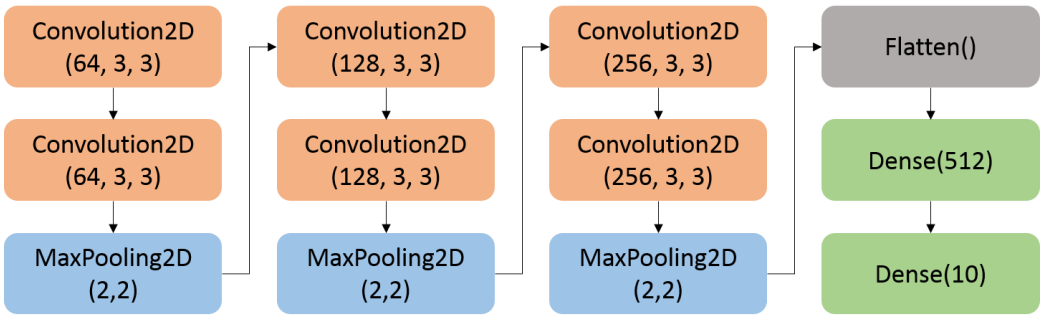
機器學習(HW3)　姓名：袁培傑　學號：B03901134

Supervised learning

整個 CNN 由 input_img 到 output classes，架構如下圖所示：



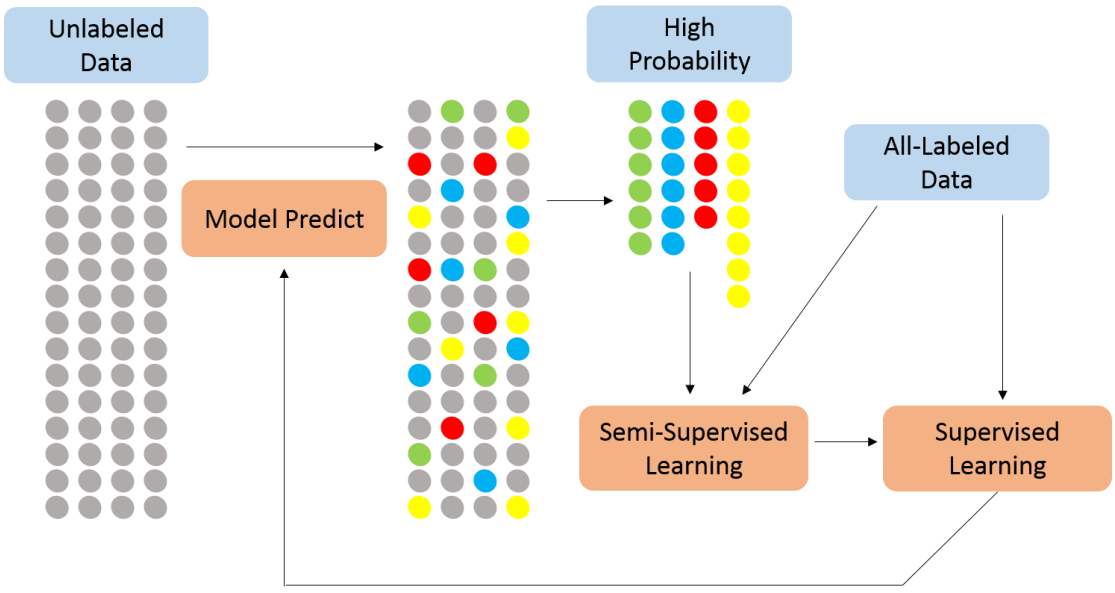
其中除了 Dense(10)的 Activation 是 sigmoid 之外，其他的 Activation 皆使用 relu。切五百筆 Data 出來做 validation，其 val_acc 的數據如下：

	1	2	3	Mean
Val_acc	0.652	0.676	0.642	0.657

除了基本架構，為有效使得 val_acc 上升，有設 checkpoint，還有將 input 做 normalize (/255)，以及利用 datagen 去增加 5 倍 data 來 train。

Semi-supervised learning (1)

用 supervised learning train 好的 Model 去 predict 45000 筆 data，其中只抓出 probability>0.999 的 data，把這些 label 當作是 unlabeled data 的 label，再將這些 data 與 labeled 的 data 一起去 train model，以此希望能夠增進 model 的正確率，並且在做完 self-learning 後，我會再對原本的 labeled data 進行 training，流程如下：

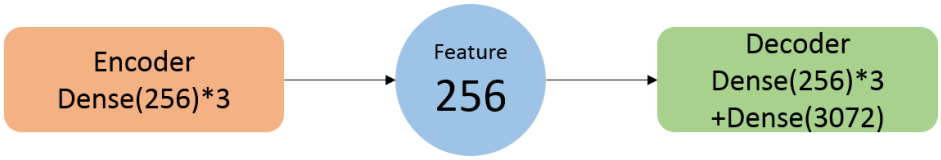


在實作上我只有從 45000 筆中隨機選取 12000 筆 data，因為用我原本的 model 去 predict 12000 筆約有 1000 多筆高度可信的 label，約是我 training data 的 1/4，我有測試一些比例，在這個比例下的 training 的 val_acc 會最好。

ula/la	1/4	1/2	1	Mean
Val_acc	0.698	0.680	0.686	0.688

Semi-supervised learning (2)

我實作了 autoencoder 做 clustering，首先 train 一個 encoder 跟 decoder 的 model，input output 都是同一張圖片，其架構如下圖：



使用 labeled data 去 train 這個 model，取他們的 feature 平均值，作為那個 class 的標準 feature。再將 unlabeled data 丟進此 model 去取得他們所有的 feature，將其中前 5000 筆(val_acc 最高)距離最小的 data label，當作正確可信的 data，將這些 data 與 labeled data 加在一起，最後在丟到我 supervised 的 CNN 架構中，進行 learning，其結果數據如下表：

ula/la	1/4	1/2	1	Mean
Val_acc	0.620	0.622	0.628	0.623

Compare and analyze your results

在我的 supervised learning 的 model 中，可以看到最高的 val_acc 達到 0.676，而加入 self-learning 的方法後，其 val_acc 可以提升到 0.698(Mean:0.657 to 0.688)，所以透過增加高度可信的 data，self-leaning 可以確實的增加有效的 labeled data，進而提升 model 的正確率。

在 method 2，autoencoder 中，與 supervised 相比掉到 0.628(Mean:0.657 to 0.623)，很明顯我的 autoencoder 在做 clustering 時的正確率並不高，所以使得在重新 train supervised model 時無法用更高的 val_acc，但我重複修改了 model，調整參數，也無法得到更好的結果，可能 encoder 跟 decoder 的複雜度必須提高進而讓 feature 可以更明顯；或者，我在 train supervised model 時，也可以加入 self-leaning 的方法，將高度可信的 data 丟到 autoencoder 裡面，去得到更好的 model，使得 feature 更明顯。