텍스트 분류를 위한 Wasserstein Autoencoder 를 활용한 생성적 데이터 증강

진교훈[1], 이준호[2], 최주환[2], 송상민[2], 장수진[1], 김영빈[1]*

1 중앙대학교 첨단영상대학원
2 중앙대학교 AI 학과

fhzh123@cau.ac.kr, jhjo32@cau.ac.kr, gold5230@cau.ac.kr, s2022120859@cau.ac.kr, sujin0110@cau.ac.kr, *ybkim85@cau.ac.kr

Generative Data Augmentation for Text Classification using Wasserstein Autoencoder

Kyo Hoon Jin, Jun Ho Lee, Ju Hwan Choi, Sang Min Song, Soo Jin Jang,
Young Bin Kim*

1Graduate School of Advanced Imaging Science, Multimedia & Flim, Chung-Ang University
2Department of Artificial Intelligence, Chung-Ang University

요 약

잠재 변수 모델은 텍스트 생성을 비롯한 다양한 분야에서 활용되고 있다. 하지만 VAE(Variational Autoencoder)와 같은 모델은 훈련 중 잠재 변수의 일부를 무시하는 사후 붕괴 문제를 겪는다. 이는 자연어 처리에 적용 시 재구성 성능 저하로 이어진다. 본 논문에서는 WAE(Wasserstein Autoencoder)와 사전 학습된 언어 모델(PLM)을 기반으로 한 데이터 증대 방법을 제안한다. WAE 를 이용하여 사후 붕괴 문제를 방지하고, 적절한 PLM 활용을 통해 증강 효과를 향상시켰다. 제안한 방법은 4 개의 벤치마크 데이터 세트에 대한 실험에서 증강 효과를 입증하였다.

Ⅰ. 서 론

최신 딥 러닝 모델은 자연어 처리 작업에 탁월하지만 데이터의 다양성과 양에 따라 성능이 달라진다. 데이터가 제한적이거나 편향된 경우 문제가 발생하여 일반화가 제대로 이루어지지 않고 과적합이 발생한다. 이미지 도메인 데이터 증강 연구는 잘 확립되어 있지만 텍스트 데이터 증강은 레이블에 직접적인 영향을 미치기 때문에 고유한 문제를 제기한다 [1].

이 논문에서는 Wasserstein Autoencoder (WAE)를 활용하는 텍스트 증강 방법을 도입하여 이러한 문제를 해결하고자 한다. 기존 방법과 달리 사후 분포를 정규화하여 Variational Autoencoder(VAE)에서 흔히 발생하는 문제인 사후 붕괴를 완화한다 [2]. 이는 VAE 가 이산 데이터 및 자동 회귀 생성기로 어려움을 겪는 NLP 작업에서 특히 중요하다 [3].

제안하는 방식은 사전 훈련된 언어 모델을 활용하여 잠재 변수를 효과적으로 추출하고 문장을 원활하게 생성한다. 제안된 방법은 4 개의 벤치마크 데이터세트에서 검증되었으며 기존 모델에 비해 1~2% 정확도 향상을 보여준다. VAE 를 사용한 비교 분석은 우리 접근 방식의 효율성을 강조하고 다양한 PLM 을 통한 확장성을 보여준다.

Ⅱ. 본론

$x^{(i)}$문장과 $z^{(i)}$로 표시된 잠재 변수 Z로 구성된 데이터 세트 X의 맥락에서 오토인코더(AE)의 목적함수는 다음과 같다.

$$J_{AE} = -\mathbb{E}_{x\sim Q(x|x)}[\log G(x|z)]$$

AE 에서 파생된, Variational Autoencoder(VAE)는 잠재 변수 하위 집합 z에 사전 분포 P_z를 도입한다. VAE 는 P_z를 계산하기 위해 사후 분포 $Q(z|x)$가 필요하며, VAE 의 쿨백-라이블러 (Kullback-Leibler; KL)발산을 최소화하는 방식으로 학습된다. KL 발산을 최소화하면 잠재 분포가 겹쳐서 잠재 변수 $z^{(i)}$에 해당 입력 $x^{(i)}$에 대한 의미 있는 정보가 부족하게 되며, 이를 사후 붕괴 (Posterior Collapse)라고 한다.

Wasserstein Autoencoder (WAE)는 Wasserstein 거리를 통해 사후 분포를 정규화하여 VAE 의 사후 붕괴 문제를 완화할 수 있다. 개별 입력에 대해 동일한 사후 분포 및 사전 분포를 요구하는 VAE 와 달리 WAE 는 집계된 사후 분포 Q_Z가 전체 데이터 세트 X에 대한 사전 분포 P_Z와 동일해야 한다는 제약 조건을 부과함으로써 디코딩 중에 더 의미 있는 잠재 변수 z를 사용할 수 있다.

Q_Z 의 정규화는 inverse multi-quadratic kernel 을 기반으로 하는 최대 평균 불일치(Maximum Mean Discrepancy; MMD)를 사용하여 계산된 Wasserstein 거리에 페널티를 적용하여 달성된다. WAE 의 훈련 목표는 재구성을 위한 표준 교차 엔트로피 손실과 Wasserstein 거리 페널티를 결합하여 하이퍼파라미터 λ를 통해 균형을 유지한다. 이에 대한 목적함수는 다음과 같다.

$$J_{WAE} = -\mathbb{E}_{Q_\phi(Z|X)}[\log G_\theta(X|Z)] + \frac{\lambda}{N(N-1)}\sum_{l\neq j}\left[k\left(z^{(l)},z^{(j)}\right)+k\left(\tilde{z}^{(l)},\tilde{z}^{(j)}\right)\right] - \frac{2\lambda}{N^2}\sum_{l,j}k\left(z^{(l)},\tilde{z}^{(j)}\right)$$

여기서 ϕ 와 θ는 인코더 및 디코더 매개변수를 나타내고 λ는 정규화 및 재구성 항의 균형을 맞추어 주는 하이퍼

파라미터이며 $\tilde{z}^{(l)}$ 및 $z^{(l)}$는 각각 집계된 사후 분포와 사전 분포의 샘플이다.

본 논문에서는 잠재 변수 추출의 성능을 향상시키기 위해 BART, T5 와 같은 사전 훈련된 언어 모델(Pre-trained Language Model; PLM)이 사용하였다. PLM 은 향상된 언어 이해를 제공하여 인코더-디코더 구조 선택의 유연성을 허용한다.

표 1. 각 모델에 대한 데이터 증강 전후 성능 비교

분류기	SST2	SST5	PC	MR
CNN	78.64	40.66	89.84	68.16
+ Ours	79.30	41.97	90.04	71.54
RNN	76.26	36.64	91.99	71.13
+ Ours	78.33	39.92	92.79	74.25
BERT	89.61	50.42	93.65	84.05
+ Ours	91.24	51.13	95.56	85.52

Ⅲ. 실험

제안된 방법을 다양한 데이터셋에 대해 테스트하기 위해 일반적으로 성능평가에 사용되는 SST2, SST5, ProsCons (PC), MR 데이터셋을 사용하였다 [4-6].

WAE 구조에서 인코더와 디코더로 BART 를 사용했다. 사전 분포는 가우스 분포를 기반으로 학습을 진행했으며, 최대 문장 길이는 300 자이다. 학습률은 5e-6 이며 AdamW Optimizer 를 사용했다 [7].

제안된 텍스트 데이터 증강 방법을 적용하기 전과 후의 벤치마크 데이터를 종합적으로 비교한 내용은 표 1 에 나와 있다. 결과는 데이터 세트 전체에서 약 0.2%p 에서 2%p 범위의 정확도가 크게 향상되었음을 확인할 수 있다. 특히, 다중 레이블이 지정된 데이터를 포함하는 SST5 데이터 세트는 2.98%p 의 성능 향상을 보여 제안된 증강 방법이 목표 달성에 성공했음을 확인했다. 표의 모든 결과는 정확성을 나타낸다.

Ⅳ. 결론

이 논문에서는 WAE(Wasserstein Autoencoder) 구조를 사용하여 텍스트 확대에 대한 새로운 접근 방식을 소개합니다. 우리의 방법은 간단하면서도 후방 붕괴로 인한 성능 저하 위험을 효과적으로 완화하여 모델 결과의 견고성을 보장합니다. 특히 PLM(사전 훈련된 언어 모델)뿐만 아니라 더 간단한 모델에서도 성능 향상이 뚜렷하게 나타납니다.

제안된 방법은 7 개의 벤치마크 데이터 세트에 대해 광범위한 테스트를 거쳤으며 지속적으로 향상된 성능을 보여주었습니다. 또한, 다양한 PLM 을 성공적으로 구현하여 확장성을 입증했습니다. 향후 연구에서는 이러한 확장성을 활용하여 텍스트 확대 기술을 더욱 발전시키기 위한 새로운 방식으로 다양한 PLM 의 조합을 탐색해야 합니다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

[2] Tolstikhin, I., Bousquet, O., Gelly, S., & Schoelkopf, B. (2017). Wasserstein auto-encoders. 6th International Conference on Learning Representations (ICLR).

[3] Lucas, J., Tucker, G., Grosse, R., & Norouzi, M. (2019). Understanding posterior collapse in generative latent variable models.

[4] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). 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).

[5] Ganapathibhotla, M., & Liu, B. (2008, August). Mining opinions in comparative sentences. In Proceedings of the 22nd international conference on computational linguistics (Coling 2008) (pp. 241-248).

[6] Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. International Journal of Advanced Computer Science and Applications, 8(6).

[7] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. 9th International Conference on Learning Representations (ICLR)..