딥러닝을 활용한 이미지 스테가노그래피 탐지 동향

윤세영*, 임세진**, 심민주**, 서화정*†

* 한성대학교 (학부생)

** 한성대학교 (대학원생)

*† 한성대학교 (교수)

Trends in Detection for Image Steganography using Deep Learning

Se-Young Yoon*, Se-Jin Lim**, Min-Joo Sim**, Hwa-Jeong Seo*†

* Hansung University(Undergraduate student)

** Hansung University(Graduate student)

*† Hansung University(Professor)

요 약

이미지 스테가노그래피(Image Steganography)는 이미지 속에 암호화된 데이터를 숨기는 방법이다. 암호문의 존재로 평문이 있다는 것을 알 수 있는 암호화(Cryptography)와는 다르게 스테가노그래피는 데이터가 숨겨져 있다는 사실조차 알 수 없다. 따라서 이미지 스테가노그래피는 겉보기에 일반적인 이미지로 보이기 때문에 그 자체로 악용될 위험성이 크다. 본 논문에서는 이미지 스테가노그래피 탐지를 위한 딥러닝 기반의 연구를 알아보고 앞으로의 방향을 살펴본다.

I. 서론

스테가노그래피는 이미지, 영상, 텍스트 파일 등의 일반적인 자료 속에 암호화된 데이터를 숨기는 방법이다. 키와 알고리즘을 가지고 평문을 암호문으로 변환하는 암호화는 평문과 암호문이 쌍을 이루기 때문에 암호문의 존재로 평문이 있다는 것을 알 수 있지만, 스테가노그래피 기법을 쓴 데이터는 겉보기에 일반적인 자료처럼 보이기 때문에 쉽게 암호화된 데이터에 대한 존재를 알기 어렵다. 따라서 스테가노그래피 기법은 악의적인 데이터를 은닉하기에 적합한 방법으로써 악용될 위험성이 크다. 이에 따라 스테가노그래피를 탐지하는 방법에 대해 연구가 활발하게 이뤄져야 할 필요성이 있다.

본 논문의 구성은 다음과 같다. 2장에서는 이미지 스테가노그래피에 대한 개념과 대표적으로 사용되는 기법에 대해 알아보고, 3장에서는 CNN 기술을 활용하여 스테가노그래피를 탐지하는 연구 동향을 살펴본다. 4장은 기존 연구의 한계를 살펴보며 본 논문을 마무리한다.

II. 관련 연구

2.1 이미지 스테가노그래피

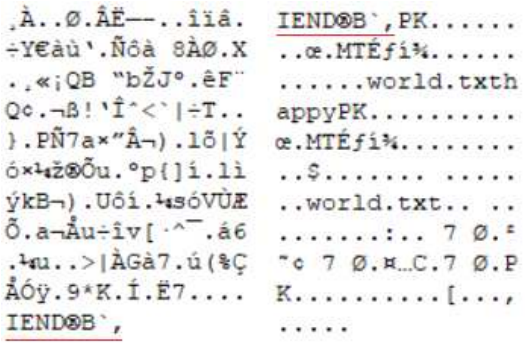

그림 1. 삽입기법 예시 (왼) Cover Image File (오) Stego Image File

정보를 숨겨주는 역할을 하는 이미지를 커버 이미지(Cover Image), 숨기려는 정보는 오리지널 데이터(Original Data), 겉으로 보기엔 커

버 이미지의 형상을 띄지만 정보가 숨겨져 있는 이미지를 스테고 이미지(Stego Image)라고 한다[1]. 이미지 스테가노그래피 기법으로는 삽입기법과 수정기법 두 가지가 주로 사용된다[2]. 삽입기법은 암호화된 데이터를 이미지 파일의 헤더나 EOF(End Of File) 뒤에 삽입하여 은닉하는 기법이다. 그림 1의 두 이미지는 PNG 파일이며, 오른쪽 이미지는 원본 이미지 파일에 텍스트 파일이 담긴 압축(zip)파일을 오픈스테고(OpenStego)를 이용하여 삽입한 것이다. 대부분의 삽입기법은 그림 1과 같이 원본 파일이 가진 헥사 값이 변하는 것이 아니라 추가되는 것이다. 또한 Hex Editor를 이용하면 삽입한 내용을 쉽게 찾아내어 분석할 수 있다. 수정기법은 각 이미지 파일의 RGB 값의 LSB(Least Significant Bit)를 수정하는 기법이다.

그림 2. 수정기법 예시 (A): Cover Image, (B): Original Image, (C): Stego Image, (D): Comparison Image

그림 2의 (D)는 Resemble.js를 이용하여 (A)와 (C)를 비교한 것으로 두 그림이 불일치 하는 부분을 분홍색으로 나타낸다. 그림 2와 같이 수정기법이 적용된 스테고 이미지 (C)는 육안으로 원본 이미지인 (A)와의 차이점을 발견할 수 없으며, 비밀 이미지인 (B)는 (C)에 전혀 나타나지 않음을 알 수 있다. 그러나 Resemble.js를 이용하여 두 이미지를 비교했을 때 다수의 RGB 값이 다르다는 것을 확인할 수 있다. 원본 이미지가 없다면 비교할 수 있는 대상이 없기 때문에 스테고 이미지만으로 은닉된 데이터가 있다는 것을 발견하기 어렵다. 수정기법 중 잘 알려진 기법으로는 HUGO(Highly Undetectable steGO), UNIWARD(UNIversal WAvelet Relative Distortion), WOW(Wavelet Obtained Weights) 등이 있다[2]. 삽입기법과 수정기법을 비교해 보았을 때, 수정기법은 삽입기법에 반해 은닉된 데이터의 존재를 알아내기 쉽지 않음을 알 수 있다.

III. 연구 동향

3.1 CNN(Convolutional Neural Network)을 이용한 스테가노그래피 분석

스테가노그래피 분석은 약 10년간 특징(Feature)을 추출하여 모델링하는 것을 기반으로 하는 기계학습 방식인 Rich Model (RM)[3]과 Ensemble Classifier (EC)[4]을 사용하여 이미지 스테가노그래피를 분석했다[5]. 이후 2015년 Qianet[6]은 딥러닝을 적용하여 기존에 쓰이던 방식보다 향상된 성능을 입증했다.

	RM+EC	CNN	FNN
평균	24.67%	7.4%	8.66%

표 1. RM+EC, CNN, FNN에 따른 스테가노그래피 분석 평균 오류 확률 [7]

이후 2016년 Pibre[7]는 표 1과 같이 RM과 EC를 함께 사용한 방법 및 CNN과 FNN(Fully connected Neural Network)을 이용한 스테가노그래피 탐지 방법에 대해 분석했으며 CNN에서 오류 확률이 가장 낮음을 확인하였다.

3.2 CNN을 이용한 범용적 스테가노그래피 분석

국내에서는 CNN을 이용하여 범용적으로 사

용이 가능한 스테가노그래피 모델을 제안하기 위해 HUGO와 UNIWARD가 적용된 스테가노그래피 데이터를 사용했다[8]. 이 연구에는 HUGO 스테고 데이터 셋만 학습시킨 모델, UNIWARD 스테고 데이터 셋만 학습시킨 모델, UNIWARD와 HUGO 스테고 데이터들이 섞여 있는 데이터 셋을 학습시킨 모델, 총 세 개의 학습 모델이 있다.

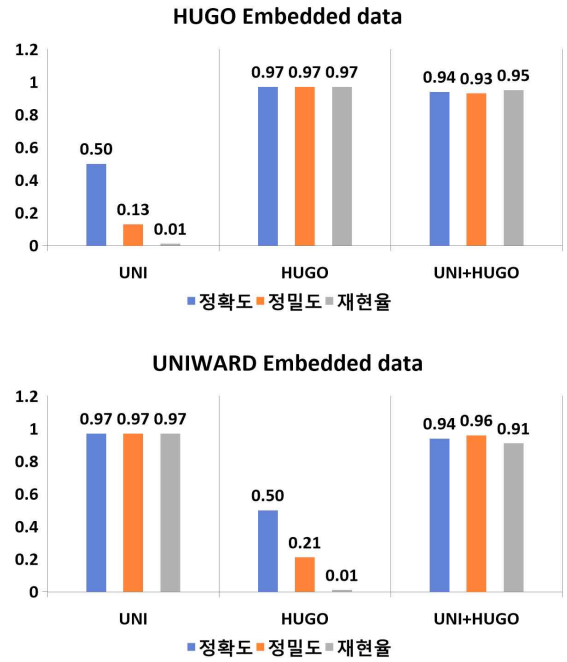

그림 3. 모델별 HUGO 및 UNIWARD의 정확도, 정밀도, 재현율 [8]

그림 3은 세 개의 모델을 이용하여 HUGO와 UNIWARD가 적용된 스테가노그래피 데이터를 검출할 때의 정확도와 정밀도, 재현율 결과를 나타낸 차트이다. HUGO로 만들어진 스테가노그래피 데이터를 각각 UNIWARD, HUGO, UNIWARD+HUGO 데이터 셋이 학습된 모델로 검출했을 때 정확도는 약 50%, 97%, 94%인 것을 확인할 수 있다. 이는 UNIWARD와 HUGO가 혼합된 형태의 모델과 HUGO 하나만 학습한 단일 모델이 비슷한 정확도로 스테가노그래피 이미지를 검출한다는 것을 알 수 있다. 마찬가지로 UNIWARD가 적용된 스테고 데이터에 대해 세 개의 모델 중 UNIWARD와 HUGO가 혼합된 모델은 정확도가 94%, UNIWARD만 학습한 단일 모델은 95%이다. 이 역시 검출 성능 면에서 혼합 모델과 단일 모델이 크게 다르지 않다는 것을 알 수 있다. 연구를 통해, 혼합된 형태의 모델도 단일 모델만큼 스테가노그래피를 검출했다는 점에서 기존의 범용성의 한계를 극복했다고 볼 수 있다.

IV. 결론

본 논문에서는 스테가노그래피 된 이미지를 분석하는 데 여러 가지 신경망 중에서 CNN이 가장 효과적임을 확인하고, CNN을 적용한 스테가노그래피 분석 연구 동향에 대해 살펴보았다. 위 연구에서는 HUGO와 UNIWARD 두 가지 기법이 혼합된 데이터 셋을 학습시킨 모델을 사용함으로써 특정 기법에 대한 범용성을 높였다. 그러나 두 가지 기법에 대해서만 학습했기 때문에 다른 기법에 대해서는 범용성을 확인할 수 없었다. 따라서 더 다양한 스테가노그래피 기법을 학습한 범용적인 스테가노그래피 분석 모델에 대한 연구가 지속적으로 이루어져야 할 필요가 있다.

V. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00540, Development of Fast Design and Implementation of Cryptographic Algorithms based on GPU/ASIC, 100%).

[참고문헌]

[1] Jae Hoon Lee, Chanran Kim, Sang Hwa Lee, and Jong-Il Park. “Image Steganography and Its Discrimination”, Journal of Broadcast Engineering, pp. 462-473, 2018.

[2] 최종석, 박종규, 김호원. “인공지능과 사물인터넷 융합 보안 기술 연구방안”, 한국통신학회, 한국통신학회지(정보와 통신) 제34권 제

3호, pp. 65-73, 2017.

[3] J. Fridrich and J. Kodovsky, “Rich Models for Steganalysis of Digital Images,” IEEE Transactions on Information Forensics and Security, Vol. 7, No. 3, pp. 868-882, 2012.

[4] J. Kodovsky, J. Fridrich, and V. Holub, “Ensemble Classifiers for Steganalysis of Digital Media,” IEEE Transactions on Information Forensics and Security, Vol. 7, 2012.

[5] Marc CHAUMONT, 『Deep Learning in steganography and steganalysis from 2015 to 2018』, Elsevier Book, 2019.

[6] Yinlong Qian, et al, “Deep Learning for Steganalysis via Convolutional Neural Networks”, Proceedings of SPIE Media Watermarking, Security, and Forensics, vol. 9409, 2015.

[7] Lionel Pibre et al. “Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source mismatch”, Proceedings of Media Watermarking, Security, and Forensics, 2016.

[8] H Kim, J Lee, G Kim, S Yoon. “Generalized Steganalysis using Deep Learning”, Korean Institute of Information Scientists and Engineers, pp. 244-249, 2017.