|  |
| --- |
| **1차원 합성곱 신경망 모델을 활용한 국문/영문 감성분석**  이창훈  한국 오라클, 오라클 컨설팅  Jackson.lee@oracle.com  **Korean/English sentiment analysis using one-dimensional convolutional neural network model**  Changhun-Lee  Oracle Korea, Oracle Consulting  **요 약**  인공 신경망 모델로 많이 활용되었던 합성곱 신경망 모델을 활용하여  Naver sentiment movie corpus v1.0 과 Freinds 데이터의 감성을 분류하는 모델을 구현하려고 합니다. 레이어의 개수에 따라서 모델을 구성하고 분류결과를 출력하여 레이어에 따른 감성분석의 결과가 어떻게 달라지는 고찰합니다. |

**1. 서론**

감성분석에 규칙기반과 통계기반의 연구는 이전 연구들에서 성능 향상에 제한이 많이 존재하였습니다. 하지만 현재는 딥러닝 기법들을 활용한 자연어처리 기법들이 많이 나왔고 성능 측면에서도 높은 성능(80%이상)에 도달하였습니다. 본 논문에서는 인공신경망의 초기 모델인 합성곱 신경망을 활용하여 실제로 어느 정도의 성능으로 범주를 분별하는지 실험하는 논문입니다. 데이터로는 네이버 영화 리뷰 코퍼스 v1.0 과 영화 프렌즈 대사 데이터를 전처리하고 모델을 학습시켜 테스트 결과를 도출하였습니다.

**2. 관련 연구**

2.1 Opinion Mining

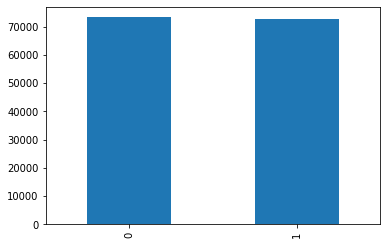
데이터는 크게 정형 데이터와 비정형 데이터로 구분된다. 데이터베이스에 일정한 형태로 저장할 수 있는 정형 데이터와 문서, 이미지, 영상 등 데이터의 필드가 정해져 있지 않은 비정형 데이터이다. 빅 데이터의 85% 이상이 비정형 데이터이며, 오피니언 마이닝은 주로 비정형 데이터를 분석한다. 오피니언 마이닝은 흔히 감성분석이라고도 불리며, 소셜 미디어 기반의 오피니언 마이닝은 트렌드의 파악, 제품/서비스 평가, 미래 예측 등에 활용된다.[1] 정치, 사회적 이슈뿐만 아니라, 연예나 스포츠 등의 분야에서의 트렌드의 파악은 매우 중요한 이슈이며, 기업의 매출과 직결되는 제품/서비스 평가는 보다 직접적인 활용도를 지닌다. 오피니언 마이닝은 의사결정의 보조 수단이며, 주가나 국가경제위기 예측 등에 다양하게 활용되고 있다[2]

2.2 Convolution Neural Network

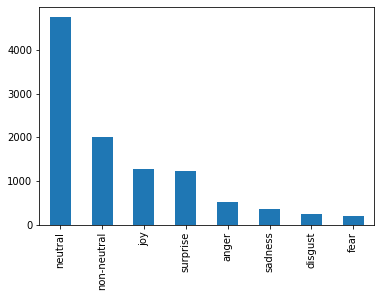
CNN은 하나의 뉴런을 여러 번 복사해서 사용하는 Neural Network라고 말 할 수 있겠다. 그렇게 함으로써 CNN은 뉴런의 행태를 보여주는 (실제 학습이 필요한)Parameter의 개수를 꽤나 작게 유지하면서도, 굉장히 많은 뉴런을 가지고 방대한 계산을 필요로 하는 모델을 표현할 수 있다.[3] 합성곱층은 합성곱 연산을 통해서 이미지의 특징을 추출하는 역할을 합니다. 우선, 합성곱 연산에 대해서 이해해봅시다. 합성곱은 영어로 컨볼루션이라고도 불리는데, 커널(kernel) 또는 필터(filter)라는 n×m 크기의 행렬로 높이(height)×너비(width) 크기의 이미지를 처음부터 끝까지 겹치며 훑으면서 n×m크기의 겹쳐지는 부분의 각 이미지와 커널의 원소의 값을 곱해서 모두 더한 값을 출력으로 하는 것을 말합니다. 이때, 이미지의 가장 왼쪽 위부터 가장 오른쪽까지 순차적으로 훑습니다.(커널(kernel)은 일반적으로 3 × 3 또는 5 × 5를 사용합니다.) 입력으로부터 커널을 사용하여 합성곱 연산을 통해 나온 결과를 **특성 맵(feature map)**이라고 합니다. CNN에서는 이 특성 맵을 최적화하는 것이 목표입니다.

**3. 개요도**

3.1 데이터 탐색



네이버 영화 리뷰 코퍼스는 15만개가 준비되었습니다. Label 칼럼의 0, 1의 개수는 각각 73,342개, 72,841개로 확인하였습니다.



프렌즈 대사 코퍼스는 총 10,561개의 대사가 준비되었습니다. Emotion의 종류는 총 8가지이고 neutral:4752, non-neutral:2017, joy:1283, surprise:1220, anger:513, sadness:351, disgust: 240, fear:185 개로 확인하였습니다.

3.2 데이터 전처리

두 종류의 코퍼스 모두 공통으로 해당 언어에 포함되지 않는 언어나 특수 문자는 모두 정규표현식을 통해서 제거하였습니다. 이후 널 값 여부를 확인하여 모두 제거하였습니다. 널 값이 데이터에 포함되어 있을 경우 모델을 훈련시키는데 문제가 발생하기 때문입니다. 세 번째 작업으로 불용어를 제거하는 작업을 하였습니다. 네이버 영화 리뷰 감성 분석의 경우 기본적인 조사를 불용어리스트를 작성하여 for문을 통해서 제거하였습니다. 프렌즈 영화 스크립트 같은 경우 nltk(Natural Language tool kit)에서 제공하는 영어 불용어 리스트를 임포트하여 마찬가지로 for문을 통해서 제거하였습니다. 불용어처리는 특히 실무적인 자연어처리 작업에 아주 중요한 처리이기 때문에 각 분야에 맞는 불용어를 추가해야 합니다. 네 번째 전처리는 특정 단어의 빈도수가 2회 이상인 희귀한 단어들을 제거하는 작업을 하고 최대 문장의 길이를 30으로 맞추는 패딩작업 이후 학습을 위한 tokenization 하고 sequence 작업으로 전처리를 마무리하였습니다. 특히 프렌즈 대사 코퍼스 같은 경우 비정형 데이터(json) 구조로 되어있어 로우 리스트를 모두 피벗 작업을 하여 한 개의 리스트로 데이터 전처리를 마무리하였습니다. [그림1] 720 x 24개의 구조이지만 구조별로 null 데이터가 존재하기 때문에 최종적으로 10561행의 list 데이터로 모델을 돌리기 위한 데이터 구조가 완성되었습니다.

720

24

**[그림1] 720 x 24의 스크립트를 10561행으로 피벗**

3.3 1차원 합성곱 모델 설계

합성곱 모델은 기본적으로 이미지처리를 위해서 고안된 모델입니다. 합성곱 신경망 모델이 자연어처리를 하는 연구에서도 높은 성능을 내기 때문에 해당 모델을 주요하게 사용합니다. 자연어처리를 위해서 사용하는 커널은 사용한 커널의 개수는 크기당(3,4,5) 128개씩 사용하였습니다. 커널의 개수를 3가지로 모델을 학습시키는 이유는 다차원 적인 학습을 통해서 모델의 성능을 향상시키기 위해서 해당 방법으로 모델을 학습시켰습니다. 본 실험에서는 신경망 모델이 과적합되는 부분을 컨트롤하기위해서 세 가지 작업을 진행하였습니다. EarlyStopping()와 ModelCheckPoint(), Dropout()함수입니다. EarlyStopping() 함수는 검증 데이터 손실(val\_loss)이 증가하면, 과적합 징후므로 검증 데이터 손실이 4회 증가하면 학습을 조기 종료(Early Stopping)합니다. 또한, ModelCheckpoint()를 사용하여 검증 데이터의 정확도(val\_acc)가 이전보다 좋아질 경우에만 모델을 저장합니다. 마지막으로 Dropout()함수는 임베딩 과정에서는 0.8, 임베딩 이후에는 0.5를 할당하였습니다. 예를 들어 드롭아웃의 비율을 0.5로 한다면 학습 과정마다 랜덤으로 절반의 뉴런을 사용하지 않고, 절반의 뉴런만을 사용합니다. 모델 설계의 마지막으로는 은닉층이 1개인 실험, 2개인 실험을 구성하여 진행합니다. 프렌즈 대사 코퍼스는 검증할 테스트 데이터가 없는 관계로 훈련데이터의 10%를 테스트 데이터로 분할하여 검증하였습니다.

**4. 결론**

실험결과를 보았을 때 은닉층의 개수가 추가되었을 때 모델의 성능이 나쁘게 도출되었습니다.

합성곱 연산과정에서 이미 충분한 연산이 완료되어서 추가적인 은닉층 추가는 오히려 모델에 악영향을 주는 것으로 보여집니다.

|  |  |  |
| --- | --- | --- |
|  | **네이버** | **프렌즈** |
| **1개의 은닉층** | **83.51%** | **45.85%** |
| **2개의 은닉층** | **83.33%** | **45.51%** |

**5. 실험 고찰**

5.1 네이버 영화 리뷰 코퍼스

성능 향상을 위해서 데이터 탐색을 통해서 불용어를 추가적으로 진행할 예정입니다. 불용어를 실무에 맞게 추가하여 불용어를 제거 한다면 모델의 성능을 개선시키는 전처리가 될 것이라 생각합니다. 추가적으로 네이버 영화리뷰 코퍼스이기 때문에 영문에 대한 단어들을 모두 제거하였었는데 해당부분 영문의 경우에도 긍정적인 리뷰가 포함될 가능성이 있기 때문에 모델의 성능을 높이기 위해서 영문도 감성분석 가능하도록 진행하도록 할 예정입니다.

5.2 프렌즈 대사 코퍼스

딥러닝 모델을 사용한 감성분석에서 해당 성능이 나와 실험에 대한 전반적인 전처리와 데이터 셋 점검을 실시할 예정입니다. 일단 사용하지는 않은 칼럼이지만 등장인물을 범주화하여 독립변수를 추가하여 전처리 작업을 업데이트를 할 예정입니다. 영화 같은 경우 등장인물에 따른 주로 발화하는 단어들이 등장인물의 핵심적인 감성을 포함되어 있을 가능성이 있기 때문입니다.

5.3 모델 설계 구조

실험에 잘못된 부분으로 은닉층이 추가되었음에도

검증 성능이 떨어짐을 확인하였습니다. 일단 다른 레퍼런스 들을 통해서 CNN 모델의 다양한 설계 구조를 개선하여 해당 감성분석 모델의 성능을 향상시킬 예정입니다. 이후에는 RNN 모델의 다양한 기법을 활용하여 감성분석기의 다양성과 성능을 향상시키는 분석기를 실험하도록 하겠습니다.

**Github 주소**

: **https://github.com/destiner90/KoreaNLPProject.git**

**참고문헌**

[1] Byungun Yoon, “Opinion Mining with Artificial Intelligence-What is Social Opinion Mining?” Samsung SDS insight reports,2017,

<http://www.samsungsds.com/global/ko/support/insight/11958882284.html>

[2] Hannah Kim, Young-Seob Jeong, “Social Issue Analysis Based on Sentiment of Twitter Users”, Journal of Convergence for Information Technology, Vol. 9. No. 11, pp.81-91, 2019

[3] Convolutional Neural Networks: 모듈 별로 쪼개서 보기  
  
출처:

<https://dgkim5360.tistory.com/entry/convolutional-neural-networks-a-modular-perspective-kr>

[개발새발로그]

[4] https://wikidocs.net/21690