퓨샷 학습을 위한 멀티모달 읽기전용 프롬프트 최적화

서정현, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

MRPO: Multi-modal Read-only Prompt Optimization for Few-shot Learning

Junghyeon Seo, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

컴퓨터 비전 과업에서 일반적인 대규모 시각-언어 모델을 특정 과업에 특화 시키기 위한 프롬프트 러닝은 효과적인 성능을 보여왔다. 우리는 프롬프트 러닝에서 모델의 성능 향상을 위해 텍스트, 비주얼 인코더에 모두 학습가능한 프롬프트를 추가하는 멀티모달 프롬프트 방식에 입력 특징 맵의 내부 표현 이동을 방지하기 위해 인코더의 셀프 어텐션 모듈에서 마스크 어텐션을 추가한 방법을 제안한다. 우리의 방법은 11 개의 데이터셋에 대해 적은 양의 데이터만을 학습에 사용하는 퓨샷 환경의 이미지 분류 과업에서 기존 연구에 비해 향상된 평균 정확도를 달성했다.

Ⅰ. 서 론

CLIP(Contrastive Language Image Pretraining)[1]과 같은 대규모 시각-언어 모델(Large Vision Language Model)은 컴퓨터 비전의 다양한 과업에서 좋은 성능을 보여주고 있다. 그러나 적은 양의 데이터만을 사용해서 모델을 미세조정하는 경우, 기존의 대규모 사전학습 과정에서 거대 모델이 얻은 유용한 정보를 잃어버린다는 단점으로 인해 성능 저하를 유발할 수 있다[4]. 또한 텍스트 인코더(Text Encoder)의 입력으로 들어가는 알맞은 프롬프트(Prompt)를 선택하는 과정은 상당한 시간이 소요되고, 데이터 집합마다 다른 프롬프트를 선택해야 하는 번거로움이 존재한다.

이러한 문제를 해결하기 위해, 자연어처리 과업의 프롬프트 러닝에서 영감을 얻어 기존 텍스트 인코더의 프롬프트에 학습가능한 프롬프트를 추가하여 성능을 향상시킨 CoOp(Context Optimization)[2]이 등장하였다. 마찬가지로 컴퓨터 비전 분야에서도 백본 네트워크의 가중치는 고정시킨 채로 텍스트가 아닌 이미지에 적은 양의 학습가능한 프롬프트를 추가하는 VPT(Visual Prompt Tuning)[3] 방식도 등장하였다.

최신 프롬프트 러닝 연구로 텍스트 인코더에 사용되는 학습가능한 프롬프트를 별도의 선형 함수를 통해 매핑(Mapping)하여 얻은 벡터를 비주얼 인코더(Visual Encoder)의 프롬프트로 사용하여 두 모달리티(Modality) 간의 상호 시너지 (Mutual Synergy)를 개선하는 MaPLe(Multi-modal Prompt Learning)[5]이 있다. 또한 학습가능한 프롬프트로 인해 발생하는 내부 표현 이동(Internal Representation Shift)이 퓨샷(Few-shot) 환경에서 악영향을 끼치는 것을 해결하기 위해 각각의 인코더에 존재하는 셀프 어텐션 모듈(Self-attention Module)에서 마스크 어텐션(Mask Attention)을 사용하여 프롬프트는 특징 맵(Feature Map)에 대한 정보를 읽기만 가능하게 하도록 하여 원래의 특징 맵에 영향을 끼치지 않도록 하는 RPO[5] 등이 존재한다.

우리의 연구는 앞서 언급한 MaPLe[4]에 RPO[5]에서 사용한 마스크 어텐션 방법을 적용한 MRPO(Multi-modal Read-only Prompt Optimization)를 제안한다. 이를 통해 퓨샷 환경에서도 좋은 성능을 보이는 모델을 만들고자 하였고, MRPO 는 기존의 방법에 비해 높은 정확도를 달성하며 성능 향상을 보였다.

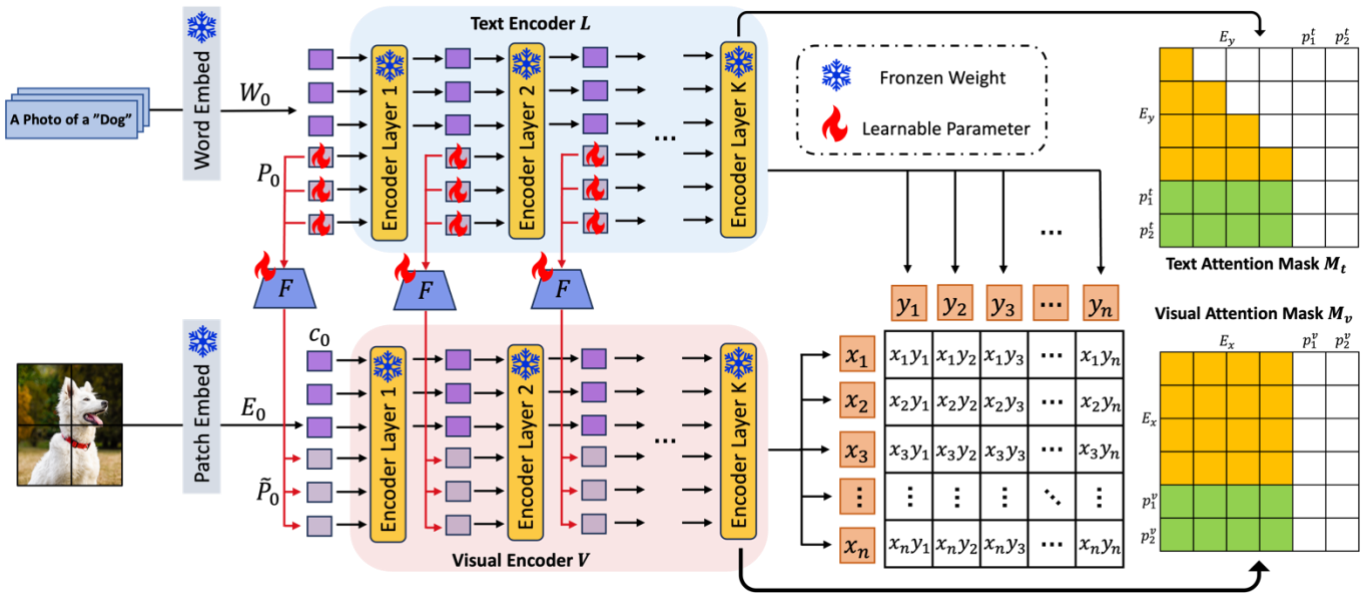

그림 1. MRPO 의 모델 구조

	Average over 11 datasets														
	16 shot			8 shot			4 shot			2 shot			1 shot		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
MaPLe[4]	81.98	**74.39**	**78**	80.07	**74.01**	76.92	77.42	73.61	75.46	74.49	**73.52**	74	72.02	**72.35**	72.19
Ours	**82.49**	73.76	77.88	**80.39**	73.79	**76.95**	**78.72**	**73.7**	**76.12**	**75.81**	72.91	**74.33**	**72.9**	72.05	**72.47**

표 1. 11 개의 데이터 집합[2]에 대한 평균 정확도

Ⅱ. 본론

[그림 1]은 모델의 전반적인 구조를 나타내고 있다. 텍스트 인코더 L 은 프롬프트 P 와 텍스트가 입력으로 들어가게 되고, 비주얼 인코더 V 는 텍스트 인코더 L 의 프롬프트 P에 선형 함수 F를 통과한 값 $\tilde{P}$을 프롬프트로 사용하여 이미지와 함께 입력으로 들어가게 된다. 이후 [그림 1]에서와 같이 모든 레이어마다 인코더의 셀프 어텐션 모듈에서 각각 마스크 M_v, M_t가 결합된 연산을 진행하게 되고, 최종적으로 레이어를 통과하여 얻은 특징 맵들 사이의 대조 학습을 통해 가장 높은 유사도를 갖는 클래스를 모델의 예측 값으로 출력하게 된다.

MRPO 의 비주얼 프롬프트 $\tilde{P}$는 [수식 1]에서와 같이 텍스트 프롬프트 P를 별도의 선형 함수 $F(\cdot)$를 통과하여 얻은 벡터를 사용한다. 이때 선형 함수는 d_l 차원의 입력을 d_v 차원의 출력으로 매핑한다. 이후 미세조정 과정에서 나머지 모델은 고정한채로 텍스트 프롬프트와 선형 함수만 학습된다. [수식 1]에서의 k 는 인코더의 레이어 순서를 의미한다.

$$\tilde{P}_k = F_k(P_k) \tag{1}$$

이때 단계별 특징 표현을 점진적으로 모델링하기 위해 더 깊은 레이어에서 프롬프트를 학습하는 것이 더 좋은 성능을 보이는 VPT[3]에서의 연구를 바탕으로 J 번째 레이어까지 학습가능한 프롬프트를 정의한다.

[수식 2]와 [수식 3]은 텍스트 인코더의 입력과 출력, [수식 4]와 [수식 5]는 비주얼 인코더의 입력과 출력을 나타낸다. W 는 워드 임베딩(Word Embedding), E 는 패치 임베딩(Patch Embedding), c 는 클래스 토큰으로 프롬프트는 임베딩 값에 결합되어 입력으로 들어간다.

$$[\_, W_i] = L_i([P_{i-1}, W_{i-1}]) \quad i = 1, 2, 3, \dots, J \tag{2}$$

$$[P_j, W_j] = L_j([P_{j-1}, W_{j-1}]) \quad j = J+1, \dots, K \tag{3}$$

$$[c_i, E_i, \_] = V_i([c_{i-1}, E_{i-1}, F_{i-1}(P_{i-1})]) \quad i = 1, 2, 3, \dots, J \tag{4}$$

$$[c_j, E_j, \tilde{P}_j] = V_j([c_{j-1}, E_{j-1}, \tilde{P}_{j-1}]) \quad j = J+1, \dots, K \tag{5}$$

학습가능한 프롬프트로 인해 발생하는 내부 표현 이동 문제를 해결하기 위해 셀프 어텐션 모듈에서 사용하는 마스크를 [수식 6], [수식 7]과 같이 정의한다. 마스크의 값은 0 혹은 $-\infty$로 설정하여 프롬프트가 학습은 되지만, 원본 특징 맵에 영향을 끼치지 못하도록 하였다.

$$M_v^{i,j} = \begin{cases} -\infty, & if\ j > 1 + N_x \\ 0, & otherwise \end{cases} \tag{6}$$

$$M_t^{i,j} = \begin{cases} -\infty, & if\ j > 1 + N_y\ or\ i > j \\ 0, & otherwise \end{cases} \tag{7}$$

각각의 인코더 내부에서의 마스크 어텐션 연산은 [수식 8], [수식 9]와 같이 기존의 셀프 어텐션 연산에 [수식 6], [수식 7]에서 정의한 마스크 값을 더한 것으로 정의한다. 이때 X 와 Y 는 셀프 어텐션 연산을 수행한 결과로, 이후에 프롬프트와 함께 결합되어 다음 레이어의 입력으로 들어가게 된다.

$$X^{(k+1)} = softmax\left(\frac{QK^T}{\sqrt{d_v}} + M_v\right) \cdot V \tag{8}$$

$$Y^{(k+1)} = softmax\left(\frac{QK^T}{\sqrt{d_t}} + M_t\right) \cdot V \tag{9}$$

Ⅲ. 실험 및 결과

실험에서 사용한 데이터 집합은 CoOp[2]에서 사용한 11 개의 이미지 분류 데이터 집합을 사용하였고, ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, DTD, EuroSAT, UCF101 로 이루어져 있다. 각각의 데이터 집합은 제로샷(Zero-shot) 설정을 위해 학습 시 사용할 기존 클래스와 사용하지 않을 새로운 클래스로 데이터 집합을 분리하였다. 모델은 퓨샷 환경에서 기존 클래스로 학습을 진행하였고, 기존 클래스와 새로운 클래스에 대해 각각 평가하여 정확도 구했다. 추가적으로 두 정확도의 조화평균을 구해 모델의 일반화 성능도 같이 평가했다.

실험 결과는 [표 1]에서와 같이 기존의 방법에 비해 우리의 방법이 여러 퓨샷 데이터 환경에서 소폭 상승한 결과를 보여주고 있고, 마스크 어텐션을 적용하여 내부 표현 이동 문제를 해결했음을 확인할 수 있다.

Ⅳ. 결론

CLIP[1]과 같은 대규모 시각-언어 모델의 등장으로 인해 여러가지 방식으로 모델을 활용하여 특정 과업에 특화 시키고자 하는 많은 연구들이 이루어지고 있다.

본 논문에서는 이러한 시각-언어 모델을 활용하여 퓨샷 환경에서도 좋은 성능을 보이는 MRPO를 제안한다. MRPO 는 기존의 멀티모달 프롬프트에서 학습가능한 프롬프트로 인해 백본 네트워크에서 발생하는 내부 표현 이동 문제를 마스크 어텐션을 적용하였고, 결과적으로 기존에 비해 여러 퓨샷 환경에서 정확도 향상을 달성할 수 있었다.

ACKNOWLEDGMENT

이 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] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[2] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." International Journal of Computer Vision 130.9 (2022): 2337-2348.

[3] Jia, Menglin, et al. "Visual prompt tuning." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[4] Khattak, Muhammad Uzair, et al. "Maple: Multi-modal prompt learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5] Lee, Dongjun, et al. "Read-only prompt optimization for vision-language few-shot learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.