CLIP 기반 색상 프롬프트 기법을 통한 제로샷 지칭 표현 이해

김성식, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

Zero-shot Referring Expression Comprehension using CLIP-based Colorful Prompt Tuning

Sungsik Kim, Jaekoo Lee*
College of Computer Science, Kookmin University.

요 약

대규모 데이터셋으로 사전 학습된 시각-언어 모델은 강력한 시각-언어 표현 능력을 학습하여, 다양한 과업을 제로샷으로 수행할 수 있다. 그 중 제로샷 지칭 표현 이해 과업에서는 비주얼 프롬프트를 추가하는 연구가 주로 진행되었다. 본 논문에서는 객체마다 고유한 색상을 표기하는 Colorful Prompt Tuning(CPT) 방식에 CLIP 모델을 적용하는 방식을 제안한다. CPT 방식에 CLIP 모델을 단순히 적용하는 것은 적절하지 않음을 보였으며, 이에 따라 서로 다른 객체 하나만을 색칠하는 방식과 크롭 방식을 합하여 개선하였다. 개선된 CLIP 기반 CPT 방식을 실험한 결과, 제로샷 지칭 표현 이해 과업에서 SOTA 방식인 Red Circle 과 성능이 우수하거나 유사함을 보였다.

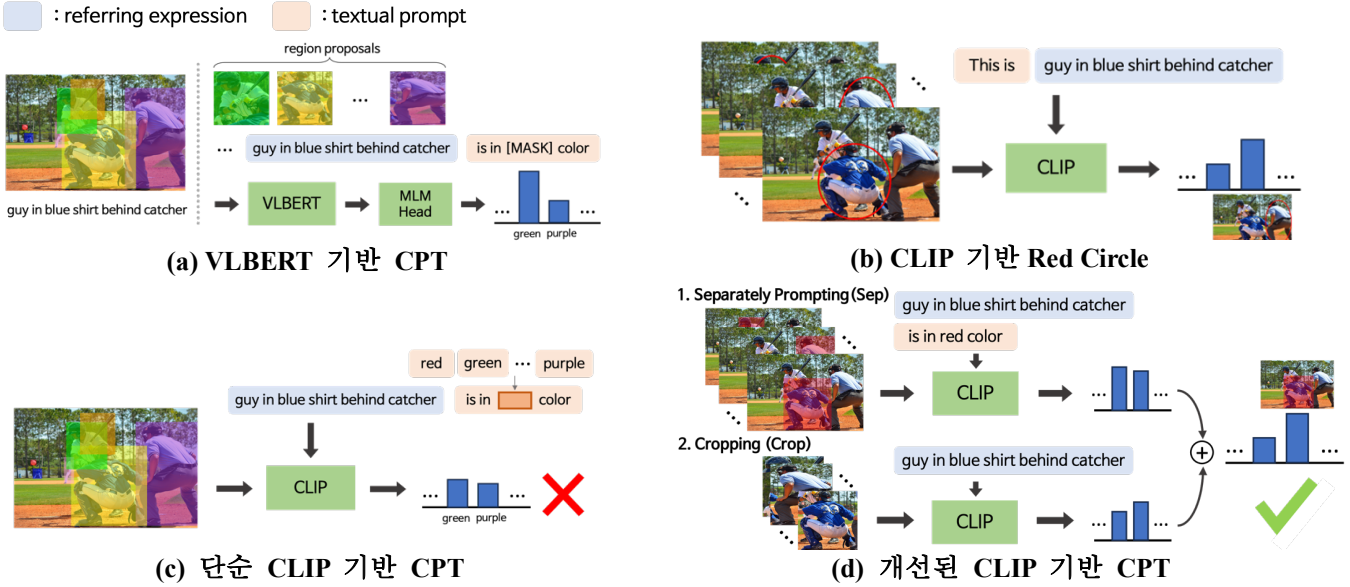
Ⅰ. 서론

최근 대형 시각-언어 모델 (Vision-Language Models, VLMs)은 시각적 질의응답과 같은 시각-언어 과업들을 크게 발전시키고 있다. 시각-언어 모델인 CLIP[1]은 거대한 데이터셋으로 이미지와 설명 문자열 임베딩이 유사해지도록 사전학습한 모델이며, 제로샷 (Zero-shot) 방식으로 이미지 분류나 언어 기반 이미지 생성 등을 수행할 수 있다. 예를 들어, "a photo of A"라는 문자열 프롬프트 (Textual Prompt)에서 A 에 원하는 클래스를 대입하고, CLIP 을 통해 문자열과 이미지의 유사도인 CLIP 스코어를 측정함으로써 제로샷 이미지 분류를 수행할 수 있다[2].

지칭 표현 이해 (Referring Expression Comprehension, REC)는 이미지와 문자열로 구성된 지칭 표현이 주어졌을 경우, 이미지 안에서 해당 지칭 표현이 가리키는 객체를 찾는 과업이다[3]. 지칭 표현 이해 과업은 로봇이나 인공지능이 카메라 입력에서 객체들을 찾는 것에서 나아가, 사용자가 원하는 특정 객체를 찾을 수 있도록 하기 때문에 중요한 기술이다[3].

하지만 시각-언어 모델을 활용하여 제로샷으로 지칭 표현 이해 과업을 수행하려 할 경우, 이미지 분류보다 훨씬 복잡하기 때문에 단순히 문자열 프롬프트를 수정하는 것으로 좋은 성능을 달성하기는 어렵다. 따라서, 이전 연구들[2, 4]에서는 이미지를 수정하는 비주얼 프롬프트 (Visual Prompt) 방식을 활용하였다.

비주얼 프롬프트 방식 중, CPT (Colorful Prompt Tuning)[4] 방식은 각 객체마다 고유한 색깔을 칠하여, 어떤 색상의 객체가 지칭 표현과 일치하는지 맞추도록 하였다. 또한, Red Circle[2]은 객체에 적색 원을 그려 해당 객체와 지칭 표현을 비교하였다. 두 방법은 모두



(a) VLBERT 기반 CPT

(b) CLIP 기반 Red Circle

(c) 단순 CLIP 기반 CPT

(d) 개선된 CLIP 기반 CPT

그림 1. 기존 제로샷 지칭 표현 이해 방식(a, b)과 제안된 CLIP 기반 CPT 방식(c, d)

비주얼 프롬프트 방법을 적용하고 있지만, 시각-언어 모델을 CPT 는 VLBERT[5], Red Circle 은 CLIP 을 적용하여 정확한 비교가 어렵다는 한계가 존재한다. 따라서, Red Circle 이 제시한 실험 결과는 Red Circle 이 CPT 보다 성능이 높았지만[2], 사용한 모델이 다르기 때문에 비교가 엄격하지 못할 수 있다. 본 논문에서는 CPT 방식에 CLIP 모델을 활용하여 기존 VLBERT 기반 방식보다 높은 성능을 달성하였고, Red Circle 방식과 유사함을 입증하였다.

Ⅱ. 본론

[그림 1]의 상단 [그림 1.(a)]와 [그림 1.(b)]는 기존 CPT 와 Red Circle 방식을 표현한 것이다. 하단 [그림 1.(c)]와 [그림 1.(d)]는 본 논문에서 제시하는 CLIP 기반 CPT 를 도식화한 것이다.

[그림 1.(a)]는 기존 VLBERT 기반 CPT 방식을 나타낸다. 이미지와 지칭 표현이 주어지면, 이미지에서 영역 제안 (Region Proposals)들을 잘라 이미지 패치들로 구성한 후, 각 패치마다 고유한 색상을 입힌다. 이후, 문자열 프롬프트의 [MASK] 위치에 있는 색상을 예측하여 제로샷 지칭 표현 이해 과업을 수행한다.

[그림 1.(b)]는 Red Circle 방식이며 이미지마다 서로 다른 객체에 적색 원을 그려 이미지와 지칭 표현 사이의 CLIP 스코어를 계산한다. 최종적으로, CLIP 스코어가 가장 높은 이미지를 선택함으로써 제로샷 지칭 표현 이해를 수행한다.

본 논문에서는 CPT 방식에 CLIP 을 적용하고자 하였으며, 그 초기 버전은 [그림 1.(c)]와 같다. 초기 버전은 한 이미지 내부 객체들에 각자 고유한 색상을 칠한 후, 문자열 프롬프트에 색상들을 바꿔가며 CLIP 스코어를 구하는 방식으로 이루어져 있다. 하지만 해당 방식은 크게 두 가지의 문제점이 존재한다. 첫번째는 [그림 1.(c)]와 같이 이미지 내부에 객체가 많을 경우, 다양한 색상들이 겹쳐 육안으로도 알아보기 힘들다는 점이다. 다음으로는 객체 위에 직접 색상을 입히는 것은 객체의 중요한 색상 정보 등을 손실할 수 있다는 점이다.

본 논문에서는 [그림 1.(d)]과 같이 두 문제점을 보완하여 개선한 CLIP 기반 CPT 방식을 제안한다. 개선된 방식은 크게 두 가지로 구성되어 있다. 첫번째로는 Red Circle 방식에 영감을 받아, 이미지마다 서로 다른 한 객체에만 색상을 입히는 방식이다. 한 객체에만 색상을 표기하는 방식을 통해, 시각-언어 모델이 해당 객체에 집중하게 하면서 글로벌한 정보를 손실하지 않도록 하였다. 하지만 해당 방식만으로는 앞서 언급한 두번째 문제점인 객체의 색상 정보 손실을 해결하지 못한다. 이를 극복하고자 크롭(Crop) 기반 방식을 추가하였다. 한 객체에만 색상을 입혀 구한 CLIP 스코어에, 동일한 객체를 크롭하여 구한 CLIP 스코어를 더해주어 최종적으로 추론을 하게 된다. 결론적으로, 시각-언어 모델이 객체에 색상을 입히는 방식을 통해 객체 간의 관계성을, 객체를 크롭하는 방식을 통해 객체의 의미론적인 정보를 포착하고자 하였다.

Ⅲ. 실험

실험을 위해, 지칭 표현 이해 과업에서 가장 많이 사용되는 RefCOCO/RefCOCO+/RefCOCOg 데이터셋을 사용하였다[2, 3, 4]. RefCOCO 와 RefCOCO+ 의 테스트 데이터셋은 "TestA"와 "TestB"로 나뉘는데 이는 각각 사람 포함 여부로 나뉜다[2]. 실험을 위해 사용된 CLIP 모델 백본은 ViT-L/14@336 이며, 영역 제안 방식은 ReCLIP[3] 방식을 따랐다. 추론된 영역과 정답 영역 사이의 IoU(Intersection over Union)가 0.5 이상일 경우 옳은 예측이며, 옳은 예측의 백분율을 평가 지표로 사용하였다.

[표 1]은 비주얼 프롬프트 방식에 따른 제로샷 지칭 표현 이해 성능 비교 결과이다. 데이터셋마다 가장 높은 성능을

방식	모델	RefCOCO			RefCOCO+			RefCOCOg	
		Val	TestA	TestB	Val	TestA	TestB	Val	Test
Random		16.0	13.4	19.3	15.8	13.9	19.6	19.2	19.1
Crop	CLIP	31.2	31.2	32.7	35.1	35.0	36.3	50.8	49.2
CPT	VLBERT	32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5
Red Circle	CLIP	38.6	**45.3**	34.0	**44.6**	**50.1**	39.4	50.6	50.0
CPT	CLIP	20.8	20.2	23.7	21.0	20.1	23.8	25.8	25.5
CPT + Sep	CLIP	31.6	36.3	30.3	32.8	35.4	30.2	34.6	34.4
CPT + Sep + Crop	CLIP	**38.8**	42.3	**36.3**	42.8	45.7	**39.7**	**52.4**	**51.7**

표 1. 비주얼 프롬프트 방식에 따른 제로샷 지칭 표현 이해 성능 비교

색상	RefCOCO			RefCOCO+			RefCOCOg	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
Red	**38.8**	**42.3**	**36.3**	**42.8**	**45.7**	**39.7**	**52.4**	**51.7**
Blue	36.5	39.7	35.7	41.2	43.7	39.0	51.0	50.9
Green	35.6	39.2	35.6	40.3	42.9	38.8	49.9	49.2

표 2. 색상 별 제로샷 지칭 표현 이해 성능 비교

굵은 글씨, 다음으로 높은 성능을 밑줄로 표현하였다. VLBERT 기반 CPT 방식은 임의 추론 방식(Random)과 영역 제안을 잘라 CLIP 스코어를 비교하는 방식(Crop)보다 큰 차이를 보이지 못했으며, Red Circle 보다 성능이 낮았다. 하지만, 단순히 CPT 에 CLIP 을 적용하는 방식은 오히려 성능 악화를 유발하였다. 본 논문에서 제시하는 서로 다른 객체에 색상을 입히는 방식 (Sep)과 크롭 방식 (Crop)을 합한 결과, Red Circle 방식보다 성능이 높거나 유사하였다. 이는 CPT 방식도 CLIP 에 맞는 적절한 기법을 적용한다면 성능이 결코 뒤떨어지지 않음을 의미한다.

또한, 제안된 방식에서 객체에 입히는 색상에 따른 성능을 비교하였다. [표 2]에서 적색으로 색상을 입히는 것이 제일 높은 성능을 보인 것을 확인할 수 있다. 이는 SOTA 방법인 Red Circle 과 동일함을 알 수 있다.

IV. 결론

본 논문에서는 CPT 방식에 CLIP 모델을 적절히 적용하는 방식을 제안하며, 제로샷 지칭 표현 이해 과업에서 SOTA 방식과 성능적으로 유사함을 보였다. 또한, CLIP 모델은 다양한 방식으로 개선되고 있고 범용성이 넓기 때문에, 원하는 상황에 맞는 CLIP 모델을 적용하여 활용할 수 있다는 장점이 있다. CPT 방식이 최근 연구에서 발표된 것과 달리 성능이 결코 뒤떨어지지 않기 때문에, 추후 더 많은 연구가 필요하다.

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)

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] Shtedritski, Aleksandar, Christian Rupprecht, and Andrea Vedaldi. "What does clip know about a red circle? visual prompt engineering for vlms." arXiv preprint arXiv:2304.06712 (2023).

[3] Subramanian, Sanjay, et al. "Reclip: A strong zero-shot baseline for referring expression comprehension." arXiv preprint arXiv:2204.05991 (2022).

[4] Yao, Yuan, et al. "Cpt: Colorful prompt tuning for pre-trained vision-language models." arXiv preprint arXiv:2109.11797 (2021).

[5] Su, Weijie, et al. "Vl-bert: Pre-training of generic visual-linguistic representations." arXiv preprint arXiv:1908.08530 (2019).