chapter 2

적대적 머신러닝 기술 동향

강대기 ‖ 동서대학교 교수

I. 서론

2016년 적대적 생성 신경망(Generative Adversarial Network: GAN)을 개발한 것으로 유명한 이안 굿펠로우는 사진에 눈에 보이지 않는 잡음을 슬쩍 넣는 것만으로도 최고 수준으로 학습된 합성곱 신경망을 기만할 수 있음을 보였다[1]. 굿펠로우는 연구에서 팬더 사진에 눈에 보이지 않는 잡음을 살짝 더했다. 여전히 팬더 사진이었지만, 신경망은 긴팔원숭이로 인식했다. 이러한 허점은 급기야 2017년 큐슈 대학교 연구진에 의해 심각한 모습으로 드러나게 된다. 딥러닝과 인공지능 기술의 금자탑과도 같은 합성곱 신경망이 주어진 사진에서 점 하나만 찍으면 배를 자동차로 인식하고, 말을 개구리나 개로 인식하고, 개를 고양이로 잘못 인식하는 오류를 보인 것이다[2]. 2017년 10월 MIT의 해커팀인 Lab6는 거북이 모형을 3D 프린터로 출력하였는데, 누가 봐도 거북이인 모형을 많은 인공지능 시스템들은 라이플총으로 잘못 인식하였다[3].

이러한 합성곱 신경망의 허점은 합성곱 신경망이 자동주행시스템이나 군사용 장비에서 사용될 경우 큰 문제를 야기할 수도 있다. 스톱 사인을 좌회전으로 오인한다든지 적 비행

* 본 내용은 강대기 교수(☎ 051-320-1724, dkkang@dongseo.ac.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

OPEN 출처표시 상업용금지 공공누리 공공저작물 자유이용허락

기를 아군으로 오인한다든지 한다면, 그건 돌이킬 수 없는 심각한 결과를 초래할 수도 있기 때문이다. 이런 식으로 합성곱 신경망이나 다른 형태의 인공지능 시스템이 사물을 인식하려 하는 경우에, 사람이 느끼지 못하는 잡음과 같은 은밀한 수단으로, 그 신경망으로 하여금 명백히 해당 사물인데도 불구하고 그것을 다른 것으로 오인 또는 착각하게 하는 공격이 가능하다. 이는 기본적으로 신경망의 허점을 이용하는 공격이며, 인공지능 보안 전문가들은 이를 기만 공격(deception attack)이라고 한다.

본 고에서는 적대적 머신러닝을 소개하고 관련 연구들에 대해 논하고자 한다. 이를 위해 적대적 머신러닝 분야의 최고 전문가인 구글 브레인의 니콜라스 칼리니 박사의 적대적 머신러닝 추천 논문들(Adversarial Machine Learning Reading List)[4]에서 소개된 연구들을 하나하나 분석하고 이를 토대로 본 고를 구성하였다.

본 고의 II장에서는 적대적 머신러닝을 이해하기 위한 예비지식에 해당하는 연구들을 소개한다. III장에서는 블랙박스 공격에 악용되는 신경망의 특성인 전이가능성(transferability)에 대한 연구들을 소개한다. IV장에서는 신경망에 대한 기만 공격 기술들에 대해 소개한다. V장에서는 기만 공격에 대해 보다 굳건한 신경망을 만들기 위한 방어 기술들에 대해 알아본다. 끝으로 VI장에서는 본 고의 결론 및 시사점을 제시한다.

II. 적대적 머신러닝 기술 연구 배경

인공지능, 컴퓨터비전, 머신러닝, 딥러닝의 종합적인 분야인 적대적 머신러닝이 연구하고자 하는 분야는 전술한 바와 같이, 공격하고자 하는 신경망이 제대로 분류해야 할 적법한 샘플(또는 예제)에 눈에 보이지 않는 잡음을 소량 가했을 때, 그 신경망이 잘못 인식하게 되는 현상과 깊은 관련이 있다. 예를 들어, 신경망이 학습한 매니폴드(manifold)의 바깥쪽 가까이에 놓여 있는 샘플들이 있는데, 이러한 샘플들을 적대적 예제(adversarial example)라고 한다. 이러한 적대적 예제들을 찾기 위한 방법으로 주로 기존의 제대로 학습된 원래의 훈련 샘플들에서 적대적 예제들이 있는 방향으로 조금씩 변화를 주는 방법들이 주로 사용되고 있다.

적대적 머신러닝에 대한 연구는 2013년 Biggio 등이 자신들의 논문[5]에서 최초로 적대적 예제를 언급하였다. 그들은 논문에서 높은 신뢰도로 만들어지는 회피 공격(evasion

attacks)을 처음으로 소개하였다. 또한, 공격자가 공격을 위해 학습하는 대리 학습자(surrogate learners)의 개념도 제시하였는데, 이는 적대적 머신러닝에서 공격자가 신경망 공격을 위해 스스로 학습시키는 대체 모델(substitute models)의 개념이다. 이 두 가지 개념은 지금까지 심층 신경망에 대한 공격 기술에서 계속 사용되고 있다.

2013년 Szegedy 등[6]은 신경망에 기존의 상식과 반하는 흥미로운 특성들이 있음을 지적하였다. 첫째는 신경망 내의 고차원 유닛 하나와 신경망 내의 고차원 유닛들의 임의의 선형적인 조합들 간에는 큰 차이가 없다는 것이다. 둘째는 심층신경망이 학습하는 입력-출력 매핑이 상당한 수준으로 불연속적이라는 점이다. 바로 이 점이 이미지에 눈에 보이지 않는 작은 섭동(perturbation)을 추가하는 것만으로도 신경망의 예측에 심각한 오류를 불러일으킬 수 있는 것이다. 게다가, 이러한 섭동은 해당 학습에만 임의로 적용되는 것이 아니고, 다른 데이터 세트로 학습된 다른 신경망에도 적용될 수 있다는 것이다. 2014년 Goodfellow 등[7]도 적대적 예제들은 오버피팅이나 비선형성의 문제라기보다는 신경망의 선형적인 본질에 있다고 주장하였다.

III. 신경망의 전이가능성

이러한 적대적 예제들은 신경망의 구조와 학습의 본질적인 문제로 보는 주장이 더 타당하다고 여겨지고 있다. 예를 들어, 현재 토론토 대학의 교수인 Papernot 등[8]은 서로 다른 구조를 가지는 두 개의 모델이 서로 다른 데이터 세트로 학습된 경우에도, 같은 섭동이 영향을 줄 수 있다는 점, 이른바 전이가능성을 이용하여 블랙박스 공격이 가능하다는 점을 지적하였다. 참고로 화이트 박스 공격과 블랙박스 공격이 있는 데, 화이트 박스 공격에서는 희생자 모델의 패러미터에 대한 정보를 알 수 있는 반면, 블랙박스 공격에서는 이를 알 수 없다. 적대적 머신러닝 공격이 무서운 점은 블랙박스인 경우에도 전이가능성을 통해 상대 모델을 효과적으로 공격할 수 있다는 점이다.

이렇게 전이가능성을 이용한 블랙박스 공격에서는 공격하고자 하는 이른바 희생자 모델이 아닌, 공격자가 직접 만들어 학습하여 그로부터 만들어 낸 대체 모델로부터 적대적 예제들을 생성한다. 이 생성된 적대적 예제들을 가지고 희생자 모델을 공격하는 것이다. 이 때, 대체 모델을 학습하기 위해 희생자 모델과 동일하거나 비슷한 학습 데이터 세트를

가지고 학습을 하거나, 심지어 인위적으로 만든 학습 데이터를 희생자 모델에 입력하여 그 출력을 사용하는 방법, 즉 희생자 모델의 출력을 오라클(oracle)로 사용하는 방법도 가능하다.

실제로 Papernot 등[8]은 이러한 기법을 활용하여 저수지 샘플링(Reservoir Sampling)을 통해 효율적으로 학습을 수행하여 아마존과 구글의 머신러닝 시스템에 대한 블랙박스 공격에 성공하였다.

Liu 등[9]은 대규모의 모델과 대규모의 데이터 세트에 대해 전이가능성을 연구하였다. 참고로 이 연구의 교신저자인 캘리포니아 대학 버클리의 Dawn Song 교수는 오랫동안 컴퓨터 보안을 연구하였다. Liu 등[9]에 따르면, 전이가능성을 통해 non-targeted 적대적 예제를 찾는 건은 그다지 어렵지 않았다. 반면 targeted 적대적 예제를 찾는 건 상당히 어려웠다.

Non-targeted 적대적 예제란, 공격자가 희생자 모델의 예측이나 분류를 특정한 클래스로 유도하지 않고, 단순히 올바른 예측과 다른 결과를 나오게 만드는 샘플을 말한다. 올바른 예측과 틀리기만 하면 되는 것이다. 반면, Targeted 적대적 예제는 공격자가 희생자 모델의 예측이나 분류를 공격자가 원하는 특정한 클래스로 유도하는 틀린 예측이나 오분류를 유발하게 만드는 샘플이다.

Targeted 적대적 예제를 찾는 문제가 쉽지 않으므로 Liu 등[9]은 이를 해결하기 위해 앙상블 기반 접근 방법을 사용하여 Targeted 적대적 예제를 찾아내는 데 성공하였다. 또한, 전이가능한 적대적 예제들에 대한 기하학적인 연구를 수행하였다.

Moosavi-Dezfooli 등[10]은 현재 일반적으로 연구되는 심층신경망 구조에서 입력 데이터에 상관 없이(즉 입력 이미지에 상관없이) 모든 자연 이미지들을 오분류하게 만드는 유니버셜 적대적 섭동(universal adversarial perturbation)이 존재함을 보이고, 이를 위한 알고리즘을 제시하였다.

이러한 전이가능성은 현존하는 모든 심층신경망들이 기본적으로 취약점을 가지고 있고, 이를 기반으로 공격자가 적대적 예제를 만들어서 공격하는 것이 어렵지 않은 일이라는 사실을 보여준다. 심지어는 모든 입력 데이터에 대해 동작하는 적대적 예제의 존재도 가능함을 시사한다.

IV. 적대적 머신러닝 공격 기술

1. 공격 기술

이제 적대적 머신러닝을 위한 핵심적인 공격 기술들에 대해서 알아본다. 기본적으로 공격이라는 것은 희생자의 심층신경망 모델이 있을 때, 주어진 데이터(또는 이미지)에 대해 희생자 모델이 오인식하도록 만드는 적대적 섭동을 생성하여 이를 주어진 데이터(또는 이미지)에 추가하여 적대적 예제를 만드는 것을 말한다. 오인식을 유도하는 방향에 있어서도 특정한 클래스 라벨로 오인식을 유도하는 공격을 targeted 공격(targeted attack)이라 하며, 단순하게 오인식을 유도하는 공격은 untargeted 공격(untargeted attack)이라 한다.

Goodfellow[1] 등은 초기의 공격 기법인 fast gradient sign method(FGSM)를 제안하였다. 그래디언트 사인(gradient sign)을 통해 최적 섭동 벡터의 예측값을 구하고, 이를 따라가는 방식을 취한다.

이로부터 파생된 주된 공격 기법들을 살펴보면, Papernot 등[11]은 심층신경망의 입력과 출력 간의 매핑에 기반하여 기존의 입력 데이터의 미분의 saliency map을 최대화하는 방향을 추가하여 적대적 예제를 만들어내는 방안으로 Jacobian-based Saliency Map Attack(JSMA)을 제시하였다.

Moosavi-Dezfooli 등[12]은 효율적으로 DeepFool이라는 알고리즘을 개발하였다. 주어진 점에서 결정경계(decision boundary)에 대한 수직으로 투영(projection)하고 적당한 상수를 더하여 적대적 예제를 만드는 방법으로, [1]의 방법이 그래디언트의 사인을 따라가는 예측값에 근거한다면, 이 방법은 보다 최적의 해를 구할 수 있다. 실은 공격하고자 하는 신경망이 비선형이므로, 이 방법은 해당 해를 구한 후, 그 해에서 다시 결정경계에 투영하고 따라가는 과정이 반복적으로 진행되게 된다.

Carlini와 Wagner[13]는 다양한 목적함수들을 제한조건(constraint)으로 하고, L_p 기반의 거리 메트릭(distance metric)을 최소화하는 좀 더 일반화된 최적화 방안을 제시하였고, 여러 목적함수들과 L_p 거리들을 테스트하였다. 실제 그들의 논문에서는 3개의 거리 메트릭과 7개의 목적함수가 테스트되었다.

2. 방어 기술에 대해 심화된 공격 기술

방어 기술에 대응하여 공격 기술을 진화하는 방안에 대한 연구도 수행되었다. Athalye 등[14]은 그래디언트를 모호하게 만드는 방식으로 적대적 공격에 대해 방어를 하는 기술들에 대한 공격 방법으로 Backward Pass Differentiable Approximation(BPDA)을 제시하였다. 이 방법에서는 그래디언트를 모호하게 하기 위해 사용되는 입력에 대한 전처리 함수를 미분 가능한 다른 함수로 대신하는 방법을 사용한다. 이렇게 미분 가능한 다른 함수로 대신하는 것이 그다지 어렵지 않은 이유는 올바른 예측을 위해서는 전처리 함수가 지나치게 입력 데이터(또는 이미지)를 훼손할 수는 없기 때문이다. 따라서 이런 전처리 함수들은 항등함수(identity function)에 가깝게 구성된다. Athalye는 Carlini와 Wagner 교수와 같이 이 BPDA 공격 기법에 대한 연구를 수행하였고, 실제로 논문이 발표된 ICML 2018에서 역시 발표된 9개의 최첨단 방어 기법들 중 7개를 우회하는 데 성공한다.

Uesato 등[15]은 적대적 위협(adversarial risk)의 개념을 최악의 입력에 대해서도 강건한 결과를 보이는 목적함수로 설정하는 데, 이는 쉽게 구하기 어려운 이론적인 함수이다. 따라서, 당장 연산 가능한 대체 함수로 근사화하기 위해 현재까지 알려진 공격 방법과 성능평가 방법을 사용하였다. 그래디언트 기반의 최적화 공격에서는 ADAM 최적화를 사용하였고, 그래디언트에 근거하지 않은 경우에는 Simultaneous perturbation stochastic approximation(SPSA) 최적화 방법을 사용하였다.

3. 제한된 위협 모델 공격

Chen 등[16]은 전이가능성에 근거한 대체 모델을 이용하는 방식이 아닌, 0차 최적화(Zeroth Order Optimization: ZOO) 기반의 공격을 제안하였다. 이 방식은 0차 스토캐스틱 좌표 하강(zeroth order stochastic coordinate descent)과 공격 공간의 차원 축소(dimension reduction), 적당한 크기로 축소된 차원 공간에서의 계층적 공격(hierarchical attack), 그리고 섭동을 추가할 결정적인 픽셀을 찾기 위한 임포턴스 샘플링(importance sampling)을 통해 목표가 되는 심층신경망의 그래디언트를 직접적으로 구한다.

Brendel 등[17]에 의하면 공격 기법들은 그래디언트 기반 공격, 스코어 기반 공격, 전이

기반 공격들로 나누어진다. Brendel 등[17]은 결정 기반 공격이라는 새로운 공격 방법인 Boundary Attack을 제시하였다. 이 공격 방법도 대체 모델을 사용하지 않고 희생자 모델의 최종적인 결정에만 근거하여 공격하는 방법이다.

Boundary Attack은 적대적 예제 공간과 그렇지 않은 원래의 샘플 공간의 경계에서 리젝션 샘플링을 수행한다. 각 단계에서 i.i.d.(independent and identically distributed, 독립항등분포) 특성을 가지는 가우시안 분포에서의 랜덤 값에 근거하여 방향을 정하고, 그 방향으로 원래 이미지 주변의 구체(sphere)에 투사시킨다. 그 다음 원래의 샘플 공간으로 투사된다.

4. 물리적 공격

지금까지 논의한 공격 시나리오는, 모델을 학습하거나 학습된 모델의 예측 및 추론 과정에서 대부분 파일 형태로 준비된 데이터(또는 이미지)에 섭동 잡음을 첨가하고, 이를 학습시키거나 추론의 입력으로 사용하는 형태이다. 이러한 시나리오는 다소 비현실적으로 보일 수 있다. 카메라 및 기타 센서의 신호를 입력으로 사용하는 시스템과 같이 물리적 세계에서 작동하는 시스템의 경우 항상 이런 식으로 섭동 잡음의 입력이 수월하지 않을 수 있다.

따라서 이른바 인공지능 연구자들이 즐겨 사용하는 용어(반면, Dijkstra 교수는 그렇게 싫어했던 용어)인 '현실세계' 즉 real world에서는, 이러한 공격 시나리오에 대해서는 크게 걱정할 필요가 없을 거라 생각할 수 있다. 앞으로 언급할 논문들은 이러한 생각이 틀렸음을 보여준다.

Kurakin 등[18]은 이미 학습된 ImageNet Inception 분류기를 이용하여 적대적 예제를 만들어 내고, 그 적대적 예제를 휴대폰 카메라를 통해 입력하였다. 휴대폰 카메라를 통해 입력된 이미지는 적대적 예제를 만들어내는 데 사용된 바로 그 ImageNet Inception 분류기로 분류를 시도하였다. 결과는 휴대폰 카메라를 통해 입력되는 경우에도 적대적 예제는 공격하고자 하는 모델을 효과적으로 속였다는 점이다. 적대적 예제를 휴대폰 카메라를 통해 입력받기 위해 적대적 예제를 종이에 출력한 경우, 종이에 출력한 것을 사진으로 찍은 경우, 그리고 특정 이미지 부분을 crop 즉 잘라낸 경우에 대해 실험을 수행하였으며, 분류기는 휴대폰 카메라의 앱에서 임베딩되어 사용되었다. 즉, 이미지의 분류는 휴대

폰 내부에서 이루어졌다.

MIT의 학생 서클인 LabSix에서는 3차원 프린터를 통해 진정한 3차원 물리적인 적대적 예제를 만들어냈다[3]. 기존의 연구에서는 적대적으로 생성된 이미지들이 이미지 변환을 거치면 적대적 예제로의 성질이 사라지는 현상이 발생하였다. 이렇게 이미지 변환에 강건한 적대적 예제를 만들기 위해 Expectation Over Transformation(EOT)이라는 방법을 제시하였다. 기본적으로 적대적 예제를 생성하는 방안을 로그우도를 최대화하는 형태로 최적화 설정을 할 수 있다. EOT에서는 하나의 데이터 예제에 대해 로그우도를 최대화하는 대신, 이미지 변환 함수를 통과한 이미지들의 기댓값을 최대화한다. 이 최대화 과정은 이미지 변환을 위한 함수 분포에 대해 이미지 변환을 통과한 적대적 예제와 역시 이미지 변환을 통과한 원래 샘플 간의 거리의 기댓값을 특정 입력값 입실론 이하로 하는 제한조건에서 수행된다.

Eykholt 등[19]은 Robust Physical Perturbations(RP2) 알고리즘을 제안하였다. RP2 알고리즘은 기본적으로 섭동 영상과 원래 영상 간의 거리를 최소화하되, 분류기의 출력은 공격자가 원하는 클래스를 유지하는 제한조건을 가지는 최적화 문제를 기반으로 구성된다. 섭동을 물리적인 대상에 넣기 위해서 마스크 개념으로 물리적 대상에 투영하였으며, 이러한 투영이 사람의 관점에서 지나치게 이상하게 느끼지 않도록 일반적인 낙서나 잡음과 비슷하게 구성하였다. 투영한 결과가 나타나는 컬러 공간에서 투영 값이 범위를 벗어날 수도 있다. 이런 경우에는 컬러 프린터가 제대로 된 마스크를 만들어내지 못할 것이다. 이러한 fabrication error를 보정하기 위해 최적화 설정에서 Non-Printability Score(NPS) 항(term)을 추가하였다.

[19]의 연구는 미시건 대학 앤 아버, 워싱턴 대학, 캘리포니아 대학 버클리, 그리고 스투니 브룩 대학에서 공동으로 연구된 것인데, 연구자들은 적대적 예제에 대해 현재 표준화된 테스트 방법이 없다는 점을 지적하였다. 따라서 실험실 및 현장 테스트로 구성된 강력한 물리적 적대적 예제에 대한 2단계 평가 방법론을 제안하였다.

V. 적대적 머신러닝 방어 기술

1. 판별 기술

이제 적대적 공격에 대한 방어 기술을 알아보고자 한다. 우선 적대적 섭동을 판별하는 판별 기술에 대해 알아본다.

Metzen 등[20]은 적대적 섭동을 포함하는 데이터와 진짜 데이터를 구별하는 이진 분류 작업을 훈련시킨 작은 '검출기'를 기존의 심층신경망의 서브 네트워크로 사용하여 해당 심층신경망을 강화하는 방법을 제안하였다. 이 방법에 따르면 일단 일반 데이터 세트로 심층신경망을 학습시킨다. 그리고 나서 학습 데이터의 각 이미지마다 학습된 심층신경망에 대한 적대적 예제를 생성해 낸다. 그리고 나서 서브 네트워크로서의 검출기를 학습시킨다. 이 검출기를 학습시킬 때 적대적 예제일 확률과 학습 데이터 레이블의 크로스 엔트로피(cross entropy)를 최소화하도록 학습시킨다.

Feinman 등[21]은 심층신경망이 적대적 예제와 정상 예제 그리고 단순히 잡음이 있는 예제를 구별할 수 있는지를 연구했다. 이를 위해 적대적 예제에 대한 모델의 신뢰도에 대해 조사하는 데, 조사하는 방법은 드롭아웃 신경망에서 구할 수 있는 베이지안 불확실성 추정치를 관측하고, 모델에 의해 학습된 심층 특징의 서브스페이스에서 밀도 추정을 수행하는 방법이다. 드롭아웃을 통해 네트워크에 무작위성을 추가함으로써, 자연 이미지는 그대로 예측되지만, 적대적 예제는 그렇게 되지 않도록 하려는 것이다. 중요한 점은 결과적인 시스템은 공격 기법에 대해 독립적인 시스템이 된다는 것이다. 적대적 예제를 생성할 때에는 적대적 예제가 서브매니폴드에서 멀리 떨어진 경우, 특정 서브매니폴드에 가깝지만 그 매니폴드에 속하지 않으며 결정 경계에서도 멀리 떨어진 경우, 특정 서브매니폴드에 가깝지만 그 매니폴드에 속하지 않으며 결정 경계와 가까운 경우로 나누었다.

그런데, Carlini와 Wagner[22]는 자신들의 논문에서 대부분의 판별 기술들이 쉽게 우회 가능함을 보였다. [22]에서는 [20]의 방어 기법에 대해서는 기본적으로 판별기의 학습이 불안하거나 용이하지 않았다고 보고하고 있다. 그러나 심지어 판별기가 학습되어도 완벽한 지식 상황에서의 공격(Perfect Knowledge Attack) 시나리오와 제한된 지식 상황에서의 공격(Limited Knowledge Attack) 시나리오 모두에서 Carlini와 Wagner의

알고리즘이 판별기를 우회하는 데 성공하였다. Feinman 등[21]의 연구는 기존의 Carlini와 Wagner의 공격을 잘 막아내는 듯 했으나, [21]의 연구 논문을 분석하고 난 후 이에 맞추어 무작위성을 고려한 손실함수를 사용하자 역시 우회하는 데 성공했다고 보고하고 있다.

2. 검증 기술

심층신경망은 이제 우리 삶에서 복잡한 문제들의 해결을 위해 널리 사용되기 시작하고 있다. 이렇게 사용되는 시스템들 중에는 안전이 지극히 중요한 시스템들이 있다. 이러한 시스템들에서는 신경망의 예측 결과나 행동에 대한 공식적인 보증이 필요하다.

Katz 등[23]은 신경망의 속성을 형식적(formal)으로 검증하기 위한(또는 이에 대한 반례를 제공하기 위한) 확장 가능하고 효율적인 기술을 제시하였다. 이 방법은 심플렉스 메쏘드를 확장하여 볼록하지 않은(non-convex) Rectified Linear Unit(ReLU) 활성화 함수까지도 다룰 수 있도록 확장된 형태이다. 그래서 제안된 시스템의 이름이 Reluplex인 것이다. 독자들도 아시다시피 ReLU는 현대 신경망 기술에서 필수적인 부분 중 하나이다. Reluplex는 효율적인 SMT Solver(Satisfiability Modulo Theories Solver)의 형태로 제공되며, 여기서 신경망은 선형 실수 산수(Linear Real Arithmetic)에 근거한 특정한 이론의 구성물로 표현되어, 이에 대해 Satisfiability(SAT)에 대한 결정 문제(decision problem)를 푸는 것이다.

Sinha 등[24]은 적대적 예제의 문제를 분산된 강건한 최적화 문제를 푸는 방법으로 해결하고자 한다. Sinha 등[24]은 Wasserstein ball에 대한 데이터 분포에 대해 섭동을 일으킬 때, 이에 소요되는 페널티를 라그랑쥬 페널티로 표현하여 학습 데이터의 최악의 섭동(이를 구하는 방법은 NP-hard임을 증명하였음)까지도 대비하여 가중치 패러미터의 업데이트를 추가하는 학습 방안을 제시하였다.

3. 방어 기술

Meng과 Chen[25]은 적대적 예제로부터 신경망 분류기를 보호하기 위한 프레임워크인 MagNet을 제안하였다. 이 방법 또한 보호하는 분류기를 별도로 수정하지 않으며 적대적

예제를 생성하는 공격 기법에 대해 미리 알아둘 필요가 없다. 즉, 기존의 분류기를 그대로 사용할 수 있으며, 공격 기법에 독립적인 방어 기술이다. MagNet 내에는 별도로 분리된 판별기 네트워크(detector networks)와 리포머 네트워크(reformer network)가 있다. 판별기는 주어진 입력이 적대적인지 아닌지를 판별한다. 리포머는 테스트 입력에서 적대적 섭동을 변경하여 분류기가 제대로 분류할 수 있도록 재구성하는 역할을 한다.

Carlini와 Wagner[26]는 자신들의 논문에서 기존의 방어기법들과 MagNet[25]도 실은 적대적 예제에 대해 강건하지 않음을 보였다. 그들은 이를 위해 자신들의 공격 기법의 최적화 세팅에서 분류기 손실함수와 판별기 손실함수를 변경하였다. 변경한 결과 MagNet이 L_2 공격에 대해 취약함을 보였다.

MIT의 Madry 등[27]은 신경망의 적대적 강건함(adversarial robustness)을 얻는 문제를 최적화의 관점에서 접근하였다. 오분류로 인해 일어나는 손실(loss)의 평균을 최소화하는 기존의 Empirical risk minimization(ERM) 패러다임이 적대적 예제를 막아내는데 완벽하지 못하다는 보고에 근거하여 ERM 모델을 확장한 연구이다. 실제 처리에서는 각각의 데이터 포인트에 대해 적대적 섭동을 고려하고, 다음 수식에서 보는 것처럼 그 중에서 가장 크기가 큰 섭동(L_∞바운드)를 최소화하는 최소-최대 안장점(min-max saddle point) 최적화를 수행한다.

$$\min_\theta \rho(\theta) \text{ where } \rho(\theta) = \mathrm{E}_{(\mathrm{x},\mathrm{y})\sim \mathrm{D}}\left[\max_{\delta \in S} L(\theta, x+\delta, y)\right]$$

이후 Madry는 자신의 홈페이지에 자신의 모델을 공격해서 우회해 볼테면 해보라는 메시지까지 올렸었다.

그러나, The Cooper Union과 IBM의 Sharma와 Chen[28]은 elastic-net attack to deep neural networks(EAD)라는 알고리즘을 통해 Madry 모델[27]을 우회하는 데 성공한다. 이것이 가능했던 이유는 [28]의 공격으로 만들어지는 적대적 예제는 사실 높은 값을 가지고 있었지만, 여전히 그 섭동이 사람의 눈으로 간파하기 어려운, 즉 정상적인 이미지로 보였던 것이다.

사실 [28]의 논문을 보면 알 수 있는 점은, EAD 공격의 결과로 만들어지는 적대적 이미지에는 눈에 보이는 잡음 픽셀들이 몇 개 존재하긴 한다는 것이다. 하지만 전체적인 적대적 이미지는 정상 이미지처럼 사람이 구별할 수 있다. 이 연구의 결과는 오히려 적대적 머신러닝에서 “사람의 눈으로 감지(perceptible)하기 어려운”이라는 계량적 기준이 과연

무엇인지에 대한 질문을 던진다고 볼 수도 있다.

4. 공격 기술에 대해 심화된 방어 기술

Wu 등[29]은 주어진 적대적으로 훈련된 모델에 대해 적대적 강건함을 강화하기 위해 Madry 등[27]의 프레임워크를 개선하여 Highly Confident Near Neighbor(HCNN) 알고리즘을 제시하였다. 해당 알고리즘은 신뢰정보와 최근접 이웃(nearest neighbor)을 검색하는 방법을 이용하였는데, 기본적인 아이디어는 "자연스러운 데이터 포인트들이 서로 다른 클래스에 속한다면 저차원 매니폴드 상에서는 서로가 분리가능하다."는 매니폴드 가정에 따른 것이다. 이에 따라 신뢰값이 특정 임계치보다 작으면 적대적 예제로, 아니면 그 신뢰값에 대한 클래스로 반환함으로써, 적대적 예제에 대한 잘못된 예측을 정정하는 것이다.

Schott 등[30]은 현재까지 가장 성공적인 방어 메카니즘이라는 Madry 모델[27]조차도, (1) L_∞에 대한 오버피팅으로 인해 상대적으로 L_2나 L_0 섭동에 취약하고, (2) 사람이 제대로 인식할 수 없는 이미지를 높은 확신도로 특정 클래스로 인식하고, (3) 단순히 이진화(binarization)된 입력에 대한 것보다도 별반 높은 성능을 보이지 않으며, (4) 사람에게는 전혀 무의미한 적대적 섭동을 중요한 특징으로 인식하는 문제들을 노출했음을 지적했다. 결국 적대적 강건성의 입장에서는 장난감 데이터에 가까운 MNIST조차도 제대로 해결하지 못했음을 지적한 것이다. 이러한 전제 위에 Schott 등[30]은 학습된 클래스 조건부 데이터 분포(learned class-conditional data distributions)에 대한 분석을 통해 더 강건한 분류 모델인 Analysis by Synthesis model(ABS) 모델을 제시하였다. 그들은 자신들이 제안한 ABS 모델을 경험적으로 평가하기 위해 우선 서로 다른 L_p norm에 대해 의사결정 기반 공격, 스코어 기반 공격, 그래디언트 기반 공격 및 전이가능성 기반 공격 등의 다양한 공격들을 적용하였다. 또한, 자신들의 방어 모델의 구조를 악용하는 새로운 공격을 설계하였고, 섭동에 의해 교란된 픽셀의 수(L_0)를 최소화하려는 새로운 의사 결정 기반 공격을 고안하여 공격을 수행하였다. 그들의 자체적 실험결과, [30]의 모델은 적대적 섭동에 대해 강건성을 보였다고 보고하고 있다.

VI. 결론 및 시사점

이상으로 예비지식, 전이가능성, 공격기술, 방어기술, 그리고 다른 도메인에서의 적대적 예까지 정리해 보았다. 향후의 연구는 다양한 방향이 있을 것으로 예상된다. 일단 산업계에 가장 큰 영향을 줄 수 있는 연구는 음성인식에 대한 적대적 사례를 연구하는 것으로 보인다. 이론적인 분야에서는 생성 모델에 대한 공격에 대한 연구가 있을 수 있다. 이미 이 분야로 국내의 몇 개 대학에서 심도 깊은 연구[31]가 이루어지고 있고, 좋은 학회에 논문이 발표된 경우도 있는 것으로 알고 있다. 이론적으로는 과연 적대적 섭동을 어떻게 정의할 것인가의 문제가 남아있는 것으로 보인다. 적대적 잡음이 의미가 있는 것은 그것이 이미지에 추가되었을 때, 인식 모델은 오분류를 일으킴에도 인간이 보기에는 원래의 이미지를 인식하는 데 문제가 없는 경우일 것이다. [28]의 경우처럼, 그런 기준을 단순히 L_p 거리로 정하기에는 논의의 여지가 있을 수 있다는 것이다. Madry의 모델을 우회한 EAD의 경우에서 보듯이 L_p 거리가 높아도 사람은 여전히 인식 가능할 수도 있다.

[참고문헌]

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. “Explaining and Harnessing Adversarial Examples.” CoRR abs/1412.6572, 2014.

[2] J. Su, D. V. Vargas, and K. Sakurai. “One Pixel Attack for Fooling Deep Neural Networks.” CoRR abs/1710.08864, 2017.

[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. “Synthesizing Robust Adversarial Examples.” ICML 2018, 284-293.

[4] N. Carlini, “Adversarial Machine Learning Reading List,” URL:https://nicholas.carlini.com/writing/2018/adversarial-machine-learning-reading-list.html

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, F. Roli. “Evasion Attacks against Machine Learning at Test Time,” Machine Learning and Knowledge Discovery in Databases, H. Blockeel, et al., 2013, 387-402.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, “Intriguing Properties of Neural Networks,” ICLR 2014.

[7] I. J. Goodfellow, J. Shlens, C. Szegedy, “Explaining and Harnessing Adversarial Examples,” ICLR 2015.

[8] N. Papernot, P. McDaniel, I. Goodfellow, “Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples,” Technical Report.

[9] Y. Liu, X. Chen, C. Liu, D. Song, "Delving into Transferable Adversarial Examples and Black-box Attacks," ICLR 2017.
[10] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard. "Universal adversarial perturbations," CVPR 2017.
[11] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, "The Limitations of Deep Learning in Adversarial Settings," IEEE European Symposium on Security & Privacy, 2016.
[12] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," CVPR 2016.
[13] N. Carlini, D. Wagner, "Towards Evaluating the Robustness of Neural Networks," 2017 IEEE Symposium on Security and Privacy(SP), 2017, 22-26.
[14] A. Athalye, N. Carlini, D. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," ICML 2018.
[15] J. Uesato, B. O'Donoghue, A. van den Oord, P. Kohli, "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," ICML 2018.
[16] P-Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.-J. Hsieh, "ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models," 10th ACM Workshop on Artificial Intelligence and Security(AISEC) 2017.
[17] W. Brendel, J. Rauber, M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," ICLR 2018.
[18] A. Kurakin, I. Goodfellow, S. Bengio, "Adversarial examples in the physical world," ICLR Workshop, 2017.
[19] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, "Robust Physical-World Attacks on Deep Learning Models," CVPR 2018
[20] J. H. Metzen, T. Genewein, V. Fischer and B. Bischoff, "On Detecting Adversarial Perturbations," ICLR 2017.
[21] R. Feinman, R. R. Curtin, S. Shintre, A. B. Gardner, "Detecting Adversarial Samples from Artifacts," CoRR abs/1703.00410, 2017.
[22] N. Carlini, D. Wagner, "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods," ACM Workshop on Artificial Intelligence and Security, 2017, 3-14.
[23] G. Katz, C. Barrett, D. Dill, K. Julian, M. Kochenderfer, "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," International Conference on Computer Aided Verification, 2017.
[24] A. Sinha, H. Namkoong, R. Volpi, J. Duchi, "Certifying Some Distributional Robustness with Principled Adversarial Training," ICLR 2018.
[25] D. Meng, H. Chen, "MagNet: a Two-Pronged Defense against Adversarial Examples," ACM Conference on Computer and Communications Security(CCS), 2017.

[26] N. Carlini, D. Wagner, “MagNet and ‘Efficient Defenses Against Adversarial Attacks’ are Not Robust to Adversarial Examples,” CoRR abs/1711.08478, 2017.
[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks,” ICLR 2018.
[28] Yash Sharma, Pin-Yu Chen, “Attacking the Madry Defense Model with L1-based Adversarial Examples,” ICLR 2018 Workshop.
[29] X. Wu, U. Jang, J. Chen, L. Chen, S. Jha, “Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training,” ICML 2018, 5334-5342.
[30] L. Schott, J. Rauber, M. Bethge, W. Brendel, “Towards the first adversarially robust neural network model on MNIST,” ICLR 2019.
[31] J. Ho, B.-G. Lee, D.-K. Kang, “Uni-Image: Universal Image Construction for Robust Neural Model,“ Neural Networks, 128, August 2020, 279-287, Elsevier B.V., ISSN: 0893-6080.