A Method Applies Transformer Globally and Convolution Locally

Hongxu Chen, USTC, HeFei, China

Abstract:

Transformer[1] is a neural network based on self-attention mechanism, which is widely used in computer vision and other fields. Due to its powerful representation ability and flexibility, Transformer has become one of the most popular neural networks currently. But as transformers lack some inductive biases inherent to CNN such as translation equivariance and locality, they do not generalize well when trained on insufficient amounts of data[2]. In addition, if the input sequence is long, it will cause a huge amount of computation in the attention mechanism. In order to get rid of those impacts, I propose a method that applies transformer globally and convolution to locally, that is mainly change the outer transformer block to Resnet[3] block while the inner transformer block still applies the Multi-attention mechanism. At the same time, in order to improve the model performance without changing the model structure, knowledge distillation methods will be applied to the training process.

Introduction:

In computer vision tasks, convolutional neural networks (CNNs) have always played an important role, but the emergence of transformers has begun to challenge the dominance of CNNs. The emergence of transformer-based models such as ViT[4], DeiT[5], and TNT[6] has provided superior performance options for downstream computer vision tasks. Unlike traditional convolutional neural networks, ViT uses self-attention mechanisms to capture global information in images, rather than using convolutional operations. This enables ViT to have stronger representation ability when processing complex and large-scale image data. To alleviate the lack of the attention on locality in ViT model, TNT was proposed, which also applies the transformer block to sub-patches. Based on ViT model as the backbone model, DeiT-B achieved top-1 accuracy 84.40% at that time with the help of a token-based distillation.

The lack of inductive biases inherent in pure-transformer models is a serious problem, pure-transformer models always need huge amounts of data to generalize well. What's more, the multi-attention mechanism's computational cost is extremely large when the input sequence is very long. How to design a model to deal with those problems is an extremely important topic.

Literature Review

"***An image is worth 16x16 words: Transformers for image recognition at scale***" **(ViT):**
Vision Transformer (ViT) is a completely transformer-based architecture for image classification. Unlike traditional CNN-based models, ViT relies solely on the transformer's attention mechanism to capture global information in images. This approach allows ViT to effectively process large images without the need for sliding windows or convolution operations.

The main idea behind ViT is to treat an image as a sequence of patches and apply the transformer directly on these patches. Each patch is then flattened into a fixed-size vector, and the entire sequence of vectors forms the input to the transformer. The transformer's self-attention mechanism allows it to capture long-range dependencies between patches, enabling it to learn more complex and global patterns in the image.

“*Transformer in Transformer*” (TNT):
TNT divides the patch into a number of sub-patches and introduces a novel transformer-in-transformer architecture which utilizes an inner transformer block to model the relationship between sub-patches and an outer transformer block for patch-level information exchange.

***“Training data-efficient image transformers & distillation through attention”* (DeiT):**
This paper introduces a teacher-student strategy specific to transformers. It relies on a distillation token ensuring that the student learns from the teacher through attention. The paper shows the interest of this token-based distillation, especially when using a convnet as a teacher.

Research Question and Objectives:

TNT divides the patch into a number of sub-patches and introduces a novel transformer-in-transformer architecture which utilizes an inner transformer block to model the relationship between sub-patches and an outer transformer block for patch-level information exchange. Its extra attention on sub-patches is the main reason that makes it get top-1 accurate score, 81.5%, at that time. But attention computation takes a huge amount of cost, and due to the lack of inductive biases inherent, pure-transformer models need sufficient amounts of data to generalize well. What if we apply the transformer model to patch-level information while CNNs model (Resnet e.g.) applies for dealing with sub-patches. In theory, compared to the pure-transformer model, this model will have more inductive biases inherent, while also alleviating the huge computational cost. My goal is to achieve higher accuracy compared to the TNT model at the same computational scale, and have a great performance to apply our model to downstream tasks.

Research Methodology:

Following contents are some methods that will be applied to my research.

Ablation Experiment: We can explore the model’s performance through an ablation experiment. For instance, we can change the form of the position code to observe the performance of this model.

Knowledge Distillation: Applying knowledge distillation method to get better performance needs us to choose a proper teacher model. A good teacher model ought to have a strong generalized ability. Choosing a proper teacher model can make our model have a better performance.

Fine-turning: Fine-tuning is necessary to train deep networks and reduce the time for model convergence. Through fine-tuning, we can leverage useful features learned in pre-trained models and combine these features with data for specific tasks to quickly adapt to new tasks and improve model performance. This approach can save a lot of time and computational resources and allow us to train efficient deep learning models on limited datasets.

Data Sets:

Plan to use ImageNet, CIFAR10, CIFAR100 and Pets in my experiments.

Expected Experiments Results:

More inductive biases inherent: Because we exchange the inner transformer block to CNNs block, it will keep the attention on locality while having more inductive biases inherent than pure-transformer block.

Relatively short training time: To train a model, Transformer block takes more computational cost than CNNs block. In our model, there's only the outer block that applies the transformer. Thus, relatively short training time can be achieved.

More lightly weight model: Applying knowledge distillation method, we can add a token-based distillation like DeiT model does. Trained student model is always lighter weight.

Higher accuracy: Through above techniques, There is a great possibility that we can achieve better performance than TNT and other models on certain tasks.

References

【1】 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

【2】 Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

【3】 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

【4】 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

【5】 Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. PMLR, 2021: 10347-10357.

【6】 Han K, Xiao A, Wu E, et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.