Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.1, Feb. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.1.89

질의 효율적인 의사 결정 공격을 통한 오디오 적대적 예제 생성 연구*

서 성 관,[1†] 문 현 준,[1] 손 배 훈,[1] 윤 주 범[2‡]
[1,2]세종대학교 (대학원생, 교수)

Generating Audio Adversarial Examples Using a Query-Efficient Decision-Based Attack*

Seong-gwan Seo,[1†] Hyunjun Mun,[1] Baehoon Son,[1] Joobeom Yun[2‡]
[1,2]Sejong University (Graduate student, Professor)

요 약

딥러닝 기술이 여러 분야에 적용되면서 딥러닝 모델의 보안 문제인 적대적 공격기법 연구가 활발히 진행되었다. 적대적 공격은 이미지 분야에서 주로 연구가 되었는데 최근에는 모델의 분류 결과만 있으면 공격이 가능한 의사 결정 공격기법까지 발전했다. 그러나 오디오 분야의 경우 적대적 공격을 적용하는 연구가 비교적 더디게 이루어지고 있는데 본 논문에서는 오디오 분야에 최신 의사 결정 공격기법을 적용하고 개선한다. 최신 의사 결정 공격기법은 기울기 근사를 위해 많은 질의 수가 필요로 하는 단점이 있는데 본 논문에서는 기울기 근사에 필요한 벡터 탐색 공간을 축소하여 질의 효율성을 높인다. 실험 결과 최신 의사 결정 공격기법보다 공격 성공률을 50% 높였고, 원본 오디오와 적대적 예제의 차이를 75% 줄여 같은 질의 수 대비 더욱 작은 노이즈로 적대적 예제가 생성 가능함을 입증하였다.

ABSTRACT

As deep learning technology was applied to various fields, research on adversarial attack techniques, a security problem of deep learning models, was actively studied. adversarial attacks have been mainly studied in the field of images. Recently, they have even developed a complete decision-based attack technique that can attack with just the classification results of the model. However, in the case of the audio field, research is relatively slow. In this paper, we applied several decision-based attack techniques to the audio field and improved state-of-the-art attack techniques. State-of-the-art decision-attack techniques have the disadvantage of requiring many queries for gradient approximation. In this paper, we improve query efficiency by proposing a method of reducing the vector search space required for gradient approximation. Experimental results showed that the attack success rate was increased by 50%, and the difference between original audio and adversarial examples was reduced by 75%, proving that our method could generate adversarial examples with smaller noise.

Keywords: Adversarial examples, Speech command classification, Decision-based attack

Received(12. 15. 2021), Modified(01. 20. 2022), Accepted(01. 20. 2022)
* 본 연구는 한국연구재단 연구과제(NRF-2021R1F1A1051251) 지원으로 수행하였습니다.
† 주저자, ths3847@gmail.com
‡ 교신저자, jbyun@sejong.ac.kr(Corresponding author)

I. 서 론

현재 딥러닝 기술은 이미지 인식, 음성 인식, 자연어 처리 등 다양한 분야에서 우수한 성능을 내고 있고 이에 따라 자율주행차, AI 스피커 등과 같이 실제 환경에 상용화되어 사용자에게 다양한 편의를 제공하고 있다. 그러나 딥러닝 기술의 상용화와 더불어 딥러닝 기술 자체의 보안 문제 또한 큰 이슈가 되고 있는데 딥러닝 모델의 잘 알려진 보안 문제로 적대적 공격이 존재한다. 적대적 공격이란 원본 데이터에 사람이 인식할 수 없는 노이즈를 추가하여 딥러닝 모델의 오인식을 일으키는 공격기법이다. 적대적 공격은 공격자에게 주어지는 정보, 공격 목표에 따라 공격 유형이 나뉜다. 먼저 공격자에게 주어지는 정보에 따라 타겟 딥러닝 모델의 정보를 모두 안다면 화이트 박스 공격, 아무 정보도 알 수 없다면 블랙박스 공격으로 나뉜다. 그리고 공격 목표에 따라 공격자가 딥러닝 모델을 특정 클래스로 오인식하게 한다면 타겟 지정 공격(targeted attack), 원본 데이터와 다른 아무 클래스로 오인식하게 한다면 타겟 미지정 공격(untargeted attack)으로 나뉜다. 적대적 공격은 주로 이미지 분야에서 활발히 연구되었는데 화이트 박스 공격의 경우 한 번의 연산으로 빠르게 공격을 수행하는 FGSM 공격[1]부터 정교한 최적화 식을 통해 높은 공격 성공률을 달성한 C&W 공격[5] 등이 연구되었고, 블랙박스 공격의 경우 분류 점수를 이용하는 점수 기반(score-based) 공격부터 모델의 분류 결과만을 이용한 의사 결정 기반(decision-based) 공격까지 다양한 공격 방법이 연구되었다. 의사 결정 기반 공격으로 대상 딥러닝 모델의 결정 경계를 근사하여 적대적 예제의 전이성을 이용한 transfer attack[10]과 목표 클래스로 분류되는 적대적 예제를 원본 데이터와 유사하게 만드는 Boundary attack[11] 등이 있다. 그러나 오디오 분야에서 적대적 예제는 비교적 연구가 더디게 진행되고 있는데 이는 오디오 데이터가 이미지 데이터와 달리 데이터 전처리 작업이 필요하고 모델의 구조가 비교적 복잡하여 공격 난이도가 높기 때문이다. 오디오 분야에서 블랙박스 공격의 경우 유전 알고리즘을 이용한 공격 기법[14, 15] 등의 점수 기반 공격기법 외에 의사 결정 기반 공격 연구가 진행되지 않고 있다. 따라서 본 논문에서는 완전한 블랙박스 환경의 공격인 의사 결정 기반 공격기법을 오디오 분야에 적용한다. 여러 의사 결정 기반 공격 중 초기 공격기법인 Boundary Attack[11]과 최신 공격기법인 HopSkipJumpAttack[12]을 적용하였다. 또한 HopSkipJumpAttack[12]의 많은 질의 수가 필요한 단점을 개선하기 위해 방향 벡터 탐색 공간 축소 방법을 제안하고 이후 실험을 통해 본 논문에서 제안한 방법이 질의 효율성 및 공격 성능을 높일 수 있음을 입증하였다.

II. 관련 연구

2.1 의사 결정 기반 공격

의사 결정 기반 공격은 타겟 딥러닝 모델의 어떠한 정보도 필요 없이 분류되는 클래스만 알면 공격이 가능하다. 따라서 가장 현실적인 공격기법이라고 볼 수 있다. 초기 의사 결정 기반 공격으로 적대적 예제의 전이성을 이용한 transfer attack[10]이 연구되었는데 transfer attack[10]은 타겟 딥러닝 모델과 유사한 대체 모델을 학습하여 대체 모델에서 적대적 예제를 생성한 후 타겟 딥러닝 모델을 공격하는 기법이다. 대체 모델의 결정 경계를 타겟 딥러닝 모델의 결정 경계와 유사하게 학습해야 하므로 많은 질의 수가 필요하고 또한 학습 데이터에 의존적이다. 그에 비해 Boundary Attack[11]은 대체 모델을 이용하지 않고 적대적 예제를 생성할 수 있어 학습 데이터의 의존 없이 공격 가능하므로 더욱 현실적인 공격기법이라고 볼 수 있다.

2.2 Boundary Attack[11]

Boundary Attack[11]은 대표적인 의사 결정 기반 공격으로 공격 목표 클래스로 분류되는 데이터를 적대적 예제로 초기화하여 분류는 유지하되 원본 데이터와의 차이를 점차 줄이는 방식으로 적대적 예제를 수정한다. Boundary Attack[11]은 기각 샘플링 (Rejection sampling)을 이용하여 적대적 예제를 생성하는데 Fig.1.과 같이 반복마다 두 가지 단계를 수행하여 원본과 유사한 적대적 예제를 생성한다. 첫 번째 단계(#1)에서는 x_t에 랜덤한 노이즈를 추가한 후에 x_t와 x_{orig}의 사이의 거리 반경 안으로 들어오도록 사영시킨다. 다음 단계(#2)에서는 적대적 예제를 x_{orig} 가까워지도록 이동시키는데 만약 타겟 분류가 변하는 경우 즉 공격이 실패하면 이

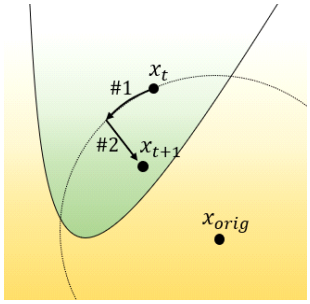


Fig. 1. Intuitive description of Boundary Attack

전 데이터 x_t로 돌아간다. 랜덤하게 노이즈를 추가하고 분류가 변하는 데이터는 버리는 방식이기 때문에 적대적 예제를 만드는데 많은 질의 수가 필요한 단점이 있다. 그리고 랜덤한 노이즈를 추가하는 방식이기 때문에 적대적 예제가 최적의 해로 수렴하는 것을 보장할 수 없는 단점 또한 존재한다.

2.3 HopSkipJumpAttack[12]

HopSkipAttack[12]은 Boundary Attack[11]의 질의 수가 많이 필요한 단점을 개선한 의사 결정 기반 공격기법이다. 본 논문에서는 HopSkipAttack[12]의 공격 알고리즘을 개선하였으므로 HopSkipAttack[12]을 자세히 설명한다. HopSkipAttack[12]은 화이트 박스 공격과 유사하게 목적 함수를 정의하고 최적화를 통해 질의 수를 줄인다. 최적화 작업을 수행하기 위해서 기울기를 구해야 하는데 HopSkipAttack[12]은 결정 경계 위의 데이터에 여러 방향의 노이즈를 추가하여 질의 한 결과를 통해 기울기를 예측한다. 공격 성공 여부에 관계없이 질의 결과를 기울기 예측에 이용하기 때문에 공격 실패 시 질의 결과를 버리는 Boundary Attack[11] 보다 효율적이다. Fig.2.에서 HopSkipJumpAttack[12]의 공격 방법을 직관적으로 확인할 수 있다. 먼저 타겟 클래스로 분류되는 데이터를 적대적 예제 $\widetilde{x_t}$로 시작하여 (#1) 단계에서 이진 탐색을 통해 적대적 예제를 결정 경계 위에 x_t로 위치시킨다. 그런 다음 (#2) 단계에서 x_t에서의 기울기를 예측하고 (#3) 단계에서 (#2) 단계에서 구한 기울기의 방향으로 분류가 변경되지 않을 만큼 $\widetilde{x_{t+1}}$로 이동한다. 마지막으로 (#4) 단계와 같이 적대적 예제가 원본 이미지와 가까워지도록 (#1)~(#3) 단계를 반복한다. HopSkipAttack의 자세한 공격 과정을 수식으로 표현하면 다음과 같다.

$$S_{x^*}(x') = F_{c^\dagger}(x') - \max_{c \neq c^\dagger} F_c(x') \tag{1}$$

$$sign(S_{x^*}(x')) = \begin{cases} 1 & (\text{if } S_{x^*}(x') > 0), \\ -1 & (otherwise) \end{cases} \tag{2}$$

$$\phi_{x^*}(x') = sign(S_{x^*}(x')) \tag{3}$$

수식 (1)에서 x^*은 원본 데이터, x'은 적대적 예제, $C^\dagger$은 목표 클래스이다. (1)은 화이트 박스 공격에 이용되는 것과 유사한 목적 함수로 $S_{x^*}(x') > 0$인 경우에 공격이 성공한다. 수식 (2)는 x'를 타겟 클래스로 분류한 경우 1, 그렇지 않은 경우 -1을 출력하는 함수로 블랙박스 환경이라도 분류 결과에 따라 수식 (2)의 값을 얻을 수 있다.

수식 (2)를 간단하게 수식 (3)으로 표현한다.

$$\min_{x'} d(x', x^*) \ \ such\ that \ \ \phi_{x^*}(x') = 1 \tag{4}$$

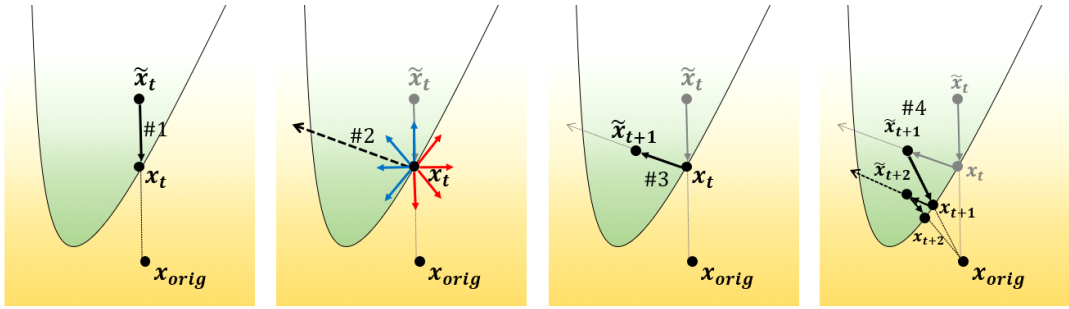


Fig. 2. Intuitive description of HopSkipJumpAttack

여기서 적대적 예제는 타겟 클래스로 분류되면서 원본 데이터와의 차이가 최소화되어야 하므로 수식(4)와 같은 최적화 문제로 정의될 수 있다.

$$x_{t+1} = \alpha_t x^* + (1-\alpha_t)\left\{x_t + \xi_t \frac{\nabla S_{x^*}(x_t)}{\| \nabla S_{x^*(x_t)} \|_2}\right\} \tag{5}$$

우선 ∇S_{x^*}에 접근할 수 있다는 가정하에 수식(5)를 반복하여 최적화 문제 (2)를 해결하는데 이는 Fig. 2.의 공격 과정과 같다. 현재 반복에서의 적대적 예제 x_t에서의 기울기 $\nabla S_{x^*}(x_t)$를 구하는 것은 (#2)단계와 같고, $\nabla S_{x^*}(x_t)$의 방향에 ξ_t 만큼 곱하여 x_t에 더하는 것은 (#3) 단계와 같다. 여기서 ξ_t는 적대적 예제의 분류가 변하지 않도록 조절해 주는데 단순히 분류가 변하는 경우 반씩 나눠 크기를 줄인다. 마지막으로 (#1)단계와 같이 적대적 예제를 결정 경계 위로 옮기는데, 원본 데이터 x^*와 $\left\{x_t + \xi_t \frac{\nabla S_{x^*}(x_t)}{\| \nabla S_{x^*(x_t)} \|_2}\right\}$ 사이의 결정 경계에 위치하는 x_{t+1}을 구한다. 이때 이진 탐색을 이용하여 적절한 값 α_t를 찾는다. 수식(5)의 경우 ∇S_{x^*}에 접근할 수 있다는 가정에서 공격이 이루어진다. 그러나 결정기반 공격 환경에서는 ∇S_{x^*}에 접근할 수 없으므로 HopSkipAttack[12]은 x_t가 결정 경계 위에 있다는 전제하에 수식(6)과 같이 몬테카를로 알고리즘을 이용하여 ∇S_{x^*}를 근사한다.

$$\widetilde{\nabla S}(x_t, \delta) = \frac{1}{B}\sum_{b=1}^{B} \phi_{x^*}(x_t + \delta u_b) u_b \tag{6}$$

여기서 u_b는 랜덤한 값, δ는 매우 작은 크기의 파라미터 값이다. x_t가 결정 경계 위에 있다는 전제하에 $\phi_{x^*}(x_t + \phi u_b)$의 결과에 따라 공격이 성공하는 방향은 $+1$, 실패하는 방향은 -1가 곱해지고 최종적으로 방향들의 평균을 구하므로 공격 성공 방향을 예측할 수 있다. 그러나 실제 환경에서 이진 탐색으로 x_t를 정확히 결정 경계에 올리는 것은 불가능하다. x_t가 결정 경계로부터 멀리 있는 경우 $\phi_{x^*}(x_t + \phi u_b)$의 결과가 모두 $+1$이 되는 문제가 발생하는데 HopSkipJumpAttack[12]은 수식(7)을 통해 구한 기준점 $\overline{\phi_{x^*}}$을 수식 (8)과 같이 $\phi_{x^*}(x_t + \phi u_b)$에 빼줌으로 u_b에 곱해지는 값이 모두 $+1$이 되는 것을 막는다.

$$\overline{\phi_{x^*}} = \frac{1}{B}\sum_{b=1}^{B} \phi_{x^*}(x_t + \delta u_b) \tag{7}$$

$$\widehat{\nabla S}(x_t, \delta) = \frac{1}{B-1}\sum_{b=1}^{B} (\phi_{x^*}(x_t + \delta u_b) - \overline{\phi_{x^*}}) u_b \tag{8}$$

III. 오디오 적대적 예제

3.1 오디오 데이터와 이미지 데이터 차이

이미지 데이터의 경우 픽셀값의 차이만으로 데이터를 구별하지만, 오디오 데이터의 경우 소리의 크기를 구별하는 진폭(amplitude)과 소리의 높낮이를 구별하는 주파수가 존재하여 이러한 특징을 반영하기 위해서 전처리 과정이 필요하다. 원시 오디오 파형의 경우 시간에 따른 진폭의 변화로 나타내는데 오디오 파형은 여러 주파수가 섞여 있는 고차원 데이터이므로 딥러닝 모델에 이용 시 오디오의 특징이 제대로 반영되기 어렵다. 따라서 주파수에 따른 진폭의 변화를 나타내기 위해 푸리에 변환을 이용하여 오디오 파형을 스펙트로그램(spectrogram)으로 바꿔주고 추가로 딥러닝 모델의 성능 개선을 위해 인간의 청각 시스템을 모방하여 중요한 특징만 추출하는 MFCC 등 전처리 과정이 이용되어 진다. 데이터 전처리 과정이 존재하는 오디의 특성에 따라 적대적 예제를 오디오 파형, 스펙트로그램(spectrogram) 혹은 MFCC 단계에서 각각 생성할 수 있다. 원시 데이터에 가까울수록 전처리 과정을 고려하여 공격해야 하므로 난이도가 높다[21].

3.2 오디오 적대적 예제 생성

본 논문에서 제안한 공격기법은 원시 음성 파일을 음성 분류 모델에 입력할 수 있는 디지털 오디오 파형으로 디코딩한 후 이용한다. 즉 공격자는 음성 분류 모델에 맞게 디코딩된 오디오 데이터에 접근할 수

있는 가정하에 공격을 수행한다. 공격자는 디코딩된 오디오 데이터를 오디오 분류 모델에 입력하고 그에 해당하는 출력을 확인하여 적대적 예제를 생성한다. 이는 음성 분류 모델이 디지털 오디오 파형을 스펙트로 그램, MFCC 등으로 어떻게 전처리하는지 알 수 없으므로 충분히 제한적인 블랙박스 환경이다. 그러나 후속 연구에서 더욱 엄격한 블랙박스 환경을 가정한다면 일반적으로 이용되는 sampling rate를 이용하여 원시 음성 파일을 디코딩하고 적대적 예제를 생성한 다음 다시 원시 음성 파일로 변환하여 음성 분류 모델에 입력하는 방법을 고려해볼 수 있다. 추가로 실제 음성 인식 기기에 대해 적대적 예제를 수행하는 경우 스피커 및 녹음 공간에서 생기는 잡음을 고려하여야 한다. 이를 해결하기 위해 이미지 분야에서 실제 환경 노이즈를 고려하여 물리적인 적대적 예제 생성을 한 EOT[7] 기법을 응용하는 것을 고려할 수 있다. 화이트 박스 오디오 적대적 예제 생성 연구의 경우 실제 환경 노이즈와 유사한 노이즈를 고려하여 적대적 예제를 생성하는 연구[20]가 진행되고 있다.

3.3 방향 벡터 탐색 공간 축소

HopSkipJumpAttack[12]은 이론적으로 공격이 성공함을 증명했지만, 실험 과정에서 예측한 기울기와 실제 기울기의 차이가 있음을 발견할 수 있었다. 본 논문에서 사용하는 오디오 데이터는 16,000 차원으로 이루어져있는데 Fig. 2. (#2)와 같이 몬테

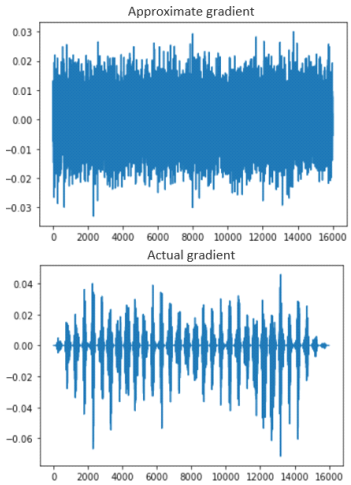


Fig. 3. Comparing approximate gradient with Actual gradient

카를로 알고리즘을 통해 정확한 기울기를 예측하는 것은 무한히 질의를 하지 않는 한 불가능할 것이다.

실제로 반복 당 7~800번의 질의를 이용해 기울기를 예측했지만 Fig. 3.과 같이 정확한 기울기를 구할 수 없었다. 그런데도 HopSkipJumpAttack[12]이 Boundary Attack[11]과 비교하여 공격 성능이 좋았던 이유는 공격이 성공하는 대략적인 방향을 알 수 있기 때문이라 판단된다. 이러한 결과는 이미지에서도 마찬가지였는데 정확한 기울기를 구할 수 없음에도 많은 질의를 소모하는 것은 매우 비효율적이다. 따라서 본 논문에서는 기울기 예측 시에 이용하는 랜덤 방향 벡터의 탐색 공간을 줄여 비교적 적은 질의로 대략적인 방향을 예측하는 새로운 알고리즘을 제안한다. Fig. 4.과 같이 x_t - x_{orig}와 직교하는 범위 내에서 랜덤 방향 벡터를 선택하므로 전체 공간에서 랜덤 방향 벡터를 추출하는 기존 방식의 질의 수를 줄인다.

$$u = u - proj_{x_t - x^*} u \tag{9}$$

정확히 랜덤 벡터 u를 선택한 다음 수식 (9)와 같이 x_t - x_{orig}와 수직이 되도록 처리해 주었다. 이후 알고리즘은 기존 방식과 마찬가지로 수식 (6)과 같이 공격이 성공하는 방향은 +1, 실패하는 방향은 -1을 곱해 더한 값의 평균을 통해 대략적인 방향을 구하고, 수식 (5)와 같이 적대적 예제를 결정 경계 위로 옮기는 과정을 반복한다. 전체적인 알고리즘은 Fig. 5.에서 확인할 수 있다. 실험 과정에서 각 반복에서 기울기를 예측하는 데 이용하는 질의를 줄이는 만큼 전체 반복 수를 증가시켜 실험을 진행하였다. HopSkipJump Attack[12]의 경우 반복 수가 늘어날수록 보다 정확한 기울기 예측을 위해 더 많은 랜덤 방향 벡터를 이용하는데 본 논문에서는 반복마다 랜덤 방향 벡터를 100개로 고정하여 실험하였다.

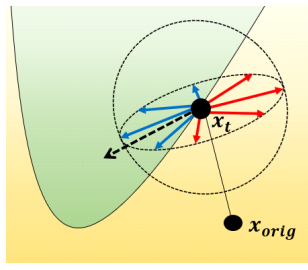


Fig. 4. Reducing direction vector search space using orthogonal vectors

Algorithm Generating audio AEs.
Inputs: Classifier C, Original audio x^*, Initial adversarial example $\tilde{x}_0$, Iterations T
Outputs: The adversarial example x_t

for t in 1,2,...,T-1 do
  # Boundary search
  x_t=Bin-search($\tilde{x}_{t-1}$, x^*) (수식 5)
  # Reducing the search space
  for i in 100 do
    Set random vector u_i
    $u_i = u_i - proj_{x_t - x^*} u_i$
  end for
  # Gradient estimation
  $\widetilde{\nabla S}$ = Grad-estimation(x^*, x_t, U) (수식 8)
  $\tilde{x}_t = x_t + \widetilde{\nabla S} / \| \widetilde{\nabla S} \|$
end for
return x_t=Bin-search($\tilde{x}_{t-1}$, x^*)

Fig. 5. Algorithm for generating audio AEs.

3.4 오디오 환경에서 노이즈 측정

이미지 적대적 예제는 원본 이미지와 적대적 예제의 차이인 노이즈를 측정하는 방법으로 $L_p norm$을 이용한다. 이미지 데이터의 경우 $L_p norm$을 이용하여 사람의 인식과 유사하도록 수치화할 수 있지만, 오디오 데이터의 경우 같은 $L_p norm$을 갖는 노이즈라도 사람의 인식 정도가 다를 수 있어 완벽한 측정 방법이라고 볼 수 없다. 따라서 기존 연구에서는 전통적으로 오디오 노이즈를 측정하는 방법인 Signal to Noise Ratio(SNR)을 추가로 이용하였다. SNR은 수식 (11)과 같이 원본 오디오의 신호 크기가 노이즈 신호 크기 보다 얼마나 큰지를 표현한다. 즉 신호 대 잡음비의 값이 클수록 뛰어난 공격 성능을 낸다고 볼 수 있다. 본 연구에서는 신호 크기로 root mean square(rms)를 이용했다.

$$\| x \|_p = \left(\sum_{k=1}^{n} x_k{}^2 \right)^{1/p} \quad (10)$$

$$SNR_{dB} = 20 \log_{10} \left(V_{x^*} / V_{\delta} \right) \quad (11)$$

IV. 실험 및 평가

4.1 실험 환경

본 연구에서는 텐서플로에서 제공하는 CNN 기반 음성 명령어 분류 모델[22]을 대상으로 공격을 실험한다. 실험에 사용되는 데이터는 음성 명령어 데이터[23]로 30개의 다른 단어를 사람들이 말하는 105,000개의 오디오 파일이다. 각각의 파일은 1초의 한 단어 오디오 클립으로 "down", "go", "left", "no", "off", "on", "right", "stop", "up", "yes" 10개의 단어를 선택하여 실험을 진행하였다.

4.2 실험 결과

Boundary Attack[11]과 HopSkipJumpAttack[12]을 이용하여 CNN 기반 음성 명령어 분류 모델[22]에 대한 L_2 기반 타겟 적대적 공격을 적용하였다. 기존 이미지 분야에 적용했던 것과 마찬가지로 질의 수를 25,000번으로 제한하고 각 클래스별로 50번의 타겟 공격을 수행하였다. 그 결과 Table 1., Tabe 2., Table 3. 와 같이 대부분 두 공격기법 모두 적대적 예제를 원본 오디오와 유사하게 생성하는 것을 확인할 수 있었다. 그러나 Boundary Attack[11]과 HopSkipJumpAttack[12] 성능 비교 시 이미지 분야와 달리 공격 성능 차이가 크지 않은데 이는 실험 대상 모델의 구조 차이와 데이터 전처리 단계의 영향 때문이다. CNN 기반 음성 명령어 분류 모델[22]의 경우 초기 입력 값을 32x32 크기로 줄여서 입력받는데 이때 입력 오디오 정보가 손실되어 기울기 예측이 더욱 어려워진다. 본 논문에서 제안한 방법을 통해 생성한 적대적 예제 결과를 확인해보면 성공률, L_2 distance, SNR 값 모두 확연한 차이로 공격 성능이 뛰어난 것을 볼 수 있다. 의사 결정 기법은 적대적 예제에서 시작하여 분류를 유지하며 원본 데이터와 유사하도록 변경하는 방식이기 때문에 성공률의 경우 L_2 distance를 기준으로 1

Table 1. Mean values of success rate

Success rate			
	BA[11]	HSJA[12]	Ours
Iterations	1,800	60	200
Success rate	44%	45%	95%

Table 2. L_2 distance on original label for targeted adversarial examples

L_2 distance			
Orinal Label	BA[11]	HSJA[12]	Ours
"down"	1.22	1.25	0.29
"go"	1.16	1.12	0.32
"left"	1.49	1.28	0.26
"no"	1.04	1.19	0.29
"off"	2.85	1.78	0.53
"on"	1.40	1.17	0.28
"right"	1.98	1.32	0.32
"stop"	2.13	2.01	0.52
"up"	2.56	2.05	0.48
"yes"	1.69	2.06	0.43
avg.	1.62	1.52	0.37

Table 3. Mean values of *SNR* on original label for targeted adversarial examples

SNR			
Orinal Label	BA[11]	HSJA[12]	Ours
"down"	21.34dB	18.38dB	29.99dB
"go"	18.29dB	20.53dB	31.47dB
"left"	16.24dB	18.49dB	30.74dB
"no"	18.43dB	18.89dB	29.73dB
"off"	15.13dB	17.24dB	29.34dB
"on"	16.45dB	18.43dB	29.71dB
"right"	12.38dB	16.42dB	27.23dB
"stop"	13.50dB	15.29dB	27.18dB
"up"	15.50dB	18.71dB	30.01dB
"yes"	14.08dB	13.63dB	25.31dB
avg.	16.13dB	17.60dB	29.07dB

보다 작은 경우 성공으로 측정했다. Table 1.에 성공률의 경우 이전 공격 들[11], [12]과 비교하여 약 50%의 성능 향상을 보였다. 반복 수를 비교하여 결과를 확인해보면 앞서 가정한 대로 HopSkipJump Attack[12]의 성능은 정확한 기울기 예측보다 반복 수에 비례한다는 것을 알 수 있다. 여기서 적절한 방향 예측 없이 반복 수만 늘리는 방법을 생각해볼 수 있는데, 완전히 랜덤한 방향으로 적대적 예제를 발전시키는 Boundary Attack[11]의 경우 반복 수가 많아도 적대적 예제가 최적의 해로 수렴을 하지 못하기 때문에 적대적 예제를 최적의 해로 수렴하기 위해 적절한 방향으로 이동시키는 방법이 효율적이라는 것을 알 수 있다. L_2 distance의 경우 HopSkipJump Attack[12]과 비교하여 적대적 예제와 원본 오디오의 차이를 약 75%를 감소시켰고, SNR의 경우 타겟 적대적 예제를 평균 SNR 값이 29dB인데 이는 데이터셋의 차이는 있지만 기존 화이트 박스 공격[16, 17, 19]의 SNR 값이 14dB~21dB로 화이트 박스 공격과 비교하여도 높은 성능을 나타낸다.

V. 결 론

본 연구에서는 오디오 분야에 의사 결정 기반 공격기법을 적용하여 성능을 비교하였다. 그리고 실험 과정에서 HopSkipJump Attack[12]의 문제점을 발견하고 방향 벡터 탐색 공간 축소 방법을 제안하여 공격 성능을 향상했다. 본 연구 결과를 통해 향후 연구에서 문장 단위의 오디오 데이터를 다루는 음성 인식 모델(ASR, Automatic Speech Recognition)에 해당 공격기법을 적용하는 연구로 확장될 수 있을 것이다. 그리고 실제 환경 노이즈를 고려하여 실제 음성 인식 기기를 대상으로 한 적대적 생성 연구로 확장할 수 있을 것이다. 또한 본 논문에서는 오디오 데이터 자체의 특징이 아닌 HopSkipJump Attack[12]의 문제점을 개선하였는데 이후 오디오 데이터의 특징인 심리음향 특성[19]을 이용하여 사람이 인식하지 못하는 정교한 노이즈를 생성하는 연구가 가능할 것이다.

References

[1] C. Szegedy, W. Zaremba, and I. Sutskever, "Intriguing properties of neural networks." International Conference on Learning Representations (ICLR), 2014.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples." International Conference on Learning Representations (ICLR), 2015.

[3] N. Papernot, P McDaniel, and S. Jha et al., "The limitations of deep learning in adversarial settings" 2016 IEEE European symposium on security and privacy (EuroS&P), pp. 372-387, Mar. 2016.

[4] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks" 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2574-2582, Jun 2016.

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

[6] A. Athalye and I. Sutskever, "Synthesizing Robust Adversarial Examples," International Conference on Machine Learning (ICML), pp. 284-293, Jul. 2018.

[7] A. Athalye, N. Carlini, and D. Wagner, "Synthesizing Robust Adversarial Examples" International Conference on Machine Learning (ICML), pp. 274-283, Jul. 2018.

[8] J. Su and D. Vasconcellos et al., "One pixel attack for fooling deep neural networks", VOL. 23, NO. 5, pp. 828-841, Oct. 2019.

[9] P. Chen, Huan Zhang, and Y. Sharma et al., "ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models" 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 15-26, Nov. 2017.

[10] N. Papernot, P. McDaniel, I. Goodfellow et al. "Practical Black-Box Attacks against Machine Learning" 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS'17, PP. 506-519, Apr. 2017.

[11] W. Brendel, J. Rauber, and M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models" International Conference on Learning Representations, May 2018.

[12] J. Chen, M. Jordan, and M. Wainwright, "HopSkipJumpAttack: A Query-Efficient Decision-Based Attack" 2020 IEEE Symposium on Security and Privacy (SP), pp. 1277-1294, May 2020.

[13] N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text" 2018 IEEE Security and Privacy Workshops (SPW), pp. 1-7, May 2018.

[14] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial Examples Against Automatic Speech Recognition" Machine Deception Workshop, Neural Information Processing Systems (NIPS), 2017.

[15] R. Taori, A. Kamsetty, and B. Chu et al.,"Targeted Adversarial Examples for Black Box Audio Systems" 2019 IEEE Security, and Privacy Workshops (SPW), pp. 15-20, May 2019.

[16] X. Yuan, Y. Chen, and Y. Zhao et al., "Commandersong: A systematic approach for practical adversarial voice recognition" 27th USENIX Security Symposium, pp. 49-64, Aug. 2018.

[17] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep Learning and Music Adversaries," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2059-2071, Nov. 2015.

[18] H. Abdullah, W. Garcia, and C. Peeters et al., "Practical hidden voice attacks against speech and speaker recognition systems" NDSS 2019, pp. 1369 - 1378, 2019.

[19] L. Schonherr, K. Kohls, and S. Zeiler et al., "Adversarial Attacks Against A

utomatic Speech Recognition Systems via Psychoacoustic Hiding" NDSS 2019.

[20] H. Yakura and J. Sakuma, "Robust Audio Adversarial Example for a Physical Attack" 28th International Joint Conference on Artificial Intelligencepp. 5334-5341, 2019.

[21] S. Hu, X. Shang, and Z. Qin et al., "Adversarial Examples for Automatic Speech Recognition: Attacks and Countermeasures" IEEE Communications Magazine, pp. 120-126, Oct. 2019

[22] Simple audio recognition: Recognizing keywords. Accessed: September 13, 2021. [Online]. Available: https://www.tensorflow.org/tutorials/audio/simple_audio

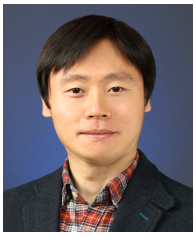
[23] P. Warden "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition" arXiv preprint arXiv:1804.

〈저자소개〉

서 성 관 (Seong-gwan Seo) 학생회원
2019년 8월: 세종대학교 정보보호학과 졸업
2020년 3월~현재: 세종대학교 정보보호학과, 지능형드론 융합전공 석사과정
〈관심분야〉 정보보호, 인공지능

문 현 준 (Hyunjun Mun) 학생회원
2020년 2월: 세종대학교 정보보호학과 졸업
2020년 3월~현재: 세종대학교 정보보호학과, 지능형드론 융합전공 석사과정
〈관심분야〉 정보보호, 인공지능

손 배 훈 (Baehoon Son) 학생회원
2021년 2월: 세종대학교 정보보호학과 졸업
2021년 3월~현재: 세종대학교 정보보호학과, 지능형드론 융합전공 석사과정
〈관심분야〉 정보보호, 인공지능



윤 주 범 (Joobeom Yun) 종신회원
1999년 2월: 고려대학교 컴퓨터학과 학사
2001년 2월: 서울대학교 컴퓨터공학과 석사
2012년 2월: KAIST 전산학과 박사
2001년 3월~2015년 2월: ETRI부설연구소 선임연구원
2015년 3월~현재: 세종대학교 정보보호학과, 지능형드론 융합전공 부교수
〈관심분야〉 네트워크 보안, 시스템 보안, 인공지능 보안