Understanding the Problem Statement

Problem statement: Defence for Adversarial attacks on object detection

Explanation: Adversarial attacks on object detection systems are a growing concern in the field of computer vision. These attacks involve manipulating the input to a object detection model in a way that causes it to make incorrect predictions. This can have serious implications in real-world applications, such as self-driving cars and surveillance systems, where incorrect predictions could lead to accidents or security breaches.

The problem of adversarial attacks on object detection can be broken down into two main categories: targeted and untargeted attacks. Targeted attacks involve manipulating the input to cause the model to detect a specific object, while untargeted attacks involve causing the model to not detect any objects at all. Both types of attacks can be executed through various methods, such as adding noise to the input image or modifying the object's appearance.

One of the main challenges in defending against adversarial attacks on object detection systems is the high dimensionality of the input space. The input to an object detection model is typically an image, which can have millions of pixels. This makes it difficult to search for adversarial examples and to design effective defenses.

Another challenge is the diversity of object detection models. Object detection models can be based on different architectures, such as CNNs or R-CNNs, and can have different levels of complexity. This means that defenses that work well for one model may not work as well for another.

There are several methods that have been proposed to defend against adversarial attacks on object detection systems. One approach is to use adversarial training, which involves training the

model on a dataset of adversarial examples. This can make the model more robust to attacks, but it can also make it less accurate on clean examples. Another approach is to use input preprocessing techniques, such as image denoising, to remove the adversarial perturbations. One other approach is to use ensemble models, which combine the predictions of multiple models. This can make the overall system more robust to attacks, as an adversarial example would have to fool all of the models in the ensemble in order to cause a misclassification.
In conclusion, adversarial attacks on object detection systems are a growing concern in the field of computer vision. These attacks can cause object detection models to make incorrect predictions, which can have serious implications in real-world applications. Defending against these attacks is challenging due to the high dimensionality of the input space and the diversity of object detection models. However, several methods have been proposed to defend against adversarial attacks, such as adversarial training, input preprocessing, and ensemble models.

Algorithms used: YOLO, SDD, R-CNN and DeblurGAN

Datasets used: PASCAL VOC and MS COCO

Research Paper:

1. Understanding Object Detection Through An Adversarial Lens by Ka-Ho Chow, Ling Liu, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu
2. Adversarial Detection: Attacking Object Detection in Real Time by Han Wu, Syed Yunas, Sareh Rowlands, Wenjie Ruan, and Johan Wahlstrom