합성곱 신경망과 비전 트랜스포머의 교차 어텐션을 통한 스테그어날리시스 모델 구축 연구

채웅*, 조영호(교신저자)**
*국방대학교 국방관리대학원 국방과학학부 사이버컴퓨터공학과 박사과정
**국방대학교 국방관리대학원 국방과학학부 사이버컴퓨터공학과 교수
e-mail:chwo501@korea.kr, younghocho@korea.kr

A Study on Steganalysis using Cross-Attention between Convolutional Neural Networks and Vision Transformer

Woong Chae*, Youngho Cho(Corresponding Author)**
*Ph.D. Course, Dept. of Cyber Security and Computer Engineering,
Korea National Defense University
**Professor, Dept. of Cyber Security and Computer Engineering,
Korea National Defense University

요 약

스테가노그래피(Steganography)는 과거부터 지금까지 다양한 불법적인 활동에 이용되어져 왔다. 특히, 이미지는 다른 디지털 매체에 비해 인터넷에서 활발히 이용되고 있는 만큼 이미지 스테가노그래피(Image Steganography) 기술이 더욱 발전했으며 이를 탐지하기 위한 이미지 스테그어날리시스(Image Steganalysis)에 대한 연구는 굉장히 중요하다고 할 수 있다. 따라서, 본 연구에서는 이미지 스테가노그래피를 더 잘 탐지하기 위해 합성곱 신경망(CNN: Convolutional Neural Network)과 비전 트랜스포머(ViT: Vision transformer)를 함께 활용한 향상된 스테그어날리시스 모델을 제안하고자 한다. 즉, 제안 모델은 비감소 웨이블릿 변환(UDWT: Undecimated Wavelet Transform)으로 분해된 저주파와 고주파의 성분을 ViT와 CNN으로 학습한 뒤, 이를 교차 어텐션(Cross-Attention)을 통해 추가 분석 학습함으로써 스테고 이미지를 더욱 정확히 탐지한다. 향후 실험을 통해 제안 모델의 검증을 수행할 예정이다.

1. 서론

스테가노그래피(Steganography)는 전달하고자 하는 정보를 다양한 매체에 은닉하는 것으로 다양한 불법적인 활동에 이용되어져 왔다. 심지어 최근에도 스테가노그래피를 활용한 악성적인 활동이 지속해서 발생하고 있다.[1]

정보를 더 잘 숨기기 위한 스테가노그래피에 관한 연구와 이를 더 잘 탐지하기 위한 스테그어날리시스에 대한 연구도 지속해서 발전되고 있다. 특히, 이미지는 오래전부터 스테가노그래피에 활용되어 왔다. 일상생활에서도 간단한 도구를 통해 쉽게 스테가노그래피를 생성할 수 있는 만큼, 이를 탐지하기 위한 스테그어날리시스 기술에 관한 연구는 매우 중요하다.

이에 본 연구에서는 기존의 이미지 스테그어날리시스의 방식을 개선하는 새로운 모델을 제안하고자 한다. 먼저, 데이터 세트를 비감소 웨이블릿 변환(UDWT: Undecimated Wavelet Transform)을 통해 저주파와 고주파 성분으로 분해한다. 이후 저주파는 비전 트랜스포머(ViT: Vision Transformer), 고주파는 합성곱 신경망(CNN: Convolutional Neural Network)으로 각각 학습을 수행한다. 마지막으로 교차 어텐션을 통해 ViT의 결과와 CNN의 결과를 상호 분석하면 전역적인 특성과 국소적인 특성을 추가 비교하여 학습함에 따라 스테고 이미지를 더욱 정확히 판별한다.

본 연구의 기대효과는 다음과 같다.

- 스테가노그래피의 주요 특성인 고주파 성분 학습뿐만 아니라 전역적인 저주파 성분을 따로 학습하는 방식의 효과 입증
- 전역적인 특성과 국소적인 특성을 비교하는 새로운 교차 어텐션 방식을 통한 스테고 이미지 탐지 능력 향상

이후 논문 구성은 다음과 같다. 2장에서는 제안하는 스테그어날리시스와 관련된 기술 및 기존 연구를 소개하고, 3장에서는 제안 스테그어날리시스 모델의 아이디어와 모델 설계 및 동작을 설명한다. 끝으로 4장에서는 결론 및 향후 연구 계획을 제시한다.

2. 관련연구

2.1 비감소 웨이블릿 변환(UDWT)

이산 웨이블릿 변환(DWT: Discrete Wavelet Transform)는 주파수의 영역에서 이미지를 분석하기 위해 총 4가지의 주파수

대역(LL, LH, HL, HH)으로 나누는 방식이다.[2] 스테가노그래피는 기본적으로 사람의 눈에 잘 띄지 않는 부분에 은닉하고자 하는 정보를 숨기기 때문에 고주파 영역(LH, HL, HH)을 분석하여 스테그어날리시스를 수행할 수 있다. 그러나, DWT의 경우 주파수 대역을 분할하며 다운샘플링을 수행하기 때문에 스테가노그래피의 미세한 신호가 손상될 수 있다. 이에 따라 해상도를 그대로 유지하는 UDWT가 스테그어날리시스에 더 적합하다고 할 수 있다.[3]

2.2 비전 트랜스포머(ViT)

ViT는 이미지를 단어처럼 처리하기 위해 우선 16×16 크기의 패치 조각으로 생성한다. 이러한 이미지 패치는 1차원으로 임베딩되고, 이 패치가 이미지의 어느 부분에 있는지에 대한 정보를 담은 위치 인코딩을 추가한다. 이후 트랜스포머 인코더의 멀티 헤드 셀프 어텐션(MHSA: Multi-Head Self-Attention)과 피드 포워드 신경망(FFN: Feed-Forward Network)을 통해 학습을 수행하게 되고, 출력 벡터를 통해 클래스를 분류한다.[4] CNN이 국소적이라면, ViT는 전역적으로 정보를 처리하기 때문에 넓은 범위에 대한 상호 의존성에 장점이 있다고 할 수 있다.

2.3 기존 연구

Bravo-Ortiz 등[5]은 전처리 단계에서 SRM 필터를 훈련 가능과 불가능의 두 가지로 나누어 결합한다. 이후 노이즈 특징 추출 및 분석단계에서는 SE-Block(Squeeze-and-Excitation Block)을 잔차 연결(Residual Connection)과 함께 사용하여 노이즈에 대한 민감도를 높인다. 마지막 분류 단계에서는 합성곱 비전 트랜스포머(CVT: Convolutional Vision Transformer)를 통해 CNN이 추출한 국소적인 특징의 전역적인 관계를 파악하여 최종 분류를 수행하였다.

Wei 등[6]은 전처리 단계에서 30개의 SRM 필터와 32개의 Gabor 필터를 합친 총 62개의 고정된 고역 통과 필터(HPF: High-Pass Filter)를 적용하여 잔차(Residual)를 추출한다. 이후 CNN을 통해 국소적인 특징을 찾아내고, 이어서 트랜스포머가 MHSA를 통해 전역적인 특징을 찾아낸다. 마지막 분류 단계에서는 GCP(Global Covariance Pooling)와 두 개의 완전 연결(Fully Connected) 레이어를 사용하여 분류를 수행하였다.

기존 연구에서는 SRM을 통해 노이즈 잔차맵을 학습에 사용하였다. 이러한 방식은 고주파는 통과하고 저주파는 제거함으로써 이미지의 전반적인 정보에 대한 특성을 감소시킨다고 할 수 있다. 또한, CNN과 ViT을 단순히 차례대로 사용하는 모습을 볼 수 있다. 국소적인 부분을 잘 학습하는 CNN과 전역적인 부분을 잘 학습하는 ViT의 특성을 고려, 최초부터 UDWT를 통해 주파수 대역을 나누어 각각 분할 학습을 수행하는 것이 스테그어날리시스에 더 효과적이라고 할 수 있다.

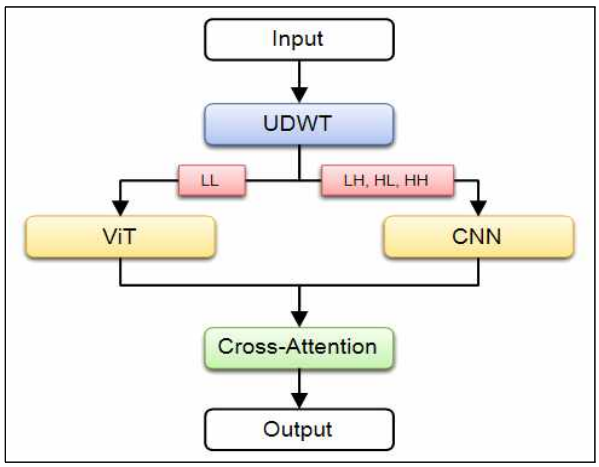
3. 제안 스테그어날리시스 모델

3.1 아이디어

앞서 2.3절에서 언급한 분할 학습 이후에 서로의 결과를 교차 분석한다면 스테고 탐지 능력을 더 향상할 수 있을 것이라는 아이디어를 추가하였다. ViT에서 나온 전역적인 결과를 CNN에서 나온 국소적인 결과에 질의를 수행하고, 질의의 결과를 통해 학습을 강화한다면 판별 능력이 더 향상될 것이라 판단하였다.

3.2 모델 설계 및 동작 설명

제안하는 모델의 구성도는 [그림 1]과 같다. 커버와 스테고 이미지를 Input 하여 UDWT를 통해 해당 이미지를 LL, LH, HL, HH의 주파수 영역으로 분리한다. LL은 ViT로 학습을 수행하고 남은 고주파 영역은 CNN을 통해 학습을 수행한다. 이후 본 연구의 핵심 제안인 교차 어텐션을 수행하여 ViT의 전역적인 정보가 CNN의 국소적인 정보에 질의를 수행하며 역전파를 수행, 이미지의 비정상적인 부분에 대한 탐지 능력을 향상시킨다. 최종적으로 분류기를 거쳐 이진 분류를 수행한다.



[그림 1] 교차 어텐션을 활용한 스테그어날리시스

4. 결론 및 향후 연구 계획

본 연구를 통해 교차 어텐션을 통한 이미지 스테그어날리시스 방식에 대한 모델 구축을 제안하였다. 또한, 실험을 통한 검증 및 결과 도출을 통해 각종 악의적인 스테고 이미지를 더욱 잘 판별하는데 기여할 것이라 확신한다.

향후 연구 계획은 다음과 같다. 첫째, 제안 모델 구현 및 각종 하이퍼파라미터 최적화를 통해 정확도를 향상시킨다. 둘째,

WOW, HILL, S-UNIWARD 등의 스테가노그래피 기법에 bpp를 달리 수행하여 비교 실험을 진행한다. 셋째, 기존에 잘 알려진 스테그어날리시스 기법을 추가 비교하여 본 연구의 완성도를 향상시킬 계획이다.[7, 8, 9]

참고문헌

[1] M. Dalal et al., "Steganography and Steganalysis (in digital forensics): a Cybersecurity guide," Multimedia Tools and Applications, vol. 80, pp. 5723-5771, 2021.

[2] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693, 1989.

[3] G. P. Nason et al., "The stationary wavelet transform and some statistical applications," in Wavelets and Statistics, vol. 103, Lecture Notes in Statistics, A. Antoniadis and G. Oppenheim, Eds. New York, NY, USA: Springer, 1995, pp. 281-300.

[4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proceedings of the International Conference on Learning Representations, 2021.

[5] M. A. Bravo-Ortiz et al., "CVTStego-Net: A convolutional vision transformer architecture for spatial image steganalysis," Journal of Information Security and Applications, vol. 81, 2024.

[6] K.-K. Wei et al., "CTNet: A convolutional transformer network for color image steganalysis," Journal of Computer Science and Technology, vol. 40, no. 2, pp. 413-427, 2025.

[7] M. Boroumand et al., "Deep Residual Network for Steganalysis of Digital Images," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181-1193, 2019.

[8] J. Ye et al., "Deep Learning Hierarchical Representations for Image Steganalysis," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2545-2557, 2017.

[9] M. Yedroudj et al., "Yedroudj-Net: An Efficient CNN for Spatial Steganalysis," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2092-2096, 2018.