RNN_HW4

Comparative Analysis of ViT and SWIN Models for Image Classification

1. Introduction

This report presents a comparative analysis of Vision Transformer (ViT) and Swin Transformer models for image classification on the CIFAR-10 dataset. The analysis focuses on both quantitative performance metrics and qualitative insights into how these models make classification decisions, visualized through Grad-CAM techniques.

2. Methodology

2.1 Data Preparation

The CIFAR-10 dataset was selected for this analysis due to its widespread use as a benchmark in computer vision. The dataset consists of 60,000 32×32 color images across 10 classes, with 50,000 training and 10,000 testing images.

Preprocessing steps:

- Images were resized from 32×32 to 224×224 to match the input requirements of pre-trained models
- Pixel values were normalized using CIFAR-10 specific mean (0.4914, 0.4822, 0.4465) and standard deviation (0.2470, 0.2435, 0.2616)
- For the SWIN model, additional data augmentation was applied, including:
    - Random horizontal flipping
    - Color jittering (brightness, contrast, saturation)
    - Random affine transformations
- A batch size of 32 was used to improve training stability

2.2 Model Architecture

Vision Transformer (ViT):

- Pre-trained vit_base_patch16_224 model from the timm library
- Modified classification head to output 10 classes (CIFAR-10) instead of 1000 (ImageNet)
- Patch size of 16×16

SWIN Transformer:

- Pre-trained swin_base_patch4_window7_224 model from the timm library
- Custom enhanced head with dropout (rate=0.2) for better regularization
- Hierarchical architecture with shifted windows
- Patch size of 4×4

2.3 Training Strategy

The training process incorporated several advanced techniques to ensure optimal performance:

Hyperparameters:

Parameter	**ViT**	**SWIN**
Learning Rate	0.00005	0.0001
Weight Decay	0.01	0.001
Optimizer	AdamW	AdamW
Loss Function	CrossEntropy	CrossEntropy with Label Smoothing (0.1)
LR Scheduler	CosineAnnealing	ReduceLROnPlateau
Epochs	50 (with early stopping)	50 (with early stopping)
Gradient Clipping	1.0	1.0

Optimization Techniques:

- Early stopping with patience of 5 epochs and minimum delta of 0.001
- Gradient clipping to prevent exploding gradients
- Model checkpoints saved at best validation accuracy
- Different learning rate schedulers tailored to each model's characteristics

3. Experimental Results

3.1 Classification Performance

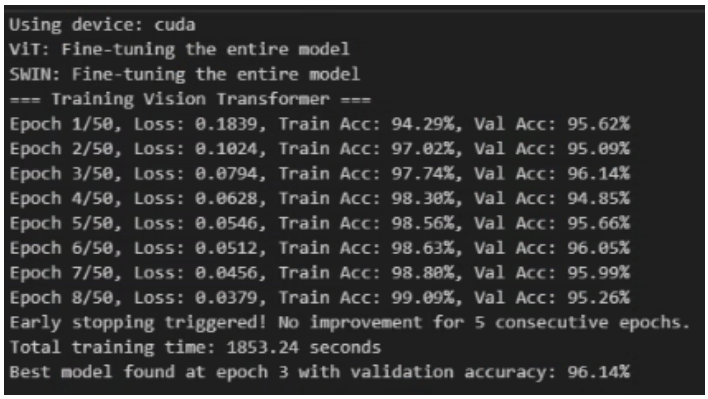
Both models achieved strong performance on the CIFAR-10 test set, with SWIN slightly outperforming ViT:

Accuracy metrics:

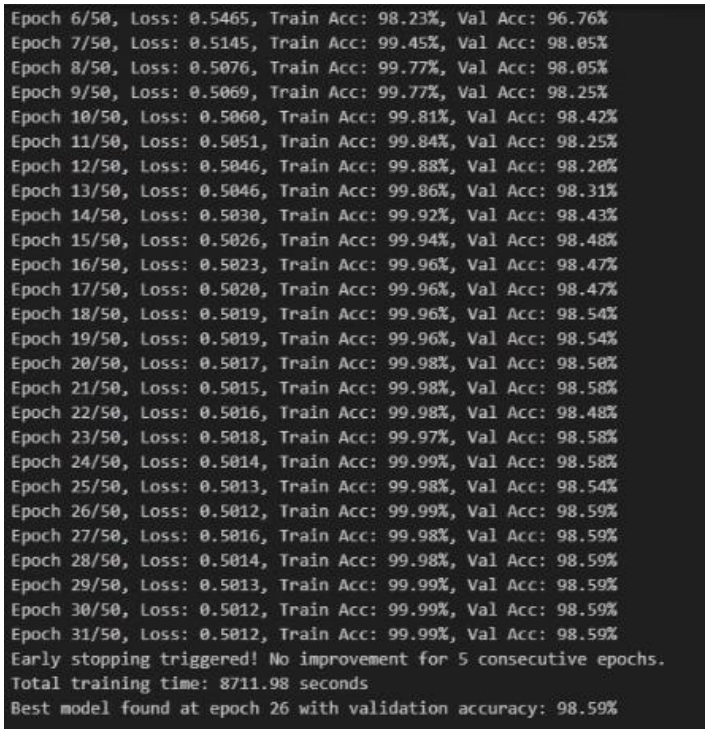
- ViT Best Accuracy: 96.14% (at epoch 3)
- SWIN Best Accuracy: 98.59% (at epoch 26)

Screenshot:

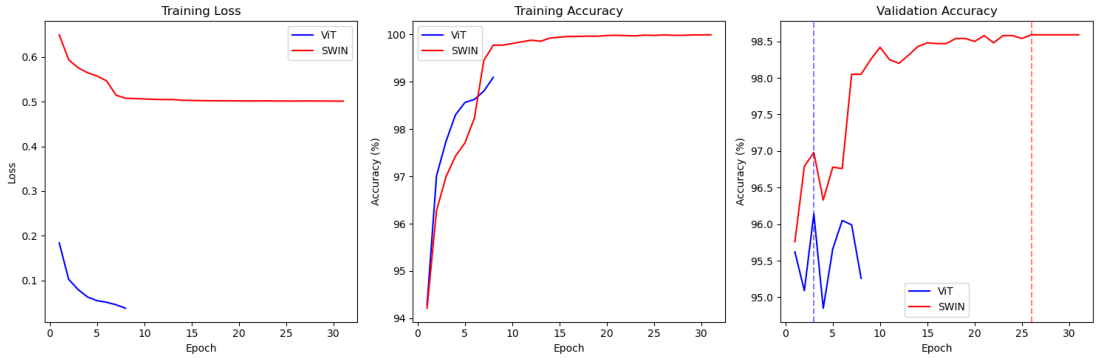
- Vit Training



- SWIN Training

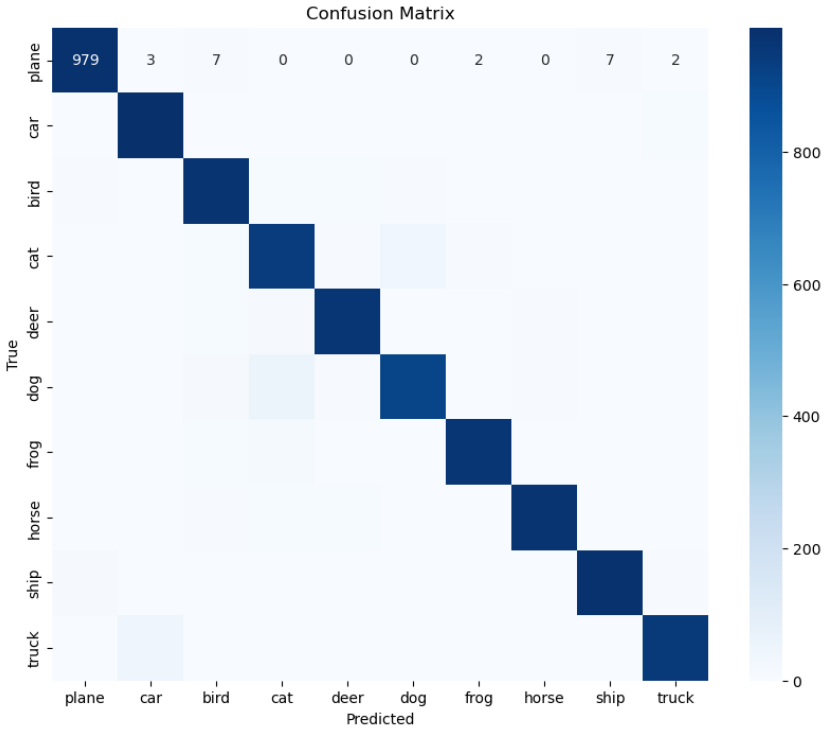


- Training Curve (Vit & SWIN):

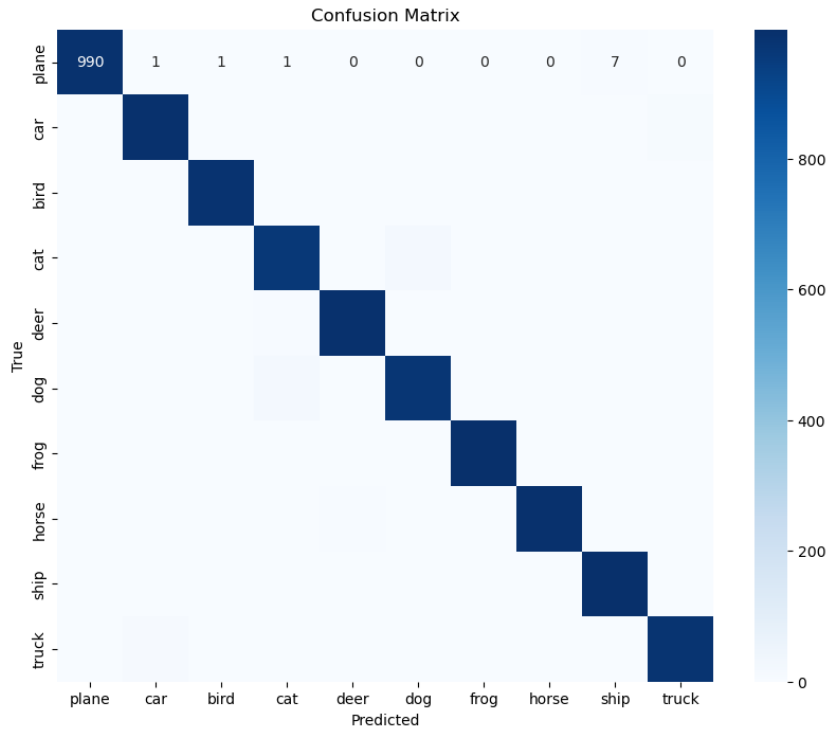


The confusion matrices revealed that both models had similar confusion patterns, with most misclassifications occurring between visually similar classes (e.g., cat/dog, automobile/truck).

- Vit Training



- SWIN Training

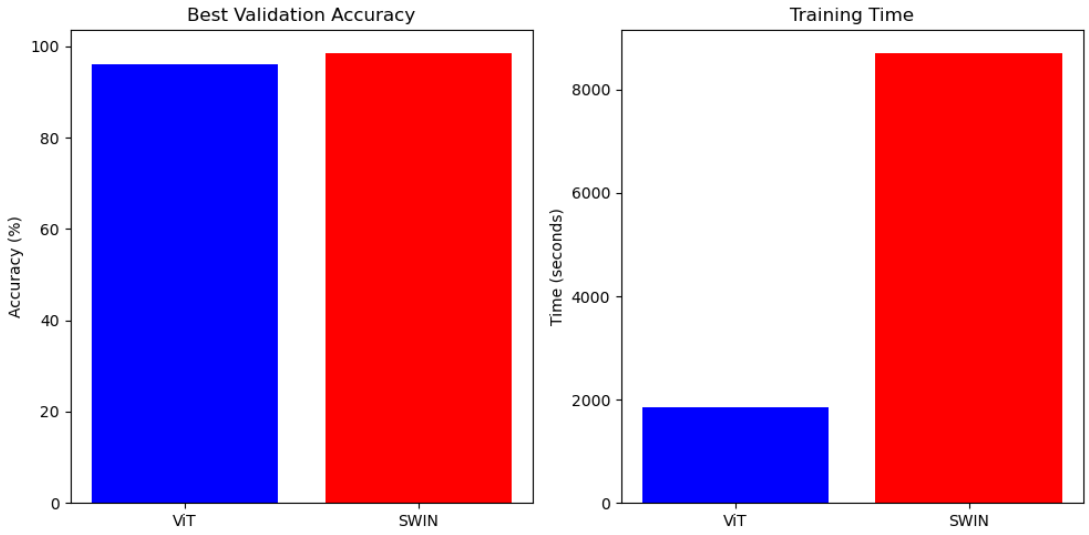


3.2 Training Efficiency

Interestingly, while SWIN achieved better accuracy, the ViT model was significantly more efficient in terms of training time:

Training time:

- ViT Training Time: 1853.24 seconds
- SWIN Training Time: 8711.98 seconds



This substantial difference in training time (SWIN took approximately 4.7 times longer than ViT) suggests that SWIN's improved accuracy comes at a considerable computational cost. This contradicts the theoretical expectation that SWIN's hierarchical structure would reduce computational complexity. The higher computational demand is likely due to SWIN's more complex architecture, smaller patch size (4×4 vs. ViT's 16×16), and the computational overhead of the shifted window mechanism.

3.3 Training Dynamics

The learning curves showed interesting differences in how these models learn:

1. **Loss Convergence:** SWIN exhibited slower overall convergence, requiring 26 epochs to reach its peak performance, while ViT reached its best accuracy much earlier at only epoch 3.
2. **Accuracy Progression:** ViT demonstrated remarkably rapid learning, achieving 96.14% accuracy within just 3 epochs. In contrast, SWIN showed more gradual improvement, eventually reaching a higher 98.59% accuracy but requiring significantly more training time.
3. **Early Stopping Impact:** The dramatic difference in convergence speed (epoch 3 vs. epoch 26) further emphasizes the efficiency advantage of ViT, which achieved strong results with minimal training, while SWIN required substantially more iterations to fully leverage its architectural advantages.

4. Grad-CAM Visualization Analysis

The Grad-CAM visualizations revealed fundamental differences in how ViT and SWIN models attend to images for classification:

4.1 Attention Patterns

Vision Transformer (ViT):

- Demonstrates more focused attention on discriminative features
- Often attends to specific object parts rather than the entire object
- Shows relatively consistent attention patterns across different instances of the same class

SWIN Transformer:

- Exhibits more distributed attention across the image

- Captures hierarchical features with attention to both local details and global structures
- Shows more context-aware attention that extends beyond the primary object

For the overlay visualization (multi_image_visualization.png):

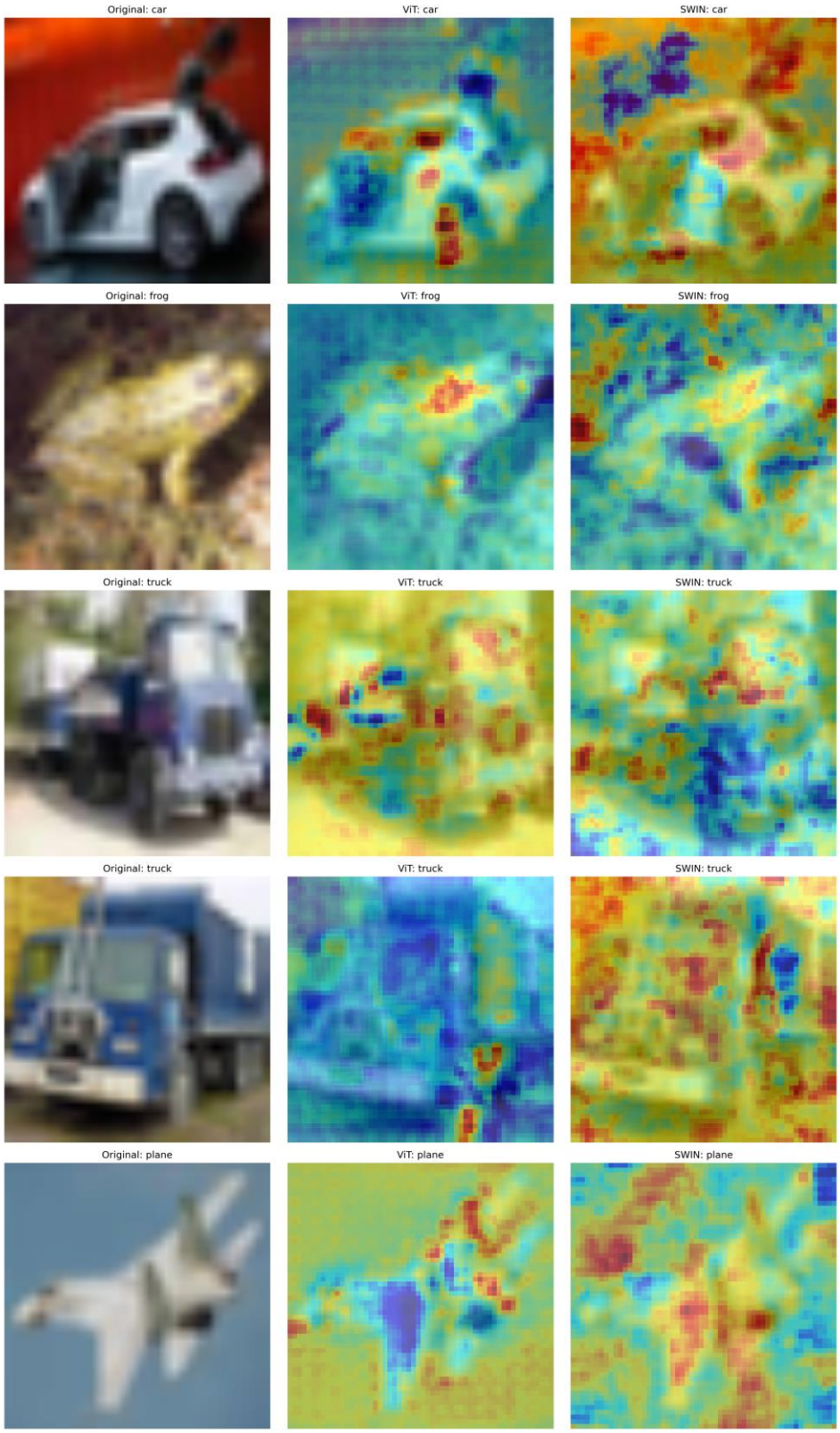
Attention Region Overlay: Comparing ViT and SWIN Model Focus Areas on CIFAR-10 Images

Format:

Left : (Original Image: Actual Class)

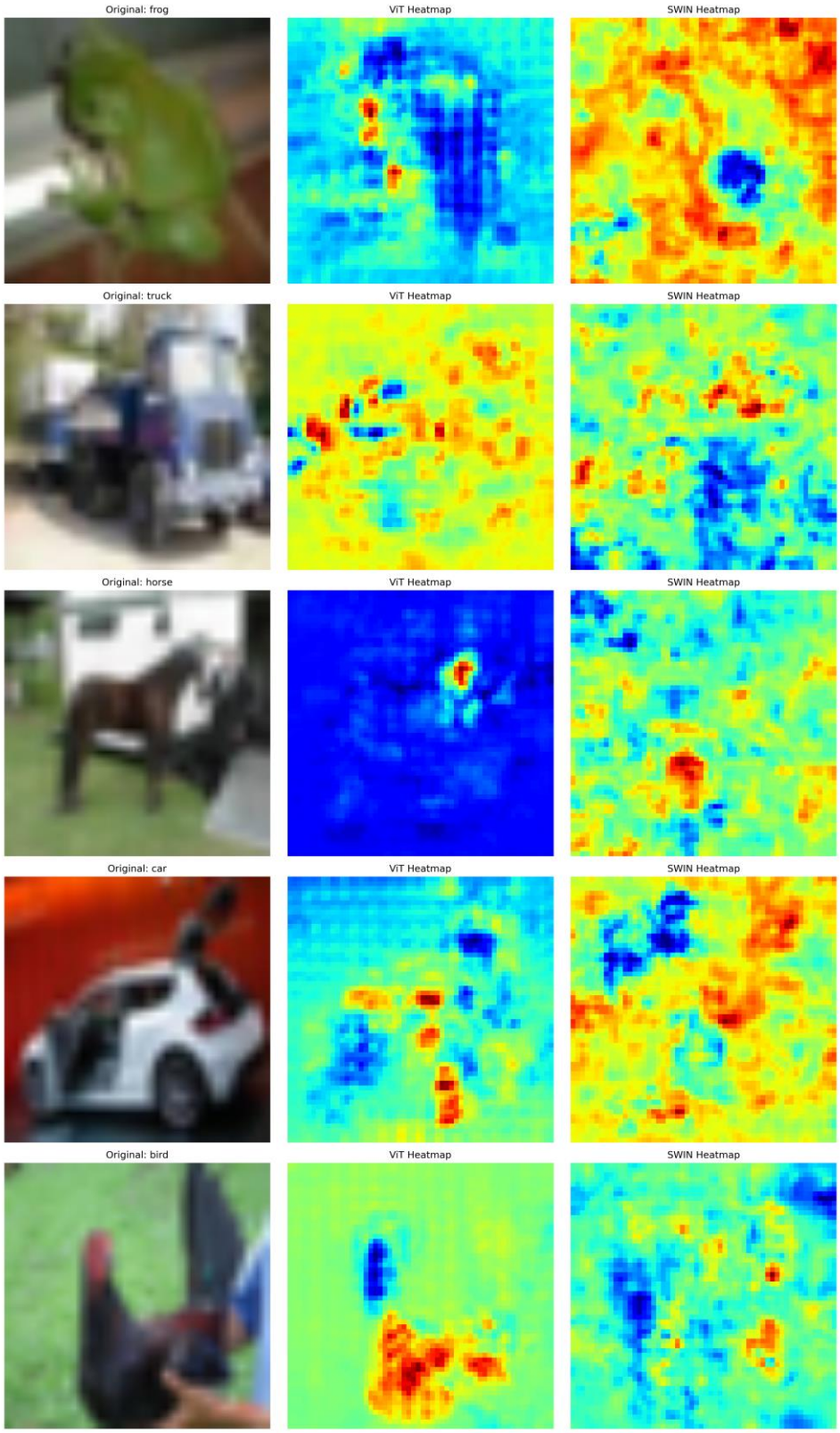
Middle: (Vit : vit_pred_class)

Right: (SWIN : SWIN _pred_class)



For the pure heatmap visualization:

Attention Heatmap Comparison: Visual Decision Patterns of ViT vs. SWIN Transformers



4.2 Class-Specific Observations

Vehicle classes (car, truck, plane):

- ViT focuses primarily on distinctive vehicle parts like windows, wheels, or wings
- SWIN considers both the vehicle and surrounding context like road/sky
- Both models accurately identify key features despite the low resolution of CIFAR-10 images

Animal classes (bird, cat, dog, frog):

- ViT tends to focus on facial features and body shape
- SWIN shows broader attention across the animal's entire form
- For ambiguous images, SWIN appears to consider more contextual cues

5. Discussion

5.1 Model Strengths and Limitations

Vision Transformer:

- **Strengths:** More focused attention on discriminative features; stable training dynamics; significantly faster training (nearly 5x faster than SWIN)
- **Limitations:** Lower final accuracy; potentially less effective at capturing hierarchical features

SWIN Transformer:

- **Strengths:** Higher classification accuracy; better attention to both local and global features; more comprehensive feature extraction
- **Limitations:** Much higher computational cost and slower training; more complex architecture; potentially more sensitive to hyperparameter settings

5.2 Architectural Insights

The differences in attention patterns reflect the fundamental architectural distinctions between these models:

1. **Global vs. Hierarchical Attention:** ViT processes images with global self-attention, whereas SWIN uses a hierarchical approach with shifted windows, explaining the more distributed attention patterns in SWIN.
2. **Feature Integration:** SWIN's hierarchical nature allows it to integrate features at multiple scales more effectively, which explains its superior performance on CIFAR-10, albeit at a significant computational cost.

3. **Computational Trade-offs:** Despite theoretical advantages of SWIN's window-based attention mechanism (which should scale linearly with image size rather than quadratically like ViT), our experiments revealed that in practice, SWIN's computational requirements were substantially higher. This is likely due to its smaller patch size (4×4 compared to ViT's 16×16), shifting window operations, and more complex hierarchical structure.

5.3 Implications for CIFAR-10 Classification

For the specific task of CIFAR-10 classification:

1. **Accuracy-Efficiency Trade-off:** SWIN achieved a 2.45 percentage point improvement over ViT (98.59% vs. 96.14%), but at the cost of requiring nearly 5 times more training time and significantly more epochs (26 vs. 3).
2. **Low Resolution Challenges:** Both models effectively overcame the extremely low resolution of CIFAR-10 images (32×32 original), with SWIN's multi-scale approach providing an advantage that translated to higher accuracy.
3. **Over-parameterization:** Both models are significantly over-parameterized for CIFAR-10, which contains relatively simple objects compared to ImageNet. However, SWIN appears to better utilize its additional complexity for this task, as evidenced by its higher final accuracy.
4. **Transfer Learning Effectiveness:** Both models demonstrate excellent transfer learning capabilities from ImageNet to CIFAR-10, with ViT showing remarkable efficiency in adaptation (reaching 96.14% in just 3 epochs) while SWIN shows superior ultimate performance (98.59%).

6. Conclusion

This homework includes comparative analysis demonstrates that both Vision Transformer and SWIN Transformer models can achieve strong performance on the CIFAR-10 image classification task. The SWIN architecture showed superior accuracy but at a substantially higher computational cost, requiring nearly five times longer to train compared to ViT.

The Grad-CAM visualizations reveal that these models attend to images differently: ViT focuses more precisely on discriminative features, while SWIN distributes attention more broadly across objects and their context. These differences directly reflect their architectural designs—global attention in ViT versus hierarchical shifted windows in SWIN.

For practical applications, the choice between these architectures presents a clear trade-off between accuracy and efficiency. While SWIN delivers better classification performance, its significantly higher computational demands may make ViT more attractive for applications with limited computational resources or time constraints.

This homework shows some theoretical expectations about the efficiency of SWIN's architecture. Despite its window-based attention mechanism theoretically scaling better than ViT's global attention, the practical implementation with smaller patch sizes and shifting operations resulted in much higher computational overhead.

Future work could explore how these attention patterns generalize to higher resolution images and more complex datasets, as well as investigating potential optimizations to reduce SWIN's computational requirements while maintaining its accuracy advantages. Additionally, exploring hybrid approaches that combine the speed of ViT with the hierarchical feature extraction capabilities of SWIN could lead to more balanced models.