學號：B07902047　系級：資工二 姓名：羅啟帆

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%)

我的小 model 主要實作了 Low Rank Approximation（DepthWise&PointWise cnn）和 knowledge distillation，最後的 model_state_dict 再透過 Quantization 壓低大小（32bit -> 8bit）。

- Low Rank Approximation：
  - 將一般的 cnn 拆成兩步驟，DepthWise Convolution 和 PointWise Convolution。DepthWise Convolution 只考慮同一個 channel 之間的關係，而 PointWise Convolution 只考慮同一個位置、不同 channel 的 pixel 之間的關係。
  - 我採用 8 層的 DW&PW cnn（每層的 filter 數量為 [16, 32, 64, 128, 256, 256, 256, 256]），加上 1 層的 fc layer。其中每一層 DW&PW cnn 的架構如下：
    - $\text{Conv2d}(\text{in\_channel}, \text{in\_channel}, \text{kernel\_size}, \ \text{padding} = 1, \text{stride} = 1, \text{groups} = 1)$
    - $\text{BatchNorm2d}(\text{in\_channel})$
    - ReLU6（ReLU + 把 ≥ 6 的值切到 6）
    - $\text{Conv2d}(\text{in\_channel}, \text{out\_channel}, \text{kernel\_size} = 1)$
- Knowledge distillation：
  - 在訓練小 model 的時候除了 ground truth label 外，再多考慮大 model 的的 logits。也就是說，小 model 不但要 minimize 自己跟 true distribution 的 cross entropy，也要 minimize 自己和大 model distribution 的 cross entropy。要注意的是大 model 的 logits 並沒有直接過 softmax，而是先同除一個常數（T）再過 softmax；這樣可以讓小 model 看到亂度比較高的 distribution（峰值比較不極端），會學得比較好。
  - 我採用的 hyperparameter 為 $T = 20, \alpha = 0.5$ （α為大 model loss 佔 final loss 的比例）
- Quantization：先將參數根據這條式子縮放：$\frac{param-min}{max-min} \times 255$，param 為參數，min/max 為參數最小/最大值。縮放之後，在把他強轉換成只有 8bit 大小的 uint8。讀取的時候就做相反的事情就好。

以下比較 validation accuracy 和 model 大小：

- 壓縮之前的 model（助教提供的 pretrained resnet18）：88.41%，model 大小為 43MB
- 壓縮之後的 model：81.92%，model 大小為 0.26MB

可以發現壓縮後的 model 只用了原本 0.6%的大小就達成原本 92%的準確率（81.41/88.41）。

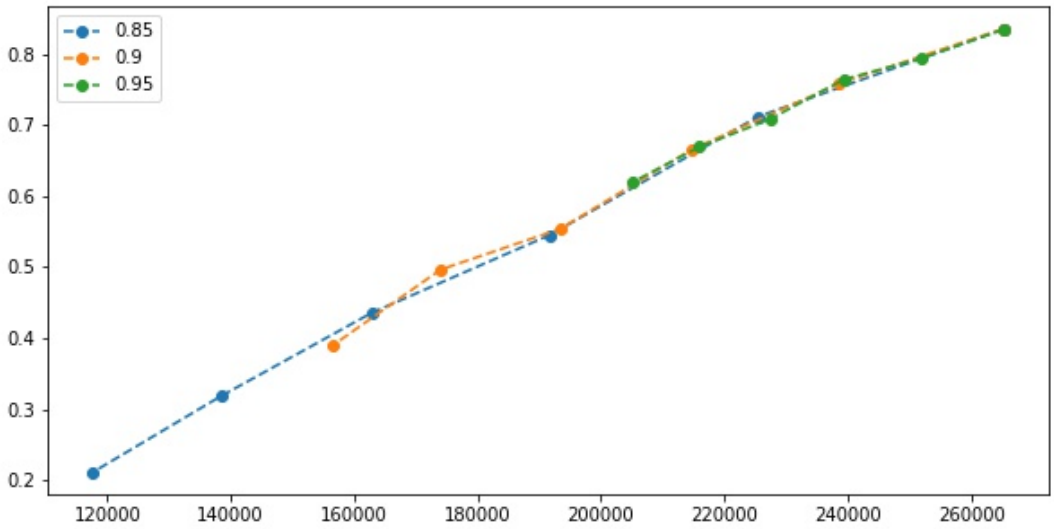
以下三題只需要選擇兩者即可，分數取最高的兩個。

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:跟 p1 的 model 一樣，這樣的參數量是 265,227。
a. Teacher net (ResNet18) from scratch: 80.09%
b. Teacher net (ResNet18) ImageNet pretrained & fine-tune: 88.41%
c. Your student net from scratch: 77.34%
d. Your student net KD from (a.): 79.91%
e. Your student net KD from (b.): 83.44%

- 可以從 cde 發現 KD 對 student net 的幫助不小，可見除了 ground truth 外 teacher net 的 logits 也是對分類來說有用的資訊。
- 由 be 和 ad 可以看出參數量較大的 teacher net 最後表現還是比較好（不過 d 其實很接近他的老師 a 了）

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%)



可以發現 pruning rate 的影響似乎不大，因在相似參數量的情況下 accuracy 也差不多。

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)