Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023312

The development of transformer in vision

Jialong Shao

Zhejiang University of Technology, Hangzhou, China, 31000

a1216574908@gmail.com

Abstract. Transformer is a deep neural network that utilizes a self-attentive technique to process data in parallel. The area of vision was neural network based before Transformer, with great frameworks like faster-RNN, YOLO, etc. Transformer, which was developed to stop this phenomena, is discussed in this article together with its accomplishments in the field of vision during the past few years and its projections for the future in order to provide some references for more research.

Keywords: transformer, Detr, Vit, self-attention mechanism.

1. Introduction

A significant area of artificial intelligence is computer vision, which primarily consists of subtasks including semantic segmentation, target identification, and target tracking. Deep learning has produced impressive research accomplishments in the area of vision over the last ten or so years.

The CNN (convolutional neural network) is the most traditional deep learning framework based on deep learning techniques [1]. Full connectivity in the CNN compensates for the loss of spatial information, and the alignment of depth directions reduces the likelihood of quick overfitting even when the model is trained with a high number of parameters. The convolution kernel, which features inductive biases like translation invariance and local sensitivity and may capture local spatio-temporal information, is the fundamental component of the CNN. The CNN has reached a whole new level as a result of Resnet's suggestion [2], and is now the industry leader in computer vision. Convolutionary processes, however, can not fully utilize contextual information since they do not have a comprehensive grasp of the picture itself and are unable to model the connections between features. Convolution's weights are also fixed; they don't alter in response to changes in the input. Researchers have thus been looking for a more thorough foundation.

In 2017, the Google team published a paper titled ATTENTION IS ALL YOU NEED [3] that introduced a straightforward network architecture dubbed Transformer in the field of sequence transcription. Transformer totally forgoes CNN and RNN in favor of attention structures. Several researchers developed the Transformer to do visual tasks as their study developed. In 2018, the idea of an image transformer was presented. Transformer was used by DETR for the visual collar for the first time in 2020, followed by deformable Transformer [5], Transformer for classification tasks (ViT) [6], and other applications.

DOI: 10.54254/2755-2721/4/2023312

2. The basic principles of the transformer

Transformer is a particular sequence to sequence model since it heavily relies on self-attention. The recurrent neural network (RNN) [7], whose topology is depicted in Figure 1, was the most commonly utilized in the field of natural language processing before Transformer was suggested. The RNN can only carry out sequential calculations in succession, which has two drawbacks: it limits the parallelism of the model since the current computation depends on the outcomes of the preceding moment's computation. Lengthy information is difficult to employ for preserving context-dependent interactions since it is difficult to retain overly long information during computation.

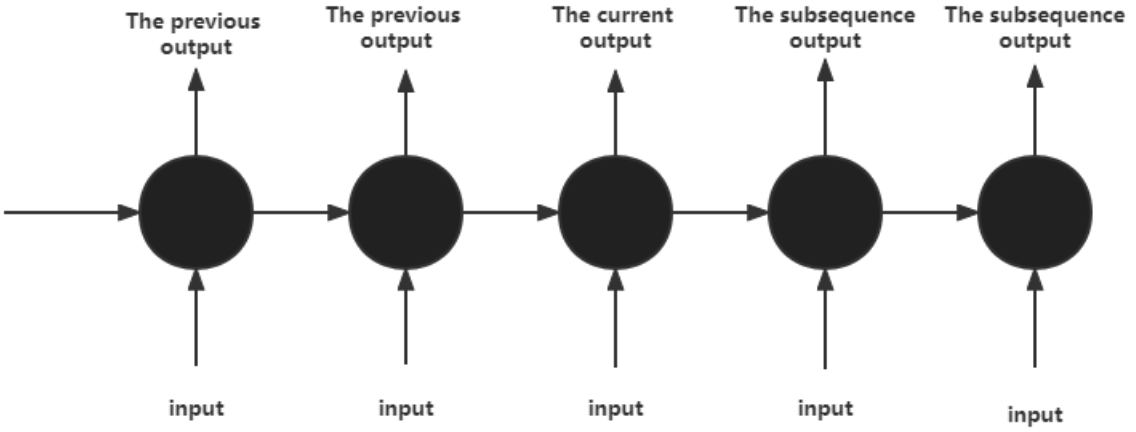

Figure 1. RNN structure diagram.

A creative solution to the RNN issue is a new layer called self-attention, which has precisely the same input and output as an RNN but can be calculated in parallel. This layer activates the attention mechanism, a system that imitates the internal workings of biological observation behavior and sharpens observation in particular areas to enable fast extraction of key aspects from sparse input. Scaled dot-product attention, which the Transformer architecture introduces, uses dot-products for similarity computation and is quicker and more space-efficient in practice than conventional attention. To create the matrices Q (query), K (key value), and V (value), which are calculated as given in equation, the input must first be linearly converted (1).

$$Attention(Q,K,V) = softmax\frac{QK^T}{\sqrt{d_k}}V \tag{1}$$

The Transformer employs an encoder-decoder architecture, which consists of six layers of individually stacked encoders and decoders. This model structure prevents loops, and the encoder's output is sent to the decoder on each layer to compute attention. The multi-head attention layer and the feed forward connection layer are the two sub-layers that make up each layer of the encoder construction. The masked multi-head attention layer and the multi-head attention layer are two of the decoder's three sub-layers. The attention action is followed by two linear transformations, a ReLU activation output, and a fully linked forward network in each layer of the encoder and decoder that applies the same operation to each positional vector. As seen in Figure 2, each sub-layer is followed by a residual connection and a layer normalization.

DOI: 10.54254/2755-2721/4/2023312

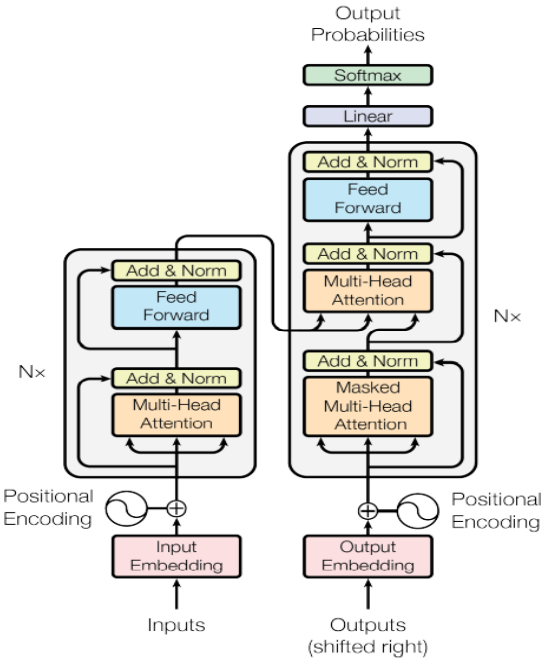

Figure 2. Transformer model structure.

The positioning information of words in the text must be intentionally inserted by positional encoding because the Transformer computation ignores recursion and convolution of cyclic structures and cannot imitate them. Each word in the sentence has a specific number assigned to it that correlates to a vector, which adds a particular amount of positional information to each word when the positional vector is combined with the word vector. In equation, the formula is shown (2).

$$\begin{cases} PE_{(pos,2i)} = \sin\left(\dfrac{pos}{10000^{\frac{2i}{d}}}\right) \\ PE_{(pos,2i+1)} = \cos\left(\dfrac{pos}{10000^{2i+\frac{1}{d}}}\right) \end{cases} \tag{2}$$

The process of the Transformer is as follows, using machine translation as an example.

Step 1. The word input procedure turns the input sentence into a vector, and the embedding position encoding method yields the word's position vector; the two are then combined to provide the input for the model.

Step 2. The word vector matrix created in step 1 is sent into the encoder, which then passes the output up to the next encoder after passing via the multi-headed attention layer and pre-neural network.

Step 3. The coded information matrix for every word in the sentence is acquired after six encoders. Each of the six decoders receives a copy of the matrix. The masked multiheaded attention layer, the multiheaded attention layer, the multihead attention layer, and the feedforward connection layer each run the matrix through each encoder in turn.

Step 4. The ultimate output of the decoder is probabilities after passing through a linear layer, a softmax layer, and a final layer.

DOI: 10.54254/2755-2721/4/2023312

3. Based on the Transformer vision model

3.1. DETR

Figure 3 depicts the overall organization of the target detection framework that Carion et al. [4] created in addition to the detection Transformer (DETR), a Transformer-based object identification framework. To get the final prediction, a feed forward network (FFN) decodes each output of the decoder separately into frame coordinates and class labels. By directly outputting the final set of predictions in parallel based on the relationships between the objects and the global context, DETR treats target detection as an ensemble prediction problem, streamlining the overall target detection process and removing the need for manually created techniques like non-maximal value suppression and anchor point removal. This enables end-to-end automatic training and learning. DETR is theoretically simpler than previous methods, does not call for specific libraries, and has a higher average precision (AP) of 42% for big target identification, which is faster and more accurate than Faster-RCNN [8]. Although DETR performs well on large targets, its accuracy is just 20.5 percent when it comes to tiny targets. Additionally, DETR requires more training time to converge as compared to other models.

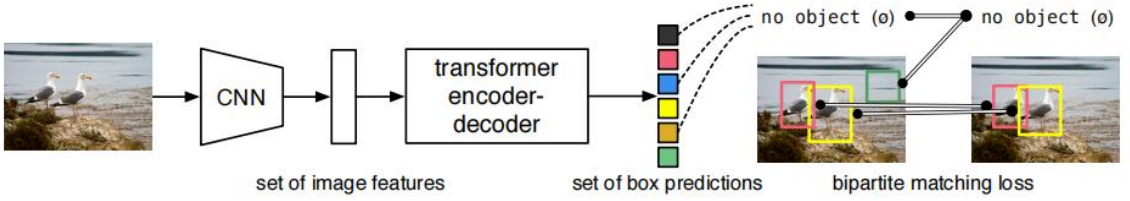

Figure 3. DETR model structure.

3.2. Deformable DETR

The target model should be able to adapt to numerous complicated deformations, which calls for more efficient algorithms or bigger data sets, in order to address the issue of DETR, which has low detection accuracy on tiny targets. Zhu et aldeformable .'s DETR, which combines the deformable convolution's [10] benefit of excellent sparse space sampling with Transformer's potent relational modeling capabilities, was developed to overcome the drawbacks of DETR. The structure is better able to adapt to object deformation thanks to the deformable convolution, which converts the fixed shape convolution into a variable convolution that can change shape depending on the shape of the object.

In Deformable DETR, the authors replace Transformer's attention module with a deformable attention module to process the feature map, pre-screen all of the feature map pixel points, and concentrate only on a small number of key sampling points around the reference point without taking the feature map's spatial size into account. This significantly reduces computational complexity and solves the convergence and feature spatial resolution issues. Deformable DETR has 10 times less training cycles than DETR and improves 5.9% APs, notably for tiny target identification, but it still struggles to find obscured targets.

3.3. ViT

Transformer's first effort, called ViT, replaces conventional convolution on sizable datasets, offering a crucial groundwork for the advancement of Transformer in computer vision applications. The idea of picture block is introduced in order to transform the image into sequence data that the Transformer structure can handle. The two-dimensional picture is split into blocks, each of which is then flattened into a one-dimensional vector. Each vector is then transformed using a linear projection while simultaneously introducing position coding and adding the sequence's position information. To properly describe the overall information, a categorization flag is also placed before the incoming

DOI: 10.54254/2755-2721/4/2023312

sequence data. ViT models are often fine-tuned for smaller downstream tasks after being pretrained on big datasets. VIT-H/14, which is more effective and scalable than conventional CNN networks, successfully overturns the dominance of convolution-dominated networks in classification tasks on the ImageNet dataset with an accuracy of 88.55 percent. Figure 4 depicts the ViT's framework.

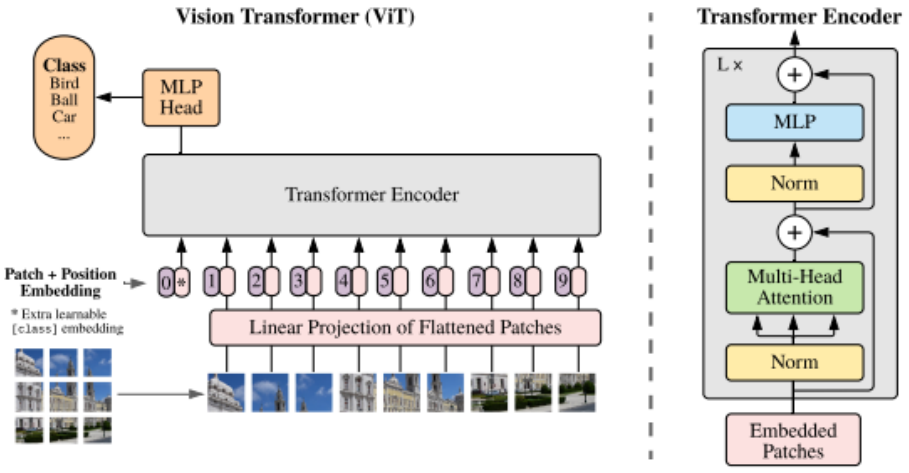

Figure 4. ViT structure.

4. Comparative experiments

Comparative tests based on open-source datasets are carried out in this study to evaluate the effectiveness of each Transformer-based target identification model, as shown in Table 1. The comparative trials reveal that although the Faster RCNN model is faster at detecting targets, the Transformer-based model is more accurate at doing so. The average response time of the algorithm based on Transformer model is found to be 2.1 times faster than that of the Faster RCNN model in the comparison test using the COCO dataset, while the average response time for larger target detection is found to be 2.8 times faster than that of the Faster RCNN model. This proves that the algorithm is based on the Transformer model. This demonstrates that the field's practical target detection application based on the Transformer model is still in its infancy.

Table 1. Comparative experiments.

Method	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_L	AP_M	Data set	FPS
Faster R-CNN + FPN	109	42.0	62.1	45.5	26.6	45.4	53.4	COCO2017	26
DETR	500	42.0	62.4	44.2	20.5	45.8	61.1	COCO2017	28
Deformable DETR	50	43.8	62.6	47.7	26.4	47.1	58.0	COCO2017	19
ViT-B/16-FRCNN	-	36	56.3	39.3	17.4	40.0	55.5	COCO2017	-

5. Conclusion

Transformer has emerged as a hub for computer vision research, and because of the model's enormous potential, academics have been paying close attention to it. Transformer does, however, have a few very clear flaws. Transformer overcomes the problem of CNNs being able to extract only local

DOI: 10.54254/2755-2721/4/2023312

dependencies and resolves the issue of concurrent RNN operations, but it also lacks local sensitivity and generalization and frequently needs big data sets and high flops to reach good accuracy. Therefore, one of the future approaches will be to figure out how to lessen the Transformer's dependency on enormous volumes of data. The combination of CNN and transformer is complimentary, and an improved method is also a highly promising approach. At the same time, the local importance of pictures is not trivial, and a single transformer will have a large redundancy problem.

References

[1] Min Lin, Qiang Chen, Shuicheng Yan, Network In Network [J] arXiv preprint arXiv: 1312.4400, 2014.

[2] Kaiming He, Xiangyu, Zhang Shaoqing, Ren Jian Sun: Deep Residual Learning for Image Recognition [J] arXiv preprint arXiv: 1512.03385, 2015.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need. In NIPS, 2017.

[4] PARMAR N, VASWANI A, USZKOREIT J., et al. Image transformer [C] International Conference on Machine Learning, 2018, pp.4055-4064.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko: End-to-End Object Detection with Transformers. In ECCV, 2020.

[6] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai: DEFORMABLE DETR: DEFORMABLE TRANSFORMERSFOR END-TO-END OBJECT DETECTION. In ICLR, 2021.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In ICLR, 2021.

[8] Wojciech Zaremba, Ilya Sutskever, Oriol Vinyals: Recurrent Neural Network Regularization. In LSTM, 2014.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In NIPS, 2015.

[10] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In ICCV, 2019.