HW8 Report

學號：B06901020 系級：電機二 姓名：張恆瑞

1. 請比較你本次作業的架構，參數量、結果和原 HW3 作業架構、參數量、結果做比較。(1%)

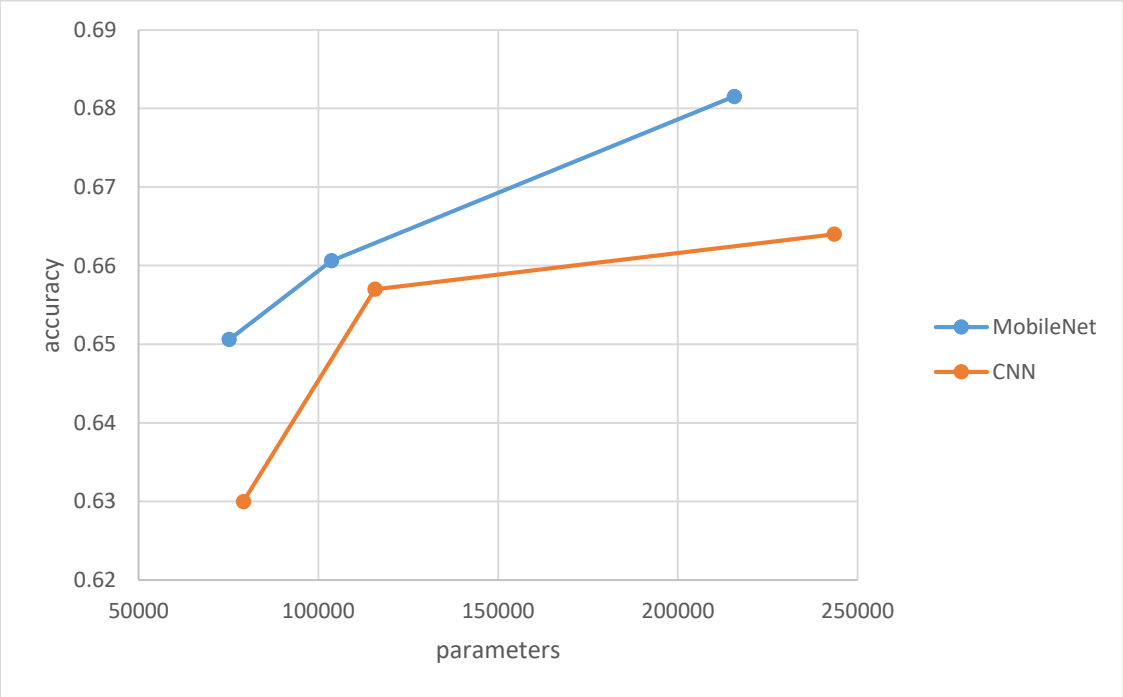
模型	CNN model	Mobile Net

架構	nn.Conv2d(1, 64, 4, 2, 1),
	nn.BatchNorm2d(64),
	nn.LeakyReLU(0.2),
	nn.Conv2d(64, 64, 3, 1, 1),
	nn.BatchNorm2d(64),
	nn.LeakyReLU(0.2),
	nn.MaxPool2d(2, 2, 0),
	nn.Dropout2d(p=0.25),
	
	nn.Conv2d(64, 128, 3, 1, 1),
	nn.BatchNorm2d(128),
	nn.LeakyReLU(0.2),
	nn.Conv2d(128, 128, 3, 1, 1),
	nn.BatchNorm2d(128),
	nn.LeakyReLU(0.2),
	nn.MaxPool2d(2, 2, 0),
	nn.Dropout2d(p=0.25),
	
	nn.Conv2d(128, 256, 3, 1, 1),
	nn.BatchNorm2d(256),
	nn.LeakyReLU(0.2),
	nn.Conv2d(256, 256, 3, 1, 1),
	nn.BatchNorm2d(256),
	nn.LeakyReLU(0.2),
	nn.MaxPool2d(2, 2, 0),
	nn.Dropout2d(p=0.25)
	
	nn.Linear(256*3*3, 1024),
	nn.BatchNorm1d(1024),
	nn.LeakyReLU(0.2),
	nn.Dropout(p=0.5),
	nn.Linear(1024, 512),
	nn.BatchNorm1d(512),
	nn.LeakyReLU(0.2),
	nn.Dropout(p=0.5),
	nn.Linear(512, 7)
		def conv_bn(inp, oup, stride):
		  return nn.Sequential(
		  nn.Conv2d(inp, oup, 3, stride, 1, bias=False),
		  nn.BatchNorm2d(oup),
		  nn.ReLU(inplace=True))
		
		def conv_dw(inp, oup, stride):
		  return nn.Sequential(
		  nn.Conv2d(inp, inp, 3, stride, 1, groups=inp, bias=False),
		  nn.BatchNorm2d(inp),
		  nn.ReLU(inplace=True),
		  nn.Conv2d(inp, oup, 1, 1, 0, bias=False),
		  nn.BatchNorm2d(oup),
		  nn.ReLU(inplace=True),)
		
		self.model = nn.Sequential(
		  conv_bn(1, 32, 2),
		  nn.Dropout(0.1),
		  conv_dw(32, 32, 1),
		  nn.Dropout(0.1),
		  conv_dw(32, 64, 1),
		  nn.Dropout(0.1),
		  conv_dw(64, 64, 1),
		  nn.Dropout(0.1),
		  conv_dw(64, 100, 2),
		  nn.Dropout(0.2),
		  conv_dw(100, 120, 2),
		  nn.Dropout(0.2),
		  conv_dw(120, 480, 2),
		  nn.AvgPool2d(7),)
		
		self.fc = nn.Linear(480, 7)

參數量	4038279	93139
Val acc	0.67816	0.67913
Public	0.66926	0.66341
Private	0.65338	0.65059

我重新 train 了幾個 hw3 的 model ensemble 成一個大的 teacher model，再利用 *knowledge distillation* (temperature : 2.5) 把 teacher model 對於 training set 的 output 餵給小的 mobile net，讓他學會如何做到同樣的事情。很明顯看得出來，參數量不到四十分之一的 student model 也可以學到同樣的事情。

2. 請使用 MobileNet 的架構，畫出參數量-acc 的散布圖（橫軸為參數量，縱軸為 accuracy，且至少 3 個點，參數量選擇時儘量不要離的太近，結果選擇只要大致收斂，不用 train 到最好沒關係。）(1%)
3. 請使用一般 CNN 的架構，畫出參數量-acc 的散布圖（橫軸為參數量，縱軸為 accuracy，且至少 3 個點，參數量選擇時儘量不要離的太近，結果選擇只要大致收斂，不用 train 到最好沒關係。）(1%)



4. 請你比較題 2 和題 3 的結果，並請針對當參數量相當少的時候，如果兩者參數量相當，兩者的差異，以及你認為為什麼會造成這個原因。(2%)

第二和三題中，我都是使用 data augmentation 和 knowledge distillation 來做訓練，目標都是讓 student model 學會 teacher model 的行為，而且都有 train 到大約收斂 (100 個 epoch)，都有盡量讓不同參數量的 model 可以 fit 在 training set 上，因此直接比較準確率應該是可行的。可以看得出來，MobileNet 和 CNN 參數量相同時，MobileNet 的效果會明顯比較好，就我觀察，不論是在 training set 還是 testing set 上，MobileNet 的準確率都較 CNN 的高。關於這個現象，我認為可能原因為 MobileNet 是使用 depthwise separable 的 convolution，跟一般 CNN 不同的是，depthwise separable 的 convolution layer 的本身架構可以用比較少的參數及比較快的速度來做到一般 CNN 可以做到的事情。因此一般 CNN 需要比較龐大的架構才能訓練出好的結果，可見 MobileNet 果然名不虛傳，model 又小又可以跑得快，相當符合”Mobile”這個稱號。