ML-HW7 Network Compression Report

學號：b06901066 系級：電機三 姓名：孟妍

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%)

--------------------------------------------------------------

˙Origin model:

原始model就是一個有8層convolution layer的CNN model，最後接上一層FC layer，參數量為2168203，架構如下：

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

(1) Architecture Design + Weight Quantization:

將原本的model的convolution Layer換成Dw & Pw後，參數量就小很多，只剩下256779，架構如下：

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

用這個小model train 200個epoch，觀察其training的validation accuracy，大約只會收斂到0.75左右，經過WQ把model壓到268471 bytes，上傳kaggle得到accuracy為0.81589

(2) Knowledge Distillation + Weight Quantization

把小model拿去對teacher\_resnet18.bin(acc= 88.41%)的 model做knowledge distillation，train 200個epoch，觀察其training的validation accuracy，就可以收斂到0.79左右，再把train好的小model再去做WQ，參數量不變。上傳kaggle得到accuracy為0.84399

\*比較：在同樣的model架構與同樣參數設定下，有做Knowledge Distillation的準確率就提升不少，而且上傳kaggle的成績竟然比validation accuracy高不少，但原本有做Architecture Design的model表現也沒有非常差，因此若需要減少參數量可以先使用Architecture Design換成Dw & Pw convolution，要得到更好的準確率可以再選擇一個好model做KD。另外，因為此次作業有限制上傳的model大小必須在300000 bytes以內，因此我都還有再用WQ壓過限制，但從結果看起來也許WQ並不會對準確率造成太大影響。

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:

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.):

--------------------------------------------------------------

y. Student net 使用的是助教提供的architecture design的colab的model架構，參數量為256779，架構如下：

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

c. 把student net從頭train 200 epoch，validation accuracy大概只會到75％左右

一張含有 文字, 窗戶, 女性, 電腦 的圖片

自動產生的描述

d. 使用student net對a.的model (teacher\_resnet18\_from\_scratch.bin acc=80.09%)做KD，同樣train 200個epoch，validation accuracy就明顯比c.好很多，大約可到79％

一張含有 窗戶, 大, 電腦, 白色 的圖片

自動產生的描述

e. 使用student net對b.的model (teacher\_resnet18.bin acc= 88.41%) 做KD，同樣train 200個epoch，validation accuracy又比d.更好一些，大約可以到82％

一張含有 窗戶, 桌, 大, 白色 的圖片

自動產生的描述

\*比較：使用accuracy最高的teacher model做knowledge distillation，validation accuracy最高，實際在kaggle的準確率也最高，因為student model本身架構較小，所以從頭train的結果並沒有很好，但做了knowledge distillation，準確率就能提高不少。

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%)

--------------------------------------------------------------

這題我用了0.95,0.9,0.85三種pruning rate，每次pruning完有再finetune兩個epoch。下圖是對助教提供的student\_custom\_small.bin做pruning所得到的結果：

一張含有 地圖 的圖片

自動產生的描述