Literature review on Transformers in object detection

Elkhan Ismayilzada[1], Minjae Lee[1], and Seungryul Baek[1]

[1]UNIST, South Korea

Abstract

- Object detection is an essential and challenging computer vision task that lately attracted a lot of attention. Convolutional neural networks (CNN) was the number one choice to use for this task due to its advantage of automatic feature extraction mechanism until the introduction of the vision transformers. Currently, transformer-based object detection models are considered as state-of-the-art due to their significant advantages over CNN-based ones and in this paper we will review recent works in this stream of works.

Introduction

- Object detection is a task in which we aim to find the objects and their bounding boxes in the images. Traditionally, convolutional neural networks (ConvNets) were extensively used due to its advantage of automatic feature extraction mechanism. However, using ConvNets alone is not sufficient for this task since it is not a basic set prediction task. Main difficulty in object detection is to get rid of near-identical bounding boxes and this requires to have handcrafted components in the model such as non-maximum suppression, anchor generation etc. Transformers can eliminate these post-processing steps as suggested first by [1] while performing better than the traditional methods. In this paper, we will explore several existing transformer-based object detection models.

Summary of the Methods

Method	Highlights	Limitations
DETR [1]	**a)** First model that uses transformers for object detection, **b)** Significantly outperforms CNN-based models in detecting large objects.	**a)** Slow convergence, **b)** Quadratical increase in computational cost as the size of feature map increases, **c)** Struggle in detecting small objects.
D-DETR [12]	**a)** Better results in detecting small objects, **b)** Faster convergence than DETR [1].	**a)** Has two-stage object detector design.
A-DETR [10]	**a)** Uses anchor points as prior for learned object queries, **b)** Better performance than D-DETR [12].	**a)** No consideration to the objects of different scales.
Pix2Seq [3]	**a)** Language modeling approach for object detection.	**a)** Slow inference due to autoregressive modeling.
YOLOS [6]	**a)** Fully transformer-based pipeline.	**a)** Poor performance compared to DETR models.
DFFT [2]	**a)** Decoder-free fully-transformer based model, **b)** 28% reduction in computation and significantly faster convergence than DETR. [1]	**a)** Poor performance in detecting small objects compared to DETR. [1]
DESTR [7]	**a)** Independent training for each subtask in Object detection, **b)** Outperforms DETR models in all measures.	**a)** Less-efficient due to the addition of another cross-attention layer.
MTTrans [11]	**a)** Support for unlabeled data in object detection using mean teacher framework, **b)** SOTA results in every scenario.	**a)** Does not work well with blurred objects.

Table 1. Summary of the highlights and limitations of the methods discussed in this paper.

Conclusion

- In this paper, we have reviewed several recent transformer-based approaches for object detection. First, we explained the basics of transformers and then introduced the earliest approach i.e., DETR [1] that detects the objects using transformers. Then we moved to the recent approaches that tackle the issue of slow convergence and high computational cost in DETR [1] such as D-DETR [2] A-DETR [3], DFFT [4], DESTR [5]. We also introduced a recent approach that deals with unlabeled data for object detection using transformers [6] and the method that treats object detection as a language modeling task [7].

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. ECCV, 2020. [2] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. ArXiv:XXXX.XXXXX, 2020. [3] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. AAAI, 2022. [4] Peixian Chen, Mengdan Zhang, Yunhang Shen, Kekai Sheng, Yuting Gao, Xing Sun, Ke Li, and Chunhua Shen. Efficient decoder-free object detection with transformers. ECCV, 2022. [5] Liqiang He and Sinisa Todorovic. Destr: Object detection with split transformer. CVPR, 2022. [6] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mttrans: Crossdomain object detection with mean teacher transformer. ECCV, 2022. [7] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. ArXiv:XXXX.XXXXX, 2021.