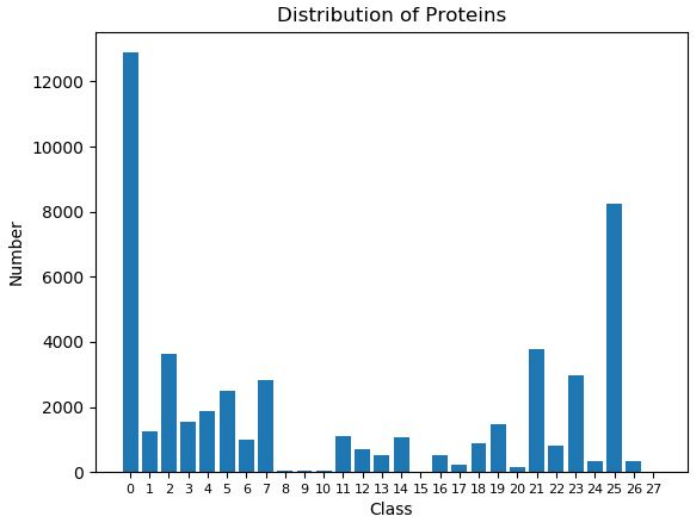
ML Final Proposal

Human Protein Atlas Image Classification

Caster 5 b04504042 劉家豪 b04203058蘇軒 b04901143陳柏瑞

Problem Study:

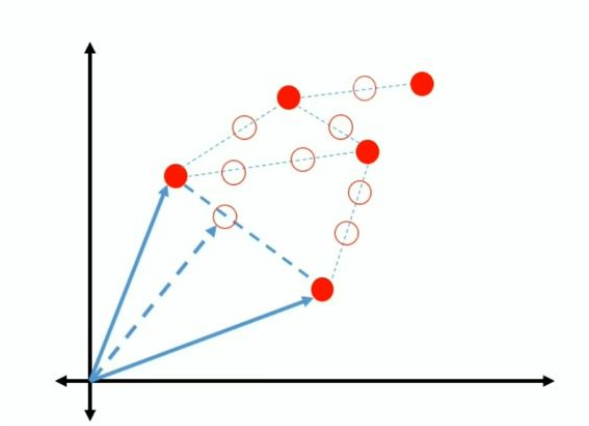
- Imbalanced Data
  Imbalanced Data 通常是指 training data class之間的數量分布相對不均勻，藉此影響 training 的準確率。
  以二元分類為例，如果兩個 class 中 class 0占了百分之95.5，而 class 1 占了百分之0.05這樣就算model直接把所有 testing data 都預測成 class 0 這樣 accuracy 也高達0.955 如此將會造成嚴重誤差。
  再以這次 final project 為例，下圖為28種 Protein 的 Distribution， 可以很明顯的發現 class 0 和 class 26 數量特別多，而 class 8 9 10 的數量特別少，如果在training的過程用 accuracy 去算的話將會忽略8 9 10的feature最後造成預測失準。



所以針對 Imbalance Data 我們上網了參考一些 Paper 得到一些改善如此狀況的方法

- ❖ SMOTE: Synthetic Minority Over-Sampling Technique
  由於 DATA Distribution 嚴重不均，所以通常有兩種方法，一種是 Under-Sampling，即是將數量較多的那個 class 刪減到一定的數量來均衡數量之間的比例，可是如此一來將有可能將重要的Data刪除; 另一種方法是 Over-Sampling， 傳統的 Over-Sampling 是將數量嚴重不均的

class的 data 用複製或 replicating 的方式來增加數據量，雖然這樣不會造成資料的流失，可是極有可能因此造成 overfit 在同樣的資訊之中。而SMOTE即是一個改善此問題的好方法，如下圖，將數據量較少的 class 中的任兩點進行 linear comnination 使兩點的權重和為一，用如此的方式來製造數據點將會合理的減少 overfitting 的問題。



❖ Macro-F1 Score

由於上述原因所以我們在 training 的過程中將不能使用 accuracy 來當作這次 evaluation 的依據，因為如此會導致 model overfitting 在數量較大的 class 上，所以我們要在這邊用 F1 score 即是判斷這個class 有或是沒有的比例來當作我們這次的依據，下圖中 TP FP TN FN 分別代表著 True/False Postive/Negative 即為是否正確預測有無的一個數字

$$Precision = \frac{TP}{TP + FP}$$ 對負樣本的區分能力

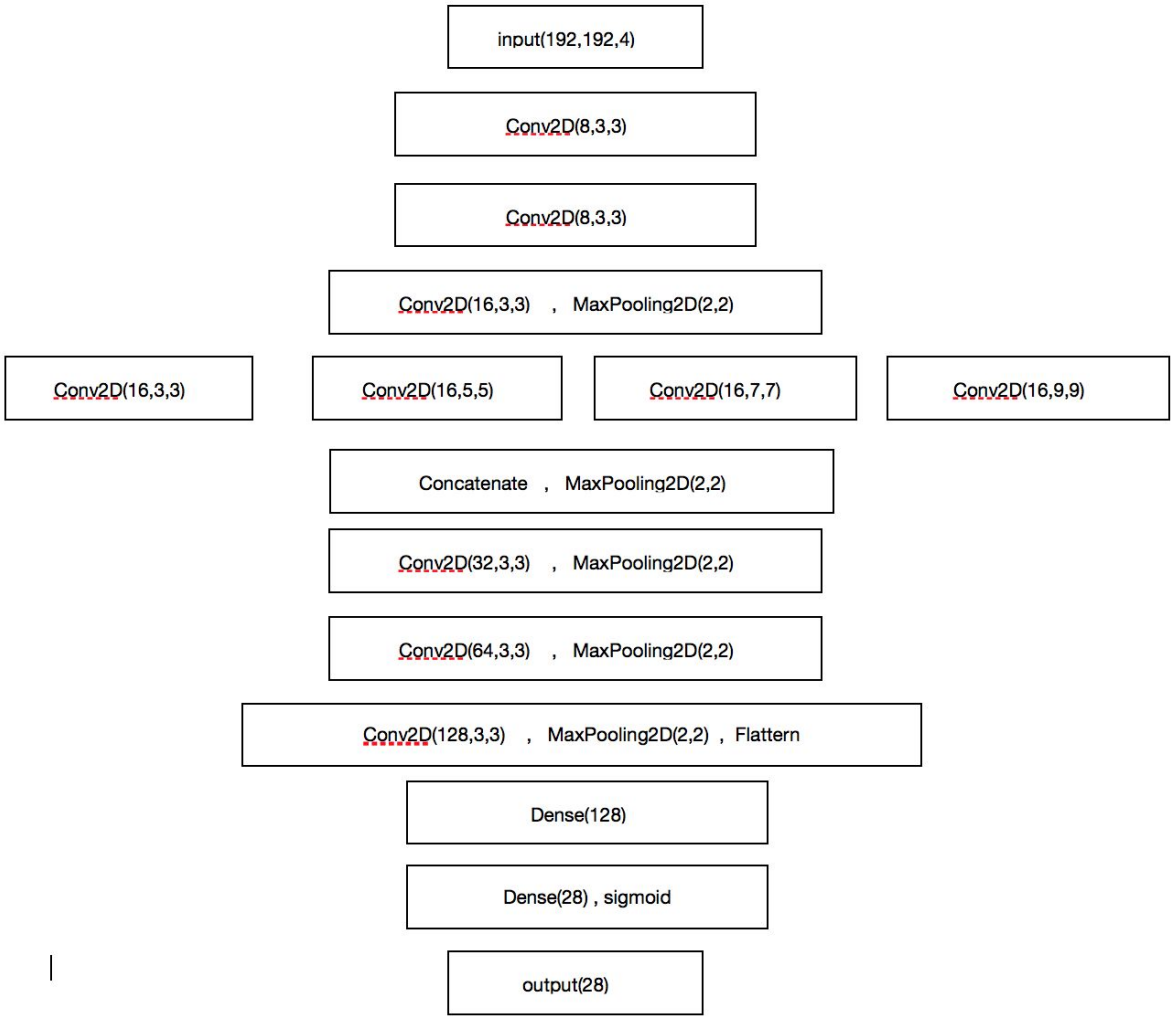
$$Recall = \frac{TP}{TP + FN}$$ 對正樣本的識別率

$$F1 = \left(\frac{Precision^{-1} + Recall^{-1}}{2}\right)^{-1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FN + FP}$$

而這次數據由於是 multiple-label 所以使用 Macro-F1 Score 來當作依據如下圖所示也就是將每一個 class 的 F1 Score 進行平均:

$$Macro - F1 = \frac{1}{c}\sum_{i=1}^{c}\frac{2TP_i}{2TP_i + FN_i + FP_i}$$

Proposed Method:



我們的model基本上是以上的架構，training過程中為了避免超過RAM的問題，每個batch從路徑去取一定量data。因為data屬於imbalanced data，我們之後打算應用data augmentation 產生圖片的翻轉和旋轉來增加data量，並且研究如何應用SMOTE在圖片上。

Reference:

1. SMOTE: Synthetic Minority Over-sampling Technique : https://arxiv.org/pdf/1106.1813.pdf?fbclid=IwAR2LsjNE0gJzStoTZw0MRnjCqL1TQvfHSAtvksI62UjtES-cnyQw2NrVX_8