Human Protein Atlas Image Classification

Mahit Tanikella

Monta Vista High School, Cupertino, CA

Abstract

Human body consists of trillions of cells. Each cell contains many proteins organized to maintain its structure and function. Proteins are “the doers” in the human cell, executing many functions that together enable life. Almost all of the medications are directed against proteins. As our knowledge about proteins is very little, today’s drugs are targeting only very small percentage of the 20,000 different proteins found in the human body. Images visualizing proteins in cells are commonly used for biomedical research, and these cells could hold the key for the next breakthrough in medicine. Automating biomedical image analysis is key to accelerating the understanding of human cells and disease.

Protein subcellular localization prediction involves the prediction of where a protein resides in a cell, its subcellular localization. This information is useful in research for cancer, neurological disorders and other diseases to develop better drugs to target proteins. In this research, a deep learning based convolutional neural network model is developed and trained to identify subcellular localization of proteins in cell microscope images. Given a microscope image, this model automatically identifies which of the 28 different organelles in the cells, the proteins are located in. The neural network is built on top of pretrained InceptionResnet50 network. The network is trained on Google Cloud machine with GPU for 150 epochs for 6 hours with cell microscope images from human protein atlas data. All image samples are represented by four filters, the protein of interest (green) plus three cellular landmarks: nucleus (blue), microtubules (red), endoplasmic reticulum (yellow).

With cross entropy loss for adjusting the neural network weights, the model achieved an accuracy of 96.5% and F1 score of 0.16. This means the model was overfitting as the data was imbalanced. By using focal loss for adjusting the neural network weights, the model has achieved 97.6% accuracy with significant increase in F1 score to 0.68. This makes it possible to use this model to build a tool integrated with smart-microscopy system to automate protein subcellular localization.

Objective

Automate analysis of cell microscope images for protein subcellular localization using deep convolutional neural networks.

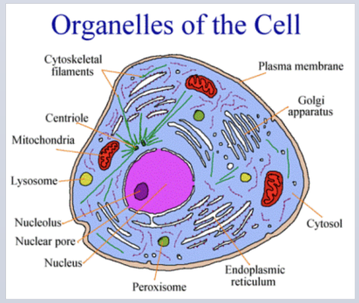
Problem

Protein subcellular localization prediction involves the prediction of where a protein resides in a cell, its subcellular localization.

This is used in research for cancer, neurological disorders and other diseases to develop better drugs to target proteins.

Microscope images of cells are analyzed manually for subcellular localization of proteins. This is slow and time consuming

.



Materials and Methods

Data

Human protein atlas image data from Kaggle website

31072 images, 512x512 pixels each. Total: 32 GB

Each sample has 4 images - Green - Protein of interest, Blue – Nucleus, Red – Microtubules, Yellow - Endoplasmic Reticulum

Protein Train	Count	Protein Train	Count
Nucleoplasm	12885	Cytosol	8228
Nuclear membrane	1254	ER network	6634
Nucleoli	3621	Golgi apparatus	2513
Nucleoli fibrillar center	1561	Intermediate filaments	1324
Nuclear speckles	1858	Actin filaments	828
Nuclear bodies	2513	Microtubules	1066
Endosomes	542	Mitotic spindle	1215
Lysosomes	504	Centrosome	312
Peroxisomes	904	Plasma membrane	3037
Mitochondria	4593	Mitochondria	4593
Aggresome	322	Aggresome	322
Cytokinetic bridge	899	Cytokinetic bridge	899
Flemming body	862	Flemming body	862

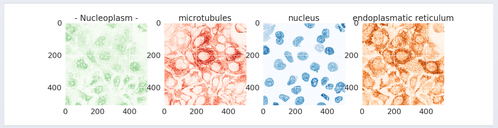
Table 1: Count of different proteins in microscope images

Count of proteins

Number of targets per sample	% of data
1	36.04
2	43.14
3	16.77
4	3.71
5	0.29

Table 2: Number of targets per percent of data

Number of targets per percent of data



Sample Image

Software

Keras (https://keras.io/),The Python Deep Learning library

Jupyter Notebook (http://jupyter.org/), Open Source Web Application for Interactive Computing

Scikit-learn (http://scikit-learn.org/), Machine Learning in Python

Numpy (http://www.numpy.org), Scientific Computing with Python

Matplotlib (www.matplotlib.org), 2D Plotting Library

Hardware

1 NVIDIA Tesla K80 GPU, 4 vCPUs, 26 GB memory, 150 GB hard disk

$300 dollars credit

1 NVIDIA Tesla K80 GPU, 4 vCPUs, 26 GB memory, 150 GB hard disk

Procedure

Built Deep Neural Network model using Pretrained Inception Resnet V2 model in Python using Keras in Jupyter notebook

Built custom data generator to combine images for each sample

Image preprocessing

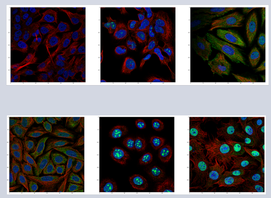
Normalization and Augmentation

Random rotation, flipping, brightness change

Data is split – 80% for training and 20% for validation

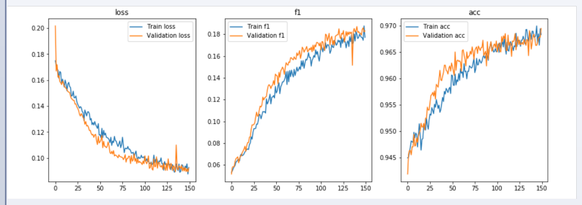
With batch size of 16, 6 hours to train the network for 150 epochs

Neural network architecture used

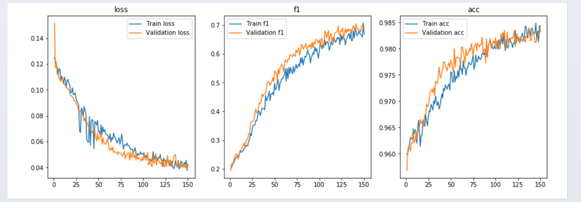


Images from generator used as input to neural network

Results



Loss, F1 score and accuracy with cross entropy loss



Loss, F1 score, accuracy with focal loss

Interpretation of results

This is a multilabel, multiclass classification problem with each sample having multiple cells and protein in multiple organelles of those cells. Data is highly imbalanced with more samples for some like nucleoplasm, cytosol etc. and very few for others like microtubule, rods and rings. Model trained with cross entropy loss had high accuracy of 96.5%, but low F1 score of 0.16. In this model, in the final layer, we have 28 different outputs for each of the organelles. Output is probability of having protein in that organelle. Probability greater than 0.5 is predicted as positive and less than 0.5 as negative.

Deep convolutional neural network model is trained by incrementally adjusting the model's parameters so that the predictions get closer and closer to actual values. Cross entropy is the most commonly used cost function in neural network model training.

$p_t = p$ $(y = 1)$, $1-p$ $(y = 0)$ (p: predicted probability, y: actual value)

$$\text{Cross entropy loss } CE(p_t) = -\log(p_t)$$

If there is a class imbalance problem i.e. we have lot more samples of one class over the other, the model trained on cross entropy loss gets better and better at predicting classes of higher frequency. In these cases, model achieves high accuracy by always predicting the same class. To better assess neural network model performance, other metrics like precision, recall and F1 score are used.

$$\text{Precision} = (\text{True Positive}) / (\text{True Positive} + \text{False Positive})$$

$$\text{Recall} = (\text{True Positive}) / (\text{True Positive} + \text{False Negative})$$

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Focal loss is calculated by applying a modulating factor γ to cross entropy loss. The idea is that, if a sample is already classified correctly by the neural network, its contribution to the loss decreases. It addresses the problem of class imbalance by making the loss implicitly focus on the problematic classes.

$$\text{Focal loss } FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

The model has achieved 97.6% accuracy with significant increase in F1 score to 0.68 by using focal loss with $\gamma = 0.5$ for training and weight adjustment.

Conclusions

This is multilabel, multiclass classification problem. Data is imbalanced with more samples for some like nucleoplasm, cytosol etc. and very few for others like microtubule, rods and rings. Initial model developed and trained with cross entropy loss had high accuracy of 96.5%, but, low F1 score of 0.16. This is because of imbalanced data. Adjusting weights of neural network using focal loss increased the accuracy to 97.6% and improved F1 score to 0.68. This makes it viable to build a tool using this model that can be integrated with smart-microscopy system to identify proteins from high throughput images.

Applications to Biotechnology

it is possible to build a tool using this model that can be integrated with smart-microscopy system to automate protein subcellular localization prediction from high throughput images.

Knowing where the protein is inside a cell helps us develop better pharmaceutical drugs to target proteins.

This will help accelerate research for cancer, neurological disorders and other diseases.

Acknowledgements

I would like to thank my Match teacher Mrs. Kathleen Koch for being my mentor and guiding me throughout.