

Machine Learning Final Report

[Team]

NTU_r06922134_台北大冒險

[Member]

組員	葉沛陽	廖彥綸	湯梵平
學號	R06922134	B05902001	R06922120

[Project Title]

Human Protein Atlas Image Classification

[Introduction & Motivation]

人體中的蛋白質是大型生物分子，幾乎所有的體內活動都需要它的參與。將蛋白質視覺化在生物研究中最常見也是最直覺的作法，因此許多的研究都是透過圖片將不同的蛋白質進行分類，以有助於快速的了解人體構造。

通常蛋白質可以依據構造跟序列進行分類，這些分類通常都會有已知的一些特色跟功能，所以如果出現新的蛋白質就可以利用分類來推測大致的功能跟特性。在以前，主流的方法大部分都會針對每一種類的蛋白質建立一個模型，一個樣本經過每個模型的預測，假如超過設定的閾值，就會被歸類為這一種類的蛋白質。而我們這次的研究是要對每一張圖片預測出裡面有哪些種類的蛋白質在內。

[Data Preprocessing/Feature Engineering]

進行資料處理時由於 training data 大小超過30GB，讀取資料前先設定 GPU RAM自動成長以保證全部的資料皆能讀進RAM中。原始資料將一張圖(大小 512 * 512)分為 RGBY 四個類別，直接將其讀入為 512 * 512 * 4 的陣列。由於資料十分龐大，為了減少運算所需時間，並未進行ImageGeneration 的擴充。testing data的資

料也經過相同處理方式輸出成 $512 * 512 * 4$ 的陣列。處理資料的流程相當費時，所以將經過處理的資料以np file的方式儲存以利之後讀取的方便性。

之後將經過處理的training data前 90% 的資料送入訓練，後10%的資料分割為valid data。

[Model Description]

Model1(public score: 0.469):

Input layer 之後接BatchNormalization()再接Dense(4) Dense(3)然後使用keras application 的pretrained model “DenseNet169”，最後再一層Dense(28)的output layer(使用softmax)。

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 512, 512, 4)	0
batch_normalization_1 (Batch Normalization)	(None, 512, 512, 4)	16
dense_1 (Dense)	(None, 512, 512, 4)	20
dense_2 (Dense)	(None, 512, 512, 3)	15
densenet169 (Model)	(None, 1664)	12642880
dense_3 (Dense)	(None, 28)	46620

Model2(public score:0.458):

Input layer 之後接BatchNormalization()再接Dense(4) Dense(3)然後使用keras application 的pretrained model “DenseNet201”，最後再一層Dense(28)的output layer(使用softmax)。

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 512, 512, 4)	0
batch_normalization_1 (Batch Normalization)	(None, 512, 512, 4)	16
dense_1 (Dense)	(None, 512, 512, 4)	20
dense_2 (Dense)	(None, 512, 512, 3)	15
densenet201 (Model)	(None, 1920)	18321984
dense_3 (Dense)	(None, 28)	53788

Model3(public score:0.451):

Input layer 之後接BatchNormalization()再接 Dense(4)Dense(3)然後使用keras application 的pretrained model “InceptionResNet_v2” , 最後再一層Dense(28)的 output layer(使用softmax)。

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 512, 512, 4)	0
batch_normalization_1 (Batch Normalization)	(None, 512, 512, 4)	16
dense_1 (Dense)	(None, 512, 512, 4)	20
dense_2 (Dense)	(None, 512, 512, 3)	15
inception_resnet_v2 (Model)	(None, 1536)	54336736
dense_3 (Dense)	(None, 28)	43036

[Experiment and Discussion]

Focal Loss^[1] vs Categorical Cross Entropy(CCE)

focal loss跟CCE只差在有alpha跟gamma的參數來調整，非log項的影響力，他會較專注在數量少的class上面，適合用在Imbalanced data，實際測試也發現focal loss的確得到了比CCE更高的分數，所以往後都使用focal loss。

下列式子為focal loss的function:

$$L_{fl} = \begin{cases} -\alpha(1 - \hat{y})^\gamma * \log \hat{y}, & y = 1 \\ -(1 - \alpha)\hat{y}^\gamma \log(1 - \hat{y}), & y = 0 \end{cases}$$

	Focal Loss	CCE
public score	0.206	0.199

Threshold(Static vs Dynamic)

Static thresholds 將所有class的threshold都設為0.5。

Dynamic thresholds採用將0~1切1000份，再使用validation set去看個別class哪個當threshold，macro f1score最高，就採用它，結果是Dynamic比較好。

	Static	Dynamic
public score	0.202	0.376

Softmax vs Sigmoid

因為使用softmax發現比sigmoid好，因此後面都使用softmax。

	Softmax	Sigmoid
public score	0.376	0.332

Pretrained models[4]

使用keras applications的pretrained models(Resnet50、InceptionV3、InceptionResnetV2、DenseNet201、DenseNet169、DenseNet121)，最後發現Densenet169是最好的。

	Resnet50	Inception V3	Inception ResnetV2	DenseNet201	DenseNet169	DenseNet121
public score	0.430	0.419	0.451	0.458	0.469	0.411

Class Weight [5]

因為data 很 imbalanced，所以要針對classes數量多寡，分別設計他們的weight，使用的是sklearn中的compute_class_weight() function，從結果可以看出有使用class weight會比較好。

	without Class Weight	Class Weight
public score	0.442	0.469

Test Time Augmentation(TTA)[6][7]

TTA為Test 階段使用augment 將test data 圖片旋轉、移動、...。使用了TTA，發現分數大幅躍升。

	without TTA	TTA
public score	0.469	0.491

Unable Pretrain model at the beginning

一開始將pretrain model 的 trainable 設為 False, 訓練少次數的 epoch (在此定為 4 次)再將其和其他的layer一起進行訓練。因為 pretrain 的 model 已經有一定程度的訓練，所以先訓練其他layer使其部會過份著重於pretrain model的比重。

Callbacks

過度訓練模型可能導致overfitting，valid 的 f1分數降低，利用callbacks可以適時保留上次較佳的model。

ReduceLROnPlateau

validation loss 如果2 epoch沒有降低，learning rate就減半，讓他走小步一點，說不定更容易到達谷底。

[Conclusion]

利用上述3個不同的pretrain model所建立的模型(並且採用softmax、dynamic thresholds、class weight、focal loss)進行 ensemble，包括Inception_resnet_v2、Densenet201和Densenet169，ensemble的模型對任一分類有超過2票即同意該分類結果。在testing data利用Test Time Augment(TTA)的方式對預測圖片進行旋轉能更有效的得到正確的預測結果。

[Reference]

- [1] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. arXiv preprint arXiv:1708.02002, 2017. 7, 8
<https://arxiv.org/abs/1708.02002>
- [2] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu. Multi-label image classification with regional latent semantic dependencies. arXiv preprint arXiv:1612.01082, 2016.<https://arxiv.org/abs/1612.01082>
- [3] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In Proc. CVPR, 2017 <https://arxiv.org/abs/1702.05891>
- [4] Keras applications

<https://keras.io/applications/>

[5] scikit-learn compute_class_weight

https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

[6] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks

<https://arxiv.org/abs/1807.07356>

[7] tta_wrapper

https://github.com/qubvel/tta_wrapper