Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.33, NO.4, Aug. 2023 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2023.33.4.621

Attention 기법에 기반한 적대적 공격의 강건성 향상 연구*

김 재 욱,[1†] 오 명 교,[1] 박 래 현,[1] 권 태 경[2‡]
[1,2]연세대학교 (대학원생, 교수)

Improving Adversarial Robustness via Attention*

Jaeuk Kim,[1†] Myung Gyo Oh,[1] Leo Hyun Park,[1] Taekyoung Kwon[2‡]
[1,2]Information Security/AI Security LAB, GSI, Yonsei University
(Graduate student, Professor)

요 약

적대적 학습은 적대적 샘플에 대한 딥러닝 모델의 강건성을 향상시킨다. 하지만 기존의 적대적 학습 기법은 입력 단계의 작은 섭동마저도 은닉층의 특징에 큰 변화를 일으킨다는 점을 간과하여 adversarial loss function에만 집중한다. 그 결과로 일반 샘플 또는 다른 공격 기법과 같이 학습되지 않은 다양한 상황에 대한 정확도가 감소한다. 이 문제를 해결하기 위해서는 특징 표현 능력을 향상시키는 모델 아키텍처에 대한 분석이 필요하다. 본 논문에서는 입력 이미지의 attention map을 생성하는 attention module을 일반 모델에 적용하고 PGD 적대적 학습을 수행한다. CIFAR-10 dataset에서의 제안된 기법은 네트워크 구조에 상관없이 적대적 학습을 수행한 일반 모델보다 적대적 샘플에 대해 더 높은 정확도를 보였다. 특히 우리의 접근법은 PGD, FGSM, BIM과 같은 다양한 공격과 더 강력한 adversary에 대해서도 더 강건했다. 나아가 우리는 attention map을 시각화함으로써 attention module이 적대적 샘플에 대해서도 정확한 클래스의 특징을 추출한다는 것을 확인했다.

ABSTRACT

Adversarial training improves the robustness of deep neural networks for adversarial examples. However, the previous adversarial training method focuses only on the adversarial loss function, ignoring that even a small perturbation of the input layer causes a significant change in the hidden layer features. Consequently, the accuracy of a defended model is reduced for various untrained situations such as clean samples or other attack techniques. Therefore, an architectural perspective is necessary to improve feature representation power to solve this problem. In this paper, we apply an attention module that generates an attention map of an input image to a general model and performs PGD adversarial training upon the augmented model. In our experiments on the CIFAR-10 dataset, the attention augmented model showed higher accuracy than the general model regardless of the network structure. In particular, the robust accuracy of our approach was consistently higher for various attacks such as PGD, FGSM, and BIM and more powerful adversaries. By visualizing the attention map, we further confirmed that the attention module extracts features of the correct class even for adversarial examples.

Keywords: Adversarial training, Attention, Adversarial robustness, Adversarial examples

Received(05. 11. 2023), Accepted(05. 31. 2023)
* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2023-00253901).

† 주저자, freak0wk@yonsei.ac.kr
‡ 교신저자, taekyoung@yonsei.ac.kr(Corresponding author)

I. 서 론

Deep neural network (DNN)은 자연어 처리, 이미지 인식 등 실생활에서 다양하게 활용되고 있다. 하지만 DNN과 인간의 의사 결정 방식의 차이가 존재함으로 실생활에서 사용되고 있는 자율 주행, 의료 진단, 악성코드 탐지 분야에서 모델의 오분류가 발생한다면 이는 치명적인 문제로 야기된다. 이에 대해 Goodfellow 등은 입력 데이터에 작은 섭동을 주입하여 모델을 속이는 적대적 공격을 소개하고 이에 대한 오분류 문제를 제시했다 [1]. 현재 적대적 공격 연구 분야에서 대표적인 공격 기법으로는 FGSM[1], PGD[2], BIM[3]이 있다. 필요한 정보보호 서비스가 수반되지 않는 상태에서 지식정보화 사회의 발전은 많은 사회적 부작용을 동반할 수밖에 없다.

적대적 공격으로 부터 모델의 안전성을 제공하기 위해 다양한 기법이 연구되었다. 적대적 방어 기법은 크게 3가지로 분류된다. 대상 모델에 샘플을 입력하기 전에 샘플에 대해 섭동을 제거하는 적대적 디노이징 기법과 적대적 샘플이 입력으로 들어왔을 때 해당 샘플을 거부하는 기법인 적대적 탐지 기법이 있다. 두개의 기법은 모델 이전에 새로운 아키텍처를 추가해야 하므로 오버헤드가 발생한다. 새로운 아키텍처의 추가 없이 적대적 공격 기법으로 생성된 샘플을 모델의 훈련 세트에 주입하여 모델을 훈련시키는 방법인 적대적 학습 기법은 현재까지 다른 두개의 기법보다 좋은 성능을 보여주고 있다.

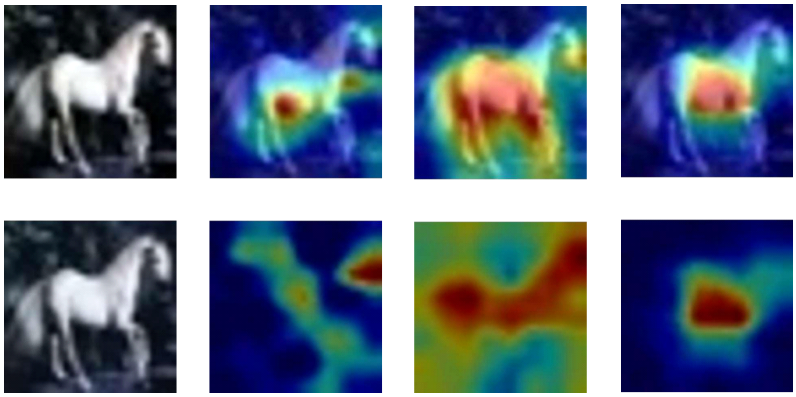
적대적 학습의 발전과 함께 computer vision에서 모델의 이미지 분류 정확도를 높이기 위해 모델의 아키텍처 수정과 데이터세트 관점에서의 연구 또한 활발히 진행되고 있다. 일반 샘플과 적대적 샘플에 대한 모델의 정확도를 높이기 위해 입력 이미지의 특징 추출 능력이 중요하다. 하지만 모델이 기존의 방식처럼 convolution layer에 크게 의존하면 다양한 이미지 영역에 대한 long-term dependency에 대한 학습을 방해하게 된다. 결국 capacity가 작은 모델은 많은 이미지에 대해 표현을 할 수 없고 이전에 입력되지 않은 이미지에 대해 잘못된 결과가 도출될 수 있다. convolution layer를 추가하게 되면 모델의 표현 능력은 증가할 수 있지만 엄청난 계산량의 증가로 인해 효율성이 떨어진다.

반면에 attention[4, 5]은 이미지의 모든 영역의 가중치를 효율적으로 계산함으로써 long-term dependency 문제를 해결할 수 있다. Attention module이 적용된 비전 모델은 오버헤드를 줄이고 다른 모델보다 더 나은 성능을 보인다 [6]. 특히 특징에서 channel의 정보만 표현하는 channel attention module과 공간적 관계를 표현하는 spatial attention module을 순차적으로 convolution layer에 적용하여 적은 오버헤드로도 유의미한 특징을 추출할 수 있다 [7]. 하지만 일반 샘플에 대한 attention의 효과는 알려졌으나 적대적 샘플에 대해서도 효과적인지 아직 확인되지 않았다. 이 점에서 우리는 하나의 질문을 제기한다: **"Attention이 적대적 샘플에 대한 long-term dependency를 해결하고 모델의 강건성을 개선하는가?"**

이 질문에 답하기 위해 본 논문에서는 모델의 특징 추출 능력을 높이는 attention module을 추가한 모델에 적대적 학습을 수행한 attention을 활용한 적대적 학습(ATVA, adversarial training via attention)를 제안한다. Attention이 적용되지 않은 일반 모델에 적대적 학습을 수행한 기법과 비교한 ATVA의 효과를 검증하기 위해 3가지 연구 질문을 수립한다. 첫 번째로 white-box, black-box, cross attack 등의 다양한 공격 시나리오를 구성하여 ATVA의 정확도를 측정한다. 두 번째로 강력한 공격으로부터의 ATVA의 성능을 확인한다. 마지막으로 attention map을 시각화함으로써 attention module의 특징 추출 능력을 확인한다. 검증 결과는 일반적으로 일반 모델보다 ATVA에서 일반 샘플 및 적대적 샘플에서 높은 accuracy를 보여주었으며 시각화에서 일반 모델보다 효과적인 특징 추출 능력을 보여주었다.

본 논문의 주요 기여는 다음과 같이 요약할 수 있다.

- 적대적 학습 기법의 long-term dependency 문제를 해결하기 위해 attention module을 추가하였다. 결국 이미지의 유의미한 특징 추출을 통해 적대적 공격에 대한 강건한 모델을 제공한다.
- 다양한 측면에서 제안된 기법의 성능 평가를 수행하였다. 실험을 통해 대상 모델과 공격 강도에 관련한 다양한 시나리오에서 ATVA가 더 우수함을 보였다. 또한 시각화를 통해 특징 추출 성능이 더 우수함을 확인했다.



(a) Image (b) NM (c) Adv+NM (d) AM

Fig. 1. Activation map of ResNet-20 models for clean and adversarial samples through Grad-CAM. The upper and lower images in (a) are a clean image and an PGD adversarial sample, respectively. (b) is the results of normal model (NM), (c) is the results of an adversarially trained normal model (d) is the results of the normally trained attention model (AM). The upper images of (b), (c), and (d) are the results for the clean image. The lower images of (b), (c), and (d) are the results for the adversarial sample.

II. Background and Motivation

2.1 Threat Model

Adversarial Attack. 본 논문에서는 모델의 강건성 평가 측정 방식으로 널리 사용되고 있는 FGSM, PGD, BIM을 선정하여 제안한 기법의 정확도를 측정한다.

Fast Gradient Sign Method (FGSM)[1] 공격은 아래의 공식을 통해 적대적 샘플을 생성한다.

$$x' = x + \epsilon \cdot sign(\nabla L(x,y,\theta)) \quad (1)$$

위의 식에서 x, x' 각각 원본 샘플, 적대적 샘플을 말하고 y, θ는 ground-truth label, model parameter를 뜻한다. $\nabla L(x,y,\theta)$는 loss function L에 대한 gradient(e.g. cross-entropy loss of softmax outputs)를 의미한다. 사람의 눈으로 인지하기 어려운 적대적 샘플을 생성하기 위해 FGSM은 입력에 대한 손실함수의 gradient의 부호에 섭동 크기를 조절하는 hyper-parameter ϵ 조정하여 공격의 강도를 설정할 수 있다.

Basic Iterative Method (BIM) [3]은 FGSM 보다 공격 성공률을 높이기 위해 FGSM에 iteration을 추가하여 좀 더 강력한 공격 샘플을 생성한다. 대신 각 단계의 섭동은 최대 iteration 횟수 T가 주어질 때 step size $\alpha = \epsilon / T$로 제한한다. BIM는 아래와 같이 공식화된다.

$$x_{t+1} = Clip_{[-\epsilon,\epsilon]}\{x_t + \alpha \cdot sign(\nabla L(x_t,y,\theta))\}$$
$$x_0 = x \quad (2)$$

Projected Gradient Descent (PGD) [2] 또한 FGSM에서 발전된 방법으로 step만큼 공격을 반복하여 정해진 ϵ범위에서 내부 최대화를 수행한다. BIM과 PGD의 차이는 PGD에서는 랜덤 포인트로 시작점을 초기화 한다는 것이며 아래와 같이 공식화된다. 현재 PGD 공격은 적대적 공격과 방어 측면에서 모두 효과적인 기법 중 하나이다.

$$x^{t+1} = \Pi_{x+S}(x^t + \alpha \cdot sign(\nabla_x L(x_t,y,\theta))) \quad (3)$$

Adversarial Defense. 적대적 방어 기법 중 적대적 학습은 모델 오분류를 유발하는 공격 샘플의 유입이 되기 전 강력한 모델을 생성하여 모델의 신뢰성을 제공하는 것을 목표로 한다. 적대적 학습은 [1]로 거슬러 올라갈 수 있으며, 적대적 공격에 보안개념을 활용한 안장점 공식을 본 논문에서 활용한다. Madry 등이 [2] 제시한 안장점 공식은 아래와 같이 내부 최대화 및 외부 최소화로 구성되어 있다.

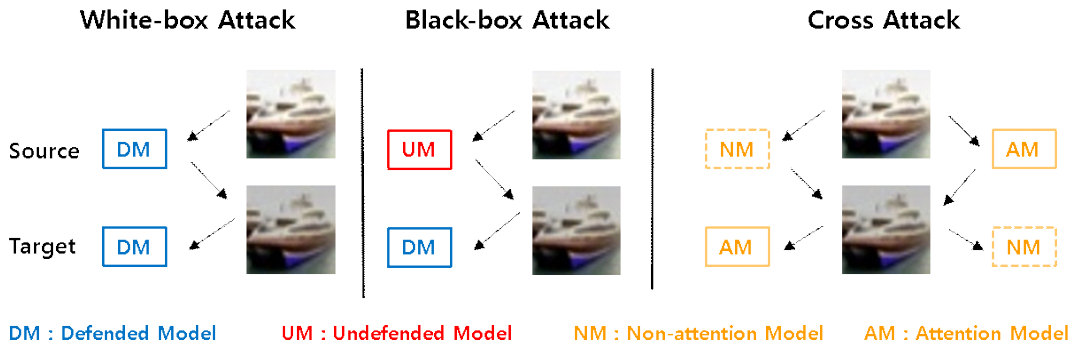


Fig. 2. Three attack scenarios in research question 1. The source is a model used to generate adversarial samples and the target is a model to inject the samples.

$$\min_{\theta}\rho(\theta)\ where\ \rho(\theta)= E_{(x,y)\sim D}[\max_{\delta\in S}L(x,\theta+\delta,y)] \quad (4)$$

PGD 공격 샘플을 생성해 내부 최대화을 달성한다. 그리고 PGD 공격 샘플을 모델의 학습단계에 주입하여 외부 최소화를 달성하는 매개변수 θ를 찾아 다른 공격에 있어서도 신뢰성 있는 모델을 만들 수 있다. Kurakin 등은 적대적 학습에 의해 달성되는 모델의 견고성은 학습에 사용된 공격 샘플의 강도에 따라 달라지며 FGSM과 같은 빠른 비반복 공격에 대한 적대적 학습은 비반복 공격에 대한 견고성만 가져온다고 언급하였다 [3]. 이와 달리, Madry는 실험을 통해 MNIST 및 CIFAR-10 데이터 세트에서 반복 및 비반복 적대적 공격에 대한 높은 수준의 모델의 견고성을 보여주었다.

2.2 Long-Term Dependency Problem

Fig 1.은 Grad-CAM(Gradient-weighted Class Activation Mapping) [8]을 통해 일반 샘플과 적대적 샘플에 대한 각 모델의 activation map을 보여준다. Fig 1.의 (b)와 (c)에서 정상 모델의 activation map이 좁고 이미지의 잘못된 부분에 초점을 맞추고 있음을 보여준다. 이는 적대적 샘플에서도 동일한 결과를 확인할 수 있다. 이를 통해, 일반모델과 적대적 훈련만으로는 long-term dependency 문제를 해결할 수 없음을 보여준다.

이를 해결하기 위해 본 논문에서는 Woo 등이 제안한 CBAM (Convolutional Block Attention Module)을 활용하여 모델을 구성한다 [7]. CBAM은 입력 이미지에서 '무엇'이 의미가 있는지 초점을 맞추는 channel attention과 '어디'에 초점을 맞추는 spatial attention을 순차적으로 수행하여 attention map을 생성하는 attention module을 제안했다. Channel attention은 특징의 channel 간의 관계를 활용하여 channel attention map을 생성하고 효율적으로 계산하기 위해 입력 feature map의 공간 차원을 압축한다. 이때 max-pooling과 average-pooling을 동시에 사용하며 오버헤드를 줄이기 위해 하이퍼 파라미터로 channel을 나누어 주어 차원을 축소 시킨다. Spatial attention은 channel axis를 따라 average-pooling과 max-pooling 적용하고 압축된 channel에 7x7 convolution layer를 적용하여 spatial attention map을 생성한다. 결국 CBAM는 channel attention과 spatial attention을 순차적으로 수행하여 attention map을 생성하는 attention module을 convolution block 사이에 위치시켜 유의미한 특징을 생성한다.

III. Study Design

본 논문에서는 적대적 공격으로 부터의 강력한 모델을 구성하기 위해 attention module이 추가된 ATVA을 제안한다. Section 3.1에서는 ATVA의 성능을 평가하기 위한 3가지의 연구 질문을 구성한다. Section 3.2에서는 ATVA에 대한 공격 및 방어 환경의 세부 사항을 소개한다.

Table 1. Robust accuracy of adversarially trained models on attack scenarios. The values in bracket are losses of adversarial samples. Bold indicates the highest accuracy for each model in a block. Underline indicates the highest accuracy for each block.

Model	Attention	White-box			Black-box		
		FGSM	BIM	PGD	FGSM	BIM	PGD
ResNet-20	O	**<u>0.3912</u>**	**0.3662**	**0.3607**	**0.6473**	**0.6506**	**<u>0.6519</u>**
		(1.7084)	(1.7861)	(1.7652)	(1.2070)	(1.2042)	(1.2054)
	X	0.3773	0.3507	0.3543	0.6091	0.6138	0.6136
		(1.7839)	(1.8221)	(1.8242)	(1.2844)	(1.2815)	(1.2810)
ResNet-56	O	**<u>0.4359</u>**	**0.4010**	**0.4050**	**0.6866**	**<u>0.6869</u>**	**0.6863**
		(1.7039)	(1.7918)	(1.7686)	(1.1386)	(1.1393)	(1.1381)
	X	0.4095	0.3831	0.3864	0.6553	0.6571	0.6559
		(1.7589)	(1.8251)	(1.8053)	(1.2093)	(1.2069)	(1.2065)
DenseNet-40	O	0.4066	0.3792	**0.3830**	**0.6676**	**<u>0.6690</u>**	**0.6700**
		(1.5771)	(1.6425)	(1.6235)	(1.0602)	(1.0577)	(1.0569)
	X	**<u>0.4109</u>**	**0.3835**	0.3566	0.6435	0.6424	0.6429
		(1.6354)	(1.7111)	(1.6879)	(1.0338)	(1.0349)	(1.0337)

3.1 Research Questions

RQ 1. ATVA는 일반모델보다 다양한 시나리오의 공격으로부터 강건성를 제공하는가?

다양한 적대적 공격에 따라 모델의 분류 성능이 달라질 수 있다. 다양하게 변환되는 공격 샘플에 대한 결과를 확인하기 위해 공격 시나리오를 Fig 2.와 같이 총 3가지 white-box attack, black-box attack, cross attack으로 구성하였다. Worst-case scenario에 대한 모델의 성능을 확인하기 위해 white-box attack을 수행한다. white-box attack은 공격 대상 모델의 정보를 알고 있는 상태에서 공격을 수행하기 때문에 적대적 학습 된 모델을 source 모델과 target model로 선정한다. 즉, 동일한 모델에서 적대적 샘플을 생성하고 공격을 수행하며 Fig 2.의 왼쪽을 통해 확인할 수 있다. 또한 현실에서 발생할 수 있는 scenario인 black-box attack을 수행하였다. Black-box attack은 공격 대상 모델의 정보를 모르고 있는 상태에서 공격을 수행하기 때문에 적대적 학습 안된 source model에서 적대적 샘플을 생성하고 적대적 학습된 target model에 공격을 수행하며 Fig 2.의 가운데를 통해 확인할 수 있다. Attention module 기법이 적용된 모델의 적대적 샘플이 강력한지 확인하기 위해 Cross attack을 수행한다. Attention module이 적용안된 model에서 생성된 적대적 샘플을 attention model에 주입하고 attention model에서 생성된 적대적 샘플을 non-attention model에 주입하여 공격 성공률을 확인한다. Cross attack에 대해서도 white-box, black-box 공격을 수행하였다. Cross attack은 Fig 3.의 오른쪽을 통해 확인할 수 있다. 각 공격에서의 eps는 $8/255$로 설정하였다.

RQ 2. ATVA는 강력한 적대적 공격으로부터 일반모델보다 강건성을 제공하는가?

이미지 섭동의 변화를 높게 설정한다면 모델은 더 많은 오분류가 발생할 것이다. 이에 따라 강한 공격으로부터의 모델의 정확도를 측정하였다. 섭동의 강도는 각 공격의 파라미터 설정을 통해 변화를 줄 수 있다. 파라미터 범위는 $0 \leq \epsilon \leq 32/255$로 설정하여 공격을 수행한다. 적대적 공격으로는 FGSM, BIM, PGD를 선정하고 대상 모델은 ATVA와 attention module이 적용안된 두 모델을 선정하여 강한 공격에 따른 성능을 비교한다.

RQ 3. 일반모델보다 attention 기법이 적용된 모델은 특징 추출 성능이 우수한가?

Table 2. Robust accuracy of adversarially trained models on cross attack scenarios. The values in bracket are losses of adversarial samples. Bold indicates the highest accuracy for each model in a block. Underline indicates the highest accuracy for each block.

Model	Attention	White-box			Black-box		
		FGSM	BIM	PGD	FGSM	BIM	PGD
ResNet-20	O	**0.4914**	**0.4858**	**0.4898**	**0.6473**	**0.6506**	**0.6511**
		(1.5116)	(1.5245)	(1.5157)	(1.2059)	(1.2030)	(1.2028)
	X	0.4792	0.4697	0.4746	0.6083	0.6119	0.6133
		(1.5709)	(1.5870)	(1.5779)	(1.2849)	(1.2834)	(1.2820)
ResNet-56	O	**0.5324**	**0.5251**	**0.5299**	**0.6874**	**0.6893**	**0.6907**
		(1.4751)	(1.4920)	(1.4822)	(1.1378)	(1.1356)	(1.1346)
	X	0.5062	0.4980	0.5018	0.6550	0.6555	0.6558
		(1.5321)	(1.5500)	(1.5402)	(1.2101)	(1.2102)	(1.2093)
DenseNet-40	O	0.4983	0.4907	**0.4953**	**0.6683**	**0.6681**	**0.6691**
		(1.3742)	(1.3878)	(1.3787)	(1.0580)	(1.0575)	(1.0570)
	X	**0.5062**	**0.4972**	0.5009	0.6396	0.6419	0.6438
		(1.3779)	(1.3994)	(1.3887)	(1.0413)	(1.0345)	(1.0327)

이미지의 오분류를 방지하고 모델의 높은 정확도를 도출하기 위해서는 이미지의 우수한 특징 추출 능력이 요구된다. 이미지에서 향상된 특징 추출 성능을 확인하기 위해 grad-cam (Gradient-weighted Class Activation Mapping)을 통한 activation map을 확인한다. Grad-cam은 Selvaraju 등이 제안했으며 CNN에 적용할 수 있는 이미지 해석 방법론이다 [8]. 다시 말해 CNN의 gradient를 가지고 이미지의 어느 부분에 가중치를 주는지 계산하는 방법이며 추가적인 구조 변경과 재훈련 없이 바로 모델에 적용이 가능하다.

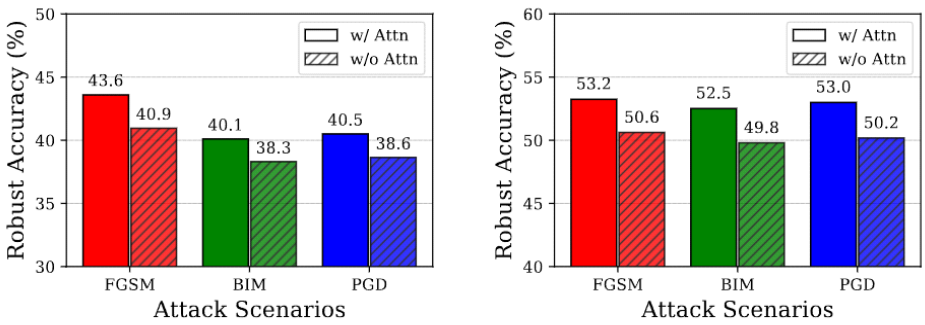
3.2 Implementation

Madry 등이 PGD를 적용하여 안장점 공식을 성공적으로 해결할 수 있음을 보여주었기 때문에 본 논문에서도 PGD 적대적 학습 기법을 활용한다. CIFAR-10의 train set으로 생성된 PGD 샘플을 모델의 학습 단계에 입력하여 안장점 공식을 안정적으로 최적화하고 강력한 분류기를 생성한다. PGD 학습 샘플 생성시 *eps*: *8/255*, *eps_step*: *2/555*, *max_iter*: *7*로 파라미터를 설정한다. 대상 모델은 ResNet-20, ResNet-56, DenseNet 총 3가지로 선정하였다. 모델 용량에 따른 정확도를 확인하기 위해 같은 아키텍처를 사용하는 ResNet-20, ResNet-56을 선정하였으며 다른 아키텍처로 구성된 모델에서도 정확도을 측정하기 위해 DenseNet을 추가하였다. 각 모델들은 convolution block 사이에 attention model을 추가된 모델과 그렇지 않은 모델로 구분되어 실험을 진행한다. 생성된 분류기에 대해 정확도을 측정하기 위해 테스트 단계는 CIFAR-10의 테스트 세트로 생성된 FGSM, BIM, PGD 공격샘플을 모델에 입력하여 각 공격으로부터의 정확도를 측정한다.

IV. Evaluation

4.1 Experiment Settings

본 논문에서는 적대적 공격에 대한 정확도를 측정하기 위해 데이터세트는 오픈소스로 제공되고 있으며 여러 딥러닝 연구 분야에서 활용 되고 있는 CIFAR-10을 사용하였다. CIFAR-10은 10개의 클래스와 각 클래스당 6000개의 이미지로 구성되어 있다. 총 60,000개의 이미지는 32x32x3의 차원수를 가진 RGB 컬러 이미지이며 각 클래스당 5000개의 학습 세트와 1000개의 테스트 세트로 구성된다. 모든 실험은 colab pro+에서 수행하였다.



(a) Attack Scenarios (b) White-box cross attack

Fig. 3. Robust accuracy of adversarially trained ResNet-56 models with and without attention module to (a) white-box attack and (b) white-box cross attack.

4.2 RQ 1. 다양한 공격 시나리오부터의 강건성

Table 1.는 PGD 적대적 학습을 수행하여 attention module이 추가된 모델과 일반 모델에서 PGD, FGSM, BIM의 white-box attack과 black-box attack으로부터 정확도를 보여준다. 거의 모든 공격에서 attention module 추가된 모델은 일반 모델보다 높은 정확도를 보여주고 있고 손실 또한 일반 모델보다 낮은 것을 확인할 수 있다. 또한 attention module이 추가된 모델과 일반 모델의 정확도 차이는 black-box attack에서 더 월등히 나타난다는 것을 확인할 수 있다. 이는 일반 모델은 모델의 파라미터를 모르고 있을 때 수행되는 black-box attack에서도 취약하다는 것으로 볼 수 있다. 또한 cross attack에 대한 실험 결과를 보여주는 Table 2.에서 attention module 추가된 모델은 일반 모델에서 생성된 적대적 샘플에 대해서도 일반 모델보다 정확도이 높으며 반대로 attention module 추가된 모델에서 생성된 적대적 샘플에 대해 일반모델은 취약한 것을 보여준다. 이는 attention module 추가된 모델이 일반 모델보다 더 강력한 샘플을 생성하는 것으로 볼 수 있다. Fig 3.은 attention 유무에 대해 가장 뚜렷한 정확도 차이가 있던 ResNet-56의 Whtie-box 실험 결과를 보여준다. 이를 통해 ATVA 정확도는 평균적으로 2% 높은 것을 확인할 수 있다. 각 Table 1. 2. 의 white-box 시나리오에서 DenseNet-40의 경우, 일반 모델이 ATVA보다 약간 더 높은 정확도를 보여준다. 하지만 그 차이는 0.008 미만에 불과하며 ATVA가 손실이 낮은 것을 확인할 수 있다. 이는 오분류 클래스에 대한 ATVA 모델의 신뢰도가 낮기 때문이다. 이는, ATVA가 일반 모델보다 결정 경계가 더 부드럽다는 것을 시사한다.

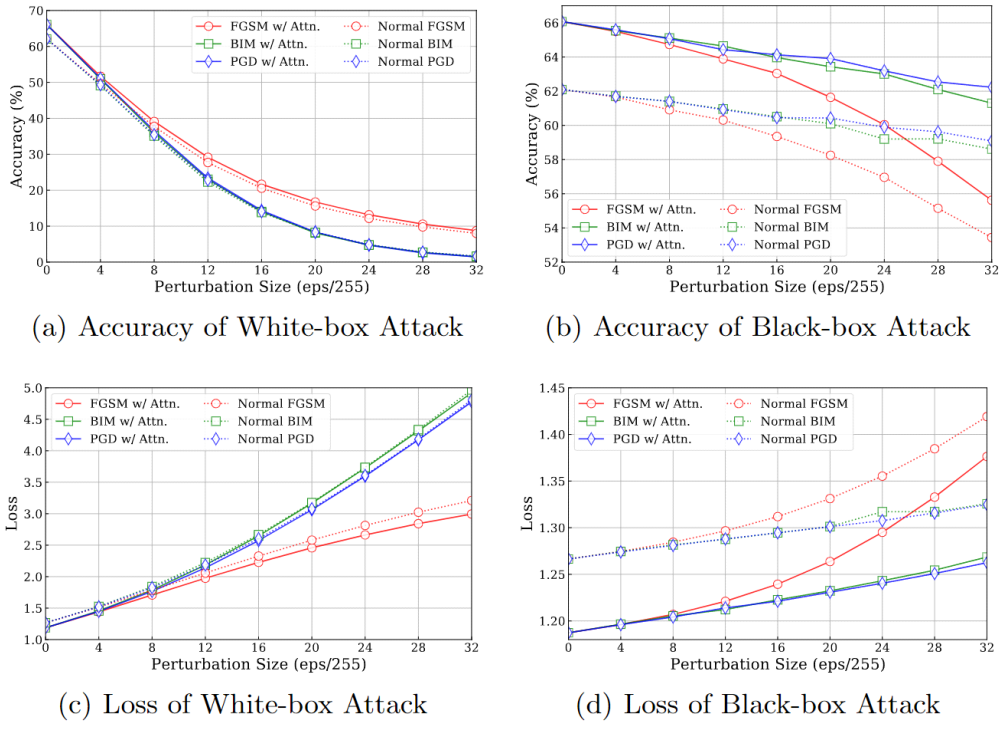
4.3 RQ 2. 강력한 공격으로부터의 강건성

강력한 공격으로부터의 ATVA 및 일반 모델에 대한 방어 성능을 비교하기 위해 실험을 진행하였다. Fig 4.를 통해 ResNet-20에서 FGSM, BIM, PGD 공격에 대한 결과를 확인할 수 있다. 섭동 크기를 $0 \leq \epsilon \leq 32/255$, step size를 $0 \leq \alpha \leq 8/255$로 변경했다. Fig 4(b).에서 볼수 있듯이 ATVA는 일반 모델보다 black-box attack에 대해 더 높은 정확도를 보여준다. White-box attack에서도 대부분 ATVA에서 $\epsilon \leq 20/255$에서 일반 모델보다 더 높은 정확도를 보여준다. 하지만 $\epsilon > 20/255$를 기점으로 일반 모델이 ATVA보다 약간 더 높은 정확도를 보여준다. 그럼에도 불구하고 그 차이는 약 0.1% 불과하며 심지어 손실에서 ATVA가 더 낮음을 확인할 수 있다.

ATVA를 통한 적대적 학습 이후 일반샘플에 대한 테스트 정확도는 ResNet-20에서 63%, ResNet-56에서 68%, DenseNet-40에서 68%로 일반 모델 대상의 적대적 학습과 큰 차이를 보이지 않았다. 적대적 학습의 고유한 한계점인 일반 샘플과 적대적 샘플간 정확도 트레이드오프가 존재하므로 향후 연구에서는 이를 해결하기 위한 기법이 ATVA에 적용되어야 한다.

4.4 RQ 3. Attention module의 효율성

일반 모델보다 ATVA에서 적대적 샘플의 특징 추



(a) Accuracy of White-box Attack (b) Accuracy of Black-box Attack

(c) Loss of White-box Attack (d) Loss of Black-box Attack

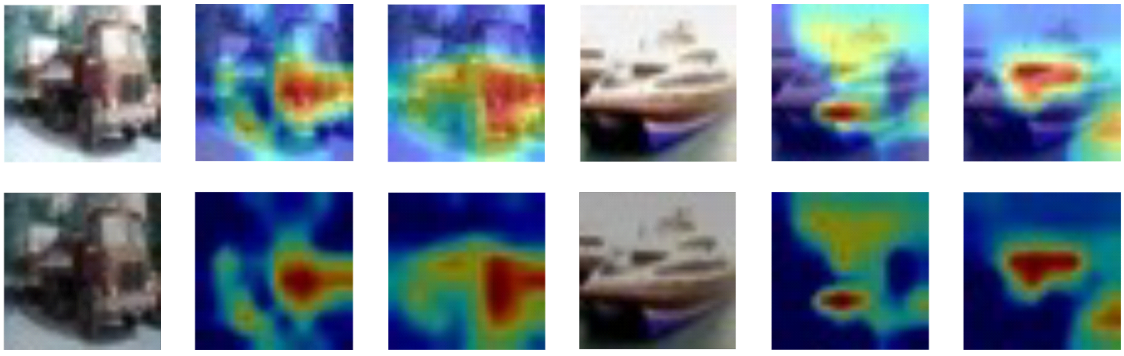
Fig. 4. Robustness of adversarially trained models on various attack strength. (a) and (b) are robust accuracy. (c) and (d) are loss of adversarial samples.

출 성능을 직관적으로 확인하기 위해 grad-cam을 사용하여 모델의 activation map을 확인하였다. Fig 5(b).를 통해 모델이 해당 객체에 대해 집중도가 낮음을 보여주고 있으며 Fig 5(e).에서 모델이 의사결정에 불필요한 부분까지 집중하고 있음을 보여준다. 이와 달리, Fig 5(c). Fig 5(f)를 통해 attention module이 포함된 모델이 더 넓은 범위의 activation map을 보여주는 것을 확인할 수 있으며 더 정확하게 객체를 타겟팅 한다는 것을 보여준다. 요약하면, attention module은 일반 샘플과 적대적인 샘플에서 객체의 의미 있는 특징을 추출하여 모델 분류 정확도를 향상시킨다.

V. Related Work

적대적 공격에 대한 모델의 오분류를 방지하기 위해 적대적 방어가 존재하며 적대적 공격과 발맞춰 활발히 연구가 진행되고 있다. 적대적 방어 기법은 적대적 학습 [1, 2], 적대적 디노이징 [9, 10], 적대적 탐지 [11, 12] 기법이 존재한다.

Adversarial denoising 기법은 대상 모델에 입력하기 전에 입력 샘플에 대해서 섭동을 제거하는 기법이다. 디노이징 기법은 입력 샘플이 모델에 입력되기 전에 섭동을 제거해야하기 때문에 대상 모델 이전에 새로운 아키텍처를 추가하는 오버헤드가 발생한다. 디노이징 기법의 대표적인 기법으로 Defense-GAN [13]과 Feature denoising [14] 기법이 있다. Defense-GAN은 WGAN을 사용하였으며 크게 3가지 단계 initialization, generation, classification으로 구분된다. Initialization, generation 단계에서는 입력 이미지로 적대적 샘플을 주입하여 적대적 샘플과 가장 유사한 이미지인 random vector를 찾고 generator는 섭동이 제거된 이미지를 생성한다. 이후 생성된 이미지를 분류 모델에 전달하여 분류 단계를 수행한다. 하지만 Defense-GAN은 GAN 모델의 성능에 많은 영향을 받는다. 이로 인해 GAN을 학습하는데 hyper parameter 조정 등 많은 변수가 존재하므로 모델 학습이 불안정하다는 문제점이 존재한다.



(a) Image 1 (b) Adv+NM (c) Adv+AM (d) Image 2 (e) Adv+NM (f) Adv+AM

Fig. 5. Activation map through Grad-CAM of the model with or without attention module after adversarial training. The upper images in (a) and (d) indicate clean samples. The lower images in (a) and (d) indicate PGD adversarial samples. (b) and (e) are the attention maps of the normal model (NM) for each image. (c) and (f) are the attention maps of the attention model (AM) for each image. Upper maps show results for clean samples, and lower maps show results for adversarial samples.

Adversarial detection 기법은 detector를 대상 모델 앞에 배치하여 적대적 샘플이 입력으로 들어왔을 때 해당 샘플을 거부하는 기법이다. Meng and Chen은 입력 샘플을 적대적 또는 깨끗한 샘플로 분류하기 위해 detector 및 denoiser를 사용하는 프레임워크를 제안했다 [15]. 훈련하는 동안 프레임워크는 다양한 깨끗한 샘플을 학습하고 테스트 단계에서 학습된 샘플의 매니폴드에서 멀리 떨어진 샘플은 적대적 샘플로 취급되며 거부된다. manifold에는 가까운 샘플은 manifold에 놓이도록 재구성되고 모델은 재구성된 샘플을 받아 입력 데이터로 사용된다. 하지만 MagNet은 깨끗한 샘플에 대해서만 학습을 진행하기 때문에 제한적인 manifold가 형성된다. 제한적인 manifold로 인해 이후 연구 Carlini, Wagner는 해당 방어 기법이 작은 섭동으로도 적대적 예제에 대해 견고하지 않다는 것을 보여주었다 [16].

Adversarial training 기법은 적대적 공격 기법으로 생성된 샘플을 모델의 훈련 세트에 주입하여 모델을 재훈련 시키는 방법이다. 적대적 공격 샘플을 간편하게 재학습을 수행하여 적대적 방어에 대한 강인함을 보여주기 때문에 현재 적대적 방어 기법에서 널리 사용되고 있는 기법이다. Goodfellow 등은 모델의 gradient를 사용하여 적대적 샘플을 생성하는 을 제안하였고 FGSM을 통해 생성된 샘플을 학습하는 기법을 제안하였다 [1]. 하지만 이후 Madry 등은 FGSM을 통해 재학습된 모델은 매우 제한적인 적대적 샘플에서 학습하기 때문에 해당 적대적 샘플에 대해서만 과적합되는 것을 보여주었다 [2]. 이를 해결하기 위해 무작위 재시작을 적용하고 여러 번의 FGSM을 수행하여 공격성을 높인 PGD 기법을 제안했다. PGD 기법은 현재까지도 적대적 훈련에 대해 기초가 되는 기법으로 사용되고 있다. 또한 Kurakin 등은 모델이 FGSM 적대적 샘플에 대해 훈련되면 적대적 이미지의 정확도가 깨끗한 이미지의 정확도보다 훨씬 높아지는 현상을 보여주었으며 이를 label leaking이라고 하였다 [3]. 적대적 샘플을 기반으로 모델을 훈련시키는 적대적 학습 기법은 강력한 공격에 대해서 제안된 다른 기법들 보다 높은 정확도를 보여주는 기법이다. 해당 논문에서는 적대적 디노이징, 탐지 기법과 달리 추가적인 아키텍처를 구성하지 않는 적대적 학습 기법을 활용한다. 공격 기법을 통해 생성된 적대적 샘플을 제안된 모델에 훈련을 수행하여 공격에 대한 높은 정확도를 보여준다.

VI. 결 론

본 논문에서는 적대적 샘플의 long-term dependency에 대한 기존 방어 기법의 한계를 지적했으며 Attention module이 있는 모델에 적대적 학습을 적용하는 ATVA를 제안했다. 평가 대상 모델은 ResNet-20, ResNet-56, DenseNet-40을 대상으로 선정하였고 FGSM, BIM, PGD 공격을 수행했다. White-box attack, black-box attack, cross attack, 섭동 크기가 큰 공격 등

다양한 공격 시나리오에서 적대적으로 학습된 attention 모델이 일반 모델보다 더 강력하다는 것을 확인했다. Attention module이 일반 모델보다 특징 추출 능력이 뛰어나며, 이는 모델의 정확도에 큰 영향을 미친다는 것을 Grad-CAM을 통해 확인했다. 향후 연구에서는 보다 다양한 타깃 모델 아키텍처를 추가하고 공격 샘플에 대한 논리적인 분석을 진행할 예정이다. 마지막으로, 본 논문의 결과가 딥러닝 모델의 견고성에 대한 추가 연구로 이어질 수 있기를 기대한다.

References

[1] Goodfellow, Ian, et al. "Explaining and harnessing adversarial examples." In Proceedings of the International Conference on Learning Representations. 2015.

[2] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." In Proceedings of the International Conference on Learning Representations. 2018.

[3] Kurakin, Alexey, et al. "Adversarial machine learning at scale." In Proceedings of the International Conference on Learning Representations. 2017.

[4] Parikh, Ankur P., et al. "A decomposable attention model for natural language inference." In Proceedings of the Empirical Methods in Natural Language Processing. 2016.

[5] Vaswani, Ashish, et al. "Attention is all you need." In Proceedings of the Neural Information Processing System. 2017.

[6] Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." In Proceedings of the Neural Information Processing System. 2019.

[7] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." In Proceedings of the European conference on computer vision. 2018.

[8] Selvaraju, Ramprasaath R., et al. "Grad- cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision. 2017.

[9] Liao, Fangzhou, et al. "Defense against adversarial attacks using high-level representation guided denoiser." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2018.

[10] Cho, Seungju, et al. "Dapas: Denoising autoencoder to prevent adversarial attack in semantic segmentation." In Proceedings of the International Joint Conference on Neural Networks. 2020.

[11] Li, Xin, and Fuxin Li. "Adversarial examples detection in deep networks with convolutional filter statistics." In Proceedings of the IEEE international conference on computer vision. 2017.

[12] Ma, Xingjun, et al. "Characterizing adversarial subspaces using local intrinsic dimensionality." arXiv preprint arXiv:1801.02613. 2018.

[13] Samangouei, Pouya, et al. "Defense-gan: Protecting classifiers against adversarial attacks using generative models." In Proceedings of the International Conference on Learning Representations. .2018.

[14] Xie, Cihang, et al. "Feature denoising for improving adversarial robustness." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[15] Meng, Dongyu, and Hao Chen. "MagNet: a two-pronged defense against adversarial examples." In Proceedings of the ACM SIGSAC

Conference on Computer and Communications Security. 2017.

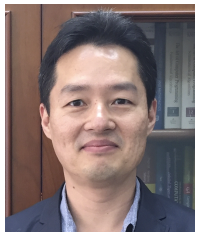
〔16〕 Carlini, Nicholas, and David Wagner. "Magnet and 'efficient defensesagainst adversarial attacks' are not robust to adversarial examples." arXiv preprint arXiv:1711.08478. 2017.

〈저자소개〉

김 재 욱 (Jaeuk Kim) 학생회원
2018년 2월: 세명대학교 정보통신학부 졸업
2018년 6월~현재: 연세대학교 정보대학원 석사과정
〈관심분야〉 정보보호, 딥페이크, 기계학습, Adversarial Machine Learning 등

오 명 교 (Myung Gyo Oh) 학생회원
2021년 2월: 세종대학교 정보보호학과 졸업
2021년 3월~현재: 연세대학교 정보대학원 석사과정
〈관심분야〉 정보보호, 언어 모델, 기계학습, 훈련 데이터 추출 공격, 프라이버시 등

박 래 현 (Leo Hyun Park) 학생회원
2017년 2월: 광운대학교 컴퓨터공학과 졸업
2017년 3월~현재: 연세대학교 정보대학원 석박사통합과정
〈관심분야〉 Adversarial Machine Learning, 딥러닝 모델 검증, 딥페이크 등



권 태 경 (Taekyoung Kwon) 종신회원
1992년 2월: 연세대학교 컴퓨터과학과 학사
1995년 2월: 연세대학교 컴퓨터과학과 석사
1999년 8월: 연세대학교 컴퓨터과학과 박사
1999년~2000년: U.C. Berkely Post-Doc
2001년~2013년 8월: 세종대학교 컴퓨터공학과 교수
2007년~2008년: Univ. Maryland at College Park 교환교수
2013년 9월~현재: 연세대학교 정보대학원 교수
〈관심분야〉 암호프로토콜, Usable Security, 소프트웨어/시스템보안, 기계학습과보안등