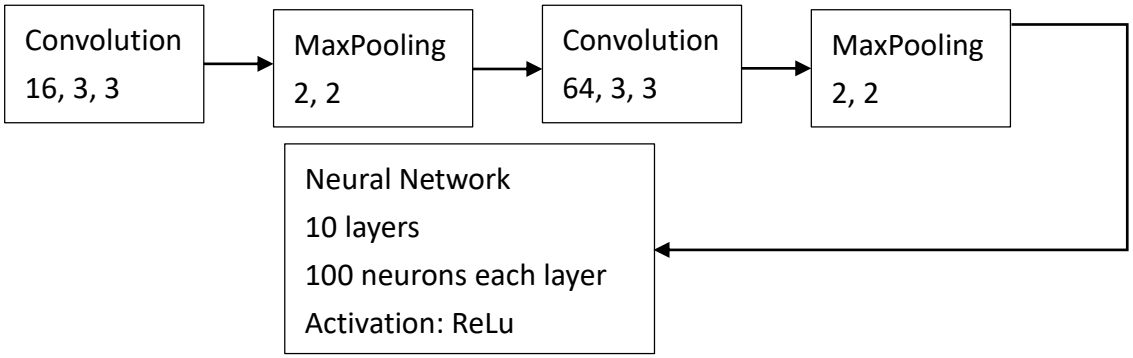
ML2016 HW3 report

姓名: 徐有慶　學號: R05922162

1. Supervised learning:

Step1. 讀取 label, unlabeled, test data

Step2. 建構 CNN



其中，Neural network 的 input layer 及 output layer 的前一層 hidden layer 皆加上 dropout = 0.25，loss function 使用 categorical cross entropy, optimizer 使用 adam，batch size = 128, number of epochs = 50。在此 model 下，利用 label data 進行 training。

2. Semi-supervised learning(1):

整體 CNN 架構同 supervised learning，先利用 label data 訓練出 cnn model，再利用其進行 self-training。

Step1. 經由 label data 先訓練出一個 cnn model(C1)

Step2. 建立一 label_flag array，紀錄 unlabeled data 是否已經被 label，若 label_flag[i] = 0 則表示第 i 筆 unlabeled data 尚未被 label, 反之則已被 label。

Step3. 利用 C1 預測所有 unlabeled data，檢查所有 label_flag 為 0 的 data，若預測其屬於 class n 且機率大於 0.85(confidence value)，則該 unlabeled data 的 label 設為 class n，並將其加入到 label data 當中，且 label_flag 設為 1，

Step4. 利用新的 label data 對 C1 再進行 training，number of epoch = 15， batch size = 128

Step5. 重複 Step3~4，直到所有 unlabeled data 都被 label 或是重複 10 次

3. Semi-supervised learning(2):

利用 auto encoder 做 clustering，整體 CNN 架構同 supervised learning

Step1. 建立 Neural network 拿來 train auto encoder 其層數為 Input layer -> 512 neurons -> 256 neurons -> 128 neurons (bottleneck layer) -> 256 neurons -> 512 neurons -> Output layer。activation = ReLu, optimizer = adam, loss function = mse

Step2. 將每筆 label, unlabeled data 經過 encoder 壓縮成 128 維的 features，5000 筆 label data 分成 10 個 clustering，並分別算出每個 clustering 的 centroid。

Step3. 計算 unlabeled data 與 10 個 centroid 的 euclidean distace，並將其加入距離最近的 clustering，同時更新 10 個 clustering 的 centroid

Step4. 重複 Step3，直到所有 unlabeled data 都有 label

Step5. 利用原先 label data 及被 label 過的 unlabeled data 去訓練 CNN

4. Compare and analyze your results:

Kaggle score

Supervise learning: 0.4762
Semi-supervise learning (self-training): 0.5216
Semi-supervise learning (auto encoder): 0.2636

Self-training 的部分設置了一個 confidence value，超過才當作是可信任的預測結果，而使用了 self-training 的 semi-supervise learning 也明顯比 supervise learning 的結果還要佳。Auto encoder 的部分，由於沒有加上 noisy，只是單純地做 deep auto encoder，且分群使用 euclidean distance 去做分群，可能較容易有分錯群的問題，而導致最後訓練出的 CNN accurate 很高，但 predict 的結果卻不是很好。