시각 언어 모델에 대한 프롬프트를 이용한 적대적 공격

송승헌, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

Prompts-Based Adversarial Attacks on Vision Language Models

Seungheon Song, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

강력한 시각 언어 모델인 CLIP 의 등장으로 시각 언어 모델은 제로샷 영상 분류, 제로샷 객체 검출 등 다양한 과업에서 성공을 보여왔다. 더불어 텍스트 인코더 입력에 프롬프트를 결합하여 시각 언어 모델의 성능을 향상시키는 연구가 진행되었다. 또한 시각 언어 모델에 대한 적대적 강건성 연구는 실제 환경에서 안전한 시각 언어 모델을 배포하기 위해 필수적이다. 본 논문에서는 시각 언어 모델의 적대적 강건성을 조사하고, 안정적인 비전 언어 모델 개발에 기여할 수 있는 프롬프트 기반 공격 방법을 제안한다. 결과적으로 제안한 프롬프트 기반 적대적 공격을 통해 CIFAR10 및 CIFAR100 데이터 세트에서 적대적 공격 성능이 향상됨을 확인하였다.

Ⅰ. 서 론

시각 언어 모델 (Vision-Language Model)은 딥러닝 (Deep Learning)의 성공과 함께 많은 관심을 받아왔다. 특히, CLIP[1]의 등장 이후 제로샷(Zero Shot) 영상 분류, 제로샷 객체 검출 등 다양한 과업에서 큰 성공을 보여주었다. 또한, 시각 언어 과업에서 텍스트 인코더 (Text Encoder)의 입력에 프롬프트 (Prompt)를 추가하여 모델의 성능을 올리고자 하는 연구 또한 이루어지고있다[2].

더불어 시각 언어 모델에 대한 적대적 강건성 (Adversarial Robustness) 연구는 실제 환경에서 외부의 공격으로부터 안전한 시각 언어 모델을 배포하기 위해 필수적이다. 하지만 프롬프트에 따른 시각 언어 모델의 적대적 공격에 대한 연구는 거의 이루어지고 있지 않다. 따라서 본 논문에서는 시각 언어 모델의 적대적 공격을 연구하여 적대적인 공격으로부터 안전한 시각 언어 모델 개발에 기여하기 위해 프롬프트를 이용한 공격 방법을 제안하고, 프롬프트에 따른 적대적 공격의 성능을 조사한다.

Ⅱ. 본론

일반적인 시각모델에 대한 적대적 공격은 FGSM[3] 이 이용된다. FGSM[3]을 수식으로 표현하면 다음과 같다.

$$X_{t\_adv} = X + \epsilon \cdot \mathrm{sign}(\nabla_X J(X, Y_{True})) \quad (1)$$

[수식 1]과 같이FGSM[3]은 입력 X와 정답 값 Y_{True}에 대한 손실 함수 J를 거쳐 나온 손실으로부터 입력 영상에 변화도를 역전파하여 손실 값이 최대화 되는 방향으로 입력 영상을 ϵ 만큼 더하여 적대적 예제 X_{t_adv} 를 생성한다.

CLIP[1]을 활용하여 분류 과업을 수행할 경우 시각 임베딩 (Embedding)과 언어 임베딩을 일치시키는 방식으로 가장 의미론적으로 유사한 언어 정답 값을 찾아 분류 과업을 수행한다. 시각 언어 모델에서는 이러한 특징을 이용하여 FGSM[3]으로 적대적 예제 생성할 때, 입력 영상에 대해 정답 값만 못 맞추게 하는 공격이 아닌 영상의 객체를 의미론적으로 다르게 보게 하는 의미론적 적대적 공격을 수행할 수 있다.

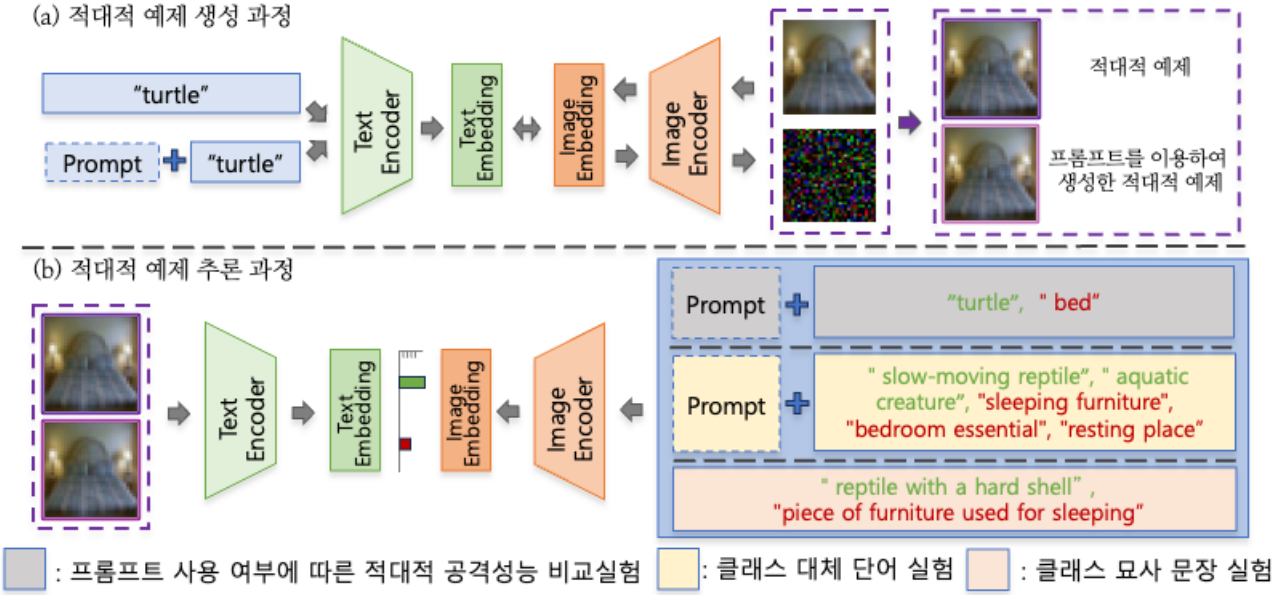

그림 1. 적대적 예제 생성과정과 적대적 예제 추론과정

실험종류	프롬프트 사용 여부에 따른 적대적 공격성능 비교실험		클래스 대체 단어 실험		클래스 묘사 문장 실험	
데이터셋	CIFAR10[5]	CIFAR100[5]	CIFAR10[5]	CIFAR100[5]	CIFAR10[5]	CIFAR100[5]
Base	85.93	82.67	63.52	25.24	45.50	29.24
Prompt	**91.69**	**83.89**	**68.30**	**26.85**	**53.38**	**34.83**

표 1. 프롬프트 사용 여부에 따른 적대적 공격성능 비교실험, 클래스 대체 단어실험 및 클래스 묘사 문장 실험에 대한 적대적 공격 정확도(Base 와 Prompt 는 적대적 예제 생성시의 프롬프트 사용 유무를 나타낸다.)

본 논문에서는 시각 언어 모델에 프롬프트를 이용하여 적대적 예제를 만드는 공격방식을 제안하고, 클래스 대체 단어 (Class Alternative Word)와 클래스 묘사 문장 (Class Descriptive sentence)을 통해 일반화된 상황에서도 프롬프트 이용한 공격이 더 강한 적대적 공격이 가능함을 확인한다.

구체적으로 정답 값을 맞추지 못하도록 하는 일반적인 FGSM[3]방식과 달리 [그림 1.(a)]에서와 같이 프롬프트를 사용하고, 정답 값이 아닌 무작위 클래스를 추출하여 추출된 정답 값을 맞추도록 갱신한다. 이를 수식으로 표현하면 다음과 같다.

$$X_{f\_adv} = X - \epsilon \cdot \text{sign}(\nabla_X J(X, Y_{Fake})) \quad (2)$$

[수식 2]에서 Y_{Fake} 는 무작위로 추출된 정답 값을 의미하며, 일반적인 FGSM[3]과 달리 입력영상 X 는 Y_{Fake}에 대한 손실을 최소화 하는 방향으로 ϵ 만큼 감하여 X_{f_adv}를 생성한다.

또한 프롬프트를 텍스트 인코더의 입력에 추가하여 예측을 수행하고, 의미론적 적대적 공격이 수행되었는지 확인하기 위해 언어 정답 값들을 클래스 대체 단어 및 클래스 묘사 문장으로 대체한다.

적대적 공격의 성능을 측정하기 위해 의도적으로 대체한 언어 정답 값에 대한 분류 정확도를 적대적 공격 정확도로 정의하고 성능을 측정한다.

Ⅲ. 실험

프롬프트를 이용한 의미론적 적대적 공격에 대한 성능을 측정 하기 위해 다음의 3 가지 실험 (1)프롬프트 사용 여부에 따른 적대적 공격성능 비교실험, (2)클래스 대체 단어 실험, (3)클래스 묘사 문장 실험을 설계한다.

클래스 대체 단어 실험과 클래스 묘사 문장 실험의 경우 언어 정답 값을 다른 값으로 대체하여 실험한다. 예를 들어 [그림 1.(b)]에서와 같이 "turtle"에 해당하는 언어 정답 값은 클래스 대체 단어 실험에서 "slow-moving reptile", "aquatic creature"와 같은 다수의 단어 중 하나로 대체되고 클래스 묘사 문장 실험에서는 "reptile with a hard shell"와 같은 문장의 형태로 대체된다. 언어 정답 값을 대체할 클래스 대체 단어와 클래스 대체 문장은 각각의 단어들에 대한 설명이 겹치지 않도록 OpenAI 의 ChatGPT[4]를 이용하여 생성하였다.

프롬프트사용 유무에 따른 적대적 공격 성능 비교 실험에서는 [그림 1.(a)]와 같이 프롬프트를 사용 유무에 따라 적대적 예제를 생성하고, 적대적 예제에 대해 추론단계에서 [그림 1.(b)]와 같이 프롬프트를 사용하였을 때와 사용하지 않았을 때의 적대적 공격 정확도를 비교한다. 프롬프트를 사용하여 적대적 예제를 만들고 적대적 공격 정확도를 평가하였을 때, [표 1]에서와 같이 프롬프트를 사용하지 않고 적대적 예제를 만들었을 때보다 평균이 CIFAR10[5] 데이터 세트 에서 5.76% , CIFAR100[5] 데이터셋에서 1.22% 더 높은 것을 확인할 수 있었다. 이러한 결과는 프롬프트를 이용하여 좀 더 정확한 적대적 예제를 생성하여 공격할 수 있음을 보여준다.

클래스 대체 단어 실험에서는 [그림 1.(b)]와 같이 각 언어 정답 값을 대신할 수 있는 단어들로 언어 정답 값을 대체하여 실험한다. 프롬프트를 이용 여부에 따라 두가지 적대적 예제를 생성하고 적대적 예제에 따른 적대적 공격의 성능을 비교하였을 때 [표 1]에서와 같이 프롬프트를 이용한 적대적 공격성능이 CIFAR10[2]에서 4.78%, CIFAR100[2]에서 1.61% 더 높은 것을 확인할 수 있었다.

클래스 묘사 문장 실험에서는 [그림 1.(b)]와 같이 각 언어 정답 값을 묘사하는 문장으로 언어 정답 값을 대체하여 실험한다. 클래스 대체 단어 실험과 마찬가지로 적대적 예제를 생성할 때 프롬프트 사용 여부에 따라 두가지의 예제를 만들어 적대적 공격 성능을 비교한다. 하지만 클래스 대체 단어 실험과 달리 대체되는 언어 정답 값이 문장이기 때문에 추론 시에는 프롬프트를 사용하지 않는다. 적대적 예제 생성시 프롬프트 사용 여부에 따라서 묘사 문장에 대한 적대적 공격 정확도를 비교하였을 때 [표 1]와 같이 프롬프트를 이용하여 적대적 공격을 생성했을 때 적대적 공격성능이 CIFAR10[5]에서 7.98%, CIFAR100[2]에서 5.59% 더 높은 것을 확인할 수 있었다.

클래스 대체 단어 실험과 클래스 묘사 문장 실험에서에서 원래의 언어 정답 값이 아닌 유사한 의미를 갖는 대체 단어 및 묘사 문장으로 언어 정답 값을 변경했음에도 프롬프트를 이용한 공격 성능이 증가함을 볼 수 있었다. 이는 프롬프트를 이용한 공격이 단순히 해당하는 언어 정답 값에만 공격이 된 것이 아니라, 의미론적 공격에도 좋은 성능을 보인다는 것을 의미한다

Ⅳ. 결론

본 논문에서는 프롬프트를 이용하여 시각 언어 모델에 대한 적대적 공격을 제안하였으며 클래스 대체 단어와 클래스 묘사 문장에 대한 적대적 공격성능을 분석하였다. 결과적으로 프롬프트를 사용한 적대적 공격이 프롬프트를 사용하지 않았을 때보다 CIFAR10[5]과 CIFAR100[5]에 대해서 높은 성능을 보여주었다. 추가로 의미론적 공격 성능을 확인하기 위해 클래스 대체 단어 실험과 클래스 묘사 문장 실험에서도 프롬프트를 사용하여 적대적 공격을 하였을 때 더 높은 적대적 공격 정확도를 달성함을 확인하였다.

또한 본 논문의 실험은 CLIP[1]과 같은 시각 언어 모델이 프롬프트를 사용한 적대적 공격에 취약할 수 있음을 보여주었다. 이는 시각 언어 모델의 많은 성공에도 여전히 적대적 공격 위험에 노출되어 있다는 것을 시사한다. 본 논문은 시각 언어 모델이 실제 세계에 배포되기 위해 보안을 강화할 수 있는 새로운 가능성을 보여준다.

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, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[2] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9), 2337-2348.

[3] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[4] OpenAI. (2024). ChatGPT (Jan 1 version) [Large language model]. https://chat.openai.com/chat

[5] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.