Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.30, NO.5, Oct. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.5.871

적대적 공격에 견고한 Perceptual Ad-Blocker 기법*

김 민 재,[1†] 김 보 민,[1] 허 준 범[2‡]
[1,2]고려대학교 (학생, 교수)

Perceptual Ad-Blocker Design For Adversarial Attack*

Min-jae Kim,[1†] Bo-min Kim,[1] Junbeom Hur[2‡]
[1,2]Korea University (Student, Professor)

요 약

Perceptual Ad-Blocking은 인공지능 기반의 광고 이미지 분류 모델을 이용하여 온라인 광고를 탐지하는 새로운 광고 차단 기법이다. 이러한 Perceptual Ad-Blocking은 최근 이미지 분류 모델이 이미지를 틀리게 분류하게끔 이미지에 노이즈를 추가하는 적대적 예제(adversarial example)를 이용한 적대적 공격(adversarial attack)에 취약하다는 연구 결과가 제시된 바 있다. 본 논문에서는 다양한 적대적 예제를 통해 기존 Perceptual Ad-Blocking 기법의 취약점을 증명하고, MNIST, CIFAR-10 등의 데이터 셋에서 성공적인 방어를 수행한 Defense-GAN과 MagNet이 광고 이미지에도 효과적으로 작용함을 보인다. 이를 통해 Defense-GAN과 MagNet 기법을 이용해 적대적 공격에 견고한 새로운 광고 이미지 분류 모델을 제시한다. 기존 다양한 적대적 공격 기법을 이용한 실험 결과에 따르면, 본 논문에서 제안하는 기법은 적대적 공격에 견고한 이미지 분류 기술을 통해 공격 이전의 이미지 분류 모델의 정확도와 성능을 확보할 수 있으며, 더 나아가 방어 기법의 세부사항을 아는 공격자의 화이트박스 공격(White-box attack)에도 일정 수준 방어가 가능함을 보였다.

ABSTRACT

Perceptual Ad-Blocking is a new advertising blocking technique that detects online advertising by using an artificial intelligence-based advertising image classification model. A recent study has shown that these Perceptual Ad-Blocking models are vulnerable to adversarial attacks using adversarial examples to add noise to images that cause them to be misclassified. In this paper, we prove that existing perceptual Ad-Blocking technique has a weakness for several adversarial example and that Defense-GAN and MagNet who performed well for MNIST dataset and CIFAR-10 dataset are good to advertising dataset. Through this, using Defense-GAN and MagNet techniques, it presents a robust new advertising image classification model for adversarial attacks. According to the results of experiments using various existing adversarial attack techniques, the techniques proposed in this paper were able to secure the accuracy and performance through the robust image classification techniques, and furthermore, they were able to defend a certain level against white-box attacks by attackers who knew the details of defense techniques.

Keywords: Adversarial example, Perceptual Ad-Blocker, Defense-GAN, MagNet

Received(08. 13. 2020), Modified(10. 07. 2020), Accepted(10. 07. 2020)

* 본 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행하였습니다.

† 주저자, 97alswo@naver.com
‡ 교신저자, jbhur@korea.ac.kr(Corresponding author)

I. 서 론

방송통신광고통계시스템에 따르면 온라인에서의 광고 산업은 2015년부터 현재까지 계속 증가하고 있다[9]. 이에 따라 Ad-Block에 대한 개발도 함께 성장해왔고 메타데이터를 활용한 기존 Ad-block과 다르게 광고 이미지 자체와 인공신경망(neural network)을 활용한 Perceptual Ad-Blocker[3]이라는 새로운 Ad-Block 기법이 등장하였다.

하지만 이러한 Perceptual Ad-Blocker이 classifier 모델의 잘못된 예측을 출력하도록 이미지에 적절한 노이즈를 추가하는 적대적 공격(adversarial attack)에 취약하다는 연구[3]가 제시되었다. 해당 논문에서는 perceptual Ad-Blocker들은 적대적 공격에 매우 취약하며, Ad-Blocker의 다양한 제약조건 및 방어의 한계점 등을 들며 방어가 힘듦을 주장하였다.

본 논문에서는 [3]에서 보인 공격 이외에도 다양한 적대적 공격에 Ad-Blocker들이 취약함을 보이고, 이에 더하여 adversarial attack에 대한 견고성을 높일 수 있는 Perceptual Ad-Blocker을 위한 방어 모듈을 제안하고자 한다. 우리는 Ad-Blocker의 다양한 제약조건들을 고려하여 MNIST[18], CIFAR-10[17] 등의 데이터에서 성공적인 방어 성능을 보인 Defense-GAN[4]과 MagNet[5]을 응용한 방어 모듈을 classifier에 탑재하였고, 다양한 공격에 대해서 실험한 결과 이들은 classifier 모델의 성능을 크게 떨어트리지 않으면서 범용적인 공격에 대해 방어가 가능하며, 추가적인 연산 비용(computational overhead)이 크지 않아 효율적이다.

또한, [14, 16] 등의 논문에서 MagNet과 Defense-GAN을 포함한 많은 방어 기법들이 보다 강력한 적응적 공격(adaptive attack)에 대해 취약함을 보였다. 본 논문에서 제시하는 모델에서 이러한 공격에 대한 실험 결과 적응적 공격이 어느 정도는 유효하나, [14, 16]에서 주장한 바와 같이 방어를 완전히 무력화할 정도는 아님을 확인할 수 있었다. 또한, 적응적 공격에는 매우 많은 시간이 소요되며, 특히 transferability[1,2]를 이용한 gray-box attack model에서는 학습 과정에 더욱 많은 시간이 소요되기 때문에 적대적 예제의 생성에 걸리는 시간이 공격 성공률을 고려할 때 비효율적임을 확인할 수 있었다. 따라서 인공신경망을 이용한 Perceptual Ad-Blocker 환경에서 적대적 공격에 대한 고려와 대응 방안에 대한 연구는 매우 중요하다고 할 수 있다.

II. 이론적 배경

2.1 Adversarial attack

적대적 공격(adversarial attack)이란 이미지 x에 작지만 의도적으로 최악의 perturbation(δ)를 추가하여 적대적 예제 x'를 생성하는 것으로, 모델이 높은 신뢰도로 잘못된 답을 출력하게 할 수 있다.

이러한 적대적 예제를 찾는 알고리즘들 중 PGD(Projected Gradient Descent)[10,11], DeepFool[12], CW(Carlini&Wagner's attack)[16]을 실험에 사용했다.

2.1.1 FGSM

Goodfellow 등[2]은 이미지와 l_∞ distance를 가지는 적대적 예제를 생성하는 untargeted FGSM을 최초로 제안하였다. FGSM은 대표적인 one-step 공격 알고리즘으로, 가장 가파른(steepest) 방향으로 optimization loss $J(\theta,x,y)$를 증가시키기 위해 loss의 gradient 방향을 따라 이미지를 갱신한다. 적대적 예제 x'은 다음과 같이 생성된다.

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

이 때 ϵ은 perturbation의 크기를 나타내고, θ은 model의 weight, x는 input image, y는 x에 대한 label을 나타낸다. $J(\theta,x,y)$는 optimization loss 혹은 adversarial loss라고 부른다.

untargeted FGSM은 optimization loss를 수정하여 targeted FGSM으로 쉽게 확장시킬 수 있다. True label에 대한 optimization loss를 극대화하는 untargeted FGSM과 달리, target label y'에 대한 optimization loss $J(\theta,x,y')$를 최소화시키는 방식으로 targeted FGSM을 정의할 수 있다. 따라서 targeted FGSM을 이용한 adversarial example x'은 다음과 같이 생성될 수 있다.

$$x' = x - \epsilon \cdot sign(\nabla_x J(\theta, x, y'))$$

FGSM은 이미지를 한 단계만 갱신시키는 방법으로, 빠르게 적대적 예제를 생성할 수 있다는 장점이 있다.

2.1.2 PGD

PGD[10,11]는 $\| \delta \|_\infty \leq \epsilon$를 만족하는 노이즈 집합에서 loss function의 gradient의 방향으로 반복적으로 FGSM 알고리즘을 이용하여 δ를 갱신하여 최악의 δ를 찾는 알고리즘이다.

2.1.3 DeepFool

Moosavi-Dezfooli 등[12]이 제안한 DeepFool은 타겟 모델이 선형이라고 가정하고 적대적 예제 x'을 찾는다. Input image x와 가장 가까운 decision boundary를 찾고, 이 방향으로 x'를 갱신한다. x'가 decision boundary를 넘어갈 때까지 해당 과정을 반복하고, 작은 크기의 perturbation으로 적대적 예제를 찾을 수 있다.

2.1.4 CW attack

Carlini와 Wagner[16]은 l_0, l_2, l_∞ 세 개의 metric을 이용하여 적대적 예제를 생성해내는 최적화 기반의 적대적 공격을 제안하였다. 최적화 목적함수는 다음과 같다.

$$\min_\delta D(x, x+\delta) + c \cdot f(x+\delta)$$
$$\text{subject to } x+\delta \in [0,1]$$

여기서 δ는 perturbation, $D(\cdot,\cdot)$은 distance metric, $f(\cdot)$은 loss term으로, classifier가 입력을 틀리게 분류하였을 때 0이하의 값을 반환한다. $f(x')$은 다음과 같이 정의된다.

$$f(x') = \max(Z(x')_{l_x} - \max Z(x')_i : i \neq l_x, -K)$$

$Z(x')$는 classifier의 logit(softmax 함수 직전의 vector)을 나타내며, K는 confidence를 의미하며, CW attack은 이 값을 조정하여 공격의 강도를 조절할 수 있다는 장점이 있다. K의 값이 클수록 classifier가 적대적 예제를 더 높은 confidence로 틀리게 분류한다.

2.2 Perceptual Ad-Blocker

Perceptual Ad-Blocker는 온라인 광고의 시각적 맥락(visual context)을 이용하여 광고를 탐지한다. Tramèr 등[3]은 Perceptual Ad-Blocker을 classifier가 광고를 판별하는 방법에 따라 (1) element-based, (2) page-based, (3) frame-based 세 가지로 분류하였다. Ad Highlighter 등의 element-based ad-blocker은 웹 페이지의 DOM tree에 접근하여 각의 HTML 요소들을 분할하여 그 중 ad-disclosure(Sponsered logo 등 광고임을 알리는 요소들)을 포함한 segment를 찾는다. Sentinel, Adblock plus 등의 page-based ad-blocker은 웹 페이지 전체의 스크린샷에서 object detection을 이용하여 광고를 직접 찾는다. Percival 등의 frame-based ad-blocker은 element-based와 유사하게 페이지를 분할하여, 각 segment들을 대상으로 classifier를 이용하여 광고를 직접 찾거나 object detection을 이용해 ad-disclosure를 찾는다.

Tramèr 등은 실험을 통해 다양한 종류의 Perceptual Ad-Blocker이 모두 적대적 공격(PGD)에 취약함을 보였으며, 대부분 100%에 이르는 공격 성공률을 보였다.

이와 같이 perceptual Ad-blocker들이 적대적 공격에는 매우 취약한 반면, 이에 대한 방어는 쉽지 않음을 제시했는데, 1) 공격자가 프로그램의 코드에 접근이 가능하기 때문에 classifier 모델의 내부구조를 활용한 공격인 Whitebox-attack이 가능하며, 2) 공격자는 미리 오프라인에서 적대적 예제를 준비할 수 있으나 Ad-Blocker 측에서는 온라인으로 광고를 분류해야 하여 시간적인 제약이 존재하는 점 등으로 인해 방어가 쉽지 않음을 주장하였다. 3) 또한, classifier 모델을 적대적 예제들을 포함하여 다시 훈련시키는 방어 기법인 adversarial training[1,11]을 수행하여도 이는 공격자가 새로운 적대적 예제를 생성하는 것보다 더 많은 비용이 들며, 현재까지 제안된 그 외의 방어 기법들은 제한된 공격자에만 대응할 수 있거나, 보다 강력한 공격

에 의해 방어가 깨질 수 있음이 증명되었다는 점[14,16] 등을 두어 방어에 한계가 있음을 주장하였다.

2.3 Adversarial Defense

현존하는 제안된 적대적 공격에 대한 방어 기법은 1) Training dataset을 수정하는 방법 2) Model의 구조를 수정하는 방법 3) 보조적인 도구(auxiliary tool)를 사용하는 방법으로 크게 세 분류로 나눌 수 있다.[15]

우리는 이 중 Ad-blocker의 제약과 다양한 특성들을 고려하여 3)에 해당하는 MagNet과 Defense-GAN을 방어에 사용하였다. 이들은 기존 classifier의 내부 구조를 수정하지 않고, 입력 데이터셋에 대한 학습이 사전에 이루어지기 때문에 모델의 성능에 영향을 미치지 않는다는 장점이 있다. 온라인으로 실시간으로 광고를 탑지해야 하는 Perceptual Ad-Blocker의 특성상 방어가 기존 classification 모델의 시간적인 성능을 떨어트리지 않아야 하므로 방어 기법으로 MagNet과 Defense-GAN을 선택하였다.

III. 연구방법

3.1 데이터셋

모델 훈련에 사용한 데이터셋은 Hussain 등[7]이 공개한 훈련 데이터셋 21,945개 (ad 8,348개, non-ad 13,597개)를 이용하였다. 3.2의 ResNet의 훈련과 3.3에서의 방어 모델의 훈련에 이 데이터셋을 사용하였다.

Classifier의 분류 성능을 측정하기 위한 테스트 데이터셋은 Tramèr 등[3]이 공개한 데이터셋을 사용하였다. [3]의 논문에서도 해당 데이터셋을 대상으로 적대적 공격에 대한 classifier의 성능을 측정했으며, 해당 데이터셋은 Alexa ranking에서 상위 10개의 뉴스 웹 사이트에서 수집된 실제 광고 데이터들이다. 총 59개의 광고 이미지들로 구성되어 있다(ad 39개, non-ad 20개).

Percival의 경우 100*100 이상의 크기의 이미지만 처리하기 때문에 percival의 성능 측정에는 이에 해당되는 데이터만 사용하였다.

3.2 Classifier

Tramèr, Florian 등은 element-based, frame-based, page-based 세 종류의 ad-blocker 8가지를 대상으로 adversarial attack을 수행하였으며, 8가지의 ad-blocker 모두 공격에 취약한 것으로 드러났다. 우리는 이들 중 classifier을 이용한 frame-based ad-blocker 두 가지 ResNet[13]을 이용한 Hussan et al's scheme[7]과 Percival이라는 model을 이용한 Percival[8] 의 분류 모델에 대한 방어를 수행하였다. Percival은 pre-trained model을 사용하였으며, ResNet은 3.1.1의 데이터셋을 이용하여 직접 훈련시켰다.

각 classifier의 성능은 Table 1과 같다. 각각의 모델에 대한 False Discovery Rate(광고로 분류하였으나 광고가 아닌 이미지의 비율), False Omission Rate(광고가 아니라고 분류하였으나 광고인 이미지의 비율), Accuracy(TP+TN/Total)를 측정하였으며 ResNet은 80%, Percival은 95%의 높은 정확도로 광고 이미지를 분류한다

Table 1. Classifier accuracy

Classifier	FDR	FOR	Acc.
ResNet	5/7	3/33	80%
Percival	2/7	0/33	95%

3.3 공격 기법

공격 기법은 PGD[10,11], FGSM[2], DeepFool[12], CW attack[16]을 사용하였다. 첫 3개의 공격 모델은 Table.2.와 같으며, CW attack의 경우 L_2 공격 모델을 사용하였으며 confidence K=1, learning rate=0.001로 두어 10000번의 iteration을 거쳐 공격을 수행하였다. 또한 모든 공격은 이미지에 가해진 노이즈를 육안 상으로 판별하기 힘든 정도이다.

Table 2. Attack algorithm and parameter

algorithm	attack parameter
PGD	$l_\infty \leq 2/255$
DeepFool	l_∞

각 공격의 파라미터는 서로 상이하여 각 공격들은 서로간의 비교가 어렵기 때문에 공격 간의 비교가 아닌, 공격에 대한 방어 전후의 성공률을 비교하는 것을 목적으로 한다.

각각의 공격 알고리즘으로 생성된 광고 이미지를 classifier에 넣은 후 공격이 적용된 광고 이미지에 대한 False Positive, False Negative, Accuracy를 측정하였다.

3.4 방어 기법

3.3장의 결과로 공격이 적용된 이미지를 방어 모듈이 탑재된 classifier에 입력으로 넣어 False Positive, False Negative, Accuracy를 측정하였다. 성능 비교를 위해 공격이 적용되지 않은 광고 이미지에 대해서도 분류 성능을 측정하였다.

방어 모듈은 MagNet을 이용한 것과 Defense-GAN을 이용한 것 두 종류에 대해서 실험하였다.

3.4.1 MagNet

Reformer과 detector로 사용한 autoencoder의 구조와 훈련 파라미터는 각각 Table 3, Table 4와 같다. autoencoder의 학습에 3.1.1의 데이터셋을 8:2의 비율로 훈련 데이터와 검증 데이터로 나누어 사용하였다.

학습 시 입력 이미지에 volume=0.1의 Gaussian noise를 추가한 후 [0, 1]의 범위로 clip하여 사용하였다.

MagNet을 제안한 논문[5]과 달리, detector network를 방어의 용도로 수정하여 사용하였다. Ad-Blocker에서 사용하는 classifier의 경우, [5]의 논문에서 실험한 multi-labael classifier가 아닌, Ad/Non-ad만을 판별하는 binary classifier이기 때문에, detector network가 적대적 예제로 판단한 이미지의 경우 label을 다른 label로 전환하는 방식으로, detector을 단순히 적대적 예제를 탐지하는 용도가 아닌, 또다른 방어의 방법으로 사용하였다.

Table 3. Structure of Magnet auto-encoder

Auto-encoder
Conv(3,3x3)
Sigmoid
Conv(3,3x3)
Sigmoid
Conv(3,4x4,2)
Sigmoid
ConvT(6,4x4,2)
Sigmoid
Conv(3,3x3)
Sigmoid

Table 4. Magnet auto-encoder parameter

Parameters	-
Optimization Method	Adam
Learning Rate	1e-3
Batch Size	256
Epochs	100
Regularization	Noise

3.4.2 Defense-GAN

사용한 Defense-GAN 구조와 훈련 파라미터는 각각 Table 5, Table 6과 같다. 훈련에 사용한 데이터는 MagNet과 동일하다.

Table 5. Defense-GAN structure

Generator	Discriminator
FC(200704)	Conv(3,3x3,2)
ReLU	LeakyReLU(0.2)
ConvT(256,2x2,2)	Conv(64,3x3,2)
ReLU	LeakyReLU(0.2)
ConvT(128,2x2,2)	Conv(128,3x3,2)
ReLU	LeakyReLU(0.2)
ConvT(64,2x2,2)	FC(1)
Tanh	-

Table 6. Defense-GAN parameter

Parameters	-
Input Latent	64
Lambda	10
Optimization Method	Adam
Optimization parameter	(0.5, 0.9)
Learning Rate	1e-4
Batch Size	32
Epochs	100
Critic Iters	5
Initial Epoch	0

IV. 실험결과

4.1 적대적 공격에 대한 방어 성능

우선, 공격이 적용된 이미지에 대한 classifier의 분류 성능은 Table 7과 같다. 높은 성능을 가지고 있던 classifier은 공격이 수행된 이미지에 대해 전반적으로 성능이 매우 하락한 바를 확인할 수 있으며, PGD와 DeepFool을 통하여 공격된 이미지에 대해서는 모든 이미지를 틀리게 분류하였다.

방어 모듈을 탑재한 classifier로 분류한 결과는 Table 8, Table 9와 같다. 공격이 적용된 이미지 뿐만 아니라 공격이 적용되지 않은 원래의 이미지에 대해서도 성능을 평가하였다. 각 공격이 적용된 이미지에 대하여 분류를 수행한 결과 성능이 전반적으로 향상되었다(40~90%).

공격이 적용되지 않은 이미지의 경우 방어 모듈을 탑재하지 않았을 때와 비교했을 때 약간의 성능 하락이 있었으나, 그 감소 폭이 유의미하게 크지 않다.

방어 모듈을 탑재하였을 때 발생하는 추가적인 실행시간은 MagNet은 거의 없고, Defense-GAN의 경우 사진 1장 당 평균 2.4초정도로 classifier의 전반적인 성능에 큰 영향을 미치지 않는다.

Table 7. Result of attack for classifiers

Attack	Classifier	FDR	FOR	Acc.
PGD	ResNet	7/7	28/33	12.5%
	Percival	7/7	33/33	0%
Deep Fool	ResNet	7/7	33/33	0%
	Percival	7/7	33/33	0%
CW	ResNet	7/7	30/33	7.5%
	Percival	7/7	31/33	5%

Table 8. Result of MagNet for adverarial attack

Attack	Classifier	FDR	FOR	Acc.
(None)	ResNet	4/7	4/33	80%
	Percival	2/7	1/33	92.5%
PGD	ResNet	2/7	6/33	80%
	Percival	0/7	1/33	97.5%
Deep Fool	ResNet	4/7	5/33	77.5%
	Percival	2/7	1/33	92.5%
CW	ResNet	4/7	4/33	80%
	Percival	2/7	5/33	82.5%

Table 9. Result of Defense-GAN for adversarial attack

Attack	Classifier	FDR	FOR	Acc.
(None)	ResNet	4/7	4/33	80%
	Percival	2/7	1/33	92.5%
PGD	ResNet	7/7	9/33	60%
	Percival	2/7	13/33	62.5%
Deep Fool	ResNet	6/7	10/33	60%
	Percival	5/7	11/33	60%
CW	ResNet	6/7	6/33	70%
	Percival	1/7	10/33	72.5%

4.2 적응적 공격에 대한 방어 성능

[14,16]에서 제시한 MagNet과 Defense-GAN에 대한 adaptive attack의 실험 결과는 Table 10, Table 11과 같다.

Defense-GAN은 그 구조의 복잡도로 인해 adaptive attack이 큰 효용이 없다. 그러나 MagNet은 사용한 autoencoder의 구조는 상대적으로 단순하여 (Table.3. 참고) adaptive attack이 높은 성공률을 보인다. 따라서 adaptive

Table 10. Result of MagNet for adaptive attack

	Classifier	FDR	FOR	Acc.
without defense	ResNet	5/7	3/33	80%
	Percival	3/7	3/33	85%
with defense	ResNet	7/7	33/33	0%
	Percival	2/7	30/33	2%

Table 11. Result of Defense-GAN for adaptive attack

	Classifier	FDR	FOR	Acc.
without defense	ResNet	5/7	3/33	80%
	Percival	2/7	0/33	95%
with defense	ResNet	6/7	9/33	62.5%
	Percival	1/7	10/33	72.5%

attack에 보다 robust한 defense를 위해서는 Defense-GAN이 적합하다.

4.3 블랙박스 공격에 대한 방어 성능

Blackbox attack에 대한 defense의 실험 결과는 Table 12와 같다. Blackbox attack으로는 가장 보편적인 PGD 방식을 채택했고, attack에 사용할 substitute classifier은 ResNet으로 설정하고 실험하였다. Defense를 탑지하지 않았을 때 공격을 수행하면 accuracy가 상당히 떨어지지만, defense-GAN과 MagNet을 탑재하면 공격을 수행해도 본래의 성능에서 크게 떨어지지 않는다. 따라서 본 논문에서 주장하는 defense는 black-box attack에도 성공적인 방어의 수행이 가능함을 알 수 있다.

Table 12. Result of defense for Black-box attack

PGD	Defense	FDR	FOR	Acc.
x	x	5/7	3/33	80%
o	x	7/7	18/33	37.5%
o	Defense-GAN	4/7	6/33	75%
o	MagNet	3/7	6/33	77.5%

4.4 방어 기법 효율성 분석

방어 모듈의 성능과 관련된 실험 결과에 따르면 공격이 적용되지 않은 이미지에 대해서는 Table 13에서 ResNet의 경우 성능이 하락하지 않고, Defense-GAN의 경우 경미한 성능 하락을 보인다.

Table 13. Performance comparison of models with defensive mechanisms for unattacked images

	None	MagNet	Defense-GAN
Percival	95%	92.5%	92.5%
ResNet	80%	80%	80%

4.5 Adversarial example 생성 시간

Defense-GAN에 대한 적응적 공격의 경우 각 알고리즘의 적대적 예제 생성 시간을 확인한 결과, 적대적 예제 생성에 걸리는 시간이 Table 14와 같이 다른 공격에 비해 상대적으로 매우 길기 때문에 공격의 효율성이 떨어짐을 확인할 수 있었다.

Table 14. Comparison of the time taken to create 40 adversarial examples

(min.)	Percival	ResNet
PGD	0.0046	0.0377
DeepFool	0.0838	0.3308
CW	99.65	58.59
MagNet	0.0071	0.3887
DefenseGAN	1672.17	1849.33

V. 결 론

Perceptual Ad-Blocker는 생성하기 쉬운 적대적 예제에 대해 매우 취약해 공격자보다 불리한 고점에 위치해 있다. 본 논문에서는 Perceptual Ad-Blocker이 다양한 종류의 적대적 공격에 대해 취약함을 실험을 통해 보이고, 이에 대한 방어 기술로서 Defense-GAN과 MagNet을 탑재한 Perceptual Ad-Blocker를 제안하였다. 안전성 분석 결과에 따르면 두 분류 모델 Percival과 ResNet을 대상으로 공격 성공률 90%정도를 보였던 적대적 예제들을 Defense-GAN과 MagNet으로 공격 성공률을 상당히 낮출 수 있음을 보였다. Defense-GAN의 경우 적응적 공격에도 일정 수준의 방어가 가능했다. 또한 효율성 실험 결과에 따르면 방어 모듈은 MagNet을 탑재했을 때 공격이 적용되지 않은 이미지에 대한 분류 모델의 정확도에는 영향을 주지 않았고, Defense-GAN의 경우 2.5% 정도로 매우 미미한 감소를 보였다. 추가적인 시간 발생 역시 MagNet은 거의 발생하지 않았고 Defense-GAN은 사진 당 평균 2.5초 정도 걸리는 결과를 보였다. 추가적으로 공격자 입장에서 걸리는 시간 비교 실험에서도 Defense-GAN adaptive attack에 걸리는 시간이 다른 적대적 예제 생성 알고리즘들 보다 많음을 보였다.

MagNet의 경우 효율성 측면에서 이점이 있고, Defense-GAN의 경우 적응적 공격에도 견고하다는 점에서 안정성 측면에서의 이점이 있다. 본 논문에서는 MagNet과 Defense-GAN의 내부 구조를 거의 수정하지 않았으며, 모델의 구조 수정과 hyperparameter의 조정을 통해 보다 나은 성능을 이끌어 내고, 각각의 단점을 보완 가능한 방어 기술을 제안하였다. 이를 통해서 본 논문에서는 Defense-GAN과 MagNet에 기반하여 적대적 공격에 견고한 Perceptual Ad-Blocker의 가능성을 보였다.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

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

[3] F. Tramèr, P. Dupré, G. Rusak, G. Pellegrino, and D. Boneh, "Adversarial: perceptual ad blocking meets adversarial machine learning," Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 2005-2021, Nov. 2019.

[4] P. Samangouei, M. Kabkab and R. Chellappa, "Defense-gan: protecting classifiers against adversarial attacks using generative models," arXiv preprint arXiv:1805.06605, 2018.

[5] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp. 135-147, Oct. 2017.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, pp. 2672-2680, 2014.

[7] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong and A. Kovashka, "Automatic understanding of image and video advertisements," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1705-1715, 2017.

[8] Z.A. Din, P. Tigas, S.T. King and B. Livshits, "Percival: making in-browser perceptual ad blocking practical with deep learning," 2020 USENIX Annual Technical Conference, pp. 387-400, 2020.

[9] Kobaco, "Broadcasting and Communications Advertising Expenses Survey," https://adstat.kobaco.co.kr/sub/expenditure_data_search.do, Feb. 2020.

[10] A. Kurakin, I. Goodfellow and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, 2016.

[11] 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.

[12] S.M. 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, 2016.

[13] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and

pattern recognition, pp. 770-778, 2016.
〔14〕 A. Athalye, N. Carlini and D. Wagner, "Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples," International Conference on Machine Learning, pp. 274-283, Jul. 2018.
〔15〕 S. Qiu, Q. Liu, S. Zhou and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," Applied Sciences, vol. 9, no. 5, pp. 909, 2019.
〔16〕 N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," IEEE symposium on security and privacy, pp. 39-57, May 2017.
〔17〕 A. Krizhevsky and G. Hinton, Learning multiple layers of features from tiny images, University of Toronto, 2009.
〔18〕 Y. LeCun, "The mnist database of handwritten digits," http://yann. lecun. com/exdb/mnist/, 1998.
〔19〕 H. Xiao, K. Rasul and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.
〔20〕 M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein generative adversarial networks," Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 214-223, Aug. 2017.

〈저자소개〉

김 민 재 (Min-jae Kim) 학생회원
2016년 3월~현재: 고려대학교 컴퓨터학과 학사과정
〈관심분야〉 정보보호, 컴퓨터공학, 인공지능

김 보 민 (Bo-min Kim) 학생회원
2016년 3월~현재: 고려대학교 바이오의공학부 학사과정
〈관심분야〉 인공지능, 컴퓨터비전

허 준 범 (Junbeom Hur) 종신회원
2001년 2월: 고려대학교 컴퓨터공학 졸업
2005년 8월: 한국과학기술원 전산학 석사
2009년 8월: 한국과학기술원 전산학 박사
2009년 9월~2011년 8월: University of Illinois at Urbana-Champaign 박사후 연구원
2011년 9월~2015년 2월: 중앙대학교 컴퓨터공학부 조교수
2015년 3월~2016년 8월: 고려대학교 컴퓨터학과 조교수
2016년 9월~현재: 고려대학교 컴퓨터학과 부교수
〈관심분야〉 응용 암호, 네트워크 보안, 클라우드 보안, 시스템 취약점