L_p Norm 기반 적대적 예제 공격에 대한 적대적 학습법의 성능 분석

박상리, 소정민
서강대학교
lp0523@sogang.ac.kr, jso1@sogang.ac.kr

On Efficacy of Adversarial Training Against L_p-norm Based Adversarial Example Attacks

Sanglee Park, Jungmin So
Sogang Univ.

요 약

본 논문에서는 딥러닝 모델의 오작동을 유발하는 적대적 예제(Adversarial example)를 생성하여, 이를 이용한 학습을 통해 L_p norm 공격에 대해 강건한 모델을 만들고 다양한 적대적 예제 공격에 대한 성능을 분석하였다. 먼저 L_∞, L_2 그리고 L_0 norm을 기반으로 하는 적대적 공격(Adversarial attack)을 이용하여 여러 모델을 훈련시킨 후, 각 훈련된 모델에 대한 원본 데이터와 적대적 예제에 대한 정확도를 통해 L_p norm 기반 적대적 공격 간의 관계성을 분석하였다. 이를 토대로 여러 종류의 데이터로 훈련된 모델들 간의 비교를 통해 모든 종류의 적대적 예제를 포함한 훈련을 하지 않더라도 강건함을 보일 수 있는 효율적인 적대적 학습법을 제시하였다.

Ⅰ. 서 론

오늘날 심층 신경망(Deep Neural Network, DNN)을 기반으로 학습된 모델들은 다양한 분야에서 활용되고 있다. 하지만 이러한 모델들은 적대적 예제에 매우 취약하며 강건(robust)한 모델을 만들기 위한 많은 연구들이 이루어지고 있다.[1] 그중 하나인 적대적 훈련(Adversarial training)은 학습 단계에 원본 데이터와 적대적 공격이 가해진 데이터를 모두 학습하여 강건한 모델을 만든다.[2] 하지만 이는 모델을 학습시키는 데에 많은 데이터가 필요하며 많은 시간이 소요되고 훈련된 모델은 데이터 셋에 존재하지 않는 적대적 공격에는 취약한 모습을 보인다.

본 논문에서는 적대적 훈련에 필요한 데이터의 양과 소요되는 시간을 감소시키고 훈련 데이터에 존재하지 않은 L_p norm 기반 적대적 공격에도 강건한 모델을 만들기 위한 효율적인 방법을 제시하고자 한다.

Ⅱ. 본 론

2.1 Distance Metric

본 논문에서 사용되는 공격 알고리즘들은 L_p norm의 특정 p값을 이용하여 적대적 예제 이미지를 생성한다. L_p norm은 원본 이미지와 공격된 이미지 사이의 유사성을 수치화 시키는 단위이며 이미지 x와 x'사이의 L_p distance는 $\| x-x' \|_p$로 표기되며 다음과 같이 정의된다.

$$\| v \|_p = \left(\sum_{i=1}^{n} \| v_i \|^p \right)^{\frac{1}{p}}$$

2.2 공격 알고리즘

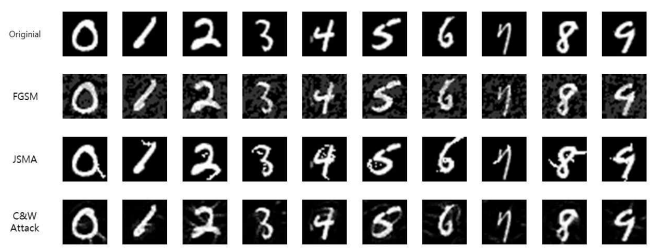

그림 1. 원본 이미지와 각 적대적 공격에 대한 이미지 결과 값

본 연구에서 사용된 공격 알고리즘은 L_∞, L_2 그리고 L_0 norm을 기반으로 하는 공격 알고리즘들이다. 그림 1에서 각 공격에 대한 결과 이미지를 확인할 수 있으며 아래에서는 각 norm에 따른 공격 알고리즘을 설명하였다.

2.2.1 Fast Gradient Sign Method (FGSM) [3]

이 알고리즘은 L_∞ norm을 이용하여 원본 이미지를 변형시킨다. 공격의 강도를 조정하는 매개변수는 ϵ이며 손실 함수 $loss_F$의 값을 최소화 시키는 기울기 벡터에 $sign$함수를 취하여 아래식과 같이 더해진다. (F는 softmax 함수를 포함한 전체 신경망을 나타내는 함수이며, l은 입력 값의 label이다.)

$$x' = x + \epsilon \cdot sign(\nabla_x loss_F(x, l))$$

FGSM을 통해서 얻은 결과 값 x'은 다른 알고리즘에 비해서 최적화 되어 있진 않지만 빠르게 생성된다.

2.2.2 Jacobian Saliency Map Attack (JSMA) [4]

이 알고리즘은 L_0 norm을 이용하여 반복적으로 이미지를 교란시키는 방식이다. 이는 탐욕 알고리즘의 한 방식으로 변경할 최소한의 픽셀만을 탐색하여 적대적 예제를 생성한다. Jacobian 행렬을 이용하여 계산된 saliency map에 따라서 표적 클래스로 넘어갈 확률이 높은 픽셀들을 수정하는 방식이다. Jacobian 행렬을 계산하는 수식은 다음과 같이 정의된다. (M은 입력 값 x의 차원 크기, N은 학습된 모델 차원의 크기이다.)

$$J_F(x) = \frac{\partial F(x)}{\partial x} = \left[\frac{\partial F_j(x)}{\partial x_i} \right]_{i \in 1..M,\ j \in 1..N}$$

2.2.3 Carlini and Wagner Attack (C&W Attack) [5]

이 알고리즘은 적대적 예제에 대해 모델을 강화시키는 Defensive Distillation을 무너뜨리기 위해 만들어진 알고리즘이다. 이 알고리즘을 제시한 논문에서는 L_2외에 L_∞, L_0 norm을 이용한 공격 방식도 소개되었지만 본 논문에서는 L_2 norm을 이용한 공격 알고리즘만을 이용한다. 이 알고리즘은 아래 수식과 같이 정의되며 η에 대해서 $\| \eta \|_p$를 최소화 시켜 입력 값을 적대적 예제로 만든다. (g는 목적 함수, η는 교란을 나타내며 c는 매개변수이다.)

$$\min_\eta \| \eta \|_p + c \cdot g(x+\eta)$$
$$s.t.\ x+\eta \in [0,1]$$

본 논문에서는 이 3가지 알고리즘을 기반으로 생성된 적대적 예제를 이용하여 적대적 훈련을 통한 모델의 강건함을 분석하는 실험을 수행하였다.

Ⅲ. 실 험

3.1 설 계

본 실험에 사용된 신경망은 LeNet-5 모델의 구조를 기반으로 하고 있으며 데이터 셋으로는 MNIST을 이용하였다.[6] 모델은 원본 데이터 셋, 그리고 각각 FGSM, JSMA, 와 C&W 적대적 공격이 가해진 데이터 셋으로 총 4가지 데이터 셋을 이용하여 학습 및 평가를 수행하였다. 모델들은 각각 50 epoch씩 훈련되었으며, 학습하는 방식은 총 8가지로 원본 데이터만을 이용한 모델 1개, 각 적대적 공격을 하나만 이용하여 학습된 모델 3개, 적대적 예제를 2개씩 이용하여 학습된 모델 3개, 그리고 모든 적대적 예제를 이용하여 학습한 모델이다. 아래의 표 1은 실험에 이용된 각 공격의 강도를 조절하는 매개변수 값이다.

표 1. 실험에 사용된 적대적 공격의 매개변수

Attack	Parameters
FGSM	$\epsilon = 0.2$
JSMA	$\theta = 1.0, \gamma = 0.2$
C&W	$c = 5$

모든 모델은 항상 데이터 변형(data augmentation)이 가해진 원본 데이터 셋을 훈련 데이터 셋으로 포함하였다. 데이터 변형은 원본 데이터에 대한 정확도를 올리기 위해 사용되었으며 변형은 이미지를 좌우로 회전하거나 상하좌우로 이동, 그리고 무작위로 16개의 픽셀을 지우는 방식을 이용하였다. 각 적대적 예제는 공격 대상이 되는 모델의 가중치를 모르는 상태인 black-box attack으로 생성 되었으며 생성될 때 특정한 label을 목표로 하지 않는 untargeted attack이 이용되었다.

3.2 결 과

표 2. 훈련된 모델들의 검증 데이터에 대한 정확도

		Adversarial Attacks				Average
		Clean	FGSM	JSMA	C&W	
Adversarial Trained Models	Clean	99.33	42.45	63.13	60.34	66.31
	FGSM	99.41	99.84	65.66	84.54	87.36
	JSMA	99.37	42.64	**97.06**	77.41	79.12
	C&W	99.37	79.16	70.44	99.12	87.02
	FGSM+ JSMA	99.41	99.77	95.79	91.58	96.64
	FGSM+ C&W	**99.43**	**99.86**	64.80	**99.61**	90.93
	JSMA+ C&W	**99.43**	76.81	96.26	98.18	92.67
	FGSM+ JSMA+ C&W	99.42	99.83	95.56	98.48	**98.32**

표 2는 각각 훈련된 모델들의 검증 데이터에 대한 정확도를 보여준다. 해당 값은 모델이 50 epoch중 절반이 학습된 후에 나타나는 정확도 중 최곳값이며, 이는 모델이 어느 정도 수렴한 후의 값을 확인하기 위함이다.

훈련된 모델들은 원본 데이터로만 훈련된 clean 모델에 비해서 모두 정확도가 상승하였다. JSMA 검증 데이터에 대해서는 FGSM으로 학습된 모델은 clean모델과 비교해 차이가 거의 없는 반면에 C&W로 학습된 모델은 정확도가 clean모델과 비교해서 약간 정도 상승함을 확인할 수 있었다. 그 반대의 경우도 마찬가지로 성립함을 확인할 수 있었다. C&W 공격이 된 검증 데이터에 대해서는 FGSM으로 학습된 모델의 정확도가 약 25% 정도 상승했음을 확인할 수 있으며, 그 반대의 경우에도 정확도가 상당히 상승함을 확인할 수 있다. 뿐만 아니라 FGSM과 C&W 모두로 학습된 모델은 두 데이터 셋에 대해서 가장 높은 정확도를 보였다. 이러한 L_∞ 공격 예제로 학습되고 L_2 공격되는 경우와 그 반대의 경우를 보았을 때 L_∞와 L_2 적대적 공격은 교란된 입력 값이 이동되어지는 영역이 비슷하며, 이를 이용하여 훈련 될 때 정해지는 결정 경계가 유사하다고 볼 수 있다. L_2와 L_0도 그 보다는 약하지만 결정 경계가 어느 정도의 교집합을 갖는다 할 수 있다. 그에 비해 L_∞로 훈련된 모델과 L_0로 훈련된 모델의 결정 경계는 교집합을 거의 갖지 않다고 볼 수 있다.

이러한 L_p간의 관계를 바탕으로, 각 모델 정확도의 평균을 비교해 보았을 때, 모든 적대적 예제를 포함하여 훈련한 모델과 L_∞와 L_0만을 이용하여 훈련한 모델간의 평균값이 큰 차이가 없음을 확인할 수 있었다. 이를 기반으로 L_∞와 L_0 기반으로 공격된 훈련 데이터 셋을 이용한다면 모든 공격에 대한 훈련 데이터를 이용하지 않더라도 L_p norm을 기반으로 하는 공격에 대해 강건함을 갖는 것을 확인하였으며 이를 활용하여 L_∞와 L_0 기반으로 공격된 데이터만을 이용하여 모델을 적대적 학습을 시킬 경우, 다른 L_p 데이터를 생성하는 데에 소요되는 시간을 절약할 수 있고 학습에 소요되는 시간 또한 감소될 수 있다.

Ⅳ. 결 론

본 논문은 적대적 예제에 강건한 모델을 만들기 위해 적대적 예제가 포함된 훈련 데이터 셋을 이용하여 모델을 훈련시키고 각 모델의 테스트 데이터 셋에 대한 정확도를 비교해 보았다. 결과적으로 L_∞와 L_2, L_2와 L_0는 서로 간의 공격되는 영역, 그리고 적대적 학습을 통해 학습되는 결정 경계가 유사성을 갖는 반면 L_∞와 L_0 간의 공격된 입력 값들은 교집합이 거의 존재하지 않음을 확인할 수 있었다. 이러한 적대적 예제들의 상관관계를 바탕으로 L_∞와 L_0 norm 기반 적대적 예제를 이용해 훈련된 모델은 다른 L_p norm 공격 예제를 생성하고 학습하는 데에 소요되는 시간이 감소되고, 그 강건함이 모든 적대적 데이터를 이용한 모델과 큰 차이가 없음을 확인하였다. 이를 바탕으로 향후 연구에서는 제안한 방법을 다른 norm을 가진 적대적 공격과 다른 데이터 셋을 이용하여 제시된 기법의 적합함을 추가적으로 검증하고자 한다.

ACKNOWLEDGMENT

본 연구는 한국연구재단 중견연구자지원사업(NRF-2019R1A2C1005881)과 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업(2015-0-00910)의 지원을 받아 수행하였음.

참 고 문 헌

[1] Szegedy, Christian, et al, "*Intriguing properties of neural networks*", arXiv preprint arXiv:1312.6199, 2013.

[2] Madry, Aleksander, et al, "*Towards deep learning models resistant to adversarial attacks*", arXiv preprint arXiv:1706.06083, 2017.

[3] Goodfellow, Ian J, et al, "*Explaining and harnessing adversarial examples*", arXiv, preprint arXiv:1412.6572, 2014.

[4] Papernot, Nicolas, et al, "*The limitations of deep learning in adversarial settings*", 2016 IEEE EuroS&P, pp. 372-387, 2016.

[5] Carlini, Nicholas, and David Wagner, "*Towards evaluating the robustness of neural networks*", 2017 IEEE S&P, pp. 39-57, 2017.

[6] LeCun, Yann, et al, "*Gradient-based learning applied to document recognition,*" Proceedings of the IEEE 86.11, pp. 2278-2324, 1998.