적대적인 공격에 대한 인증 가능한 방어 방법의 성능 향상

*고효준 **박병준 ***김창익
한국과학기술원
*gohyojun15@kaist.ac.kr **pbj3810@kaist.ac.kr ***changick@kaist.ac.kr

Improving the Performance of Certified Defense Against Adversarial Attacks

*Hyojun Go **Byeongjun Park ***Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)

요약

심층 신경망은 적대적인 공격으로 생성된 적대적 예제에 의해 쉽게 오작동할 수 있다. 이에 따라 다양한 방어 방법들이 제안되었으나, 더욱 강력한 적대적인 공격이 제안되어 방어 방법들을 무력화할 가능성은 존재한다. 이러한 가능성은 어떤 공격 범위 내의 적대적인 공격을 방어할 수 있다고 보장할 수 있는 인증된 방어(Certified defense) 방법의 필요성을 강조한다. 이에 본 논문은 인증된 방어 방법 중 가장 효과적인 방법의 하나로 알려진 구간 경계 전파(Interval Bound Propagation)의 성능을 향상하는 방법을 연구한다. 구체적으로, 우리는 기존의 구간 경계 전파 방법의 훈련 과정을 수정하는 방법을 제안하며, 이를 통해 기존 구간 경계 전파 방법의 훈련 시간을 유지하면서 성능을 향상할 수 있음을 보일 것이다. 우리가 제안한 방법으로 수행한 MNIST 데이터 셋에 대한 실험에서 우리는 기존 구간 경계 전파 방법 대비 인증 에러(Verified error)를 Large 모델에 대해서 1.77%, Small 모델에 대해서 0.96% 낮출 수 있었다.

1. 서론

심층 신경망의 적용은 다양한 분야에서 성공을 거두고 있으며 자율주행과[1] 의료 분야처럼[2] 안전이 중요한 분야에 활발히 적용되고 있다. 하지만 심층 신경망은 적대적인 공격으로 생성된 적대적 예제에 오작동할 수 있으며[8, 9] 보안 문제를 초래할 수 있다. 심층 신경망이 적대적인 공격에 취약하다는 것에 대한 반응으로, 적대적인 공격에 대한 다양한 방어 방법들이 제안되어왔다. 하지만 이에 따라 더욱더 강력한 공격들이 새롭게 제안되면서 기존에 있던 방어 방법들을 무력화시켰다. 예를 들면 Distillation[3]을 기반으로 한 방어 방법은 더욱더 강력한 C&W 공격[4]으로 인해 무력화되었으며, 난독화된 구배(Obfuscated gradient)에 의존하던 방어 방법들은 적응적 공격[5]으로 인해 무력화되었다.

더욱더 새롭고 강력한 적대적인 공격이 심층 신경망을 위협할 수 있다는 가능성은 공격자가 조작할 수 있는 어떤 이미지의 픽셀 범위 공간 내에서 적대적인 공격에 대한 방어를 보장할 수 있는 인증된 방어(Certified defense) 방법의 필요성을 강조한다. 이에 따라 많은 인증된 방어 방법들은 제안되었으나, 그중에서 심층 신경망의 층마다 간단한 전파 규칙을 사용하는 구간 경계 전파(Interval Bound Propagation) 기반 방법[6, 7]이 가장 효과적인 방법으로 알려져 있다. 본 논문에서는 구간 경계 전파 방법에서의 훈련 과정을 연구하여 심층 신경망이 학습을 더욱 잘할 수 있는 훈련 과정을 제안하며, 기존 구간 경계 전파 방법의 훈련 시간을 유지하면서 성능을 향상할 수 있음을 보일 것이다.

2. 구간 경계 전파

구간 경계 전파는 심층 신경망의 입력 x_k과 공격자의 공격 범위 $S(x_k,\epsilon)=\{x\,|\,\|x-x_k\|_\infty \le \epsilon\}$에 대해 심층 신경망 출력 $f(x)$의 요소별 하한과 상한을 계산한다. 입력 x_k에 대한 정답 레이블을 y_k라하고, 심층 신경망 출력 $f(x)$의 k번째 요소의 공격 범위 $S(x_k,\epsilon)$에 대한 상한과 하한을 $\overline{f(x)_k}$, $\underline{f(x)_k}$라고 하면, y번째 여분인 $m_y=\underline{f(x)_{y_k}}-\overline{f(x)_y}$ 을 계산할 수 있다. 심층 신경망 출력의 벡터 요소 개수 d에 대해 여분 m_y가 $y=1,\ldots,d$에 대해서 모두 0보다 크거나 같다면 심층 신경망 $f()$는 입력 x_k와 공격 범위 $S(x_k,\epsilon)$안에서 정답인 레이블 y_k을 출력한다. 이때 공격자는 입력 x_k에 대하여 공격 범위 $S(x_k,\epsilon)$내에서 심층 신경망 $f()$가 정답 레이블 y_k과 다른 레이블을 출력하는 입력을 찾지 못하며, 이는 공격자의 적대적인 공격이 공격 범위 $S(x_k,\epsilon)$안에서 실패하는 것을 보장한다. 따라서 구간 경계 전파의 적대적 손실함수는 식 (1)과 같이 공식화된다.

$$L_{robust}=L(-\boldsymbol{m},y_k) \tag{1}$$

L은 교차 엔트로피 손실함수를 나타내고, $\boldsymbol{m}$은 심층 신경망의 출력 벡터 요소 개수 d에 대해 $y=1,\ldots,d$에 대응하는 여분 m_y를 원소로 가지는 여분 벡터이다. 더불어 구간 경계 전파의 전체 목적함수는 일반적인 분류기의 교차 엔트로피 손실함수 L_{CE}와 식 (1)의 적대적 손실함수를 사용하여 식 (2)와 같이 공식화된다.

$$L_{total} = KL_{CE} + (1-K)L_{robust} \quad (2)$$

여기서 K는 L_{CE}와 L_{robust} 사이의 균형을 맞추기 위한 하이퍼 파라미터이다.

구간 경계 전파의 학습 과정은 공격자의 공격 범위 $S(x_k,\epsilon)$에 대하여 ϵ을 0부터 목표 범위인 ϵ_{train}까지 증가시키면서 진행된다. 구간 경계 전파의 학습 과정은 3가지 단계로 진행되는데, 순서대로 예열 단계, 준비 단계, 안정화 단계이다. 첫 번째로 예열 단계에서는 심층 신경망 모델을 $\epsilon = 0$인 상태로 훈련한다. 이때 구간 경계 전파의 전체 목적함수는 일반적인 분류기의 교차 엔트로피 손실함수 L_{CE}와 같아지며, 이에 따라 심층 신경망 모델은 일반적으로 훈련된다. 두 번째로 준비 단계에서는 ϵ을 0부터 목표 범위인 ϵ_{train}까지 증가시키며, 식 (2)의 K도 $K_{initial}$에서부터 K_{final}까지 변화시키면서 목적함수 내에서 적대적 손실함수의 비중을 증가시킨다. 이러한 준비 단계에서는 하이퍼 파라미터들을 초깃값에서부터 최종값까지 변화시키며 안정화 단계를 준비하는 역할을 한다. 마지막으로 안정화 단계에서는 준비 단계에서 변화한 하이퍼 파라미터들을 최종값으로 고정하며, 심층 신경망 모델이 ϵ_{train}에 대해 학습한다.

이러한 일련의 훈련 과정은 심층 신경망의 구간 경계 전파의 학습 과정을 안정화하는 역할을 한다. 만약 ϵ을 0에서부터 ϵ_{train}까지 증가시키는 이 훈련 과정을 거치지 않고 ϵ_{train}으로 심층 신경망 모델을 학습시킨다면 모델은 종종 발산하고 때때로는 최적점으로 수렴하는 것에 실패하며 구간 경계 전파의 학습 과정이 불안정해질 수 있다.

3. 제안 방법

구간 경계 전파의 학습 과정의 준비 단계에서는 ϵ을 0에서부터 ϵ_{train}까지 증가시키며 심층 신경망 모델을 학습시킨다. 이는 심층 신경망 모델이 옳게 분류해야 하는 영역이 작은 $\epsilon = 0$에서 심층 신경망 모델이 옳게 분류해야 하는 영역이 큰 $\epsilon = \epsilon_{train}$으로 변화하며, 심층 신경망 모델이 적응하는 과정으로 볼 수 있다. 하지만 단순히 ϵ을 증가시키며 심층 신경망 모델을 훈련하는 것은 오로지 그 순간의 ϵ만 집중하여 모델을 훈련하기만 할 뿐 더 높은 ϵ을 훈련할 것이라는 가이드라인을 제시할 수 없다는 단점이 있다. 따라서 우리는 기존 준비 과정에서의 ϵ에다 조금 더 큰 ϵ을 입력 배치에 혼합하여 더 높은 ϵ을 훈련할 것이라는 가이드라인을 심층 신경망 모델에 제시하여 성능을 향상하는 방법에 대해서 제안할 것이다.

학습하는 ϵ이 커질수록 심층 신경망 모델이 옳게 분류해야 하는 영역이 커지기 때문에 학습 난도는 증가한다고 할 수 있다. 만약 기존 준비 과정에서의 ϵ보다 너무 큰 ϵ을 입력 배치에 혼합한다면 더 높은 ϵ을 훈련할 것이라는 가이드라인을 잘 제시하긴 하지만, 학습 난도가 너무 증가하기 때문에 심층 신경망 모델이 학습하는데 너무 큰 부하가 걸리게 된다. 반면에 기존의 준비 과정에서의 ϵ보다 약간 큰 ϵ을 입력 배치에 혼합한다면 학습 난도는 조금 증가하지만, 더 높은 ϵ을 훈련할 것이라는 가이드라인을 심층 신경망 모델에 효과적으로 제시하지 못할 것이다.

우리는 준비 과정에서의 ϵ의 차이에 따라 구간 경계 전파의 성능 변화를 확인하기 위해 MNIST 데이터 셋에 대해 실험을 진행하였다. 우리는 기존 구간 경계 전파의 MNIST 실험 환경을 따랐으며, 입력 배치의 절반을 준비 단계 동안 증가하는 ϵ으로 할당하고 또 다른 입력 배치의 절반에는 $\min(\epsilon+\alpha,\epsilon_{train})$을 할당했다. MNIST 데이터의 이미지는 최솟값 0, 최댓값 1로 정규화했고, $\epsilon_{train} = 0.4$로 설정하였으며 기존 구간 경계 전파 논문[6]에서의 large 모델을 사용했다. 그리고 훈련된 심층 신경망 모델의 인증된 방어 성능을 테스트 데이터 셋에 대해 틀리게 분류한 샘플의 비율인 일반 에러(Clean error)과 식 (1)의 $\epsilon = 0.4$에 대해 $\boldsymbol{m}$의 요소가 전부 0보다 크지 않은 샘플의 비율인 인증 에러(Verified error) 두 개의 기준으로 평가했다.

α	Verified Error	Clean Error
0	15.55%	2.27%
0.05	15.47%	2.43%
0.10	14.52%	2.18%
0.15	**14.19%**	2.17%
0.20	14.96%	2.16%
0.25	14.37%	2.12%
0.30	17.42%	3.50%
0.35	17.33%	3.57%
0.40	17.06%	3.68%

표 1. MNIST 데이터셋에 대한 실험에서의 α에 따른 심층 신경망 모델의 인증 에러와 일반 에러.

표 1에는 α에 따른 구간 경계 전파의 인증 에러와 일반 에러를 확인할 수 있다. 여기에서 α= 0은 기존의 구간 경계 전파의 훈련 방식으로 훈련한 모델의 결과이고, α가 0.05부터 0.40까지 증가할수록 입력 배치에 더욱 어렵고 큰 ϵ를 가지는 샘플들을 혼합했을 경우의 결과이다. α가 0.05일 때처럼 차이가 크지 않은 샘플들을 입력 배치에 혼합한 경우에는 구간 경계 전파의 성능이 크게 변하지 않은 것을 볼 수 있으며 이는 가이드라인을 효과적으로 제시하지 못했기 때문이라고 볼 수 있다. 반면에 α가 0.10, 0.15, 0.20, 0.25 정도로 적절하게 ϵ보다 큰 샘플들을 입력 배치에 혼합한 경우에는 인증 에러가 기존보다 크게는 1.36% 작게는 0.59%까지 작아져 기존보다 성능이 향상됨을 볼 수 있다. 그리고 α가 0.30, 0.35, 0.40인 경우는 너무 큰 ϵ을 입력 배치에 혼합해 학습 난도가 너무 많이 증가했기 때문에, 일반 에러와 인증 에러가 상승한 것을 볼 수 있다.

종합하면, α를 0.10, 0.15, 0.20, 0.25일 때 구간 경계 전파의 성능은 향상되는 결과를 보였다. 우리는 이번 단원에서 α가 하나인 경우에서 실험했지만, 더 나아가 실험에서 효과적이었던 α들의 경우들을 혼합하여 ϵ에 따라 부드럽게 변화하게 만들어준다면 더 높은 ϵ을 훈련해야 한다는 가이드라인을 더욱 부드럽게 심층 신경망에 제시할 수 있다. 따라서 우리는 입력 배치의 절반을 ϵ으로 할당하고, 효과적이었던 α인 0.10, 0.15, 0.20, 0.25에 각각 입력 배치의 1/8을 할당할 것이다. 우리의 방법은 기존 구간 경계 전파의 훈련 시간에서 입력 배치를 구성하는 시간만이 추가될 뿐이며, 이는 전체적으로 소요되는 심층 신경망 모델 훈련 시간에 크게 영향을 끼치지 않는다. 따라서 우리가 제안한 방법은 기존 방법의 훈련 시간을 유지하면서 성능을 향상할 수 있다.

Model	Method	Verified Error	Clean Error
Large	IBP	15.55%	2.27%
	Ours	**13.77%**	2.17%
Small	IBP	17.93%	3.48%
	Ours	**16.97%**	3.59%

표 2. 기존 구간 경계 전파 방법과 제안 방법의 비교. IBP는 기존 구간 경계 전파 방법을 의미하고 Ours는 우리의 제안한 방법을 의미한다.

4. 실험

우리는 기존의 구간 경계 전파 방법으로 훈련된 심층 신경망과 우리가 제안한 훈련방법으로 구간 경계 전파 방법을 훈련한 심층 신경망의 인증된 방어 성능을 비교하기 위해 MNIST 데이터 셋에 대해 실험을 진행하였다. 이번 단원에서 우리는 심층 신경망의 구조를 바꾸면서 우리가 제안한 방법을 통해 구간 경계 전파 방법의 성능을 향상할 수 있음을 보일 것이다.

먼저 우리는 ϵ_{train} =0.4로 기존 구간 경계 전파 논문[6]에서의 심층 신경망인 large 모델과 small 모델을 다시 학습시켰으며 실험 환경은 기존 논문의 환경과 동일하게 구성하였다.

우리가 제안한 훈련방법으로 구간 경계 전파 방법을 훈련할 때는 기존 방법과 같이 large 모델과 small 모델을 ϵ_{train} =0.4로 학습시켰으며, 예열 단계 10 세대(epoch), 준비 단계 50 세대, 안정화 단계 140 세대로 진행하였다. 아담 최적화(Adam optimizer)를 사용했으며 초기의 학습률(learning rate)는 5e-4로 총 130번째 세대 190번째 세대 때 학습률을 1/10으로 줄였다. 총 입력 배치의 크기는 200으로 설정하였다.

표 2에서는 기존 구간 경계 전파 방법으로 훈련된 모델의 ϵ =0.4일 때의 인증 에러와 일반 에러, 우리가 제안한 방법으로 훈련된 모델의 ϵ =0.4일 때 인증 에러와 일반 에러를 볼 수 있다. Large 모델과 Small 모델의 경우 둘 다 우리가 제안한 방법으로 훈련된 모델의 인증 에러가 낮아지는 것을 볼 수 있다. 구체적으로는 Large 모델의 경우에는 인증 에러가 1.78% 낮아졌으며, Small 모델의 경우에는 인증 에러가 0.96% 낮아졌다. 이러한 결과는 Large model에서 효과적이었던 α 값이 구조가 다른 모델인 small model에서도 효과적이라는 것을 알 수 있다.

5. 결론

본 논문에서는 인증 가능한 방어 방법 중에서 효율적인 기존의 구간 경계 전파 방법의 훈련 과정을 수정함으로써 훈련에 걸리는 시간을 유지하면서 성능을 향상할 방법을 제안하였다. 심층 신경망 모델을 효율적으로 학습하기 위해 입력 배치에 더 높은 ϵ에 대한 가이드라인을 제시하는 방법을 고안했으며, MNIST 데이터 셋에 대한 실험에서 가이드라인과 학습 난도에 따른 심층 신경망의 인증 가능한 방어 성능에 대한 영향을 분석하였다. 앞으로는 인증된 방어 방법의 성능을 더욱 개선하기 위해 적절한 α 값을 찾는 방법과 입력 배치를 구성하는 방법에 관한 연구가 필요할 것이다.

참고문헌

[1] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Finer, Beat Flepp, Prasoon Goyal, Lawrence DJackel, Mathew Monfort, Urs, Muller, Jiakai Zhnag, et al. “End to end learning for self-driving cars.” arXiv preprint arXiv:1604.07316,2016

[2] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Black-well, Harry Askham, Xavier Glorot, Brendan O’Donoghue,Daniel Visentin, et al. “Clinically applicable deep learning fordiagnosis and referral in retinal disease.” Nature medicine,24(9):1342-1350, 2018.

[3] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha,and Ananthram Swami. “Distillation as a defense to adversarial perturbations against deep neural networks.” In 2016 IEEE Symposium on Security and Privacy (SP), pages 582–597. IEEE, 2016.

[4] Nicholas Carlini and David Wagner. “Towards evaluating the robustness of neural networks.” In 2017 ieee symposium on security and privacy (sp), pages 39–57. IEEE, 2017

[5] Anish Athalye, Nicholas Carlini, and David Wagner. “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples.“ In International Conference on Machine Learning, pages 274–283, 2018.

[6] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. “On the effectiveness of interval bound propagation for training verifiably robust models.” arXiv preprint arXiv:1810.12715, 2018.

[7] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and ChoJui Hsieh. “Towards stable and efficient training of verifiably robust neural networks.“ arXiv preprint arXiv:1906.06316, 2019.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples.” arXiv preprint arXiv:1412.6572, 2014.

[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, JoanBruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. “Intriguing properties of neural networks.” arXiv preprint arXiv:1312.6199, 2013.