강화학습과 이미지 패칭을 통한 신뢰도 독립적 적대적 공격

고대화, 김재현, 정재훈*
한국항공대학교

daehwa001210@gmail.com, kjh990127@kau.kr, jhjung@kau.ac.kr

Confidence-Independent Adversarial Attacks Via RL with Image patching

Ko Dae Hwa, Kim Jae Hyun, Jung Jay Hoon
Korea Aerospace Univ.

요 약

본 연구는 신뢰도 점수(Confidence score)에 의존하지 않는 새로운 강화학습(RL)을 이용한 적대적 공격 방법론을 제안한다. 이 방법은 강화 학습(RL)을 기반으로 하여 이미지의 패치들을 조작함으로써 공격을 수행한다. 특히, Vision Transformer (ViT)의 패칭을 이용하여 각 이미지 패치에 대한 공격 전략을 최적화한다. 이 접근 방식은 기존 적대적 공격 방법과 비교하여 실제 환경에서의 적용 가능성을 높이며, 적대적 예제 생성의 효과를 극대화한다.

Ⅰ. 서 론

최근 인공지능 시스템, 특히 이미지 분류 모델의 적대적 공격에 대한 취약성이 크게 부각되었다. 이러한 공격은 모델의 예측을 오도하기 위해 입력 데이터에 소량의 변형을 가하는 기법이다. 전통적인 적대적 공격 방법들은 주로 신뢰도 점수를 활용하거나 모델의 내부 구조에 대한 상세한 정보를 필요로 하였다. 하지만, 이러한 정보에 접근할 수 없는 '블랙박스' 상황에서의 공격 가능성에 대한 연구는 상대적으로 제한적이었다. 본 연구는 강화 학습(RL)을 활용하여, 신뢰도 점수에 의존하지 않고도 효과적인 적대적 예제를 생성할 수 있는 새로운 방법론을 제안한다. 이 방법은 이미지 패치를 활용하여, 공격의 정밀도를 높이고, 다양한 머신러닝 모델에 대한 적응성을 향상시킨다. 본 논문은 이 새로운 접근 방식의 개념적 기반과 이론적 배경, 그리고 실제 적용 결과에 대해 논의한다.

Ⅱ. 본론

본 연구에서 제안하는 모델은 강화학습의 강력한 성능을 순수 블랙박스 적대적 공격에 적용하기 위해 새로운 접근 방식인 ViT 와 통합한다. 이 절에서는 모델 설계에 대한 세부 사항을 설명하고, 적대적 기계학습 분야에서의 기여를 강조한다.

적대적 공격 분야에서 신뢰도 점수를 활용하는 기존 연구들은 주로 공격자가 모델의 신뢰도 출력을 통해 입력 이미지의 작은 변형이 예측 결과에 미치는 영향을 파악하고 이를 이용하는 방식으로 진행되었다. 이러한 접근 방식은 공격의 성공 가능성을 높이지만, 동시에 공격자가 모델의 내부 정보에 접근할 수 있어야 한다는 큰 제약이 따른다.

예를 들어, Fast Gradient Sign Method(FGSM)[3]와 Projected Gradient Descent(PGD)[4]는 모두 모델의 오차 기울기를 사용하여 입력 이미지에 변형을 가하며, 이를 통해 적대적 예제를 생성한다. FGSM 은 간단하고 효율적인 접근 방식을 제공하지만, 그 실행은 모델의 기울기 정보에 대한 접근을 필요로 한다. PGD 는 다단계 접근 방식을 통해 보다 정교한 예제를 생성하며, 각 단계의 입력 변형이 결과에 미치는 영향을 측정하고 최적화한다. 이러한 방법들은 모두 완전한 블랙박스로 접근하지 않는다.

이와 대비되는 본 연구는 신뢰도 점수에 전혀 의존하지 않으며, 이는 공격자가 모델의 내부 정보에 접근하지 못하는 현실적인 상황에서도 적용 가능한 방법을 제공한다. 본 연구는 신뢰도 점수를 활용하는 기존 방법들과는 달리, 공격을 순수한 '블랙박스' 방식으로 수행하며, 이는 보다 범용적이고 현실적인 적용을 가능하게 한다.

본 연구는 강화 학습 알고리즘인 근접 정책 최적화(PPO, Proximal Policy Optimization)[1]와 이미지의 공간적 계층을 파싱하기 위해 특별히 조정된 주의 기반 메커니즘(Attention based)인 ViT-like model[2] 시너지 조합이다. 이 혁신적인 조합은 모델이 이미지 분류기의 취약점을 식별하고 이용하는 체계적인 방법을 제공한다.

강화 학습 구성 요소: 우리 방법론의 핵심은 PPO 에 의해 조종되는 RL 에이전트다. 이 에이전트의 주요 기능은 잘 훈련된 이미지 분류기를 속이기 위한 적대적 예제를 생성하는 것이다. 이 전략은 환경과의 지속적인 상호작용을 통해 점차적으로 접근 방식을 정제하는 정책 네트워크에 의해 결정된다.

ViT-like 구성 요소: 변조의 영향을 이해하기 위한 구조적 기반으로서, 이미지를 패치로 분석한다. 이는 교차-주의(Cross-Attention) 메커니즘을 사용하여 특정 이미지 세그먼트의 변경이 분류기의 최종 결정에 어떻게 영향을 미치는지에 대한 자세한 관점을 제공한다.

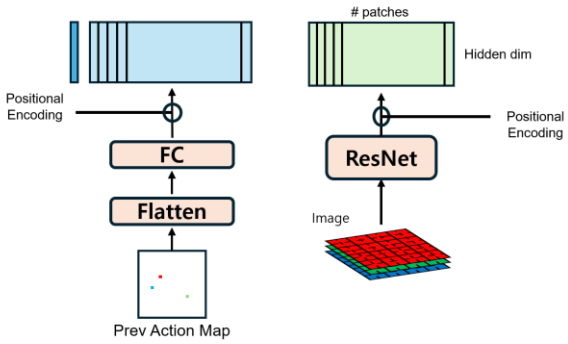

그림 1 모델 개요 1

이 아키텍처는 우리 모델이 타겟과 상호 작용하며 분류기의 취약점을 효과적이고 효율적으로 탐색하고 이용할 수 있게 해준다. 이는 이미지 분류 시스템의 견고성을 평가하고 향상시키는 데 강력한 도구다.

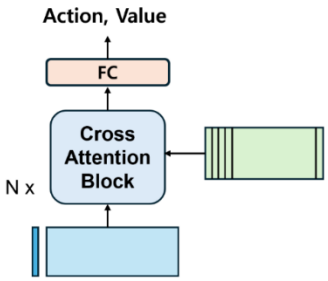

그림 2 모델 개요 2

패칭 전략을 RL 프레임워크 내에서 활용하는 결정은 엄격한 블랙 박스 조건 하에서 효과적인 적대적 전략을 식별하는 과정을 정제하기 위한 우리의 목표에서 비롯되었다. 패칭 기반 공간 관계 캡처 능력은 적대적 환경 탐색을 안내하는 데 이상적이다. 또한, PPO 의 사용은 정책 네트워크가 점점 더 정교한 공격 전략에 적응함에 따라 학습 과정이 안정적이고 효율적으로 유지되도록 보장한다.

보상 구조를 세밀하게 조정하여 RL 에이전트에 대한 적응성과 뉘앙스 있는 피드백에 초점을 맞추어, 벌점과 성공 보상의 가중 조합을 도입한다.

- 벌점 보상 ($R_{penalty}$): 에이전트의 조치가 타겟 모델의 오분류로 이어지지 않을 때마다 부여되는 부정적 보상이다. 이 벌점의 크기는 계수 α를 사용하여 조정할 수 있으며, 비효율적인 전략을 피하도록 유도한다.

- 성공 보상 ($R_{success}$): 에이전트의 조치가 타겟 분류기를 성공적으로 속여 오분류를 유발할 때 부여되는 긍정적 보상이다. 이 보상은 계수 γ를 사용하여 보상의 강도를 결정한다, 따라서 에이전트가 분류기의 취약점을 찾아내고 이용하도록 동기를 부여한다.

각 시간 단계에서 에이전트가 받는 총 보상 R 은 이 두 구성 요소의 합으로, 각각의 계수에 의해 조정된다:

$$R = \alpha R_{penalty} + \gamma R_{success} \quad (1)$$

이 가중 보상 구조는 RL 에이전트가 생산적이지 않은 조치를 취하는 것을 방지하는 동시에, 성공적인 적대적 전략을 발견하고 구현하도록 충분히 동기를 부여한다.

Ⅲ. 결론

본논문에서는 CIFAR10 이미지를 학습한 ResNet-18 모델을 적대적 공격을 위한 target model 로 두었다.

강화학습 모델에 최대 1024(32 * 32)개의 공간 내에서 점을 32 회 찍을 기회를 부여하였다. 32 회를 초과하는 경우, 해당 공격은 실패로 간주하였다. 공격이 성공한 경우의 평균 공격 횟수와 공격 성공률은 아래 표에서 제시하였다.

Target Model	공격 성공률	평균 공격 성공 스텝 수
ResNet-18	85.5%	8.0

표 1 Adversarial Attack model 성능 표

그림 3 적대적 공격 예제

ACKNOWLEDGMENT

본 과제(결과물)은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업(차세대통신)의 연구 결과입니다.

Following are results of a study on the "Convergence and Open Sharing System (NCCOSS)" Project, supported by the Ministry of Education and National Research Foundation of Korea.r acknowledgments.

참 고 문 헌

[1] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, July 2017.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Uszkoreit, J. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, Oct. 2020.

[3] Goodfellow, I. J., Shlens, J., & Szegedy, C. "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572, Dec. 2014.

[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. "Towards Deep Learning Models Resistant to Adversarial Attacks," arXiv preprint arXiv:1706.06083, June 2017.