

ML HW3 REPORT

b02902041 徐朝駿

1. (1%) Supervised learning:

這一部分只使用了 cnn 配上 labeled data 來做 training，主要是使用連續通過兩層的 convolutional layer 來獲得較好的效果。雖然從 loss 上來看 30 個 epoch 就已經收斂，但是發現繼續下去到 60 個 epoch 還可以有更好的成果，大約是 64%到 66%的進步。

2. (1%) Semi-supervised learning (1):

在這階段使用的是 self training，一開始用了第一提的 model 對 unlabeled data 做 predict，取 softmax 後機率大於 0.99 的就當做他為那個分類在進入 training。但是後來找到了一個 imagedata generator，這個 function 的功能就是使你放入的 data 進行些微小的變化，像是旋轉，上下左右移動之類，這樣的等於間接增加了很多的 training data，也是因為用了這個方法後，就讓正確率大幅的上升，來到了 80%。

3. (1%) Semi-supervised learning (2):

我採用 autoencoder/kmeans clustering 當做這題的作法，Autoencoder 主要是對於 unsupervised data 有不錯的作用。我採用的是 convolutional layer 疊成的 autoencoder，中間輸出的 bottle neck feature 為 (8, 4, 4)。Training 完 autoencoder 後就把 labeled data 和 unlabeled data 丟入 encoder 獲得 feature，之後把 label data 的 mean 放入 kmean 作為 init。在江全部的 feature 放入做 training。之後就可以使用 testing data 做 prediction。但是最終的結果竟然只有 24%，雖然比亂猜的 10%來得高，但是效果很差，不知道是我哪裡有出現問題了嗎？

4. (1%) Compare and analyze your results

因為 autoencoder 成績有點糟糕，所以暫時不列入評價。我想對於 semi-supervised 和 supervised 來做比較的話，最簡單的就是茲量料的差距，在擁有夠多的資料之後，可以容許誤差的空間就會變得比較大。再加上我在之後又加上了 image generator 所以資料量的差距可能不只有 10 被之差。原先沒有使用 image generator 的時候最高達到 72%。可以觀察得出，果然在做 deep learning 時，data 的數量決定你 model 的高度。