Transformers in Computer Vision: Architecture, Comparative Analysis, and Recent Advances

Sandipan Chakraborty, 2205412

November 3, 2025

1 Introduction

The transformer architecture serves as a revolutionary force in computer vision because it enables machines to handle visual data through new processing methods. The transformer architecture which was created for natural language processing sequential data now processes images through new architectural designs. The research examines transformer architecture basics by studying Vision Transformers against conventional Convolutional Neural Networks through their distinct characteristics and their operational limitations and resulting design improvements.

2 Transformer Architecture: Core Components

2.1 Encoder Design

The encoder uses a stack of identical layers to process input sequences through two essential components which operate together in each layer. The first component uses multi-head self-attention to determine which input elements should receive more importance relative to others. Following this, a position-wise feed-forward network processes each position independently using the same learned parameters. The sub-components receive residual connections and layer normalization which enables stable gradient flow and allows training of deep networks. The encoder receives positional encodings for input representations which enable the model to understand sequence order information. The attention mechanism requires positional encodings because it handles input data as an unstructured collection of elements. The original implementation depends on sinusoidal functions with different frequencies yet learnable positional embeddings have shown similar effectiveness in actual use.

2.2 Decoder Design

The decoder architecture follows the same structure as the encoder but adds more complexity to perform sequence generation operations. The encoder contains self-attention and feed-forward components but the decoder contains an additional cross-attention layer which focuses on encoder output information. The decoder applies this intermediate attention mechanism to select vital input segments for producing each output element. The decoder self-attention mechanism requires masking as a crucial difference because it blocks positions from accessing future positions in the sequence. The model maintains causal relationships through its masked attention mechanism which blocks it from accessing future tokens during training.

2.3 Self-Attention Mechanism

The self-attention operation calculates all sequence position relationships at once through a process that differs from traditional sequential methods. The mechanism generates three vectors for each position through learned linear transformations which produce queries and keys and values. The attention calculation starts by performing dot product operations between queries and keys

followed by key dimension square root scaling and softmax normalization and ends with value weighting using attention weights. Multi-head attention extends this by performing multiple parallel attention operations with different learned projections. Each attention head can focus on different aspects of the relationships between elements, with their outputs concatenated and projected to produce the final result. The model uses parallel structure to process information from multiple representation subspaces at the same time.

3 Vision Transformers Versus Convolutional Neural Networks

3.1 Architectural Philosophy

The Vision Transformer system transforms image processing through its method of treating images as sequential discrete patch elements instead of using traditional continuous spatial grids. An image is partitioned into fixed-size non-overlapping patches, which are flattened, linearly embedded, and processed as tokens analogous to words in text. The processing method of CNNs differs from this approach because they use hierarchical layers with localized filters to expand their receptive fields step by step. The self-attention mechanism allows each patch to establish direct connections with all other patches during the first layers which creates global connectivity from the start. The global understanding of CNNs develops through multiple convolutional layers which first detect local patterns before merging them into more complex features at increasing depths. The different architectural structure of Vision Transformers enables better natural modeling of distant relationships yet it loses the spatial learning patterns which convolutions provide.

3.2 Computational Characteristics

The two paradigms require different computational resources for their operations. The computational complexity of convolutional operations stays the same between layers because the operations depend on kernel size and channel dimensions and spatial extent. Standard self-attention requires quadratic computational time that scales with sequence length which makes it impractical for processing large high-resolution images because their sequence length equals the product of height and width divided by patch area. The CNN architecture maintains translation equivariance and local information through its convolutional structure which implements strong architectural priors. The inductive biases of CNNs enable them to learn image data efficiently because they produce excellent results when trained with small amounts of data. The built-in assumptions of Vision Transformers disappear when the model needs to learn visual representations at the same level as other models, thus requiring extensive additional training data. When provided sufficient training data, however, transformers can discover more flexible and general patterns unconstrained by predetermined architectural choices. The memory footprint also differs substantially. Quadratic attention complexity necessitates significantly more memory for processing high-resolution inputs compared to CNNs with fixed-size kernels. This disparity has important practical implications for deployment scenarios with constrained computational budgets or real-time processing requirements.

4 Challenges and Architectural Advances

4.1 Fundamental Limitations

Early Vision Transformer implementations revealed several practical challenges. Data efficiency emerged as a primary concern, with these models requiring massive pre-training datasets to achieve competitive performance. Without extensive pre-training on hundreds of millions of images, Vision Transformers frequently underperformed compared to similarly-sized CNNs, particularly on smaller target datasets (Touvron et al. [2]).
Computational scalability posed another significant challenge. The quadratic relationship between sequence length and computational cost creates prohibitive resource requirements for high-resolution imagery (Liu et al. [1]). Processing images at native resolution becomes impractical, forcing compromises in either patch granularity or input resolution. Furthermore, standard Vision Transformers

produce single-scale feature representations, limiting their effectiveness for dense prediction tasks like segmentation and detection that benefit from multi-scale feature hierarchies.

4.2 Hierarchical Windowed Attention: Swin Transformer

Liu et al. [1] introduced the Swin Transformer to address computational limitations through hierarchical architecture and localized attention windows. Rather than computing attention globally across all patches, this approach restricts self-attention to non-overlapping local windows, reducing computational complexity from quadratic to linear with respect to image size. The architecture achieves cross-window communication through shifted window partitioning between consecutive layers, maintaining the ability to model relationships across the entire image while controlling computational costs. The hierarchical design progressively merges patches in deeper layers, producing multi-scale feature maps analogous to CNN feature pyramids. This structural choice makes Swin Transformer directly applicable to dense prediction tasks requiring features at multiple spatial resolutions. The model achieved state-of-the-art results across diverse computer vision benchmarks, demonstrating that careful architectural design can overcome the scalability limitations plaguing standard Vision Transformers while maintaining reasonable computational efficiency.

4.3 Knowledge Distillation Approaches: DeiT

Touvron et al. [2] tackled Vision Transformers' data hunger through sophisticated training strategies embodied in Data-efficient Image Transformers. The central innovation involves a distillation token that enables transformer students to learn from teacher models through the attention mechanism itself. This token-based distillation proves particularly effective when using convolutional teachers, allowing the student transformer to inherit beneficial inductive biases from the teacher while maintaining its own architectural advantages.

5 Experiment and Analysis

This section details the comparative analysis between a Vision Transformer (ViT) and a standard Convolutional Neural Network (CNN) trained from scratch on the CIFAR-10 dataset.

5.1 Model Configuration and Parameter Impact

The experiment parameters were set based on the seed (roll number 2205412). The key hyperparameters for the ViT were:

- **Hidden Dimension (D):** 192
- **Number of Heads (k):** 6
- **Patch Size (P):** 8×8
- **Number of Epochs:** 12

The **8×8 patch size** is the most critical parameter. For a 32×32 image, this results in a sequence of only $N = (32/8)^2 = 16$ patches. This low-resolution tokenization severely limits the ViT's ability to discern fine-grained details, which is a major disadvantage on a dataset like CIFAR-10.

5.2 Performance Comparison

Both models were trained for 12 epochs. The quantitative results, summarized in Table 1, show a significant performance disparity.

Table 1: Performance and Complexity Comparison

Metric	ViT	Simple CNN
Best Test Accuracy	58.92%	**79.92%**
Total Parameters	2,712,010	**919,306**
Parameter Ratio (vs CNN)	2.95x	1x

The CNN achieves a **21% higher accuracy** while being almost **three times more parameter-efficient**. The learning curves in Figure 1 visually confirm this. The CNN (red) converges faster, at a lower loss, and to a much higher accuracy than the ViT (blue).

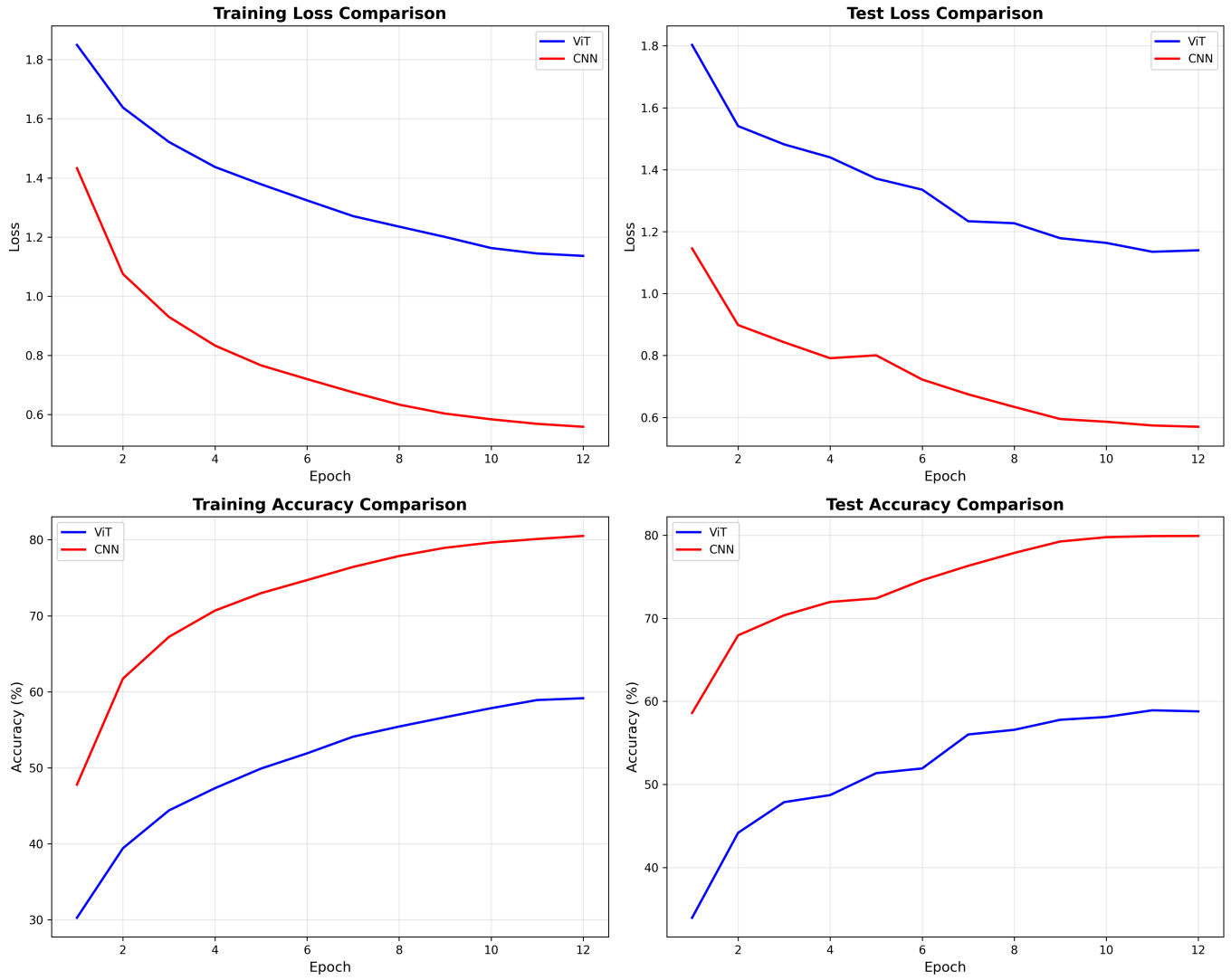

Figure 1: Comparison of training/test loss and accuracy for ViT (blue) and CNN (red) over 12 epochs. The CNN clearly outperforms the ViT in all metrics.

This gap is attributed to the **inductive bias** of CNNs (locality and translation invariance), which are ideal for image data. The ViT, lacking this bias, cannot learn these fundamental spatial concepts from the small CIFAR-10 dataset.

5.3 Confusion Matrix Analysis

The confusion matrices for both models highlight the difference in classification confidence.

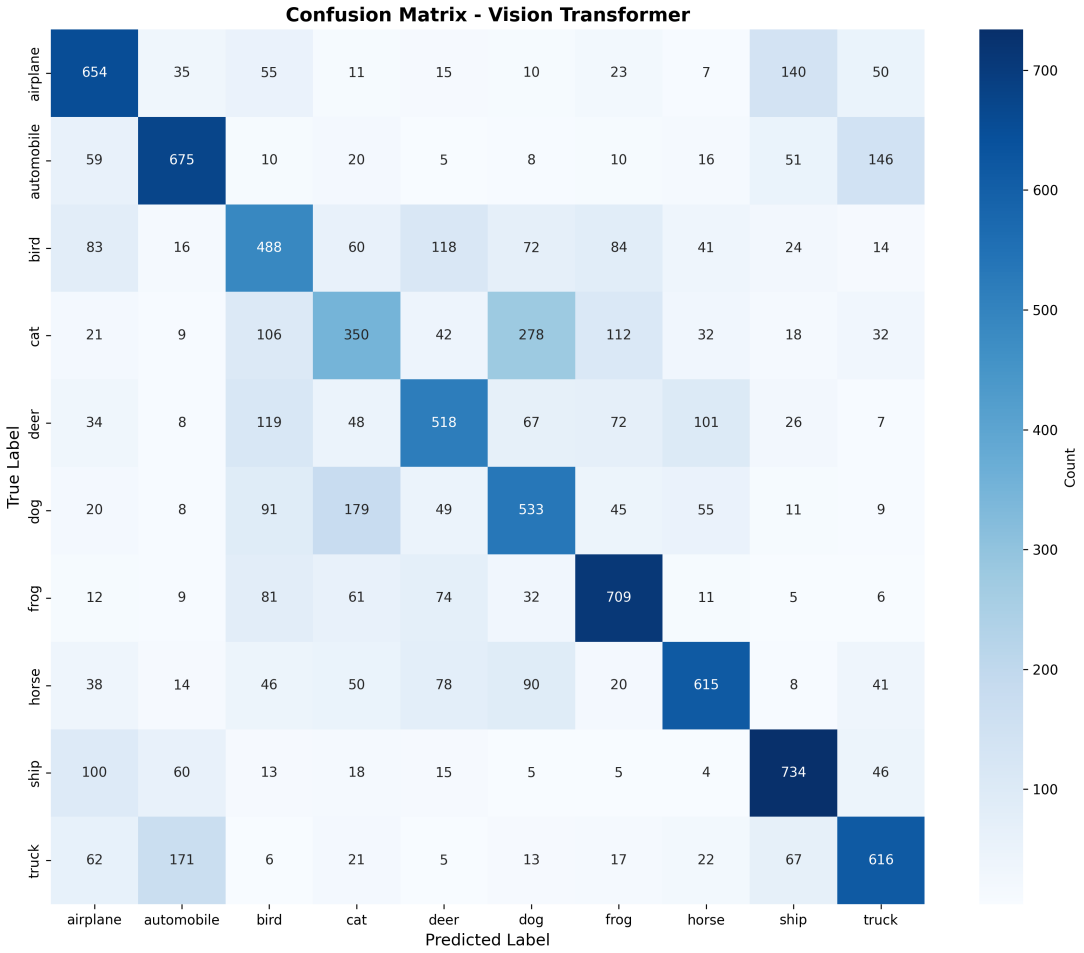

Figure 2: ViT Confusion Matrix (Accuracy: 58.92%). Note the significant off-diagonal noise, indicating widespread confusion (e.g., True 'cat' predicted as 'dog' 278 times).

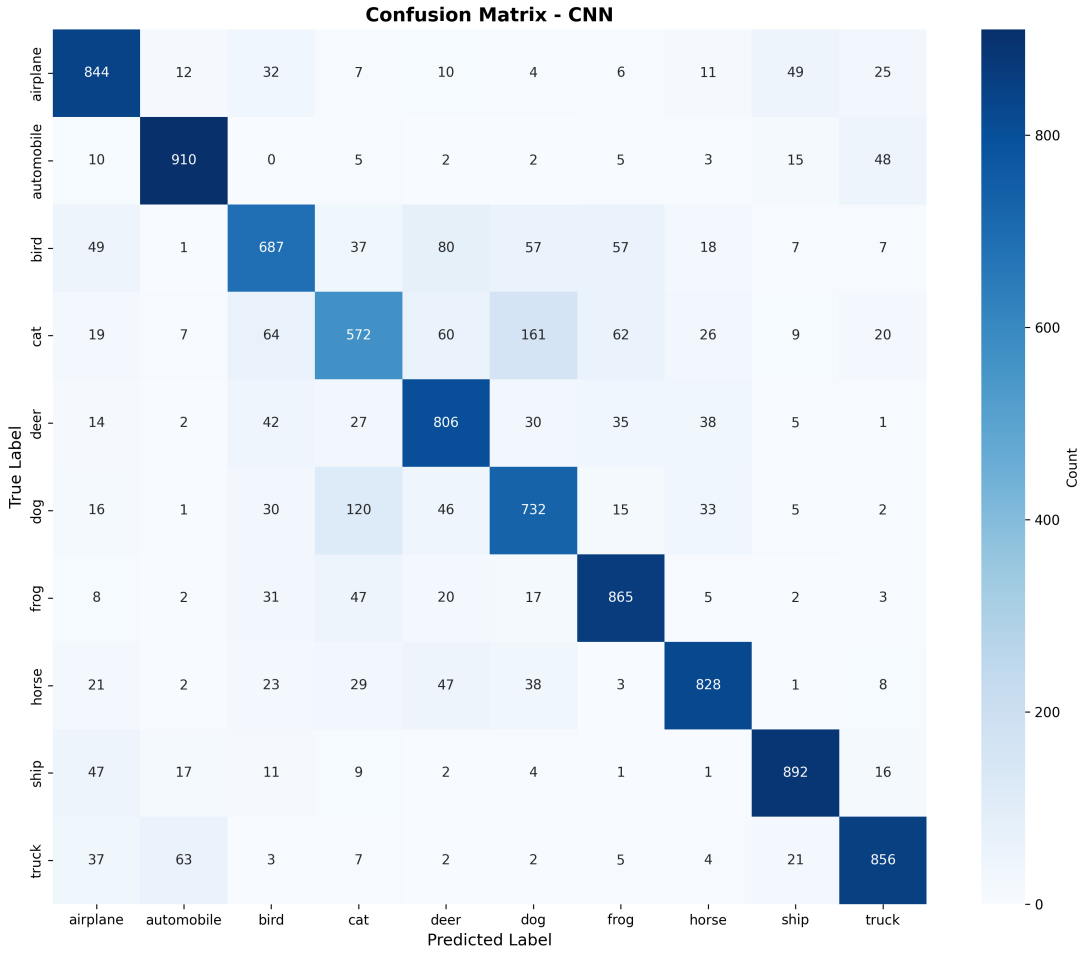

Figure 3: CNN Confusion Matrix (Accuracy: 79.92%). The diagonal is much stronger, showing better per-class accuracy. Confusion still exists (e.g., True 'cat' as 'dog' 161 times) but is far less frequent.

The ViT matrix (Figure 2) shows high confusion between visually similar classes. For example, 'cat' is frequently misclassified as 'dog' (278 instances) and 'bird' (106 instances). In contrast, the CNN matrix (Figure 3) shows a much stronger diagonal, corresponding to its higher accuracy.

5.4 Model Interpretability (Explainability)

To understand *how* each model made its decision, we visualized their focus on a sample image of a 'cat'.

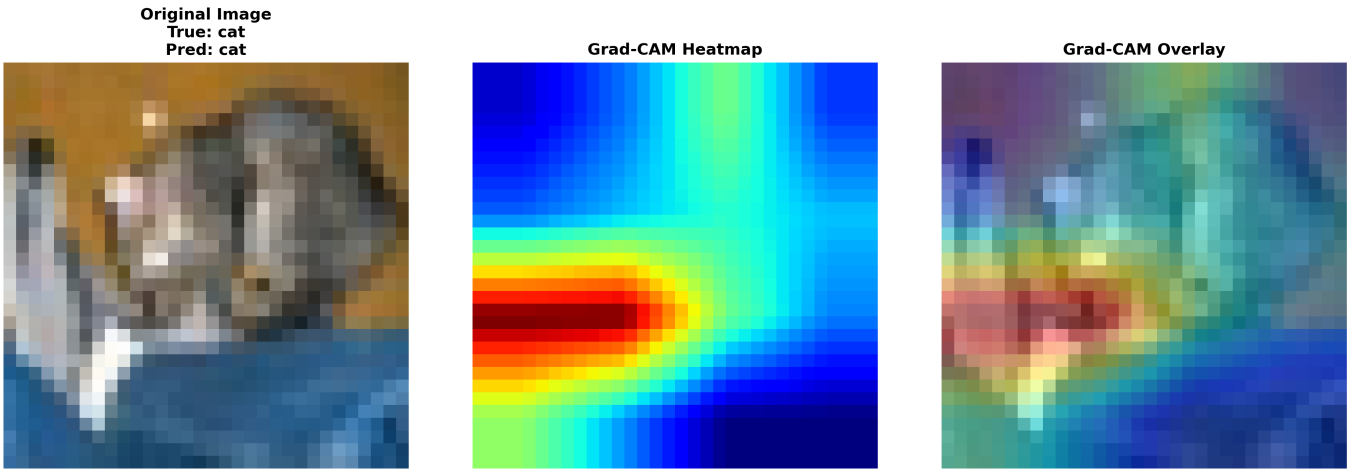

Figure 4: Grad-CAM visualization for the CNN. The model correctly classifies the 'cat' by focusing its attention (red heatmap) on the cat's head and body.

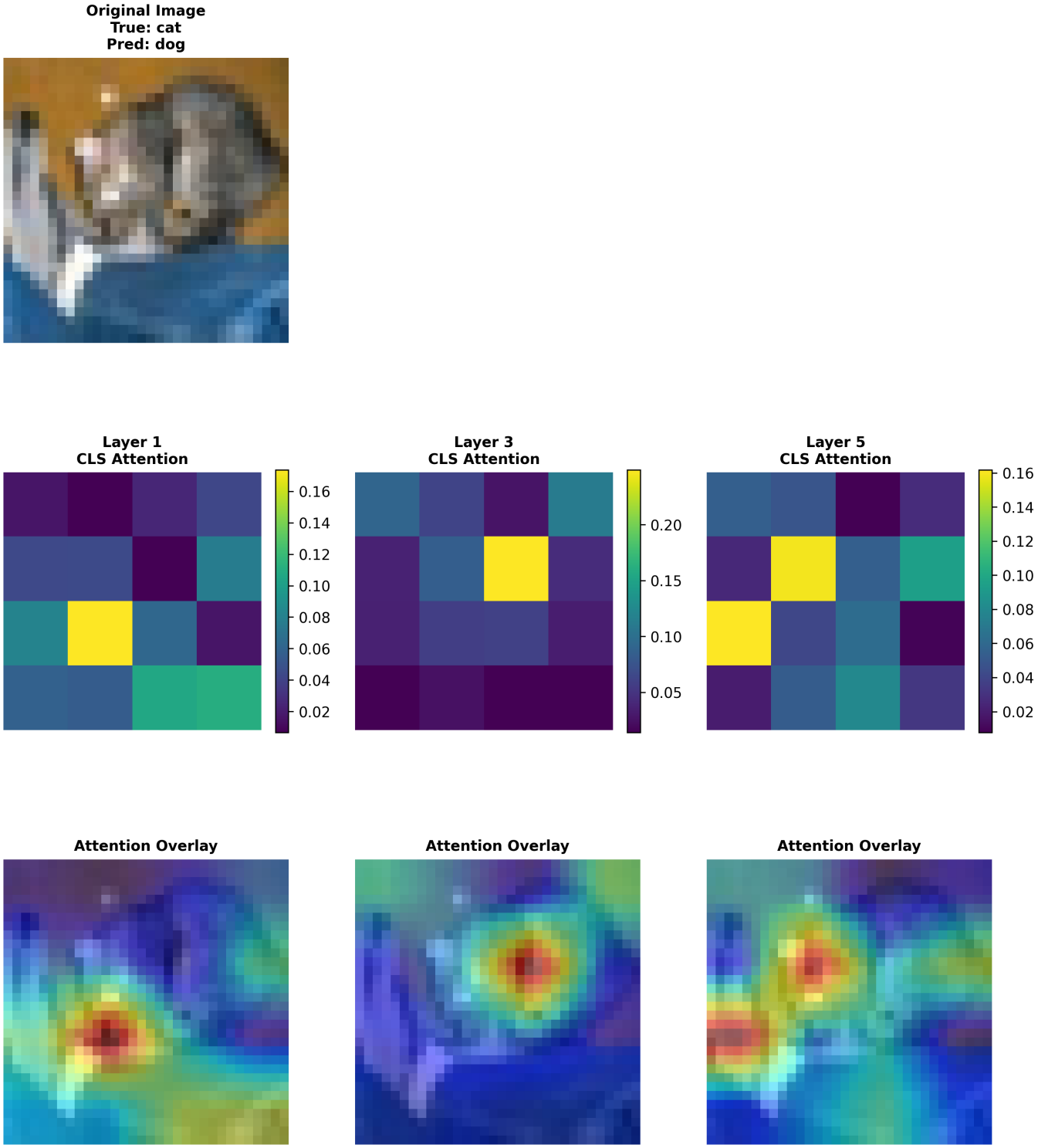

Figure 5: ViT CLS token attention maps (Layers 1, 3, 5). The model *incorrectly* classifies the 'cat' as a 'dog', despite its attention (yellow patches) being on the animal's head.

The explainability plots are very revealing:

- **CNN (Grad-CAM):** As seen in Figure 4, the CNN correctly identifies the 'cat'. The heatmap shows it is looking at the most salient features (the animal's head and torso) to make this decision.
- **ViT (Attention):** As seen in Figure 5, the ViT *incorrectly* classifies the 'cat' as a 'dog'. The attention maps show the model is also looking at the correct patches (the animal's head). However, this demonstrates that even when "looking" at the right place, the coarse 8×8 patch representation did not provide enough detail for the model to distinguish a cat from a dog, leading to misclassification.

6 Conclusion

This experiment provided a clear and definitive comparison between a Vision Transformer (ViT) and a standard Convolutional Neural Network (CNN) for image classification when trained from scratch on the CIFAR-10 dataset.

The results demonstrate that the **CNN significantly outperformed the ViT** in this low-data, "from-scratch" regime. The CNN achieved a best test accuracy of **79.92%**, whereas the ViT's performance plateaued at **58.92%**. This 21% accuracy gap was achieved despite the CNN being far more **parameter-efficient**, utilizing approximately 3x fewer parameters (919k vs. 2.7M).

The primary reason for this disparity lies in the **inductive bias** inherent to CNNs. The architectural principles of locality and translation invariance provide a crucial "head start" for learning from image data, making CNNs highly effective even on smaller datasets.

Conversely, the ViT, which lacks these built-in priors, must learn all spatial relationships from the data. Without large-scale pre-training, it struggles to generalize. This weakness was further compounded by the roll-number-based **8×8 patch size**, which resulted in a very short sequence of only 16 patches. This coarse tokenization discarded fine-grained details essential for distinguishing between classes in CIFAR-10, a fact highlighted by the ViT's misclassification of a 'cat' despite attending to the correct patches.

In summary, for small-scale datasets like CIFAR-10, the classic CNN architecture remains the vastly superior and more practical choice when training is performed from scratch.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[2] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021.