정 보 보 호 학 회 지
제 31 권 제2호, 2021. 4

딥러닝 모델에 대한 적대적 사례 기술 동향

권 현*, 김 용 철**

요 약

이미지 인식, 음성 인식, 텍스트 인식 등에서 딥러닝 모델이 좋은 성능을 보여주고 있다. 하지만 이러한 딥러닝 모델은 적대적 사례에 대하여 취약점을 갖고 있다. 적대적 사례는 원본 데이터에 최적의 노이즈를 추가하여 생성되며 사람이 보기에는 문제가 없지만 딥러닝 모델에 의해서 잘못 오인식되는 데이터를 의미한다. 적대적 사례에 대한 연구는 인공지능 분야와 보안 분야에서 관심을 받고 있으며 이미지, 음성, 텍스트 등으로 다양하게 연구가 진행 되고 있다. 이 연구에서는 적대적 사례에 대한 전반적인 기술 동향에 대해서 살펴보고자 한다.

Ⅰ. 서 론

최근 컴퓨팅 기술이 발전되고 많은 수의 데이터를 저장할 수 있는 클라우드 환경과 개인별 모바일 환경으로 데이터 수집이 용이하게 되었다. 이러한 환경 속에서 많은 수의 데이터를 학습하고 분류나 예측의 성능 향상을 가져온 딥뉴럴네트워크[1]에 대한 관심도 많아졌다. 특히, 환자의 데이터를 이용하여 병의 원인이나 진료에 활용 될 수 있는 딥러닝 모델[2]이나 표지판을 인식해야하는 자율주행차량에서의 딥러닝 기술 등이 각광받고 있다. 또한, 텍스트기반에서도 딥러닝 연구[3]가 진행되어 특정 문구를 적으면 관련 소설이나 관련 문장을 생성해주는 기술도 연구가 활발히 진행되고 있다.

하지만 이러한 딥뉴럴네트워크 기반의 딥러닝 모델은 보안상 취약점[4]이 존재한다. 딥러닝 모델에 대한 보안은 크게 causative attack과 exploratory attack으로 구분이 된다. causative attack[5][6]은 모델이 학습하는 데이터에 직접적으로 악의적인 데이터를 추가하여 모델의 정확도를 하락시키는 공격이 된다. 반면에 exploratory attack[7]은 이미 학습이 끝난 모델에 대하여 테스트 데이터를 조작하여 모델의 오인식을 유발하는 공격방법이다. causative attack은 모델이 학습해야하는 학습데이터에 접근해야하는 권한이 있어야하기 때문에 다소 현실성이 떨어지지만 exploratory attack은 테스트 데이터를 조작하는 방법으로 모델에 의해 잘못 오인식이 되는 샘플이기 때문에 현실성이 높은 공격방법이 된다. 이러한 exploratory attack의 대표적인 예로 적대적 사례가 있으며 보안학회나 인공지능학회에서 많은 관심을 받고 있는 연구주제이다.

본 연구에서는 적대적 사례에 관련한 연구내용의 정리와 향후 연구에 대한 내용을 다룬다. 2장에서는 이미지 기반의 적대적 사례에 대한 연구를 소개하고 3장에서는 다양한 도메인에서의 적대적 사례에 대한 연구를 다룬다. 그리고 마지막 4장은 결론으로 구성하였다.

Ⅱ. 이미지 기반의 적대적 사례

적대적 사례연구는 Szegedy 연구진[8]에 의해서 처음으로 제안이 되었다. 적대적 사례는 원본 데이터에 최소한의 노이즈를 추가하여 사람이 보기에는 문제가 없지만 딥러닝 모델에 의해서 잘못 오인식되게 하는 샘플이다. 적대적 사례의 기본 생성방법은 타겟 모델에 여러 번의 쿼리를 통해서 최소한의 노이즈를 업데이트 하여 최소한의 노이즈를 가지면서 모델의 오인식을 일으키는 샘플을 생성하는 것이다. 최소한의 노이즈는 L_0, L_2, L_∞ 방식을 이용하여 적대적 사례와 원본 샘플간의 차이를 측정한다[9]. 따라서 적대적 사례의 전제조건은 원본 샘플에 추가되는 최소한의 노이즈이어야 하고 모델의 오

본 연구는 육군사관학교 화랑대연구소의 2021년도(21-군학-5) 연구활동비 지원을 받아 연구되었음.
* 육군사관학교 전자공학과 (조교수, hkwon.cs@gmail.com)
** 육군사관학교 전자공학과 (정교수, kyc6454@kma.ac.kr), 교신저자

인식을 일으키는 조건을 만족해야한다. 최소한의 노이즈 기준은 보통 사람에 의해서 식별될 수 있는 노이즈인지가 중요한 점이고 보통 컬러이미지의 경우 거의 사람의 눈으로 식별되기 어려운 특징이 있다. 모델의 오인식을 일으키는 조건은 통상 해당 원본 클래스의 decision boundary를 벗어나는 지점이 된다. 따라서 적대적 사례는 decision boundary 근처 밖에 생성되어 원본 샘플간의 왜곡을 최소화하면서 모델에 의해 잘못 오인식이 된다. 접근 방법에 따라 적대적 사례 연구에 있어서 원본 샘플간의 왜곡 중점을 두는 연구들이 있고 반대로 모델의 decision boundary 측면에서 잘못 오인식 되는 적대적 사례에 중점을 둔 연구들이 있다.

이외에도 다양한 접근방법으로 적대적 사례를 연구하는 공격방법과 방어방법들이 있다. 2.1장에서는 적대적 사례의 분류에 대한 내용을 다루고 2.2장에서는 구체적인 적대적 사례 생성방법을 소개하며 2.3장에서는 적대적 사례의 대표적인 방어방법을 소개한다.

2.1. 적대적 사례의 분류

적대적 사례의 분류를 여러 가지 관점에 따라 분류할 수 있지만 이 기고문에서는 공격하고자 하는 모델의 정보양과 공격 목표에 따라서 분류하였다.

2.1.1 모델의 정보양

공격하고자 하는 모델의 정보양에 따라서 적대적 사례는 화이트 박스 공격과 블랙 박스 공격으로 구분된다. 화이트 박스[10]는 공격자가 타겟 모델에 대한 모든 정보를 알고 있는 상황에서의 공격을 의미한다. 따라서 타겟 모델의 아키텍처, 파라미터, 결과값의 확률값 등에 대한 모든 정보를 알고 있는 상태에서의 공격을 의미한다. 반면에 블랙박스 공격[11]은 공격자가 타겟 모델에 대한 정보가 없이 공격하는 방법이다. 블랙박스 공격의 경우 논문에 따라 입력값에 대한 결과값의 확률값을 아는 것까지 블랙박스로 보는 연구들도 있고 결과값의 확률값도 모르고 단순히 결과만 아는 가정을 블랙박스로 가정하기도 한다. 통상 블랙박스에서 입력값에 대한 각 클래스의 확률값을 알면 적대적 사례 생성 시 용이하기 때문에 확률값을 제공하지 않는 가정이 좀 더 어려운 가정이 된다.

현실적인 측면에서 화이트박스보다 블랙박스가 좀 더 가까우며, 이미지 부분에 있어서 화이트 박스 공격은 거의 100% 공격성공률을 가지고 있기 때문에 블랙박스 환경에서 높은 공격성공률을 갖는 연구들이 소개되고 있다. 블랙박스 공격의 경우 대표적으로 universal perturbation, transfer attack, substitute network 방법들이 있다. 먼저 universal perturbation 방법[12]은 다소 노이즈가 강하지만 모든 원본 데이터에 특정 noise를 추가하면 공격자가 원하는 타겟클래스로 오인식 되게 하는 방법이다. 이 방법은 여러 모델의 gradient 손실함수를 극대화하는 노이즈를 이용하여 딥러닝 모델에서 마지막 softmax layer에서 다른 클래스가 나올 확률이 크게 나오도록 노이즈를 원본 샘플에 추가함으로써, 일반적 공격이 가능한 적대적 사례를 생성한다.

두 번째로 transfer attack[13][14]은 임의의 한 모델에서 오인식 되도록 하는 적대적 사례는 알려지지 않은 모델에 대해서도 어느정도 공격효과가 있다는 방법이다. 이 방법은 기존에 한 개 모델에서 생성된 적대적 사례는 다른 모델에 공격효과가 있었지만 좀더 개선하여 여러 개의 모델을 동시에 오인식 하는 앙상블적 적대적 사례[15]은 임의의 다른 모델에 대해서 보다 높은 공격성공률로 오인식 공격이 가능한 것을 볼 수 있었다. transfer attack은 상당히 높은 공격성공률을 갖고 있고 이러한 이유는 특정 데이터에 최적화된 모델들은 어느정도 높은 정확도를 가지면 거의 유사한 decision boundary로 형성이 되기 때문에 이러한 패턴을 가지는 것을 확인할 수가 있다.

세 번째로는 substitute network를 이용한 방법[16]이다. 이 방법은 타겟 모델이 black box 모델인 경우에 유사한 substitute network를 먼저 생성한 후에 그 생성된 substitute network를 대상으로 생성된 적대적 사례는 black box 모델에 대하여 어느정도 공격 효과가 있다는 특징을 이용한 방법이다. 관련 논문에서는 MNIST의 경우 200번정도의 쿼리를 통해서 유사모델을 생성할 수 있다고 발표하였고 실제 이미지 머신러닝 서비스에도 어느 정도 공격효과가 있다는 것을 보여주었다. 이외에도 black box 환경에서 공격하는 방법들이 다양하게 연구가 되고 있다.

2.1.2 공격목표

적대적 사례에 대한 공격목표에 따른 targeted attack[17][18]과 untargeted attack[19][20]이 있다. 목표 적대적 사례는 공격자가 정한 특정 클래스로 오인식하는 적대적 사례를 의미한다. 반면에 비목표 적대적 사례는 원본 클래스가 아닌 임의의 잘못된 클래스로 오인식 하는 샘플을 의미한다. 비목표 적대적사례가 목표 적대적 사례보다 좀 더 쉬운 공격이며 적은 왜곡을 가지는 특징이 있다. 반면에 목표 적대적 사례는 공격자가 정한 클래스로 오인식 할 수 있는 정교한 공격이다. 통상적으로 적대적 사례의 연구는 비목표 적대적 사례를 먼저 연구를 한 후에 어느정도 연구성과가 많이 나오면 다음 단계로 목표 적대적 사례를 연구하는 순으로 연구가 된다.

2.2. 적대적 사례의 생성방법

적대적 사례의 생성방법은 다양한 방법이 있으나 통상적으로 fast gradient sign method (FGSM) [21] Deepfool[22], jacobian-based saliency map attack (JSMA)[23], carlini wagner (CW)[24]의 4가지 방법들이 대표적인 벤치마크로 사용이 된다. 이외에도 한 개의 픽셀을 조작하여 오인식을 일으키는 one pixel 공격[25]과 decision boundary 측면에서 black box attack일지라도 노이즈를 주면서 찾아가는 방법[26]이 있다. 최근에는 backward pass differentiable approximation (BPDA)방법[27]과 adaptive 방법[28]이 최신 공격방법으로써, 적대적 사례의 방어방법들에 대해서도 상당히 높은 확률의 공격성공률을 보여준다. 위의 언급된 방법 말고도 많은 공격 연구들이 소개되고 있다. 위의 언급된 공격방법들을 소개하면 아래와 같다.

먼저, fast gradient sign method (FGSM) [21]은 딥뉴럴네트워크의 gradient 손실함수를 계산한 후에 absent 방식으로 손실함수를 극대화하도록 epsilon 크기만큼 노이즈를 추가하여 딥러닝 모델이 적대적 사례를 잘못 오인식 하도록 하는 방법이다. 이 방법은 모델이 학습하는 gradient descent 방식의 반대방향으로 데이터에서 gradient absent 방식으로 잘못 오인식되는 샘플을 생성하는 간단한 원리이지만 상당히 좋은 성능을 보여주고 있고 컬러이미지의 경우에는 노이즈가 잘 식별되지 않는 특징이 있다. Deepfool 방법[22]은 FGSM 방법보다는 복잡한 구조로 비선형 뉴럴네트워크 구조에 여러번의 쿼리를 통하여 무표적 공격을 제안한 방법이다. gradient 계산이 아닌 여러 개의 점에서 decision boundary에 대해 수직으로 투영하고 적당한 노이즈를 추가하여 적대적 사례를 만드는 방법이다. 생성과정 간에 많은 시간과 타겟 모델에 의한 피드백이 필요한 점이 있다. jacobian-based saliency map attack (JSMA) 방법[23]은 뉴럴네트워크에서 입력과 결과값에 대한 맵핑을 통하여 적대적 사례를 생성하는 방법이다. 입력데이터를 미분하고 이에 대한 값을 saliency map에 맵핑한 후에 최대한 잘못 오인식 되는 방향으로 노이즈를 추가하여 적대적 사례를 생성하는 방법이다. 이 방법은 gradient 기반의 FGSM를 좀 더 개선한 방법으로 볼 수가 있다. Carlini Wagner (CW) 방법[24]은 화이트 공격으로 100% 성공률을 가지는 좋은 방법이다. 이 방법은 approximate 방법으로 최소한의 노이즈를 담당하는 손실함수와 공격 성공률을 높이는 손실함수의 합을 최소화함으로써 최적의 적대적 사례를 찾는 방법을 제안하였다. 원본 샘플간의 왜곡과 공격성공률 간의 가중치를 적절히 찾도록 설계를 하였으며 목표공격과 비목표 공격에 대하여 구현한 방법이다. 또한, 적대적 사례에 대한 방어가 있는 모델에 대해서도 신뢰값 k를 증가시켜 공격성공률을 높일 수 있도록 구성이 되어 있다.

위의 언급된 4가지 방법들은 대표적인 방법이고 그 이후에는 응용이 되거나 개선된 방법들이 소개되었다. one pixel 방법[25]은 CIFAR10 데이터셋을 대상으로 한 개의 pixel을 조작함으로써 잘못 오인식 하는 방법을 제안하였다. 다양한 모델에 대하여 1개 픽셀, 3개 픽셀, 5개 픽셀 등에 따라서 공격성공률에 대하여 분석을 하였고 공격원리에 대해서 설명하였다. 이 방법은 differential evolution (DE) 알고리즘을 사용하여 가장 classification score를 크게 변동시키는 픽셀 한 개를 찾는 알고리즘으로 다른 방법에 비해서 공격성공률이 낮다. 실제 코드를 구현하여도 타겟 대상이 되는 모델에 따라서 50% 이하의 공격성공률을 가지며 한 개 픽셀이 사람 눈에 많은 왜곡으로 보이는 특징이 있다. decision boundary 방법[26]은 블랙박스 공격으로 공격하고자 하는 클래스 이미지를 지정한 후에 그 이미지를 픽셀을 이동시키면서 잘못 오인식 되는 지점에서 더 이상 변경을 하지 않는 방법이다. 별도의 classification score와

같은 확률값이 필요하지 않고 단순히 잘못 오인식 되는 지점을 찾는 효율적인 방법이지만 최소한의 왜곡이라고 보기 어려운 점이 있다. 잘못 오인식하는 이미지가 같은 이미지 일지라도 공격하고자 하는 이미지의 선택에 따라서 decision boundary 근처에 원본 샘플과 가까운 지점에 이미지가 있다면 왜곡이 적을 수가 있지만 원본 샘플와 거리가 있는 이미지라면 왜곡이 많을 수도 있다.

적대적 사례에 대한 강건성을 가진 딥러닝 모델에 대해서도 무력화시킬 수 있는 공격방법이 제안되고 잇다. 먼저, gradient 계산을 어렵게 하여 적대적 사례를 생성하지 못하도록 하는 방어방법에 대하여도 공격할 수 있는 backward pass differentiable approximation (BPDA) 방법[27]이 있다. 이 방법은 이미지 훼손을 별로 주지 않는 전처리 함수를 미분이 가능한 함수로 대체함으로서 gradient 기반의 방어방법을 무력화하는 방법을 제안하였다. 이 방법을 통해서 ICML 인공지능 학회에서 발표되었던 9개 방어기법에 대하여 무력화 시켰다.

adaptive 방법[28]은 13가지 Top 컨퍼런스인 ICLR, ICML, NeurIPS에서 제안하였던 적대적 방어방법에 대하여 다양한 적대적 사례 공격으로 무력화하는 방법을 분석하였다. 한가지 공격이 모든 방어방법을 무력화 시킬 수는 없지만 특정 방어방법에 잘 맞는 공격방법을 선택하여 무력화 시킬 수 있는 것을 보여주었다. adaptive 방법의 향후연구는 어떻게 하면 좀 더 강건한 딥러닝 모델을 구축할지에 대한 분석도 소개되어있다.

2.3. 적대적 사례의 방어방법

적대적 사례의 방어방법[29][30][31]은 입력 데이터를 조작해서 노이즈를 제거[32]하거나 강건한 딥뉴럴네트워크를 구축하는 방법[33]으로 구분된다. 입력 데이터를 조작하여 노이즈를 제거하는 방식은 generative adversarial net(GAN)[34]을 이용하여 제거하는 방법이 있다. 또한 앙상블 방식으로 magnet 방식[35]으로 적대적 사례를 방어하는 방법이 있다. 이 방법은 적대적 사례 중에 왜곡이 심한 부분은 decision boundary에서 거리 측정하여 1차적으로 제거한다. 그 이후에 reformer를 통하여 적대적 사례와 가장 가까이 있는 원본 샘플과 대체하는 식으로 적대적 사례를 제대로 인식하는 방법이다. feature squeeze 방식[36]은 입력데이터를 조작하여 적대적 사례를 탐지하는 방법이다. 이 방법에서 공격 공간을 8bit 컬러비트로 공간을 줄이고 노이즈를 median smoothing과 non-local smoothing을 이용하여 입력데이터에 변조시켜 타겟 모델에게 제공한다. 이를 통해 입력데이터와 변조된 입력데이터의 출력값 차이를 계산하여 적대적 사례를 탐지하는 방법을 제안하였다.

반면에 강건한 딥뉴럴네트워크를 구축하는 것은 적대적 사례 생성을 어렵게 하여 방어하는 방법이다. distillation[33]은 두 개의 뉴럴네트워크를 이용하여 구성한다. 첫 번째 뉴럴네트워크에서 원본 샘플에 대한 각 클래스의 확률값을 라벨링하여 두 번째 뉴럴네트워크에 제공한다. 직접 label를 제공하는 대신 뉴럴네트워크를 통해 각 클래스의 확률값을 제공함으로써, 적대적 사례의 gradient 계산을 하지 못하게 하여 적대적 사례 생성을 방지하는 방법을 제안하였다. adversarial training 방법[37][38]은 간단한 방법이지만 효과적인 방어방법으로 소개되고 있다. 이 방법은 방어자가 알고 있는 임의의 모델을 대상으로 생성한 적대적 사례를 타겟모델에 추가적으로 제대로 인식되도록 학습하면, 알려지지 않은 적대적 사례에 대해서 타겟 모델이 강건성을 갖게 되는 특징을 이용한 방어방법이다. 이 adversarial training 방법는 간단하지만 방어 방법 중에 효과성이 있는 방법으로 알려져있다. 이 방법의 고려사항은 많은 적대적 사례를 학습하게 되면 원본 샘플에 대한 인식률이 저하가 될 수 있기 때문에 원본 샘플의 정확도를 유지하는 범위 내에서 적정한 양의 적대적 사례를 학습해야 한다. 이외에도 decision boundary를 수학적으로 계산하하여 적대적 사례 생성을 방지하는 방법[39]들이 소개되고 있다. 하지만 화이트박스로 적대적 사례를 생성할 경우 잘못 오인식 되는 경우가 많이 있으며, 방어 측면보다 공격측면에서 좀 더 유리한 측면[40]이 있다. 방어 측면에서 고려해야하는 점은 원본 샘플의 정확도를 유지하면서 적대적 사례를 잘 인식하거나 탐지하는 측면이 중요하다. 공격 측면에서는 모델의 오인식과 원본 샘플간의 왜곡 최소화로 2가지 측면이 중요하다. 아무리 좋은 방어방법이 있더라도 왜곡을 좀 더 주어서 모델의 오인식을 일으킬 수 있는 방법이 있으므로 공격이 유리한 측면이 있다.

Ⅲ. 다양한 도메인에서의 적대적 사례

적대적 사례에 대한 연구는 이미지 뿐만 아니라 음성

[41], 텍스트, 의료분야, CAPTCHA[42] 등에서도 확장되어 연구가 진행되고 있다.

3.1. 음성 분야

음성분야에서 적대적 사례에 대한 연구는 사람이 들을 수 없는 작은 노이즈에서 적대적 사례를 생성한 후에 음성 인식 모델에 의해서 잘못 오인식 되는 적대적 사례를 찾는 방법[43][44]을 사용한다. 이 연구는 화이트박스를 이용해서 컴퓨터 시뮬레이션상에서 적대적 사례를 연구 하였고 이후에 블랙박스 환경에서 적대적 사례를 생성하거나 외부 잡음이 있는 실제 환경에서 적대적 사례에 대한 연구가 진행되고 있다. 적대적 사례 방법은 아니지만 음성분야에서 오인식 공격 방법으로써, 특정 이상한 음으로 기계가 오작동을 하거나 사람이 들을 수 없는 영역대의 주파수를 이용하여 실제 휴대폰을 오인식 시키는 연구[45][46]들도 진행되고 있다.

3.2. 텍스트 분야

적대적 사례 연구는 텍스트 분야에서도 연구가 되고 있다. 이 분야에서는 문장에 대한 긍정 또는 부정에 대한 판별을 해주는 모델에 대하여 연구가 이뤄지고 있다. 따라서 사람이 보기에는 동일한 문장인데, 모델에 의해서 잘못 분류하는 방법으로 연구가 되고 있다. 문장에서 중요 단어에 대한 후보군을 선정 한 후에 대체함으로써 문법상 문제 여부, 유사도 등을 통과 후에 모델에 의해서 잘못 오인식 되는 확률값이 높은 단어로 대체함으로써 적대적 사례 공격이 가능하다. 최근에는 텍스트 분야에 BERT 기법을 이용한 방법[47][48]들이 소개가 되고 있다. 이러한 BERT 기반을 속이는 적대적 사례에 대한 연구가 진행되고 있다.

3.3. 의료 분야

의료분야에서 적대적 사례는 주로 segmentation 영역이나 판별하는 딥러닝 모델을 대상으로 한다. segmentation을 하는 딥러닝 모델에 대해서 segmentation이 잘 되지 않도록 적대적 노이즈를 추가하거나 잘못 되도록 하는 방법[49][50]들이 제안되었다. 또한 질병을 보고 어떤 종류인지 분별하는 문제에 있어서도 적용하였고 이미지 기반의 적대적 사례 방법을 응용하여 적대적 사례를 생성하기도 하였다.

Ⅳ. 결 론

이 연구에서는 적대적 사례의 공격과 방어 그리고 적용할 수 있는 도메인에 대해서 적대적 사례의 기술 동향을 살펴보았다. 전반적으로 적대적 공격이 방어에 비해서 좀 더 유리한 특징이 있지만 점차적으로 적대적 사례에 대한 방어기법에 대한 연구 필요성이 강조되고 있다. 또한, 적대적 사례연구는 여러 도메인에서 공통적으로 적용될 수 있는 공격연구와 방어솔루션에 대한 연구가 동시에 이뤄지고 있다.

참 고 문 헌

[1] Liu, Weibo, et al. "A survey of deep neural network architectures and their applications." Neurocomputing 234 (2017): 11-26.

[2] Parvathy, Velmurugan Subbiah, Sivakumar Pothiraj, and Jenyfal Sampson. "Optimal Deep Neural Network model based multimodality fused medical image classification." Physical Communication 41 (2020): 101119.

[3] Jahangir, Rashid, et al. "Text-independent speaker identification through feature fusion and deep neural network." IEEE Access 8 (2020): 32187-32202.

[4] Xue, Mingfu, et al. "Machine learning security: Threats, countermeasures, and evaluations." IEEE Access 8 (2020): 74720-74742.

[5] Yang, Chaofei, et al. "Generative poisoning attack method against neural networks." arXiv preprint arXiv:1703.01340 (2017).

[6] Kwon, Hyun, Hyunsoo Yoon, and Ki-Woong Park. "Multi-targeted backdoor: Indentifying backdoor attack for multiple deep neural networks." IEICE Transactions on Information and Systems 103.4 (2020): 883-887.

[7] Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vi-

sion: A survey." Ieee Access 6 (2018): 14410-14430.

[8] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[9] Carlini, Nicholas, et al. "Provably minimally-distorted adversarial examples." arXiv preprint arXiv:1709.10207 (2017).

[10] Athalye, Anish, and Nicholas Carlini. "On the robustness of the cvpr 2018 white-box adversarial example defenses." arXiv preprint arXiv:1804.03286 (2018).

[11] Jiang, Linxi, et al. "Black-box adversarial attacks on video recognition models." Proceedings of the 27th ACM International Conference on Multimedia. 2019.

[12] Zhao, Yuhang, et al. "An Universal Perturbation Generator for Black-Box Attacks Against Object Detectors." International Conference on Smart Computing and Communication. Springer, Cham, 2019.

[13] Huang, Qian, et al. "Enhancing adversarial example transferability with an intermediate level attack." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[14] Kwon, Hyun, et al. "Advanced ensemble adversarial example on unknown deep neural network classifiers." IEICE TRANSACTIONS on Information and Systems 101.10 (2018): 2485-2500.

[15] Hang, Jie, et al. "Ensemble adversarial black-box attacks against deep learning systems." Pattern Recognition 101 (2020): 107184.

[16] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017.

[17] Kwon, Hyun, et al. "Multi-targeted adversarial example in evasion attack on deep neural network." IEEE Access 6 (2018): 46084-46096.

[18] Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.

[19] `Wu, Aming, et al. "Untargeted adversarial attack via expanding the semantic gap." 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019.

[20] Kwon, Hyun, et al. "Random untargeted adversarial example on deep neural network." Symmetry 10.12 (2018): 738.

[21] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. CoRR, abs/1412.6572, 2014

[22] Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[23] Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016.

[24] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). IEEE, 2017.

[25] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." IEEE Transactions on Evolutionary Computation 23.5 (2019): 828-841.

[26] He, Warren, Bo Li, and Dawn Song. "Decision boundary analysis of adversarial examples." International Conference on Learning Representations. 2018.

[27] Athalye, Anish, Nicholas Carlini, and David Wagner. "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples." International Conference on Machine Learning. PMLR, 2018.

[28] Tramer, Florian, et al. "On adaptive attacks to adversarial example defenses." arXiv preprint arXiv:2002.08347 (2020).

[29] Kwon, Hyun, Hyunsoo Yoon, and Ki-Woong Park. "Acoustic-decoy: Detection of adversarial examples through audio modification on speech recognition system." Neurocomputing 417 (2020): 357-370.

[30] Kwon, Hyun, and Jun Lee. "AdvGuard: Fortifying Deep Neural Networks against Optimized Adversarial Example Attack." IEEE Access (2020).

[31] Kwon, Hyun, et al. "Classification score approach for detecting adversarial example in deep neural network." Multimedia Tools and Applications (2020): 1-22.

[32] Liang, Bin, et al. "Detecting adversarial image examples in deep neural networks with adaptive noise reduction." IEEE Transactions on Dependable and Secure Computing (2018).

[33] Papernot, Nicolas, et al. "Distillation as a defense to adversarial perturbations against deep neural networks." 2016 IEEE symposium on security and privacy (SP). IEEE, 2016.

[34] Samangouei, Pouya, Maya Kabkab, and Rama Chellappa. "Defense-gan: Protecting classifiers against adversarial attacks using generative models." arXiv preprint arXiv:1805.06605 (2018).

[35] Meng, Dongyu, and Hao Chen. "Magnet: a two-pronged defense against adversarial examples." Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. 2017.

[36] Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." arXiv preprint arXiv:1704.01155 (2017).

[37] Tramèr, Florian, et al. "Ensemble adversarial training: Attacks and defenses." arXiv preprint arXiv:1705.07204 (2017).

[38] Kwon, Hyun, and Jun Lee. "Diversity Adversarial Training against Adversarial Attack on Deep Neural Networks." Symmetry 13.3 (2021): 428.

[39] Shumailov, Ilia, et al. "Towards certifiable adversarial sample detection." Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security. 2020.

[40] Carlini, Nicholas, and David Wagner. "Adversarial examples are not easily detected: Bypassing ten detection methods." Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017.

[41] Kwon, Hyun, et al. "Selective audio adversarial example in evasion attack on speech recognition system." IEEE Transactions on Information Forensics and Security 15 (2019): 526-538.

[42] Kwon, Hyun, Hyunsoo Yoon, and Ki-Woong Park. "Robust CAPTCHA Image Generation Enhanced with Adversarial Example Methods." IEICE TRANSACTIONS on Information and Systems 103.4 (2020): 879-882.

[43] Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.

[44] Kwon, Hyun, Hyunsoo Yoon, and Ki-Woong Park. "POSTER: Detecting audio adversarial example through audio modification." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.

[45] Zhang, Guoming, et al. "Dolphinattack: Inaudible voice commands." Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.

[46] Carlini, Nicholas, et al. "Hidden voice commands." 25th {USENIX} Security Symposium ({USENIX} Security 16). 2016.

[47] Garg, Siddhant, and Goutham Ramakrishnan. "Bae: Bert-based adversarial examples for text classification." arXiv preprint arXiv:2004.01970 (2020).

[48] Jin, Di, et al. "Is bert really robust? a strong baseline for natural language attack on text classification and entailment." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 05. 2020.

[49] Finlayson, Samuel G., et al. "Adversarial attacks against medical deep learning systems." arXiv preprint arXiv:1804.05296 (2018).

[50] Ozbulak, Utku, Arnout Van Messem, and Wesley De Neve. "Impact of adversarial examples on deep learning models for biomedical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.

〈저자소개〉

권 현 (Hyun Kwon)

종신회원

2010년 2월 : 육군사관학교 수학과 학사 졸업

2015년 8월 : 한국과학기술원 전산학부 석사 졸업

2020년 2월 : 한국과학기술원 전산학부 박사 졸업

2020년 2월~현재 : 육군사관학교 전자공학과 조교수

<관심분야> 인공지능 보안, 시스템 보안, 머신러닝

김 용 철 (Yongchul Kim)

1998년 2월 : 육군사관학교 전자공학 학사 졸업

2001년 11월 : University of Surrey 전자공학과 석사 졸업

2012년 1월 : North Carolina State University 전자공학과 박사 졸업

2012년 1월~현재 : 육군사관학교 전자공학과 교수

<관심분야> 무선통신네트워크, 통신공학. 전자공학