應用於耳穴辨識的模型壓縮

Model compression for ear acupuncture

指導教授：藍崑展

專題成員：劉耿宏、洪翊豪

開發工具：Python、PyTorch、MMPose

測試環境：Ubuntu 24.04.1 LTS

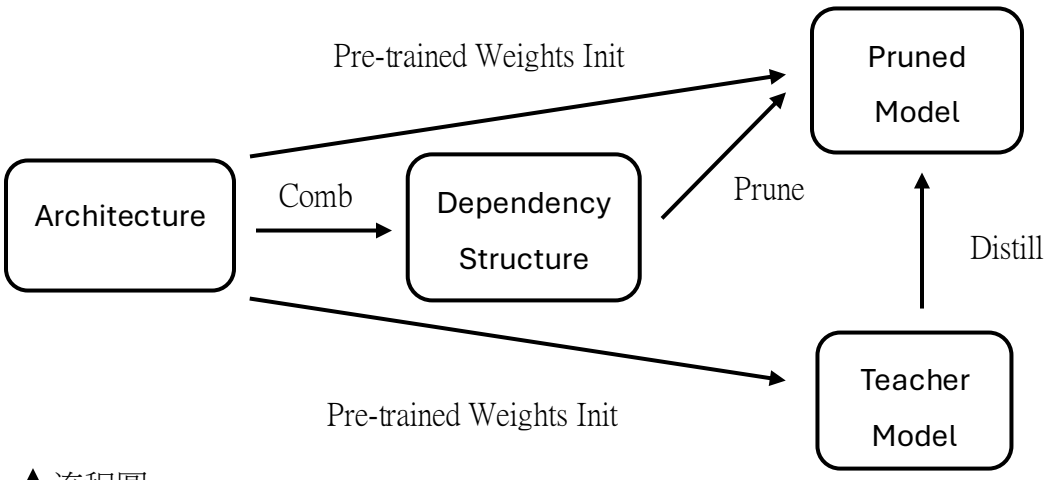
一、 簡介：

本專題旨在針對耳朵穴位點判斷模型，比較壓縮其比例多寡對於其表現影響。由於耳穴分布細密，模型需具備高準確度輸出；但在實際應用中，如行動裝置或即時診療輔助，也需兼顧模型體積與運行效率。

我們採用 MMPose 為基礎架構，並引入 CPD（Comb, Prune, Distill）壓縮技術，結合：

- 結構整合（Comb）：合併類似模組結構；
- 模型剪枝（Prune）：移除冗餘通道以降低運算量；
- 知識蒸餾（Distill）：以原模型為教師網路輔助訓練壓縮後模型。

此流程有效減少模型大小與推論時間，同時維持預測準確度，提升實際部署的可行性。



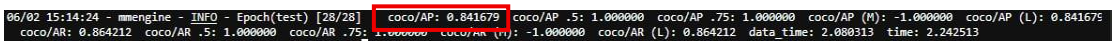
▲流程圖

二、 測試結果：

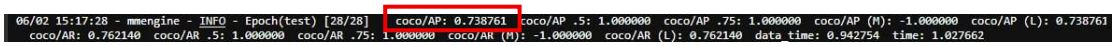
在本次實驗中，我們觀察到在未對模型進行操作前，所測出的準確度為 0.841679，而在進行完 prune 後，所測出的準確度為 0.738761，而後我們又將 prune 後的模型作為學生，以原模型作為老師進行 knowledge distillation，並觀察到準確度為 0.798534。

因此符合我們預測，在進行 prune 前的原模型的準確度為最高，而在進行 prune 後，剪枝掉一些參數使整體的判斷結果變差，而在進行 knowledge distillation 後，使得 KD 後的模型判斷結果與 prune 後相比，相對較高，但仍差於原模型的判斷結果。

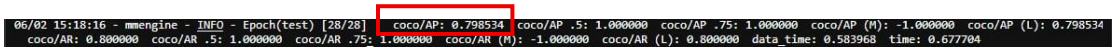
	Accuracy
原模型	0.841679
prune 後的模型	0.738761
KD 後的模型	0.798534



▲ 原模型測出的準確度



▲ prune 後模型測出的準確度



▲ KD 後模型測出的準確度