Paper: *Image Transformer*

Razwan Ahmed Tanvir

Date: 03/20/2022

Quote *"... For both the encoder and decoder, the Image Transformer uses stacks of self-attention and position-wise feed-forward layers, similar to (Vaswani et al., 2017). In addition, the decoder uses an attention mechanism to consume the encoder representation"*

Overview Generation of images is considered to be a problem of auto regressive sequence generation. This family of the problem is also known as transformation problem. The authors of this paper got the motivation from the architecture which is based on the self attention mechanism. This attention based architecture is known as the Transformer. The authors proposed a generalized self attention architecture known as transformer to a sequence modeling for the generation of images with tractable likelihood. The authors claimed to increase the capability of the model to deal with larger size of the images. Authors at first, mentioned about the drwabacks of Convolutional Neural Networks in comparison to Recurrent Neural Networks. They mentioned that CNNs have limited receptive layers which can negatively affect the occulation and symmetry of images. This phenomenon is frequently noticeable within a smaller number of layers. In this work, it has been shown that the self attention mechanism can find the optimal trade-off between the infinitely many receptie field and limited receptive field. The authors considered the pixel intensities as distinct categories and each of the three color channels of input pixels is encoded by a 256-d embedding vectors. The output is another 256 d- dimentional embedding. The generation of the images is done using an encoder-decoder model. Moreover, they have incorporated local self attention to render around 3072 positions. The authors used !d and 2D local attentions to partition the images.

The authors claimed that their proposed self attention based sequence modeling formulation outperforms the state-of-the-art image generation on the ImageNet. This model showed best results in negative log likelihood on ImageNet with values 3.83 to 3.77. Furthermore, they claimed that the generated image from their model fooled human observers three times which is an improvement from the previous state of the art.

Intellectual Merit Transformer models were proposed for sequence to sequence learning to process natural languages. However, this paper proposed a similar model to generate images and produced state-of-the-art output. This research carried their experiments with clear explanations and mathematical notations. The whole research was well organized in terms of writing and presentation. The authors used the state-of-the-art benchmark to compare the results.

Broader Impact This research presented an implication of the powerful transformer model to generate images utilizing the sequence to sequence learning. Authors of this paper are from google brain team. The code provided by the authors is available in tensor-to-tensor. Among the seven writers, one of them was women. Later three of the authors did their PhD.

Keywords Image Processing, Attention, Transformers

Discussion Questions

- It is due a discussion why the authors used the loss function (maximum likelihood) that they used.
- The authors did not clearly mentioned the training speed ups for image generation. I think there could be room for the speed up during training.

Table 1: Grade deductions by section

Overview	Intellectual M.	B. Impact	Keywords	Questions	Is Online?