Machine Learning Engineer Nanodegree
Capstone Proposal

Prabhat Sharma
June 15th, 2019

Detecting Malaria with Deep Learning

Domain Background

Malaria is a mosquito-borne infectious disease caused by Plasmodium parasites. Malaria causes symptoms that typically include fever, tiredness, vomiting, and headaches. In severe cases it can cause yellow skin, seizures, coma, or death. In 2016, there were 216 million cases of malaria worldwide resulting in an estimated 445,000 to 731,000 deaths.

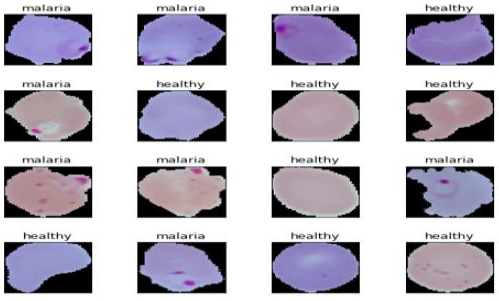
DETECTION

According to WHO protocol, detection of malaria involves intensive examination of the blood sample at 100X magnification, where people manually count red blood cells that contain parasites out of 5000 cells. According to the paper by Rajaramen *"Thick blood smears assist in detecting the presence of parasites while thin blood smears assist in identifying the species of the parasite causing the infection (Centers for Disease Control and Prevention, 2012). The diagnostic accuracy heavily depends on human expertise and can be adversely impacted by the inter-observer variability and the liability imposed by large-scale diagnoses in disease-endemic/resource-constrained regions (Mitiku, Mengistu & Gelaw, 2003). Alternative techniques such as polymerase chain reaction (PCR) and rapid diagnostic tests (RDT) are used; however, PCR analysis is limited in its performance (Hommelsheim et al., 2014) and RDTs are less cost-effective in disease-endemic regions (Hawkes, Katsuva & Masumbuko, 2009)."*

Thus, malaria detection is an intensive manual process which be automated using AI deep frameworks like keras to build robust, scalable and effective deep learning solution which being open-sourced and free, enable us to build solutions which can be cost effective.

Problem Statement

This is an image classification problem. Inputs is balance set of parasitized and uninfected image cells and the goal is to predict if cell is infected or uninfected of malaria.



I will be tackling problem by using Convolution neural network. I will build three deep learning models, train them with training dataset. Each model will be saved and evaluated. Model 1 will be basic CNN model. Model 2 will be pre-trained model as a feature extractor and Model 3 will be Fine-tuned pre-trained model with image augmentation.

Datasets and Inputs

This Dataset is taken from Kaggle datasets mentioning that the data is available on official NIH Website: https://ceb.nlm.nih.gov/repositories/malaria-datasets/. They are free to download.

Input Data files are in two folder

- *Parasitized* – folder containing infected file
- *Uninfected* – folder containing uninfected file

Solution Statement

The solution will be predictions of either infected or not in the test dataset. First I will build a basic CNN model to process all the images and do accuracy calculation using test data to get some understanding. Then I will create more advance CNN model using transfer learning algorithm VGG-19 model as a feature extractor. Lastly a fined-tuned pre-trained model with image augmentation will be created using VGG-19 and ImageDataGenerator.

Finally, each model performance will be evaluated using F1 score.

Benchmark Model

For this problem, the benchmark model will be CNN model with at least 80% F1-score on test data.

Evaluation Metrics

Prediction results are evaluated on the F1 score between the predicted infected images and the actual infected image in testing data. We will also check accuracy, precision and recall of each trained model.

Project Design

Before even start training models, I will first take observation of the both infected and uninfected image see what the basic difference in the cell structure and difference in structure. Then I will label the data as malaria and healthy. Next process will be splitting the data for train, validation and test datasets. If images are not of uniform dimension than I will select an optimal dimension based on statistical interpretation.

To train models, I plan to choose 3 different models to compare. Because this is a classification problem, a few approaches in my head would be as mentioned above. With Basic CNN, pre-trained model as a feature extractor and Fine-tuned pre-trained model with image augmentation.

I expect to spend 20% of the time on data processing part and 60% of the time on training models and tweaking parameters and also 20% in model evaluation.

Reference

1. https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria
2. https://ceb.nlm.nih.gov/repositories/malaria-datasets/
3. https://en.wikipedia.org/wiki/Malaria