THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Human Protein Atlas Classification

Group 2:

Voratham Thiabrat

Aashish Nair

Tran Hieu Le

Content

01 Introduction

02 Dataset

03 Methodology

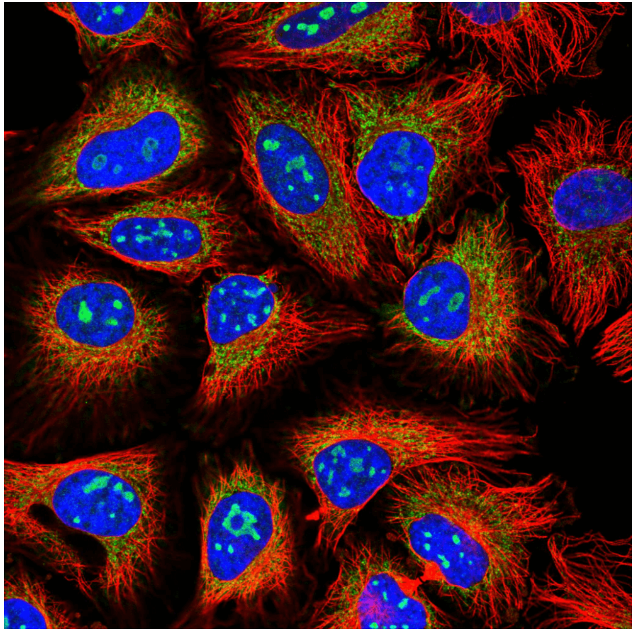
04 Model

05 Improvement

06 Conclusion

Introduction

- Deep Learning in medical research:
  - Faster diagnosis of disease
  - Development of new medicine
  - Improvement in treatment
- Objectives:
  - Classify mixed patterns of proteins
  - Handle 4 channel microscope images
  - Address strong imbalance level

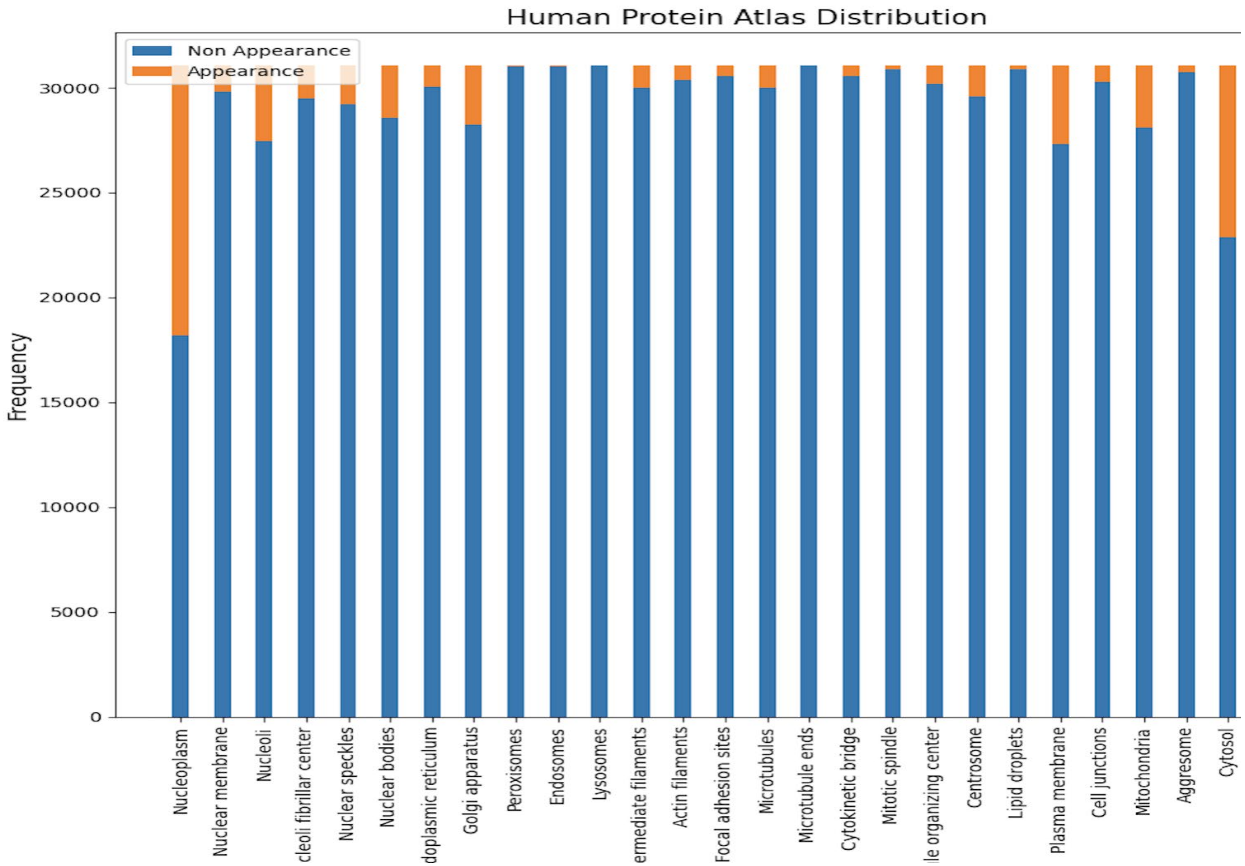


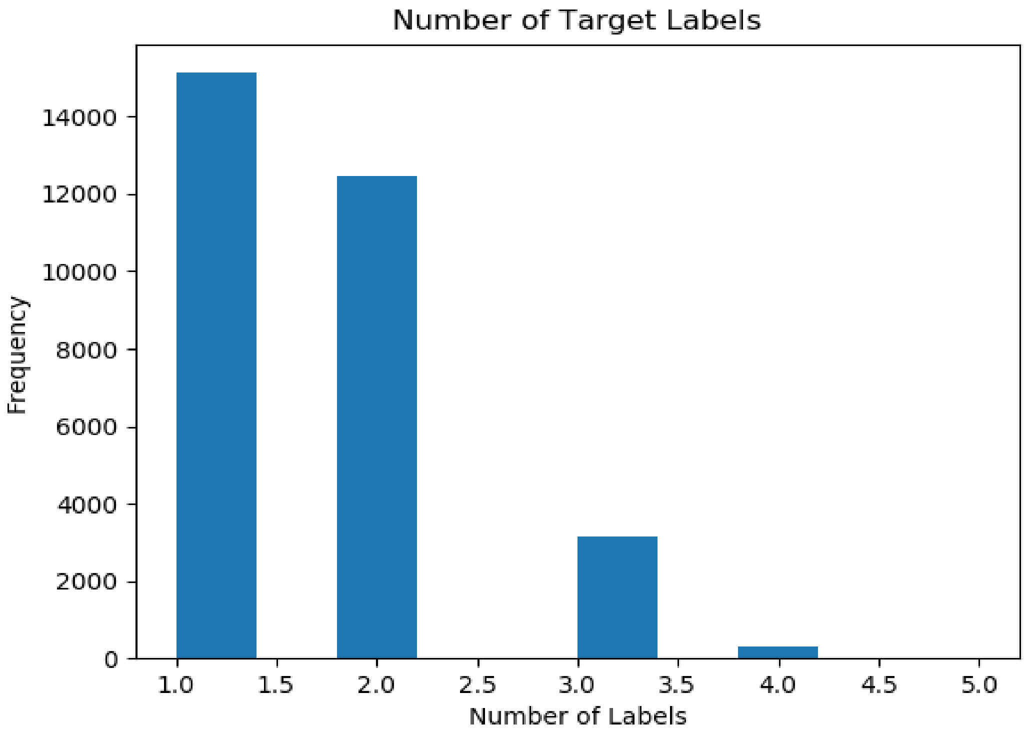
Dataset

- Data obtained from a Kaggle competition hosted by the Human Protein Atlas

  https://www.kaggle.com/c/human-protein-atlas-image-classification/data

- Provided data:
  - csv file containing image ids of 31072 images and 28 target labels.
  - 4 channels RGBY for each image.
  - Rare classes appears only 10-15 times in total.


Human Protein Atlas Distribution
Non Appearance
Appearance
Frequency
30000
25000
20000
15000
10000
5000
0
Nucleoplasm
Nuclear membrane
Nucleoli
cleoli fibrillar center
Nuclear speckles
Nuclear bodies
doplasmic reticulum
Golgi apparatus
Peroxisomes
Endosomes
Lysosomes
ermediate filaments
Actin filaments
Focal adhesion sites
Microtubules
Microtubule ends
Cytokinetic bridge
Mitotic spindle
le organizing center
Centrosome
Lipid droplets
Plasma membrane
Cell junctions
Mitochondria
Aggresome
Cytosol


Number of Target Labels
Frequency
14000
12000
10000
8000
6000
4000
2000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Number of Labels

Key Obstacles

Highly Imbalanced Data

Data is quite different from ImageNet

Methodology

- Performance Metrics and Loss function:
  - Macro F1 score as suggested by the competion.
  - Binary Cross-Entropy Loss.
- Transfer learning using VGG, ResNet and DenseNet.
- Keras, Tensorflow and Pytorch frameworks.
- Experiment with various parameters and architecures.
- Processing with data augmentation.
- Implementing different ideas and techniques to address imbalance.

Model

VGG16 (Keras framework)	
Parameter	**Assigned Values**
Batch Size	64
Learning Rate	1e-4 with 5e-5 decay
Activation function (hidden layers)	relu
Dropout	0.2
Optimizer	Adam

Evaluation Metric	**Value**
Validation Loss	0.1724
Macro f1-score (validation)	0.1421
Macro f1-score (testing)	0.1103

Model	**Activation Function and Optimizer**	**Macro F1-score (Validation set)**	**Macro F1-score (Test set)**
CNN (4 Conv2d)	Relu and Adam	0.1979	0.1582
Resnet50 (Base model)	Relu and Adam	0.2866	0.2783
Resnet50 with data augmentation	Relu and Adam	0.5305	0.4894
Resnet152 with data augmentation	Relu and Adam	0.5700	0.5133

Model	Macro F1 - validation	Macro F1 - testing
Baseline ResNet34(4)	0.166794	0.153524
Improved ResNet34(4) (focal loss, oversampling, augmentation, differential lr)	0.612564	0.601216
ResNet34(3) & 1x1 Conv	0.615857	0.601246
DenseNet121(3) & 1x1 Conv	0.628404	0.609551
Best ResNet34 (3 stages of training)	0.635510	0.639354

Future Improvement

- External dataset.
- Weighted Focal Loss.
- K-fold stratification.
- Fine tuning thresholds or F1 score for different classes.
- Higher image resolution.
- More complex model architecture.
- Ensemble learning.

Conclusion

- Different CNN architectures and frameworks to classify mixed patterns of proteins.

- Application of resampling, data augmentation and focal loss to handle class imbalance

- Different ideas such as differential learning rate, freezing/ unfreezing layers to train model to improve model performance.