Proceedings of CECNet 2021
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210461

Detection Method of Computer Room Personnel Based on Improved DERT

Dan SU [a] , Qiong-lan NA [a1], Hui-min HE [a] and Yi-xi YANG [b]
[a] *State Grid Jibei Information and Telecommunication Company, Beijing 100053, China*
[b] *State Grid Information and Telecommunication Branch, Beijing 100761, China*

Abstract. Recently developed methods such as DETR [1] apply Transformer [2] structure to target detection. The performance of using Transformers for target detection (DETR) is similar to that of two-stage target detector. First of all, this paper attempts to apply Transformer to computer room personnel detection. The contributions of the improved DETR include: 1) in order to improve the poor performance of small target detection. Embed Depthwise Convolution in the encoder. When the coding feature is reconstructed, the channel information is retained. 2) in order to solve the problem of slow convergence in DETR training. This paper improves the cross-attention in DECODE and adds the spatial query module. It can accelerate the convergence of DETR. The convergence speed of the improved method is six times faster than that of the original DETR, and the mAP0.5 is improved by 3.1%.

Keywords. Transformer, DETR, Personnel detection, Cross attention, Depthwise Convolution

1. Introduction

DETR proposes an end-to-end framework based on codec Transformer architecture and two-part matching, which directly predicts a set of bounding boxes without post-processing. However, there are several problems: (1) it takes longer training time to converge. DETR needs 10 to 20 times more training time than modern mainstream detectors to converge. (2) the performance of DETR in small target detection is relatively poor. Current target detectors usually use multi-scale features to detect small targets from high-resolution feature images. For DETR, high-resolution feature maps mean high complexity. The improved DETR can improve these two problems: (I) improve the cross-attention module of the decoder, better learn high-quality content embedding [3], and add spatial query to speed up the convergence speed of training. (II). Embed Depthwise Convolution in the encoder. When the coding feature is reconstructed, the channel information is retained. This can not only improve the detection accuracy, but also improve the detection effect of small targets.

This paper summarizes the research status of general object detection and DETR variants in section 2. Every part of the improved DETR method is described in section 3.1, especially the detailed improvement method is given in the Transformer Encoder

[1] Corresponding Author: Qionglan Na, E-mail: 81885883@qq.com ; This work is supported by the Science and Technology Project of State Grid Jibei Power Company Limited (No. B3018E210000).

part. In section 3.2, we analyze the reasons for the slow convergence of DETR, and introduce the improvement of Transformer Decoder in detail. At section 4.1 we introduced the dataset and the parameter settings during training. In section 4.2, we compare the detection performance of the improved DETR and other algorithms, which shows the effectiveness of the proposed method. In section 4.3, we visualize the output of the cross-attention module. In section 5, we summarize this paper and look forward to the future.

2. Related Work

General object detection. General object detectors can be divided into two categories, namely, two-stage detection and one-stage detection, and one-stage detector is mainly based on anchor point. Anchor points are generated at the center of each sliding window position, providing candidates for objects. For example, SSD[4], RetinaNet[5], YOLO [6]and FCOS[7], the network directly predicts the types and offsets of anchor points in the whole feature graph. Usually, multiple label targets are generated for each anchor point, and then the repeated prediction of the object is removed by post-processing techniques such as NMS. The two-stage detector does not directly output the final prediction. Such as Faster RCNN[8] and Mask RCNN[9], they first generate prospect proposals through the Regional proposal Network (RPN). The ROI Pool[8] or ROI Align [10] layers are used to extract the features generated by RPN from the backbone features. Then NMS and other processing techniques are used to remove repeated predictions to get the final results. In general, the accuracy of the two-stage detector is higher than that of the first-stage detector.

DETR and its variants. DETR successfully applies Transformer to object detection without the need for additional manual design components and performs object detection end-to-end. For example, non-maximum suppression or the generation of an initial anchor. DETR has two problems: slow convergence and poor small target detection. To solve these problems, multi-scale features are applied to help detect small objects in Deformable DETR[11]. And the use of deformable attention module can solve the problem of high computational complexity caused by the self-attention of the global encoder. The adaptive clustering converter[12] clusters the key in the self-attention mechanism and improves the convergence speed by reducing the complexity. The TSP (transformer-based ensemble prediction) method[13] eliminates the cross-attention module and combines FCOS and R-CNN detection heads. The spatial modulation common attention (SMCA) method[14] uses a Gaussian graph around several (shift) centers learned from the decoder embedding to modulate the global crossover attention of the DETR multiple heads to pay more attention to several areas within the estimation frame. Conditional DETR[3] learns conditional space queries from decoder content embedding and predicts spatial attention weights without manual attention decay, highlighting the four extremes of box regression and different regions within the objects used for classification.

this paper attempts to apply Transformer to personnel detection in the computer room. In order to solve the problem of slow convergence of DETR, the spatial query module is introduced to make the algorithm converge quickly. Embed Depthwise Convolution into the Encoder of Transformer. The coding features are reconstructed. Retain channel information as much as possible. This can improve the detection accuracy of the algorithm. The improved method has high convergence speed and detection

accuracy, and has achieved excellent results in the detection of personnel in the computer room.

3. Proposed Method

3.1. Depthwise Convolution in transformer network

The improved DETR, as shown in figure 1, consists of CNN backbone for feature extraction, location coding, six Transformer encoders for feature coding and reconstruction, six transformer decoders for object query, and a simple feedforward network for detection and prediction.

Feature extractor. ResNet50 is used as the backbone of the feature extractor. We use ResNet50 as the feature extraction backbone network. First of all, the image of $x \in \mathbb{R}^{3\,x\,H_0\,x\,w_0}$ is used as input. Then Outputting an advanced feature map $h \in \mathbb{R}^{B\,x\,C\,x\,H\,x\,w}$. Where $C = 2048$ and $H, W = \frac{H_0}{32}, \frac{W_0}{32}$. The input feature map is down sampled using the convolution of 1x1 to reduce the dimension of the number of channels to get the feature map $m \in \mathbb{R}^{256\,x\,\frac{H_0}{32}\,x\,\frac{W_0}{32}}$. Then the feature graph m is serialized and transformed into a serialized feature with a shape size of $(\frac{H_0\,x\,W_0}{1{,}024}, 256)$.

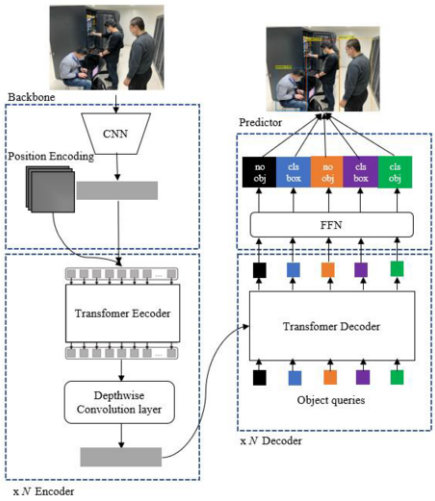

Figure 1. improved overall structure of DETR

Positional Embedding. The author provides two methods of absolute position coding. One is trigonometric function position coding, which is also known as Sinusoidal position coding; the other is training position coding, which directly takes the position coding as a trainable parameter, which is initialized randomly and updated with the training process. Sinusoidal location coding is used in this article.

Encoder. Contains six Transformer encoders with built-in Depthwise Convolution. Our encoder consists of a standard Transformer encoder and Depthwise Convolution layer, each Transformer encoder follows the standard architecture, and consists of a multi-head self-attention module and a feedforward network. Positional embedding is introduced into the input of each self-attention layer, and each Depthwise Convolution

layer consists of a 3 × 3 convolution, batch normalization, and correction linear unit. As shown in figure 1, the location embedded and extracted features are used as inputs to the Transformer encoder. After Transformer Encoder, the sequence characteristics of $(\frac{H_0 \, x \, W_0}{1{,}024}, 256)$ are obtained. The feature is reconstructed and deformed to get the shape($256, \frac{H_0}{32}, \frac{W_0}{32}$) of feature graph. Using 256 convolution kernel of 3x3 for Depthwise Convolution. In this way, the feature information of different channels in the same spatial location can be preserved to the maximum extent. Then through batch normalization and GELU activation unit. Finally, the feature graph m is serialized and transformed into $(\frac{H_0 \, x \, W_0}{1{,}024}, 256)$ serialized features.

Decoder. Contains six Transformer decoders. The main difference from the original DETR is the input query and cross-attention module. It will be introduced in Section 3.2.

Object predictor. The object predictor is a feedforward network with two fully connected layers, namely, a box regression layer to predict the location of the target (x, y, w, h) and a frame classification layer to predict the target score. Therefore, N object queries are independently decoded into frame coordinates and class labels by the feedforward network, resulting in N final predictions, including object (defect) and no object (background) prediction.

3.2. Cross-attention module based on spatial Query

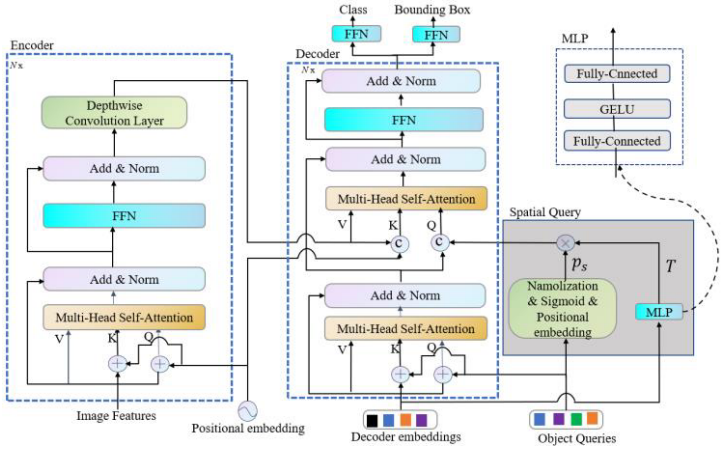

Figure 2. improved DETR codec structure

Due to the slow convergence of DETR, various variants of DETR have been improved. We think that Conditional DETR has given the answer to the reason why DETR converges slowly. DETR highly relies on high-quality content embedding to locate the extremity region of an object, which is the key to locating and identifying objects. In order to make our algorithm converge quickly, as shown in figure 2, we first modify the DETR Cross-Attention, and the output of the original Attention module is shown in formula (1).

$$c_q^T c_k + c_q^T p_k + p_q^T c_k + p_q^T p_k \quad (1)$$

Where c_q is content query, p_q is spatial query, c_k is content key, p_k is spatial key. Among them, c_q and c_k contain more information on the image (color, texture, etc.). $p_q$and $p_k$contain more spatial information.

As shown in figure 2, the Query of the decoder is stacked by the c_q and the p_q. The key is stacked by c_k and p_k. The output of the decoder becomes the result of the formula (2).

$$c_q^T\ c_k\ +\ p_q^T\ p_k \tag{2}$$

Where $c_q^T\ c_k$ represents the similarity of image information, and $p_q^T\ p_k$ represents the similarity of computing spatial information.

Depu Meng et al found that decoder embedding and reference point contain spatial location information. This paper also introduces Conditional spatial query prediction and makes simple improvements, as shown in the gray module in the figure.

First of all, the reference point corresponding to query is normalized and mapped to the same bit coding space as spatial key to get reference p_s in formula (3).

$$p_s\ =\ sinusoidal((simgmoid(s)) \tag{3}$$

Then, we map the offset information contained in decoder embedding to high-dimensional space through an MLP. Get the "offset" for p_s. MLP, we're using the GELU activation function. GELU is not only nonlinear activation, but also introduces the idea of random regularity. The experimental effect is better than that of RELU. Finally, we get p_q by doing the inner product of p_s and T. It is important to note that when we use T as the inner product, we use the values on the diagonal matrix to calculate.

4. Results and Discussion

4.1. Experiment setup

Dateset: A total of 4192 images were marked by professionals, including 10674 tags in one category. The ratio of training sets to test sets is 8:2.

Training: the improved DETR structure uses ResNet-50 pre-trained by ImageNet[15] as the backbone network, the batch norm layer is fixed[16], and the Transformer parameters are initialized using Xavier initialization scheme[17]. Weight falloff is set to 10-4. Use the ADAMW[18] optimizer[19]. The learning rates of backbone and converter are initially set to 10-5 and 10-4, respectively. The shedding rate of the transformer is 0.1. For 50 training cycles, after 40 cycles, the learning rate is reduced 10 times; object query is set to 300. for other default parameters that use the original DETR, and two 12G GTX-1080 are used to complete the training.

4.2. Comparison with DETR

The improved DETR uses the same architecture as DETR, and the improved DETR is negligible in terms of computational cost and training time for each period. Compared with DETR, the improved DETR is due to DETR in accuracy and convergence speed. The improved DETR can achieve better overall performance. the accuracy of our method for training 75 epoch is a little higher than that of the original DETR training 300 epoch. Moreover, the convergence rate of the improved DETR is 6 times faster than that of the original DETR method. As shown in Table 1, DETR needs to train 300 epoch to achieve good performance. In the case of significant improvement in convergence speed and performance, the increase in the number of FLOPS of the improved DETR is very small. Flops increased by 4G. The FPS is reduced by 2 frames. Compared with the Faster RCNN using FPN, the improved method can achieve the performance comparable to the two-

stage algorithm. However, the detection effect of DETR algorithm for small targets is not as good as that of two-stage algorithm, and the detection effect of medium targets and large targets is better than that of Faster RCNN.

Table 1. performance comparison between DETR and improved DETR

Method	epoch	AP	AP50	AP75	APs	APm	APl	FPS	FLOPs
Faster RCNN+FPN	110	61.8	89.3	67.6	26.7	48.2	72.1	13	181
DETR	100	56.2	86.9	60.1	14.2	46.8	69.5	15	86
DETR	300	61.4	88.1	66.5	24.4	50.8	74.1	15	86
Ours	75	61,4	**90.0**	67.1	24.6	51.1	74.2	13	90

4.3. Visualization

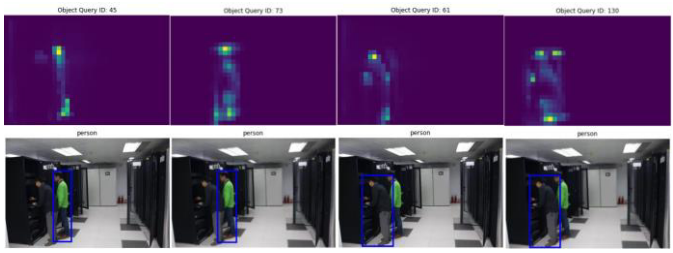

Figure 3. Multi-Attention visualization of DETR Decoders and improved DETR Decoder

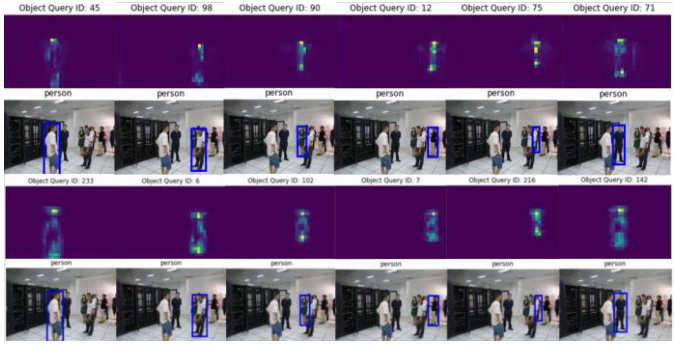

Figure 4. Multi-Attention visualization of DETR Decoders and improved DETR Decoder

As shown in figure 3, the first line is the id of the object query, and the second line is the attention matrix diagram of the decoder's Multi-Attention. The third line is the currently detected category, and the fourth line is the corresponding bounding box. Figure 3 shows a photo that detects two targets. The left (DETR) and the right (Ours) are visualization results of the same location. Interesting things can be found by comparison. The improved DETR seems to focus on more areas, which can improve the detection effect of small targets. This is more obvious in figure 4, where the visualization methods of the improved DETR and the original DETR are shown above and below, respectively.

5. Results and Discussion

For the first time, we use Transformer to detect the personnel in the computer room. The detection accuracy can be improved by embedding Depthwise Convolution in the encoder and preserving the channel information when the coding features are reconstructed. In order to make the DETR algorithm converge faster, we improve the cross-attention module of the decoder and mine the location information of the bounding box of the previous decoder as much as possible before spatial query. The improved DETR is used to achieve the detection performance comparable to that of the two-stage detector. The improved detection method has better detection accuracy and convergence speed. The detection method of computer room personnel in this paper provides a good idea to improve the migration to other fields. In the future, we will continue to study the detection of dense occlusion personnel and model compression.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End object detection with transformers. InECCV, 2020.1,2,5,6,7,9
[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser,L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017).
[3] Meng D , Chen X , Fan Z , et al. Conditional DETR for Fast Training Convergence. 2021.
[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. InECCV, 2016.2
[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. InICCV,2017.1,2,6,8,9
[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and AliFarhadi. Y ou only look once: Unified, real-time object detection. InCVPR, 2016.1,2,6
[7] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS:Fully convolutional one-stage object detection. InICCV,2019.2,4,8,9
[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun.Faster R-CNN: Towards real-time object detection with region proposal networks. InNeurIPS, 2015.1,2,6,7,9
[9] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. InICCV, 2017.2,6,7
[10] Zhaowei Cai and Nuno V asconcelos. Cascade R-CNN: Delving into high quality object detection. InCVPR, 2018.2
[11] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection.CoRR, abs/2010.04159, 2020.1,2,6,7,8
[12] Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer.CoRR, abs/2011.09315, 2020.2
[13] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris Kitani.Rethinking transformer-based set prediction for object detection.CoRR, abs/2011.10881, 2020.2,7
[14] Jifeng Dai, Haozhi Qi, Y uwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. InProceedings of the IEEE international conference on computer vision, pages 764–773, 2017.3
[15] Xavier Glorot and Y oshua Bengio. Understanding the difficulty of training deep feedforward neural networks. InAISTATS, 2010.6
[16] Sobti, P ., Nayyar, A., & Nagrath, P . (2021). EnsemV3X: a novel ensembled deep learning architecture for multi-label scene classification. PeerJ Computer Science, 7, e557.
[17] Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In Journal of physics: conference series (Vol. 1142, No. 1, p. 012012). IOP Publishing.
[18] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. InICLR, 2017.6
[19] Jain, A., & Nayyar, A. (2020). Machine learning and its applicability in networking. In New age analytics (pp. 57-79). Apple Academic Press.