DS-265 DLCV 2025: Assignment 2

Kaniki Bhikshapathi - 24489

1 Experiment

Experimental Setup

The model configuration for Experiment 1 was as follows:

- **Input Image Size:** 32×32 (CIFAR-10)
- **Patch Size:** 4×4 (non-overlapping patches)
- **Embedding Dimension:** 64
- **Transformer Depth:** 6 layers
- **Number of Attention Heads:** 4
- **Optimization:** Adam optimizer with learning rate 1×10^{-3}
- **Loss Function:** Cross-Entropy Loss
- **Epochs:** 20

The CIFAR-10 dataset consists of 50,000 training images and 10,000 test images across 10 classes. No pre-training was used, and the model was trained from scratch.

Results

The final test accuracy achieved is **64.98%**.

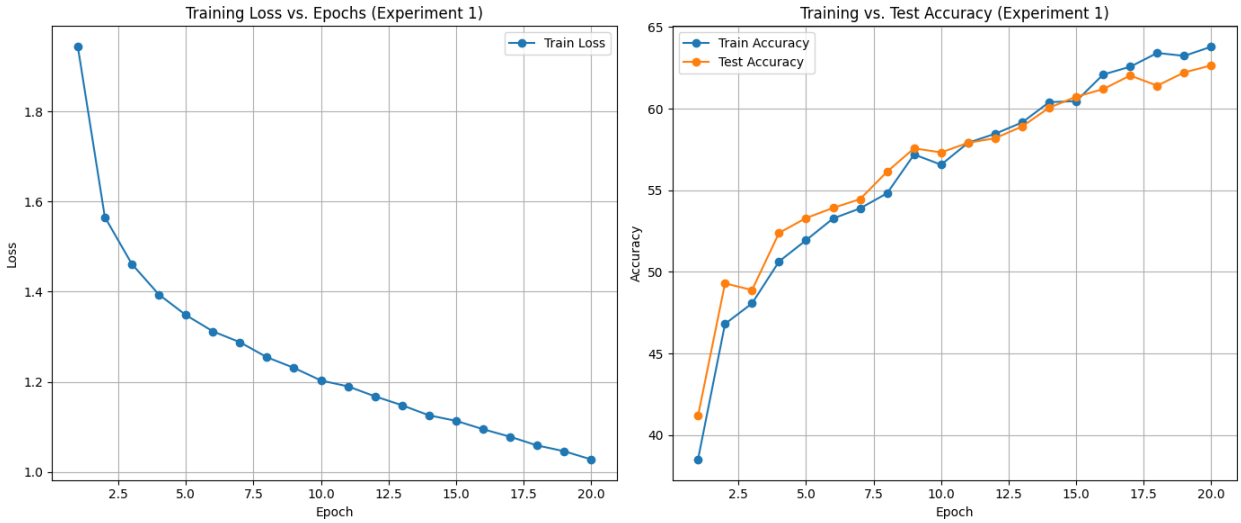


Figure 1: Training and Test Metrics over 20 Epochs for Experiment 1

Experiment 2: Data Fraction Analysis

Percentage of Data	Final Test Accuracy
5%	40.26%
10%	44.61%
25%	52.98%
50%	57.71%
100%	61.40%

Table 1: Final Test Accuracies at Different Data Percentages

Key Observations:

- **Data Dependent Model:** We see that Accuracy improves more significantly as more and more data is used for training.
- **Rapid initial improvement:** we can see a major improvement in Accuracy in the intial increase in the data .For 5% to 25% increase in data we can see at net of 12% increase in accuracy.
- **Diminishing Returns:** After reaching 50% of the accuracy we see that accuracy will increase very slowly.
- **Implication:** Extra techniques (e.g., augmentation or longer training) may help in increasing the accuracy

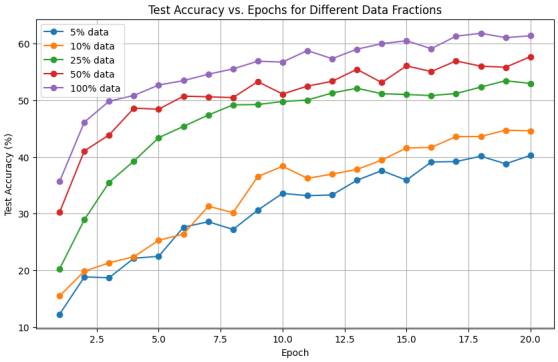


Figure 2: Test Accuracy for data fractions

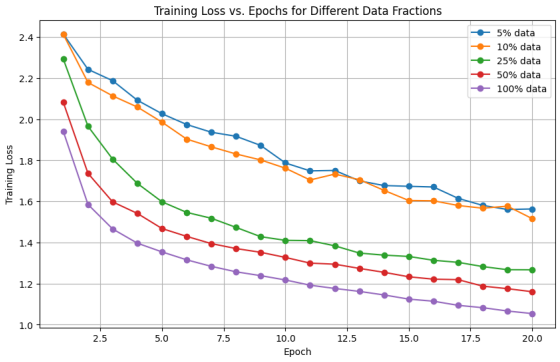


Figure 3: Train loss for different data fractions

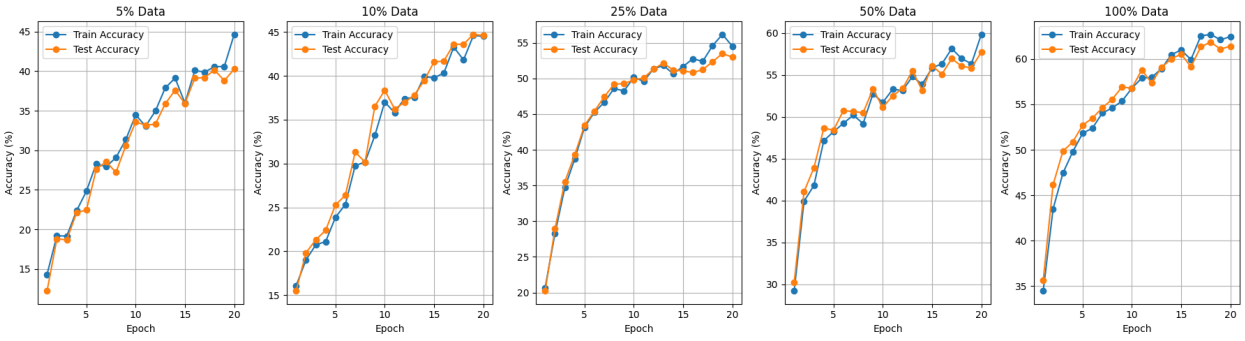


Figure 4: CHeck Overfitting Test and Train Accuracy

Experiment 3: Patch Size and Overlap Analysis

Configuration	Final Test Accuracy
4x4 non-overlapping	62.94%
4x4 with 50% overlap	68.46%
8x8 non-overlapping	58.76%
8x8 with 50% overlap	66.05%
16x16 non-overlapping	54.53%
16x16 with 50% overlap	60.05%

Table 2: Final Test Accuracies for Different Configurations

Key Observations:

- **Patch Granularity:** Finer patches (4x4) capture more local details, leading to higher accuracy compared to coarser patches (8x8 and 16x16).
- **Overlap Benefit:** Adding 50% overlap consistently improves performance across all patch sizes by providing richer contextual information and reducing boundary effects.
- **Optimal Setting:** The best performance is achieved with 4x4 patches with overlap (68.46%), indicating that detailed and redundant patch coverage is beneficial for classification.

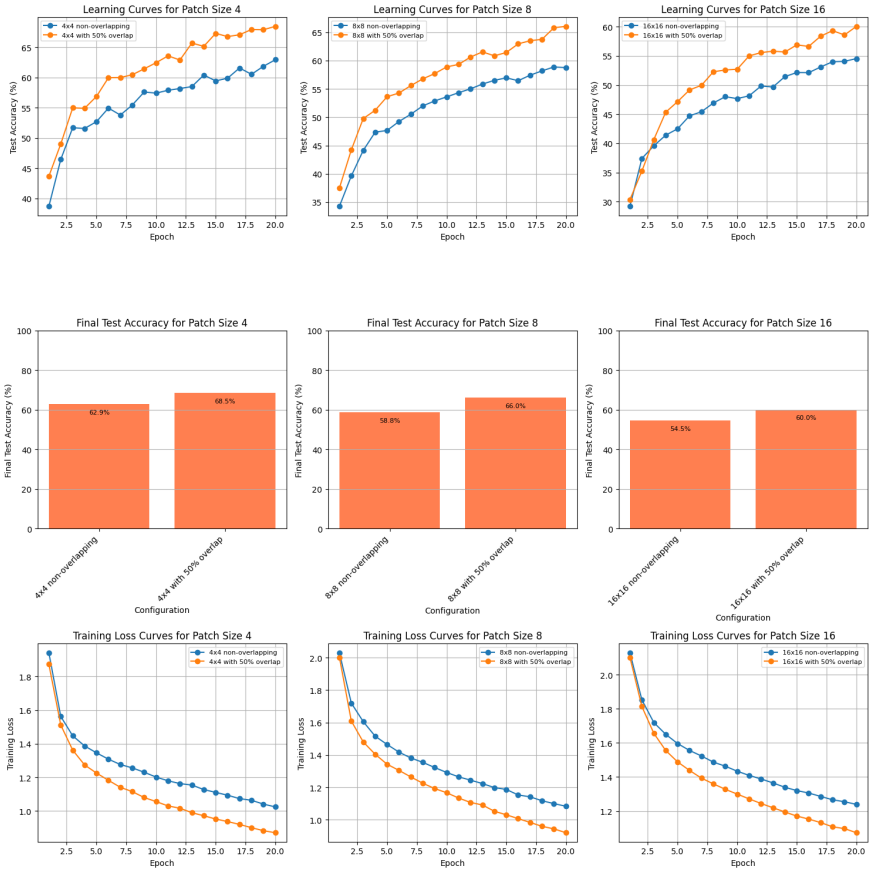


Figure 5: Overlaping and Non-overlaping Analysis

Experiment 4: Varying Number of Attention Heads

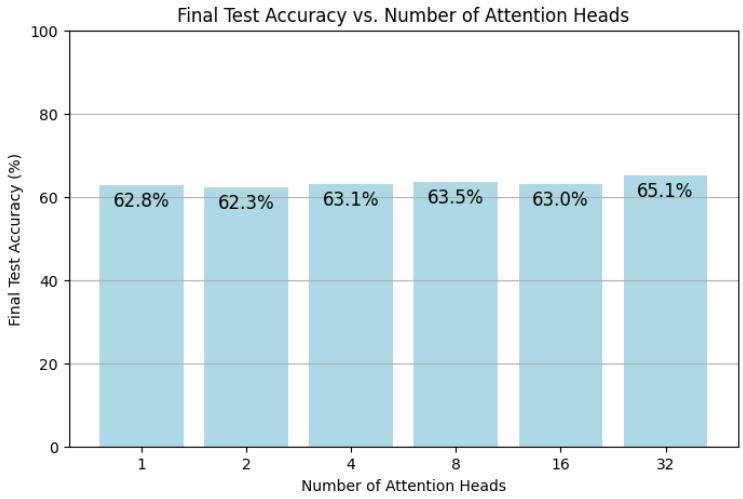


Figure 6: Final Test Accuracy vs. Number of Attention Heads

Key Observations

- **General Trend:** Increasing heads generally improves accuracy, but gains can fluctuate if each head's dimension becomes too small or hyperparameters are not tuned.
- **Trade-Off:** More heads require greater compute; balancing head count and embedding size is essential.
- **Implication:** While multiple heads enrich attention patterns, simply adding heads without increasing embedding dimension can limit each head's expressive power.

Experiment 5: Evaluating CLS Token Outputs from Each Layer (Base Model)

Results

Layer	CLS Token Accuracy
Layer 1	28.10%
Layer 2	46.23%
Layer 3	53.40%
Layer 4	57.30%
Layer 5	59.96%
Layer 6	60.74%

Table 3: CLS Token Accuracies at Each Layer of the Base Model

Observations

- **Progressive Improvement:** The classification accuracy increases steadily from the first to the last layer. Early layers (Layer 1) yield low accuracy, indicating that the initial layers primarily capture low-level, local features.

- **Mid-Layer Enhancement:** By the third and fourth layers, the accuracy improves significantly (53.40% and 57.30%, respectively), suggesting that these layers start integrating more global context.
- **Final Layer Superiority:** The highest accuracy is achieved at the final layer (Layer 6, 60.74%), confirming that deeper layers are most effective in aggregating information across the entire image for robust classification.

Experiment 6: Attention Map Visualization Across Transformer Layers

Results & Observations

- **Early Layers (Layer 1–2):**
  The attention maps in the initial layers are relatively diffuse, indicating that the model is primarily capturing low-level, local features without strong emphasis on specific regions.
- **Intermediate Layers (Layer 3–4):**
  As we move deeper into the network, the attention maps begin to show a more focused pattern, with attention gradually concentrating on regions that correspond to the object of interest.
- **Deeper Layers (Layer 5–6):**
  The final layers display highly localized and discriminative attention, highlighting the salient parts of the image that are most relevant for classification. This progression suggests that the transformer successfully integrates global context as the data passes through successive layers.

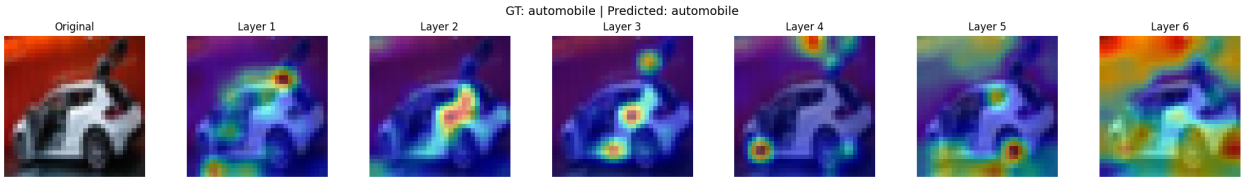


Figure 7: Attention across differnt layerss

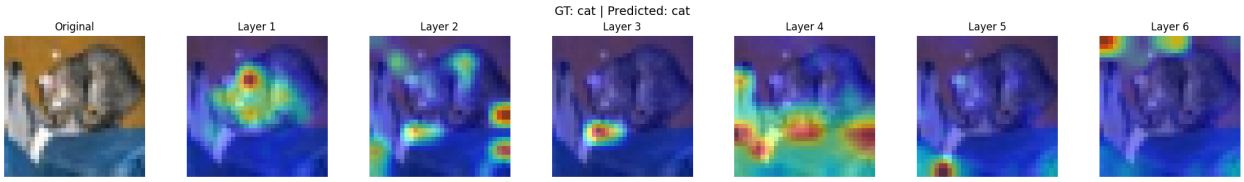
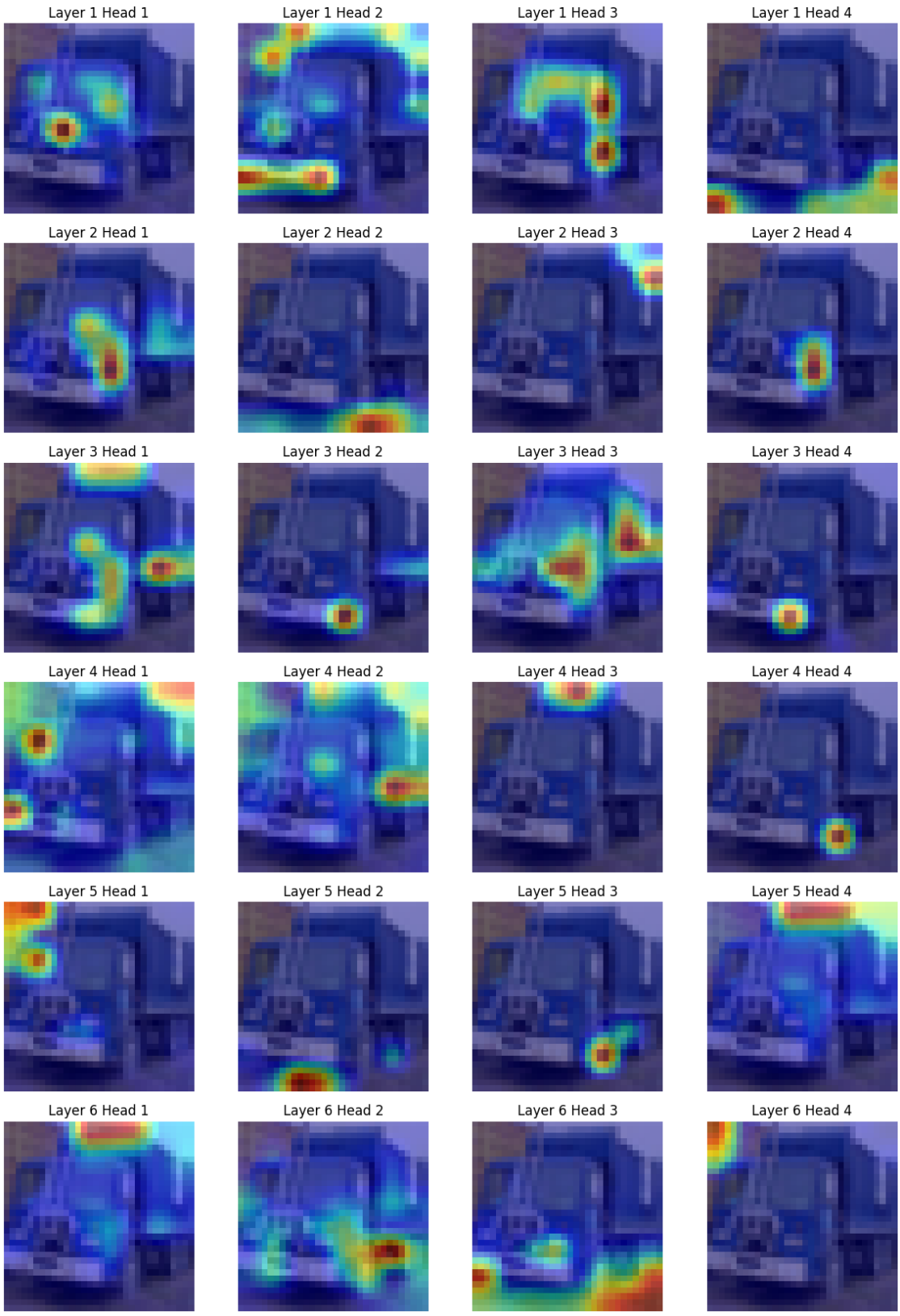


Figure 8: Attention across differnt layers


Attention Maps for All Heads and Layers (CLS Token)
Layer 1 Head 1
Layer 1 Head 2
Layer 1 Head 3
Layer 1 Head 4
Layer 2 Head 1
Layer 2 Head 2
Layer 2 Head 3
Layer 2 Head 4
Layer 3 Head 1
Layer 3 Head 2
Layer 3 Head 3
Layer 3 Head 4
Layer 4 Head 1
Layer 4 Head 2
Layer 4 Head 3
Layer 4 Head 4
Layer 5 Head 1
Layer 5 Head 2
Layer 5 Head 3
Layer 5 Head 4
Layer 6 Head 1
Layer 6 Head 2
Layer 6 Head 3
Layer 6 Head 4