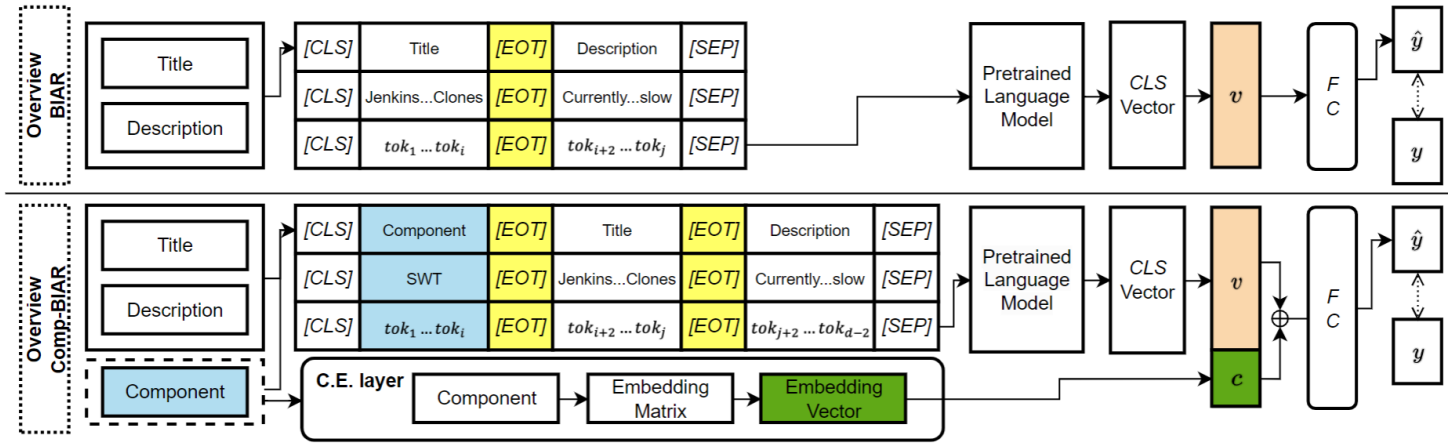


Figure 3: The structure of BIAR and Comp-BIAR. *C.E. layer: Component Embedding layer

1-7. Joint Contrastive and Supervised Learning in Human Activity Recognition

- 기여: Supervised Contrastive Learning과 Supervised Learning을 결합
  - 핵심 기술: Supervised Contrastive Learning
  - 모델의 representation을 강화하기 위해 supervised contrastive learning(SupCon) 도입
  - 기존 SupCon이 contrastive learning과 supervised learning을 2-stage로 학습한 것과 달리 두 학습 방식을 결합해 동시에 학습하는 Joint SupCon 방식을 제안
  - 비슷한 human activity 사이의 구별을 더 명확히 할 수 있음을 보임
  - https://github.com/dongin1009/joint_supcon_har

Table 2. Performance comparison of our methods on a PAMAP2 dataset. We report the averaged accuracy and f1-score for five distinct random seeds.

Model	PAMAP2					
	Original		SupCon		Joint SupCon	
	ACC	F1	ACC	F1	ACC	F1
M1	$91.16_{\pm 0.12}$	$90.95_{\pm 0.14}$	$\mathbf{92.02_{\pm 0.36}}$	$\mathbf{91.75_{\pm 0.38}}$	$91.80_{\pm 0.40}$	$91.58_{\pm 0.50}$
M2	$95.65_{\pm 0.24}$	$95.45_{\pm 0.30}$	$96.12_{\pm 0.23}$	$96.02_{\pm 0.25}$	$\mathbf{96.36_{\pm 0.19}}$	$\mathbf{96.16_{\pm 0.21}}$
M3	$97.22_{\pm 0.50}$	$97.12_{\pm 0.53}$	$97.51_{\pm 0.27}$	$\mathbf{97.33_{\pm 0.25}}$	$\mathbf{97.80_{\pm 0.12}}$	$97.11_{\pm 0.17}$

Table 3. Performance comparison of our methods on a WISDM dataset. We report the average and standard deviation of accuracy and f1-score for five different random seeds in each model.

Model	WISDM					
	Original		SupCon		Joint SupCon	
	ACC	F1	ACC	F1	ACC	F1
M1	$\mathbf{93.63_{\pm 0.65}}$	$\mathbf{91.35_{\pm 0.80}}$	$91.32_{\pm 0.36}$	$88.05_{\pm 0.28}$	$93.62_{\pm 0.98}$	$91.11_{\pm 1.11}$
M2	$90.97_{\pm 0.65}$	$87.40_{\pm 0.97}$	$91.33_{\pm 0.22}$	$88.37_{\pm 0.46}$	$\mathbf{92.46_{\pm 0.50}}$	$\mathbf{89.78_{\pm 0.82}}$
M3	$95.65_{\pm 0.30}$	$94.21_{\pm 0.51}$	$97.05_{\pm 0.40}$	$96.08_{\pm 0.47}$	$\mathbf{97.23_{\pm 0.25}}$	$\mathbf{96.44_{\pm 0.41}}$

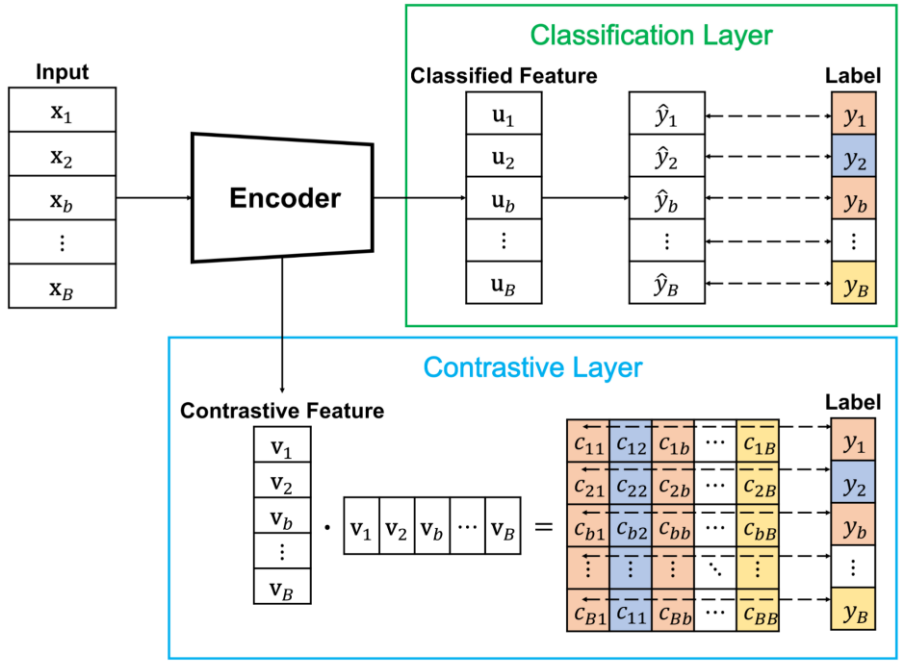
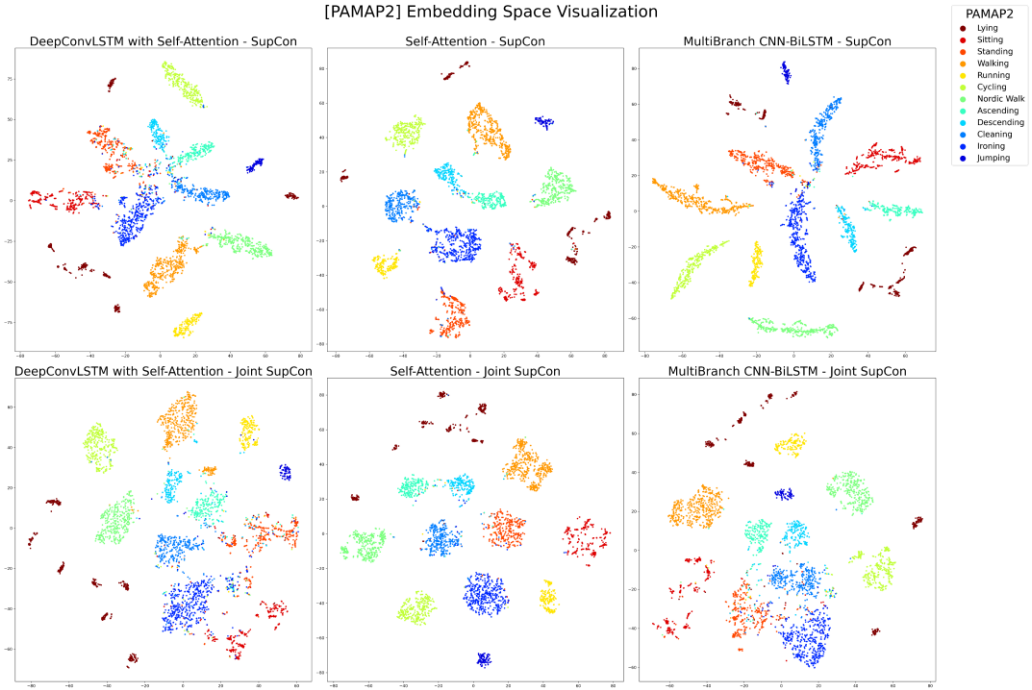
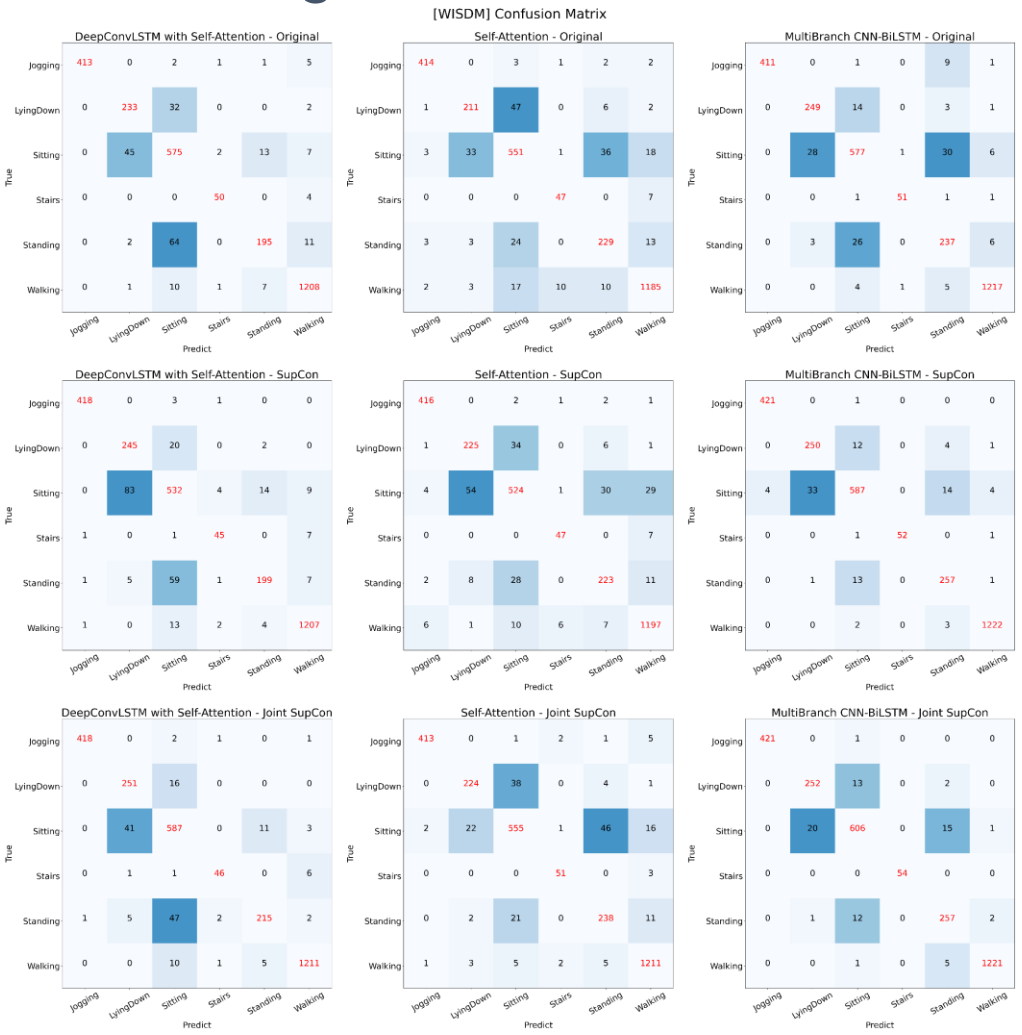


Fig. 2. Overview of joint contrastive and supervised learning method. Our method comprises the encoder, contrastive layer, and classification layer.

1-7. Joint Contrastive and Supervised Learning in Human Activity Recognition

- 기여: Supervised Contrastive Learning과 Supervised Learning을 결합
  - SupCon은 입력 데이터 표현의 alignment에만 집중을 하는 반면, Joint SupCon은 cross entropy loss를 동시에 학습
  - cross entropy loss만 사용하는 original method 대비 Joint SupCon method에서 클래스간 예측 모호성을 완화





2. Projects

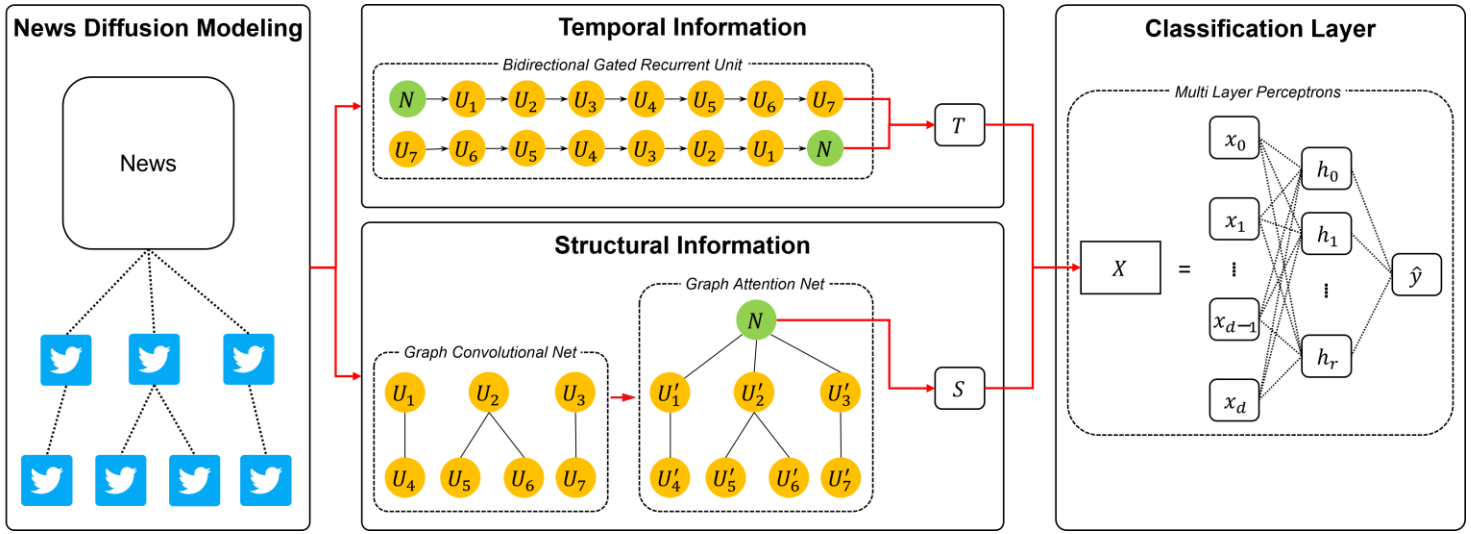
- 2-1. ‘실감형 뉴스’를 위한 빅데이터 분석 참여형 통합 플랫폼 구축
  - 한국콘텐츠진흥원 연구과제
  - 2021.03 ~ 2022.12
- 2-2. 뇌인지-행동 데이터 기반 비대면 정신건강문제 실시간 인터랙티브 셀프 모니터링 HW/SW 플랫폼 개술개발
  - 정보통신기획평가원 연구과제
  - 2021.03 ~ 2021.12
- 2-3. Polyglot: Multilingual Large Language Models
  - EleutherAI 프로젝트: Polyglot - Romance group
  - 2022.10 ~ 진행중
- 2-4. KR3: Korean Restaurant Review with Ratings dataset
  - DIYA 동아리 프로젝트
  - 2021.04 ~ 2021.12
- 2-5. Development of an Online Learning Management System
  - 원광대학교 소프트웨어중심대학사업단 내주 과제
  - 2020.06 ~ 2020.12

2-1. '실감형 뉴스'를 위한 빅데이터 분석 참여형 통합 플랫폼 구축

- Text Fake News Detection
  - **기여: 가짜뉴스를 일관성 불일치 뉴스, 허위 정보 뉴스, 2가지 유형으로 정의해 탐지 방법을 제안**
- 허위 정보 뉴스 탐지 (Non-Fact News Detection)
  - Fake News(Non-Fact News)와 Real News는 SNS를 통해 독자들에게 확산되는 유형이 다름
  - 핵심 기술: GNN을 통해 확산의 구조적 정보를, bi-GRU를 통해 확산의 시간적 정보를 모델링
  - User Preference-aware Diffusion Structural and Temporal (UPDST) fake news detection model 제안
  - 구조적 정보에 leafGNN과 supernode를 활용해 기존 모델 대비 더 높은 성능을 보임

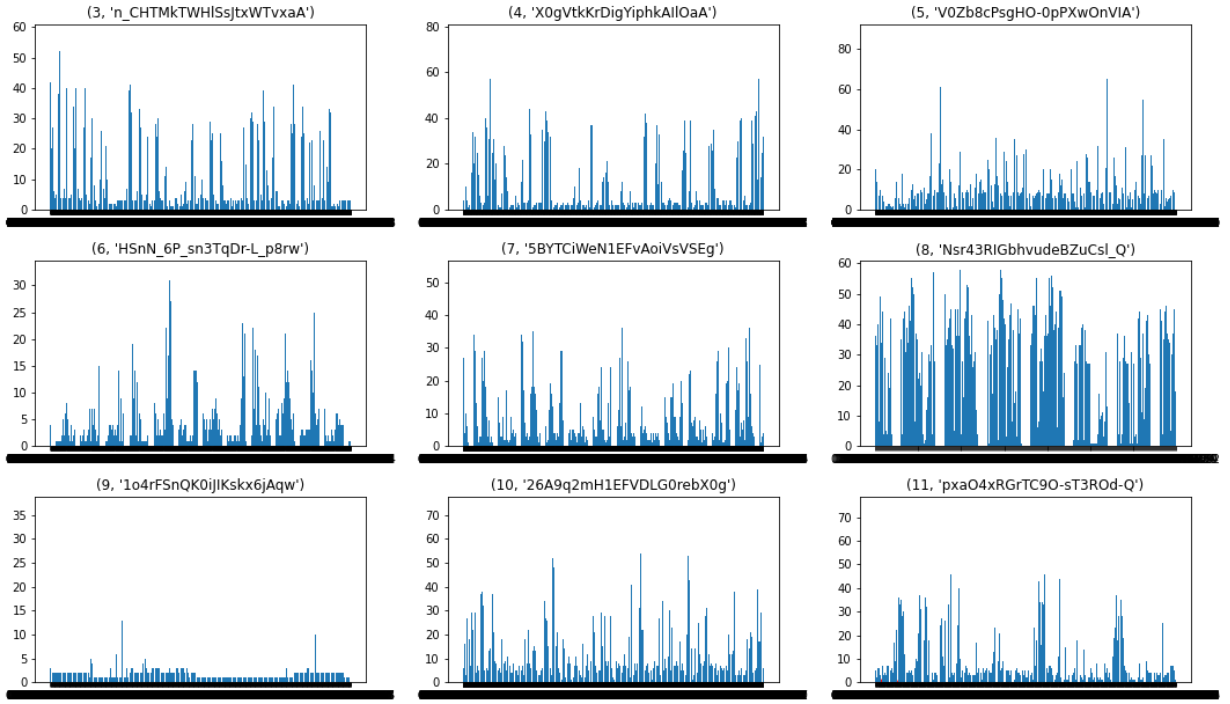
Model	*PolitiFact*		
	Accuracy	F1-score	AUROC
UPDST (ours)	**88.61**	**88.54**	**96.22**
UPFD-SAGE [4]	82.28	82.28	89.81
UPFD-GAT [4]	84.81	84.81	88.27
UPFD-GCN [4]	86.08	86.07	92.63
Bi-GCN [1]	83.54	83.53	89.68
GCN-FN [14]	87.34	87.34	92.37

Model	*GossipCop*		
	Accuracy	F1-score	AUROC
UPDST (ours)	**97.00**	**96.96**	98.54
UPFD-SAGE [4]	96.71	96.67	**99.18**
UPFD-GAT [4]	95.75	95.70	98.78
UPFD-GCN [4]	**97.00**	**96.96**	98.72
Bi-GCN [1]	96.56	96.52	98.99
GCN-FN [14]	96.63	96.59	98.90



2-2. 뇌인지-행동 데이터 기반 비대면 정신건강문제 실시간 인터랙티브 셀프 모니터링 HW/SW 플랫폼 개술개발

- **기여: 스마트폰 센서 정보로 우울증 예측**
  - 사용자의 GPS, Bluetooth, sms, call 등의 feature를 통해 우울증을 예측
  - 대면 진료가 아닌 비대면 사용자의 정보를 통해 진단 예측이 가능
  - 각 feature별 특성에 맞는 feature selection
  - 데이터 양이 적은 문제를 해결하기 위해 feature analysis 및 ML 모델로 classification



2-3. Polyglot: Multilingual Large Language Models

- 한국어 text generation model 개발 - Korean Pretraining Team (진행중)
  - **기여: Polyglot 모델의 한국어 데이터 전처리 진행**
  - 핵심 기술: token count-based dataset curation, pySpark
  - LLM 학습에 필요한 데이터 구성
  - 여러 유형의 tokenizer 토큰 개수 기반 데이터 추출
  - polyglot-ko-v2 배포를 목표로 모델 학습 예정

- 다국어 text generation model 개발 - Romance language group
  - **기여: 다국어를 지원하는 Polyglot 모델의 Romance language (Spanish, French, Italian, Portuguese, Romanian) 그룹에 속해 데이터 수집 및 전처리 진행**
  - 핵심 기술: Data collection & preprocessing
  - 스페인 계열 외국 NLP 연구자, 개발자와 협업 진행
  - GPT-NeoX 구조 기반의 text generation model 개발 예정
  - Korean, East Asia, Romance 언어 계열의 특징 및 상관관계를 language model의 관점에서 분석 예정
  - https://github.com/EleutherAI/polyglot-data

2-4. KR3: Korean Restaurant Review with Ratings dataset

- 한글 맛집 리뷰 데이터 구축
  - **기여: 맛집 소개 플랫폼의 리뷰 및 평점 데이터를 크롤링해 맛집 리뷰 긍부정 데이터 구축**
  - https://github.com/yejoon-lee/kr3

Distribution

label	#(samples)
0 (Negative)	70910
1 (Positive)	388111
2 (Ambiguous)	(+182741)
Total	**459021(+182741)**

- **Task-Adaptive PreTraining 실험**
  - 핵심 기술: Task-Adaptive PreTraining, Data Analysis
  - BERT를 한글 리뷰 데이터로 추가 학습해 리뷰 domain의 task를 잘 수행하도록 representation 강화
  - 다른 task 리뷰 데이터인 NSMC(영화), Naver Shopping(쇼핑), Steam(게임) 데이터셋 집단과 cross-task로 추가 pretraining 및 sentiment classification fine tuning 수행
  - pretraining 데이터가 cross-task임에도 추가 pretraining을 수행한 모델의 긍부정 분류 정확도가 더 높음을 확인

Effect of additional pretraining

fine-tune \ additional pre-train	KR3		NSMC	
	F1 (macro)	Acc.	F1 (macro)	Acc.
no pre-training (original bert)	0.8709	0.9348	0.8616	0.8617
KR3	-	-	0.8748	0.8749
NSMC + Naver Shopping + Steam	0.9325	0.9653	-	-

2-5. Development of an Online Learning Management

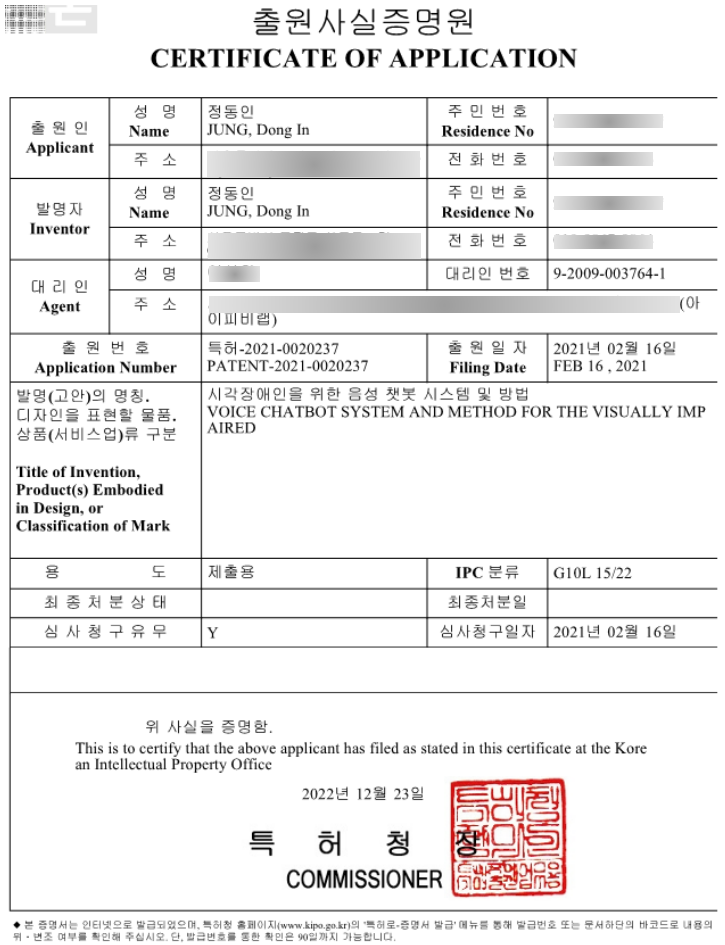
- 교내 온라인 강의 LMS 플랫폼 구축
  - 기여: 코로나19 비대면 수업의 확산으로 온라인 강의의 수요 증가에 따라 교내 특강을 서비스할 수 있는 LMS 플랫폼 구축
  - 핵심 기술: **moodle** 플랫폼을 기반으로 커스터마이징 수행
  - **Apache Tomcat, PHP, MariaDB, HTTPS/SSL**
  - 서비스화에 따른 보안 취약점 개선, 부하 테스트, 유지보수 등 수행
  - https://swpride.wku.ac.kr/

3. Patents

- 3-1. 시각장애인을 위한 음성 챗봇 시스템 및 방법
  - 국내 특허 출원 (2022)
  - 발명자: **정동인**
- 3-2. 스마트 화분
  - 국내 특허 등록 (2021), SW 저작권 등록 (2020)
  - 발명자: **정동인** 외 5인

3-1. 시각장애인을 위한 음성 챗봇 시스템 및 방법

- STT를 적용한 seq2seq 챗봇
  - 특허출원 제10-2021-0020237호
  - 기존 챗봇 시스템의 seq2seq 구조에서 입출력 단에 **Speech-to-Text(STT)**와 **Text-to-Speech(TTS)**를 결합해 시각장애인도 음성으로 사용이 가능한 챗봇 시스템 아이디어 제안
  - https://patents.google.com/patent/KR20220116859A



출원사실증명원

CERTIFICATE OF APPLICATION

출원인 **Applicant**	성 명 **Name**	정동인 JUNG, Dong In	주민번호 **Residence No**	[redacted]
	주 소	[redacted]	전화번호	[redacted]
발명자 **Inventor**	성 명 **Name**	정동인 JUNG, Dong In	주민번호 **Residence No**	[redacted]
	주 소	[redacted]	전화번호	[redacted]
대리인 **Agent**	성 명	[redacted]	대리인 번호	9-2009-003764-1
	주 소	[redacted] (아이피비랩)		
출원번호 **Application Number**		특허-2021-0020237 PATENT-2021-0020237	출원일자 **Filing Date**	2021년 02월 16일 FEB 16 , 2021
발명(고안)의 명칭. 디자인을 표현할 물품. 상품(서비스업)류 구분 **Title of Invention, Product(s) Embodied in Design, or Classification of Mark**		시각장애인을 위한 음성 챗봇 시스템 및 방법 VOICE CHATBOT SYSTEM AND METHOD FOR THE VISUALLY IMPAIRED		
용도		제출용	IPC 분류	G10L 15/22
최종처분상태			최종처분일	
심사청구유무		Y	심사청구일자	2021년 02월 16일

위 사실을 증명함.

This is to certify that the above applicant has filed as stated in this certificate at the Korean Intellectual Property Office

2022년 12월 23일

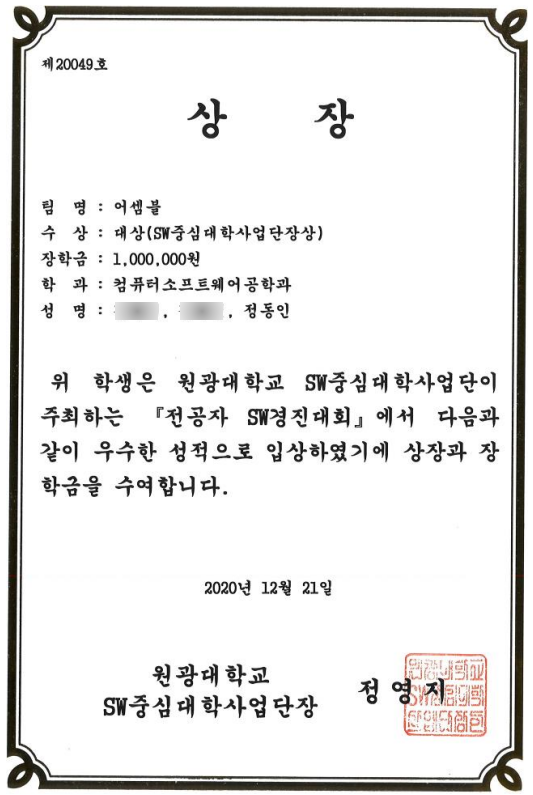
특 허 청 장

COMMISSIONER

◆ 본 증명서는 인터넷으로 발급되었으며, 특허청 홈페이지(www.kipo.go.kr)의 '특허로-증명서 발급' 메뉴를 통해 발급번호 또는 문서하단의 바코드로 내용의 위·변조 여부를 확인해 주십시오. 단, 발급번호를 통한 확인은 90일까지 가능합니다.

3-2. 스마트 화분 장치

- IOT 모듈과 앱을 통한 독거 노인 관리형 스마트 화분
  - 특허등록 제10-2170452호, SW 저작권 등록 제 2020-038703호
  - 식물을 키울 수 있는 화분에 각종 **인터랙티브 모듈을 추가**해 스마트 화분 제안
  - 화분 장치의 센서 정보를 추적할 수 있는 앱을 통해 **독거 노인을 관리**할 수 있는 시스템 제안
  - https://patents.google.com/patent/KR102170452B1



제20049호

상 장

팀 명 : 어셈블
수 상 : 대상(SW중심대학사업단장상)
장학금 : 1,000,000원
학 과 : 컴퓨터소프트웨어공학과
성 명 : [redacted], [redacted], 정동인

위 학생은 원광대학교 SW중심대학사업단이 주최하는 『전공자 SW경진대회』에서 다음과 같이 우수한 성적으로 입상하였기에 상장과 장학금을 수여합니다.

2020년 12월 21일

원광대학교
SW중심대학사업단장 정 영 지

제 C-2020-038703 호

저작권 등록증

1. 저작물의 제호(명칭)	IOT 시스템을 활용한 독거노인 관리 프로그램 (아이오티 시스템을 활용한 독거노인 관리 프로그램)
2. 저작물의 종류	컴퓨터프로그램저작물>시스템프로그램>데이타통신>유.무선 통신 응용 S/W
3. 저작자 성명(법인명)	정동인 외 4인(별지기재) 전라북도
4. 생년월일(법인등록번호)	1998년07월26일 (별지기재)
5. 창작연월일	2020년06월01일
6. 공표연월일	-
7. 등록연월일	2020년10월30일
8. 등록사항	공동저자 : 정동인, [redacted] 창작 : 2020.06.01

「저작권법」 제53조에 따라 위와 같이 등록되었음을 증명합니다.

2020년 11월 02일

한국저작권위원회

특허증
CERTIFICATE OF PATENT

특 허 Patent Number 제 10-2170452 호
출원번호 Application Number 제 10-2020-0029729 호
출원일 Filing Date 2020년 03월 10일
등록일 Registration Date 2020년 10월 21일

발명의 명칭 Title of the Invention
스마트 화분 장치

특허권자 Patentee
정동인(980726-*******)
전라북도 [redacted]

발명자 Inventor
등록사항란에 기재

위의 발명은 「특허법」에 따라 특허등록원부에 등록되었음을 증명합니다.
This is to certify that, in accordance with the Patent Act, a patent for the invention has been registered at the Korean Intellectual Property Office.

2020년 10월 21일

특허청
Korean Intellectual Property Office

특허청장
COMMISSIONER,
KOREAN INTELLECTUAL PROPERTY OFFICE
김 용 래

4. Awards

- 4-1. Dacon '동서발전 태양광 발전량 예측 AI 경진대회'
  - 11등 / 234명
  - 한국동서발전 (2021.06.09 ~ 2021.07.09)
- 4-2. Dacon '물류 유통량 예측 경진대회'
  - 29등 / 237명
  - 국토연구원 (2021.12.06 ~ 2021.12.20)

- 4-3. MIND 창의역량상
  - 원광대학교 (2021)
- 4-4. 전공자 SW경진대회 대상
  - 원광대학교 SW중심대학사업단 (2020)
- 4-5. 학생 포트폴리오 경진대회 동상
  - 원광대학교 (2019)
- 4-6. 창의소프트웨어 아이디어 경진대회 동상
  - 원광대학교 (2019)
- 4-7. 기업연계 캡스톤 디자인 경진대회 동상
  - 원광대학교 (2019)

4-1. Dacon '동서발전 태양광 발전량 예측 AI경진대회'

- 동서발전 태양광 발전량 예측 AI경진대회 (2021.06.09 ~ 2021.07.09)
  - https://dacon.io/competitions/official/235720/overview
  - 당진, 울산 지역의 과거 날씨, 발전량을 기반으로 미래의 발전량을 회귀 예측
  - https://github.com/dongin1009/Solar_Predict

동서발전 태양광 발전량 예측 AI 경진대회

알고리즘 | 정형 | 회귀 | 에너지 | NMAE

11등 / 234

- **기여: 시계열 데이터 가공**
  - NaN 값 처리를 위해 1) linear interpolation, 2) observation-based interpolation 등 시도.
  - 일정 구간의 연속적인 NaN 값(ex. 한달 간 기온 측정이 안된 경우)을 선형 보간으로 채울 경우 보간 오류를 범할 수 있어 관측된 값(ex. 온도가 아닌 풍량, 습도 등) 을 기준으로 보간 규칙을 선정해 column 기준이 아닌 row 기준 보간을 사용
- **기여: LSTM 기반 베이스라인 개발**
  - 최종 모델 XGB 기반 모델과 비교를 위한 LSTM 기반 발전량 예측 모델 개발
  - 1시간(6*10분)을 timestep으로 하는 timeseries LSTM 구현
  - 최종 모델의 하이퍼파라미터 튜닝 수행

4-2. Dacon '물류 유통량 예측 경진대회'

- 물류 유통량 예측 경진대회 (2021.12.06 ~ 2021.12.20)
  - https://dacon.io/competitions/official/235867/overview
  - 제주도 택배 운송 데이터를 이용해 물류 유통량을 회귀 예측

물류 유통량 예측 경진대회

알고리즘 | 정형 | 회귀 | 운송량 | RMSE

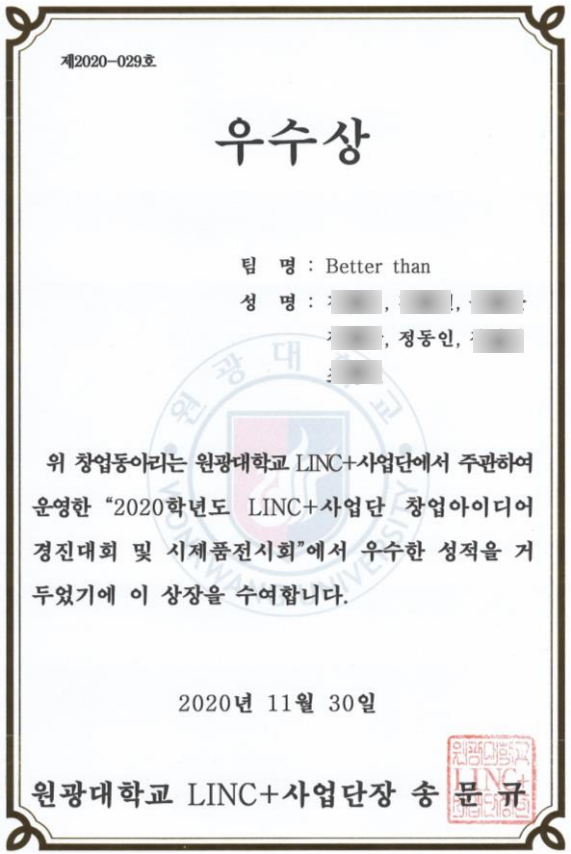
29등 / 237

- **기여: 데이터 전처리**
  - 물품 카테고리를 여러 형태로 표현해 유용한 feature로 인코딩
  - BinaryEncoding, OneHotEncoding 중 OneHotEncoding이 7종류의 카테고리를 더 잘 표현

- 모델링
  - **최종 모델은 XGBoost가 test case에서 가장 좋은 성능**
  - GridSearch로 하이퍼파라미터 튜닝
  - XGB, RandomForest, LightGBM을 **soft voting**한 모델 구현, 그러나 단일 케이스에서 가장 좋은 결과

4. 기타 수상 내역

- LINC+ 창업아이디어 경진대회 및 시제품 전시회 우수상 수상
- 학생 포트폴리오 경진대회 동상 수상
- 창의소프트웨어 아이디어 경진대회 동상 수상
- 기업연계 캡스톤 디자인 경진대회 동상 수상



제2020-029호

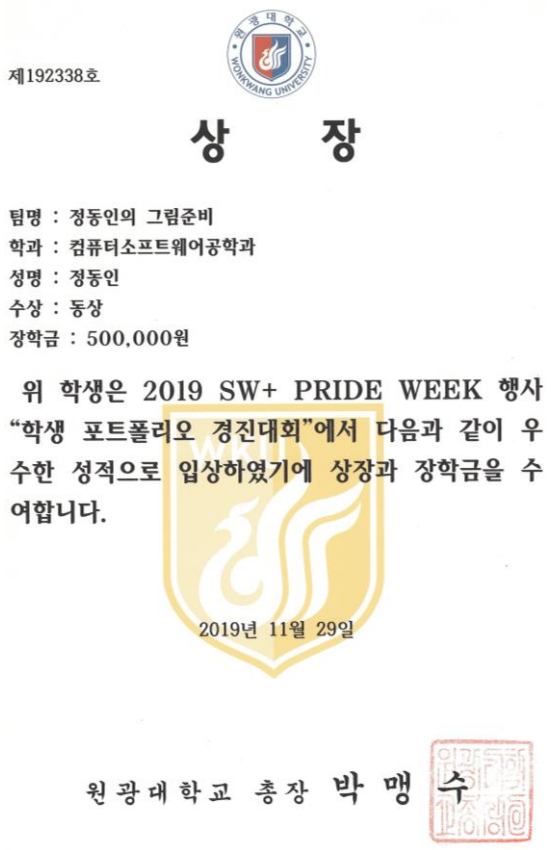
우수상

팀 명 : Better than
성 명 : [redacted], 정동인, [redacted]

위 창업동아리는 원광대학교 LINC+사업단에서 주관하여 운영한 "2020학년도 LINC+사업단 창업아이디어 경진대회 및 시제품전시회"에서 우수한 성적을 거두었기에 이 상장을 수여합니다.

2020년 11월 30일

원광대학교 LINC+사업단장 송 문 규



제192338호

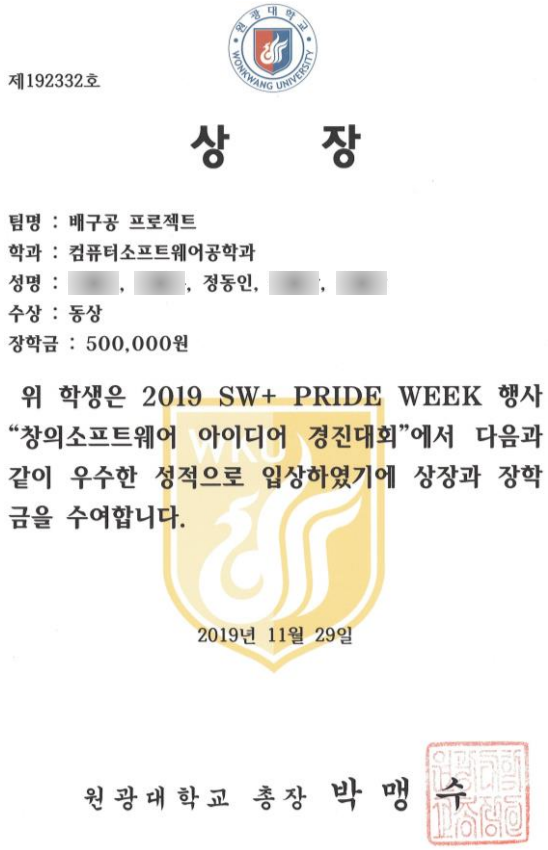
상 장

팀명 : 정동인의 그림준비
학과 : 컴퓨터소프트웨어공학과
성명 : 정동인
수상 : 동상
장학금 : 500,000원

위 학생은 2019 SW+ PRIDE WEEK 행사 "학생 포트폴리오 경진대회"에서 다음과 같이 우수한 성적으로 입상하였기에 상장과 장학금을 수여합니다.

2019년 11월 29일

원광대학교 총장 박 맹 수



제192332호

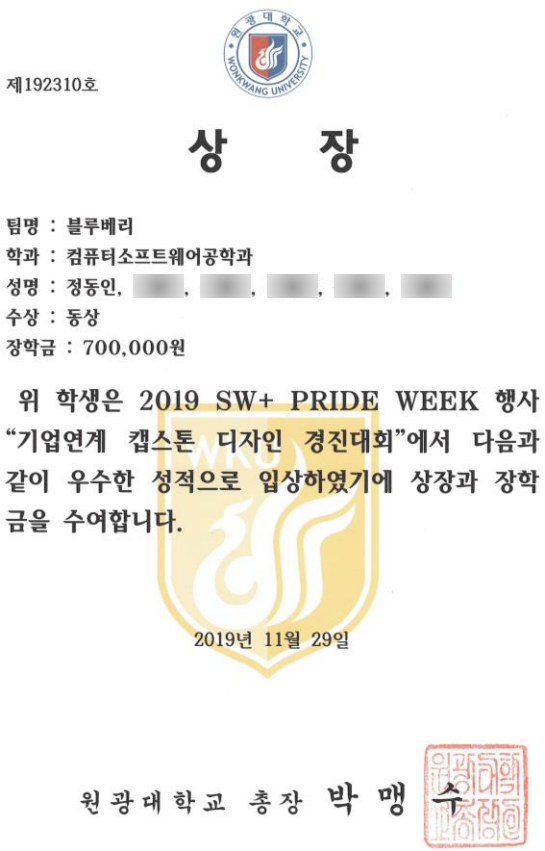
상 장

팀명 : 배구공 프로젝트
학과 : 컴퓨터소프트웨어공학과
성명 : [redacted], 정동인, [redacted]
수상 : 동상
장학금 : 500,000원

위 학생은 2019 SW+ PRIDE WEEK 행사 "창의소프트웨어 아이디어 경진대회"에서 다음과 같이 우수한 성적으로 입상하였기에 상장과 장학금을 수여합니다.

2019년 11월 29일

원광대학교 총장 박 맹 수



제192310호

상 장

팀명 : 블루베리
학과 : 컴퓨터소프트웨어공학과
성명 : 정동인, [redacted]
수상 : 동상
장학금 : 700,000원

위 학생은 2019 SW+ PRIDE WEEK 행사 "기업연계 캡스톤 디자인 경진대회"에서 다음과 같이 우수한 성적으로 입상하였기에 상장과 장학금을 수여합니다.

2019년 11월 29일

원광대학교 총장 박 맹 수

Thanks for your attention.