Multi-channel 과 Densely Connected Convolution Networks 을 이용한 한국어 감성분석

윤민영*, 구민재*, 이병래*
* 방송통신대학교 정보과학과
e-mail : roundlifeymy@knou.ac.kr and cgkmj@knou.ac.kr and brlee@knou.ac.kr

Korean Sentiment Analysis using Multi-channel and Densely Connected Convolution Networks

Min-Young Yoon*, Min-Jae Koo*, Byeong Rae Lee*
*Dept. of Computer Science, Korea National Open University

요 약

본 논문은 한국어 문장의 감성 분류를 위해 문장의 형태소, 음절, 자소를 입력으로 하는 합성곱층과 DenseNet 을 적용한 Text Multi-channel DenseNet 모델을 제안한다. 맞춤법 오류, 음소나 음절의 축약과 탈락, 은어나 비속어의 남용, 의태어 사용 등 문법적 규칙에 어긋나는 다양한 표현으로 인해 단어 기반 CNN 으로 추출 할 수 없는 특징들을 음절이나 자소에서 추출 할 수 있다. 한국어 감성분석에 형태소 기반 CNN 이 많이 쓰이고 있으나, 본 논문에서 제안한 Text Multi-channel DenseNet 모델은 형태소, 음절, 자소를 동시에 고려하고, DenseNet 에 정보를 밀집 전달하여 문장의 감성 분류의 정확도를 개선하였다. 네이버 영화 리뷰 데이터를 대상으로 실험한 결과 제안 모델은 85.96%의 정확도를 보여 Multi-channel CNN 에 비해 1.45% 더 정확하게 문장의 감성을 분류하였다.

Keywords: 한국어 감성분석, Korean Sentiment Analysis, Multi-channel DenseNet, Text Classification, DenseNet

1. 서론

자연어 처리는 사람이 사용하는 언어를 컴퓨터가 인식하여 처리할 수 있도록 하는 기술이다. 감성분석(sentiment analysis)은 자연어 분석 기술 중 하나로 대중이 생산한 비정형 문서를 대상으로 문서 속에 내포된 감성 및 감정을 추출해 내는 기술이다. 감성분석은 문서에서 가장 작은 의미적 단위인 단어의 감성 극성(sentiment polarity)에 기반을 두고 있다.

감성분석은 상품에 대한 만족도 조사, 정치적 이슈에 대한 여론조사, 다양한 산업분야, SNS(social network service)에서 생산되는 데이터를 기반으로 의견분석(opinion mining)에 활용된다.

기존의 감성분석에는 Naive Bayes 와 Logistic Regression 등의 전통적인 기계학습 방법이 많이 사용되었으나, 최근에는 딥러닝 기반의 기술들이 많은 분야에서 높은 성능을 기록했다[1] [6].

대부분의 감성분석 연구들은 대용량 영어에 초점이 맞추어져 있다[4][9][6]. 하지만 영어와는 달리 교착어인 한국어는 조사와 어미가 다양해서 어근을 추출해야 한다. 즉 영어에 적용한 기술이나 모델을 그대로 사용하면 한국어 감성분석에는 정확하지 않은 결과가 나올 수 있다. 높은 성능의 결과를 기대하기 위해서는 언어의 특성에 맞는 기술과 모델을 사용해야 한다.

CNN 은 성능 향상을 위해 합성곱층을 다양한 방법을 통해 깊게 구성하였다. 그러나 합성곱층이 깊어짐으로써 Vanishing Gradient 등의 문제로 멀리 떨어진 합성곱층까지 정보가 전달되지 않아 원하는 결과를 얻기 위해 학습시키는 것이 점점 어려워졌다. DenseNet(Densely Connected Convolutional Networks)은 크기가 동일한 특징 맵 사이의 연결을 통해 멀리 떨어진 합성곱층까지 정보를 전달하도록 하였다.

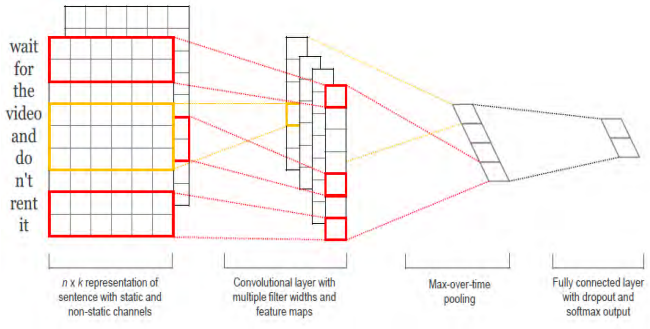
본 논문에서는 한국어 문장의 감성 분류에 효과적인 Text Multi-channel DenseNet 모델을 제안한다. 깊은 합성곱 층을 사용하여 정보 전달 가능한 네트워크 구성이 목표이다. 연구에 사용된 데이터는 네이버 영화 리뷰 평점을 기준으로 긍정과 부정 2 개의 클래스로 나누어진 데이터 셋이다. 영화 리뷰 데이터를 세 개의 다른 합성곱층을 이용하여 형태소, 문자, 자소 모두 입력 값으로 받은 후 DenseNet 을 통하여 특징을 추출한다.

본 논문의 구성은 다음과 같다. 서론에 이어 제 2 장에서는 제안 Text Multi-channel DenseNet 모델과 관련된 연구를 소개한다. 제 3 장과 4 장 에서는 제안한 모델 아키텍쳐, 데이터셋, 실험의 결과를 분석한다. 마지막으로 제 5 장에서는 연구의 결론 및 향후 연구를 기술한다.

2. 관련연구

전통적으로 문서 분류를 하기 위해서 형태소를 최소 단위로 사용해왔다. 하지만 형태소 단위로 문서를 분해하는 과정에서 맞춤법 오류, 음소나 음절의 축약과 탈락으로 인해 정보가 손실될 가능성이 있다.

Kim, Yoon 은 컴퓨터 비전을 위해 고안된 자연어 처리에 CNN 을 적용하였다[1]. Kim, Yoon 이 제안한 어절 기반의 CNN 모델은 (그림 1)과 같은데, 교착어인 한국어의 특성상 심각한 OOV(Out-of-Vocabulary)가 발생하는 문제가 있다.



(그림 1) 어절 기반 CNN [1]

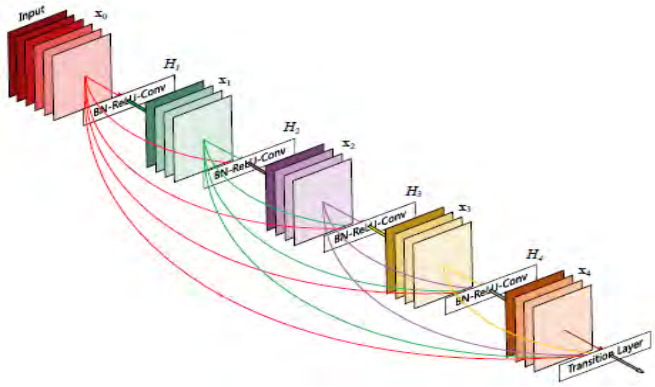
한국어에서 음절 기반의 CNN 모델을 제안한 연구로는 [10]이 있다. [10]에서는 학습되지 않은 새로운 단어들에 대해서도 예측을 할 수 있도록 형태소 기반이 아닌 음절 기반의 CNN 모델을 사용하였다.

모경현[5]은 한국어에서 단어 단위와 자소 단위 특성을 결합하여 문서 분류를 수행하는 방법론을 제안하였다. 자소 정보를 활용하는 방식은 국소적인 정보를 보존할 수 있음과 동시에 문법 파괴적인 문서 분류에도 이점이 있다. 맞춤법 오류, 음소나 음절의 축약과 탈락은 온전한 단어로서 표현하기에 어려움이 있는 문장 정보들이지만 자소 단위의 분석은 손실된 정보를 추출할 수 있다.

김민[7]은 형태소, 음절, 자소 세가지 모두를 이용하는 Multi-channel 모델을 통하여 감성분류를 하였다. 음절을 고려하지 않은 모델보다 Multi-channel 모델이 좋은 성능을 기록했다.

기존 연구에서는 얕은 합성곱층 기반으로 연구가 진행되었으며, 깊은 합성곱 층을 사용하면 성능이 나오지 않았다[4].

DenseNet 모델은 Dense Block 과 Transition layer 가 반복적으로 연결된다[11]. Dense Block 은 합성곱층의 출력이 이후에 오는 모든 합성곱층들의 입력에 연결이 되도록 Dense connectivity 을 구축함으로써, 더 깊게 층을 쌓을 수 있게 한다. 합성곱층의 수가 늘어나면 전체 파라미터가 증가하여 훈련이 제대로 이루어지지 않는다. 이런 현상을 방지하기 위해 합성곱층의 출력 특징맵 수를 의미하는 성장률을 사용하여 파라미터 증가를 제한했다.



(그림 2) 5-layer dense block, growth rate k = 4 [2]

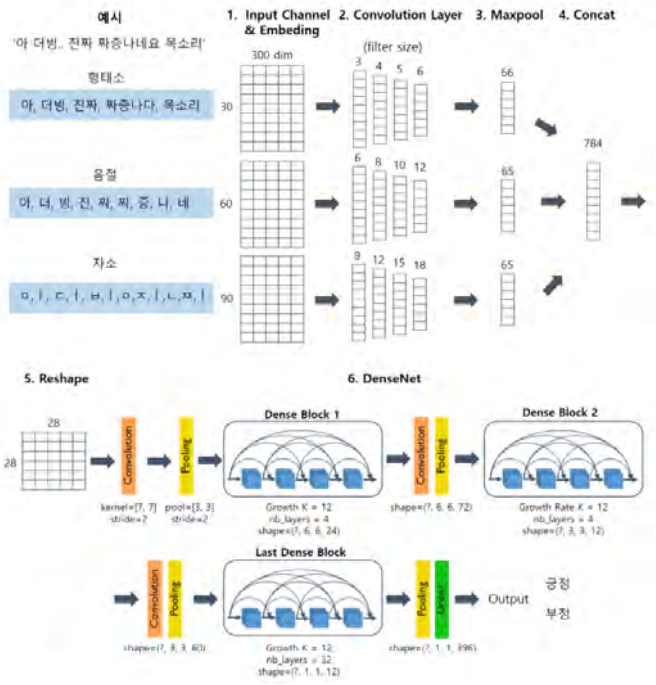
Lee[6]는 캐릭터 레벨 특징 추출에서 Densely Connected Networks 을 함께 적용하여 대용량 데이터셋으로 슬롯 태깅, 품사 태깅(POS), 개체명 인식(NER)에서 성능 평가를 진행하여 검증하였다.

본 논문에서는 한국어 문장의 특징을 사용할 수 있는 형태소, 음절, 자소를 동시에 고려하고, DenseNet 을 사용하여 특징 정보를 밀집시켜 전달하여 문장의 감성을 분류 성능을 개선 하고자 한다.

3. 제안한 방법

3.1 모델

본 논문에서는 Kim, Yoon 의 CNN 모델[1]과 Huang 의 DenseNet[2]를 결합한 Text Multi-channel DenseNet 모델을 제안한다. 한국어의 특징을 나타낼 수 있는 형태소, 음절, 자소를 입력 값으로 사용하였고, 이후 넓고 얕은 합성곱의 한계를 극복하기 위해 각각의 입력을 maxpool 하고 연결한 정보를 DenseNet 을 사용하여 입력된 정보를 깊은 합성곱층에 전달 후 문장 분류를 하였다.



(그림 3) Text Multi-channel DenseNet 모델

3.2 데이터셋

공개된 네이버 영화 리뷰 코퍼스 v1.0[13] 데이터를 이용하여서 모델들을 학습시키고 테스팅 하였다.

<표 1> 네이버 영화 리뷰 코퍼스 v1.0

평균 문장길이(글자)	35.24
감성 개수	2
학습 단위 크기	150000
테스트 단위 크기	50000
형태소 Vocab 크기	49309
음절 Vocab 크기	2830
자소 Vocab 크기	516
형태소+음절+자소 Vocab 크기	50157

<표 2> 영화 리뷰 감성 예제

문장	감성
"아 더빙.. 진짜 짜증나네요 목소리"	부정
"너무재밋었다그래서보는것을추천한다"	긍정
"교도소 이야기구먼 ..솔직히 재미는 없다.. 평점 조정"	부정
"액션이 없는데도 재미 있는 몇안되는 영화"	긍정

3.3 환경

개발 환경은 윈도우 10 OS 보급형 노트북을 사용하였으며 사양은 CPU i7-7700HQ, RAM 16GB, 그래픽카드 GTX 1060 6GB, SSD 256 GB 에서 개발되었다. 프로그램 환경은 아나콘다 가상환경에 파이썬 3.6, 텐서플로우 1.13.0 에서 개발되었다.

4. 실험

4.1 실험 모델

- 형태소 CNN

Kim, Yoon 의 CNN 모델[1]을 사용하였다. 어절 단위로 문장을 분리하여서 CNN 의 입력 값으로 사용하였다. konlpy[15]를 사용하여 형태소 분석을 하였고, gensim[14]의 Word2Vec 모델을 사용하였다. 특징을 추출하고 maxpooling 한 후 문장 분류를 하였다.

- 음절 CNN

전처리 과정에서 단어에 Space 를 추가하여 음절로 나누어 사용하였다. konlpy 를 사용하여 형태소 분석을 하였고, gensim 의 Word2Vec 모델을 사용하였다. 특징을 추출하고 maxpooling 한 후 문장 분류를 하였다.

- 자소 CNN

전처리 과정에서 단어를 hgtk[16]를 사용하여 한글 자모 분리 작업을 진행하여 자소로 분리하였다. konlpy 를 사용하여 형태소 분석을 하였고, gensim 의 Word2Vec 모델을 사용하였다. 특징을 추출하고 maxpooling 한 후 문장 분류를 하였다.

- Multi-channel CNN (형태소 + 음절 + 자소)

형태소, 음절, 자소를 각각 임베딩한 합성곱층을 생성하고 특징을 추출한 후 각 정보를 연결 후, 특징을 추출하고 maxpooling 한 후 문장 분류를 하였다.

- Multi-channel DenseNet (형태소 + 음절 + 자소)

형태소, 음절, 자소를 각각 임베딩한 합성곱층을 생성하고 특징을 추출한 후 각 정보를 연결 하였다. 특징을 추출하고 maxpooling 한 후 DenseNet 을 거친 후 문장 분류를 하였다.

모델 별 파라미터는 다음과 같다.

<표 3> 모델 별 파라미터

모델	Epoch	Batch size	Sequence length	Filters	Learning Rate
형태소 CNN	15	256	30	3, 4, 5	1e-4
음절 CNN	15	256	30	3, 4, 5	1e-4
자소 CNN	15	256	30	3, 4, 5	1e-4
Multi-channel CNN (형태소+음절+자소)	15	256	30, 30, 30	3, 4, 5	1e-4
Text Multi-channel DenseNet (형태소+음절+자소)	20	256	30, 60, 90	3, 4, 5, 6	1e-4

4.2 실험 결과

실험결과는 모델 별 CNN 의 감성 분석하여 정확도를 <표 4>에 정리하였다

<표 4> 감성분석 실험 결과

모델	정확도
형태소 CNN	83.55%
음절 CNN	81.73%
자소 CNN	75.30%
Multi-channel CNN (형태소+음절+자소)	84.51%
Text Multi-channel DenseNet (형태소+음절+자소)	85.96%

실험결과 Multi-channel 의 정확도 84.51%, Text Multi-channel DenseNet 85.96% 로 1.45% 더 정확하게 문장의 감성을 분류하였다.

형태소 CNN 모델이 음절 CNN, 자소 CNN 모델 보다 성능이 높고, Multi-channel CNN 이 형태소 CNN 보다 높다. 본 논문에서 제안한 형태소, 음절, 자소를 동시에 사용하는 Text Multi-channel DenseNet 모델이 영화 리뷰 데이터에서 가장 높은 정확도로 문장을 감성분석 하였다.

형태소가 의미를 가지는 최소 단위이기 때문에 음절이나 자소 기반 CNN 보다 감성분류 성능이 더 높았다. 그러나 형태소 기반 CNN 이 올바르게 분류하지 못한 문장들을 음절 기반 CNN 이나 자소 기반 CNN 이 올바르게 분류 한 경우도 있다.

본 논문에서 제안한 Text Multi-channel DenseNet 은 형태소뿐만 아니라 음절과 자소를 동시에 이용함으로써 Multi-channel CNN 에 비해 문장들의 감성을 정확하게 분류할 수 있었다. 음절과 자소를 같이 이용함으로써 형태소 기반 CNN 의 큰 문제점인 OOV(Out-

of-Vocabulary) 문제를 개선할 수 있었다. Multi-channel CNN 과 달리 제안 모델에서는 DenseNet 을 사용였는데, 정보 전달에 있어 성능이 향상됨을 알 수 있다. 감성분석에 사용된 타 모델에서는 넓고 얕은 CNN 을 사용하였으나, 제안 모델은 깊은 학습을 통해 특징 정보 유실의 문제를 개선하였다.

5. 결론 및 향후 연구

본 논문에서 여러 실험을 통해서 한국어 구어체 감성 분석에 효과적인 형태소, 자소, 음절 기반 Text Multi-channel DenseNet 을 제안하였다. OOV(Out-of-vocabulary) 문제를 가지는 형태소 기반 CNN 의 문제점을 개선하였으며, 음절과 자소에서 추출한 특징벡터를 형태소 기반의 특징벡터와 상호보완적으로 사용할 수 있음을 확인하였다. 깊은 신경망을 사용한 DenseNet 을 사용하여 기존에 많이 사용된 형태소 및 음절 기반 CNN 보다 높은 감성분석 정확도를 보여주어 한국어 구어체를 분류하는 연구에도 활용 가능성이 있음을 확인하였다. 향후 연구는 임베딩 단계에서 한국어 감성분석 성능 향상을 위한 방법 연구와 CRF(Conditional Random Fields)를 사용한 한국어 감성분석에 성능을 향상 방법을 연구하고자 한다.

6. 참고 문헌

[1] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[2] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).

[3] Conneau, Alexis, et al. "Very deep convolutional networks for text classification." arXiv preprint arXiv:1606.01781 (2016).

[4] Le, Hoa T., Christophe Cerisara, and Alexandre Denis. "Do convolutional networks need to be deep for text classification?." Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[5] 모경현, 박재선, 장명준, & 강필성. 단어와 자소 기반 합성곱 신경망을 이용한 문서 분류. 대한산업공학회지, 44(3), 180-188. (2018).

[6] Lee, Chanhee, et al. "Character-Level Feature Extraction with Densely Connected Networks." arXiv preprint arXiv:1806.09089 (2018).

[7] 김민, 변증현, 이충희, 이연수. Multi-channel CNN 을 이용한 한국어 감성분석, 제 30 회 한글 및 한국어 정보처리 학술대회 논문집, (2018): 79-83.

[8] 조휘열, et al. "컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술." 한국정보과학회 학술발표논문집 (2015): 792-794.

[9] Zhang, Ye, and Byron Wallace. "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification." arXiv preprint arXiv:1510.03820 (2015).

[10] S. Choi et al. A Syllable-based Technique for Word Embeddings of Korean Words, Proceedings of the First Workshop on Subword and Character Level Models in NLP, pp. 36-40, (2017).

[11] Convolutional Neural Network for Text Classification in Tensorflow, https://github.com/dennybritz/cnn-text-classification-tf

[12] Densenet-Tensorflow, https://github.com/taki0112/Densenet-Tensorflow

[13] Naver sentiment movie corpus v1.0, https://github.com/e9t/nsmc

[14] gensim, https://radimrehurek.com/gensim/index.html

[15] konlpy, https://konlpy-ko.readthedocs.io/ko/v0.4.3/

[16] hgtk, https://github.com/bluedisk/hangul-toolkit