

一、Data

1. Dataset：CIFAR-10，包含 10 個類別，每類別有 6000 張 32x32 彩色圖片
2. 前處理：所有圖片統一 resize 至 224x224，以符合 Vision Transformer 與 SWIN 模型的輸入需求。
3. Normalization：利用 Train dataset，計算得出 Mean = [0.4914, 0.4822, 0.4465]，STD = [0.2023, 0.1994, 0.2010]

二、模型設定

1. Vision Transformer (ViT)
 - i. 使用 timm 提供的 vit_base_patch16_224 Pretrain model
 - ii. 設定 head 為 nn.Linear(vit_model.head.in_features, 10)
2. SWIN Transformer
 - i. 使用 timm 提供的 swin_base_patch4_window7_224 Pretrain model
 - ii. 設定 head 為 in_channel=1024, num_class=10

三、Training

1. 在 CIFAR-10 訓練集上微調整個模型
2. 優化器：AdamW
3. 損失函數：CrossEntropyLoss
4. lr=1e-5
5. 學習率策略：每過 step_size=10 個 epoch，就將學習率乘上 gamma=0.1
6. 訓練周期：共訓練 5 個 epoch
7. 批次大小：32
8. ViT

```
Training Vision Transformer...
Training Epoch 1/5: 100%|██████████| 1563/1563 [03:11<00:00, 8.17it/s, Loss=0.11, Acc=96.67%]
[Epoch 1] Loss: 0.1097 | Accuracy: 96.67%
儲存最佳模型: ./checkpoints\best_model_epoch_1.pth
Training Epoch 2/5: 100%|██████████| 1563/1563 [03:11<00:00, 8.17it/s, Loss=0.0241, Acc=99.23%]
[Epoch 2] Loss: 0.0241 | Accuracy: 99.23%
儲存最佳模型: ./checkpoints\best_model_epoch_2.pth
Training Epoch 3/5: 100%|██████████| 1563/1563 [03:11<00:00, 8.15it/s, Loss=0.0148, Acc=99.51%]
[Epoch 3] Loss: 0.0148 | Accuracy: 99.51%
儲存最佳模型: ./checkpoints\best_model_epoch_3.pth
Training Epoch 4/5: 100%|██████████| 1563/1563 [03:11<00:00, 8.15it/s, Loss=0.0136, Acc=99.58%]
[Epoch 4] Loss: 0.0136 | Accuracy: 99.58%
儲存最佳模型: ./checkpoints\best_model_epoch_4.pth
Training Epoch 5/5: 100%|██████████| 1563/1563 [03:11<00:00, 8.14it/s, Loss=0.0114, Acc=99.64%]
[Epoch 5] Loss: 0.0114 | Accuracy: 99.64%
儲存最佳模型: ./checkpoints\best_model_epoch_5.pth
```

9. Swin

```
Training SWIN Transformer...
Training Epoch 1/5: 100%|██████████| 1563/1563 [04:00<00:00, 6.50it/s, Loss=0.149, Acc=96.09%]
[Epoch 1] Loss: 0.1495 | Accuracy: 96.09%
儲存最佳模型: ./checkpoints\best\_model\_epoch\_1.pth
Training Epoch 2/5: 100%|██████████| 1563/1563 [04:01<00:00, 6.48it/s, Loss=0.0296, Acc=99.11%]
[Epoch 2] Loss: 0.0296 | Accuracy: 99.11%
儲存最佳模型: ./checkpoints\best\_model\_epoch\_2.pth
Training Epoch 3/5: 100%|██████████| 1563/1563 [04:01<00:00, 6.48it/s, Loss=0.015, Acc=99.58%]
[Epoch 3] Loss: 0.0150 | Accuracy: 99.58%
儲存最佳模型: ./checkpoints\best\_model\_epoch\_3.pth
Training Epoch 4/5: 100%|██████████| 1563/1563 [04:01<00:00, 6.48it/s, Loss=0.0113, Acc=99.67%]
[Epoch 4] Loss: 0.0113 | Accuracy: 99.67%
儲存最佳模型: ./checkpoints\best\_model\_epoch\_4.pth
Training Epoch 5/5: 100%|██████████| 1563/1563 [04:00<00:00, 6.49it/s, Loss=0.008, Acc=99.77%]
[Epoch 5] Loss: 0.0080 | Accuracy: 99.77%
儲存最佳模型: ./checkpoints\best\_model\_epoch\_5.pth
```

四、實驗結果

	ViT	Swin
Training loss	0.0114	0.0080
Training Accuracy	99.64%	99.77%
Testing Accuracy	97.92%	98.69%

1. 兩個模型都在短短 5 個 epoch 內達到了超過 99% 的準確率，顯示模型初始化與微調策略設定良好。
2. 在 CIFAR-10 上表現皆接近完美，Loss 持續下降。
3. SWIN 模型在訓練集和測試集的最終準確率都略高（99.77% vs. 99.64%）、（98.69% vs. 97.92%）
4. ViT 的訓練速度較快(約 1 分鐘左右)
5. 如果想追求訓練速度可以使用 ViT，追求準確率可以使用 SWIN

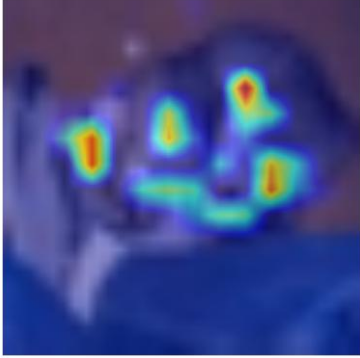
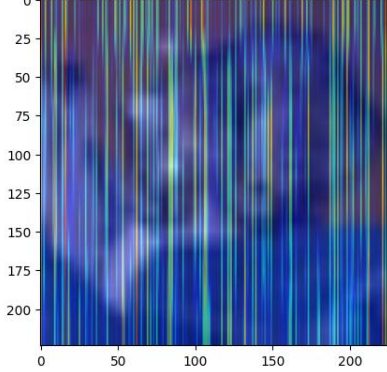
五、Grad-CAM (Gradient-weighted Class Activation Mapping)視覺化分析

1. 使用目的：透過 Grad-CAM，可以觀察模型在分類任務中「關注」的圖像區域，有助於解釋模型判斷依據與注意力分布。
2. 實作方式
 - i. 使用 CIFAR-10 測試集中隨機選取一張圖像作為分析對象，但由於測試集圖像有先經標準化處理，因此在視覺化前需進行反標準化（denormalization）。
 - ii. 使用 pytorch-gradcam 套件產生 Grad-CAM 熱區圖
 - iii. 為方便觀察，將熱區圖（CAM）疊加在原始圖像上顯示
 - iv. 對於 Vision Transformer，設定最後一層 block 的 norm1 層作為目標層，並實作 reshape_transform() 函數，將 patch token 轉為空間

結構，以便生成空間熱圖。

- v. 對於 SWIN Transformer，選取最後一個 block 的 norm1 層作為目標層（target_layer）。

3. 結果

ViT	Swin
	
塊狀（Block-like）的注意力熱區	條狀（Stripe-like）的注意力熱區
這是因為 ViT 採用固定大小的 patch（如 16×16）將圖像切分，並將每個 patch 作為一個 token 輸入至 Transformer 編碼器。此架構使得模型對於各個 patch 的輸出保持相對獨立，並能在全局語意理解的基礎上集中注意於具體且具語意性的區域，從而在 Grad-CAM 可視化上形成整塊式的關注模式。	SWIN Transformer 的 Grad-CAM 熱圖呈現出明顯的「直條紋狀熱區」，為縱向的線條分布。SWIN 採用移動視窗（Shifted Window Attention）機制進行區域性注意力運算，因此在每個 block 中僅能聚焦於特定的空間分割，導致其注意力圖無法完整集中於特定物體，較易出現規律性的條紋分布。此現象可能也與 SWIN 架構中層與層之間的區塊對齊方式有關，導致最終 Grad-CAM 在高層特徵中出現「帶狀關注」的模式。
ViT 模型的 Grad-CAM 熱圖偏好整體結構性區塊，顯示其強調語意層次的整合能力，適合需要全圖語意理解的場景。ViT 著重全局序列建模。	SWIN 模型則較偏好線條式的區域注意力，顯示其更著重於空間區域細節與局部運算，適合需要精細區域分析的任務。SWIN 更強調區域性與分層結構。