적대적 공격에 대한 강건성: 연구 동향, 문제점, 그리고 미래

고등과학원 ▌이성윤

1. 서 론

딥러닝(deep learning)은 2012년 ImageNet 이미지 분류(classification) 대회에서 기존의 방법론과 비교하여 큰 성능 향상을 가져온 AlexNet [1]을 시작으로 다양한 분야에서 큰 관심을 받으며 급격한 발전을 이루었다. 이때부터 딥러닝을 이해하고자 하는 많은 시도들이 있었고, 바로 이듬해인 2013년, 입력값(input)에 아주 작은 섭동(perturbation)을 더해주어 모델의 성능을 현저히 낮출 수 있다는 연구 결과가 알려지면서 새로운 측면에서 딥러닝을 바라보는 시각들이 생기게 되었다 [2, 3]. 이때 작은 섭동이 더해진 새로운 입력값을 적대적 입력(adversarial example)이라 부른다. 이러한 적대적 입력의 존재성은 자율주행 자동차 [4], 얼굴 인식 [5], 의료 이미징 [6] 등 안전이 중요한(safety-critical) 응용 분야에 딥러닝을 활용하는 데에 있어서 기존의 딥러닝을 통한 학습 방법론이 가진 문제점을 제기하였다.

본 글에서는 이후 약 10년간 이루어진 연구들을 따라가 보면서, 적대적 입력에 대한 강건성 연구들과, 이들의 문제점, 그리고 미래 연구 방향에 대해 알아볼 것이다.

2. 적대적 입력(Adversarial Example)

먼저 적대적 입력의 정의를 명확히 살펴보자.

> [정의] 입력값 x와 모델 f가 주어졌을 때, 적대적 입력 x^*은 (i) 기존 입력값 x와 "구분이 잘 되지 않으면서" (ii) f의 출력값(output)을 "사용자가 원치 않는 방향"으로 이끄는 입력값이다.

이 정의에서 입력값의 종류에 따라 "구분이 잘 되지 않는다"의 의미가 달라지고, 모델 f가 수행하는 태스크(task)에 따라 "사용자가 원치 않는 방향"의 의미가 달라진다. 이미지 분류 문제의 경우, 이미지 입력값을 l_p-norm($p=0,1,2,\infty$)으로 측정한 거리 $\|x-x^*\|_p$가 어느 특정 $\epsilon>0$보다 작다는 것을 "구분이 잘 되지 않는다"라는 의미로 사용할 수 있고, 분류 모델의 출력값이 다른 것 ($f(x)\neq f(x^*)$)을 "원치 않는 방향"의 의미로 사용할 수 있다. 이미지 검출(detection) [7], 음성 인식 [8], 강화학습 [9] 등의 예시를 생각하면, 위와는 또 다른 다양한 방법으로 (i)과 (ii)의 정의가 가능하다. 본 글에서는 이미지 분류에만 집중해 다루어 볼 것이다.

3. 적대적 공격(Adversarial Attack)

먼저 적대적 입력을 얻는 구체적인 방법을 몇 가지 적대적 공격을 통해 알아보자.

Szegedy et al. [2]은 주어진 이미지 x에 대응하는 라벨(label) y와 다른 출력값을 얻기 위해 x 근방에서 손실(loss) 함수 $l(x',y;\theta)$를 최대화하는 x'을 찾고자 하였고, 이를 위해 아래의 최적화 문제를 풀었다. 이때 다소 복잡한 BFGS 알고리즘이 사용되었다.

$$\max_{\|x'-x\|_p\leq\epsilon} \ell(x',y;\theta)$$

Goodfellow et al. [10]은 x 근방에서 손실 함수를 선형(linear)으로 근사하여, [2]에서 사용한 BFGS 알고리즘 대신, 한번의 기울기(gradient) 계산만으로 적대적 입력을 얻었다. 예를 들어 l_∞-norm($p=\infty$)으로 제약된 섭동을 고려하는 경우, 아래와 같이 적대적 입력 x^*을 얻을 수 있고, 이러한 공격 방법론을 FGSM (Fast Gradient Sign Method)라 부른다.

$$x^* = x + \epsilon\text{sign}(\nabla_x \ell(x,y;\theta))$$

Papernot et al. [11]은 아래의 또 다른 최적화를 통해 공격자가 원하는 방향(targeted)으로 $y^*\neq y$을 출력값으로 얻는 방법론을 제안하였고, 이에 해당하는 최적화 문제를 풀기 위해 Jacobian-based saliency map을 사용한다. 이 방법론을 JSMA (Jacobian-based Saliency Map Attack)이라 부른다.

$$\min_{\delta_x} \|\delta_x\| \cdot \text{s.t.} f(x+\delta_x) = y^*$$

이외에도 대표적으로 CW [12], PGD [13], AutoAttack [14] 등 다양한 공격 방법론들이 있다. 추후에 살펴보겠지만, 유한개의 가능한 공격을 아는 것은 진정한 강건성을 얻기 위해서 큰 의미는 없기에 이들 각각 공격 방법론에 대해서 자세히 다루지는 않을 것이다.

[Disclaimer]

본 글에서는 10년간의 적대적 입력 관련 모든 연구 방향을 다루지 않는다. 예를 들어, 공격자가 공격하고자 하는 모델을 모르는 가정에서의 black-box attack 방향으로도 많은 연구가 진행되어 왔고 [15, 16, 17, 18, 19] 실제 응용 시나리오에 적합한 연구 방향일지도 모르나, 궁극적인 강건성을 위해, 모델이 알려지더라도 강건할 수 있는 white-box attack에 대한 연구에만 집중하였다. 또한 이미지 검출, 음성 인식, 강화학습 등 다양한 태스크에 대해서도 연구가 진행되었으나, 적대적 공격에 대한 이해는 크게 다르지 않으므로 가장 연구가 많이 된 이미지 분류 태스크에 대해서만 살펴볼 것이다.

4. A Typical Example - Defensive Distillation

앞으로의 논의를 위한 단적인 예시로 초기의 연구인 defensive distillation [20]에 대해서 알아보자. defensive distillation은 다음의 과정을 통해 output-input sensitivity를 줄이는 방향으로 강건한 모델을 학습한다. 먼저, 학습 데이터 $\{(x_i, y_i)\}$로부터 모델 f를 학습하고, 이로 얻어진 soft-target $f(x_i)$를 사용하여 새로운 학습 데이터 $\{(x_i, f(x_i))\}$를 만든다. 그리고 이 새로운 학습 데이터를 이용해 (f와 같은 구조의) 다른 모델 f^d를 학습한다. 이때 중요한 점은 두 학습에서 softmax layer에 높은 temperature T를 사용한다는 점이다. 높은 T값을 사용하면, softmax를 통해 얻은 확률 출력값(output probability)이 더 균등(uniform)해지고, 결국 output-input 기울기가 작아지고, 입력값 변화에 대한 민감도를 낮추는 효과를 가져온다. 그리고 실험을 통해 모델 f^d가 JSMA 공격에 대해 강건함을 보였다.

그러나 이후의 연구에서 위 방법론은 단순히 수치적으로 기울기를 0에 가깝게 만들어서 공격을 회피하는 방법(gradient masking)이라는 것이 밝혀지고, 공격자가 temperature T를 고려하여 제대로 된 기울기를 얻어서 공격할 수 있었다 [21].

5. Milestone – Obfuscated Gradients

이후 2017년 NIPS에서 Competition Track (defense against adversarial attack) [22]을 통해 다양한 방어 방법론들이 제안되었고 (1등 [23], 2등 [24]), 뒤이어 2017년 말 ICLR 2018 Openreview에서도 많은 방어 방법론들이 게시되었다. 이때 제안된 대부분의 방어 방법론들을 크게 (i) 입력값 변환(input transformations) 기반 방어 방법론, (ii) 무작위화(randomization) 기반 방어 방법론의 두 가지 방법론으로 나눌 수 있다.

첫번째, 입력값 변환 기반 방법론은 디노이징(denoising) 등의 입력값 변환을 통해 모델에 들어가기 전에 입력값에 전처리(preprocessing)를 해주어서 적대적 입력의 섭동을 제거하고자 하는 시도이다 [25, 26, 27]. 입력값 변환 g를 이용한 경우 수식으로는 아래와 같이 표현된다.

$$f(g(x))$$

두번째, 무작위화 기반 방법론은 무작위한 이미지 리사이징(resizing)을 적용하거나 뉴런(neuron)을 무작위로 pruning시키는 등 모델 자체를 무작위화 시켜 적대적 공격에 필요한 섭동 계산을 무력화시키는 방법이다 [28, 29, 24]. 무작위화된 모델의 분포를 F라 표현하면 수식으로는 아래와 같이 표현된다.

$$f(x), where\ f \sim F$$

이 방법론들을 제안한 저자들은 공격자가 무작위화된 모델의 분포 F는 완벽하게 알고 있다 하더라도, 어떤 f가 샘플링되어 추론(inference) 때 사용되는지는 무작위화 때문에 알 수 없기 때문에 기존 공격 방법에서 주로 사용하는 기울기를 얻을 수 없어 "계산이 불가능"하고 최적화 문제가 "수렴하지 않을 수 있다"고 주장하였다.

그러나 Athalye et al. [30]은 2017년 말 ICLR 2018 Openreview에 게시된 많은 방어 방법론들을 공격하는 방법론을 제안하면서, 위의 두 방법론 모두 진정한 의미의 강건성을 얻지 못했다는 것을 보였다. (비슷한 때에 다른 논문도 비슷한 주장을 하였다 [31].)

첫째로, 입력값 변환 기반 방어 방법론은 BPDA (Backward Pass Differential Approximation)이라 불리는 공격에 의해 공격당했다. BPDA는 공격자가 기울기를 계산할 수 없는 입력값 변환 g에 대해서 단순히 이 변환을 무시하고 기울기를 계산해서 공격하는 방법이다. 예를 들어, 디노이징의 경우, 디노이징 변환

을 항등함수(identity)로 근사하는 것과 같다. 즉, 아래와 같이 좌변의 기울기를 우변과 같이 근사하여 g의 기울기 계산 없이 공격이 가능하다.

$$\nabla_x f(g(x))|_{x=\hat{x}} \approx \nabla_x f(x)|_{x=g(\hat{x})}$$

(모델 f에 대한 기울기를 알면 손실 함수 l에 대한 기울기를 계산할 수 있다.)

두번째로, 무작위 기반 방어 방법론은 EOT (Expectation over Transformation)이라는 공격 방법론에 의해 공격되었다 [32]. EOT는 공격자가 기울기를 계산할 때, 모델 분포 F에서 샘플링 된 하나의 f_0에 대해서만 $\nabla_x f_0(x)$를 계산하는 것이 아니라, 여러 샘플 $\{f_i\}_{i=1}^N$를 사용해서 평균 기울기를 이용해 적대적 섭동을 계산하는 방법이다.

$$\nabla_x f(x) \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_x f_i(x)$$

6. 문제점 – Arms-race

최근까지도 새로운 방어 방법론을 제안한 논문들이 많이 나오고 있으나, 고정된 몇몇의 유한한 적대적 공격 방법론들에 대해서만 강건성을 실험적으로 보일 뿐, 제대로 된 의미의 강건성을 보이지는 못한다 [14, 33]. 앞서의 defensive distillation과 obfuscated gradients의 예시로 살펴본 방법론들은 강건성을 주장한 실험에 사용되었던 고정된 몇몇 적대적 공격이 아닌 새로운 공격에 의해 공격 되었다. 이처럼 방어에 맞춤형 공격을 맞춤 공격(adaptive attack)이라 한다. 진정한 강건성을 위해 우리는 가능한 모든 공격에 대해 강건함을 보여야 한다.

최근의 단적인 예로 “Geometry-aware Instance-reweighted Adversarial Training” [34]는 ICLR 2021 oral presentation까지 한 높은 평가를 받은 논문이지만, 몇몇 고정된 적대적 공격들에 대해서만 강건성을 보였다. 리뷰를 통해 제안된 AutoAttack [14]에 대해서는 큰 성능 향상을 보여주지 못하며, 심지어 logit scaling attack(LSA) [35]에 의해 성능이 현저히 떨어지는 것이 알려졌다. 이처럼 아직까지도 적대적 강건성 커뮤니티는 일관된 이해가 부족해 진정한 의미의 강건성까지 가는 길은 멀기만 하다.

물론 다양한 방어 방법론들을 통해 적대적 공격에 대한 이해가 깊어지고, 더 강한 공격 방법론과 더 강한 방어 방법론이 서로 경쟁하며(arms-race) 더 나은 방어 방법론이 나오게 될지도 모른다. 하지만 이렇게 얻어진 방어 방법론은 영원한 강건성을 보장할 수 없으며 새로운 맞춤 공격이 나오기 전 까지만 유효할 뿐이다.

7. 검증 가능한 방어 방법론(Certifiable/ Certified Defense)

이러한 문제점을 해결하고 진정한 강건성을 얻기 위해 검증 가능한 방어 방법론이 등장하였다. 검증 가능한 방어 방법론은 강건성을 평가할 때, 기존 방법론들이 몇몇 특정 공격 방법론(e.g., FGSM [10], CW [12], PGD [13], AutoAttack [14])에 대해서만 평가하는 것과 달리, 정해진 섭동 영역 $\{x' : \|x' - x\|_p \leq \epsilon\}$을 고정시켜 놓고 해당 영역에서 가능한 모든 적대적 섭동이 모델의 결정(decision)을 바꿀 수 없다는 것을 증명하는 과정을 통해 강건성을 검증한다. 기존 방법론과 비교해서, 검증 가능한 방어 방법론의 가장 중요한 차이점은 실험적인 평가가 아닌, 수학적인 증명을 통해 강건성을 보인다는 것이다. 이 때문에 우리는 검증 가능한 방어 방법론을 통해서 진정한 의미의 강건성을 보일 수 있다.

가장 처음으로 등장한 방법론 [36]은 간단한 커널 방법론(kernel method)이나 2층 신경망(2-layer neural networks)정도에만 적용 가능했지만, 이후 등장한 선형 완화(linear relaxation) 기반의 방법론 [37, 38, 39, 40]과 영역 전파(bound propagation) 기반의 방법론 [41, 42]을 통해 점점 더 큰 심층 신경망(deep neural networks)에도 적용이 가능하게 되었다.

두 방법 모두 입력 공간에서 norm 제한 조건 $\|x' - x\|_p \leq \epsilon$을 만족하면서 출력값이 가질 수 있는 범위를 계산해서 최악의 경우(worst-case) 손실 함수 $\max_{\|x'-x\|_p \leq \epsilon} \ell(x', y; \theta)$보다 큰 상한 손실 함수(upper bound loss) l^{UB}를 사용하여 학습을 한다.

$$\max_{\|x'-x\|_p \leq \epsilon} \ell(x', y; \theta) \leq \ell^{UB}(x, y; \theta)$$

첫번째, 선형 완화 기반 방법론은 비선형 활성함수(nonlinear activation)를 선형 상/하한 함수(linear upper/lower bound)로 제한(bound)시켜, 상한 손실 함수의 계산을 선형 계획법(linear programming)으로 완화(relax)시켜서 계산한다.

두번째, 영역 전파 기반 방법론은 입력 공간에서의 섭동 영역에서부터 층별로 영역(bound)을 전파(propagate)시켜가면서 가장 마지막 층까지 도달하여 상한 손실 함수를 계산하는 방법이다.

둘을 비교하였을 때, 일반적으로 선형 완화 방법론이 더 느리지만, 더 밀착된(tight) 상한 손실 함수를 얻을 수 있다. 실제로 상한 손실 함수의 밀착성(tightness)을 얻기 위한 다양한 방법론들이 제안되어 왔다 [43, 40, 44]. 따라서 계산량을 무시한다면 선형 완화 방법론이 더 유망한 방법론이라 기대할 수 있다. 하지만 실제로는 영역 전파 기반 방법론인 IBP(Interval Bound Propagation) [41]가 선형 완화 기반 방법론보다 상한 손실 함수의 밀착성 측면에서 부족하지만 큰 섭동에 대해 성능이 잘 나오는 경향이 있다. 이는 상한 손실 함수의 밀착성 뿐만 아니라, 목적 함수의 최적화 측면에서 상한 손실 함수의 매끈함(smoothness) 또한 중요하기 때문이다. 따라서 IBP가 밀착성 측면에서는 불리하지만, 상한 손실 함수의 매끈함이 우수하기 때문에 최적화 측면에서 유리해 좋은 성능을 얻을 수 있다 [45, 46, 47].

8. 현재까지는 유효한 휴리스틱 방법론

검증 가능하지 않은 방어 방법론들을 통틀어 휴리스틱 방어 방법론(heuristic defense)이라 부르기도 한다. 많은 휴리스틱 방어 방법론들이 맞춤 공격에 의해 공격 되었지만, 대표적으로 adversarial training (AT) [13], TRADES [48] 등과 같은 휴리스틱 방어 방법론들은 아직까지는 좋은 성능을 보여주고 있다. 이러한 방법론들을 기반으로 여러가지 하이퍼파라미터 튜닝(hyperparameter tuning)을 통한 다양한 시도들을 통해 성능을 다소 향상시켰지만, 원인을 파악하기는 쉽지 않다 [49, 50].

그리고 적대적 강건성의 일반화(generalization)를 위해서는 일반적인 학습보다 더 많은 데이터가 필요하다는 것이 알려진 이후 [51], 최근에는 AT나 TRADES 등 기본적인 방법론을 기반으로 일반화 성능을 향상시키기 위해, 추가로 라벨이 없는(unlabeled) 데이터를 사용하거나 [52], 생성 모델(DDPM [53] 등)을 사용하여 데이터를 생성해서 사용하는 방법론 [54] 등이 많이 등장하였다. 뿐만 아니라 일반화 성능을 향상시키기 위해 adversarial weight perturbation (AWP) [55]라 불리는 손실 함수의 기하적 특성 기반의 연구까지 진행되었다.

현재까지는 휴리스틱 방법론들의 강건성을 측정하기 위해 충분히 강한 것으로 알려진 AutoAttack [14]을 일반적으로 많이 사용하고, RobustBench [56]에 이를 사용하여 측정한 가장 성능이 좋은(state-of-the-art) 방법론들의 결과가 잘 정리되어 있다. 그러나 이는 단순히 현재의 관례일 뿐 진정한 강건성을 측정하는 것과는 다르다.

9. 결론 및 앞으로의 연구 방향

진정한 의미의 강건성을 위해 우리가 궁극적으로 가야할 방향은 검증 가능한 방법론이다. 하지만 해당 방향으로 연구는 상대적으로 부족하고 검증 성능(certified accuracy)도 매우 낮은 상황이다. 최근 들어 무작위 평활화(randomized smoothing) [57]라는 확률적으로 검증 가능한 방어 방법론이 주목을 많이 받고 있다. 무작위 평활화는 기존의 결정적(deterministic)인 출력값을 주는 모델을 사용하지 않고, 여러 출력값을 내뱉는 무작위 모델을 사용해서 Neyman-Pearson lemma를 통해 (유의 수준 개념에서) 높은 확률로 강건성을 검증할 수 있다. 이처럼 단순히 학습 알고리즘 측면에서의 연구보다는, 새로운 신경망 구조 또는 신경망을 벗어난 새로운 모델을 고려를 하는 것이 미래에 강건성을 위해 더 가능성 있는 방향이지 않을까 생각한다.

또한 현재 주로 고려하고 있는 l_p-norm 섭동 가정은 실제와 맞지 않고, 특정 하나의 l_p-norm에 대해서만 안전한 것 또한 문제가 될 수 있다. 최근 들어 "잘 구분되지 않는다"는 의미에 더 적합한 인지적 섭동(perceptual perturbation) 방향의 연구도 진행되고 있다 [58].

적대적 강건성은 10년의 길을 걸어왔지만 여전히 갈 길이 멀다. 그럼에도 불구하고, 적대적인 세팅 외에도 다른 세팅에서의 일반화 성능에 대한 연구가 많이 진행되고 있는 만큼, 여러 분야에서의 연구들이 얽혀 이해가 더 깊어지면서 서로 발전을 이룰 수 있을 것으로 기대해 본다.

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks," *International Conference on Learning Representations,* 2014.

[3] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European*

conference on machine learning and knowledge discovery in databases, 2013.

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[5] 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*, 2016.

[6] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science,* vol. 363, p. 1287–1289, 2019.

[7] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[8] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018.

[9] S. Huang, N. Papernot, I. Goodfellow, Y. Duan and P. Abbeel, "Adversarial attacks on neural network policies," *International Conference on Learning Representations,* 2017.

[10] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations,* 2015.

[11] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*, 2016.

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in *International Conference on Learning Representations*, 2018.

[14] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*, 2020.

[15] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

[16] C. Guo, J. Gardner, Y. You, A. G. Wilson and K. Weinberger, "Simple black-box adversarial attacks," in *International Conference on Machine Learning*, 2019.

[17] A. Ilyas, L. Engstrom, A. Athalye and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International Conference on Machine Learning*, 2018.

[18] A. Ilyas, L. Engstrom and A. Madry, "Prior Convictions: Black-box Adversarial Attacks with Bandits and Priors," in *International Conference on Learning Representations*, 2019.

[19] M. Andriushchenko, F. Croce, N. Flammarion and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *European Conference on Computer Vision*, 2020.

[20] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016.

[21] N. Carlini and D. Wagner, "Defensive distillation is not robust to adversarial examples," *arXiv preprint arXiv:1607.04311,* 2016.

[22] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie and others, "Adversarial attacks and defences competition," in *The NIPS'17 Competition: Building Intelligent Systems*, Springer, 2018, p. 195–231.

[23] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[24] C. Xie, J. Wang, Z. Zhang, Z. Ren and A. Yuille, "Mitigating adversarial effects through randomization," *International Conference on Learning Representations,* 2018.

[25] J. Buckman, A. Roy, C. Raffel and I. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in *International Conference on Learning Representations*, 2018.

[26] P. Samangouei, M. Kabkab and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *International*

Conference on Learning Representations, 2018.

[27] Y. Song, T. Kim, S. Nowozin, S. Ermon and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," *International Conference on Learning Representations,* 2018.

[28] G. S. Dhillon, K. Azizzadenesheli, Z. C. Lipton, J. Bernstein, J. Kossaifi, A. Khanna and A. Anandkumar, "Stochastic activation pruning for robust adversarial defense," *International Conference on Learning Representations,* 2018.

[29] C. Guo, M. Rana, M. Cisse and L. van der Maaten, "Countering adversarial images using input transformations," *International Conference on Learning Representations,* 2018.

[30] A. Athalye, N. Carlini and D. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," in *International Conference on Machine Learning*, 2018.

[31] J. Uesato, B. O¡¯donoghue, P. Kohli and A. Oord, "Adversarial risk and the dangers of evaluating against weak attacks," in *International Conference on Machine Learning*, 2018.

[32] A. Athalye, L. Engstrom, A. Ilyas and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*, 2018.

[33] F. Tramer, N. Carlini, W. Brendel and A. Madry, "On adaptive attacks to adversarial example defenses," *Advances in Neural Information Processing Systems,* vol. 33, p. 1633-1645, 2020.

[34] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama and M. Kankanhalli, "Geometry-aware Instance-reweighted Adversarial Training," in *International Conference on Learning Representations*, 2020.

[35] D. Hitaj, G. Pagnotta, I. Masi and L. V. Mancini, "Evaluating the robustness of geometry-aware instance-reweighted adversarial training," *arXiv preprint arXiv:2103.01914,* 2021.

[36] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Advances in Neural Information Processing Systems*, 2017.

[37] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*, 2018.

[38] E. Wong, F. Schmidt, J. H. Metzen and J. Z. Kolter, "Scaling provable adversarial defenses," in *Advances in Neural Information Processing Systems*, 2018.

[39] K. Dvijotham, S. Gowal, R. Stanforth, R. Arandjelovic, B. O'Donoghue, J. Uesato and P. Kohli, "Training verified learners with learned verifiers," *arXiv preprint arXiv:1805.10265,* 2018.

[40] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh and L. Daniel, "Efficient neural network robustness certification with general activation functions," in *Advances in neural information processing systems*, 2018.

[41] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, T. Mann and P. Kohli, "Scalable verified training for provably robust image classification," *In Proceedings of the IEEE/CVF International Conference on Computer Vision,* 2019.

[42] M. Mirman, T. Gehr and M. Vechev, "Differentiable abstract interpretation for provably robust neural networks," in *International Conference on Machine Learning*, 2018.

[43] S. Lee, J. Lee and S. Park, "Lipschitz-certifiable training with a tight outer bound," *Advances in Neural Information Processing Systems,* vol. 33, p. 16891-16902, 2020.

[44] H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning and C.-J. Hsieh, "Towards Stable and Efficient Training of Verifiably Robust Neural Networks," in *International Conference on Learning Representations*, 2019.

[45] S. Lee, W. Lee, J. Park and J. Lee, "Towards Better Understanding of Training Certifiably Robust Models against Adversarial Examples," *Advances in Neural Information Processing Systems,* vol. 34, 2021.

[46] M. Balunovic and M. Vechev, "Adversarial training and provable defenses: Bridging the gap," in *International Conference on Learning Representations*, 2019.

[47] N. Jovanović, M. Balunović, M. Baader and M. Vechev, "On the Paradox of Certified Training," *Transactions on Machine Learning Research,* 2022.

[48] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui and M. Jordan, "Theoretically Principled Trade-off between Robustness and Accuracy," in *International Conference on Machine Learning*, 2019.

[49] S. Gowal, C. Qin, J. Uesato, T. Mann and P. Kohli, "Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples," *arXiv preprint arXiv:2010.03593,* 2020.

[50] T. Pang, X. Yang, Y. Dong, H. Su and J. Zhu, “Bag of tricks for adversarial training,” *International Conference on Learning Representations,* 2021.

[51] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar and A. Madry, “Adversarially robust generalization requires more data,” *Advances in neural information processing systems,* vol. 31, 2018.

[52] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi and P. S. Liang, “Unlabeled data improves adversarial robustness,” *Advances in Neural Information Processing Systems,* vol. 32, 2019.

[53] J. Ho, A. Jain and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems,* vol. 33, p. 6840–6851, 2020.

[54] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian and T. A. Mann, “Improving robustness using generated data,” *Advances in Neural Information Processing Systems,* vol. 34, p. 4218–4233, 2021.

[55] D. Wu, S.-T. Xia and Y. Wang, “Adversarial Weight Perturbation Helps Robust Generalization,” *Advances in Neural Information Processing Systems,* vol. 33, 2020.

[56] F. Croce, M. Andriushchenko, V. Sehwag, N. Flammarion, M. Chiang, P. Mittal and M. Hein, “RobustBench: a standardized adversarial robustness benchmark,” *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2),* 2021.

[57] J. Cohen, E. Rosenfeld and Z. Kolter, “Certified Adversarial Robustness via Randomized Smoothing,” in *International Conference on Machine Learning*, 2019.

[58] C. Laidlaw, S. Singla and S. Feizi, “Perceptual adversarial robustness: Defense against unseen threat models,” *arXiv preprint arXiv:2006.12655,* 2020.

[59] Z. Lyu, C.-Y. Ko, Z. Kong, N. Wong, D. Lin and L. Daniel, “Fastened crown: Tightened neural network robustness certificates,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[60] G. Singh, T. Gehr, M. Mirman, M. Püschel and M. Vechev, “Fast and effective robustness certification,” in *Advances in Neural Information Processing Systems*, 2018.

약 력

이 성 윤

현재 고등과학원 AI기초과학센터 Research Fellow
2021 서울대학교 수리과학부 박사 졸업
2016 서울대학교 재료공학부/수리과학부 학사 졸업
Email: sungyoonlee@kias.re.kr