소프트 라벨을 적용한 규칙 기반 텍스트 데이터 증강 기법

최주환[1], 이준호[1], 송상민[1], 진교훈[2], *김영빈[2]
[1]중앙대학교 AI학과
[2]중앙대학교 첨단영상대학원
e-mail: gold5230@cau.ac.kr, jhjo32@cau.ac.kr,
s2022120859@cau.ac.kr, fhzh123@cau.ac.kr, ybkim85@cau.ac.kr

Applying Soft Label to Rule-based Text Data Augmentation Methods

Juhwan Choi[1], Junho Lee[1], Sangmin Song[1], Kyohoon Jin[2], and *Youngbin Kim[2]
[1]Department of Artificial Intelligence, Chung-Ang University
[2]Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University

Abstract

The use of rule-based text data augmentation is common in various NLP tasks because of its simplicity. However, this method may have negative effects on the text's original meaning, which can potentially hurt the performance of the model as a result. We propose a simple method for applying soft labels to the augmented data to address this problem. We conducted experiments on seven different text classification tasks and found evidence supporting the effectiveness of our proposed approach.

I. 서론

최근 이미지, 텍스트, 음성 등의 다양한 형태의 데이터를 처리하기 위해 가장 많이 활용되는 방법은 딥러닝 기법이다. 이 중에서도 특히 자연어 데이터를 처리하는 데 있어서 순환 신경망 기반의 모델에서부터 Transformer 모델의 고안, 그리고 이에 기반한 사전 학습 언어 모델의 등장까지 딥러닝 기술의 발전에 힘입어 빠른 발전과 높은 성능을 거둘 수 있었다.

이와 같이 딥러닝 기술이 발전하면서 모델의 크기 역시 증대되었는데, 주어진 데이터에 기반하여 학습을 진행하는 딥러닝 기법의 특성 상 더욱 큰 규모의 모델을 학습시키기 위해서 학습 데이터 역시 더욱 많이 요구되었다. 모델의 규모에 비해 학습에 사용할 수 있는 데이터가 부족한 상태에서 학습을 진행했을 경우, 딥러닝 모델이 주어진 데이터에 대해서 지나치게 의존하는 과적합 현상이 발생할 수 있다.

이렇게 데이터가 부족하여 발생하는 문제를 완화하기 위하여 다양한 데이터 증강 기법이 고안되었다. 데이터 증강이란 학습에 사용할 수 있는 데이터를 추가적으로 확보하기 위해 기존에 가지고 있는 데이터를 가공하여 새로운 데이터를 얻는 작업을 말한다. 이러한 데이터 증강을 통해 딥러닝 모델의 일반화 성능을 높이고, 최종적으로는 과적합을 방지하는 효과를 기대할 수 있다[1].

이미지 데이터의 경우, 주어진 이미지의 일부만을 잘라서 사용하거나 이미지를 좌우로 반전시켜서, 혹은 이미지를 회전시키는 단순한 규칙을 통해서 데이터를 증강시켜 결과적으로 성능 향상을 얻을 수 있음이 잘 알려져 있다[2]. 이와 같이 미리 정해져 있는 규칙을 기반으로 보유한 데이터를 일부 수정하는 방법론을

규칙 기반의 데이터 증강 기법이라고 한다.

이미지 처리 분야에서 활용되는 규칙 기반 증강 기법에 영향을 받아 텍스트 데이터에 대해서도 단순하며 효율적인 규칙 기반의 데이터 증강 기법이 자주 활용된다. 가장 대표적인 규칙 기반의 텍스트 데이터 증강 기법은 Easy Data Augmentation (EDA) [3]로, 이는 주어진 문장에서 무작위로 단어를 선택하여 동의어로 교체하거나, 문장을 구성하는 단어 간의 순서를 바꾸거나, 문장에서 일부 단어를 삭제하거나 문장에 무작위로 단어를 삽입하는 네 가지의 세부 기법으로 구성된 방법이다. 이러한 기법을 통하여, 간단한 규칙을 통해 원본 데이터와 유사하면서도 일부 변형된 새로운 데이터를 얻을 수 있다. 변형을 가한 새로운 데이터에는 변형 이전의 원본 데이터와 동일한 라벨 값을 할당하여 학습에 활용한다. 그러나, 이러한 기법은 문장 내에서 단어를 삭제하는 등의 과정을 통해 주어진 문장이 가지고 있는 본래 의미를 일부 훼손시킬 수 있다는 단점이 있다.

본 연구에서는 소프트 라벨을 적용해 기존의 규칙 기반 텍스트 데이터 증강 기법을 개선하는 방법을 제안한다. 소프트 라벨이란 데이터의 라벨 값에 변형을 가해 딥러닝 모델의 일반화 성능을 높이고자 하는 기법 중 하나이다. 정답 클래스에 대해서는 1, 그 외에 대해서는 0으로 값을 나타내는 원-핫 인코딩 라벨과는 달리 소프트 라벨은 정답 이외의 다른 클래스에도 값을 할당하여, 더욱 다양한 정보를 전달할 수 있다. 기존 데이터의 원-핫 인코딩 라벨을 소프트 라벨로 변환하기 위해서 주로 활용되는 방법은 라벨 스무딩 기법[4]이다. 라벨 스무딩을 통해 모든 클래스에 동일하게 값을 할당시킨 균일 분포를 반영함으로써, 원-핫 인코딩 라벨을 소프트 라벨로 변환시킬 수 있다.

주어진 이미지를 좌우 반전시키는 등의 규칙 기반 이미지 데이터 증강 기법의 경우, 해당 규칙을 적용한 이후의 증강된 데이터는 원본 데이터의 의미적 정보를 그대로 유지하고 있다. 예를 들어, 강아지 이미지를 좌우 반전시킨 결과물 역시 강아지로 인식할 수 있는 이미지이기 때문에, 증강된 데이터에 강아지에 대한 원-핫 인코딩 라벨 값을 유지시킬 수 있다. 그러나 이와는 달리, 텍스트 데이터를 증강시키는 기존의 EDA 기법은 원래 문장에 규칙을 적용하는 과정에서 원본 데이터에서 의미적인 변형이 발생할 가능성이 있다. 그러나 기존의 EDA 기법은 이러한 의미 변형에 의해 라벨 값에 불확실성이 발생함에도 불구하고, 증강된 데이터에 원본 데이터의 라벨 값을 그대로 부여한다. 이는 모델이 학습 과정에서 의미적인 변형이 발생한 데이터와 원본 데이터를 구별하기 어렵게 만들고, 결과적으로 최종적인 성능을 떨어뜨릴 수 있다.

본 연구는 이러한 문제를 완화하기 위해 증강된 데이터의 라벨 값에 대해 라벨 스무딩 기법을 적용하여 증강된 데이터가 원본 데이터와 다른 소프트 라벨 값을 갖도록 한다. 이러한 방법을 통해 딥러닝 모델은 학습 과정에서 증강된 데이터를 학습할 때 소프트 라벨에서 원본 데이터의 원-핫 인코딩보다 상대적으로 약한 신호를 학습할 수 있게 되어, 결과적으로 일반화 성능을 향상시킬 수 있다.

본 논문에서 제안하는 방법을 검증하기 위해 텍스트 분류 작업에 대해서 실험을 진행한 결과, 기존의 규칙 기반 텍스트 증강 기법과 비교했을 때 최대 2.51%p의 성능 향상을 확인할 수 있었다. 또한, 기존 기법을 적용했을 때 성능이 하락하는 경우에도 본 논문에서 제안하는 기법을 적용했을 때에는 성능을 향상시킬 수 있음을 확인하였다.

II. 본론

이 장에서는 본 논문에서 제안하는 방법인 소프트 라벨을 적용한 규칙 기반 텍스트 데이터 증강 기법을 구성하는 요소에 대하여 설명한다.

2.1 Easy Data Augmentation (EDA)

본 논문에서는 기존의 규칙 기반 텍스트 데이터 증강 기법에 소프트 라벨을 적용하여 이를 개선시키는데, 이를 위한 대상으로 가장 많이 활용되는 EDA를 활용하였다.

EDA는 원본 문장을 단어 수준에서 변형시키는 네가지의 세부 기법으로 구성되어 있다. 가장 먼저, 동의어 교체 기법은 원본 문장에서 무작위로 단어 하나를 선택하여 이를 동의어 중 하나로 교체한다. 무작위 단어 삽입 기법의 경우, 원본 문장에서 무작위로 단어 하나를 선택하여 이 단어의 동의어를 문장의 무작위 위치에 삽입한다. 무작위 순서 변경의 경우 문장에서 무작위로 두 단어를 골라 이들끼리 순서를 변경한다. 무작위 삭제 기법은 문장에서 단어 중 일부를 미리 정해진 확률에 따라 삭제한다.

2.2 라벨 스무딩

라벨 스무딩이란 원-핫 인코딩 형태의 라벨 값을 모든 클래스에 동일하게 값이 부여되는 균일 분포를 활용해 소프트 라벨 값으로 변환하는 방법이다. 이를 통해, 모델의 일반화 성능을 향상시킬 수 있다. 본 연구에서는 EDA 기법이 원래 문장에서 변화를 일으키는 과정에서 원래 문장이 가지고 있는 라벨 값에 대한 의미가 손상될 수 있음을 고려하여, 증강된

데이터에 라벨 스무딩을 적용하여 원-핫 인코딩 값 대신 노이즈가 주입된 소프트 라벨 값을 부여한다. K개의 클래스를 가지고 있는 기존의 원-핫 인코딩 라벨값 y에 대해 라벨 스무딩을 적용하여 얻는 소프트 라벨 값은 아래와 같이 정의된다.

$$\hat{\boldsymbol{y}} = (1 - \alpha)\boldsymbol{y} + \frac{\alpha}{K}$$

이때 α는 라벨 스무딩의 강도를 결정하는 스무딩 매개변수이다. α가 클수록 소프트 라벨에서 정답 라벨의 값은 낮아지고, 균일 분포에 해당하는 값의 비중이 커진다.

2.3 제안 기법

본 논문에서는 규칙 기반 데이터 증강 기법의 단순성을 유지하면서도, 기존의 규칙 기반의 텍스트 증강 기법이 가지고 있었던 의미 손실로 인한 성능 하락의 가능성을 완화하기 위한 방안을 제시한다. 제안 기법은 기존의 규칙 기반 텍스트 데이터 증강 기법인 EDA를 활용하되, 이 과정에서 기존의 EDA가 증강 이전 데이터의 라벨 값을 그대로 유지했던것과 달리 증강된 데이터의 라벨 값에 라벨 스무딩을 적용한 소프트 라벨을 할당한다. 이와 같은 단순한 과정을 추가함으로써, 규칙 기반 데이터 증강 기법의 장점인 단순성을 유지하면서도 기존의 규칙 기반 텍스트 증강 기법이 가지고 있었던 의미 손실 문제를 완화할 수 있다. 기존의 EDA 기법과 본 논문에서 제안하는 기법의 라벨 값 차이를 표 1에 나타내었다.

증강된 데이터 문장	기존 기법 라벨값	제안 기법 라벨값
I really enjoyed watching the new movie. → I watching really enjoyed the new.	긍정 1 부정 0	긍정 0.9 부정 0.1
The new horror movie was an awful, boring disaster. → The boring horror was an, new movie disaster.	긍정 0 부정 1	긍정 0.1 부정 0.9

표 1. 기존 기법과 제안 기법의 라벨값 비교

Ⅲ. 실험

이 장에서는 본 논문에서 제안하는 기법의 효과를 검증하기 위한 실험 구성과 실험 결과에 대해 설명한다.

3.1 실험 환경

텍스트 분류 작업을 통해 제안하는 방법의 효과를 검증하기 위해 사전 학습된 BERT [5] 모델을 바탕으로 실험을 진행하였다. 세 가지의 텍스트 분류 모델을 학습시키기 위한 Batch 크기는 32로 지정하였다. 또한, Optimizer로는 AdamW를 사용하였고 이 때의 Learning Rate는 1e−4로 하였다.

실험을 위한 데이터셋으로는 총 7개의 데이터셋을 사용하였다. 이진 감정 분류 데이터셋으로 SST2 [6], CR [7], MR [8]을 활용하였다. 또한, 이외 다양한 분야에 대한 이진 분류 데이터셋으로 SUBJ [9], PC [10], CoLA [11]를 활용하였다. 마지막으로, 이진 분류가 아닌 다중 분류 데이터셋으로 TREC [12] 데이터셋을 활용하였다.

소프트 라벨을 부여하기 위해 라벨 스무딩을 적용할 때, 스무딩 강도를 결정하는 매개변수 α는 0.1, 0.15, 0.2, 0.25, 0.3으로 설정하여 가장 좋은 결과를 활용하였다.

3.2 실험 결과

각 실험 데이터셋마다 BERT 모델에 대해 원본 데이터만을 통해 학습을 진행시킨 Baseline과 제안하는 기법을 통해 증강된 데이터를 추가적으로 사용한 학습 결과, 그리고 기존의 규칙 기반 데이터 증강 기법인 EDA와 EDA를 수정하여 문장 부호의 무작위 삽입만으로 구성된 AEDA [13] 를 적용한 결과를 비교 대상으로 하여 각 방법에 대해 학습 이후 성능을 표 2로 나타내었다. 제안하는 기법의 경우, 라벨 스무딩의 강도에 따라 차이를 보이는 결과 중 가장 좋은 결과를 나타내었다.

실험 결과, 본 논문에서 제안하는 방법이 대부분의 경우에서 다른 방법과 비교했을 때 성능을 크게 향상시킴을 확인하였다. 또한, 제안 기법은 EDA 등의 다른 기법이 성능 하락을 보이는 경우에도 모델의 성능을 향상시킬 수 있었다. 특히, CR 데이터셋에서 EDA 기법을 적용하였을 때 Baseline과 비교해 성능이 0.41%p 감소하였던 것과 달리, 제안하는 방법을 적용하였을 때에 성능이 2.10%p 향상되어 EDA 기법과 2.51%p의 차이를 보였다.

기법/데이터셋	SST2	CR	MR	TREC
Baseline	89.74%	89.08%	84.28%	95.47%
EDA	+0.71%p	−0.41%p	−	+0.51%p

			0.92%p	
AEDA	+0.22%p	+1.84%p	**+0.19%p**	−0.67%p
Proposed Method	**+0.83%p**	**+2.10%p**	**+0.19%p**	**+1.17%p**
기법/데이터셋	SUBJ	PC	CoLA	
Baseline	96.18%	93.44%	75.38%	
EDA	−0.35%p	+0.58%p	−0.45%p	
AEDA	−0.30%p	−0.15%p	−0.34%p	
Proposed Method	**+0.15%p**	**+0.67%p**	**+1.50%p**	

표 2. 데이터셋 별 실험 결과

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 기존의 규칙 기반 텍스트 증강 기법의 한계점을 보완하기 위해 증강된 데이터에 라벨 스무딩을 통해 소프트 라벨을 적용하는 방법을 제안하였다. 또한, 제안한 방법을 적용했을 때 기존 방법과 비교해 높은 수준의 성능 향상을 이뤄낼 수 있음을 확인할 수 있었다. 향후에는 증강 기법을 적용하는 강도에 따라 라벨 스무딩의 강도 또한 조절하는 기법에 대해서 연구해보고자 한다.

Acknowledgements

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF−2022R1C1C1008534)과 정보통신기획평가원의 지원(2021−0−01341, 인공지능대학원지원(중앙대학교))의 지원을 받아 수행된 연구임.

참고문헌

[1] Sebastien C. Wong, Adam Gatt, Victor Stamatescu and Mark D. McDonnell. "Understanding Data Augmentation for Classification: When to Warp?" 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1−6, 2016.

[2] Luke Taylor and Geoff S. Nitschke. "Improving Deep Learning with Generic Data Augmentation." 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1542−1547. 2018.

[3] Jason Wei and Kai Zou. "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] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens and Zbigniew Wojna. "Rethinking the Inception Architecture for Computer Vision." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818−2826, 2016.

[5] Jacob Devlin, Ming−Wei Chang, Kenton Lee and Kristina Toutanova. "BERT: Pre−training of Deep Bidirectional Transformers for Language Understanding." In North American Association for Computational Linguistics (NAACL), pp. 4171−4186, 2019.

[6] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank." In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631−1642, 2013.

[7] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168−177, 2004.

[8] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79−86, 2002.

[9] Bo Pang and Lillian Lee. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts." In Proceedings of the 42nd Annual Meeting of the Association for Computational

Linguistics (ACL-04), pp. 271-278, 2004.

[10] Murthy Ganapathibhotla and Bing Liu. "Mining Opinions in Comparative Sentences." In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 241-248, 2008.

[11] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. "Neural Network Acceptability Judgments." Transactions of the Association for Computational Linguistics, 7:625-641, 2019.

[12] Xin Li and Dan Roth. "Learning Question Classifiers." In COLING 2002: The 19th International Conference on Computational Linguistics, pp. 1-7, 2002.

[13] Karimi Akbar, Rossi Leonardo and Prati Andrea. "AEDA: An Easier Data Augmentation Technique for Text Classification." In Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2748-2754, 2021.