그래프-문자 멀티모달 기반 분자 그래프 분류 과업에서 프롬프트의 효과 분석

조우성, 김민성, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

Analyzing the Effect of Prompt on Graph-Text Multimodal Based Molecule Graph Classification Task

Wooseong Cho, Minsung Kim, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

최근 거대 언어 모델 및 시각 인지 모델의 발전은 자연어 처리 과업 뿐 아니라 시각 인지 과업에도 탁월한 성능을 보여주고 있다. 사전 학습된 (Pretrained) 거대 언어 모델 전체를 미세 조정 (Finetuning)하기도 하지만, 거대 언어 모델은 하위 과업 (Downstream Task)으로의 미세 조정이 쉽지 않다. 이런 한계를 극복하기 위해 거대 언어 모델은 고정시킨 채 입력에 붙인 프롬프트 (Prompt)만을 학습시키는 연구가 진행되고 있다. 이러한 발전과 같은 흐름으로, 본 논문에서는 그래프 분류 (Graph Classification) 과업에 거대 언어 모델과 프롬프트를 사용하는 모델을 제안한다. 본 논문은 세 가지 분자 그래프 데이터 집합에서 프롬프트 개수에 따른 성능 변화와 함께 기존 모델과의 비교를 통해 거대 언어 모델 및 프롬프트를 그래프에 적용하는 것에 대한 가능성을 시사한다.

Ⅰ. 서론

최근 거대 언어 모델 및 사진 모델의 발전은 자연어 처리 과업 뿐만 아니라 사진 분류, 객체 분할, 객체 탐지 등의 시각 인지 과업에서도 높은 성능을 달성할 수 있게 해주었다[1]. 여기에는 CLIP[2]이 큰 영향을 주었다.

CLIP 은 자기 지도 학습 (Self-Supervised Learning)의 한 방법인 대조 학습 (Contrastive Learning) 기반의 모델이다. 일반적인 대조 학습과 달리 CLIP 은 멀티모달 (Multimodal)로 사진과 그에 해당하는 문자열을 긍정적인 쌍 (Positive Pair)으로, 서로 다른 사진과 문자열은 부정적인 쌍 (Negative Pair)으로 설정한다. 이때 사진과 문자열은 각각 사진 인코더 (Encoder), 문자 인코더를 통해 임베딩 (Embedding) 된다. 대조 학습 방식으로 사전 학습 (Pretrain)을 진행한 이후 입력 사진 임베딩에 대해 여러 클래스 (Class) 문자열(단어)들 중 정답 값에 대한 임베딩은 가깝게, 다른 것은 멀게 하여 사진 분류 과업에 대한 미세 조정 (Finetuning)을 진행한다. 이와 같은 방법으로 CLIP 은 당시 높은 퓨샷 (Few-Shot) 예측 성능을 달성했으며, 미세 조정 없이 진행하는 제로 샷 (Zero-Shot) 예측 성능 또한 높았다. 한편 CoOp[3]은 클래스 문자열 앞에 프롬프트 (Prompt)를 추가하여 모델 전체가 아닌 프롬프트만을 학습시켜 사진 분류 과업에서 당시 높은 성능을 달성하였다.

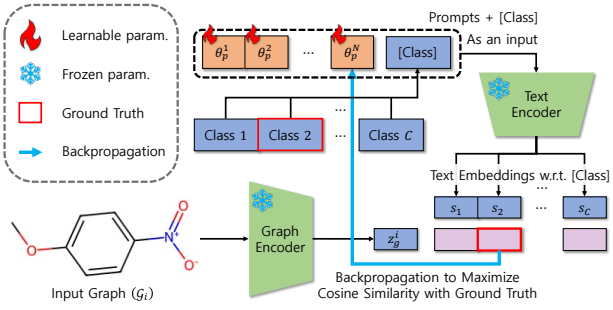

그림 1. 제안 모델 구조

이렇듯 시각 인지 과업은 거대 언어 모델과 프롬프트의 도입으로 높은 성능을 달성하였다. 하지만 그래프 과업은 화학 분야, 소셜 네트워크 (Social Network), 인용 네트워크, 생물 정보학 등 다양한 분야에서 응용됨에도 해당 연구가 미비한 상태이다[4]. 따라서 본 논문은 CoOp 과 같이 분자 그래프 과업을 위해 사전 학습된 그래프 인코더와 문자 인코더를 이용하여 입력에 붙인 프롬프트만을 학습시키는 모델을 제안한다. 우리는 OGB[5]의 세 가지 데이터 집합에서 프롬프트 개수에 따른 성능을 비교하여 그래프 과업에 거대 언어 모델과 프롬프트 도입의 효과를 보이고, 가능성을 제시하였다.

Ⅱ. 관련 연구

본 논문에서는 MoMu[6]에서 사전 학습된 그래프 인코더와 문자 인코더를 사용하였다. MoMu 는 그래프 모델(GIN[7])과 언어 모델(SciBERT[8])을 CLIP 과 같이 분자 그래프와 분자에 관련된 설명을 이용하여 멀티모달로 사전 학습한 모델이다. 해당 논문에서는 사전 학습된 그래프 모델만을 미세 조정하여 하위 과업을 진행한다. 따라서 본 논문은 그래프에서는 CoOp 에서와 같이 하위 과업을 위해 사전 학습된 그래프 인코더와 문자 인코더를 고정시키고 프롬프트만을 학습시키는 모델을 제안한다.

Ⅲ. 실험

본 논문에서 제안한 모델은 [그림 1]과 같다. MoMu 에서 사전 학습된 그래프 인코더 (Graph Encoder)와 문자 인코더 (Text Encoder)는 고정시키고 입력 클래스 문자열 ([Class]) 앞에 추가한 N 개의 프롬프트 $\left(\{\theta_p^n\}_{n=1}^N\right)$ 만을 학습시킨다. z_g^i 를 입력 그래프 $(\mathcal{G}_i)$ 에 대한 임베딩이라 하고, 전체 클래스 개수 C 에 대하여 $\{s_j\}_{j=1}^C$ 를 프롬프트가 추가된 입력 문자열에 대한 임베딩이라 하자. 이때

표 1. 데이터 집합 별 [Class]로 사용한 입력 문장

데이터 집합 / 입력 문자	Class 1	Class 2
BBBP [5]	Permeable to blood-brain barrier	Non-permeable to blood-brain barrier
HIV [5]	Active to HIV	Inactive to HIV
BACE [5]	Inhibits human beta-secretase 1	Does not inhibit human beta-secreatase

입력 그래프가 클래스 c를 가질 확률은 [수식 1]과 같이 계산된다.

$$p(y=c|\mathcal{G}_i)=\frac{\exp\left(\cos(s_c,z_g^i)/\tau\right)}{\sum_{j=1}^{C}\exp\left(\cos(s_j,z_g^i)/\tau\right)} \quad (1)$$

이때 τ는 MoMu 에서 학습된 온도 (Temperature) 변수이며 $\cos(\cdot,\cdot)$은 코사인 유사도 (Cosine Similarity)이다. 모델은 정답 (Ground Truth)에 대한 확률이 가장 높아지도록 교차 엔트로피 (Cross-Entropy) 손실 함수를 통해 학습한다. [그림 1]에서는 입력 그래프에 대해 두 번째 클래스가 정답 값인 상황을 묘사한다. 따라서 z_g^i와 s_2의 유사도가 가장 높아지도록 프롬프트가 학습된다.

하위 과업 데이터 집합은 OGB[5]의 BACE, BBBP, 그리고 HIV 의 총 세 가지 분자 이진 그래프 분류 데이터 집합을 사용하였다. BACE 는 특정 효소 활성의 저해 여부, BBBP 는 뇌혈관장벽 통과 유무, HIV 는 에이즈 바이러스 (HIV)에 대한 효과 여부를 분류한다. 이렇듯 사진과 달리 클래스 정보를 한 단어로 표현할 수 없어 각 데이터 집합에 대한 [Class]는 [표 1]과 같이 문장 형태로 정의하였다.

사전 학습된 그래프 인코더와 문자 인코더에 프롬프트를 도입했을 때의 분류 성능 (Aera Under Curve, AUC, %)은 [표 2]에 나와있다. $N=0$은 프롬프트를 전혀 사용하지 않았을 때, 즉 제로 샷 성능을 의미하며 기존 모델 (Baseline)은 MoMu 에서 그래프 인코더만을 미세 조정하여 측정한 성능을 의미한다.

대체적으로 프롬프트 개수가 증가함에 따라 성능이 높아지는 것을 확인할 수 있다. 이는 CoOp 논문에서도 보였던 것으로, 학습 가능한 매개변수 개수가 많아졌기 때문이다. 하지만 그럼에도 기존 모델보다는 성능이 낮음을 볼 수 있다. 물론 매개변수 개수의 차이가 있지만, 사전 학습에 사용한 데이터 집합에서 원인을 찾아볼 수 있다. CoOp 에서 사용한 CLIP 은 인터넷에서 수집한 사진과 그에 대한 직접적인 설명 글들을 데이터로 사용하여 사진에 대한 [Class] 단어가 직접적으로 연결되어 있을 가능성이 높다[2]. 하지만 MoMu 에서는 분자와 해당 분자가 등장한 논문을 데이터로 사용하여 그래프와 문자 간의 연결이 약하다[6]. 또한 CLIP 은 4 억 개[2]의 데이터 쌍을 수집하였지만 MoMu 는 15,613 개[6]의 데이터 쌍을 수집하였다. 따라서 사전 학습된 그래프 인코더만을 학습시켰을 때의 성능이 더 좋은 것으로 보인다.

표 2. 실험 결과

데이터 집합 / 모델	BBBP [5]	HIV[5]	BACE [5]
기존 모델[6]	70.5	75.9	76.7
$N=0$	47.6	46.9	66.6
$N=4$	59.3	66.3	64.0
$N=8$	58.9	68.4	64.2
$N=16$	60.5	72.0	65.2

또한 BACE 를 제외하고 프롬프트가 없을 때보다 있을 때 성능이 훨씬 높은 것을 확인할 수 있다. BACE 에서 프롬프트 도입으로 성능이 떨어지는 것은 데이터 집합의 불균형 (Imbalance) 때문인 것으로 보인다. 학습에 사용한 손실 함수인 교차 엔트로피는 클래스 불균형이 발생할 경우 개수가 적은 클래스에 대한 학습량이 줄어든다[9]. BACE 는 훈련 데이터 집합에는 [Class 1]이 약 40:60 으로 많고, 검증 및 테스트 데이터 집합에는 [Class 2]가 약 14:84 로 많다. 따라서 학습 시 [Class 2]에 대한 학습량이 상대적으로 감소하였고, 결과 검증 및 데이터 집합에서의 성능이 낮아지게 된 것이다. 실제로 학습 중에 훈련 데이터 집합에 대한 성능은 증가하지만 검증 데이터 집합에 대한 성능은 감소하는 현상을 관찰하였다.

IV. 결론

본 논문에서는 그래프 분류 과업에서 거대 언어 모델과 프롬프트의 도입의 효과를 연구하였다. BACE 에서는 프롬프트를 사용하지 않을 때의 성능이 더 높은 등의 한계가 있었지만, 사전 학습을 위한 충분한 데이터와 불균형 데이터 집합을 위한 적절한 방법을 도입한다면 시각 인지 모델에서처럼 성능 향상을 기대할 수 있을 것이다.

ACKNOWLEDGMENT

본 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00167194,미션 크리티컬 시스템을 위한 신뢰 가능한 인공지능)

본 연구는 2022 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음 (2022-0-00964)

참 고 문 헌

[1] Zhang, J., et al. (2023). Vision-language models for vision tasks: A survey. arXiv preprint arXiv:2304.00685.

[2] Radford, A., et al. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[3] Zhou, K., et al. (2022). Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9), 2337-2348.

[4] Jin, B., et al. (2023). Large Language Models on Graphs: A Comprehensive Survey. arXiv preprint arXiv:2312.02783.

[5] Hu, W., et al. (2020). Open graph benchmark: Datasets for machine learning on graphs. Advances in neural information processing systems, 33, 22118-22133.

[6] Su, B., et al. (2022). A molecular multimodal foundation model associating molecule graphs with natural language. arXiv preprint arXiv:2209.05481.

[7] Xu, K., et al. (2018). How powerful are graph neural networks?. arXiv preprint arXiv:1810.00826.

[8] Beltagy, I., et al. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

[9] Lin, T. Y., et al. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).