機器學習實務與應用

Homework #11
Due 2019 May 20 9:00AM

Exercise 1: CPU Time of DNN model inference

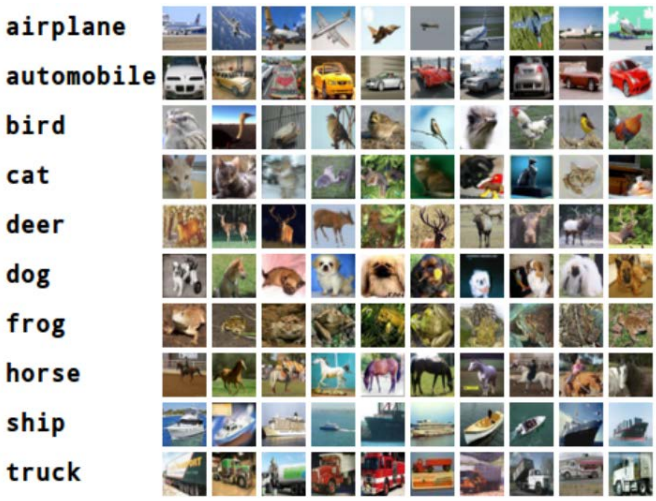
Test Model: VGG16、Resnet50、mobilenet

請利用提供的程式碼，比較各個 Key CNN models 進行多張圖片 Inference 的時間長短，比較參數量對於執行時間的影響。(請利用 time 函式進行時間測量)

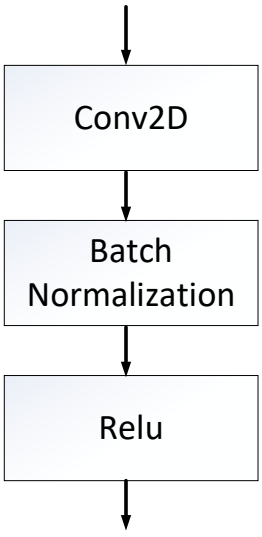
Exercise 2:Model Compression

Dataset:

Cifar-10 由 60000 張 32*32 的 RGB 彩色圖片，共 10 個分類。training data 共 50000 筆，testing data 共 10000 筆。



Conv Block:



type	Kernel size(or pooling size)	Output channel	stride	Padding
Conv Block	3x3	32	1x1	The same
Conv Block	3x3	32	1x1	The same
MaxPooling	2x2	--	2x2	The same
Dropout	25%			
Conv Block	3x3	64	1x1	The same
Conv Block	3x3	64	1x1	The same
MaxPooling	2x2	--	2x2	The same
Dropout	25%			
Conv Block	3x3	128	1x1	The same
Conv Block	3x3	128	1x1	The same
MaxPooling	2x2	--	2x2	The same
Dropout	25%			
Conv Block	3x3	128	1x1	The same
Conv Block	3x3	128	1x1	The same
MaxPooling	2x2	--	2x2	The same
Flatten				
Dence(FC)	Output size=256			
Dropout	25%			
Dence(FC)	Output size=128			
Dropout	25%			
Dence(FC	Output size=10			

請回答:

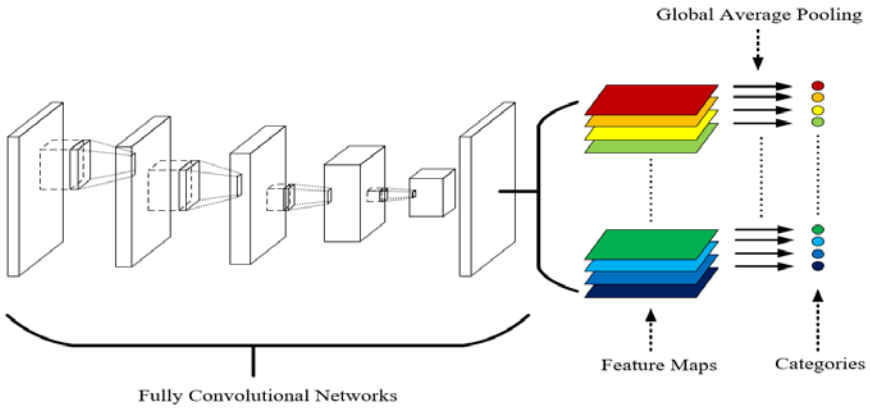
(1) 請依據上面表格建立模型架構，並且計算參數量的總個數(含 kernel 和 bias)，將計算過程列出來，並與 model.summary()的結果比對是否吻合。

(2) 觀察 training 後的結果圖，training &validation data 的 accuracy 和 loss，將結果截圖下來。

(3) 請利用下述說明的壓縮方法，**將模型架構重新訓練(Re-training)後**，在精確度在 80%以上的情況下，將問題(1)的模型架構進行模型壓縮，並且比較(1)壓縮前後的模型精確度，(2)計算參數量的總個數(含 kernel 和 bias)，(3)觀察 training 後的結果圖，training &validation data 的 accuracy 和 loss，將結果截圖下來，(4)在 CPU Inference 執行時間 。

一. 使用 Global Average Pooling 取代 FC

傳統的神經網路始由兩個部分所構成，特徵提取的卷積層與種類分類運算的全連接層，由 model.summary() 可以看出權重值主要是集中在 FC，利用 GAP 取代 FC 來降低權重值數量，來加快神經網路的訓練。



二. Reduced Number of Output Channels

影響卷積層的權重值數量，有三個個要素，分別為 Input Channel、Output Channel 與 Kernel size，因此減少 output channels 數量能夠降低整體模型的參數量。

三. 加入 1x1 Convolution layers

如果持續裁減每一層卷積層的輸入與輸出影像數量(Reduced Number of Output Channels)，很有可能會因此損失許多影像特徵，導致最後深度學習神經網路辨識準確度下降，1x1 的卷積運算將多張輸入特徵圖上的特徵結合後輸出較少張數的特徵圖，達到減少影像張數的效果同時不減小特徵圖尺寸，此作法比起直接裁減影像數量的做法對整體神經網路的辨識準確度影響較小，加上 1x1 的 Conv 可使神經網路架構變得更深，有可能使得最後的影像辨識能力有所提升。