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

**Faisal Alamri[1], Sinan Kalkan[2], Nicolas Pugeault[3]**

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

ICPR 2020 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION Milan, Italy 10 | 15 January 2021

## 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, only the Transformer-Encoder is adapted, as shown in Figure 1. First, an image is passed to the baseline detector (i.e., Faster-RCNN) for feature extraction, where the features representing image regions for the detected objects are extracted. The dimension of the features extracted is 4,096, they are mean-pooled to 2,048 to suit further processing. The pooled features and the boundary boxes obtained from the detector are passed into 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.

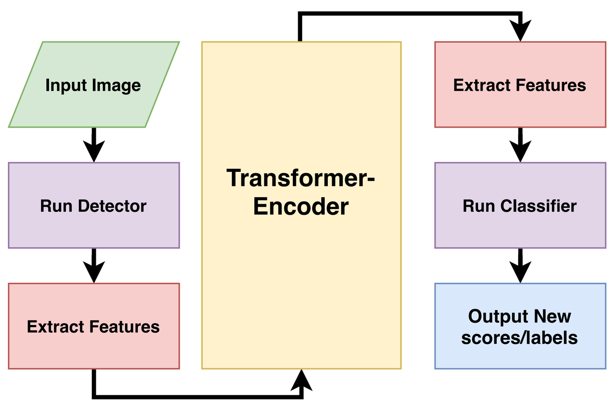


**Figure 1:** Transformer-Encoder Detector Module Architecture.

This method, to the best of our knowledge, is the first to include the Transformer-Encoder to benefit from the use of attention and the positional encoding operation to apply for object detection performance in both natural and perturbed images.

## 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.

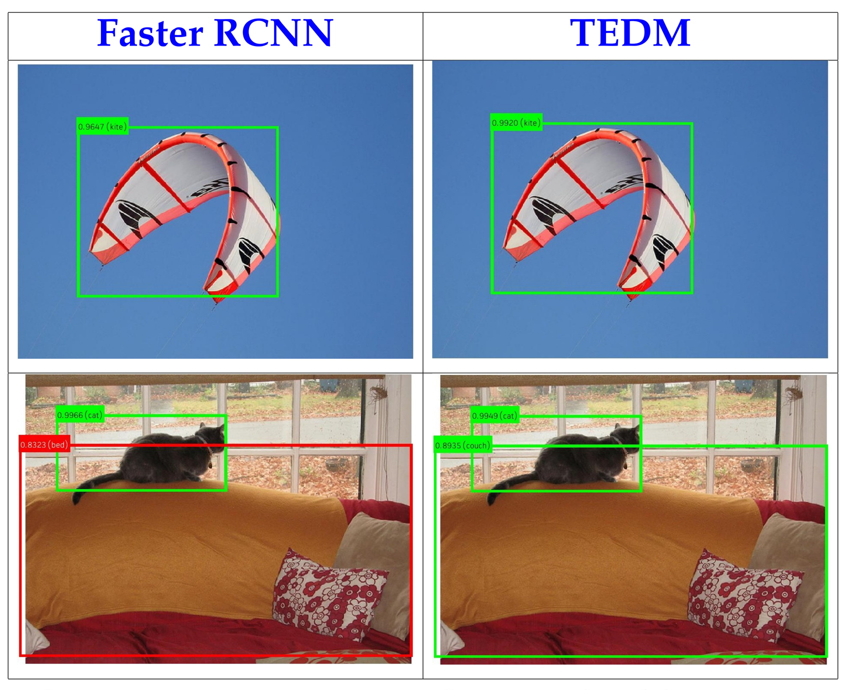


**Figure 2:** Results: Faster RCNN vs. TEDM outputs: Green and red boxes represent correct detection and incorrect detection, respectively.

Note that images used in these experiments are taken from MS COCO 2017 *val* dataset. It can be seen in the first row that a `kite` is detected by Faster RCNN as the only object and scored as 0.9647. However, applying *TEDM* increases the confidence to 0.9920. This can lead us to the result that *TEDM* is a compatible method producing comparable results as Faster RCNN.

In the second row, an image with more than one object is used. Such an image includes the contextual information that cannot be found when one object is presented. In this image, we see that *TEDM* outperforms Faster RCNN predictions. It corrects the `bed`, which is incorrectly detected, to a `couch` as correct detection. This clearly shows how *TEDM* does not only contribute to removing false detections but also how it rescores and relabels them.

**Table 1:** mAP and F1 scores in percentages [%] for Faster RCNN and TEDM.

| Model | $mAP_{0.5}$ | F1 |
|---|---|---|
| Faster RCNN | 62.82 | 57.34 |
| Relabelling Model [1] | 65.50 | 58.95 |
| Transformer-Encoder Detector Module | **69.07** | **63.27** |

***Summary***: We can clearly see that *TEDM* performs better scoring higher mAP and F1 scores compared with Faster RCNN and the *Relabelling Model*. From the reported results, we can say that *TEDM* is an excellent tool to overcome some of the errors that Faster RCNN attempts.

### 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, as presented in Figure 3, which are Universal Adversarial Perturbations (UAP) [2] and Fast Feature Fool (FFF) [3], are applied.

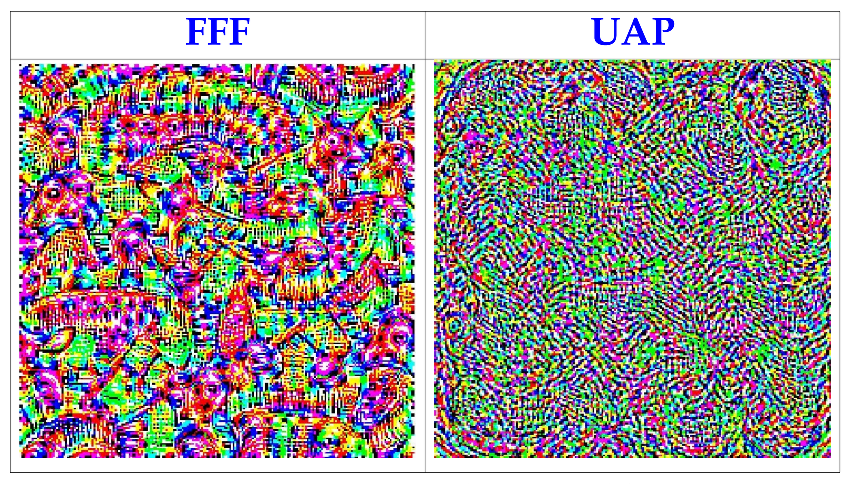


**Figure 3:** FFF vs. UAP perturbations added to images

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 be observed that the larger the regions are the more likely they are to be impacted.

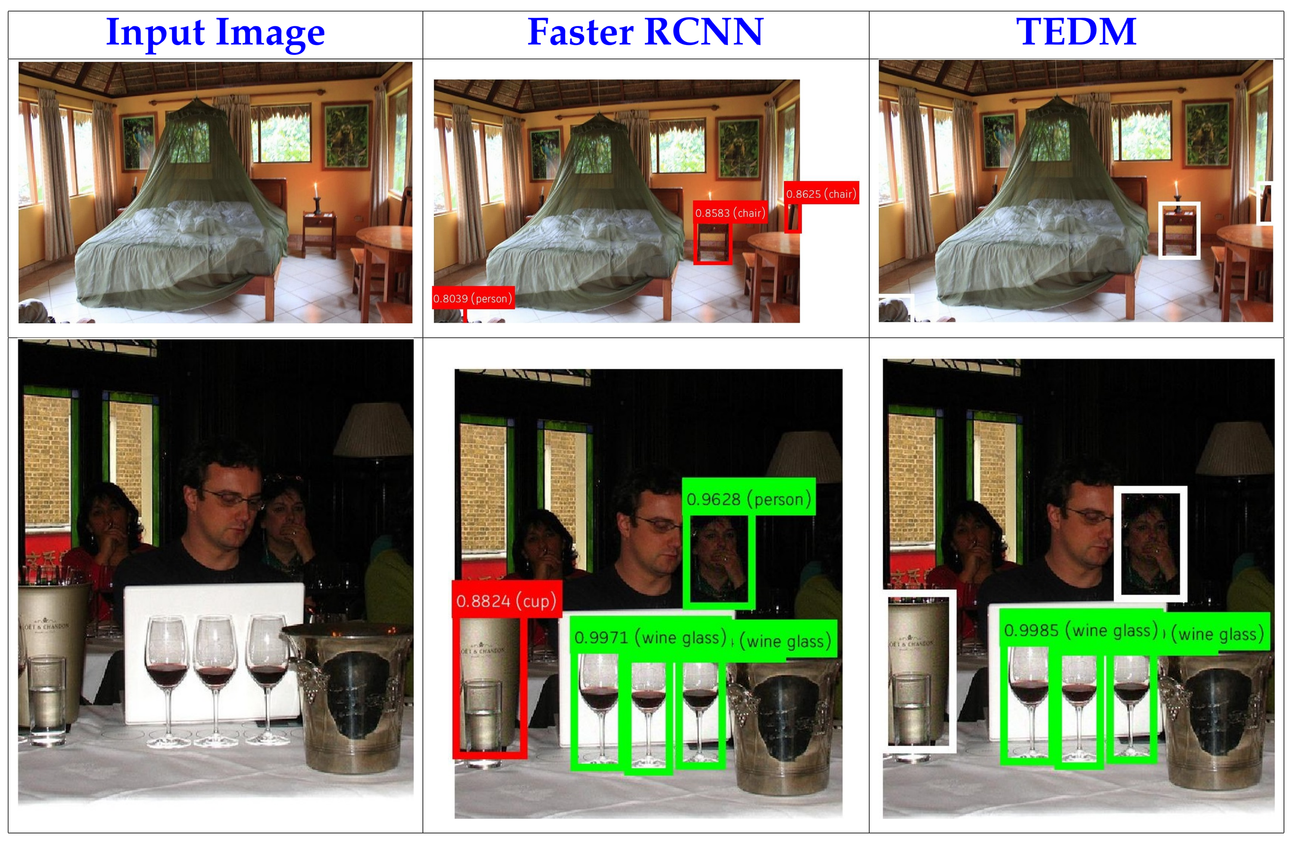


**Figure 4:** Results: Faster RCNN and *TEDM* outputs for FFF and UAP perturbed images: Green, red and white boxes represent correct detection, incorrect detection, and objects removed and relabelled as background, respectively

**Table 2:** mAP and F1 scores in percentages [%] for Faster RCNN and TEDM.

| Model | $mAP_{0.5}$ | F1 |
|---|---|---|
| Faster RCNN | 26.25 | 20.78 |
| Relabelling Model [1] | 29.05 | 23.14 |
| Transformer-Encoder Detector Module | **35.28** | **26.25** |

**Summary**: *TEDM* is performing better than Faster RCNN to tackle adversarial perturbations. This is because of the features encoding during the encoder process, as it learns the spatial and visual features.

## 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

[1] F. Alamri and N. Pugeault. Improving object detection performance using scene contextual constraints. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1, 2020.

[2] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 86–94, 2017.

[3] Konda Reddy Mopuri, Utsav Garg, and Venkatesh Babu Radhakrishnan. Fast feature fool: A data independent approach to universal adversarial perturbations. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

**To access the full paper: Please scan this code**