불용어 빈도수에 강건한 BERT 기반 문장 자동 분류기 학습

박도영 ○ 김영훈

한양대학교, 컴퓨터공학과 바이오인공지능융합전공

enter1994@naver.com, nongaussian@hanyang.ac.kr

Robust to Stop words Frequency on the Sentence Classifier based on BERT

Doyeong Park○ Younghoon Kim

Hanyang Univ.Dept. of Computer Science&Engineering. Maj. in Bio Artificial Intelligence

요 약

머신러닝 알고리즘은 비약적으로 성장하며 좋은 성능을 보이고 있지만, 알고리즘을 통해 산출된 결과물에 대한 해석이 불분명하다. 자연어 처리에서 뛰어난 성능을 보이는 BERT 는 대규모 데이터로 사전 학습된 후, 사용자의 의도에 맞는 미세조정을 거쳐 성능을 개선하여 이용된다. 하지만 BERT 는 학습시에 빈도수가 높은 단어(온점, “the”)와 같은 단어가 변형되었을 때 결과값에 큰 영향을 받는다. 이러한 문제점을 해결하기 위해 본 논문에서는 BERT 의 취약점을 개선하기 위해 데이터 증강을 통한 모델 학습 방안을 제시한다.

1. 서 론

머신러닝은 비약적인 성장을 보이며 이미지 분류, 의사결정 및 자연어 처리 등 다양한 분야에서 좋은 성능을 보여주고 있다. 하지만, 머신러닝은 계층이 깊어질수록 산출된 결과물에 대한 해석이 어려워지고, 블랙박스화 된다. 그래서 최근 머신러닝 알고리즘이 결과를 어떻게 도출하는지 설명하려는 ‘XAI(eXplainableAI)’의 필요성이 부각되고 있다.

트랜스포머(Transformer)[1]는 최근 자연어 처리에서 뛰어난 성능을 보이면서, 트랜스포머의 구조를 변형한 다양한 모델이 연구되었다. 특히, BERT(Bidirctional 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 ‘Wikipidia’ 데이터와 ‘BooksCorpus[3]’ 데이터로 사전 학습된 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 학습된 BERT는 번역기, 문장 분류 또는 문장 주제 찾기, 질의응답 등 다양한 주제에 접목되고 있다.

하지만 특정 데이터에서 미세 조정된 BERT는 학습 데이터에서 빈번하게 등장하는 단어에 의해 출력값이 크게 변하는 현상이 존재한다[4]. [4]에서는 감정 분류 데이터로 미세 조정된 BERT에 대해 온점(.), “at”, “the”와 같이 학습 데이터에서 빈도수가 높은 단어를 마스킹했을 때, 감정분류 결과값의 변화가 가장 크다는 사실을 보여준다. 이는 학습 데이터에서 크게 의미가 없지만, 빈번하게 등장하는 단어가 BERT의 미세조정에 크게 영향을 미친다고 볼 수 있다.

본 논문에서는 데이터 증강(Data Augmentation)을 통해 감정분류 모델로 미세 조정된 BERT가 빈도수가 높은 단어를 마스킹 하더라도 정답을 맞출 확률이

표 1. 학습 데이터 증강 이후, 실험 데이터에 빈도수 높은 데이터를 마스킹 했을 때 정확도 비교

	학습 데이터		
실험 데이터		Original	Augmentation
	Original	92%	92.5%
	4 token Masking	88%	92.8%
	5 token Masking	86.2%	92.2%

떨어지지 않도록 한다. 위 실험을 통해 BERT를 특정 분류기로 미세 조정 시 강건하게 학습시키는 방향을 제시한다. 그 결과, 빈도수가 높은 단어를 마스킹하더라도 정확도가 떨어지지 않는 모델을 학습하였다.

2. 관련연구

[5, 6]은 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 논문에서 BERT를 해석하는 방법과 유사한 연구로는 [7]이 있다. [7]에서는 입력값을 지워가면서 입력의 어떤 부분이 분류결과에 큰 영향을 주는지 찾아내는 기법을 연구한 논문이다. 본 논문에서는 [7]에서 연구된 섀플리 가치를 응용하여 입력 문장 속 단어에서 BERT가 중요하게 여기는지 단어를 마스킹 후, 마스킹된 문장을 기존 데이터셋과 합하여 실험을 진행한다.

3. 마스킹으로 인한 모델 예측값의 변화

[4]는 입력 문장에서 사전 학습된 BERT의 분류 성능에 영향을 많이 주는 단어가 무엇인지 확인하기 위해 두가지 실험을 진행한다.

첫째, 원본 문장을 넣었을 때 [CLS] 출력값과 임의의 단어 하나를 마스킹한 문장을 넣었을 때 [CLS] 출력값을 코사인 유사도를 통해 비교한다. 이때, 문장 속 단어를 마스킹하는 방법은 단어 하나를 [MASK] 토큰, [UNK] 토큰, 또는 0벡터로 대치시키는 방법을 사용한다. 그 결과, 문장의 의미를 결정하는 단어보다는 사전 학습 시에 빈도수가 높은 단어를 마스킹했을 때, [CLS]토큰의 임베딩 벡터에 많은 영향을 줄 수 있다는 사실을 보인다.

둘째, [4]는 사전 학습된 BERT 보다는 감정 분류기로 미세 조정된 BERT를 이용하여 BERT가 정말 학습 데이터에 많이 등장하는 단어에 영향을 크게 받는지 확인하고자 한다. 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 확인한다. 첫번째 실험과 마찬가지로 3가지 마스킹 방법을 사용하여 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 카운트한다. 그 결과, 3가지 마스킹 방법 모두 특정 감정을 나타내는 단어 토큰보다는 데이터셋에서 빈도수가 높은 단어를 마스킹했을 때 잘못된 예측을 하는 경우가 많았다.

4. 데이터 증강을 통한 강건한 모델 학습

본 논문에서는 빈도수가 높은 단어가 오타나 오류에 의해 변형되어 모델에 입력될 경우 출력에 큰 영향을 주는 것을 보완하기 위해 높은 빈도수의 단어를 마스킹하여 학습 데이터에 추가한다. 높은 빈도수의 단어가 마스킹되어 실험 데이터에 들어왔을 때 모델의 성능을 확인하여 학습 데이터 추가 전후를 비교한다. 즉, 마스킹에 취약한 토큰이 마스킹 되어 입력되더라도, 정확도를 유지할 수 있도록 실험하였다.

감정 분류 모델은 “huggingface”[8]에서 미세 조정된 BERT 모델을 가져와 이용했으며, 실험 데이터로는 Emotion 데이터셋[9]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.

본 논문에서는 높은 빈도수의 단어를 마스킹한 데이터를 기존 데이터셋과 합치는 데이터 증강기법을 이용하였다. 기존 학습 데이터 16000 문장과 임베딩 벡터 변화를 크게 만든 토큰인 ‘feel’, ‘feeling’, ‘and’를 마스킹한 문자열 데이터 15645개의 문장을 합하여 총 31645 문장을 이용하여 학습 데이터로 이용했다. 실험 데이터는 임베딩 벡터 변화를 크게 만든 4개의 토큰인 ‘feel’, ‘feeling’, ‘and’, ‘to’를 마스킹했을 때와 추가로 ‘i’를 마스킹한 경우, 즉 5가지 토큰을 마스킹했을 때와 기존 실험 데이터만을 이용하였을 때를 비교하였다.

표 1에서 볼 수 있듯이 기존 학습 데이터에 마스킹 된 문장을 증강하여 학습 데이터에 추가했을 때 정확도가 더 높은 것을 본 연구팀은 표 1에서 볼 수 있듯이 데이터 증강이 적용되지 않은 데이터(original)로 학습한 BERT와 데이터 증강이 적용된 데이터(augmentation)로 학습한 BERT, 총 2가지 BERT 모델을 학습한 후

성능을 비교했다. 실험 데이터는 마스킹을 하지 않은 데이터(original)과 4개의 토큰(“feel”, “feeling”, “and”, “to”)을 마스킹한 ‘4 token Masking’ 데이터와 5개의 토큰(“i”, “feel”, “feeling”, “and”, “to”)을 마스킹한 ‘5 token Masking’ 데이터를 사용했다.

결과적으로, 데이터 증강이 적용된 모델(augmentation)은 빈도수 높은 단어를 지운 학습 데이터에 대해서 데이터 증강이 적용되지 않은 데이터(original)로 학습한 모델에 비해 성능 하락이 없이 높은 정확도를 유지하는 것을 확인했다. 즉, 우리는 빈도수가 높은 단어가 미세 조정된 BERT에 미치는 영향을 최소화한 모델을 만들었다.

5. 결론

[4]는 빈도수가 높은 단어가 BERT의 출력에 큰 영향을 줄 수 있음을 보였으며, 이를 바탕으로 본 논문은 빈도수가 높은 단어를 마스킹한 데이터로 BERT를 미세 조정하여 빈도수가 높은 단어에 강건한 BERT를 만들었다. 이를 통해 미세조정할 때 학습데이터의 단어의 빈도를 고려하는 학습 방향을 제시한다.

사사의 글

이 논문은 2022년도 정부(산업통상자원부) 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0008691, 2021년 산업혁신인재성장지원사업)

또한 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1G1A1011471).

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] 박도영, 이문현, and 김영훈. "불용어의 BERT 기반 문장 자동분류기에 대한 영향." 한국정보과학회 학술발표논문집 (2021): 715-717.

[5] 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.

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

[7] 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.

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

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