딥러닝 기반 스테가노그래피 생성 및 판별 시스템

안지수, 고가은, 곽지현, 박진성
상명대학교 정보보안공학과

Deep Learning-Based Steganography Generation and Detection System

Ahn-Ji Su, Go-Ga Eun, Kwak-Ji Hyeon, Jin Sung Park
Department of Information Security Engineering, Sangmyung University

Abstract - 정보 보안과 저작권 보호 분야에 중요한 역할을 하는 스테가노그래피(Steganography)는 숨기고 싶은 정보를 기존 데이터에 숨기는 기법이다. 디지털 포렌식 및 보안 분야에서 많이 활용되는 스테그아날리시스(Steganalysis)는 스테가노그래피에 의해 숨겨진 데이터를 감지하여 탐지하는 기법이다. 최근 딥러닝 기술의 발전으로 인해, 본 논문에서는 GAN 기반 스테가노그래피 생성 시스템과 CNN 기반 검출 시스템을 설계하고, 각 모델의 성능을 평가한다. 두 모델을 활용하여 통합 프로그램을 구현해 향후 딥러닝 기반 스테가노그래피 연구의 기초 자료로 활용될 수 있을 것으로 기대된다.

1. 서 론

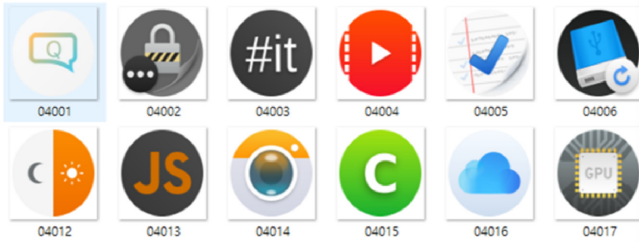
최근 사이버 보안 위협은 점차 정교해지고 있으며, 이미지에 악성 코드를 은닉하거나 생성형 AI의 보안 장치를 회피하는 새로운 기법들이 등장하고 있다. 이 가운데 스테가노그래피는 정보 은닉 기술로 다시 주목받고 있으며, 악용 가능성 또한 커지고 있다. 다양한 공격 방식이 병행되는 추세 속에서, 단일한 탐지 방법만으로는 효과적으로 대응하기 어렵다는 지적도 있다[1][2].

정보 보안 및 저작권 보호 분야에서 중요한 역할을 하는 스테가노그래피는 디지털 콘텐츠에 정보를 삽입하여 그 존재 자체를 은폐하는 기술이다[2]. 반대로, 스테그아날리시스는 이러한 은닉 정보를 탐지하는 기법으로, 디지털 포렌식이나 정보 유출 방지와 같은 분야에서 핵심적인 탐지 수단으로 사용되고 있다. 실제로 랜섬웨어 배포 등 악성 소프트웨어 유포 방식 중 일부는 스테가노그래피 기법을 활용하고 있어, 이에 대응하는 탐지 기술 개발의 필요성이 제기된다. 전통적인 스테가노그래피 방식으로는 LSB(Least Significant) 방식으로 이미지의 최하위 비트를 조작하는 방식이 대표적이나, 최근에는 GAN(Generative Adversarial Network)을 활용한 은닉 기법이 주목받고 있다[3]. GAN은 생성자와 판별자의 경쟁적 학습을 통해 원본 이미지와 시각적으로 유사한 stego 이미지를 생성할 수 있어 은닉성과 뛰어나다. 한편, 스테그아날리시스 분야에서는 CNN(Convolutional Neural Network)을 이용한 접근이 효과적인 것으로 보고되고 있다. CNN은 이미지의 공간적 특성을 학습하여 스테가노그래피로 인한 미세한 변화를 탐지하는 데 강점을 보인다[4]. 이처럼 스테가노그래피와 스테그아날리시스는 서로의 발전을 자극하는 상호 보완적 관계에 있으며, 이러한 경쟁은 정보 보안 기술 전반의 향상을 이끌고 있다. 본 연구에서는 GAN 기반 스테가노그래피 생성 시스템과 CNN 기반 탐지 시스템을 설계하고, 이를 통합한 GUI 기반 응용 프로그램을 구현함으로써, 스테가노그래피와 스테그아날리시스의 상호 발전을 위한 기반을 마련하고자 한다.

2. 본 론

2.1 데이터셋 구성 및 특성

본 연구에는 Kaggle의 Stego Image Dataset을 사용하여 진행하였다. 해당 데이터 세트의 모든 이미지는 512x512 해상도로 스테가노그래피가 삽입된 Stego Image와 스테가노그래피가 삽입되지 않은 Clean Image로 구성되어 있다. 이 중 Clean 8,000장과 Stego 6,000장을 사용하였다.



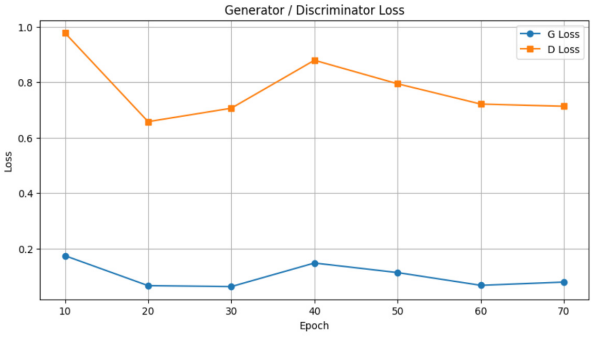
〈그림 1〉 Kaggle Stego Image Dataset

2.2 GAN 기반 스테가노그래피 생성 모델 설계

GAN(Generative Adversarial Network)은 딥러닝 기반 생성 모델로 생성자(Generator)와 판별자(Discriminator)가 경쟁적으로 학습하는 구조이다. 스테가노그래피에 적용할 경우 stego 이미지를 생성할 수 있다. 생성자는 이미지에 숨길 텍스트를 자연스럽게 은닉하고, 판별자는 은닉 여부를 구별하도록 학습된다[5].

본 논문에서는 GAN을 이용하여 스테가노그래피 생성 시스템을 구축하였다. 각 계층에 CBAM(Convolutional Block Attention Module)을 통합하였고, U-Net 기반의 생성기 구조를 설계해 정보를 은닉하였다. 입력은 데이터 세트의 이미지로 구성되며, 네트워크는 다단계 컨볼루션과 업샘플링을 통해 정보를 통합하고, 최종적으로 Tanh 함수를 거친 3채널 RGB 이미지를 출력한다.

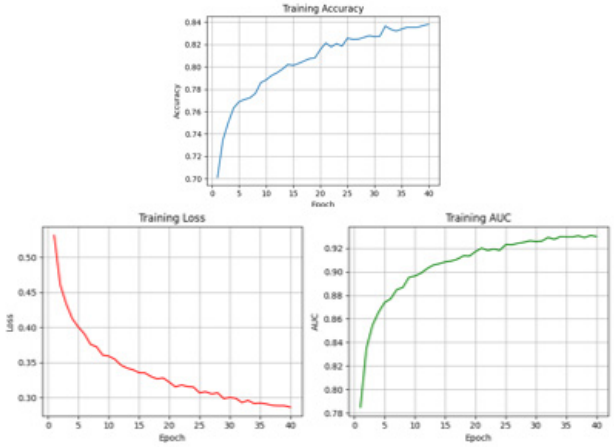
성능 평가 결과 G Loss 0.0669, D Loss 0.7214, PSNR 31.81dB, SSIM 0.9760의 성능을 기록하였다. 여기서 PSNR은 이미지 화질의 손실양을 평가하기 위해 사용되는 지표로 클수록 좋고, SSIM은 변형된 이미지의 화질 손실양을 평가하기 위해 사용한다. 해당 값이 높을수록 좋은 수치이다. 따라서 두 네트워크 간의 학습 균형이 적절히 유지되었음을 확인할 수 있다.



〈그림 1〉 생성자와 판별자 학습 손실 변화 추이

2.3 CNN 기반 스테가노그래피 판별 모델 설계

스테가노그래피 이미지와 clean 이미지를 판별하기 위해 구조화된 데이터인 이미지를 처리하는 데 적합한 구조를 가진 CNN(Convolutional Neural Network)을 기반으로 하여 모델을 설계하였다. 스테가노그래피를 판별하기 위해 CNN은 ResNet-18을 기반으로 하였다. 손실 함수와 Adam Optimizer을 사용해 학습 안정성을 확보했다. Accuracy는 학습 초반 76.4%를 기록하여, 최종적으로 83.8%까지 상승했다. Loss는 지속적으로 감소해 0.29를 기록하였다. AUC는 학습 초반에 0.82를 기록하여, 0.92로 최종 학습을 종료했다. 제안된 모델이 Stego 이미지와 clean 이미지의 판별 성능이 지속적으로 개선되었으며, 학습이 안정적으로 진행되었음을 알 수 있다.



〈그림 2〉 Accuracy, Loss, AUC의 변화 추이

2.4 스테가노그래피 생성 및 판별 통합 시스템

GAN 기반 스테가노그래피 생성 시스템과 CNN 기반 판별시스템을 하나의 통합된 프로그램을 개발하였다. PyQt5 기반의 GUI 형태로 직관적인 버튼과 이미지 미리보기 기능을 통해 쉽게 사용할 수 있도록 구현하였다. 생성 시스템의 경우 사용자는 clean 이미지를 업로드하고 16자의 텍스트의 입력을 통해 stego 이미지를 생성하도록 하였고, 생성된 stego 이미지는 자동 저장이 되도록 설계하였다. 또한, stego 이미지 또는 clean 이미지를 업로드하여 판별할 수 있도록 하여 직관적인 결과를 보여준다. 이러한 스테가노그래피 생성 및 판별 통합 시스템을 통해 정보 은닉과 탐지를 하나의 시스템으로 통합함으로써, 정보 보안 분야에서 실용적으로 활용될 수 있다.

〈그림 3〉 스테가노그래피 생성 및 판별 통합 프로그램

3. 결 론

GAN 기반 스테가노그래피 생성 시스템은 이미지에 텍스트를 은닉할 수 있도록 구현하였고, CNN 기반 스테가노그래피 탐지 시스템은 stego 이미지와 clean 이미지를 판별할 수 있도록 구현하여, 은닉과 판별 기능을 하나로 통합한 프로그램을 개발하였다. 통합 프로그램은 이미지 입력, 텍스트 은닉, 탐지 결과 확인까지 처리할 수 있도록 구성되어 있고 실시간으로 탐지할 수 있다. 해당 시스템은 디지털 저작권 보호, 정보 유출 방지와 같은 사이버 보안 분야에서 활용 가능성을 보여줄 것이라 기대한다. 다만, 한계점이 존재한다. 고정된 텍스트 길이와 학습 데이터가 512×512 해상도로 한정되어 있어 길이가 정해지지 않은 텍스트 입력과 다양한 크기의 일반화 성능은 추가 검증이 필요하다. 또한, 현재 개발한 시스템은 생성 및 판별에 중점을 두고 있어, 이미지에 은닉된 텍스트를 원래대로 복원하는 기능을 포함하지 않는다. 향후 연구에는 다양한 해상도의 이미지에 대한 학습과 비정형 텍스트의 은닉이 가능하도록 시스템을 개선할 필요가 있고, 디코더 개발이 필요할 것으로 보인다. 더불어 텍스트뿐만 아니라 음성과 같은 다양한 형태의 데이터를 숨기는 기능도 향후 연구에 고려할 수 있다. 그리고 프로그램뿐만 아니라 모바일이나 웹에서도 사용할 수 있도록 플랫폼 확장도 가능할 것이다.

[참 고 문 헌]

[1] A. Mishra, “New Steganographic Malware Hides in JPG Files to Deploy Multiple Password Stealers,” *GBHackers on Security*, online, March 17, 2025. [Online]. Available: https://gbhackers.com/new-steganographic-malware-hides-in-jpg-files/

[2] Divya, “New Jailbreak Technique Bypasses DeepSeek, Copilot, and ChatGPT to Generate Chrome Malware,” *GBHackers on Security*, online, March 19, 2025. [Online]. Available: https://gbhackers.com/new-jailbreak-technique-bypasses/

[3] 김현지,임세진,김덕영,윤세영,서화정, “인공지능 기반 스테가노그래피 생성 기술 최신 연구 동향”, Smart Media Journal, Vol.12,No.4, 2023

[4] 김현재, 이재구, 김규완, 윤성로, “딥러닝을 이용한 범용적 스테그아날리시스”, 정보과학회 컴퓨팅의 실제 논문지, 23(4), p.244-249, 2017

[5] Ian J.Goodfellow, Jean Pouget-Abadie, Mehd Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, “Generative Adversarial Nets”, Advances in Neural Information Processing Systems, vol.27, pp.2672-2680, 2014