ViT Sequential Tuning Results on CIFAR-10

Siddharth S

April 21, 2025

Tuning Summary

The following table summarizes the best configuration found at each stage of the sequential hyperparameter tuning process for a standard Vision Transformer (ViT) on the CIFAR-10 dataset. Each tuning run within a stage used 20 epochs. The overall best configuration from the tuning phase is used for a final 60-epoch training run.

Table 1: Best Results per Tuning Stage (20 Epochs/Run)

Stage	Goal	Best Config ID	Selected Parameter	Best Val Acc	Test Acc
1	Patch Size	`Stage1_Patch_4`	`patch_size=4`	0.7134	0.7106
2	Model Params	`Stage2_Model_wider`	`arch=wider`*	0.8030	0.7952
3	Data Augmentation	`Stage3_Aug_Mild`	`aug=Mild`**	0.8008	0.7994
4	Pos. Embedding	`Stage4_Pos_sinusoidal`	`PE=sinusoidal`	0.8154	0.8165
Overall Best Config from Tuning (Stage 4)				**0.8154**	**0.8165**

*Best architecture from Stage 2 (`wider`): `embed_dim=384`, `depth=8`, `num_heads=12`.

**Best augmentation from Stage 3 (`Mild`): `['random_crop', 'horizontal_flip']`.

Note: Test accuracy corresponds to the model weights achieving the best validation accuracy during that specific run.

Final Model Training

The overall best configuration identified during tuning (`Stage4_Pos_sinusoidal`) was then trained for a longer duration (60 epochs). The results of this final training run should be reported separately (e.g., from the `results_vit_concat_aug_final_train` directory outputs).