Gaberial Campese

Machine Learning II

December 9, 2019

Final Project Individual Report

Introduction

The purpose of this project was to utilize the complex Diabetic Retinopathy Dataset from Kaggle to create an effective image classification model. It is recognized that Diabetic Retinopathy is a serious, prominent health issue and it is also the leading cause of blindness. There is great appeal in this research in that this is an important issue that we can apply our learning from Machine Learning II to ultimately help medical professionals and patients. My partner and I decided to use the pretrained ResNet50 Convolutional Neural Network (CNN) in Keras to train our model because of the interest we had when it was discussed late in the semester in class.

Individual Work

The purpose of this section of the report is to provide an overview detailing my individual efforts in our Final Project in Machine Learning II. Each of our contributions were equal and we met in person to collaborate for nearly all of this project so that we could pair program and share ideas with immediate feedback. Our approach was quite collaborative and an overall group effort, yet both of us did take responsibility in certain areas of the research.

In terms of coding, I provided the overarching framework to be used in this network, Fernando focused more on the fine tuning and preprocessing. I had originally gotten our code to execute and we continued to work together on problems we ran into such as memory errors and augmentation. I had helped significantly in the training, fitting, and graphing for our network.

Specifically, I helped in the experimentation with different activation functions and optimizers and pivotal in consolidating the results. Moreover, I believe that my Linux expertise truly helped in this project when we were having problems in the instance itself such as moving folders and copying the data effectively. We each used each other's strengths to our advantage!

My more rigorous contributions were focused mainly on the deliverables for this project. I had initially set up the GitHub Repository so that my partner and I could collaborate and store the assignments for the project. Also, I had stood up Google Drive and took charge in the creations of the report and presentation. Moreover, I constructed the README in our GitHub homepage.

Aside from my previous contributions, I was also able to store the subset data in a Google Cloud Platform Bucket. The purpose of this is so that any user, interested in researching the subset data further, can easily pull it into their environment using the wget command.

Results

The accuracies for the models appeared to be good at 74.31% for training and 82.88% for validation. The accuracies were constant among the majority of epochs. The produced Cohen Kappa Score from our model was 0.0 in all trials. The F1 score that was generated was 0.1812. I learned quite a bit about these metrics including the fact that Cohen Kappa at 0 equates to "chance", but the ranges are up for opposition by many.

Summary & Conclusions

Despite the high accuracies, we realized that accuracy does not automatically mean that a model is successful. We did not intend to blindly communicate and conclude that the accuracy was the determining success factor. Our accuracies were also static when training through each epoch, which is a problem. With that said, this is why we looked into Cohen Kappa and F1

scores. These scores were not promising, which means that the model is not as good as the accuracy potentially makes it out to be. The distribution of the number of images in each class does not help with this because the model may be accurate in one class and not so accurate in the sparse classes.

The magnitude of the data was an overwhelming challenge and I would like to better utilize it in the future. It is 82GB and our capabilities in our GPU node could not process the entirety of the data. In the future, I would like to have more robust capabilities to use the whole dataset. I would also like to further research network architectures running in ensemble.

References

Chollet, François. (2018). Deep learning with Python. Shelter Island, NY: Manning Publications Co.

Diabetic Retinopathy Detection (dataset). (2015). Retrieved from https://www.kaggle.com/c/diabetic-retinopathy-detection/.

Dwivedi, P. (2019, January 04). Understanding and Coding a ResNet in Keras. Retrieved from https://towardsdatascience.com/understanding-and-coding-a-resnet-in-keras-446d7ff84d33.