Transformer-Encoder Detector Module: Using Context to Improve Robustness to Adversarial Attacks on Object Detection

Faisal Alamri, Sinan Kalkan, Nicolas Pugeault
1. Department of Computer Science, The University of Exeter. Email: fa269@exeter.ac.uk
2. Department of Computer Engineering, Middle East Technical University. Email: skalkan@ceng.metu.edu.tr
3. School of Computing Science, University of Glasgow. Email: nicolas.pugeault@glasgow.ac.uk

Introduction
This paper describes a contextual detection module, proposed as an add-on to classical object detection architectures such as Faster-RCNN, named Transformer-Encoder Detector Module (TEDM). This module implicitly encodes contextual statistics of objects and uses attention mechanism to improve the labelling of image regions. It improves the detection performance of a state-of-the-art object detector, as evaluated on natural images and perturbed images. This module does rescore, relabel, and correct detector predictions. It can also be applied to any object detector to (i) improve the labelling of object instances; and (ii) improve the detector’s robustness to adversarial attacks.

Proposed Method
The proposed module is built upon the success of the Transformer model proposed by [4]. However, the proposed method is the first to include the Transformer-Encoder with a dimension of 2,048. The Transformer-Encoder then takes the features and processes them with the attention mechanism to output features with a dimension of 512. A feed-forward NN classifier is then applied, which produces the new scores and labels for each region.

Experiments
Experiment One: Natural Images
We are presenting how effective TEDM is in comparison with Faster RCNN when applied on natural images, as shown in Figure 2.

In the first image, where FFF perturbation is added to regions TEDM predicts no objects. All objects detected by Faster RCNN after attacked are false detection. TEDM helps to prevent false detections that Faster RCNN outputs. This is a good example of how negatively the perturbation impacts the model. It is found that when the regions are large in size, they are likely to be impacted more by perturbation resulting in not being detected. In terms of UAP perturbation, ten regions are detected by Faster RCNN before the attack, and only half of them are detected after. Faster RCNN detects four objects correctly but fails in detecting the cup. The TEDM fails to detect the person that Faster RCNN already detects, which can be due to the perturbation and lighting conditions. We can see that the person in the large region is not detected, which can be observed that the larger the regions are the more likely they are to be impacted.

In the second row, an image with more than one object is used. Such an image includes the person due to the perturbation and lighting conditions. We can see that the person is detected after attacked. Faster RCNN detects four objects correctly but fails in detecting the cup. TEDM helps to prevent false detections that Faster RCNN outputs. This is a good example of how negatively the perturbation impacts the model. It is found that when the regions are large in size, they are likely to be impacted more by perturbation resulting in not being detected. In terms of UAP perturbation, ten regions are detected by Faster RCNN before the attack, and only half of them are detected after. Faster RCNN detects four objects correctly but fails in detecting the cup. The TEDM fails to detect the person that Faster RCNN already detects, which can be due to the perturbation and lighting conditions. We can see that the person in the large region is not detected, which can be observed that the larger the regions are the more likely they are to be impacted.

Experiment Two: Adversarial Images
Adversarial images are used to examine the impact of TEDM (i.e., the use of contextual information) against adversarial attacks. Such images may have different visual features due to the addition of adversarial perturbations leading to an effect on contextual information especially when some objects are misdetected. Two different approaches of adversarial perturbations are presented in Figure 3, which are Universal Adversarial Perturbations (UAP) and Fast Feature Fool (FFF) [3], are applied.

In the first image, where FFF perturbation is added to regions TEDM predicts no objects. All objects detected by Faster RCNN after attacked are false detection. TEDM helps to prevent false detections that Faster RCNN outputs. This is a good example of how negatively the perturbation impacts the model. It is found that when the regions are large in size, they are likely to be impacted more by perturbation resulting in not being detected. In terms of UAP perturbation, ten regions are detected by Faster RCNN before the attack, and only half of them are detected after. Faster RCNN detects four objects correctly but fails in detecting the cup. The TEDM fails to detect the person that Faster RCNN already detects, which can be due to the perturbation and lighting conditions. We can see that the person in the large region is not detected, which can be observed that the larger the regions are the more likely they are to be impacted.

Conclusion
TEDM, which when combined with an object detection architecture improves both performance and robustness to adversarial attacks. As experimented, the impact of adversarial attacks was reported to be higher when applied on regions, which we believe is due to the size of the regions. Surprisingly, UAP perturbation affects the performance of the examined models when added to the entire image less than FFF does. Future work will involve developing an end-to-end model, to refine not only predictions but also boundary boxes from both contextual and visual features.

References