Convolutional Neural Networks for Protein Image Classification

Nick Littlefield
*University of Southern Maine*

Follow this and additional works at: https://digitalcommons.usm.maine.edu/thinking_matters

Part of the Computer Sciences Commons, and the Medicine and Health Sciences Commons

**Recommended Citation**

https://digitalcommons.usm.maine.edu/thinking_matters/208

This Poster Session is brought to you for free and open access by the Student Scholarship at USM Digital Commons. It has been accepted for inclusion in Thinking Matters Symposium by an authorized administrator of USM Digital Commons. For more information, please contact jessica.c.hovey@maine.edu.
Abstract

A solution to the Kaggle competition: Human Protein Atlas Image Classification (HPA). Using microscopic images of cells provided by the Human Protein Atlas, convolutional neural networks, CNNs, were used to analyze and predict the location of protein patterns. Challenges included working with an unbalanced dataset, finding a correct learning rate, and choosing a correct architecture to solve the problem.

Background

Kaggle is an online platform for predictive modelling and analytics competitions, that hosts the HPA competition. In this competition, competitors try to build models to classify mixed patterns of proteins found in microscopic images. The Human Protein Atlas is based in Sweden and aimed at mapping all human proteins in cells, tissues, and organs. They will use models from the competition to help classify these patterns.

To solve this problem, convolutional neural networks are used to analyze the images. By feeding an image into the network, complex patterns and features can be extracted that aren’t easily detected.

Methods

Several different techniques were used to try and train the best CNN for the data. These included:

• Transfer learning
• Training networks from scratch
• Ensemble methods

Along with this we used different:

• Learning rates
• Augmentations
• Learning rate schedulers

Evaluation Metric

The models submitted were evaluated using the F1 score. This is defined as:

\[
F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

Where precision is calculated as:

\[
\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}
\]

and recall is calculated as:

\[
\text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\]

Results

Using different deep learning libraries effective models were able to be built. The results in Table 1, show the F1 scores both locally (during training) and when submitted to Kaggle.

Locally, the scores are high and good. When submitted to the competition the scores would be significantly lower (roughly half the local F1).

Discussion

The most difficult challenge with the competition was the difference between the training and test sets. As can be seen in Figure 2, there are labels that have over 12,000 occurrences and then some, such as rods and rings that barely had any. This made it difficult to classify the very rare labels. In the test set this distribution wasn’t the same. The models, therefore, had difficulty making accurate predictions.

The use of the discussion boards and kernels (sample code) allowed for an understanding of both the data and the problem with the distributions. It also provided potential solutions to be used in the next steps to improve the models. Solutions included: stratification of the data so the distributions in the training set is equal and using external data from Human Protein Atlas.

Links

Human Protein Atlas: https://www.proteinatlas.org/