1 단계 객체 탐지 모델에 대한 전이적인 적대적 공격

김중수, 서영주
포항공과대학교

joongsukim@postech.ac.kr, yjsuh@postech.ac.kr

Transferable Adversarial Attacks for One-Stage Object Detector

Joongsu Kim and Young-Joo Suh
Pohang University of Science and Technology (POSTECH)

요 약

최근 인공지능의 발전으로 인해 딥러닝 기반 객체 탐지 모델이 높은 성능을 달성함으로써 자율주행, 불량검사 등 다양한 산업 분야에서 활용되고 있다. 특히 신속한 추론 속도를 강점으로 하는 1단계 객체 탐지 모델(One-Stage Object Detector)이 주로 활용되고 있다. 한편, 딥러닝 모델이 작은 노이즈에도 취약하다는 것이 밝혀 짐에 따라 객체 탐지 모델을 대상으로 한 적대적 공격(Adversarial Attack) 연구가 활발히 진행되고 있다. 하지만 해당 연구들은 2단계 객체 탐지 모델(Two-Stage Object Detector)에 치중되어 있으며, 또한 모델의 구조를 이미 파악하고 있다는 가정이 선행된다는 한계가 있다. 본 논문은 1단계 객체 탐지 모델에 대한 전이적인 적대적 공격 기법을 제안하며, 접근 불가능한 객체 탐지 모델에 관한 공격을 위해 다른 구조를 갖는 객체 탐지 모델을 활용하여 적대적 예제들을 생성하였다는 점에서 의의가 있다. 다양한 실험을 통해 제안한 적대적 예제들이 1단계 객체 탐지 모델을 성공적으로 공격하는 것을 확인하였다.

Ⅰ. 서 론

최근 객체 탐지 기술은 딥러닝의 발전으로 인하여 많은 발전을 이루었다. 딥러닝 기반 객체 탐지 모델은 크게 2단계 객체 탐지 모델과 1단계 객체 탐지 모델로 분류된다. 2단계 객체 탐지 모델은 객체로 인식될 만한 후보 영역을 추출하는 과정과 해당 후보 영역에 속한 객체가 어떤 객체인지 판별하는 과정이 순차적으로 이루어진다. 반면에 1단계 객체 탐지 모델은 후보 영역 산출과 판별 과정이 동시에 이루어져 상대적으로 추론 속도가 빠르다는 장점을 가지고 있다.

한편, CNN(Convolution Neural Network) 기반의 딥러닝 모델들이 적대적 공격에 취약하다는 것이 밝혀지면서 객체 탐지 모델 또한 다양한 공격 기법[1]이 제안되었다. 하지만 제안된 기법들은 모델의 구조를 완전히 파악 가능하다는 비현실적인 전제하에 진행되었고 2단계 객체 탐지 모델에 대해서만 공격이 가능하다는 한계가 있다.

이에 본 논문에서는 모델의 구조를 파악할 수 없는 전제하에 1단계 객체 탐지 모델에 대한 적대적 공격 기법을 제안한다. 적대적 공격 중에서 사람 눈에 보이지 않은 노이즈를 가해서 모델의 성능을 하락 시키는 회피 공격(evasion attack)을 사용하였으며 이를 위해 2단계 객체 탐지 모델을 사용하였다.

Ⅱ. 본론

Ⅱ.1. 공격 기법 개요

본 논문에서는 그림 1과 같이 1단계 객체 탐지 모델에 대한 공격 기법을 제안한다. 정상 이미지를 2단계 객체 탐지 기반 모델인 Faster R-CNN[2]의 입력 데이터로 넣는다. 입력 데이터에 대한 Faster R-CNN의 비용 그레이디언트(Cost Gradient)를 추출한 후 공격 방법론을 통해 노이즈를 생성한다.

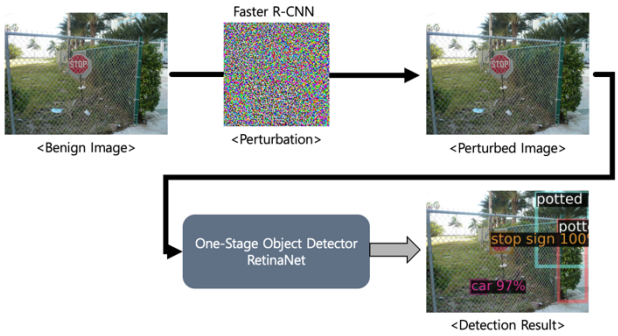

그림 1. 객체 탐지 모델에 대한 공격 기법 개요

비용 그레이디언트는 손실 함수에 따라서 다르게 추출할 수 있다. Faster R-CNN의 손실 함수는 아래 수식과 같다.

$$L_{total} = L_{cls}^{RPN} + L_{reg}^{RPN} + L_{cls}^{Fast\ R-CNN} + L_{reg}^{Fast\ R-CNN}$$

L_{cls}^{RPN}과 L_{reg}^{RPN}는 객체가 있을 후보 영역을 산출하는 과정에서 사용되는 부분 손실 함수이다. 또한 $L_{cls}^{Fast\ R-CNN}$은 앞의 과정에서 제안된 후보 영역 안의 객체를 판별하는 과정에서 사용되는 부분 손실 함수이다. 객체의 위치를 보여주는 경계 상자의 크기는 $L_{reg}^{Fast\ R-CNN}$을 통해서 정확하게 조절할 수 있다. 1단계 객체 탐지 모델은 후보 영역을 찾는 과정과 객체를 판별하는 과정이 동시에 이루어지므로 모든 부분 손실 함수를 사용하는 것은 무분별한 공격으로 이어져 공격의 전이성(Transferability)을 낮추는 역효과를 초래할 수 있다. 따라서 두 종류의 모델 모두 어떤 객체인지 판별하는 부분은 동일하다고 판단하여 $L_{cls}^{Fast\ R-CNN}$만을 사용하여 공격을 가한다. 손실 함수에 따

른 공격 성능은 실험을 통하여 비교하였다. 적대적 예제는 기존 이미지에 노이즈를 더하여 생성한다. 생성된 적대적 예제는 타인이 소유한 1단계 객체 탐지 모델에 입력되고 1단계 객체 탐지 모델은 객체를 탐지하지 못하거나 다른 객체로 판별하는 결과를 도출한다.

Ⅱ.2. 실험 및 성능 평가

Method	Loss Function	mAP (%)
Benign	-	35.3
FGSM	L_{total}	20.88
	$L_{cls}^{Fast\ R-CNN}$	19.97
	$L_{reg}^{Fast\ R-CNN}$	23.37
PGD	L_{total}	26.31
	$L_{cls}^{Fast\ R-CNN}$	25.26
	$L_{reg}^{Fast\ R-CNN}$	28.25

표 1. 적대적 예제에 대한 객체 탐지 성능

본 논문에서 제안된 공격 기법의 성능을 평가하기 위해 MS-COCO 데이터셋 5000장을 사용하였다. 공격할 1단계 객체 탐지 모델은 RetinaNet[5]을 사용하였으며 mAP (mean average precision)로 성능을 측정하였다. 대표적인 회피 공격 방법론인 FGSM[3], PGD[4] 을 통해 노이즈를 생성하였으며, 노이즈의 크기를 결정할 ε값은 16으로 정하여 인간의 눈에 보이지 않게끔 지정하였다. PGD는 반복적인 노이즈를 가하는 방식으로 공격이 진행되므로 반복 횟수는 20회로 지정하였다. 모든 실험의 연산은 NVIDIA GeForce 2080 Ti를 사용해 진행되었다.

각각의 공격 방법론과 손실 함수 값을 통해 만든 적대적 예제에 대한 객체 탐지 성능은 위 표 1와 같다. RetinaNet은 적대적 공격을 하지 않은 샘플들인 Benign 데이터에 대해서는 38.6의 높은 성능을 보여주었으나 적대적 공격을 가한 샘플들에 대해서는 크게는 44%의 성능하락이 있음을 확인할 수 있었다. 또한 L_{total}을 사용하는 것보다는 $L_{cls}^{Fast\ R-CNN}$를 사용하는 것이 mAP를 최대 3%까지 하락 시키는 효과가 있었다. 각각의 공격 방법론에 대해서는 FGSM이 PGD보다 우수한 효과를 가짐을 확인했다. 이는 PGD가 Faster R-CNN에 대해 반복적인 공격을 가함으로써 overfitting 효과가 생겨 단일 단계의 공격을 하는 FGSM보다 상대적으로 전이성이 낮기에 생기는 결과이다. 적대적 공격 결과는 아래 그림 2와 같다.

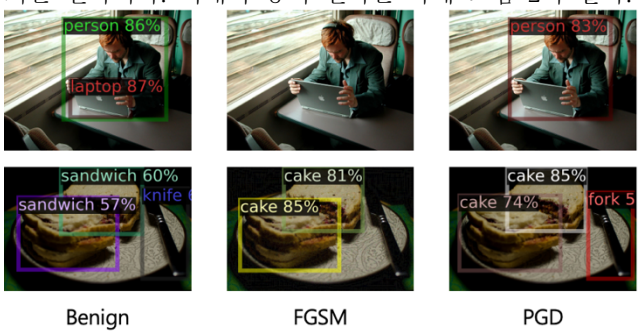

그림 2. 적대적 공격 결과

첫 번째 열은 정상 이미지, 두 번째 열은 FGSM으로 생성한 이미지, 세 번째 열은 PGD로 생성한 이미지로 RetinaNet을 통해 객체를 검출하였다. 첫 번째 이미지의 경우 FGSM은 노트북과 사람을 미검출하는 효과가 있었다. PGD는 사람은 검출할 수 있었지만 노트북은 검출할 수 없었다. 두 번째 이미지에 대해서는 FGSM과 PGD 모두 샌드위치대신 케이크를 검출하게끔 한 것을 확인할 수 있다. 또한 FGSM으로 생성한 이미지는 나이프를 검출할 수 없었다. 이는 제시된 공격 기법이 아무 정보가 없는 1단계 객체 탐지 모델에 대해서 해를 가할 수 있는 것으로 해석할 수 있다.

Ⅲ. 결론

딥러닝의 발전에 따라 객체 탐지 기술도 빠르게 발전되었다. 특히 1단계 객체 탐지 모델은 빠른 추론 속도를 강점으로 다양한 산업 분야에서 사용되고 있다. 이와 더불어 최근 객체 탐지 모델에 대한 적대적 공격이 활발히 연구되고 있다. 하지만 현재까지 제시된 공격 기법은 모델의 구조를 파악 가능하다는 비현실적 전제를 두고 있거나 특정 모델에 대해서만 효과적이라는 한계가 있었다. 본 논문에서는 객체 탐지 모델 중에서는 공격의 시도가 적었던 1단계 객체 탐지 모델에 대한 공격 기법을 제안하였다. 모델을 직접적으로 접근할 수 없는 현실적인 상황을 고려하여 2단계 객체 탐지 모델을 대신 사용하여 공격을 진행하였다. 전이성을 높이기 위해서 2단계 공격 기법의 객체 판별 손실 함수만을 이용하여 노이즈를 생성하였고 다양한 실험을 통해 전체 손실 함수를 사용하는 것보다 효과적인 것을 확인할 수 있었다.

추후 연구에서는 본 연구에서 진행되었던 회피 공격을 발전시켜 전이성을 더 높이는 공격 기법을 제안할 것이다. FGSM 이나 PGD 이외의 다른 방법론을 도입하고 특정 손실 함수를 단순히 선택하는 것이 아니라 새로운 손실 함수를 제안하여 공격할 모델의 성능을 더욱 효과적으로 낮출 계획이다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2019-0-01906,인공지능대학원지원(포항공과대학교))과 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원(No. 2022R1A6A1A03052954, 기초연구사업)을 받아 수행한 연구 과제입니다.

참 고 문 헌

[1] Y. Wang, K. Wang, Z. Zhu, F. Wang. “Adversarial attacks on Faster R-CNN object detector,” Neurocomputing, pp. 87-95, 2020

[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” In Advances in Neural Information Processing Systems, pp. 91-99, 2015

[3] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples” In International Conference on Learning Representations, 2015

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” In International Conference on Learning Representations, 2018

[5] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, “Focal Loss for Dense Object Detection,” International Conference on Computer Vision, 2017