Capstone Proposal

Human Protein Atlas Image Classification

Feb. 22, 2019
Dongmyunghee Kim
kedarnath6970@gmail.com

Domain Background

Localizing a specific protein in a human cell is essential for understanding cellular functions and biological processes of underlying diseases. A promising, low-cost, and time-efficient biotechnology for localizing proteins is high-throughput fluorescence imaging(HTI). Together with images of stained proteins or cell organelles and the annotation by the Human Protein Atlas project, these images provide a rich source of information on the protein location which can be utilized by computational methods. However, it is yet unclear how precise such computational methods are and whether they can compete with human experts.

The focus of this project is to implement convolutional neural networks("GapNet-PL") based on "Human-level Protein Localization with Convolutional Neural Networks[1] and compare this method with other approaches.

Problem Statement

This is an image classification problem. Inputs are four images per sample and the goal is to predict organelle localization labels for each sample. There are in total 28 different labels present in the dataset. All image samples are represented by four filters (stored as individual files), the protein of interest (green) plus three cellular landmarks: nucleus(blue), microtubules (red), endoplasmic reticulum (yellow). The green filter should hence be used to predict the label, and the other filters are used as references.

Datasets and inputs

The datasets are obtained on Kaggle competition website[1]. They are free to download.

- train.csv – filenames and labels for the training set.

- sample_submission.csv – filenames for the test set, and a guide to constructing a working submission.
- train.zip – All images for the training set.
- test.zip – All images for the test set.

Date fields:

- Id – the base filename of the sample. As noted above all samples consisted of four files – blue, green, red, and yellow.
- Target – in the training data, this represents the labels assigned to each sample.

Solution Statement

I would like to implement the model capable of classifying mixed patterns of proteins in microscope images based on the paper[1]. Up until now, Convolutional Neural network outperforms other methods in image classification and object detection. One of those methods is ResNet which has reached human level performance in image classification of general images. The architecture used in [1] is still based on CNN. However, they use Scaled Exponential Linear Unit(SELU) rather than RELU with batch normalization. They claimed that this approach lowers memory consumption to process high resolution images and computation time. The architecture is called “GapNet-PL”. This approach is compared with the result with ResNet.

Benchmark Model

There are two benchmark models. First, this is a Kaggle competition[1], the best Kaggle score for the test set will be used to compare with. The top scorer has F1 score 0.65602. The second benchmark model is the one the author[1] implemented. The performance of GapNet-PL is summarized as follows: 0.91 accuracy ,0.82 F1 score, 0.75 Precision, and 0.95 Recall. I will compare the performance of the model I implemented with Kaggle top leader and the performance result from the study[1].

Evaluation Metrics

The model prediction for this problem can be evaluated in several ways. Since this is a Kaggle competition project. I will use the same evaluation metrics; F1 score as well as precision and recall.

Project Design

There are three approaches I will take. The first one is to use ResNet and transfer learning and measure the performance. The second is to implement GapNet-PL with RELU. The last one is to adjust GapNet-PL with SELU[2]. ResNet will be a baseline architecture for this project since ResNet is well established method for image classification. CNN with RELU needs to downscale high-resolution images due to memory consumption and computational time. With GapNet-PL and SELU activation function, the model will take original resolution images without downscaling and therefore capture fine features of high-resolution images.

References

1. *Elisabeth Rumetshofer*, *Markus Hofmarcher*, *Clemens Röhrl*, *Sepp Hochreiter*, *Günter Klambauer* (2018), Human-level Proten Localization with Convolutional Neural Networks (under review)

2. Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems 30, pages 971–980. Curran Associates, Inc.

3. Kaggle, Human Protein Atlas Image Classification, Classify subcellular protein patterns in human cells, https://www.kaggle.com/c/human-protein-atlas-image-classification