學號：B05901168 系級：電機四　姓名：陳冠豪

1. 請從 Network Pruning/Quantization/Knowledge Distillation/Low Rank Approximation 選擇兩個方法(並詳述)，將同一個大 model 壓縮至同等數量級，並討論其 accuracy 的變化。 (2%)

   使用 Quantization 及 Knowledge Distillation 來做實驗.
   Quantization 的部分使用 HW3 已經 train 好的 model (4 layer CNN + 1 layer full-connected layer, 86% accuracy) 並將每個參數壓成 8bit, 把參數壓成 8bit 後還有做 fine tune, 方法為將 quantization 完的模型在去訓練 50 個 epoch, 每 10 個 epoch 會做一次 quantization.

   Knowledge Distillation 部分則是使用助教提供之 student net, teacher net 同樣使用 HW3 已經 train 好的 model 來做訓練. 共訓練 200 epoch 結果如下

	Quantization	Knowledge Distillation
Training accuracy	83.8%	85.51%
Validation accuracy	81.9%	82.37%

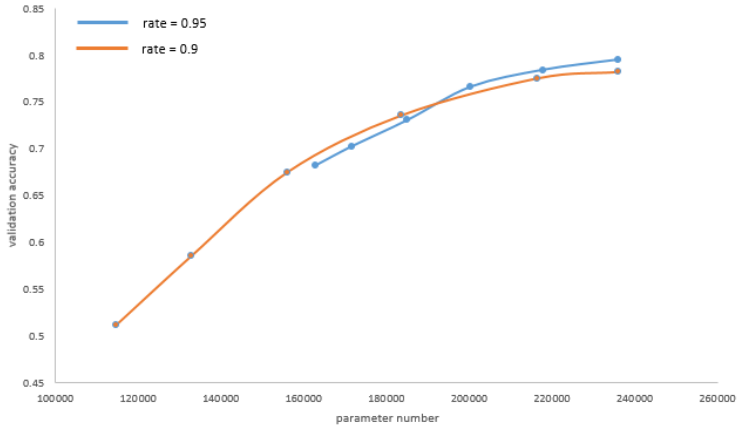
   Quantization 的結果跟原始的模型相比下降幅度較大, 而 Knowledge Distillation 則可以訓練到接近原本模型的正確率, 因此在模型大小有限制下,先使用 knowledge distillation 讓正確率上去再做 quantization 應該會是不錯的選擇.

   以下三題只需要選擇兩者即可，分數取最高的兩個。

2. [Knowledge Distillation] 請嘗試比較以下 validation accuracy (兩個 Teacher Net 由助教提供)以及 student 的總參數量以及架構，並嘗試解釋為甚麼有這樣的結果。你的 Student Net 的參數量必須要小於 Teacher Net 的參數量。 (2%)
   x. Teacher net architecture and # of parameters: torchvision's ResNet18, with 11,182,155 parameters.
   y. Student net architecture and # of parameters:
   a. Teacher net (ResNet18) from scratch: 80.09%
   b. Teacher net (ResNet18) ImageNet pretrained & fine-tune: 88.41%
   c. Your student net from scratch:
   d. Your student net KD from (a.):
   e. Your student net KD from (b.):

3. [Network Pruning] 請使用兩種以上的 pruning rate 畫出 X 軸為參數量，Y 軸為 validation accuracy 的折線圖。你的圖上應該會有兩條以上的折線。 (2%)

   從結果來看不管是用 rate = 0.95 or 0.9, 作前兩次對正確率的影響沒這麼大, 而如果使用 rate = 0.9 做 2 次 pruning 約可以把整體模型參數數量減少 20%, 算是蠻大的減少了



4. [Low Rank Approx / Model Architecture] 請嘗試比較以下 validation accuracy，並且模型大小須接近 1 MB。 (2%)
   a. 原始 CNN model (用一般的 Convolution Layer) 的 accuracy
   b. 將 CNN model 的 Convolution Layer 換成參數量接近的 Depthwise & Pointwise 後的 accuracy
   c. 將 CNN model 的 Convolution Layer 換成參數量接近的 Group Convolution Layer (Group 數量自訂，但不要設為 1 或 in_filters)

   All train with 150 epoch and same preprocess

	CNN	DP&PW	GC
Training acc	70.15%	79.20%	76.87%
Validation acc	68.22%	74.50%	75.60%

   結果 CNN 表現蠻差得, 因為要在 1MB 得限制下模型變得非常淺, 而 DP&PW 及 GC 則能夠有較完整得模型架構, 因此表現得較好一點,