앙상블 기법을 활용한 NLP 의 잘못된 상관관계 감소 통합 접근법

진교훈, 최주환, 윤정민, 이준호, 장수진, 김영빈
중앙대학교

fhzh123@cau.ac.kr, gold5230@cau.ac.kr, cocoro357@gmail.com, jhjo32@cau.ac.kr,
sujin0110@cau.ac.kr, ybkim85@cau.ac.kr

Unified Approach for Spurious Correlation Reduction in NLP via Ensemble Techniques

Kyohoon Jin, Juhwan Choi, Jungmin Yun, Junho Lee, Soojin Jang and Youngbin kim
Chung-ang Univ.

요 약

본 연구는 자연어 처리(NLP)에서 대규모 딥러닝 모델이 학습하는 잘못된 상관관계를 식별하고 해결하는 방법을 제안합니다. 다양한 매개변수와 훈련 방법을 가진 여러 모델들이 서로 다른 잘못된 상관관계를 나타내는 현상을 관찰하였으며, 이를 해결하기 위해 앙상블 기법을 사용하여 중요한 단어를 집계하고 분석하는 방법을 개발했습니다. IMDb, SST2, TREC6 데이터셋을 사용한 실험 결과, 제안된 방법이 기존의 대조 데이터 생성 기술보다 성능을 크게 향상시킴을 확인했습니다. 이는 여러 모델의 장점을 결합한 접근 방식이 잘못된 상관관계를 효과적으로 포착하고 완화할 수 있음을 입증합니다.

Ⅰ. 서 론

최근 기술의 발전으로 데이터 수집이 용이해지면서, 대규모 딥러닝 모델의 훈련이 가능 해졌다. 특히 자연어 처리(NLP) 분야에서 딥러닝 모델의 크기 증가로 성능이 향상되었다. 그러나 대규모 데이터셋으로 훈련된 딥러닝 모델에서는 "잘못된 상관관계"가 발생하는 문제가 있다. 이는 모델이 의도하지 않은 잘못된 연관성을 학습하여 의도한 결과를 생성하는 능력을 저해한다. 이러한 잘못된 상관관계는 모델의 견고성과 사회적 영향을 악화시킬 수 있다 [1].

많은 NLP 연구는 이러한 잘못된 상관관계를 식별하고 해결하는 것을 목표로 한다. 예를 들어, 영화 리뷰 데이터셋에서 코미디와 드라마 장르는 긍정적인 평가를, 공포 장르는 부정적인 평가를 받는 경향이 있다. 이로 인해 모델은 코미디와 드라마 장르에 대해 긍정적인 결과를, 공포 장르에 대해 부정적인 결과를 생성하는 경향이 있다 [2]. 이러한 잘못된 상관관계를 해결하기 위해 데이터 증강, 인과 개입, 대조 학습 등의 다양한 방법이 제안되었다.

본 연구에서는 단일 모델에만 집중하지 않고, 다양한 모델에서 관찰된 잘못된 상관관계를 분석한다. 우리의 연구는 다양한 매개변수와 훈련 방법을 가진 모델들이 서로 다른 잘못된 상관관계를 나타낸다는 것을 밝혀냈다. 이를 해결하기 위해 앙상블 기법을 사용하여 여러 모델에서 중요한 단어를 집계하고 분석하여 강력한 대조 데이터를 생성한다. 실험 결과, 제안된 방법이 기존의 대조 데이터 생성 기술에 비해 성능을 크게 향상시킴을 확인했다. 이는 여러 모델에서 잘못된 상관관계를 효과적으로 포착하고 완화하는 접근 방식이 효과를 보인것으로 파악된다.

Ⅱ. 본론

이 연구에서는 각 모델이 감정 분류에 영향을 미치는 최고 중요 단어를 어떻게 결정하는지와 이러한 선택이 모델 간에 차이가 있는지 조사했다. 이 접근법은 모델이 텍스트를 해석하고 처리하는 방식의 차이를 제시하는 것을 목표로 한다. 해당 실험을 위해 주요 단어 선정 방식으로 Lime 과 Integrated Gradient 를 사용하였다 [3-4].

Lime	Integrated Gradient
8.64	7.00

표 1. 데이터별 주요 단어의 겹침 정도

표 1 은 IMDb 데이터셋에 대해서 모델이 선택한 최고 중요 단어의 겹침을 정량화한 것으로, 상위 10 개 단어에 중점을 둔다 [5]. 네 가지 모델 중 적어도 하나가 동일한 단어를 포함하는 경우를 카운트하며, 겹침이 없는 경우는 카운트하지 않는다. 표는 전체 데이터에 대한 카운트 비율을 보여준다. 모든 모델이 BERT-Base 를 기반으로 하고 있음에도 불구하고, 선택된 주요 단어는 상당한 변동성을 보였다 [6]. 놀랍게도, 상위 10 개 주요 단어를 선택할 때조차도 네 가지 모델 모두 동일한 단어를 선택한 경우는 7%에 불과했다. 이러한 결과는 모델들이 유사한 정확도를 보이더라도, 중요한 특징으로 간주되는 특정 단어가 크게 다를 수 있음을 시사한다.

주요 단어는 문장의 레이블에 영향을 미치는 결정적 속성이므로 이 단어들을 수정해야 합니다. 반면, 잘못된 패턴을 유발하는 단어는 레이블이 변경되더라도 유지되어야 한다. 먼저, 반사실적 데이터를 생성하기

위해 최우선 중요한 단어에 마스킹을 적용한다. 또한, 반사실적 데이터 생성을 보장하기 위해 무작위로 선택된 단어도 마스킹한다. 이 과정에서 잘못된 패턴을 유발하는 것으로 식별된 단어는 무작위 선택에 포함되지 않는다. 마스킹된 단어는 언어 모델을 사용하여 채워진다. 본 논문에서는 라벨별 마스크 채우기를 수행할 수 있도록 언어 모델을 미세 조정하기 위해 프리픽스-튜닝(prefix-tuning)을 사용한다 [7]. 본 논문에서는 언어 모델로 BART-Base 를 사용했다 [8].

우리의 방법은 잘못된 패턴을 유발하는 단어를 효과적으로 식별했다. 이러한 차이가 실제 모델 성능에 얼마나 영향을 미치는지 평가하기 위해 다양한 분류 데이터셋에서 실험을 진행했다. 해당 실험을 위해 데이터셋으로는 SST2, IMDb 그리고 TREC6 데이터를 사용하였으며, 비교대상으로는 Fine-tuning 을 진행한 BERT 를 Baseline 으로, SentimenCAD 와 본 논문에서 제안하는 방식으로 증강된 데이터로 학습을 진행한 BERT 모델을 사용했다. 그리고 성능 평가 지표로는 정확도를 사용하였다 [9-10].

	SST2	IMDb	TREC6
Baseline	92.8	91.5	94.6
SentimentCAD	89.8	90.1	97.3
Ours	94.6	93.9	97.8

표 2. 각 데이터에 대한 제안하는 방식과 비교 방식간의 성능 차이 비교

표 2 는 제안된 방법이 대부분의 데이터셋에서 일관되게 가장 높은 정확도를 달성했음을 보여준다. 이는 감정 분석뿐만 아니라 주제 분류에서도 강력한 성능을 보였다. 이러한 개선은 기존 방법들이 단일 모델 특징에 집중하여 최적의 주요 단어를 선택하지 못했기 때문일 가능성이 크다. 그러나 여러 모델을 결합함으로써, 우리는 주요 단어를 더 효과적으로 선택할 수 있었고, 이는 성능 향상으로 이어졌다.

Ⅲ. 결론

본 연구는 여러 딥러닝 모델을 활용한 앙상블 기법을 통해 자연어 처리에서 발생하는 잘못된 상관관계를 효과적으로 해결하는 방법을 제안했다. 다양한 모델의 상위 중요 단어를 집계하고 분석함으로써, 모델 간 변동성을 줄이고 더욱 강력한 대조 데이터를 생성할 수 있었다. 실험 결과, 제안된 방법은 감정 분석, 주제 분류 등 다양한 분야에서 기존 방법들보다 높은 정확도를 일관되게 달성했다. 이는 여러 모델의 장점을 결합한 접근 방식이 잘못된 상관관계를 포착하고 완화하는 데 효과적임을 입증한다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Ye, W., Zheng, G., Cao, X., Ma, Y., Hu, X., & Zhang, A. (2024). Spurious Correlations in Machine Learning: A Survey. arXiv preprint arXiv:2402.12715.

[2] Kaushik, D., Hovy, E., & Lipton, Z. (2019). Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In International Conference on Learning Representations.

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[4] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR.

[5] Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).

[6] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

[7] Li, X. L., & Liang, P. (2021, August). Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 4582-4597).

[8] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020, July). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871-7880).

[9] Li, X., & Roth, D. (2002). Learning question classifiers. In COLING 2002: The 19th International Conference on Computational Linguistics.

[10] Yang, L., Li, J., Cunningham, P., Zhang, Y., Smyth, B., & Dong, R. (2021, August). Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 306-316).