Machine Learning Capstone Project Proposal

Aayush Bhaskar

May 14th, 2018.

Domain Background:

The project that I am undertaking is Diabetic Retinopathy Detection, which falls under the healthcare sector. Since the Machine Learning wave has swept across the world, they have carved their niche in the healthcare domain too. IBM's artificial intelligence system Watson has been used in the healthcare and medical sector since 2013, and Google's DeepMind Health has also make lot of advances in the field of medical support. Microsoft's Inner Eye initiative, started in 2010, is presently working on image diagnostic tools.

Diabetic retinopathy, also known as diabetic eye disease, is a medical condition in which damage occurs to the retina due to diabetes mellitus and is a leading cause of blindness. There have been several works on the employment of machine learning to the detection of diabetes retinopathy. Employment of machine learning would speed up the rather slow process of detection of this disease. A few years ago, Google research team studied whether ML could be applied for detection of Diabetic Retinopathy or not. Then on 29th November 2016 Google published an article stating: "Today, in the Journal of the American Medical Association, we've published our results: a deep learning algorithm capable of interpreting signs of DR in retinal photographs, potentially helping doctors screen more patients, especially in underserved communities with limited resources." This article can be referenced here: https://www.blog.google/topics/machine-learning/detecting-diabetic-eye-disease-machine-learning/. Therefore I also want to work on this problem statement to solve this problem.

Problem Statement:

The presence of diabetic retinopathy (DR) in each image is rated on a scale of 0 to 4, according to the following scale:

0 - No DR
1 - Mild
2 - Moderate
3 - Severe
4 - Proliferative DR

The problem here to be solved is to create an automated analysis system capable of assigning a score of diabetic retinopathy based on this scale. This system will take the aid of machine learning to solve the problems by classifying the input images accordingly, and assigning the required score to it.

Dataset and Inputs:

The dataset being used here is a part of the dataset that was hosted on Kaggle, as a part of the online competition on DR detection, sponsored by California Healthcare Foundation. The dataset is provided by EYEPACS, a free platform for DR screening. The dataset contains about a 1000 high quality scans of the eye, one left and one right for each patient, and a label attached to it, based on the score scale described above.

The input should also contain 2 scans, 1 for each eyeball, and the output will be the scores.

The dataset can be obtained from here: https://storage.googleapis.com/kaggle-competitions-data/kaggle/4104/train.zip.005?GoogleAccessId=web-data@kaggle-161607.iam.gserviceaccount.com&Expires=1526547455&Signature=FczWaZgcg5EsxoGHH0v7uINOMFqOTiTW8BVeHFeAdaqSGl3NsP9xHncp0nZBw7JoVBnZzdZc4y4Fr2LNi7ANL18vKndgx2iWo2Ovnq11%2F%2FzhR%2Fcu4n%2FDExe7x2%2BcJ48wBL68O%2B4HuVr6DUorS7HZY%2F7s0mgLY1UZDeWm29ifd5klUCQCQeptPnsfK9cnzGPkju%2BbiUbQj9UoHKa84JFhTahItZWxAK0PRmL7A%2FMHUXYg5j2jHxaMlhHD1HjLxFrIjmb3y5FoQW1uxz79X4ghcAvcFZev19tpfv9JPadXRGPlGFrLcMFSgCPEhIvHrXxOp6jHNh7PQ6Pmy3S2bgxa%2Fg%3D%3D.

The algorithm will read in the images, and will analyse its properties to generate the classifier.

Solution Statement:

The problem dataset contains images, and looking at its basic concept, this problem, at its core, is an image classification problem. So it would be logical to solve this problem using an efficient CNN model. So, the solution would be to create a relevant CNN model, and get the scores for the new images using our classifier.

Benchmark Model:

Considering the Kaggle contest held 3 years ago, the best solution of the contest had an accuracy of 86% on test data, and 55% on new data (from a different source). This would serve as the benchmark for the solution, considering the use of only one part of the dataset. The best solution to the problem can be found here: https://www.kaggle.com/kmader/vgg16-640hr-nloss-retinopathy/code.

Evaluation Metrics:

The evaluation metrics for this project as well as the benchmark solution is checking the accuracy of classification of the images based on their score. A confusion matrix will also be plotted, which will help compare the scores. A confusion matrix shows how much data has been correctly classified, and how many classified into other classes.

Project Design:

The most important part in any Machine Learning project is the data. We need clean data to build an efficient machine learning model. So, the first step in this project would be to analyse the dataset.

The dataset contains images, so we need to create arrays of the image's pixels, using OpenCV, and then store it in csv format using the pandas library. I will employ data cleaning mechanisms to fill in the missing data, or to remove the ambiguous data or outliers.

The next step to the project would be to create a machine learning model so that the classifier can be generated. As stated earlier, this is an image classification problem at its core, so an efficient CNN will do the task. So, my next goal would be to design a CNN and train the model using the dataset. Once this task is done, we store the weights onto the computer so that it can be employed later on to test the model. The CNN model would be created using Tensorflow model. Since my system has an AMD graphics card, so I would be using the CPU version of Tensorflow only.

Next comes validation and testing of data. We would do validation of the model, so that we can improve upon it, by tweaking the parameters. Then, I will test the model on the test data. Finally I will complete the project by creating the confusion matrix, and writing a report on it. The evaluating metrics, accuracy and confusion matrix will be included in the report.