비지도 대조 학습에서 삼중항 손실 함수 도입을 위한 토큰 컷오프 기반 데이터 증강 기법

한명수 [1], 정유현 [2], 채동규 [3]
[1] 한양대학교 인공지능학과 석사과정
[2] 한양대학교 인공지능학과 석박통합과정
[3] 한양대학교 컴퓨터소프트웨어과 교수

myngsoo@hanyang.ac.kr, robo0725@hanyang.ac.kr, dongkyu@hanyang.ac.kr

Data Augmentation Strategy based on Token Cut-off for Using Triplet Loss in Unsupervised Contrastive Learning

Myeongsoo Han[1], Yoo Hyun Jeong[1], Dong-Kyu Chae[2]
[1]Dept. of Artificial Intelligence, Hanyang University
[2]Dept. of Computer Science, Hanyang University

요 약

최근 자연어처리 분야에서 의미론적 유사성을 반영하기 위한 대조 학습 (contrastive learning) 관련 연구가 활발히 이뤄지고 있다. 이러한 대조 학습의 핵심은 의미론적으로 가까워져야 하는 쌍과 멀어져야 하는 쌍을 잘 구축하는 것이지만, 기존의 손실 함수는 문장의 상대적인 유사성을 풍부하게 반영하는데 한계가 있다. 이를 해결하기 위해, 이전 연구에서는 삼중 항 손실 함수 (triplet loss)를 도입하였으며, 본 논문에서는 이러한 삼중 항을 구성하기 위해 대조 학습에서의 효과적인 토큰 컷오프 (cutoff) 데이터 증강 기법을 제안한다. BERT, RoBERTa 등 널리 활용되는 언어 모델을 이용한 실험을 통해 제안하는 방법의 우수한 성능을 보인다.

1. 서론

최근 자연어처리 (NLP, natural language processing) 분야는 의미론적 유사성을 반영한 문장 표현 학습을 통해 감정 분석, 질의 응답 등의 분야에서 성능 향상을 이루고 있다. 문장 표현을 학습하는 것은 단어, 문장 등을 고정된 크기의 벡터로 임베딩하여 의미론적으로 유사한 문장들이 잠재 공간에서 서로 가까운 위치에 존재하도록 만드는 것을 의미한다. 그 중 널리 활용되는 방법으로 NT-Xent 손실 함수 (Normalized Temperature-scaled Cross Entropy Loss) 기반의 대조 학습 (contrastive learning) 이 있다. 대조 학습의 성능에 영향을 미치는 요소들 중 하나는 의미론적으로 가까워져야 하는 쌍 (양성 쌍)과 멀어져야 하는 쌍 (음성 쌍)을 잘 구성하는 것이다. 그러나 긍정과 부정 표현 쌍의 구성에만 초점을 맞추고 NT-Xent 손실 함수의 훈련 목적을 충분히 고려하지 않는 것은 문장 간 상대적 유사성을 모델링 할 수 없다는 한계가 있다 [1].

최근 자연어처리 분야에서 삼중 항 손실 함수 (triplet loss) [2]를 기존 대조 학습에 도입하는 방법 [1]과 토큰 단위에서의 데이터를 증강하는 기법 [3] 들이 발표되었다. 본 연구는 [1], [3]의 영감을 받아 대조 학습에서 삼중 항 손실 함수를 도입하고, 문장의 상대적 유사성을 반영하기 위한 토큰 컷오프 (cutoff) 방식의 데이터 증강 기법을 제안한다. 여기서 삼중 항으로 구성되는 문장 쌍은 {원본 문장, 약한 변형 문장, 강한 변형 문장}으로 구성되며 문장 간의 관계를 더 잘 모델링 하기 위해 설계되었다.

제안하는 방법의 우수성을 평가하기 위해 비지도 대조 학습에서 널리 활용되는 SimCSE [4]를 기준으로 하여, 의미론적 유사성을 평가하는 작업 (task)으로 구성된 STS (Sentence Textual Similarity) 데이터셋 [5]을 통해 성능을 평가한다. 실험 결과를 통해 제안하는 컷오프 방식으로 토큰화가 진행된 후 데이터 증강을 적용했을 때 성능 향상을 확인한다.

2. 관련 연구

2.1 비지도 문장 표현 학습

최근 보다 나은 문장 표현 학습을 위하여 사전 학

습 언어 모델을 대조 학습을 통해 성능을 개선하는 연구들이 발표되고 있다. 대표적으로 SimCSE [4]는 임베딩 공간에서 의미가 유사한 데이터는 가까워지도록, 유사하지 않은 데이터는 멀어지도록 함으로써 비지도 및 지도 문장 표현 학습 모두에서 성능 향상을 이루었다. ArcCSE [1]는 기존 대조 학습 기반의 손실 함수에 삼중 항 손실 함수를 추가하여 문장 표현 학습의 성능을 개선하였다.

2.2 데이터 증강

자연어처리 분야에서 모델의 성능을 높이기 위한 데이터 증강 관련 다양한 방법론들이 연구되었다. EDA (Easy Data Augmentation) [6]는 의미론적으로 유사성을 보유한 데이터 증강을 위해 문장 내 단어 제거, 유의어 대체, 단어 위치 변경 등의 다양한 증강 방법을 제시하고 실험적으로 성능을 측정하였다. Back Translation [7]은 인코더-디코더 구조의 언어 모델을 활용하여 문장을 재번역하는 데이터 증강을 다뤘다.

문장 표현을 위한 대조 학습의 경우, SimCSE [4]의 경우 데이터 증강 기법으로 드롭아웃 (dropout)을 활용하여 대조 쌍을 구성했으며, ConSERT [8]은 적대적 공격 (adversarial attack), 토큰 셔플링 (token shuffling), 컷오프 (cutoff) 등의 입력 및 토큰 단위의 데이터 증강 기법을 제시했다. ArcCSE [1]는 문장 간의 의미론적 순서를 반영할 수 있도록 문장 내에서 연속되는 단어 일부를 제거하는 방법으로 삼중 항 쌍을 구성했다.

3. 제안하는 방법

본 장에서는 삼중 항 손실 함수를 활용한 비지도 문장 표현 학습을 위해 토큰 컷오프 기반 데이터 증강 기법을 소개한다. 이전 연구 [1]와 유사하게, 삼중 항을 앵커 문장 (anchor), 약한 변형 문장, 강한 변형 문장으로 구성하였다. 이 때, 기존 논문 [1]이 입력 수준에서 변형을 준 것과 달리, 토큰 수준에서 변형을 주어 효과적인 유사성 학습이 가능하도록 한다. 토큰 컷오프 방법을 위해 사전 학습 언어 모델에서 제공하는 특수 토큰 (special token)을 활용하며, 모델의 마지막 층에 있는 MLP layer 내의 활성화 함수 변형을 통해 삼중항 쌍을 구성한다.

3.1 특수 토큰을 활용한 토큰 컷오프 기법

BERT [9]와 같은 사전 학습 언어 모델은 학습을 위해 [PAD], [CLS], [SEP], [UNK], [MASK]와 같이 서로 목적이 다른 특수 토큰들을 사전 (vocabulary) 에 포함하고 있다. 본 논문에서는 사전 내 미등록 단어를 위해 사용되는 [UNK] (unknown 을 의미) 토큰을 활용한다. 토큰화가 진행된 문장의 토큰들을 특정 비율로 [UNK] 토큰으로 마스킹 (masking) 하여 의미론적으로 유사한 문장을 여러 만들고, 변경되는 비율을 다르게 설정하여 강한 변형 문장 (s**, 20%) 과 약한 변형 문장 (s*, 10%) 을 생성한다. 그 결과, 아래 예시와 같은 삼중 항 (s, s*, s**) 을 구성할 수 있다.

s Then perhaps you just need to increase your water intake to revive your brain function.

s* Then perhaps you just need to increase your water ***[UNK]*** to revive your brain function.$

s** Then perhaps you just ***[UNK]*** to increase your water intake to revive your ***[UNK]*** function.

3.2 MLP Layer 활용

SimCSE [5]에서 드롭아웃을 노이즈로 사용하여 긍정 쌍을 구성한 것과 유사하게, 본 논문에서는 BERT의 마지막 층에 있는 MLP 층 내의 활성화 함수를 다르게 하여 강한 변형 문장 (s**) 의 임베딩에 노이즈 효과를 더한다. 이와 같은 방법은 임베딩 단에서 진행되기 때문에, 추가적인 전처리 작업 없이 삼중항 쌍을 구성할 수 있는 장점을 지닌다. 강한 변형 문장 (s**) 의 활성화 함수로는 기존에 사용되던 Tanh 함수 대신 LeakyReLU 함수를 사용하며, 기존 문장 (s) 과 약한 변형 문장 (s*) 은 Tanh 함수를 활용한다.

(1) [UNK] 토큰 컷오프 적용 (2) 활성화 함수 노이즈 적용

(그림 1) 데이터 증강 방법

3.3 최종 손실 함수

(그림 1)의 (1)과 (2)는 토큰 컷오프와 활성화 함수 노이즈를 적용했을 때 앵커 문장과의 거리를 나타낸다. 제안된 삼중 항 손실 함수의 목적은 임베딩 공간에서 앵커 문장과 다양한 차원의 변형 문장들 간 거리를 줄여 의미적 유사성을 풍부하게 하는 것이다. 기존 대조 학습의 NT-Xent 손실 함수에 삼중 항 손실 함수를 더한 최종 손실 함수는 아래와 같다.

$$L_{NT-Xent} = -\log \frac{e^{sim(s_i,\ s_i)/\tau}}{\sum_{j=1}^{n} e^{sim(s_i,\ s_j)/\tau}}$$

$$L_{triplet} = \max\left(0, sim(s_i, s_i^*) - sim(s_i, s_i^{**})\right)$$

$$L = L_{NT-Xent} + L_{triplet}$$

4. 실험

본 논문은 BERT-base, RoBERTa-base 를 활용하여 실험을 진행하였다. 해당 모델들을 통해 제안하는 토큰

컷오프 방법을 적용한 삼중 항 손실 함수의 도입이 SimCSE 에서 얼마나 효과적인지를 평가하였다. 기존 연구 [4]를 참고하여 English Wikipedia 데이터셋 (1M)을 통해 모델을 훈련시켰으며, 평가 데이터로는 STS 벤치마크 데이터셋 [5]을 사용하였다. 또한 [CLS] 토큰을 MLP 층에 통과시킨 것을 문장 표현 벡터로 사용했으며, 드롭아웃 비율은 0.1 로 설정하였다. 문장의 최대 길이는 32 로 고정하였으며, BERT-base 와 RoBERTa-base 에서 temperature 는 각각 0.07, 0.05 를, learning rate 는 3e-5, 1e-5 를, 배치 사이즈는 64, 512 로 설정하였다.

<표 1> 데이터 증강을 통한 SimCSE 의 실험 결과

Method	[UNK] Cutoff	MLP(s**)	STS Task Average
SimCSE BERT-base	-	-	0.7625
	√	Tanh	0.7654
	√	LeakyReLU	**0.7691**
SimCSE RoBERTa-base	-	-	0.7657
	√	Tanh	0.7765
	√	LeakyReLU	**0.7783**

<표 1>은 대조 학습에서 각 데이터 증강 기법을 다양하게 적용한 모델의 실험 결과이다. [UNK] 토큰 마스킹을 적용했을 때는 체크 표시를, 강한 변형 문장(s**)에 MLP 층 내의 활성화 함수를 활용한 노이즈를 추가했을 때는 탑재된 활성화 함수를 기록하였다. 그 외에는 두 방법이 모두 적용되지 않음을 의미한다. <표 1>에서 볼 수 있듯이, 제안하는 [UNK] 컷오프와 LeakyReLU 활성화 함수가 탑재된 MLP layer 를 활용한 노이즈를 추가하는 방법을 모두 사용했을 때, BERT-base (76.25→76.91)와 RoBERTa-base(76.57→77.83)에서 가장 효과가 우수한 것을 확인하였다. 이는 제안하는 데이터 증강 기법이 삼중 항 손실 함수와 대조 학습의 훈련 목표에 적절하게 적용되었다고 보여진다.

<표 2> 다양한 특수 토큰을 활용한 실험 결과

Method	Special Token	STS Task Average
SimCSE BERT-base	[PAD]	0.7651
	[MASK]	0.7650
	[UNK]	**0.7691**
SimCSE RoBERTa-base	[PAD]	0. 7760
	[MASK]	0.7773
	[UNK]	**0.7783**

[UNK] 토큰 이외에도 [PAD], [MASK]와 같은 다른 특수 토큰들을 활용하여 대조 학습에서의 성능 차이를 기록하였다. 이에 대한 실험 결과는 <표 2>에 정리되어 있다. 실험 결과, [UNK] 토큰이 좋은 임베딩을 학습하기에 가장 적합한 특수 토큰임을 알 수 있었다. 이는 [UNK] 토큰이 '미지의 단어'를 의미하기 때문에 데이터 증강에서 마스킹의 역할을 잘 수행했을 것이라 해석된다.

5. 결론 및 향후 연구 방향

본 논문은 문장 간의 상대적인 유사성을 학습에 도움을 주는 삼중 항 손실 함수를 위한 토큰 컷오프 방법을 제안하였다. 본 연구에서 제안된 [UNK] 토큰 컷오프 방법은 문장의 상대적 유사성을 반영하기 위해, MLP 층 내의 활성화 함수는 변형 문장의 출력에 노이즈를 주기 위해 사용되었다. 기존 대조 학습을 통한 문장 표현 학습에서 데이터 증강은 문장의 의미가 바뀔 수 있는 위험이 있어 지양되었으나, 이와 달리 본 논문에서 제안한 방법들은 우수한 성능을 보였다는 점에서 향후 관련 연구 방향의 가능성을 넓힐 것으로 예상된다. 특히, 대조 학습 방법론에서 그 유효성을 실험적으로 입증했으며, 추후 다양한 모델에 활용될 뿐만 아니라, 자연어 처리의 광범위한 응용 분야에도 활용되기를 기대한다.

감사의 글

이 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-01373, 인공지능대학원지원(한양대학교))을 받아 수행되었음.

참고문헌

[1] Zhang, Yuha, et al. "A contrastive framework for learning sentence representation from pairwise and triple-wise perspective in angular space." in ACL, 2022.

[2] Weinberger, Kilian.Q and Saul, Lawrence "Distance Metric Learning for Large Margin nearest Neighbor Classification." Journal of Machine Learning Research 10.2, 2009.

[3] Shen, Dinghan, et al. "A Simple but tough-to-beat data augmentation approach for natural language understanding and generation." arXiv:2009.13818 (2020).

[4] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple contrastive learning of sentence embeddings." in EMNLP, 2021.

[5] Cer, Daniel, et al. "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation." arXiv:1708.00055 (2017).

[6] Wei, Jason, and Kai Zou. "EDA: Easy data augmentation techniques for boosting performance on text classification tasks." in EMNLP, 2019.

[7] Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving neural machine translation models with monolingual data." arXiv: 1511.06709 (2015).

[8] Yan, Yuanmeng, et al. "ConSERT: A contrastive framework for self-supervised sentence representation transfer." arXiv:2105.11741 (2021).

[9] Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv:1810.04805 (2018).