국어 문자열 감성 분석기

정확도 92.73%!

이건 뭐 사람 아니냐고!!

알파쿠

2020 건국대학교 컴퓨터공학과 졸업 작품

김주언, 류한길, 유병찬, 최보라

프로젝트 개요

프로젝트 목표

국어 문자열의 감성을 긍정 또는 부정으로 분류하는 딥러닝 모델을 설계하고, 기존 시스템보다 정확도를 개선하는 것: 최신 논문[1]의 NSMC[2] Task 최종 정확도인 90.51% 를 넘기는 것을 목표로 한다.

사용한 딥러닝 언어 모델: Electra

이 프로젝트에선, Electra 모델을 기반으로 사전학습과 미세조정[3]을 거쳐 성능을 향상시킨다. Electra 는 BERT 의 학습 방식을 바꾸어서, 좀 더 가벼우면서도 성능을 개선한 언어 모델이다:

- 가중치 공유를 통해 파라미터를 감소시켜서 가볍다
- Dynamic Masking 으로 학습 효과가 증가하였다
- Is Next Prediction 을 Sentence Order Prediction 으로 변경하여 학습 효과가 증가하였다
- Generator - Discriminator 로 기존처럼 사전학습 후에, 옳게 추측했는지 여부를 바탕으로 한 번 더 학습시킨다. 이를 바탕으로, 학습 효과가 증가하였다

사용된 데이터 셋

Electra 사전 학습 데이터: 뉴스기사, 나무위키, 위키피디아 문자열 데이터 (자체 구축, 총 50GB)
미세조정 학습 데이터: NSMC 리뷰 15만 개 (공개), 네이버 영화평 정제 데이터 350만 개 (자체 구축)
Feature 추가 학습 데이터 1: 군산대 한국어 감성사전 (공개)
Feature 추가 학습 데이터 2: 네이버 영화평 word-piece 별 감성 레이블링 데이터 (자체 구축)

모델 설명

개발 과정

이 프로젝트의 알고리즘은, 다음과 같은 단계로 개발되었다:

1. 공개되어 있는 Electra 를 기반으로 사전학습한, 자체 개발 Electra 모델을 개발
2. 웹크롤링을 이용하여 학습에 사용되는 자체구축 데이터 증량 (350만 건)
3. 그 모델에 군산대 감성사전을 활용하여, 각 단어의 감성을 별도로 추가 학습
4. 그 모델 위에 Self-Attention[4] 과 Bi-LSTM[5] 계층을 쌓고, CLS[6] 토큰과 Bi-LSTM 의 출력 양단의 토큰을 Affine[7] 하여 최종 예측 값을 산출 (Fine-tuning, 미세조정)

1 - Electra 모델 사전 학습 (약 40일 소요, 0.17% 성능향상)

연구실에서 자체구축한 데이터로 Electra 모델을 사전학습

2 - 학습용 데이터 자체 구축 (약 3일 소요, 1.97% 성능향상)

- 네이버에서 공개한 NSMC 데이터 15만건 외에도, 웹크롤링을 사용하여 350만건의 영화평 데이터를 구축
- 긍정과 부정 문자열이 1:1 의 비율로 구성되도록 하였고, 빈 문자열이나 의미 없는 특수문자 또는 크롤시에 변환된 ASCII 코드 등을 제거하거나 수정하여 데이터의 품질을 높임

3 - 감성 사전 적용 (약 5일 소요, 0.04% 성능향상)

감성 사전 토크나이징

군산대 감성사전은 해당 1 ~ 8 어절의 단어 쌍을, 긍부정 정도에 따라 -2, -1, 0, 1, 2 의 5 개의 점수로 구분하였다. 그런데 우리 모델은 어절이 아닌 word-piece[8] 단위로 문자열을 분석하기 때문에 그대로는 사용할 수가 없어서, 기존 감성 사전의 인덱스를 전부 Electra 토크나이저를 사용해 우리가 개발하는 모델과 호환되는 word-piece 단위로 변경하였다.

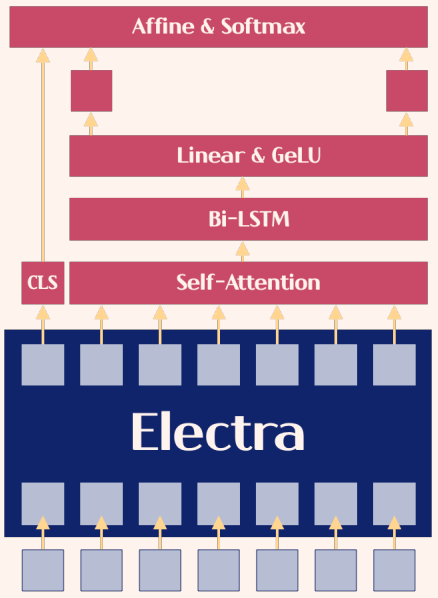
감성 사전 학습

입력 문자열을 구성하는 각 토큰 쌍에, 감성 사전을 바탕으로한 감성 값을 대응시킨다. 그 감성 값들의 벡터를 정답 시퀀스로, 각 단어의 감성 값을 추론하도록 Electra 를 추가 학습시킨다.

*이 기술은, 학습 효과가 미미하여, 최종 모델에서는 제외되었습니다.

4 - 미세조정 레이어 설계 : Bi-LSTM 신경층 (약 60일 소요, 0.38% 성능향상)

- Electra 의 출력 벡터 중, 문장 전체의 정보를 갖고 있는 CLS 토큰을 따로 분리하고 나머지는 Self-Attention Layer 와 Bi-LSTM 을 거쳐 학습시킨다.
- 문장 내에서 각 단어의 연관 관계 정보가 감성 분석에 유효할 수 있으므로, Self-Attention Layer 를 삽입하였다.
- 또한, 문장을 구성하는 각 단어의 감성 정보는 앞 단어 뿐 아니라 뒤의 단어에도 영향을 받을 것이므로 문맥 정보를 포함하는 Bi-LSTM 을 선택하였다.
- Bi-LSTM 출력의 양단 토큰은 각각 앞에서 뒤 방향으로의 문맥 정보와 뒤에서 앞 방향으로의 문맥 정보를 갖고 있다. 따라서 이 두 토큰과 CLS 토큰을 Affine 계층을 사용해 결합하면 문장 전체의 정보를 고려한 벡터를 얻을 수 있다.
- 최종적으로 얻은 이 토큰을 긍정(1)과 부정(0)의 두 클래스로 분류한다.



[그림1] 사전학습된 Electra 모델과 미세조정 레이어의 구조

직접 문장을 입력하고, 감성을 분석해볼 수 있어요!!

모델 시연용 데모 웹앱
Jolp.kimjooeon.com

성능 평가 및 분석

성능 평가

시스템 분류	Base (고성능 모델)	성능 향상 폭
koElectra (기존 시스템)	90.21%	
50GB 말뭉치 Electra 사전학습	90.38%	0.17%
+ Bi-LSTM 레이어 추가	90.76%	0.38%
+ 365만 개 영화평 감성 말뭉치	92.73%	1.97%

기존 모델	우리 모델	성능 향상 폭
(koElectra base) 90.21%	92.73%	2.52%
(최신 BERT 논문) 90.51%	92.73%	2.22%

성능 평가 결과 해석

- 각 단계 별로 실제 성능이 향상되었음을 알 수 있다.
- 해결하려는 문제(여기서는 감성분석)에 맞게 미리 레이블링된 구축 데이터가 성능에 가장 큰 영향을 준다는 것을 알 수 있다.

논문[1] : <BERT 기반 Variational Inference와, RNN을 이용한 한국어 영화평 감성 분석(박천음, 이창기 2019)>
NSMC[2] : 네이버에서 구축하여 공개한 네이버 영화평 데이터로, 학습용 15만 건, 성능 평가용 5만 건의 리뷰로 구성되어 있다.
미세조정[3] : 미세 조정(Fine-tuning)이란, 기반이 되는 특정 딥러닝 모델 위에 별도의 신경층을 설계하거나, 입력 데이터를 정제하는 등의 수단으로 모델을 변경하여 성능을 향상시키는 것을 의미한다.
Self-Attention[4] : 각 단어가 서로에 대한 어떤 관계성을 갖고 있는지를 저장하는 신경층.

Bi-LSTM[5] :게이트 기법을 통해 순환 신경 회로망(RNN)의 한계를 극복한 모델인 LSTM을 순방향뿐 아니라 역방향의 결과를 함께 이용하는 모델이다. 문맥(Context)을 기반으로 하는 연관성 분석에 유리하여 속도를 고려한 다양한 NLP 문제에 널리 활용되고 있다.
CLS[6] : Electra 의 최종 출력층의 첫 번째 토큰으로, classification 의 약자이다.
Affine[7] : Affine 변환을 수행하는 신경층으로, Linear 계층이라고 보면 쉽다.
Word-piece[8] : 단어를 구성하는 일종의 하위어(sub-word)로, 자연어 처리 분야에서 어절도 음절도, 형태소도 아닌 단위로 문장을 분석하기 위해 사용한다. 사용하는 모델에 따라 별도의 Vocab.txt 파일에 저장되어 있다.