PSNR과 SSIM을 활용한 NMS 알고리즘 대상 Adversarial Examples 분석

김광남[1], 이한주[1], 이한진[1], 최석환[2]
[1]연세대학교 전산학과 석사과정
[2]연세대학교 소프트웨어학부 교수
crazynam@yonsei.ac.kr, hanleju@yonsei.ac.kr, han-.-jin@yonsei.ac.kr,
sh.choi@yonsei.ac.kr

Analysis of Adversarial Examples for NMS Algorithms Using PSNR and SSIM

Gwang-Nam Kim[1], Han-Ju Lee[1], Han-Jin Lee[1], Seok-Hwan Choi[2]
[1]Dept. of Computer Science, Yonsei University
[2]Dept. of Software, Yonsei University

요 약

딥러닝 모델이 다양한 분야에 적용되면서, 딥러닝 모델에 대한 보안이 큰 이슈가 되고 있다. 특히, 입력 데이터에 섭동(perturbation)을 추가하여 모델의 정상적인 추론을 방해하는 적대적 공격(Adversarial Attack)에 대한 연구가 활발하게 진행되고 있다. 본 논문에서는 객체 탐지 모델의 NMS(Non-Maximum Suppression) 알고리즘에 대한 적대적 공격 기법 중 하나인 Phantom Sponges 공격을 수행하여 적대적 예제(Adversarial Example)를 생성하고, 원본 이미지와의 유사성을 측정하여 분석하고자 한다.

1. 서론

최근 다양한 분야에서 딥러닝 모델을 활용함에 따라 딥러닝 모델과 관련된 보안도 큰 이슈가 되고 있다. 그중에서 딥러닝 모델의 입력 이미지에 섭동(perturbation)을 추가하여 딥러닝 모델의 성능 감소 및 잘못된 결정을 유도하는 공격인 적대적 공격(Adversarial Attack)과 관련한 다양한 연구가 등장하고 있다. 현재까지의 적대적 공격은 대부분 이미지 분류 모델을 대상으로 연구가 수행되었다[1]. 그러나 최근 객체 탐지 모델이 발전됨에 따라, 객체 탐지 모델에 대한 적대적 공격 연구가 활발하게 진행되고 있다[2]. 객체의 클래스만 고려하는 이미지 분류 모델에 대한 적대적 공격과는 다르게 객체 탐지 모델에 대한 적대적 공격은 객체의 존재 여부 및 개수, 객체의 클래스, 객체의 위치와 같은 다양한 정보를 고려해야 한다. 따라서 객체 탐지 모델에 대한 적대적 공격은 모델의 성능 감소를 위해 다양한 대상에 공격이 수행될 수 있다. 특히, 객체 탐지 모델의 핵심 알고리즘 중 하나인 NMS(Non-Maximum Suppression)에 대한 기존 적대적 공격은 높은 공격 성공률에만 초점을 두고 있으며, 원본 이미지와의 유사성은 고려하지 않는다. 따라서, 본 논문에서는 PSNR 및 SSIM을 이용해 NMS 대상의 공격이 생성한 적대적 예제와 원본 이미지 간의 차이를 계산하고 분석하고자 한다.

2. 본론

본 장에서는 Phantom Sponges 공격을 통한 적대적 예제를 생성하고, 원본 이미지와의 유사성을 실험을 통해 분석한다.

2.1 NMS(Non_Maximum Suppression)

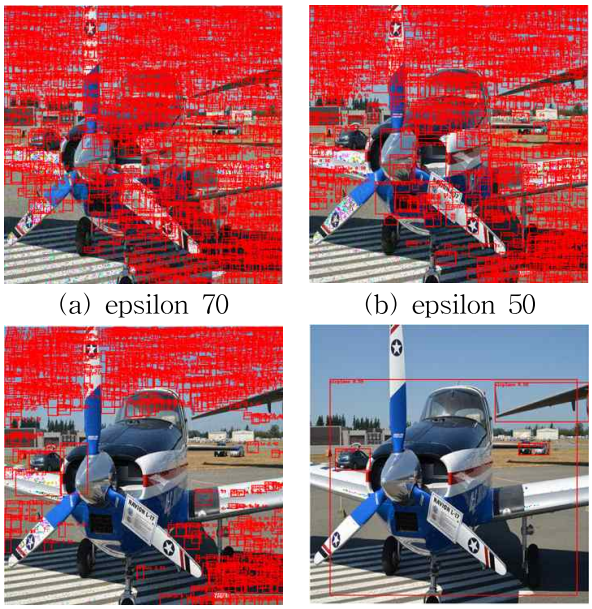
NMS 알고리즘은 객체 탐지 모델에서 활용되며, 객체 탐지 모델이 예측한 bounding box 중 가장 정확한 bounding box를 선택하기 위한 알고리즘이다. NMS 알고리즘은 다음과 같은 절차로 수행된다. 우선 confidence score threshold 값을 선정하여 해당 threshold 값보다 작은 bounding box들을 제거한다. 이후 가장 높은 confidence score를 가지는 bounding box를 기준으로 나머지 bounding box와의 IoU(Intersection over Union)를 계산한다. 마지막으로, IoU 값이 사전에 정의된 IoU threshold보다 큰 bounding box들을 제거한다. 위의 과정을 반복하여 객체 탐지 모델에 입력 이미지의 각 객체는 하나의 bounding box를 가지게 된다.

2.2 Phantom Sponges

Phantom Sponges 공격은 객체 탐지 모델의 NMS 알고리즘을 공격하는 대표적인 공격 기법 중 하나이다[3]. 해당 공격의 목적은 NMS 알고리즘의 정상적인 수행을 방해하여, 많은 bounding box 후보들이 정상적으로 삭제되지 않는 적대적 예제를 생성하는 것이다. Phantom Sponges 공격은 NMS 알고리즘의 IoU 값이 IoU threshold보다 큰 bounding box들을 제거하는 점을 활용하여 공격을 수행한다.

(a) (b)

그림 1. Phantom Sponges 공격 수행 결과



(a) epsilon 70 (b) epsilon 50

(c) epsilon 30 (d) epsilon 10

그림 2. epsilon 변화에 따른 적대적 예제 탐지 결과

구체적으로, 가장 높은 confidence score를 가지는 bounding box와의 IoU값을 낮추어 후보 bounding box 들을 제거하지 못하게 한다. 식(1)은 Phantom Sponges 공격의 손실함수이다.

$$\min_p E_{x\sim D}[\lambda_1 \cdot \ell_{\max\,obects} + \lambda_2 \cdot \ell_{bbox\,area} + \lambda_3 \cdot \ell_{\max\,IoU}] \quad (1)$$

여기서 p는 섭동을 의미하며, D는 데이터셋 분포, $\ell_{\max objects}$, $\ell_{bbox\,area}$, $\ell_{\max\,IoU}$ 는 각각 confidence score, bounding box의 크기, 원본과 적대적 예제 사이의 IoU 값에 대한 손실함수를 의미한다. 손실함수를 최적화함으로써, 객체의 confidence score는 높이며, IoU 값은 줄여 많은 후보 bounding box들이 제거되지 못하고 존재하게 된다.

2.3 PSNR과 SSIM

본 연구에서는 Phantom Sponges로 생성한 적대적 예제들이 원본 이미지와 얼마나 차이나는지 확인하기 위한 지표로 PSNR(Peak Signal-to-Noise ratio)과 SSIM(Structural similarity index measure)

표 1. 원본 이미지와 적대적 예제의 유사성 측정

Epsilon	PSNR	SSIM
70	33.24	0.7583
50	34.06	0.8268
30	36.30	0.9029
10	44.27	0.9730

을 활용하였다[4]. PSNR은 생성 혹은 압축된 영상의 손실 정보를 평가하며, 값이 높을수록 원본 이미지와 비슷한 이미지이다. PSNR은 식 2와 같이 계산할 수 있다.

$$PSNR = 10\log_{10}(\frac{R^2}{MSE}) \quad (2)$$

여기서, R은 픽셀의 최대값을 의미하며, MSE는 원본 이미지와 비교 이미지의 픽셀 평균 제곱 오차를 말한다. SSIM은 두 이미지 간의 상관계수를 휘도, 대비, 구조 측면에서 평가한다. SSIM은 식 3와 같이 계산할 수 있다.

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \quad (3)$$

여기서, l은 휘도, c는 대조, s는 구조에 대한 요소를 의미한다. 원본 이미지와 적대적 이미지와의 SSIM 값이 높을수록 원본 이미지와 유사한 적대적 예제이다.

2.4 실험 및 검증

본 논문에서는 객체 탐지 모델의 NMS를 공격하기 위한 Phantom Sponges 공격을 수행하여 생성된 적대적 예제의 눈에 띄는 정도를 측정하고 분석하기 위해 COCO Dataset 2000장과 YOLOv5 모델을 사용하였다. 그림 1은 Phantom Sponges 공격을 수행한 결과이며, (a)는 원본 이미지, (b)는 공격이 수행된 이미지이다. (b)를 보면 원본 이미지와 육안으로 구분되는 섭동이 추가 되어있는 것을 확인할 수 있다. 따라서 원본 이미지와의 유사성을 측정할 필요가 있다. 그림 2는 Phantom Sponges 공격의 섭동을 제어하기 위한 변수인 epsilon을 70, 50, 30, 10으로 설정하여 적대적 예제를 생성한 후 모델에 입력한 결과이다. Epsilon 값이 높을수록 섭동의 크기가 커서 epsilon이 높은 적대적 예제들은 NMS 알고리즘이 정상적인 역할 수행을 불가능하게 하며, epsilon이 10인 경우에는 공격이 정상적으로 수행되지 않은 것을 확인할 수 있다. 표 1은 원본 이미지와 적대적 예제의 유사성 측정을 위해 epsilon 변화에 따른 PSNR과 SSIM 측정 결과이다. 섭동의 크기가 가장 큰 epsilon 70의 경우 PSNR과 SSIM은 각각 33.24, 0.7583을 보였으며, 섭동의 크기가 작아 원본과 큰 차이가 나지 않는 epsilon 10의 경우 PSNR과 SSIM은 각각 44.27, 0.9730을 보였다. Epsilon 70과 10을 비교하였을 때 PSNR은 33.18% 감소하였으며, SSIM은 28.31% 감소하였다.

3. 결론

본 논문에서는 PSNR와 SSIM 지표를 활용해 객체 탐지 모델의 NMS 알고리즘을 대상으로 하는 적대적 예제를 분석하였다. NMS 알고리즘을 대상으로하는 대표적인 공격 기법인 Phantom Sponges 및 COCO Dataset을 활용해 생성된 적대적 예제들과 원본 이미지들의 유사성을 측정하였으며, 섭동이 강해질수록 SSIM의 값이 급격히 떨어지는 것을 확인하였다. 따라서, 향후 연구에서는 섭동을 효과적으로 생성하여 공격 성능은 유지하며, 원본 이미지와 구별하기 힘든 적대적 예제들을 생성하기 위한 연구를 수행할 것이다.

사사문구

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단(RS-2023-00243075) 및 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업(IITP-2023-RS-2023-00259967)의 지원을 받아 수행된 연구임.

참고문헌

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

[2] Xie, Cihang, et al. "Adversarial examples for semantic segmentation and object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[3] Shapira, Avishag, et al. "Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[4] Hore, Alain, and Djemel Ziou. "Image quality metrics: PSNR vs. SSIM." 2010 20th international conference on pattern recognition. IEEE, 2010.