Discussion Quiz #9

Ashutosh Tiwari (ashutiwa@iu.edu)

April 4, 2022

1 Papers

1.1 Paper 1

CvT

1.2 Paper 2

LocalViT

2 Questions

2.1 Describe the architecture of the Convolutional Transformer Block shown in Fig. 2 (b). (3 pts)

Convolutional Transformer Block is an attempt to introduce local receptive fields and shared weights in transformer architecture. A Convolutional Block is collection of Convolutional Projection followed by multi-head self attention which is followed by layer normalization. Each block ends with a set of fully connected layers called MLP. This block is repeated several times in the transformer architecture.

2.2 True or False. (3 pts)

2.2.1 CvT uses positional encoding like all other Transformer based models.

False, authors point out that they didn't notice any significant improvement with positional embeddings and thus decided to remove them.

2.2.2 CvT uses MLP for projection for attention.

False, CvT uses a depth wise convolutional operation for convolutional projetion, which is fed to attention block.

2.2.3 Just like ViT, CvT also uses non-overlapping token embedding.

No, CvT uses overlapping token embeddings.

2.3 CvT architecture design introduces convolutions to two core sections of the ViT architecture. Describe in detail these two core section modifications. (4 pts)

The two core sections where CvT architecture introduces convolutions are:

1. Beginning of each stage has a convolutional token embedding that performs an overlapping convolutional operation. This allows model to capture spatially local information and achieve spatial downsampling while increasing the number of feature maps.

2. Secondly linear projection is replaced by convolutional projection. This allows both decrease in computational complexity as well as capturing of local spatial context and reduce attention ambiguity.

2.4 Describe briefly. (4 pts)

2.4.1 Squeezed convolutional projection

A squeezed convolutional projection is the process of scaling an input channel to a single numerical value. In the paper we are discussing it is implemented as 1×1 convolutional layers through which input is passed then allowed to expand. However in general, it can take many forms. In some papers, a different value less than the previous layer is used and in some cases an Average pooling is used.

2.4.2 Why the original position-wise linear projection for multi-head self-attention is replaced with depth-wise separable convolutions in the convolutional projection layer.

This allows both decrease in computational complexity as well as capturing of local spatial context and reduce attention ambiguity. The third benefit is a simplified design at almost no additional cost.

2.4.3 Global context fusion

Global context fusion is the fusion of global and local features to achieve greater performance. This can be done in many ways. In the paper we are discussing it is done by taking the advantage of pyramid structures. In several other papers this is done by concatenating global and local context together.

2.4.4 Dynamic attention

Dynamic attention is a mechanism that allows the model to learn to focus on the most relevant parts of the input. This is done by using a softmax function to calculate the attention weights.

2.5 Describe the two efficiency benefits from the design of the Convolutional Projection layer. (3 pts)

There are two goals of the Convolutional Projection layer.

1. adds to accurately modelling local spatial context.
2. provide a more efficient way of computing attention by permitting the undersampling of attention matrices.

2.6 Paper 2 experimentally determines the effectiveness of the locality mechanisms. State at least four conclusions that are drawn from these experiments. (2 pts)

The four conclusion authors mention are:

1. Depth wise convolutional projection is more efficient than position-wise linear projection and can alone give a major boost in the performance of the network.
2. Activation function after the depth wise convolution is very important in determining the quality of results.
3. This locality mechanism used by paper is more important for lower layers.
4. Expanding hidden dimension of network results in higher classification accuracy.

2.7 Explain how depth-wise convolution is an efficient way to introduce locality into networks? (3 pts)

In depth wise convolution same convolution filter is applied to all the channels of the input. This is done to reduce the number of parameters and increase the computational efficiency. This strategy can provide precisely the mechanism to capture local context accurately.

2.8 Why do we need to split and concatenate the class token for every transformer layer in paper 2? (3 pts)

Since output of 2D depth convolution cannot be used as a native input to the next layer, we need to split and concatenate the class token for every transformer layer. This is done to preserve the locality of the input.

2.9 Explain briefly, why introducing the locality to the lower layers is more advantageous compared with higher layers. (3 pts)

Paper proposes that the locality mechanism is more advantageous for lower layers. They suggest that as locality meves to higher layers from lower layers, the accuracy of the network decreases. Intuition behind this is that if depth wise convolution is used in lower layers, then the locality information can be transported to higher layers and thus they can be more accurate. This is also corraborated by the their experiments.

2.10 Answer briefly: (3 pts)

2.10.1 What is depth-wise convolution?

In depth wise convolution same convolution filter is applied to all the channels of the input. In it we use we use eah filter channel for just one input channel.

2.10.2 Locality is an intrinsic property of CNNs, Explain.

The most important property of CNNs is locality. The locality of a feature is the extent to which it is localized in the input. Convolution operation is a way to capture the locality of the input which is controlled by filter size and stride.

2.10.3 What are inverted residual blocks?

An inverted residual block is a residual block which has an inverted structure. In most cases that is done for computational efficiency. It starts and ends with a convolutional 1×1 layer and in between it has a depth wise convolutional 3×3 layer.