Vision Transformer Architectures in CVPR2022

June 30, 2022

1 Introduction

With the recent efficient computation, Transformers are gradually used and known as SOTA approaches in many vision tasks. In fact, Transformer-based approaches are potential to replace Convolutions Neural networks (CNNs) as the powerful tool, which is seen in visual benchmarks. In particular, the top of leaderboards in video objects segmentation is SwinB-AOTv2-L(MS) [16], a Transformer-based method. Attention paradigms show human-like ability, which focuses on the region of interest and transformer-based methods have global receptive fields, which is beyond to CNNs. This report summaries some promising Transformer-based architectures in CVPR2022, explaining why designs of attention mechanisms are much more robust than that of CNNs.

2 Methodology

Convention robust visual Transformer like ViT [5], DeiT [12] captures global receptive field, high model capacity. However, simply enlarging receptive field in conventional Transformer also rises to several concerns or using dense attention e.g. in ViT leads to excessive memory and computational cost. Features can be influenced by irrelevant parts of which are beyond the region of interests. Sparse attention adopted in PVT or Swin Transformer is data-agnostic and may limit the ability to model long range relation.

In this report, many novel Transformer architectures will be considered and contributed to address the drawbacks. Regarding architectures, attention paradigms will be categorized as follos:

2.1 Adaptive tokens

ViT, DeiT are easily vulnerable in training and causes to over-fitting, slow convergence and high computation costs because each query attends to all the other tokens. As a result, the author [14] proposes deformable self-attention the position of key and value pairs in self-attention are selected in a data-dependent way, thus to focus on relevant regions. They propose a general backbone model with deformable attention for both classification and detection.

2.2 Window-based attention

Recently, a surge of interest in visual Transformers is to reduce the computational cost by limiting the calculation of self-attention to a local window. Most current work uses a fixed single-scale window for modelling by default, ignoring the impact of window size on model performance. However, this may limit the modeling potential of these window-based models for multi-scale information. The author [9] proposes to use dynamic multi-scale windows to explore the upper limit of the effect of window settings on model performance. In DW-ViT, multi-scale information is obtained by assigning windows of different sizes to different head groups of window multi-head self-attention. Then, the information is dynamically fused by assigning different weights to the multi-scale window branches. Compared with related state-of-the-art (SoTA) methods, DW-ViT obtains the best performance. Specifically, compared with the current SoTA Swin Transformers [8], DW-ViT has achieved consistent and substantial improvements on all three datasets with similar parameters and computational costs. In addition, DW-ViT exhibits good scalability and can be easily inserted into any window-based visual Transformers.

2.3 Adaptive attention

The author [3] proposes a Transformer decoder, which contains self and cross-attentions. Besides, the author [4] introduces the modification in self-attention block with spatial window attention and channel group attention. Although the accuracy of the proposed method is not much significant, they introduce an interesting insight in the attention block.

2.4 Multi-scale attention

MViTv2 has state-of-the-art performance in 3 domains: 88.8% accuracy on ImageNet classification, 58.7 APbox on COCO object detection as well as 86.1% on Kinetics-400 video classification. The author [7] proposes decomposed relative positional embeddings and residual pooling connections in attention blocks. Besides, MViT backbone uses with FPN, in which global self-attention and local window attentions which calculate windows inside every token. Pooling attention and window attention both control the complexity of self-attention by reducing the size of query,key, value.

- Pooling attention pools features by downsampling them via local aggregation, but keeps a global self-attention computation.
- Window attention keeps the resolution of tensors but performs self-attention locally by dividing the input (patchified tokens) into non-overlapping windows and then only compute local self-attention within each window.

In conclusion, they mention that combining pooling attention and Hwin achieves the best performance for object detection.

Author paper MeMViT [13] uses multi-scale in the proposed architecture. In details, memory augmented MViT is built on the top of MViT video model, it consists of multiple self-attentions. The method is good at exploiting temporal information. It caches feature map at each iteration. The author mentions temporal support (40s) while others use less than 5s of a video, which is without hitting the computation or memory bottlenecks. They have online and cache memory like the paper [10]. Multi-scale vision Transformer uses strided pooling to extract features at different scales, pooling attention that pools spatialtemporal dimensions of Q, K, V to reduce computational costs of an attention layer. Many methods consider two modules including memory and online video modeling, that is efficient in computation. However, these methods capture only final layer features and require two backbones, two rounds of training and inference computation. MeMViT flexibly models features at arbitrary layers with minimal changes to standard training methods and only requires one standalone backbone.

2.5 Hybrid networks

In this paper, the author [6] has 3 contributions to Transformer and train from scratch, which make quicker convergence than that for RCNN. 3 contributions:

- From RoIPooling to RoIAlign. We observe that RoIPooling hinders the gradient from being smoothly back-propagated to backbone layers, and address this problem by replacing RoIPooling with RoIAlign.
- From T-T-T-T to C-C-T-T. We replace the first two stages of vision Transformers with convolution blocks, namely, from T-T-T-T to C-C-T-T.
- Gradient Calibration. Since it is better to adjust all of the layers a little rather than to adjust just a few layers a large amount.

2.6 Spatialtemporal attention

The author [15] proposes a nearly convolution-free, which contains a Transformer backbone and a query-based video instance segmentation head. In the backbone stage, they propose a nearly parameter-free messenger shift mechanism for early temporal context fusion. In the head stages, they propose a parameter-shared spatiotemporal query interaction mechanism to build the one-to-one correspondence between video instances and queries. Thus, TeViT fully utilises both framelevel and instance-level temporal context information and obtains strong temporal modeling capacity with negligible extra computational cost. [13] also consider this attention in their architecture.

2.7 Techniques

The author [1] illustrates data augmentation is crucial to vision Transformers. Mixup and cutmix are mentioned for showing class attention improvement. The author [11] introduces (pre) training with less data for Transformer-based methods. The technique is list as follows:

- Data augmentation
- Better regularization
- Distillation (CNN teacher)
- Optimization

Specifically, Transformer-based methods do not require much large datasets, even when training Transformers from scratch. On the contrary, we see that CNNs backbones usually have to be pre-trained on datasets like ImageNet dataset.

3 Visual Benchmarks

This section will provide the outlook over SOTA methods in segmentation tasks, which is well reported in visual benchmarks on image datasets like COCO, Cityscapes, Pascal Context, ADE20K and video datasets like DAVIS17, YouTube-VOS in Table. 1. Overall, the SOTA methods on different datasets under the considered tasks belong to transformer-based methods. As a result, the potential with computation improvements in attention and transformer mechanisms as shown in previous section make this technique gain superior accuracies in comparison with that of CNNs.

Table 1: Quantitative results of SOTA methods on different datasets are reported by paperswithcode.com in the year 2022.

Dataset	Method	Accuracy	Metric
COCO test-dev	DINO (Swim-L,multi-scale) [17]	63.3	boxAP
Cityscapes test	ViT-Adapter-L (Mask2Former, BEiT pretrain, Mapillary) [2]	85.20	mIOU
Pascal Context	ViT-Adapter-L (Mask2Former, BEiT pretrain) [2]	68.2	mIOU
DAVIS16	SwinB-AOTv2-L (MS) [16]	93	J&F
DAVIS17	SwinB-AOTv2-L (MS) [16]	87	J&F
YouTube-VOS	SwinB-AOTv2-L [16]	86.5	J&F

References

[1] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2022.

[2] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[4] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. *arXiv preprint arXiv:2204.03645*, 2022.

[5] 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.

[6] Weixiang Hong, Jiangwei Lao, Wang Ren, Jian Wang, Jingdong Chen, and Wei Chu. Training object detectors from scratch: An empirical study in the era of vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4662–4671, 2022.

[7] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[9] Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, and Xiaojun Chang. Beyond fixation: Dynamic window visual transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11987–11997, 2022.

[10] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7406–7415, 2020.

[11] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[12] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[13] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.

[14] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022.

[15] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2885–2895, 2022.

[16] Zongxin Yang, Jiaxu Miao, Xiaohan Wang, Yunchao Wei, and Yi Yang. Associating objects with scalable transformers for video object segmentation. *arXiv preprint arXiv:2203.11442*, 2022.

[17] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.