Intro Deep Learning Homework 6

Jaskin Kabir
Student Id: 801186717
GitHub:
https://github.com/jaskinkabir/Intro_Deep_Learning/tree/master/HM6

April 2025

1 Problem 1: ViT From Scratch

1.1 Model Architectures

This section focuses on training four variations of a vision transformer based network for image classification on the CIFAR-100 Dataset. The four models shared the same parameters other than the patch size and number of attention layers.
The shared parameters are as follows:

- Embedding Size: 192
- MLP Hidden Size: 384
- Number of Attention Heads: 4
- Classifier Head Hidden Layers: 384, 192

The four variations are combinations of a patch size of 4 or 8 and either 4 or 8 attention layers. The models were trained on the CIFAR-100 dataset with a batch size of 64, and learning rate of 5e-4. They were also compared to a baseline ResNet-18 pretrained on ImageNet. The Resnet-18 model was fine-tuned for CIFAR-100 for 10 epochs with a batch size of 64 and a learning rate of 1e-3.

1.2 Results

Model	Parameters	MACs	Avg Epoch Time	Accuracy
ViT Patch4 Attn4	1,377,508	756,480	11.76s	43.9%
ViT Patch4 Attn8	2,565,604	756,480	15.07s	44.6%
ViT Patch8 Attn4	1,395,940	756,480	10.58s	37.9%
ViT Patch8 Attn8	2,584,036	756,480	11.53s	37.7%
ResNet-18	11,227,812	1,816,096,768	30.07s	56.9%

Table 1: ViT Model Comparisons

The vision transformer models were soundly outperformed by the ResNet-18 model. This is due to a number of reasons. Firsly, the Resnet is much more complex in terms of parameters and MACs. The ResNet-18 model has 11 million parameters and 1.8 billion MACs, while the largest ViT model has only 2.5 million parameters and 756 million MACs. This means that the ResNet-18 model is able to learn more complex features from the data, which is important for image classification tasks. Additionally, the vision transformers were trained for 50 epochs on the CIFAR-100 dataset, while the ResNet-18 model was trained on the much larger ImageNet dataset. With their lack of inductive bias, the vision transformers were unable to learn the features of the dataset as well as the ResNet-18 model.

2 Problem 2: Pretrained Swin Transformers

Microsoft's Swin Transformer is a vision transformer model trained on the ImageNet dataset. For this experiment, the large and tiny models were loaded and fine-tuned for the CIFAR-100 dataset for 5 epochs. The training used a batch size of 32 and a