합성곱 신경망을 이용한 뉴스 분류

An Approach to Classify News Using Convolutional Neural Network

저자
(Authors)

유승재, 이수영, 권나영, 이희수

출처
(Source)

한국정보과학회 학술발표논문집 , 2018.12, 1815-1817(3 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07614103

APA Style

유승재, 이수영, 권나영, 이희수 (2018). 합성곱 신경망을 이용한 뉴스 분류. 한국정보과학회 학술발표논문집, 1815-1817

합성곱 신경망을 이용한 뉴스 분류

유승재·이수영·권나영·이희수

서강대학교 컴퓨터공학과

·An Approach to Classify News Using Convolutional Neural Network

요 약

합성곱 신경망(Convolutional Neural Network)은 이미지 처리에 쓰이는 학습 알고리즘이다. 이 알고리즘을 이용하여, 텍스트 분류에 적용한 논문들이 현재 진행형으로 되고 있다. 본 논문에서는 합성곱 신경망을 이용하여 뉴스 기사를 학습시키고, 학습 된 모델을 이용하여 분류를 테스트 해 보았다. 총 23개의 카테고리로 분류를 한 결과 88.9%의 정확도가 나왔다. 이는 이미지 처리 알고리즘이 자연어 처리에도 적용될 수 있음을 시사한다.

1. 서 론

CNN 알고리즘은 이미지 분석에 많이 쓰이는 알고리즘이다. 최근 들어 CNN알고리즘을 이용하여 자연어 처리에(영문)도 매우 뛰어난 성능을 보인다는 연구결과가 있다. 따라서 이를 실험해보기 위하여 직접 CNN알고리즘을 구현한 후, 테스트를 해 보았다.

먼저 학습용 자료는 쉽게 구할 수 있는 데이터로 뉴스 데이터를 선택하였고, 그 중 네이버 뉴스에 있는 뉴스를 크롤링 하여 데이터 셋을 얻었다. 약 100만 건에 가까운 뉴스를 크롤링 하였으며, 네이버의 분류 기준으로 뉴스는 총 23개 카테고리로 분류하였다.

카테고리는 정치, 경제, 금융, 부동산, 교육, 사회, 취업, 음식, 건강, 문화, 생활, 자동차, 패션/뷰티, 여행, 종교, 세계, e스포츠/게임, IT/인터넷/통신, 모바일, 스포츠, 엔터테인먼트, 영화/뮤직, 컴퓨터로 구성되어 있으며, 학습용 데이터를 크롤링 할 때, 너무 한 쪽에만 데이터가 크롤링 되지 않도록 비율을 적절히 유지하며 크롤링을 진행 하였다.

크롤링 된 데이터는 23개의 카테고리로 총 907,933건을 준비 하였다.

먼저 크롤링 된 뉴스를 학습용 데이터 셋으로 만들어야 한다. 학습용 데이터 셋으로 만들기 위하여, 약 9만개의 뉴스를 형태소분석기를 이용하여 뉴스를 학습용 데이터 셋으로 변환하였다. 이 때, 영문 분류에는 명사(Noun)만을 추출하였을 때 가장 분류를 잘 하였다는 결과가 있어, 학습용 데이터 셋을 두 가지 분류로 나누었다. 첫 번째는 모든 문장들을 형태소 분석하여 저장한 데이터이고, 두 번째는 그 데이터 중 명사만 추출하여 저장한 데이터이다.

2. 모 델

학습 모델을 만들기 위하여 인용한 논문의 모델을 소개한다.

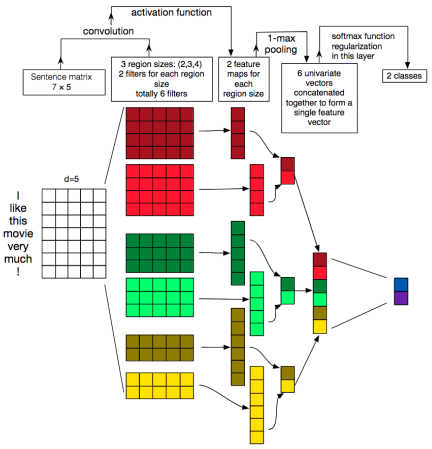

Figure 1: Illustration of a CNN architecture for sentence classification. We depict three filter region sizes: 2, 3 and 4, each of which has 2 filters. Filters perform convolutions on the sentence matrix and generate (variable-length) feature maps; 1-max pooling is performed over each map, i.e., the largest number from each feature map is recorded. Thus a univariate feature vector is generated from all six maps, and these 6 features are concatenated to form a feature vector for the penultimate layer. The final softmax layer then receives this feature vector as input and uses it to classify the sentence; here we assume binary classification and hence depict two possible output states.

위 이미지는 A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification 논문에[1] 등장한 문장 분류를 위한 합성곱 신경망(CNN) 아키텍처다. 초기에 단어를 벡

터로 임베딩하는 과정이 진행되어야 함을 제외하면 이미지 분류와 크게 다르지 않다. 3개의 필터 사이즈 2, 3, 4를 각 두 개씩 총 6개를 문장 매트릭스에 합성곱을 수행하고 피쳐 맵을 생성한다. 이후 각 맵에 대해 맥스 풀링을 진행하여 각 피쳐 맵으로부터 가장 큰 수를 남긴다. 이들 6개 맵에서 단변량 univariate 벡터가 생성되고, 이들 6개 피쳐는 두 번째 레이어를 위한 피쳐 벡터로 연결한다. 최종적으로 소프트맥스 레이어는 피쳐 값을 받아 문장을 분류한다. 여기서는 이진 binary 분류를 가정했고 따라서 두 가지 가능한 출력 상태를 묘사했다.

추가적으로, 한국 뉴스 분류 모델을 생성하기 위하여 이 과정 전에 형태소 분석기가 들어갔으며 모델을 구현할 때 사용한 핵심 변수들은 다음과 같다. 이 변수들은 김윤 박사의 논문을[2] 참고하여 설정하였다.

batch_size - 한 번에 처리하는 뉴스의 수

embedding_dim - embedding의 차원

filter_size s- convolutional filter에서 다루기를 원하는 단어의 수

num_filters - 필터 크기 당 필터 수

num_epochs - 세대 수

3. 실 험

3.1 데이터 셋

앞서 언급했듯이, 학습용 데이터 셋으로는 네이버 뉴스에서 크롤링 된 907,933건에서 0.77대 0.23으로 무작위 분류하여 0.77인 699,108을 사용하였고, 나머지 208,825건의 뉴스는 생성한 모델로 테스트하기 위하여 남겨두었다.

총 분류해야 하는 카테고리는 네이버의 뉴스 분류 기준으로 잡아 총 23개 카테고리로 하였으며, 각각 카테고리의 뉴스 비율은 동일하게 잡고 크롤링을 하였다.

3.2 실험 환경

3.2.1 학습 환경

학습 환경으로는 형태소분석기 Mecab이 윈도우즈는 지원이 되지 않아 Ubuntu 16.04에서 진행하였다. 언어는 알고리즘을 구현하기 쉬운 python3.5로 하였고 사용한 모듈 중 핵심이 되는 모듈인 텐서플로우 모듈은 CUDA가속이 지원되는 tensorflow-gpu==1.4.1을 사용하였다.

학습에 사용한 컴퓨터의 환경은 다음과 같다.

	CASE 1	CASE 2
CPU	E5-1680 v4 @ 3.40GHz	E5-1680 v4 @ 3.40GHz
GPU	Geforce GTX TITAN X	Geforce GTX 1080
RAM	512GB	512GB

학습은 이렇게 세종대학교 클라우드 컴퓨팅 서비스에서 서버를 빌려 두 가지 케이스로 나누어 실험해 보았다.

학습은 최고 성능까지 두 케이스 전부 약 2시간 이내에 도달하였다.

3.2.2 결과 비교

적용 모델	정확도	비고
영화 감상평 약 547,413건 (단문)	91.52%	논문[3]
뉴스 907,933건 (장문 및 복수 문장)	88.9%	

논문 [3]의 결과에서는 단문의 문장들로 이루어진 영화 평점 데이터셋을 가지고 총 5가지로 분류를 하였다. 그 결과 91.52%라는 매우 높은 성능을 보였다.

본 논문의 실험결과로는 단문이 아닌 여러 문장으로 된 뉴스와 5건이 아닌 총 23건으로 학습하여 분류 하였다. 실험결과로 볼 때, CNN알고리즘은 단순한 문장 분류가 아닌, 복잡한 문장에서도 분류 성능이 매우 뛰어남을 알 수 있고 더 나아가 여러 분류 기준에 대해서도 효과적으로 분류 한다는 것을 알 수 있다.

3.3. 실험 결과

형태소분석기와 CNN알고리즘을 이용하여 뉴스를 분류한 결과, 가장 높은 정확도는 학습용 데이터 셋 중 명사만을 추출한 것이 가장 높게 나왔다.

예측 정확도는 88.9%가 나왔으며, 오차가 나온 것의 대부분은 정치, 사회 분야였다. 다시 말해, 태깅 된 데이터는 정치이지만 모델은 사회로 분류하였거나 원래 태깅 된 것은 사회이지만 모델은 정치로 분류 된 것이 오류의 거의 대부분을 차지하였다.

모델을 생성할 때, CNN 알고리즘 앞에 단어를 벡터화하는 알고리즘인 Word2Vec 알고리즘을 사용했었다. 하지만 한국어는 기존의 영어로 대상으로 한 Word2Vec 알고리즘 보다 자모를 이용한 Sub-Word-level-Vector Representation을 통해 자연어처리 연구를 수행하는 것이 효과적임이 알려져 있다.

CNN알고리즘을 이용하여 한국어 텍스트분류에 관한 논문은 많이 있다. 하지만 전부 다 한국어 단문 글들을

분류한 결과이며 분류의 기준이 되는 범주도 5개 내외이다. 본 논문에서는 단문위주의 글이 아닌 장문의 한국어 문단과 많은 카테고리에 대해서도 적용이 가능하다는 것을 입증했다.

4. 맺음말

본 논문에서는 뉴스기사를 데이터셋으로 하여 CNN알고리즘의 자연어처리의 적용에 대한 가능성을 연구하였다. 결과는 매우 긍정적인 결과가 나왔으며, 추후 Sub-Word-Level-Vector 알고리즘을 이용하면 한국어 처리 능력이 더욱 개선될 것이다.

논 문 사 사

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음. (2015-0-00910)

참 고 문 헌

[1] Ye Zhang and Byron C. Wallace “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification“. page 4. 2016.
[2] Yoon Kim. “Convolutional Neural Networks for Sentence Classification“ EMNLP, page 1750. 2014.
[3] Geonyeong Kim and Changki Lee “Korean Movie Review Sentiment Analysis Using Convolutional Neural Network“. page 747-749. 2016.