저작권 보호를 위한 딥러닝 기반 비가시적 워터마크 생성 모델*

양보성†, 허준범§

† 고려대학교 (대학원생)

§ 고려대학교 (교신저자)

Deep Learning-based Imperceptible Watermark Generation Model for Copyright Protection

Bo-Sung Yang†, Jun-Beom Hur§

† Korea University(Graduate student)

§ Korea University(Corresponding author)

요 약

비가시적 워터마크는 이미지 속에 원작자의 정보 혹은 이미지 소유자의 정보를 보이지 않게 삽입하여 저작권을 보호하는 기법이다. 비가시적 워터마크는 은닉된 데이터를 바탕으로 저작권의 주장 및 불법 유포자를 식별할 수 있다. 하지만, 기존 비가시적 워터마크는 데이터가 숨겨져 있는 이미지가 훼손된 경우 삽입된 워터마크의 추출이 정확하게 되지 않는다는 문제가 있다. 따라서 본 논문에서는 이미지 훼손에 강건한 딥러닝 기반 비가시적 워터마크 생성 모델을 제안한다. 제안한 모델은 인코더-디코더 구조로, 워터마크가 삽입된 이미지를 생성하고, 생성한 이미지로부터 삽입된 데이터를 추출한다. 제안된 모델은 99.6%의 데이터 추출 정확도를 달성하였으며, 훼손된 이미지에 대해서 기존 딥러닝 기반 데이터 은닉 모델(HiDDeN)보다 평균 약 9% 높은 정확도를 달성하였다.

I. 서론

비가시적 워터마크는 디지털 콘텐츠의 저작권을 보호하기 위한 기술로, 원작자의 정보를 디지털 콘텐츠안에 삽입하여, 해당 콘텐츠의 원작자를 명시하여 저작권을 주장할 수 있다. 또한, 데이터 전송 시 콘텐츠 수신자의 정보를 은닉한 뒤 추후 불법 유포된 이미지에서 콘텐츠 수신자 정보를 추출함으로써 최초 이미지 불법 유포자를 식별을 통해 저작권을 보호할 수 있다. 하지만, 데이터가 은닉된 이미지가 불법 이미지 복제 및 배포 과정에서 훼손될 경우, 훼손된 이미지로부터의 데이터 추출 성능은 크게 저하된다. 최근 딥러닝 기반 데이터 은닉-추출 기법들은 인위적으로 압축, 크기 변화, 블러 등이 적용된 훼손된 이미지를 생성 및 학습하여 이미지 변형에 데이터 추출 성능 저하를 완화하였다. 하지만 훼손된 이미지를 함께 학습하는 것은 학습을 불안정하게 만들어 훼손되지 않은 이미지에 대한 데이터 추출 성능을 저하시킨다. 또한, 이미지 불법 복제에서 일어날 수 있는 이미지 훼손을 고려하지 않아, 광고 삽입 등의 이미지 훼손에는 여전히 데이터 추출 성능이 크게 저하된다.

따라서, 본 논문에서는 저작권 보호를 위해 여러 종류의 이미지 훼손에도 강건한 데이터 추출 성능을 갖는 비가시적 워터마크 생성 모델을 제안한다. 제안한 모델은 이미지에 데이터를 은닉시키는 인코더의 표현력을 높이고,

* 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2019-0-00533, IITP-2021-0-01810)과 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2021R1A6A1A13044830)

은닉된 데이터를 추출하는 디코더의 저수준 특징과 고수준 특징을 결합하여 데이터 추출 정확도를 향상시킨다. 그 결과, 제안 모델은 훼손되지 않은 이미지에 대한 데이터 추출 정확도를 저하하지 않으면서 다양한 종류의 이미지 훼손에 대한 데이터 추출 정확도 하락을 크게 완화하였다. 제안 모델은 이미지 훼손이 없는 상황에서 99.6%의 데이터 추출 정확도를 기록하였으며, 다양한 종류의 훼손된 이미지에 대해서 기존 딥러닝 기반 데이터 은닉 모델인 HiDDeN[1]보다 평균 약 9%의 데이터 추출 정확도 향상을 달성하였다.

II. 관련 연구

2.1 저작권 침해 사이트 행위 분석 연구

저작권 침해 사이트들은 웹툰, 드라마, 영화 등의 디지털 콘텐츠를 불법적으로 배포하여 수익을 창출한다. 최근 연구들은 저작권 침해 사이트들의 생태계 및 행위를 분석하고 탐지하는 방법을 제시한다. Rafique 외 4명[2]은 저작권 침해 사이트의 네트워크 트래픽을 분석하여 저작권 침해 사이트들이 불법 광고 게시, 악성코드 유포 및 링크 하이재킹 등의 행위를 바탕으로 수익모델을 구성하였음을 보인다. 최슬기 외 1명[3]은 저작권 침해 사이트들의 특징을 분석하고 탐지하는 방법을 제안했다. 저작권 침해 사이트들은 광고와 관련된 특징이 다수 존재하기 때문에, 광고 정보를 주요 특징으로 활용하여 저작권 침해 사이트를 탐지할 수 있다. 하지만, 저작권 침해 사이트들을 적발하고, 접속을 차단하는 등의 사후적인 조치는 재범 방지의 효과가 미비하다. 따라서 본 논문은 재범 방지를 위해 이미지 불법 유포자를 식별할 수 있도록 광고 삽입 같은 불법 유포에서 발생하는 이미지 훼손에 강건한 비가시적 워터마크 생성 모델을 제안한다.

2.2 딥러닝 기반 비가시적 워터마킹 생성 연구

비가시적 워터마킹은 데이터 은닉 기법인 스테가노그래피의 일종으로, 디지털 콘텐츠 속에 원작자 혹은 구매자 정보를 숨기는 기법이다. Hayes 외 1명[4]은 최초로 뉴럴넷을 활용하여 이미지에 데이터를 숨기는 모델을 제안하였다. Baluja[5]는 CNN을 사용하여 데이터 은닉 모델을 개선하였으며, SteganoGAN[6]은 GAN을 도입하여, 기존의 스테가노그래피 분석 기법들의 탐지를 우회할 수 있으면서 많은 양의 데이터를 숨길 수 있는 모델이다. HiDDeN[1]은 메시지가 숨겨진 이미지가 훼손되었을 경우, 숨겨진 메시지를 복구가 어렵다는 문제를 해결하였다. end-to-end 학습에서 고의로 이미지를 훼손시키는 Noise layer를 추가하여, 강건한 메시지 복구 성능을 얻을 수 있음을 선보였으며, Luo 외 4명[7]은 adversarial training을 통해 학습되지 않은 이미지 훼손에도 강건한 메시지 복구 성능을 달성했다. 하지만, 이러한 딥러닝 기반 비가시적 워터마크 생성 모델들은 실제 이미지 불법 유포 시나리오에서 나타나는 이미지 훼손 유형들에 대해서는 데이터 추출 성능이 여전히 저하된다. 따라서, 본 연구에서는 이미지 불법 유포 시에 발생하는 이미지 훼손에 강건한 데이터 복구 성능을 갖는 비가시적 워터마크 생성 모델을 제안한다.

III. 비가시적 워터마크 생성 모델

본 논문에서 제안하는 딥러닝 기반 비가시적 워터마크 생성 모델은 데이터를 은닉하는 인코더, 데이터를 복구하는 디코더 그리고 인코더의 은닉 성능을 돕기 위한 판별자 네트워크로 구성된다. 비가시적 워터마크 모델의 전체적인 구조는 그림 1과 같다.

3.1 인코더

인코더(Encoder)는 커버 이미지(I_{COVER})와 비트 배열(M_{ENC})을 입력으로 받아, Convolution 연산과 ReLU 연산을 통해 데이터와 이미지를 가공하여 인코딩된 이미지(I_{ENC})를 생성하고, 학습을 통해 I_{COVER}와 I_{ENC} 사이의 이미지 차이를 최소화한다. 인코더는 M_{ENC}를 복제하여 $(W \times H \times L)$형태의 메시지 블록을 만든다. 이때 W와 H는 각각 I_{COVER}의 넓이와 높이이며, L은 M_{ENC}의 비트 수다. 이후, 메시지 블록과 이미지를 합쳐 $(W \times H \times L+3)$ 차원의 데이터

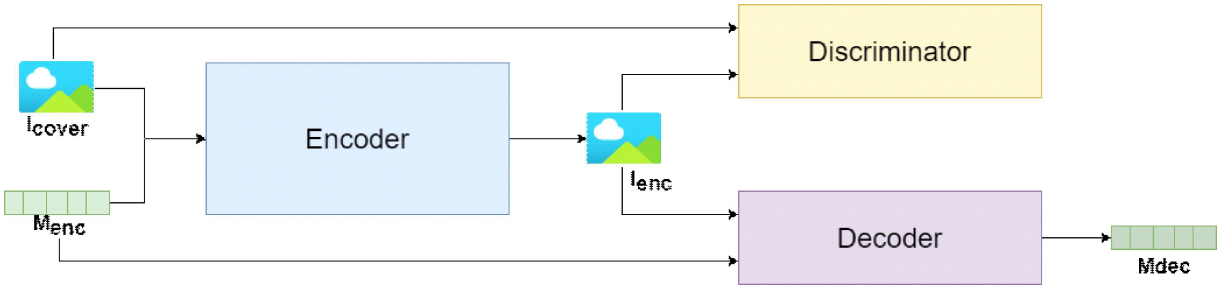

그림 1 비가시적 워터마크 생성 모델 개요

블록으로 가공하고, convolution, 배치 정규화, ReLU연산을 5번 반복하여 특징 맵을 생성한다. 그 후 생성된 특징맵에 1×1 convolution을 적용하여 I_{ENC}를 생성한다.

3.2 디코더

디코더(Decoder)는 인코딩된 이미지(I_{ENC})을 입력으로 받아 숨겨진 메시지(M_{ENC})를 추출한다. 디코더는 U-net[8] 구조로 이루어져 convolution 필터의 수를 32에서 256까지 늘렸다가 줄이면서 저수준 특징맵과 고수준 특징맵을 결합한다. 그 후, 완전연결층을 통해 데이터의 각 비트 추출한 데이터(M_{DEC})를 생성한다.

3.3 판별자 네트워크

판별자 네트워크(Discriminator)는 주어진 이미지의 데이터 은닉 여부를 판단한다. 인코더 학습 시, 잘 훈련된 판별자 네트워크를 속이도록 인코더를 학습시킴으로써, I_{COVER}와 I_{ENC} 사이의 시각적인 차이를 감소시켜 데이터의 비가시성이 향상된다.

3.4 손실 함수

본 논문에서 제안하는 모델은 인코딩된 이미지의 데이터에 의한 변화를 나타내는 이미지 손실과 복구한 데이터와 원본 데이터 간의 차이를 계산하는 손실 함수를 최소화한다. 수식 1은 이미지 손실을 최소화하기 위한 손실 함수이다.

$$L_{img} = \alpha \frac{\| I_{COVER} - I_{ENC} \|^2}{CHW} - \beta \log(P_{dis}(I_{ENC}, I_{COVER}))$$

수식 1. 이미지 손실 함수

이때, α와 β는 각 항 사이의 중요도를 반영한 하이퍼 파라미터로, 본 논문에서는 각각 0.8, 0.01로 설정하였다. $P_{dis}(I_{ENC}, I_{COVER})$는 판별자 네트워크가 I_{ENC}를 I_{COVER}로 예측할 확률로, 판별자를 속이는 데 필요한 손실 함수이다.

데이터의 추출 성능을 향상시키기 위해 원본 데이터와 추출한 데이터 간의 L2 손실 함수를 계산한다.

$$L_{msg} = \frac{\| M_{ENC} - M_{DEC} \|^2}{len}$$

수식 2. 데이터 손실 함수

이때, M_{ENC}는 원본 데이터이고, M_{DEC}는 추출한 데이터이다. len은 데이터의 길이로, 디코더는 각 데이터의 비트들의 차이를 최소화하는 방향으로 학습한다.

IV.실험 및 결과

4.1 실험 세팅

본 실험은 $3\times128\times128$ 이미지에 30비트를 은닉시키고, 숨겨진 30비트를 추출하는 것을 목표로 한다. 이때 사용된 데이터는 COCO[9] 데이터셋이며, 학습에는 10,000장, 결과 검증에는 1,000장의 이미지를 사용하였다.

4.2 실험 결과

데이터 추출 정확도는 전체 비트 중 올바르게 예측한 비트의 비율이다.

$$Accuracy = (1 - \frac{1}{len} \sum_{n=0}^{len} o_n \oplus e_n) * 100\ (\%)$$

수식 3. 데이터 추출 정확도

이때, o_n과 e_n은 각각 원본 데이터와 추출한 데이터의 n번째 비트를 의미한다. 표 1은 다양

	원본	광고 삽입	크기 변화	이미지 잘림
Identity[1]	98.83%	60.67%	50.99%	53.18%
Combined[1]	83.20%	73.31%	71.89%	**71.99%**
SteganoGAN[6]	84.59%	73.29%	51.50%	64.02%
제안 방법	**99.60%**	**86.81%**	**85.67%**	71.78%

표 2 데이터 추출 정확도 비교

한 종류의 이미지 훼손에 대한 데이터 추출 정확도를 기존 딥러닝 기반 데이터 은닉 모델과 비교한 결과이다. Identity는 HiDDeN을 이미지 훼손 없이 학습시킨 모델이고, Combined는 HiDDeN을 다양한 종류의 이미지 훼손을 고려하여 학습한 모델이다. 광고 삽입은 전체 이미지의 가로, 세로 각 10% 면적에 해당하는 작은 광고가 삽입된 경우이고, 크기 변화는 이미지가 전체 이미지의 81% 크기로 축소된 경우이다. 이미지 잘림은 전체 이미지의 가로, 세로 각 10%씩 이미지가 잘린 경우이다.

제안 모델은 훼손되지 않은 이미지에 대해서 99.6%의 정확도를 기록하였으며, 광고 삽입 및 크기 변화에 대해서 기존 딥러닝 기반 데이터 은닉 모델들보다 13% 이상 높은 데이터 추출 정확도를 기록하였으며 이미지가 잘린 경우에도 기존 모델과 비슷한 수준의 정확도를 달성하였다.

V.결론

본 논문에서는 불법 이미지 복제 및 유포에서 발생하는 이미지 훼손에 강건한 비가시적 워터마크 생성 모델을 제안했다. 제안 모델은 훼손되지 않은 이미지에 대해서 99.6%의 데이터 추출 정확도를 기록하였으며, 광고 삽입, 크기 변화 등의 이미지 훼손에 대한 데이터 추출 정확도 저하를 크게 개선하였다. 따라서 제안 모델은 불법 복제 및 유포에서 발생하는 이미지 훼손에 강건한 비가시적 워터마크를 생성함으로써 저작권 주장 및 불법 유포자 식별을 통한 저작권 보호에 이바지할 수 있을 것으로 기대된다.

[참고문헌]

[1] Zhu, J., Kaplan, R., Johnson, J., & Fei-Fei, L. (2018). Hidden: Hiding data with deep networks. In Proceedings of the European conference on computer vision (ECCV) (pp. 657-672).

[2] M Zubair Rafique, Tom Van Goethem, Wouter Joosen, Christophe Huygens, andNick Nikiforakis. 2016. It's free for a reason: Exploring the ecosystem of freelive streaming services. InProceedings of the 23rd Network and Distributed SystemSecurity Symposium (NDSS 2016). Internet Society, 1 - 15.

[3] Jin Kwak Seul-Ki Choi. 2020. Feature Analysis and Detection Techniques forPiracy Sites. In KSII Transactions on Internet and Information Systems, vol. 14, no.5. 2204 - 2220.

[4] Jamie Hayes and George Danezis. 2017. Generating steganographic images viaadversarial training.Advances in Neural Information Processing Systems 30 (2017).

[5] Shumeet Baluja. 2017. Hiding images in plain sight: Deep steganography. InProceedings of the 31st International Conference on Neural Information ProcessingSystems. 2066 - 2076.

[6] Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu, and Kalyan Veeramachaneni.2019. SteganoGAN: High capacity image steganography with GANs.arXivpreprint arXiv:1901.03892(2019).

[7] Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar.2020. Distortion agnostic deep watermarking. InProceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition. 13548 - 13557.

[8] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[9] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.