Project Update

Kains Praveen
Nagamma Patil
18 September 2021

Dear Ma’am,

For researching on models to implement for multi-modal learning, I’ve found out that Attention mechanism is very important for multi-modal models. So for that I’ve done preliminary work implementing Attention models (Transformers), and tested on **CIFAR-100** Dataset. Due to computation limitations I chose to train on CIFAR-100 [**50,000 Training and 10,000 Testing Images of size: (32,32,3)**]. And after I’ve come up with a **new architecture (hybrid)** consisting of **progressive convolutions** in addition to transformer blocks. The performance of the novel model is higher than the **normal transformer and ResNet50** (without pre-training on bigger datasets like Imagenet 1k).

Validation Accuracy on Transformer is around (**51%**) but validation accuracy of novel hybrid model is around (**57%**) which is a high jump with small resolution image like **CIFAR-100**.

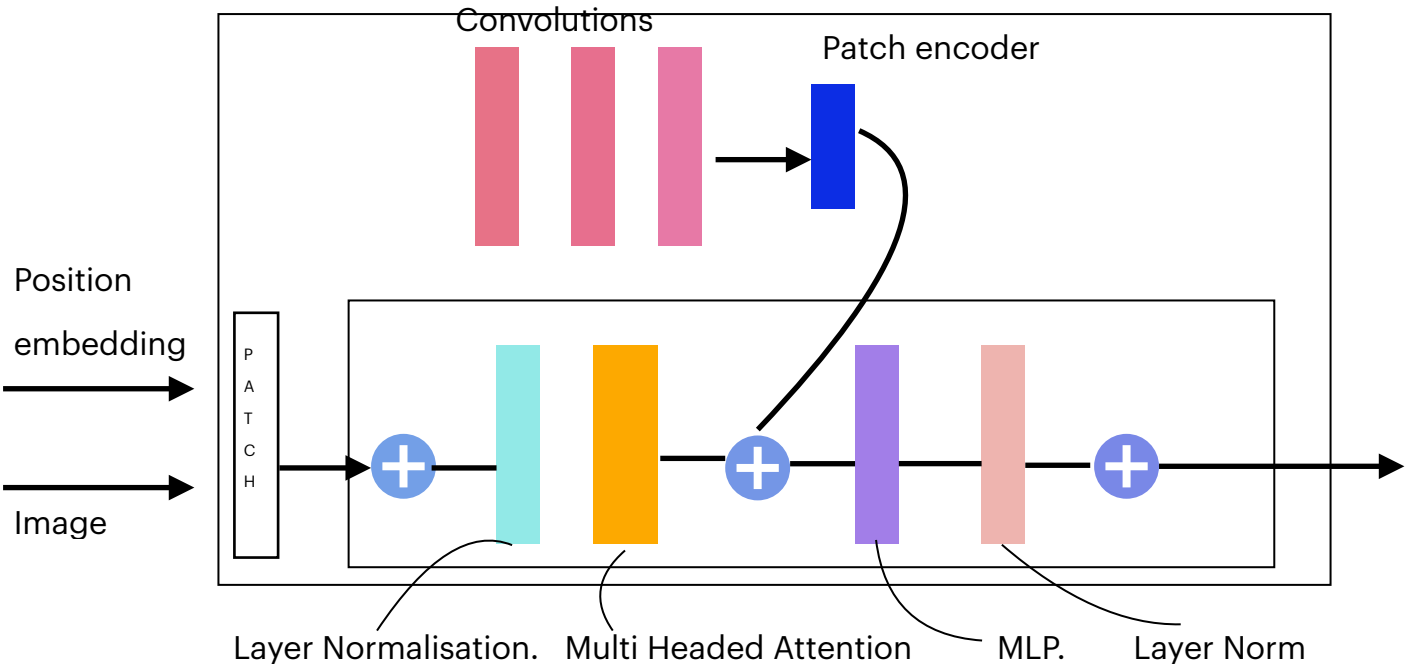
However I think **if we could pre-train the model on ImageNet1K** we could heavily boost the model performance further and maybe compare with **SOTA model baselines too**. But due to the lack of computational power at my end I was not able to do the **heavy pre-training the SOTA models do**.

I yet to do further experiments on the model, but I just want to update you on the progress.

[Next page has Model Overview Diagram.]

Thanks Ma’am

Kains Praveen


Convolutions
Patch encoder
Position
embedding
Image
PATCH
Layer Normalisation.
Multi Headed Attention
MLP.
Layer Norm