Vision-Language Pre-training 모델에서의 전이 가능성 기반 적대적 공격 연구 동향

김용재*, 송민혁**, 이연준***
*한양대학교 ERICA 컴퓨터학부(학부생), **한양대학교 정보보호학과(대학원생),
***한양대학교 컴퓨터공학과(교수)
*likethestars@hanyang.ac.kr, **thd1008@hanyang.ac.kr, ***yeonjoonlee@hanyang.ac.kr

A Survey on Transferability-based Adversarial Attack in Vision-Language Pre-training Models

Yongjae Kim*, Minhyeok Song**, Yeonjoon Lee***
*Hanyang University ERICA, Major in Computer Science (Undergraduate Student)
**Hanyang University, Department of Computer and Information Security (Master Student)
***Hanyang University, Computer Science and Engineering (Associate Professor)

요 약

Vision-Language Pre-training(VLP) 모델에 대한 적대적 공격 연구는 증가하는 추세다. 초기에는 구조적 복잡성과 모달리티 간 상호작용으로 인해 효과적인 공격이 어려웠으나, 최근 연구들은 전이 가능성을 활용하여 Black-box 상황에서도 성공적인 공격을 달성하였다. 특히 다양한 구조의 VLP 모델을 넘나들며 전이성을 확보하려는 시도가 이어지고 있으며, 대규모 비전-언어 모델(LVLM)까지 공격 가능성이 확장되고 있다. 이에 본 논문은 VLP 모델의 구조와 전이 가능성 기반 공격 기법들을 정리하고, 향후 연구 방향을 제시한다.

Ⅰ. 서론

최근 CLIP[1], ALBEF[2], BLIP[3] 등 대규모 Vision-Language Pre-training (VLP) 모델의 등장으로, 이미지-텍스트 검색 (image-text retrieval), 시각적 추론 (visual entailment), 비주얼 그라운딩 (visual grounding) 등 다양한 멀티모달 과제에서 눈에 띄는 성능 향상이 이루어졌다. 이러한 성능 향상은 대규모 데이터와 트랜스포머 기반 아키텍처의 결합에 기인하며, VLP 모델은 실제 서비스와 애플리케이션에도 빠르게 확산되고 있다.

그러나 딥러닝 기반의 다른 기술들과 마찬가지로 VLP 모델 역시 적대적 공격에 대한 견고성이 부족하다. 특히 실제 배포 환경에서는 모델 내부 구조나 파라미터에 접근하기 어려워 white-box 가정이 성립하지 않는 경우가 많으므로, black-box 상황에서의 전이 가능성 (transferability)이 훨씬 더 현실적이고 위협적인 공격 방식으로 주목받고 있다.

따라서 본 논문에서는 멀티모달 상호작용을 적극적으로 활용하거나, 모달리티 일관성/불일치 (modality consistency/discrepancy) 특성을 조정하는 등 전이가능성을 높이기 위한 다양한 공격 연구들을 VLP 모델의 구조와 연관 지어 종합적으로 정리하고, 향후 연구 방향을 제시한다.

Ⅱ. 배경

VLP 모델은 두 가지 구조로 구분된다. Aligned 구조는 이미지 인코더와 텍스트 인코더가 독립적으로 작동하며, 동일한 임베딩 공간 (embedding space)에서 각 모달리티의 특징 (feature)을 정렬하도록 학습된다. 대표적 예인 CLIP [1]은 대규모 웹 데이터셋을 활용해 대조학습 (contrastive learning) 기반의 이미지-텍스트 정렬을 달성한다. 이 구조는 대규모 학습에 효율적이고 이미지-텍스트 검색에서 우수한 성능을 보인다. 그러나 모델 내부에서 교차 모달 상호작용 (cross-modal interaction)이 직접적으로 일어나지 않으므로, 정교한 멀티모달 추론이나 모달리티 간 깊은 상호작용이 필요한 과제에는 한계가 있다. 이러한 특성은 적대적 공격에서도 취약 지점으로 작용하며, 임베딩 공간 자체를 교란하는 방식이 주요 공격 전략이 된다.

Fused 구조는 이미지와 텍스트를 각각 인코더로 처리한 뒤 멀티모달 인코더 (multimodal encoder)에서 융합하여 하나의 통합 표현을 학습한다. ALBEF [2], BLIP [3], TCL [4] 등이 여기에 속한다. Fused 구조는 self-attention 과 cross-attention 을 통해 이미지와 텍스트 간 상호작용을 적극 활용하므로, 시각적 추론 (visual entailment)이나 비주얼 그라운딩 (visual grounding) 등 복잡한 과제에서 강점을 지닌다. 다만 연산 비용이 크고, 이미지-텍스트 검색 기반 과제에서는 Aligned 구조보다 효율성이 낮을 수 있다. 공격 관점에서는 교차 모달 상호작용을 교란할 수 있는 지점이 다양하여 보다 정교한 공격 기법 설계가 가능하다.

Ⅲ. 전이 가능성 기반 공격

VLP 모델은 크게 Aligned 구조와 Fused 구조로 구분되지만, 실제로는 모델마다 세부 아키텍처, 학습 방식, 모달 상호작용 메커니즘이 상이하다. 일부는 임베딩 공간 정렬을 중심으로, 다른 일부는 cross-attention 을 통해 모달 상호작용을 활용한다. 이러한 다양성 때문에 특정 모델에만 특화된 white-box 공격은 실제 배포 환경을 충분히 설명하기 어렵다. 현실적

시나리오에서는 서로 다른 구조의 VLP 모델에도 효과적인 전이 가능성이 핵심 요건으로 요구된다.

초기 연구 중 하나인 X. Xu et al.[5]는 이미지와 텍스트 모달리티를 독립적으로 혹은 단순 병합해 공격했을 뿐, 두 모달리티가 결합하여 형성하는 공유 임베딩 공간 자체의 교란·최적화는 고려하지 못했다. 이에 따라 초기 멀티모달 공격인 Co-attack [6]은 이미지와 텍스트를 순차적으로 공격해 멀티모달 상호작용을 부분적으로 고려했다는 의의가 있다. 그러나 단일 쌍 (single-pair)에 국한된 교차 모달 상호작용만을 다루었기에, white-box 에서는 강력하더라도 black-box 에서는 전이 가능성이 크게 제한되었다.

이 한계를 보완한 Set-level Guidance Attack (SGA) [7]은 단일 쌍 의존을 넘어서 세트 수준 (set-level)의 다양한 image- text pairs 를 활용해 교차 모달 상호작용을 확장하였다. 동일 이미지에 대해 여러 캡션을 고려하거나, 이미지의 다양한 스케일 변환을 도입함으로써 정렬을 유지하면서도 더 풍부한 상호작용 정보를 공격에 활용했다. 특히 cross-attention 기반 상호작용이 강한 Fused 구조에서 전이 가능성을 유의미하게 향상시켰으며, Aligned 구조에서도 기존 대비 개선을 보였으나 효과는 상대적으로 제한적이었다.

최근 제안된 Transferable Multi-Modal (TMM) Attack [8]은 set-level 확장을 넘어, VLP 모델의 핵심 특성인 모달리티 일관성 특징과 모달리티 불일치 특징을 동시에 조정하는 전략을 도입하였다. 우선 어텐션 지향 특징 교란 (attention-directed feature perturbation)으로 여러 VLP 모델에 공통적인 일관적 특징을 교란하고, 직교 지향 특징 이질화 (orthogonal-guided feature heterogenization)로 모달 간 차별적 특징을 증폭시킨다. 이 접근은 Fused 구조에서 전이 가능성을 극대화하는 데 효과적이었고, Aligned 구조에서도 전이를 확보했다. 더 나아가 TMM 은 대규모 비전- 언어 모델 (LVLM; BLIP-2, LLaVA, GPT-4V 등)에서도 black-box 공격 성능을 입증하며, 구조 차이를 넘어서는 전이 가능한 공격 가능성을 보여주었다.

IV. 결론

본 논문에서는 VLP 모델의 구조를 정리하고, Co-attack, SGA, TMM 으로 이어지는 일련의 공격 연구를 비교함으로써 각 기법이 전이 가능성에 어떤 향상을 가져왔는지 논의하였다. Co-attack 은 멀티모달 협력적 공격의 가능성을 최초로 제시했으나 전이 가능성 확보에는 한계가 있었다. SGA 는 세트 수준에서 다변화된 가이던스를 활용하여 black-box 환경에서 전이 가능성을 개선하였다. 마지막으로 TMM 은 모달리티 일관성 특징과 모달리티 불일치 특징을 동시에 조정하는 정교한 전략을 통해 다양한 VLP 모델은 물론 대규모 비전-언어 모델 에서도 강력한 전이 가능성을 달성하였다.

향후 연구는 다음과 같이 확장될 수 있다. 첫째, LVLM 을 대상으로 한 체계적 전이 가능성 연구가 필요하다. 실세계 활용이 빠르게 증가하는 모델들에 맞춘 공격 및 방어 기법의 표준화된 벤치마크와 프로토콜이 요구된다. 둘째, Aligned 구조와 Fused 구조의 차이를 넘어서는 범용 전이 가능한 공격 기법 설계가 필요하다. 현재 기법들은 특정 구조에서 효과가 두드러지는 경향이 있으므로, 구조적 차이에 구애받지 않는 보편적 전략이 요구된다. 셋째, 단일 모달 환경에서의 텍스트 의미 보존과 이미지 은밀성 문제는 멀티모달 결합 상황에서 더욱 복잡해진다. 생성된 적대적 텍스트가 원문과 의미적으로 불일치하거나, 이미지의 미세 변조가 쉽게 눈에 띄는 경우 공격의 자연스러움과 설득력이 저하될 수 있다. 따라서 향후에는 이미지-텍스트 양쪽의 교란을 균형 있게 설계하여 의미론적 정합성과 시각적 은밀성을 동시에 보장하는 균형 잡힌 멀티모달 공격 기법이 요구된다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2024-00438056)

참 고 문 헌

[1] A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," Proc. 38th Int. Conf. on Machine Learning (ICML), pp. 8748- 8763, 2021.

[2] J. Li et al., "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[3] J. Li et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Proc. 39th Int. Conf. on Machine Learning (ICML), pp. 12888- 12900, 2022.

[4] J. Yang al., "Vision-Language Pre-Training with Triple Contrastive Learning," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10647- 10657, 2022.

[5] X. Xu et al., "Fooling Vision and Language Models Despite Localization and Attention Mechanism," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4957- 4967, 2018.

[6] J. Zhang et al., "Towards Adversarial Attack on Vision-Language Pre-training Models," Proc. 30th ACM Int. Conf. on Multimedia (ACM MM), pp. 5428- 5436, 2022.

[7] D. Lu et al., "Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models," Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1234- 1244, 2023.

[8] H. Wang et al., "Transferable Multimodal Attack on Vision-Language Pre-training Models," in Proc. IEEE Symposium on Security and Privacy (S&P), pp. 1-17, 2024.