POP: 확산 모델의 국소 복원 민감도를 활용한 적대적 샘플 탐지 지표 제안

POP: Adversarial Sample Detection Metric utilizing Local Restoration Sensitivity of Diffusion Models

이한진 [1].정재환 [1].정현진 [1].박수연 [2].최석환 [2*]
Han-Jin Lee, Jae-Hwan Jeong, Hyeon-Jin Jung, Su-Yeon Park and Seok-Hwan Choi

[1]연세대학교 전산학과
E-mail: {han-.-jin, jjh021101, wjdehdtod}@yonsei.ac.kr
[2]연세대학교 소프트웨어학부
E-mail: {swyk213, sh.choi}@yonsei.ac.kr

요 약

기존 적대적 예제 탐지 기법은 대상 모델에 종속적이라는 한계가 있다. 이에 본 논문은 사전 학습된 확산 모델을 활용한 Model-agnostic 탐지 지표 POP(Perturbation-to-Output Propagation)를 제안한다. 본 연구는 적대적 이미지가 인위적 노이즈로 포화되어 추가적인 미소 교란에 둔감하게 반응하는 '적대적 경직성(Adversarial Stiffness)' 현상을 새롭게 규명하였다. POP 는 이러한 복원 거동의 차이를 정량화하여, CIFAR-10 환경에서 타겟 모델 정보 없이 단일 이미지 연산만으로 FGSM 0.9880, PGD 0.9280 의 우수한 탐지 성능(AUROC)을 달성하였다.

키워드: Adversarial Example Detection, Diffusion Model, Adversarial Stiffness, Model-Agnostic Method, Local Sensitivity

1. 서론

딥러닝 기반의 시각 인식 시스템은 육안으로 식별할 수 없는 미세한 노이즈가 더해진 적대적 예제에 매우 취약하다 [1]. 대다수의 기존 탐지 기법은 대상 분류기의 내부 특징맵이나 출력 신뢰도 변화에 의존하므로 [2], 모델 정보에 접근할 수 없는 블랙박스 환경에서는 적용이 불가능하다는 한계가 있다.

따라서 대상 분류기의 정보 없이, 입력된 이미지 자체가 지니는 매니폴드 고유의 특성만을 분석하여 공격 여부를 판별하는 범용적인 탐지 기법이 요구된다. 최근 생성 모델 분야에서 괄목할 성과를 보이는 확산 모델은 방대한 데이터 매니폴드 구조를 학습하여 강력한 사전 지식을 제공하며, 이를 탐지에 활용하려는 최신 연구들이 등장하고 있다. 대표적으로 Lorenz 등 [3]은 확산 모델의 변환 과정을 거쳤을 때 적대적 예제가 정상 데이터의 매니폴드와 정렬되지(Misaligned) 않는 현상을 분석하여 탐지에 활용하였다. 하지만 이러한 방식은 역방향 변환 과정을 거친 후 별도의 분류기를 다시 학습시켜 판별해야 하므로 연산 복잡도가 높다는 단점이 있다.

본 논문에서는 사전 학습된 확산 모델 [4]을 이용하되, 전체 복원 과정을 추적할 필요 없이 이미지의 국소 복원 민감도만을 단일 연산으로 측정하는 새로운 탐지 지표 POP(Perturbation-to-Output Propagation)를 제안한다. 특히 본 연구는 적대적 예제가 확산 모델의 복원 과정에서 보이는 비정상적 둔감성인 '적대적 경직성'을 실험적으로 규명하고, 이를 기반으로 고성능 적대적 탐지가 가능함을 증명한다.

감사의 글 : 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2026-25472714).

2. 연구 배경 및 제안 기법

확산 모델은 원본 데이터에 점진적으로 가우시안 노이즈를 추가하는 정방향 과정과, 노이즈를 제거하며 원본 데이터를 복원하는 역방향 과정으로 구성된다. 확산 모델은 이 과정을 통해 데이터의 복잡한 확률 분포를 학습하며, 자연 이미지가 위치한 매니폴드로 데이터를 투영하는 성질을 지닌다.

직관적으로, 적대적 예제는 정상 데이터의 매니폴드 경계 밖으로 밀려나 있으므로 노이즈 주입 시 더 불안정한 복원 궤적을 보일 것으로 예상하기 쉽다. 그러나 실제 관측 결과, 적대적 이미지는 분류기를 속이기 위한 인위적 교란으로 이미 포화되어 있어 정반대의 현상을 보였다. 미소 교란이 추가되더라도 노이즈 예측 네트워크는 이를 기존의 거대한 교란에 묻힌 것으로 인식하여 복원 방향을 크게 바꾸지 않으며, 본 논문은 이러한 현상을 '적대적 경직성'으로 정의한다. 반면, 매니폴드 내부에 안정적으로 안착해 있는 자연 이미지는 미세한 노이즈에도 즉각적으로 반응하여 큰 복원 변화량을 보인다.

본 논문은 이러한 국소 민감도 차이를 측정하기 위해 탐지 지표 POP를 제안한다. 입력 이미지를 x, 주입된 미소 교란을 δ라 하고, 특정 타임스텝 t에서 확산 모델의 노이즈 예측 네트워크를 $\epsilon_\theta(\cdot, t)$라 할 때, POP 지표는 다음과 같이 정의된다.

$$POP(x) = \mathbb{E}_{\delta \sim \mathcal{N}(0,\sigma^2 \mathbf{I})}\left[\frac{\|\epsilon_\theta(x,t) - \epsilon_\theta(x+\delta,t)\|_2}{\|\delta\|_2}\right] \quad (1)$$

POP 지표는 입력 이미지에 미소 교란 δ를 가했을 때, 확산 모델이 예측하는 노이즈 벡터의 변화량을 교란의 크기로 정규화하여 계산한다. 전체 역방향 과정을 수행할 필요 없이 단 한 번의 타임스텝 t 에서의 예측 변화량만으로 매니폴드의 국소 기울기 변화를 측정할 수 있어 연산 효율성이 뛰어나다.

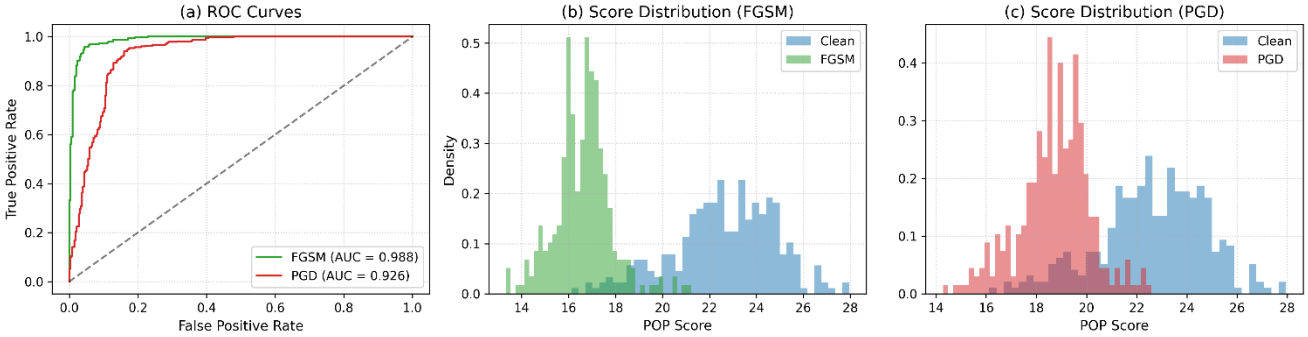

그림 1. 제안하는 POP 지표의 탐지 성능 및 분포. (a) FGSM 및 PGD 공격에 대한 ROC 곡선, (b, c) 정상 및 적대적 이미지 간의 POP 점수 분포 차이.

Fig. 1. Detection performance and distribution of the POP metric. (a) ROC curves for FGSM and PGD attacks. (b, c) POP score distributions between clean and adversarial images.

3. 실험

제안하는 POP 지표의 효과를 검증하기 위해 CIFAR-10 데이터셋을 활용하여 실험을 수행하였다. 방어자가 보호해야 할 대상 모델로는 사전 학습된 ResNet-20 을 사용하였으며, POP 지표 계산을 위한 확산 모델로는 외부 데이터 추가 학습 없이 CIFAR-10 으로 사전 학습된 DDPM 을 채택하였다. 적대적 공격 기법으로는 단일 스텝 공격인 FGSM [1]과 다중 스텝 반복 공격인 PGD [5]를 적용하였으며, 최대 섭동 크기는 8/255 로 설정하였다.

탐지 성능의 엄밀한 평가를 위해, 타깃 모델이 원래 정답을 맞힌 이미지 중에서 공격 기법을 통해 성공적으로 오분류를 유도한 '유효 적대적 샘플'만들 필터링하여 실험쌍을 구성하였다. POP 수식의 하이퍼파라미터는 타임스텝 $t = 10$, 노이즈 표준편차 $\sigma = 0.005$ 로 설정하였으며, 안정적인 기댓값 산출을 위해 샘플당 3 회의 몬테카를로 샘플링을 수행하여 평균값을 취하였다. 적대적 샘플과 정상 샘플 간의 분포 차이를 통해 AUROC 를 측정하였다.

4. 결과 및 결론

실험 결과, 제안하는 POP 지표는 정상 이미지와 적대적 이미지 간의 뚜렷한 분포 차이를 나타냈다. POP 점수의 분포를 시각화한 결과, 정상 이미지는 미소 교란에 민감하게 반응하여 높은 점수 대역(주로 22~25 점)에 분포한 반면, 적대적 이미지는 적대적 경직성으로 인해 매우 둔감하게 반응하여 낮은 점수 대역(18~20 점)에 밀집하는 것을 확인하였다 (그림 1 의 b, c 참조).

이러한 분포 분리를 바탕으로 탐지 성능을 측정한 결과, POP 지표는 FGSM 공격에 대해 0.9880, 보다 강력한 화이트박스 공격인 PGD 공격에 대해서도 0.9280 이라는 탁월한 성능을 달성하였다 (그림 1 의 a 참조).

또한, 확산 모델의 타임스텝 t에 따른 탐지 성능의 변화를 분석하기 위해 Ablation Study 를 수행하였다 (그림 2 참조). 실험 결과, 노이즈가 과도하게 누적된 큰 타임스텝 ($t \geq 50$)보다는 입력 매니폴드의 국소 민감도를 가장 잘 포착할 수 있는 초기 타임스텝($t = 10$) 부근에서 탐지 성능이 가장 높게 측정되었다. 이는 본 논문에서 정의한 적대적 경직성이 이미지의 고주파(High-frequency) 성분이 지배적인 초기 노이즈 주입 단계에서 가장 명확히 드러남을 시사한다.

결론적으로, 본 연구는 적대적 예제가 지니는 매니폴드 상의 특이점인 '적대적 경직성'을 새롭게 발견하고, 이를

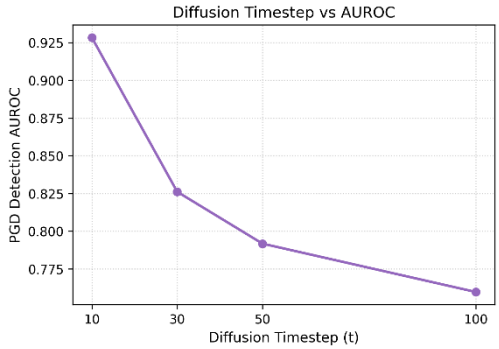

그림 2. 확산 모델의 타임스텝 변화에 따른 탐지성능.

Fig. 2. Performance across varying timesteps.

실용적으로 측정할 수 있는 POP 지표를 제안하여 모델 독립적 환경에서 매우 높은 수준의 탐지가 가능함을 증명하였다. 제안 기법은 이론적으로는 입력 공간에서의 국소적 Jacobian norm 을 근사하며, 구조가 단순하고 연산이 가벼워 다양한 시각 인식 시스템의 전처리 구간에 쉽게 통합될 수 있다. 다만 본 연구에서 진행한 실험은 CIFAR-10 에만 국한되어, 향후 연구로 고해상도 대용량 데이터셋(ImageNet 등)과 고도화된 공격 기법을 통해 본 지표의 일반화 성능을 추가 검증할 필요가 있다.

참 고 문 헌

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," Proc. Int. Conf. Learn. Represent., San Diego, USA, 2015.

[2] D. Meng and H. Chen, "MagNet: a two-pronged defense against adversarial examples," Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Dallas, USA, pp. 135-147, 2017.

[3] P. Lorenz, R. Durall, and J. Keuper, "Adversarial examples are misaligned in diffusion model manifolds," Proc. Int. Joint Conf. Neural Netw., Yokohama, Japan, 2024.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Adv. Neural Inf. Process. Syst., vol. 33, pp. 6840-6851, 2020.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," Proc. Int. Conf. Learn. Represent., Vancouver, Canada, 2018.