Capstone Proposal

Kenichi Iwata

I choose a image classification of types of cultured cells and their gene expression as a capstone exam.
Because I am interested in life science field and applying AI/ML to research projects in this field.
There are a lots of tasks and problems that AI/ML can solve and contribute.
The classification/visualization of cell images which contains useful information is one of them.

The project's domain background

In life science field, a lots of cell line are used for researching molecular biology, disease and drug discuvery.
The RNA has a critical role for lives and deepl related to diseases but whole network of RNA world has not been revealed yet.
To understand RNA role in a morphology of cells is really important to diagnose disease and investigate drugs.

A problem statement

The problem is that to detect slight morphologicla change is really difficult for human eyes.
The system which can recognize and classify cell types by slight morphological change related to RNA expression with high accuracy is needed.

The datasets and inputs

I will use open source data "RxR1" which is provided from here.
This data source includes the 6-channel fluorescent microscopy images of different organelles - the nucleus, endoplasmic reticulum, actin cytoskeleton, nucleolus, mitochondria, and golgi apparatus.
There are 125,510 images in dataset and the size of each image is 512 x 512 x 6.
These images were gathered from lots of experiment which was carried out to investigate specific RNA function using siRNA, which is a popular tool for this kind of study.

A solution statement

I will try to build machine learning model that can califfier what RNA is "knock downed" from morphological change of the cell.

A benchmark model

A benchmark model is ResNet18 because it is representative of image classifier.
So I will create pretrained ResNet18 model to make a baseline for this proposal.

A set of evaluation metrics

I use CrossEntropyLoss as a loss function because this dataset includes many types of cell line, so it is multi class classification.
And I will use cross validation method to evaluate a model to ignore data split bias.

An outline of the project design

First, I download them and upload to s3 bucket. After obtaining dataset, I am going to explore the dataset and see detail of image of the cells.
Here I will check the size of images and visualize them to understand dataset.
Then, I will preprocess data and separate data into training set and test set.
I may not need to argument dataset because dataset I will use includes enough data to train.

After taht, I am going to try some models like SVM and CNN also pretrained ResNet as a base line model.
Finally, I make each model infference and evaluate model performance.
I will use cross validation method to evaluate and compaire each model.