1. 怎麼判斷哪些 neurons 是比較沒用的?

A：

APoZ 是判斷神經元重要性的一種有效方法，能夠準確找到貢獻較低的神經元。

高 APoZ 值，神經元輸出多數為 0，對模型結果貢獻較低。

低 APoZ 值：神經元輸出非 0，對模型結果有較大貢獻。

2. 簡單介紹一下知識蒸餾的概念以及 soft label 有甚麼優點

A：

知識蒸餾 (Knowledge Distillation) 概念：

知識蒸餾是一種**模型壓縮**技術，其目的是通過一個大型、複雜的「**老師模型 (Teacher Model)**」來訓練一個較小、更高效的「**學生模型 (Student Model)**」，使學生模型在保持接近老師模型性能的同時，具有更小的參數量、更低的運算成本。

soft label 優點：

1.提供更多信息。

2.減少過擬合。

3.有效傳遞知識。

4.提升小模型性能。

3. 甚麼是假量化（quantized-aware training）

A：

通過在訓練過程中模擬量化操作，幫助模型適應量化後的運算限制，使得最終量化模型在性能和準確率上接近浮點模型。

在模型訓練期間引入量化操作（如截斷和取整）來模擬量化效應，讓模型逐步適應低精度的數據表現，從而在量化後性能損失最小。

優點：減少量化對準確度的負面影響， 適用於精度敏感應用。

缺點：訓練成本增加，且需要大量資料 和計算資源。

4. 為甚麼早停（early exiting）有用?

A：

1. 早停根據數據難度提前停止，避免為簡單樣本進行不必要的深層計算。
2. 簡單樣本可能在早期層已經足夠被正確分類，無需進入更深層。
3. 早停跳過部分層的計算，顯著降低整體推論時間。
4. 在資源受限的環境中，早停能減少計算和能耗。