DATS6203 – Machine Learning II, Fall 2020
Final Project Proposal
Group 2: Tran Hieu Le, Voratham Tiabrat, Aashish Nair

Human Protein Atlas Image Classification

Introduction

Proteins are complex molecules made of thousands of amino acids, being responsible for many important functions in human body such as execution and regulation of issues and organs. Identifying the pattern and organelle localization of proteins would provide more insights about human living cells and accelerate the diagnostic of diseases. Moreover, understanding the complexity of cell structure plays a key role in developing medicine and treatment. Therefore, the classification of proteins has become a field of interest for many scientists and biomedical researchers.

Objective

The primary objective is to classify mixed patterns of proteins from microscope images of different human cell types. Since the complexity and highly various morphology of human cells, it is difficult to identify the structure and the number of protein patterns in organelles. Another challenge is the imbalanced distribution of proteins. While coarse grained cellular classes such as nucleus, plasma membrane and cytosol are very popular, small components such as endosomes, lysosomes and microtubule ends are rarely observed in cell structures. Consequently, the classification would project to the majority classes. The project will explore different techniques and methods to handle these two difficulties.

Dataset

The dataset for the project comes from the Kaggle competition which is originally provided by the Human Protein Atlas – a Sweden-based program researching proteins in cells, tissues and organs:

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

The "train.csv" data contains 31072 samples of 27 different living cells. Each sample is represented by four images, the protein of interest which is the main filter and three cellular landmarks as references: nucleus, microtubules and endoplasmic reticulum. The target is the protein pattern including 28 categories labeled from 0 to 27. Each sample may have one or more labels.

Platform

Due to the complexity of the dataset and the requirement of image augmentation, the project needs to be executed in a virtual machine with powerful GPU.

As the majority of image classifications, Convolutional Neural Network or CNN will be the approach for this project. The main utilized frameworks are Keras and Tensorflow since they provide a comprehensive set of functions and commands to build, train and evaluate CNN models. Pytorch would be used to experiment other pre-trained models that are not in Keras.

Methodology

In general, the original dataset will be split into training, validation and testing sets. The training set is used to train the model, whereas the validation set is used for fine-tuning hyperparameters. The final model is evaluated on the testing data.

Due to the imbalance of the target variable, accuracy is not a good measurement for model performance. The macro F-1 score, Cohen's Kappa and confusion matrix are the metrics to evaluate the models. However, the macro F-1 score will be the primary score as the competition also uses this metric for evaluation.

Schedule

No.	Date	Task
1	Nov 1 – Nov 7	Understand the topic. Read references and notebooks.
2	Nov 7 – Nov 14	Explore and visualize data. Data preprocessing. Base-line CNN model. Identify issues and learn to solve them.
3.	Nov 14 – Nov 21	Improve CNN model by changing parameters. Transfer learning using pre-trained model. Implement tricks and creative methods to improve the score.
4.	Nov 21 – Nov 28	Continue to improve the model performance. Compare the best models and previous models. Prepare final report.