Analyzing the Performance of DETR for Object Detection with Occlusions

Jaira Mathena B. Angeles
CoE 197 Deep Learning
College of Engineering
University of the Philippines-Diliman

ABSTRACT

Carion et al. [1] released the DEtection TRansformer (DETR) in 2020. One major difficulty in object detection is occlusion handling. This project sheds light on the underlying attention mechanism of the network when faced with this challenge. The pre-trained DETRs loaded from Facebook's model zoo leverages the global image context in order to output predictions that are nearly as accurate as the ones for the COCO dataset they were trained on. This study demonstrates how adaptive this new approach is when presented with noisier images.

1. INTRODUCTION

Transformers and sequence-to-sequence learning has largely been applied in natural language processing tasks. There have, however, been more recent advancements in computer vision. In particular, there have been some research being done in the image recognition and object detection domains. One of the most notable innovations in the latter is the DEtection TRansformer (DETR) released by Facebook in early 2020 [1]. Given the novelty of this architecture, there are limited resources that can provide insight on its underlying attention mechanism. There are still many aspects of this approach that must be evaluated in order to guide future studies. One such feature is how it handles occlusions. As such, the main contribution of this project is to visualize the weights and self-attention mechanism of baseline DETR models, specifically in inferring partially occluded objects. The subsequent experiments validate the recognition ability of DETR in what is considered to be a more context-aware manner.

2. RELATED WORK

Carion et al. [1] demonstrated how DETR is able to achieve an accuracy on par with SOTA models like Faster R-CNN. It does so through a transformer encoder-decoder architecture and via bipartite matching. The former is inspired by the concepts introduced by Vaswani et al. [2] in their paper Attention Is All You Need. Among the benefits of using transformers is that the performance of these models does not seem to saturate when increasing the number of parameters [3]. An extension of the DETR, the Deformable DETR, that is better suited for detecting small objects was later released by Zhu et al. [4]. As for previous works related to occlusion, a vast majority of them are focused on building out existing CNN-based models [5] [6].

3. EXPERIMENTS

All baseline models provided by Facebook were included. Namely, these are the DETR and DETR-DC5 with a ResNet50 backbone, along with the DETR and DETR-DC5 with a ResNet100 backbone. These pretrained models were loaded from TorchHub. The datasets were split into two. One contained examples of occluded objects, while the other consisted of samples taken from COCO 2017. Regardless, all images within scope maintained that $x_{img} \in \mathbb{R}^{3 \times H_0 \times W_0}$. They also all represented at least one of the classes the models were trained to infer.

Qualitatively, both the multi-head attention weights from the last layer of the decoder, as well as the the self-attention weight from the decoder were visualized. The former is deliberately done at that stage in order to most precisely identify what part of the image was responsible for the networks' predictions. In other words, the mean of the attention weights over all the heads of the transformer was obtained. The latter, on the other hand, shows the state of the self-attention module at each reference point throughout the input. These all contribute to the eventual box coordinates and class labels decoded by the Feed Forward Network (FNN). Figures 1 and 2 illustrate the results of these components. One sample was taken from each of the aforementioned datasets for comparison's sake.

Figure 1: The occluded image (left) and the COCO image (right) received prediction scores of 0.99 and 1.0 respectively. The weights from the decoder are localized.

The Average Precision (AP) is used as the metric to quan-

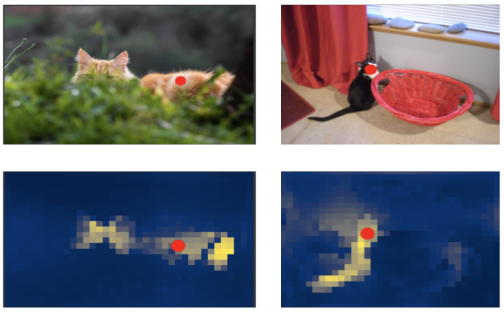

Figure 2: The encoder is able to separate the objects from their surrounding environments, which belong to the "no object" class.

titatively evaluate the performance of the DETR architecture. In calculating these values, the conventional definition of precision P and recall R is followed. The variables t_p, f_p, and f_n denote true positions, false positives, and false negatives respectively.

$$P = \frac{t_p}{t_p + f_p} \tag{1}$$

$$R = \frac{t_p}{t_p + f_n} \tag{2}$$

Precision-recall value pairs are derived by changing the threshold for what probabilities to consider. The table below summarizes the AP of the four models tested. AP_O is the measure for the occluded images, while AP_C is for the non-occluded COCO images.

Model	Backbone	AP_O	AP_C
DETR	ResNet50	38.5	41.6
DETR-DC5	ResNet50	30.4	43.5
DETR	ResNet101	39.7	44.2
DETR-DC5	ResNet101	40.3	46.2

4. ANALYSIS

According to the table above, the AP of the DETR models are only a few marks below their reported AP on the COCO dataset, which served as the control set. These minor deviations indicate that this architecture is adequately equipped to handle occlusions. It is able to do so by concentrating on specific features that it has learned to attribute to given classes. This is evident in Figure 2 where the demarcated regions are the ears and tails of the cats in both cases. These features are indeed fairly distinct in terms of form. Investigating different reference points in given images revealed that the encoder was performing instance segmentation across the input. Figure 1 shows that although a big part of the object is partially obscured in the first image, the model is still able to discern enough to compute a reasonable prediction, albeit with relatively less certainty. The fixed positional embedding, referred to as object queries, inform succeeding attention layers. The fact that the encoder preemptively undertakes this separation process makes it easier for the decoder to extract and localize areas of interest.

The slightly diminished performance can be attributed to circumstances in which the visible parts of the object is either too small or seemingly blends into the background. This impairs the transformer's ability to distinguish boundaries. It therefore reinforces the extent of the model's generalization ability as one that is heavily reliant on clear relationships between objects rendered.

5. CONCLUSION

It appears that the DETR's ability to globally reason about all object through pair-wise relations while still considering the whole image as context is especially advantageous in this problem domain. The intuition gained on how this network encodes occluded objects can significantly improve the empirical performance of later iterations. Although the observed accuracy was satisfactory in this case, larger studies must be done in order to ascertain its robustness in noisier environments. This research opens the door for future extensions that are even better at discovering the unique features of objects in the world around us.

6. REFERENCES

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," May 2020.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Dec. 2017.

[3] T. B. Brown et al. "Language Models are Few-Shot Learners," Jul. 2020.

[4] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for Emd-to-End Object Detection," Nov. 2020.

[5] A. Wang, Y. Sun, A. Kortylewski, and A. Yuille, "Robust Object Detection Under Occlusion With Context-Aware CompositionalNets," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd," Jul. 2018.