Large-scale Protein Atlas Compartmentalization Analysis

Zijian Zhang Kuangcong Liu Wen Zhou {zzhang7, cecilia4, zhouwen}@stanford.edu
Department of Chemical and Systems Biology, Department of Computer Science, Stanford University

Introduction

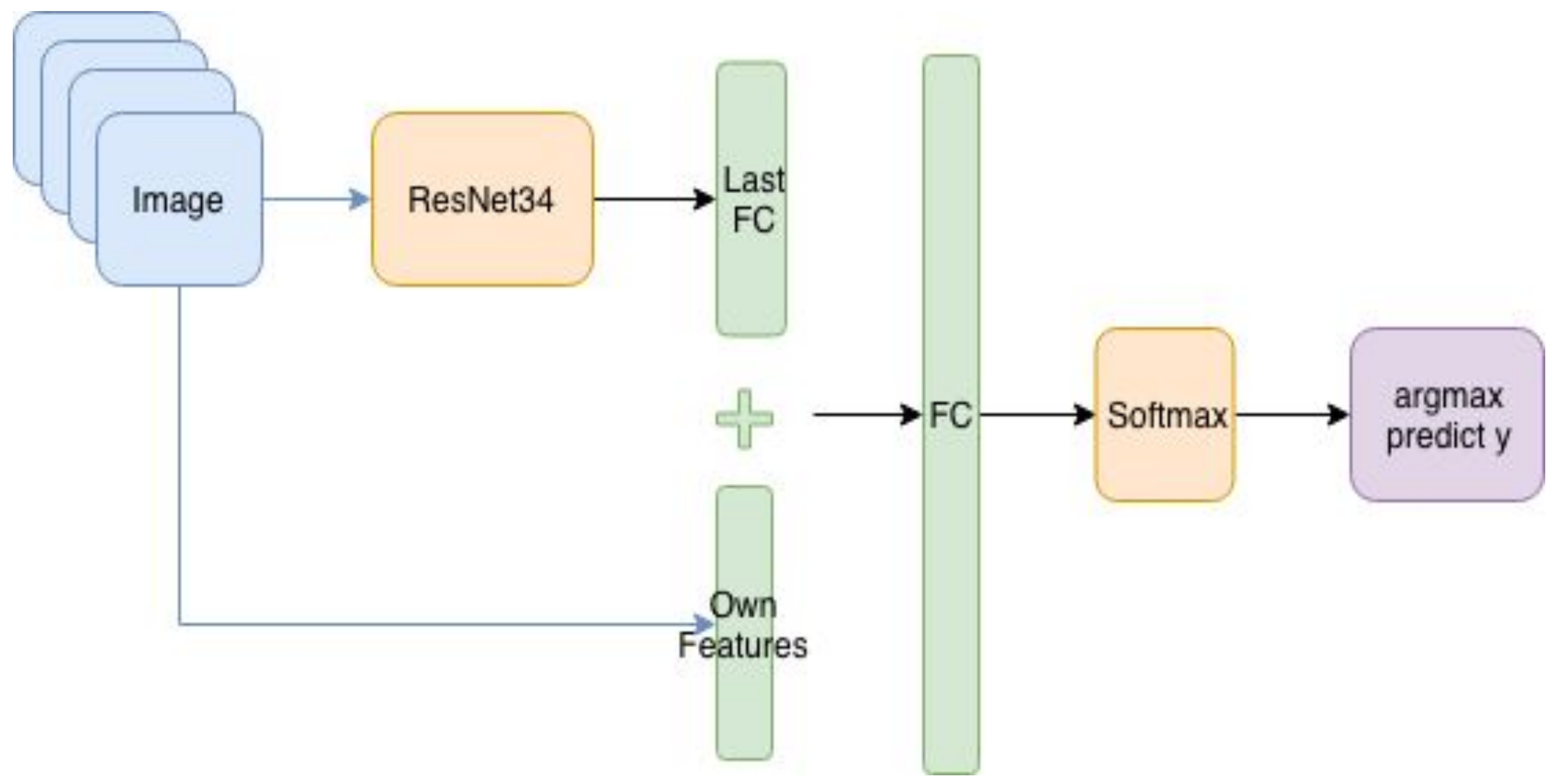
We aim to establish a model to **predict the localization of specific proteins** in the cells, which can help biology researchers to gain more insight into the regulation of protein function, interactions as well as their roles in human diseases.

We focus on analysis of a Kaggle problem, Human Protein Atlas Image Classification. To accomplish our goal, we use Computer Vision approaches to extract images features selected by biological knowledge about proteins, and then use **multiple Resnet models** combined with our **extracted image feature scoring matrix**, to tackle this problem.

The results show that we could achieve decently accurate prediction of protein localization across various cell types. Our feature scoring matrix however, needs more fine tuning in order to boost the effect.

Models

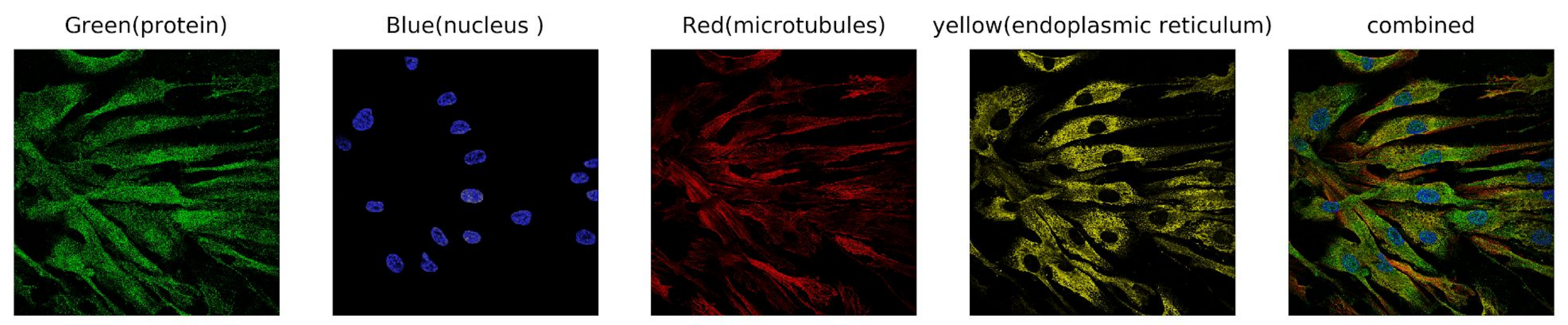
- Treat the last fully connected layer output from Residual Network [1] as a graph representation of each the image sample.
- Combine this graph representation with our designed features for each image sample to add more expertise -based protein classification features on locations and morphology.
- Features are normalized and given individual weights.
- We use another fully connected layer after the combination to capture more non-linear relations between resnet representations and our image features.
- In the end a softmax layer to select the class labels with highest probabilities.

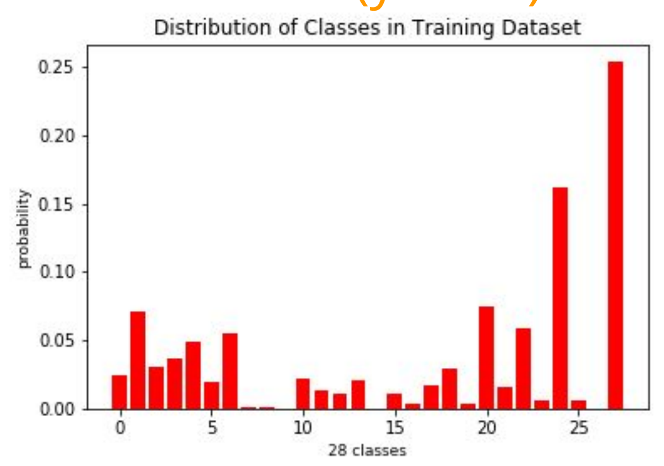


Dataset

The dataset is provided by Kaggle. There are 31072 samples in the train dataset, and we also perform data augmentation technique on it. Images were resized to 512x512 or 224x224.

In this dataset, each sample is a microscope image of a type of cells that contain a certain type of protein. The image are shown in 4 filters: the protein of interest (green), and three cellular landmarks: nucleus (blue), microtubules (red), and endoplasmic reticulum (yellow):





The protein organelle localization is represented as integers 0-27. The right distribution shows the unbalanced probability distribution of 28 classes in our training data.

Due to the overlapping information between the yellow and red channels, we removed the yellow channel from the input, but extracted its critical features during data preprocessing [2].

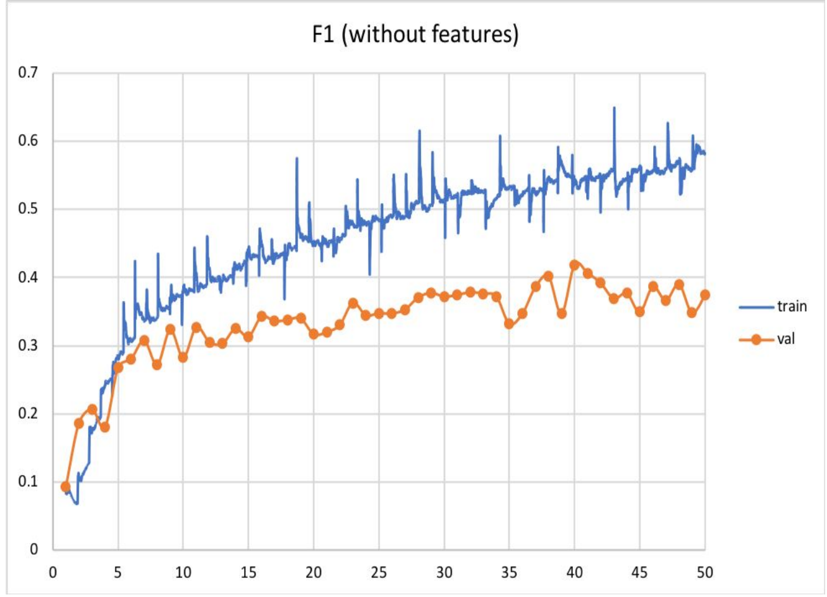
Results

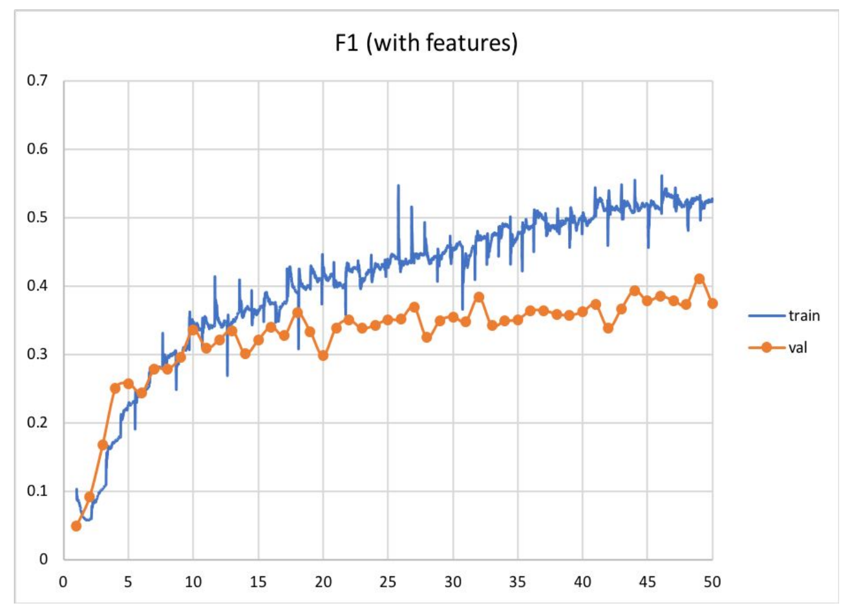
1. We trained both Resnet50 and Resnet34 (without our extracted features) for 20 epochs, but Resnet50 would be badly overfitting as its validation set loss remained high while its train set loss was continuously decreasing. Hence we continued with Resnet34.
2. Compared macro-f1 score for (1) Basic Resnet34 and (2) Resnet34 + Features for 50 epochs.

$$precision = \frac{tp}{tp+fp}$$

$$recall = \frac{tp}{tp+fn}$$

$$F_1\,score = 2 * \frac{p*r}{p+r}$$





3. We did multiple experiments on different parameters of our model:

Model	Epoch	Kaggle Score
Resnet34 + RGBY 224 + zero initial weight for 4th layer of input	**20**	**0.366**
Resnet34 + RGBY 224 + zero initial weight for 4th layer + threshold	**40**	**0.387**
Resnet34 + RGBY 224 + pretrained weight for 4th layer + threshold	**40**	**0.403**
Resnet34 + RGBY 224 + pretrained weight for 4th layer + threshold + additional dropout	**40**	**0.394**
Resnet34 + RGBY 224 + pretrained weight for 4th layer + threshold + TTA	**40**	**0.401**
Resnet34 + RGB 512 + pretrained weight + threshold + own features	**50**	**0.351**

Data Features

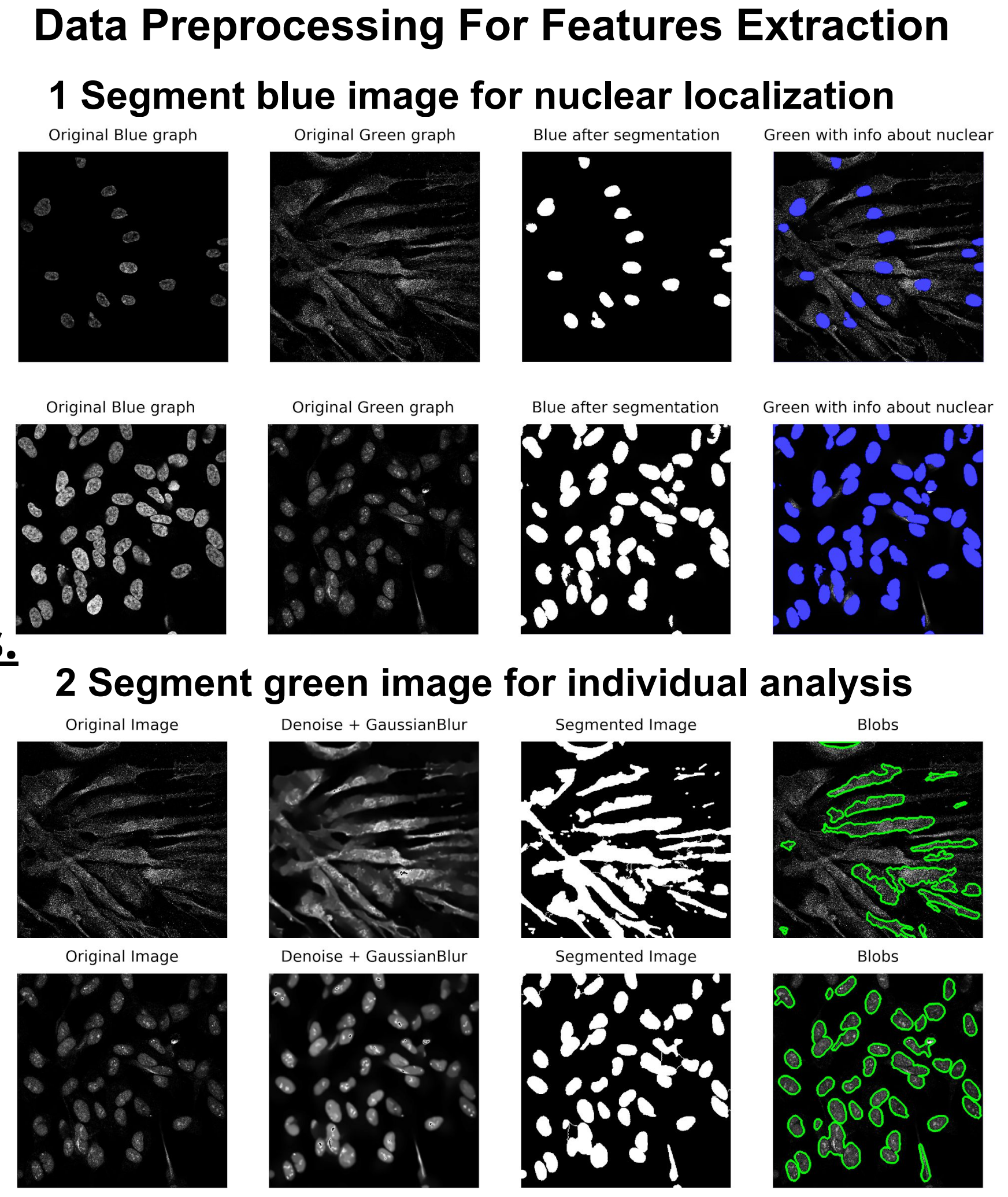
We have developed 10 data features:

Overall properties:

1. Relative ratio of green in blue **(localation)**
2. Structural similarity between green & red
3. Structural similarity between green & yellow
4. Structural similarity between green & blue
5. Total intensity of green / yellow **(Intensity)**
6. Area size of green above background **(Size)**

Information about individual protein segments.

7. Averaged **compactness** of protein segments **(shape)**
8. Averaged **eccentricity** of protein segments **(shape)**
9. Average area size of each protein segments / nucleus **(distribution)**
10. Average distance of each protein blob to the closest nucleus **(distribution)**



Analysis

❖ **Problem:**

➢ From the figure in Results section we can see that val f1 is lower than train f1.

➢ Our feature scoring matrix did not significantly boost the prediction accuracy.

❖ **Analysis:**

➢ Gaps between train f1 and val f1 shows that our model is overfitting. In the future, we will try to resolve this issue by: (1) leaning rate annealing: use periodic learning rate that first increase and then slowly decrease to drive the model out of steep minima; (2) add more external training image data from other sources.

➢ Besides, small scale of our own features, which is now mostly within range of (0, 1), may prevent the gradient being effective in back-propagation. So in the future, we will try to scale our own features by higher factors to make back propagation capture more information about it.

❖ **More future improvement work includes:**

- assign weights to each of our designed features according to their importance
- develop more features
- use different classification thresholds for validation set and test set according to their data statistics.
- more design about the fully connected layers after we combine our own data features with the resnet outputs, which need more explore on this research topic of combining features.

References:
[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016.
[2] Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., ... & Smith, K. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology*, 36(9), 820, 2016.