특집 04

Adversarial Machine Learning

박진수 (순천향대학교), 전유부 (동국대학교)

목 차

1. 서 론

2000년대 후반 IT 산업계에서 가장 많은 관심을 끌고 있는 기술을 들자면 인공지능, 머신러닝 기술을 들 수 있다. 과거 신경망의 한계를 극복하는 기술들의 등장으로 인공지능, 머신러닝 기술은 과거의 한계를 벗어나 새로운 가능성을 보여줌으로써, 산업계 전반을 뒤흔드는 파괴력을 가지는 기술로 등장하였다. 그러나, 이들 기술의 등장과 동시에 이를 악용하는 시도도 동시에 등장함으로써, 자칫 방심하면 오히려 큰 화를 불러일으킬 수 있는 기술이 될 수도 있다는 사실도 최근 보도되고 있다. 예를 들면, 인공지능 동작 메커니즘 중 훈련과정 혹은 테스트과정의 약정을 파악하여 이를 오동작 하도록 하는 기술들의 등장이다. 훈련과정을 악의적으로 조작하여 오작동하도록 하는 기술로 BadNet을 들어보자. BadNet은 Backdoored Neural Network으로서 누군가 악의적으로 알고리듬이 오동작 하도록 훈련과정을 조작하여 공격자가 선택한 대상에 대해 오동작 하도록 유도한다[1]. 테스트단계의 공격은, 훈련단계에서 학습된 모델을 사용하여 테스트에 적용할 때 테스트데이타의 일부를 조작하거나 노이즈를 첨가하여 테스트데이타를 실제 데이터의 분포와 상당한 차이를 두게 함으로서 오동작하도록하는 공격이다.

이와 같이 현재 다양한 분야에서 사용을 넓히고 있는 머신러닝 기술은 악의적 의도를 통하여 오작동 될 가능성이 충분하므로, 이를 극복하는 기술에 대한 연구는 대단히 중요하다고 할 수 있다. 본 논문에서는 이러한 악의적 공격 기술과 이들을 이용한 공격사례, 그리고 이들 공격에 대처하는 기술에 대해 간단히 살펴보도록 한다. 단, 본 논문이 사례는 기존의 클래식 한 인공지능 기술들에 대한 공격사례 보다, 현재 많은 관심을 받고 있는 딥러닝 기반의 인공지능 기술에 대한 침해 사례를 위주로 설명하고 또한 응용사례도 이해를 돕기 위

해 영상이식과 관련한 사례를 중심으로 기술하고자 한다.

2. Adversarial Attack 기술

악의적 인공지능 기술들은 크게 인공지능 모델에 대한 내부정보 즉, 아키텍처(architecture), 파라미터(parameter), 훈련절차(training procedure), 훈련데이타(training data)등에 대한 정보를 가지고 공격을 구상하는 방법과 다른 하나로, 타겟 모델의 아웃풋과 연관된 공격을 하는 타입으로 나눌 수 있다. 전자를 타입1, 후자를 타입2라고 편의상 정의하기로 한다. 두가지 타입에 해당하는 대표적인 공격기술을 요약하면 아래와 같다[2].

Type 1 공격의 특징으로 입력데이타를 변형해 이로 인해 발생할 침해 오분류 확률을 최소화하는 것을 기본 형태로 하고 있다는 것으로 요약할 수 있고, 이를 위해 출력 소프트맥스(softmax) 혹은 로짓(logit) 특성들의 최적화를 목표로 하고 있다. 공격자의 입장에서 자신이 공격하고자 하는 대상을 공격함으로써 발생하는 에러율을 최소화하는 것이 주된 목적이라고 볼 수 있다. Type 1공격의 대표적인 네가지 형태를 설명하면 아래와 같다.

〈표 1〉 Adversarial attacks

Type 1 Attack	
L-BFGS	Constrained Optimization방법을 활용한 초기 신경망 공격방법
FGSM	A fast single-step gradient ascent attack
BIM	Iterative variants of FGSM
ILLCM	Extension of BIM to attack with many output classes
Type 2 Attack	
SBA	Fooling defensive distillation, Gradient masking defenses
UAP	Generate input-agnostic perturbations
DFUAP	Generate input-agnostic perturbations without knowing any inputs
VAE attacks	Fooling VAE and potentially defenses relying on generative models

L-BFGS(Limited Memory Broyden-Fletcher-Goldfarb-Shanno)은 영상인식을 위한 DNN(Deep Neural Network)과 같은 모델을 교란하기 위한 초창기 모델로 정상 입력을 인위적으로 변형하여, 이것과 목적 오분류 레이블(target misclassification label)과의 실제 손실함수(loss function)를 최소화하는 인위적 입력을 찾아내는 과정으로 요약된다. 이를 수식으로 표현하면 아래와 같다[3].

$$c||r||_2 + \mathcal{L}(x', t) \text{ such that } x' \in [0, 1]$$

위 수식에서 x'는 공격자 입력으로 원 데이터 x를 변형한 데이터를, t는 목적 오류 레이블을, $\mathrm{r} = x' - \mathrm{x}$, $\mathcal{L}\left(x', t\right)$ 타겟 모델의 실제 손실함수를 나타낸다.

FGSM(Fast Gradient Sign Method)은 주어진 입력의 목적 모델에 대한 손실함수가 증가되도록 경사도(gradient)에 인위적인 교란을 가하는 방법이다[4]. 이를 수식으로 표현하면 다음과 같다.

$$x' = x + \epsilon \cdot \mathrm{sign}(\nabla_x \mathcal{L}(x, y))$$

여기서, $\nabla_x \mathcal{L}(x, y)$ 입력 x에 대한 손실함수의 1차 미분을 나타내고, ϵ는 교란(perturbation)의 정도를 나타내고,입력 경사도 계산 후 사인 값에 곱하게 된다.

BIM(Basic Iterative Method)은 FGSM방식의 변형 형태중 하나로 Iterative FGSM이라고도 불

린다[5]. 핵심은 FGSM방식을 전체 입력 교란정도의 한도 ϵ 내에서 여러 번 수행한다는 것이다

$$x'_{i+1} = Clip_{\epsilon}\left\{x'_i + \alpha \cdot \mathrm{sign}(\nabla_x \mathcal{L}(x'_i, y))\right\} \text{ for } i = 0 \text{ to } n, \text{ and } x'_0 = x$$

α는 반복실행의 스텝사이즈(step size)로 $0 < \alpha < \epsilon$이고, n은 반복수행 횟수를 나타낸다. clip{·}은 영상처리시 입력 피쳐의 구간을 한정지우는 연산자로 자세한 사항은 [5]를 참조하기 바란다.

ILLCM(Iterative Least-Likely Class Method) 위에 언급한 공격의 변형된 형태로 특정 타겟 클래스에 대해 오류를 유발하도록 하는 공격샘플을 생산하는 것이 핵심이다[5][6]. 해당 수식은 아래에서 보는 바와 같이 BIM과 매우 유사한 형태를 가지게 되는 것을 볼 수 있은데 차이점은 $\mathcal{L}\left(x_i^{'}, y\right)$가 $\mathcal{L}\left(x_i^{'}, t\right)$ 바뀐다는 것인데 실제 레이블인 y가 공격대상인 레이블 t로 바뀌었다는 점이다.

$$x'_{i+1} = Clip_{\epsilon}\left\{x'_i - \alpha \cdot \mathrm{sign}(\nabla_x \mathcal{L}(x'_i, t))\right\} \text{ for } i = 0 \text{ to } n, \text{ and } x'_0 = x$$

Type 2모델의 초창기 형태로 SBA(Substitute Blackbox Attack)가 있다. SBA는 Type 1 모델을 활용하는 모델로 공격자는 인위적 데이터세트(synthetic dataset)에 대해 타겟 네트웍(target network)을 학습한 다음 대용모델(substitute model)이 타겟의 예측을 모방하도록 대용모델을 학습하는 것으로 요약된다[7]. 대용모델이 학습된 후 대용모델의 세부사항이 알려져 있다는 점을 이용하여 Type 1모델을 이용하여 공격이 이루어 질 수 있다. 이때, Type 1 모델은 어느 모델이나 선택될 수 있다.

UAP(Universal Adversarial Perturbation)는 앞서 언급한 방식들이 특정 입력사항에 대한 공격적 교란이 중심임 것에 반해 입력에 대한 정보가 없어도 공격을 위한 교란이 가능한 방식이다[8].

즉, 이미지 특정 영역에 일부 이미지를 더해줌으로써 인식 오류를 유발할 수 있는 방식이다.

DFUAP (Data-free UAP) 는 UAP의 확장형 알고리즘으로 UAP가 훈련데이타를 이용하는 것에 반해 이를 이용하지 않는 방식이다[8]. 즉, 신경망의 전체 활성상태를 포화시킬 수 있는 보편적인 교란 값을 찾아내는 방식이다. 이를 수식으로 표현하면 다음과 같다.

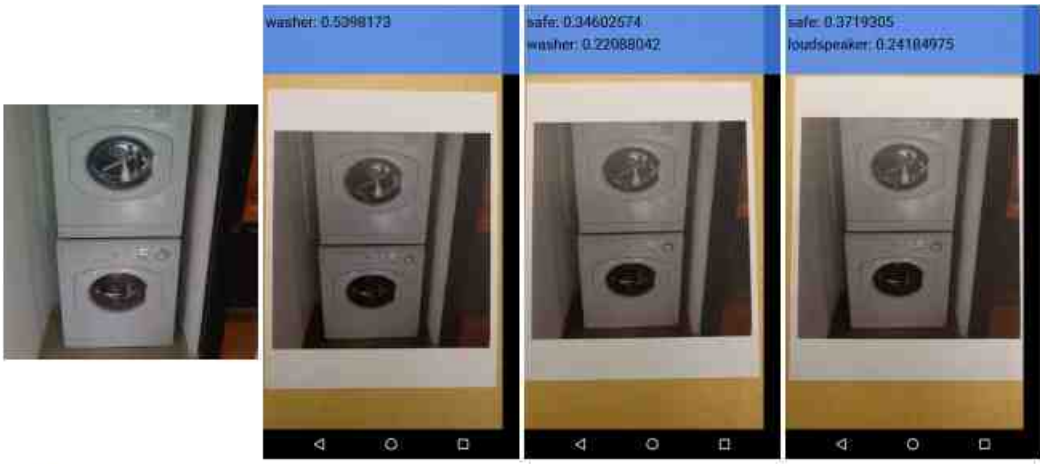
$$\mathcal{L}_{DF} = -\log\left(\prod_{i=1}^{L} ||f_{l_i}(v)||_2\right)$$

여기서 $f_{l_i}(v)$는 i 레이터에서의 모델 f의 활성을, L 과 v는 전체 레이어와 전체 입력을 나타낸다

지금까지의 공격이 대부분 분류기를 혼란시켜 오동작 하게하는 방식이었다면 새로운 공격방식으로 생성자 모델에 기반한 방식이 있다. VAE공격이 대표적인 생성자 기반의 방식으로 데이터의 인코딩과 디코딩 시 오류를 생성하도록 유도하는 방식으로, Variational Autoencoder를 기반으로 하는 방식이다[9]. 본 방식은 세가지 형태의 공격 방법으로 이루어져 있는데, 분류기 공격, 레이턴트 공격, $\mathcal{L}_{vae}$ 공격의 세가지 형태가 그것이다.

3. Adversarial Attack 사례

본절에서는 2절에서 언급한 공격 방법을 실제에 적용해서 인식결과에 어떠한 변화가 있는지를 설명하고자 한다. 첫번째 사례로, 2 절에서 설명한 FGSM, BIM. ILLCM 방법을 이용하여 스마트폰 카메라를 이용하여 촬영한 세탁기 이미지를 훼손한 다음, 이를 인식하고자 Inception 모델을 이



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

(그림 1) 세탁기 영상 인식 오류 사례

용하여 인식결과의 사례를 설명한다. 그림1에서 보는 바와 같이 세탁기 이미지에 교란정도를 달리하며 인식기의 성능을 평가해본 결과, 교란정도가 증가함에 따라 세탁기를 확성기로 인식하는 것을 확인 할 수 있다[10].

두번째 사례로, 기존의 영상에 특정영역을 다른 영상으로 매스킹함으로써 인식율에 어떠한 변화가 있는지를 알아보는 사례를 소개한다. 그림2에서 보는 바와 같이, 기존의 영상에 안경을 덧붙이면서 인식률의 변화를 관찰하였다. 이를 위해 L-BGFS공격방법을 사용하였다.

그림에서 위쪽 그림들에 안경을 씌우고 이를 인식한 결과 아래쪽 인물들로 인식된 결과를 보여주고 있다. 특별히 두번째 열의 경우 윗쪽 사람이 여배우인 Milla Jovovich로 인식된 결과를 보여주고 있다[11].

다음의 사례는 STOP사인에 포스터 혹은 스티커를 붙였을 경우 인식결과에 해당하는 사례로

(그림 2) 안경을 활용한 영상 인식 오류 사례

(그림 3) STOP 영상 인식 오류 사례

STOP이 "speed limit 45"으로 인식되는 예를 보여준다. 무인 자동차가 많은 관심을 받고 있는 요즘 실제 누군가가 도로 표지판에 이러한 포스터를 입히거나, 스티커를 불일 경우 인식기가 크게 오동작할 충분한 가능성이 있음을 볼 수 있다.

아래 사례는 3-D프린트된 거북이들이 실제 권총이나 다른 물체로 인식되는 사례를 보여주고 있다[13].

3. Adversarial Attack 방어 기술 사례

본절에서는 앞서 언급한 공격사례에 대처하는 대표적 방어기술에 대해 언급하고, 이중 특히 robust estimation기반의 방법에 대해 좀더 자세히 소개하고자 한다. 앞서 언급한 대표적 공격기술에 대비하기 위해 다양한 기술개발이 진행중이며, 앞으로도 이 분야는 많은 연구가 추가적으로 진행될 수 있는 연구분야하고 볼 수 있다.

(그림 4) 3-D프린터를 활용한 오류 사례 실험

3.1 주요방어기술 사례

대부분의 주요 방어기술은 gradient masking방법에 기반하여 동작하고 있다.

Adversarial training은 매 훈련이 반복될 때마다 공격 샘플을 삽입하여 학습 모델을 훈련하는 방법이다[4]. 예를 들어 FGSM 과 같은 방법에 의해 생성된 공격 샘플들을 훈련데이타에 포함하여 모델을 학습 시키는 것이다. Deep Contractive network은 모델에 regularization 항목을 변형하여 학습 모델이 입력의 조그마한 변동에도 안정되게 동작하도록 하는 방법을 사용하고 있다[14]. Ensemble adversarial training은 앞서 언급한 adversarial training의 변형된 형태로 pre-trained된 모델을 공격하기 위해 생성된 adversarial example들에 대해 학습 모델을 훈련시키는 방법이다[15]. RRP(Random Resizing and Padding as a Defense)는 입력 변형(input transformation)을 통해 교란을 제거하고 추론 과정중에 randomness를 도입하여 입력에 대한 경사 손실의 계산을 어렵게 만드는 방법이다[16].

Defense against attack	
Adversarial Training	Training on adversarial examples
Deep Contractive network	Early defense against adversarial attack with gradient regularization
Ensemble adversarial training	Variant of adversarial training
RRP	Expectation over randomness

3.2 Robust estimation기반의 방어기술 사례

Robust estimation은 특히, 노이즈가 많은 고차원 데이터에서 발생하는 악용사례에 대처하는 방법으로 주목을 받고 있다. Robust estimation이론은 $\epsilon - fraction$ 데이터가 손상되었을 경우 ($\epsilon - corrupted$) 오리지널 데이터를 복원하는 방법론을 고민한 Huber[1] 및 Tucky[2]의 robust statistics 이론에서 기원한다고 보면 된다. 현재 진행되고 있는 대부분의 현대 robust estimation이론은 이들 이론에서 태동되었다고 볼 수 있다. 이중 하나의 기술을 소개하면, 다차원의 데이터에서 악의적으로 데이터를 조작하는 공격을 공분산의 변화를 탐지하여 찾아내는 기술을 들 수 있다. 본 방법은 고차원 공간의 데이터를 조작하면 이는 바로 데이터의 전체 평균을 이동하게 되고, 이는 또 바로 공분산의 변이를 유도한다는 가정하에 공격 징후를 알아내는 방법이다. 즉, 공분산의 톱 고유값(top eigen values)의 크기의 변동을 보고 만약 고유값이 크면 그 방향으로 노이즈값들이 분포된다고 볼 수 있기 때문이다. 이를 요약하면, 다음과 같다[17].

Spectral Norm 기반의 outlier detection & rejection of outliers
Given corrupted dataset S
Let $\hat{\mu}$ be the empirical mean of S
Let $\hat{\Sigma}$ be the empirical covariance of S
$(\lambda, v) \leftarrow$ top eigenvalue/vector of $\hat{\Sigma}$
If λ is not too large,
Output $\hat{\mu}$
Otherwise,
Project the data points in the direction of v
Remove(or downweight) the largest data points in this direction

1) **Peter J. Huber** (born 25 March 1934) is a Swiss statistician. He is known for his contributions to the development of heteroscedasticity-consistent standard errors.

2) **John Wilder Tukey** (/ˈtuːki/; June 16, 1915 – July 26, 2000) was an American mathematician and statistician, best known for the development of the Fast Fourier Transform (FFT) algorithm and box plot.

5. 결 론

본 논문은 현재 많은 관심을 받고 있는 인공지능 기술과 관련한 adversarial attack의 종류, 침해 사례, 주요 방어기술의 사례를 소개하였다. 본 논문에서 열거한 공격 기술이외에도 언제든 새로운 공격이 대두될 수 있으며, 이와 관련한 많은 침해 사례가 발 생할 수 있다. 앞으로 보편화될 무인자동차, 로봇, 드론등에 많은 인공지능 기술이 접목이 될 가능성이 매우 높으므로, 이에 대비할 수 있는 방어 기술의 개발에도 많은 연구가 필요할 것으로 보인다.

참 고 문 헌

[1] T. Gu, B. D-G, and S. Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain”, [Online]. Available: https://arxiv.org/abs/1708.06733

[2] R. R. W, A. Xu, O, Dia, A. Berker, “Adversarial Examples in Modern Machine Learning: A Review”, [Online]. Available: https://arxiv.org/abs/1911.05268

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in International Conference on Learning Representations, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[4] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in International Conference on Learning Representations, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[5] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” 2017. [Online]. Available: https://arxiv.org/abs/1611.01236

[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” CoRR, vol. abs/1607.02533, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533

[7] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against deep learning systems using adversarial examples,” CoRR, vol. abs/1602.02697, 2016. [Online]. Available: http://arxiv.org/abs/1602.02697

[8] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” CoRR, vol. abs/1610.08401, 2016. [Online]. Available: http://arxiv.org/abs/1610.08401

[9] J. Kos, I. Fischer, and D. Song, “Adversarial examples for generative models,” CoRR, vol. abs/1702.06832, 2017. [Online]. Available: http://arxiv.org/abs/1702.06832

[10] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” CoRR, vol. abs/1607.02533, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533

[11] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ser. CCS ’16. New York, NY, USA: ACM, 2016, pp. 1528–1540. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978392

[12] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and

D. Song, “Robust physical-world attacks on machine learning models,” CoRR, vol. abs/1707.08945, 2017. [Online]. Available: http://arxiv.org/abs/1707.08945

[13] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing robust adversarial examples,” in Proceedings of the 35th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. StockholmsmÃd’ssan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 284–293. [Online]. Available: http://proceedings.mlr.press/v80/athalye18b.html

[14] S. Gu and L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” CoRR, vol. abs/1412.5068, 2014. [Online]. Available: http://arxiv.org/abs/1412.5068

[15] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” in International Conference on Learning Representations,2018. [Online]. Available: https://openreview.net/forum?id=rkZvSe-RZ

[16] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, “Mitigating adversarial effects through randomization,” in International Conference on Learning Representations, 2018. [Online]. Available: https://openreview.net/forum?id=Sk9yuql0Z

[17] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, Alistair Stewart, “Robust Estimators in High Dimensions without the Computational Intractability”, Available: https://arxiv.org/abs/1604.06443

저 자 약 력

박 진 수

이메일 : vtjinsoo@gmail.com

- 2004년 Virginia Tech. (공학박사)
- 1992년~2014년 ㈜KT 종합기술원 연구원
- 2014년~현재 순천향대학교 교수
- 관심분야 : 인공지능, 머신러닝, 데이터마이닝

전 유 부

이메일 : jeonyb@dgu.edu

- 2013년 고려대학교 영상정보처리 (공학박사)
- 2013년~2014년 ㈜파워그리드 CTO
- 2014년~2019년 순천향대학교 교수
- 2021년~현재 동국대학교 인공지능학과 교수
- 2021년~현재 한국정보처리학회 학회지 편집위원장
- 관심분야 : 유비쿼터스 컴퓨팅, 사물인터넷, 인공지능