VLM(Vision-Language Model)의 구성적 추론 문제 해결 및 향상

윤경윤 [1], 조영준 [2]
1 전남대학교 인공지능융합학과 석사과정
2 전남대학교 인공지능융합학과 부교수
kyungyoon201@gmail.com, yj.cho@jnu.ac.kr

Addressing and Improving Compositional Inference in Vision-Language Model(VLM)

Kyung-Yoon Yoon[1], Yeong-Jun Cho[1]
[1]Dept. of AI Convergence, Chonnam National University

요 약

본 논문은 Vision-Language Model(VLM)의 성능을 향상시키고, 구성적 추론 문제를 해결하는 새로운 접근을 제시한다. VLM 은 시각적 정보와 언어적 정보를 결합하여 다양한 다운스트림 작업에서 뛰어난 성능을 보였지만, 여전히 이미지와 텍스트 간의 복잡한 관계를 완전히 이해하지 못하는 문제를 안고 있다. 특히, VLM 이 텍스트와 이미지의 구조적 차이를 인식하고 올바르게 매칭하는 데 한계가 있으며, 이는 주로 학습 데이터의 불균형과 손실 함수의 한계로 인해 발생한다. 이 문제를 해결하기 위해 다양한 연구들이 데이터셋과 손실 함수의 개선에 집중해 왔다. 본 논문에서는 제안하는 아키텍처는 두 가지 주요 구성 요소를 통해 문제를 해결한다. 첫 번째는 노이즈가 많은 Raw 데이터를 전처리하는 모델로, 잘못된 이미지-텍스트 쌍이나 단일 데이터를 처리하여 정제된 데이터를 출력한다. 두 번째는 하드 네거티브 데이터를 생성하여 VLM 의 구성적 추론 능력을 향상시키는 모델이다. 이를 통해 이미지와 텍스트 간의 구조적 차이를 더욱 명확히 구별할 수 있으며, 대조 학습을 통해 모델의 성능을 최적화한다.

1. 서론

VLM(Vision-Language Model)의 발전[1]은 시각과 언어 정보를 동시에 처리하는 기술을 크게 향상시켰다. VLM 은 텍스트와 이미지를 각각 임베딩하여 이미지 분류, 검색 등 다양한 작업에서 우수한 성능을 입증했다. 특히, 대규모 데이터로 학습된 VLM 은 높은 일반화 능력을 보이며, 새로운 도메인에서도 뛰어난 제로샷(Zero-shot) 성능을 발휘한다. 또한, VLM 은 단순한 객체 인식을 넘어서 이미지와 텍스트 간의 복잡한 의미적 관계를 이해하는 능력을 갖추고 있어, 다양한 다운스트림 작업에 기반 모델로 활용되고 있다. 이에 따라 VLM 의 성능을 향상시키고 구성적 능력을 강화하는 것이 중요한 과제로 대두되고 있다.

VLM 은 시각적-언어적 이해에서 탁월한 성능을 보였으나, 몇 가지 근본적인 문제를 안고 있다. 먼저, 이미지와 텍스트 매칭에서 VLM 이 보여주는 유사도는 기대보다 낮거나 불안정한 경우가 있다. 이는 주로 손실 함수가 이미지와 텍스트 간의 복잡한 관계를 충분히 포착하지 못하고, 학습 데이터가 다양한 사례를 충분히 포함하지 못하는 데서 기인한다. 이러한 한계는 VLM 의 성능을 최적화하는 데 장애가 되고 있으며, 새로운 상황에 적응하는 능력을 저하시킬 수 있다.

구성성(Compositionality)은 전체의 의미가 부분들의 의미와 관계의 함수로 결정된다는 개념으로, 인간의 인지 능력에서 핵심적인 역할을 한다. VLM 역시 이러한 능력을 필요로 하지만, 현재의 사전 학습된 모델들은 구성성 측면에서 한계를 보이고 있다. 예를 들어, "나무가 쇼핑 카트에 있다"와 "쇼핑 카트가 나무에 있다"는 동일한 단어로 구성된 문장이지만, 시각적으로는 완전히 다른 의미를 전달한다. 인간은 이 두 문장을 쉽게 구별할 수 있으나, Transformer 기반 모델들은 텍스트의 단어 순서에 덜 민감하여 이러한 차이를 정확히 인식하지 못하는 경우가 많다. 이는 VLM 이 복잡한 문장과 장면을 이해하고 해석하는 데 어려움을 겪게 한다.

또한, VLM 은 방향 인지, 계수, 시점 파악 등의 시각적 패턴에서 한계가 두드러진다. 이러한 패턴들은 이미지와 텍스트 간의 복잡한 관계를 정확히 이해하고 표현하는 데 중요한 도전 과제로 남아 있다. 따라서 VLM 의 구성적 추론 능력과 시각적 패턴 인식 능력을 강화하는 것이 필수적이다. 이를 통해 VLM 이 더 복잡한 문맥과 장면을 정확하게 해석하고, 보다 안정적인 성능을 발휘할 수 있을 것이다.

이러한 한계점을 해결하기 위해, 본 논문에서는 두 가지 주요 구성 요소로 이루어진 모델 아키텍처를 제안한다. 첫 번째는 노이즈가 많이 포함된 불완전한 입력 데이터에 대응할 수 있는 Raw 데이터 전처리 모델이며, 두 번째는 기존 VLM 의 문제를 보완하기 위한 하드 네거티브 이미지-텍스트 생성 모델이다. 제안된 모델은 다양한 입력 데이터에 유연하게 대응할 수 있도록 설계되었으며, VLM 의 성능을 향상시키고 보다 안정적인 결과를 도출할 수 있다.

2. 관련 연구

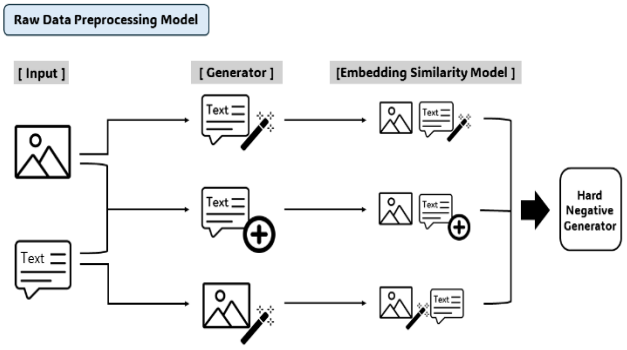
최근 VLM 에서 나타나는 구성적 추론 문제를 해결하기 위한 연구들은 주로 데이터셋과 손실 함수의 개선을 통해 성능 향상에 중점을 두고 있다. 예를 들어, Pyramid CLIP[2]은 계층적 특징 정렬을 통해 사전 학습을 수행함으로써 기존 CLIP 에 비해 성능을 향상시켰으며, Soft CLIP[3]은 레이블 스무딩(Label Smoothing) 기법을 활용하여 유사한 성과를 얻었다. 또한, Structure CLIP[4]은 Scene Graph 를 활용해 의미적으로 부정적인 예시를 생성하고, 지식 강화 인코더를 도입하여 모델의 구조적 표현력을 강화하였다. 이러한 접근들은 VLM 의 성능 개선에 기여했지만, 입력 데이터가 충분히 정제되어야 한다는 전제 조건 때문에 대규모 학습 데이터를 처리하는 데 어려움이 있다. 이로 인해 학습된 모델의 성능이 여전히 제한적이라는 한계가 남아 있다.

구성적 추론 문제와 시각적 패턴 인식 문제를 평가하기 위해 다양한 데이터셋과 벤치마크가 제안되었다. Winoground[5]는 동일한 단어 세트를 포함하되, 순서가 다른 두 개의 캡션을 사용하여 최신 비전-언어 모델의 성능을 평가하는 데이터셋이다. CREPE[6]는 모델의 체계성(Systematicity)과 생산성(Productivity)을 평가하기 위한 새로운 구성성 평가 벤치마크를 제시한다. 또한, 연구에서는 이미지와 텍스트 모두에서 하드 네거티브 샘플을 생성하는 프레임워크를 통해 모델의 성능을 개선하고자 하였다. SPEC[7]은 VLM 성능 평가를 위한 고품질 데이터 생성을 목표로 한 프레임워크로, 하드 네거티브 손실 함수(Hard Negative Contrastive Loss)를 도입하여 모델 최적화를 목표로 한다. 이러한 연구들은 VLM 모델들의 구성성과 시각적 패턴 이해 능력을 평가하고, 모델의 한계를 명확하게 드러내는 데 기여하였다. 그러나 여전히 이러한 한계점을 완벽하게 해결할 수 있는 모델은 부족하며, 구성적 추론 문제와 시각적 패턴 인식을 향상시키는 연구는 지속적인 도전 과제로 남아 있다.

3. 모델 아키텍처

본 논문에서 제안하는 모델 아키텍처는 두 가지 주요 구성 요소로 이루어져 있다. 첫 번째 모델은 노이즈가 많이 포함된 Raw 데이터를 전처리한다. 이때 고려한 노이즈는 두 가지 유형으로 잘못된 이미지-텍스트 쌍이 입력된 경우와 이미지 또는 텍스트가 단일 데이터로 입력된 경우를 포함한다. 첫 번째 모델의 출력 결과는 두 번째 모델로 전달된다.

두 번째 모델은 VLM 의 구성적 추론 문제를 해결하기 위한 하드 네거티브 데이터를 생성하는 역할을 담당한다. 이 모델은 두 가지 주요 구조로 이루어져 있다. 첫 번째는 이미지와 텍스트의 순서를 변경하는 구조로, 이를 통해 VLM 이 이미지와 텍스트의 구조를 더욱 효과적으로 이해할 수 있도록 한다. 두 번째는 구성성을 향상시키기 위해 이미지와 텍스트에 대한 하드 네거티브 데이터를 생성하는 구조이다. 생성된 하드 네거티브 데이터를 활용하여 대조적 학습을 수행함으로써, 보다 우수한 이미지 및 텍스트 임베딩을 학습할 수 있도록 설계되었다.



(그림 1) Raw 데이터 전처리 모델 아키텍처

4. Raw 데이터 전처리 모델

본 모델은 이미지-텍스트 쌍 또는 단일 이미지와 단일 텍스트를 입력으로 받아, 하드 네거티브 데이터 생성 모델이 처리하기 적합한 형태의 전처리된 이미지-텍스트 쌍을 출력한다. 노이즈가 많이 포함된 입력에 대해서 안정적인 처리를 위해 생성 모델과 유사도 검출 모델을 함께 사용한다.

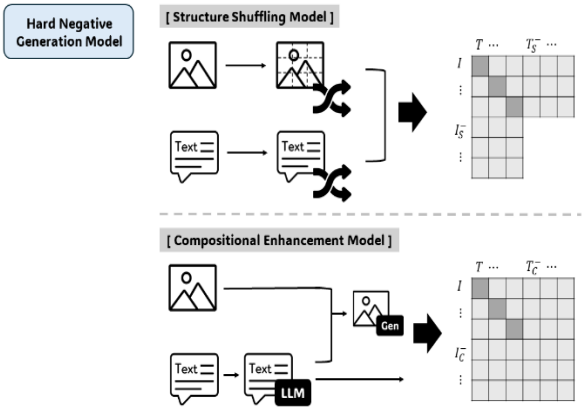
잘못된 이미지-텍스트 쌍이 입력될 경우, 기존 VLM 모델을 통해 텍스트가 이미지를 올바르게 설명할 수 있도록 내용을 수정하거나 보완한다. 최종적으로 수정된 이미지-텍스트 쌍이 생성되면, 이를 유사도 검증 모델에 입력하여 학습에 활용할지 여부를 결정한다.

단일 이미지가 입력으로 들어온 경우, VLM 모델에 해당 이미지를 입력하여 적절한 텍스트 설명을 생성한다. 이 과정에서 프롬프트 엔지니어링을 통해 가장 성능이 좋은 텍스트가 생성되도록 조정한다. 생성된

이미지-텍스트 쌍은 유사도 검증 모델을 통해 학습에 활용할지 여부가 결정된다.

단일 텍스트가 입력된 경우에는 텍스트를 반영하는 이미지를 생성하기 위해 텍스트 기반 이미지 생성기를 활용한다. 이렇게 생성된 이미지-텍스트 쌍은 유사도 검증 모델에 입력되어 학습 활용 여부가 평가된다.

유사도 검증 모델에서는 이미지를 ViT(Vision Transformer) 모델과 CNN(Convolutional Neural Network) 모델을 통해 임베딩하고, 텍스트는 Transformer 모델을 통해 임베딩한다. 생성된 이미지-텍스트 쌍의 임베딩 값을 바탕으로 VLM 모델 학습에 사용할지 여부를 결정하며, 임계값은 실험을 통해 사용자가 설정한다.



(그림 2) 하드 네거티브 생성 모델 아키텍처

5. 하드 네거티브 이미지-텍스트 생성 모델

본 모델은 이미지-텍스트 내부 구조 변경 모델과 구성성 향상 모델로 구성되어 있다. 먼저, 이미지-텍스트 내부 구조 변경 모델은 이미지와 텍스트의 내부 순서가 뒤바뀌더라도 이를 온전히 이해할 수 있는 능력을 학습하는 것을 목표로 한다. 입력으로 이미지-텍스트 쌍이 주어지면, 이미지는 N x N 크기로 분할하여 순서를 무작위로 섞고, 텍스트는 트라이그램을 생성하여 이들의 순서를 무작위로 뒤섞는다. 여기서 트라이그램은 연속된 세 개의 단어나 문자로 이루어진 그룹을 의미한다. 이렇게 순서가 뒤바뀐 이미지와 텍스트는 각각 I_S^- 및 T_S^- 로 표현된다. 이에 대해 VLM 은 대조 학습(Contrastive Learning)을 수행한다.

구성성 향상 모델은 이미지-텍스트 쌍이 입력되면, 먼저 텍스트를 대형 언어 모델(LLM, Large Language Model)에 입력하여 색상, 수량, 위치, 동사, 긍정/부정 등의 요소를 변환한 하드 네거티브 텍스트를 생성한다. 이때, 효과적인 텍스트 생성을 위해 프롬프트 엔지니어링 기법을 활용한다. 이후 원본 이미지와 생성된 하드 네거티브 텍스트를 텍스트 기반 이미지 생성 모델에 입력하여, 이미지를 텍스트에 맞게 수정 및 생성한다. 이렇게 생성된 하드 네거티브 이미지와 텍스트는 각각 I_C^- 및 T_C^- 로 표현된다. 이에 대해 VLM 은 대조 학습을 수행하여, 모델이 하드 네거티브 데이터를 잘 구분할 수 있도록 학습한다.

최종적으로, 유사한 이미지-텍스트 쌍에 대해서는 가까운 임베딩 값을, 다른 이미지-텍스트 쌍에 대해서는 먼 임베딩 값을 가지도록 학습된 VLM 모델이 생성된다. 이를 통해 기존 모델에 비해 구조와 구성성이 향상된 결과를 확인할 수 있다.

6. 적용 사례

본 논문에서 제안된 VLM 아키텍처는 다양한 응용 분야에서 성능을 검증할 수 있다.

첫 번째 적용 사례는 이미지 검색 시스템에서의 활용이다. 노이즈가 많이 포함된 이미지-텍스트 데이터를 전처리하는 모델을 통해, 잘못된 이미지-텍스트 쌍이나 단일 이미지 또는 텍스트의 입력 문제를 해결할 수 있다. 이를 통해 사용자는 텍스트를 입력하여 이미지 검색을 수행하거나, 이미지를 입력해 관련된 텍스트를 검색할 때, 정확한 결과를 얻을 수 있다. 특히, 전처리 과정을 통해 정제된 이미지-텍스트 쌍은 VLM 모델의 성능을 극대화하여 다양한 도메인에서의 검색 정확도를 높일 수 있다.

두 번째 적용 사례는 이미지 설명 생성 분야이다. VLM 아키텍처는 단일 이미지가 입력되었을 때, 해당 이미지를 적절하게 설명하는 텍스트를 생성하는 데 사용될 수 있다. 이를 위해 대형 언어 모델(LLM)을 활용하여 프롬프트 엔지니어링을 통해 최적의 텍스트 설명을 생성한다. 이 방법은 이미지 캡셔닝 시스템이나 이미지 분석을 위한 자연어 설명 생성 작업에서 큰 이점을 제공하며, 생성된 이미지-텍스트 쌍을 대조 학습을 통해 학습하여 보다 높은 품질의 설명을 가능하게 한다.

세 번째 적용 사례는 구성적 추론 능력을 요구하는 분야이다. 예를 들어, 복잡한 장면을 설명하거나 텍스트와 이미지의 순서를 변화시키는 작업에서는 구성적 추론 능력이 필수적이다. 본 논문에서 제안한 하드 네거티브 데이터 생성 모델은 이미지와 텍스트의 내부 순서가 뒤바뀌더라도 이를 정확하게 이해할 수 있는 능력을 학습하며, 이를 통해 모델이 다양한 이미지-텍스트 조합에서 정확한 의미를 추론할 수 있도록 한다.

7. 결론

본 연구에서 제안된 VLM 아키텍처는 이미지와 텍스트 데이터의 복잡한 구성적 관계를 효과적으로 처리하기 위한 새로운 접근을 제시하였다. 특히, Raw 데이터를 전처리하는 과정에서 노이즈를 줄이고, 하드 네거티브 데이터를 생성하여 대조 학습을 수행함으로써, VLM 의 구성적 추론 능력과 시각적 패턴 인식 능력을 크게 향상시킬 수 있음을 보여주었다. 이를 통해 다양한 이미지 검색, 이미지 설명 생성, 복잡한 장면 분석 등의 응용 분야에서 성능 향상이 기대된다.

본 연구에서 제시한 아키텍처는 여전히 몇 가지 한계점을 가지고 있다. 첫째, 대규모 학습 데이터를 처

리하는 과정에서 실용적인 적용이 중요한 도전 과제로 남아 있다. 또한, VLM 의 구성적 추론 능력을 강화하기 위해서는 보다 다양한 구성적 추론 작업에 대한 추가 연구가 필요하다. 향후 연구에서는 이러한 한계를 극복하기 위한 효율적인 데이터 처리 방법과 학습 전략을 제안하여, VLM 의 성능을 최적화하고 다양한 도메인에서의 일반화 능력을 더욱 향상시키는 데 중점을 둘 필요가 있다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었으며(IITP-2023-RS-2023-00256629), 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2022-00165919).

참고문헌

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

[2] Gao, Yuting, et al. "Pyramidclip: Hierarchical feature alignment for vision-language model pretraining." *Advances in neural information processing systems* 35 (2022): 35959-35970.

[3] Gao, Yuting, et al. "Softclip: Softer cross-modal alignment makes clip stronger." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 3. 2024.

[4] Huang, Yufeng, et al. "Structure-clip: Enhance multi-modal language representations with structure knowledge." *arXiv preprint arXiv:2305.06152* 2.3 (2023).

[5] Thrush, Tristan, et al. "Winoground: Probing vision and language models for visio-linguistic compositionality." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[6] Ma, Zixian, et al. "Crepe: Can vision-language foundation models reason compositionally?." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[7] Peng, Wujian, et al. "Synthesize Diagnose and Optimize: Towards Fine-Grained Vision-Language Understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.