HW4 experiment Report

Name:李崇楷 Student ID: 313834006

Code: https://github.com/kailee0422/RNN-Transformer/tree/main/HW4

Introduction

In recent years, transformer-based architectures have shown remarkable success in various computer vision tasks. This assignment aims to compare two such architectures—Vision Transformer (ViT) and Swin Transformer (SWIN)—on the CIFAR-10 image classification dataset. By fine-tuning pre-trained models and visualizing their decision-making using Grad-CAM, we seek to understand not only their performance in terms of accuracy but also how each model interprets image features. Through this comparative analysis, we aim to highlight the strengths and limitations of ViT and SWIN for small-scale image classification tasks.

Method

Data Preprocessing

- Dataset: **CIFAR-10** (using PyTorch's built-in dataset).
- Preprocessing: Normalize images using ImageNet statistics (mean and standard deviation) and resize images from 32×32 to 224×224 to match input size required by pre-trained models.
- Data Splitting: Use the standard CIFAR-10 train/test split and further split the training set into **80% training** and **20% validation**.

Data Preprocessing

- Models: Models are loaded from the **timm** library
  Vision Transformer: vit_base_patch32_224
  Swin Transformer: swin_base_patch4_window7_224
- Modification: Replace the final classification head to output 10 classes instead of the original 1000.
- Fine-tuning Options:
  Option 1: Fine-tune the entire model.
  Option 2: Freeze early layers and fine-tune only the classification head and later layers to reduce computational cost.

Training

- Optimizer: AdamW
- Learning Rate Scheduler: Cosine Annealing
- Epoch: 20
- Monitoring: Track loss and accuracy on both training and validation sets and save checkpoints based on best validation accuracy.

Evaluation

- Evaluate both models on the CIFAR-10 test set.
- Metrics:

  Classification accuracy

  Top-1 error rate

  Confusion matrix

Grad-CAM Visualization

- Input: One randomly selected image from the test set.
- Visualization for ViT: Initially targeted class token / last attention layer. Due to poor results, switched to **last transformer block's norm1 layer**.
- Visualization for Swin: Initially used feature maps from model.layers[3]. Also switched to **last transformer block's norm1 layer** due to better results.
- Output: Grad-CAM heatmaps showing model attention on the selected image.

Analysis and Comparison

- Performance Comparison:

  Compare accuracy and computational efficiency between ViT and Swin.
- Attention Analysis:

  Use Grad-CAM to understand differences in attention patterns.

  Discuss how each model interprets image features.
- Reflection:

  Evaluate strengths and limitations of Vision Transformer and Swin for the CIFAR-10 classification task.

Result

I conducted four experiments by training four different model configurations: fully fine-tuned Vision Transformer and Swin Transformer models, as well as Vision Transformer and Swin Transformer models where only the classification head and later layers were fine-tuned.

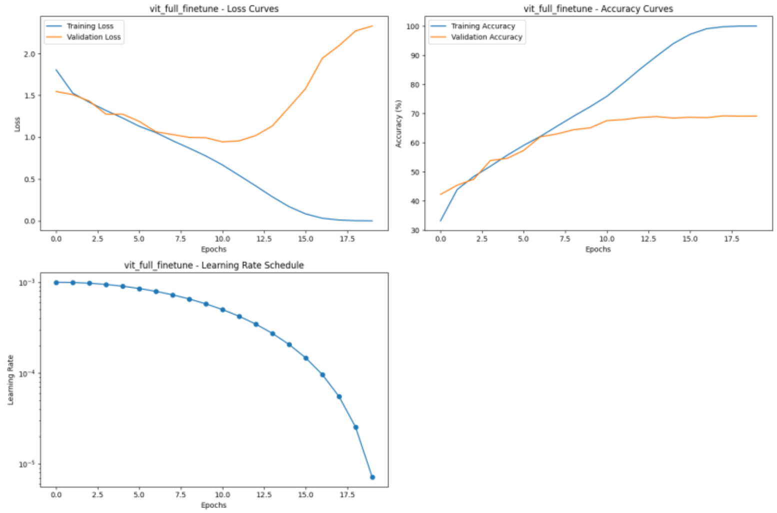


Figure 1. Training loss, accuracy and learning rate schedule over epochs for using fully fine-tuned Vision Transformer

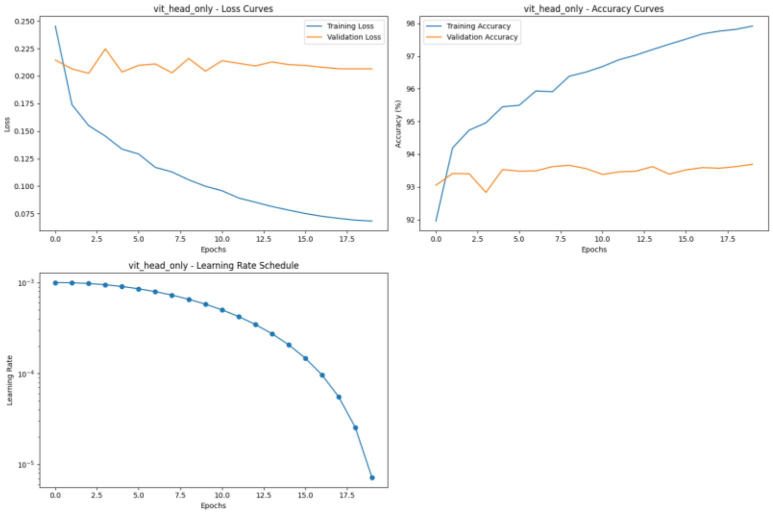


Figure 2. Training loss, accuracy and learning rate schedule over epochs for using only head fine-tuned Vision Transformer.

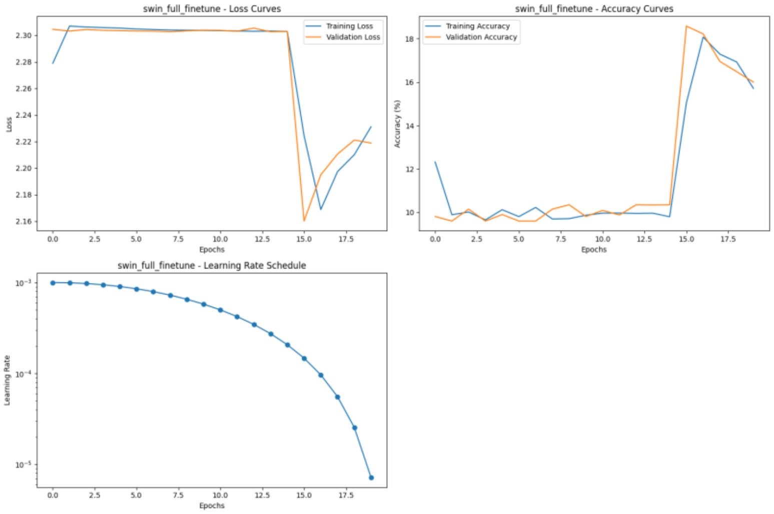


Figure 3. Training loss, accuracy and learning rate schedule over epochs for using fully fine-tuned Swin.

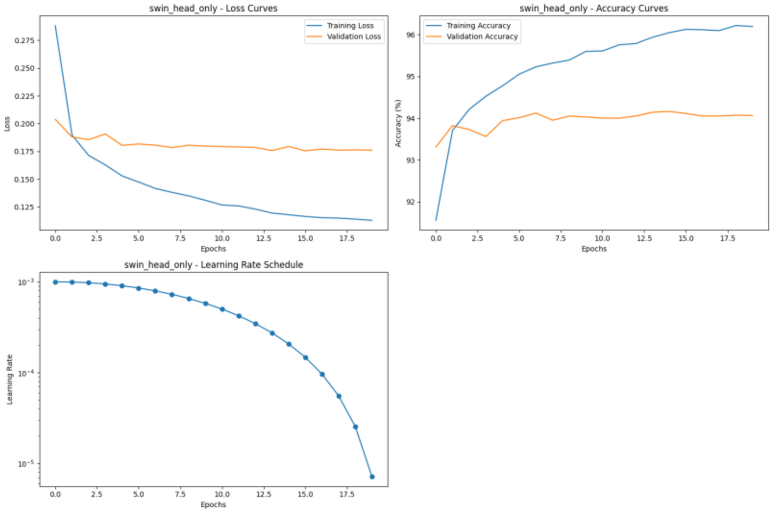
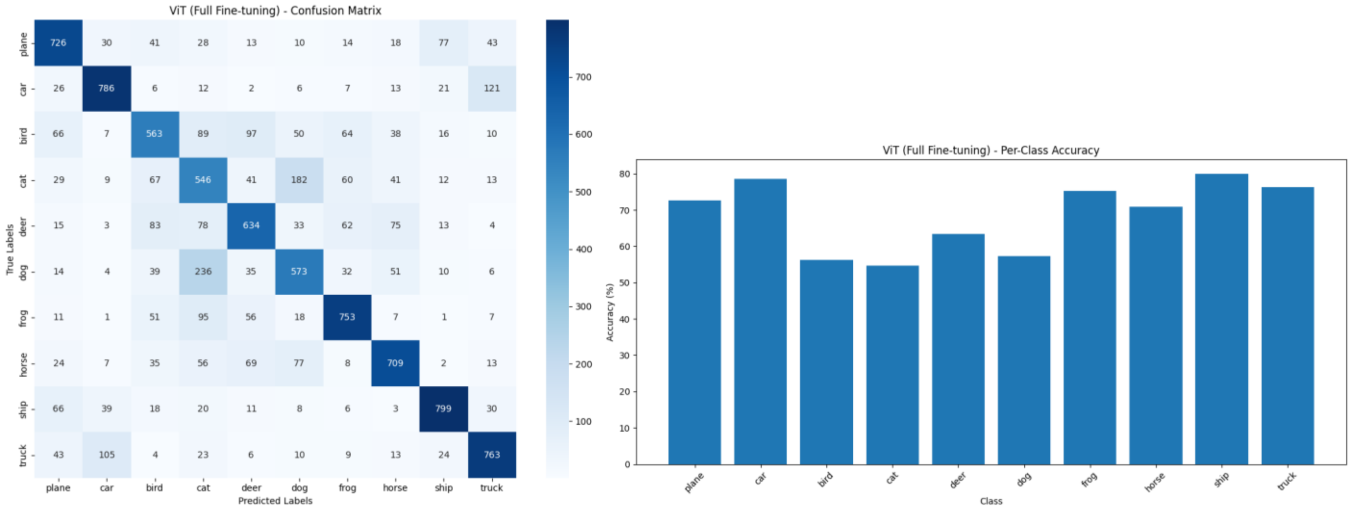


Figure 4. Training loss, accuracy and learning rate schedule over epochs for using only head fine-tuned Swin.

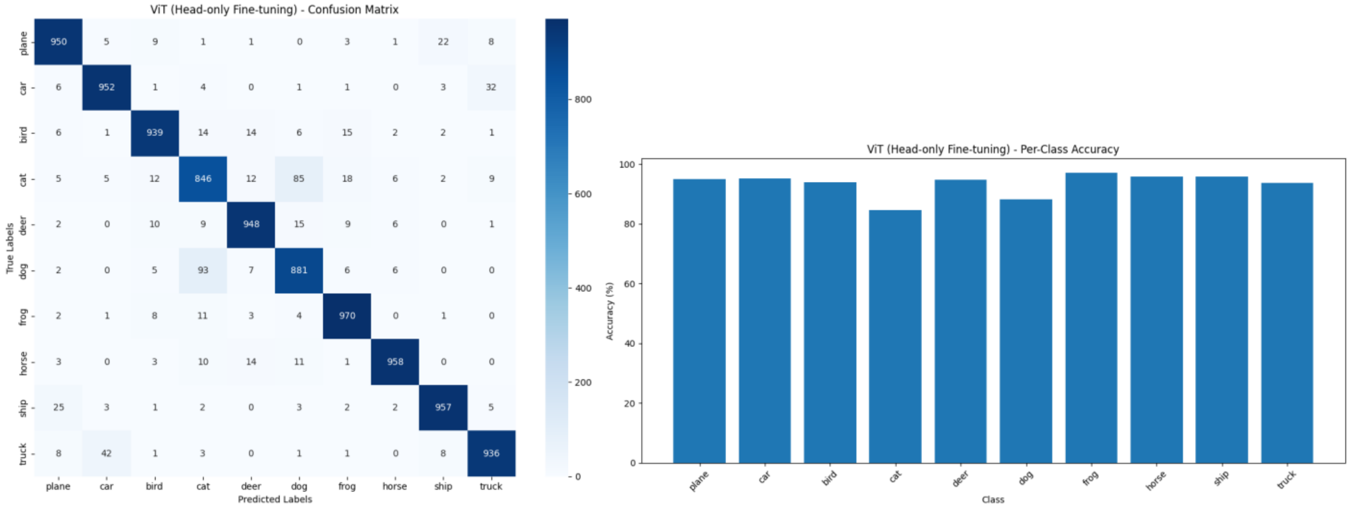
	Accuracy↑	Top-1 Error Rate↓	Average inference time per batch↓
ViT(full)	68.52%	31.48%	**3.18 ms**
ViT(head only)	93.37%	6.63%	3.26 ms
Swin(full)	18.03%	81.97%	13.37 ms
Swin(head only)	**94.21%**	**5.79%**	13.74 ms

Table 1. Accuracy, top-1 error rate and Average inference time per batch compared with four different models.



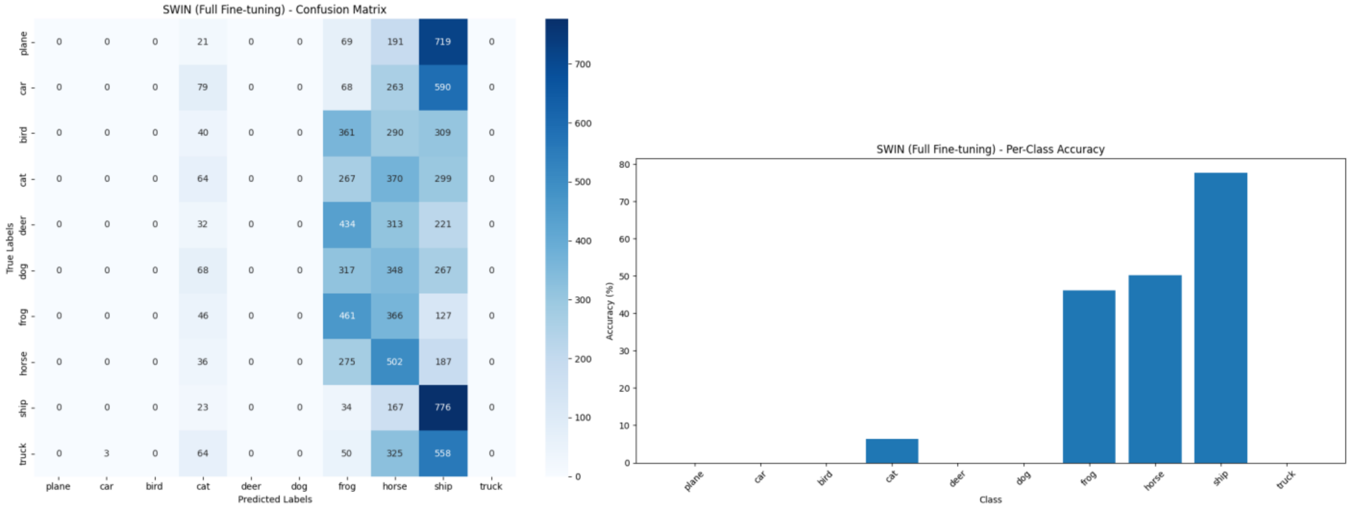
	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
plane	726	30	41	28	13	10	14	18	77	43
car	26	786	6	12	2	6	7	13	21	121
bird	66	7	563	89	97	50	64	38	16	10
cat	29	9	67	546	41	182	60	41	12	13
deer	15	3	83	78	634	33	62	75	13	4
dog	14	4	39	236	35	573	32	51	10	6
frog	11	1	51	95	56	18	753	7	1	7
horse	24	7	35	56	69	77	8	709	2	13
ship	66	39	18	20	11	8	6	3	799	30
truck	43	105	4	23	6	10	9	13	24	763

Figure 5. Confusion matrix and per class accuracy for using fully fine-tuned Vision Transformer



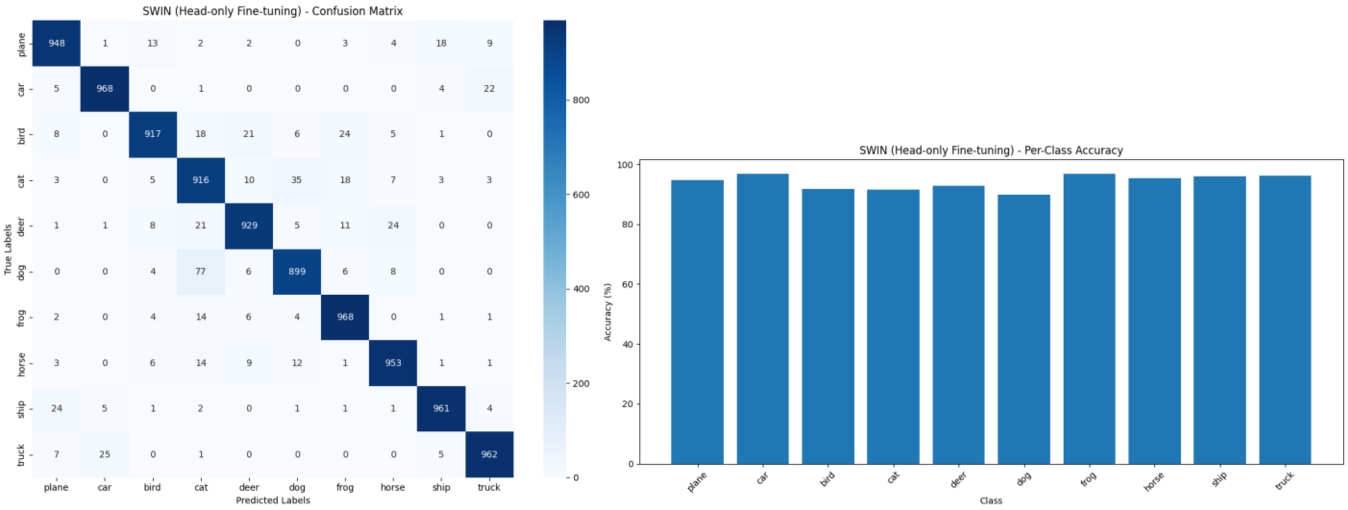
	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
plane	950	5	9	1	1	0	3	1	22	8
car	6	952	1	4	0	1	1	0	3	32
bird	6	1	939	14	14	6	15	2	2	1
cat	5	5	12	846	12	85	18	6	2	9
deer	2	0	10	9	948	15	9	6	0	1
dog	2	0	5	93	7	881	6	6	0	0
frog	2	1	8	11	3	4	970	0	1	0
horse	3	0	3	10	14	11	1	958	0	0
ship	25	3	1	2	0	3	2	2	957	5
truck	8	42	1	3	0	1	1	0	8	936

Figure 6. Confusion matrix and per class accuracy for using only head fine-tuned Vision Transformer.



	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
plane	0	0	0	21	0	0	69	191	719	0
car	0	0	0	79	0	0	68	263	590	0
bird	0	0	0	40	0	0	361	290	309	0
cat	0	0	0	64	0	0	267	370	299	0
deer	0	0	0	32	0	0	434	313	221	0
dog	0	0	0	68	0	0	317	348	267	0
frog	0	0	0	46	0	0	461	366	127	0
horse	0	0	0	36	0	0	275	502	187	0
ship	0	0	0	23	0	0	34	167	776	0
truck	0	3	0	64	0	0	50	325	558	0

Figure 7. Confusion matrix and per class accuracy using fully fine-tuned Swin.



	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
plane	948	1	13	2	2	0	3	4	18	9
car	5	968	0	1	0	0	0	0	4	22
bird	8	0	917	18	21	6	24	5	1	0
cat	3	0	5	916	10	35	18	7	3	3
deer	1	1	8	21	929	5	11	24	0	0
dog	0	0	4	77	6	899	6	8	0	0
frog	2	0	4	14	6	4	968	0	1	1
horse	3	0	6	14	9	12	1	953	1	1
ship	24	5	1	2	0	1	1	1	961	4
truck	7	25	0	1	0	0	0	0	5	962

Figure8. Confusion matrix and per class accuracy using only head fine-tuned Swin.

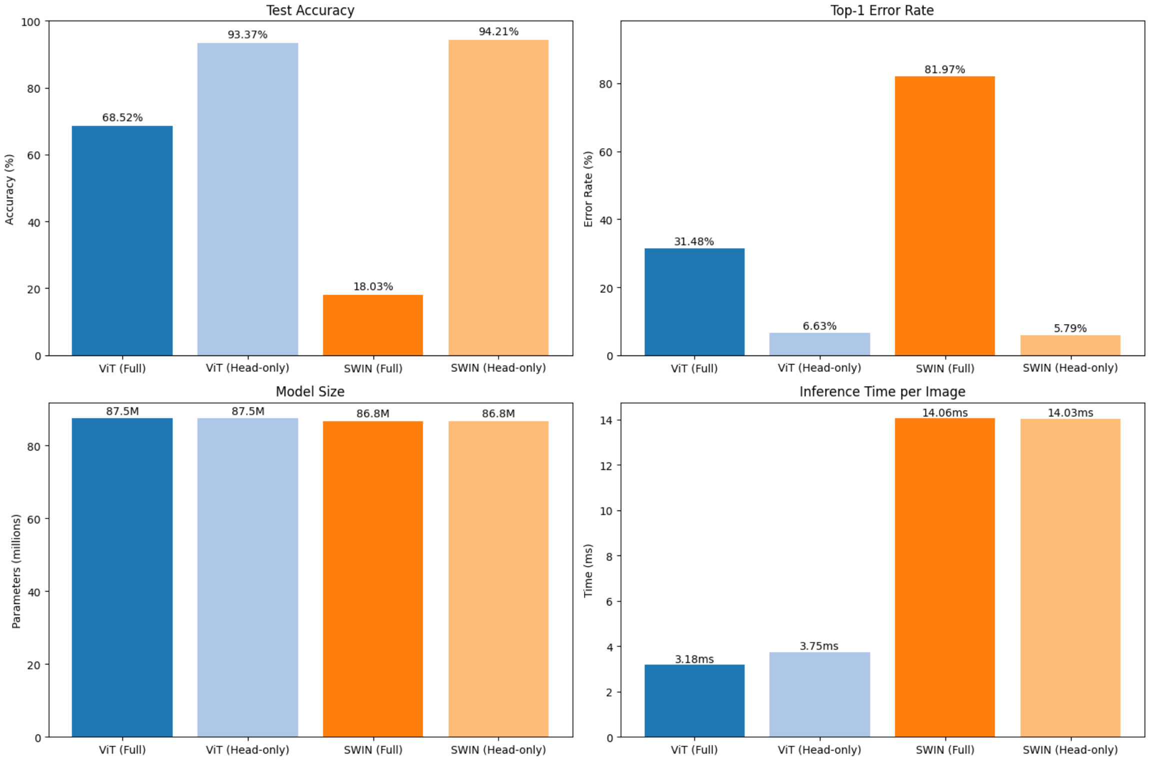


Figure9. Overall comparison with test accuracy, top1 error rate, model size and inference time per image.

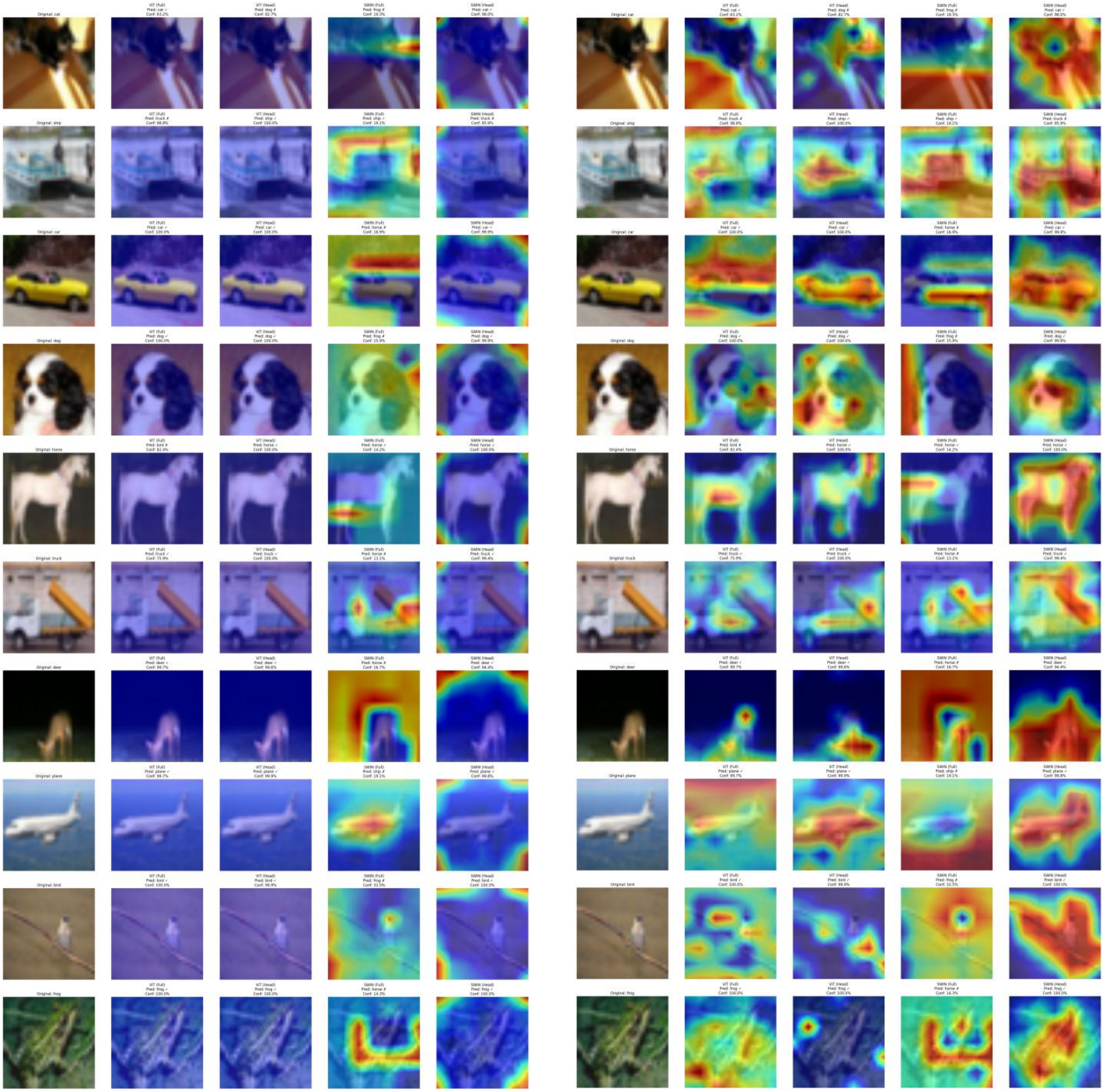

Original: cat
ViT (Full) Pred: cat ✓ Conf: 63.2%
ViT (Head) Pred: dog ✗ Conf: 82.7%
SWIN (Full) Pred: frog ✗ Conf: 19.3%
SWIN (Head) Pred: cat ✓ Conf: 98.0%
Original: ship
ViT (Full) Pred: truck ✗ Conf: 98.8%
ViT (Head) Pred: ship ✓ Conf: 100.0%
SWIN (Full) Pred: ship ✓ Conf: 19.1%
SWIN (Head) Pred: truck ✗ Conf: 85.9%
Original: car
ViT (Full) Pred: car ✓ Conf: 100.0%
ViT (Head) Pred: car ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✗ Conf: 16.9%
SWIN (Head) Pred: car ✓ Conf: 99.9%
Original: dog
ViT (Full) Pred: dog ✓ Conf: 100.0%
ViT (Head) Pred: dog ✓ Conf: 100.0%
SWIN (Full) Pred: frog ✗ Conf: 15.9%
SWIN (Head) Pred: dog ✓ Conf: 99.9%
Original: horse
ViT (Full) Pred: bird ✗ Conf: 82.4%
ViT (Head) Pred: horse ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✓ Conf: 14.2%
SWIN (Head) Pred: horse ✓ Conf: 100.0%
Original: truck
ViT (Full) Pred: truck ✓ Conf: 75.9%
ViT (Head) Pred: truck ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✗ Conf: 13.1%
SWIN (Head) Pred: truck ✓ Conf: 99.4%
Original: deer
ViT (Full) Pred: deer ✓ Conf: 99.7%
ViT (Head) Pred: deer ✓ Conf: 99.6%
SWIN (Full) Pred: horse ✗ Conf: 16.7%
SWIN (Head) Pred: deer ✓ Conf: 94.4%
Original: plane
ViT (Full) Pred: plane ✓ Conf: 99.7%
ViT (Head) Pred: plane ✓ Conf: 99.9%
SWIN (Full) Pred: ship ✗ Conf: 19.1%
SWIN (Head) Pred: plane ✓ Conf: 99.8%
Original: bird
ViT (Full) Pred: bird ✓ Conf: 100.0%
ViT (Head) Pred: bird ✓ Conf: 99.9%
SWIN (Full) Pred: frog ✗ Conf: 32.5%
SWIN (Head) Pred: bird ✓ Conf: 100.0%
Original: frog
ViT (Full) Pred: frog ✓ Conf: 100.0%
ViT (Head) Pred: frog ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✗ Conf: 14.3%
SWIN (Head) Pred: frog ✓ Conf: 100.0%
Original: cat
ViT (Full) Pred: cat ✓ Conf: 63.2%
ViT (Head) Pred: dog ✗ Conf: 82.7%
SWIN (Full) Pred: frog ✗ Conf: 19.3%
SWIN (Head) Pred: cat ✓ Conf: 98.0%
Original: ship
ViT (Full) Pred: truck ✗ Conf: 98.8%
ViT (Head) Pred: ship ✓ Conf: 100.0%
SWIN (Full) Pred: ship ✓ Conf: 19.1%
SWIN (Head) Pred: truck ✗ Conf: 85.9%
Original: car
ViT (Full) Pred: car ✓ Conf: 100.0%
ViT (Head) Pred: car ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✗ Conf: 16.9%
SWIN (Head) Pred: car ✓ Conf: 99.9%
Original: dog
ViT (Full) Pred: dog ✓ Conf: 100.0%
ViT (Head) Pred: dog ✓ Conf: 100.0%
SWIN (Full) Pred: frog ✗ Conf: 15.9%
SWIN (Head) Pred: dog ✓ Conf: 99.9%
Original: horse
ViT (Full) Pred: bird ✗ Conf: 82.4%
ViT (Head) Pred: horse ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✓ Conf: 14.2%
SWIN (Head) Pred: horse ✓ Conf: 100.0%
Original: truck
ViT (Full) Pred: truck ✓ Conf: 75.9%
ViT (Head) Pred: truck ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✗ Conf: 13.1%
SWIN (Head) Pred: truck ✓ Conf: 99.4%
Original: deer
ViT (Full) Pred: deer ✓ Conf: 99.7%
ViT (Head) Pred: deer ✓ Conf: 99.6%
SWIN (Full) Pred: horse ✗ Conf: 16.7%
SWIN (Head) Pred: deer ✓ Conf: 94.4%
Original: plane
ViT (Full) Pred: plane ✓ Conf: 99.7%
ViT (Head) Pred: plane ✓ Conf: 99.9%
SWIN (Full) Pred: ship ✗ Conf: 19.1%
SWIN (Head) Pred: plane ✓ Conf: 99.8%
Original: bird
ViT (Full) Pred: bird ✓ Conf: 100.0%
ViT (Head) Pred: bird ✓ Conf: 99.9%
SWIN (Full) Pred: frog ✗ Conf: 32.5%
SWIN (Head) Pred: bird ✓ Conf: 100.0%
Original: frog
ViT (Full) Pred: frog ✓ Conf: 100.0%
ViT (Head) Pred: frog ✓ Conf: 100.0%
SWIN (Full) Pred: horse ✗ Conf: 14.3%
SWIN (Head) Pred: frog ✓ Conf: 100.0%

Figure10. Grad Cam’s comparison with baseline(left) and change layer with method(right).

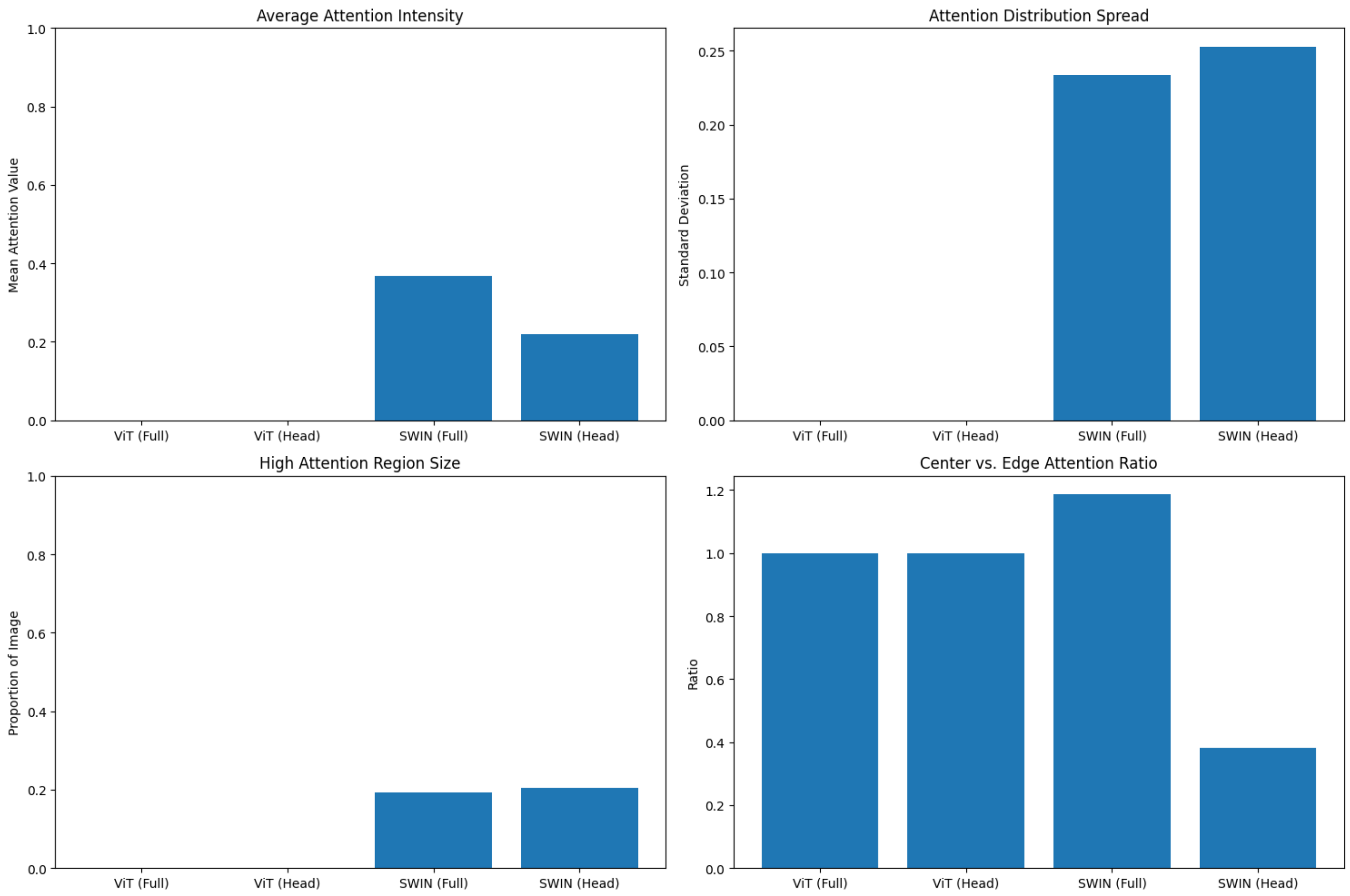


Figure11. Attention pattern analysis compared with four different models.

	Attn mean	Attn Std	High Attn Region Size	Center/Edge Ratio
ViT(full)	39.34%	25.99%	**20.41%**	1.1743
ViT(head only)	23.54%	26.64%	19.39%	**1.9129**
Swin(full)	**46.95%**	27.84%	**20.41%**	1.0716
Swin(head only)	44.54%	**32.87%**	**20.41%**	**1.6750**

Table 2. Attention mean, attention std, high attention region size and center/edge ratio compared with four different models.

Conclusion

Among the four model configurations, partially fine-tuned models (head and later layers only) outperformed fully fine-tuned ones in both ViT and Swin. Swin with head-only fine-tuning achieved the highest accuracy (94.21%), while fully fine-tuned models suffered from lower performance, likely due to overfitting on the small CIFAR-10 dataset. ViT was more computationally efficient, showing faster inference time due to model selection. Grad-CAM visualizations confirmed that head-only fine-tuned models preserved better attention patterns. Overall, partial fine-tuning is more effective for adapting large pre-trained models to small datasets like CIFAR-10.