Training and Validation results

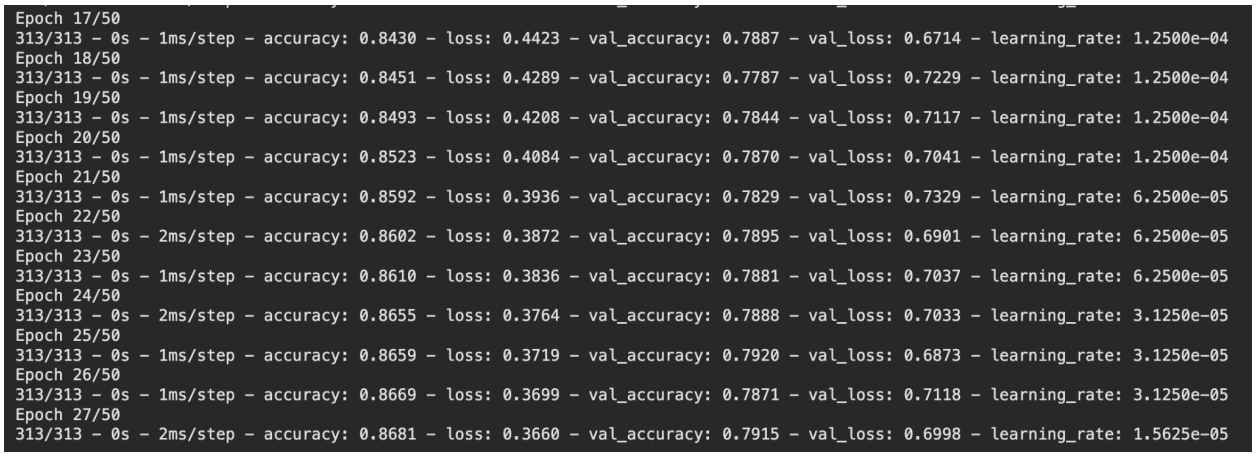

Figure 1: Last 10 epochs ran for CNN model

Referring to figure 1, we can observe that the final training accuracy and loss reached 0.8681 and 0.3660, and the final validation accuracy and loss reached 0.7915 and 0.6998.

With the CNN reaching the smallest validation loss of 0.6714 and validation accuracy of 0.7887 in epoch 17.

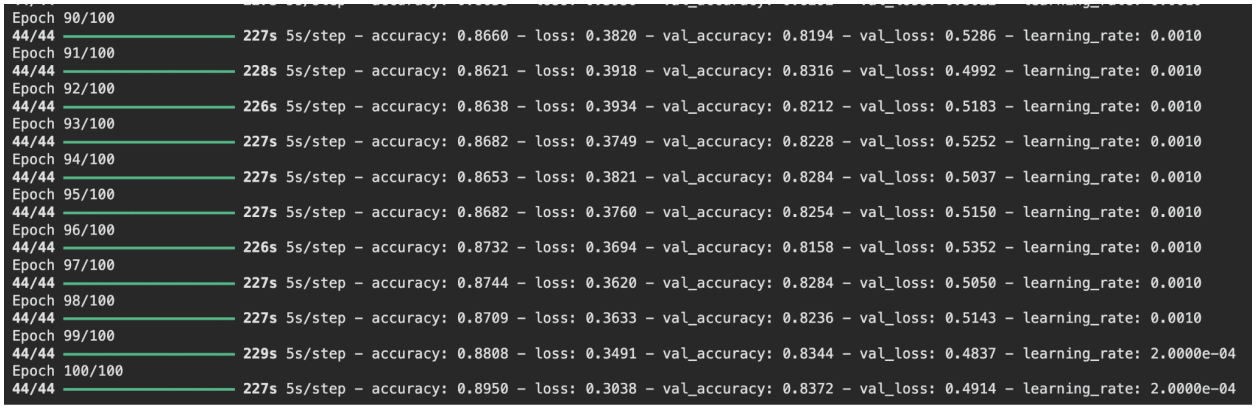

Figure 2: Last 10 epochs ran for ViT model

Referring to figure 2, we can observe that the final training accuracy and loss reached 0.8950 and 0.0.3038, and the final validation accuracy and loss reached 0.8372 and 0.4914.

With the ViT reaching the smallest validation loss of 0.4914 and validation accuracy of 0.8372 in epoch 100.

Performance comparison

The ViT performed better than the CNN both in terms of a moderately lower validation loss score and a slightly higher validation accuracy score.

Training efficiency and learning dynamics

The CNN was able to reach its peak and plateaued early on at epoch 17, and was stopped prematurely by the implemented early stopping procedure.

Whereas we can observe that even in its final few runs, the ViT displayed continued growth, and likely could have gained an even better validation accuracy and loss score, if allowed to run for more epochs. However, the ViT is limited by its much slower training time (average of 227s per epoch in ViT vs average of 300ms per epoch in CNN).

This is explained by the nature of CNN, utilising convolutional filters that while allowing the model to quickly learn local relationships between neighbouring pixels. This allows CNN to quickly reach its peak but struggle to see much gains in performance.

But in a ViT, utilising self attention and being able to extract global complex patterns (not just limited to local relationships) in images allows the ViT to outperform the CNN in the long run. But without the spatial assumption in ViT when it first performs patching, the ViT has to learn position embeddings of the image patches. This is observed in the training results where the ViT requires many epochs (16 epochs on ViT) to hit a validation accuracy of 0.70, while the CNN easily achieves that in 5 epochs.