學號：B05902010　系級：資工四 姓名：張頌平

ML HW7 report

1. 請從 Network Pruning/Quantization/Knowledge Distillation/Low Rank Approximation/Design Architecture 選擇兩者實做並詳述你的方法，將同一個大 model 壓縮至接近相同的參數量，並紀錄其 accuracy。 (2%)

About my big model:
- architecture: VGG16 from hw3
  - linear layer(25088 -> 4096, 4096->1024, 1024->128, 128->11)
- parameters: 121,815,627
- validation acc: 0.90430
- test acc: 0.90974(public)/ 0.91517(private)

Design Architecture:
- parameters: 63,133,483
- total epoch: 120
- optimizer: 80(Adam) + 40(SGD)
- learning rate: 7e-5(first 80 epoch), 2e-3(last 40 epoch)
- ***architecture:***
  - 將原先VGG16的每層convolution都改成
    - depthwise convolution(每一個channel針對對應的Kernel各自做convolution)
    - batchnorm2D
    - ReLU
    - pointwise convolution(對Depthwise Convolution的結果做1*1的卷積計算)
  - 將linear layer改成(25088 -> 2400, 2400-> 512, 512->11)
- validation acc: 0.82519

Knowledge Distillation:
- Student model architecture: VGG11
  - linear layer(25088 -> 2048, 2048->512, 512->11)
- parameters: 61,662,987
- total epoch: 120
- optimizer: 80(Adam) + 40(SGD)
- learning rate: 9e-5(first 80 epoch), 2e-3(last 40 epoch)
- a = 0.5
- K = 20
- validation acc: 0.89266
- Method: 在訓練時不僅計算logits和原本hard labels的cross entropy loss，也希望從大model那邊學習到它是如何去預測分類的，因此在訓練時會freeze大model的參數，並且計算小\model和大model之間預測分類機率的KL Divergence當成 soft loss，兩者相加後再進行optimize。

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 from TA, 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: 77.988%
d. Your student net KD from (a.): 79.942%
e. Your student net KD from (b.): 82.507%

Hyperparameter:
- total epoch: 200
- optimizer: 150(AdamW) + 50(SGD)
- learning rate: 7e-4(first 150 epoch), 2e-3(last 50 epoch)
- a = 0.5
- K = 20

我使用的student架構為助教提供的student net，也就是MobileNet v1，其架構原本總共有七層convolution 2D，但為了節省參數量，在每一層的convolution 2D都改成了Depthwise Convolution(每一個channel針對對應的Kernel各自做convolution) + Pointwise convolution(對Depthwise Convolution的結果做1*1的卷積計算)。

從上表可以看到 a, b, c, d, e分別的 validation accuracy，可以發現d的結果非常接近原model(a) 所達到的正確率，而 e 的結果雖然較c, d來的好上一些，卻離他原本的teacher model(b)的準確率還有一定的距離。造成這樣的差別，我覺得主要原因是因為b是有經過ImageNet pretrained的model，因此他學習了很多這個food dataset裡面沒有的東西，而我們在做Knowledge Distillation的時候，只讓student model透過food dataset的內容來更新參數，因此，雖然我們已經有soft label來幫助我們更接近原本的teacher model，但終究還是不會像都用同一個dataset訓練出來的model一樣，這麼接近teacher model的準確率。

在我自己實際傳上kaggle的結果當中，其中我使用的teacher model是我在hw3所訓練出來的VGG16和VGG13，因為hw3也都是用food dataset所訓練，因此最後我經過Knowledge Distillation和Quantization將其壓縮到8bit後，validation 和 public test set的準確率都僅下跌了1%左右，這也是我前面會覺得在正常的Knowledge Distillation訓練過程，和teacher用同一個dataset訓練出來的model會更接近原本teacher的準確率的原因。

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%)

Ans:
原model是我經過Knowledge Distillation後訓練出的student model，其參數量為 256,779，validation accuracy為0.8914，pruning rate分別設為0.95和0.9，對於每個pruning rate分別進行5次的pruning，並且在每個pruning的當下fine-tune 5個epoch，並且將其最好val acc的model存下來並記錄其acc，詳細的model參數和準確率如下圖～