Computer Vision_DiscussionQuiz_10

1. The process of generalizing the network's dot product output is known as pairwise self-attention. This operator behaves in a similar way to the set operator. The pairwise operator is essentially a set operator, and it differs from another sequence operators in this regard. The Hadamard operator is the spatial value of the feature space that is supplied to the function in equation 2 from paper 1. The beta function generates the input vectors' aggregated weights. The weights for the altered features are computed using the alpha function. The outcomes of summation, subtraction, concatenation, Hadamard product, and dot product are obtained without changing the transformer functions' dimensionality.

2. Because self-attention is independent of the position of the feature in the input data, positional encoding is utilized in paper 1. The following are the steps involved in position encoding: In [-1,1], normalize the horizontal and vertical feature space. For each trainable layer, adjust the 2d coordinates accordingly. Each feature map's outputs are linearly mapped, and a positional feature is added to each place I in the feature map. We encode the relative location of each mapping pair as the difference of the spatial position of the base reference point. Prior to the insertion of feature difference sets, the result is enhanced by concatenation.

3. Self-attention performed over a patch of feature vectors in the footprint under examination is known as patchwise self-attention. Input of feature vectors xr of the footprint is shown in the patchwise self-attention equation. The tensor of the patch in the same dimension is the patch's alpha output. The position of this tensor will be j, which is the alpha function's output. The weight vector is constructed using a beta function that is identical to that of paired self-attention and is Hadamard producted with the alpha result.

4. In comparison to the vision transformer model (ViT), an architectural modification is introduced by replacing attention layer by feed-forward neural network. The author's premise is that transformer models trained by substituting attention with a feed-forward neural network performed similarly to a ViT transformer with similar hyper-parameters. As a result, the author contends that the excellent performance of transformers on vision tasks, which is commonly attributed to self-attention blocks, may be due to other factors such as the inductive bias provided by patch embedding and augmentation techniques.

5. Transformers' attention mechanism offers major advantages over CNN's convolution process. The following are the details.
    i. Unlike the fixed weight convolution kernel, the spatial location interacts in the self-attention mechanism depending on their distinct properties.

ii. Transformers try to create a worldwide receptive field. They attempt to capture both long and short-range dependencies. As opposed to convolution, which is limited to local reception.

6. A relatively small number of channels are moved in all four top, down, left, and right directions during the shift operation for an input tensor. The shift block has three operations shift, layer normalization, and MLP network. The first layer raises the input characteristics to a greater value, while the second layer reduces it to its original size. The overhead pixels are lost after the shift. The output layer is normalized and transferred to the MLP network with two linear layers after shifting.

7. K x K convolution is approximated by the shift operation as shift operation consists of two 1x1 convolutions stacked together with a shift operation in between, essentially simulating a KxK convolution action. The 1x1 convolution is followed by a K-shift and another 1x1, effectively doing a KxK convolution.

8. In 3rd paper, the shift block executes a half shift operation. A relatively small number of channels are moved in all four top, down, left, and right directions during the shift operation for an input tensor. The overhead pixels are lost after the shift. In the paper, the shift step is set to 1.

9.
   a. Changing the expand ratio of MLP: The MLP expand ratio is a parameter that controls the network's depth. It is set to 2 by default. It is discovered in the paper that deeper models, i.e., those with a smaller expand ratio would function better.

   b. Percentage of shifted channels: It's the proportion of channels in the input tensor in the shifting block that the shift operation is performed on. By default, it is set to 33 percent. Based on testing, it was determined that this hyper parameter had a negligible impact on the final model outcome.

   c. Shifted pixels: There is no shift if the shifted pixel is equal to 1, hence the Top-1 accuracy on the Image Net is rather poor because when there is no shift, there is no interaction between the spatial characteristics.

10.
   a. Rotation invariance: The term "rotation invariance" refers to the fact that the rotation of the picture in question has no impact on its categorization or detection. The CNN's rotation invariance is a flaw since it prevents the model from detecting the picture as it rotates. The CNNs in this example are not rotation invariant.

b. Permutation invariant: The ability of a neural network to perform well even when the features are permuted before training is known as permutation invariance. There is no assumption of a spatial link between the characteristics. Because one of the essential assumptions of convolutional neural networks is the existence of significant collinearity between nearby neurons, they are not invariant to permutation.

c. Cardinality invariant: It is cardinality invariant, which means it is unaffected by the size and dimensions of the input vector. CNNs are cardinality variable and hence reliant on the size of the input vector.

Note: https://pyimagesearch.com/2021/05/14/are-cnns-invariant-to-translation-rotation-and-scaling/

Papers:

Exploring self-attention for image recognition

Do you even need Attention? A stack of feed-forward layers does surprisingly well on ImageNet

When Shift Operation Meets Vision Transformer: An Extremely Simple Alternative to Attention Mechanism (partial shift operation)