Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.33, NO.6, Dec. 2023 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2023.33.6.1087

이미지 분류를 위한 오토인코더 기반 One-Pixel 적대적 공격 방어 기법

심 정 현,[1†] 송 현 민[2‡]
[1,2]단국대학교 (학생, 교수)

Autoencoder-Based Defense Technique against One-Pixel Adversarial Attacks in Image Classification

Jeong-hyun Sim,[1†] Hyun-min Song[2‡]
[1,2]Dankook University (Undergraduate student, Professor)

요 약

인공지능 기술의 급격한 발전으로 다양한 분야에서 적극적으로 활용되고 있으나, 이와 함께 인공지능 기반 시스템에 대한 공격 위협이 증가하고 있다. 특히, 딥러닝에서 사용되는 인공신경망은 입력 데이터를 고의로 변형시켜 모델의 오류를 유발하는 적대적 공격에 취약하다. 본 연구에서는 이미지에서 단 하나의 픽셀 정보만을 변형시킴으로써 시각적으로 인지하기 어려운 One-Pixel 공격으로부터 이미지 분류 모델을 보호하기 위한 방법을 제안한다. 제안된 방어 기법은 오토인코더 모델을 이용하여 분류 모델에 입력 이미지가 전달되기 전에 잠재적 공격 이미지에서 위협 요소를 제거한다. CIFAR-10 데이터셋을 이용한 실험에서 본 논문에서 제안하는 오토인코더 기반의 One-Pixel 공격 방어 기법을 적용한 사전 학습 이미지 분류 모델들은 기존 모델의 수정 없이도 One-Pixel 공격에 대한 강건성이 평균적으로 81.2% 향상되는 결과를 보였다.

ABSTRACT

The rapid advancement of artificial intelligence (AI) technology has led to its proactive utilization across various fields. However, this widespread adoption of AI-based systems has raised concerns about the increasing threat of attacks on these systems. In particular, deep neural networks, commonly used in deep learning, have been found vulnerable to adversarial attacks that intentionally manipulate input data to induce model errors. In this study, we propose a method to protect image classification models from visually imperceptible One-Pixel attacks, where only a single pixel is altered in an image. The proposed defense technique utilizes an autoencoder model to remove potential threat elements from input images before forwarding them to the classification model. Experimental results, using the CIFAR-10 dataset, demonstrate that the autoencoder-based defense approach significantly improves the robustness of pretrained image classification models against One-Pixel attacks, with an average defense rate enhancement of 81.2%, all without the need for modifications to the existing models.

Keywords: Adversarial Attack, One-Pixel Attack, Autoencoder, Deep Neural Networks, Differential Evolution

Received(10. 17. 2023), Modified(11. 28. 2023), Accepted(11. 28. 2023)

† 주저자, tlawjdgus0425@naver.com
‡ 교신저자, hyunminsong@dankook.ac.kr(Corresponding author)

I. 서 론

딥러닝 기반의 컴퓨터 비전(Computer vision) 기술은 자율 주행 자동차, 의료 등 다양한 분야에서 이미지 인식 문제를 해결하기 위한 기술로 널리 활용되고 있다[1, 2]. 그러나 딥러닝에서 사용되는 심층 신경망(DNNs, Deep Neural Networks)이 이미지에 작은 왜곡을 만들어 인공지능 모델이 이미지를 오분류하도록 만드는 적대적 공격에 취약하다는 문제점이 제기되고 있다[3, 4]. Madry 등[5]은 이미지 분류 모델 학습 시 적대적 학습(Adversarial training)을 사용하여 생성된 모델이 적대적 공격에 대해 저항성을 갖게 할 수 있다는 연구 결과를 발표하였다. 일부 연구자들은 이미지 분류 모델에 대한 적대적 공격이 단순한 우려가 아니라 감시 카메라, 도로 표지판, 얼굴 인식 시스템 등 실생활에서 당장 실현 가능한 위협이며, 이에 대한 대비책 마련이 시급함을 주장했다[6, 7, 8].

적대적 공격을 위해 주어진 이미지를 왜곡시키는 기법은 FGSM (Fast Gradient Sign Method), PGD (Projected Gradient Descent) 등 다양한 기법이 존재한다[9]. 차분 진화(DE, Differential Evolution) 알고리즘 기반의 One-Pixel 공격은 입력 이미지에서 단 하나의 픽셀값만을 변형시켜도 모델의 오동작을 일으킬 수 있다[10]. 이러한 One-Pixel 공격은 단순하지만, 이미지 분류 모델의 성능을 현저하게 감소시킬 수 있는 위험성을 지니고 있어, 본 연구에서는 다양한 적대적 공격 기법 중 One-Pixel 공격에 대해 집중하고자 한다.

하지만, 기존에 제안된 적대적 공격에 대한 방어 기법들은 예방이 아닌 탐지를 목적으로 하거나 모델의 학습 단계에서 알려진 적대적 공격 기법으로 생성된 샘플들을 제공하여 모델이 이를 학습하도록 하는 방식을 사용한다. 이 경우 기존에 생성된 모델을 폐기하고 새로 모델을 생성해야 하는 단점이 존재한다. 이에 본 연구에서는 일반적인 방식으로 학습되어 사용되고 있는 사전 학습된 이미지 분류 모델을 수정하지 않고, 오토인코더 기반의 여과 모델을 사용하여 새로운 이미지가 분류 모델에 전달되기 전에 정제 과정을 통해 이미지에 포함된 잠재적인 위협 요소를 제거하는 적대적 공격 방어 기법을 제안한다.

본 논문에서 제안하는 오토인코더 기반의 적대적 공격 방어 기법을 CIFAR-10 데이터셋에서 학습된 Lenet, DenseNet, WideResNet 등의 이미지 분류 모델에 적용한 결과 One-Pixel 공격에 대한 모델의 강건성이 평균적으로 81.2% 향상됨을 확인할 수 있었으며, 오토인코더 기반 이미지 여과 모델의 재구성 이미지에 PSD (Patch Selection Denoiser) 기법을 추가로 적용하였을 때 이미지 분류 모델의 성능 손실 없이 One-Pixel 공격을 100% 방어할 수 있었다.

II. 관련 연구

S.A.A. Shah 등[11]은 이미지를 구성하고 있는 수많은 픽셀 중 소수의 픽셀값만을 변조하는 픽셀 단위의 적대적 공격을 탐지하기 위해 ADNet (Adversarial Detection Network)이라는 적대적 예제 탐지 모델을 제안하였으며, CIFAR-10 이미지 데이터셋을 이용한 실험을 통해 적대적 예제를 효과적으로 탐지할 수 있음을 보였다.

Ian J. Goodfellow 등[12]은 심층 신경망 모델에서 나타나는 적대적 공격에 대한 취약성은 비선형성이 아닌 선형성에 있음을 설명하였고, 모델의 그래디언트(Gradient)를 이용하여 주어진 이미지로부터 대상 모델에 대한 적대적 예제를 효율적으로 생성할 수 있는 FGSM를 제안하였다. 하지만 FGSM 방식은 모델의 구조와 그래디언트를 계산해 내야 한다는 점에서 강한 공격자 모델이라는 한계점이 남아있다. Y. Dong 등[13]은 모델의 구조와 그래디언트를 얻지 못하는 블랙박스 공격에서 적대적 공격의 성공률을 향상시키기 위한 모멘텀 반복 알고리즘을 제안하였다. 모멘텀 반복 알고리즘을 이용하면 적대적 공격에 강건성을 지니고 있는 모델들에 대해서도 적대적 공격이 가능함을 보여주었다.

MA. Husnoo 등[23]은 사물 인터넷에서 사용되는 딥러닝 시스템에서 적대적 공격의 위험성을 제기하며, 픽셀 단위의 이미지 왜곡을 식별하기 위한 APG (Accelerated Proximal Gradient) 기반의 적대적 공격 방어 기법을 제안하였다. Su 등[10]은 적대적 공격 샘플을 만들기 위한 왜곡 픽셀의 수를 단 한 개로 제한한 One-Pixel 공격이 가능한 것을 보였다. Nguyen-son 등[24]은 One-Pixel 공격에서 왜곡된 한 개의 픽셀값이 주변 픽셀들에 비하여 RGB 값의 큰 차이를 가지기 때문에 시각적으로 눈에 띄는 문제를 개선하기 위해 주변 픽셀들과 차이가 적은 왜곡 값을 찾으며, 공격 성공률이 높은 주요 픽셀을 탐색하는 OPA2D-ATK 기

법을 제안하였으며, 공격 기법을 역이용하는 탐지 및 방어 기법을 함께 제안하였다. 또한, D. chen 등[25]은 공격받은 이미지로부터 원본 이미지를 재구성할 때 이미지 선명도가 떨어지는 문제를 해결하기 위해 Patch Selection Denoiser 기법을 적용하는 방어 기법을 제안하였다.

[표 1]에 여러 선행 연구에서 제안한 적대적 공격 기법들을 요약하였다. 각 공격 기법을 모델의 구조와 그래디언트 정보를 필요로 하는 화이트박스 공격과 모델에 대한 정보가 필요하지 않은 블랙박스 공격으로 구분하였으며, 적대적 공격이 단순히 모델의 오동작만을 목적으로 하는 대상 미지정 공격과 이미지를 공격자가 목표로 하는 특정 클래스로 잘못 분류하도록 만드는 지정 공격인지를 나타내었다. 마지막으로 적대적 공격이 이미지 데이터에서만 적용이 가능한 기법인지 다른 유형의 데이터에도 적용할 수 있는지를 나타내었다.

Table 1. Summary of types of adversarial attacks

Method	Black/ White box	Targeted /Non-targeted	Image-specific/Universal
One-pixel [10]	White box	Targeted	Image-specific
FGSM [12]	White box	Targeted	Image-specific
L-BFGS [14]	White box	Non-targeted	Image-specific
BIM & ILCM[15]	White box	Targeted	Image-specific
JSMA [16]	Black box	Non-targeted	Image-specific
C&W attacks [17]	White box	Targeted	Image-specific
DeepFool [18]	White box	Non-targeted	Image-specific
Universal perturbations[19]	White box	Non-targeted	Universal
UPSET [20]	Black box	Targeted	Universal
ANGRI [20]	Black box	Targeted	Image-specific
Houdini [21]	Black box	Targeted	Image-specific
ATNs[22]	White box	Targeted	Image-specific

III. 오토인코더 기반의 One-Pixel 공격 방어

적대적 공격이란 딥러닝의 인공신경망 모델의 입력 데이터에 고의적인 왜곡을 만들어 인공신경망 모델이 해당 입력 데이터를 본래 모델의 예측값과 다른 예측값을 출력하도록 만들어 인공지능 시스템의 오류를 일으키는 것을 목적으로 한다.

본 논문에서는 다양한 적대적 공격 기법 중 이미지에서 단 하나의 픽셀만을 왜곡시키는 One-Pixel 공격과 이에 대한 방어 기법을 다루며, 공격 샘플 생성을 위해서 차분 진화 알고리즘 기반의 One-Pixel 공격 기법을 사용하였다. 차분 진화 알고리즘은 일반적으로 벡터 최적화 문제를 해결하기 위한 최적화 알고리즘으로 간단하고 계산이 효율적이라는 특징이 있다. [그림 1]은 이미지 분류 모델인 LeNet에 대한 One-Pixel 공격 예제들을 보여준다. 각 사진의 하단에는 원본 이미지의 실제 정답 클래스와 해당 적대적 예제 이미지에 대한 LeNet 모델의 예측 클래스를 표기하였다. 예를 들어, 첫 번째 사진은 고양이 사진으로 원본 이미지에 대한 분류 모델의 예측값은 cat이지만 고양이의 얼굴 영역에 하나의 픽셀이 왜곡된 적대적 예제 이미지에 대해서는 분류 모델이 dog라는 잘못된 클래스 예측된 것을 알 수 있다.

[그림 2]는 본 논문에서 제안하는 오토인코더 기반의 적대적 공격 방어 기법을 나타낸다. 먼저, 원본

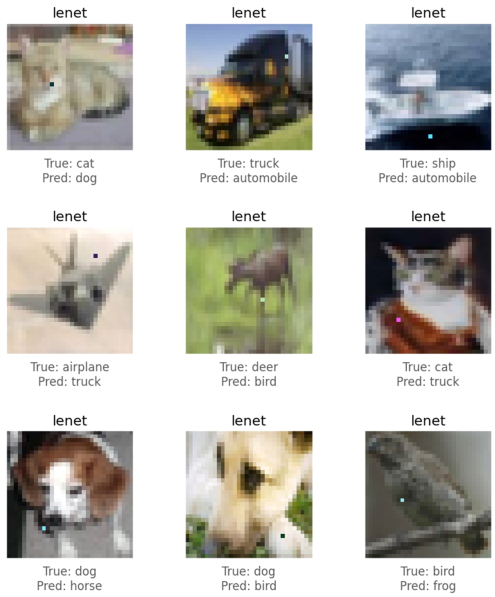

Fig. 1. One-Pixel Attack Examples on LeNet

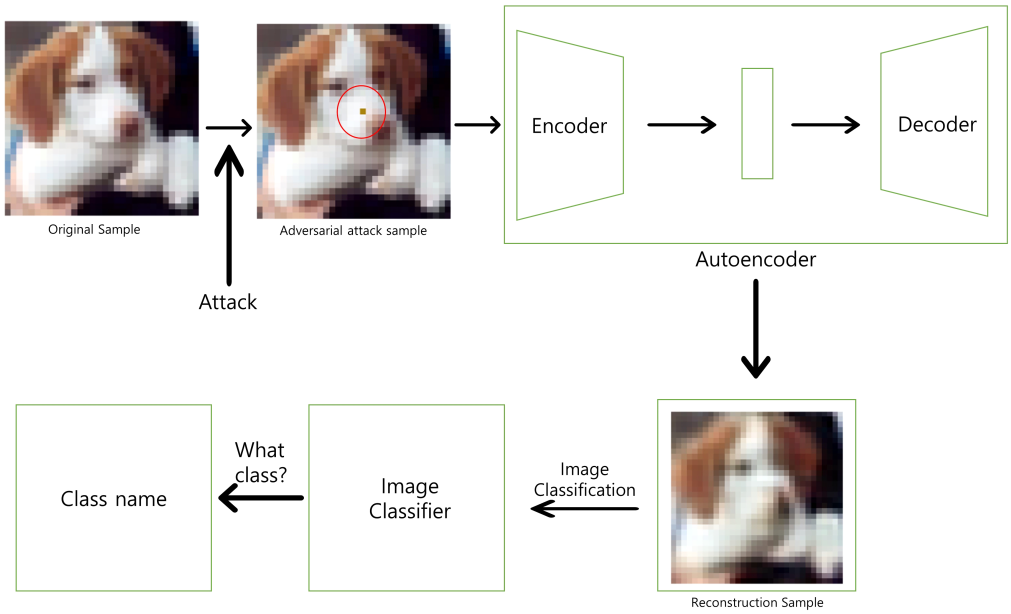

Fig. 2. Defense Architecture Against Adversarial Attacks Using Autoencoder-Based Approaches

샘플 이미지에 대해서 차분 진화 알고리즘 기반의 One-Pixel 공격을 수행하여 적대적 예제 샘플을 생성한다. 적대적 예제 샘플이 이미지 모델 분류 모델에 앞서 오토인코더 기반의 여과 모델에 전달되고, 여과 모델이 재구성한 이미지를 이미지 분류 모델에 전달한다. 이 과정에서 왜곡되었던 픽셀의 정보가 회복되어 분류 모델에 대한 적대적 공격을 무력화시킨다.

3.1 차분 진화 알고리즘 기반의 One-Pixel 공격

적대적 공격 이미지를 생성하기 위해서 주어진 벡터에서 값을 변경할 차원과 각 차원에 대한 수정의 강도를 어떻게 할 것인지를 정해야 한다. 이미지 데이터의 경우 이미지가 벡터에 해당하며, 차원은 이미지 내 각 픽셀을 의미한다.

인공지능 기반의 이미지 분류 문제는 다음과 같이 정의될 수 있다.

$$f_t(x) = P(T=t|X=x) \quad (1)$$

여기서 f는 n차원의 입력을 받는 대상 이미지 분류기이며, x는 모델에게 주어지는 입력 이미지로서 클래스 t로 올바르게 분류된다. 따라서 x가 클래스 t에 속할 확률은 $f_t(x)$이다.

원본 이미지 x에 대한 적대적 공격은 아래와 같이 원본 이미지 x에 적대적 왜곡이 일어났을 때 잘못된 클래스로 분류할 가능성을 최대화하는 문제로 나타낼 수 있다.

$$\max_{e^*(x)} f_{adv}(x+e(x)) \quad s.t. \| e(x) \| \leq L \quad (2)$$

이때, $e(x)=(e_1, \ldots, e_n)$는 원본 이미지 x에 대한 n개 픽셀의 왜곡이며, $\| e(x) \|$은 벡터 $e(x)$의 길이로 원본 이미지에서 왜곡된 픽셀 수를 나타낸다.

본 연구에서 적대적 공격으로 사용한 One-Pixel 공격은 왜곡시킬 픽셀의 수를 1로 제한하므로 아래와 같이 $L=1$ 인 경우이다.

$$\max_{e(x)^*} f_{adv}(x+e(x)) \quad s.t. \| e(x) \| = 1 \quad (3)$$

픽셀의 왜곡은 주어진 이미지 내 n개의 픽셀 중 임의의 픽셀을 선택하여 임의의 값으로 수정함으로써 만들어진다.

차분 진화 알고리즘은 최적해를 찾기 위한 모집단 기반 최적화 알고리즘으로 일반적인 진화 알고리즘에 속한다. 차분 진화는 다양성을 유지하는 모집단(population) 선택 과정을 통해 경사 하강법 기반

알고리즘이나 다른 종류의 진화 알고리즘보다 더 높은 품질의 최적해를 효과적으로 찾을 수 있으며, 최적해를 찾기 위해 모델의 정보가 필요하지 않은 블랙박스 모델에 해당한다.

차분 진화 알고리즘은 먼저 현재의 모집단에 따라 후보 솔루션(candidate solution) 집합을 생성한다. 그 후, 후보 솔루션들을 해당 모집단과 비교하여 모집단보다 더 적합한 경우에만 후보 솔루션을 채택하고, 모집단이 더 적합한 솔루션인 경우 후보 솔루션을 폐기한다. 즉, 모집단으로부터 생성된 후보 솔루션 집합 P는 모집단으로부터 무작위 변형을 통해 생성된 n개의 후보 솔루션으로 구성된다.

$$P = X_i\ ,\ (1 \le i \le n) \tag{4}$$

X_i는 각 후보 솔루션을 의미하며, 다음과 같이 정의된다.

$$X_i = (x_i,\ y_i,\ r_i,\ g_i,\ b_i),\ \ (1 \le i \le d) \tag{5}$$

x_i, y_i는 각각 이미지 내 픽셀의 가로, 세로 좌표를 나타내며, r_i, g_i, b_i는 해당 픽셀을 구성하고 있는 빨강, 초록, 파랑의 색상 값을 나타낸다.

후보 솔루션은 아래와 같이 돌연변이 함수 F를 통해 생성된다.

$$X_i = X_{r1} + F(X_{r2} - X_{r3}),\ \ (r1 \ne r2 \ne r3) \tag{6}$$

본 연구에서는 돌연변이 확률을 0.5로 설정하여 사용하였다. 이는 50%의 확률로 기존의 후보 솔루션을 그대로 사용하고, 나머지 50% 확률로 다른 임의의 두 후보 솔루션의 차이를 더해줌으로써 새로운 후보 솔루션을 생성한다. $r1$, $r2$, $r3$는 후보 솔루션 집합 P에서 임의의 후보 솔루션 인덱스를 나타낸다.

위 과정을 통해 새로운 후보 솔루션 집단이 생성되면 각 후보 솔루션들과 모집단의 적합도를 평가한다. 본 연구의 적대적 공격에서는 이미지 분류 모델이 주어진 이미지에 대해 잘못된 클래스에 대한 확률 예측이 높을수록 적합한 해라고 판단할 수 있다.

차분 진화 알고리즘은 위 과정을 반복함으로써 모집단과 후보 솔루션을 비교하여 해의 다양성을 유지하면서 적합도 값을 향상시키는 목표를 동시에 달성할 수 있다.

3.2 오토인코더

오토인코더는 비지도 학습 알고리즘의 한 종류로, 학습 데이터셋의 데이터 분포를 모델에 학습시키기 위한 목적으로 사용된다. 오토인코더 모델은 고차원의 입력 데이터를 저차원의 벡터로 압축하는 인코더(Encoder) 모듈과 이를 본래의 고차원 데이터로 복원하는 디코더(Decoder) 모듈로 구성되어 있다. 이미지 데이터의 경우, 입력된 이미지를 저차원의 벡터로 인코딩한 후 다시 본래의 이미지로 복원해내는 기능을 수행한다. 즉, 올바른 이미지 데이터셋으로 학습된 오토인코더 모델은 입력된 이미지에 일부 왜곡이 발생하였을 때 학습된 데이터 분포를 기반으로 원본 이미지를 근사하게 추정할 수 있다.

오토인코더의 이러한 특성을 이용하면 적대적 공격을 당한 이미지가 입력되었을 때 원본 이미지로 복원함으로써 이미지 분류 모델이 올바르게 동작할 수 있도록 할 수 있다. 본 논문에서 사용한 오토인코더의 구조는 〔그림 3〕과 같으며, 오토인코더 모델은

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 32, 32, 3)]	0	[]
conv2d (Conv2D)	(None, 32, 32, 32)	896	['input_1[0][0]']
batch_normalization (BatchNormalization)	(None, 32, 32, 32)	128	['conv2d[0][0]']
max_pooling2d (MaxPooling2D)	(None, 16, 16, 32)	0	['batch_normalization[0][0]']
dropout (Dropout)	(None, 16, 16, 32)	0	['max_pooling2d[0][0]']
conv2d_1 (Conv2D)	(None, 16, 16, 32)	9248	['dropout[0][0]']
leaky_re_lu (LeakyReLU)	(None, 16, 16, 32)	0	['conv2d_1[0][0]']
batch_normalization_1 (BatchNormalization)	(None, 16, 16, 32)	128	['leaky_re_lu[0][0]']
max_pooling2d_1 (MaxPooling2D)	(None, 8, 8, 32)	0	['batch_normalization_1[0][0]']
dropout_1 (Dropout)	(None, 8, 8, 32)	0	['max_pooling2d_1[0][0]']
conv2d_2 (Conv2D)	(None, 8, 8, 64)	18496	['dropout_1[0][0]']
batch_normalization_2 (BatchNormalization)	(None, 8, 8, 64)	256	['conv2d_2[0][0]']
max_pooling2d_2 (MaxPooling2D)	(None, 4, 4, 64)	0	['batch_normalization_2[0][0]']
conv2d_transpose (Conv2DTranspose)	(None, 8, 8, 64)	36928	['max_pooling2d_2[0][0]']
batch_normalization_3 (BatchNormalization)	(None, 8, 8, 64)	256	['conv2d_transpose[0][0]']
dropout_2 (Dropout)	(None, 8, 8, 64)	0	['batch_normalization_3[0][0]']
conv2d_transpose_1 (Conv2DTranspose)	(None, 16, 16, 32)	18464	['dropout_2[0][0]']
batch_normalization_4 (BatchNormalization)	(None, 16, 16, 32)	128	['conv2d_transpose_1[0][0]']
dropout_3 (Dropout)	(None, 16, 16, 32)	0	['batch_normalization_4[0][0]']
conv2d_transpose_2 (Conv2DTranspose)	(None, 16, 16, 32)	9248	['dropout_3[0][0]']
add (Add)	(None, 16, 16, 32)	0	['conv2d_transpose_2[0][0]', 'conv2d_1[0][0]']
leaky_re_lu_1 (LeakyReLU)	(None, 16, 16, 32)	0	['add[0][0]']
batch_normalization_5 (BatchNormalization)	(None, 16, 16, 32)	128	['leaky_re_lu_1[0][0]']
conv2d_transpose_3 (Conv2DTranspose)	(None, 32, 32, 3)	867	['batch_normalization_5[0][0]']

Fig. 3. Autoencoder Architecture

CIFAR-10 학습 데이터셋에 포함된 50,000개의 정상 이미지를 사용하여 학습되었다.

3.3 이미지 분류 모델

이미지 분류 모델은 적대적 공격에 대해 학습되지 않은 사전에 생성된 이미지 분류 모델을 사용하며, 새로운 이미지가 주어졌을 때 바로 이미지 분류 모델에 입력하지 않고, 여과 모델을 통해 재구성한 이미지를 입력하여 이미지에 대한 클래스를 예측한다.

본 연구에서는 CIFAR-10 학습 데이터셋만으로 사전 학습된 이미지 분류 모델들을 수정하지 않고 사용하여 One-Pixel 공격에 대한 강건성을 측정하였다.

Table 2. Accuracy and Performance Degradation Rate of an Image Classification Model

Model Name	Accuracy	Accuracy after Reconstruction	Performance Loss
PureCnn [10]	0.8877	0.8514	4.1%
Lenet[26]	0.7487	0.7360	1.7%
NiN[27]	0.9074	0.8728	3.81%
ResNet [28]	0.9231	0.8685	5.9%
DenseNet [29]	0.9467	0.8856	6.5%
WideResNet [30]	0.9534	0.8983	5.8%

IV. 실 험

4.1 이미지 데이터셋

본 연구에서는 이미지 분류 모델 학습 및 적대적 공격 실험을 위해 [표 3]의 CIFAR-10 데이터셋을 사용하였다. CIFAR-10 이미지 데이터셋에는 해상도는 (32, 32, 3) 이미지 샘플들과 각 이미지의 종류를 나타내는 10개의 클래스 정보가 기록되어 있으며, 이미지 분류 모델 학습용 샘플 50,000개, 테스트용 샘플 10,000개로 구성되어 있다.

Table 3. CIFAR-10 Information

Image	Resolution	Class	Train set	Test set
CIFAR -10	(32, 32, 3)	10	50,000	10,000

4.2 작업 환경

본 연구의 실험은 Google Colab 환경에서 수행되었으며, Colab에서 제공된 환경은 CPU는 Intel(R) Xeon(R) CPU@2.00GHz이며, GPU는 Nvidia Tesla T4, RAM 12.68GB로 구성되었다.

4.3 사전 학습 이미지 분류 모델

본 논문에서 이미지 분류 모델은 기본적인 CNN (Convolutional Neural Network) 구조를 가장 처음 제안한 모델인 Lenet-5[26], PureCnn[10], NiN (Network-in-Network)[27], ResNet (Residual Network)[28], DenseNet (Densely Connected Convolutional Networks)[29], WideResNet (Wide Residual Networks)[30]을 사용하였다.

[표 2]는 CIFAR-10 테스트셋 10,000개에 대한 각 이미지 분류 모델의 분류 정확도, 오토인코더 기반 여과 모델 적용 후의 분류 정확도, 분류 모델로 인한 분류 정확도 성능 손실을 나타낸다. 여과 모델 적용 후의 성능 변화와 관련하여서는 아래의 실험 결과에서 설명하도록 한다.

4.4 적대적 공격 이미지 생성

적대적 공격 이미지를 생성하기 위해 CIFAR-10 테스트셋에서 1,000개의 이미지를 임의로 선택하였다. 이때, 공격으로 인한 오분류 여부를 식별하기 위하여 이미지 분류 모델이 올바르게 분류하는 샘플들만을 공격 대상 샘플로 사용하였다. 선택된 1,000개의 이미지에 대하여 차분 진화 알고리즘 기반의 One-Pixel 공격을 수행하여 적대적 공격 이미지 1,000개를 생성하였다. 이미지 분류 모델에 따라 테스트셋 샘플들 중 오분류 샘플이 다르므로 공격 샘플도 모델마다 다르게 구성하였다.

4.5 적대적 공격에 대한 강건성 평가

적대적 공격 이미지에 대한 각 이미지 분류 모델

의 강건성을 평가하기 위해서 생성된 1,000개의 적대적 공격 이미지에 대한 이미지 분류 모델의 정확도를 측정하였다. 이미지 분류 모델이 주어진 1,000개의 적대적 공격 이미지 중 오분류한 샘플의 비율을 사용하여 적대적 공격 성공률을 측정한다. 사용된 공격 이미지의 원본 이미지는 모두 각 모델이 올바르게 분류하는 샘플들이므로 공격 이미지에 대한 오분류는 모두 적대적 공격으로 인한 오류이다.

$$Rate_{\text{success}} = \frac{\text{A}_{\text{success}}}{\text{A}_{\text{total}}} \tag{7}$$

$Rate_{\text{success}}$는 적대적 공격의 성공률이며, 이는 총 공격 이미지 개수 A_{total}에 대한 공격에 성공하여 오분류된 이미지의 개수 A_{success}의 비율이다.

이후, 적대적 공격 이미지를 오토인코더 기반의 여과 모델을 이용하여 원본 이미지와 근사한 이미지로 복원한 후 이미지 분류 모델의 분류 성능을 재측정하였다. 이를 통해, One-Pixel 공격에 대한 오토인코더 기반 여과 모델의 방어율 D_{suc}을 계산한다.

$$D_{suc} = 1 - \frac{\text{ReconA}_{\text{success}}}{A_{\text{success}}} \tag{8}$$

$ReconA_{\text{success}}$는 앞서 공격에 성공하여 이미지 분류 모델이 오분류 하였던 이미지 중 오토인코더 모델을 통해 재구성된 복원 이미지에 대해서도 분류 모델이 여전히 오분류한 이미지의 개수를 나타낸다. 즉, 여과 모델을 사용하였음에도 적대적 공격이 성공한 이미지의 개수이며, 이를 이용해 여과 모델을 통해 무효화된 적대적 공격의 비율을 산출하였다.

여과 모델을 사용함으로써 왜곡된 이미지를 복원하여 이미지 분류 모델이 오분류하였던 이미지들을 다시 정확하게 분류할 수 있게 되었으나 일부 이미지들의 경우 적대적 공격이 실패하였음에도 여과 모델로 인해 추가적인 분류 오류가 발생하여 분류 성능의 손실이 발생할 수 있다. 여과 모델로 인한 분류 성능의 손실 PL (Performance Loss)은 다음과 같이 측정하였다.

$$PL = \frac{\text{ReconA}_{\text{fail}}\text{error}}{A_{\text{total}} - \text{ReconA}_{\text{success}}} \tag{9}$$

$ReconA_{fail}error$은 적대적 공격에 실패한 공격 이미지 중 재구성 후 이미지 분류 모델이 오분류한 이미지 개수를 나타낸다.

4.6 실험 결과

〔표 4〕는 CIFAR-10 학습 데이터셋으로 사전 학습된 각 이미지 분류 모델에 대하여 CIFAR-10 테

Table 4. Experimental Results Based on Image Classification Models

Attack Method	One-Pixel Attack〔10〕, A_total = 1,000			
Result	Attack Success Rate (Rate_success)	Attack Success Rate after Reconstruction (ReconA_success)	Defense Rate (D_suc)	Performance Loss (PL)
PureCnn〔10〕	162/1,000 (16.2%)	23/1,000 (2.3%)	139/162 (85.8%)	131/977 (13.4%)
Lenet〔26〕	612/1,000 (61.2%)	217/1,000 (21.7%)	395/612 (64.5%)	151/783 (19.2%)
NiN〔27〕	286/1,000 (28.6%)	51/1,000 (5.1%)	235/286 (82.2%)	126/949 (13.2%)
ResNet〔28〕	300/1,000 (30%)	48/1,000 (4.8%)	252/300 (84%)	121/952 (12.7%)
DenseNet〔29〕	252/1,000 (25.2%)	45/1,000 (4.5%)	207/252 (82.1%)	153/955 (16%)
WideResNet〔30〕	221/1,000 (22.1%)	24/1,000 (2.4%)	197/221 (89.1%)	105/976 (10.7%)

스트셋에서 임의로 선택된 1,000개의 이미지를 이용한 One-Pixel 공격 실험 결과이다.

Lenet은 여과 모델을 사용하기 전 612개의 공격 이미지를 오분류하였으나 여과 모델을 통해 복원된 이미지에 대해서는 217개를 오분류하였다. 공격에 성공했던 이미지 612개 중 395개 이미지는 여과 모델을 사용하였을 때 본래의 클래스로 올바르게 분류하여 64.5%의 방어율을 보였다. 하지만 공격에 실패한 이미지 783개 중 여과 모델로 인해 추가적인 오분류를 일으킨 이미지가 151개 발생하여 성능 손실이 19.2% 발생하였다. 사용된 이미지 분류 모델 중 가장 분류 성능이 뛰어났던 WideResNet의 경우 공격에 성공한 이미지가 221개에서 24개로 감소하여 89.1%의 적대적 공격을 방어하였으며, 여과 모델로 인한 성능 손실도 10.7%로 가장 적게 나타났다. 실험 결과를 통해 사전 학습 이미지 분류 모델의 본래 성능이 우수할수록 여과 모델을 적용하였을 때 적대적 공격에 대한 강건성이 크게 향상되며, 여과 모델로 인한 성능 손실도 적게 나타나는 경향이 있었다. PureCnn 모델의 경우에는 본래 모델의 분류 성능은 88.7% 수준으로 Lenet 모델에 이어 두 번째로 좋지 않았지만, 여과 모델 적용 후 강건성 향상 폭은 85.8%로 WideResNet 다음으로 두 번째로 높은 수치를 보였다. 다만, 성능 손실은 13.4%로 나타나 Lenet 다음으로 큰 성능 손실을 보였다.

〔표 2〕에서 이미지 분류 성능이 좋은 모델인 DenseNet, WideResNet이 다른 모델에 비해 One-Pixel 공격에 대해 더 강건한 경향이 있으며, 오토인코더 기반의 여과 모델을 사용했을 때 공격 방어의 효과가 좋게 나타났다. 결과적으로 본 논문에서 제안하는 오토인코더 기반의 적대적 공격 방어 기법은 6가지 사전 학습 이미지 분류 모델에서 평균적으로 81.2% 강건성이 향상되었고, 성능 손실은 14.2%가 발생하였다.

4.7 방어 기법 비교

본 논문의 방어 기법과 OPA2D-DEF, PSD 방어 기법의 비교를 하였다〔24, 25〕.

OPA2D-DEF 방어 기법은 적대적 공격으로 오분류된 이미지에 대하여 다시 한번 적대적 공격을 수행하는 경우 본래 클래스로 예측될 확률이 높다는 성질을 이용하여 공격 이미지에 대해 One-Pixel 공격을 재수행하여 이미지 분류 모델이 원래의 클래스로 예측하도록 만든다.

PSD 방어 기법은 3×3 크기의 패치로 입력 이미지와 재구성한 이미지의 영역을 나누고, 각 영역 픽셀의 RGB 값의 평균 차이를 비교한다. RGB 값의 차이가 임계값보다 큰 패치 영역은 왜곡된 픽셀이 포함되었다고 판단하여 재구성한 이미지의 RGB 값을 사용하고, 임계값보다 작은 영역은 왜곡된 픽셀이 없는 것으로 간주하여 입력 이미지의 RGB 값을 사용한다. 이는 One-Pixel 공격이 하나의 픽셀만을 왜곡시킨다는 특징을 이용한 것으로 공격받은 픽셀이 포함된 패치 영역만을 재구성 값으로 대체하고, 공격받지 않은 영역은 입력 이미지의 픽셀값을 그대로 사용함으로써 이미지 분류 모델의 오분류 가능성을 낮추는 효과가 있다.

이미지 분류 모델은 〔표 4〕에서 One-Pixel 공격에 대한 강건성이 가장 높았던 WideResNet을 사용하였으며, One-Pixel 공격 샘플 1,000개를 동일하게 사용하여 비교를 수행하였다.

PSD 방어 기법의 경우 이미지 분류 모델의 학습 단계에서 SRResNet 모델을 이용하여 노이즈가 포함된 샘플들을 생성하여 이미지 분류 모델의 학습 단계에서 함께 활용함으로써 이미지 분류 모델이 적대적 공격에 대한 강건성을 갖도록 하는 방식이지만, 본 연구에서 목표로 하는 기존 학습된 모델의 수정 없이 적대적 공격에 대한 방어 관점에서 직접적인 비교를 위하여 이미지 분류 모델을 재학습하지 않고 본 논문의 연구에서 사용한 오토인코더 모델과 동일한 조건으로 SRResNet 모델만을 CIFAR-10 학습 데이터셋을 이용하여 50 에포크 학습한 후 PSD 기법을 적용하였다.

본 연구에서 제안하는 오토인코더 기반 여과 모델 기법, OPA2D-DEF, SRResNet/PSD 기법을 이용한 WideResNet 이미지 분류 모델의 One-Pixel 공격에 대한 방어 효과를 〔표 5〕에 나타내었다. 방어 기법이 적용되지 않고, CIFAR-10 학습 데이터로만 학습된 WideResNet 이미지 분류 모델은 1,000개의 One-Pixel 공격 샘플 중 221개를 오분류하였으나 OPA2D-DEF 기법을 적용하였을 때 94개 샘플만을 오분류하여 221개 공격 성공 샘플 중 127개 샘플에 대해 성공적으로 방어할 수 있음을 보였으며, 이때 공격 실패 샘플 중 오분류를 일으킨 샘플이 906개 중 132개로 약 14.5%의 분류 성능 손실을 보여주었다. 이는 본 연구에서 제안하는 오토인코더 기반의 여과 모델에 비하여 약 64% 수

Table 5. Comparison of Defense Techniques

Method	ReconA_ success	D_suc	PL
Our Method	24/1,000 (2.4%)	197/221 (89.1%)	105/976 (10.7%)
OPA2D-DEF	94/1,000 (9.4%)	127/221 (57.4%)	132/906 (14.5%)
SRResNet with PSD	828/1,000 (82.8%)	0/221 (0%)	607/779 (77.9%)
Our Method with PSD	0/1,000 (0%)	221/221 (100%)	0/1000 (0%)

준의 방어율과 약 35.5% 많은 성능 손실로써 본 연구에서 제안하는 오토인코더 기반 여과 모델 기법이 One-Pixel 공격 방어에 더 효과적임을 나타낸다.

SRResNet 기반의 PSD 방어 기법의 경우 본 실험에서 유의미한 방어 성능을 보여주지 못하였으며, 오히려 SRResNet으로 재구성된 이미지의 품질이 크게 낮아 분류기의 성능이 크게 감소하는 역효과가 발생하였다. 이는 이미지 분류 모델 학습 단계에서 SRResNet을 이용하여 생성된 노이즈가 포함된 데이터를 함께 사용하는 경우 분류 모델이 이를 분류할 수 있으나 본 실험에서는 분류 모델을 재학습하지 않고 기존의 분류 모델을 사용하여 SRResNet이 생성하는 노이즈 데이터에 대한 성능이 낮은 것으로 추정된다.

다만, 해당 논문에서 제안한 PSD 기법은 오토인코더 모델이 재구성하는 이미지의 품질을 보정하는 역할을 하므로 본 연구의 오토인코더 모델에 PSD 기법을 적용하여 성능 변화를 측정하였다. PSD 기법 적용 전후 재구성된 이미지 품질 차이는 〔그림 4〕에 보이는 바와 같이 왜곡된 픽셀 영역을 제외한 나머지 영역에서 이미지 선명도가 개선되었다. 이를 통해 이미지 분류 모델의 성능 손실이 감소할 것이라 기대하였으나 성능 손실이 0%를 달성하였으며, 더 나아가 방어하지 못하였던 24개의 공격 샘플에 대해서도 모두 방어에 성공하는 결과를 보여주었다. 이는

Fig. 4. Our method with PSD Comparative Image

오토인코더 기반 여과 모델에 PSD 기법을 함께 사용하였을 때 이미지 분류 모델의 성능 손실을 방지해 줄 뿐만 아니라 적대적 공격에 대한 강건성을 더욱 향상하는 데 이바지한다는 것을 보여준다.

V. 결 론

본 논문에서는 오토인코더 기반의 여과 모델을 활용하여, 기존 모델을 수정하지 않고도 사전 학습된 이미지 분류 모델을 One-Pixel 공격으로부터 효과적으로 보호하는 방어 기법을 제안하였다. CIFAR-10 데이터셋을 이용한 실험을 통해 본 연구의 방어 기법이 적대적 공격에 대한 강건성을 평균 81.2% 향상시키며, 여과 모델로 인한 분류 모델들의 성능 손실을 평균 14.2%로 제한하는 것을 확인하였다. 이러한 결과는 제안된 방어 기법이 One-Pixel 공격에 대응하는 데 효과적임을 시사하며, 이는 기존 연구와의 비교에서도 더 높은 방어율과 더 낮은 성능 손실을 달성함으로써 본 연구 방법의 우수성을 나타낸다. 특히, PSD 기법을 본 연구의 방어 기법에 함께 적용했을 때, 분류 모델의 성능 손실 없이 모든 One-Pixel 공격 샘플을 방어하는 결과를 보여주었다.

그러나 본 연구는 일부 한계점을 가지고 있다. 실험에서 사용된 CIFAR-10 데이터셋은 저해상도 이미지로 구성되어 있어, 실제 애플리케이션에서 요구하는 고해상도 이미지에 대한 적용 가능성에 대한 연구가 추가적으로 필요하다. 또한, 다양한 적대적 공격 유형 중에서 One-Pixel 공격만을 대상으로 방어 효과성을 확인하였다. 이러한 제한사항은 향후 연구에서 고해상도 이미지 데이터셋과 다양한 적대적 공격 기법에 대한 연구, 그리고 적대적 공격 탐지에 적합한 개선된 오토인코더 모델 구조에 대한 추가적인 연구를 필요로 한다.

결론적으로, 본 연구는 One-Pixel 공격에 대한 효과적인 방어 기법을 제시함으로써, 적대적 공격에 대한 이미지 분류 모델의 강건성을 높이는 데 중요한 기여를 하였으며 이는 향후 적대적 공격에 대한 방어 메커니즘 개발에 중요한 발판이 될 것이다.

References

〔1〕 A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with

deep convolutional neural networks," Communications of the ACM, vol. 25, pp. 1097-1105, 2012.
[2] K. He, X. Zhang, and S. Ren, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," Proceedings of the IEEE international conference on computer vision 2015, pp. 1026-1034, 2015.
[3] Huang, Sandy, et al. "Adversarial attacks on neural network polices," arXiv preprint arXiv:1702.02284, Feb. 2017.
[4] K. Ren, T. Zheng, and Z. Qin, "Adversarial attacks and defenses in deep learning," Engineering, vol. 6, no. 3, pp. 346-360, Mar. 2020.
[5] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, Sep. 2019.
[6] Qiu, Shilin, et al. "Review of artificial intelligence adversarial attack and defense technologies," Applied Sciences, vol. 9, no. 5, Jun. 2017.
[7] Akhtar, Naveed, and Ajaml Mian, "Threat of adversarial attacks on deep learning in computer vision: a survey," IEEE Access, vol. 6, pp. 14410-14430, Feb. 2018.
[8] J. Ito, JL. Zittrain, and AL. Beam, "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287-1289, Mar. 2019.
[9] H. Liang, E. He, and Y. zhao, "Adversarial attacks and defense: a survey," Electronics 2022, vol. 11, no. 8, Apr. 2022.
[10] J. Su, D. V. Vargas and K. Sakurai, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 828-841, Oct. 2019.
[11] Shah, Syed Afaq Ali, et al. "Efficient detection of pixel-level adversarial attacks," 2020 IEEE International Conference on Image Processing, pp. 718-722, Oct. 2020.
[12] Goodfellow, Ian J., and Jonathon Shlens, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, Dec. 2014.
[13] Y. Dong, F. Liao, and T. Pang, "Boosting adversarial attacks with momentum," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9185-9193, Mar. 2018.
[14] C. Szegedy et al. "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, Dec. 2013.
[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," Artificial intelligence safety and security, pp. 99-112, 2018.
[16] N. Papernot, P. McDaniel, and S. Jha, "The limitations of deep learning in adversarial settings," 2016 IEEE European symposium on security and privacy, pp. 372-387, Mar. 2016.
[17] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017 IEEE symposium on security and privacy, pp. 39-57, May. 2017.
[18] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574-2582, Jul. 2016.
[19] S. M. Moosavi-Dezfooli, A. Fawzi, and O. Fawzi, "Universal adversarial perturbations," Proceedings of the IEEE conference on computer vision

and pattern recognition, pp. 1765-1773. Oct. 2017.
〔20〕 S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," IEEE transactions on evolutionary computation, vol. 15, no. 1, pp. 4-31, Feb. 2011.
〔21〕 F. Tramer, N. Papernot, and I. Goodfellow, "The space of transferable adversarial examples," arXiv preprint arXiv:1704.03453, Apr. 2017.
〔22〕 S. Baluja and I. Fischer, "Adversarial transformation networks: learning to generate adversarial examples," arXiv preprint arXiv:1703.09387, Mar. 2017.
〔23〕 MA. Husnoo and A. Anwar, "Do not get fooled: defense against the one-pixel attack to protect IoT-enabled deep learning systems," Ad Hoc Networks 122, vol. 122, 2021.
〔24〕 Nguyen-Son, Hoang-Quoc, and T. P. Thao, "OPA2D: One-pixel attack, detection, and defense in deep nueral networks," 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1-10, Sep. 2021.
〔25〕 Chen, D., Xu, R., and Han, B., "Patch selection denoiser: an effective approach defending against one-pixel attacks," 26th International Conference on Neural Information Processing (ICONIP 2019), pp. 286-296, Dec. 2019.
〔26〕 Y. Lecun, L. Bottou, and Y. Bengio, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 89, no. 11, pp. 2278-2324, Nov. 1998.
〔27〕 Lin, Min, and Qiang Chen, "Network in network," arXiv preprint arXiv:1312.4400, Dec. 2013.
〔28〕 K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
〔29〕 Gao Huang, Zhuang Liu, and Laurens van der Maaten, "Densely connected convolutional networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017.
〔30〕 Zagoruyko, Sergey, and Nikos Komodakis, "Wide residual networks," arXiv preprint arXiv:1605.07146, 2016.

〈저자소개〉

심 정 현 (Jeong-hyun Sim) 학생회원
2020년 3월~현재: 단국대학교 산업보안학과 학사과정
〈관심분야〉 인공지능보안, 정보보호

송 현 민 (Hyun-min Song) 종신회원
2014년 2월: 한양대학교 전자통신컴퓨터공학부 졸업
2021년 2월: 고려대학교 정보보호대학원 정보보호학과 박사
2021년 1월~2021년 6월: 하나금융융합기술원 책임연구원
2021년 7월~2022년 8월: 국가안보기술연구소 연구원
2022년 9월~현재: 단국대학교 산업보안학과 조교수
〈관심분야〉 정보보호, 인공지능보안, 네트워크 및 시스템보안