Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution

August 2023

This paper aims at building a Vision Transformer model that does not rely on the need for resizing the input images to fixed size before being fed to the first layer. Traditionally, in order to feed images to deep learning models people used to resize the images to a given size, which would then allow more efficient compute through parallelization, this was designed with CNNs in mind. Another popular approach is padding the images to a fixed size in order to allow batching on GPUs, but this is highly inefficient. It is clear that the resizing is suboptimal because the model is biased into seeing only the fixed size images, which sometimes incurs distortions, blur, aspect ratios not being represented correctly, ... etc while the original images content is highly related to its resolution ; resizing a medical image of a tumor might cause the tumor to become invisible in the resulting image for example. But since transformer based models can take varying-length sequences as their input, then a Vision Transformer should be able to process images in their native resolutions. The authors of the paper, propose NaViT, a novel of approach that consists in grouping multiple patches from different images at the same time during training into a single sequence, this approach is termed Patch n' Pack. Some more technical modifications need to be made : mainly the final sequence that contains many examples needs to be padded to fixed length and from a single sequence we need to be able to extract a feature vector per example and not only one single vector per sequence. This requires modifying how the loss is computed and how pooling feature representations is done for a single sequence.

In order to make this work, the authors had to make some changes on top of the original ViT architecture : 1- Masked self attention and masked pooling which consists in adding additional self-attention masks and masked pooling on top of the encoder are done for each example separately in order to prevent examples from attending to each other. 2-Factorized and fractional positional embeddings, which are a new positional embedding aimed at handling arbitrary resolutions, they achieve this through decomposing the positional embedding into two different embeddings for x and y coordinates then they are summed to obtain the final embedding.

This new method allows more efficient training by allowing continuous token dropping and resolution sampling. The first one consists in dropping tokens

from the examples present in a batch, which results in faster training and acts as a regularization technique by making the model aware of the whole image and not just some specific patch or token. Another interesting idea is being able to modify the token dropping distribution during training using some fixed schedule (like it's done for learning rate, for example). The second method consists in being able to expose the model to a variable range of resolutions during pre-training, unlike what is done classically where the pre-training is done at a small resolution to allow higher throughput and the fine-tuning is done at a much higher resolution which creates discrepancies.

Finally, all these modifications make up for a more natural variant of image processing that does not require resizing and results in better performance compared to the classical approach. One downside of this method could be the cost of the self-attention on the long sequence that is constituted from many examples, this could be an issue for smaller models but for bigger models the authors showed that the attention overhead from the packing becomes negligee as the model is scaled up.