Towards Improving Object Detection with End-to-End Transformers

Rituraj Singh
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
riturajs@andrew.cmu.edu

Tushar Vatsa
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
tvatsa@andrew.cmu.edu

Vivek Aswal
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
vaswal@andrew.cmu.edu

Wallace Dalmet
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
wdalmet@andrew.cmu.edu

Abstract

Convolutional neural networks have been proven to be extremely effective for object detection. Lately, the onset of transformers have opened new dimensions in improving the performance of the existing object detection networks. Although the application of transformers has been in natural language processing predominantly, recent research works have suggested that they are capable of achieving comparable, if not better, performance on computer vision problems. In this project, we will develop pure transformer networks for object detection and measure their effectiveness on popular large scale datasets.

1 Introduction

Transformers have been fairly effective in the field of natural language processing. The BERT and GPT breakthroughs in NLP are both based on the fundamentals of transformers. Recently, there have been attempts to take advantage of the self attention mechanism in transformers for computer vision problems, one of them being object detection. Transformers have the potential to achieve better or comparable performance when pitted against convolutional neural networks or recurrent neural networks. Transformers have an upper hand when compared to recurrent networks because unlike RNNs, transformers aren't bound by the constraint of being computed sequentially. Instead, a transformer's self attention layer and fully connected layers can be computed in parallel.

In this project, we will focus on the effectiveness of pure transformers for object detection, a topic that is relatively untapped. Naturally, there are many challenges in using a transformer for a computer vision task. For example, object detection using transformers requires much longer training time compared to traditional CNN networks. However, transformers perform surprisingly well in case of large datasets, in the order of 15 million to 300 million data points. Detecting smaller objects in images is another challenge faced by a transformer based network.

In this work, we will attempt to create a pure transformer network for object detection and compare the results with popular architectures.

2 Literature Review

The authors in [1] present an overview of the progress of transformers in computer vision. The survey goes through the formulation of a transformer, the encoder and decoder modules and the self attention mechanism. It also states the application of multi-head attention and residual connection in encoders and decoders which strengthen the flow of information in the network. The vision transformer (ViT) [2] by Dosovitskiy et al. achieved state of the art results on multiple image recognition tasks. An important conclusion of the paper was that transformers perform exceptionally well on large datasets.

In addition, the survey covers the potential bottlenecks in building transformer networks for computer vision tasks and their proposed solutions, for example, the deformable detection transformer (DETR) [3]. The deformable DETR requires 10x less training time and is 1.6x faster during inference.

The most common object detectors relied on various region proposal algorithms such as Selective search [4] and Edgeboxes [5], however, these algorithms contribute significantly to the increased computational cost. By introducing a Region Proposal Network (RPN) [6], a fully convolutional network which utilizes sharing of features by an attention mechanism with Fast R-CNN [4] as the detection network, that gives cost-free region proposals. The combined object detector network is known as Faster R-CNN [7] that was alternatively trained end-to-end and achieved state-of-the-art results on PASCAL VOC 2007, 2012, and MS COCO datasets.

In [8], the authors have implemented an end-to-end multi-object tracking and segmentation model (TrackFormer) based on encoder-decoder Transformer architecture by introducing the concept of track query embeddings that follow objects through a video sequence in an autoregressive manner. The model achieves a data association between frames by self- and encoder-decoder attention mechanisms that reason about location, occlusion, and object identity. TrackFormer yields state-of the-art performance on the tasks of multi-object tracking (MOT17) and segmentation (MOTS20).

3 Datasets

The goal of this project is to create an end to end object detection network using a transformer. We will be implementing DETR and ViT transformers for improving object detection and assessing their accuracy on datasets like COCO, ImageNet and VOC. The Common Objects in COntext (COCO) dataset contains 91 common object categories with 82 of them having more than 5,000 labeled instances. In total the dataset has 2,500,000 labeled instances in 328,000 images. The PASCAL VOC dataset contains 20 object categories spread over 11,000 images. Over 27,000 object instance bounding boxes are labeled, of which 7,000 have detailed segmentations.

References

[1] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D. (2020). A Survey on Visual Transformer. ArXiv, abs/2012.12556.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[3] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.

[4] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” International Journal of Computer Vision (IJCV), 2013.

[5] C. L. Zitnick and P. Dollar, “Edge boxes: Locating object proposals from edges,” in European Conference on Computer Vision (ECCV), 2014.

[6] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, June 2017.

[7] R. Girshick, “Fast R-CNN,” in IEEE International Conference on Computer Vision (ICCV), 2015.

[8] Meinhardt, T., Kirillov, A., Leal-Taixé, L., Feichtenhofer, C. (2021). TrackFormer: Multi-Object Tracking with Transformers. ArXiv, abs/2101.02702.