

Machine Learning HW7

學號: B06902060 系級: 資工三 姓名: 鄒宗霖

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%)

	Original	Design Architecture	Network Pruning
accuracy	0.8052	0.7805	0.1493

本題我實作了 Design Architecture 以及 Network Pruning 將同一個大 model (2168203 parameters) 壓縮至同等數量級 (Design Architecture : 256779 parameters , Network Pruning : 254038 parameters) , 上表為 Original, Design Architecture, Network Pruning 三種不同的模型在 validation set 上的準確率。其中實作 Design Architecture 的方法為利用 Depthwise Separable Convolution 將參數量壓縮到原來的 1/9 , self.cnn 的架構如同第二題中的程式碼所示 , 把原本的 convolution layer 拆成 depthwise & pointwise layer ; 而實作 Network Pruning 的方法為將原本模型中第 3 ~ 6 層的 channels 數量減少為 0.23 倍。從上表中我們可以看到經過 Design Architecture 後的模型雖然準確率有所下降 , 但與原模型差異不大 ; 然而經過 Network Pruning 後的模型準確率掉的很誇張 , 因此 Design Architecture 既能有效減少參數量 , 又能使模型保持原來的準確率 , 為實作 Network Compression 較佳的選擇。

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 11182155 parameters

y. Student net architecture and # of parameters: 256779 parameters

```
# Student net architecture
bandwidth = [16, 32, 64, 128, 256, 256, 256, 256]
# self.cnn 1-th layer
nn.Conv2d(3, bandwidth[0], 3, 1, 1),
nn.BatchNorm2d(bandwidth[0]),
nn.ReLU6(),
nn.MaxPool2d(2, 2, 0),
# self.cnn 2~8-th layer (with i = 1 to 7)
nn.Conv2d(bandwidth[i], bandwidth[i], 3, 1, 1, groups = bandwidth[i]),
nn.BatchNorm2d(bandwidth[i]),
nn.ReLU6(),
nn.Conv2d(bandwidth[i], bandwidth[i+1], 1),
nn.MaxPool2d(2, 2, 0),
```

- o 上面程式碼為 Student net 的架構 , 利用 Depthwise Separable Convolution 的技術把原本第 2 ~ 8 層的 convolution layer 拆成 depthwise & pointwise layer , 將參數量壓縮到原來的 1/9

a. Teacher net (ResNet18) from scratch: 80.09 %

b. Teacher net (ResNet18) ImageNet pretrained & fine-tune: 88.41 %

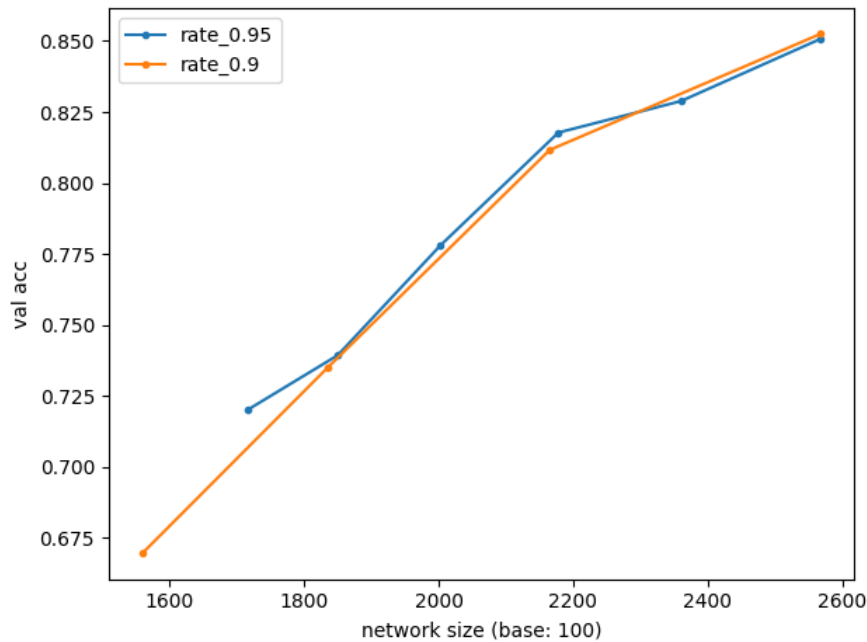
c. Your student net from scratch: 78.29 %

d. Your student net KD from (a.): 80.56 %

e. Your student net KD from (b.): 81.52%

上述為各個模型在 validation set 上的準確率，可以看到沒有經過 Knowledge Distillation 的 Student net 準確率較低，可能的原因是 Teacher net 不只告訴學生正確答案是甚麼，還加以解釋哪些類別間有比較接近的關係、接近的程度為何，因此學生學習成果較佳；從 d, e 我們可以看到 Knowledge Distillation from Teacher net (ImageNet pretrained & fine-tune) 的學生準確率較高，可能的原因是經過 ImageNet 預訓練以及微調參數的 Teacher net 給出的機率分布較準確，學生自然學的比較好。

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%)



上圖為對 testing accuracy 最高的 student net 實作 Network Pruning (pruning rate 0.95 , 0.9) 的結果，可以看到在相同的參數量下，pruning rate 0.95 的模型 validation accuracy 高了一些，因此可以推論：每次剪枝一些再微調再剪枝，比起一次剪枝很多效果來的佳。