딥러닝 기반의 적대적 사례를 통한 보안성 높은 무선통신 연구

서중하, 박찬호, 김상현, 최진호, 강준혁
한국과학기술원
junghaa.seo@kaist.ac.kr, klmsang@kaist.ac.kr, cyberz@kaist.ac.kr, jkang@kaist.ac.kr

Secured Wireless Communication via Deep-Learning based Adversarial examples

Junghaa Seo, Chanho Park, Sanghyun Kim, Jinho Choi, Joonhyuk Kang
Korea Advanced Institute of Science and Technology

요 약

본 논문은 도청자가 신호의 변조기법을 정확히 분류하지 못하게 하여 무선통신의 보안성을 향상시킨다. 딥러닝 기술의 취약점인 적대적 사례(adversarial example)를 이용하여 DNN 기반의 협력적이고 정당한 수신기에서의 분류정확도는 보장하는 동시에 DNN 기반 도청자는 제대로 분류하지 못하도록 하는 기술을 제안하였으며, 실험을 통해 확인하였다..

Ⅰ. 서 론

무선통신은 전송용량 증가, 신뢰성 향상, 지연시간 감소 등 지속적인 발전으로 주요 통신수단이 되었다. 이에 따라 유선통신을 기반으로 운영되던 다양한 서비스 및 애플리케이션이 무선통신을 기반으로 변화하고 있으며 또한 비선형적 채널환경, 기존 통신이론 적용의 제한 등에서 더욱 확장하기 위해 딥러닝 기반의 연구가 활발히 진행되고 있다. 기존의 무선통신 시스템에서 보안성이 중요한 것처럼 딥러닝을 활용한 무선통신 시스템에서도 가장 중요한 요소인데, 특히 군사, 의료, 공공 보호 및 재난 구호에서 보안성은 최우선 요구사항이라 할수 있다. 무선통신 보안에 대한 일반적인 접근방식은 전송할 데이터를 암호화하는 것이지만 암호화가 항상 완전한 보안을 보장하지는 않으며(부채널 공격 등), 사물인터넷(Internet of Things, IoT) 환경에서의 소형 장치들이 제한된 능력으로 인해 강력한 암호화를 사용하지 못할 수도 있다. 보안성을 향상시키는 또 다른 방법은 협력적인 수신기 이외의 사용자, 즉 도청자가 신호를 복조할 수 없도록 하는 것이다. 이러한 목적을 위한 무선통신 분야의 기존 연구에서는 DNN 기반의 수신기에서는 오분류되지만, DNN을 활용하지 않은 일반적인 수신기에서는 정상범위의 BER을 나타내도록 적대적 사례를 설계하였다[1]. 본 논문에서는 일반적인 수신기 및 DNN 기반의 협력적이고 정당한 수신기에서 정확한 분류를 하는 동시에 DNN 기반 도청자는 정확히 분류하지 못하도록 하는 적대적 사례 제작방법을 제안하고 실험을 통해 확인한다.

Ⅱ. 본 론

본 장에서는 DNN 기반의 변조기법 분류기와 DNN 모델의 취약점 중 하나인 적대적 사례에 대해 소개하고, 적대적 사례 손실함수 제안 및 실험방법 소개, 그리고 실험결과를 분석한다.

2-1 DNN 기반 변조기법 분류기 및 적대적 사례

신호의 동상 및 직교 성분(In-phase and Quadrature-phase)을 활용하는 DNN 기반 신호 분류기는 수식 (1)과 같이 신호의 스냅샷 x를 입력하여 변조기법으로 이루어진 클래스 y를 출력한다.[2]

$$\underset{\theta}{\operatorname{argmin}}\mathcal{L}(f(\theta,x),y)) \tag{1}$$

이 때, f는 DNN 구조를 나타내며, θ는 네트워크 파라미터, x와 y는 각각 입력 데이터셋과 출력 클래스를 나타낸다. $\mathcal{L}$ 은 손실함수이며, θ를 학습하기 위해 주로 categorical cross-entropy가 사용된다.

통상적인 적대적 사례를 제작하는 방법은 수식 (2)와 같이 수식 (1)에서 학습된 DNN을 고정시킨 채, 적대적 사례의 손실함수를 최대화하여 입력 데이터 x'를 만든다.

$$\underset{x'}{\operatorname{argmax}}\mathcal{L}(f(\theta,x'),y)) \tag{2}$$

수식 (2)는 풀기 어려운 문제이기 때문에 근사방법으로 FGSM (Fast gradient sign method)[1] 또는 C&W(Carlini-Wagner attack)[3]로 적대적 사례를 제작한다. FGSM 은 수식 (3)으로 표현되며, 근사방법이 간단하면서도 높은 성능을 보인다.

$$x' = x + \alpha \cdot \operatorname{sign}(\nabla_x \mathcal{L}(f(\theta,x),y)), \quad \|x'-x\|_\infty \le \epsilon \tag{3}$$

이 때, α는 스텝사이즈이며, 원 신호 x와 오류 범위내에서 손실함수를 증가시키는 방향으로 x'를 변화시켜 클래스 y가 출력되지 않도록 적대적 사례 x'을 제작한다.

C&W 는 수식 (4)와 같이 손실함수와 왜곡도간의 균형을 파라미터 $\lambda > 0$를 조절해가며 x'을 제작한다. 이 방법은 명확한 DNN을 대상으로 FGM보다 높은 공격 성능을 보인다.

$$\underset{x'}{\operatorname{argmax}}(\mathcal{L}(f(\theta,x'),y)) - \lambda \cdot \|x'-x\|_p) \tag{4}$$

2-2 적대적 사례 제작을 위한 손실함수 설계

컴퓨터 비전(Computer Vision, CV) 분야에서 주로 사용되는 적대적 사례는 인지영역에서는 정상으로 분류되지만, 대상이 되는 DNN 모델에서는 오분류가 되도록 제작된다. 무선통신 분야에서의 기존 연구는 DNN 모델에서는 오분류되지만, DNN을 활용하지 않은 일반 수신기에서 정상범위의 BER이 나오도록 적대적 사례를 제작하였다. 본 논문에서는 일반적인 수신기뿐만 아니라 송신기와 협력적이고 정당한 DNN 기반 수신기에서 정확한 분류를 하는 동시에 이외의 DNN 기반 수신기에서는 오분류되는 목적을 가진 적대적 사례를 수식 (5)와 같이 제안한다.

$$\underset{x'}{\operatorname{argmax}}(-\mathcal{L}(f^{bob}(\theta,x'),y) + \mathcal{L}(f^{eve}(\theta,x'),y) - \lambda \cdot \|x'-x\|_2), \tag{5}$$

이때, f^{bob}은 송신기에 협력적이고 정당한 수신기의 DNN, f^{eve}는 그 이외 수신기의 DNN, 즉 도청자이다. f^{bob}의 손실함수는 입력

된 적대적 사례 x'을 정상적으로 분류하기 위해 감소하는 방향으로 최적화를 시키고, f^{eve}의 손실함수는 증가하는 방향으로 최적화시켜 정상적인 클래스 y를 출력하지 못하도록 한다. 동시에 원 신호와의 왜곡을 줄여 DNN 기반이 아닌 일반 수신기에서 정상범위의 BER을 보일수 있도록 파라미터 $\lambda > 0$를 통해 균형을 맞춘다.

2-3 실험 환경

본 실험에서는 24가지 변조 기법을 라벨로 가진 약 2백만개의 데이터셋을 사용하였다.[4] 이 데이터셋은 1024 × 2 의 형태로 각각 in-phase와 quadrature phase로 구분하여 표현되며, -20dB부터 30dB 까지의 SNR 로 구분되어 있다. 두 DNN 모델은 동일한 VGG(Visual geometry group) 구조를 사용하였으며 [2], 학습에 사용된 데이터셋의 batch-size와 순서를 다르게 해 서로 다른 하이퍼 파라미터 θ를 가지도록 하였다. 학습된 분류기는 각각 95.9%, 97.1%의 분류성공률을 가진다.

2-4 실험 결과

먼저 두가지 DNN 모델을 SNR 별로 학습시킨 후 각각 원 신호와 수식 (5)에 의해 제작된 적대적 사례를 입력하였을 때 분류정확도에 대해 측정하였다. 그림 1과 같이 원 신호를 입력하였을 때는 두가지 모델 모두 SNR이 증가함에 따라 분류정확도가 증가하였으나 적대적 사례를 입력하였을 때는 모델 #1은 5dB 이상에서 50% 이상의 분류정확도를 보이는데 반해, 모델 #2는 전 SNR 구간에서 10% 내외의 분류정확도를 보였다. 그림 1의 결과를 토대로 효과적인 측정을 위해 0dB 이상의 신호들로 다시 학습을 시킨 후 적대적 사례의 제작 반복횟수를 증가시키며 각 모델의 분류정확도를 측정하였다.

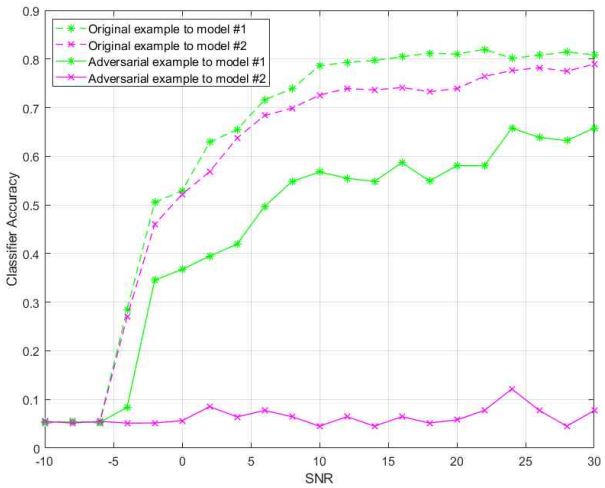

그림 1 원 신호와 적대적 사례를 입력하였을 때
SNR에 따른 두 모델의 분류정확도 변화

그림 2의 우측 작은 그래프와 같이 반복횟수가 5번까지는 두 모델 모두 분류정확도가 하락하였다. 이후 반복횟수가 증가함에 따라 모델 #1의 분류정확도는 증가하여 500번일 때 88.7%인데 반해, 모델 #2는 100번 이후로는 3% 미만의 분류정확도를 보였다.

그림 3은 적대적 사례 제작 반복횟수를 10번, 300번으로 하였을 때 신호의 형태와 원 신호와의 왜곡도, 정당한 수신기 모델과 도청자 모델에서 분류되는 클래스를 측정한 것으로, 반복횟수가 10번일때는 두 모델 모두 원 신호의 분류와 다른 클래스로 분류되었으나 300번일때는 정당한 수신기는 정상적으로 분류한 반면, 도청자는 여전히 오분류된 것을 볼수 있다.

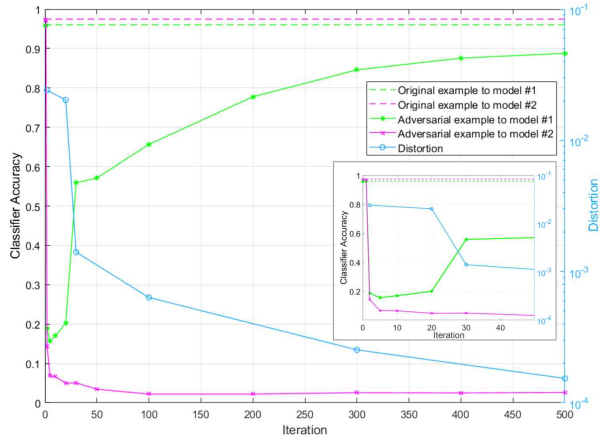

그림 2 원 신호와 적대적 사례를 입력하였을 때
적대적 사례 제작 반복횟수에 따른 두 모델의 분류정확도 변화

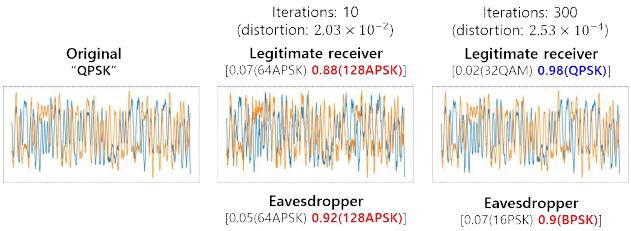

그림 3 적대적 사례 제작 반복횟수가 10번, 300번일 때 신호의 형태와
이때 정당한 수신기와 도청자에서의 왜곡도 및 분류된 변조기법 비교

Ⅲ. 결 론

무선통신 신호를 복조하지 못하도록 하는 기술은 데이터를 암호화하는 방법과 함께 무선통신의 보안성을 향상하기 위한 중요한 기술이다. 본 논문에서는 적대적 사례(adversarial example)를 활용하여 딥러닝 모델로 변조 기법을 탐지하는 도청자가 무선통신에 사용된 변조 방식을 분류하지 못하도록 하는 동시에 협력적이고 정당한 수신기에서는 정상적으로 분류되도록 하는 기술을 제안하였고, 실험을 통해 성능을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2018-0-00831, 이종 무선 네트워크를 위한 물리 계층 보안 기술 연구).

참 고 문 헌

[1] M. Sadeghi and E. Larsson, "Adversarial attacks on deep-learning based radio signal classification," IEEE WCL, vol. 8, no. 1, pp. 213-216, Feb. 2019.

[2] T. O'Shea et al.,"Over-the-air deep learning based radio signal classification," IEEE J. Sel. Topics Signal Process, vol. 12, no. 1, pp. 168-179, Feb. 2018.

[3] N. Carlini and D. Wagner,"Towards evaluating the robustness of neural networks," Security and Privacy (SP) 2017 IEEE Symposium on, pp. 39-57, 2017.

[4] Deepsig dataset: RADIOML 2018.01A 2018. (http://deepsig.io).