

Report on DLCV Assignment-2

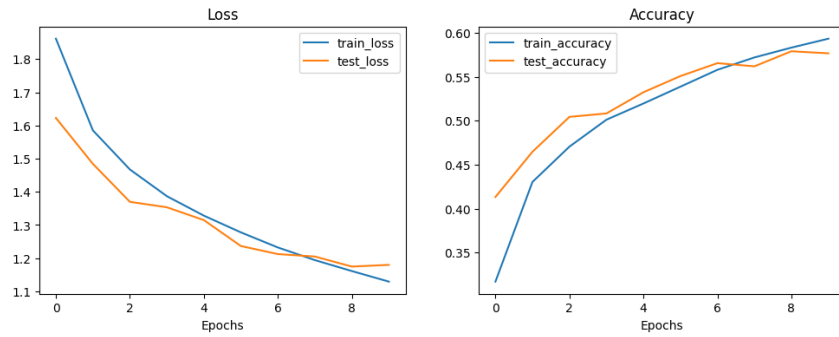
Raja Sekhar M (SR No: 600014806)

1 Experiment-1

ViT model is implemented and trained on CIFAR10 dataset for 10 epochs.

Hyperparameters:

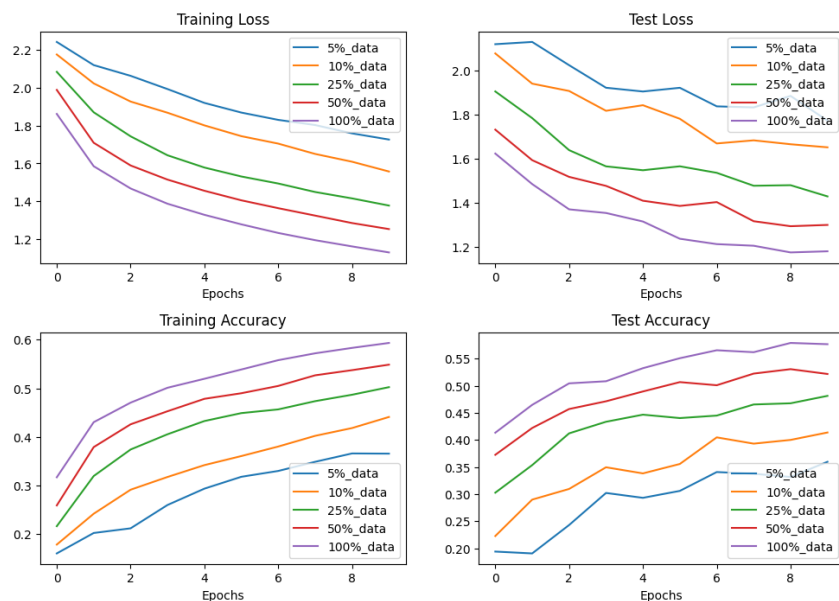
patch size = 4x4, transformer layers = 6, embedding dim = 64, attention heads = 4.



2 Experiment-2

Model is trained for 10 epochs with 5%, 10%, 25%, 50%, and 100% of training data while hyperparameters remain the same as in experiment 1.

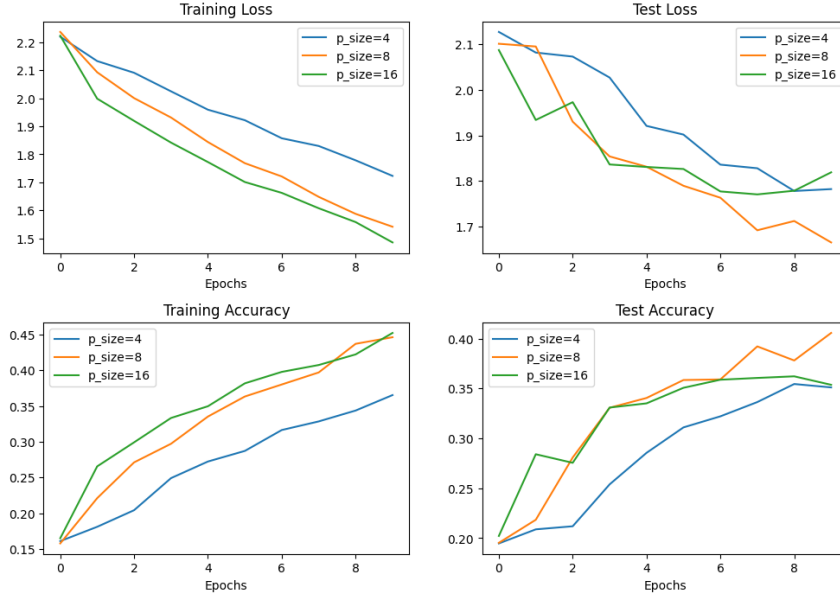
Observation: Performance of the model improved as the amount of training data increased.



3 Experiment-3

The model is trained for 10 epochs with different patch sizes - 4x4, 8x8, and 16x16. The other hyperparameters remain the same as in experiment 1.

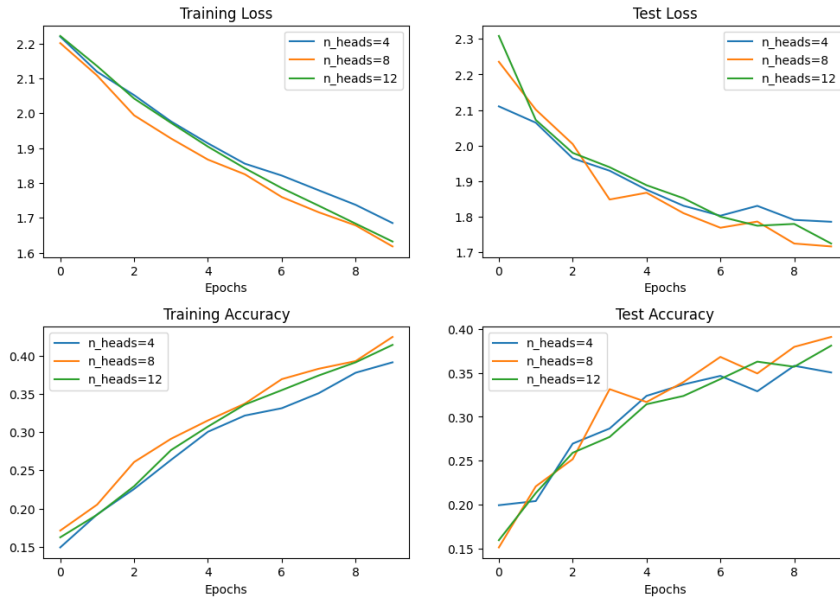
Observation: Model exhibits better test performance with patch size = 8x8.



4 Experiment-4

The model is trained for 10 epochs with different number of attention heads - 4, 8, and 12. Embedding dimension = 96. The other hyperparameters remain the same as in experiment 1.

Observation: Model exhibits better test performance with number of attention heads = 8.













5 Experiment-5

The classification is performed using CLS token from various transformer layers. The predictions for a randomly chosen 7 different test images are plotted below.

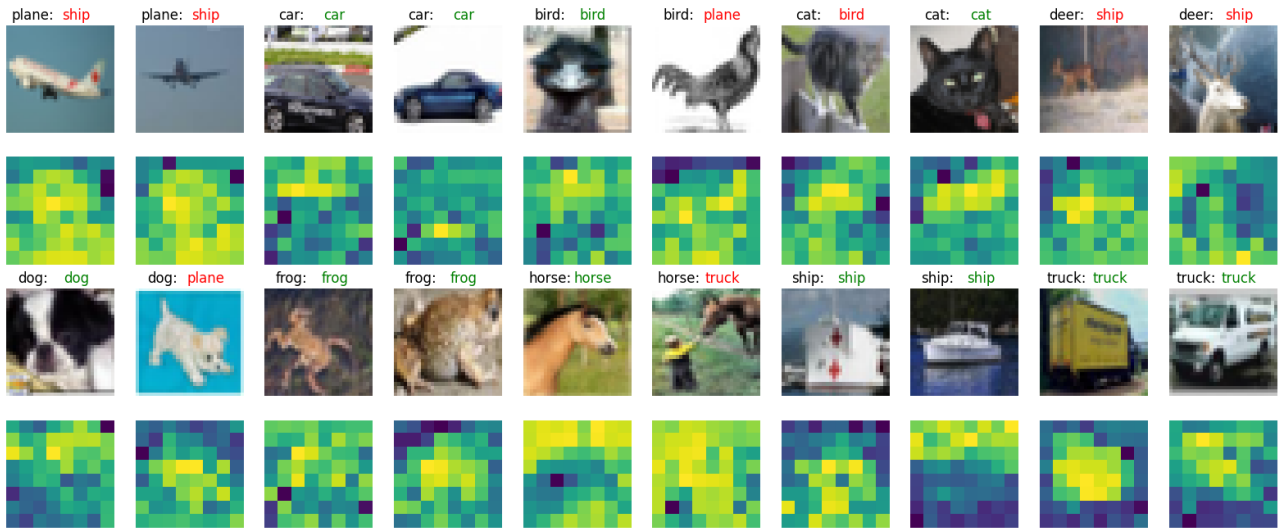
Observation: Predictions are usually more accurate for the CLS token from deeper transformer layers than from shallow layers. This can be explained by the fact that the deeper layers yield richer feature representations compared to the shallow layers.

Predictions using CLS token from each layer

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
 horse	plane	horse	horse	horse	horse	horse
 horse	plane	horse	horse	horse	horse	horse
 ship	plane	horse	car	truck	truck	truck
 truck	plane	ship	car	car	car	car
 truck	plane	plane	ship	car	car	car
 car	plane	car	car	car	car	car
 dog	plane	horse	horse	dog	dog	dog
 dog	plane	bird	cat	cat	dog	dog
 horse	plane	plane	horse	horse	horse	horse
 bird	plane	horse	horse	horse	horse	bird

6 Experiment-6

Two images per class are chosen at random from the test set, and their attention maps are visualized.



Attention matrices from multiple heads (4 heads) within a transformer layer are averaged, forming a single attention matrix per layer. Attention matrices from different transformer layers are combined through *attention rollout*.

Since the deeper layers do not directly attend to the image patches (only the first transformer layer does), attention rollout is used to combine and interpret the attention flow across layers of the model, highlighting which parts of the input image are influential in the for the output.