Experimentation with Vision Transformers

Q1)https://colab.research.google.com/drive/1E9CVIolljxRI0vpVmccflA58yWrXoaaQ?usp=sharing in this google Collaboration Notebook I initialized the pre-trained CNN model (ReSNet-18) with weights from ImageNet dataset and initialized the ViT model on 'google/vit-base-patch16-224' dataset. Later, I fine-tuned both models on CIFAR-10 subset using train_model function and then evaluated them on CIFAR-10 test loader to see their performance and then again evaluated the fine-tuned models on CIFAR-100 subset function.

Instead, I should have pre-trained both CNN and ViT models on the same dataset which would have been better as this allows both models to learn general visual features from the same dataset and then fine-tune both on the CIFAR-10 subset and then evaluate the fine-tuned model on the cifar-10 test loader.

And here I got good performance of the CNN model with respect to the ViT model as the dataset I fine-tuned was not large and the ViT model works well for large datasets as it is good at generalizing global features and the CNN model is better at finding local patterns.