Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.6, Dec. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.6.1081

얼굴 인식 모델에 대한 질의 효율적인 블랙박스 적대적 공격 방법*

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

Query-Efficient Black-Box Adversarial Attack Methods on Face Recognition Model*

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

요 약

얼굴 인식 모델은 스마트폰의 신원 인식에 활용되는 등 많은 사용자에게 편의를 제공하고 있다. 이에 따라 DNN 모델의 보안성 검토가 중요해지고 있는데 DNN 모델의 잘 알려진 취약점으로 적대적 공격이 존재한다. 적대적 공격은 현재 DNN 모델의 인식 결과만을 이용하여 공격을 수행하는 의사결정 공격기법까지 발전하였다. 그러나 기존 의사결정 기반 공격기법[14]은 적대적 예제 생성 시 많은 질의 수가 필요한 문제점이 있다. 특히, 기울기를 근사하는데 많은 질의 수가 소모되는데 정확한 기울기를 구할 수 없는 문제가 존재한다. 따라서 본 논문에서는 기존 의사결정 공격기법의 기울기를 근사할 때 소모되는 질의 수 낭비를 막기 위해서 직교 공간 샘플링과 차원 축소 샘플링 방법을 제안한다. 실험 결과 섭동의 크기가 L2 distance 기준 약 2.4 적은 적대적 예제를 생성할 수 있었고 공격 성공률의 경우 약 14% 향상할 수 있었다. 실험 결과를 통해 본 논문에서 제안한 적대적 예제 생성방법의 같은 질의 수 대비 공격 성능이 우수함을 입증한다.

ABSTRACT

The face recognition model is used for identity recognition of smartphones, providing convenience to many users. As a result, the security review of the DNN model is becoming important, with adversarial attacks present as a well-known vulnerability of the DNN model. Adversarial attacks have evolved to decision-based attack techniques that use only the recognition results of deep learning models to perform attacks. However, existing decision-based attack technique[14] have a problem that requires a large number of queries when generating adversarial examples. In particular, it takes a large number of queries to approximate the gradient. Therefore, in this paper, we propose a method of generating adversarial examples using orthogonal space sampling and dimensionality reduction sampling to avoid wasting queries that are consumed to approximate the gradient of existing decision-based attack technique[14]. Experiments show that our method can reduce the perturbation size of adversarial examples by about 2.4 compared to existing attack technique[14] and increase the attack success rate by 14% compared to existing attack technique[14]. Experimental results demonstrate that the adversarial example generation method proposed in this paper has superior attack performance.

Keywords: Black-box attack, Adversarial attack, Face recognition model

Received(07. 06. 2022), Modified(10. 11. 2022), Accepted(10. 17. 2022)
* 본 논문은 2022년도 한국정보보호학회 하계학술대회에 발표한 우수논문을 개선 및 확장한 것임.
† 주저자, ths3847@gmail.com
‡ 교신저자, jbyun@sejong.ac.kr(Corresponding author)

I. 서 론

최근 DNN 모델(deep neural networks)의 발전으로 자연어 처리, 컴퓨터 비전 등 여러 도메인에 DNN 모델이 적용되어 우수한 성능을 내고 있다. 컴퓨터 비전 분야는 DNN 모델의 상용화가 활발하게 이루어지고 있는데 특히 얼굴 인식 분야의 경우 스마트폰의 신원 인식에 활용되어 사용자에게 편의를 제공하고 있다. 그러나 상용화가 활발히 이루어진 만큼 보안성 검토에 대한 필요가 커지고 있는데 DNN 모델은 적대적 공격(adversarial attack)이라는 취약점이 존재한다. 적대적 공격이란 정상 입력 데이터에 인지할 수 없는 정도의 미세한 섭동(perturbation)을 추가하여 DNN 모델의 오인식을 일으키는 것을 말한다. 여기서 조작된 입력 데이터를 적대적 예제(adversarial examples)라 한다. 현재 적대적 예제는 DNN 모델의 출력 결과만을 이용하는 의사결정 공격기법으로 발전하였는데 적대적 예제 생성 시 DNN 모델에 많은 질의를 해야 하는 문제점이 있다. 특히 기존 최첨단 의사결정 공격기법[14]의 경우 적대적 예제 생성을 위해 기울기를 근사하는 과정에서 많은 질의를 필요로 하는데 근사한 기울기가 정확하지 않은 문제가 있다. 이러한 문제를 해결하기 위해 본 논문에서는 직교 공간 샘플링과 차원 축소 샘플링을 이용하여 적은 질의로 적대적 예제를 생성하는 방법을 제안한다. 이 논문의 기여는 다음과 같다.

얼굴 인식 모델에 대한 의사결정 공격기법을 적용한다. 기존 적대적 예제의 경우 대부분 이미지 분류 모델에 대해서[16][17] 이루어져 왔다. 얼굴 인식 모델[15][18]의 경우 일반적인 이미지 분류 모델과 다른 손실함수를 사용하여 공격 성능의 차이가 있다. 본 논문의 실험 환경 설정과 결과를 이후 논문에서 활용할 수 있다.

기존 최첨단 의사결정 공격기법의 공격 성능을 개선한다. 기존 의사결정 공격기법의 기울기를 근사하기 위해서 공격 대상 모델로 많은 데이터를 질의하는 문제를 본 논문에서 제안하는 직교 공간 샘플링 방법과 차원 축소 샘플링 방법으로 개선한다. 실험 결과를 통해 본 논문에서 제안한 방법의 공격 효율성을 입증하고 향후 발전 방향에 대해서도 논의한다.

II. 관련 연구

2.1 적대적 공격

적대적 공격은 공격자의 공격 목표에 따라 표적 공격(targeted attack)[5]과 비표적 공격(non-targeted attack)[2]으로 나뉜다. 표적 공격[5]은 공격자가 원하는 클래스로 오인식을 일으키는 공격이고 비표적 공격[2]은 원본 데이터의 클래스 외에 다른 아무 클래스로 오인식을 일으키는 공격이다. 또한, 공격자에게 주어지는 정보에 따라 화이트박스 공격(White-box attack)[5]과 블랙박스 공격(Black-box attack)[11]으로 나뉘는데 공격자가 대상 DNN 모델의 모든 정보(모델 구조, 하이퍼 파라미터, 기울기 등)를 알 수 있다면 화이트박스 공격[5], 공격자가 대상 DNN 모델의 출력 결과만 알 수 있다면 블랙박스 공격[11]이라 한다. 블랙박스 공격[11]의 경우 DNN 모델의 출력 점수를 알 수 있는지에 따라 점수 기반 공격(score-based attack)[10]과 의사결정 기반 공격(decision-based attack)[12]으로 다시 나뉜다. 실제 상용화된 DNN 모델은 사용자에게 모델의 인식 결과만을 제공하므로 의사결정 기반 공격[12]이 가장 현실적인 공격이라고 볼 수 있다. 그러나 의사결정 기반 공격[12]은 DNN 모델의 인식 결과만을 이용하기 때문에 많은 질의 수가 필요한 문제가 있다. 이에 따라 최근 연구는 의사결정 기반 공격[12]의 질의 수를 줄이는 문제에 초점이 맞춰져 있는데 최첨단 의사결정 기반 공격인 HopSkipJump 공격[14]의 경우 기울기 근사 방법을 이용하여 공격이 성공함을 증명하고 기존 의사결정 기반 공격[12]의 질의 수를 줄였지만 기울기 근사 과정에서 여전히 질의가 낭비되는 문제가 존재했다. 따라서 본 논문에서는 HopSkipJump 공격[14]의 질의 낭비를 줄이기 위한 축소 샘플링 방법을 제안하고 DNN 모델의 강건성 향상을 위해 의사결정 기반 공격의 특징을 이용한 방어 방법 또한 제안한다. 또한, 더욱 현실적인 시나리오를 구성하기 위해서 얼굴 인식 모델[15][18]에 대해 적대적 공격을 수행한다. 얼굴 인식은 스마트폰 신원 인증 서비스 등에 적용되어 많은 사용자가 이용하고 있으므로 더욱 취약점 및 보안성 검토가 중요하다.

2.2 블랙박스 적대적 공격

블랙박스 적대적 공격은 DNN 모델의 출력 점수를 알 수 있는지에 따라 점수 기반 공격(score-based attack)[10]과 의사결정 기반 공격(decision-based attack)[12]으로 나뉜다. 점수 기반 공격[10]은 입력 데이터의 변화에 따른 DNN 모델의 출력 점수 변화를 이용하여 DNN 모델이 더욱 오인식을 일으키는 적대적 예제를 생성하는 공격기법이다. 출력 점수 변화를 바탕으로 기울기를 근사하여 화이트박스 공격[5]과 같은 목적 함수를 최적화하여 적대적 예제를 생성하거나 출력 점수를 적합도 점수로 이용하여 진화 알고리즘을 통해 적대적 예제를 생성할 수 있다. 점수 기반 공격[10]은 화이트박스 공격[5]만큼 높은 공격 성공률을 달성했지만, 상용화된 DNN 모델의 경우 사용자에게 출력 점수를 제공하지 않는 경우가 대부분이므로 실용적이지 못한 단점이 있다. 의사결정 기반 공격[12]은 DNN 모델의 인식 결과만 이용하기 때문에 현실적인 가정의 공격 기법이지만 적은 정보를 이용하는 만큼 많은 질의 수를 이용하는 문제가 있다. DNN 모델로의 질의 수를 제한하는 것만으로 공격을 막을 수 있으므로 의사결정 기반 공격에서 적대적 예제 생성 시 질의 수를 줄이는 문제는 중요하고 최근 연구들에서 해당 문제를 해결하기 위해 다양한 방법을 제안하고 있다. 의사결정 기반 적대적 예제 생성 방법은 일반적인 적대적 예제 생성 방법과 다른데 일반적인 적대적 예제 생성방법이 원본 데이터에 작은 섭동을 추가하여 DNN 모델의 결정 경계를 넘기는 방식이라면 의사결정 기반 적대적 예제 생성 방법은 적대적 예제에서 시작하여 인식 결과는 유지하되 원본 데이터와 가까워지도록 만드는 방식이다. 초기 의사결정 기반 공격인 Boundary 공격[12]은 적대적 예제와 원본 데이터와 같은 거리 반경 내에 랜덤 방향으로 이동하는 1단계와 이동 후 적대적 예제를 원본 데이터와 가깝게 이동하는 2단계를 반복 수행하여 적대적 예제를 생성한다. 반복 과정에서 랜덤 방향으로 이동하고 공격이 실패한다면 해당 데이터는 버리는 방식이기 때문에 질이 소모가 크고 적대적 예제 생성의 수렴을 보장할 수 없는 문제가 있다. 해당 문제를 해결하기 위해 HopSkipJump 공격[14]이 제안되었는데 HopSkipJump 공격[14]은 몬테카를로 알고리즘을 통한 기울기 근사 방식으로 기존의 적대적 예제 생성의 목적함수를 그대로 이용하고 최적화를 수행하여 적대적 예제를 생성한다. 기울기 근사를 이용하여 최적화를 수행하므로 적대적 예제의 수렴을 보장하고 기존 목적함수를 그대로 이용하므로 공격 성능 또한 뛰어난 장점이 있다. 그러나 HopSkipJump 공격[14]은 기울기 근사 시에 많은 질의를 소모하는데 이때 기울기를 근사하는 방식이 결정 경계 위의 적대적 예제에서 가능한 모든 방향을 고려하는 몬테카를로 기반 방식을 이용하여 이미지 데이터와 같이 차원이 큰 데이터를 다룰 때 정확한 기울기를 근사할 수 없는 문제가 존재한다. 이러한 문제를 해결하기 위해서 본 논문에서는 직교 공간 샘플링과 차원 축소 샘플링 방법을 제안한다. 정확한 기울기를 근사 시에 랜덤 벡터를 샘플링 할 때 모든 방향, 모든 차원에 대해 랜덤 벡터를 샘플링 하지 않고 대략적인 공격 성공 방향 내에서 랜덤 벡터를 샘플링 하는 것으로 질의 낭비를 줄인다. 절약한 질의 수 만큼 반복 수를 늘리는 것으로 질의 수가 같은 조건에서 더욱 원본 데이터와 유사한 적대적 예제를 생성할 수 있다.

2.3 얼굴 인식 모델(Face recognition model)

얼굴 인식은 신원 인증, 생체인식 등에서 널리 사용되는 분야로 최근 DNN의 발전으로 성능이 월등히 개선되었다. 얼굴 인식은 해결하는 문제에 따라 크게 두 가지 유형으로 나뉜다. 먼저 얼굴 식별(face identification)은 학습한 데이터를 기반으로 입력된 얼굴을 특정 클래스로 식별하는 문제이고 얼굴 검증(face verification)은 입력된 두 얼굴이 같은지를 판단하는 문제이다. 입력된 이미지를 별도의 전처리 없이 바로 사용하는 일반적인 이미지 분류 모델과 달리 얼굴 인식 모델은 입력 이미지에서 얼굴을 검출(face detection)하고 얼굴의 각도를 정면으로 돌리는 작업(face alignment)이 전처리로 수행되어 진다. 본 논문에서는 얼굴 인식 모델 중 얼굴 검증 작업을 수행하는 모델을 대상으로 공격을 수행하였다. 얼굴 검증 작업의 경우 학습 데이터와 관계없는 새로운 얼굴 데이터에 대해서도 검증해야 하므로 일반적인 이미지 분류 모델과 다르게 동작한다. 두 이미지가 객체가 같은지 여부를 검증하기 위해 객체들 사이의 유사성을 표현하기 위한 메트릭러닝(Metric learning)을 수행하는데 모델에 이미지를 입력하면 그에 해당하는 임베딩 값을 출력하여 임베딩 값을 기반으로 두 이미지가 유사한지를 검사한다. 이러한 유사도 표현을 위해서 같은 분류의 이미지는

더 가깝게 다른 분류의 이미지는 더 멀도록 임베딩 시켜주어야 하는데 기존의 Softmax 손실함수로는 한계가 있다. Sphereface[15]는 이러한 문제를 해결하기 위해 수식 1과 같이 Angular softmax 손실함수를 제안하였다.

$$\frac{1}{N}\sum_{i}-\log\left(\frac{e^{\|x_i\|\cos(m\theta_{y_i,i})}}{e^{\|x_i\|\cos(m\theta_{y_i,i})}+\sum_{j\neq y_i}e^{\|x_i\|\cos(\theta_{j,i})}}\right) \quad (1)$$

기존 Softmax 손실함수에서 가중치 W와 x의 행렬곱 W^Tx을 $\|WRIGHT\|\|xRIGHT\|\cos(\theta)$와 같이 내적으로 나타낸 다음 W를 크기 1로 정규화하여 $\|x\|\cos(\theta)$로 표현한다. 수식 1에서 m은 margin으로 같은 분류는 더 작은 각도 θ를 갖고 다른 분류는 더 큰 θ를 갖도록 한다. 해당 손실함수를 이용하여 학습을 한 후 추론 과정에서는 손실함수 이전의 전 연결계층(Fully connected layer)에서 나온 값을 입력 데이터의 임베딩 값으로 이용한다. 얼굴 검증 작업을 수행할 때 두 얼굴 이미지에서 나온 임베딩 값의 코사인 유사도(Cosine similarity)를 구하여 두 얼굴 이미지가 같은 사람인지를 검증한다. Sphereface[15]는 손실함수에 각도를 이용하는 아이디어를 제안하였지만 제안된 손실함수를 이용하는 경우 모델 학습이 불안정해져 기존 Softmax 손실함수와의 Hybrid 방식을 사용해야 했다. Arcface[18]는 이러한 학습 불안정 문제를 해결하고 얼굴 이미지를 분별하는 능력을 향상하기 위해 Addictive Angular margin 손실함수를 제안하였다. 해당 손실함수는 수식1과 유사하나 θ에 직접 margin을 빼서 페널티를 부여하여 모델을 안정적으로 학습시키고 뛰어난 얼굴 검증 성능을 낸다.

III. 얼굴 인식 모델에 대한 적대적 공격

3.1 얼굴 인식 모델에 대한 적대적 공격 시나리오

본 논문에서는 두 장의 얼굴 이미지를 입력받아 같은 사람인지 아닌지를 출력하는 얼굴 검증(face verification) 모델을 대상으로 공격을 수행한다. 공격자의 목표는 서로 다른 사람의 이미지를 얼굴 인식 모델이 같게 인식하도록 만드는 것이다. Fig. 1과 같이 모델에 질의를 통해 더욱 원본 이미지와 같은 적대적 예제를 생성한다. 기존 이미지 분류 모델

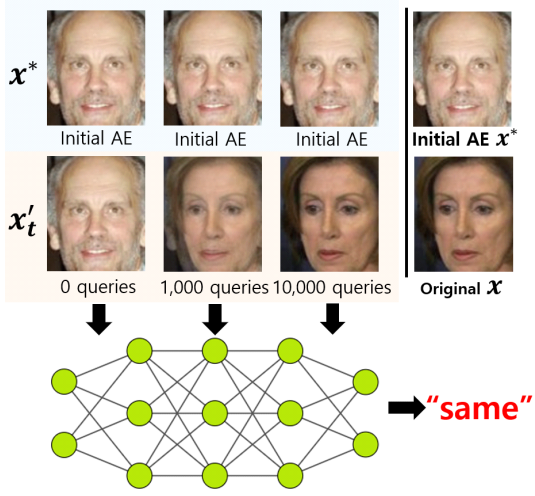


Fig. 1. Demonstration of Adversarial attack on face recognition model

에서 적대적 예제 생성 방식[14]과 달리 얼굴 검증 작업을 수행하는 얼굴 인식 모델은 입력 데이터로 비교할 얼굴 이미지 2장을 입력받는다. 따라서 적대적 예제 생성 과정에서 초기 적대적 예제 얼굴 이미지 x^*를 검증 모델에 넣어주어야 한다. 얼굴 인식 모델은 신원 확인 시스템에 상용화되어 활발히 이용되고 있는데 해당 공격기법을 통하여 시스템에 인가된 얼굴이 아니지만 같은 얼굴로 인식하도록 하여 신원 확인을 우회할 수 있다.

3.2 질의 효율적인 적대적 예제 생성방법

적대적 예제 생성방법은 Fig. 2와 같이 세 단계로 구성되어 있다. 1단계에서 공격 성공 방향을 구하기 위해 적대적 예제를 결정 경계와 인접하게 위치시키고 2단계에서 공격 성공 방향을 구하고 3단계에서 구해진 공격 성공 방향으로 적대적 예제를 얼마나 이동할지 결정한다. 여기서 2단계의 공격 성공 방향을 구하

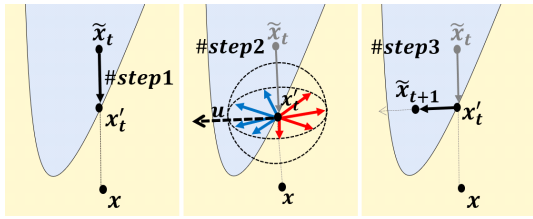


Fig. 2. Description of generating adversarial examples

는 새로운 방법을 제안하여 기존 HopSkipJump 공격[14]의 성능을 개선한다. 그 외 1,3단계는 HopSkipJump 공격[14]과 같게 동작한다.

적대적 예제의 목적함수는 원본 데이터와 차이를 최소화하면서 공격자가 원하는 인식 결과를 출력해야 하도록 구성한다. 의사결정 기반 공격의 경우 모델의 인식 결과에만 접근할 수 있으므로 의사결정 기반 공격의 적대적 예제 생성을 위한 목적함수를 수식으로 표현하면 다음과 같다.

$$\min_{x'} d(x',x) \quad such\ that \quad C_{x^*}(x') = 1 \tag{2}$$

여기서 x는 유사하게 만들고자 하는 원본 데이터이고 x'는 적대적 예제, x^*는 초기 적대적 예제로 공격자는 초기 적대적 예제 x^*와 같은 이미지로 인식되는 것을 목표로 한다. 얼굴 검증 모델의 경우 입력 데이터로 비교할 얼굴 이미지 2장을 입력받으므로 초기 적대적 예제의 분류를 유지하기 위해 검증 모델에 초기 적대적 예제 x^*와 t번째 반복의 적대적 예제 x'_t를 넣어준다. 이후 검증 모델에 비교를 위해 넣는 초기 적대적 예제는 고정된 값이기 때문에 x^*와 같이 따로 표기해주었다. $C_{x^*}(\cdot)$는 초기 적대적 예제 x^*와 입력 값이 같은 이미지인지 여부를 판단하는 함수이다. 같은 이미지라면 +1, 아니라면 -1을 출력한다. 다음으로 Fig. 2에서 묘사된 3단계의 적대적 예제 생성방법에 대해 자세히 서술한다. 먼저 1단계에서 적대적 예제를 결정 경계 위로 이동시키는데 이때 이진 탐색을 이용한다. 이를 수식으로 표현하면 다음과 같다.

$$x'_t = \alpha_{t-1}x + (1-\alpha_{t-1})\left\{x'_{t-1} + \xi_t \frac{A(x'_{t-1})}{\| A(x'_{t-1}) \|_2}\right\} \tag{3}$$

여기서 $A(\cdot)$는 2단계의 공격 성공 방향을 구하는 함수이다. 현재 시점의 적대적 예제 x'와 원본 데이터 x를 시작점과 끝점으로 설정하고 이진 탐색을 수행하는데 중간지점의 데이터를 모델에 입력한 후 인식 결과가 초기 적대적 예제와 같다면 중간지점을 시작점으로 바꾸고 그렇지 않으면 중간지점을 종료지점으로 바꾼다. 해당 과정을 반복하면 시작점과 종료지점이 가까워지는데 일정 임계치 이하로 가까워지면 이진 탐색을 종료한다. 다음으로 2단계인 공격 성공 방향 근사 단계는 본 논문에서 제안하는 핵심 아이디어로 직교 공간 샘플링과 차원 축소 샘플링을 적용하여 기존 HopSkipJump 공격[14]의 질의 낭비를 개선하고 같은 질의 수 대비 공격 성공률을 높인다. 기존 HopSkipJump 공격[14]의 경우 Fig. 2의 2단계에서 모든 방향에 대해 랜덤 벡터를 생성한 후에 모델에 질의 하여 인식이 같은 경우 +1, 인식이 변하는 경우 -1을 랜덤 방향 벡터에 곱한 후 평균 내어 기울기를 근사하는데 이때 다루는 이미지 데이터의 차원이 고차원이므로 모든 방향을 고려하는 것은 불가능하다. 따라서 정확한 기울기를 구할 수 없음에도 많은 질의 수를 사용하여 질의 낭비가 발생한다. 해당 문제를 해결하기 위해 본 논문에서는 직교 공간에서만 랜덤 벡터 샘플링을 수행한다. 여기서 직교 공간이란 결정 경계 위의 적대적 예제 x'와 원본 데이터 x를 뺀 벡터와 직교하는 공간을 의미한다.

Algorithm Generating AEs.

Inputs: Face recognition model C,
Original data x,
Initial adversarial example x^*
Number of iterations T

Outputs: Adversarial example x'

```
for t=1 to T-1 do
    # Search Boundary AE
    x'_t = Bin-search(x̃_{t-1}, x)
    for i in 100 do
   # Dimension Reduction sampling
        Set random vector r_i
        u_i = Dim-Upscale(r_i)
        # Orthogonal Space sampling
        u_i = u_i − proj_{x'_t − x} u_i
    end for
    # Success Direction
    A = Dir-estimation(x, x'_t, U)
    x̃_t = x'_t + A/‖A‖
  end for
return x' = Bin-search(x̃_{T-1}, x)
```

Fig. 3. Algorithm for generating adversarial examples

HopSkipJump 공격[14]의 동작 과정을 살펴보면 실질적으로 적대적 예제를 원본 데이터와 가깝게 만드는 단계는 1단계, 적대적 예제를 결정 경계에 가깝게 이동시키는 단계이다. 따라서 공격 성공 방향을 구하는 2단계에서는 다음 반복의 1단계에서 적대적

예제를 결정 경계 위로 올렸을 때 원본 데이터와 더욱 가까워질 수 있도록 하는 방향을 찾으면 된다. 마지막으로 3단계에서 2단계에서 구한 성공 방향으로 이동 얼마만큼 이동할지 결정한다. 공격 성공 방향을 구하더라도 너무 많이 이동하면 인식 결과가 바뀔 수 있다. 만약 인식 결과가 바뀌는 경우 크기를 등비로 축소해 인식이 같아질 때까지 반복한다. 이후 실험에서는 등비를 1/2로 하였다. 앞선 1~3단계를 반복하여 적대적 예제를 원본 데이터와 더욱 유사해지도록 만들어 준다. 전체 알고리즘은 Algorithm 1과 같다. Algorithm 1의 $\widetilde{x_t}$는 t번째 반복에서 3단계를 거친 후의 적대적 예제 또는 1단계를 거치기 전의 적대적 예제를 의미한다. 즉 결정 경계 위에 있지 않은 상태의 적대적 예제를 의미한다. 여기서 초기 적대적 예제 x^*와 $\widetilde{x_0}$은 같은 값을 나타낸다.

3.3 직교 공간 샘플링

앞서 언급된 직교 공간 샘플링 방법을 수식으로 표현하면 다음과 같다.

$$A(x'_t,\delta)=\frac{1}{B}\sum_{b=1}^{B}C_{x^*}(x'_t+\delta u_b)u_b \tag{4}$$

여기서 u는 직교 공간 샘플링을 통해서 샘플링된 랜덤 벡터이고 B는 생성할 랜덤 벡터의 수이다. 샘플링 한 랜덤 벡터 u를 결정 경계 위의 적대적 예제에 더하고 모델에 질의하여 $C_{x^*}(\cdot)$의 결과에 따라 인식이 같은 벡터는 +1, 인식이 다른 벡터는 -1을 곱해준 후에 평균 내어 공격 성공 방향을 근사한다.

$$u=r-\mathrm{Proj}_{x'_t-x}r \tag{5}$$

직교 벡터는 생성방법은 Fig. 4와 같다. 먼저 랜덤 벡터 r을 생성하고 적대적 예제 x'와 원본 데이터 x를 뺀 벡터에 사영한다. 그다음 랜덤 벡터 r과 사영한 벡터 $\mathrm{Proj}_{x'_t-x}r$를 빼주는 것으로 직교 벡터 u를 구할 수 있다.

여기서 기존 HopSkipJump 공격[14]의 2단계 기울기 근사 과정을 대체하기 위해 직교 공간 샘플링을 사용하는 이유는 다음과 같다. 먼저 HopSkipJump 공격[14]의 동작 과정을 살펴보면 실질적으로 적대적 예제가 원본 데이터와 가까워지는

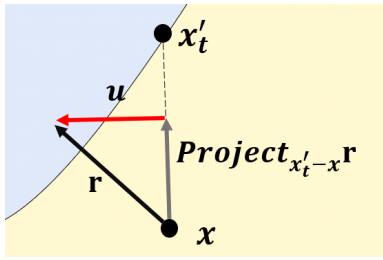


Fig. 4. Description of generating orthogonal vector

단계는 1단계, 적대적 예제를 결정 경계를 넘지 않으면서 원본 데이터와 최대한 가깝게 이동시키는 단계이다. 따라서 2단계에서는 다음 반복의 1단계에서 적대적 예제가 이전 반복의 적대적 예제보다 더 원본 데이터와 가까워지는 방향만 찾으면 된다. 이때 직교 공간 샘플링을 이용하면 효율적으로 해당 방향을 찾을 수 있다. 전체 방향에서 랜덤 벡터를 샘플링 하는 것과 달리 범위를 줄여 직교 공간 내에서 성공 방향을 구하므로 줄여진 범위 내에서 더 정확한 성공 방향을 구할 수 있고 여기서 구한 방향으로도 적대적 예제를 원본 데이터와 가깝게 이동시킬 수 있다. 이전 연구[2]에서 언급된 것과 같이 딥 러닝 모델은 학습의 용이함을 위해 Rectified Linear Units(ReLU) 활성화 함수를 이용하는 등 고차원에서 선형적인 특성을 갖는다. 또한, 이후 실험을 통해서 직교 공간 샘플링을 통해 찾은 방향의 효율성을 입증하였다. 직교 공간 샘플링 기법을 사용하는 또 다른 이유로 적대적 예제의 수렴을 확인할 수 있음이 있다. 기존 HopSkipJump 공격[14]은 적대적 예제 수렴 조건은 수식 6와 같다.

$$\cos\angle(x'_t-x,\widetilde{\nabla S})=1 \tag{6}$$

결정 경계 위의 적대적 예제x'_t와 원본 데이터 x를 뺀 벡터와 x'_t에서 근사한 기울기$\widetilde{\nabla S}$의 코사인 유사도가 1인 경우를 나타낸다. 이와 마찬가지로 1단계를 거친 결정 경계 위에 적대적 예제가 수렴한 해라면 수식 7과 같이 적대적 예제 x'_t에서 구한 모든 직교 공간 벡터 u_b는 적대적 예제 x'_t와 원본 데이터 x를 뺀 벡터와 코사인 유사도가 0일 것이다.

$$\sum_{b=1}^{B}\cos\angle(x'_t-x,u_b)=0 \tag{7}$$

HopSkipJump 공격[14]의 경우 반복마다 약 700개의 랜덤 벡터를 생성하고 질의하는데 본 논문에서는 직교 공간 샘플링을 이용하여 반복마다 100개의 랜덤 벡터를 생성하는 것으로 반복마다 질의 소모를 줄인다. 반복마다 절약한 질의 수로 반복 수를 늘려서 같은 반복 수 대비 공격 성능을 향상할 수 있다. 반복마다 샘플링 하는 벡터의 수는 실험을 통해 최적의 성능을 낸 벡터 수를 이용하였다. 너무 작은 수의 경우 제대로 된 방향을 찾기 힘들고 너무 많은 수의 벡터를 샘플링하는 경우 기존 연구와 차별점을 두기 힘들다.

3.4 차원 축소 샘플링

얼굴 이미지 데이터의 경우 인접한 픽셀끼리는 유사한 값을 갖는 특징이 있다. 이러한 특징을 이용하여 Fig 2. 2단계 랜덤 벡터 생성 시 모든 픽셀 값에 대하여 랜덤 값을 생성하는 것이 아니라 이미지의 부분 부분에 랜덤 값을 생성하고 인접한 픽셀은 생성한 랜덤 값과 유사하게 바꿔주는 것으로 축소된 차원에서 랜덤 벡터를 샘플링 할 수 있다. 이미지 차원보다 작은 차원의 랜덤 벡터 생성하고 이미지에 랜덤 값을 더할 때 인접 값들끼리 유사한 값을 갖도록 사이즈를 늘려준다. 이때 이중 선형 보간법을 이용하여 작은 차원의 랜덤 벡터를 원본 데이터와 같은 차원으로 늘려준다. 이중 선형 보간법은 값이 정해지지 않은 부분을 인접 값들의 사이 값으로 채워준다. 차원 축소 샘플링 방법은 기울기 근사 단계에서 랜덤 벡터 생성 시 우선 적용된다.

IV. 성능 평가

4.1 실험 환경

본 연구에서는 얼굴 검증 모델인 Sphereface[15]와 Arcface[18]을 대상으로 적대적 공격의 성능 평가를 수행하였다. 얼굴 검증 작업에서 적대적 예제를 생성하며 얼굴 정렬(face alignment)을 거친 후의 데이터를 사용한다. 실험에 이용된 데이터로는 LFW(labeld faces in the wild) 데이터셋[13]의 100개의 얼굴 이미지 쌍을 사용하였다. LFW 데이터셋[13]은 얼굴 이미지 데이터셋으로 5,749명의 인물 이름이 라벨링된 13,233개의 이미지이다. LFW 데이터셋[13]에 대한 각 모델의 얼굴 검증 성능은 Sphereface[15]가 99.42%, Arcface[18]가 99.78로 뛰어난 검증 성능을 갖고 있다. 공격 성능 평가를 위해 측정 기준으로 공격 성공률과 적대적 예제와 원본 데이터가 얼마나 유사한지를 측정하였다. 공격 성공률의 경우 원본 데이터와 적대적 예제의 L_2 거리 차이가 10 이하인 경우 성공으로 측정하였고, 원본 데이터와 적대적 예제의 차이는 L_2-norm을 기반으로 측정하였다. 적대적 예제 생성 시 질의 수 제한은 선행 연구[14]와 동일하게 25,000으로 설정하였다.

4.2 Sphereface 실험 결과

본 논문의 성능 평가를 위해 대표적인 의사결정 공격기법인 Boundary 공격[12]과 HopSkipJump 공격[14]을 비교하여 실험을 진행하였다.

실험 결과 Fig. 5와 같이 모든 질의 수에 대해서 본 논문에서 제안한 방법이 원본 데이터와 더 적은 차이로 적대적 예제를 생성하는 것을 확인할 수 있었

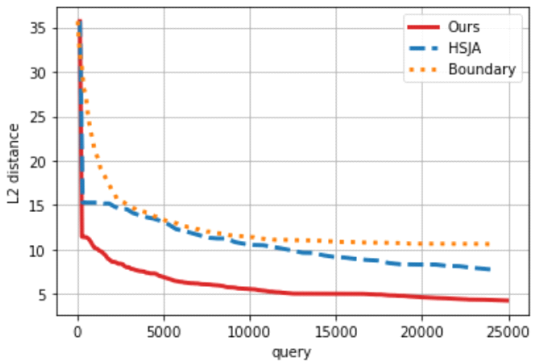


Fig. 5. L_2-distance versus the number of queries on LFW dataset with Sphereface

Table 1. Attack success rate versus the number of queries on LFW dataset with Sphereface

Query	Attack success rate(ASR)		
	BA	HSJA	Ours
1000	0.0	0.12	**0.48**
10000	0.5	0.46	**0.87**
15000	0.75	0.75	**1.00**
20000	0.75	0.86	**1.00**
25000	0.75	0.86	**1.00**

다. 이는 같은 질의 수 대비 본 논문에서 제안한 방법의 더욱 원본 얼굴 이미지와 유사한 적대적 예제를 생성하여 질의 효율성이 높음을 입증한다. 또한, Table 1의 실험 결과에서 볼 수 있듯이 더 적은 질의 수만으로 공격이 성공 공격함을 확인할 수 있었는데 약 14% 공격 성공률 향상을 할 수 있었다. 이를 통해 본 논문에서 제안한 방법의 공격 성능 또한 뛰어남을 알 수 있다.

4.3 Arcface 실험 결과

앞선 Sphereface 실험과 마찬가지로 Boundary 공격[12]과 HopSkipJump 공격[14]을 비교하여 실험을 진행하였다.

실험 결과 Fig. 6와 같이 Arcface 모델[18]에 대한 공격 역시 같은 질의 수 대비 본 논문에서 제안한 방법의 적대적 예제 생성 법이 더 작은 섭동을 갖는 것을 확인할 수 있었다. 또한, Table 2의 질의 수별 공격 성공률 역시 Sphereface[15]의 실험 결과와 마찬가지로 본 논문에서 제안한 방법이 공격 성공률이 가장 높은 것을 확인할 수 있다. 실험 결과를 통해 본 논문에서 제안한 아이디어인 HopSkipJump 공격[14]의 적대적 예제 생성 시 기울기 근사 과정에서 질의 비효율성을 개선하기 위한 직교 공간 샘플링과 차원 축소 샘플링 방법을 같이 이용하는 방법의 효율성을 입증하였다.

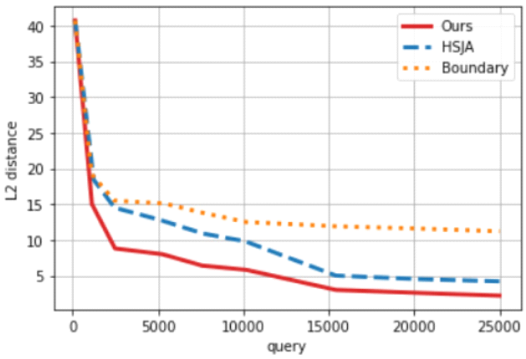


Fig. 6. L_2-distance versus the number of queries on LFW dataset with Arcface

Table 2. Attack success rate versus the number of queries on LFW dataset with Arcface

Query	Attack success rate(ASR)		
	BA	HSJA	Ours
1000	0.0	0.19	**0.53**
10000	0.32	0.52	**0.89**
15000	0.56	0.92	**0.98**
20000	0.68	0.92	**0.98**
25000	0.68	0.92	**0.98**

V. 결 론

본 논문에서는 의사결정 기반 공격기법을 이용하여 얼굴 인식 모델을 공격하는 연구를 수행하였다. 기존 의사결정 기반 공격기법의 질의 효율성 향상을 위해 직교 공간 샘플링 방법과 차원 축소 샘플링 방법을 적용하는 공격기법을 제안하였다. 실험 결과를 통해 더 적은 질의 수로 적대적 예제를 생성하고, 같은 질의 수 대비 더욱 원본 데이터와 유사한 적대적 예제를 생성하여 본 논문에서 제안한 적대적 예제 생성방법의 효율성을 입증하였다. 또한, DNN 모델의 강건성 향상을 위해 의사결정 기반 공격의 특징을 이용한 방어 방법을 제안하였다. 본 연구 결과를 통해 상용화된 얼굴 인식 모델에 적대적 공격을 수행하는 연구를 진행할 수 있을 것이다. 또한, 얼굴 인식 모델의 손실함수의 특징을 적용하여 더욱 정교한 적대적 예제를 생성하는 연구로 확장될 수 있을 것이다.

References

[1] C. Szegedy, W. Zaremba, and I. Sutskever, "Intriguing properites of neural networks." arXiv preprint arXiv:1312.6199, 2013.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples." arXiv preprint arXiv: 1412.6572, 2014.

[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, March 2016.

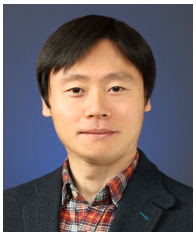
[4] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural

networks" 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2574-2582, June 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. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083, 2017.
[7] A. Athalye and I. Sutskever, "Synthesizing Robust Adversarial Examples," International Conference on Machine Learning (ICML), pp. 284-293, July 2018.
[8] A. Athalye, N. Carlini, and D. Wagner, "Synthesizing Robust Adversarial Examples" International Conference on Machine Learning (ICML), pp. 274-283, July 2018.
[9] J. Su and D. Vasconcellos et al., "One pixel attack for fooling deep neural networks", VOL. 23, NO. 5, pp. 828-841, October 2019.
[10] 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, November 2017.
[11] 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, April 2017.
[12] 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.
[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1, 2, 4
[14] 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.
[15] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6738 - 6746.
[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385, 2015.
[18] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," arXiv:1801.07698, 2018.

〈저자소개〉

서 성 관 (Seong-gwan Seo) 학생회원
2019년 8월: 세종대학교 정보보호학과 졸업
2022년 8월: 세종대학교 정보보호학과, 지능형드론 융합전공 석사
〈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