강화학습 기반 적대적 워터마크 생성을 통한 AI 무단 학습 방지에 대한 연구

김지훈[1], 이종호[2], 이지은[3], 신용태[4]
[1,2]숭실대학교 컴퓨터학과 석사과정
[3]숭실대학교 컴퓨터학과 박사과정
[4]숭실대학교 컴퓨터학부 교수

bizkjh9827@gmail.com, leejongho@soongsil.ac.kr, lhsgsse10@soongsil.ac.kr, shin@ssu.ac.kr

A Study on Reinforcement Learning-based Adversarial Watermark Generation for Copyright Protection against AI Training

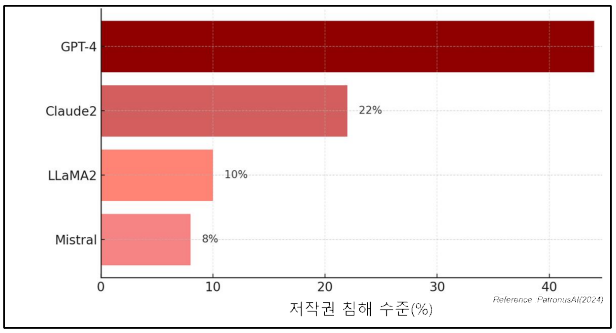
Ji-Hun Kim[1], Jong-Ho Lee[2], Ji-Eun Lee[3], Young-Tae Shin[4]
[1,2,3]Dept. of Computer Science and Engineering, Soong-Sil University
[4]School of Computer Science and Engineering, Soong-Sil University

요 약

본 논문은 AI 모델의 무단 학습으로 인한 저작권 침해를 방지하기 위해, 강화학습 기반 적대적 워터마크 생성 기법을 제안한다. 에이전트는 이미지에 인간이 인식하기 어려운 노이즈를 삽입하여, 학습 모델의 성능을 저하시킬 수 있는 워터마크를 학습한다. 보상 함수는 공격 성공률(ASR)과 시각 유사도(SSIM)를 동시에 고려하며, 반복 학습을 통해 은밀하고 효과적인 워터마크를 생성한다.

1. 서론

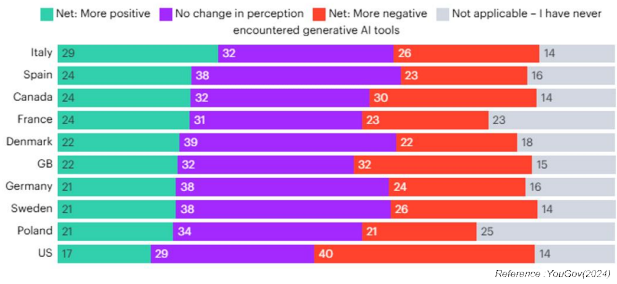
인공지능(AI)의 발전은 콘텐츠 생성, 자연어 처리, 예술 분야 등에서 획기적인 가능성을 열었지만, 동시에 저작권 침해와 창작자 권리 침해에 대한 우려 또한 빠르게 증가하고 있다. 특히 생성형 AI 모델의 등장 이후, 창작자의 스타일이나 문장, 시각적 언어 등을 무단으로 학습하여 그대로 재현하는 행위가 현실화되고 있다.



(그림 1) AI 모델 저작권 침해 평가

이를 보여주는 대표적인 사례가 대형 언어모델(LLM)에 대한 저작권 침해 실태다. (그림 1)은 인기 저서의 첫 구절을 입력한 뒤 AI가 이를 얼마나 그대로 복제하는지를 측정한 결과를 나타낸다. GPT-4의 경우 전체 입력 중 무려 44%에서 원문을 그대로 재현한 것으로 나타났으며, Claude2, LLaMA2, Mistral 등 타 모델에서도 유사한 복제 행위가 보고되었다. 이는 단순한 참고 수준을 넘어 명백한 저작물 복제로 해석될 수 있는 행위이며, AI가 저작권을 무시한 학습을 수행하고 있음을 보여준다.

이러한 행태는 이미지 생성 분야에서도 유사하게 나타나고 있다. 최근 SNS에서 큰 화제를 모은 '지브리 화풍 AI 그림' 사례는, 스튜디오 지브리의 고유한 스타일이 AI 모델을 통해 그대로 재현되며 급속히 확산된 사건이었다. 이에 대해 지브리의 감독 미야자키 하야오는 "삶에 대한 모독이다"라는 강한 표현으로 자신의 철학과 창작의 가치가 침해당하고 있다는 비판을 제기했다.



(그림 2) AI에 대한 인식 변화 추이

이와 같은 배경 속에서 AI에 대한 인식 역시 빠르게 변화하고 있다. (그림 2)는 2024년 7월 YouGov가 발표한 대상 설문 결과로, 생성형 AI에 대한 인식 변화 추이를 보여준다. 전체 응답자의 약 22%가 AI에 대해 전보다 부정적으로 인식하게 되었다고 답변했으며, 특히 미국의 경우 40%에 달하는 응답자가 부정적 인식으로 전환되었다. 이러한 흐름은 최근 'AI Slop'이라는 신조어의 등장으로 이어졌으며, 이는 생성형 AI가 인터넷 생태계를 오염시키고 있음을 풍자하는 비판적 인식의 확산을 반영한다.

이처럼 AI의 콘텐츠 침해와 사회적 반감이 동시에 심화되고 있는 상황에서, 창작자가 자신의 저작물을 AI로부터 방어할 수 있는 기술적 수단이 절실히 요구되고 있다. 이러한 배경 속에서 최근 주목받고 있는 대응 기술이 바로 '적대적 워터마크(Adversarial Watermarking)'이다. 이 기술은 사람이 인식할 수 없는 미세한 노이즈를 콘텐츠에 삽입하여, AI 모델이 해당 콘텐츠를 학습하거나 모방하지 못하도록 하는 방식이다[1]. 대표적인 사례로 Nightshade와 Glaze가 있으며, 이들은 이미지의 시각적 품질을 유지하면서도 AI 모델의 학습 성능을 효과적으로 저해하는 기술로 평가되고 있다.

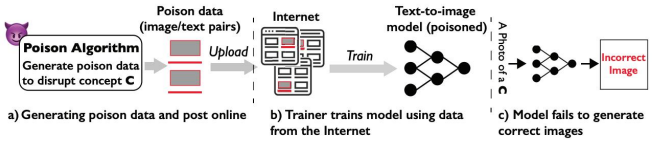
그러나 기존 대부분의 적대적 워터마크 기법은 gradient 기반의 고정된 손실 함수 최적화 방식에 의존하고 있으며, 다양한 환경 변화나 AI 모델의 적응에 유연하게 대응하지 못하는 한계를 지닌다[2].

본 논문에서는 이와 같은 문제를 해결하기 위해, 강화학습(Reinforcement Learning, RL)을 활용하여 적대적 워터마크를 생성하는 새로운 방식을 제안한다. 본 방식은 공격 성공률(Attack Success Rate, ASR)과 시각적 유사도(Structural Similarity Index Measure, SSIM)를 동시에 고려하는 다중 목적 보상 함수(multi-objective reward)를 기반으로하여 AI가 콘텐츠를 학습하지 못하도록 하면서도 사람 눈에는 변화를 거의 인지할 수 없는 워터마크를 삽입하는 것을 목표로 한다. 본 논문을 통해 기존의 gradient-based 방법에서 벗어나 강화학습을 통한 동적이고 적응적인 워터마킹 정책 학습 구조를 제안하고 시각적 품질과 공격 효과를 동시에 고려한 보상 함수 구조를 설계하며, 이를 통해 향후 디지털 콘텐츠 보호 기술의 발전에 기여할 수 있을 것으로 기대한다.

2. 선행연구

2.1 NightShade: Gradient 기반 적대적 워터마킹

"Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models"는 생성형 AI 모델의 학습을 방해하기 위해 프롬프트 특화 적대적 예제를 생성하는 기법으로 주목받고 있다. 이 방법은 multi-objective optimization을 통해 시각적 변화는 최소화하면서도, 모델 내부의 feature representation을 왜곡시키는 방식으로 동작한다. 특히, 인간의 눈에는 원본 이미지와 거의 동일하게 보이지만, AI 모델은 이를 전혀 다른 이미지로 인식하게 된다[3].

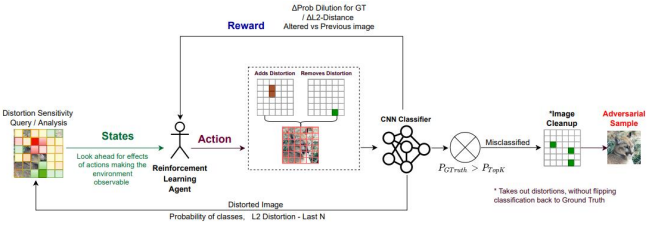


(그림 3) Nightshade의 개념

그러나 Nightshade는 gradient 기반의 고정된 손실 함수 최적화 방식에 의존하고 있어, 다양한 환경 변화나 AI 모델의 적응에 유연하게 대응하지 못하는 한계를 지닌다. 이러한 고정된 최적화 방식은 다양한 콘텐츠나 학습 모델 변화에 능동적으로 대응하지 못하는 문제를 야기하며, 이는 실제 환경에서의 적용 가능성을 제한한다.

2.2 기존 강화 학습기반 적대적 공격

강화학습을 활용한 적대적 공격 연구 중 하나인 "Robustness with Query-Efficient Adversarial Attack using Reinforcement Learning"에서는 강화학습 에이전트를 활용하여 입력 이미지에 최소한의 가우시안 노이즈를 추가하여 모델의 오분류를 유도하는 방법(RLAB)을 제안하였다 . 이 방법은 쿼리 효율성을 높이고, 민감한 영역에 집중하여 공격을 수행하는 데 초점을 맞추었다[4].



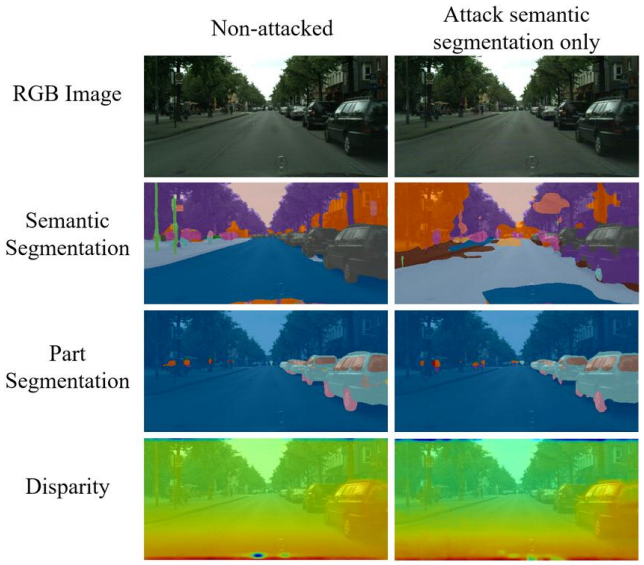
(그림 4) RLAB의 개념

그러나 해당 연구는 시각적 품질에 대한 고려가

부족하였다. 즉, 인간의 눈에 노이즈가 어떻게 인식되는지에 대한 평가나, 시각적 유사성에 대한 보상 함수 설계가 이루어지지 않았다. 이는 실제 환경에서의 탐지 회피(stealthiness) 측면에서 한계를 가지며, 시각적 품질을 고려한 적대적 공격의 필요성을 시사한다.

2.3 Stealthy Attack

Stealthy 공격은 인간의 눈에 인식되지 않는 미세한 노이즈를 추가하여 모델의 오분류를 유도하는 방식으로, 탐지 회피에 중점을 둔 적대적 공격 기법이다. 예를 들어, "Stealthy Multi-Task Adversarial Attacks" 연구에서는 다중 작업 환경에서 타겟 작업의 성능을 저하시킴과 동시에 비타겟 작업의 성능을 유지하거나 향상시키는 공격을 제안하였다[5].



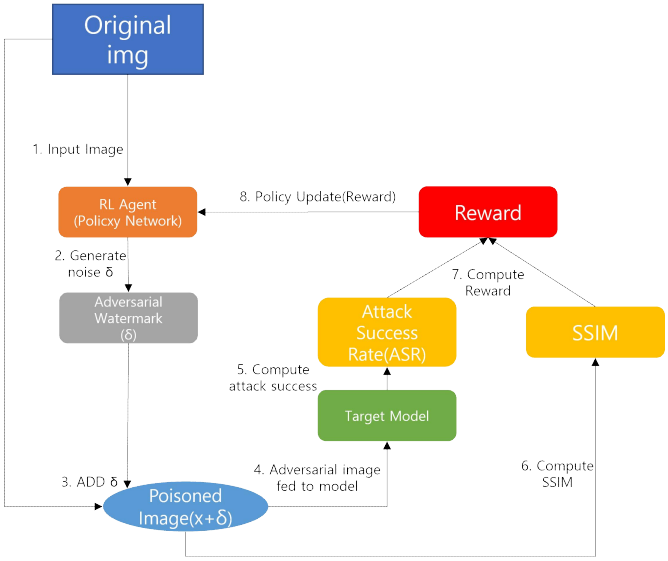
(그림 5) Stealthy Attack의 예시

그러나 이러한 Stealthy 공격은 대부분 일회성(one-shot) 예제 생성에 초점을 맞추고 있어, 지속적인 워터마크 삽입이나 장기적인 모델 학습 방해에는 한계가 있다. 즉, 훈련 데이터에 지속적으로 삽입되어 모델 학습 자체를 방해하는 형태의 공격에는 적합하지 않다.

3. 제안하는 방식

본 연구에서는 인간이 인식하기 어려운 수준의 미세한 노이즈를 삽입하여, AI 모델의 학습 성능을 저하시킴으로써 저작권 보호를 가능하게 하는 적대적 워터마크를 생성하는 강화 학습 기반 프레임워크를 제안한다.

3.1 프레임워크 개요



(그림 6) 제안하는 방식 프레임 워크

본 논문의 프레임 워크는 다음과 같은 단계로 구성된다.

1. 입력 이미지
2. 정책 네트워크를 통한 노이즈 생성
   : RL Agent(정책 네트워크)는 입력 이미지 x를 기반으로 전체 노이즈 맵 δ를 생성한다.
3. 이미지 변형
   : 생성된 노이즈는 원본 이미지에 더해져, 변형된 적대적 이미지 $x' = x + \delta$ 를 구성한다.
4. 타겟 모델 평가
   : x'는 공격 대상 모델에 입력되어 공격 성공 여부가 평가된다. 이는 공격 성공률(Attack Success Rate, ASR)로 정량화 된다.
5. 시각 유사도 계산
   : 동시에 x와 x'사이의 시각적 유사도를 계산하여 이미지의 시각적 보존 정도를 측정한다. 이는 SSIM(Structural Similarity Index Measure)로 정량화 된다.
6. 보상 계산
   : ASR과 SSIM을 종합한 보상 함수를 기반으로 정책의 성과를 평가한다.
7. 정책 업데이트
   : 최종 보상에 따라 정책 네트워크는 업데이트되며, 반복 학습을 통해 더 정교하고 효과적인 적대적 워터마크를 생성 할 수 있도록 강화된다.

3.2 보상 함수 설계

본 논문의 핵심 기여는 공격 성공률과 시각 유사도를 동시에 고려하는 다중 목적 보상 함수 설계이다. 보상 함수는 다음과 같이 정의된다.

$$Å = \alpha \cdot R_{ASR} + \beta \cdot R_{SSIM} \quad (1)$$

수식(1)에서 R_{ASR}은 적대적 공격 성공 여부(misclassification 여부)를 의미하며 타겟 공격과 비타겟 공격으로 나누어진다. 이 때
타겟 공격은 $R_{ASR} = 1$ if $f(x+\delta) = y_{target}$
비타겟 공격은 $R_{ASR} = 1$ if $f(x+\delta) \neq f(x)$ 로 정의된다.
R_{SSIM}은 시각적 유사도에 대한 보상으로 x와 $x+\delta$ 간의 SSIM 값으로 결정된다. 이 때 공격 강도와 시각 품질 간의 가중치는 α,β로 조절 할 수 있다.
이 설계를 통해, 에이전트는 모델을 효과적으로 공격하면서도 시각적으로는 거의 변화가 없는 워터마크를 생성하도록 학습된다.

4. 결론

본 논문에서는 생성형 AI 모델에 대한 저작권 침해를 방지하고자, 인간의 눈에는 인식하기 어려우나 모델 학습에는 치명적인 영향을 미치는 적대적 워터마크를 생성하는 새로운 방법을 제안하였다. 제안한 방법은 기존 gradient-based 방식의 한계점을 극복하고, 강화학습(Reinforcement Learning)을 기반으로 공격 성공률(ASR)과 시각적 유사도(SSIM)를 동시에 고려하는 다목적 보상 함수를 통해 은밀하고 효과적인 워터마크 생성 정책을 학습한다.

이러한 접근은 단순한 one-shot 적대적 예제 생성에 그치지 않고, 정책(policy)을 통한 지속 삽입형 워터마크 생성을 가능하게 하며, 향후 다양한 콘텐츠 보호 시나리오에 적용 가능한 유연성을 지닌다. 특히, 본 방법은 공격력과 시각 품질 간의 균형을 보장함으로써, AI의 학습 성능을 저해하는 동시에 인간 사용자에게는 자연스럽게 보이는 워터마크 생성을 실현한다는 점에서 기존 연구들과 비교해 차별화된 가능성을 보여준다.

본 논문에서는 프레임워크 설계와 보상 함수 구조를 제시하였으며, 구체적인 실험 및 성능 검증은 향후 연구로 이어질 예정이다. 향후에는 실제 이미지 생성 모델 또는 분류기를 대상으로 제안한 정책의 효과를 정량적으로 검증하고, 탐지 우회 능력, 범용성, 방어 회피력 등 다양한 측면에서 확장 평가를 수행할 예정이다.

사사문구

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 지원을 받아 수행되었음 (2024-0-00071)

참고문헌

[1] 김지훈, 홍석민, 신용태, "적대적 워터마킹 기술의 비교 연구와 성능 분석", 한국소프트웨어감정평가학회 논문지, 제20권, 제4호, pp. 81 - 89, 2024.

[2] Zhang, J., Qin, Y., Zhang, C., Ren, K., "Invisible Watermarks for Resisting Deep Learning Based Copyright Infringement", Proceedings of the 28th ACM Conference on Computer and Communications Security (CCS), Seoul, Korea, 2021, pp. 1132 - 1145.

[3] Carlini, N., Lee, S., Tay, Y., Kolter, J. Z., Erlingsson, Ú., Goodfellow, I., "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models", arXiv preprint, arXiv:2310.13828, 2023.

[4] Sarkar, D., Yang, L., You, Y., "Robustness with Query-Efficient Adversarial Attack using Reinforcement Learning", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, Canada, 2023, pp. 2324 - 2333.

[5] Sharma, D., Wei, X., Sato, M., Lee, W., "Stealthy Multi-Task Adversarial Attacks", arXiv preprint, arXiv:2411.17936, 2024.