人工智慧模型設計與應用 Lab4

NM6121030 余振揚

1. Outline:

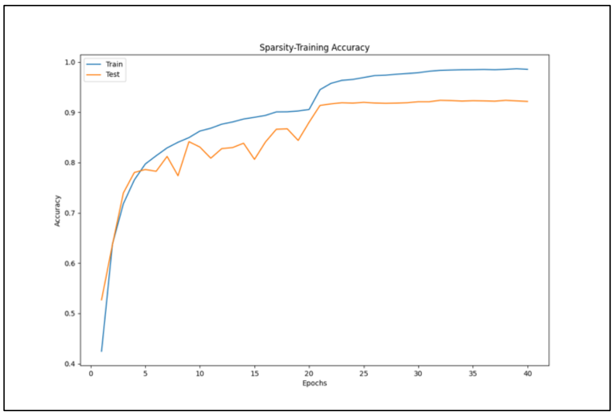
這次的 lab 主要是將 model 做 pruning，目的是在不犧牲效能的情況下盡量縮小模型。

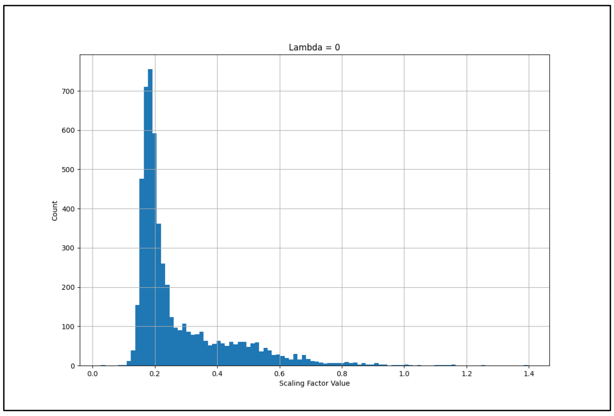
這次 train model 調整 λ 的順序為 0 → 1e-4 → 1e-5。

並將 λ=1e-5 的模型進行 prune ratio = 0.5 及 prune ratio = 0.9 的剪枝。

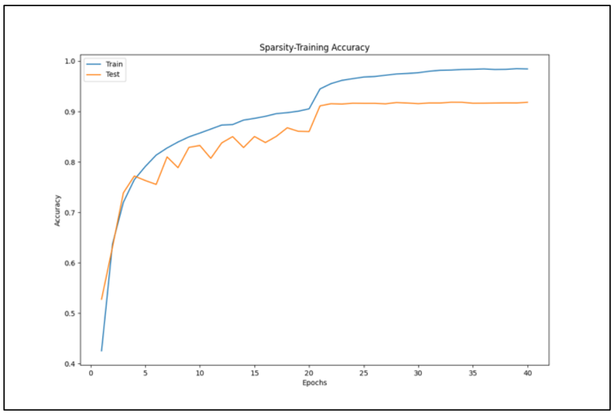
2. Sparsity-Training Accuracy and Scaling Factor Distribution with 3 Different λ value:

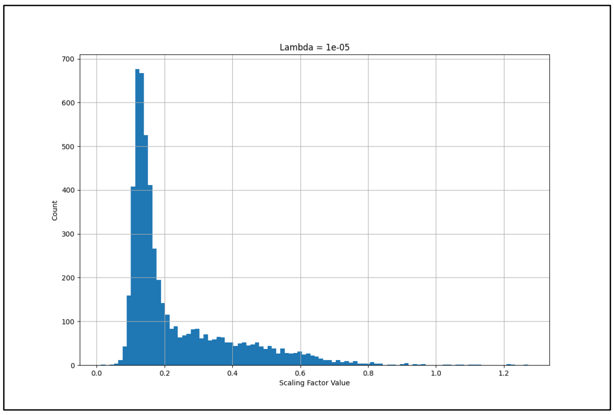
- λ = 0 :



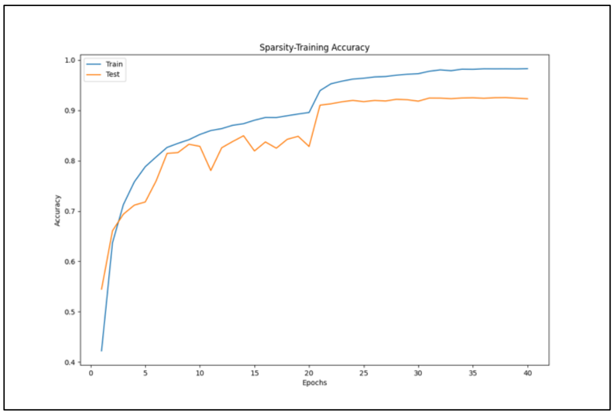


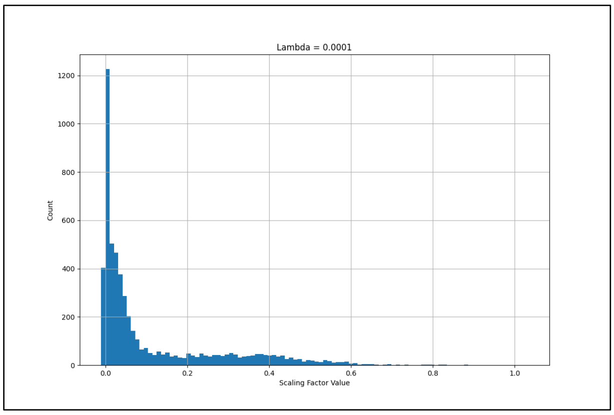
- λ = 1e-5 : （將以此模型結果進行 pruning）





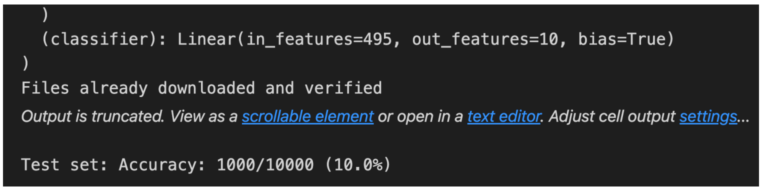
- λ = 1e-4 :





3. Model Test Accuracy with Different Prune Ratio:

- 50% Prune Ratio:

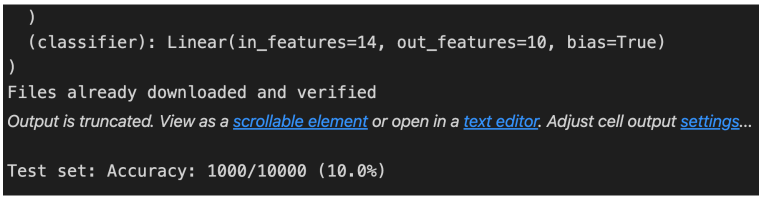


```
  )
  (classifier): Linear(in_features=495, out_features=10, bias=True)
)
Files already downloaded and verified
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
Test set: Accuracy: 1000/10000 (10.0%)
```

- 90% Prune Ratio:



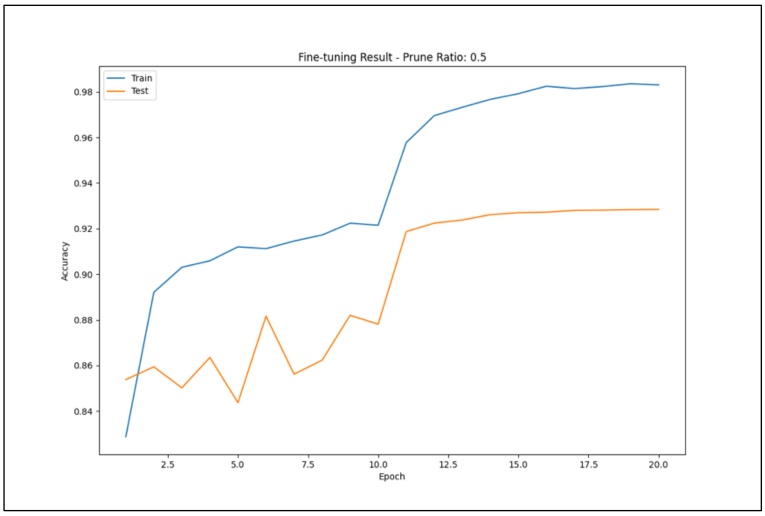
```
  )
  (classifier): Linear(in_features=14, out_features=10, bias=True)
)
Files already downloaded and verified
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

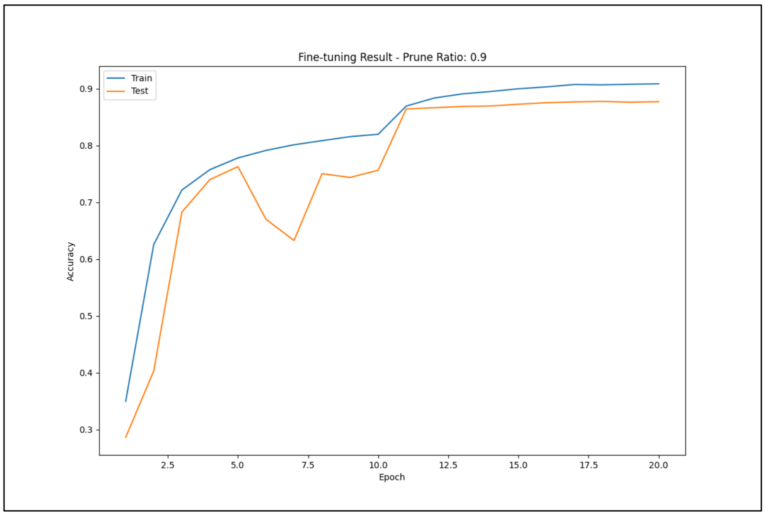
```
Test set: Accuracy: 1000/10000 (10.0%)
```

4. Accuracy of Fine-tuned Model with Different Prune Ratio:

- 50% Prune Ratio: (Test Accuracy Around 0.91)



- 90% Prune Ratio: (Test Accuracy Around 0.88)



5. **Model File Size (λ = 1e-5, Prune Ratio = 0.9):**

	model_best.pth	**model_prune.pth**	**model_prune_finetune.pth**
Size	160.4MB	1.8MB	3.5MB

6. **Feedback and Problem Encounter:**

可以明顯感覺到當 model 被 pruned 的比例越高，精度就會隨之下降。

在 pruning model(vggprube.ipynb)時，不曉得為何最終 Test set 的 Accuracy 永遠都是 1000/10000 (10%)，我與 Project 的組員也都很納悶這個問題，但還是找不太出原因。

7. **Reference:**

- https://github.com/foolwood/pytorch-slimming
- https://github.com/Eric-mingjie/network-slimming