정 보 보 호 학 회 지
제 30 권 제5호, 2020. 10

GAN 기반 은닉 적대적 패치 생성 기법에 관한 연구

김 용 수*, 강 효 은*, 김 호 원*

요 약

딥러닝 기술은 이미지 분류 문제에 뛰어난 성능을 보여주지만, 공격자가 입력 데이터를 조작하여 의도적으로 오작동을 일으키는 적대적 공격(adversarial attack)에 취약하다. 최근 이미지에 직접 스티커를 부착하는 형태로 딥러닝 모델의 오작동을 일으키는 적대적 패치(adversarial patch)에 관한 연구가 활발히 진행되고 있다. 하지만 기존의 적대적 패치는 대부분 눈에 잘 띄기 때문에 실제 공격을 받은 상황에서 쉽게 식별하여 대응할 수 있다는 단점이 있다. 본 연구에서는 GAN(Generative Adversarial Networks)을 이용하여 식별하기 어려운 적대적 패치를 생성하는 기법을 제안한다. 실험을 통해 제안하는 방법으로 생성한 적대적 패치를 이미지에 부착하여 기존 이미지와의 구조적 유사도를 확인하고 이미지 분류 모델에 대한 공격 성능을 분석한다.

Ⅰ. 서 론

최근 딥러닝 기술의 급속한 발전으로 인해 얼굴 인식, 객체 탐지, 질병 진단 등 다양한 지능형 시스템에 적용되고 있다. 하지만, 최근 연구에 따르면 딥러닝 기술은 공격자가 입력 데이터를 조작하여 의도적으로 오작동을 일으키는 적대적 공격(adversarial attack)에 취약한 것으로 밝혀졌다.

적대적 공격 기법 중 이미지에 작은 노이즈를 추가하여 딥러닝 모델을 오작동하게 만드는 적대적 교란 신호(adversarial perturbation)에 관한 연구가 활발히 진행되고 있다. 적대적 교란 신호에 관한 대부분의 연구들은 다양한 최적화 기법을 통해 딥러닝 모델을 오작동하게 만드는 작은 노이즈를 찾는 방법이며, 대표적으로 FGSM(Fast Gradient Sign Method)[1], PGD(Projected Gradient Descent)[2], C&W(Carlini & Wagner)[3] 공격 기법이 있다. 하지만 적대적 교란 신호는 이미지의 픽셀 값을 직접적으로 조작하는 방법이기 때문에, 카메라 시스템을 해킹하여 공격하는 것과 같이 디지털 환경에서만 공격이 가능하다는 한계점이 있다.

적대적 교란 신호와 비교하여, 이미지에 직접 스티커를 부착하는 형태로 딥러닝 모델의 오작동을 일으키는 적대적 패치(adversarial patch)[4]가 있다. 하지만 적대적 패치에 관한 기존 연구들은 대부분 패치가 눈에 잘 띄기 때문에 실제 공격을 받은 상황에서 쉽게 식별하여 대응할 수 있다는 단점이 있다.

본 연구에서는 이러한 문제를 해결하기 위해 GAN(Generative Adversarial Networks)[5]을 이용하여 식별하기 어려운 적대적 패치를 생성하는 기법을 제안한다. 제안하는 기법은 이미지의 일부 영역에 교란 신호를 추가하여 이를 적대적 패치로 활용하는 것이며, 원하는 영역의 크기 및 위치에 해당하는 일부분의 적대적 교란 신호를 통해 기존 이미지에 은닉할 수 있는 적대적 패치를 생성할 수 있다. 실험에서는 기존의 적대적 패치 및 적대적 교란 신호 기법과 구조적 유사도 및 공격 성능을 비교하며, 제안하는 기법이 다른 기법들에 비해 공격 성능이 크게 떨어지지 않으면서도 구조적 유사도가 매우 높은 것을 확인할 수 있다.

Ⅱ. 관련 연구

이 장에서는 적대적 교란 신호 및 적대적 패치에 관한 연구들을 소개하고, 생성 모델인 GAN과 GAN을 이용한 적대적 공격 기법을 소개한다.

이 논문은 국토교통부의 스마트시티 혁신인재육성사업으로 지원되었습니다.
* 부산대학교({대학원생, dkgoggog0329, hyoeun0915, 교수, howonkim}@pusan.ac.kr)

2.1. 적대적 교란 신호

적대적 교란 신호는 딥러닝 모델을 높은 신뢰도로 오작동하게 만드는 목적으로 이미지에 추가되는 작은 노이즈를 의미한다[6]. 적대적 교란 신호는 다음과 같이 정의할 수 있다. 입력 이미지 x를 실제 클래스 y로 분류하는, 즉 $M(x)=y$를 만족하는 학습된 딥러닝 모델 M이 있을 때, 사람이 구별할 수 없을 정도의 작은 크기를 가지는 노이즈 η를 이미지 x에 추가하여 만든 x'을 클래스 y로 분류하지 못하는 경우, 즉 $M(x')\neq y$를 만족할 때의 η를 적대적 교란 신호라고 정의한다.

적대적 공격은 공격자의 지식 정보에 따라 크게 화이트박스 공격, 블랙박스 공격으로 나눌 수 있다[7,8]. 화이트박스 공격은 공격자가 딥러닝 모델의 구조 및 파라미터, 학습 데이터 등의 정보를 모두 알고 있을 때의 공격을 의미한다. 대표적인 화이트박스 공격 기법으로는 L-BFGS[6], FGSM[1], DeepFool[9] 등이 있다. 블랙박스 공격은 화이트박스 공격과 반대로 공격자가 딥러닝 모델의 정보를 모를 때의 공격을 의미한다. 이때의 공격자는 입력 데이터에 대한 딥러닝 모델의 출력값만을 가지고 공격을 수행해야 한다. 적대적 교란 신호는 같은 작업을 수행하는 다른 딥러닝 모델도 어느 정도 공격이 가능한 전이성을 가지고 있기 때문에, 블랙박스 공격에서는 주로 이러한 전이성을 이용한 공격이 이루어진다[12].

2.2. 적대적 패치

적대적 교란 신호는 실제 공격 상황에서 디지털 환경에서만 적용할 수 있다는 한계점이 있다. 이와 반대로, 적대적 패치는 물리적 환경에서 실제 이미지에 부착하여 공격할 수 있다는 장점이 있다. Brown 등[4]은 실제 물리적 환경에서 발생할 수 있는 변형 및 왜곡에도 강한 적대적 패치를 제안하였다. 또한, DPatch[10]는 Faster R-CNN, YOLO와 같은 객체 탐지 딥러닝 모델을 공격할 수 있는 적대적 패치이다. 이러한 적대적 패치는 물리적 환경에서 공격 성능을 높이기 위해 기존 이미지에 비해 왜곡 현상이 심하며, 이로 인해 사람이 쉽게 식별할 수 있다는 단점이 있다. 본 연구에서는 물리적 환경에서 적용할 수 있음과 동시에 식별하기 어려운 적대적 패치를 생성하는 기법을 제안한다.

2.3. GAN(Generative Adversarial Networks)

Goodfellow 등[5]은 생성자(generator) 및 판별자(discriminator)로 구성된 두 종류의 모델을 적대적으로 훈련함으로써 실제 데이터와 매우 유사한 데이터를 생성할 수 있는 프레임워크를 제안하였다. 판별자 D는 실제 데이터와 생성자가 생성한 데이터를 분류하도록 학습되며, 생성자 G는 데이터의 분포를 모델링하여 실제 데이터와 유사한 데이터를 생성함으로써 판별자 D를 속이도록 학습된다. 두 개의 모델이 경쟁적으로 학습됨으로써 결과적으로 생성자 G는 실제 데이터와 매우 흡사한 데이터를 생성할 수 있게 된다.

Xiao 등[11]은 GAN을 이용한 적대적 교란 신호를 생성하는 기법을 제안하였다. 여기서의 생성자 G는 판별자를 속임과 동시에, 공격 대상 모델 T를 오작동하게 만드는 적대적 교란 신호를 생성하도록 학습된다. 결과적으로 생성자 G는 원본 이미지와 매우 유사하면서도 딥러닝 모델의 성능을 크게 떨어뜨리는 적대적 교란 신호를 생성할 수 있게 된다.

Liu 등[12]은 GAN을 이용하여 시각적으로 자연스러운 적대적 패치를 생성하는 기법을 제안하였다. 해당 연구에서는 GAN 모델에 patch-to-patch translation 기법을 적용하여, 원본 이미지에 자연스러운 seed patch를 입력하여 적대적 패치로 변환하는 작업을 진행하였다. 하지만 적용한 seed patch 또한 원본 이미지와 뚜렷한 차이가 있기 때문에, 여전히 식별하기 어려운 적대적 패치와는 거리가 멀다. 본 연구에서는 GAN 모델로 적대적 교란 신호를 생성하여 일부분의 영역을 활용함으로써 은닉된 적대적 패치를 생성하는 기법을 제안한다.

Ⅲ. 본 론

이 장에서는 본 연구에서 제안하는 GAN 기반의 은닉 적대적 패치를 생성하는 기법 및 과정을 소개한다.

3.1. 문제 정의

데이터세트 X와 결과 레이블 Y 집합이 있을 때, 딥러닝 모델 T는 분류 함수 $F: X\rightarrow Y$를 만족하도록 학습된다. 본 연구에서 정의하는 문제는 $F(x)=y$를 만족하는 데이터 $x\in X, y\in Y$에 대해서, 적대적 사례

$x_A = x + p$를 분류 함수에 입력하면 레이블 y로 분류하지 못하게 하는, 즉 $F(x_A) \neq y$를 만족하는 적대적 패치 p를 생성하는 것이 목적이다. 또한, 적대적 패치 p를 추가한 x_A는 원본 데이터 x와 구분할 수 없을 정도로 유사해야 한다. 본 연구에서는 구조적 유사도 (structural similarity index measure, SSIM)[21]를 통해 적대적 패치의 은닉성을 측정한다.

3.2. 은닉 적대적 패치 생성 모델 구조

제안하는 적대적 패치 생성 모델 구조는 그림 1과 같으며, 크게 생성자 G, 판별자 D, 그리고 공격 대상 모델 T로 구성되어 있다. 모델 구조는 GAN 기반의 적대적 교란 신호를 생성하는 기존 연구[11]와 유사하다. 본 연구에서는 생성자 G를 통해 적대적 교란 신호를 생성한 후, 원하는 영역을 지정하여 해당하는 부분만 기존 이미지에 더하여 적대적 패치로 활용하는 기법을 제안한다. 제안하는 기법을 통해 결과적으로 적대적 교란 신호의 원하는 영역을 잘라내어 직접 이미지에 부착할 수 있는 적대적 패치로 사용할 수 있다.

판별자 D는 원본 이미지 x와 적대적 사례 $x_A = x + p$를 구분하도록 학습되며, 생성자 G와 경쟁적으로 학습하여 결과적으로 생성자 G가 원본 이미지와 구분하기 어려운 적대적 패치를 생성할 수 있게 도와주는 역할을 한다. 또한, 생성자 G는 공격 대상 모델 T를 오작동하게 만드는 적대적 패치를 생성할 수 있어야 한다. 따라서, 본 연구에서 제안하는 적대적 패치 생성 모델은 다음과 같은 세 가지 종류의 손실 함수를 포함한다.

$$L_{adv} = E_x(\log D(x) + \log(1 - D(x_A))) \quad (1)$$

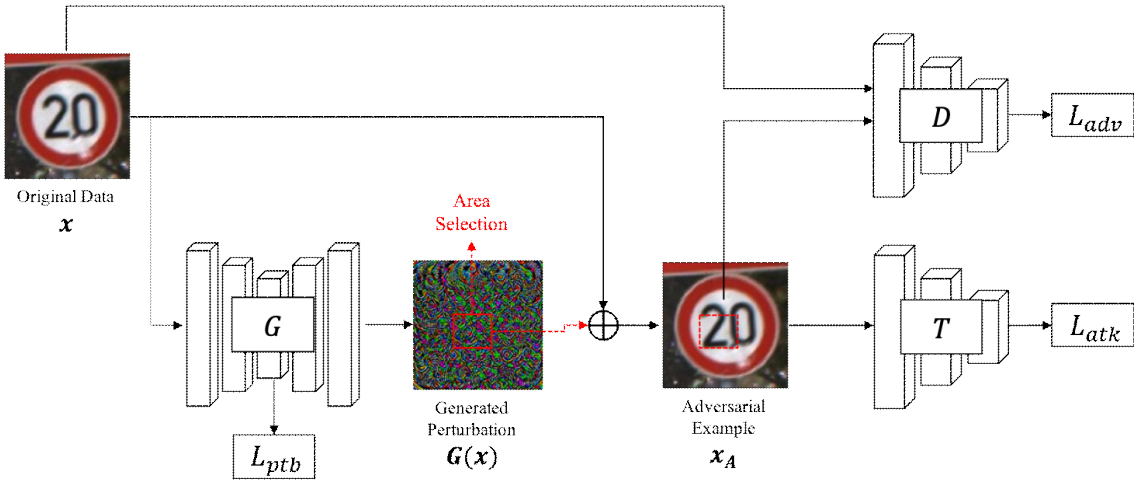
$$L_{atk} = -E_x l_T(x_A, y) \quad (2)$$

$$L_{ptb} = E_x \max(0, \| G(x) \|_2 - c) \quad (3)$$

수식 1은 판별자 D가 원본 이미지 x와 적대적 사례 $x_A = x + p$를 구분하도록 학습하는 데 사용되는 손실 함수를 나타낸다. 생성자 G의 입장에서는 해당하는 손실 함수를 최소화하도록 학습되어, 원본 이미지와 생성한 적대적 패치를 부착한 적대적 사례를 구분하기 어려워진다.

수식 2는 공격 대상 모델 T를 오작동하게 만드는 적대적 패치를 생성하는 데 사용되는 손실 함수를 나타낸다. l_T는 공격 대상 모델 T에 적용된 cross-entropy와 같은 손실 함수이다. 생성자 G는 수식 2를 최소화하도록 학습되며, 이는 적대적 사례 x_A를 x의 레이블인 y로 분류하도록 하는 손실 함수를 최대화하는 것과 같다. 즉, 공격 대상 모델 T를 오작동하게 만드는 적대적 사례를 생성하기 위한 손실 함수라고 할 수 있다.

수식 3은 생성자 G가 생성하는 적대적 교란 신호의 L_2 norm의 임계값을 c로 설정하는 단계를 나타낸다. 즉, 적대적 교란 신호의 크기가 특정한 값을 넘지 못하도록 학습하여 생성하는 적대적 패치의 은닉성을 강화



〔그림 1〕 제안하는 은닉 적대적 패치 생성 모델 구조도

한다. 세 가지 종류의 손실 함수를 통합한 최종 손실 함수는 다음과 같다.

$$L_{GAN} = L_{adv} + \alpha L_{atk} + \beta L_{ptb} \quad (4)$$

$$\min_G \max_D L_{GAN} \quad (5)$$

수식 4의 $\alpha > 0$와 $\beta > 0$는 각각 공격 대상 모델의 공격에 기여하는 정도와 은닉성에 대한 파라미터를 나타낸다. 제안하는 모델의 생성자 G와 판별자 D는 수식 5와 같이 각각 L_{GAN}을 최소화하고 최대화하도록 경쟁적으로 학습하여, 최종적으로 생성자 G는 식별하기 어려운 특성과 공격 대상 모델을 오작동하도록 공격하는 특성을 가지는 적대적 패치를 생성할 수 있게 된다.

Ⅳ. 실 험

4.1. 실험 환경

본 연구에서 제안하는 은닉 적대적 패치 생성 모델을 구성하기 위해, image-to-image translation[13,14]에 사용되는 GAN과 유사한 모델 구조를 사용하였다. 생성자 G는 인코더와 디코더로 구성된 U-Net 구조를 사용하였고, 이는 모델 각 층에 skip connection 기법을 적용하여 고해상도 이미지를 생성하는 데 이점이 있다. 판별자 D는 이미지 분류 정확도가 높은 CNN(Convolutional Neural Network) 구조를 사용하여 생성자 G의 적대적 교란 신호의 생성 능력 향상에 기여하였다.

본 연구에서는 GTSRB[15]와 CIFAR-10[16] 데이터세트를 사용하여 실험을 진행하였다. GTSRB(German Traffic Sign Recognition Benchmark) 데이터세트는 43가지 종류의 교통표지판 이미지 50,000장으로 구성되어 있다. CIFAR-10 데이터세트는 컴퓨터 비전 알고리즘 실험에 주로 사용되며, 10가지 종류의 클래스와 60,000장의 이미지로 구성되어 있다.

또한, 실험에서의 공격 대상 모델은 이미지 분류 문제에 높은 성능을 보이는 대표적인 CNN 모델인 VGG16[17], ResNet34[18], SqueezeNet1.1[19], MobileNetV2[20]을 사용하였다. 공격 대상 모델을 GTSRB, CIFAR-10 데이터세트에 대해 각각 학습하여 공격 실험을 진행하였다. 위의 데이터세트에 대한 각 공격 대상 모델의 분류 정확도는 표 1과 같다. 실험 환경으로 운영체제는 Ubuntu 18.04 LTS, CPU는 Intel Core i9-7980XE, 메모리 64GB, GPU는 NVIDIA Titan XP를 활용하였다. 또한 딥러닝 모델 학습을 위해 Python 3.6 기반의 PyTorch 라이브러리를 사용하였다.

〔표 1〕 GTSRB, CIFAR-10 데이터세트에 대한 각 공격 대상 모델의 분류 정확도

Target Model	Classification Accuracy	
	GTSRB	CIFAR-10
VGG16	91.9%	92.3%
ResNet32	92.1%	94.8%
SqueezeNet1.1	90.8%	91.2%
MobileNetV2	92.3%	94.5%

4.2. 실험 결과

실험은 먼저 공격 대상 모델의 정보를 알고 있는 상황인 화이트박스 공격 실험, 그리고 전이성을 이용하여 특정 공격 대상 모델을 통해 생성한 적대적 패치로 다른 공격 대상 모델을 공격하는 블랙박스 공격 실험, 마지막으로 생성한 적대적 패치를 실제 교통표지판 모형에 부착하여 공격을 확인하는 물리적 공격 실험을 진행하였다.

4.2.1. 화이트박스 공격 실험

화이트박스 공격 실험에서는 제안하는 적대적 패치 생성 기법의 은닉성과 공격 성능을 비교하기 위해 관련된 기존 연구[4,11,12]와의 비교 실험을 진행하였다. 은닉성은 구조적 유사도(SSIM), 공격 성능은 공격 대상 모델의 분류 정확도를 보고 판단한다. 제안하는 기법의 적대적 패치의 크기는 원본 이미지의 약 6.7%의 크기로, 위치는 임의로 지정하여 생성하였다. 각 데이터세트에서 무작위로 100장씩의 이미지를 선택하여 공격 실험을 진행하였으며, 구조적 유사도와 분류 정확도는 각각의 평균값을 기록하였다. 비교 실험 결과는 표 2와 같다.

표 2에서 구조적 유사도(SSIM)가 높을수록 은닉성이 높고, 분류 정확도(Acc)가 낮을수록 공격 성능이 높다고 볼 수 있다. 표 2에서 제안하는 기법의 공격 성능이 기존 연구들에 비해 크게 떨어지지 않으면서도, 구조

〔표 2〕 제안하는 적대적 패치 생성 기법의 은닉성 및 공격 대상 모델 분류 정확도 비교 실험 결과

Method \ Dataset	GTSRB		CIFAR-10	
	SSIM	Acc.	SSIM	Acc.
Our Method	**0.956**	**16.4%**	**0.981**	**9.8%**
Adv. Patch[4]	0.824	6.3%	0.839	2.7%
AdvGAN[11]	0.925	14.1%	0.931	6.7%
PS-GAN[12]	0.871	12.5%	0.895	4.9%

적 유사도가 매우 높은 것을 확인할 수 있다. 이는 기존 이미지에 적대적 패치를 부착하더라도 식별하기 어려운 특징을 나타내며, 다른 기법들에 비해 은닉성이 높다는 의미로 볼 수 있다. 그림 2는 제안하는 기법과 기존 연구들을 실제 이미지에 적용한 예시이며, 제안하는 기법의 적대적 패치가 상대적으로 식별하기 어려운 특징을 지님을 알 수 있다.

〔그림 2〕 기존 연구들과 제안하는 은닉 적대적 패치 기법 비교

4.2.2. 블랙박스 공격 실험

화이트박스 공격 실험과는 달리, 블랙박스 공격 실험에서는 공격자가 공격 대상 모델의 정보를 모르는 상황으로 가정하여 실험을 진행한다. 따라서, 여기서는 특정 공격 대상 모델을 대상으로 생성한 적대적 패치를 다른 공격 대상 모델에 적용하는, 즉 전이성을 이용한 실험을 진행하였다. 표 3은 GTSRB 데이터세트에서 특정 공격 대상 모델을 대상으로 생성한 적대적 패치를 다른 공격 대상 모델에 적용한 후의 평균 분류 정확도를 나타낸다. 표 3에서 다른 공격 대상 모델의 분류 정확도 또한 매우 낮은 양상을 보이며, 이는 제안하는 기법의 전이성이 높으며 블랙박스 공격 환경에서도 잘 동작한다는 것을 알 수 있다.

〔표 3〕 제안하는 적대적 패치 생성 기법의 전이성을 이용한 블랙박스 공격 실험 결과

Source \ Target	VGG	ResNet	Squeeze Net	Mobile Net
VGG	**19.4%**	31.3%	28.4%	34.5%
ResNet	28.8%	**14.5%**	30.9%	24.5%
SqueezeNet	22.4%	28.4%	**15.1%**	27.3%
MobileNet	31.1%	26.7%	23.9%	**17.7%**

4.2.3. 물리적 공격 실험

제안하는 기법의 적대적 패치를 실제 환경에서 적용할 수 있는지 검증하기 위해, 교통표지판 모형에 부착하여 물리적 공격 실험을 진행하였다. 교통표지판 모형의 크기는 약 15×15 cm이며 10가지 종류로 구성되어 있다. 실험을 진행하기 위해 실제 교통표지판 모형 사진을 100장씩 촬영하여 ResNet 모델을 학습하였고, 평균 약 97.8%의 정확도로 교통표지판 모형을 분류하였다. 그런 다음 교통표지판 사진 중 무작위로 100장의 이미지를 추출하여 제안하는 기법을 통해 적대적 패치를 생성

〔그림 3〕 제안하는 기법의 적대적 패치를 통한 교통표지판 모형 분류 모델 공격 예시

한 결과, 학습한 ResNet 모델의 평균 분류 정확도가 97.8%에서 약 39.0%로 하락하였다. 이는 화이트박스 공격 실험에 비해 실제 환경에서는 공격 성능이 약간 떨어지지만, 여전히 기존 딥러닝 모델의 분류 정확도를 크게 떨어뜨리기 때문에 치명적인 공격이라고 볼 수 있다. 그림 3은 제안하는 기법의 적대적 패치로 교통표지판 모형 분류 모델 공격에 성공한 예시이다.

Ⅴ. 결 론

본 연구에서는 GAN을 이용하여 식별하기 어려운 적대적 패치를 생성하는 기법을 제안하였다. 제안하는 기법은 기존의 GAN 기반 적대적 교란 신호 생성 모델에서 착안하여, 이미지의 일부 영역에만 교란 신호를 생성하여 이를 적대적 패치로 활용하여 실제 물리적 환경에서도 공격이 가능하다. 화이트박스 및 블랙박스 공격 환경의 실험 결과에서, 제안하는 기법으로 생성한 적대적 패치의 은닉성 및 공격 성능이 모두 높은 것을 확인할 수 있다. 또한, 실제 교통표지판 모형을 대상으로 공격 실험을 진행하여, 물리적 환경에서도 제안하는 기법이 치명적인 공격이 될 수 있음을 보여주었다. 향후 연구로는 물리적 환경에서의 촬영 거리, 조명 등과 같은 왜곡 정보에 대해서도 강한 적대적 패치를 생성하도록 제안하는 기법을 보완하는 추가적인 연구를 진행할 예정이다.

참 고 문 헌

[1] I. Goodfellow, et al., “Explaining and Harnessing Adversarial Examples,” *International Conference on Learning Representations*, 2015.

[2] A. Madrym et al., “Towards Deep Learning Models Resistant to Adversarial Attacks,” *arXiv preprint* arXiv:1706.06083, 2017.

[3] N. Carlini, D. Wagner, “Towards Evaluating the Robustness of Neural Networks,” *IEEE Symposium on Security and Privacy*, 2017.

[4] T. B. Brown, et al., “Adversarial Patch,” *arXiv preprint* arXiv:1712.09665, 2018.

[5] I. Goodfellow, et al., “Generative Adversarial Nets,” *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. 2, pp. 2672-2680, 2014.

[6] C. Szegedy, et al., “Intriguing Properties of Neural Networks,” *International Conference on Learning Representations*, 2014.

[7] S. Qiu, et al., “Review of Artificial Intelligence Adversarial Attack and Defense Technologies,” *Applied Sciences*, 9(5), 2019.

[8] X. Yuan, et al., “Adversarial Examples: Attacks and Defenses for Deep Learning,” *IEEE Transactions on Neural Networks and Learning Systems*, 30(9), pp.2805-2824, Sep 2019.

[9] S. Moosavi-Dezfooli, et al., “DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks,” *arXiv preprint* arXiv:1511.04599, 2015.

[10] X. Liu, et al., “DPatch: Attacking Object Detectors with Adversarial Patches,” *arXiv preprint* arXiv:1806.02299, 2018.

[11] C. Xiao, et al., “Generating Adversarial Examples with Adversarial Networks,” *arXiv preprint* arXiv:1801.02610, 2018.

[12] A. Liu, et al., “Perceptual-Sensitive GAN for Generating Adversarial Patches,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, pp. 1028-2035, 2019.

[13] P. Isola, et al., “Image-to-Image Translation with Conditional Adversarial Networks,” *arXiv preprint* arXiv:1611.07004, 2016.

[14] J. Y. Zhu, et al., “Unpaired Image-to-Image Translation using Cycle-consistent Adversarial Networks,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 2242-2251, 2017.

[15] J. Stallkamp, et al., “Man vs. Computer: Benchmarking machine learning algorithms for traffic sign recognition,” *Neural Networks : the official journal of the International Neural Network Society*, 32, 2012.

[16] A. Krizhevsky, et al., “CIFAR-10 (Canadian Institute For Advanced Research),” *http://www.cs.toronto.edu/~kriz/cifar.html*.

[17] K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-scale Image

Recognition," *arXiv preprint* arXiv:1409.1556, 2015.

[18] K. He, "Deep Residual Learning for Image Recognition," *arXiv preprint* arXiv:1512.03385, 2015.

[19] F. N. Iandola, et al., "SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and <0.5MB Model Size," *arXiv preprint* arXiv:1602.07360, 2016.

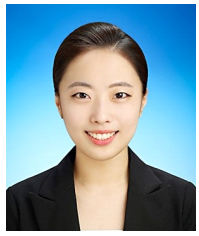
[20] M. Sandler, et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *arXiv preprint* arXiv:1801.04381, 2019.

[21] Z. Wang, et al., "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, 13(4), pp. 600-612, 2004.

〈저 자 소 개〉

김 용 수 (Yongsu Kim)

2019년 2월 : 부산대학교 전기컴퓨터공학부 졸업
2019년 3월~현재 : 부산대학교 정보융합공학과 석사과정
<관심분야> 딥러닝, 인공지능, AI 보안



강 효 은 (Hyoeun Kang)

2017년 2월 : 부산대학교 IT응용공학과 졸업
2018년 3월~현재 : 부산대학교 정보융합공학과 석박통합과정
<관심분야> 딥러닝, 인공지능, AI 보안

김 호 원 (Howon Kim)

1993년 2월 : 경북대학교 전자공학과 졸업
1995년 2월 : 포항공과대학교 전자전기공학과 석사
1999년 2월 : 포항공과대학교 전자전기공학과 박사
2008년 3월~현재 : 부산대학교 전기컴퓨터공학부 교수
<관심분야> 인공지능, 정보보호, 블록체인, IoT 등