오류 정정 출력 부호와 변분 정보 병목을 활용한 강건 신경망

장우람, 박호성*

전남대학교 지능전자컴퓨터공학과

wjang@jnu.ac.kr , *hpark1@jnu.ac.kr

Robust Neural Networks via Error-Correcting Output Codes and the Variational Information Bottleneck

Wooram Jang, Hosung Park*

Chonnam National University

요 약

본 논문에서는 신경망의 적대적 강건성 강화를 위해 오류 정정 출력 부호(Error Correcting Output Codes: ECOCs)와 변분 정보 병목(Variational Information Bottleneck: VIB)을 결합한 ECOC-VIB 모델을 제안한다. 이전 연구 [1]에서 채널 잡음과 적대적 예제의 유사성에 착안해 ECOCs를 심층 신경망(Deep Neural Networks: DNNs)에 적용함으로써 최첨단 수준의 적대적 강건성을 달성한 바 있다. 본 연구의 핵심은 이에 VIB를 결합하여 ECOCs 앙상블의 잠재 벡터 분포를 정규 분포에 가깝게 압축함으로써 적대적 강건성을 개선하는 것이다. 그 결과, MNIST 데이터셋에서 적대적 훈련(Adversarial Training: AT) [2] 모델 대비 1/14 수준의 모델 용량만으로 약 4% 높은 적대적 정확도를 달성하였다.

Ⅰ. 서 론

심층 신경망(Deep Neural Networks: DNNs)은 다양한 분야에서 최첨단 수준의 성능을 입증하고 있으나, 여러 연구에서 적대적 예제에 취약하다는 문제점이 지적되고 있다 [3, 4]. 적대적 예제란 원본 입력에 작고 정교하게 설계된 교란을 추가하여 모델이 오분류하도록 유도한 데이터를 말한다. 이러한 취약성은 모델이 실제 환경에서 신뢰성을 유지하는데 중대한 위협이 된다. 이를 해결하기 위해 적대적 예제 탐지 및 적대적 학습 [2]과 같은 다양한 접근법이 제안되었으나, 현재까지 적대적 학습 기법만이 일반적인 공격에 대해 실질적인 방어 능력이 있는 것으로 확인되었다 [5].

적대적 학습 기법의 뛰어난 적대적 강건성에도 불구하고 한계는 명확하다. 훈련 단계에서 적대적 예제를 반복 생성해야 하는 계산 복잡도와 원본 정확도와 적대적 정확도사이의 trade-off가 존재한다. 적대적 예제를 생성하는 과정은 이미지 크기에 영향을 받으며, 적대적 정확도 향상은 원본 정확도의 하락을 초래한다.

이러한 한계를 극복하기 위해, 부호 이론에서 영감을 받은 ECOCs [6]를 도입하기 위한 시도가 있었으며, 최근 연구들은 ECOCs를 활용하여 최첨단 수준의 강건성을 달성하였다 [1]. 정보와 레이블 간의 유사성, 그리고 잡음 채널과 적대적 교란 간의 유사성은 적대적 학습 분야에서 ECOCs를 적용하게 된 동기가 되었다. ECOCs는 다중 분류 문제를 간단한 이진 분류 문제로 분해하여 해결하는 기법으로 신경망 구조와 레이블에 중복성을 추가하는 방식으로 구현한다. DNNs을 앙상블로 설계하여 구조 자체에 중복성을 추가하며, 레이블을 원-핫 벡터가 아닌 오류 정정 부호, 즉 코드워드로 인코딩한다. 따라서 ECOCs는 특징 학습을 위한 다양성과 오류 정정 능력을 확보할 수 있다. 각 코드워드들은 오류 정정을 위해 해밍 거리가 크도록 설계되어야 한다.

ECOCs는 적대적 훈련 기법에 비해 계산 비용이나 모델 크기 측면에서도 매우 효율적으로 높은 수준의 적대적 강건성을 유지할 수 있다. 본 연구에서는 ECOCs와 VIB의 결합을 통해 적대적 학습 모델 대비 약 1/14 수준의 모델 용량만으로 약 4% 높은 적대적 정확도를 달성하였다.

Ⅱ. 본론

2.1. 적대적 공격

DNNs이 적대적 예제에 취약하며, 공격자에 의해 신뢰성을 위협받을 수 있다는 것은 심각한 문제이다. 이러한 적대적 공격은 블랙박스 공격과 화이트박스 공격으로 구분할 수 있다. 블랙박스 공격은 공격자가 DNNs의 구조나 가중치 값에 접근할 수 없으며, 오직 입력과 출력만을 이용해 적대적 예시를 생성한다. 반면 화이트박스 공격은 DNNs의 구조와 가중치 값을 포함한 모든 정보를 활용할 수 있다. 화이트 박스 공격은 모델 정보를 이용해 보다 정교한 적대적 예시를 만들 수 있다. 화이트박스 공격은 DNNs의 신뢰도에 큰 영향을 미치기 때문에, 적대적 강건성 연구에서 널리 사용되어 왔다. 해당 분야에서 주로 사용되는 화이트 박스 공격의 예로는 투영 경사 하강법(Projected Gradient Descent: PGD) 공격 [4]이 있다.

PGD 공격은 빠른 기울기 부호 방법(Fast Gradient Sign Method: FGSM)을 반복해서 수행하는 정교한 공격 기법이다. PGD 공격은 식 (1)과 같이 정의된다.

$$\mathbf{x}^{t+1} = \Pi_{x+S}\left(\mathbf{x}^{t} + \alpha\,\mathrm{sgn}(\nabla_x L(\theta, \mathbf{x}, \mathbf{y}))\right). \quad (1)$$

x와 y는 신경망의 입력과 정답 레이블, α는 각 단계의 변형 정도를 의미한다. 또한 $sgn(\circ)$은 부호 함수, $L(\circ)$은 DNNs의 손실 함수, Π는 투영을 의미한다. PGD 공격은 각 t 단계에서 신경망의 손실함수 기울기를 따라 α만큼의 변형을 이전 단계의 이미지에 추가한다. FGSM 공격보다 정밀한 공격이 가능하여, DNNs의 정확도에 치명적인 영향을 줄 수 있다.

2.2. Error Correcting Output Codes

오류 정정 부호(Error Correction Codes: ECCs)은 본래 통신 분야에서 오류를 탐지·수정하기 위해 고안된 기술이다. 최근 통신상의 오류와 적대적 공격의 유사성에서 착안하여 DNNs의 강건성을 높이는데 활용되고 있다. 다중 분류에서 ECOCs는 DNNs이 출력한 비트열과 각 클래스에 미리 지정된 코드워드 간 해밍 거리를 계산해 거리가 가장 작은 클래스를 예측

값으로 결정한다. 해밍 거리가 작은 경우, 몇 비트만 반전되더라도 잘못 분류되므로, 코드북 설계 시 코드워드 간 거리를 크게 확보하는 것이 핵심이다. 본 연구는 코드 길이의 절반에 해당하는 최대 최소 해밍 거리를 갖는 하마다드 행렬을 코드북으로 사용해 오류 정정 능력을 보장한다. 예를 들어, 길이 16, 최소 거리가 8인 코드에서는 최대 3비트까지 오류를 정정할 수 있어, 하다마드 코드북이 적대적 환경에서 특히 유리함을 보여준다.

2.3. Variational Information Bottleneck

정보 병목(Information Bottleneck: IB)는 입력 신호 X가 목표 신호 Y에 대해 의미 있는 정보는 보존하면서, 그 외의 불필요한 세부 정보는 일반화 성능을 위해 최대한 압축하는 표현 Z를 찾는 기법이다. 즉 $X \rightarrow Z \rightarrow Y$ 마르코프 연쇄(Markov Chain)을 가정하고, 표현 Z와 Y 사이의 상호 정보 $I(Z;Y)$는 최대화하되, X와 표현 Z 사이의 상호 정보 $I(X;Z)$는 최소화한다. 라그랑주 승수 β를 도입하여 최대화 목적함수로 표현하면 식 (2)와 같다.

$$f_{IB}(\theta) = I(Z,Y;\theta) - \beta I(Z,X;\theta) \quad (2)$$

θ는 DNNs의 매개변수를 의미한다. 직관적으로 첫 번째 항은 Z가 Y를 잘 예측하도록 유도하고, 두 번째 항은 Z가 X에 대한 불필요한 정보를 잊도록 만든다는 것을 알 수 있다. 또한 β로 정보 압축 강도를 조정할 수 있으며, β가 클수록 Z에는 보다 압축된 정보가 남게 된다.

하지만 실제 데이터 분포를 알 수 없으며, IB 목적식은 상호정보량 추정 및 최적화가 어렵다는 한계가 있다. 따라서 DNNs에 적용하기 위해서는 최적화 가능한 형태로 변환하는 것이 필수적이다. 이전 연구 [7]에서는 upper bound와 reparameterization trick을 통해 식 (3)을 정의한다.

$$J_{IB} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\epsilon\sim p(\epsilon)}\left[-\log q(y_n|f(x_n,\epsilon))\right] + \beta\,\mathrm{KL}\left[p(Z|x_n), r(Z)\right] \quad (3)$$

N은 학습 데이터 개수, $f(x_n,\epsilon)$은 잠재 벡터 z_n, $q(y_n|f(x_n,\epsilon))$은 잠재 벡터 z로부터 y_n의 조건부 확률을 추정하는 디코더, $p(Z|x_n)$은 인코더가 출력한 잠재 벡터의 분포를 의미한다. 또한 $r(Z)$은 사전 분포로써, 인코더 잠재 벡터 분포와의 KL 발산항을 통해 정보 압축을 달성한다.

2.4. 실험 결과

본 연구에서는 10개의 클래스로 구성된 MNIST 데이터셋을 사용하였다. 코드워드 길이는 16으로, 하다마드 행렬의 상위 10개 행을 선택해 코드북을 구성하였다. MNIST 데이터셋에 대하여 epoch=500, batch size=200, learning rate=1e-03로 설정하여 최적화하였다. 모델의 강건성을 평가하기 위해 l_∞-PGD 공격을 수행하였으며, 허용 최대 노름 ϵ=0.3, 반복 횟수=100, 스텝 크기=0.0075로 설정하였다. ECOC 구조는 TanhEns [1] 구조와 동일하며, ECOC-VIB에 대해 β는 1e-01로 설정하였다.

Madry는 적대적 훈련을 적용한 모델로써, 훈련 과정 중 적대적 예제를 생성하는 방식으로 모델의 강건성을 향상시킨다. 반면 ECOC, ECOC-VIB 모델의 경우 적대적 훈련을 수행하지 않았으며, 표1의 결과는 오직 ECOC와 VIB 기법 여부에 따른 적대적 정확도이다.

MNIST (ϵ=0.3)

Models	Clean	PGD	# Params
Madry [2]	.985	.925	3,274,634
ECOC [1]	.982	.934	215,632
ECOC-VIB	**.980**	**.964**	**232,272**

표1. MNIST ECOC-VIB 모델 정확도

Ⅲ. 결론

본 연구에서는 오류 정정 출력 부호(ECOCs)의 오류 정정 능력과 변분 정보 병목(VIB)의 잠재 벡터 표현 압축 효과를 결합한 ECOC-VIB 모델을 제안하였다. 제안 모델은 ECOCs 기반 앙상블 구조를 통해 이진 분류기 사이의 다양성을 확보하고, VIB 정규화 항을 도입해 정보량을 조절함으로써 과적합을 억제하였다. MNIST 실험 결과, Madry 적대적 학습 모델에 비해 파라미터 수를 약 1/14로 줄이면서도 약 4% 높은 적대적 정확도를 입증하였다. 이는 VIB를 통한 잠재 벡터 압축이 ECOC의 오류 정정 효과와 결합되어, 적대적 학습 없이도 적대적 강건성을 향상 시킬 수 있음을 시사한다.

향후 연구에서는 CIFAR-10, CIFAR-100 등 복잡한 데이터셋으로의 확장 가능성을 검증하고, VIB 정보 압축 강도(β)의 동적 조정 및 코드북 설계를 통한 추가적인 경고성 및 효율성 개선을 목표로 한다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2022M3C1A3090857), (No. RS-2024-00410005)과 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구입니다.(IITP-2025-RS-2022-00156287, 20%).

참 고 문 헌

[1] Verma, Gunjan, and Ananthram Swami. Error correcting output codes improve probability estimation and adversarial robustness of deep neural networks. Advances in Neural Information Processing Systems vol. 32, 2019.

[2] Madry, Aleksander, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[3] Szegedy, Christian, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[4] Goodfellow, Ian J., Jonatho1'n Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[5] Tramer, Florian, et al. On adaptive attacks to adversarial example defenses. Advances in neural information processing systems vol. 33, 2020: 1633-1645.

[6] Dietterich, Thomas G., and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. Journal of artificial intelligence research 2, 1994: 263-286.

[7] Alemi, Alexander A., et al. Deep variational information bottleneck. arXiv preprint arXiv:1612.00410, 2016.