낯선 데이터를 활용한 과잉신뢰 완화 텍스트 증강 기법

Text Augmentation to Mitigate Overconfidence through Unfamiliar Data

이준호
Junho Lee
중앙대학교
AI 학과
Department of Artificial Intelligence, Chung-Ang University
jhjo32@cau.ac.kr

송상민
Sangmin Song
중앙대학교
AI 학과
Department of Artificial Intelligence, Chung-Ang University
s2022120859@cau.ac.kr

최주환
Juhwan Choi
중앙대학교
전자전기공학부
School of Electrical and Electronics Engineering, Chung-Ang University
gold5230@cau.ac.kr

박주형
Juhyoung Park
중앙대학교
컴퓨터공학부
School of Computer Science and Engineering, Chung-Ang University
hynciath51@gmail.com

진교훈
Kyohoon Jin
중앙대학교
첨단영상대학원
Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University
fhzh123@cau.ac.kr

김영빈
Youngbin Kim
중앙대학교
첨단영상대학원
Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University
ybkim85@cau.ac.kr

요약문

최근 자연어 처리 모델은 대용량 데이터를 기반으로 사전학습(pretrain) 후 미세조정(fine tuning)을 하는 방식을 통해서 좋은 성과를 보이고 있다. 미세조정 과정에서 사용되는 데이터 수가 부족할 때 학습 데이터에 지나치게 의존하는 과적합 문제를 데이터 증강을 통해서 완화 할 수 있다. 하지만 기존의 데이터 분포와 크게 벗어나지 않는 증강은 모델이 높은 확신을 가지고 잘못된 예측을 하는 문제를 발생시킬 수 있다. 본 논문에서는 과잉확신 문제를 해결하기 위해 기존 데이터세트와 유사도가 적은 다른 데이터세트을 추가로 활용하는 데이터 증강 기법을 제안한다. Sentence BERT(SBERT)를 활용하여 한글 딥러닝 데이터 세트간의 유사도를 측정하는 방식을 통하여 낯선 데이터 세트를 규정한다. 제안된 기법은 다른 데이터 증강 기법들에 비해 과잉확신 완화에 효과가 있음을 확인할 수 있었다. 해당 기법을 통해 Korean hate speech 분류 작업에서 기준치 대비 3.98%p 향상되었으며 기존 기법과 비교해 2.39%p 의 성능 향상을 확인할 수 있었다.

주제어

딥러닝, 자연어처리, 데이터 증강

1. 서 론

최근 딥러닝 기반 자연어처리 모델은 대규모 데이터를 활용해 언어 표현을 학습시킨 사전학습 언어 모델을 구축한 후 이를 원하는 작업에 적용시키기 위해 미세조정을 하는 방식이 주를 이루었다. 그러나 미세조정 데이터 세트의 크기가 충분히 크지 못할 때 주어진 데이터에만 지나치게 의존하는 과적합 문제가 발생할 수 있다[1]. 이러한 문제는 미세 조정 데이터 세트에 데이터 증강 기법을 적용하여 완화할 수 있다.

데이터 증강 기법은 크게 규칙 기반 기법과 모델 기반 기법으로 구분된다. 모델 기반 기법의 경우 역번역(back-translation)[2]기법과 같이 추가적인 딥러닝 모델을 활용하여 성능을 안정적으로 향상시킬 수 있지만, 증강 과정에서 많은 비용이 소요될 수 있다. 반면, EDA[3], AEDA[4]와 같은 규칙 기반 기법의 경우 적용이 단순한 장점이 있지만, 다른 기법에 비해 기존의 데이터 분포에서 크게 벗어나지 않는 데이터가 생성된다[5].

기존의 데이터 분포와 차이가 적은 데이터만 학습할 경우 친숙하지 않은 평가 데이터에 대해서 과잉확신을 가지고 잘못 분류하는 문제가 발생할 수 있다 [6]. 이러한 과잉확신의 예시는 표 1 을 통해서 확인할 수 있다.

표 1 은 한국어 BERT[7]모델을 활용해 댓글을 긍/부정 이진 분류 작업에 대해 미세조정 시켰을 때 각 예시별

긍/부정 점수를 나타낸다. 예시 문장을 분석하였을 때, 긍정적인 문장에 대해서 높은 예측 확률을 가지고 부정으로 예측하는 예시가 있음을 확인할 수 있다. 이는 Spurious 패턴[8] 연구에서와 같이 비슷한 문장에 대해서 학습을 하게 된다면 딥러닝 모델이 반복적으로 학습하게 되어 과잉확신이 발생하고 전체적인 모델 성능 하락으로 이어질 수 있다.

표 1. 문장 별 긍/부정 점수 비교

문장	긍정	부정
지금까지 본 영화중 마음이 가장 따뜻해지는 영화	97.17	2.83
평점 1 점도 주기 싫어지는 영화	0.06	99.94
10 점 만점에 9 점을 주고싶은 딱 좋은 영화	3.08	96.92

본 논문에서는 비용 효율화를 위해 비교적 간단한 방식으로 데이터 증강을 진행하면서, 미세조정 학습 방식에서 과잉확신 문제를 완화시키기 위해 학습 도메인에서 벗어난 낯선 데이터를 사용하는 방법을 제안한다. 낯선 데이터는 주어진 작업과 별개의 데이터를 활용하여 레이블에 중립성이 있다고 가정한다. 예를 들어 악성 댓글 분류 작업에서 댓글이 아닌 기사, 소설과 같은 내용은 악성 여부 판단이 모호하기 때문에 이를 중립적인 레이블로 판단한다. SBERT[9] 를 이용하여 데이터 세트의 유사도를 측정하여 낯선데이터를 정의하고, 유사도가 낮을수록 기존의 데이터 세트 와 별개의 정보를 가진다는 가설을 가지고, 유사도가 낮은 데이터의 라벨에 대해 균일 분포로 가정하여 학습 데이터 세트로 추가하여 학습한다. 해당 기법을 통해 미세조정 학습 방식에서 발생하는 과잉확신 문제를 완화하고, 미세조정 학습의 성능을 개선하였다.

2. 낯선 데이터 세트 증강 제안

본 논문에서는 과잉확신 방지를 위한 데이터 증강 기법을 제안한다. 증강을 위한 데이터 세트를 구분하기 위하여 데이터 세트간 유사도를 측정해 유사도가 높은 데이터 세트를 익숙한 데이터 세트로 구분하고, 유사도가 낮은 데이터 세트를 낯선데이터 세트로 구분한다.

2.1 낯선 데이터 구분

본 연구에서는 바깥 도메인 데이터 세트를 정의하기 위해서 SBERT 를 이용하였다. SBERT 는 BERT 의 문장 임베딩 성능을 자연어 추론(Natural Language Inferencing)에 대하여 학습시켜 문장 임베딩 성능을 개선한 모델이다. 각각의 데이터 세트간 유사도를 계산하기 위해서 SBERT 를 이용하여 라벨 정보와 상관없이 데이터 세트(D)를 구성하는 문장들($D = \{x_1, x_2, \dots, x_n\}$) 의 임베딩 값을 이용해 데이터 세트의 임베딩 평균값을 구하였다. 앞서 계산된 데이터 세트의 평균 임베딩 값을 내적하여 유사도를 측정하였다. N 은 샘플링된 데이터 개수이고, D^A_{emb}와 D^B_{emb}는 서로 다른 데이터 세트에 대한 임베딩 평균일 때, 데이터 세트간 유사도는 다음 수식으로 표현할 수 있다.

$$D_{embedding} = \frac{\sum SBERT(x_i)}{|N|}^{D} = Similarity(D^A_{emb}, D^B_{emb}) = D^A_{emb} \cdot D^B_{emb} (A \neq B)$$

본 논문에서는 한국어 데이터로 사전 학습된 Ko-Sentence-BERT [10] 를 사용하였고, 측정된 유사도가 낮은 데이터 세트를 out domain 데이터 세트로 구분지었다.

2.2 낯선 데이터 활용

일반적으로 분류문제에 사용하는 손실함수는 교차 엔트로피이고 다음과 같은 수식으로 표현할 수 있다.

$$L_{CE} = -\frac{1}{N}\sum_{j}^{N}\sum_{i}^{C} y_{ij} \log(\widehat{y_{ij}})$$

N 은 데이터의 개수이고, C 는 데이터 세트의 분류 종류의 개수이다. y 는 라벨에 대한 정답이며, $\hat{y}$ 은 모델의 예측값으로 0 과 1 사이의 값을 가진다.

과잉확신을 완화시키기 위해 학습 데이터와 의미적 유사성이 낮은 낯선데이터 세트를 학습에 포함시킨다. 낯선 데이터의 의미 분포는 학습 데이터와 유사하지 않기 때문에 학습에 활용하기 위해 모든 라벨을 균일 분포로 가정한다. 예를 들어, 7 개의 라벨이 있는 데이터 세트의 경우 낯선 데이터의 라벨 확률은 각 1/7 로 균일하게 부여하였다. 이를 통해 딥러닝 모델은 학습 데이터의 정보와 낯선정보를 적절하게 활용하여 데이터 분포를 더 잘 추정할 수 있으며, 더 나아가 과잉확신을 줄일 수 있다.

3. 실험

3.1 실험 환경

제안하는 기법의 성능을 측정하기 위해서 KCBERT-Base 를 사용하며, 미세 조정을 진행할때 배치 크기는 240 으로, 학습률은 5e-4 로 10 에폭동안 학습을 진행하였으며 최적화를 위해 AdamW 를 사용하였다. 테스트 세트에서 성능을 파악하기 위하여 정확도를 기준으로 학습된 모델의 성능을 평가하였으며, 매 실험마다 발생하는 무작위성을 고려하기 위해 5 회 실험을 진행한 평균값을 사용하였다.

본 논문에서 제안하는 기법을 평가하기 위해서 데이터증강에서 규칙 기반 기법 중 일반적으로 사용되는 EDA, AEDA 와 딥러닝 모델을 활용하는 기법 중에서 일반적으로 사용되는 역번역 기법, 그리고 딥러닝 모델의 과잉확신을 개선하기 위한 라벨 스무딩 방식과의 비교를 통해서 성능을 검증한다. 해당 실험에서 라벨 스무딩의 epsilon 은 0.1 로 설정하였다.

실험을 위해 사용한 데이터 세트로는 Naver Sentiment Movie Corpus (NSMC)[11], Korean hate speech(Hate)[12], KLUE Topic classification(KLUE-TC)[12], 구어체 번역 말뭉치(AI-HUB)[13], Korpora[14]를 사용하였다. NSMC 는 영화 리뷰에 관련한 데이터 세트로 리뷰를 긍정과 부정으로 분류하였다. Hate 는 인터넷 댓글에 관련한 데이터 세트로 댓글을 편견이 있는 댓글, 모욕이나 혐오가 있는 댓글, 그 외 댓글 총 3 가지로 분류하였다. KLUE-TC 는 뉴스 헤드라인 데이터 세트로 뉴스 주제에 따라 7 가지로 분류하였다. 번역을 학습하기 위한 AI-Hub 의 구어체 데이터와 한영 번역 말뭉치인 Korpora 데이터를 사용하였다.

3.2 데이터 세트 유사도

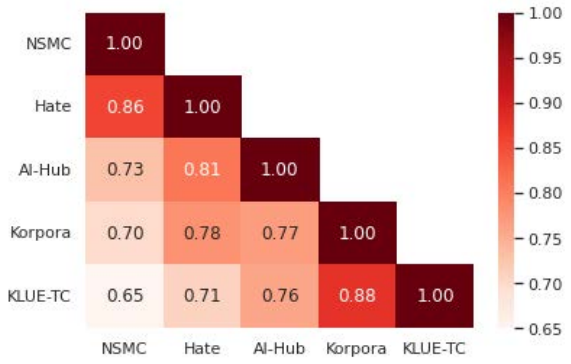

그림 1 데이터 세트간 유사도

낮선 데이터써의 활용을 위해 각 데이터 세트의 임의의 500 개 표본에 대해서 SBERT 를 통해 각 데이터 세트간의 유사도를 측정하였다. 이에 대한 결과는 그림 1 에서 확인할 수 있다. NSMC 와 Hate 데이터 세트는 둘 모두 댓글을 활용하는 데이터이며 부정적인 댓글에 대한 정보를 포함하고 있기 때문에 유사도가 높게 측정되었다. 뉴스 주제 분류를 위한 KLUE-TC 는 감정 분류를 위한 NSMC 와 다르게 감정 정보를 주로 다루지 않기 때문에 서로 낮은 유사도를 보였다.

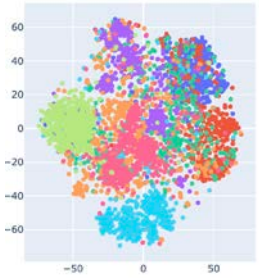

그림 2 Hate(연두)와 KLUE-TC 의 t-SNE 시각화

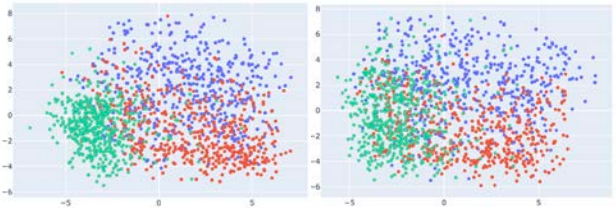

그림 3 왼쪽 NSMC(긍정 : 파랑, 부정 : 빨강) 와 Kopora(연두), 오른쪽 NSMC 와 Hate(연두) 의 t-SNE 시각화

이를 시각적으로도 살펴보기 위해 본 논문에서는 NSMC, Hate, KLUE-TC 와 같은 데이터들을 t-SNE 를 통해 시각화를 진행했다. 그림 2 의 연두색은 Hate 로 KLUE-TC 의 다른 클래스들과는 독립적으로 군집되어있는 모습을 보여주고 있다. 또한 그림 3 왼쪽에서도 마찬가지로 낮선 데이터 세트는 기존 데이터 세트와 독립적인 분포를 확인할 수 있다. 반면, 그림 3 오른쪽에서는 유사도가 높은 데이터 세트인 NSMC 와 Hate 간의 분포가 비슷함을 관찰할 수 있다. 이를 통해 유사도 차이가 나는 데이터 세트들을 낮선 데이터 세트로 활용할 수 있음을 확인하였다. 본 논문에서는 낮선 정보 반영을 위해 NSMC, Hate 와 같은 감정 정보가 포함된 데이터에서는 Korpora, KLUE-TC 와 같은 데이터를 낮선 데이터로 활용하였으며, KLUE-TC 에서는 반대로 NSMC 데이터를 낮선 데이터 세트로 활용하였다.

3.3 실험 결과

표 2. 다양한 데이터 증강 및 label smoothing 기법과 본 논문에서 제안하는 방식과의 비교

데이터 세트 / 기법	Hate	NSMC	KLUE_TC
원본(기준치)	64.57	88.92	85.66
라벨 스무딩	66.30	88.70	85.49
역번역	65.89	88.52	84.88
EDA	65.49	87.95	85.31
AEDA	66.16	88.45	85.43
익숙한 데이터	64.43	88.88	84.59
낯선 데이터	**68.55**	**89.04**	**85.68**

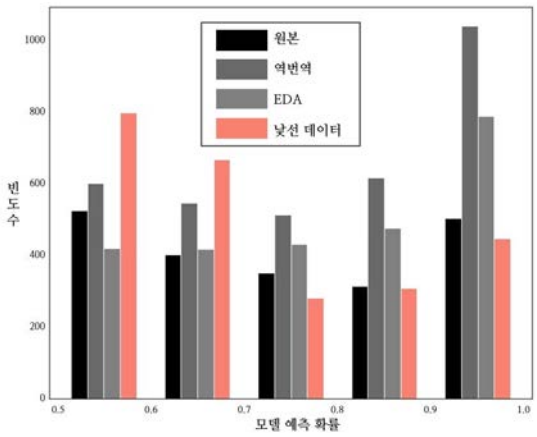

그림 4 증강 기법에 따른 모델 예측 확률 도식화

기존 기법들의 과잉확신 현상을 확인하기 위해 각 기법의 NSMC 평가 데이터 세트 오답에 대한 모델 예측 확률을 도식화하였다.

그림 4 에서 확인할 수 있듯이 본 논문에서 제안하는 방식인 낯선 데이터 세트를 활용한 기법이 오답에 대해 높은 예측 확률에서의 가장 낮은 빈도를 보이고 있다. 특히나 데이터 증강 없이 원본 데이터만으로 학습시켰을 때 보다 과잉확신에 대한 문제를 완화하고 있음을 확인할 수 있다. 뿐만 아니라 EDA, 역번역 기법의 경우 기존 증강을 진행하지 않은 원본 데이터 세트보다 더 높은 과잉확신을 보여주고 있다.

다음으로, 각 실험 데이터 세트에 대해서 원본 데이터 세트만으로 학습을 진행한 기준치와 다양한 데이터 증강기법을 적용하여 증강된 데이터를 추가적으로 사용한 결과를 비교하였다. 모델을 학습할 때 마다 발생하는 편차를 고려하기 위해 모든 실험을 5 회 진행한 후에 평균 값을 표 2 에 나타내었다. 각 학습데이터의 수 만큼 데이터를 증강하여 실험을 진행하였다.

실험 결과, 모든 데이터 세트에서 본 논문에서 제안된 기법을 통해서 증강된 데이터를 추가로 학습했을 때 제일 높은 정확도를 보였다. 특히나 데이터가 제일 적은 Hate 에서는 기준치 대비 3.98%p 의 성능 향상을 확인할 수 있으며, 다른 데이터 증강기법 중 가장 좋은 성능을 보인 AEDA 대비 2.39%p 의 성능 향상을 확인할 수 있었다. 규칙 기반의 데이터 증강기법 뿐만 아니라 딥러닝 모델을 활용하는 역번역 기법과 비교해보아도 본 논문이 더 많은 성능 향상 폭을 보임을 확인할 수 있었다. 기존의 데이터 증강 방식들을 적용할 경우 기존 데이터와 비슷한 분포의 데이터를 추가로 사용하게 되는 것이므로 이러한 기법이 과잉확신을 발생시켜서 본 논문에서 제안하는 기법과 비교해 낮은 성능 향상 폭을 보인다고 판단하였다.

과잉확신이 성능에 끼치는 영향을 알아보기 위해, 각 실험에 사용되는 데이터 세트와 유사도가 높은 익숙한 데이터들에 대해서도 본 논문에서 적용한 기법을 똑같이 접목시켜 보았다. 예를 들어서 NSMC 데이터 세트의 경우 높은 유사도를 가진 Hate 데이터 세트가 익숙한 데이터이다. 모든 실험에서 데이터 증강 없이 진행한 모델보다 오히려 성능이 떨어지는 것을 확인하였다. 이를 통해 본 논문에서 제안하는 방식이 미세 조정 과정에서 과잉확신을 완화시켜 더 좋은 성능을 보이는 것을 확인하였다.

4. 결론

본 논문에서는 낯선 데이터 세트를 규정하여 딥러닝 모델의 과잉확신 현상을 완화하고 성능을 개선하는 기법을 제안한다. SBERT 를 활용하여 한글 딥러닝 데이터 세트 간의 유사도를 측정하는 방식을 통하여 낯선 데이터 세트를 규정했다. 다음으로, 규정된 데이터 세트에 중립적인 라벨 값을 부여하여 새로운 학습 데이터로서 활용했다. 실험을 통해 EDA, 역번역과 같은 데이터 증강 기법들은 과잉확신에 대한 문제를 야기하는 것을 확인한 반면, 제안하는 기법은 과잉확신을 완화시킴을 관찰할 수 있었다. 실험을 통해 해당 기법이 분류 작업에서 기준치 및 다른 데이터 증강 기법보다 성능이 향상됨을 확인하여 낯선 데이터를 학습에 활용할 수 있음을 보였다.

제안하는 기법은 낯선 데이터 세트에 대해 결정론적 관측치로서 일정한 라벨을 할당한다. 이러한 기존 데이터간의 레이블 할당 차이는 손실함수 계산시 로짓값에서 큰 차이가 있게 되고, 수치적 불안정성을 야기할 수 있다. 향후 연구에서는 낯선 데이터에 대해 확률론적 예측을 통해 수치적 불안정성을 완화하는 기법으로 발전시키고자 한다.

사사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2022R1C1C1008534)을 받아 수행된 연구임.

참고 문헌

1. Li, D. and Zhang, H. Improved Regularization and Robustness for Fine-tuning in Neural Networks. Advances in Neural Information Processing Systems 34. 2021.
2. Sennrich, R., Haddow, B. and Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 86-96. 2016.
3. Wei, J. and Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6383-6389. 2019.
4. Karimi, A., Rossi, L. and Prati, A. AEDA: An Easier Data Augmentation Technique for Text Classification. Findings of the Association for Computational Linguistics: EMNLP 2021. 2021.
5. Zhang, L. and Ma, K. A Good Data Augmentation Policy Is Not All You Need: A Multi-Task Learning Perspective. IEEE Transactions on Circuits and Systems for Video Technology. 2022.
6. Li, Z. and Hoiem, D. Improving Confidence Estimates for Unfamiliar Examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2683-2692. 2020.
7. Devlin, J., Chang, M., Lee, K. and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171-4186. 2019.
8. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. and Smith, N. A. Annotation Artifacts in Natural Language Inference Data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 107-112. 2018.
9. Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp.3982-3992. 2019.
10. https://github.com/BM-K/KoSentenceBERT-ETRI December 27, 2022.
11. https://github.com/e9t/nsmc December 27, 2022.
12. Moon, J., Cho, W. I., and Lee, J. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media. pp. 25-31. 2020.
13. Park, S., Moon, J., Kim, S., Cho, W., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S, Park, E., Oh, A., Ha, J. and Cho, K. KLUE: Korean Language Understanding Evaluation. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021). 2021.
14. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aih

ubDataSe=realm&dataSetSn=126 December 27, 2022.

15. Park, J., Hong, J. and Cha, J. Korean Language Resources for Everyone. In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers. pp. 49-58. 2016

