學號：b07902006 系級：資工二 姓名：林則仰

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%)

	大 model	model a (knowledge distallation)	model b (low rank approxiamtion)
Validation acc	87.23%	79.94%	81.82%
# of parameters	54.34M	4.32M	5.92M

在此題中我的大模型是hw3的模型（54.34M）改良而成，在validation 上的準確率有87.23%左右。

而所使用的第一個方法是knowledge distillation 的方法，使用一個較小且簡單的CNN模型，參數量約僅有4.32M，訓練過程中使用Adam來作為optimizer，並將其與大model的 KL Divergence Loss 加入損失函數中，來達到知識蒸餾的目的。最後在訓練了250個epoch之後在validation set上的準確率達到了79.94%，跟原本的大model（87.23%）還是有一段不小的差距。我認為會有這樣的原因可能是與teacher net的參數量差距過大，導致「學不好」的現象出現，亦或是小模型的參數量不足以讓它學習的起這麼多的資訊，因此我猜測或許需要使用TAKD的方法來解決這個問題。

第二個方法是 low rank approximation的方法，將CNN模型中convolution layer改為使用Depthwise & Pointwise 的方式來取代，以降低參數量，參數量大約為大model 的 1/9。在同樣訓練了250個epoch之後，在validtion set上得到的準確率為81.82%，較大模型來說準確率掉了將近5.5%左右。由這個實驗可以知道，雖然model b是由大model 壓縮而成，但其準確率還是會受其參數量的大小影響，但相較於使用一般的CNN+KD來說，效果還是比較好一些。

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: similar to mobile net, with 655,8172 parameters.

```python
def conv_bn(inp, oup, stride):
    return nn.Sequential(
        nn.Conv2d(inp, oup, 3, stride, 1, bias=False),
        nn.BatchNorm2d(oup),
        nn.ReLU(inplace=True)
    )

def conv_dw(inp, oup, stride):
    return nn.Sequential(
        nn.Conv2d(inp, inp, 3, stride, 1, groups=inp, bias=False),
        nn.BatchNorm2d(inp),
        nn.ReLU(inplace=True),

        nn.Conv2d(inp, oup, 1, 1, 0, bias=False),
        nn.BatchNorm2d(oup),
        nn.ReLU(inplace=True),
    )

self.cnn = nn.Sequential(
    conv_bn(  3,  32, 2),
    conv_dw( 32,  64, 1),
    conv_dw( 64, 128, 2),
    conv_dw(128, 128, 1),
    conv_dw(128, 256, 2),
    conv_dw(256, 256, 1),
    conv_dw(256, 512, 2),
    conv_dw(512, 512, 1),
    conv_dw(512, 512, 1),
    conv_dw(512, 512, 1),
    conv_dw(512, 512, 1),
    conv_dw(512, 512, 1),
    nn.AdaptiveAvgPool2d((1, 1)),
)
self.fc = nn.Sequential(
    nn.Linear(512, 11),
)
```

a. Teacher net (ResNet18) from scratch: 80.09%
b. Teacher net (ResNet18) ImageNet pretrained & fine-tune: 88.41%
c. Your student net from scratch: 79.88%
d. Your student net KD from (a.): 82.05%
e. Your student net KD from (b.): 84.73%

以上三個結果是我各只有train 60 個 epoch 後在validation set 上得到的準確率。可以發現很明顯地 [沒有teacher net 輔助的] < [有一個較弱的teacher net 輔助] < [有一個較強的teacher net輔助]。會有這樣的結果，我認為是因為較好的teacher net，在差不多的參數量中正確的資訊量本來就比較龐大，student net 比較不會出現「學到錯誤的東西」的情形，而是在一邊訓練的同時一邊將 teacher net 所學到的一些有關於class 之間的關聯性都納入了學習當中，因此可以在training data這樣的hard label 以外學到一些 teacher net 所給予的 soft label。而teacher net 越是強大，所給予student net訓練用的soft label 通常品質也會更好（連結不同class之間資訊的能力通常更好），student net 學得更好的機率也就會上升。

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%)

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)

	model a (group = in_filters)	model b (group = 1)	model c (group = 2)
validation acc	0.7635	0.8011	0.7859
model size	1.08MB	1.13MB	1.07MB

在此實驗中，很明顯的可以看出在大致相同的參數量之下，group數量越小的模型擁有越好的validation accuracy。因為實際上在相同參數量之下，group < in_filter 的模型所學到的東西是等同於一個更大架構的模型所學到的。以model a 以及 model b為例，由 knowledge distillation的理論可知在kernal = 3之下，model b 其實跟 9 倍大小的 model a學到的東西可能是差不多的，而顯然地group 數量介在1 跟 in_filter 之間的表現會介在兩者之間（model c 實際上跟一個大小介於 model a 跟 9倍 model a之間的模型學到的東西差不多），也與我的實驗結果相符。