介紹

隊名與隊員

NTU_r07921078_4000 盃
張廷維 r07921078
李沂倫 r05621110

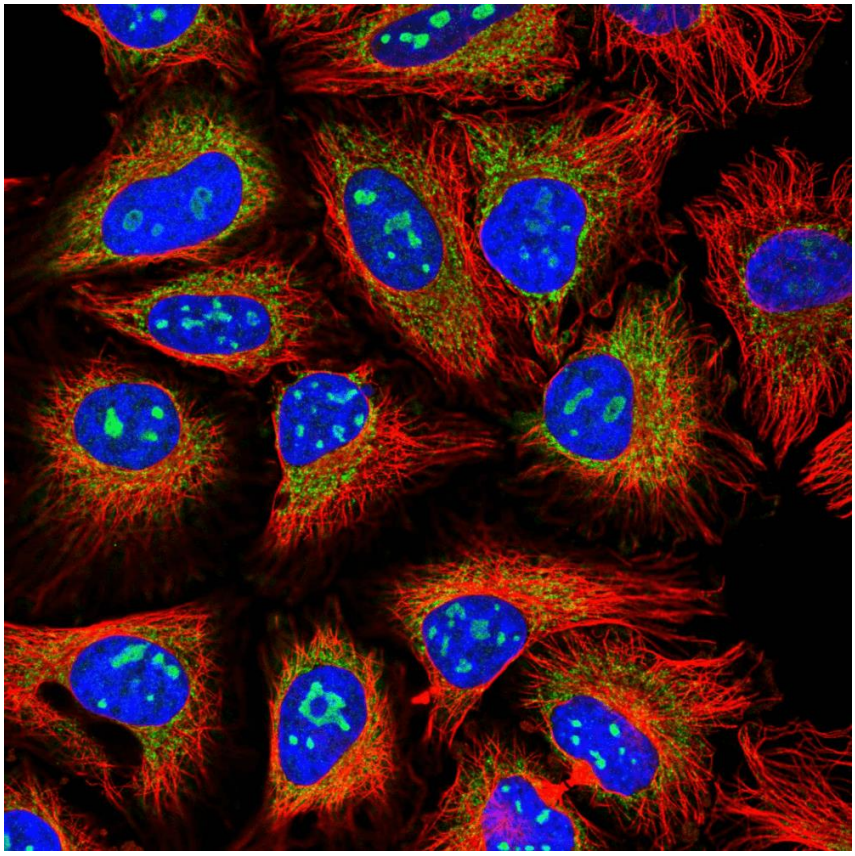
競賽標題:

Human Protein Atlas Image Classification
Classify subcellular protein patterns in human cells

簡介:

本次競賽主要是針對細胞電顯圖片做分類，有 28 個類別，分別代表蛋白質(螢光綠) 在細胞的位置，資料主要由 Human Protein Atlas 提供，希望可以利用程序化的方式快速隊蛋白質在細胞的分布進行鑑定。

總共有 31072 個 Training dataset 與 11702 個 Testing dataset，每個樣本由四個代表不同顏色(RGBY)的 PNG 組成，大小為 512*512 個像素點，與先前不同的是，每個樣本可能屬於多個類別。

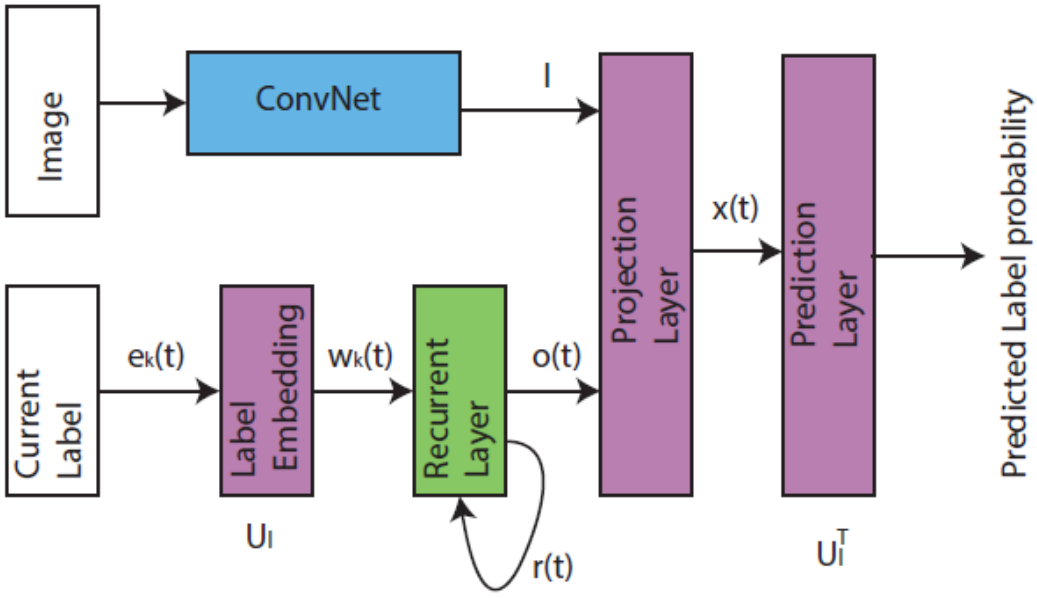


Problem study

由於本次選定的 Final 題目與 multi-label 相關，所以 paper study 選了 2016 CVPR 的一篇 paper ”CNN-RNN: A Unified Framework for Multi-label Image Classification”。

一張真實的 Image 上往往會有很多 labels，傳統上會將 multi-label classification 分成多個 binary classifier 去 training，然而這樣分類沒有考慮到同時存在的 labels 它們彼此之間的 dependency。因此這篇 paper 提供一個 CNN+RNN 的 framework。

Model 架構:



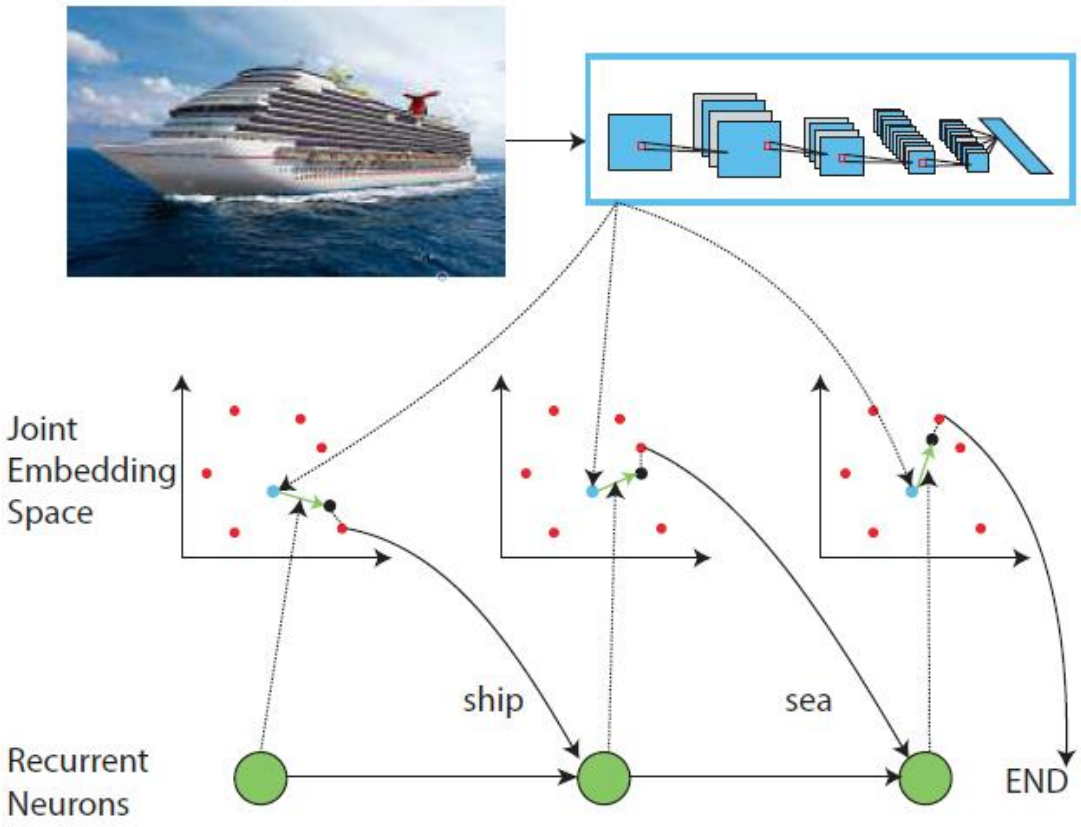
Paper 想法:

The CNN part extracts semantic representations from images.

The RNN part models image/label relationship and label dependency.

當 CNN 及 RNN 都抽完 feature 後，將它們 merge 到同一層(可視為一個抽象的空間)，最後在 predict 出一個 label，然而目前 predict 出來的 label 會當成 predict 下一個 label 的 input，也就是說每次預測 label 時，都會考慮到之前的 label，將 labels 之間的相依性考慮進去。

Inference 示意圖:



CNN 及 RNN 的 output vectors 分別經過 transform matrix 之後會 mapping 到同一個 Joint Embedding Space，然後兩個 vectors 再相加，最後將一個最近的 vector 再經過 transform matrix 後 predict 第一個 label(ship); 預測下一個 label 時，model 會將上一次的預測結果納入考量，通過 Recurrent layer 之後 mapping 到同一個 Space 時會發現 ship 跟 sea 有較高的關聯，因此預測 label 為 sea，以此類推。

Preprocess:

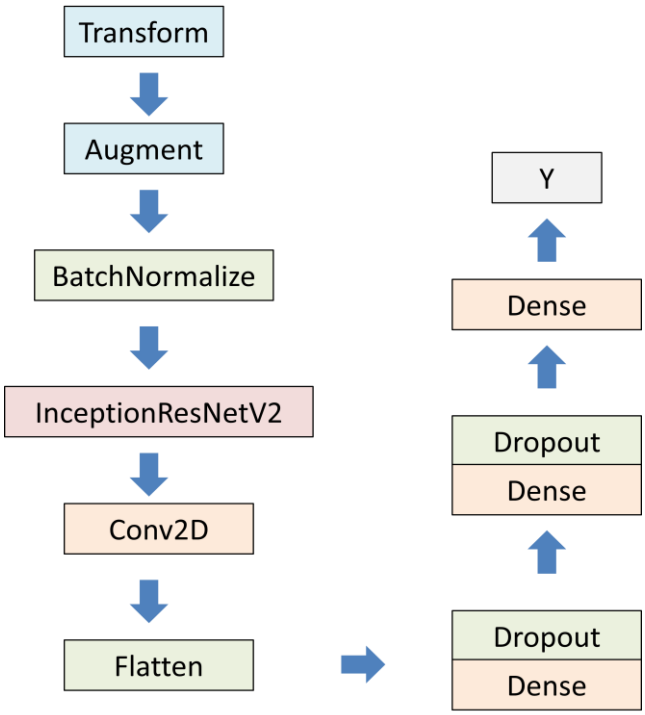
1. A pretrained CNN model
2. Label embedding matrix
3. One-hot encoding label

Proposed method

1. **利用已知 model 進行探勘**

   首先先用已知有效的模型進行訓練，觀察其訓練結果，並嘗試改進。模型如下圖。參考自:

   https://www.kaggle.com/byrachonok/pretrained-inceptionresnetv2-base-classifier



   最後得出來的 testing f1 為 0.29

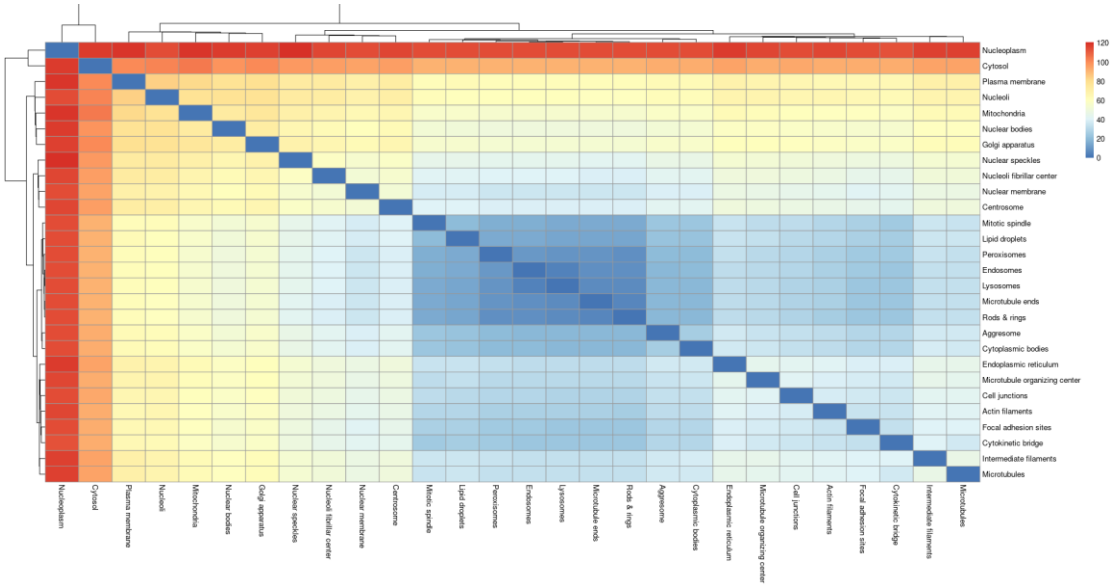
2. **觀察 class 分布並進行分群**

   對 training data 進行 predict 以預估各個樣本的 score 以及準確率。

   並用 label 的 Y 做 distance matrix (下圖) 以及計算準確度，從圖中可以看出有幾類 (eg. Nucleoplasm) 跟其他很多類十分相關，而這幾類的準確度也有一定的水準，並作為 A 群。

   針對樣本數小的，準確度低的類群，分成 B 群。

   剩下的，分做 C 群。



3. 分群訓練

針對 A 群，利用原本的 model 就可以了

針對 B 群，目前有想到的是可以用 resampling 的方式把各個 class 的樣本弄成差不多的數量，或者利用 penalty，修改 loss function 針對不同的 class 給予不同懲罰的權重，可能根據每個 class 的樣本數量。

針對 C 群，對 distance matrix 進行 cluster 或 kmeans 分群，希望可以分成 3 群，每一群個別進行訓練，一樣是利用原本的 model 進行修改以及訓練。

4. 其他想法

A. 如果有餘力的話，希望可以把細胞一顆一顆分開再訓練試試看，不但可以增加樣本數量，pattern 也會明確些，雖然 ResNet 應該有類似的功能。有餘力的話可能會利用 attention, rcnn 等方法嘗試。

另外也有一篇文章對於與我們相似的情境，利用 seeded watershed 的演算法對細胞進行 segmentation，並用 random forest 以及 SVM 進行分群，目前也打算利用 OpenCV 嘗試。 目前想先用 kmeans + dbscan 去切看看，原理上就是先利用 kmeans 做顏色上的分群，dbscan 把距離相近且分成同群的點連結成一個大群，應該可以把核拿出來，細胞整顆應該比較難。下周會去詢問會做電顯蛋白質分類的朋友，請教以人的角度是怎麼進行分類的，希望對我們的模型有啟發。

B. 希望可以將 paper 中 RNN 的想法經過調整後加入到目前的 model 中來提高 performance

Reference

1. J.Y. Newberg, J. Li, A. Rao, F. Ponten, M. Uhlen, E. Lundberg, et al. Automated analysis of human protein atlas immunofluorescence mages Proc IEEE Int Symp Biomed Imaging, 5193229 (2009), pp. 1023-1026
2. Ester, M., Kriegel, H.P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA: AAAI Press, pp. 226–231.
3. J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, “CNN-RNN: A Unified Framework for multi-label Image Classification” CVPR 2016