學號：b06901153 系級：電機三 姓名：林瑩昇

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%)

   (1) 原先 resnet18_fine_tuned 有 11,182,155 parameters，我用 weight quantization 的方式把 parameters 變成 8 bit，也就是 model 小了 4 倍。Val acc 也從原先的 88.41%掉到了 88.25%。
   (2) 我建了一個新的小 model，是把 modileNet 每一層改成一般的 convolution layer，有 2,233,995 parameters，我用 knowledge distillation 的方式讓他學習 resnet18_fine_tuned 的輸出，val acc 是 82.65%
   (3) 從上面兩個實驗可以看出 weight quantization 其實不會使原先的 model 正確率下降太多，主要是因為 trained model 的 weight 其實數字都蠻小的，當我們把原先 32bit quantize 成 8bit 其實只是犧牲掉少許的精準度，對正確率其實沒有太大的影響，而用 kd 雖然能夠用小一點的 model 學習大 model 的輸出，但正確率下降的比較多，畢竟 model 結構簡化了許多。

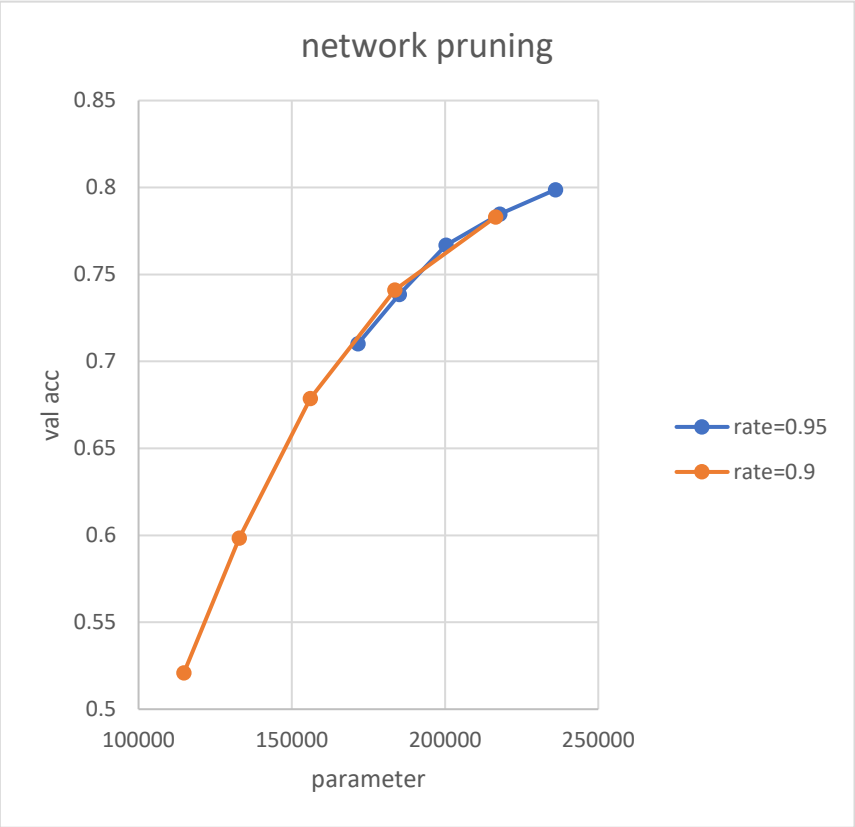
2. [Knowledge Distillation] 請嘗試比較以下 validation accuracy (兩個 Teacher Net 由助教提供)以及 student 的總參數量以及架構，並嘗試解釋為甚麼有這樣的結果。你的 Student Net 的參數量必須要小於 Teacher Net 的參數量。(2%)

x. Teacher net architecture and # of parameters:
torchvision's ResNet18, with 11,182,155 parameters
y. Student net architecture and # of parameters:
mobileNet, with 256,779 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: 78.16 %
d. Your student net KD from (a.): 79.79%
e. Your student net KD from (b.): 83.5%

(1) x & y: 因為用 mobileNet 的緣故，將 convolution layer 改成 depth-separable convolution，能夠大大降低計算量以及參數量
(2) a & b: 用 pretrained model 的表現自然比從頭訓練來的好
(3) c & d & e: student net 是一個相對比較小的 model，所以如果從頭訓練的效果可能比不上學習 teacher's net 的效果。而用不同正確率的 teacher's net knowledge distillation 出來的結果當然也有差別。

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%)

rate	parameter	val acc	rate	parameter	val acc
0.95	236066	0.7988	0.9	216518	0.7831
0.9025	217850	0.7848	0.81	183578	0.7411
0.8574	200354	0.7668	0.729	156084	0.6787
0.8145	185091	0.7385	0.6561	132858	0.5985
0.7738	171606	0.7102	0.5905	114810	0.521



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)