이미지 군집 기반 임베딩 기법 선택을 통한 적대적 스테가노그래피 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 Adversarial Steganalysis via Image Clustering Based Embedding Method Selection

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

이미지 스테가노그래피를 탐지하기 위한 스테그어날리시스(Steganalysis) 모델은 지속해서 발전을 거듭해왔으며, 이를 회피하기 위한 적대적 스테가노그래피 연구도 활발히 진행되고 있다. 그러나, 이러한 스테가노그래피 기법이나 적대적 스테가노그래피 기법의 경우 대부분 모든 이미지에 같은 기법을 적용함으로써 이미지의 특성을 고려하지 않는다는 한계점이 존재한다. 따라서, 본 연구에서는 이미지 군집 기반 임베딩 기법 선택을 통한 적대적 스테가노그래피를 제안하고자 한다. 즉, 제안 모델은 커버 이미지를 SRM(Spatial Rich Model)을 통해 GMM(Gaussian Mixture Model)으로 군집화를 수행하여 군집별 최적의 임베딩 기법을 선정한다. 새로운 이미지는 어느 군집에 속할지를 판단하여 이미 선정된 임베딩 기법과 공격하고자 하는 스테그어날리시스의 Gradient를 연계하여 비대칭 방식(ASYM: Asymmetric)으로 적대적 공격을 수행한다. 초도실험을 통해 제안하는 방법이 탐지 오류율을 향상시킬 수 있다는 결과를 확인하였다.

1. 서론

이미지 스테가노그래피(Image Steganography)는 전달하고자 하는 정보를 이미지에 은닉하는 것으로 다양한 활동을 위해 이용됐다.

이러한 이미지 스테가노그래피를 탐지하기 위한 스테그어날리시스(Steganalysis) 모델은 지속해서 발전을 거듭해왔으며, 특히 딥러닝 기반 스테그어날리시스 모델의 경우, 잘 알려진 임베딩 기법인 WOW[1], S-UNIWARD[2], HILL[3], MiPOD[4]을 통해 0.4bpp(bit per pixel)로 숨긴 스테고 이미지에 대하여 약 90%의 탐지정확도를 나타내기도 하였다.

또한, 탐지정확도의 향상에 따라 이를 회피하기 위하여 적대적 공격 개념을 도입한 적대적 스테가노그래피 연구도 활발히 진행되고 있으며, 이러한 방식의 경우 스테그어날리시스 모델의 탐지정확도를 유의미하게 감소시키기도 하였다.

그러나, 이러한 스테가노그래피 기법이나 적대적 스테가노그래피 기법의 경우 대부분 모든 이미지에 같은 기법을 적용함으로써 이미지의 특성을 고려하지 않는다는 한계점이 존재하며, 이러한 부분을 개선하기 위하여 본 연구에서는 이미지 군집 기반 임베딩 기법 선택을 통한 적대적 스테가노그래피를 제안하고자 한다.

먼저 딥러닝 스테그어날리시스에 많이 사용되는 SRM(Spatial Rich Model) 필터를 통해 추출된 고주파 잔차 정보를 이용, 커버 이미지를 GMM(Gaussian Mixture Model)으로 군집화한다. 이후 사전 학습된 스테그어날리시스 모델의 결과를 통해 군집 내 스테고 이미지를 이진 탐지 결과 기반으로 분석하여 군집 내 최적 임베딩 기법을 판단한다. 이러한 사전 학습된 결과를 통해 새로운 이미지를 적대적 스테고로 생성하여 공격하고자 하는 스테그어날리시스 모델에 입력하여 회피능력이 향상되는 것을 확인한다.

이후 논문 구성은 다음과 같다. 2장에서는 제안하는 모델과 관련된 기술 및 기존 연구를 소개하고, 3장에서는 제안하는 적대적 스테가노그래피 모델의 설계 및 동작과 초도실험 결과를 설명한다. 끝으로 4장에서는 결론 및 향후 연구 계획을 제시한다.

2. 관련연구

2.1 공간영역 적응형 스테가노그래피(Spatial Ada

ptive Steganography)

공간영역 적응형 스테가노그래피는 이미지의 평탄한 영역보다는 복잡한 영역에 수정을 집중시켜 통계적 왜곡을 최소화하는 것이다. 이러한 공간영역 적응형 스테가노그래피의 잘 알려진 4가지 방법은 아래와 같다.

WOW(Wavelet Obtained Weights): 방향성 웨이블릿 필터를 사용하여 픽셀의 복잡도를 측정하고, 엣지나 질감이 풍부한 영역에 낮은 비용을 할당[1]

S-UNIWARD(Spatial Universal Wavelet Relative Distortion): 다양한 웨이블릿 계수의 상대적 왜곡을 합산하여 비용을 계산[2]

HILL(High pass, Low pass, Low pass): 고주파 필터와 두 번의 저주파 필터를 결합하여 임베딩 비용을 확산시킴으로써, 국부적인 수정을 부드럽게 연결하여 탐지 내성 향상[3]

MiPOD(Minimizing the Power of Optimal Detector): 가우시안 노이즈 모델을 기반으로 픽셀의 분산을 추정하고, 최적 탐지기의 성능을 최소화하는 방향으로 비용 함수를 설계[4]

2.2 SRM 필터와 딥러닝 스테그어날리시스 모델

SRM 필터는 다양한 고주파 필터를 통해 이미지 내 픽셀 간의 잔차 성분을 추출한다. 3×3, 5×5와 같은 필터를 통해 실제값과 예측값의 차이(잔차)를 추출하며, 스테가노그래피가 만들어내는 미세한 왜곡을 극대화하여 보여주는 역할을 한다.[5]

대부분의 딥러닝 스테그어날리시스 모델들은 이러한 SRM 필터를 첫 번째 레이어에 사용하고 있으며, SR-Net 계열의 딥러닝 스테그어날리시스 모델은 SRM 필터를 사용하진 않지만 3×3 필터를 사용하여 학습을 통해 픽셀 간의 잔차 성분을 추출하여 학습에 사용하고 있다.[6, 7, 8, 9]

2.3 기존 연구

Lu 등[10]은 이미지를 공출현 행렬(CM: Co-occurrence Matrix) 기반으로 사전 군집화 후, 군집마다 더 적합한 스테그어날리시스를 선택하여 탐지정확도를 향상시켰다.

Wang 등[11]은 스테가노그래피에 유리한 이미지와 아닌 이미지가 있으므로 이를 임베딩 왜곡과 이미지 유사성을 고려하여 커버 이미지를 선택하는 방법을 통하여 스테그어날리시스 탐지 오류율을 증가시켰다.

Li 등[12]은 이미지를 여러 임베딩 기법을 통해 픽셀별 Cost를 계산한 후, 주변 픽셀의 수정 방향을 고려하여 Cost를 보정하고 가장 큰 Cost를 선택하여 스테그어날리시스 탐지 오류율을 증가시켰다.

기존 연구들은 군집별 탐지기 선택, 커버 이미지 선택, 픽셀 수준 Cost 보정 및 결합을 통해 스테그어날리시스와 스테가노그래피의 성능을 향상시켰다. 이러한 결과는 이미지마다 같은 기법을 적용하는 전략이 최적이 아님을 보여준다. 이에 따라 커버 이미지를 군집화하고 각 군집에 적합한 임베딩 기법을 선택하는 것이 스테가노그래피의 성능 향상에 기여할 수 있으며 또한, 이러한 임베딩 기법을 스테그어날리시스의 탐지 결과와 연계하여 분석하고 선택한다면 더 나은 화이트박스 기반 적대적 공격을 수행할 수 있을 것으로 판단하였다.

3. 제안 적대적 스테가노그래피 모델

3.1 모델 설계 및 동작 설명

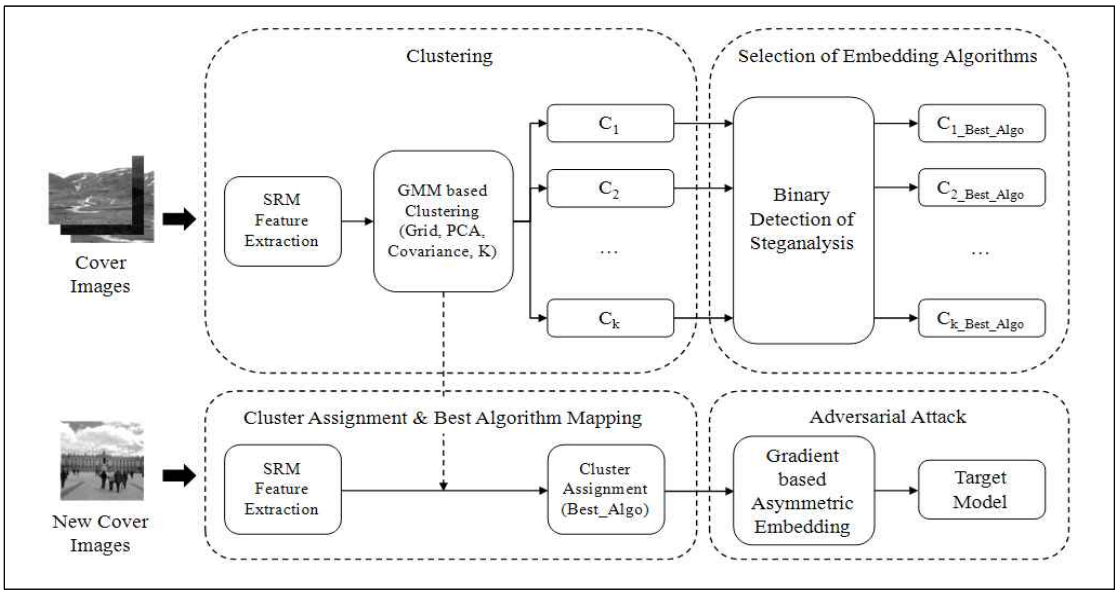
제안하는 모델의 구성도는 [그림 1]과 같다. 커버 이미지에 SRM 필터를 적용하여 잔차 맵을 획득한다. 잔차 맵을 패치로 나누고 이를 계산 가능한 하나의 숫자로 표현하기 위해 표준편차로 변환하고 표준화(평균 0, 표준편차 1)를 적용한다. 그리드와 주성분 분석(PCA), 군집 개수(K), 공분산을 다양하게 변화하여 GMM 기반으로 군집화를 수행한다.

군집별 최적의 임베딩 기법 도출을 위하여 먼저, 커버 이미지를 스테고 이미지로 생성한다. 또한, 학습 4, 검증 1, 테스트 5의 비율로 스테그어날리시스 모델을 사전학습한다. 사전학습된 스테그어날리시스 모델에 WOW, S-UNIWARD, HILL, MiPOD의 학습 및 검증 스테고 이미지를 입력하여 이진 결과를 획득한다. 이진 결과에는 4가지 임베딩 기법을 통해 생성된 이미지를 스테그어날리시스 모델이 맞췄는지 틀렸는지에 대한 결과가 표시되어 있다.

커버 이미지 기반으로 생성된 군집에 이진 결과를 통해 가장 많이 스테고를 틀린 임베딩 기법을 군집별 최적의 임베딩 기법으로 선정한다. 이후 검증 이미지를 군집에 배치하여 일반화가 정상적으로 이루어졌는지 확인하고 최적을 성능을 나타낸 군집을 최종 후보로 선정한다.

이진 결과 중 커버는 맞추고, 스테고는 틀린 결과를 대상으로 실험을 수행하였다. 이렇게 실험을 구성한 이유는 커버와 스테고를 동시에 맞추면 임베딩이 효과적으로 수행되지 않았다고 판단하였으며, 커버가 틀렸는데 스테고를 맞추거나 틀리면 이것이 커버의 노이즈 때문인지 스테고의 왜곡 때문인지 확인이 제한되기 때문이다.

위와 같이 도출된 결과를 통해 총 2가지의 방식을 평가에 활용하였다. 먼저, 군집 내에서 최적으로 선정된 1가지 기법으로 임베딩을 수행하는 Best, 이미지가 군집에 속할 확률을 계산하여 Best로 선정된 기법을 가중 합으로 임베딩을 수행하는 Mix 방식이다. 이러한 3가지 방식을 사전 학습된 스테그어날리시스의 Gredient를 활용하여 비대칭 방식으로 커버 이미지에 임베딩하여 화이트박스 공격을 수행한다.



[그림 1] 이미지 군집 기반 임베딩 기법 선택을 통한 적대적 스테가노그래피

3.2 초도실험 방법

BOSSbase 1.01의 해상도 256×256, 총 10,000장의 그레이스케일 이미지(PGM)를 기반으로 하며, 제안하는 모델이 정상적으로 동작하는지 확인하기 위해 학습 및 평가 모두 Train 이미지를 통해 수행하였다. 스테그어날리시스 모델은 GBRAS-Net을 사용하였으며, WOW 0.4bpp 기반으로 학습하였다. Google Colab을 통해 Python으로 실험을 수행하였으며, 자원은 Colab의 A100 GPU를 사용하였다.

3.3 초도실험 결과 및 분석

결과는 [표 1]과 같다. 초도실험 결과 제안하는 모델이 정상적으로 동작하였으며, Train 이미지 기반에서 공격하고자 하는 스테그어날리시스 모델의 탐지 오류율을 유의미하게 증가시킬 수 있다는 결과를 획득하였다.

[표 1] 초도실험 결과

Model	Embedding	MDR
GBRAS-Net	WOW	0.4970
	S-UNIWARD	0.5142
	HILL	0.5150
	MiPOD	0.5175
	Proposed(Best)	**0.5198**
	Proposed(Mix)	**0.5195**

4. 결론 및 향후 연구 계획

본 연구를 통해 이미지 군집 기반 임베딩 기법 선택을 통한 적대적 스테가노그래피 모델을 제안하였다. 또한, 초도실험을 통한 검증 및 결과를 도출하여 성능이 향상될 수 있음을 확인하였다.

향후 연구 계획은 다음과 같다. 첫째, 제안하는 모델의 성능 및 다른 이미지에 대해 적용 가능토록 일반화 능력 향상을 위해 하이퍼파라미터 변경 및 군집 평가 방법에 대한 추가 실험을 진행한다. 둘째, 스테그어날리시스 모델의 사전 학습을 위한 bpp의 다양화를 통해 추가적인 환경에서의 실험을 진행한다. 셋째, 다양한 스테그어날리시스 모델에 대한 적용을 통해 제안하는 기법이 단일 모델이 아닌 다양한 모델에서도 적용될 수 있다는 것을 증명하여 본 연구의 완성도를 향상할 계획이다.

참고문헌

[1] V. Holub and J. Fridrich, “Designing steganographic distortion using directional filters,” in Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), pp. 234-239, 2012.

[2] V. Holub et al., “Universal distortion function for steganography in an arbitrary domain,” EURASIP Journal on Information Security, vol. 2014, no. 1, pp. 1-13, 2014.

[3] B. Li et al., “A new cost function for spatial image steganography,” in Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 4206-4210, 2014.

[4] V. Sedighi et al., “Content-Adaptive Steganography by Minimizing Statistical Detectability,” IEEE Transactions on Information Forensics and Security, vol. 11, no. 2, pp. 221-234, 2016.

[5] J. Fridrich and J. Kodovský, “Rich Models for Steganal

ysis of Digital Images," IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 868-882, 2012.

[6] M. Yedroudj, F. Comby, and M. Chaumont, "Yedroudj-Net: An efficient CNN for spatial steganalysis," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2092-2096, 2018.

[7] R. Tabares-Soto et al., "GBRAS-Net: A convolutional neural network architecture for spatial image steganalysis," IEEE Access, vol. 9, pp. 14337-14350, 2021.

[8] J. He, S. Weng, L. Yu, and D. Chen, "Steganalysis Network With Two-Branch Preprocessing for Spatial and JPEG Domains," IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 2, pp. 1451-1466, 2025.

[9] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181-1193, 2019.

[10] J. Luo, G. Zhou, C. Yang, Z. Li, and M. Lan, "Steganalysis of content-adaptive steganography based on massive datasets pre-classification and feature selection," IEEE Access, vol. 7, pp. 21702-21713, 2019.

[11] Z. Wang, G. Feng, L. Shen, and X. Zhang, "Cover selection for steganography using image similarity," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 3, pp. 2328-2340, 2023.

[12] F. Li, Z. Yu, K. Wu, C. Qin, and X. Zhang, "Multi-modality ensemble distortion for spatial steganography with dynamic cost correction," IEEE Transactions on Dependable and Secure Computing, vol. 21, no. 4, pp. 1557-1569, 2024.