

Object Detection Based on Transfer Learning Techniques and Transformer

Ke Xu^a

College of Computer, Mathematical and Natural Science, University of Maryland, College Park, U.S.A.

Keywords: Object Detection, Transformer, Transfer Learning, Set Prediction Loss.


Abstract: At present, researchers treat object detection as a hot topic especially using transformer. The goal is to categorize individual objects and determine their location by means of a bounding box. Object detection is the basis for many applications such as surveillance, image retrieval, automatic driving and face recognition. This paper proposes a newly designed transformer model to cope with the object detection task based on migration learning techniques. Specifically, the model consists of several parts: a Resnet-101 model as a backbone, an encoder with an attention mechanism, a decoder with an object query input, and a network that can fetch the output banding box. In addition, an ensemble prediction loss function for bipartite frame matching is developed. With respect to the experiment results, the paper shows it is useful to implement transfer learning technique between COCO 2007 and PASCAL VOC 2012 dataset. The Fsater Region Convolutional Neural Network (R-CNN) model was used for the model comparison. This research demonstrates the broad promise of using migration learning for object detection, enabling many downstream tasks in this area. It also lays a solid foundation for future semantic segmentation tasks using other improved converter models, which will have a real impact on applications in computer vision.

1 INTRODUCTION

A key task named object detection which contains image identification and object locating plays a vital roll in computer vision. This process encompasses not only recognizing various objects but also determining their boundaries or locations represented through bounding boxes. The traditional approaches to object detection have heavily relied on convolutional neural networks (CNNs) due to their strong performance in extracting hierarchical features from images. However, the advent of transformer models, originally developed for natural language processing tasks, has introduced a paradigm shift in object detection techniques (Vaswani, 2017). The adaptation of transformer models to computer vision, and specifically to object detection, leverages this ability to process information globally across an image. In contrast to CNNs, which extract features through local receptive fields and build up representations through successive convolutional layers, transformers can theoretically attend to any part of the image directly, regardless of spatial distance. This global view of the

image allows for more effective integration of contextual information, which is crucial for accurately identifying and localizing objects.

Transformer models utilize a technique called self-attention function which is a special layer designed to gather variety of features across the entirety of an input sequence (Bahdanau, 2014). With the advent of transformers, self-attention layers were introduced. Every aspect of features are examined by these layers which ordered by a sequence and after continuously gathering the whole body of features within a sequence, they then do self-updating in a manner akin to Non-Local Neural Networks (Wang, 2018). One of the first and most influential works to adapt transformers for object detection is the Vision Transformer by Dosovitskiy (Dosovitskiy, 2020). He showed that image patches ordered like a sequences could be plugged in to a plain transformer which allows it to classify images successfully. Building on this, Carion introduced the Detection Transformer (DETR) model for object detection. DETR simplifies the object detection pipeline by eliminating residue parts, which are common in traditional object detection systems (Carion, 2020). A primary benefit

^a <https://orcid.org/0009-0003-6685-7197>

of models that utilize attention mechanisms is their capability for global computation and flawless memory retention, rendering more adept at handling lengthy sequences compared to Recurrent Neural Networks (Graves, 2020). Consequently, different types of RNNs are increasingly supplanted within transformers.

Contemporary approaches to object detection typically base their predictions on initial estimations. Two-stage detectors generate box predictions in relation to proposals, while single-stage methods predict in reference to anchors or a predefined grid that represents potential centers of objects (Cai, 2019). Recent studies have shown that the ultimate effectiveness of these systems is significantly influenced by the precise manner in which these initial estimates are configured (Zhang, 2020).

This research is meant to introduce Transformer-based object detection through transfer learning. Specifically, the authors added a transformer module to the encoder and decoder by utilizing the DETR-based attention mechanism and the structure of the transformer model. Meanwhile, this paper proposes an enhanced feedforward network module. The authors first trained the model on COCO 2017 dataset and then froze the parameters in the encoder and decoder and analyzed it on PASCAL VOCAL 2012 dataset to implement migration learning (Lin, 2014) (Everingham, 2020). The transformer model achieves relatively high average precision (AP) on the PASCAL VOCAL 2012 dataset. The prediction performance of different models is also analyzed and compared. The author compares model performance with Fast-RCNN, and it is found that the transformer achieves an average accuracy of 3.0 higher than that of Fast-RCNN, which indicates that the transformer performs well through the process of migration learning. The experimental results show that using Resnet-101 as the backbone and predicting losses using the object detection set, the transformer model achieves 42 APs on the COCO 2007 dataset, which is 3.0 higher than Fast-RCNN. This research may lay the groundwork for the transformer to perform other downstream tasks based on migration learning or pre-trained models.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

This paper is conducted using the COCO 2017 datasets for detection which comprise 118,000

images for training and 5,000 for validation (Lin, 2014). The COCO 2017 dataset is a comprehensive collection designed for tasks like object detection, segmentation, and captioning, which are central to advancing computer vision technologies. Originating from the COCO project, this dataset is pivotal for evaluating algorithm performance in detecting and segmenting objects across a wide array of categories in varied and complex images. Each image in these datasets is equipped with annotations for bounding boxes. According to the model that is trained before, author analyzes the transfer learning performance on Pascal VOC 2012 which contains a total of 11,530 image and are divided into a training/validation set of around 5,717 images and a test set with the rest (Everingham, 2020). It is an annual competition designed to foster advancements in the recognition of visual object classes in digital images. The sample is shown in Figure 1.



Figure 1: The sample of the dataset. The left column is COCO 2017 and the right column is PASCAL VOC 2012 (Photo/Picture credit: Original).

In the approach, author employs scale augmentation by adjusting the size of the input images so that their shortest side measures no less than 480 pixels and no more than 800 pixels, while ensuring the longest side does not exceed 1333 pixels. Additionally, author incorporates random crop augmentations during the training phase.

2.2 Proposed Approach

In this paper, author focuses on utilizing transformer model to do object detection through transfer learning technique. The main transformer model contains four specific parts: backbone model, encoder, decoder and feed-forward network (FFN). The backbone extracted variety of features. Then as input for the transformer, it decodes after adding positional embedding to output object queries. The decoder processes the object queries then transmits to FFN to acquire final

predication results. In addition, set prediction loss is used to optimize the model to get the best prediction results (Carion, 2020). The system architecture is illustrated in Figure 2.

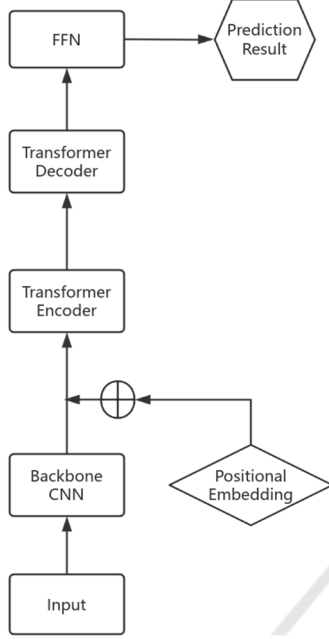


Figure 2: The pipeline of the model (Photo/Picture credit: Original).

2.2.1 Backbone

Beginning with an initial image x that belongs to $R^{3 \times H \times W}$ (where the image has 3 color channels), a standard CNN backbone produces an activation map of lower resolution, denoted as f in $R^{C \times H \times W}$. Commonly, author selects $C = 2048$, with H and W being one thirty-second of H_0 and W_0 , respectively.

2.2.2 Transformer Encoder

Initially, the activation map f , with dimensions C , is reduced to d lower dimension through a 1×1 convolution, generating a new feature map $Z_0 \in R^{d \times H \times W}$. The encoder is designed to process sequences, prompting the conversion of Z_0 's spatial dimensions into a single dimension, resulting in a feature map of dimensions $d \times HW$. In this paper, author adds two more transformer blocks in order to increase the output match precision between object queries and the predicted bounding box. Given the transformer architecture's invariance to permutations, encodings with location which are fixed are

incorporated and for every attention layer they are plugged into the input to maintain sequence order.

2.2.3 Transformer Decoder

With the standard design, the transformer decoder processes large number of embeddings. Unlike the model which decodes output sequences one element at a time in an autoregressive manner, model deciphers a number of objects simultaneously. Due to the decoder's permutation invariance, it is essential that the N input embeddings are distinct to yield varied outcomes. For each attention layer's input, they are combined with learned location embeddings which are called object queries, similar to the encoder's approach. These large number of object queries are transforms by the decoder into output embeddings. Then through large number of predictions of FFN, the bounding box and labels are decoded from object queries. The model utilizes encoder-decoder attention across large number of features to understand the whole parts in relation to one another through pairwise relationships, while leveraging the entire image for context.

2.2.4 FFN

In order to generated the final predictions, the author constructs a neural network with ReLU, a fully-connected layer. This FFN estimates every single element of each bounding box, while the class label is determined by the linear layer. Given that the number of stable sets of bounding boxes produce a unique category, denoted as \emptyset , is introduced to signify slots. It is similar to the "background" category.

2.2.5 Set Prediction Loss

The set prediction Loss is illustrated as follows (Carion, 2020). Let y represent the set of actual objects present in the image, and let $\hat{y} = \{\hat{y}_i\}$ denote the set of N predictions made by the model. Given that N typically exceeds the count of objects, y is treated as a similarly sized set of N to indicate the absence of an object. To establish a two sets correspondence, author aims to identify a permutation of a large number of features, σ , from the symmetric group SN , that results in the minimum matching cost:

$$\sigma = \operatorname{argmin} \sum_i^N L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \quad (1)$$

In this context, $L_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is denoted by the cost of matching a ground truth instance y_i with a predicted instance identified by index $\sigma(i)$.

In the subsequent step, author calculates the loss function, which refers to as the loss function (Stewart, 2016). Each ground truth element i is represented as $y_i = (c_i, b_i)$, where c_i denotes the intended category and b_i is a four-dimensional vector. Every element of the ground truth bounding box is denoted by this vector which is similar to the overall dimensions. In order to make predictions of index $\sigma(i)$, the class c_i 's predicted probability is given by $\hat{p}_{\sigma(i)}(c_i)$, and the associated predicted bounding box is denoted by $\hat{b}_{\sigma(i)}$. This loss is conceptualized in a manner akin to those found in typical object detectors: it is a weighted sum of the negative log-likelihood for class prediction and a box loss, which will be specified later on:

$$L_{Hungarian}(y, \hat{y}) = \sum_{i=1}^N [-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + L_{box}(b_i, \hat{b}_{\hat{\sigma}(i)})] \quad (2)$$

where $\hat{\sigma}_i$ is the optimal assignment.

2.3 Implementation Details

This study used Python 3.9.10 and Pytorch deep learning framework to conduct the experiment. For the hyperparameters, the research is conducted with the AdamW optimizer, configuring the initial learning rate for the transformer at 1e-4, for the backbone at 1e-5, and setting the weight decay to 1e-4. Author initializes all transformer weights using Xavier initialization and employ a ResNet model pre-trained on ImageNet, sourced from torchvision, for the backbone initialization. Additionally, the batch normalization layers within the backbone are kept frozen during training. This adjustment doubles the resolution, enhancing the detection of small objects, but it results in a sixteenfold increase in the computational expense for the encoder's self-attentions. Consequently, this leads to a twofold rise in the overall computational cost.

3 RESULTS AND DISCUSSION

In this section, author conducts a various experiment to illustrate that when training during transfer learning from COCO2007 to PASCAL VOC 2012, the transformer model achieves relatively high AP on the target dataset. Moreover, author utilizes Faster R-CNN as the comparison model and demonstrates that the transformer outperforms Faster R-CNN on both of these two datasets which proves the effectiveness

of this transfer learning process. Author also provides detailed evaluation about the whole experiment.

3.1 Trained on COCO 2007

Scale augmentation is employed, adjusting the input images so that their shortest side measures no less than 480 and no more than 800 pixels, while ensuring the longest side does not exceed 1333 pixels. Additionally, to enhance the learning of the whole picture of encoder attention mechanism, random crop augmentations are applied during training, which has been found to boost performance. A training image is cropped to a random square, which is subsequently resized to the range of 1333 pixels.

Table 1: Comparison with Faster R-CNN on COCO 2007 dataset.

Model	The number of parameters	AP
Faster R-CNN	166M	39.0
Transformer	41M	42.0

Table 1 illustrates when training on COCO 2007 the transformer model outperforms Faster R-CNN by 3.0 AP. Since the author trains the model with Adam and it takes a very long time for training to complete, this means the training epoch is longer and it can fetch more useful features. Faster R-CNN is trained without sufficient data preprocessing technique. Therefore, it makes sense that Faster R-CNN has lower performance.

3.2 Trained on PASCAL VOC 2012

The PASCAL VOC 2012 is a widely recognized benchmark in computer vision field. It is widely used for evaluating the performance of models in object detection, classification, and segmentation tasks. This widespread use allows for easy comparison of model's performance among well-developed models. Given its moderate size and complexity, the dataset is suitable for transfer learning, especially when computational resources are limited or when one aims to fine-tune pre-trained models quickly. It's an excellent choice for demonstrating the efficacy of transfer learning techniques. Therefore, the experiment is conducted on this target datasets. The mean AP is used to evaluate the model performance.

Table 2: Comparison with Faster R-CNN on PASCAL VOC 2017 dataset.

Model	The number of parameters	mAP
Faster R-CNN	166M	65.7
Transformer	41M	68.8

Table 2 illustrates that with less parameter, the transformer model achieves higher mAP which outperforms Faster R-CNN by 3.0. Also in the meantime, the AP acquired on PASCAL VOC is strictly higher than that of COCO 2007 which shows the effectiveness for transfer learning.

4 CONCLUSIONS

This study proposes a new object detection based on transformer modelling. In addition, this paper sets up a bidirectional matching loss for prediction. The model contains a Resnet-101 model as a backbone, an encoder part with an attention mechanism, a decoder with an object query input, and a feedforward network. The loss function is a two-step set prediction loss carefully designed for object detection. In addition, migration learning techniques are invoked to demonstrate the effectiveness of improving model performance through two baseline object detection datasets. The paper then conducts various experiments to analyse the performance of the model on these two datasets. The authors implement Faster R-CNN model for comparison. On both datasets, the transformer model outperforms the Faster R-CNN and has a higher AP by 3.0. Meanwhile, the transformer trained on the PASCLA VOC maintains AP of 68.8, which is significantly higher than that of COCO 2007. the effectiveness of transfer learning is well demonstrated. This redesigned approach to the detection system presents a number of challenges, particularly in the areas of training, optimization, and small-object performance. Previous detection models have been improved over the years to address similar problems. In the future, semantic segmentation tasks for transformers will be considered as the next phase of research.

REFERENCES

- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in neural information processing systems*, p. 30.
- Bahdanau, D., Cho, K., Bengio, Y., (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint:1409.0473.
- Wang, X., Girshick, R., Gupta, A., et al. (2018). Non-local neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp: 7794-7803.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
- Carion, N., Massa, F., Synnaeve, G., et al. (2020). End-to-end object detection with transformers. *European conference on computer vision*. Cham: Springer International Publishing, pp: 213-229.
- Graves, A., Graves, A., (2012). Long short-term memory. Supervised sequence labelling with recurrent neural networks, pp: 37-45.
- Cai, Z., Vasconcelos, N., (2019). Cascade R-CNN: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, vol. 43(5). pp: 1483-1498.
- Zhang, S., Chi, C., Yao, Y., et al. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp: 9759-9768.
- Lin, T, Y., Maire, M., Belongie, S., et al. (2014). Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, pp: 740-755.
- Everingham, M., Van, Gool, L., Williams, C, K, I., et al. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, vol 88. pp: 303-338.
- Stewart, R., Andriluka, M., and Ng, Y, A., (2016). End-to-End People Detection in Crowded Scenes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp: 2325-2333.