

Portfolio for Dongin Jung

Dongin Jung, 정동인

dongin1009@gmail.com

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
 - Cumulative GPA: 4.34 / 4.5
 - 학부과정: 원광대학교 (Wonkwang University)
 - 컴퓨터소프트웨어공학과 (주전공), Bachelor of Science, 2017.03 ~ 2021.02
 - Cumulative GPA: 3.91 / 4.5
 - 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
 - 디지털포렌식 (복수전공), Bachelor of Science, 2018.03 ~ 2021.02
 - 블록체인, 보안 관련 교과목 수강
- Contact
 - Email: dongin1009@gmail.com
 - Github: <https://github.com/dongin1009>

1. Publications

- 1-1. [Detecting Documents With Inconsistent Context](#)
 - IEEE Access
 - **Dongin Jung**, Misuk Kim, and Yoon-Sik Cho
- 1-2. [Automatic Conversation Turn-Taking Segmentation in Semantic Facet](#)
 - International Conference on Electronics, Information, and Communication (ICEIC) 2023
 - **Dongin Jung**, and Yoon-Sik Cho
- 1-3. [이미지 검색을 위한 대조 학습 모델의 한국어 학습 방법](#)
 - 대한전자공학회 2022년도 하계종합학술대회
 - **정동인**, 김응엽, 강성민, 조윤식
- 1-4. 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
- 1-5. Joint Contrastive and Supervised Learning in Human Activity Recognition
 - (in preparation)
 - **Dongin Jung**, Jaehyeok An, and Yoon-Sik Cho
- 1-6. Exploiting Component Information in Bug Triage
 - (in preparation)
 - Dae-Sung Wang, **Dongin Jung**, Yoon-Sik Cho, and Chan-Gun Lee

1-1. Detecting Documents With Inconsistent Context

- 기여: 그래프 기반의 문서 내용 불일치 탐지 (Graph-based Document Inconsistent Detection)
 - 핵심 기술: Graph Attention Network, 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>

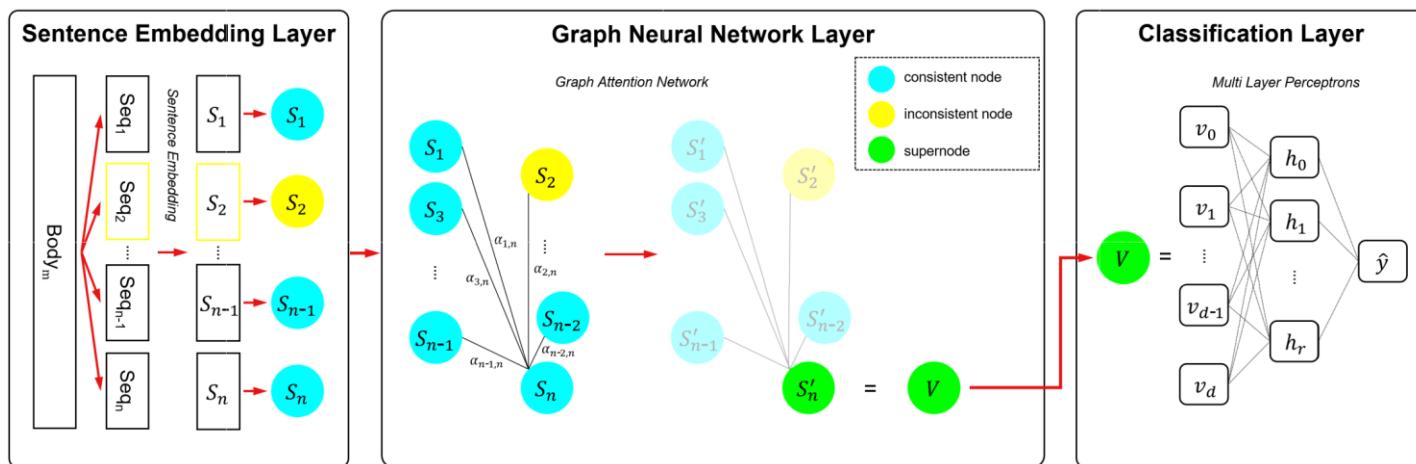


FIGURE 2. Overview of our GraDID model structure applied to inconsistent document. Sentence embedding layer encodes n sentences to n fixed-sized(d dim) vectors by sentence-transformer model. Graph neural network layer performs propagating information between nodes and makes supernode V . The edges have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

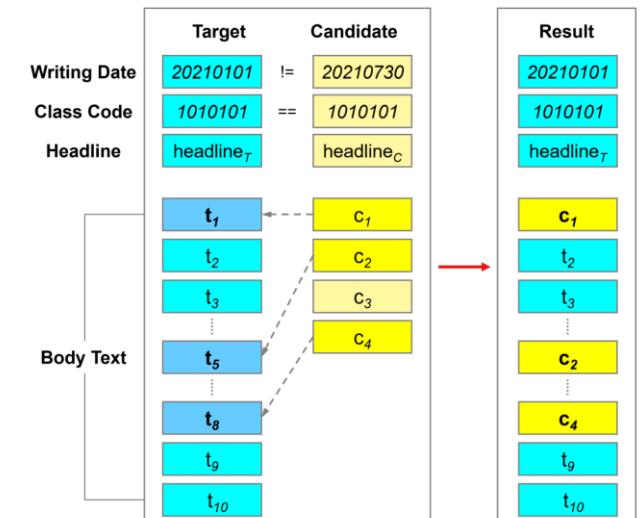


FIGURE 3. Our sentence substitution rules with some conditions. Left (blue) article is selected as target article, middle (yellow) article is chosen as candidate. These have different writing date with at least one common class code. We substitute 30% of the longer articles sentences with those from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

1-1. Detecting Documents With Inconsistent Context

- 기여: 그래프 기반의 문서 내용 불일치 탐지 (GraDID)
 - 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

1-2. 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
 - Recurrent Neural Networks 대비 GPT-2가 실제 전환점을 더 잘 구분함을 보임
 - <https://github.com/dongin1009/semantic-turn-taking-prediction>

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

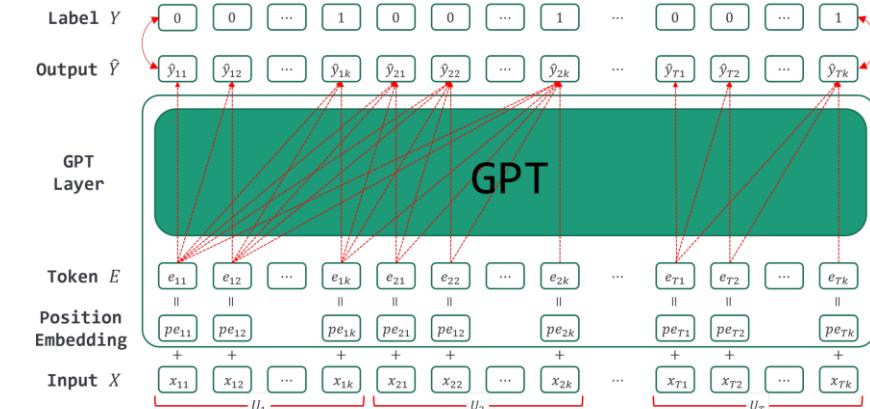


Fig. 2. A pipeline of token-level turn-taking segmentation. Input X is tokens from the tokenizer, and the model predicts in each token level whether the token is end-of-turn or not. In the current time point prediction, the future time point tokens are not used.

	actual conversation	GRU	LSTM	GPT-2
Multi WoZ	i need train reservations from nor wich to can 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 can you help me find an attraction called c in eworld cinema // yet 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 can bridge // [89.71%] 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 // [72.22%] 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 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 // [71.95] yes that was all i needed thank you very much // thank you for using our system	i need train reservations from nor wich to can 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 now // [72.77%] 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 // [69.79%] 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 // [71.95] yes that was all i needed thank you very much // [78.43%] thank you for using our system	i need train reservations from nor wich to can 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 now // [72.77%] 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 // [71.95] yes that was all i needed thank you 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 // 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 // [58.68%] 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.

1-3. 이미지 검색을 위한 대조 학습 모델의 한국어 학습 방법

- 기여: 한국어 Lanugage Model을 사용한 Ko-CLIP
 - 핵심 기술: Korean Language Model, multimodal
 - Language-Image 쌍을 contrastive learning으로 학습한 CLIP의 Language 파트의 transformer를 한국어 모델로 대체
 - 데이터셋이 현저히 적은 한국어 텍스트-이미지 쌍을 효과적으로 학습하기 위해 language, visual 모델에 각각 pretrained model을 사용해 fine-tuning
 - Language model에는 KLUE-RoBERTa-Large, Visual model에는 Vision Transformer(ViT)와 ResNet 버전으로 나누어 Ko-CLIP 제안
 - <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



1-4. 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

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	<u>69.4</u>	<u>79.3</u>	<u>16.5</u>	42.9	69.5	<u>80.1</u>	15.8
Ours (Div4)	43.6	69.8	79.7	15.5	43.7	69.5	80.4	<u>16.5</u>
Test in Global(Div2)	<u>43.5</u>	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	<u>42.9</u>	68.8	78.8	17.6

Table 2. Text-to-video retrieval performance on MSR-VTT [35] dataset with 9k-training data through data division of [9]. 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]	42.4	69.7	80.8	15.7	43.7	71.0	81.2	14.7
Ours (Div2)	44.1	70.8	<u>80.6</u>	13.9	44.4	<u>72.5</u>	<u>81.8</u>	<u>13.4</u>
Ours (Div4)	<u>43.4</u>	72.2	80.2	<u>14.4</u>	<u>44.3</u>	72.8	82.9	13.3
Test in Global(Div2)	43.3	69.2	79.5	15.2	44.4	71.8	81.3	<u>13.4</u>
Test in Global(Div4)	43.2	<u>71.0</u>	80.4	15.4	43.4	72.3	81.7	14.0

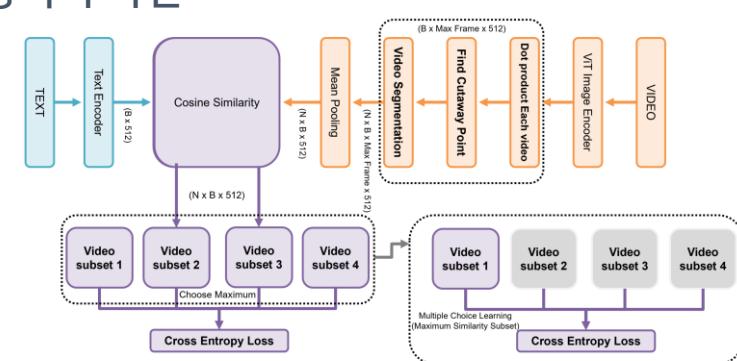


Fig. 1. The overall structure of proposed model based on CLIP4Clip [24]. Our methods are highlighted in a box with dotted lines, where other modules are borrowed from CLIP4Clip [24]. The content in parentheses is the dimension of the video. The output for final learning uses the concept of multiple choice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

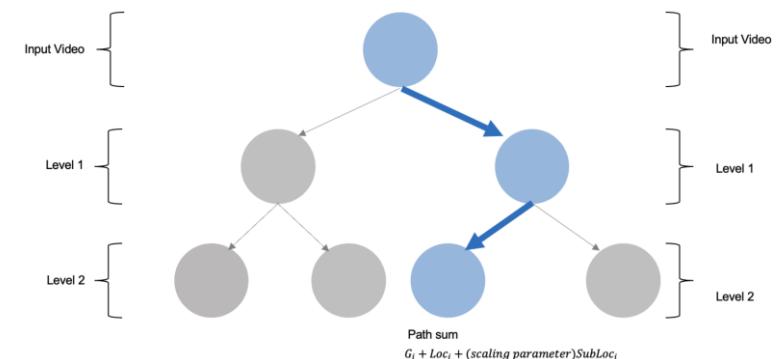


Fig. 3. Tree-based video segmentation and its global-local aggregation method

1-5. 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	92.02 \pm 0.36	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	96.36 \pm 0.19	96.16 \pm 0.21
M3	97.22 \pm 0.50	97.12 \pm 0.53	97.51 \pm 0.27	97.33 \pm 0.25	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	93.63 \pm 0.65	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	92.46 \pm 0.50	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	97.23 \pm 0.25	96.44 \pm 0.41

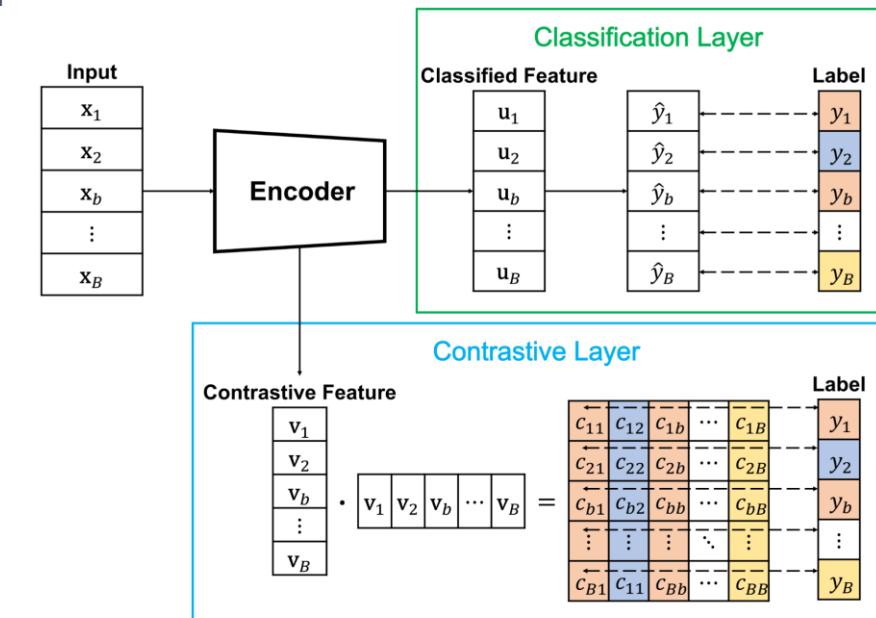
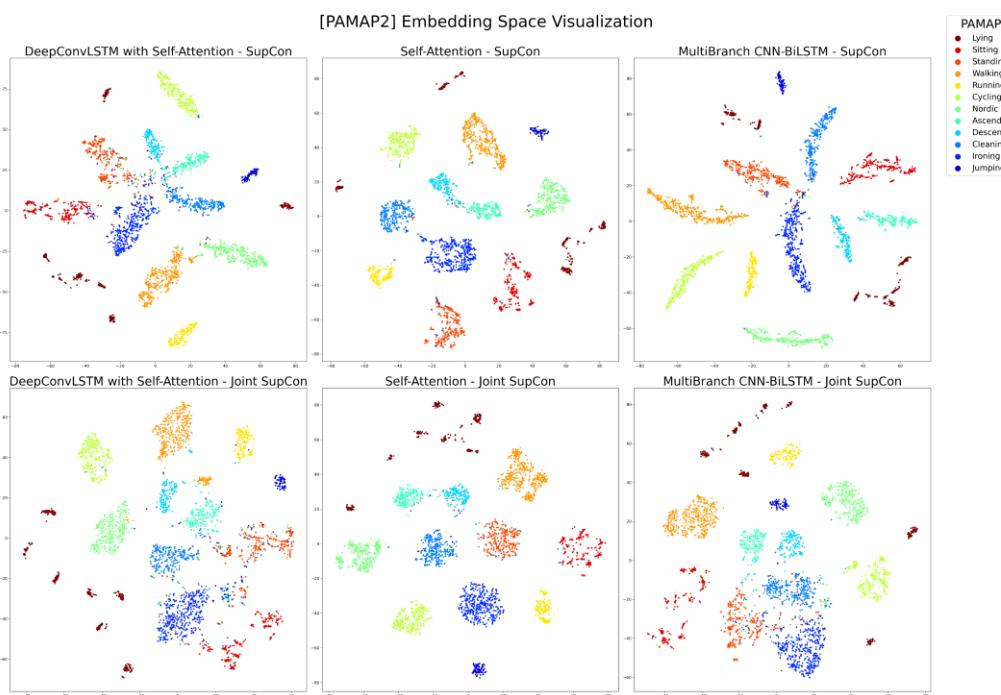


Fig. 2. Overview of joint contrastive and supervised learning method. Our method comprises the encoder, contrastive layer, and classification layer.

1-5. 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



[WISDM] Confusion Matrix

		Self-Attention - Original						MultiBranch CNN-BiLSTM - Original					
		Jogging	LyingDown	Sitting	Stairs	Standing	Walking	Jogging	LyingDown	Sitting	Stairs	Standing	Walking
True	Jogging	413	0	2	1	1	5	414	0	3	1	2	2
	LyingDown	0	233	32	0	0	2	1	211	47	0	6	2
True	Sitting	0	45	575	2	13	7	3	33	551	1	36	18
	Stairs	0	0	0	50	0	4	0	0	0	47	0	7
True	Standing	0	2	64	0	195	11	3	3	24	0	229	13
	Walking	0	1	10	1	7	1208	2	3	17	10	10	1185
		Self-Attention - SupCon						MultiBranch CNN-BiLSTM - SupCon					
		Jogging	LyingDown	Sitting	Stairs	Standing	Walking	Jogging	LyingDown	Sitting	Stairs	Standing	Walking
True	Jogging	418	0	3	1	0	0	416	0	2	1	2	1
	LyingDown	0	245	20	0	2	0	1	225	34	0	6	1
True	Sitting	0	83	532	4	14	9	4	54	524	1	30	29
	Stairs	1	0	1	45	0	7	0	0	0	47	0	7
True	Standing	1	5	59	1	199	7	2	8	28	0	223	11
	Walking	1	0	13	2	4	1207	6	1	10	6	7	1197
		DeepConvLSTM with Self-Attention - SupCon						Self-Attention - SupCon					
		Jogging	LyingDown	Sitting	Stairs	Standing	Walking	Jogging	LyingDown	Sitting	Stairs	Standing	Walking
True	Jogging	418	0	2	1	0	0	416	0	2	1	2	1
	LyingDown	0	251	16	0	0	0	1	225	34	0	6	1
True	Sitting	0	41	587	0	11	3	4	54	524	1	30	29
	Stairs	0	1	1	46	0	7	0	0	0	47	0	7
True	Standing	1	5	47	2	215	2	2	8	28	0	223	11
	Walking	0	0	10	1	5	1211	6	1	10	6	7	1197
		DeepConvLSTM with Self-Attention - Joint SupCon						Self-Attention - Joint SupCon					
		Jogging	LyingDown	Sitting	Stairs	Standing	Walking	Jogging	LyingDown	Sitting	Stairs	Standing	Walking
True	Jogging	418	0	1	2	1	5	413	0	1	2	1	5
	LyingDown	0	224	38	0	4	1	0	224	38	0	4	1
True	Sitting	0	22	555	1	46	16	2	22	555	1	46	16
	Stairs	0	0	0	51	0	3	0	0	0	51	0	3
True	Standing	0	2	21	0	238	11	0	2	21	0	238	11
	Walking	1	3	5	2	5	1211	1	3	5	2	5	1211
		MultiBranch CNN-BiLSTM - SupCon						MultiBranch CNN-BiLSTM - Joint SupCon					
		Jogging	LyingDown	Sitting	Stairs	Standing	Walking	Jogging	LyingDown	Sitting	Stairs	Standing	Walking
True	Jogging	421	0	1	0	0	0	421	0	1	0	0	0
	LyingDown	0	250	12	0	4	1	0	250	12	0	4	1
True	Sitting	4	33	587	0	14	4	4	54	587	0	14	4
	Stairs	0	0	1	52	0	1	0	0	1	52	0	1
True	Standing	0	1	13	0	257	1	2	8	28	0	257	1
	Walking	0	0	2	0	3	1222	0	0	2	0	3	1222
		DeepConvLSTM with Self-Attention - Joint SupCon						Self-Attention - Joint SupCon					
		Jogging	LyingDown	Sitting	Stairs	Standing	Walking	Jogging	LyingDown	Sitting	Stairs	Standing	Walking
True	Jogging	418	0	2	1	0	0	413	0	1	2	1	5
	LyingDown	0	251	16	0	0	0	0	224	38	0	4	1
True	Sitting	0	41	587	0	11	3	2	22	555	1	46	16
	Stairs	0	1	1	46	0	3	0	0	0	51	0	3
True	Standing	1	5	47	2	215	2	0	2	21	0	238	11
	Walking	0	0	10	1	5	1211	1	3	5	2	5	1211
		MultiBranch CNN-BiLSTM - SupCon						MultiBranch CNN-BiLSTM - Joint SupCon					
		Jogging	LyingDown	Sitting	Stairs	Standing	Walking	Jogging	LyingDown	Sitting	Stairs	Standing	Walking
True	Jogging	421	0	1	0	0	0	421	0	1	0	0	0
	LyingDown	0	252	13	0	2	0	0	252	13	0	2	0
True	Sitting	0	20	606	0	15	1	0	20	606	0	15	1
	Stairs	0	0	0	54	0	0	0	0	0	54	0	0
True	Standing	0	1	12	0	257	2	0	1	12	0	257	2
	Walking	0	0	1	0	5	1221	0	0	1	0	5	1221

1-6. Exploiting Component Information in Bug Triage

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

TABLE II

SUMMARY OF EXPERIMENTS ON PREVIOUS AND OUR DATASET. THE BASIC DENOTES OUR BASIC MODEL AND THE FULL DENOTES OUR FULL MODEL. THE METRICS ARE ON TOP- k ACCURACY AND THE BOLD VALUE IS THE HIGHER ACCURACY. OUR METHODS ARE OUTPERFORMED IN ALL DATASETS.

Dataset	Method	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8	Top-9	Top-10
Google Chromium	DBRNN-A	0.205	0.296	0.354	0.394	0.426	0.453	0.476	0.496	0.513	0.531
	BASIC	0.286	0.404	0.470	0.521	0.558	0.588	0.610	0.629	0.647	0.660
Mozilla Core	DBRNN-A	0.286	0.398	0.463	0.514	0.553	0.582	0.608	0.630	0.648	0.664
	BASIC	0.360	0.483	0.551	0.601	0.636	0.668	0.697	0.715	0.732	0.746
Mozilla Firefox	DBRNN-A	0.285	0.397	0.467	0.529	0.574	0.600	0.625	0.656	0.674	0.689
	BASIC	0.344	0.455	0.536	0.583	0.625	0.662	0.692	0.707	0.728	0.741
Eclipse Platform	BASIC	0.518	0.635	0.709	0.752	0.790	0.813	0.827	0.840	0.848	0.868
	FULL	0.584	0.697	0.748	0.790	0.824	0.846	0.856	0.873	0.880	0.882
Eclipse Foundation	BASIC	0.714	0.840	0.882	0.904	0.922	0.931	0.939	0.949	0.953	0.956
	FULL	0.728	0.848	0.886	0.914	0.929	0.943	0.953	0.959	0.968	0.969
Bugzilla Firefox	BASIC	0.527	0.625	0.691	0.724	0.757	0.782	0.796	0.810	0.819	0.826
	FULL	0.560	0.627	0.691	0.729	0.747	0.750	0.769	0.784	0.799	0.806

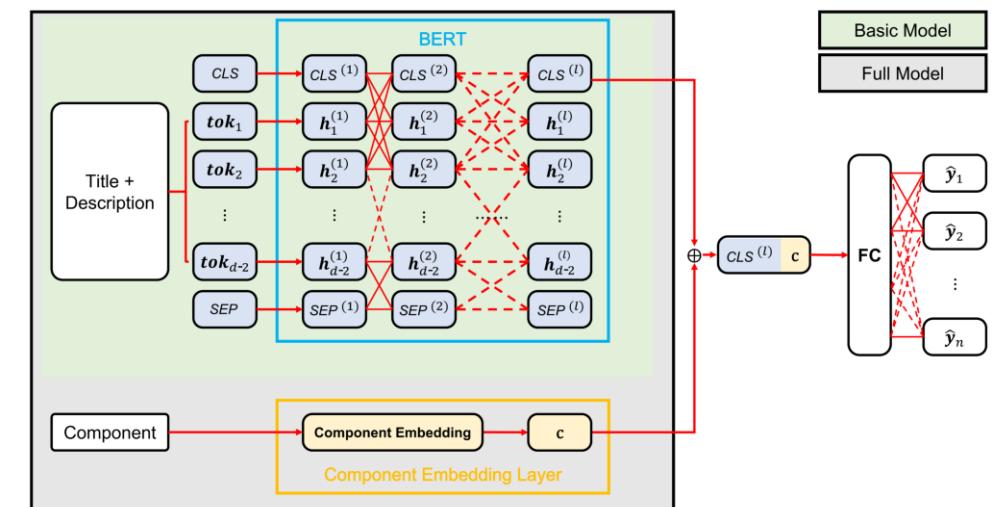


Fig. 1. Overview of our full model structure. The title and description are combined and separated to tokens by tokenizer. Each token with special tokens through to a BERT model. Simultaneously, the component through the Component Embedding Layer and generate component embedding c . The CLS token of the last layer of BERT and component embedding are concatenated and through the fully-connected layer.

2. Projects

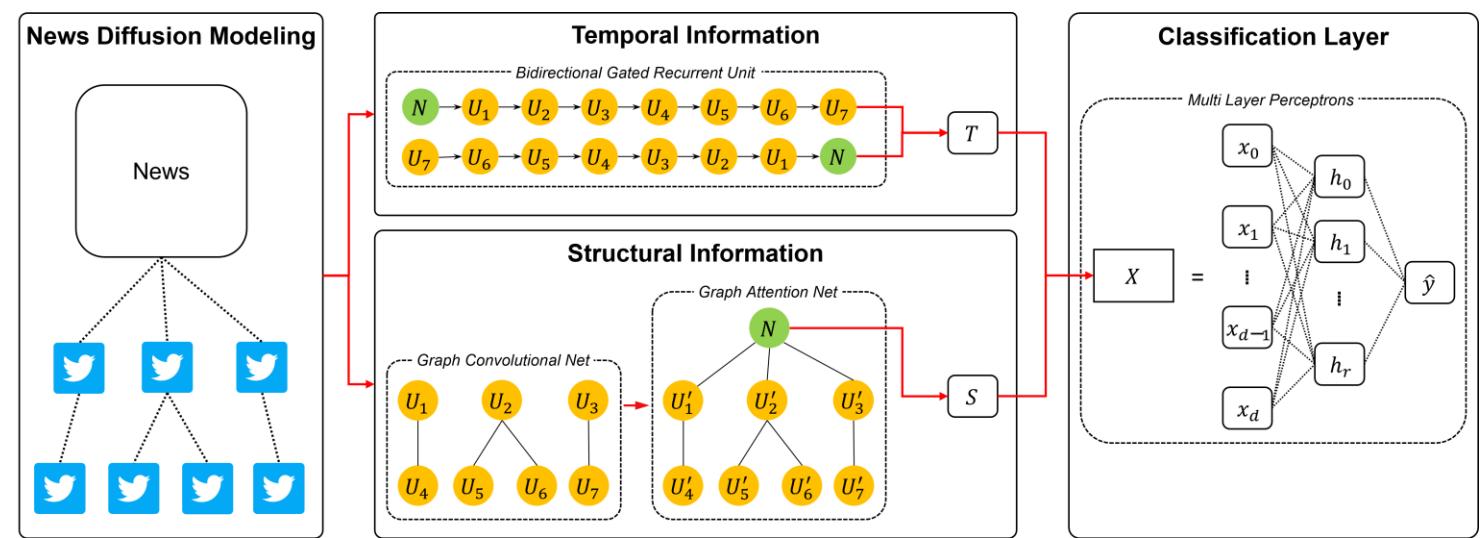
- 2-1. '실감형 뉴스'를 위한 빅데이터 분석 참여형 통합 플랫폼 구축
 - 한국콘텐츠진흥원 연구과제
 - 2021.03 ~ 2022.12
- 2-2. 뇌인지-행동 데이터 기반 비대면 정신건강문제 실시간 인터랙티브 셀프 모니터링 HW/SW 플랫폼 개발
 - 정보통신기획평가원 연구과제
 - 2021.03 ~ 2021.12
- 2-3. Polyglot: Multilingual Large Language Models
 - EleutherAI 프로젝트: Polyglot - Romance group
 - 2022.10 ~ 진행중
- 2-4. 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	<i>PolitiFact</i>		
	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

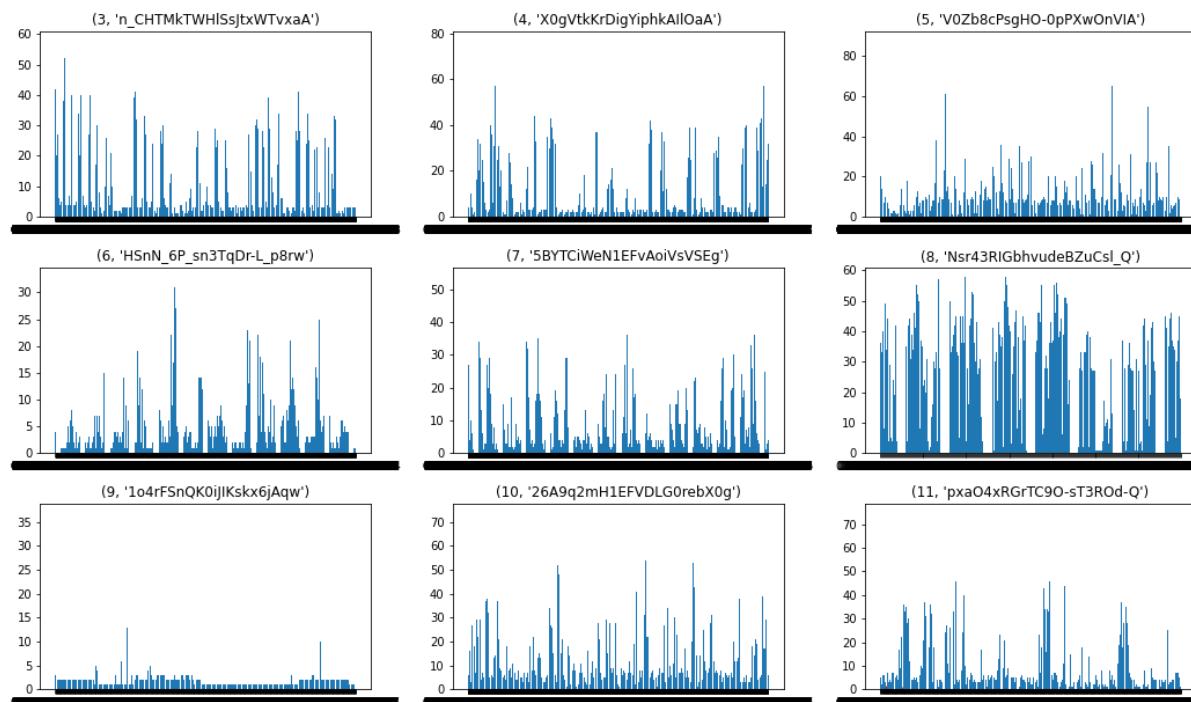
Model	<i>GossipCop</i>		
	Accuracy	F1-score	AUROC
UPDST (ours)	97.00	96.96	98.54
UPFD-SAGE [4]	96.71	96.67	99.18
UPFD-GAT [4]	95.75	95.70	98.78
UPFD-GCN [4]	97.00	96.96	98.72
Bi-GCN [1]	96.56	96.52	98.99
GCN-FN [14]	96.63	96.59	98.90



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

- **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

- 다국어 text generation model 개발 - Romance language group (진행중)
 - 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/Wittgensteinian/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, 강현지, 김수빈, 윤지환, 장형국, 조영진

3-1. 시각장애인을 위한 음성 챗봇 시스템 및 방법

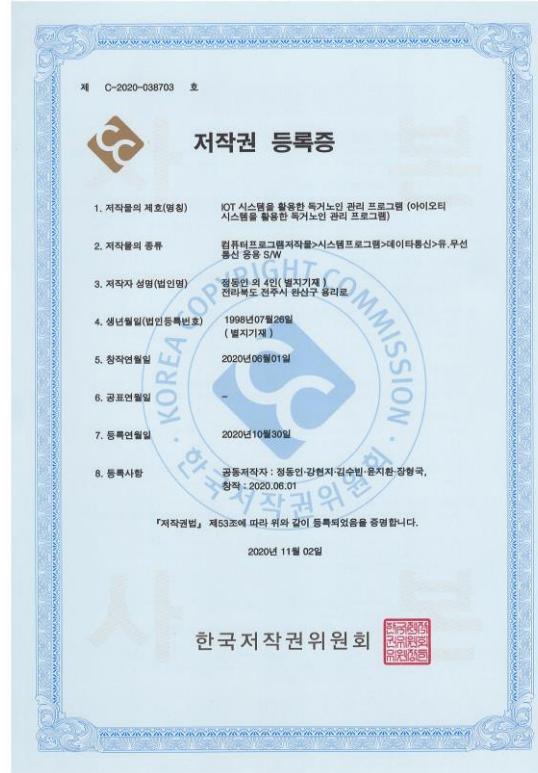
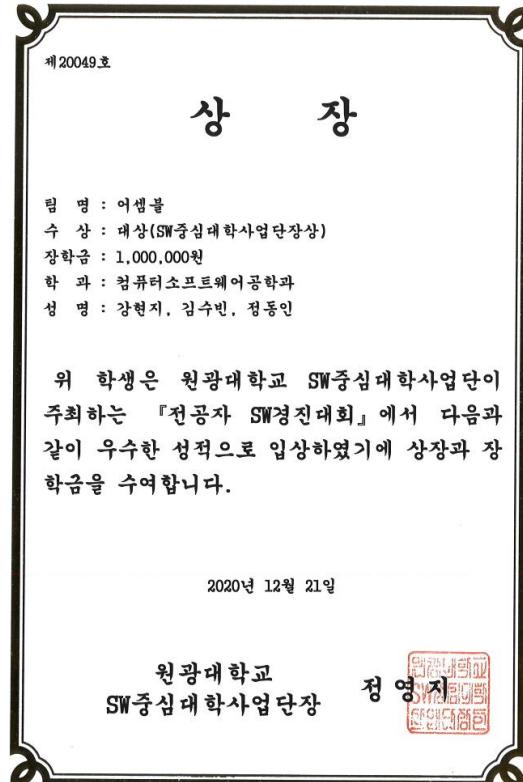
- STT를 적용한 seq2seq 챗봇
 - 특허출원 제10-2021-0020237호
 - 기존 챗봇 시스템의 seq2seq 구조에서 입출력 단에 Speech-to-Text(STT)와 Text-to-Speech(TTS)를 결합해 시각장애인도 음성으로 사용이 가능한 챗봇 시스템 아이디어 제안
 - <https://patents.google.com/patent/KR20220116859A>

출원사실증명원 CERTIFICATE OF APPLICATION				
출 원 인 Applicant	성 명 Name	정동인 JUNG, Dong In	주 민 번 호 Residence No	980726-1*****
	주 소	서울특별시 동작구 상도로47길 116-9 (상도1동)	전 화 번 호	010-9867-9864
발명자 Inventor	성 명 Name	정동인 JUNG, Dong In	주 민 번 호 Residence No	980726-1*****
	주 소	서울특별시 동작구 상도로47길 116-9 (상도1동)	전 화 번 호	010-9867-9864
대 리 인 Agent	성 명 Name	이성원	대리인 번 호	9-2009-003764-1
	주 소	서울특별시 성동구 성수일로 77 1512호 (성수동1가, 서울숲IT밸리)(아이피비밸)		
출 원 번 호 Application Number	특허-2021-0020237 PATENT-2021-0020237	출 원 일 자 Filing Date	2021년 02월 16일 FEB 16 , 2021	
Title of Invention, Product(s) Embodied in Design, or Classification of Mark	발명(고안)의 명칭. 디자인을 표현할 물품. 상품(서비스업)류 구분 VOICE CHATBOT SYSTEM AND METHOD FOR THE VISUALLY IMPAIRED			
용 도	제출용	IPC 분류	G10L 15/22	
최 종 처 분 상 태		최종처분일		
심 사 청 구 유 무	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 COMMISSIONER				

● 본 증명서는 인터넷으로 발급되었으며, 특허청 홈페이지(www.kipo.go.kr)의 '특허증명서 발급' 메뉴를 통해 발급번호 또는 문서번호의 위변조 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 추가해 스마트 화분 제안
 - 화분 장치의 센서 정보를 추적할 수 있는 앱을 통해 독거 노인을 관리할 수 있는 시스템 제안
 - <https://patents.google.com/patent/KR102170452B1>



Dongin Jung



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 풍향, 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

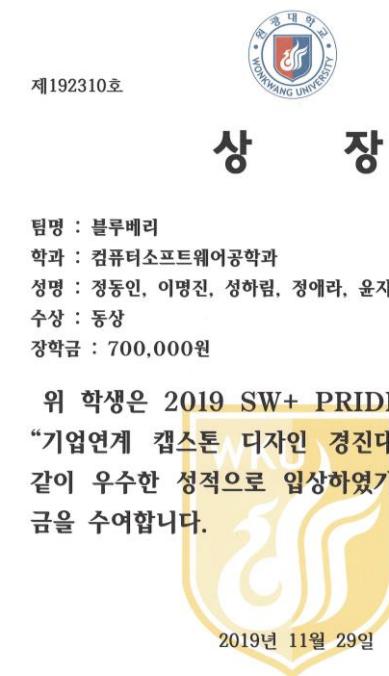
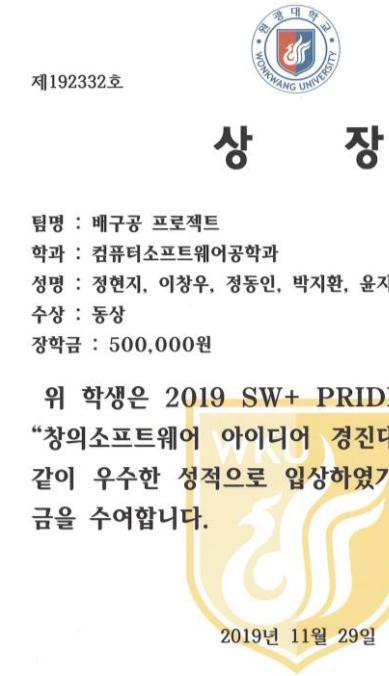
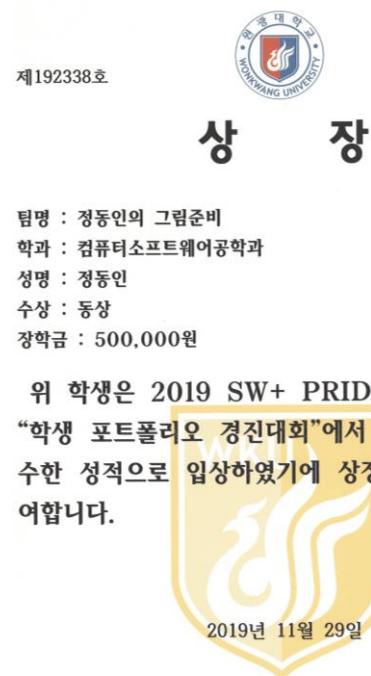
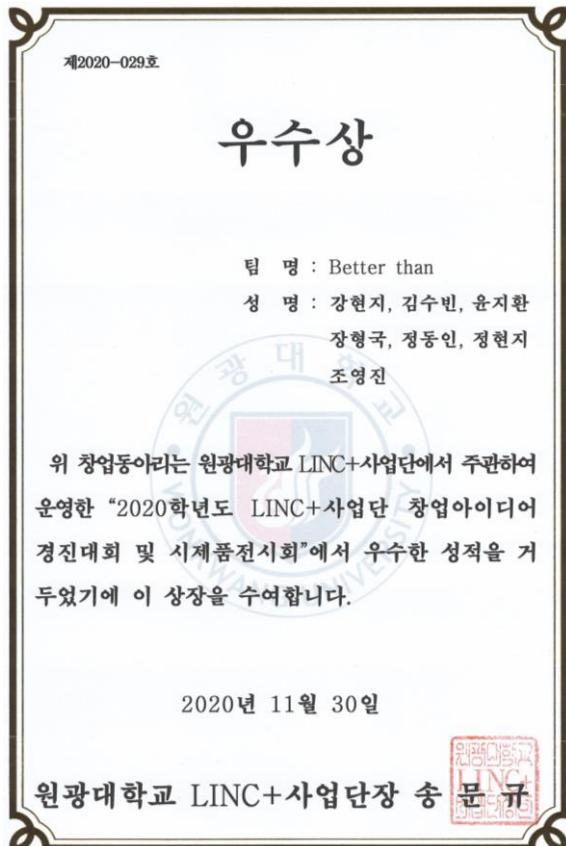


29등 / 237

- 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 voting한 모델 구현, 그러나 단일 케이스에서 가장 좋은 결과

4. 기타 수상 내역

- LINC+ 창업아이디어 경진 대회 및 시제품 전시회 우수 상 수상
- 학생 포트폴리오 경진대회 동상 수상
- 창의소프트웨어 아이디어 경진대회 동상 수상
- 기업연계 캡스톤 디자인 경진대회 동상 수상



Thanks for your attention.
