Swin Transformer V2: Scaling Up Capacity and Resolution
Reading Report

Tianyu Shi

May 21, 2022

1 Introduction

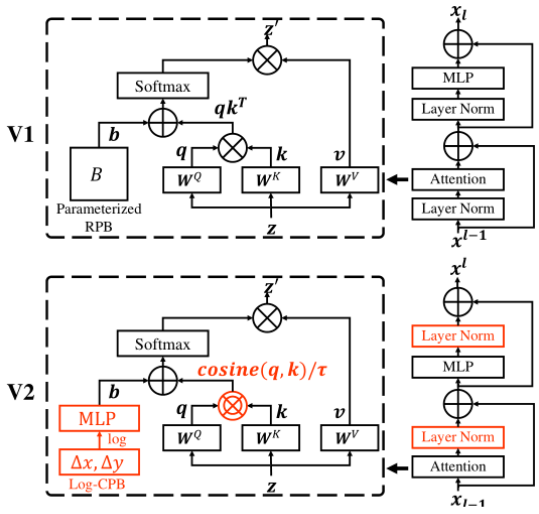
Today, more and more pre-trained models are being proposed in Computer Vision, most of them mainly with a Transfomer backbone. Swin Transformer V2 is a work by the Swin-T team on scaling up vision models based on the Swin-T model, refreshed to new SOTA on 4 datasets. The starting point of the article is that there has not been a better exploration of increasing model scale in the field of vision, as there has been in NLP. Possible reasons for this are: (1) The potential for training instability while increasing the size of the visual model. (2) For many downstream tasks that require high resolution, there is no well explored way to migrate trained models at low resolution to larger scales. (3) GPU memory costs too much.

The authors therefore propose three improvement points on the Swin-T based backbone in response to the above observed problems.

2 Method

2.1 Post normalization

The first improvement the authors made was to replace the pre-norm operation in the transformer block with a post-norm operation. The authors found that after increasing the Swin Transformer model from small size to large size, the activation values of deep layers of the network became very large and had a large gap with the activation values of shallow layers. They also found that the performance was significantly improved by using the post-norm operation, and to further stabilize the training of the largest Swin V2, a layer normalization was added after every 6 transformer blocks.



2.2 Scaled cosine attention

The similarity between features in self-attention is measured using the inner product. The

authors observed that when replaced with a post-norm operation, the attention map in some blocks or heads would be dominated by some features in a large model. To improve this problem, the authors replaced it with cosine similarity, since the result of the cosine function is containing normalization.

$$Sim(q_i, k_j) = cos(q_i, k_j)/\gamma + B_{ij}$$

2.3 Log-spaced contiguous position bias

When scaling the model Windows-size, the performance drop is very severe. The authors speculate that the reason for this may be that the relative position encoding used by the Swin-Transformer is weakly generalized to the model scale. Therefore, one of the first improvements proposed by the authors is to replace the original set of learnable relative position parameters defined in Swin-Transformer with a smaller network. But when the Windows-size increases, the predicted target space, which is a linear space, increases considerably as well. By improving the predicted relative position coordinates from linear space to log space, the range of relative position coordinates that need to be predicted when the windows-size is expanded is greatly reduced compared to the previous linear space. The smoother prediction range will also increase the stability of the training and improve the generalization ability. With CPB and log-spaced CPB, the performance when expanding the model size is significantly improved.

3 Experiment

The authors achieved SOTA on the ImageNet 22K, COCO, ADE20K, and Kinetics-400 datasets, and performed ablation experiments for analysis.

Backbone	post-norm	scaled cosine attention	ImageNet top-1 acc
Swin-T			81.5
	✓		81.6
	✓	✓	81.7
Swin-S			83.2
	✓		83.3
	✓	✓	83.6
Swin-B			83.6
	✓		83.8
	✓	✓	84.1

Table 6. Ablation on post-norm and cosine attention.

Backbone	L-CPB	ImageNet* W8, I256	ImageNet† W12, I384	ImageNet† W16, I512
SwinV2-S		83.7	81.8/84.5	79.4/84.9
	✓	83.7	84.1/84.8	82.9/85.4
SwinV2-B		84.1	82.9/85.0	81.0/85.3
	✓	84.2	84.5/85.1	83.8/85.6

4 Conclution

The pre-trained model Swin Transformer V2 has received a further performance boost compared to Swin Transformer. However, some of the less elegant designs in Swin Transformer have not been further improved. Transformer is applied to CV from NLP, perhaps in combination with CNN-based models, such as VGG, to potentially improve performance further. I recently heard that the large convolutional kernel model proposed by Tsinghua surpassed Swin, so perhaps I can continue reading and thinking about it in the future.

References

[1] Ze Liu and Han Hu and Yutong Lin and Zhuliang Yao and Zhenda Xie and Yixuan Wei and Jia Ning and Yue Cao and Zheng Zhang and Li Dong and Furu Wei and Baining Guo. *Swin Transformer V2: Scaling Up Capacity and Resolution*, 2021.