Human Protein Atlas challenge : Finding individual human cell differences in microscope images

Rohan Bhukar[1,2]

[1] BMED-6517 : Machine Learning in Biosciences, Georgia Institute of Technology
[2] Bioinformatics, School of Biological Sciences, EBB, Georgia Institute of Technology, GA-30332

rbhukar3@gatech.edu
https://bioinformatics.gatech.edu/graduate-students

Abstract

Proteins have important roles in almost all cellular processes. Different proteins interact at a specific location to perform a task, the exact outcome of that task depends on which proteins are present. Differences in subcellular distributions of one protein can give rise to functional heterogeneity between cells. Not only finding differences, but figuring out how and why they occur, is important for understanding how cells function, how diseases develop. Present machine learning models for classifying protein localization patterns in microscope images summarize the entire population of cells. Single-cell studies require precise classification models. The specific aim of this project is to develop models capable of segmenting and classifying each individual cell with accurate labels. In this paper, I attempt to propose a deep learning pipeline for instance segmentation of microscope images of proteins using the Kaggle HPA challenge-2 dataset. With recent advancements in deep-learning algorithm development and single-cell revolution, models that can precisely classify patterns in each individual cell in image, will fundamentally improve our understanding of how cells function and how diseases develop. Experimental results show that the proposed solution is able to classify protein localization patterns with 0.355 mAP score achieved in the Kaggle challenge.

Keywords: Machine Learning, Image Classification, Protein localization, Deep Learning, Single-cell analysis, EfficientNet, Cell segmentation,

1 Introduction

More than 7.674 billion humans exist on this earth, and each of us is made up of trillions of cells. Like every individual is unique including identical twins, scientists observe differences between the genetically identical cells in our bodies. Differences in the location of proteins can give rise to cellular heterogeneity. Proteins are critical to almost all cellular processes, with many coming together at a specific location to perform a role, resulting in an outcome depending on protein composition at location. Different subcellular distributions of one protein can give rise to great functional heterogeneity between cells.

Images from cell microscopy provide a lot of information, and are a crucial source of high throughput biological data of cells. Intracellular proteins when tagged with fluorescent markers can provide information about multiple states of living cells, and can be annotated by the specific gene function while measuring the spatial variation in localization of those proteins. Imaging can be performed on single living cells and the process of acquiring data can be manual as well as automated, the later allowing thousands of micrographs per hour to be produced in arrays. With technological advancements and rapid screening of fluorescent tagged proteins, researchers have attempted to look for mutant effects on protein presence (Albert *et al.* 2014; Parts *et al.* 2014) and localization (Chong *et al.* 2015), gene function determination (Hériché 2014), and changes in cell morphology (Ohya *et al.* 2005).

In the past decade, the area of deep learning has proven to be a very promising method for the scientific community in multiple fields, including natural language processing and computer vision (Krizhevsky et al., 2012; Sutskever et al., 2014). Deep neural networks are popular for image processing tasks, providing an advancement to the feature selection problem. There exists multiple examples of methods based on deep learning to address problems of object detection (He *et al.* 2015), image captions (Vinyals *et al.* 2015), as well as applications in biological sciences (Rampasek and Goldenberg 2016) for fields ranging from genomics to electron microscopy (Cireşan *et al.* 2012). Given a large training dataset, these methods can automatically learn features that are most critical to the classification problems of interest.

This is a weakly supervised multi-label classification problem and a code competition. Solving single-cell image classification challenge will help to characterize single-cell heterogeneity in the dataset provided, by generating more accurate annotations of the subcellular localizations for thousands of human proteins in individual cells. Previous solutions exist (Ouyang, Wei et al 2019), which attempt to assign multiple labels to individual single cell images in the HPA 2019 challenge. Here, the proposed solution applies the deep learning paradigm to high-throughput single-cell images for instance segmentation. The network can be used to extract useful features and provide precise information to classify patterns in individual images, aiding in further improvement of our understanding of how cells function.

2 Methods

Data

The Human Protein Atlas is an initiative based in Sweden aimed at mapping proteins in all human cells, tissues, and organs. The

data in the Human Protein Atlas database (http://www.proteinatlas.org) is freely accessible. The data from the Kaggle challenge provides the following formats :

1) The data page on Kaggle provides a set of full size original images (a mix of 1728x1728, 2048x2048 and 3072x3072 PNG files) in train.zip and test.zip.
2) The image level labels from train.csv and the filenames for the test set from sample_submission.csv.
3) Public HPA images available to download in the notebook.
4) The 16-bit version of the training images are available here.
5) Each sample consists of four files, with every file representing a different filter on the subcellular protein patterns represented by the sample. The format is [filename]_[filter color].png for the PNG files. Colors are red for microtubule channels, blue for nuclei channels, yellow for Endoplasmic Reticulum (ER) channels, and green for protein of interest, as mentioned.

The unlabelled dataset consists of training and testing images in .png format along with tensorflow records for both train and test datasets, a label file as .csv for the training set with multilabel label codings, and sample submission file to prepare the participants results in acceptable formats for scoring. A total of 21,806 training IDs across 87,224 train samples were provided. Each image has 4 channels. To use high confidence training examples tf records were dropped from the datasets. The final dataset comprised of 21,806 train IDs with 4 channels per microscopy image from the 19 classes (nucleoplasm, nuclear membrane, nucleoli, nucleoli fibrillar center, nuclear speckles, nuclear bodies, endoplasmic reticulum, golgi apparatus, intermediate filaments, actin filaments, microtubules, mitotic spindle, centrosome, plasma membrane, mitochondria, aggresome, cytosol, vesicles and punctate cytosolic patterns, and negatives).

Further, python modules were written to resize whole images to 512 x 512 pixel images due to a mix of original image sizes available. Stratified shuffle split was used to split the train set with 80:20 split ratio into training and validation sets resulting in 17,445 and 4,361 image IDs in train and validation sets respectively. Image rescaling was done to convert the data from 0-255 to 0-1 scale in the float format which makes the computational steps convenient and at the same time the model converges faster with rescaled images. Images were processed in batch sizes of 20 and one hot encoding of labels was performed. The distribution of labels across train images (Figure 1), provides insights into how frequent are the individual labels in images, where nucleoplasm and cytosol being the top 2 observed classes.

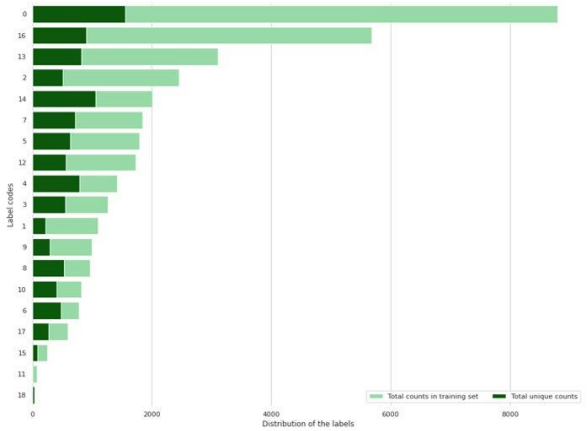

Figure 1. Distribution of labels across the training dataset.

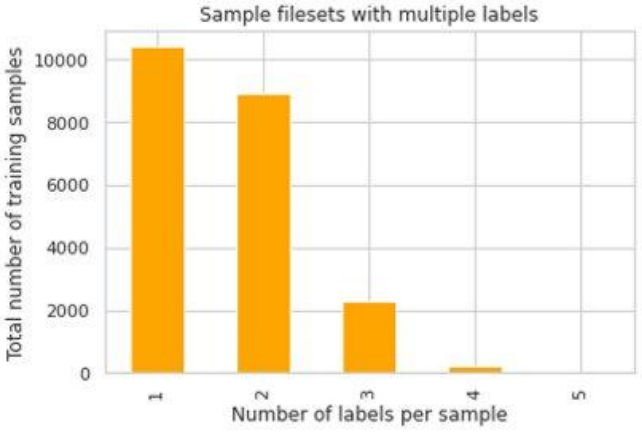

Figure 2. Multi-class combinations in microscopy images from competition data. A total of 4 classes was observed to be maximum label per image with more than 50% of the image IDs being single class labels.

Nucleoplasm (class 0) and cytosol (class 16), were found to co occur most frequently as compared to other class labels in the microscopy images (Figure 3).

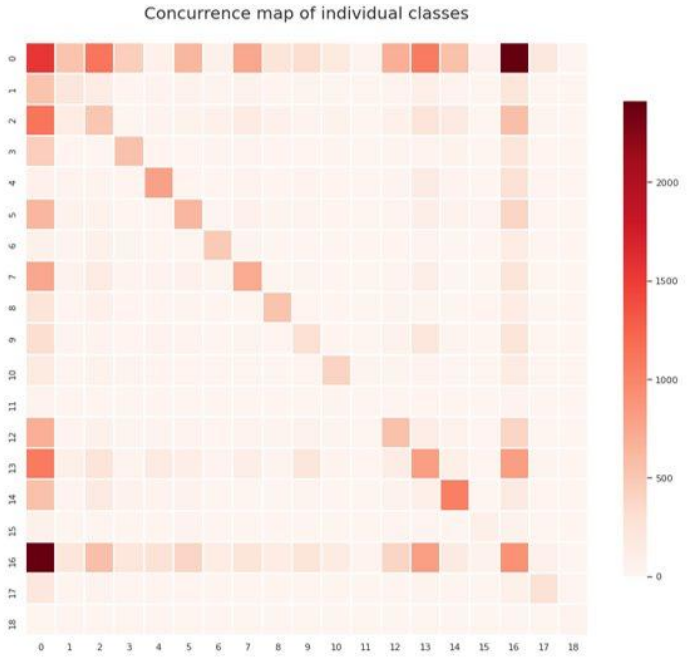

Figure 3. A map of concurrence of individual classes of features of biological importance.

Convolutional neural network

A convolutional neural network (CNN) is a class of deep neural networks which is commonly applied to analyzing visual image data. When analysing images, the mask (or filter) is a critical construct. The convolution involves an operation with an initial image and the mask. Here, the mask is flipped both vertically and horizontally, while placing it over every pixel in turn. Summation of the pixel wise product of the sub-image and the mask, is the output. Convolution between 2 f and g is provided by the following equation,

$$(f * g)(t) \triangleq \int_{-\infty}^{+\infty} f(\tau)\, g(t-\tau) d\tau \qquad (1)$$

While processing an image, a convolution between an image I and kernel K of (d x d) size and centered at pixel (x,y), is given as,

$$(\mathbf{I} * \mathbf{K})(x, y) = \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbf{I}(x + i - d/2, y + j - d/2) \times \mathbf{K}(i, j) \qquad (2)$$

CNNs are a family of neural network architectures with at least 1 convolutional layer.

Transfer learning was used in constructing the deep learning architecture proposed as part of the solution to the HPA problem statement. It is essentially the improvement of learning in a new designed task through the transfer of knowledge from a related task that is already being learnt. In simple terms, a model trained on task 1 is re-used with additional layers as per task requirement with updated input and output size of outer layers, on a task 2. When a state-of-the-art model is developed, it usually takes researchers a lot of time in training and optimizing the model. On tasks that are addressed through models built upon transfer learning, results in saving time and provides better performance.
EfficientNet is a family of CNNs built by the researchers at Google (Mingxing Tan and Quoc V Le. et al 2019). They not only provide higher accuracy but also improve the efficiency of models by reducing the number of parameters as compared to other state-of-the-art models (Figure 4).

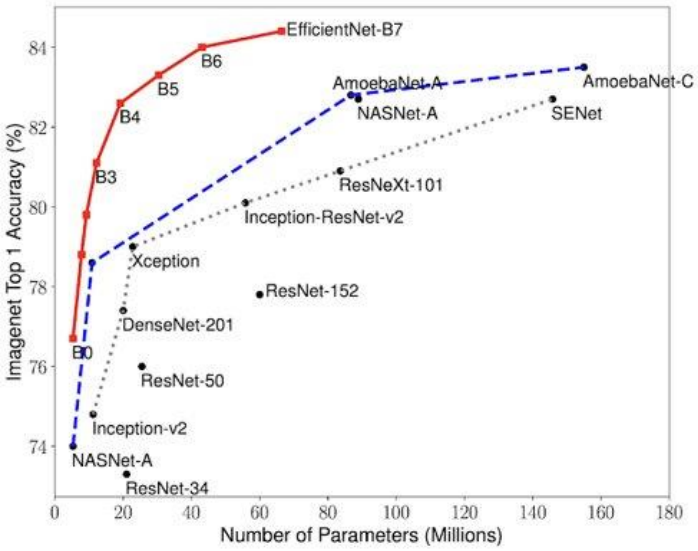

Figure 4. Total layers in EfficientNet-B0 is 237 while EfficientNet-B7 has 813 layers.

In overview, each of the state-of-the-art models follow a general architecture design, with basic stem and final layers as described in figure below (Figure 5).

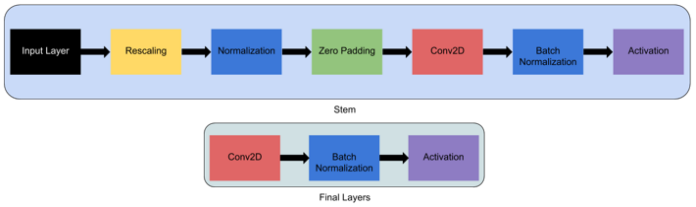

Figure 5. EfficientNet and ResNet family of CNNs in overview have Input layer, followed by rescaling, normalization, zero padding, Conv2D, batch normalization and activation layers as the stem layer. While the final layers consist of Conv2D, Batch normalization and activation layers.

As suggested by the competition organisers, all 4 channels (RGBY) were used for the model training. Since EfficientNet requires 3 channels, the final model was initialized with weights being reused for Yellow channel from the blue-channel weights. Input and output layers were removed from the pre-trained model and added these layers with 4-channel dimensions. The model summary from the 4-channel DNN architecture is available in Figure 6.

```
top_bn (BatchNormalization)     (None, 16, 16, 1280) 5120        top_conv[0][0]
top_activation (Activation)     (None, 16, 16, 1280) 0           top_bn[0][0]
avg_pool (GlobalAveragePooling2 (None, 1280)         0           top_activation[0][0]
dense_1 (Dense)                 (None, 18)           23058       avg_pool[0][0]
==================================================================================
Total params: 4,072,910
Trainable params: 4,030,894
Non-trainable params: 42,016
```

Figure 6. EfficientNet model initialised with 4 channels, with 4,030,894 total trainable parameters.

Cell segmentation

The HPA Kaggle challenge organisers suggested to use HPA cell segmentator and the module was released for the participants to incorporate in the image processing pipeline. As per the authors, the module is made for programmatically iterating through a list of images and returning the segmentations. There are two models as part of the module, nuclei model and cell model both containing weights for the respective models as part of the package. Incorporating the weights in the cell segmentation steps of the pipeline, provides masks to individual features with a scaling factor of 0.25 and multi channel model set to True. The resulting masks for unique features separately (Figure 7), facilitated further the cell-level predictions.

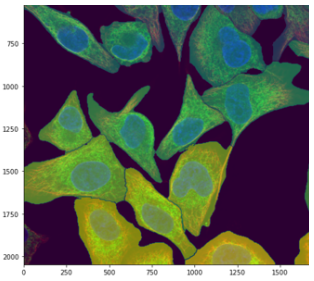

Figure 7. Masks for features in individual microscopy images.

Integrated Gradients

The addition of integrated gradients (IGs) as a concept to the image processing pipeline helps in addressing the explainability aspect of the deep learning solution proposed. IG is an explainability technique for DNNs which helps in visualizing its input feature importance that essentially contributes to the model's predictions. In simple terms it computes the gradient of the model's prediction output to its input features. It can be applied to any differentiable model like image, text, etc.and no modification is made to the original DNN. Here, for the instance segmentation problem, IGs were incorporated for debugging the performance of the deep learning model developed as described above and further to understand the feature importance of the individual classes while segmenting and extracting the rules from network.

3 Results

To perform the task of accurate classification of protein localization in single-cells and populations, a deep convolutional neural network incorporating transfer learning from EfficientNet, resulting in 4 channel classifier model that learned from 21,806 single cell images from the HPA Kaggle challenge. Each of the images record the microtubule signal in the red channel, nuclei in the blue channel and the protein of interest in the green channel. The network consists of 237 layers with over 4,000,000 parameters in total.

The model pre-training was done with only RGB channels and early stopping was introduced to monitor validation loss and prevent the overfitting of model to the training set. The PR AUC was used as a metric to assess the model performance, and results for training and validation PR AUC are presented in Figure 8. The training and validation loss for the model training steps can be visualised from Figure 9.

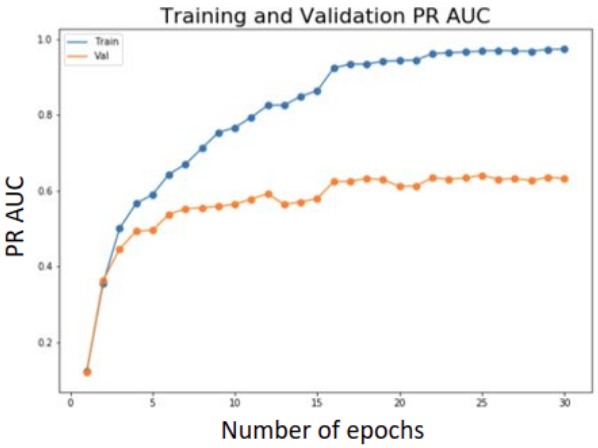

Figure 8. PR AUC curve for the proposed model training.

Figure 9. Binary crossentropy was used as the loss function.

Mean average precision (mAP), was used as the metric from the competition organisers to assess the final submissions from participants. mAP score is calculated by taking the mean AP over all classes and/or over all the IoU thresholds, and mostly used in object detection competitions. The mAP scores for the submissions to HPA Kaggle challenge over multiple iterations can be assessed from Table 1.

Model	mAP score
EfN B7 + Cell segmentation	0.355
EfN B5 + Cell segmentation	0.326
EfN B2 + Cell segmentation	0.317
EfN B0 + Cell segmentation	0.301

Table 1. mAP scores from the competition submissions

The final DNN architecture proposed here for instance segmentation and multi-label classification of single cell microscopy images, has the EfficientNet B7 model with 4 channel classifier created. It resulted in the highest score mAP of 0.355 achieved over the multiple iterations to the problem statement. Though multiple attempts were made to address the problem statement including building a merged 19-layer convolutional neural network for multi-class classification, the performance of the models was best achieved when EfficientNet B7 was used for transfer learning.

Correlation between multiple pair of channels was performed to assess the importance of any individual channel to identify their predictive power (Figure 10). The green channel was found to be highly correlated with other channels, indicating a significant predictive power of this channel alone. Though from prior knowledge, we understand that this channel provides color coding for the proteins of interest.

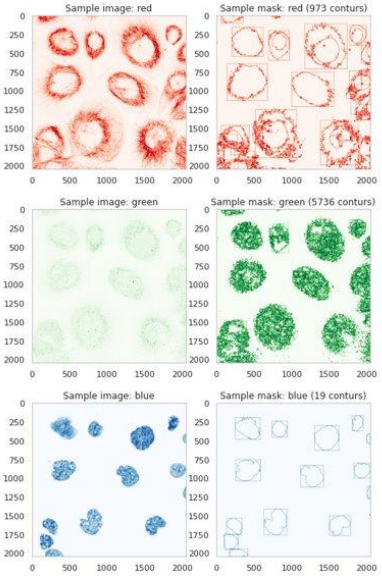

Figure 10. Cell level contours for the green channel.

4 Discussion

Here, with the proposed solution for the HPA challenge, I have demonstrated that the transfer learning model with 237 layers and reconfigured to accommodate the 4 channels (RGBY), can achieve mAP scores of 0.355 for individual cells over 18 subcellular localizations. Though the concept of integrated gradients is used as part of the image processing pipeline to understand the model explainability, further improvement needs to be made to the proposed image processing pipeline to enhance the interpretability of the results and model's prediction, while stepping away from the general black box model approach.

5 References

1. **https://www.kaggle.com/c/hpa-single-cell-image-classification**
2. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5427497/#bib1**
3. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5427497/#bib40**
4. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6976526/**
5. **http://proceedings.mlr.press/v97/tan19a.html**