學號：R08922167 系級：資工碩一 姓名：曾民君

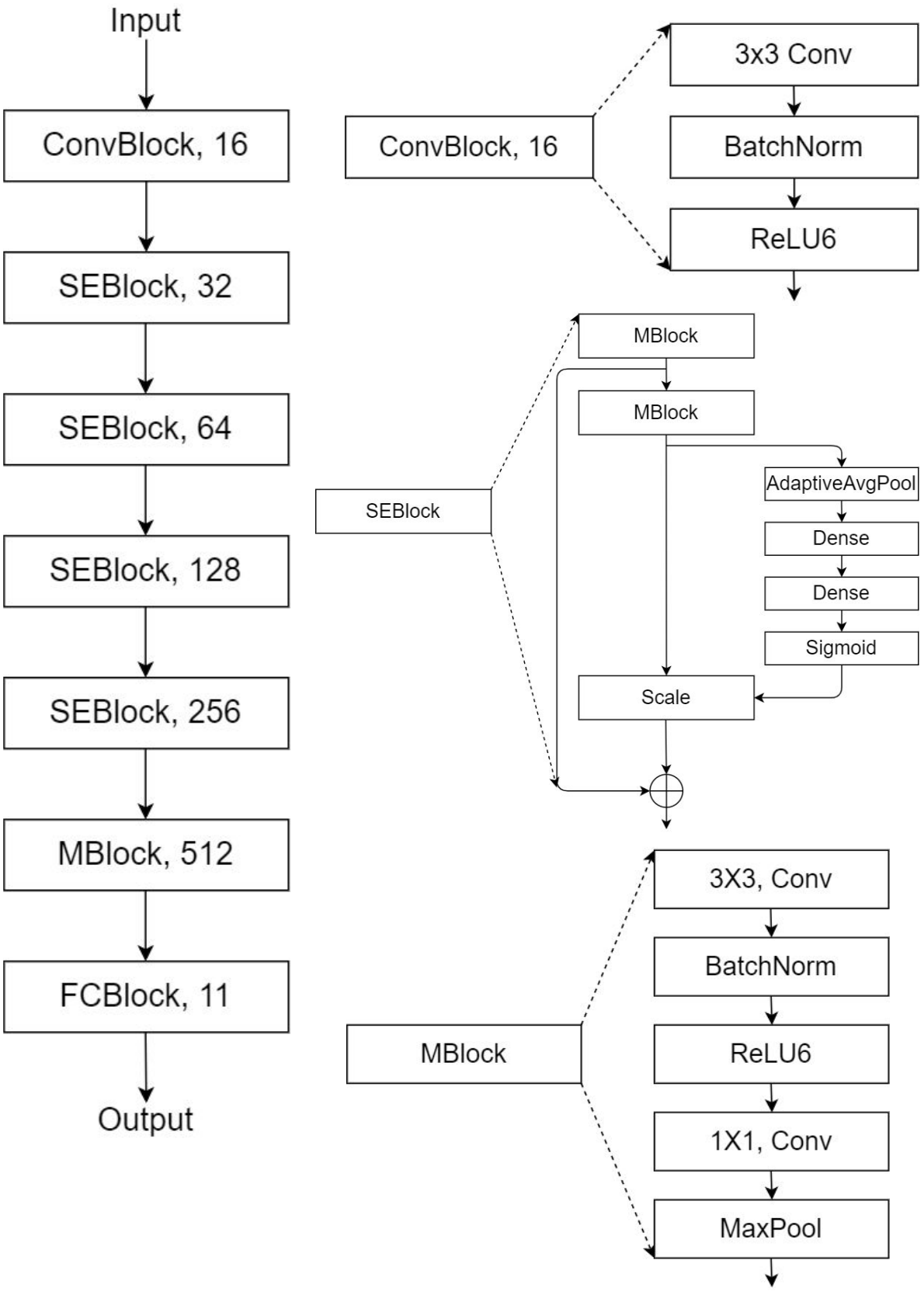
1. 請從 Network Pruning/Quantization/Knowledge Distillation/Low Rank Approximation 選擇兩個方法(並詳述)，將同一個大 model 壓縮至同等數量級，並討論其 accuracy 的變化。 (2%)

   Ans: 選擇使用 Low Rank 與 Knowledge Distillation，其中 Low Rank 部分坊校 mobile net 與 seresidual net，以 densewise 與 pointwise 組合成的 mobile block 取代原本的 convolutional block，此方法若都只使用 3x3 kernel 情況下，大約可以省下 75% ~ 85% 的空間 (單層與原conv比)，另外再後面幾層的 conv layers 再使用 grouping，大約又可再降 50 ~ 75% (單層)。但上述做法無法達到 300 k 以下，所以我在training 完後只存 float 16 為單位的 model，大約整體可以降 40% ~ 45 %的空間。 Knowledge Distillation 部分，實作 squeeze and excitation layers 加入到模型中，希望這部分可以加強因為使用 inception layer 數較多可能導致較早資訊遺忘的問題，另外 loss function 策略為前半段使用 soft/hard 混和的 loss，後半段就單純使用 hard loss 來 fine tune model，learning rate 也會隨時間進行遞減。以上兩者合用後的結果再 test accuracy 可以到 0.84339。

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%)

Ans:

Student Net 架構如下



a. Teacher net (ResNet18) from scratch: 80.09%
b. Teacher net (ResNet18) ImageNet pretrained & fine-tune: 88.41%

另外用此 model 分別 KD from a 跟 b 的配置如下:

- batch_size: 64
- n_epochs: 400
- optimizer: sgdm (lr = 0.1 -> 0.01, m=0.9)
- loss strategy: hard + soft (1 ~ 250epochs), hard (251 ~ 400epochs)

c. Your student net from scratch: 78.48%
d. Your student net KD from (a.): 81.83%
e. Your student net KD from (b.): 83.14%

實做過程中，有發現若 loss 方面只是單純只是依照 teacher net 進行學習，最終結果會表現不好，改善方式為 training 前半段的 epochs 使用 hard loss 與 soft loss 混和，後半段在使用 hard only 的 loss 進行訓練，這樣的方式大概會提昇 1% ~ 2% 的 accuracy。 Optimizer 的learning rate 部份為了加速訓練，開頭使用 0.1 的 learninmg rate，隨著時間遞減到 0.01，這樣原本訓練大約要 500 個 epochs 收斂結果可以壓到 400 個 epochs（實際上 500 epochs 結果好一點點）

3. 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)

   Ans: 統一的配置如下
   - Teacher net (ResNet18) ImageNet pretrained & fine-tune: 88.41%
   - batch_size: 64
   - n_epochs: 100
   - optimizer: adam(lr=1e-3)
   - loss strategy: hard + soft

Model	Accuracy
原始 CNN	69.37%
Depthwise & Pointwise	80.42%
Group Convolution	72.10%

原始 CNN 相對於使用壓縮過後的 models ， 在表現上差很多，可能的原因為要壓到 1MB 左右的大小的話，原始 CNN 層數會很少，導致整張影像資料萃取不完整，另外跟老師上課提到的比較多層的 model 和一層很寬的 fully connect 一樣的關係。這個實驗再次證明了，在差不多參數量的狀況下，較深的 model 表現的比較淺的 model 好。另外有發現如果重從第一層開始就使用壓縮過的 layer 會 train 不太起來，可能的原因為前面 cnn layers 在取特徵時，使用壓縮過的 layer似乎會提取不夠資訊量。