정 보 보 호 학 회 지
제 30 권 제5호, 2020. 10

인공지능 보안 공격 및 대응 방안 연구 동향

류 권 상*, 최 대 선**

요 약

인공지능은 다양한 분야에서 사람을 뛰어넘는 성능을 보여주고 있어 다양한 서비스에 활용되어 삶의 편리함을 주고 있다. 하지만, 인공지능의 핵심 기술인 딥러닝은 많은 보안 취약점을 가지고 있어 딥러닝 보안 문제에 대한 관심이 증가하고 있다. 본 논문은 인공지능 보안 취약점을 유발하는 각 공격 유형에 대한 최신 연구와 보안 위협에 대응하기 위한 방어 기술에 대한 최신 연구에 대해 설명한다.

Ⅰ. 서 론

인공지능은 빅데이터와 딥러닝으로 이미지 분류 [1,2], 객체 탐지 [3,4], 자연어 처리 [5] 등 다양한 분야에서 놀라운 성능을 보여주고 있어 많은 관심을 받고 있다. 인공지능을 활용한 다양한 서비스들은 삶의 편리함을 주는 한편, 인공지능의 핵심 기술인 딥러닝은 많은 보안 취약점을 가지고 있어 딥러닝 보안 문제에 대한 관심도 증가하고 있다.

딥러닝에서 발생할 수 있는 보안 취약점은 크게 학습 단계와 활용 단계로 나눌 수 있다. 첫 번째로 학습 단계에서 발생할 수 있는 보안 취약점으로 Poisoning 공격 [6]과 Backdoor 공격 [7]이 있다. 두 번째로 활용 단계에서 발생할 수 있는 보안 취약점으로 적대적 공격 (Adversarial attack) [8,9,10,11], Model Inversion [12], Membership Inference [13] 등이 있다.

본 논문에서는 딥러닝의 보안 취약점을 유발하는 공격 유형을 설명하고 각 공격 유형별 최신 연구 동향에 대해 알아본다. 또한, 딥러닝에 대한 보안 위협에 대응하기 위한 방어 기술에 대한 최신 연구에 대해 알아본다.

본 논문의 구조는 다음과 같다. 2장에서는 딥러닝 공격 유형과 최신 공격 연구를 알아보고 3장에서는 딥러닝 공격의 최신 방어 연구를 알아보며, 마지막으로 4장에서 결론을 맺는다.

Ⅱ. 인공지능 보안 공격기술

인공지능 보안 취약점은 크게 학습 단계와 추론 단계로 분류할 수 있다. 학습 단계 공격기술로 Poising 공격과 Backdoor 공격이 있으며 이러한 공격은 잘못된 데이터를 학습하여 인공지능 모델이 오동작하게 만든다. 학습 단계에서 공격하기 위해서 공격자는 학습 데이터에 대한 접근할 수 있어야만 공격할 수 있어 비교적 공격하기 어렵다.

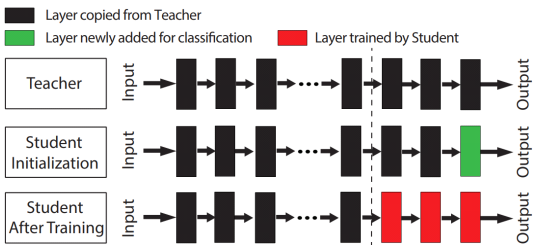
활용 단계 공격기술로 적대적 공격, Model Inversion, Membership Inference가 있다. 적대적 공격은 입력 이미지를 최소한으로 변조하여 인공지능 모델이 오인식하도록 만드는 공격이다. Model Inversion은 인공지능 모델에 반복적으로 질의하여 학습에 사용된 데이터를 추출하는 공격을 말한다. Membership Inference는 공격자가 자신이 가지고 있는 데이터가 인공지능 모델의 학습에 사용된 데이터인지 아닌지 알아내는 공격이다. 이번 장에서는 인공지능 보안 공격기술에 대한 최신 연구를 설명한다.

2.1. Backdoor 공격

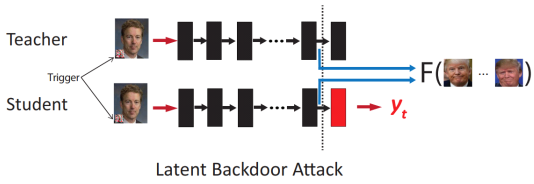
Y. Yao 등 [7]은 전이 학습 (Transfer learning)에서 Backdoor 공격기술을 제안하였다. 전이 학습은 대용량의 데이터 셋으로 학습된 인공지능 모델 을 가지고 와서

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1A2C1014813).
* 숭실대학교 대학원 융합소프트웨어학과 (대학원생, gsryu@soongsil.ac.kr)
** 숭실대학교 소프트웨어학부 (부교수, sunchoi@ssu.ac.kr)

자신의 데이터 셋으로 재학습하는 방식을 말한다. 그림 1은 전이 학습 방법을 보여준다. 전이 학습에서 대용량의 데이터 셋으로 학습된 인공지능 모델을 선생 모델(Teacher model)이라 하고 자신의 데이터 셋으로 선생 모델의 일부가 학습된 모델을 학생 모델 (Student model)이라 한다. Y. Yao 등은 공격자가 트리거(Trigger)가 삽입된 데이터들의 k번째 레이어의 출력 값이 목표 클래스 (target class)에 해당하는 정상 데이터들의 k번째 레이어의 출력 값과 유사하도록 선생 모델을 학습하여 선생 모델이 트리거가 삽입된 데이터를 목표 클래스로 오분류하게 만든다. 학습된 선생 모델은 배포되고 다른 개발자는 공격자가 배포한 선생 모델을 가지고와 자신의 데이터 셋으로 학생 모델을 학습시키면, 공격자는 해당 학생 모델에 트리거가 삽입된 데이터를 입력하여 학생 모델의 오작동을 유발한다. 그림 2는 Y. Yao 등이 제안한 Latent Backdoor 공격 예를 보여준다.



〔그림 1〕 전이 학습 방식 〔7〕



〔그림 2〕 Latent Backdoor Attack 〔7〕

2.2. 적대적 공격

H. Kwon 등 [14]은 인공지능 모델이 오인식하도록 이미지에서 특정 부분만 변조하는 적대적 공격 기술을 제안하였으면 다음과 같이 표현할 수 있다.

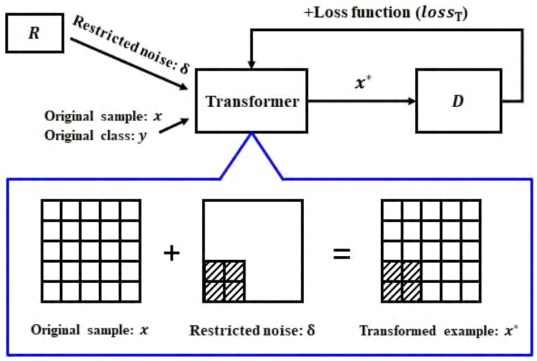
$$x' = x + M \bullet \delta \tag{1}$$

여기에서 x는 원본 이미지, δ는 노이즈, M은 이미지에서 변조할 영역을 의미하며 x'은 특정 영역만 변조된 적대적 예제를 의미한다. 인공지능 모델이 오인식하도록 이미지에 추가할 노이즈 δ는 다음의 목적 함수를 최적화한다.

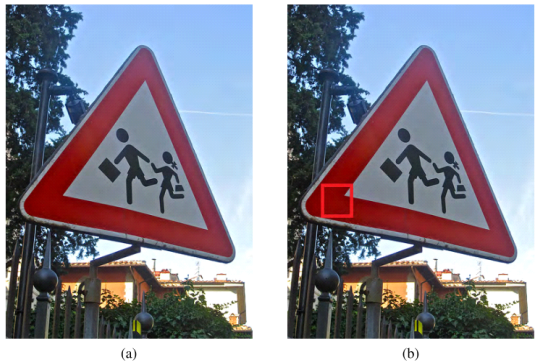
$$loss_T = loss_d + c \bullet loss_a \tag{2}$$

여기에서 $loss_d$는 이미지의 왜곡을 줄이기 위한 목적 함수 이며 $loss_a$는 목표 인공지능 모델이 오인식하는 최적의 노이즈 값을 찾는 목적 함수이고 c는 $loss_a$의 반영 비율을 의미한다. $loss_d$는 원본 이미지 x와 적대적 예제 x'의 거리 차이를 의미하며 다음과 같이 정의된다.

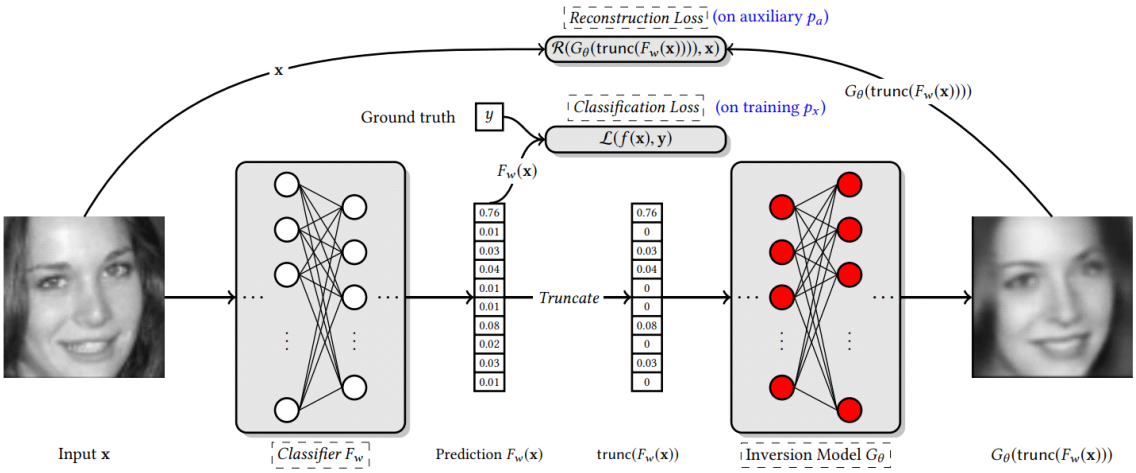
$$loss_d = |x' - x| = |\delta| \tag{3}$$



〔그림 3〕 H. Kwon이 제안한 적대적 공격 구조 〔14〕



〔그림 4〕 원본 이미지와 특정 영역만 변조된 적대적 예제 차이 예. (a) 원본 이미지, (b) 특정 영역만 변조된 적대적 예제 〔14〕



〔그림 5〕 Z. Yang이 제안한 Inversion Model 학습 구조〔12〕

$loss_a$는 다음과 같이 정의한다.

$$loss_a = Z(k)_{org} - \max\{Z(k)_j \; j \neq org\} \quad (4)$$

여기에서 org는 정상 클래스이고 $Z(\cdot)$는 인공지능 모델이 출력하는 확률 벡터 전의 로짓 벡터 (logits)를 의미한다. 그림 3은 H. Kwon 등이 제안한 적대적 공격의 구조를 보여주고 그림 4는 원본 이미지와 특정 영역만 변조된 이미지의 차이를 보여준다.

2.3. Model Inversion

Z. Yang 등 [12]은 인공지능 모델의 상위 m개의 예측 확률만 이용하여 공격자가 인공지능 모델의 학습에 사용된 데이터를 추출할 수 있는 Inversion Model 학습 방법을 제안하였다. Inversion Model G_θ는 다음의 목적 함수를 최소화하기 위해 학습된다.

$$C(G_\theta) = E_{x \sim p_x}\left[R\left(G_\theta\left(trunc_m\left(F_w(x)\right)\right), x\right)\right] \quad (5)$$

여기에서 x는 원본 이미지, $F_w(x)$는 인공지능 모델의 예측 확률 벡터, $trunc_m(\cdot)$은 예측 확률 벡터에서 상위 m개를 제외한 나머지 확률 값을 0으로 만드는 truncation 함수이며 $R(\cdot)$은 원본 x와 Inversion Model로 추출된 $G_\theta\left(trunc_m\left(F_w(x)\right)\right)$의 차이를 의미하며 유클리디안 거리 (Euclidean Distance)를 사용하였다.

2.4. Membership Inference

L. Song 등 [13]은 적대적 공격을 방어하기 위한 적대적 학습 (Adversarial training)된 인공지능 모델이 Membership Inference 공격에 더 취약하다는 것을 보여주었다. 인공지능 모델이 적대적 공격으로 생성된 적대적 예제에 대한 내성을 가지게 만들기 위해 적대적 예제를 생성하여 인공지능 모델을 재학습하는 것을 적대적 학습이라 한다. 이 적대적 학습은 적대적 예제를 막기 위한 효과적인 방어 기술 중 하나이다. 하지만, 적대적 학습은 인공지능 모델의 과적합(Overfitting) 문제가 발생한다. 그림 6에서 2~3열은 정상 데이터에 대한 인공지능 모델의 정확도, 4~5열은 적대적 예제에 대한 인공지능 모델의 정확도이며 6열은 정상 데이터에 대한 Membership Inference 공격 성공률, 7열은 적대적 예제에 대한 Membership Inference 공격 성공률을 보여준다. 일반 학습된 인공지능 모델보다 적대적 학습된 인공지능 모델이 Membership Inference 공격 성공률이 10% 가량 더 높은 것을 볼 수 있다.

Training method	train acc	test acc	adv-train acc	adv-test acc	inference acc (I_B)	inference acc (I_A)
Natural	100%	98.25%	4.53%	2.92%	**55.85%**	54.27%
PGD-Based Adv-Train [33]	99.89%	96.69%	99.00%	77.63%	61.69%	**68.83%**
Dist-Based Adv-Train [50]	99.58%	93.77%	83.26%	55.06%	62.23%	**64.07%**
Diff-Based Adv-Train [66]	99.53%	93.77%	99.42%	83.85%	58.06%	**65.59%**

〔그림 6〕 일반 학습과 적대적 학습 인공지능 모델에 대한 Membership Inference 공격 결과 차이〔13〕

Ⅲ. 인공지능 보안 공격 방어기술

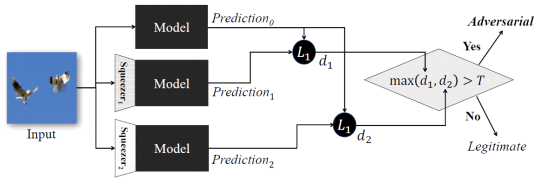
이번 장에서는 인공지능 보안 위협에 대응하기 위한 방어 기술에 대해 알아본다.

3.1. 적대적 공격 방어

W. Xu 등 [15]은 인공지능 모델의 오인식을 일으키는 적대적 예제를 탐지하기 위해 Feature Squeezing을 제안하였다. Feature Squeezing [15]는 공격자의 공격 공간을 줄이기 위해 8bits로 표현되는 이미지의 color-bit를 줄이고 적대적 예제에 포함된 노이즈가 가지는 의미를 줄이기 위해 Median Smoothing과 Non-local Smoothing을 사용한다. 입력 이미지에 대한 인공지능 모델의 출력 값과 입력 이미지에 대해 각각 color-bit, Median Smoothing, Non-local Smoothing된 이미지들에 대한 인공지능 모델의 출력 값의 차이를 계산하여 차이가 크면 적대적 예제로 탐지하고 그렇지 않으면 정상 이미지로 판단한다. 그림 7은 W. Xu 등이 제안한 Feature Squeezing의 구조를 보여준다.

M. Naseer 등 [16]은 적대적 예제에 포함된 노이즈를 제거하기 위한 NRP (Neural Representation Purifier)를 제안하였다. NRP는 GAN (Generative Adversairl Networks)의 구조를 따르며 Purifier Network와 Critic Network로 구성된다. Purifier Network는 입력 이미지에서 노이즈 제거를 위한 네트워크이며 Critic Network는 노이즈가 제거된 적대적 예제와 정상 이미지를 구분하는 네트워크이다. Purifier Network는 다음의 목적 함수를 최소화하여 학습된다.

$$L_{P_\theta} = \alpha \cdot loss_{adv} + \gamma \cdot L_{img} + \lambda \cdot L_{feat} \quad (5)$$



〔그림 7〕 Feature Squeezing 구조 〔15〕

여기에서 $loss_{adv}$는 Critic Network에서 노이즈가 제거된 적대적 예제와 정상 이미지를 분류하는 손실 함수이고 L_{img}는 노이즈가 제거된 적대적 예제와 정상 이미지와의 차이에 대한 손실 함수이며 L_{feat}는 노이즈가 제거된 적대적 예제와 정상 이미지에 대한 사전에 학습된 인공지능 모델의 n 번째 레이어의 출력값 차이를 나타낸다. L_{adv}는 다음과 같이 정의된다.

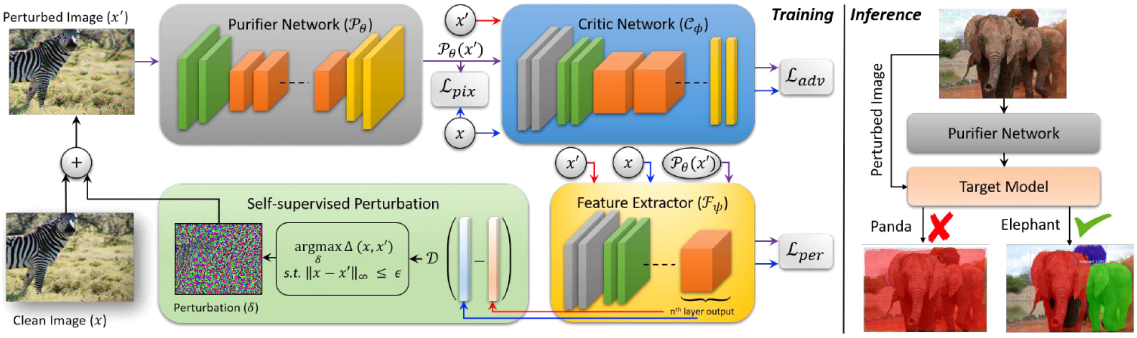
$$L_{adv} = -\log(\sigma(C_\phi(P_\theta(x')) - C_\phi(x))) \quad (6)$$

여기에서 x는 원본 이미지, x'은 적대적 예제, $C_\phi(\bullet)$는 Critic Network, $P_\theta(\bullet)$는 Purifier Network를 의미하며 σ는 시그모이드 레이어 (Sigmoid layer)를 의미한다. L_{img}는 다음과 같이 정의된다.

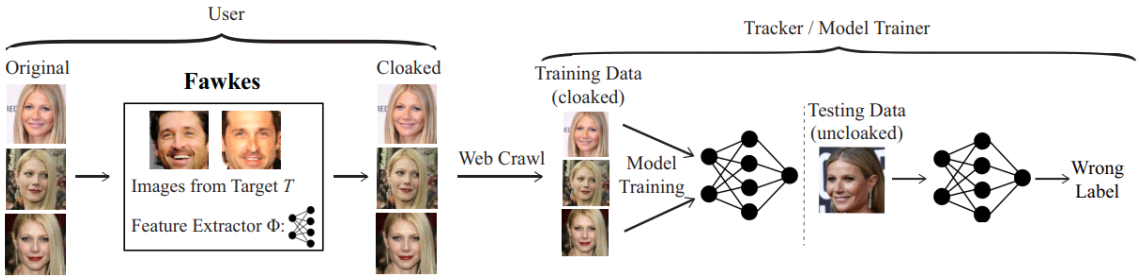
$$L_{img} = \| P_\theta(x') - x \|_2 \quad (7)$$

여기에서 $\| \bullet \|_2$는 l_2-norm을 의미한다. L_{feat}는 다음과 같이 정의된다

$$L_{feat} = \Delta(F_\psi(x), F_\psi(P_\theta(x'))) \quad (8)$$



〔그림 8〕 NRP 구조 〔16〕



〔그림 9〕 Fawkes 시스템 구조〔18〕

여기에서 $F_{\psi}(\bullet)$는 사전에 학습된 인공지능 모델의 n번째 레이어의 출력 값을 의미하며 Δ는 평균 절대 오차 (Mean Absolute Error)를 의미한다. 그림 8은 NRP 구조를 보여준다.

3.2. Membership Inference 방어

J. Jia 등 [17]은 적대적 예제 생성 원리를 기반으로 Membership Inference 공격을 방어하기 위한 MemGuard를 제안하였다. 공격자는 인공지능 모델의 학습 데이터로 사용되었는지 아닌지 판단하는 이진 분류 공격 모델을 가지고 있다. 하지만, 방어자는 공격자의 공격 모델을 알지 못하기 때문에 J. Jia 등은 방어자 스스로 공격 모델을 생성하여 자신의 공격 모델을 속이도록 인공지능 모델의 예측 값에 노이즈를 추가하며 안전성을 위해 다음 식을 최소화하여 공격자가 Membership Inference를 정확하게 하지 못하도록 최적의 노이즈를 찾는다.

$$M^{*}=\arg\min_{M}\left|E_{M}(g(s+n))-0.5\right| \tag{9}$$

여기에서 s는 인공지능 모델의 예측 값, n은 노이즈, $g(\bullet)$는 공격 모델, M은 노이즈 추가 메커니즘을 의미한다. 방어자는 데이터 유용성 (Utility)를 보장하기 위해 다음과 같이 원래 인공지능 모델의 예측 값과 노이즈가 추가된 예측 값의 차이가 ϵ을 넘지 않아야 한다.

$$D(s,s+n)\leq\epsilon \tag{10}$$

여기에서 $D(\bullet)$는 거리 함수이다. 따라서, MemGuard는 데이터의 안전성과 유용성 모두 보장하기 위한 방어 기술이다.

3.3. 프라이버시 보호

S. Shan 등 [18]은 인공지능 기반 얼굴인식 시스템에서 사용자의 얼굴에 대한 프라이버시를 보호하기 위해 사용자 얼굴 이미지에 다른 사람의 특징을 삽입하는 Fawkes를 제안하였다. Fawkes는 적대적 예제처럼 얼굴 이미지에 사람이 알아차릴 수 없는 노이즈를 삽입하는 방법으로 다음을 만족하는 노이즈를 찾는다.

$$\max_{\delta}D(\Phi(x),\Phi(x+\delta(x,x_{T}))) \tag{11}$$

여기에서 x는 원본 얼굴 이미지, x_{T}는 원본 얼굴이미지 x를 보호하기 위해 사용되는 다른 사람의 얼굴 이미지, $\delta(\bullet)$는 얼굴 이미지 x에 추가되는 노이즈, $\Phi(\bullet)$는 특징 벡터 추출 함수이며 $D(\bullet)$는 거리 함수이다. 또한, 노이즈는 다음의 조건을 만족해야한다.

$$\left|\delta(x,x_{T})\right|<\rho \tag{12}$$

여기에서 ρ는 최대 노이즈 량을 의미한다. 노이즈가 삽입된 얼굴 이미지는 온라인에 공개하고 공격자 혹은 인공지능 개발자는 노이즈가 삽입된 얼굴 이미지를 수집하여 인공지능 모델 학습에 사용한다. 인공지능 모델에 정상 이미지를 입력하면 잘못된 결과를 예측한다. 그림 9는 Fawkes 시스템 구조를 보여준다.

Ⅳ. 결 론

본 논문에서는 최신 인공지능 보안 공격 및 방어 기술에 대해 살펴보았다. 인공지능 보안 취약점을 공격하는 공격 기술도 다양하며 이를 막기 위한 방어 기술 또한 다양하고 활발히 연구되고 있다. 하지만, 다양한 인공지능

공격 유형을 모두 고려한 방어 기술은 아직 연구 개발이 미흡하여 모든 공격 유형에 대응 가능한 방어 기술에 대한 적극적인 연구가 필요하다.

참 고 문 헌

[1] K. He, X. Zhang, S.Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” European Conference on Computer Vision, Springer, pp. 630-645, 2016.

[2] C. Szegedy, S. Ioffe, and V. Vanhoucke, “Inception-v4, Inception-Resnet and The Impact of Residual Connections on Learning,” arXiv preprint arXiv:1602.07261, 2016.

[3] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” Proceedings of the IEEE International Conference on Computer Vision, IEEE, pp. 2961-2969, 2017.

[4] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, “Focal Loss for Dense Object Detection,” Proceedings of the IEEE International Conference on Computer Vision, IEEE, pp. 2980-2988, 2017.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All you Need,” Proceedings of the Neural Information Processing Systems, pp. 5998-6008, 2017.

[6] J. Matthew, A. Oprea, B. Biggio, C. Liu, C.N. Rotaru, and B. Li, “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning,” 2018 IEEE Symposium on Security and Privacy, IEEE, pp. 19-35, 2018.

[7] Y. Yao, H. Li, H. Zheng, and B.Y. Zhao, “Latent Backdoor Attacks on Deep Neural Networks,” Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 2041-2055, 2019.

[8] 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.

[9] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” arXiv preprint arXiv:1412.6572, 2014.

[10] 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, IEEE, pp. 2574-2582, 2016.

[11] N. Carlini, D. Wagner, “Towards Evaluating the Robustness of Neural Networks,” 2017 IEEE Symposium on Security and Privacy, IEEE, pp. 39-57, 2017.

[12] Z. Yang, J. Zhang, E.C. Chang, and Z. Liang, “Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment,” Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 225-240, 2019.

[13] L. Song, R. Shokri, and P. Mittal, “Privacy Risks of Securing Machine Learning Models against Adversarial Examples,” Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 241-257, 2019.

[14] H. Kwon, H. Yoon, and D. Choi, “Restricted Evasion Attack: Generation of Restricted-Area Adversarial Example,” IEEE Access, 7, pp. 60908-60919, 2019.

[15] W. Xu, D. Evans, and Y. Qi, “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks,” Symposium on Network and Distributed Systems Security, 2018.

[16] M. Naseer, S. Khan, M. Hayat, F.S. Khan, and F. Porikli, “A Self-supervised Approach for Adversarial Robustness,” roceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 262-271, 2020.

[17] J. Jia, A. Salem, M. Backes, Y. Zhang, and N.Z. Gong, “MemGuard: Defending against Black-Box

Membership Inference Attacks via Adversarial Examples," Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 259-274, 2019.

[18] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B.Y. Zhao, "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models," In 29th USENIX Security Symposium, pp. 1589-1604, 2020.

〈저 자 소 개〉

류 권 상 (Gwonsang Ryu)

학생회원

2016년 2월 : 공주대학교 응용수학과 학사

2018년 2월 : 공주대학교 대학원 융합과학과 석사

2018년 3월~2020년 8월 : 공주대학교 대학원 융합과학과 박사과정

2020년 9월~현재 : 숭실대학교 대학원 융합소프트웨어학과 박사과정

<관심분야> 인증, 이상거래탐지, 인공지능 보안

최 대 선 (Daeseon Choi)

종신회원

1995년 2월 : 동국대학교 컴퓨터공학과 학사

1997년 2월 : 포항공과대학교 컴퓨터공학과 석사

2009년 1월 : 한국과학기술원 전산학과 박사

1997년 1월~1999년 6월 : 현대정보기술 선임

1999년 7월~2015년 8월 : 한국전자통신연구원 인증기술연구실 실장/책임연구원

2015년 9월~2020년 8월 : 공주대학교 의료정보학과 부교수

2020년 9월~현재 : 숭실대학교 소프트웨어학부 부교수

2016년~현재 : 정보보호학회 차세대인증연구회장

<관심분야> 인증, 개인정보보호, 이상거래탐지, 의료정보보안, 머신러닝, AI보안