說明：請各位使用此 template 撰寫 report，如果想要用其他排版模式也請註明題號以及題目內容（請勿擅自更改題號），最後上傳至 GitHub 前，請務必轉成 PDF 檔，並且命名為 report.pdf，否則將不予計分。

中英文皆可，但助教強烈建議使用中文。

----------------------------------------閱讀完以上文字請刪除---------------------------------------

學號： 系級： 姓名：

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%)

|  |  |  |  |
| --- | --- | --- | --- |
| Orig bin | Acc on val | Compressed acc on val | Compressed acc on kaggle |
| student\_custom\_small | 0.8137 | 0.8087 | 0.8344 |
| student\_model\_202ep | 0.8207 | 0.8155 | 0.8332 |
| student\_model\_deeper\_200ep | 0.8283 | 0.8277 | 0.8446 |
| student\_model\_deeper\_203ep | 0.8388 | 0.8362 | 0.8494 |
| student\_model\_deeper\_205ep | 0.8429 | 0.8376 | 0.8470 |

以下三題只需要選擇兩者即可，分數取最高的兩個。

1. [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:

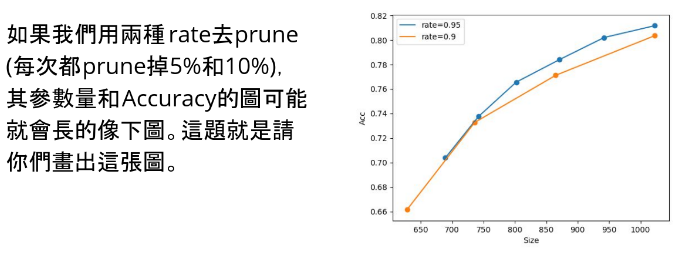
d. Your student net KD from (a.):

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | Validation acc | Model size | Parameters |
| a. Teacher net (ResNet18) from scratch | 0.8012 | 44,788,712 | 11,182,155 |
| b. Teacher net (ResNet18) ImageNet pretrained & fine-tune | 0.8851 | 44,788,712 | 11,182,155 |
| Provided student\_custom\_small | 0.8137 | 1047430 | X |
| Student net from scratch |  | 1047574 |  |
| Student net KD from a |  | 1047574 |  |
| Student net KD from b | 0.8213 | 1047574 | student\_model\_200ep\_512 |

1. [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%)

(Skip) (refer to hw slide)

被prune掉的neuron越多，雖然model的大小變小了運算速度也會比較快，但accuracy也會有顯著的下降，因為被prune過的model並不能完全代表原本的model，且刪掉neuron的依據是按重要性，刪越多(rate越大)越有可能刪到重要的，因此當一直迭代下去越剪枝越多(rate\*rate\*rate...)，pruned model和原model的失真就越大。

1. [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)

再勸三個一樣大的model比較acc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Validation acc | Model size | Paramters | Model structure |
| Standard Convolution Layer |  |  |  |  |
| Depthwise & Pointwise Convolution Layer |  |  |  |  |
| Group Convolution Layer |  |  |  |  |