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

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 1024, and learning rate of 5e-4. They were also compared to a baseline ResNet-18 pretrained on ImageNet.

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