College of Arts & Science / OSU / Physics

Identifying Metastatic Tissue in Histopathologic Scans of Lymph Node Sections Using Convolutional Neural Networks

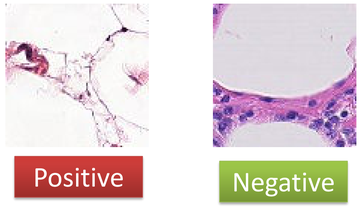
Daniel Sanchez-Rosales

INTRODUCTION

Machine learning has caused a tremendous rise in big data exploration. From face recognition to earthquake prediction, machine learning algorithms have helped us dramatically improve the lives of people. In this project we set to develop an algorithm that analyzes histopathologic scans of lymph node sections and predicts with high confidence whether cancer is present in the tissue. Here we have 120,000 samples which include training and test samples.

This was originally a competition in Kaggle. All the images and data is available in their website. Some of the best accuracy achieved by people around the world was between 88-91% accuracy. Below are some example of the images provided and their labels.

Figure 1: Samples from the data showing their respective labels.

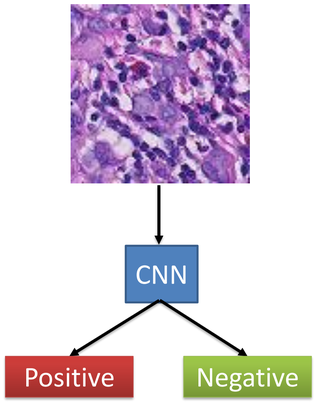


AIM

The goal of in this project was to develop a convolutional neural network, which aids in the diagnosis of cancer patients by predicting whether histopathologic scan contains traces of cancerous tissue.

The algorithm would simply take an image (histopathologic scan of a lymph node section), then analyze it in a trained convolutional neural network, and output a prediction: a positive or a negative. The process is shown below:

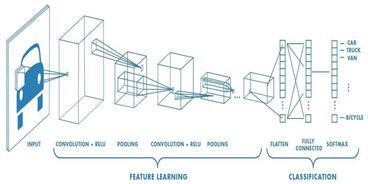
Figure 2: Diagram showing the order in which the tasks were performed



METHODS

In this project the algorithm uses a convolutional neural network (CNN) to classify the images as positive(1) or negative (0). CNNs are useful because they allow us to identify structure in the images. CNNs also help us to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction, as shown in the image below:

Figure 3: Diagram showing how CNNs typically work



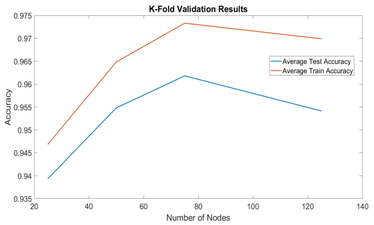
The first step of a CNN is to take the input image (in our case an RGB image) and pass a filter (in our case a 3 x 3 x 3 filter), creating a convolutional layer, which contains information about structure in the image. This layer is then reduced further in size using Max Pooling (in our case we us3 a 2 x 2 x 3 filter). However, this new max pooled layer contains no new information about the structure in the image. This is simply done to ease the computational requirements of the algorithm. This completes a single convolutional net; we can add as many of these as we want. In our case we use 2, each containing 32 convolutional layers.

We then take the output of these convolutional nets and flatten it (turn it into a 1- dimensional array). Finally, the flattened input is fed to a fully connected neural network (FCN). The output of the FCN is either a 1 or a 0.

The FCN contains parameters that need to be optimized, such as the number of hidden nodes, or the number of epochs. To optimize this, we use K-Fold validation (5 folds) and use the average results to finally train our algorithm with 100,000 samples and test in 20,000).

RESULTS

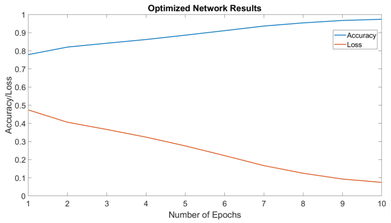
5-Fold validation returned the following results:



Clearly, the best results are achieved with about 75 nodes in the FCN. The average number of epochs was 10 for this example.

Using these optimized results, we then take the entire 100,000 train samples and labels and train on them using the optimized parameters. Achieving about 97% training accuracy, as shown below:

Figure 5: Optimized Network accuracy and loss



Finally we take the 20,000 samples that the network has not seen before and test our model. We achieve 81.6% accuracy. This means that given an image of a histopathologic scan of a lymph node section the algorithm can predict the correct diagnosis about 82 times out of 100, which is much better than guessing

CONCLUSIONS

The algorithm used in this project can predict the correct diagnosis 81.6 % of the time. This is already much more accurate than guessing and can still be improved upon.

CONCLUSIONS

There seems to be a case of overfitting here since our test accuracy deferred quite a bit from the training accuracy. This can be solved by training with more data, or by optimizing other parameters, such as the layer sizes. In the case of this algorithm, size of the kernels (filters) used was arbitrarily picked to be 3 x 3; similarly, the number of layers was arbitrarily picked to be 32; and we only used two convolutional nets. All of these parameters can be modified and optimized to give better results. Finally, the images themselves can also be altered (i.e. change contrast) to allow for the algorithm to identify features more clearly.

BIBLIOGRAPHY

1 https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

2 https://www.kaggle.com/c/histopathologic-cancer-detection/data

3 https://www.youtube.com/watch?v=KTB_OFoAQcc&index=6&list=PLkDaE6sCZn6GI29AoE31iwdVwSG-KnDzF

4 https://towardsdatascience.com/a-beginners-guide-to-convolutional-neural-networks-cnns-14649dbddce8

5 https://cv-tricks.com/tensorflow-tutorial/training-convolutional-neural-network-for-image-classification/

ACKNOWLEDGEMENTS

Prof. Richard Hughes.

THE OHIO STATE UNIVERSITY