Discussion Quiz #7

Ashutosh Tiwari (ashutiwa@iu.edu)

March 18, 2022

1 Papers

1.1 Paper 1

Transformers

1.2 Paper 2

ViT

2 Questions

2.1 What is the architectural similarities and difference between encoder and decoder modules of the transformer model as described in paper 1? (3pts)

They have following similarities,

1. Both encoder and decoder are composed of a stack of 6 identical layers
2. Authors employ residual connections across each sub-layer and then a layer normalization in each of those cases.

2.2 Define Attention as per paper 1. Describe briefly the two types of attention discussed in paper 1. (3pts)

An attention function is similar to a mapping between a query and a set of key value pairs to an output, where each of these are tensors. This output is a weighted sum of values computed by a compatibility function of the query with the corresponding key.

Paper describes two types of attention mechanisms which differ in the way they compute the output:

1. **Scaled Dot Product Attention**

   In this case input is a query and key is a set of keys. The output is the dot product of the query with the key and then scale it by a parameter $\sqrt{d_k}$.

2. **Multi-Head Attention**

   This is different from the previous case in the sense that the query is split into multiple heads using learnt linear projections. The output is the weighted sum of the heads.

2.3 Discuss the reasons presented by the author for applying scaling in dot-product attention. (3pts)

Authors talk about two attention functions, dot product and additive attention. Because authors find dot product method much faster and efficient, they decide to use it. They use scaling for dot product attention because they suspect that large values of dot product push softmax into very high regions where values do not change and therefore additive attentions yeilds better results. To fix this they use a parameter $\frac{1}{\sqrt{d_k}}$ to scale the dot product value which is the square root of the dimension of the key.

2.4 State the three different ways the transformer uses multi-head attention? (3pts)

Transformer uses multi-head attention in following three different ways:

1. In encoder decoder attention queries come from previous decoder layer and keys, values come from encoder layer. This is helpful to attend over all possible previous decoder states.
2. Encoder contains self attention layers where each layer gets keys, values and queries from same place, which is output of previous layer in encoder.
3. Very similar to encoder self attention, decoder contains self attention layers where each layer gets keys, values and queries from same place, which is output of previous layer in decoder upto that position.

2.5 Why does the author use a sinusoid for positional encoding. Give at least two reasons. (4pts)

Authors talk about two ways of encoding positional information. One is learned and the other is fixed. Authors did try to learn the positional information but results were same as fixed. They describe two reasons why they use sinusoid.

1. Because sinusoid is a periodic function, it is easy to learn as it is a linear function, i.e. PE_{pos+k} is a linear function of PE_{pos}.
2. Because it is periodic it allows to incorporate positional encoding of sequences of lengths longer than encountered during training.

2.6 Training a transformer model for classification task on ImageNet dataset underperforms compared to the state-of-the-art ResNet models, why? Explain the reasons. (3pts)

Authors suggest that when transformers are trained on small datasets like Imagenet, they are not able to generalize well. They suggest that this is the case because when compared to CNNs, they lack the ability to have inductive biases related to locality of features and translation invariance.

However they also suggest that when transformers are trained on large datasets, they perform better than CNNs.

2.7 What are positional embeddings as discussed in paper 2? The author prefers 1D positional embedding over 2D positional embedding for images. What is the argument of the author in paper 2 for not using hand-crafted 2D-aware embedding for images.? (4pts)

Since this transformer does not have CNNs, authors flatten the image patches and feed then to the transformer. Further in the process position embeddings which are learned, are used. Authors prefer 1D positional embedding over 2D positional embedding for images, since they did not find any significant difference in performance. These are fed into the encoder. Authors point out that the probable reason for this might be the fact that a row column structure appears which inspecting the image patches in 2D space, which is much more revealing in larger grids. This means that same row and column in image have similar embeddings.

2.8 To check how the transformer model uses self-attention capabilities, they compute average distance in image space based on attention weights – attention distance. Explain how 'attention distance' is analogous to receptive field size in CNNs. (5pts)

In order to verify their claim that self attention allows ViT to integrate information across image or in other words, to act similar to convolution they compute average distance in image space based on attention weights.

This is analogous to receptive field size in CNNs because self attention in every layer can learn pair of representations in every layer. This essentially means that we don't need convolutional layers to have an increased receptive field size. In some sense this is also better, since in case of convolutional layers, this increase is linear, while it can be done more efficiently in self attention. In their experiments authors found that there is a proportional relationship between network depth and attention distance.

2.9 Describe briefly. (a) Auto-regressive model (paper 1), (b) Positional encoding (paper 1), (c) patch embedding (paper 2), (d) Inductive bias (paper 2) (2pts)

1. **Auto-regressive model**
   Auto-regressive models are models that are trained to predict the next element in sequence. They are trained to predict the next element in sequence. In this paper, authors compare their encoder-decoder architecture to an auto-regressive model, since decoder uses all previous elements in sequence to make predictions.

2. **Positional encoding**
   Positional encoding is a way of encoding the position of elements in sequence. Authors in this paper use sinusoid positional encoding, which is a linear function of the position of the element in sequence.

3. **Patch embedding**
   Authors flatten out the image patches and feed them to the transformer. They match them to D dimensions by using a traninable linear projection layer. Output of this layer is called patch embedding by authors.

4. **Inductive bias**
   Inductive biases are relations or beliefs that are learnt by models (mostly CNNs) because of their ability to learn invariant representations of the input. Authors point out this fact to assert their point that when image transformers are trained on small datasets, they are not able to generalize well. They suggest that this is the case because when compared to CNNs, they lack the ability to have inductive biases related to locality of features and translation invariance.

References