적대적 훈련과 엔트로피 최소화를 통한 추론 시기 적응의 성능 및 견고성 향상

강민준, 조우성, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

Enhancing Performance and Robustness in Test-Time Adaptation through Adversarial Training and Entropy Minimization

Minjun Kang, Wooseong Cho, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

실제 상황에서 심층신경망은 학습 중에 접하지 못한 환경 변화나 센서 품질 저하와 같은 도메인 이동(Domain Shift)으로 인해 성능이 저하될 수 있다. 이러한 도메인 이동에 대응하기 위해 추론 시기 적응 방법론이 등장하였다. 또한, 심층 신경망 모델이 잘 학습되었더라도 적대적 예제 공격에 취약할 수 있으며, 이는 추론 시기 적응에 고려해야 할 요소이다. 본 논문에서는 엔트로피 최소화를 기반으로 한 추론 시기 적응 방법에 적대적 훈련을 통합하여, 적대적 공격에 대한 모델의 견고성을 높이면서 성능을 향상시키는 연구를 수행하였다. 실험 결과, 추론 시기 적응에 적대적 훈련을 함께 수행하는 것이 모델의 견고성을 향상하면서 추론 시기 적응에도 효과적임을 확인하였다.

Ⅰ. 서 론

잘 학습된 심층 신경망 모델이라도 실제 환경에 배포하는 것은 여러 어려움을 수반한다. 실제 환경에서 심층 신경망은 훈련 중에 보지 못한 날씨 변화, 센서 품질 저하와 같은 자연적인 변이를 마주할 수 있다[1]. 이처럼 모델이 학습된 환경과 추론 환경의 분포가 달라 성능 감소를 겪는 현상은 도메인 이동(Domain Shift) 문제로 알려져 있다[2].

추론 시기 적응(Test Time Adaptation, TTA)은 도메인 이동 문제에 대응하기 위해 추론 시기의 데이터를 활용하여 심층 신경망을 지속적으로 적응시키는 방법론이다[1]. 즉, TTA 는 훈련에 사용한 기준(Source) 도메인 데이터의 활용 없이, 오로지 추론 시기의 목표(Target) 도메인 데이터만을 활용한다.

한편, 잘 사전 학습된 심층 신경망일지라도 적대적 예제 공격에 취약할 수 있으며, 모델이 가진 취약성은 하위 과업(Downstream Task)에도 전이될 수 있다[3]. 적대적 예제 공격은 모델이 높은 신뢰도로 틀린 예측을 하도록 유도하는 공격이며, 이러한 공격은 모델의 예측을 왜곡해 잘못된 결과를 매우 확신하면서 내리게 한다. 이는 엔트로피(Entropy) 최소화를 기반으로 모델을 적응하는 TTA 방법에도 치명적이다[4].

따라서, 우리는 엔트로피 최소화를 기반으로 하는 대표적인 TTA 방법인 Tent[2]에 적대적 훈련(Adversarial Training)을 적용하기 위해 대표적인 적대적 예제 생성(Adversarial Attack) 방법인 FGSM(Fast Gradient Signed Method)[5]을 이용하였다. 이를 통해, 적대적 예제 공격에 대한 모델의 견고성을 향상하면서 추론 환경에 모델이 성공적으로 적응할 수 있는지 탐색하였다.

Ⅱ. 본 론

본 논문에서는 사전 학습된 합성곱 신경망에서 엔트로피 최소화 기반 TTA 방법인 Tent 와 적대적 훈련을 효과적으로 적용하는 방법을 연구하였다.

엔트로피 최소화는 그림 1 의 초록색 화살표와 같이, 목표 데이터 x^t에 대한 모델의 예측 확률 $\hat{y}^t$를 이용하여 엔트로피 손실 $L_{Entropy} = -\sum_c \hat{y}_c^t \log(\hat{y}_c^t)$ 을 최소화하는 방식으로 수행된다. 여기서, c 는 각 클래스(Class)를 의미하며, 모델이 x^t 에 대해 예측할 수 있는 가능한 클래스의 집합을 나타낸다. 이후 그림 1 의 회색 점선 화살표에서 보이듯이, FGSM 을 활용하여 적대적 훈련을 위한 적대적 예제x_{adv}^t를 생성한다. TTA 에서 FGSM 을 이용한 적대적 예제 생성을 위해, 우리는 데이터의 정답값을 사용하지 않고, 모델의 예측 $\hat{y}^t$ 에서 $\mathrm{argmax}(\hat{y}^t)$를 통해 가상 정답값(Pseudo Label)으로 삼아 적대적 예제를 생성하였다. 이후, 그림 1 의 빨간색 화살표와 같이, 생성한 적대적 예제 x_{adv}^t에 대한 모델의 예측 확률 $\hat{y}_{adv}^t$ 와 $\hat{y}^t$ 의 KL 발산(Kullback-Leibler divergence) 손실 $L_{KL} = \sum_c \hat{y}_c^t \log(\frac{\hat{y}_c^t}{\hat{y}_{adv,c}^t})$을 최소화하도록 적대적 훈련을 수행하였다. 따라서, 전체적인 학습 목표는 $L_{Entropy}$ 와 L_{KL}을 결합한 $L = \alpha \cdot L_{Entropy} + (1-\alpha) \cdot L_{KL}$ 와 같다. 여기서 α는 엔트로피 최소화 손실 $L_{Entropy}$ 와 적대적 훈련의 KL 발산 손실 L_{KL} 사이의 균형을 조절하는 초매개변수이다.

우리는 제안한 학습 방법이 모델을 실제 환경에서 발생할 수 있는 변형에 효과적으로 적응시키면서 적대적 예제 공격에 대한 모델의 견고성도 향상시킬 수 있는지 분석하였다. 또한, 적대적 학습에서 KL 발산 손실 L_{KL}과 크로스 엔트로피(Cross Entropy) 손실 L_{CE}에 따른 실험 비교를 통해 TTA 에서 효과적인 적대적 훈련을 위한 손실 함수를 탐구하였다.

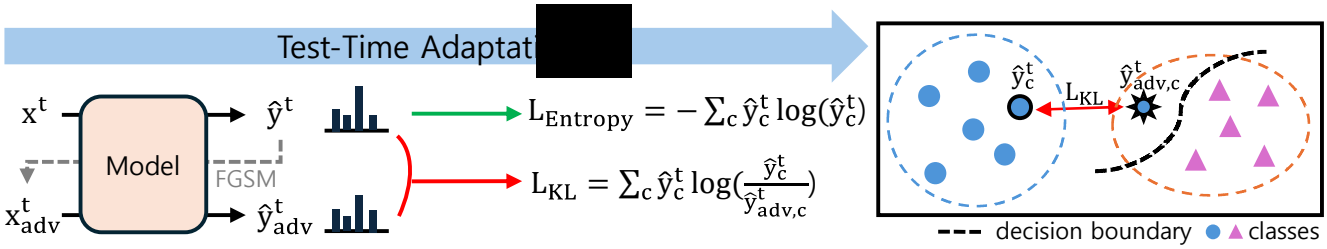

그림 1 적대적 훈련과 엔트로피 최소화 기반 추론 시기 적응

Ⅲ. 실 험

실험에는 CIFAR10[6] 데이터 집합을 기준 도메인으로 사전 학습된 ResNet18[7] 모델을 사용하여, CIFAR10-C[8] 데이터 집합을 목표 도메인으로 TTA 을 수행하였다. CIFAR10-C 은 15 가지의 다양한 변형을 포함하고 있으며, 각 변형의 정도인 심각도(Severity)에 따라 모델을 다양한 변형 강도로 평가할 수 있다. 실험의 성능은 사진 분류 과업에 대한 오류율(Error, %)을 이용하여 측정하였다. 실험 결과는 표 1 에 요약되어 있으며, CIFAR10-C 의 15 가지 변형에 대한 각 실험 결과의 평균을 나타냈다. 실험에 사용한 초매개변수는 $\alpha = 0.1$, 적대적 예제 생성을 위한 잡음 $\epsilon = 0.01$ 이며, 굵은 글씨는 가장 높은 성능을 나타낸다.

표 1 의 Method 열에서 Baseline 은 사전 학습된 Resnet18 을 어떠한 적응 없이 CIFAR10-C 에 대해 평가한 결과를 나타낸다. TENT 는 선행 연구인 Tent 방법을 이용하여 엔트로피 최소화 기반의 TTA을 수행한 것이다. Ours(CE)과 Ours(KL)은 Tent 방법과 적대적 훈련을 통합한 우리의 제안 방법이다.

효과적인 적대적 훈련을 위한 손실 함수 탐색 실험으로, Adv.tr. + CE 는 적대적 훈련에서 크로스 엔트로피(Cross Entropy) 손실 L_{CE} 을 사용한 방법이며 Ours 는 KL 발산 손실 L_{KL} 을 사용한 방법이다. 구체적으로, Adv.tr. + CE 는 목표 도메인 데이터 x^t 에 대한 예측 확률 분포 $\hat{y}^t$ 을 $\text{argmax}(\hat{y}^t)$ 을 통해 가장 확률이 높은 클래스를 가상 정답값으로 삼아, 적대적 예제 x_{adv}^t 에 대한 예측 분포 y_{adv}^t 와 크로스 엔트로피 손실 L_{CE}을 최소화하도록 한 실험이다. 반면, Ours 은 두 예측 확률 분포 $\hat{y}^t$ 와 y_{adv}^t 간의 KL 발산 손실 L_{KL} 을 최소화하는 방식으로 적대적 훈련을 수행한 것이다.

표 1 의 Target 열은 TTA 에 사용한 데이터를 나타낸다. 표 1 의 CIFAR10-C 열은 CIFAR10-C 데이터 집합에 대한 TTA 후 모델의 예측 오류율을 측정한 값으로, 추론 시점에서 모델이 입력 데이터에 대해 잘못 예측한 비율이다. Adv. Attack 열은 TTA 후, 모델의 견고성을 측정하기 위해 적대적 예제 공격을 수행하여 적대적 예제에 대해 측정한 오류율을 나타낸다. 이때, 적대적 예제 공격은 사전 학습된 원본 모델 ResNet18 을 이용하여 CIFAR10 테스트 데이터 집합에 대해 적대적 예제를 생성하여 수행하였다.

실험 결과 엔트로피 최소화와 적대적 훈련을 통합한 우리의 방법이 변형된 데이터와 적대적 예제 공격에 대한 오류율을 Baseline 대비 각 3.70%p 와 15.66%p 낮추었다. 이는 TTA 에서 적대적 훈련이 모델의 견고성을 향상시키면서 모델 적응 성능 또한 증진시킬 수 있음을 보여준다. 또한, 적대적 훈련에서 목표 데이터와 적대적 예제를 동일한 클래스로 분류하도록 하는 크로스 엔트로피 손실 L_{CE} 보다 적대적 예제에 대한 모델 예측이 목표 데이터에 대한 모델 예측을 따르도록 유도하는 KL 발산 손실 L_{KL}이 더 효과적임을 확인할 수 있다. 이는 KL 발산 손실 L_{KL} 을 이용한 학습이 모델이 목표 도메인 데이터와 적대적 예제에 대해 일관된 예측 출력을 생성하도록 도와, 테스트 시점에서 더 강건한 성능을 발휘할 수 있음을 시사한다.

표 2 방법별 추론 시기 적응과 적대적 공격 오류율

Method	Target	CIFAR10-C	Adv. Attack
		Error(%, ↓)	Error(%, ↓)
Baseline	x^t	22.20	51.08
TENT[2]	x^t	18.88	40.15
Adv.tr. + CE	x^t, x_{adv}^t	18.64	35.72
Ours	x^t, x_{adv}^t	**18.50**	**35.42**

Ⅳ. 결 론

본 논문에서는 엔트로피 최소화 기반의 TTA 방법에서 적대적 훈련이 모델의 데이터 변형에 대한 적응과 견고성을 향상시킬 수 있음을 확인하였다. 특히, 적대적 예제에 대한 모델의 예측이 목표 데이터의 예측 분포를 따르도록 유도하는 접근이 더욱 효과적임을 확인하였다. 이러한 결과는 실제 환경에 모델을 배포하고 활용할 때, TTA 에 적대적 훈련을 적용하는 것이 매우 효과적임을 시사하며, 적대적 훈련의 도입이 필수적임을 강조한다.

ACKNOWLEDGMENT

본 연구는 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00167194,미션 크리티컬 시스템을 위한 신뢰 가능한 인공지능)

참 고 문 헌

[1] Wang, Dequan, et al. “Tent: Fully Test-Time Adaptation by Entropy Minimization.” *International Conference on Learning Representations*, 2021.

[2] QUIONERO-CANDELA, Joaquin, et al. Dataset Shift in Machine Learning. The MIT Press, 2009.

[3] Zhou, Ziqi, et al. "Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning." *Proceedings of the 31st ACM International Conference on Multimedia*. 2023.

[4] Wu, Tong, et al. “Uncovering Adversarial Risks of Test-Time Adaptation.” *Proceedings of Machine Learning Research*, vol. 202, 2023, pp. 37456-37495.

[5] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples.” *International Conference on Learning Representations*, 2015.

[6] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[7] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[8] Hendrycks, Dan, and Thomas G. Dietterich. “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations.” *International Conference on Learning Representations*, 2019.