Defending Object Detection Networks Against Adversarial Patch Attacks

Thomas Gittings[1], Steve Schneider[2], John Collomosse[1,3]

[1]CVSSP, University of Surrey – Guildford, UK

[2]SCCS, University of Surrey – Guildford, UK

[3]Adobe Research, Creative Intelligence Lab – San Jose, CA

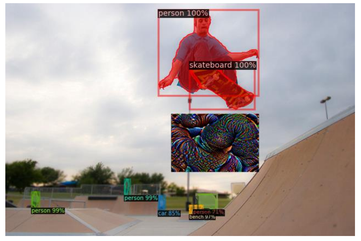
2021 ICCV OCTOBER 11-17 VIRTUAL

Introduction

Object detection is a fundamental computer vision capability underpinning applications including robotics, security, and content analytics. Contemporary object detection methods are enabled via convolutional neural networks (CNNs) and so are prone to adversarial patch attacks (APAs); that introduce visible regions or 'stickers' that induce a significant change in the network prediction. Due to the increasing viability of physical attacks of this kind, there is an emerging threat to autonomous systems relying upon visual sensing.

Undefended Mask R-CNN

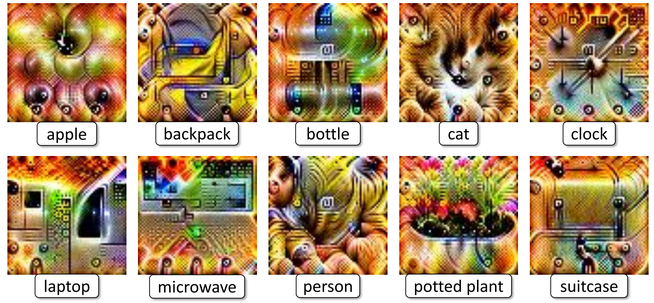
Defended Mask R-CNN



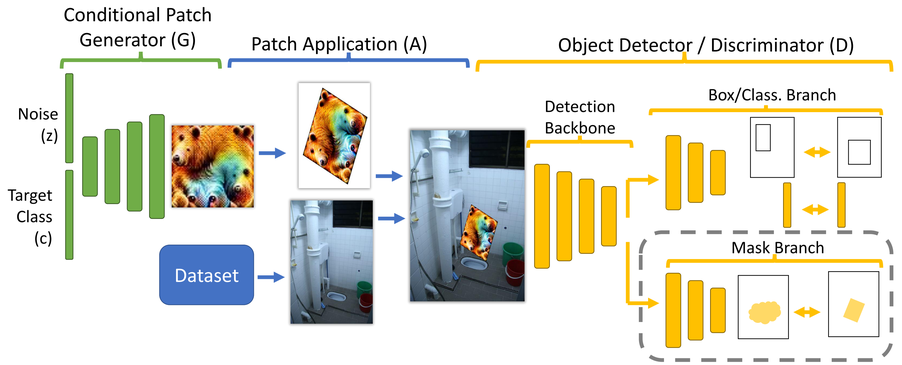
This paper contributes the first training-time defence against APAs for object detection networks. APAs targeting object detection have been sparsely researched and, consequently, few defences exist. Our core technical contribution is to harness adversarial training to improve the resilience of object detection models at training time. Such training need not be applied from scratch, enabling pre-trained models to be fine-tuned via our method in order to confer protection against APAs.

Method

We restrict our attention to attacks which cause the detector to identify an object of some target class in the location of the patch, when no such object is present in the image. Our aim is to defend the networks against these attacks at training time. Our training architecture is inspired by Vax-A-Net; an adversarial training method used to mitigate APAs on image classification networks. The architecture synthesises the patches using a conditional generative network, and applies an adversarial training process to update the generator while simultaneously training the detector to build resistance to the patches.

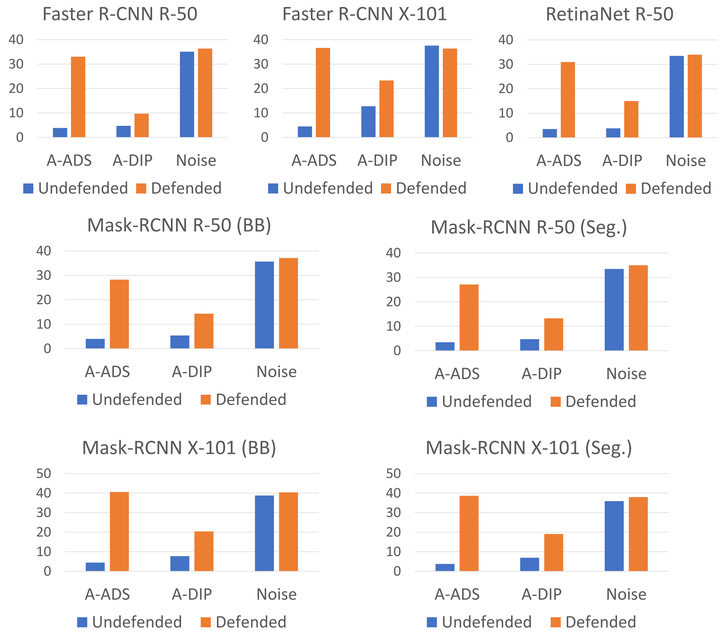


The architecture we are using is similar to that of a generative adversarial network (GAN). The object detection or segmentation network that we are defending takes the place of the discriminator in the GAN architecture. Instead of the discriminator acting as a tool to make the generator better, we are using the generator in order to produce a more effective discriminator. When training the networks we train alternately the discriminator and generator at each iteration, the same as for a normal GAN.



Results

We evaluate using the MSCOCO 2017 dataset. We report the Mean Average Precision (mAP) of the detectors, averaged across the set of 10 test classes. For Mask R-CNN networks we report the mAP for both detection (BB) and segmentation (Seg) tasks. We evaluate the efficacy of our defence by subjecting it to two different adversarial patch attacks (A-ADS and A-DIP). We also include a 'noise' patch for comparison; this is a patch filled with uniform random noise, which provides a baseline for an optimal defence, since it is occluding the image in the same way as the adversarial patches but without any adversarial component. We test our defended network with a form of grey box attack, in which the patches are trained on the network with the publicly available weights before any adversarial training is applied.



In the paper we also include results for white box attacks, in which the patches are trained on the final defended network.