Deep Learning Project: Pay Attention to MLPs

Valsaraj, Indrajitt

f20181019@goa.bits-pilani.ac.in

Jain, Harsh

f20190065@goa.bits-pilani.ac.in

May 2022

Abstract

In this report, we take a look at the paper 'Pay attention to MLPs'. The paper proposes a simple network architecture, gMLP, based on MLPs with gating. The authors propose that self-attention is not critical for Vision Transformers as gMLP can achieve same or better accuracy on Image Classification Tasks. During our study of this paper, we implemented the gMLP model from scratch and compare it's performance with the Vision Transformer model. We train all models on the CIFAR-10 dataset. gMLP results in a test accuracy of 64% on the dataset while a ViT implemented from scratch gives us a test accuracy of only 61% on the dataset. However on adding a little bit of attention to the gMLP model, the test accuracy increases to 71%.

1. Introduction

The transformer architecture combines two important concepts:

1. A recurrent-free architecture which computes the representations for each individual token in parallel multi-head self attention.

2. Blocks which aggregate spatial information across tokens.

The gmLP model proposed in this paper[1] is a MLP-based alternative to Transformers without self-attention. It consists of spatial and channel projections with static parameterization.To investigate the necessity of the Transformer's self-attention mechanism, the team designed gMLP using only basic MLP layers combined with gating, then compared its performance on vision tasks to previous Transformer implementations.On the ImageNet image classification task, gMLP achieves an accuracy of 81.6, comparable to Vision Transformers (ViT) at 81.8, while using fewer parameters and FLOPs.

As a part of experimentation, we implemented both the gMLP model and the ViT model[2] and compared their performance on the CIFAR-10 dataset. We then performed an ablation to study to see the usefulness of self-attention in transformer models. The details are outlined in the next section.

2. Experiments and Results

We chose the CIFAR-10 dataset for the purpose of this Image Classification task. The CIFAR-10 dataset is a standard balanced dataset used for several image classification tasks like ImageNet which is used in the paper. The limitations of this dataset is that it consists of a lesser number of images that are of lower resolution than the images that were used in this paper. However it is an appropraite dataset for image classification taking into consideration the amount of compute resources that are available to us. We implemented the gMLP model first and observed that the model gave a test accuracy of 64%.

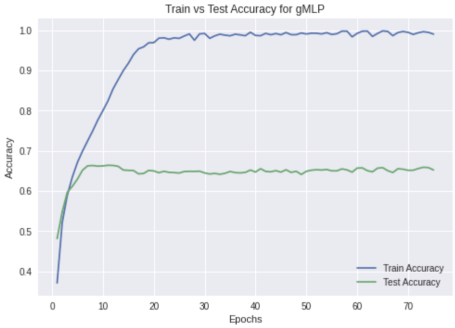

Figure 1: Train vs Test accuracy for gMLP on CIFAR-10 dataset

Next, we implemented the ViT model from scratch to compare its performance with the gMLP model. The model achieves an accuracy of around 60% on the same dataset. We can conclude that for smaller datasets and lesser, gMLP performs better.

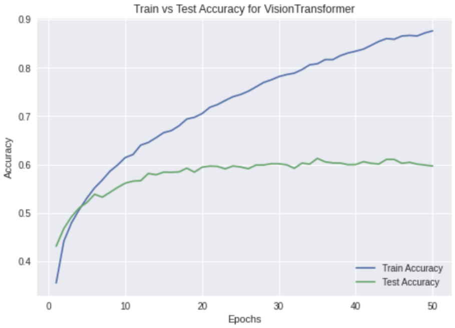

Figure 2: Train vs Test accuracy for ViT on CIFAR-10 dataset

Finally to perform an ablation study on the effectiveness of self-attention for Image Classification tasks we will implement gMLP with self attention.

To isolate the effect of self-attention, we experiment with a hybrid model where a tiny self-attention block is attached to the gating function of gMLP. gMLP itself is capable of capturing spatial relationships, the authors hypothesize that this extra self-attention module's presence is more relevant than its capacity. In our model we use a small attention head which has size 64. This is much lesser in comparision to transformer models like ViT which usually have 8 heads with larger size. The resulting model gives the 71% accuracy on the test set, the highest amongst all the models. The paper states that *With increased data and compute, models with simpler spatial interaction mechanisms such as gMLP can be as powerful as Transformers and the capacity allocated to self-attention can be either removed or substantially reduced.* The significant difference in performance can be attributed to the lack of data and compute and hence, self-attention improves performance.

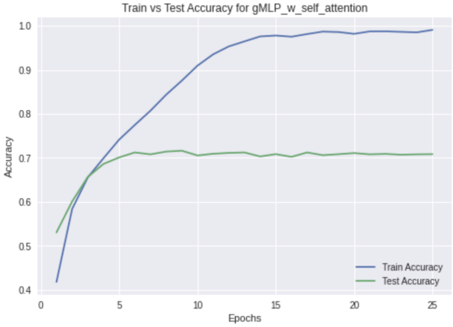

Figure 3: Train vs Test accuracy for aMLP on CIFAR-10 dataset

References

[1] Hanxiao Liu et al. “Pay Attention to MLPs”. In: *CoRR* abs/2105.08050 (2021). arXiv: 2105.08050. URL: https://arxiv.org/abs/2105.08050.

[2] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv preprint arXiv:2010.11929* (2020).