GAN과 FGSM을 결합한 적대적 워터마킹 : AI 모델의 무단 학습 방지

김지훈[1], 신용태[2]
[1]숭실대학교 컴퓨터학과 석사과정
[2]숭실대학교 컴퓨터학과 교수

jihunthank@soongsil.ac.kr, shin@ssu.ac.kr

Adversarial Watermarking Combining GAN and FGSM: Preventing Unauthorized Learning of AI Models

Ji-Hun Kim[1], Young-Tae Shin[2]
[1,2]Dept. of Computer Science and Engineering, Soong-Sil University

요 약

본 논문은 디지털 콘텐츠의 무단 사용을 방지하기 위해 GAN과 FGSM을 결합한 적대적 워터마킹 기법을 제안하는 것을 목표로 한다. 이 기법은 GAN을 사용해 시각적으로 인지하기 어려운 워터마크를 생성하고, FGSM을 통해 AI 모델이 이 워터마크를 학습하지 못하도록 방해한다. 제안된 기법의 효과를 SSIM과 Probability Shift & MAX Probability Shift 지표를 통해 분석하여, 디지털 콘텐츠 보호에 대한 새로운 접근 방식을 제시하고자 한다.

1. 서론

디지털 콘텐츠는 오늘날 경제와 사회 전반에 걸쳐 중요한 자산으로 자리 잡고 있으며, 이를 보호하는 문제는 점점 더 중요한 과제로 떠오르고 있다. 저작권 침해와 불법 복제는 디지털 미디어의 가치와 창작자의 권리를 위협하며, 특히 인공지능(AI)의 발달은 모델의 학습 과정 혹은 모델을 악용한 저작권 침해 행위에서 콘텐츠가 무단으로 사용될 위험이 높아지고 있다. AI 모델이 대량의 데이터를 학습하면서 저작권이 있는 이미지나 영상, 텍스트 등을 무단으로 학습하는 상황은 더 이상 이론적인 문제가 아니라 실제로 발생하는 문제로 대두되고 있다[1].

이를 해결하기 위해 최근 적대적 워터마크(Adversarial Watermarking) 기법이 디지털 콘텐츠 보호의 중요한 도구로 주목받고 있다. 적대적 워터마크는 AI 모델이 정확하게 학습하지 못하도록 원본 콘텐츠에 노이즈를 주어 모델의 성능을 저하시키는 기술이다[2].

이 중에서도 FGSM(Fast Gradient Sign Method)와 같은 기법은 간단하면서도 강력한 성능을 보이며, AI 모델의 학습에 혼란을 줄 수 있는 대표적인 방법으로 사용되고 있다[3].

한편, GAN(Generative Adversarial Network)은 생성자와 판별자가 경쟁적으로 학습하면서 데이터를 생성하는 기술로, 다양한 이미지 생성 및 변형 작업에 널리 사용되고 있다. GAN의 생성 능력을 적대적 워터마킹에 적용하면, 디지털 콘텐츠에 삽입된 워터마크가 원본 콘텐츠에 변형을 최소화 하여 인간에게는 인지되지 않으면서도 AI의 저작권 침해 행위를 방지해 줄 수 있다[3].

본 논문에서는 GAN과 FGSM을 결합한 새로운 적대적 워터마킹 기법을 제안한다. 이 기법은 AI 모델이 저작권이 있는 디지털 콘텐츠를 무단으로 이용하지 못하도록 방해하는 것을 목표로 한다. 최종적으로 본 연구는 AI를 이용한 저작권 침해에 대한 보호론을 새로운 접근 방식으로 제안하며, 향후 디지털 콘텐츠 보호 기술의 발전에 기여할 수 있을 것으로 기대한다.

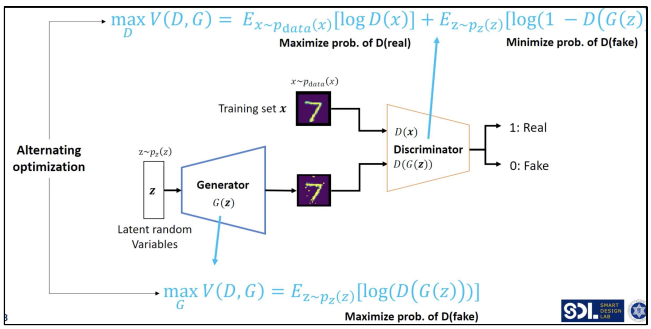
2. 선행연구

디지털 콘텐츠 보호를 위한 적대적 워터마킹 기술은 최근 다양한 연구에서 주목받고 있다. 특히, GAN과 FGSM 같은 기법은 개별적으로 또는 결합

하여 디지털 콘텐츠의 무단 학습을 방지하는 데 효과적인 방법으로 제안되고 있다. 본 장에서는 GAN과 FGSM을 이용한 적대적 워터마킹 기술의 연구들을 중심으로 관련된 선행 연구들을 검토한다.

2.1 GAN을 이용한 적대적 워터마크

GAN(Generative Adversarial Network)은 Goodfellow (2014)에 의해 처음 제안된 이후, 다양한 이미지 생성 및 변형 작업에서 활용되어 왔다. GAN은 생성자(Generator)와 판별자(Discriminator)라는 두 네트워크 간의 경쟁을 통해 데이터를 생성하며, 이러한 구조는 디지털 콘텐츠에 적용할 수 있는 워터마크를 생성하는 데 유용하다. 특히, GAN을 활용한 적대적 워터마크 기법은 원본 콘텐츠에 변형을 최소화하면서도 AI 모델에 혼란을 줄 수 있는 워터마크를 생성할 수 있다[4].



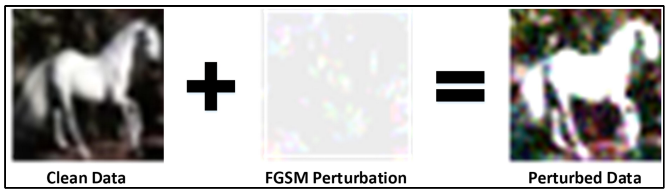
(그림 1) GAN의 개념[4]

Zhu et al.(2019)은 GAN을 이용해 적대적 워터마크를 생성하고 이를 이미지에 삽입하는 방법을 제안했다. 이 연구는 GAN이 생성한 워터마크가 인간에게는 거의 인지되지 않지만, AI 모델에는 혼란을 줄 수 있음을 보여주었다[5]. 이러한 연구들은 GAN이 적대적 워터마킹에 효과적인 도구임을 보여주고 있으며, 본 논문에서 제안하는 기법의 기초가 된다.

2.2 FGSM[6]

FGSM(Fast Gradient Sign Method)은 Goodfellow 등(2015)에 의해 제안된 방법으로, 신경망 모델에 대한 적대적 공격(adversarial attack)을 수행하는 데 사용된다. FGSM은 입력 데이터에 대해 손실 함수의 그래디언트를 이용해 최소한의 변형을 가하여, 모델이 잘못된 출력을 내도록 유도한다. 이 기법은 간단하면서도 효과적으로 AI 모델에 혼란을 줄 수 있어, 적대적 워터마킹에 적합한 도구로 사용될 수 있다[6].

이와 같은 연구들은 FGSM이 적대적 워터마킹 기법의 주요 구성 요소로 사용될 수 있는 가능성을 제시하며, 본 논문에서는 이러한 기법을 GAN과 결합하여 더욱 효과적인 워터마킹 기법을 제안하고자 한다.



(그림 2) FGSM의 개념[6]

3. 제안하는 방식

본 논문에서 제안하는 적대적 워터마킹 기법은 GAN(Generative Adversarial Network)과 FGSM(Fast Gradient Sign Method)를 통합하여, AI 모델의 학습을 방해하는 동시에 원본 디지털 콘텐츠의 변형을 최소화하는 새로운 접근 방식을 제시한다. 이 기법은 AI의 무단 학습을 막기 위해 두 가지 주요 아이디어를 결합하여 연구되었다. 먼저 GAN을 활용하여 시각적으로 인지하기 어려운 워터마크를 생성하고, 그 후 FGSM을 사용해 이 워터마크를 AI 모델이 정상적으로 학습하지 못하도록 하는 적대적 노이즈로 강화하는 방식이다.

3.1 GAN을 활용한 적대적 워터마크 생성

GAN은 주어진 랜덤 노이즈 벡터를 입력으로 받아, 이 노이즈를 시각적으로 인지하기 어려운 형태로 변환해 원본 이미지에 워터마크로 삽입하는 데 활용되었다. 구체적으로, 생성자(Generator)는 무작위로 생성된 노이즈를 입력받아, 이 노이즈를 점차 이미지 크기에 맞게 변형하여 워터마크 역할을 할 수 있는 노이즈 패턴을 생성한다.

생성된 워터마크 노이즈는 원본 이미지에 더해져, 사람의 눈으로는 쉽게 인지할 수 없지만 AI 모델에는 혼란을 줄 수 있는 형태의 이미지로 변환된다. 이 과정에서 워터마크의 강도를 조절하여 원본 이미지의 시각적 품질을 유지하면서도 AI 모델이 이를 학습하기 어렵게 만들었다.

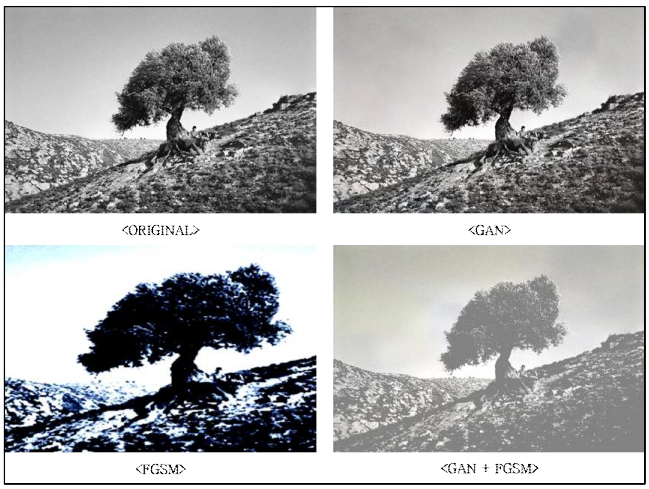
3.2 FGSM을 이용한 적대적 강화

GAN으로 생성된 워터마크를 AI 모델에 대한 적대적 공격의 효과를 강화하기 위해, FGSM 기법을 적용하였다. FGSM을 통해 강화된 워터마크는 AI 모델이 인식하기 어려운 패턴을 포함하게 되어, AI의 학습 과정에서 무단으로 활용되는 것을 방지하는

데 효과적이다. 이 기법은 생성된 워터마크의 강도를 최적화하여, 원본 이미지의 변형을 최소화하면서도 AI 모델의 인식 성능에 큰 영향을 미치도록 설계되었다.

4. 실험 및 결과

본 장에서는 제안된 GAN과 FGSM을 결합한 적대적 워터마킹 기법의 성능을 평가하기 위해, 세 가지 워터마킹 기법(GAN, FGSM, GAN+FGSM)을 비교 분석한 결과를 제시한다. 실험은 두 가지 주요 평가 지표인 SSIM(Structural Similarity Index Measure)과 Probability Shift & MAX Probability Shift를 사용하여 수행되었다. 모델 평가를 위해 Pytorch 에 기본 내장된 ResNet18 모델을 활용하였다.



(그림 3) 생성된 적대적 워터마크

4.1 실험 설정

실험은 각 워터마킹 기법에 대해 원본 이미지와의 유사성(SSIM) 및 AI 모델의 예측 확률 변화(Probability Shift & MAX Probability Shift)를 측정하였다. ResNet18 모델을 통해 각 워터마킹 이미지에 대한 예측 확률 변화를 평가하였다.

- SSIM(Structural Similarity Index Measure): 이미지의 시각적 품질을 평가하는 지표로, 원본 이미지와 워터마킹 이미지 간의 유사성을 측정한다. 값이 1에 가까울수록 원본 이미지와 유사한 것으로 간주된다.
- Probability Shift: 워터마킹 이미지에 대해 AI 모델의 예측 확률이 원본 이미지와 비교하여 얼마나 변화했는지를 측정한다. 값이 클수록 예측 확률이 크게 변화한 것을 의미한다.
- MAX Probability Shift: 워터마킹 이미지에 대해 AI 모델이 예측한 클래스의 확률 중 최대 확률이 원본 이미지에 비해 얼마나 변화했는지를 측정한다. 값이 클수록 AI 모델이 크게 혼란스러워하는 것을 의미한다.

4.2 SSIM 분석 결과

<표 1> SSIM 비교

워터마킹 기법	SSIM 값
GAN	0.9898
FGSM	0.3180
GAN+FGSM	0.7192

<표 1>은 SSIM 지표를 통해 각 워터마킹 기법이 원본 이미지와 얼마나 유사한지를 측정한 결과이다. GAN을 단독으로 사용한 경우 SSIM 값이 0.9898로 가장 높게 나타났으며, 이는 원본 이미지와의 유사성이 가장 높음을 의미한다. 반면, FGSM을 단독으로 사용한 경우 SSIM 값이 0.3180으로 가장 낮게 나타났으며, 이는 원본 이미지와의 변형이 가장 크다는 것을 의미한다. GAN과 FGSM을 결합한 경우 SSIM 값은 0.7192로, 원본 이미지와의 유사성이 GAN 단독보다는 낮으나 FGSM 단독보다는 높은 수준을 유지하였다.

4.2 Probability Shift 분석 결과

<표 2> Probability Shift 비교

워터마킹 기법	Probability Shift 값
GAN	0.3804
FGSM	1.4681
GAN+FGSM	1.5537

<표 2>는 Probability Shift 지표를 통해 AI모델의 예측 확률 변화를 측정한 결과이다. GAN을 단독으로 사용한 경우 Probability Shift 값이 0.3804로 가장 낮았으며, 이는 AI 모델의 예측 확률 변화가 상대적으로 적었음을 의미한다. 반면, FGSM을 단독으로 사용한 경우 Probability Shift 값이 1.4681로 나타나 AI 모델의 예측 확률이 크게 변했음을 보여준다. 제안된 GAN과 FGSM을 결합한 방식에서는 Probability Shift 값이 1.5537로 가장 높았으며, 이는 AI 모델에 가장 큰 혼란을 야기했음을 의미한다.

4.2 MAX Probability Shift 분석 결과

<표 3> MAX Probability Shift 비교

워터마킹 기법	MAX Probability Shift 값
GAN	0.0888
FGSM	0.1260
GAN+FGSM	0.4270

<표 3>은 AI모델이 가장 확신하는 예측 클래스의 확률 변화를 MAX Probability Shift를 통해 측정한 결과이다. GAN을 단독으로 사용한 경우 MAX Probability Shift 값이 0.0888로 가장 낮게 나타났으며, 이는 AI 모델이 여전히 원본 이미지와 유사한 클래스로 예측하고 있음을 의미한다. FGSM을 단독으로 사용한 경우 MAX Probability Shift 값이 0.1260으로 증가하였으며, GAN과 FGSM을 결합한 방식에서는 MAX Probability Shift 값이 0.4270으로 가장 크게 나타났다. 이는 제안된 기법이 AI 모델의 예측 확률을 가장 크게 변화시켰음을 나타낸다.

6. 결론

본 연구에서는 GAN과 FGSM을 결합한 새로운 적대적 워터마킹 기법을 제안하고, 그 성능을 SSIM과 Probability Shift & MAX Probability Shift 지표를 통해 평가하였다. 제안된 기법은 AI 모델이 디지털 콘텐츠를 무단으로 학습하는 것을 방지하는 효과적인 방법으로, 원본 이미지의 시각적 품질을 최대한 유지하면서도 AI 모델의 학습을 혼란시키는 데 성공적인 결과를 보였다.

SSIM 분석 결과, GAN을 단독으로 사용한 경우가 가장 높은 유사성을 유지하며 원본 이미지와의 변형이 최소화되었음을 확인할 수 있었다. 반면, GAN과 FGSM을 결합한 방식은 추가적인 노이즈로 인해 SSIM 값이 일부 낮아졌으나, 수치적으로 크게 밀리지 않았고 여전히 시각적 품질에서 우수한 성능을 보였다. 또한, Probability Shift와 MAX Probability Shift 지표에서는 GAN과 FGSM을 결합한 방식이 AI 모델의 예측 확률에 가장 큰 변화를 야기하여, 모델의 학습을 효과적으로 방해함을 증명하였다.

이러한 결과는 제안된 GAN과 FGSM을 통합한 기법이 디지털 콘텐츠 보호를 위한 적대적 워터마킹의 효과적인 해결책임을 시사한다. 특히, 이 기법은 AI 모델의 무단 학습을 방지하는 동시에 원본 콘텐츠의 시각적 품질을 유지할 수 있어, 실제 디지털 콘텐츠 보호 환경에서 중요한 역할을 할 수 있을 것으로 기대된다.

향후 연구에서는 다양한 AI 모델에 대한 적용 가능성을 평가하고, 워터마킹 기법의 강인성을 더욱 향상시키는 방향으로 확장할 수 있을 것이다. 이러한 연구는 디지털 콘텐츠의 저작권 보호뿐만 아니라, AI 기술의 윤리적 사용을 촉진하는 데에도 중요한 기여를 할 것으로 기대된다.

사사문구

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

참고문헌

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NeurIPS), Vol. 27, 2014, pp. 2672-2680.

[2] Goodfellow, I., Shlens, J., Szegedy, C., "Explaining and Harnessing Adversarial Examples," International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015, pp. 1-11.

[3] Kurakin, A., Goodfellow, I., Bengio, S., "Adversarial Machine Learning at Scale," International Conference on Learning Representations (ICLR), Toulon, France, 2017, pp. 1-17.

[4] Zhu, J., Kaplan, R., Johnson, J., Fei-Fei, L., "Hidden: Hiding Data with Deep Networks," Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 1-9.

[5] Luo, X., Zhu, J., Duan, J., "Robust Adversarial Watermarking for Deep Neural Networks," IEEE Transactions on Multimedia, Vol. 22, No. 12, 2020, pp. 2993-3004.

[6] Chen, C., Weng, R., "Robust Data Hiding Using Inverse Gradient Attention," IEEE Access, Vol. 7, 2019, pp. 168179-168191.