한국정보전자통신기술학회논문지(jkiiect)'22-2, Vol.15 No.1
논문 22-15-01-27 http://dx.doi.org/10.17661/jkiiect.2022.15.1.27

영상 인식을 위한 딥러닝 모델의 적대적 공격에 대한 백색 잡음 효과에 관한 연구

이영석*, 김종원**

Study on the White Noise effect Against Adversarial Attack for Deep Learning Model for Image Recognition

Youngseok Lee*, Jongweon Kim**

요 약 본 논문에서는 영상 데이터에 대한 적대적 공격으로부터 생성된 적대적 예제로 인하여 발생할 수 있는 딥러닝 시스템의 오분류를 방어하기 위한 방법으로 분류기의 입력 영상에 백색 잡음을 가산하는 방법을 제안하였다. 제안된 방법은 적대적이든 적대적이지 않던 구분하지 않고 분류기의 입력 영상에 백색 잡음을 더하여 적대적 예제가 분류기에서 올바른 출력을 발생할 수 있도록 유도하는 것이다. 제안한 방법은 FGSM 공격, BIM 공격 및 CW 공격으로 생성된 적대적 예제에 대하여 서로 다른 레이어 수를 갖는 Resnet 모델에 적용하고 결과를 고찰하였다. 백색 잡음의 가산된 데이터의 경우 모든 Resnet 모델에서 인식률이 향상되었음을 관찰할 수 있다. 제안된 방법은 단순히 백색 잡음을 경험적인 방법으로 가산하고 결과를 관찰하였으나 에 대한 엄밀한 분석이 추가되는 경우 기존의 적대적 훈련 방법과 같이 비용과 시간이 많이 소요되는 적대적 공격에 대한 방어 기술을 제공할 수 있을 것으로 사료된다.

Abstract In this paper we propose white noise adding method to prevent missclassification of deep learning system by adversarial attacks. The proposed method is that adding white noise to input image that is benign or adversarial example. The experimental results are showing that the proposed method is robustness to 3 adversarial attacks such as FGSM attack, BIN attack and CW attack. The recognition accuracies of Resnet model with 18, 34, 50 and 101 layers are enhanced when white noise is added to test data set while it does not affect to classification of benign test dataset. The proposed model is applicable to defense to adversarial attacks and replace to time- consuming and high expensive defense method against adversarial attacks such as adversarial training method and deep learning replacing method.

Key Words : Deep learning, Adversarial attack, FGSM attack, BIM attack, CW attack, White noise, Perturbation

1. 서 론

CNN(convolutional neural network)의 출현으로 대표되는 신경망 기술의 진보는 더 복잡하고 더 방대한 양의 데이터가 필요한 추론 분야로 적용 범위를 넓혀가고 있다[1]. 또한, 딥러닝 기술은 자체의 추론 기술을 발전시켰을 뿐만 아니라 기존의 기술에 딥러닝 기술을 응용하여 다양한 분야에서 발전을 이루고 있다. 특히 자동차 산업 분야에서는 전기 자동차의 출현을 앞두고 딥러닝 기술에 의한 자율주행에 관한 관심이 높아지고 있으며 상당한 기술 수준에 도달한 것으로 알려져 있다[2].

딥러닝 기술의 기본 모델은 주어진 데이터를 이

This work was conducted with the support of the National Research Foundation of Korea (NRF-2020RIA2C1101938)
*Department of Electronic Engineering, Chungwoon University
**Corresponding Author : Department of Electronic Engineering, Sangmyung University (jwkim@smu.ac.kr)
Received February 04, 2022 Revised February 08, 2022 Accepted February 18, 2022

용하여 학습한 신경망 파라미터를 기반으로 현재의 입력에 대한 출력을 추론하는 형식이다. 학습된 딥러닝 시스템은 테스트 데이터를 이용하여 해당 시스템의 성능을 평가한다. 이때 테스트 데이터의 조작으로 딥러닝 시스템이 잘못 인식되는 경우, 시스템의 성능 저하는 물론 치명적인 결과로 이어질 가능성도 있다. 예를 들어 카메라를 이용하여 외부의 장애물을 인식하는, 차량의 딥러닝 기반 자율주행 시스템의 경우 물리적 조작을 받은 교통표지판 영상이 카메라에 입력되어 잘못된 조향 명령을 내리는 경우, 치명적인 사고로 이어질 가능성이 높다[3]. [3]에서 제기된 딥러닝 시스템의 오분류 가능성은 2016년 Ian Goodfellow에 의해 사진에 인간의 시각 시스템으로는 알아챌 수 없을 정도의 잡음 (perturbation)을 삽입하고 시스템의 오 분류를 유도하여 당시 가장 앞선 수준의 딥러닝 시스템을 기만할 수 있다는 것을 보여줌으로서 현실적인 문제로 대두되었다[4].

적대적 예제 (adversarial example)는 [4]의 예에서 나타낸 바와 같이 공격하고자 하는 신경망이 정확히 분류해야 하는 샘플에 인간의 시각 시스템에는 탐지되지 않는 적은 양의 잡음을 추가하여 해당 신경망이 잘못 인식하게 되는 샘플을 의미하여 이와 같은 예제를 이용하여 딥러닝 시스템을 무력화시키는 공격을 적대적 공격(adversarial attack)이라고 정의한다[5]. 적대적 공격이 가능하게 하는 적대적 예제의 생성 원인은 통상적으로 신경망 자체의 고차원 유닛 하나와 신경망 내의 고차원 유닛의 선형 결합들 사이에 큰 차이가 없는 점과 심층 신경망의 입출력 간의 매핑이 매우 비선형적이라 작은 잡음의 추가에도 신경망의 분류에 심각한 오류를 발생시킨다는 것이다.

또한, 적대적 공격의 심각한 점은 서로 다른 구조를 갖는 두 개의 신경망 모델이 서로 다른 학습 데이터를 이용하여 학습된 경우에도 동일한 잡음에 의하여 영향을 받을 수 있는 전이 가능성 (transferability) 때문이다[6]. 즉 모델의 파라미터가 알려진 상황에서 적대적 공격으로 얻어진 모델의 정보를 이용하여 모델의 파라미터가 알려지지 않는 시스템에 대한 공격 즉, 블랙박스 공격 (black box attack)이 가능하다는 점이다[7].

본 연구에서는 서술된 적대적 공격을 무력화시키기 위한 방어 기술로서 백색 잡음을 데이터에 삽입하는 방법을 제안하였다. 제안한 방법은 적은 양의 잡음을 추가하여 딥러닝 시스템의 오분류를 유도하는 적대적 공격에 대항하기 위하여 적당한 양의 백색 잡음을 임의의 샘플에 추가하는 것이다. 이때 공격을 받지 않은 샘플에 백색 잡음을 추가한 경우에 분류 결과에는 영향을 미치지 않는다는 것을 조건으로 한다. 제안한 방법은 MNIST 데이터셋을 적용한 Resnet 신경망 모델[8]을 이용하여 대표적인 적대적 공격들인 FGSM(Fast Gradient Sign Method), BIM(Basic Iterative Method) 그리고 CW(Carlini and Wagner) 공격에 적용하였고 실험 결과는 백색 잡음 적용 전과 후의 적대적 공격에 대한 Resnet 시스템의 분류 정확도를 측정하고, 원본 데이터에 적용된 백색 잡음의 영향을 관찰하기 위하여 백색 잡음이 추가된, 공격받지 않은 샘플의 분류 정확도를 측정하는 것으로, 제안한 방법의 효용성을 평가하였다.

2. 이론적 배경

영상의 분류와 관련된 분류 시스템 모델은 영상 공간 X의 임의의 영상 x가 분류기 f에 입력되어 거쳐 1에서 k까지의 출력 값 가운데 하나를 발생 시키는 것으로 식 (1)과 같이 표현한다.

$$f: x \in X \rightarrow \{1,2,3,\cdots,k\} \tag{1}$$

이때 적대적 예제는 두 입력 영상 x와 x'간의 임의로 정의된 거리 값 D가 ϵ보다 작고 두 값의 분류기 출력이 서로 다른 식 (2)와 같이 정의된다.

$$\begin{cases} D(x,x') < \epsilon \text{ for } \epsilon > 0 \\ f(x) \neq f(x') \end{cases} \tag{2}$$

위 식에서 거리 값 D는 L_p norm의 거리 측도로서 원본 영상과 적대적 영상 사이의 차이 $\| x-x' \|_p$를 나타낸다. 특히 적대적 예제를 생성할 때에는 각 영상에서 동일한 원소의 개수 나타내는 0-norm, 두 영상 사이의 유클리드 거리(Euclidean distance)를 나타내는 2-norm 그리고 두 영상의 차이 값이 가장 큰 값을 표현하는 ∞-norm을 가장 많이 사용한다.

Goodfellow 등에 의하여 처음으로 제안된 FGSM 공격은 L_∞ 기반의 공격 방법으로서 손실함수 $J(\theta, x, y)$의 그레디언트(gradient) 방향으로 한 번의 갱신(updating)을 통하여 적대적 예제를 생성하는 공격 기법으로 식 (3)과 같이 정의된다[9].

$$x' = x + \epsilon \cdot sign[\nabla_x J(\theta, x, y)] \quad (3)$$

FGSM 공격은 전형적인 비타깃 공격(non-target attack)의 일종으로서 손실함수 $J(\theta, x, y)$의 그레디언트의 기울기를 임의의 타깃 방향 y'으로 변경하여 타깃 공격(target attack)으로 확장 가능한 특성이 있다. 또한 불규칙하게 ϵ이 결정되는 경우 적대적 공격의 성능이 향상되는 것으로 알려져 있다.

FGSM 공격이 한 번의 갱신에 의해 적대적 예제를 발생시키는데 비하여 더 최적화된 적대적 예제를 생성하기 위하여 연속적인 순환을 이용하는 공격 방법이 BIM 공격이다[10]. BIM 공격의 갱신 식은 식 (4)와 같이 나타낼 수 있고 반복 횟수 T에 대하여 적절한 영역을 설정하고 적대적 예제를 생성한다.

$$x_{t+1}' = Clip\{ x + \alpha \cdot sign[\nabla_x J(\theta, x_t', y)] \} \quad (4)$$

위 식에서 $\alpha T = \epsilon$이고 α는 각 순환에서 잡음의 크기이다. 일반적으로 잡음에 제한 조건을 주기 위하여 적대적 예제는 각 순환에서 잡음의 크기 ϵ에서 L_∞-norm의 관점에서 투사(projection)하여 식 (5)와 같이 잡음의 크기를 더 작게 조절함으로서 인간의 시각시스템 뿐만 아니라 적대적 공격 검출 시스템에 강건한 정밀한 공격이 가능하다.

$$x_{t+1}' = Proj\{ x + \alpha \cdot sign[\nabla_x J(\theta, x_t', y)] \} \quad (5)$$

Carlini와 Wagner에 의해 제안된 CW 공격은 식 (6)의 목적 함수를 최적화시키는 적대적 공격으로 L_0, L_2 및 L_∞ norm을 기반으로 한다[11].

$$\min D(x, x+\epsilon) + c \cdot f(x+\epsilon) \quad (6)$$
$$subject\ to\ x+\epsilon \in [0,1]$$

위 식에서 $f(\cdot)$은 손실부로서 분류기가 잘못된 분류를 수행하였을 때 0이하의 값을 반환한다. CW 공격은 화이트 박스 공격으로 100%의 성공률을 갖고 있으며 근사화를 이용하여 최소한의 잡음을 담당하는 손실 함수와 공격 성공률을 담당하는 손실함수의 합을 최소화함으로써 최적의 적대적 예제를 생성하도록 설계되어 있다. FGSM 공격, BIM 공격 그리고 가장 강력한 CW 공격은 적대적 예제에 대한 방어 기술을 평가하기 위한 지표로 널리 사용되고 있다.

3. 백색 잡음 가산 방법의 제안

적대적 공격에 대한 방어 기술로서 가장 널리 알려져 있는 방식이 적대적 훈련(adversarial training)이다. 적대적 훈련 방식은 딥러닝 시스템에 사전에 다양한 적대적 공격으로 적대적 예제를 생성하고 이를 학습하여 시스템이 예측되는 적대적 공격에 대하여 면역성을 기르는 방법이다. 그러나 이 방법은 예측할 수 있는 공격의 범위가 매우 방대하기 때문에 모든 적대적 공격에 대하여 학습할 수 없는 단점이 있다. 또한 다른 방어 기술로는 딥러닝 모델이 데이터를 분류하는데 사용하는 알고리즘을 지속적으로 변경하여 적대적 공격으로 인한 시스템의 교란 가능성을 줄

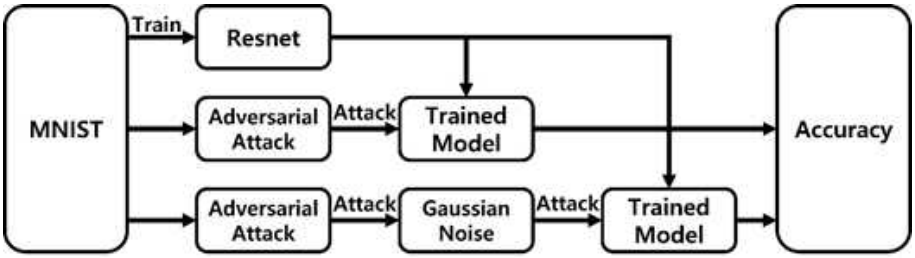


그림 1. 적대적 예제 생성 및 백색 잡음 가산을 위한 실험 흐름도
Fig. 1. Experimental flow chart of generation of adversarial example and adding white noise

이는 것이다. 그러나 이 방법은 알고리즘을 교체하는 비용과 시스템의 재훈련 과정 (retraining) 문제로 현실적인 어려움이 있다. 그러므로 적대적 공격에 대응하는 방어 기술은 현실적이면서 적은 비용과 시간을 들이는 효율성이 중요시 되는 관점에서 연구되어야 한다.

본 연구에서는 이론적 배경에서 언급된 적대적 예제에서 원본 영상과 적대적 예제 영상사이의 차이를 실험적으로 비교한 결과 백색 잡음과 유사한 패턴을 나타내고 있음을 관찰하였다. 이과 같은 관찰은 원본 영상에 백색 잡음을 가산하여 올바른 인식을 출력한다는 가정 하에 적절한 양의 백색 잡음을 적대적 예제에 가산하면 실험적 관찰로부터 얻어진 백색 잡음 형태의 원본 영상과 적대적 예제 영상 사이의 차이를 상쇄하여 올바른 인식 결과를 낼 수 있다고 가정하였다. 즉. 식 (1)과 식 (2)를 기반으로 다음의 두 가지 조건을 만족하는 백색 잡음을 삽입하여 정확한 인식 결과를 얻을 수 있다.

$$\begin{cases} c_i = f(x + N(0,\sigma^2)) \\ d_j = f(x' + N(0,\sigma^2)) \end{cases} \tag{7}$$

$$Find\ N(0,\sigma^2)\ \text{for}\ c_i = d_j \tag{8}$$

식 (7)으로부터 원본 영상에 백색 잡음을 가산하는 경우에도 원본 영상에 대한 분류기의 결과 c_i에는 영향을 미치지 않으며 적대적 예제의 분류 결과 d_j에 영향을 주어 식 (8)과 같이 적대적 예제의 올바른 출력 c_i를 분류기가 출력하도록 하는 백색 잡음을 찾는 알고리즘을 구성하였다.

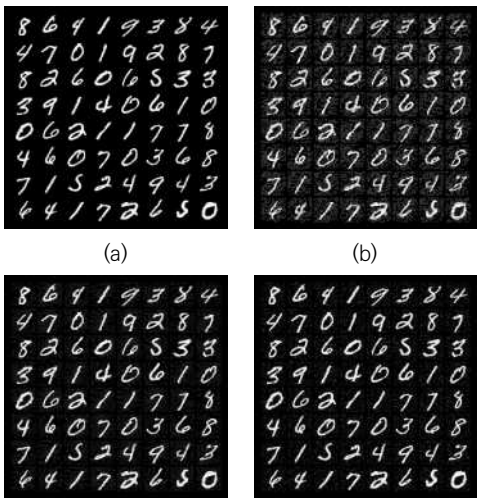
4. 실험 및 고찰

본 연구에서는 적대적 예제에 대한 방어 기술로서 백색 잡음을 가산하는 실험을 수행하기 위하여 그림 1과 같은 실험 흐름 도를 구성하였다. 실험을 위하여 사용된 데이터 셋은 MNIST의 손글씨 숫자데이터를 이용하였다.

그리고 사용된 딥러닝 모델은 Resnet으로 레이어의 수에 따라 Resnet 18, Resnet 34, Resnet 50, Resnet 101 모델이 사용되었다. 그리고 적대적 공격으로는 FGSM 공격, BIM 공격 및 표 1은 그림 1의 실험 흐름도에 적용된 딥러닝 모델과 적대적 공격들과 데이터 셋을 나타낸다.

표 1. 실험을 위한 딥러닝 모델, 데이터 셋 및 공격 타입
Table 1. Deep learning model, dataset and adversarial attacks

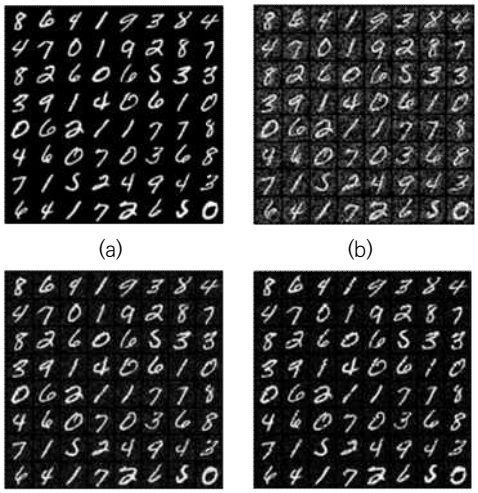
Deep learning model	Resnet 18	Resnet 34
	Resnet 50	Resnet 101
Attack types	FGSM attack	
	BIM attack	
	CW attack	
Dataset	MNIST	



(a) (b) (c) (d)

그림 2. 원본 MNIST 데이터와 공격을 받은 MNIST 데이터의 일례들 (a) 원본 MNIST 데이터 (b) FGSM 공격을 받은 MNIST 데이터 (c) BIM 공격을 받은 MNIST 데이터 (d) CW 공격을 받은 데이터

Fig. 2. Original MNIST data and examples of attacked MNIST data (a) original MNIST data (b) FGSM attacked MNIST data (c) BIM attacked data (d) CW attacked data



(a) (b) (c) (d)

그림 3. 백색 잡음이 가산된 원본 MNIST 데이터와 공격을 받은 MNIST 데이터의 일례들 (a) 원본 MNIST 데이터 (b) FGSM 공격을 받은 MNIST 데이터 (c) BIM 공격을 받은 MNIST 데이터 (d) CW 공격을 받은 데이터

Fig. 3. Original MNIST data and examples of attacked MNIST data with additive noise (a) original MNIST data (b) FGSM attacked MNIST data (c) BIM attacked data (d) CW attacked data

표 2 딥러닝 모델의 하이퍼 파라미터

Table 2. Hyper parameters of Deep learning model

Hyper parameters	
Epoch	10
Optimization method	Adam
Learning rate	1e−3
Batch size	64

실험의 순서는 먼저 MNIST 데이터 셋을 이용하여 각각의 Resnet에 대한 학습을 진행한다. 학습에 완료된 Resnet 모델과는 별도로 테스트용 MNIST 데이터 셋과 서술된 3개의 적대적 공격을 받은 데이터 셋에 적절한 양의 백색 잡음을 추가한다. 최종적으로 공격을 받은 테스트용 데이터 셋과 공격을 받은 후 백색 잡음을 추가한 데이터 셋 사이의 인식 율을 비교하였다. 표 2는 실험을 위하여 설정한 딥러닝 모델의 하이퍼 파라미터를 나타내고 있다.

그림 2는 원본 MNIST의 손글씨 숫자 영상과 3 종류의 공격을 받은 MNIST 손글씨 영상의 일례를 보여주고 있다. FGSM 공격을 받은 영상은 한 번의 갱신 식을 이용하여 적대적 예제를 생성하기 때문에 인간의 시각 시스템에 노출되는 경우도 존재하는 것을 확인할 수 있다. 즉, 그림 2(a)의 원본과 비교할 경우 잡음이 추가된 것 같은 패턴을 눈으로 인식할 수 있다.

표 3. 파라미터의 변화에 따른 정확도 비교
Table 3. Accuracy comparison in change of parameters

Attacks	Parameter and Accuracy					
FGSM	ϵ- value	0	0.1	0.2	0.3	0.4
	Accuracy	99.3	86.2	41.5	14.3	11.2
BIM	ϵ- value	0	0.1	0.12	0.14	0.16
	Accuracy	99.5	48.6	46.1	27.5	24.3
CW	c- value	0	0.2	0.4	0.6	0.8
	Accuracy	99.1	75.9	23.7	14.4	11.0

그러나 BIM 공격을 받은 그림 2의 (c)는 원본 영상과 비교하여 구분이 불가능할 정도로 영상의 품질이 향상된 것을 확인할 수 있으며 그림 2(d)의 CW 공격을 받은 영상은 원본과의 비교할 때, 인간의 시각 시스템으로 구분할 수 없을 정도이다. 그림 3은 그림 2의 원본 영상 및 공격받은 영상들에 대하여 백색 잡음이 가산된 영상을 나태내고 있다. 원본 영상에 대한 백색 잡음의 추가는 원본 영상의 의미(context)를 왜곡되지 않을 정도에 추가되었으며 그림 2의 영상과 백색 잡음이 가산된 그림 3의 영상은 숫자를 구분하는 인식의 의미에서 큰 차이를 보이지 않는다.

3종류의 공격에서 파라미터에 따른 인식률을 측정하기 위하여 딥러닝 모델을 Resnet18로 고정하고 각 공격에서 파리미터의 변화에 따른 인식율의 정확도를 측정하여 표 3에 나타내었다. 표 3으로부터 FGSM 공격을 받은 Resnet18의 정확도는 ϵ값이 0.1에서 0.4로 증가하면서 인식률이 급격하게 감소하는 것을 나타내고 있으며 $\epsilon = 0.4$이하에서는 0.4일 경우와 큰 차이가 없는 인식률을 나타내었다. 또한 BIM 공격에서는 ϵ값에 민감하게 반응하는 인식율의 감소를 보여주었다. FGSM 공격에서 $\epsilon = 0.1$에서 0.2 사이의 인식률 하락이 BIM 공격에서는 $\epsilon = 0.1$에서 0.12 사이에 나타는 것을 관찰할 수 있다.

또한 c값에 의해 결정되는 CW 공격에서는 $c = 0.4$ 이후로 급격한 인식율의 변화를 관찰할 수 있다. 전체적인 인식율의 변화는 그림 4에서 관찰할 수 있다. 모든 공격에서 파라미터의 값이 증가하면서 Resnet18의 인식률이 떨어지는 것을 관찰할 수 있으며 마지막에 표시된 파라미터 값들은 그 이후의 값들에서 명확한 인식율의 변화가 나타나지 않는 것을 의미한다.

본 연구에서는 각각의 공격에서 가장 심각한 경우를 고려하여 FGSM 공격과 BIM 공격에서는 표 3에 근거하여 ϵ값을 각각 0.4와 0.16으로 설정하였고 CW 공격에서는 c값을 0.8로 설정하고 실험이 진행되었다.

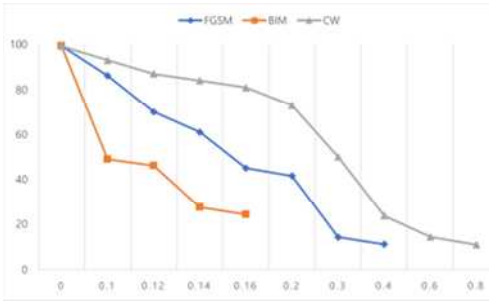


그림 4. 파라미터 변화에 따른 공격들의 인식 정확도 비교
Fig. 4. Comparison of recognition accuracy in change of parameters

최종적으로 백색 잡음을 테스트 데이터 셋과 공격을 받은 데이터 셋에 동시에 적용하고 표 1에서 기술된 4개의 Resnet 모델들에 대하여 인식율의 변화를 관찰하였다.

표 4. 공격받은 모델의 인식율의 정확도
Table 4. Recognition Accuracy of attacked models with layers

Attacks / Model	FGSM	BIM	CW
Resnet 18	14.4	10.5	7.9
Resnet 34	12.1	12.1	10.6
Resnet 50	6.5	7.0	15.2
Resnet 101	5.4	6.3	17.6

표 5. 백색 잡음이 가산된 공격받은 모델의 인식률 정확도
Table 5. Recognition Accuracy of attacked models with additive white noise

Attacks / Model	FGSM	BIM	CW
Resnet 18	87.6	18.7	71.7
Resnet 34	90.6	18.6	75.0
Resnet 50	96.3	12.2	79.2
Resnet 101	95.6	12.6	80.8

표 4와 표 5는 가장 심각한 오인식율을 나타내는 파라미터 값에서 4개의 Resnet 모델에 대하여 10번의 반복 실험을 통하여 얻어진 평균 인식률을 나타내고 있다. 이미 예측한 바와 같이 모든 Resnet 모델에서 매우 낮은 인식율의 정확도를 나타내고 있음을 관찰할 할 수 있다. 또한 그림 5는 공격 받은 테스트 영상에 경험적인 방법으로 분산 0.5인 가우시안 백색 잡음을 가산하고 인식율의 정확도를 동일하게 10회 실험의 평균으로 나타낸 것이다. 표 4에 비하여 모든 모델에서 인식률이 개선된 것을 확인할 수 있다.
그림 5는 FGSM 공격에 대한 각 모델별 인식율의 정확도를 나타낸 그래프이다. 모든 모델에서 백색 잡음을 가산하였을 경우 인식율의 증가를 관찰할 수 있고 레이어의 수가 증가할수록 인식률이 증가하는 것을 확인할 수 있다. 그림 6은 BIM 공격에 대한 각 모델별 인식률을 나타낸 그래프이다.

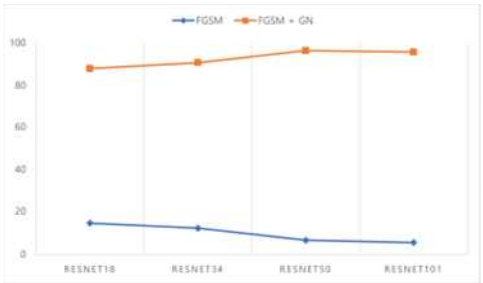


그림 5. FGSM 공격에 대한 모델들의 인식률 정확도 비교
Fig. 5. Recognition accuracy comparison against FGSM attack

그림 6. BIM 공격에 대한 모델들의 인식률 정확도 비교
Fig. 6. Recognition accuracy comparison against BIM attack

BIM 공격에서는 백색 잡음을 가산하지 않은 경우에 비하여 인식률이 더 향상된 것을 확인할 수 있으나 FGSM 공격에서와 같은 뚜렷한 인식율의 향상을 나타내지 않았다.

그림 7. CW 공격에 대한 모델들의 인식률 정확도 비교
Fig. 7. Recognition accuracy comparison against CW attack

그림 7은 CW 공격에 대한 각 모델별 인식률을 나타낸 그래프이다. 백색 잡음을 가산하였을 경우 FGSM 공격에서와 동일하게 높은 인식율의 변화가 모든 모델에서 나타난 것을 관찰할 수 있다. 특히 CW 공격은 화이트 박스 공격에서 100%의 성공률을 갖고 있기 때문에 백색 잡음을 가산하는 방법이 화이트 박스 공격에 우수한 방어 기술이라 할 수 있다.

표 6. 적대적 훈련 방법과 제안한 방법의 정확도 비교
Table 6. Comparison of Recognition Accuracy of proposed and adversarial training method

Model \ Attacks	FGSM	BIM	CW
Proposed	95.6	X	83.99
Adversarial training	80.8	X	77.15

표 6은 [12]에서 연구된 결과로서 본 연구에서 사용한 동일한 데이터 셋에 대하여 Lenet6 구조를 갖는 딥러닝 모델을 이용하여 적대적 학습 방법을 이용하였을 경우 인식율의 정확도를 보여주고 있다. 제안한 방법은 적대적 학습 방법에 비하여 FGSM과 CW 공격에서 더 우수한 것을 보여주고 있으며 BIM의 경우는 비교 자료가 없어 비교 할 수 없었다. 제안한 방법은 기존의 적대적 학습 방법과 훈련 시간에 대하여 비교하는 경우 제안한 방법은 기존의 데이터 셋을 학습하는 방법과 동일한 시간을 소비하지만 적대적 학습 방법은 각각의 공격에 대한 적대적 예제를 모두 훈련해야 하므로 더 많은 시간을 소비한다.

5. 결 론

본 논문에서는 영상 데이터에 대한 적대적 공격으로부터 생성된 적대적 예제로 인하여 발생할 수 있는 딥러닝 시스템의 오분류를 방어하기 위한 방법으로 분류기의 입력 영상에 백색 잡음을 가산하는 방법을 제안하였다. 제안된 방법은 적대적이든 적대적이지 않던 구분하지 않고 분류기의 입력 영상에 백색 잡음을 더하여 적대적 예제가 분류기에서 올바른 출력을 발생할 수 있도록 유도하는 것이다. 제안한 방법은 FGSM 공격, BIM 공격 및 CW 공격으로 생성된 적대적 예제에 대하여 서로 다른 레이어 수를 갖는 Resnet 모델에 적용하고 결과를 고찰하였다. 백색 잡음의 가산된 데이터의 경우 모든 Resnet 모델에서 인식률이 향상되었음을 관찰할 수 있다. FGSM 공격의 경우에는 Resnet 101 모델에서 백색 잡음을 가산하여 인식률이 5.4%에서 95.6%로 증가하였고 강력한 공격인 CW 공격에서는 같은 모델에 대하여 17.6%의 인식률이 80.8%로 증가하였다.

제안된 방법은 단순히 백색 잡음을 경험적인 방법으로 가산하고 결과를 관찰하였으나 이에 대한 엄밀한 분석이 추가되는 경우 기존의 적대적 훈련 방법과 같이 비용과 시간이 많이 소요되는 적대적 공격에 대한 방어 기술을 제공할 수 있을 것으로 사료되며 추가적인 이론 연구가 필요하다.

REFERENCES

[1] Alzubaidi, Laith, et al. "Review of deep learning: Concepts, CNN architectures challenges, applications, future directions," Journal of big Data, Vo; 8, no. 1pp. 1-74, 2021.
[2] Singh, Kanwar Bharat, and Mustafa Ali Arat. "Deep learning in the automotive industry: Recent advances and application examples," arXiv preprint :1906.08834, 2019.
[3] W. Sultani, J. Choi,"Abnormal traffic detection using intelligent driver model," 2010 20th Int. Conf. Pattern Recognit. pp. 324-327, 2010.
[4] I. Goodfellow, S, Jonathon Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
[5] Wiyatno, Rey Reza, et al. "Adversarial examples in modern machine learning: A review," arXiv preprint arXiv:1911.05268, 2019.
[6] Chakraborty, Anirban, "Adversarial attacks and defences: A survey," arXiv preprint arXiv:1810.00069, 2018.
[7] N. Papernot, M. McDaniel, and I. Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," arXiv :1605.07277, 2016.
[8] Targ, Sasha, Diogo Almeida, and Kevin Lyman. "Resnet in resnet: Generalizing residual architectures," arXiv:1603.08029, 2016.
[9] Huang, Sandy, et al. "Adversarial attacks on

neural network policies," arXiv preprint arXiv:1702.02284, 2017.
[10] Kurakin Alexey, Ian Goodfellow, and @Samy Bengio. "Adversarial examples in the physical world," 2016.
[11] N. Carlini and David Wagner. "Towards evaluating the robustness of neural networks," 2017 ieee symposium on security and privacy (sp). IEEE, 2017.
[12] Sanglee Park and Jungmin So, "On the Effectiveness of Adversarial Training in Defending against Adversarial Example Attacks for Image Classification," Applied Science, Vol. 10, pp. 2-16, 2020.

저자약력

이 영 석 (Youngseok Lee) **[정회원]**

- 1995년 2월 : 서울시립대학교 대학원 전자공학과 (공학석사)
- 1998년 2월 : 서울시립대학교 대학원 전자공학과 (공학박사)
- 1998년 3월 ~ 현재 : 청운대학교 인천캠퍼스 전자공학과 교수

〈관심분야〉 딥러닝 모델, 신경망 모폴로지, 기계학습

김 종 원 (Jongweon Kim) **[정회원]**

- 1989년 2월 : 서울시립대학교 대학원 전자공학과 (공학석사)
- 1995년 2월 : 서울시립대학교 대학원 전자공학과 (공학박사)
- 2009년 3월 ~ 현재 : 상명대학교 SW융합학부 지능IOT융합 교수

〈관심분야〉 디지털신호처리, 컴퓨터비전, 인공지능