오디오 적대적 예제를 통한 결제 시스템 공격 사례 연구

박태정*, 최태정* ,하주현*, 윤주범**

*세종대학교 (학부생) **세종대학교(교수)

Research on Possibility of Adversarial Attack in Payment Systems

Taejeong Park*, Teajeong Choi*, Juhyun Ha*, Joobeom Yun**

*Sejong University(Undergraduate student) **Sejong University Professor

요 약

딥러닝을 활용한 인공신경망(Neural Network) 기술들이 발전하면서 머신러닝은 산업적으로 높은 가치를 갖고 발전하고 있다. 하지만 심층 신경망(DNN-Deep Neural Network) 모델에 적대적 교란(Adversarial Pertubation)이 적용되었을 경우, 오분류를 발생하는 적대적 공격(Adversarial Attack)이 가능하다는 것이 다양한 연구를 통해 발견되고 있다. 실제 음성, 이미지 인식 기술이 산업적으로 고도화될 경우, 음성을 입력받거나 이미지를 클릭하는 것만으로도 적대적 공격이 가능하게 되어 가까운 미래에 적대적 공격이 커다란 사회적 문제로 발전될 가능성이 있다고 판단한다. 본 논문에서는 오디오 입력에 적대적 공격을 수행하였을 때, 특정한 결제 사이트로 접근하여 금전적 피해를 주는 실험을 진행하고 검증하였다. 실험은 간편결제를 사용하는 e-커머스와 Speech To Text(STT) 모델을 사용한다. 입력받은 오디오에 pertubation을 추가하는 공격을 실행하여, 공격에 성공한 텍스트를 e-커머스로 전달하였을 때 자동결제가 실행되는가를 실험하였다. 실험 결과 공격이 성공하는 것이 확인되었다. 실험의 결과를 바탕으로 분석과 결론을 기술한다.

Ⅰ. 서론

인공지능과 딥러닝 기술이 빠르게 발전하면서, 이 기술을 이용하여 금융, 자동차, 사물인터넷 등 실생활과 밀접한 산업들에 적용되는 모습을 볼 수 있다. 하지만 인공지능과 딥러닝 기술의 보안 취약점을 파악하여 위협 및 정보를 탈취하는 문제점이 발생하고 있다. 따라서, 딥러닝을 활용하는 기술들에 내부 보안을 강화하는 것이 중요하다. 최근에 이미지, 오디오 관련 인공지능 모델에 대한 적대적 예제를 통한 공격들에 관한 연구들을 통해 적대적 공격(Adversarial Attack)을 통한 심층 신경망(DNN)의 취약점을 확인할 수 있다. 하지만, 발전하는 위협 기술들은 이미지, 오디오 정보 변환에 국한되지 않고, 다양한 영역으로 발전할 것이라 예측된다. 본 논문에서는 적대적 공격의 회피(Evasion)공격을 바탕으로 결제 시스템과 관련하여 적대적 공격을 통해 보안사고의 가능성에 대해 연구한다. 실험에서 사용하는 회피 공격은 머신러닝 모델 추론 시점에 공격을 수행하여 데이터를 교란시키는 방식이다 [1]. 2장에서는 Adversarial Attack에 관련된 배경 지식에

대해 기술한다. 3장에서는 실제 실험 환경과 방법, 그리고 결과를 기술한다. 4장에서는 실험 결과 분석, 본 논문의 실험 결론 및 의의를 기술한다.

Ⅱ. 배경

2.1 관련 연구

인공 신경망의 적대적 공격에 관한 이전의 연구는 다음과 같이 수행되었다.

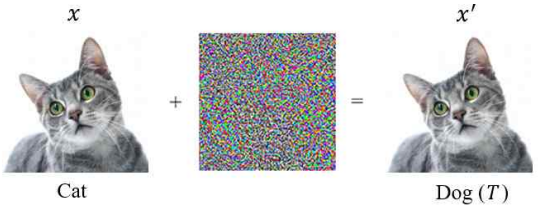

그림1. 인공 신경망 적대적 공격의 이전 연구

인스턴스 x' 은 자연적 원형 인스턴스 x와 비슷하지만 신경망에 x'을 인식시킬 경우 공격자가 선택한 잘못된 대상 t 로 분류한다. 즉 입력값 x 를 제공할 시, x와 x' 사이의 거리를 최소화하는 x'을 찾는다. 이 때 x' 은 원형 인스턴스와 같은 형태이고, 신경망에 넣을 시 공격자가 목표로 하는 T로 인식한다. 그림1은 기존 연구의 과정을 간략화하여 나타낸 것이다. 이를 표현하면 다음과 같다 [2][3].

$$minimized(x, x')$$
$$s.t. F(x') = T, x' \ is \ valid$$

적대적 공격 사례에 대한 기존 연구는 이미지 분류, 이미지에 대한 생성 모델, 이미지 분할, 얼굴 감지 등 이미지 공간에 주로 초점을 맞추었다 [3]. 오디오 공간에서 적대적 공격을 시행할 시 이미지 공간의 적대적 공격과 같은 원리를 사용할 수 있으나 음성의 특성으로 인해 이미지와 다른 속성을 가진다.

2.2 음성인식 적대적 공격의 구성

자동 음성 인식에서 신경망은 오디오 파형 x를 제공할 시 이에 대응하는 문구 y 로 출력하는 음성-텍스트 변환을 수행한다. (예: apple siri, google assistant) 이전의 오디오 적대적 공격에 관한 연구 [Carlini and Wagner, 2018]에서는 최첨단 음성-텍스트 전사 신경망인 DeepSpeech를 공격하여 오디오 영역에 표적화된 적대적 사례가 존재함을 보여준다. DeepSpeech는 end-to-end 뉴럴 네트워크를 사용하는 자동 음성 인식 시스템으로, 이러한 뉴럴 네트워크 사용은 google, apple, amazon 의 음성 인식 assistant에도 일반적으로 적용된다. 때문에 DeepSpeeh에서의 적대적 공격은 다른 종류의 뉴럴 네트워크 시스템의 적대적 공격 접근과도 유사성을 가지고 적용할 수 있다.

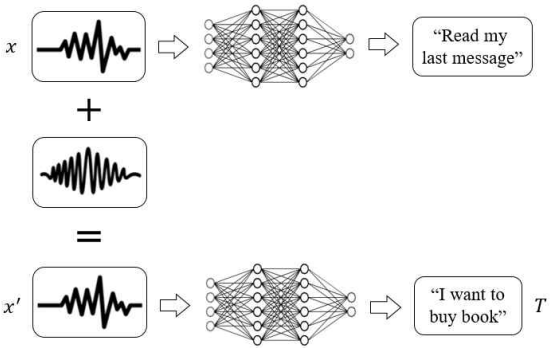

그림2. 음성 인식 적대적 공격의 구조

공격 방식은 다음과 같다. 그림2에서 보이는 바와 같이 음수1과 1 사이의 부동 소수점 값을 가지는 Wav 포맷의 원형 오디오 x를 준비한다. 이 오디오 x를 공격해 왜곡시키고 순환 뉴럴 네트워크에 제공하면 공격자가 원하는 텍스트 T가 도출된다. 이 공격은 end-to-end 방식이기 때문에 원형오디오에서 출력 텍스트로 바로 도출된다. 이와 같은 적대적 예시를 구성하기 위해서는 오디오와 대응되는 문구 사이의 정렬을 알 필요가 있다. 그러나 오디오의 어느 부분에서 해당 문구가 발생하는지 정확히 아는 것은 어렵기 때문에 이 과정에서 CTC loss를 사용한다. CTC loss 를 통해 신경망의 출력과 목표 문구의 차이를 알려주는 미분가능한 행렬을 생성해낼 수 있다. 따라서 오디오를 목표 문구와 가깝게 인식하도록 만

들기 위해 훈련 오디오와 대응되는 문구 사이의 CTC loss를 최소화하는 방법을 사용한다 [3].

공격은 공식화하여 다음과 같이 수행된다. 자연 파형 x 가 주어지면 x'과 x사이의 거리와, x'과 목표 문구T 사이의 CTC loss가 최소화되는 오디오 x'을 찾는다. 이때 x'은 원형 오디오와 비슷한 형태의 오디오이다. 이는 다음과 같이 표현된다.

$$minimize\ d(x, x') + CTC\ loss(x', T)$$
$$s.t\ x'\ is\ valid$$

이러한 음성 적대적 공격은 End-to-end 신경망에 적용되며 약간의 왜곡을 추가하여 모든 오디오 파형을 목표하는 문구로 변환한다.

Ⅲ. 실험 환경 및 결과

서론에서 언급한 것처럼, 본 논문의 실험에서는 적대적 예제가 결제 시스템에 적용되는 보안사고를 시뮬레이션 한 것이며, 실험의 시나리오에 적용된 커머스의 실제 산업적 보안 문제가 아님을 밝힌다.

3.1 실험 환경

간편결제가 도입된 e커머스 쿠팡을 대상으로 자동결제 시스템을 구축하여, 적대적 공격과 연계되도록 실험 환경을 구성하였다. 적대적 공격에 사용된 DNN (Deep Neural Network) 모델은 STT (Speech To Text) 모델 DeepSpeech v0.4.1이며 , 인공지능 스피커는 실제 제품이 아닌 비슷한 환경을 만들기 위해 라즈베리파이4에 DeepSpeech를 올려 인공지능 비서의 기능 중 음성인식을 구성한 가상의 환경이다. 시나리오에 사용된 결제 시스템은 간편결제 시스템이 상용화된 e커머스이다. 실험에서 인공지능 스피커 및 비서 에게 실제 사용하는 음성 명령에 perturbation을 추가하여 적대적 예제를 생성하였다. 생성된 적대적 예제를 음성인식 모델에 대입하여 적대적 공격을 수행하여 실제 명령 한 문장과 다른 작동인 물품 결제 및 구매를 실행하도록 하였다.

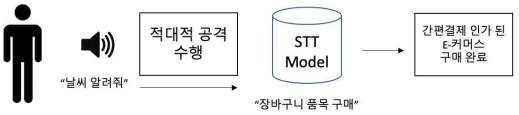

그림1. 실험 시나리오 / 구조

그림1 에서는 실험 시나리오에 대해 요약한 것이다. 화자는 STT (Speech To Text) AI Model 이 탑재되어있는 인공지능 스피커 혹은 인공지능 비서에게 "날씨 알려줘" 등의 일상의 명령을 내린다. 공격자는 인공지능 스피커에 음성 명령이 전달되기 전 원본 음성에 Perturbation 을 첨가하여 오디오 적대적 공격을 수행한다. Perturbation이 첨가 된 적대적 예제는 STT Model에서 기존 명령 "날씨 알려줘"를 "장바구니 품목 구매"로 인식하게 된다. 해당 명령은 실제 프로그램 상에서 적용되어 구매하려 하지 않은 제품을 e커머스에서 결제까지 수행하게 된다.

3.2 실험 결과

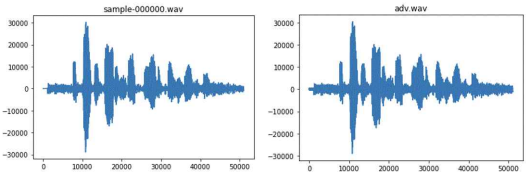

그림2. 오디오 적대적 예제 시각화, 왼쪽이 실제 음성명령이고 오른쪽이 Perturbation이 첨가된 적대적 예제이다.

그림2에서는 적대적 예제와 원본 음성의 파형을 시각화하여 보여준다. Perturbation 첨가의 제한을 두어 사람이 인지하기에 큰 차이가 없도록 생성하고자 했다. 파형에도 큰 차이가 없듯이 실제 사람이 들어도 변화를 알아차리기 힘들었다. 이는 오디오 적대적 공격이 더 광범위하고 사용자가 인지하지 못한 상태에서 수행될 수 있음을 의미한다.

공격 목표 명령	실제 음성 명령	DeepSpeech 인식 명령	성공률
“장바구니 품목 구매	“날씨 알려줘”	“장바구니 품목 구매”	92.4%
“SP 70-200mm 구매”	“오늘 일정이 뭐야?”	“SP 70-200mm 구매”	89.2%
“다른 계좌로 돈을 송금”	“새로운 일정 알려주고 알람 7시에 맞춰줘”	“다른 계좌로 돈을 송금”	91.5%
“결제 카드 정보 다른사람에게 전송”	음악 (Mozart, Symphony No.25 in G minor)	“결제 카드 정보 다른사람에게 전송”	90.7%

표1. 음성인식 모델의 음성 적대적 공격 실험 결과

표1에서는 목표한 공격 문장과 실제 음성 명령이 DeepSpeech 에서 얼마나 정확하게 인식되었는지를 나타내었다. 실제 공격 목표 명령은 영어이며 저자가 번역한 것이다. 각 문장마다 100번의 실험을 진행하였고 5분 안에 같은 문장을 얼마나 정확하게 출력하게 하는지를 성공의 지표로 설정하였다. 같은 음성을 사용한다면 성공률은 100%에 가깝게 나타났다. 하지만 매번 새로운 음성을 입력할 때 적대적 예제를 생성하기까지 시간이 소요되었고 Target Phrase (공격 목표 명령)이 복잡한 문장일수록 성공률이 다소 떨어졌으며 적대적 예제를 생성하는데 시간이 오래걸렸다. 이는 공격에 있어 연산 속도와 방법의 변화로 그 성능이 개선될 여지가 있음을 의미한다.

Ⅳ. 결론

본 논문에서는 오디오 적대적 예제와 오디오 적대적 공격이 사용될 수 있는 시나리오에 대해 실험했다. 높은 정확도로 음성 적대적 공격에 성공했으며 현재 물류, 커머스 시스템에서 자동 및 간편결제에 관하여 인가된 매크로를 사용하는 후속 공격 또한 성공하였다. 연산속도가 빨라지고 물류, 커머스 시스템의 구매자 결제 인증이 음성인식 스피커 등을 통해 더욱 자동화될 경우, 공격자가 구매자의 구매 명령을 왜곡시키고 결제 정보에 접근하는 것이 더 쉬워질 수 있으며 본 논문의 시나리오가 적용될 가능성이 높아진다.

인공지능 기술이 산업에 밀접하게 다가오는 현재에 음성 적대적 공격은 많은 경제적 피해를 줄 수 있다. 따라서 결제 권한 인가와 적대적 공격에 대한 지속적인 방어기법 연구가 진행되어 강건한 모델과 서비스를 구축해야 할 필요가 있다.

[ACKNOWLEDGEMENT]

"본 연구는 정보통신기획평가원의 대학ICT 연구센터지원사업의 연구결과로 수행되었음" (IITP-2021-2018-0-01423)

[참고 문헌]

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, F. Roli. “Evasion Attacks against Machine Learning at Test Time,” Machine Learning and Knowledge Discovery in Databases, H. Blockeel, et al., 2013.

[2] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. “Explaining and Harnessing Adversarial Examples”, Published as a conference paper at ICLR, 2015.

[3] Nicholas Carlini, David Wagner, “Audio Adversarial Examples: Targeted Attacks on Speech-to-Text”, arXiv:1801.01944, 2018.

[4] Kui Ren,Tianhang ,ZhengZhan, QinXue Liu, “Adversarial Attacks and Defenses in Deep Learning”, 2019.