불용어의 BERT 기반 문장 자동분류기에 대한 영향

박도영 이문현 김영훈
한양대학교, 바이오인공지능융합전공 인공지능융합학과
enter1994@naver.com greenzip2004@naver.com nongaussian@hanyang.ac.kr

The influence of stop words on the sentence classifier based on BERT

Doyeong Park○ Moonhyun Lee Younghoon Kim
Hanyang Univ. Maj. in Bio Artificial Intelligence, Dept. of Applied Artificial Intelligence

요 약

머신러닝 알고리즘은 비약적으로 성장하며 좋은 성능을 보이고 있지만, 알고리즘을 통해 산출된 결과물에 대한 해석이 불분명하다. 자연어 처리에서 뛰어난 성능을 보이는 BERT는 대규모 데이터로 사전 학습된 후, 사용자의 의도에 맞는 미세조정을 거쳐 성능을 개선하여 이용된다. 하지만 BERT 또한 산출된 결과물이 어떻게 계산되는지 명확하게 알려져 있지 않다. 본 논문에서는 BERT가 출력하는 결과에 대해 해석하기 위해 마스킹을 통해 주로 어떠한 단어가 문장의 임베딩 벡터에 큰 영향력을 끼치는지 알아본다. 그리고 BERT 모델이 문장에서 중요하게 여기는 단어와 사람이 생각했을 때 중요하게 생각되는 단어가 다르다는 것을 증명한다.

1. 서 론

머신러닝은 비약적인 성장을 보이며 이미지 분류, 의사결정 및 자연어 처리 등 다양한 분야에서 좋은 성능을 보여주고 있다. 하지만, 머신러닝은 계층이 깊어질수록 블랙박스화 되고 산출된 결과물에 대해서 어떻게 결과를 산출하는지에 대해 명확히 알려져 있지 않다. 그래서, 최근 머신러닝 알고리즘이 결과를 어떻게 도출하는지 설명하려는 'XAI(eXplainableAI)'의 필요성이 부각되고 있다.

트랜스포머[1]는 최근 자연어 처리에서 뛰어난 성능을 보이면서, 트랜스포머의 구조를 변형한 다양한 모델이 연구되었다. 특히, BERT(Bidirectional Encoder Representations from Transformers)[2]는 트랜스포머의 인코더 부분만 취하고 강화한 모델로, 자연어 처리 분야에서 의미 있는 정보를 추출하는 용도로 많이 이용되고 있다. 하지만 트랜스포머나 BERT 역시 산출된 결과물이 어떻게 계산되는지 명확하게 알려져 있지 않다.

BERT는 일반적으로 '위키피디아' 데이터와 'BooksCorpus[3]' 데이터로 사전 학습 된 후, 사용자가 원하는 형태의 데이터로 미세조정해서 이용하게 되는데, 사전 학습에서는 BERT가 MLM(Masked Language Model)문제와 NLP(Next Sentence Prediction)문제를 해결하도록 학습된다.

BERT에는 사전 학습 시에 사용된 특별 토큰인 [CLS], [SEP], [MASK], [UNK] 토큰이 있다. 문장을 BERT에 입력으로 넣기 전에 [CLS] 토큰은 문장의 전반적인 문맥을 학습하여 분류 모델을 미세 조정할 때 사용되는 분류 토큰이고, [SEP]는 문장을 구분하는 역할을 하는 토큰이다. [MASK] 토큰은 사전 학습 시 입력되는 문장의 15%의 단어를 마스킹하여 BERT 모델로 하여금 가려진 단어를 예측하게 만드는데 사용된다. [UNK] 토큰은 BERT가 사전 학습을 통해 저장한 단어 토큰에서 찾을 수 없는 단어가 나왔을 때 사용되는 토큰이다.

사전 학습된 BERT에 추가적인 계층을 구성하고, 라벨이 부여된 데이터를 추가적으로 훈련하는 과정을 통해 특정 과제에 최적화된 모델을 만드는데, 이를 미세조정(fine-tuning)이라고 한다. 미세 조정을 통해 번역기, 문장 분류 또는 문장 주제 찾기, 질의응답 등 다양한 주제에 접목되고 있다.

본 논문에서는 BERT 모델이 문장에서 중요하게 여기는 단어가 사람이 생각했을 때 중요하게 생각되는 단어가 다르다는 것을 증명하고 원인을 분석한다. 본 연구팀은 사전 학습된 BERT가 어떤 단어를 중요하게 보는지 확인하기 위해 입력 문장의 각 단어를 하나씩 마스킹하여 BERT의 [CLS]값의 변화를 비교해보았다. 실험에서 "온점(.)", "at", "the"와 같이 문장 의미에 별로 관련이 없는 단어를 마스킹했을 때, 원본 문장의 [CLS]값의 변화가 가장 크다는 사실을 관찰했다. 이러한 단어는 일반적인 문장에 거의 항상 포함되는 단어임을 확인할 수 있었다.

BERT가 실제로 데이터셋에 가장 많이 나타나는 단어들을 중요하게 여기는지 확인하기 위해, 감정분류모델로 미세 조정된 BERT에서 어떤 단어를 마스킹했을 때 정답을 맞출 확률이 가장 떨어지는지

확인해보았다. 그 결과, 감정분류 BERT 모델 역시 학습데이터에서 빈도수가 가장 높은 단어가 영향력이 크다는 사실을 알 수 있었다.

위 두 실험을 통해 어떤 단어가 감정 분류 결과에 영향을 미치는지 확인하고, BERT가 중요하게 생각하는 단어가 무엇인지 확인했다.

2. 관련 연구

[4][5]는 문장에서 분류에 큰 영향을 미치는 단어를 찾아 사람이 보았을 때, 원래 문장의 의미를 유지하지만, 모델은 틀린 분류를 하게끔 단어를 대치시킨다. 이러한 연구들은 분류에 큰 영향을 주는 중요한 단어를 찾는 방법을 제안했다.

본 논문이 BERT를 해석하는 방법과 유사한 연구로는 [6]이 있다. [6]에서는 입력값을 지워가면서 입력의 어떤 부분이 분류결과에 큰 영향을 주는지 찾아내는 기법을 연구한 논문이다. 본 논문에서는 [6]에서 연구된 섀플리 가치를 응용하여 입력 문장 속 단어를 지워가면서 BERT가 문장의 어느 단어를 중요하게 여기는지 파악하고자 했다.

3. 마스킹으로 인한 [CLS] 토큰의 변화

입력 문장에서 사전 학습된 BERT의 분류 성능에 영향을 많이 주는 단어가 무엇인지 확인하기 위해, 원본 문장을 넣었을 때 [CLS] 출력값과 그림1과 같이 마스킹 된 문장을 넣었을 때 [CLS] 출력값을 비교했다. [CLS] 출력값은 분류기에 직접적으로 사용되는 벡터로, 분류 성능에 영향이 큰 단어를 제거한 문장으로 계산한 [CLS]는 기존 문장으로 계산한 [CLS]와 차이가 클 것이라 예상했기 때문이다.

그림 1과 같이 입력 문장의 단어 하나를 마스킹한 후, BERT에 넣어 출력되는 [CLS] 토큰을 기존 문장을 BERT에 넣었을 때 출력된 [CLS] 토큰과 코사인 유사도를 통해 비교해보았다. 이때, 문장 속 단어를 마스킹하는 방법은 단어 하나를 [MASK] 토큰, [UNK] 토큰, 또는 0벡터로 대치시키는 방법을 사용했다.

이를 226개의 문장에 있는 모든 단어에 그림1과 같이 마스킹 후에 각각 원본 문장과 마스킹 된 문장의 [CLS] 토큰을 비교한 결과, 사람이 보았을 때 문장의 의미를 결정하는 단어와 BERT모델의 분류 확률값에 영향을 주는 단어가 다르다는 사실을 확인할 수 있었다. 그림2에서 볼 수 있듯이 일반적으로 문장의 의미를 파악하는데 중요하지 않다고 생각되는 관사나 전치사, 혹은 'is', 'would' 와 같은 동사를 마스킹했을 때 코사인 유사도가 상대적으로 더 낮아지는 것을 볼 수 있었다. 이러한 단어들은 모든 문장에서 자주 사용되는 단어들이라는 공통점이 있는데, 데이터셋을 학습할 때 빈도수가 높은 단어들이 [CLS]토큰의 임베딩 벡터에 많은 영향을 줄 수 있다는 가설을 수립하였다.

i	feel	a	little	mellow	today
[MASK]	feel	a	little	mellow	today
i	[MASK]	a	little	mellow	today
i	feel	[MASK]	little	mellow	today
i	feel	a	[MASK]	mellow	today
i	feel	a	little	[MASK]	today
i	feel	a	little	mellow	[MASK]

그림 1. [MASK] 토큰으로 문장의 모든 단어를 한 번씩 마스킹한 예시

4. 미세 조정된 BERT를 이용한 단어의 영향력 측정

그렇다면 BERT의 분류 결과에 영향을 많이 주는 단어들이 정말 학습 데이터에서 많이 등장하는 단어들인지 확인할 필요가 있다. 본 연구에서는 사전 학습된 BERT 보다는 감정 분류기로 미세 조정된 BERT를 이용하여 BERT가 정말 학습 데이터에 많이 등장하는 단어에 영향을 많이 받는지 확인하고자 했다. 감정분류 BERT의 입력으로 원본 문장을 넣었을 때의 감정 분류 결과 확률값과 문장에서 각각의 단어를 마스킹한 문장을 넣었을 때의 감정 분류 확률값을 비교하고 문장 내에서 어떤 단어를 마스킹했을 때 감정분류 확률이 가장 많이 변하는지를 확인했다.

감정 분류 모델은 "huggingface"[8]에서 미세 조정된 BERT 모델을 가져와 이용했으며, 실험 데이터로는 Emotion 데이터셋[7]을 사용했다. Emotion 데이터셋은 학습 데이터 16000문장, 검증 데이터 2000문장, 실험 데이터 2000문장, 총 20000개의 영어 문장으로 이루어져 있으며 6가지의 감정(sadness, joy, love, anger, fear, surprise)을 분류할 수 있도록 라벨링이 되어있다.

그림 1과 마찬가지로 실험 데이터 2000개의 각 문장에서 단어 하나씩을 [MASK], [UNK], 임베딩 벡터에 0을 곱해주는 방법을 통해 마스킹하였다. 그리고 미세 조정된 감정 분류 BERT에 넣었을 때, 마스킹한 문장을 계산한 감정 분류 확률값과 정상 문장을 감정 분류 BERT로 계산한 감정 분류 확률값을 비교한 후, 확률값 차이를 가장 크게 만드는 단어를 카운트했다.

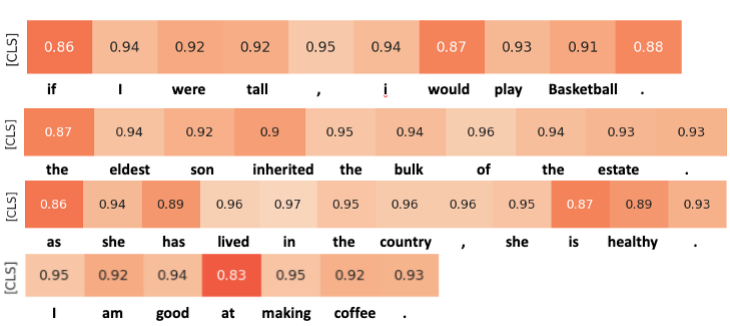

그림 2. 각 단어를 [UNK] 토큰으로 마스킹 후 BERT모델을 통과시켜 출력된 [CLS]토큰과 기존 문장을 BERT모델에 통과시켜 출력된 [CLS] 토큰의 코사인 유사도를 계산한 예시

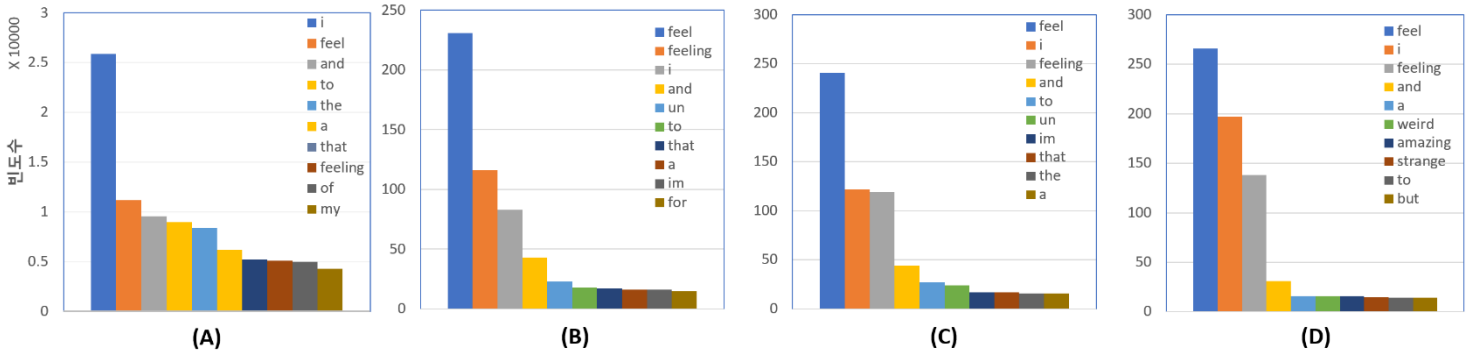

그림 3. (A)는 Emotion 데이터셋에서 학습 데이터 16000문장의 단어 토큰 빈도수. (B), (C), (D)는 Emotion 데이터셋에서 실험 데이터 2000문장을 [MASK] 토큰, [UNK] 토큰, 단어 토큰의 입력 임베딩 벡터에 0을 곱하는 방법으로 각각 마스킹 후, 기존 문장의 라벨과 다른 라벨을 출력하게 만드는 단어의 빈도수

그림 3에서 (A)는 학습 데이터 16000문장의 단어 토큰 빈도수 중 상위 10개를 출력한 그래프를 나타내고, (B), (C), (D)는 각각 [MASK] 토큰, [UNK] 토큰, 입력 단어의 임베딩 벡터를 0으로 곱해줘서 마스킹했을 때 기존 문장의 라벨과 다른 예측 라벨을 출력하는 단어들을 카운트하여 상위 10개를 출력한 그래프이다.

(B), (C), (D) 그래프에서 볼 수 있듯이 3가지 마스킹 방법 모두 특정 감정을 나타내는 단어 토큰과는 관련이 없는 단어를 마스킹했을 때 기존 라벨과 다른 예측을 하는 경우가 많았다. (A) 그래프와 나머지 그래프를 비교해 보면 학습된 데이터셋의 단어 빈도수가 높은 단어 토큰들을 마스킹했을 때 다른 예측 라벨을 출력할 확률이 높다는 것을 알 수 있다.

5. 결론

BERT는 많은 문장을 통해 사전 학습된 모델에 사용자의 의도에 맞게 미세조정을 거쳐 성능을 개선하여 사용하는 경우가 많다. 하지만 BERT가 중요하게 여기는 단어와 사용자가 중요하게 여기는 단어 간 의미상 차이가 있음을 실험을 통해 관찰하였다. 관찰한 결과, BERT가 중요하게 여기는 단어에는 학습 데이터에서 빈도수가 많은 단어가 선택될 확률이 높다는 것을 알 수 있다.

따라서, 빈도 수가 높은 단어가 오타나 오류에 의해 변형되어 모델에 입력되었을 경우 BERT 모델의 출력에 큰 영향을 줄 수 있을 것이라고 예측할 수 있다.

이러한 위험을 예방하기 위해 학습 데이터의 단어의 빈도를 고려한 미세 조정을 진행하여, 강건한 BERT모델을 학습시키는 연구를 계획 중이다.

사사의 글

이 논문은 2021년도 정부(산업통상자원부) 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0008691, 2021년 산업혁신인재성장지원사업)

참고 문헌

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[3] Zhu, Yukun, et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." Proceedings of the IEEE international conference on computer vision. 2015.

[4] Jin, Di, et al. "Is bert really robust? a strong baseline for natural language attack on text classification and entailment." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 05. 2020.

[5] Garg, Siddhant, and Goutham Ramakrishnan. "Bae: Bert-based adversarial examples for text classification." arXiv preprint arXiv:2004.01970 (2020).

[6] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Proceedings of the 31st international conference on neural information processing systems. 2017.

[7] Saravia, Elvis, et al. "Carer: Contextualized affect representations for emotion recognition." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[8] Wolf, Thomas, et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019).