Computer Vision_DiscussionQuiz_07

1. Architectural similarities between encoder and decoder are, in encoder and decoder modules, residual links are established in every sub layer, succeeded by the normalizing layer that helps in rationalizing the output. And there are six identical layers in both the encoder and decoder units.
   Differences between encoder and decoder modules are, unlike in encoder module, across the encoding module stack decoder module conducts multi-head self-attention. This self-attention layer in the decoder module has been adjusted to prevent prior layer locations from propagating to succeeding layers.

2. Because the attention function's inputs are all vectors, it's a mapping query with a key value pair. The attention takes into account the weights from the compatibility function that is generated as an output.

3. To keep the gradients from being too tiny, the author employs scaling after dot-product attention. We obtain very little values when we evaluate the softmax of the query and key matrix multiplication. We divide the significant value of the matrix multiplication with the root of the dimension of the key to address this significant value of the matrix multiplication.

4. Three different ways the transformer uses multi-head attention are:

   a. The encoder could attend to every point from the output of the preceding layer using the self-attention layer.
   b. Decoders exercise self-attention in the same way as encoders do by attending all points from the decoder's output to the present layer.
   c. Any place in the decoder may go to every position in the encoder output, thanks to the transformer's architecture, that is similar to a classic encoder-decoder architecture.

5. Author used a sinusoid for positional encoding for the following reasons.
   a. Because throughout training, it enables to extend series lengths that seem to be larger than the original sequence of input vectors.
   b. Author utilized positioning encoding to convey the position of the input pattern as well as its absolute positions.

6. The following are the reasons why training a transformer model for classification on the ImageNet dataset underperforms state-of-the-art ResNet models.

   Because transformers doesn't possess inductive biases such as translational equivariance. When only a few photos are used to train it, this makes the model futile to adapt to this. Training that isn't well-structured as there's no perfect regularization.

7. Positional embeddings act as identifiers for the transformer, informing it of the sequence in which patches should be applied for the model to learn sequential information. Because positional encodings are used to save the locations of each pixel in feature space. Two-dimensional embedding was developed to cope with two-dimensional pictures. However, the vision transformer studies in article 2 show a 1-dimensional position embedding may learn to describe two-dimensional topology, implying that utilizing handmade 2-dimensional embedding for images may not provide a meaningful benefit.

8. As we travel deeper in a typical CNN, the receptive field grows. Starting layers identify low-level features, and as we progress deeper, high-level characteristics are detected. Lower levels have a shorter attention distance. It's in a little section of the image at first. As we travel further into the image, the attention distance becomes more important and affects a larger portion of the image. This is how, attention distance is similar to the receptive field size in CNNs.

9.
    a. Auto-regressive model: It is a model that utilizes before created data as an input to create new information.

    b. Positional encoding: In order to grasp the sequence's semantics, it's critical to maintain track of the absolute locations of adjacent elements. So, we use positional encodings to encode the elements, to save sequential location data.

    c. Patch embedding: A two-dimensional picture is compressed to a series of updates and fed into the transformer as an input. With a trainable linear projection, these patches are assigned to a fixed latent vector utilized by the transformer. Patch embedding is the result of this effort.

    d. Inductive bias: Some estimates are taken in NN architecture in order to anticipate a result while new, unknown data is presented as an input. Inductive bias is caused by such assumptions. In CNNs, for example, there is indeed a positional inductive bias, which implies CNNs presume the data has a specific spatial structure. Inductive bias is not present in transformers.

Note:
https://en.wikipedia.org/wiki/Inductive_bias#:~:text=The%20inductive%20bias%20(also%20known,predict%20a%20certain%20target%20output

http://jalammar.github.io/illustrated-transformer/

Papers - An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Attention is all you need