Machine Learning Homework#3

b02901120 羅志軒

實驗流程

本次作業我們利用cifar-10提供的5000張labeled圖片和45000張unlabeled圖片分別用三種方式（surpervised, self-training, clustering）訓練辨識圖片種類的模型

Supervised Learning:

在這次作業中，我將labeled的5000筆資料中的500筆資料當作validation data（每個class取50筆），並利用keras內建函式庫中的ImageDataGenerator來做data augmentation，將原本labed data擴增為45000筆，另外，所有的圖片都經過normalization。

結構上，我以cifar10_cnn範例為基礎做修改，最主要的改進是在每層layer的activation後加上BatchNormalization，使最終output的結果不會特別受到某一層參數的影響。

Self-training:

self-training利用以上已經訓練好的模型來分類 unlabeled data，若某一筆unlabeled data的分類結果足夠明確（即predict得到的信任值趨近於1），則會將此筆unlabeled data加到原training data中，產生數量更多的training data，利用新的training data訓練夠多次後，再繼續利用原training data訓練出新的模型（減少新的training data中label錯誤的情況對模型的傷害），重複以上步驟以提昇模型正確率，這裡我使用的信任值為0.99，並重複以上步驟3次。

Clustering (by autoencoder):

Clustering的基本概念是透過分析資料利用資料的特性將資料群事先分成一些類別，方便模型在訓練時能更簡單和迅速的完成資料分類，在這裡我使用autoencoder將圖片經過encode的方式取出feature，這類似於將圖片投影到另一個不同維度的feature空間，利用這些feature搭配DNN結構以簡化訓練過程，最後再結合上面的self-training，將預測結果大於信任值的unlabeled data經過encode加到training data中，重複訓練。

實驗分析

Supervised Learing:

這裡我們就模型結構、data augmentation和batch normalization的有無來討論validation accuracy的高低，這裡的validation accuracy為50 個epoch中最高的：

(**mp: maxpooling**)

model架構(每層 layer filter 數)	data augmentation	batch normalization	validation accuracy
(32, mp, 64, mp)	no	no	0.582
(32, mp, 64, mp)	yes	no	0.671
(32, mp, 64, mp)	no	yes	0.558
(32, 32, mp, 64, 64,	yes	yes	0.678

mp)			
(64,64,mp,128,128, mp)	yes	yes	0.748
(64,64,mp,128,128, mp, 256, 256, mp)	yes	yes	0.754

Self-training:

這裡我們就self-training重複的次數和信任值的高低做討論，如下：

iteration	reliant value	accuracy
3	0.95	0.78
3	0.99	0.788
3	0.999	0.772
4	0.99	0.764
5	0.99	0.75

Clustering:

這裡我們就autoencoder的結構和training model的結構做討論，以下為20個iteration後的結果，如下：

(**mp: maxpooling**)

autoencoder(每層layer filter數)	training model(DNN中每層的neuron數)	accuracy
(16, mp, 8, mp, 8)	(128, 64, 32)	0.396
(32, mp, 16, mp, 8)	(128, 64, 32)	0.422
(64, mp, 32, mp, 16)	(256, 128, 64, 32)	0.398

結論

總體來說，surpervised learning正確率上升的最穩定，且加上data augmentation後可以一定程度的降低overfitting的機率，在這個基礎上使用self-training若信任值適當也可以略為增進正確率，但若過於信賴原model對於unlabeled data的預測，可能會影響原model的判斷，造成正確率下降。

而autoencoder方面，或許是嘗試的encode方式不夠多，並沒有利於分類的feature產生，且大部分時候連最基礎的surpervised learning也無法讓validation正確率上升（train data accuracy上升迅速許多），這方面可能還需要類似於PCA的方法來改進。