배치 정규화 레이어 파인튜닝을 통한 적대적 학습의 일반화 성능 및 견고성 개선

이정민 [1], 김광수 [2]
[1] 성균관대학교 소프트웨어학과 석사과정
[2] 성균관대학교 소프트웨어학과 교수

starjungmin@skku.edu, kim.kwangsu@skku.edu

Improving generalization performance and robustness of adversarial learning through fine-tuning batch normalization layers

Jungmin Lee[1], Kwangsoo Kim[2]
[1]Dept. of Computer Science and Engineering, Sungkyunkwan University
[2]Dept. of Computer Science and Engineering, Sungkyunkwan University

요 약

딥러닝 모델은 이미지 분류를 비롯한 다양한 분야에서 우수한 성능을 보여주고 있으나, 적대적 공격(adversarial attack)에 취약하다. 이를 해결하기 위해 적대적 학습(adversarial training)을 통해 적대적 공격에 대한 견고성을 강화하지만, 정상 데이터에 대한 성능 저하가 동반되는 문제점이 있다. 이러한 일반화 성능의 하락이라는 문제를 완화하기 위해, 본 연구는 견고성 보존과 일반화 성능 향상에 있어서 배치 정규화(Batch Normalization, BN) 레이어의 중요성을 알아보고, RiFT(Robustness Critical Fine-Tuning) 기법을 확장하여 파인튜닝 대상에 포함하는 방안을 제안한다. 또한, BN 레이어를 포함하여 파인튜닝하는 경우 적대적 학습된 모델의 통계적 특성을 효과적으로 보정함으로써 견고성과 일반화 성능을 동시에 개선할 수 있음을 실험적으로 검증한다.

1. 서론

딥러닝 모델은 이미지 분류, 객체 탐지 등 다양한 컴퓨터 비전 분야에서 탁월한 성능을 달성했지만, 적대적 공격에 대한 취약성[1]이 중요한 문제로 남아있다. 적대적 공격이란 입력 데이터에 인간의 눈으로는 거의 감지하기 힘든 수준의 미세한 변화(perturbation)를 의도적으로 추가하여, 딥러닝 모델이 잘못된 예측을 하도록 유도하는 공격을 의미한다. 이러한 공격은 모델의 오작동을 유발할 수 있으며, 특히 자율 주행, 의료 영상 분석, 보안 시스템과 같이 신뢰성과 안전성이 매우 중요한 응용 분야에서는 심각한 위험을 초래할 수 있다. 따라서 적대적 공격에 대한 모델의 견고성을 확보하는 것은 딥러닝 기술의 신뢰도를 높이고 안전한 활용을 위해 반드시 해결해야 할 핵심 과제이다. 적대적 학습은 이러한 공격에 대한 모델의 견고성을 높이는 가장 효과적인 방법의 하나로 알려져 있으나[2], 정상 데이터에 대한 모델의 일반화 성능을 희생시키는 트레이드오프를 치른다[2, 3].

적대적 학습에 있어서 일반화 성능-견고성 트레이드오프 문제를 완화하기 위해 MRC(Module Robust Criticality) 기반의 RiFT 기법[4]이 제안되었다. RiFT 는 적대적 학습된 모델 내에서 각 모듈(레이어)이 견고성에 기여하는 정도를 MRC 로 정의하고, 이 값이 낮아 견고성에 덜 중요한 모듈만을 선택적으로 파인튜닝한다. 이를 통해 Zhu et al.[4]은 전체 가중치를 업데이트하는 방식에 비해 견고성 손실을 최소화하면서도 저하되었던 일반화 성능을 효과적으로 회복할 수 있음

을 보였다.

적대적 학습 환경에서 딥러닝 모델의 견고성 및 일반화 성능은 정상 데이터와 적대적 예제 간의 통계적 분포 차이에 민감하게 반응한다.[5, 6], 이는 모델의 최종 견고성 및 일반화 성능에 직접적인 영향을 미칠 수 있다. 하지만 RiFT 기법은 BN 레이어를 MRC 측정 대상에서 제외함에 따라, 정상 데이터와 적대적 예제 간의 통계적 분포 차이와 그 영향력을 고려하지 못하게 된다.

이러한 배경에서, 본 연구는 기존 RiFT 기법이 BN 레이어를 고려하지 않음으로써 발생하는 한계를 극복하고 일반화 성능과 견고성 간의 트레이드오프를 개선하고자 한다.

2. 관련 연구

2.1 적대적 학습과 일반화 성능의 트레이드오프

Madry et al.[2]이 제안한 PGD(Projected Gradient Descent) 기반 적대적 학습은 현재까지도 가장 대중적인 적대적 방어 기법의 하나로 평가받는다. 그러나 이 방법은 모델의 견고성을 크게 향상시키는 대신, 정상 데이터에 대한 정확도를 상당 부분 감소시키는 경향이 있다. Tsipras et al.[3]은 이러한 현상이 단순한 경험적 관찰을 넘어, 견고성과 일반화 성능이 특정 조건에서는 근본적으로 상충될 수 있음을 이론적으로 보이기도 했다.

2.2 MRC 기반 RiFT 기법

Zhu et al.[4]은 적대적 학습된 모델의 일반화 성능 저하 문제를 해결하기 위해 RiFT(Robust Critical Fine-Tuning)를 제안했다. 이 방법은 모델의 각 모듈(레이어)이 견고성에 기여하는 정도를 'MRC(Module Robust Criticality)'로 측정하고, 중요도가 낮은 모듈만을 선택적으로 파인튜닝한다. 이를 통해 전체 가중치를 업데이트하는 완전 파인튜닝 방식보다 견고성 손실을 최소화하면서 일반화 성능을 효과적으로 회복할 수 있음을 보였다.

2.3 BN 레이어의 역할과 적대적 환경에서의 중요성

Ioffe 와 Szegedy[7]가 제안한 배치 정규화(Batch Normalization, BN) 레이어는 미니배치 내 데이터의 평균과 분산을 기반으로 활성화 값을 정규화함으로써 내부 공변량 변화(internal covariate shift)를 줄이고 학습 안정성과 수렴 속도를 향상시키는 데 기여한다. 이후 다양한 연구들[5, 6, 8]은 BN 이 단순한 학습 안정화 기법을 넘어서, 입력 분포의 변화에 민감하게 반응하며, 특히 분포가 변하는 환경에서는 BN 통계가 성능에 결정적인 영향을 미친다는 점을 밝혔다.

특히 적대적 학습 환경에서는 정상 이미지와 적대적 예제 간의 통계적 분포 차이가 존재하며, 이러한 분포 불일치는 BN 의 실행 평균(running mean)과 분산(running variance)에 왜곡을 유발하여 모델의 일반화 성능 또는 견고성을 저하시킬 수 있다[7].

Guo et al.[9]는 적대적 학습이 BN 통계치를 변화시킴으로써, 그 결과 BN 이 특정 입력 유형에 과적합되어 도메인 간 일반화 성능이 저하될 수 있다고 보고하였다. 또한 Walter et al.[10]은 BN 레이어가 '취약한 특징(fragile features)'에 과도하게 의존하게 만들 수 있으며, 오직 BN 레이어만 미세 조정하더라도 모델의 견고성이 향상됨을 입증하였다. Awais et al.[11]은 BN 통계의 재조정이 도메인 간 분포 불일치를 완화하고, 결과적으로 적대적 학습 모델의 견고성을 높이는 데 핵심적인 역할을 한다고 밝혔다.

3. 제안 기법: BN 을 포함한 RiFT (RiFT with BN)

3.1 MRC 평가 범위 확장

본 연구에서 제안하는 방식으로, MRC 계산 시 BN 레이어도 후보군에 포함한다. 이후 MRC 가 낮은 상위 k 개의 레이어(BN 또는 Conv/Linear)를 선택하여 파인튜닝을 진행한다. 이는 BN 레이어를 조정함으로써 도메인 간 정규화 불일치를 완화하고, 일반화 성능을 향상시키기 위한 접근이다.

3.2 통계치 재적응 전략

BN 레이어의 경우 학습 가능한 파라미터(gamma, beta)뿐만 아니라, 실행 통계(running mean, running var)도 모델의 동작에 큰 영향을 미친다. 따라서 본 연구에서는 선택된 BN 레이어의 실행 통계 또한 파인튜닝 과정에서 업데이트되도록 하여, 적대적 분포와 정상 분포 간의 불일치를 완화할 수 있도록 한다. 이는 일반적으로 적대적 학습 환경에서 BN 이 한 도메인에 과적합되는 현상을 완화하는 데 효과적이다.

4. 실험 설정

4.1 데이터셋 및 모델

본 연구에서는 CIFAR-10 과 CIFAR-100 데이터셋을 사용하였다. 두 데이터셋은 각각 10 개 및 100 개의 클래스로 구성된 32×32 크기의 컬러 이미지로 이루어져 있으며, 각각 학습용 50,000 장, 테스트용 10,000 장의 이미지를 포함한다. 실험에 사용된 모델은 ResNet-18 으로, PGD 기반의 적대적 학습(AT)[2]을 통해 사전 학습된 가중치를 사용하였다.

4.2 적대적 공격 방법

모델의 적대적 견고성을 평가하기 위해 PGD (Projected Gradient Descent) 공격 [2]을 사용하였다. PGD 는 입력 이미지에 대한 손실 함수의 그래디언트를 이용해 반복적으로 입력 이미지를 교란하는 대표적인 1 차 최적화 기반 공격 방법이다.

본 연구에서는 RiFT [4]의 설정을 따르며, 최대 허용 교란 범위 $\varepsilon = 8/255$, 이동 크기 $\alpha = 2/255$, 반복 횟수는 10 회(PGD-10)로 설정하였다.

4.3 실험 구성

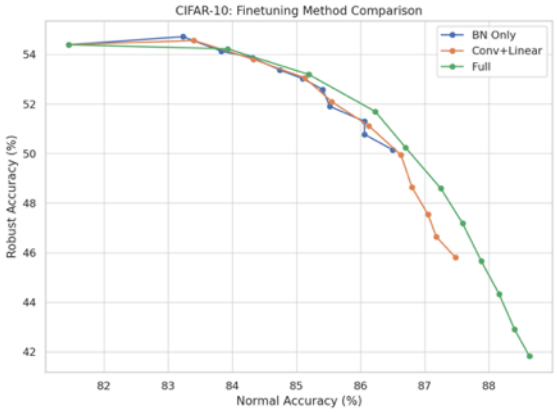
I. **실험 1: 레이어 유형별 파인튜닝 결과 비교**
BN-only, Conv/Linear-only, Full fine-tuning 세

가지 방식의 비교를 통해, BN 레이어의 견고성 보존 효과와 일반화 성능 회복 가능성을 비교한다.

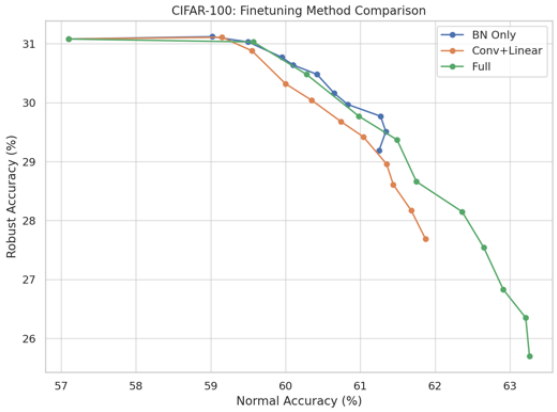
II. **실험 2: RiFT 와 RiFT with BN 성능 비교**
기존 RiFT 와 RiFT with BN 간의 비교를 통해, BN 레이어를 포함한 선택적 파인튜닝의 유효성을 Top-k 설정(k=1~5)에 따라 평가하였다. 이 실험은 견고성과 일반화 성능 간의 균형 유지 측면에서 BN 레이어의 조정이 기여하는 정도를 분석하는 데 초점을 둔다.

5. 결과 및 분석

5.1 파인튜닝 레이어에 따른 견고성 보존



(그림 1) CIFAR-10 에서 파인튜닝 방식에 따른 결과 비교



(그림 2) CIFAR-100 에서 파인튜닝 방식에 따른 결과 비교

그림 1 과 그림 2 는 각각 CIFAR-10, CIFAR-100 데이터셋에서 실험 결과를 나타낸 것이다. 적대적 학습[2]된 ResNet-18 모델에서 파인튜닝 레이어에 따른 정확도(Normal Accuracy)와 견고성(Robust Accuracy)의 변화를 보여준다. BN 레이어만 파인튜닝하는 경우와 Conv/Linear 레이어만 파인튜닝하는 경우 모두 초기에는 정확도와 견고성의 동시 향상을 확인할 수 있다. 그러나 파인튜닝이 진행될수록 Conv/Linear 레이어만 파인튜닝하는 경우는 BN 레이어만 파인튜닝하는 경우에 비해 견고성의 손실이 커진다. 전체 레이어를 파인튜닝하는 방식은 가장 높은 일반화 성능을 달성하지만, 견고성이 현저하게 저하되는 전형적인 트레이드오프 현상을 보여준다. 이는 파인튜닝 대상 레이어 유형에 따라 모델의 일반화 및 견고성에 기여하는 정도가 다름을 보여준다. 또한, 두 데이터셋에서 일관된 경향은 BN 레이어와 Conv/Linear 레이어가 파인튜닝 시 모델 성능에 미치는 영향이 상이함을 재확인시켜준다.

5.2 RiFT 기반 파인튜닝 결과 분석

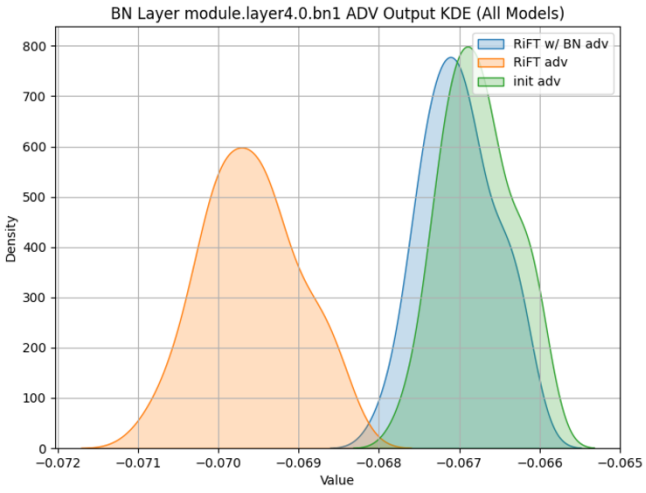
Top k	Method	Normal Acc	Adv Acc
	Base[2]	81.45%	53.75%
Top 1	RiFT	83.14%	53.91%
	Δ RiFT	+1.69%	+0.16%
	RiFT with BN	**82.64%**	**53.86%**
	Δ RiFT with BN	+1.19%	+0.11%
Top 2	RiFT	81.93%	53.63%
	Δ RiFT	+0.48%	-0.12%
	RiFT with BN	**82.46%**	**53.95%**
	Δ RiFT with BN	+1.01%	+0.20%
Top 3	RiFT	82.00%	53.63%
	Δ RiFT	+0.55%	-0.12%
	RiFT with BN	**82.62%**	**53.98%**
	Δ RiFT with BN	+1.17%	+0.23%
Top 4	RiFT	82.06%	53.63%
	Δ RiFT	+0.61%	-0.12%
	RiFT with BN	**82.48%**	**53.82%**
	Δ RiFT with BN	+1.03%	+0.07%
Top 5	RiFT	82.12%	53.56%
	Δ RiFT	+0.67%	-0.19%
	RiFT with BN	**81.90%**	**53.89%**
	Δ RiFT with BN	+0.45%	+0.14%

<표 1> Top-k 에 따른 RiFT 결과 비교

표 1 은 CIFAR-10 데이터셋을 기반으로 한 RiFT 와 제안 기법인 RiFT with BN 의 효과를 비교한다. 각각에 대해 일반화 성능(Normal Accuracy)과 PGD 공격에 대한 견고성(Robust Accuracy)을 Top-k (k=1~5) 설정에 따라 측정하였다.

측정 결과, RiFT with BN 은 기존 RiFT 대비 전반적으로 더 안정적인 성능 균형을 유지했다. 특히 k 값이 2 이상으로 증가하는 경우, 기존 RiFT 는 일반화 성능의 향상 폭이 감소하고 견고성 또한 하락(-0.12% ~ -0.19%)하는 경향을 보였다. 이는 MRC 점수가 낮은 레이어라 하더라도, 다수의 레이어를 동시에 파인튜닝할 경우 견고성 손실이 누적되어 전체적인 견고성이 감소할 수 있음을 보여준다.

반면, RiFT with BN 은 레이어 수가 증가하더라도 견고성과 일반화 성능 모두 향상(+0.07% ~ +0.23%)시키는 모습을 보였다. 이러한 결과는 BN 레이어가 모델 내부의 통계적 특성을 효과적으로 조정함으로써, 정상 데이터는 물론 잠재적인 적대적 분포 변화에까지 적응할 수 있도록 구조적 유연성을 제공했기 때문이다.



(그림 3) Top-5 실험에서 파인튜닝 방식에 따른 출력 분포 비교

추가적으로, BN 레이어 파인튜닝이 모델의 분포 안정화에 미치는 영향을 분석하기 위해, RiFT with BN 모델과 기존 RiFT 모델, 그리고 파인튜닝 이전의 대적 학습된 모델(init)의 적대적 예제의 출력값에 대해 KDE 분석을 수행하였다.

그림 3은 중요 레이어로 식별된 BN 레이어의 출력 분포를 각 모델별로 비교한 결과이며, RiFT with BN 의 경우 초기 모델(init)과 유사한 분포를 유지하는 반면, 기존 RiFT 는 분포가 뚜렷하게 이동된 양상을 보였다.

BN 레이어를 함께 파인튜닝할 경우, 적대적 학습을 통해서 학습된 통계적 특성을 그대로 유지할 수 있으며, 파인튜닝 이후에도 견고성의 하락을 효과적으로 억제할 수 있음을 의미한다.

6. 논의

본 연구는 적대적 학습 모델에서 일반화 성능과 견고성 간의 트레이드오프를 완화하기 위한 파인튜닝 전략으로서, BN 레이어의 중요성에 주목하였다. 실험 결과, BN 레이어만을 단독으로 파인튜닝하는 경우에도 상당한 수준의 견고성을 유지하며 일반화 성능을 향상시킬 수 있음이 확인되었으며(그림 1, 그림 2), 이는 BN 레이어가 단순한 정규화 기능을 넘어 모델의 견고성에 직접적으로 기여하고 있음을 보여준다.

또한, 기존 MRC 기반 파인튜닝 기법인 RiFT 에 BN 레이어를 통합한 'RiFT with BN' 전략을 적용한 결과, 기존 방식에 비해 다중 레이어 파인튜닝 시 견고성과 일반화 성능 모두에서 더 나은 성능을 보였다. 제안 기법을 사용하면 모델이 정상 데이터와 적대적 예제 간의 통계치 변화에 보다 유연하게 대응할 수 있다. 이는 BN 레이어의 분포 안정화 특성이 Conv/Linear 레이어 파인튜닝 과정에서 발생하는 견고성 저하를 효과적으로 완화시킬 수 있기 때문이다.

제안 기법은 기존 RiFT 방식이 BN 레이어의 중요성을 충분히 반영하지 못함으로써, 다중 레이어 파인튜닝 과정에서 성능 개선 기회를 놓쳤음을 보여준다. 아울러, BN 레이어를 통합하는 접근은 단일 레이어 중요도 기반 평가의 한계를 넘어서, 네트워크 내 다양한 레이어 유형 간의 상호작용을 고려한 더 효과적인 파인튜닝 전략으로 이어질 수 있다.

ACKNOWLEDGEMENT

이 논문은 2025 년도 정부(개인정보보호위원회)의 재원으로 한국인터넷진흥원의 지원을 받아 수행된 연구임 (RS-2023-00231200, 자율주행 환경에서 AI 학습 가능한 개인영상정보 프라이버시 보존 기술개발)

참고문헌

[1] I. Goodfellow et al., Explaining and harnessing adversarial examples. ICLR, 2015.

[2] A. Madry et al., Towards deep learning models resistant to adversarial attacks. ICLR, 2018.

[3] D. Tsipras et al., Robustness may be at odds with accuracy. ICLR, 2019.

[4] K. Zhu et al., Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning. ICCV, 2023.

[5] P. Benz et al., Revisiting batch normalization for improving corruption robustness. WACV, 2021.

[6] C. Xie, A. Yuille, Intriguing Properties of Adversarial Training at Scale. ICLR, 2020.

[7] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML, 2015.

[8] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, Theoretically principled trade-off between robustness and accuracy. ICML, 2019.

[9] C. Xie et al., Adversarial examples improve image recognition. CVPR, 2020.

[10] N. Walter, D. Stutz, and B. Schiele, On fragile features and batch normalization in adversarial training. arXiv preprint, arXiv:2204.12393, 2022.

[11] M. Awais, F. Shamshad, and S. Bae, Towards an adversarially robust normalization approach. arXiv preprint, arXiv:2006.11007, 2020.