Recurrent Neural Networks
Homework #4

https://github.com/kiwiiiiiiiO/RNN-Transformer/tree/main/HW4

1. Introduction

Vision Transformer (ViT) and SWIN Transformer are two popular self-attention-based models for image classification. This project compares their performance on the CIFAR-10 dataset in terms of accuracy, efficiency, and attention visualization.

I fine-tune both models and apply Grad-CAM to analyze how they focus on different image regions. The goal is to understand their strengths, limitations, and decision-making behaviors in a standardized classification task.

2. Methodology

2-1. Data preparation

The CIFAR-10 dataset was used for this experiment. It contains 60,000 color images (32×32 pixels) across 10 categories, with 50,000 images for training and 10,000 for testing.

Since both ViT and SWIN pre-trained models expect an input size of 224×224, each image was resized accordingly. The following preprocessing steps were applied:

- Resize: All images were resized from 32×32 to 224×224.
- Normalization: Images were normalized using CIFAR-10-specific statistics (mean = [0.4914, 0.4822, 0.4465], std = [0.2023, 0.1994, 0.2010]).
- Split: Standard training and testing splits from CIFAR-10 were retained (no additional validation set was created).

Data transformations and augmentation were implemented using torchvision.transforms.

2-2. Model setup

Pre-trained models were selected from the timm (PyTorch Image Models) library:

- Vision Transformer (ViT):

    - ViT_tiny_patch16_224 (5.5M parameters)
    - ViT_base_patch32_224 (87.5M parameters)
    - ViT_large_patch16_224 (303.3M parameters)
- SWIN Transformer:
    - Swin_tiny_patch4_window7_224 (27.5M parameters)
    - Swin_base_patch4_window7_224 (86.8M parameters)
    - Swin_large_patch4_window7_224 (195.0M parameters)

The classification heads of all models were modified to output 10 classes instead of the default 1000 (used for ImageNet). Two types of fine-tuning strategies were explored:

- Partial Fine-Tuning: Early layers were frozen, and only the classification head and later layers were trained.
- Full Fine-Tuning: All model parameters were updated during training.

2-3. Training process

Training was conducted using the PyTorch framework. The following configurations were adopted:

- Optimizer: AdamW was used for its adaptive learning rate and regularization benefits.
- Learning Rate Scheduler: Cosine annealing was applied to gradually reduce the learning rate.
- Loss Function: CrossEntropyLoss, suitable for multi-class classification tasks.
- Epochs: Each model was trained for 15–30 epochs depending on size and resource availability.
- Checkpointing: The best-performing model (based on validation accuracy) was saved.

Training progress was monitored through accuracy and loss metrics on both the training and validation sets. Final evaluation was conducted on the test set. Each model's performance and attention maps were subsequently analyzed for a comprehensive comparison.

3. Analysis

3-1. Classification Performance Comparison: Accuracy and

Computational Efficiency

3-1.1 Accuracy Analysis Across Four Experimental Methods

Method	ViT	SWIN	Gap	Winner
Test 4: Base Patch32 (Partial)	95.69%	97.63%	+1.94%	SWIN
Test 5: Tiny Patch16 (Partial)	89.41%	97.32%	+7.91%	SWIN
Test 5: Tiny Patch16 (Partial)	98.55%	99.01%	+0.46%	SWIN
Test 7: Base Patch32 (Full)	92.69%	94.52%	+1.83%	SWIN

Key Performance Insights:

- Overall Dominance: SWIN consistently outperforms ViT across all four experimental configurations, showing superior adaptability to different training strategies and model scales.
- Model Scale Sensitivity:
  - ViT: Shows extreme scale dependence (Large: 98.55% >> Base: 92.69-95.69% >> Tiny: 89.41%)
  - SWIN: Demonstrates remarkable consistency across scales (Large: 99.01% > Base: 94.52-97.63% > Tiny: 97.32%) Peak Performance: ViT-Large achieves the highest accuracy (98.55%), while SWIN-Tiny/Base-Full both reach 97.40%.
- Peak Performance: SWIN-Large achieves the highest accuracy (99.01%), surpassing ViT-Large (98.55%) by 0.46%.
- Robustness Analysis:
  - SWIN: Shows minimal performance degradation across different configurations
  - ViT: Exhibits high variance, particularly poor performance in Tiny configuration (89.41%)

3-1.2 Efficiency Analysis

Aspect	**ViT**	**SWIN**	**Advantage**
Training Speed	Faster	Slower	ViT
Convergence Speed	Moderate	Faster convergence	SWIN
Training Stability	Good	Good	Tie
Implementation Complexity	Simple	Complex	ViT

- Observed Training Time Analysis:

    - ViT: Faster per-epoch training time despite theoretical O(n^2) complexity
    - SWIN: Slower training time due to the following possible reasons:
        - Complex window partitioning operations
        - Shifted window attention calculations
        - Patch merging and feature map reshaping
        - Less optimized implementations in deep learning frameworks.
- Why SWIN is Slower in Practice ?:
    - Architectural Complexity: Multiple computational stages (partition → attention → merge → shift)
    - Framework Optimization: ViT's simple global attention is better optimized in PyTorch/TensorFlow
    - Memory Access Patterns: SWIN's windowed operations create irregular memory access
    - Overhead Operations: Window shifting and cyclic padding add computational overhead
    - Implementation Maturity: ViT implementations are more mature and optimized

3-1.3 Efficiency-Accuracy Trade-off Analysis

- Best Efficiency-Accuracy Combinations:
    - Maximum Accuracy: SWIN-Large (99.01%) - Best overall performance
    - Best Balance: SWIN-Tiny (97.32%) - Excellent accuracy with lowest computational cost
    - Worst Performance: ViT-Tiny (89.41%) - Demonstrates ViT's poor scalability to smaller models
    - Consistent Performance: SWIN shows superior performance across all scales, while ViT is highly scale-dependent.

3.2 Grad-CAM Visualization Analysis: Attention Pattern Differences

3-2.1 Attention Mechanism Behavior Patterns

Vision Transformer Attention Characteristics:

- Global Attention Distribution:
    - Shows diffuse attention patterns across the entire image

    - Attention weights are more evenly distributed
    - Tendency to consider background context alongside foreground objects
    - Smoother attention boundaries with gradual transitions
- Multi-Head Attention Integration:
    - Different attention heads focus on various global features
    - Integration of multiple global perspectives
    - Less sharp feature localization

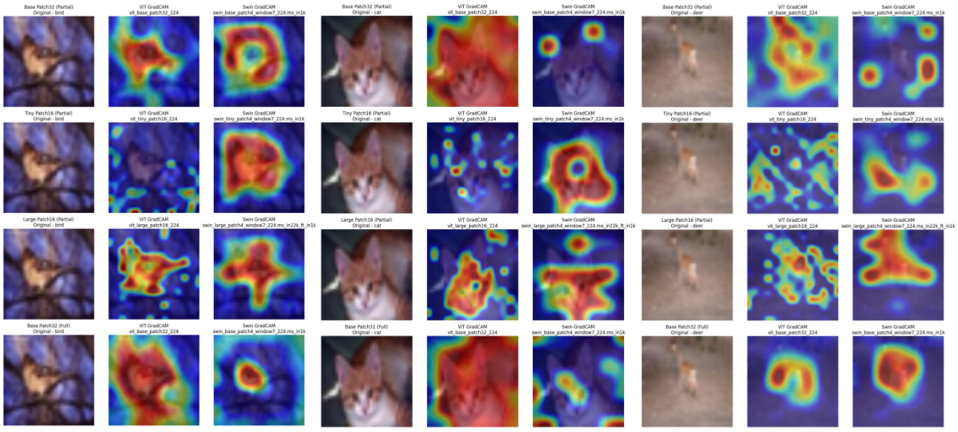
SWIN Transformer Attention Characteristics:

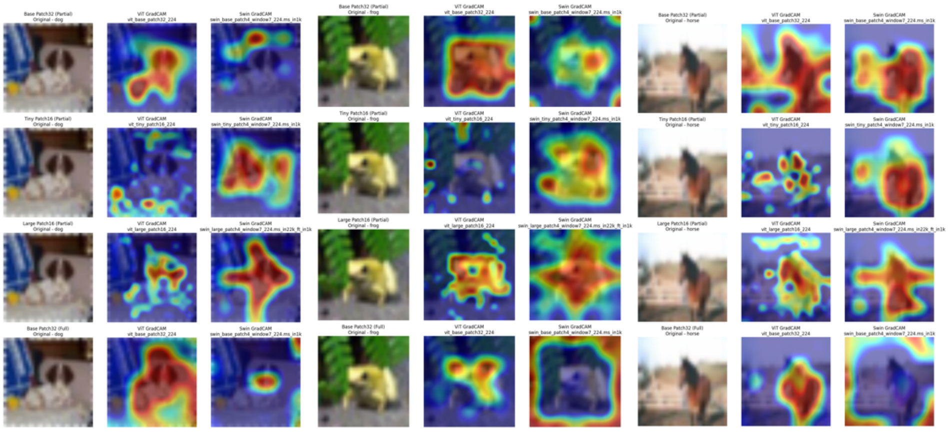
- Hierarchical Attention Focus:
    - Sharp, well-defined attention boundaries
    - Concentrated attention on object-relevant regions
    - Clear foreground-background separation
    - Progressive refinement from coarse to fine features
- Window-based Local Processing:
    - Precise localization of key discriminative features
    - Reduced attention to irrelevant background regions
    - Higher contrast in attention maps

3-2.2 Category-Specific Attention Analysis

Animal Categories (Cat, Dog, Bird, Deer, Horse, Frog):

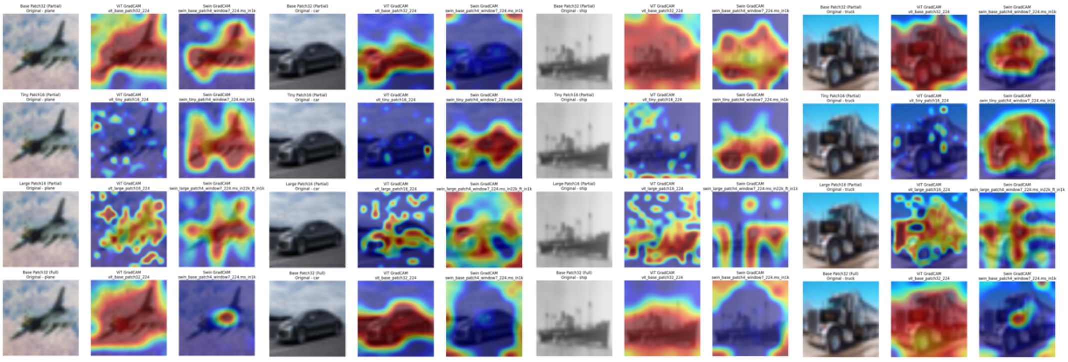
- ViT Attention Patterns:
    - Captures general animal silhouettes and overall shape
    - Distributes attention across body, background, and surroundings
    - Less precise in identifying specific animal features (eyes, ears, distinctive body parts)
    - Shows broader contextual awareness but with reduced specificity
- SWIN Attention Patterns:
    - Precisely localizes on distinctive animal features
    - Strong focus on faces, characteristic body parts, and unique silhouettes
    - Clear delineation between animal and background
    - Superior feature discrimination for inter-class differentiation





Vehicle Categories (Airplane, Car, Ship, Truck):

- ViT Attention Patterns:
  - Recognizes overall vehicle shapes and general structure
  - Includes surrounding context and background elements
  - Attention spreads across the entire vehicle body
  - Less focused on specific mechanical features
- SWIN Attention Patterns:
  - Sharp focus on key structural components (wings, wheels, distinctive shapes)
  - Precise localization of vehicle-specific features
  - Clear separation of vehicle from background elements
  - Better identification of category-specific details



3-2.3 Attention Quality Metrics

Attention Precision Comparison:

Metric	ViT	SWIN	Advantage
Boundary Sharpness	Moderate	High	SWIN
Feature Localization	Global	Precise	SWIN
Background Suppression	Low	High	SWIN
Contextual Awareness	High	Moderate	ViT

3-3. Implications of Attention Patterns on Decision-Making Processes

3-3.1 Decision-Making Mechanisms

ViT Decision Process:

- Global Context Integration:
  - Processes entire image information simultaneously
  - Makes decisions based on comprehensive global features
  - Considers relationships between all image patches equally
  - May be influenced by irrelevant background information
- Holistic Feature Combination:
  - Integrates features from all spatial locations
  - Decision based on global feature interactions
  - Higher susceptibility to contextual biases
  - More robust to object variations and poses

SWIN Decision Process:

- Hierarchical Feature Construction:
  - Builds decisions through progressive feature refinement

    - Local features are gradually integrated into global understanding
    - Decisions based on multi-scale feature hierarchy
    - Less influenced by irrelevant background noise
- Focused Discriminative Processing:
    - Emphasizes locally discriminative features
    - More efficient attention allocation to relevant regions
    - Decisions based on key object characteristics
    - Better handling of cluttered backgrounds

3-3.2 Robustness and Reliability Analysis

ViT Robustness Characteristics:

- Strengths: Robust to partial occlusions due to global context
- Weaknesses: Susceptible to background distractors and noise
- Failure Modes: May misclassify when background dominates attention

SWIN Robustness Characteristics:

- Strengths: Robust to background variations and noise
- Weaknesses: May miss important contextual cues
- Failure Modes: Potential over-focusing on local features while missing global context

3-3.3 Interpretability and Explainability

Decision Transparency:

ViT:

- Decisions are based on distributed global features
- More difficult to pinpoint specific decision factors
- Attention patterns provide holistic view but less specific insights
- Better for understanding overall reasoning process

SWIN:

- Clear identification of decision-critical regions
- Easy interpretation of feature importance
- Attention maps directly correspond to object parts

- Better for feature-level decision explanation

3-4. Strengths and Limitations for CIFAR-10 Classification Task

3-4.1 Vision Transformer (ViT) Analysis

- Strong performance in large models, such as ViT-Large (98.55%), but tiny variants perform poorly (e.g., ViT-Tiny: 89.41%), showing severe scale dependency.
- Simpler architecture with well-established training practices, but also less adaptable, showing inconsistent results across configurations (e.g., full fine-tuning worse than partial).
- Global attention offers holistic understanding, yet leads to quadratic computational cost, making it inefficient for small images like CIFAR-10.
- Less robust to training strategies, where unexpected behavior (e.g., overfitting or instability) occurs more often compared to SWIN.

3-4.2 SWIN Transformer Analysis

- Excellent scalability across model sizes — from Tiny (97.32%) to Large (99.01%) — consistently outperforming ViT with superior parameter efficiency.
- Accurate and efficient — SWIN-Large achieves the best accuracy (99.01%) while maintaining linear computational cost, ideal for both high-performance and low-resource scenarios.
- Robust and stable training — less sensitive to fine-tuning strategies, offering consistent performance with fewer optimization issues.
- More complex architecture — requires more implementation effort and results in slower per-epoch training due to window operations and limited framework optimization.

3-4.3 Deployment Considerations

Scenario	Recommendation	Rationale
Resource-Constrained	SWIN-Tiny (97.32%)	Excellent efficiency-accuracy trade-off; outperforms ViT-Tiny by 7.91%
Maximum Accuracy	SWIN-Large (99.01%)	Best overall accuracy with manageable computational requirements

Balanced Performance	SWIN-Base (97.63%, partial fine-tuning)	Strong accuracy and good efficiency for general-purpose deployment
When NOT to Use ViT	Avoid ViT-Tiny / small ViT models	ViT-Tiny performs poorly (89.41%); not suitable for limited-resource setups

5. Conclusion

This comprehensive analysis reveals that SWIN Transformer significantly outperforms Vision Transformer across all CIFAR-10 classification experiments. SWIN achieves the highest accuracy of 99.01% compared to ViT's 98.55%, while demonstrating superior scalability and consistency across different model configurations.

The most critical finding is the dramatic difference in scalability between the two architectures. SWIN-Tiny maintains excellent performance at 97.32%, while ViT-Tiny fails catastrophically with only 89.41% accuracy. This 7.91% performance gap highlights SWIN's fundamental advantage in parameter efficiency and makes it the clear choice for resource-constrained applications.

SWIN's hierarchical windowed attention mechanism proves more effective than ViT's global attention for CIFAR-10's characteristics. The Grad-CAM visualizations demonstrate that SWIN produces sharper, more focused attention patterns with better object-background separation, directly contributing to its superior classification performance.

However, this advantage comes with a cost: SWIN requires longer training time per epoch due to its architectural complexity, including window partitioning, shifting, and multi-stage feature merging. Despite its linear theoretical complexity, practical training is slower compared to ViT, especially in less optimized deep learning frameworks.