붙임3

연구논문/작품 최종보고서

2021 학년도 제 1 학기

제목 : 적대적 예제를 통한 딥페이크 방지

정다솔(2017310367)
(※팀원 이름 모두 기재)

2021년 10월 22일

지도교수: o o o 서명

계획(10)	주제(20)	개념(20)	상세(30)	보고서(20)	총점(100)

* 지도교수가 평가결과 기재

평가 배점

계획 (10)

무리 없는 진행.

주제 (20)

신규성, 진보성, 창의성, 현실성

개념설계 (20)

주제에 명시된 바대로 필요한 기능이 모두 포함되었는가?

상세설계 (30)

계획에 명시된 바대로 결과물과 일치하는가?

각각의 practical constraints에 기술된 사항을 만족하는가?

보고서 (20)

짜임새 있는 정리 및 표현

■ 요약

딥페이크(Deepfake)란 인공지능(AI)을 사용하여 특정 인물의 얼굴을 다른 사진이나 영상과 합성하여 실제처럼 보이게 하는 기술이다. 이러한 딥페이크 기술이 발전하여 사진 한장으로도 딥페이크 영상을 만들 수 있게 되면서, 연예인과 같은 유명인사를 대상으로 했던 딥페이크 음란물 제작 범죄는 일반인들을 대상으로 확대되었다. 그러나 기존의 딥페이크 방어 기술은 딥페이크 탐지에 집중되어 있어 이러한 일반인 대상의 딥페이크 성범죄 피해를 막기 힘들다는 한계가 있다. 이에 딥러닝을 공격하는 공격 기법이었던 적대적 예제를 사용하여 딥페이크 성범죄를 예방하는 기술을 제시하고 구현하고자 한다. 본 논문에서는 딥페이크 예방 기술로서 적대적 예제를 제시한다. 적대적 예제는 심층신경망의 선형적인 특징을 이용하여 특정 레이블을 가진 이미지가 해당 레이블로 분류되지 못하도록 만드는 공격이다. 이에 딥페이크와 같은 인공지능이 적대적 예제에 취약한 이유를 설명하고, Cifar-100 데이터셋과 ResNet 모델을 사용하여 사람의 얼굴을 인식하는 심층신경망을 제작한 후, 이를 기반으로 적대적 예제 이미지를 생성한다. 해당 이미지는 원래는 사람 레이블에 분류되어야 하나, 적대적 공격을 통해 사람 레이블이 아닌 다른 레이블로 오분류되도록 만들어진 이미지이다. 이렇게 만들어진 적대적 예제를 실제 상용화된 딥페이크 어플리케이션에 넣어 적대적 예제가 딥페이크 예방 기술로서 효과가 있는지 확인한다. 연구 결과, 이렇게 만들어진 적대적 예제가 어플리케이션에서 인식되지 않는 것을 확인할 수 있었다. 이는 적대적 예제를 통해 딥페이크 영상 제작을 막고 디지털 성범죄를 예방할 수 있음을 의미한다.

■ 서론

가) 제안배경 및 필요성

딥페이크란 컴퓨터 심층학습을 일컫는 딥러닝(deep learning)과 가짜를 뜻하는 페이크(fake)의 합성어로 인공지능(AI) 기술을 사용하여 특정 인물의 얼굴을 다른 사진이나 영상과 조합하여 실제처럼 만든 영상을 말한다. 이러한 딥페이크 영상은 실제 존재하지 않는 사진 혹은 영상을 만드는 데에 최적화되어 있으며 일반인의 눈으로는 그 진위여부를 간파하기 쉽지 않다는 특징이 있다. 이러한 특징 때문에 딥페이크 기술은 가짜 뉴스, 선동, 성범죄에 이용되며 인터넷에서 공신력 있는 정보를 가려내는 것을 어렵게 만들 뿐만 아니라 사이버 보안 문제를 일으키고 정치적, 국가적 문제를 발생시키기도 한다. 딥페이크는 사회 전반에서 커다란 문제를 일으키며 주요한 위험요인으로 여겨지고 있다.

특히 인터넷 기술이 발전하며 사람들은 인터넷에서 정보를 검색하고, 이미지를 얻으며, 그렇게 얻은 자료에 의존한다. 인터넷의 보급이 현대인의 삶을 유익하고 편리하게 만들어 준 것은 분명 사실이지만 동시에 인터넷을 이용한 범죄 또한 활발하게 발생하고 있으며 그 중 대표적인 것이 바로 음란물의 등장이다. 인터넷에는 수많은 유해한 음란물들이 돌아다니고 있으며 가장 대표적인 음란물 사이트인 '폰허브'에는 하루 평균 1 억 5 천만 명이 접속한다는 통계자료 또한 존재한다. 최근에는 딥페이크 기술을 이용하여 특정 사람의 얼굴을 음란물에 삽입하여 동영상을 제작하는 범죄 또한 발생하고 있다. 실제로 네덜란드 디지털 보안 연구소의 2019 년 보고서에 따르면, 온라인에 유포되고 있는 딥페이크, 즉 가짜 합성영상물 가운데 96%가 불법 음란 영상물이다.

딥페이크는 몇 년 전부터 사회적 문제로 지적되어 왔지만 최근 문제는 점점 심각해지고 있다. 해외에 서버를 둔 범죄자들이 연예인의 얼굴을 성착취물에 합성해 판매하고 돈을 받는 경우가 증가하였으며, 심지어는 평소에 알고 지내던 평범한 일반인 지인들로 딥페이크 성착취물을 만드는 경우 또한 증가하였다. 기존 딥페이크 성범죄의 피해자는 연예인과 정치인과 같은 유명인사가 대부분이었으나 기술의 발전으로 사진 한 장만으로도 딥페이크 제작이 가능해짐에 따라 딥페이크 성범죄 피해자는 일반인으로 확장되고 있다. 최근 소셜미디어에서는 돈을 내고 지인의 얼굴을 보내주면 딥페이크를 기술을 사용하여 음란한 사진이나 영상에 실제처럼 합성해주겠다는 제안이 올라오기도 하며, 사진 한 장으로 간단한 딥페이크 영상을 만들 수 있는 어플리케이션들 또한 이미 상용화가 된 상태이다. 이흥규 KAIST 전산학부 교수는 "누구든지 관심을 갖고 약간의 시간을 투자하면 딥페이크 합성 영상을 만들 수 있는 상황"이라고 전했다. 이러한 상황에서 인터넷과

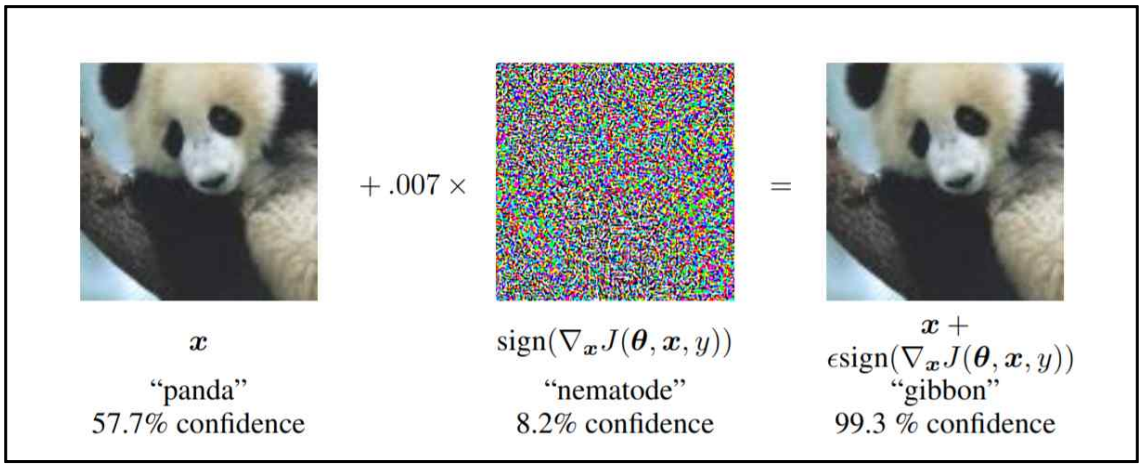
스마트 기기에 익숙한 10 대 청소년들이 이러한 딥페이크 제작에 접근할 수 있게 되면서 성폭력을 넘어 '왕따' (집단 괴롭힘) 목적으로 불법 합성물을 유포하는 경우 또한 발생했다. 방송통신위원회의 '2020 년 사이버폭력 실태 조사'에서 불법 합성물 제작·유포 등 디지털 성범죄를 목격한 경험이 있는 학생 가운데 16%가 "전혀 문제 되지 않는다"라고 답할 정도로 10 대들 사이에서는 딥페이크를 사용한 성범죄에 대한 윤리의식이 부족하다. 경찰에 따르면 지난해 12 월부터 올해 4 월까지 딥페이크 불법 합성물을 제작하거나 유포해 검거된 피의자 94 명 가운데 10 대가 무려 69.1%(65 명)라고 한다. 또한 인적 사항이 확인된 딥페이크 합성 피해자 가운데 57.9%(66 명)는 19 세 미만이었다. 딥페이크 범죄가 활발해지며 딥페이크 피해자들의 고통 또한 심해지자 올해 초 '딥페이크 영상 제작·유포자를 강력히 처벌해 달라'는 청와대 국민청원이 30 만 명 이상 동의를 얻기도 했다. [1]

이러한 딥페이크를 막는 방법에는 주로 법에 의한 규제와 딥페이크 탐지 기술이 언급된다. 실제 페이스북은 2019 년에 인공지능(AI) 기술로 조작된 딥페이크 영상을 탐지하고 구별하기 위한 '딥페이크 탐지 챌린지' 프로젝트를 진행했다. 페이스북은 1 천만 달러(약 120 억원)를 딥페이크 영상 탐지 기술 연구에 투자하기로 했으며 이 프로젝트에 페이스북을 비롯해 마이크로소프트, 알파벳(구글 모기업), 애플 등이 후원하였다. 인공지능이 만든 딥페이크 이미지를 인공지능을 통해 막아보겠다는 생각이다. 이러한 딥페이크 탐지에는 GAN 으로 생성한 이미지에 다시 GAN 알고리즘을 적용해 위조 여부를 탐지하는 알고리즘, 이미지 픽셀의 고유값을 학습하는 인공지능 알고리즘을 활용한 위조 탐지 알고리즘, 정상 이미지와 위조 이미지의 픽셀 패턴을 비교해내는 기계학습 알고리즘 등이 연구되고 있다. 그러나 이러한 딥페이크 탐지 기술에는 결정적인 한계가 있다. 딥페이크 탐지 기술이 사용되는 것은 이미 딥페이크 영상이 만들어진 이후라는 것이다. 물론 가짜 뉴스와 같은 범죄에서는 이러한 딥페이크 구분과 탐지가 중요한 역할을 할 것이다. 그러나 성착취물에서 딥페이크 영상의 진위여부는 피해자와 가해자들에게 중요하지 않기 때문에 딥페이크 성범죄에서는 이러한 딥페이크 탐지 기술을 의미가 크지 않다. 성착취물은 영상이 만들어졌다는 데에 그 목적과 의의가 있으며 이미 딥페이크 영상이 만들어져 유포되는 순간 피해자들에게는 지울 수 없는 상처로 남게 된다. 따라서 이러한 딥페이크 성범죄 피해를 최소화하는 방법은 딥페이크 영상을 탐지하는 기술이 아니라 아예 딥페이크 영상 제작을 막는 기술이다. 즉, 일반인을 상대로 딥페이크 범죄가 발생하는 것을 막기 위해서는 딥페이크 영상 제작을 막기 위한 기술이 필요하다.

나) 연구논문/작품의 목표

따라서 작성할 논문의 목표는 딥페이크 영상 제작을 불가능하게 만드는 기술을 제시하는 것이다. 딥페이크 영상 제작을 막는 기술을 생각하기 위해 가장 먼저 고려한 것은 딥페이크 기술의 한계였다. 심층신경망의 발전으로 사진 한 장만으로도 충분히 딥페이크 영상을 만들 수 있게 되면서 일반인들 또한 딥페이크 범죄로부터 자유로울 수 없게 되었지만, 이는 반대로 말하면 일반인들을 대상으로 딥페이크 성범죄를 저지르는 데에 한계가 확실하다는 것이다. 일반인들은 유명 인사들과는 달리 사진과 영상 데이터가 한정되어 있다. 이러한 데이터의 한계를 이용하여 해당 사진과 영상들이 딥페이크에 악용되지 못하도록 한다면 일반인 대상 딥페이크 범죄를 예방할 수 있다.

이를 위해 본 논문에서 제시하는 기술은 바로 적대적 예제이다. 적대적 예제는 딥러닝을 공격하는 가장 대표적이며, 아직까지도 완벽한 해결법이 나오지 않은 공격 방법이다. 2014년도에 처음 발견되어 인공지능 보안에 커다란 영향을 주었으며, 아직까지도 수많은 공격 방법과 방어 방법이 제시되며 창과 방패의 싸움이 계속되고 있는 인공지능 공격 방법이다. 적대적 예제는 2014년 ICLR에서 Szegedy 등에 의해 처음으로 제기되었으며, 이들은 원본 이미지에 왜곡을 추가함으로써 적대적 예제를 만들 수 있음을 주장했다.[2]



[그림 1] 적대적 예제의 원리(출처: [2])

적대적 예시의 내용은 다음과 같다. 심층신경망으로 만들어진 이미지 분류 모델은 '판다 이미지'를 '판다'로 분류한다. 그러나 이 '판다 이미지'에 특정 왜곡을 결합하면 인공지능은 이 이미지를 '긴팔 원숭이'로 구분하게 된다. 여기서 중요한 점은 사람의 육안으로는 '판다 이미지'와 특정 왜곡이 결합한 이미지가 '판다'로 보인다는 것이다. 즉, 사람의 눈은 제대로 인식이 가능하지만 인공지능은 잘못 인식하게 만드는 것이 바로 적대적 예제의 내용이다.

적대적 예제의 조금 더 자세한 원리는 다음과 같다. 원본 이미지에 계산된 작은 왜곡을 추가하면 심층신경망의 이미지 분류 결과를 임의로 바꿀 수 있다. 이러한 왜곡은 아주 작은 값이기 때문에 사람의 눈으로는 원본 이미지와 왜곡이 추가된 이미지를 구분할 수 없지만 심층신경망은 왜곡이 추가된 이미지를 원본 이미지와 다르게 분류한다. 이렇게 왜곡이 추가된 이미지를 '적대적 예제(Adersarial Example)'라 하며 지금까지 심층신경망의 취약점이자 문제점으로 다루어졌다. 적대적 예제는 지금까지 대부분 딥러닝의 취약점이자 문제점으로 여기고 다루어졌지만 본 논문에서는 적대적 예제를 딥페이크 범죄의 예방 방법으로 제시하고자 한다.

논문의 기본적인 아이디어는 이러한 적대적 예제를 사용하여 딥페이크 범죄를 사전에 예방하는 것이다. 만약 일반인들의 사진을 적대적 예제로 만든다면, 사람은 육안으로 일반인들의 사진을 원래 사진처럼 감상할 수 있다. 그러나 인공지능의 경우 사람의 얼굴 이미지들을 제대로 인식하지 못하여 이미지를 토대로 딥페이크 영상물을 만들지 못하게 된다. 따라서 이 논문의 목표는 실제로 사람의 이미지를 통해 적대적 예제를 제작하고, 이렇게 만들어진 적대적 예제들을 실제 상용화된 딥페이크 어플리케이션에 넣어 적대적 예제가 딥페이크 예방 기술로서 효과가 있는지 확인하는 것이다.

다) 연구논문/작품 전체 overview

결론부터 말하자면, 본 논문에서는 적대적 예시를 사용한 딥페이크 방지의 가능성을 확인하였으며 실제로 딥페이크 어플리케이션이 오작동하도록 만드는 데에 성공했다. Cifar-100 데이터셋과 ResNet 모델을 통해 사람의 얼굴을 분류하는 심층신경망을 제작하였으며, 해당 심층신경망을 토대로 적대적 예제를 만들고 정확도를 측정하는 데에 성공하였다. 이렇게 만들어진 적대적 예제들을 실제 상용화된 어플리케이션에 넣어 테스트를 진행하였으며, 어플리케이션이 적대적 예제들을 인식하지 못하는 것을 확인하였다.

전체적인 논문의 구성은 다음과 같다. 1 장에서 해당 연구의 서론과 필요성을 설명하고, 2 장에서 딥페이크, 적대적 예제, ResNet 과 같은 관련 연구에 대해 설명한다. 3 장에서는 실제로 구현한 심층신경망과 적대적 예시를 설명한다. 4 장에서는 구현 결과와 실험 결과를 분석한다. 그 후 마지막 5 장에서는 결론 및 소감을 작성한다.

■ 관련연구

가) 적대적 예제

적대적 예제로 인한 심층신경망의 취약성은 2014 년 ICLR 에서 Szegedy 등에 의해 처음으로 제기되었다. 이들은 원본 이미지에 왜곡을 추가함으로써 적대적 예제를 만들 수 있음을 제시했다.[2] 이미지 인식과 생성에는 주로 심층신경망(Deep Neural Network, DNN)이 사용된다. 그러나 이러한 DNN 에는 보안상의 취약점이 존재하며 그 중 대표적인 취약점이 다름 아닌 적대적 예제이다.

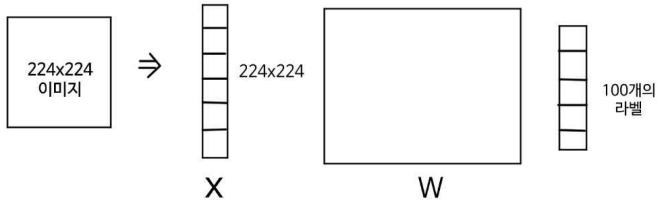
조금 더 자세하게 설명하자면, 인공지능은 신경망을 통해 학습하고 결과값을 분류한다. 이 과정에서 손실함수를 통해 가중치를 조절해가며 가장 정확하게 인식하는 모델을 만드는 것이 인공지능의 목표라고 할 수 있다. 적대적 예제는 이미지에 왜곡을 추가하면 인공지능이 이미지를 잘못 인식할 수 있다는 점을 악용한 공격으로 사람은 이러한 왜곡을 제대로 식별할 수 있지만 픽셀 단위로 학습하는 기계는 해당 이미지를 잘못 식별하게 된다. 따라서 만들어진 모델과 반대되는 방향으로 왜곡을 넣음으로써 손실함수의 값을 키워 활성화함수가 잘못된 분류값을 출력하도록 하는 것이 적대적 예제의 원리이다.

예를 들어, 자동자가 유턴 표지판을 인식해야 하는 상황이라고 가정해보자. 이미 학습되어 있는 모델은 유턴 표시판을 읽고 활성화함수를 통해 유턴 표시판일 확률이 제일 높다는 것을 알아낸 후 유턴 표시판이라고 인식하게 된다. 그러나 공격자가 악의적으로 표지판에 왜곡을 넣으면, 자율주행 인공지능은 해당 표지판을 유턴 표지판이 아닌 다른 표지판으로 인식하게 된다.

또한 I. Goodfellow 는 적대적 예제의 원인은 심층신경망의 지나친 선형성 때문이라고 주장하며, 이를 기반으로 Fast Gradient Sign Method(FGSM)이라는 적대적 예제 생성법을 만들어냈다. 원본 이미지 x 에 대해, 적대적 예제 $\tilde{x} = x + \eta$ 를 만드는 식은 다음과 같다.

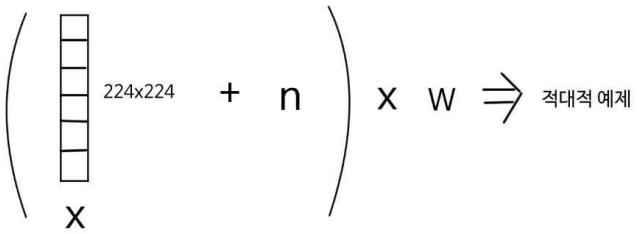
$$\eta = \varepsilon \, \mathrm{sign}\left(\nabla_x J(\theta, x, y)\right)$$

.



[그림 2] 적대적 예제의 원리 설명 1

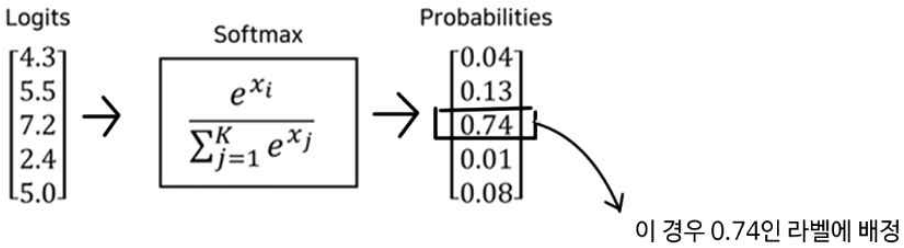
가장 먼저, 224x224 픽셀의 이미지를 벡터 형태(X)로 변환하고 가중치(W)를 곱해 100개의 레이블 중 하나에 분류하는 형태의 모델이 있다고 하자. 따라서 해당 모델의 가중치(W)는 224X224의 input을 가지고 100의 output을 가진 형태이다.



[그림 3] 적대적 예제의 원리 설명 2

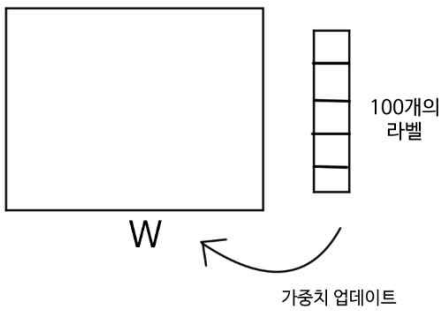
이런 벡터(X)에 왜곡(η)를 추가하여 가중치(W)를 통해 학습하면 적대적 예제가 만들어진다. 즉, 적대적 예제의 수식은 다음과 같다.

$\tilde{x} = x + \eta$ (이 때, η 은 ε 보다 작은 값이며 각 픽셀은 ε 만큼 이동할 수 있다.)



[그림 4] 적대적 예제의 원리 설명 3

이러한 왜곡을 만드는 방법은 다음과 같다. 학습을 통해 얻어낸 결과값을 softmax 함수를 통해 가장 높은 확률값을 가진 레이블에 배정한다.



[그림 5] 적대적 예제의 원리 설명 4

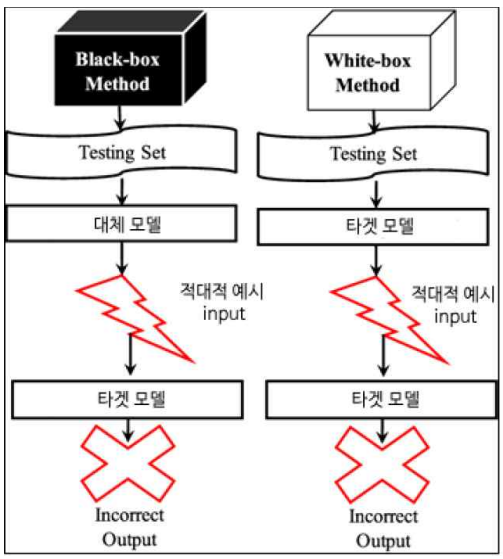
이 과정에서 Loss 함수가 사용되는데, Loss 함수란 데이터를 토대로 산출한 모델의 예측값과 실제값과의 차이를 구해 가중치 W 를 값을 업데이트하는 함수이다. 적대적 예시는 이 Loss 함수값에 영향을 주어 다른 레이블로 배정되도록 조작하는 것이다. Loss 함수의 그래프의 기울기를 구한 후, x 값을 변화시켜 y 값에 영향을 주는 방식이다. 따라서 기울기가 양수면 ε 값을 양수로 주고, 기울기가 음수면 ε 값을 음수로 주어 loss 값을 크게 만든다.

이러한 방법으로 적대적 예제를 만들 수 있다. 이를 수식으로 표현하면 다음과 같다.

$$\eta = \varepsilon \, \mathrm{sign}\,(\nabla_x J(\theta, x, y))$$

이들은 이러한 방식으로 손실함수의 기울기 부호를 사용하여 왜곡을 추가하는 것만으로 신경망의 예측 오류를 일으킬 수 있음을 증명했다.[3]

나) 적대적 학습의 전이성



[그림 6] 블랙 박스와 화이트 박스 방식

특정 심층신경망을 공격하기 위해 만들어진 적대적 예제는 유사한 구조로 학습된 다른 심층신경망을 공격할 때에도 효과적이다. 이러한 특징을 적대적 예제의 전이성(Transferability)이라고 하며, 전이성의 존재로 인해 적대적 예제에 관한 연구는 심층신경망의 근본적인 문제로 간주되기 시작했다. 적대적 예제가 존재하는 이유와 전이성과 같은 특징을 지닌 이유의 명확한 이론적 규명은 아직 이루어지지 않았다. Szegedy등은 심층신경망의 과적합으로 인한 지나친 비선형성이 적대적 예제의 원인이라고 주장하였으나[2], Goodfellow등은 오히려 심층신경망의 선형적 특징이 적대적 예제의 원인이 된다고 주장하였다.[3] Ilyas등은 적대적 예제는 심층신경망의 오류가 아니라 신경망 분류 성능에 영향을 주는 비강건한(non-robust) 특징이라는 의견을 펼쳤다.[4]

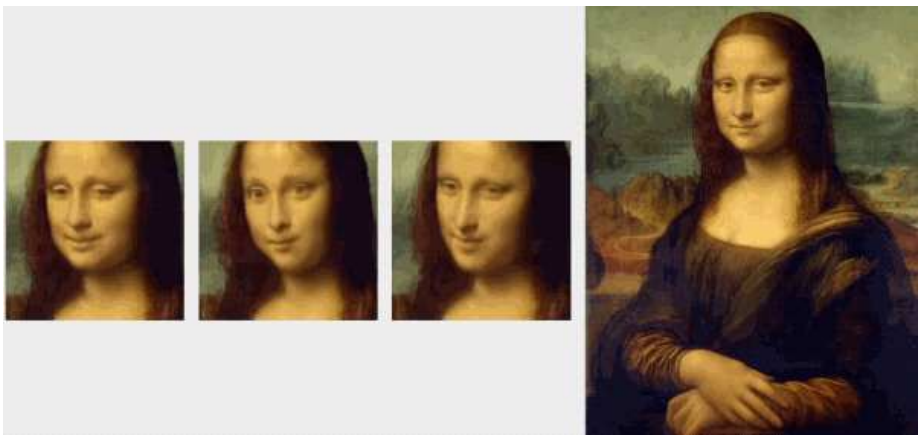
적대적 예시를 사용한 공격은 크게 화이트 박스 방식과 블랙 박스 방식, 두 가지로 구분된다. 화이트 박스 방식은 파라미터, 모델 구조, 활성화함수, 손실함수, 가중치

등과 같은 모델에 대한 정보를 이미 모두 알고 있는 상황에서 test값을 넣어가며 학습을 진행하는 방식이다. 이러한 화이트 박스 방식의 성공률은 거의 100%에 달한다. 반면 블랙 박스 방식은 모델에 대한 정보가 주어지지 않는 상황에서 대체 모델과 쿼리 값을 통해 test값을 넣어가며 학습을 진행하는 방식이다.

이러한 블랙 박스 방식 또한 효과를 보이는데 이는 적대적 예제의 특징인 '전이성(Transferability)' 때문이다. 심층신경망은 데이터의 일반적인 특성을 학습하기 때문에 '전이학습' 이 가능하다. 전이학습이란 제대로 만들어진 심층신경망 모델의 가중치를 비슷한 과제를 해결하는 심층신경망에서도 재활용하여 사용할 수 있다는 심층신경망 학습 기법이다. 이러한 '전이학습'과 비슷하게, 하나의 심층신경망에서 공격에 성공한 적대적 예제는 다른 심층신경망에서도 공격에 성공할 확률이 높다. 이를 적대적 예제의 특징인 전이성이라고 한다. 이러한 적대적 예제의 전이성 덕분에 블랙 박스 방식 또한 상당히 높은 성공률을 보인다. 해당 논문에서는 어플리케이션이 어떠한 모델을 사용하는지에 대한 정보가 전혀 없기 때문에 블랙 박스 방식을 사용한다.[5]

다) 딥페이크

딥페이크(deepfake)란 딥러닝(deep learning)과 가짜(fake)를 합친 말로 인공지능을 기반으로 한 이미지 합성 기술로 생성적 적대 신경망(GAN)라는 인공지능 알고리즘을 사용하여, 특정 인물의 사진을 기존의 사진이나 영상에 합성하여 만드는 기술이다. 이러한 딥페이크 기술은 이미 사망한 사람을 가상으로 되살리거나 초상권 보호 등을 위해 사용할 수 있는 한편, 가짜 뉴스나 성착취물과 같은 범죄에 악용되기도 한다.

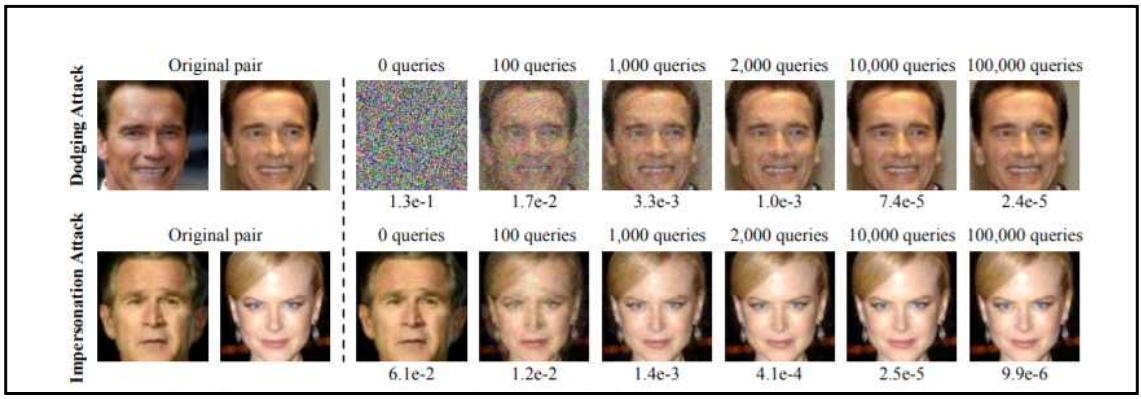


[그림 7] 사진 한장으로 만들어진 딥페이크(출처: [5])

특히 최근에는 사진 한 장만으로 다양한 얼굴 표정과 움직임이 담긴 가짜 동영상을 만들 수 있는 인공지능(AI) 기술이 삼성전자에 의해 개발되었다. 삼성전자의 러시아 모스크바 AI 연구센터가 2019 년에 발표한 논문 'Few-Shot Adversarial Learning of Realistic Neural Talking Head Models'에 따르면 이 기술은 기존의 AI 영상합성 기술인 '딥페이크(deepfake)'와는 달리 별도의 3 차원 모델링 과정 등이 필요 없는 것이 특징이다. 1 장 이상의 사진으로 얼굴 윤곽(랜드마크)을 잡아내 실제와 구분하기 어려울 정도의 가상 동영상을 만들 수 있고, 이를 애니메이션 등으로도 변환할 수 있다. [5, 6]

라) 얼굴 인식 기술

얼굴인식 기술은 얼굴의 영상을 보고 해당 인물이 어떤 인물인지 판별하는 기술로 출입국 심사, 결제 시스템, 단말 잠금 해제 등과 같은 실생활 서비스에 활용된다. 실제 수집된 얼굴 영상은 다양한 표정 및 조명 변화, 원거리 촬영, 해상도, 블러(blur) 등으로 인해 얼굴인식 성능이 떨어진다는 문제점을 가지고 있으나 딥러닝 기술을 사용하여 다양한 데이터 환경에서도 높은 성능의 얼굴인식이 가능하게 되며 이러한 문제들은 해결되었다. [7]



[그림 8] 적대적 예제를 통한 얼굴인식 공격 (출처:[8])

이러한 얼굴 인식 분야 또한 딥러닝 기술을 사용하여 학습되었기 때문에 적대적 공격(adversarial attack)에 취약하다. '딥러닝 기반 얼굴인식 모델에 대한 변조 영역 제한 기만공격'과 'Efficient Decision-based Black-box Adversarial Attacks on Face Recognition' 논문에서는 적대적 공격을 통해 생성된 적대적 예제를 이용하여 딥러닝 기반 얼굴인식 모델에 대해 공격을 수행하는 데에 성공했다. [6, 7]

즉, 해당 논문들을 통해 얼굴 인식(Face Recognition) 또한 적대적 공격이 가능한 것을 확인하였다. 따라서 이 논문에서 주장하는 얼굴 이미지에 적대적 공격을 가해

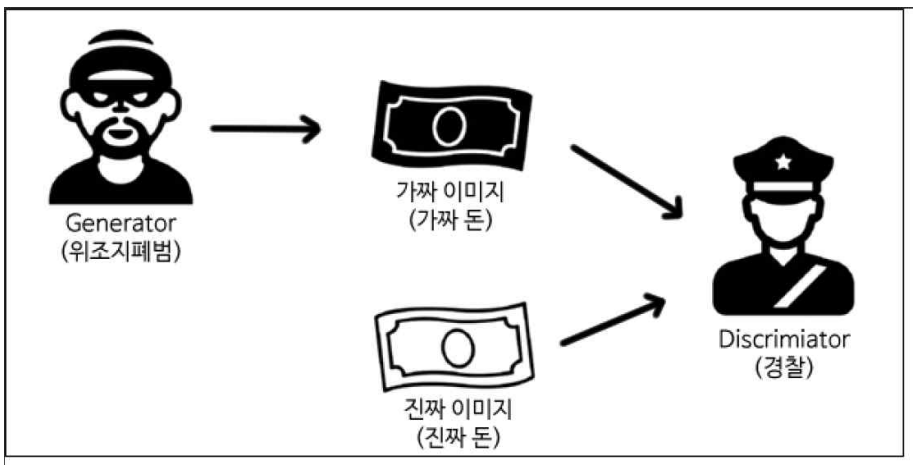
인공지능이 인식하지 못하도록 하는 기술 역시 이론 상으로 구현이 가능한 것을 확인할 수 있었다.

마) ResNet

이미지를 분석하기 위해 패턴을 찾는 알고리즘인 Convolutional Neural Network(CNN)의 일종인 ResNet 은 이미지 인식과 구분에 사용되는 알고리즘으로 2015 년 Kaiming He 에 의해 제시되어 ImageNet Large Scale Visual Recognition Challenge(ILSVRC)에서 우승을 차지하며 그 성능을 입증하였다. ResNet 은 VGGNet 의 구조를 토대로, 기존의 CNN 망과는 달리 입력값을 출력값에 바로 더해줄 수 있도록 지름길(shortcut)을 만들어 학습하는 Residual Block 방식을 사용한다. 일반적인 CNN 심층신경망은 입력이 들어왔을 때 은닉층을 거쳐 출력이 나오는 방식을 사용한다. 이러한 구조에서는 Vanishing graient 이 발생하여 학습이 제대로 되지 않는 한계점이 있다. ResNet 의 Residual Block 은 입력이 들어왔을 때 출력으로 연결시키는 방법을 통해 Vanishing graient 문제를 해결하여 좋은 성능을 얻을 수 있다.[9]

■ 제안 작품 소개

가) GAN 을 사용한 적대적 예제



[그림 9] 생성적 적대 신경망 GAN의 원리

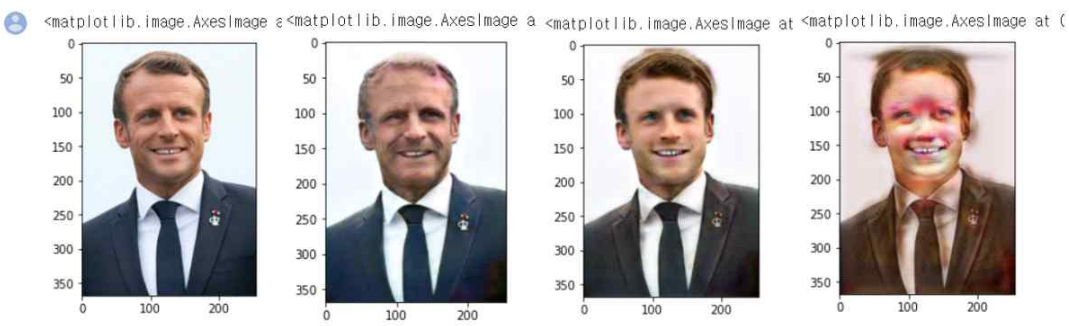
GAN 이란 생성적 적대 신경망(Generative Adversarial Network)의 줄임말로 비지도 학습에 사용되는 인공지능 알고리즘을 의미한다. 두 개의 심층신경망을 서로 경쟁시켜 진짜 같은 가짜 이미지를 생성해내는 인공지능 알고리즘이다. 예를 들어 설명하면, 위조지폐범은 위조지폐를 진짜 지폐와 비슷하게 만들어야 하고,

경찰은 위조지폐와 지폐를 구분할 수 있어야 한다. 이 과정에서 위조지폐범은 점점 더 좋은 위조지폐를 만들어내고, 경찰은 점점 더 위조지폐를 잘 구분해내며 위조지폐범과 경찰 모두 성장하게 된다. GAN 은 이러한 원리를 사용하여 가짜 이미지를 만들어내는 인공지능과 가짜 이미지와 진짜 이미지를 구분하는 인공지능, 두 개의 인공지능을 만들고 서로 경쟁시키며 더 정확한 이미지를 생성해내는 알고리즘이다. 현재 대부분의 이미지 생성 알고리즘에는 GAN 이 주로 사용된다.

참고자료로 삼기 위해 논문들을 찾아본 결과 현재 논문 주제들과 비슷한 주제의 논문들[10][11]은 모두 GAN 을 기반으로 연구되고 있음을 알게 되었다. 이에 가장 먼저 GAN 을 기반으로 논문을 쓰고 Adversarial Example 을 만들고자 시도하였다. GAN 기반 적대적 예제의 장점은 레이블 정보가 필요 없다는 것이다. 적대적 예제은 해당 이미지가 판다 레이블이 아닌 토이푸들 레이블로 인식하도록 만들어야 한다. 즉, 적대적 예제을 만들 때에는 레이블 정보가 필수적이다. 그러나 이미지를 생성하는 기술인 딥페이크 기술에서는 레이블 정보라고 할만한 것이 없다.

[그림10] GAN을 이용한 적대적 예제 (출처:[12])

따라서 해당 논문에서는 이미지의 레이블을 잘못 분류하도록 유도하는 것이 아니라 결과물로 나온 이미지에 왜곡이 많이 생기게 하는 방식을 사용한다. 엡실론(ε) 값이 커질수록 원래 이미지에는 변화가 없어 보여도 결과물 이미지에는 심하게 왜곡이 생겨 이미지 생성의 효과를 없애는 원리이다. [10][11]

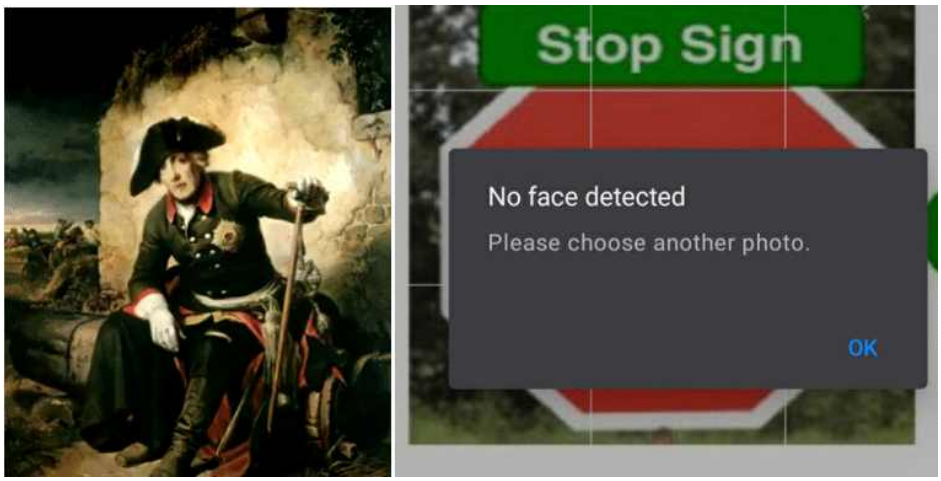


[그림 11] 실제 구현한 GAN을 사용한 적대적 예제

실제로 이를 구현하기 위해 가장 먼저 StarGAN 을 사용하여 프랑스의 마크롱 대통령(왼쪽)을 늙은 마크롱(가운데) 어린 마크롱(오른쪽)으로 만들었다. 이후 적대적 예제를 만들자 결과물을 알아볼 수 없을 정도로 왜곡이 끼는 것을 확인할 수 있었다.

이러한 GAN 을 사용한 적대적 예제는 가장 많은 논문 결과가 있으며 주로 사용된다. 그러나 이 적대적 예제의 단점은 사람의 육안 또한 이미지의 왜곡을 너무 쉽게 인식한다는 것이다. 따라서 GAN 을 사용한 적대적 예제는 '사람의 육안으로는 인식이 가능하나 인공지능은 인식하지 못하는 이미지를 만들어 딥페이크를 예방한다'는 목표에 알맞지 않았다.

나) Face recognition 을 사용한 적대적 예제



[그림 12] 실제 어플리케이션에서 만든 딥페이크 영상

Face recognition 을 기반으로 하는 적대적 예제는 GAN 에 비해 실패와 성공이 분명하다는 장점이 있다. GAN 기반 적대적 예제는 인식 실패가 아니라 이미지 생성을 방해하는 원리이다. 즉, 사람 또한 이미지의 왜곡을 육안으로 쉽게 인식할 수 있으며 정확한 성공의 여부를 가리기 어렵다. 그러나 Face recognition 을 기반으로 하는 적대적 예제의 경우 실용화된 어플리케이션에서 'No face detected' 알림이 뜬다는 확실한 성공 여부가 있으며 아예 이미지 생성 자체가 불가능하다는 장점을 가진다. 다만 레이블을 사용해야 하기 때문에 딥페이크 기술 예방에 적당하지 않으며, 따라서 관련 논문 또한 거의 없다는 단점이 있다.

적대적 예제에서는 특정 레이블을 가지고 있는 레이블을 다른 레이블로 인식하게 만드는 과정을 거친다. 그러나 딥페이크 예방을 위한 Face recognition 의 경우 특정

레이블이라고 할 만한 것이 없다. 그래서 생각해낸 아이디어가 바로 'Person 레이블'과 Person 이 아닌 다른 레이블'을 가진 데이터셋을 활용하여 적대적 예제를 만드는 방식이다. 예를 들어, 'Person'이라는 레이블을 'Car'라는 레이블로 인식하도록 적대적 예제를 만드는 것이다.

이를 실제로 구현하기 위해 가장 먼저 ImageNet 데이터셋과 MobileNetV2 모델을 사용하여 적대적 예제를 만들어보았다. 해당 ImageNet 데이터셋과 MobileNetV2 모델은 텐서플로우 홈페이지에서 제공하는 것을 사용하였으며, 레이블 분류를 방해하는 적대적 예시에 가장 대표적으로 사용되는 데이터셋과 모델이라 이미 학습된 모델 또한 많아 인공지능을 학습하는 시간을 줄일 수 있다는 장점이 있다.

[그림 13] 적대적 예제 구현 1

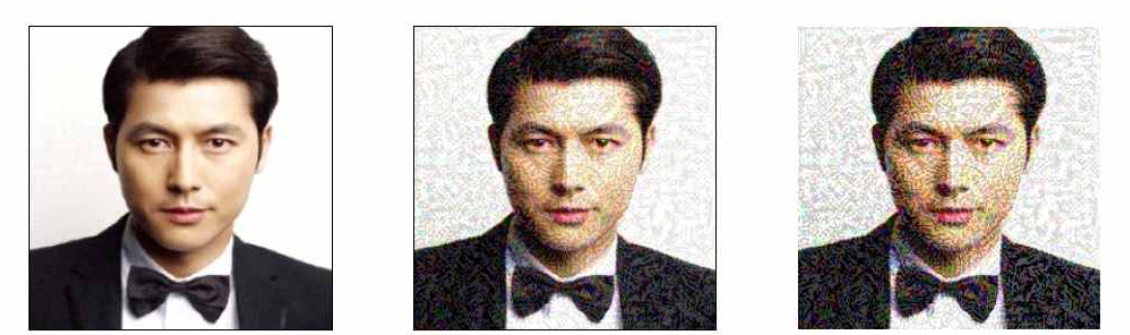
가장 먼저 텐서플로우 홈페이지에서 ImageNet 데이터셋과 MobileNetV2 모델을 다운로드받은 후 적대적 예제를 만드는 데에 사용할 정우성씨의 이미지를 업로드 하였다.

```
[ ]  loss_object = tf.keras.losses.CategoricalCrossentropy()

    def create_adversarial_pattern(input_image, input_label):
      with tf.GradientTape() as tape:
        tape.watch(input_image)
        prediction = pretrained_model(input_image)
        loss = loss_object(input_label, prediction)


      gradient = tape.gradient(loss, input_image)
      signed_grad = tf.sign(gradient)
      return signed_grad
```

[그림 14] 적대적 예제 구현 2

그 후 적대적 예제 코드를 구현하였다. 적대적 예제의 원리는 이미지 x 가 y 가 아닌 다른 레이블에 분류되도록 loss 함수의 gradient, 즉 기울기 값을 변화시키는 것이다. 이를 위해 기울기 값을 구하고 이동시켜야 할 방향을 구해야 한다.

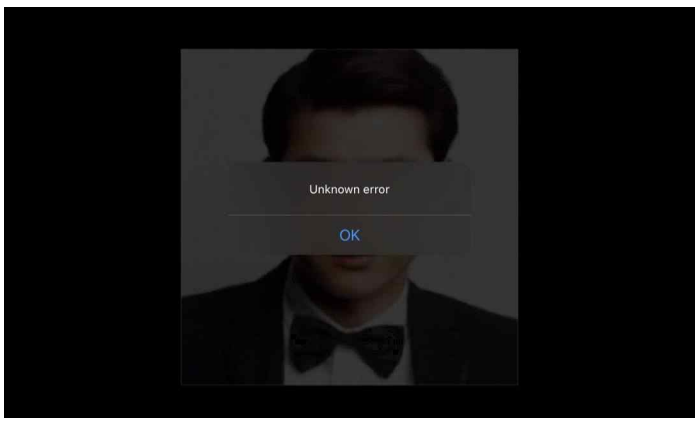
엡실론 값(왜곡값)을 점점 증가시켜 이미지의 왜곡을 더 심하게 만들 수 있다. 엡실론 값이 크면 클수록 이미지는 원래의 레이블에서 점점 멀어진다는 장점이 있지만 이미지에 왜곡이 심해져 원래의 이미지가 왜곡된다는 단점을 가진다.



[그림 15] 적대적 예제 구현 3

해당 이미지는 실제 엡실론 값이 0.01 일 때의 정우성씨, 0.1 일 때의 정우성씨, 0.15 일 때의 정우성씨이다. 엡실론 값이 커질수록 이미지에 왜곡이 심하게 끼는 것을 확인할 수 있다.

실제 상용화된 어플리케이션에도 만들어진 적대적 예제를 테스트해보았다.



[그림 16] 적대적 예제 구현 4

왜곡이 들어간 정우성씨의 이미지는 오른쪽과 같이 오류가 발생하며 어플리케이션에서 딥페이크 영상을 만들 수 없었다. 이렇게 적대적 예제를 만드는 데에 성공하였다.

그러나 해당 구현에는 문제점이 있었다. ImageNet 에는 '사람'에 대한 레이블이 없다는 것이다. 다른 적대적 예제 구현에는 상관이 없으나 논문의 목표인 '사람의 얼굴 이미지를 다른 레이블로 인식하도록 만든다'에는 부합하지 않았다. 이에 'person' 레이블이 있는 다른 데이터셋으로 실험을 진행하는 것이 맞다고 판단하여, 다른 데이터셋을 찾아 실험을 진행하였다.

이에 다른 데이터셋을 찾아 학습을 진행하고 적대적 예제를 만들어야 하는 상황이 되었다. 이를 위해 찾아본 논문들에서 사용한 데이터셋들은 다음과 같다.

- 1. MNIST
- 2. CIFAR
- 3. ImageNet
- 4. WIDER FACE, MS Celeb 1M, The 'Celebrity Together' Dataset, LFW dataset
- 5. Pascal VOC (object detection)
- 6. COCO Dataset (object detection)

원래는 Object detection 용도로 주로 이용되었기에 사용하지 않았던 Pascal VOC 과 COCO 데이터셋을 사용해 보기로 하였다.

Pascal VOC 데이터에서는 레이블 정보를 Xml 형태로 제공한다. 이에 Xml 파싱을 한 후 심층신경망 학습을 실행하였다.

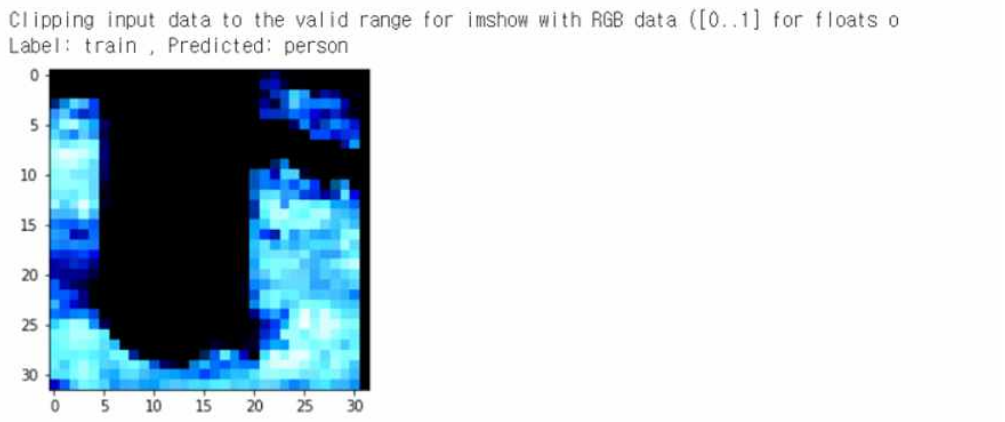
```
with torch.no_grad():
    for data in trainloader:
        images, labels = data[0].to(device), data[1].to(device)
        outputs = model(images)
        _, predicted = torch.max(outputs.data, 1)
        total += labels.size(0)
        correct += (predicted == labels).sum().item()

print(correct)
print(total)
print('Accuracy of the network on the 10000 test images: %d %%' % (100 * correct / total))
```

```
885
5717
Accuracy of the network on the 10000 test images: 15 %
```

[그림 17] 적대적 예제 구현 5

학습이 끝난 후 테스트를 진행하자 정확도가 15%라는 결과가 나왔다. 이 심층신경망을 토대로 사람의 얼굴 이미지를 분류하고 적대적 예제를 만드는 것은 어렵다. 이러한 결과가 나온 이유를 생각해보니 다음 2 가지로 원인이 좁혀졌다.

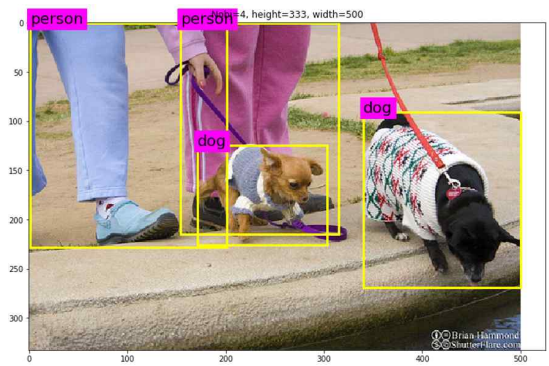
첫째, 이미지 전처리가 잘못되었다. 이미지 전처리는 크기 조절, 이미지 자르기, Normalize 를 진행하였다. 그러나 해당 이미지 전처리는 Classification 에 최적화되어있는 것으로 Object detection 에 주로 사용되는 Pascal VOC 는 이러한 이미지 전처리 과정에 적합하지 않았다. 예를 들어, Classification 데이터셋에서는 이미지를 크기에 맞춰 잘라도 중앙에 항상 레이블의 대상이 되는 이미지가 존재한다. 그러나 Object detection 데이터셋의 경우 레이블의 대상이 되는 이미지의 위치가 정해져있지 않다.



[그림 18] 적대적 예제 구현 6

실제로 이미지를 하나 분류해본 결과 원래의 레이블은 train 이었으나 분류기는 이를 person 으로 분류하였다. 그러나 사진을 보면 해당 사진은 train 으로도 person 으로도 보이지 않았다.

둘째, 레이블을 지정하는 데에 문제가 있었다. 해당 분류기에서 레이블은 하나만 결정된다. 그러나 Object detection 을 위해 Pascal VOC 데이터셋에서는 여러 개의 레이블을 제공한다. 이 중 하나만을 선택해야 하니 쓸데없는 라벨이 결정되는 일이 많았다.



[그림 19] 적대적 예제 구현 7

예를 들어 해당 사진에서 중요한 라벨은 dog 임에도 불구하고 person 으로 라벨링이 되었다. 이러한 문제를 해결하기 위해 Pascal VOC Xml 에 쓸데없는 라벨링을 없애주는 difficult 라는 항목이 있다는 사실을 알아내서 적용해주었다.

```
print(correct)
print(total)
print('Accuracy of the network on the 10000 test images: %d %%' % (100 * co

1384
5717
Accuracy of the network on the 10000 test images: 24 %
```

[그림 20] 적대적 예제 구현 8

그 결과 정확도가 24%로 증가하였다. 그러나 해당 심층신경망 역시 분류기로 사용하기에는 지나치게 낮은 수치였다.

해당 문제들은 Pascal VOC 데이터셋의 본질적인 요소에서 발생한 문제였으며, 해결하기에도 쉽지 않았다. 이에 데이터셋을 바꿔 심층신경망을 학습하고자 하였다. 그렇게 찾은 데이터셋이 바로 Cifar-100 이다. Cifar-100 은 8 만 개의 classification 학습용 데이터로, 100 개의 레이블 중 man 레이블이 있는 것을 확인하였다. 해당 데이터셋의 장점은 classification 학습용 데이터이며, people 레이블이 있다는 것이다. 다만 미리 제공되는 학습된 모델이 없어 직접 인공지능 학습을 진행해야 하며, 이미지의 크기가 32x32 픽셀이기 때문에, 이미지 사이즈를 224x224 픽셀로 증가시켜야 한다는 단점이 있다.

따라서 직접 인공지능 학습을 진행하기 위해 알고리즘을 선정했다. Cifar-100 관련 논문에서 주로 사용하는 Classification 알고리즘인 ResNet 을 사용하였다. ResNet 은 이미지 분류용 CNN 알고리즘으로 마이크로소프트에서 개발하였으며 많은 논문에서 사용되는 기본적인 알고리즘 중 하나이다. 이 알고리즘은 input 을 output 에 더할 수 있도록 지름길(shortcut)을 만들어 예측값과 실제 관측값의 차이를 최소화하는 방식을 사용하며, 2015 이미지넷 이미지 인식 대회(ILSVRC)에서 우승을 차지한 바 있다.

그렇게 Cifar-100 데이터셋과 ResNet 알고리즘을 사용하여 인공지능 학습을 진행하였다.

■ 구현 및 결과분석

가) 적대적 예제 구현

적대적 예제 이미지를 제작하기 위해, 사람의 얼굴을 인식하는 심층신경망 분류기를 만들었다. 해당 분류기는 많은 수의 클래스 레이블 중 사람의 레이블을 구분하는 업무를 수행해야 하기 때문에, 총 100 개의 클래스 레이블에 대해서 예측하며 사람 레이블을 포함하고 있는 Cifar-100 이미지 데이터셋을 사용하여 학습했다. 이미지의 크기는 적대적 예제 생성과 맞춰 224×224 로 통일시켰다.

```
learning_rate = 0.1
weight_decay = 1e-4

model = resnet32(num_classes=100)

model.to(device)

model.train()

# base optimizer with following parameters:
import torch.optim as optim

criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(), lr=learning_rate, momentum=0.9, weight_decay=weight_decay)
```

[그림 21] 적대적 예제 구현 9

또한 Pytorch 공식 홈페이지에서 이미지를 정규화하기 위해 공식적으로 사용하는 값인 mean=[0.485, 0.456, 0.406] 및 std=[0.229, 0.224, 0.225]를 사용하여 전처리를 진행하였다. 그 후 이미지 인식과 분류에 주로 사용되는 알고리즘인 ResNet-56 알고리즘을 사용하여 심층신경망을 학습시켰다. 여러 개의 클래스를 구분하는 다중 분류를 위해 CrossEntropyLoss 손실함수를 사용하였으며 최적화로는 SGD 를 사용하였다. Batch 사이즈는 32 이었으며 심층신경망의 학습 Epoch 횟수는 100 회 진행하였다.

```
plt.imshow(img.permute(1, 2, 0))
print('Predicted:', predict_image(img, model))
```

```
Clipping input data to the valid range for imshow with R
Predicted: man
```

```
img, label = testset[9]
plt.imshow(img.permute(1, 2, 0))
print('Label:', trainset.classes[label], ', Predict
```

```
Clipping input data to the valid range for imshow w
Label: apple , Predicted: apple
```

[그림 22] 적대적 예제 구현 10

학습 완료 후 test 결과이다. Apple 을 정확히 apple 로 인식하고 있으며 정우성씨 또한 정확히 man 으로 인식하고 있는 것을 확인할 수 있다. 이어 적대적 예제의 대상이 될 사람 얼굴 이미지들 또한 test 를 진행하였다.

심층신경망 학습 결과 테스트와 적대적 예제 생성을 위한 테스트 데이터셋으로는 대표적인 사람 얼굴 데이터셋인 CelebA를 사용하였다. CelebA는 약 20만 개의 사람 얼굴 이미지를 제공하는, 가장 대표적인 사람 얼굴 이미지 데이터셋 중 하나이다. 20만개 중 임의로 3295개의 이미지를 추려 테스트를 진행하였다.

```
print(len(file_list))
print(sum)
print(int(sum)/int(len(file_list))*100)
```

```
3295
3232
98.08801213960547
```

[그림 23] 적대적 예제 구현 11

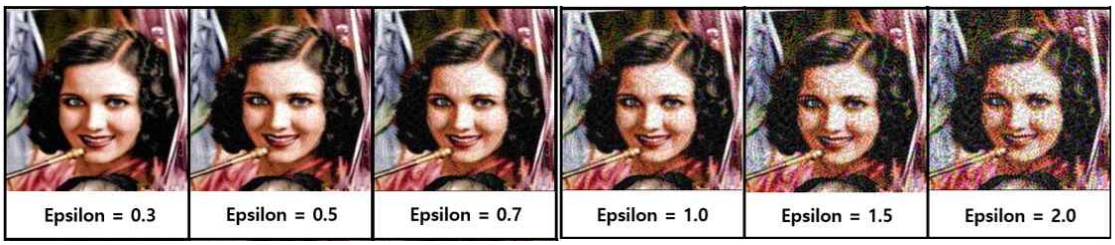
CelebA 데이터셋에서 제공하는 3295 개의 사람 얼굴 이미지를 토대로 만들어진 심층신경망을 테스트한 결과, 3295 개의 이미지 중 3232 개의 이미지가 사람 클래스 레이블에 분류되며 약 98.1%의 정확도를 보였다.

즉, 해당 분류기는 사람의 얼굴 이미지를 제대로 사람 클래스 레이블의 분류하고 있는 것을 확인할 수 있었다. 심층신경망이 제대로 학습된 것을 확인한 후 만들어진 심층신경망을 기반으로 적대적 예제를 만들었다.

```
# FGSM 공격 코드
def fgsm_attack(image, epsilon, data_grad):

    sign_data_grad = data_grad.sign()
    perturbed_image = image + epsilon*sign_data_grad
    perturbed_image = torch.clamp(perturbed_image, 0, 1)
    return perturbed_image
```

[그림 24] 적대적 예제 구현 12

적대적 예제를 만드는 데에는 FGSM 방법을 사용하였다. 심층신경망 손실함수의 기울기 부호를 구한 후, 그 부호대로 ε 값만큼 왜곡을 발생시켜 이미지에 추가하는 함수를 만든다. 그 후 미리 제작해둔 Test dataset 을 모델에 넣고 모델의 손실과, Gradient 을 계산하여 FGSM 공격을 실행한다. 그 후 이미지의 레이블을 재분류한다.



[그림 25] 구현된 적대적 예제

그렇게 만들어진 적대적 예제는 다음과 같다. 왜곡이 들어갔지만, 사람의 눈으로는 충분히 사람 얼굴로 인식이 가능하다. Epsilon 값이 커질수록 왜곡 또한 점점 심한 것을 확인할 수 있다.

이러한 방법으로 각 엡실론 값마다 3295 장의 적대적 예제가 만들어졌으며, 이렇게 만들어진 적대적 예제를 사람 얼굴 이미지 분류기에 넣고 사람 레이블로 판단하는지의 여부를 살펴보았다.

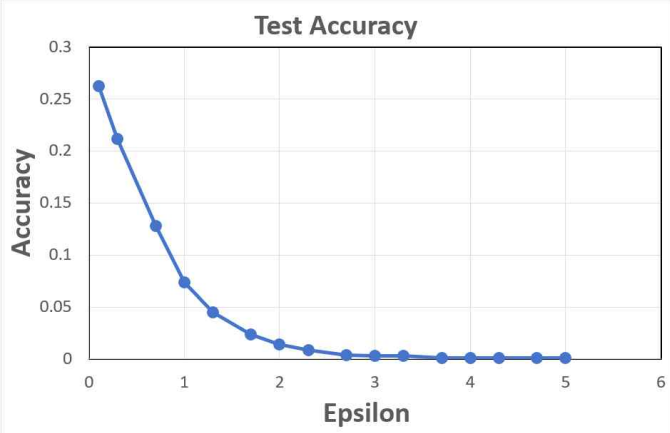
ε값은 0.01, 0.03, 0.07, 0.1, 0.13, 0.17, 0.2, 0.23, 0.27, 0.3, 0.33, 0.37, 0.4, 0.43, 0.47, 0.5로 상세하게 구분하여 어느 지점에서 적대적 예제가 심층신경망을 속이는데에 성공하는지 살펴보고자 하였다.

Epsilon	people image / total iamge	Test Accuracy
0.1	864 / 3295	0.262215478
0.3	698 / 3295	0.211836115
0.7	421 / 3295	0.127769347
1	243 / 3295	0.073748103
1.3	149 / 3295	0.04522003
1.7	78 / 3295	0.023672231
2	47 / 3295	0.014264036
2.3	28 / 3295	0.008497724
2.7	13 / 3295	0.003945372
3	11 / 3295	0.003338392
3.3	10 / 3295	0.003034901
3.7	4 / 3295	0.001213961
4	4 / 3295	0.001213961
4.3	4 / 3295	0.001213961
4.7	2 / 3295	0.00060698
5	2 / 3295	0.00060698

그 결과는 다음 [표 1]과 같다. ε값이 0.01일 때 3295개의 사람 얼굴 이미지 중 864개만이 사람 클래스 레이블로 분류되며 26%의 정확도를 보였으며, ε값이 점점 커질수록 정확도는 점점 낮아져 ε값이 0.2가 넘은 후에는 정확도가 1% 미만으로 떨어졌다. ε값이 0.3을 넘은 후에는 10장 이하의 이미지만이 사람 레이블로 구분되었다. 이를 그래프로 나타내면 다음과 같다.

[표 1] 적대적 예제의 심층신경망 테스트 결과

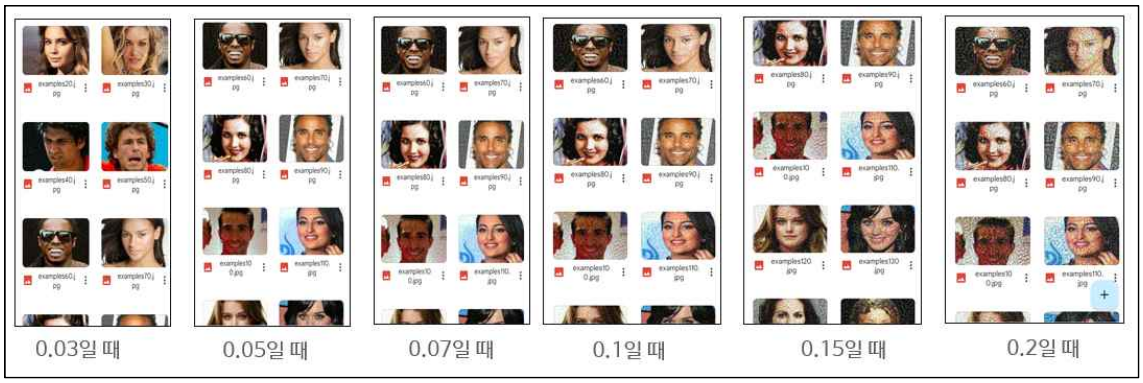
Epsilon 값이 커질수록 Accuracy는 낮아졌다.



[그림 26] 적대적 예제의 심층신경망 테스트 결과

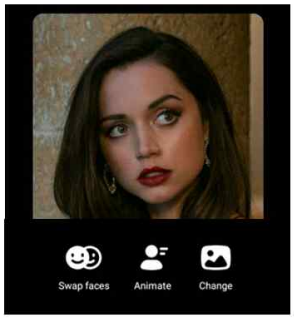
초반부에 급격하게 Accurary가 감소하며, Epsilon 값이 점점 커질수록 Accurary은 완만하게 감소하였다. 또한 ε값이 0.2가 넘어가면 대부분의 이미지들이 사람 레이블이 아닌 다른 레이블로 분류되었다. 즉, ε값이 점점 커질수록 분류기는 사람의 이미지를 사람 클래스 레이블이 아니라 다른 클래스 레이블로 구분하는 것을 확인하였다.

이렇게 만들어진 적대적 예제들을 실제 상용화된 어플리케이션에 넣어 딥페이크 방지 방법으로 효과가 있는지 확인한다. 이러한 방법을 사용할 수 있는 이유는 적대적 예제는 '전이성'이라는 특징을 갖기 때문이다. 특정 심층신경망을 공격하기 위해 만들어진 적대적 예시는 다른 심층신경망을 공격하는 데에도 효과적이다. 이러한 전이성은 심층신경망이 데이터의 일반적인 특성을 학습했기에 나타난다. 즉, 사람의 얼굴 이미지를 분류하는 심층신경망을 공격하는 적대적 예제는 전이성에 의해 사람 얼굴을 인식하는 딥페이크 어플리케이션에서도 효과를 볼 가능성이 높다.



[그림 27] 어플리케이션 적대적 예제 테스트 1

적대적 예제 테스트를 위해 사용한 어플리케이션은 구글 플레이 스토어에서 높은 다운로드 횟수를 기록한 Reface, FaceMasic이다. Reface는 다운로드 횟수가 1억이 넘은 대표적인 딥페이크 어플리케이션이며, FaceMasic은 다운로드 횟수 100만, 2만개의 리뷰와 4.7의 높은 별점을 받은 딥페이크 어플리케이션이다. 적대적 예제는 CelebA 데이터셋을 사용하여 만들어졌으며, ε값은 0.03, 0.05, 0.07, 0.1, 0.15, 0.2로 각 ε값마다 동일한 이미지로 100장의 적대적 예제를 만들어 테스트를 진행하였다.



얼굴 이미지 인식 성공한 경우

얼굴 이미지 인식 실패한 경우

Reface 어플리케이션에서 얼굴 이미지 인식을 성공한 경우라면 [그림 28]의 왼쪽과 같이 딥페이크 영상을 제작할 수 있다. 그러나 만약 얼굴 이미지 인식을 실패한 경우라면 오른쪽과 같이 'No face detected'라는 알림이 뜨게 된다. Reface 어플리케이션의 실험 결과는 다음과 같다.

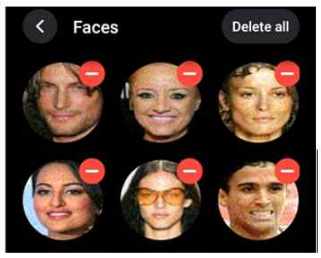
[그림 28] 어플리케이션 적대적 예제 테스트 2

엡실론값	0.03	0.05	0.07	0.1	0.15	0.2
인식o	97	65	16	0	0	0
인식x	3	35	84	100	100	100

[표 2] Reface 어플리케이션 적대적 예제 테스트 결과

Epsilon 값이 커질수록 이미지를 인식하지 못했다.

Epslion 값이 0.03인 경우에는 100장의 얼굴 이미지 중 97장이 인식되었다. 그러나 Epslion 값이 0.05로 증가하자 65장이 인식되었으며 0.07인 경우 16장만이 인식되었다. Epslion 값이 0.1을 넘은 순간부터는 그 어떠한 이미지도 인식되지 않았다.



얼굴 이미지 인식 성공한 경우

얼굴 이미지 인식 실패한 경우

[그림 30] 어플리케이션 적대적 예제 테스트 3

FaceMagic 어플리케이션에서 얼굴 이미지 인식을 성공한 경우라면 [그림 30]의 왼쪽과 같이 딥페이크 영상을 제작할 수 있도록 갤러리 리스트에 사진이 들어간다. 그러나 만약 얼굴 이미지 인식을 실패한 경우라면 오른쪽과 같이 'No face found' 라는 알림이 뜨게 된다. Reface 어플리케이션의 실험 결과는 다음과 같다.

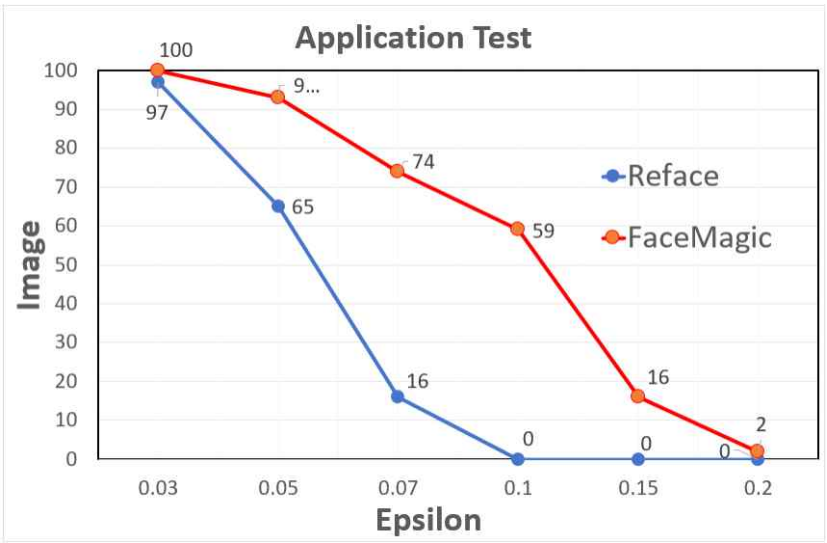
엡실론값	0.03	0.05	0.07	0.1	0.15	0.2
인식O	100	93	74	59	16	2
인식X	0	7	26	41	84	98

[표 3] FaceMagic 어플리케이션 적대적 예제 테스트 결과

Epsilon 값이 커질수록 이미지를 인식하지 못했다.

Epslion 값이 0.03인 경우에는 100장의 얼굴 이미지 전부가 인식되었다. 그러나 Epslion 값이 0.07로 증가하자 74장이 인식되었으며 0.1인 경우 59장만이 인식되었다. Epslion 값이 0.15인 경우 16장만이 인식되었으며 0.2인 경우 단 2장만이 인식되었다.

즉, Reface의 경우 ε값이 0.03일 경우 97장의 얼굴 이미지를 인식했으나, 0.07일 경우 16장의 이미지만을 인식하고 0.1 이상인 경우 단 한 장의 이미지도 인식하지 못하였다. FaceMagic의 경우 ε값이 0.03일 경우 100장의 얼굴 이미지를 모두 인식했으나, 0.1일 경우 59장의 이미지만을 인식하고 0.2일 경우 단 2장만의 얼굴 이미지를 인식하였다. 이를 그래프로 보면 다음과 같다.



[그림 2] 적대적 예제의 어플리케이션 테스트 결과

가로축은 Epsilon값, 세로축은 사람으로 인식된 이미지 개수로 Epsilon값이 커질수록 어플리케이션은 이미지를 인식하지 못했다.

어플리케이션마다 차이는 있지만 전체적으로 ε값이 커질수록 낮은 정확도를 보였으며, 적대적 예제 이미지의 얼굴 인식에 실패하여 영상을 제작하지 못하였다.

나) 결론

Cifar-100 데이터셋과 ResNet 모델을 기반으로 만들어진 심층신경망은 사람의 얼굴 이미지를 98% 정확하게 라벨링하는 데에 성공하며 높은 정확도를 보였다. 이 심층신경망을 기반으로 FGSM 방법을 사용하여 적대적 예제를 만들었으며, 적대적 예제의 ε값이 점점 커질수록 분류기가 사람의 이미지를 사람 레이블이 아니라 다른 레이블로 구분하는 것을 확인했다.

이렇게 만들어진 적대적 예제를 딥페이크 기술을 사용하여 실제 상용화된 영상 제작 어플리케이션인 Reface와 FaceMagic에 넣어 테스트했다. 그 결과, 사람의 눈으로는 적대적 예제의 이미지를 사람으로 인식할 수 있었음에도 불구하고, 어플리케이션은 얼굴 인식에 실패했다. 즉, 적대적 예제의 원리를 이용하여 이미지에 왜곡을 넣는 방법으로, 사람의 눈으로는 인식 가능하지만 어플리케이션은 인식하지 못하는 딥페이크 방지용 이미지가 만들어진 것을 확인했다.

■ 결론 및 소감

인공지능에도 관심이 많고 사이버보안에도 관심이 많아 주제를 무엇으로 할지에 대한 고민이 매우 컸다. 초반에는 DDoS 공격이나 블루투스 해킹 쪽에 대한 논문을 쓸 예정이었으나 친구들과 어플리케이션으로 재미로 만들어본 딥페이크 영상에 경각심을 가지게 되어 이를 막는 방법을 고민하던 도중 적대적 예제를 생각하게 되었다. 그리고 이 기술로 딥페이크의 악용을 막을 수 있을 거 같아 해당 주제를 선택하게 되었다.

논문 주제를 정할 때 가장 중요한 것은 자신이 흥미를 가진 주제여야 한다는 것을 이 경험으로 알게 되었다. 블루투스 해킹을 연구할 때는 스트레스를 굉장히 심하게 받았는데 적대적 예시를 연구할 때는 그래도 조금은 즐기면서 연구를 진행할 수 있었다. 주제를 정한 이후에도 문제는 계속되었다. 계속 데이터셋에 문제가 발생한 것이다. ImageNet 데이터셋에 person 레이블이 없는 것도 모자라 Cifar-100 데이터셋은 크기가 너무 작았기 때문이다. 결국 계속해서 인공지능을 학습시켜야만 했기 때문에 상당히 오랜 시간이 걸렸다. 만약 해당 기술을 정말 상용화하기 위해서는 이러한 문제 또한 고려해야할 것이다. 특히 크기 문제의 경우 실생활에서 사진은 정말 다양한 크기를 가지기 때문에 이러한 여러 픽셀을 모두 다룰 수 있는 적대적 예제 모델을 만들어야 할 것이다.

이번 논문을 통해 논문 하나를 쓰는 데에 정말 엄청난 시행착오가 들어간다는 사

실을 깨달았다. 처음에 주제를 선정했을 때에는 해당 분야에 관심도 있었고 주제와 관련된 논문과 자료들도 조금 찾아놨기 때문에 그렇게 어렵지 않을 것이라고 생각했으나, 논문이 진행될수록 얼마나 어리석은 생각이었는지 알게 되었다. 어떤 것은 코랩 환경에서는 따라 구현하기 너무 어려웠고 어떤 것은 나의 논문 주제에 적합하지 않았다. 결국 수많은 시행착오 끝에 이 자료에서 아이디어를 조금 인용하고 저 논문에서 아이디어를 조금 인용하여 연구를 계속 진행해야했다. 정말 중요한 것은 얼마나 자료가 많은지가 아니라 얼마나 쓸모있는 자료가 많은지였다는 것을 알게 되었다.

또한 중요한 것은 계획을 세우고 미리미리 논문을 준비하는 습관이었다. 사실 처음 계획을 세웠을 때는 이렇게 시행착오를 많이 겪을지 몰랐기 때문에 처음의 계획대로 된 것은 하나도 없었다. 그러나 저번 학기부터 미리미리 논문 작성을 시작하지 않고 논문작성일이 다가와 연구를 시작했다면 이러한 시행착오를 거칠 여유가 없었을 것이며 절대 연구를 끝낼 수 없었을 것이다. 격주 혹은 매주 조교님과의 발표가 있었기 때문에 억지로라도 매주 꾸준히 연구를 진행할 수 있었다. 즉, 만약 다음번에 다시 한번 논문을 쓰게 된다면 논문주제변경과 시행착오를 고려하여 전체적인 논문일정은 널널하게 잡되 매주 꾸준히 연구를 진행하는 것이 좋을 것 같다. 왜 이렇게 논문 준비를 일찍하는건지 궁금했는데 교수님과 조교님들은 여러번의 논문 준비로 다 경험이 있으셔서 학부생들을 배려하여 이러한 일정을 잡아주셨다는 것을 이해할 수 있었다.

조교님들과 매주 미팅을 진행한 것 또한 많은 도움이 되었다. 매주 꾸준히 연구를 진행할 수 있었던 것도 좋았지만 조교님들에게 발표하고 질문에 대답하며 내가 잘 이해하고 있지 않은 것이 무엇인지, 무엇을 더 보완해야할지 이해할 수 있었다. 특히 파라미터와 모델의 경우 왜 해당 파라미터와 모델을 선택했는지 그 이유를 적어야한다는 것과, 결과를 측정하기 위해 어떠한 방법을 사용해야 하는지와 같은 전혀 생각해보지 않았지만 논문을 쓰는 데에 중요한 요소들을 조교님들이 짚어주셨다. 또한 논문의 논리력이 부족한 부분들도 알려주셨기 때문에 해당 부분의 논리를 보충할 수 있었으며 작성된 논문의 피드백도 굉장히 상세하게 해주셨다.

그렇지만 연구는 역시 아무나 하는 것이 아니라는 것을 느꼈다. 좋아하는 주제로, 좋은 환경에서, 좋은 피드백을 받으며 연구를 진행하였음에도 스스로의 역량이 부족함을 많이 느꼈다. 그래도 하나의 논문을 작성하게 되어 굉장히 뿌듯했으며 잊지 못할 좋은 경험이 되었다.

■ 참고문헌

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition" 2015

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks.", the International Conference on Learning Representations (ICLR), 2014

[3] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples.", the International Conference on Learning Representation (ICLR), 2015

[4] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," Neural Information Processing Systems (NeurIPS), 2019

[5] Few-Shot Adversarial Learning of Realistic Neural Talking Head Models, Samsung AI Centor, 2019

[6] "사진 한장으로 인터뷰 동영상 제작"...삼성, AI 신기술 개발, 동아사이언스, 이승관 기자,https://m.dongascience.com/news.php?idx=28962

[7] 딥러닝 기반 얼굴인식 모델에 대한 변조 영역 제한 기만공격, 류권상, 정보보호학회지, 2019

[8] Efficient Decision-based Black-box Adversarial Attacks on Face Recognition, CVPR, 2019

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition" 2015

[10] Defending against GAN-based Deepfake Attacks via Transformation-aware Adversarial Face, Chaofei Yang, Duke University, cs.CV, 2020

[11] Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems, Nataniel Ruiz, Boston University, cs.CV, 2020

[12] Efficient Decision-based Black-box Adversarial Attacks on Face Recognition, CVPR, 2019