

자연어처리 모델을 이용한 이커머스 데이터 기반 감성 분석 모델 구축

최준영¹, 임희석^{2*}

¹고려대학교 컴퓨터정보통신대학원 석사과정, ²고려대학교 컴퓨터학과 교수

E-commerce data based Sentiment Analysis Model Implementation using Natural Language Processing Model

Jun-Young Choi¹, Heui-Seok Lim^{2*}

¹Master's Course, Graduates School of Computer & Information Technology, Korea University

²Professor, Department of Computer Science and Engineering, Korea University

요약 자연어 처리 분야에서 번역, 형태소 태깅, 질의응답, 감성 분석 등 다양한 영역의 연구가 활발히 진행되고 있다. 감성 분석 분야는 Pretrained Model을 전이 학습하여 단일 도메인 영어 데이터셋에 대해 높은 분류 정확도를 보여주고 있다. 본 연구에서는 다양한 도메인 속성을 가지고 있는 이커머스 한글 상품평 데이터를 이용하고 단어 빈도 기반의 BOW(Bag Of Word), LSTM[1], Attention, CNN[2], ELMo[3], KoBERT[4] 모델을 구현하여 분류 성능을 비교하였다. 같은 단어를 동일하게 임베딩하는 모델에 비해 문맥에 따라 다르게 임베딩하는 전이학습 모델이 높은 정확도를 낸다는 것을 확인하였고, 17개 카테고리 별, 모델 성능 결과를 분석하여 실제 이커머스 산업에서 적용할 수 있는 감성 분석 모델 구성을 제안한다. 그리고 모델별 용량에 따른 추론 속도를 비교하여 실시간 서비스가 가능할 수 있는 모델 연구 방향을 제시한다.

주제어 : 자연어처리, 감성 분석, 이커머스, 워드 임베딩, 센텐스 임베딩, 전이학습

Abstract In the field of Natural Language Processing, Various research such as Translation, POS Tagging, Q&A, and Sentiment Analysis are globally being carried out. Sentiment Analysis shows high classification performance for English single-domain datasets by pretrained sentence embedding models. In this thesis, the classification performance is compared by Korean E-commerce online dataset with various domain attributes and 6 Neural-Net models are built as BOW (Bag Of Word), LSTM[1], Attention, CNN[2], ELMo[3], and BERT(KoBERT)[4]. It has been confirmed that the performance of pretrained sentence embedding models are higher than word embedding models. In addition, practical Neural-Net model composition is proposed after comparing classification performance on dataset with 17 categories. Furthermore, the way of compressing sentence embedding model is mentioned as future work, considering inference time against model capacity on real-time service.

Key Words : NLP, BERT, KoBERT, ELMo, LSTM, CNN

*Corresponding Author : Heui-Seok Lim(limhseok@korea.ac.kr)

Received September 10, 2020
Accepted November 20, 2020

Revised October 12, 2020
Published November 28, 2020

1. 서론

매년 이커머스 온라인 쇼핑물 사용률이 가파르게 성장하고 있다. 고객은 상품 정보보다 상품 후기 댓글을 통해 상품 구매를 결정하고 있어 이커머스 기업들이 자연어처리 기술을 연구 개발하는데 많은 투자를 지원하고 있다.

현재, 자연어처리 감성 분석 연구가 활발히 진행되고 있고 고무적인 분류 정확도를 보여주고 있지만 대부분 영어 데이터셋에 기반한 단일 도메인 연구 결과이다.

본 논문에서는 다양한 도메인(카테고리)가진 한글 상품평 데이터셋으로 6가지 감성 분석 모델 구현하여 성능을 비교하고자 한다. 그리고 비정제된 실제 한글 이커머스 온라인 데이터를 적용하기 위한 전처리 방안을 언급하고 다양한 도메인(카테고리) 데이터셋의 감성 분석 모델을 고성능으로 구현하는 방안을 제시한다. 본 논문의 2장에서는 기존 자연어처리 감성 분석 연구 내용을 언급하고 3장에서는 구현한 6가지 감성 분석 모델 구조를 설명한다. 4장에서는 데이터셋과 감성 분석 모델 구현 방법을 상세히 소개하고 5장에서는 모델 성능 결과 및 추론 속도에 대해 논의한다. 마지막 6장에서는 결론과 향후 과제 방향 제시한다.

2. 관련 연구

언어와 같은 Sequential Data를 문장 의미를 잘 반영한 context vector로 변환하기 위해 LSTM[1], CNN[2], ELMo[3], BERT[5]와 같은 언어 모델이 연구되고 있고 BERT인 경우 특정 과제(QnA)에서 이미 인간보다 더 높은 성능을 내고 있다. 국내에서는 김한민[6]이 캐글에서 공개한 영어 음식 상품 리뷰 데이터로 기본적인 머신러닝 기법(Naive Bayes, Logistic, Support Vector Machine, Decision Tree)을 앙상블(Boosting, Bagging, Random Subspace, Random Forest)하여 성능을 비교 분석하였다. 그리고 박호연[7]은 영어 영화 리뷰 데이터(IMDB)로 CNN-LSTM 조합모델을 구현하여 단순 CNN 및 LSTM 모델과 성능을 비교하였다. 선행 연구들은 단일 도메인의 영어 단어들을 TF-IDF로 변환하거나 동의어를 동일하게 임베딩하는 기법을 사용하여 감성 분석 모델을 구현하였으나, 다양한 도메인(카테고리)을 가진 한글인 경우, 교차어라는 특성과 동의어를 문맥에 따라 다르게 임베딩하는 기법을 고려해야 할 것이다. 따라서 본 논문에서는 이를 고려하여 한글 데이터셋

를 형태소 분석기 및 정규 표현식을 이용하여 토근화하였고, 구현화한 6가지 모델 중에서 KoBERT[4] 모델이 가장 좋은 성능을 내는 것을 확인하였다. 그리고 이커머스 온라인 상품평 17가지 카테고리 별 성능을 비교하여, 실제 서비스를 하기 위한 모델 구성을 제시한다.

3. 감성 분석 모델

해당 논문에서는 한글 이커머스 데이터셋으로 BOW, BDLSTM, BDLSTM+Attention, CNN[2], ELMo[3], KoBERT[4] 총 6가지 감성 분석 모델을 구현하여 성능을 비교 분석하였다.

3.1 BOW(Bag of Word) 모델

Bag Of Word란 문장의 단어 순서를 고려하지 않고 단어 출현 빈도로 문장을 벡터화하는 방법이다. 본 논문에서 상품평 단어 사전을 구축하고 index를 붙여 각 상품평들을 아래 Fig 1 와 같이 카운트 기반 multi-hot 벡터로 변환하였다. 그리고 단순 2-Fully Connected Layers로 Fig 2 와 같이 구성하였고 음/부정에 대한 분류 문제를 풀기 위해 최종 Activation Function에 아래 식 1 sigmoid를 적용하였다. 이후 언급되는 모델들의 최종 Activation Function은 동일하게 적용한다.


$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$




Fig. 1. BOW InputData Vectorization

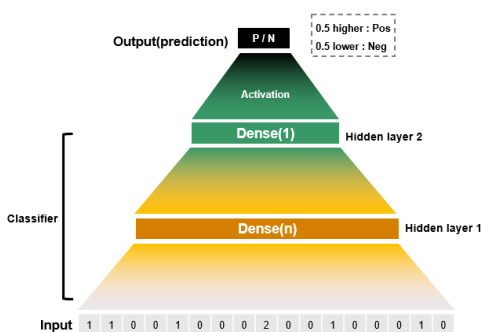


Fig. 2. BOW Sentiment Model

3.2 BDLSTM(BiDirectional Long Short Term Memory) 모델

BDLSTM 모델은 자연어처리 분야에 일반적으로 많이 사용되는 LSTM(Long Short Term Memory)[1]을 각 단어에 2개 적용하여 문장의 정/반 방향의 정보를 학습하여 단일 LSTM에 비해 더 정교한 context vector를 생성하는 모델이다.

본 논문에서 BDLSTM 모델을 아래 Fig 3과 같이 구성하였다.

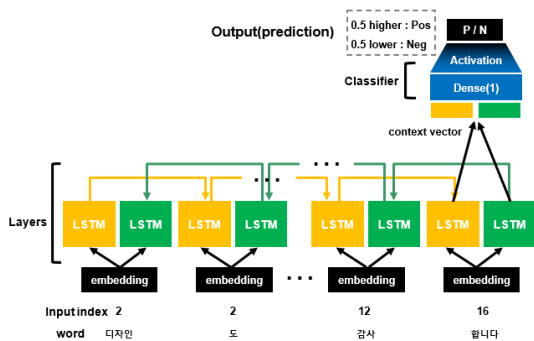


Fig. 3. BDLSTM Sentiment Model

3.3 BDLSTM(BiDirectional Long Short Term Memory) + Attention 모델

BDLSTM + Attention 모델은 3.2절에 언급한 BDLSTM output cell vector에 Bahdanau Attention[8]를 적용한 모델이다. 해당 모델은 각 단어를 임베딩 시, 전체 입력 단어를 동일한 비율로 학습하지 않고 해당 단어와 연관 있는 단어에 더 집중하여 임베딩되기 때문에 더 정교한 context vector가 생성된다.

Bahdanau Attention은 식 2와 같이 정의된다.

해당 식의 V는 최종 긍/부정을 판별하는 Dense lay의 가중치 행렬이고, W₁, W₂는 각 단어들의 관계 강도를 계산하는 가중치 행렬이다.

$$score(query, key) = V^T \tanh(W_1 key + W_2 query) \quad (2)$$

즉, 기존 BDLSTM 모델은 마지막 hidden layer context vector를 사용하는 반면 해당 구현 모델은 Bahdanau Attention[8]기법을 이용하여 모든 hidden layer context vector를 고려하기 때문에 정보 손실을 최대한 줄일 수 있게 된다.

본 논문에서 BLSTM + Attention 모델을 Fig 4와 같이 구성하였다.

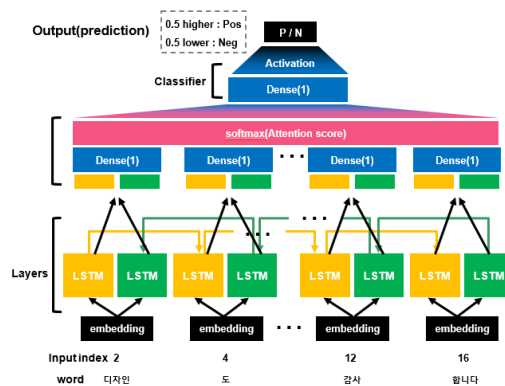


Fig. 4. BDLSTM + Attention Sentiment Model

3.4 CNN(Convolutional Neural Net) 모델

CNN은 Lecun이 1998년 LeNet[2]으로 처음 제시되었고 Convolution이라는 Filter를 여러층 쌓아서 Receptive Field가 가장 넓은 마지막 Filter를 이용하여 이미지를 분류하는 Neural Net이다. 김윤은 CNN 방법으로 언어를 학습할 수 있는 구조를 논문[9]에서 Fig 5와 같이 제시하였다.

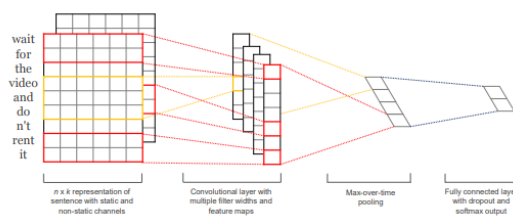


Fig. 5. CNN Sentence Classification Model Example

본 논문의 CNN 감성 분석 모델 구조는 아래 Table 1과 같이 구성하여 구현하였다.

Table 1. CNN Sentiment Model

Layer No	Layer Name
1	Embedding
2	Dropout
3	Conv1D
4	Global Max Pooling1D
5	Fully Connected
6	Dropout
7	Fully Connected

3.5 ELMo(Embeddings from Language Models) 모델

AllenNLP에서 공개한 ELMo[3]는 동의어를 다르게 임베딩하는 문맥 반영 워드 임베딩 Pretrained 모델이다. 임베딩은 문자 단위로 CNN[2]을 적용하여 생성되고 Bidirectional Language Model의 구조로 각 임베딩 및 LSTM의 출력값들을 가중합하여 학습된다.

BERT[5] 논문에서 ELMo 구조를 아래 Fig 6과 같이 제시하였고 본 논문에서는 Tensorflow Hub에서 제공하는 Pretrained ELMo 모델을 사용하여 Fig 7과 같이 모델을 구성하였다.

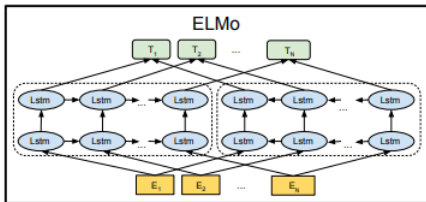


Fig. 6. ELMo structure

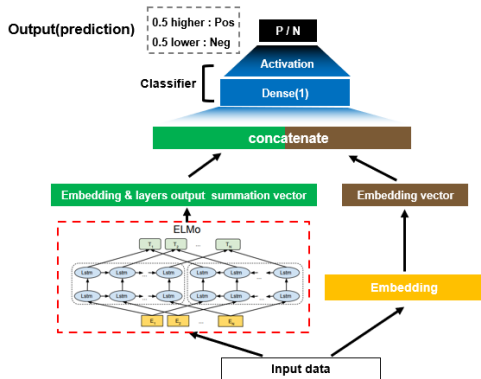


Fig. 7. ELMo Sentiment Model

3.6 KoBERT(Korean Bidirectional Embedding Representation Transformer) 모델

Google이 공개한 BERT[5]는 Attention is all you need[10]에서 제시된 Fig 8의 Transformer Encoder 부분을 이용해 만든 Language Representation Model 이다.

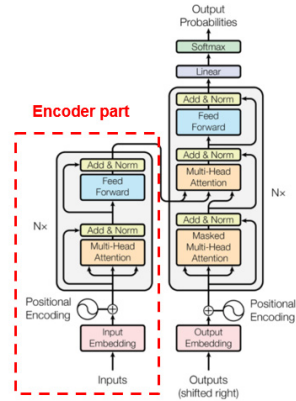


Fig. 8. Transformer architecture

BERT는 BookCorpus(8억 단어)와 Wikipedia(25억 단어) 말뭉치를 이용하여 MLM(Masked Language Model) 과 NSP(Next Sentence Prediction) 방식으로 학습되고 입력 데이터는 Transformer와 달리 word token embedding, segment embedding, position embedding이 element wise sum하여 만들어진다.

word token embedding은 sentencepiece[11]로 토큰화되고, BERT[5] 논문에서 제시된 아래 Fig 9와 같이 문장 시작에는 [CLS], 개별 단어의 분리는 ##, 문장 종료는 [SEP] 토큰을 붙여 입력 데이터가 구성된다.

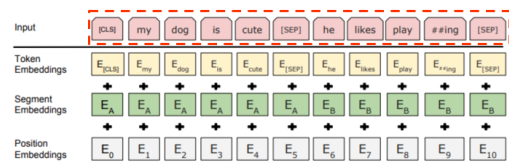


Fig. 9. BERT Input data structure

본 논문에서는 BERT[5] 논문에서 제시된 Fig 10의 모델 구조에 한국어 위키/뉴스를 추가 학습시킨 KoBERT[4]를 이용하였고, 아래 Fig 11과 같이 Classifier을 붙여 모델을 구성하였다.

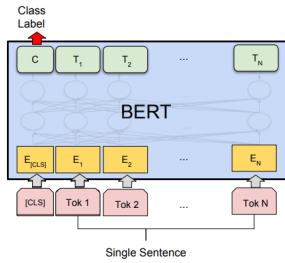


Fig. 10. BERT single sentence classification

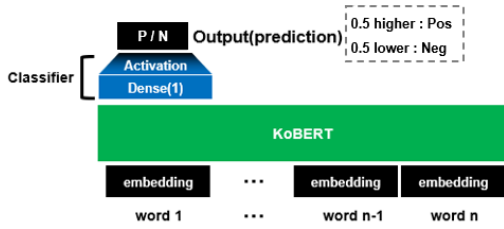


Fig. 11. KoBERT Sentiment Model

4. 모델 구현

4.1 Dataset 설명

데이터셋은 롯데닷컴 이커머스 상품평과 별점을 수집하여 아래 Table 2와 같이 준비하였다. 수집된 상품평 수는 26,782,700건이고 17가지 카테고리를 가진 별점 별, 상품평 수는 아래 Table 3와 같다.

Table 2. Dataset Example

No	reply comment	star score
1	아이들이 너무 좋아해요~~	5
2	배송도 빠르고 맛도 짱짱 !!	4
3	예상보다 작아서 놀랐어요^^	1
4	물건이....부서져 있어요 ㅜㅜ	2

Table 3. Dataset Information

star score	reply comment number	Ratio(%)
1	154,161	0.58
2	143,888	0.54
3	724,555	2.71
4	3,006,772	11.22
5	22,753,824	84.96

별점 1,2점을 부정 상품평으로 4,5점은 긍정 상품평으로 가정하였고 상품평 긍/부정 일치 여부를 조사한 결과, random sampling 10,000건 기준 긍/부정 불일치 비

율이 35% / 50%임을 확인하였다. 따라서 긍/부정 Tagging 작업을 수행하여 긍/부정 상품평을 8만건씩 총 16만건을 준비하였다. 현 실험에서 불용어는 지정하지 않았고 Train / Validation / Test 데이터셋의 비율은 8:1:1으로 분리하였다

4.2 형태소 분석기

본 논문에서 다루는 이커머스 온라인 상품평 데이터셋은 맞춤법, 띄어쓰기, 비속어, 등 정리되지 않은 문장이고 한글은 교착어이기 때문에 모델 입력 전, 전처리 작업이 필요하였다.

먼저 정규 표현식으로 한글을 제외한 이모티콘, 특수 기호, 영어를 제거했고 KoNLpy package에서 제공하는 형태소 분석기를 적용하여 아래 Table 4와 같이 성능을 비교 분석하였다.

Table 4. KoNLpy POS Tagger Comparison

POS Tagger	komoran	mecab	okt	hannanum	twitter
Tokenization time (sec)	48	5	183	223	181
Dictionary size	27,653	27,696	65,846	111,525	65,864
Max sentence length	1,140	1,032	799	1,097	799
Average sentence length	34	31	23	27	23
Auto spacing words	0	0	0	X	0

표제어와 어간 추출 성능 및 속도를 고려하여 Mecab을 형태소 분석기로 선택하였고, KoBERT[4]인 경우 Mecab 형태소 분석기로 토큰화된 문장을 이어 붙인 후, sentencepiece[11]로 재토큰화 하였다.

4.2 모델 실험 시스템 환경

Machine Learning Framework은 Tensorflow v1.14, GPU는 Single Nvidia 2080ti, Memory는 32GB, OS는 ubuntu 18.04 인 환경에서 모델을 개발하여 실험 결과를 도출하였다.

4.3 구현 감성 분석 모델 구성

3절에 언급한 모델 중, BOW 모델을 Baseline으로 설정하고, BDLSTM, BDLSTM+Attention, CNN 모델의 Embedding Layer는 페이스북에서 개발한

FastText[12]를 적용하여 Embedding size 300으로 생성 후, freezing 하여 학습시켰다.

그리고 각 모델에 적용한 Layer Activation Function, Loss Function, Optimizer은 아래 Table 5와 같다.

Table 5. Activation Function, Loss Function, Optimizer

No	Model	layer activation function	loss function	optimizer
1	BOW	Relu[13]	Binary Cross Entropy	Adam[14]
2	BDLSTM	Relu	Binary Cross Entropy	Adam
3	BDLSTM+Attention	Relu	Binary Cross Entropy	Adam
4	CNN	Relu	Binary Cross Entropy	Adam
5	ELMo[7]	Relu	Binary Cross Entropy	Adam
6	KoBERT[11]	Gelu[15]	Binary Cross Entropy	Adam

5. 모델 성능 결과

모델 별, parameters 개수 및 학습시간은 아래 Table 6과 같다.

Table 6. Model Parameter number & Training Time

No	Model	Parameters (#)	1 epoch time (min)
1	BOW	27,366,401	6
2	BDLSTM	8,174,901	5
3	BDLSTM+Attention	9,112,528	7
4	CNN	7,646,301	4
5	ELMo[7]	13,862,657	56
6	KoBERT[11]	178,271,741	61

서비스 측면에서 KoBERT[4]는 parameter 개수가 타 모델에 비해 상대적으로 많고 학습 및 예측 시간이 상당히 길어 실시간 서비스에 대한 제약이 있게 된다. 따라서 가지치기, 가중치 분해, 지식 증류, 가중치 공유, 양자 화등 경량화에 대한 많은 연구[16]가 진행 중이다

모델별, 분류 성능 비교는 아래 Table 7과 같다.

Table 7. Model Accuracy & Loss

No	Model	Accuracy (%)	BCE loss
1	BOW	92.6	0.245
2	BDLSTM	93.2	0.184
3	BDLSTM+Attention	94.4	0.173
4	CNN	93.8	0.175
5	ELMo	94.6	0.122
6	KoBERT	96.2	0.102

KoBERT[4]가 가장 좋은 성능을 보였고 BCE loss 측면에서 상대적으로 낮은 값을 가짐을 확인할 수 있었다.

그리고 최고 성능을 낸 KoBERT[4]모델의 17가지 카테고리별 분류 정확도 성능을 비교 분석하여 rental, pets, culture service의 분류 정확도가 낮음을 확인하였고, 해당 카테고리를 별도 모델로 학습하여 아래 Table 8과 같이 성능 향상을 도출할 수 있었다.

Table 8. KoBERT Accuracy per Category

No	Category	KoBERT Accuracy (%)	rental/pets/culture service discrete KoBERT Accuracy(%)
1	clothes	98.5	98.5
2	beauty	98.1	98.2
3	bag	98.2	98.2
4	furniture	95.4	96.1
5	digital appliance	96.3	97.5
6	health	97.2	97.7
7	miscellaneous goods	96.3	97.4
8	daily necessities	95.8	96.9
9	shoes	96.1	96.0
10	sports	97.8	97.7
11	food	97.7	97.7
12	hobby	95.7	96.8
13	kitchen appliance	96.5	97.3
14	child	95.8	96.8
15	rental	93.3	96.5
16	pets	93.5	97.2
17	culture service	92.7	96.9

6. 결론 및 향후 계획

본 논문에서 BOW, LSTM, Attention, CNN[2], ELMo[3], KoBERT[4] 모델을 구현하여 분류 성능을 비교하였는데 문맥에 따라 동의어 임베딩을 달리 처리하는

전이학습 모델 KoBERT가 가장 좋은 성능을 보였다.

또한 17가지 카테고리 별, 상품평 분류 성능을 비교 분석하여 분류 성능이 좋지 않은 특정 카테고리의 상품 평 단어 패턴이 상이함을 확인하였고 해당 카테고리의 감성 분석 모델을 별도 구성하여 고성능을 확보하는 방안을 제시하였다.

마지막으로 모델 용량 및 학습시간을 비교하여 KoBERT가 성능은 우수했지만 추론 속도가 타 모델에 비해 다소 느리다는 점을 확인하여 경량화하는 연구 내용[16]을 적용해 보는 것을 향후 계획으로 제시해 보고자 한다

REFERENCES

[1] Hochreiter & Schmidhuber. (1997). *LONG SHORT-TERM MEMORY*. Neural Computation, DOI: 10.1162/neco.1997.9.8.1735

[2] Lecun. (1998). Gradient-Based Learning Applied to Document Recognition. *IEEE, 86(11)*, 2278-2324. DOI: 10.1109/5.726791

[3] M. Peters. (2018). *ELMo-Deep contextualized word representations*. NAACL 2018. <https://arxiv.org/abs/1802.05365>

[4] SKTBrain, KoBERT. (2019). <https://github.com/SKTBrain/KoBERT>

[5] J. Devlin, K. Lee & K. Toutanova. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. <http://arxiv.org/abs/1810.04805>

[6] H. M. Kim & K. B. Park. (2019). Sentiment analysis of online food product review using ensemble technique. *Journal of Digital Convergence, 17(4)*, 115-122. DOI: 10.14400/JDC.2019.17.4.11

[7] H. Y. Park & K. J. Kim. (2019). Sentiment Analysis of Movie Review Using Integrated CNN-LSTM Model. *Journal of Intelligence and Information Systems, 25(4)*, 141-154. DOI : 10.13088/jiis.2019.25.4.141

[8] D. Bahdanau, K. H. Cho & Y. Bengio (2014) *Neural Machine Translation by Jointly Learning to Align and Translate*. ICLR 2015. <https://arxiv.org/abs/1409.0473>

[9] Y. Kim. (2014) *Convolutional Neural Networks for Sentence Classification*. EMNLP 2014. <https://arxiv.org/abs/1408.5882>

[10] Vaswani et al. (2017). *Attention is all you need*. <https://arxiv.org/abs/1706.03762>

[11] T. Kudo & J. Richardson. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.

EMNLP2018, 66-71
<https://arxiv.org/abs/1808.06226>

[12] A. Joulin. (2016). *FastText.zip: Compressing text classification models*. ICLR 2017. <https://arxiv.org/abs/1612.03651>

[13] A. F. Agarap. (2018). *Deep Learning using Rectified Linear Units (ReLU)*, 1, 2-8. <https://arxiv.org/abs/1803.08375>

[14] D. P. Kingma & J. Ba. (2014). *Adam: A Method for Stochastic Optimization*. 1-15. <https://doi.org/http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503>

[15] D. Hendrycks & K. Gimpel. (2016). *Gaussian Error Linear Units (GELUs)*. <https://arxiv.org/abs/1606.08415>

[16] M. A. Gordon - All The Ways to Compress <http://mitchgordon.me/machine/learning/2019/11/18/all-the-ways-to-compress-BERT.html>

최 준 영(Jun-Young Choi)

[학생회원]



- 2013년 2월 : 한국항공대학교 항공전자학부 (공학사)
- 2018년 3월 ~ 현재 : 고려대학교 컴퓨터정보통신대학원 석사과정
- 관심분야 : 자연어처리, 개인화 추천
- E-Mail : geni01@korea.ac.kr

임 희 석(Heui-Seok Lim)

[종신회원]



- 1992년 2월 : 고려대학교 컴퓨터학과 (공학사)
- 1994년 2월 : 고려대학교 컴퓨터과 (공학석사)
- 1997년 2월 : 고려대학교 컴퓨터학과 (공학박사)
- 2008년 3월 ~ 현재 : 고려대학교 정보대학 컴퓨터학과 교수
- 관심분야 : 자연어처리, 정보검색
- E-Mail : limhseok@korea.ac.kr