디지털 스테그아날리시스 기술에 대한 적대적 예제 연구 동향

김자운, 조영호(교신저자)
국방대학교 사이버전협동과정
e-mail : ju0927@mnd.go.kr, youngho@kndu.ac.kr

Adversarial Example Research Trend on Digital Steganalysis

JaUn Kim, Youngho Cho
Dept. of Computer Science, Korea National Defense University

요 약

스테가노그래피(Steganography)란 은닉하려는 데이터를 육안으로 식별하기 어려울 정도로 미세한 노이즈(noise)의 형태로 변환하여 이미지, 비디오, 오디오와 같은 디지털 멀티미디어 매개체 내에 은닉함으로써, 정보의 은닉 여부를 알 수 없게 하는 것이 주 목적인 고도의 은닉 기술이다. 다양한 매개체를 활용할 수 있으며 은닉성이 높기 때문에 간첩 활동 같은 범죄 활동에 많이 사용되고 있다. 이에 따라 은닉된 정보를 탐지 및 추출하기 위한 스테그아날리시스(Steganalysis) 기술이 개발되고 있으며, 최근에는 머신러닝을 활용해 성능을 향상시켜 은닉된 데이터를 쉽게 추출 및 탐지할 수 있게 되었다. 따라서 공격자들은 머신러닝 혹은 딥러닝 기반의 스테그아날리시스 기술을 우회하는 방법으로 적대적 예제를 적용하는 방법을 고안하였다. 본 연구에서는 매체별 스테그아날리시스 유형과 방식을 확인하고 스테그아날리시스 기법에 대한 적대적 예제 연구 동향을 살펴보고자 한다.

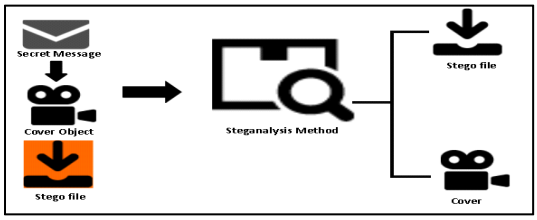
1. 서론

최근 외교·국방 분야에 종사하는 관계자를 겨냥한 이메일 해킹 시도가 연이어 발생하고 있다. 북한 해커 조직으로 알려진 라자루스는 이미지에 몰래 악성코드를 은닉하는 스테가노그래피(steganography) 기법을 활용했다[1]. 스테가노그래피는 데이터 은폐 기술 중 하나로, 제 3자에게 데이터의 존재를 인지하지 못하도록 숨기고자 하는 정보를 매체(image, audio, video 등) 내부에 은닉하는 것을 말한다.

스테그아날리시스(Steganalysis)는 매체에 은닉된 정보를 탐지하는 방법이다. 최근에는 머신러닝(Machine Learning)과 딥러닝(Deep Learning)을 활용한 스테그아날리시스가 연구되고 있다. 그러나 대부분의 스테그아날리시스는 스테고의 정상여부만 식별하게 되는데, 이러한 과정에서 적대적 예제를 이용하여 우회할 수 있는 취약점이 발생한다. 본 연구에서는 디지털 매체별 스테그아날리시스의 탐지 기술에 대해 이해하고 적대적 예제 연구 동향을 알아본다. 본 눈문의 구조는 다음과 같다. 2장에서는 디지털 매체의 스테그아날리시스의 유형과 방식에 대해서 알아보고, 3장에서는 스테그아날리시스에 대한 현재까지의 적대적 연구를 다룬다. 그리고 마지막 4장에서는 결론과 향후계획으로 구성하였다.

2. 스테그아날리시스(Steganlysis)

Steganalysis는 이미지 등 일반적인 자료에 암호화된 정보를 은닉하는 스테가노그래피에 대한 검출 및 분석방법이다. 기존의 스테그아날리시스는 커버와 스테고의 통계적 특성을 이용해 특징(Feature)를 추출하고 추출된 특징을 통해 스테고 분류기를 학습하는 수동적 방식이었으나 최근에는 머신러닝과 딥러닝을 활용하여 특징 추출 및 분류의 성능을 향상시키고 있다.



[그림 1] 스테그아날리시스 개념

2.1 영상(Image) Steganalysis

Image Steganlysis는 다른 매체에 비해 많은 연구가 진행되고 있으며, Qian-Net 등[2]에 의해 딥러닝을 통한 특징을 자동으로 분석하는 방법이 제안되었고, 이후 CNN(Convolution Neural Network) Layer의 고밀도화, DNN(Depp Neural Network) 적용한 연구 등 다양한 개선을 통해 발전중에 있다.

2.2 음성(Audio) Steganalysis

오디오 스테그아날리시스는 오디오 신호에 숨겨진 메시지를 확인한다. 오디오 신호는 압축과 비압축 형태로 나뉘어 있다. 비압축 오디오 형식(WAV 등)은 잡음을 제거하거나, 통계적 특징을 포함하여 특징을 추출한다. 압축 오디오(MP3, AAC등) 형식은 높은 압축률과 작은 파일 크기로 오디오 스테가노그래피에 많이 이용되고 있다. 최근 Ren 등[3] 은 DNN을 기반으로 다른 도메인의 압축 오디오를 스테그아날리시스하는 연구를 제안했다.

2.3 동영상(Video) Steganalysis

비디오 스테그아날리시스는 주로 MV의 증감, 프레임 간의 삽입 등을 통해 추출하여 분석하였다.

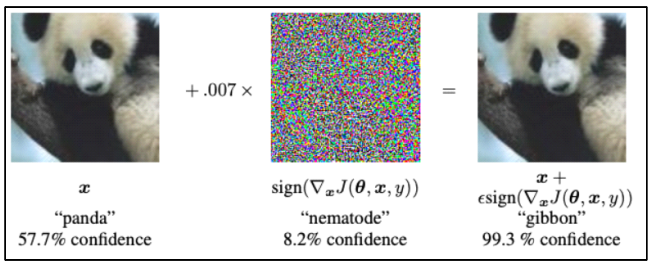
Kancherla[4] 등은 시간적, 공간적 중복성을 탐색하여 신경망과 SVM을 사용하여 비디오 스테가 분석을 위한 방법론을 최초로 제시하였다. Li 등[5] 은 차세대 동영상 부호화 기술인 HEVC(High efficiency Video Coding)에 대한 연구로 HEVC의 단위 중 예측 단위(PU)의 특징을 추출하여 딥러닝에 적용하는 연구를 진행하고 있다.

3. 적대적 예제 공격 연구 동향

3.1 적대적 예제

적대적 예제(Adversarial Example)는 역전파 변화도를 기반으로 가중치를 조정하여 손실을 최소화하기보다는 공격이 동일한 역전파 변화도를 기반으로 손실을 최대화하는 방향으로 입력 데이터를 조정함으로써 머신러닝 모델들의 취약점을 공격하는 방법이다.

적대적 공격으로 특별히 머신러닝 모델을 속이기 위해 제작된 입려 값을 적대적 예제라고 하며, 대표적인 공격방법으로 FGSM(Fast Gradient Sign Method), PGD(Projected Gradient Descent) 등이 있다. 최초 소개된 적대적 공격은 Goodfellow 등[6]이 소개한 FGSM로서 그림2와 같이 정상 이미지에 노이즈를 추가한다. 육안으로는 구별할 수 없는 노이즈를 통해 분류기를 공격하는 방법이다. 최근에는 AI 모델의 발전과 더불어 이미지 분류 모델, 객체 탐지 모델, 객체 추적 모델 등으로 적대적 예제에 대한 연구의 범위를 넓히고 있다.[7]



[그림 2] Adversarial Attack 예시[6]

3.1.1 DNN 기반의 디지털 스테그아날리시스에 대한 적대적 예제 연구

딥러닝을 적용한 디지털 스테그아날리시스가 적대적 예제 공격에 취약한지에 많은 연구가 진행되고 있다. 본 논문에서는 이에 관한 문헌 등을 정리하였다.

Liu 등[8]은 C&W(Carlini & Wagner), SRM(Spatial Rich Model) 등 다양한 적대적 예제를 통해 DNN 기반의 스테그아날리시스에 대한 취약점이 있음을 밝히고 적절한 확률 수정을 통해 스테그아날리시스를 강화하는 방법을 제안했으나, 2차 적대적 공격에서는 해당 방법이 고차원적 인공 특징과 피셔 선형 분석을 기반으로 하기때문에 효과적이지 않음을 입증하였다.

Hyunsik Na [9] 등은 왜곡이나 이미지를 과도하게 수정하던 기존 연구 방법에서 등고선 공격 방법을 제안한다. 스테그아날리시스 기반의 탐지시스템이 주변 픽셀 간의 차이를 이용해 탐지하는 원리에 기반해 원래부터 차이가 있는 경계선 부분에 noise를 추가하여 탐지를 회피한다. 실험 결과 이미지 스테그아날리시스 탐지 회피율이 최대 19.9%가 높다는 것을 보여주며 윤곽선이 있는 형태의 영상은 약 100%의 효과가 높다는 것을 입증했다.

Yiwei Zhang [10] 등은 DNN 신경망에 저항할 수 있는 강력한 강화 커버 이미지를 반복적으로 구성하는 방법을 제안한다.

적절한 ϵ를 선택함으로써 적대적 예제의 값을 변형시키며, 아래와 같은 수식으로 정의된다.

$$C' = C + \epsilon\eta \quad (1)$$

커버 이미지가 스테가노그래피의 noise에 저항할 수 있도록 반복적으로 강화된 커버를 구성하여 내구성을 높인다. 강력한 커버 이미지를 반복적으로 생성함으로

써 스테그아날리시스를 우회할 뿐만 아니라 노이즈의 강도를 조절할 수 있다. 또한 네트워크 기반의 스테그아날리시스 탐지 측면에서도 효과적인 것을 입증하였다.

3.1.2 CNN 기반의 디지털 스테그아날리시스에 대한 적대적 예제 연구

Sai Ma 등[11]은 기존의 FGSM 방법에서 새로운 방법을 추가하여 아래 수식과 같이 적대적 예제를 정의하였다.

$$\tilde{x} = x + \eta \quad (2)$$

해당 논문에서는 η을 커버미이지로 생성시키고, 메시지에 단일 레이어 STC(Syndrome-Trellis Code)를 적용하여 변경할 픽셀을 결정한다. 적은 수의 픽셀을 LSB(Least significant bit)방식을 통해 변경시키고 경사도에 따라 픽셀의 방향을 변화시킴으로써 커버 영상으로 보이는 스테고 이미지를 완성시켜 CNN기반의 스테그아날리시스를 우회시켰다. Li[12] 등은 두가지 방법을 제안하는데, LSB-Jstgeo Gradient based attck과 LSB-Jstego Evolutionary Algorithms Based Attack으로 구분한다. Gradient Based Attack은 CNN의 모든 세부 사항을 요구하는 White-Box 공격이며, Evolutionary Algorithms Based Attack은 신경망의 출력만 필요로 하지만, 대상 이미지가 복잡할 경우 시간이 오래 걸린다는 단점이 있으나 두 방법 모두 스테그아날리시스를 우회시키는 데에 성공하였다.

4. 결론 및 향후연구 계획

본 논문에서는 현재까지 디지털을 매체로 한 스테그아날리시스에 대한 적대적 예제의 공격 동향을 살펴보았다. 전반적으로 딥러닝을 적용한 영상(image) 스테그아날리시스에 대한 적대적 예제 공격 등 이미지 스테그아날리시스의 취약점을 탐지하기 위한 연구와 대응 방안은 다른 디지털 매체에 대해 많이 제시되었지만, 오디오와 동영상에 대한 스테그아날리시스의 적대적 예제 공격에 대한 연구는 그에 비해 발표되지 않고 있다. 따라서 향후 연구에서는 딥러닝을 알고리즘을 적용한 오디오 스테그아날리시스에 대해 적대적 예제를 활용하여 취약점과 장・단점을 분석하고 궁극적으로는 다양한 디지털 매체를 이용한 스테그아날리시스에 대한 적대적 공격을 통해 취약점을 찾고 대응 방안에 대한 연구를 수행할 예정이다.

참고문헌

[1] 라이센스뉴스, “이스트시큐리티, 사이버 표적 공격 급증 주의보탈륨, ‘라자루스’ 지목, LCD 모듈 테스터 설계” https://www.lcnews.co.kr/news/articleView.html?idxno=16531(검색일자:2021.09.23.)

[2] Yinlong Qian, et al., "Deep Learning for Steganalysis via Convolutional Neural Networks," Proc. of SPIE Media Watermarking, Security, and Forensics, pp. 94090J-94090J-10, 2015.

[3] Yanzhen Ren, et al, “Spec-ResNet: a general audio steganalysis scheme based on deep residual network of spectrogram”, CoRR,arXiv:1901.06838, 2019

[4] K. Kancherla, S. Mukkamala, Video steganalysis using spatial and temporal redundancies, in: Proceedings of the 2009 International Conference on High Performance Computing and Simulation, HPCS 2009, 2009, pp. 200 - 207

[5] Li et al, “A HEVC Video Steganalysis Algorithm Based on PU Partition Modes”, Teeh Science Press, CMC Vol59, no/2, pp.563~574, 2019

[6] Ian J Goodfellow el al, “Explaining and harnessing adversarial examples”. arXiv preprint arXiv:1412.6572, 2014

[7] 김호원, 컴퓨터 비전 분야에서 AI 보안에 대한 연구 동향, 2021 KISA Report, 7월, 2021

[8] Jiayang Liu et al, “Detection based Defense against Adversarial Examples from the Steganalysis Point of View”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.4825~4834, 2019

[9] Hyunsik Na el at, “Adversarial Attack Based on Perturbation of Contour Region to Evade Steganalysis-Based Detection”, 01 September 2021

[10] Yiwei Zhang et al, “Adversarial Examples Against Deep Neural Network based Steganalysis” IH&MMSec’18, Innsbruck, Austria, June 20 - 22, 2018

[11] Sai Ma et al, “Weakening the Detecting Capability of CNN-based Steganalysis” arXiv:1803.10889, 2018

[12] Li, S., et al. “Anti-steganalysis for image on convolutional neural networks.” Multimed Tools Appl 79, 4315 - 4331, 2020