Machine Learning Engineer Nanodegree

Capstone Proposal

Huan-Wei Wu May 13rd, 2019

Proposal

Domain Background

Malaria is a mosquito-borne infectious disease that affects humans and other animals. Malaria causes symptoms that typically include fever, tiredness, vomiting and headaches. In severe cases it can cause yellow skin, seizures, coma, or death[2]. According to WHO data: In 2017, an estimated 219 million cases of malaria occurred worldwide and cause 435 000 deaths. Most cases were in the WHO Africa Region[3].

Malaria is caused by Plasmodium parasites. The parasites are spread to people through the bites of infected female Anopheles mosquitoes. [1].

According to CDC U.S., microscopy examination remains the "gold standard" for laboratory confirmation of malaria. By visually inspecting the blood smear specimen collected from patient under the microscopy , the laboratorian can determined if this patient is infected. However, the proficiency of laboratorian becomes a problem.

ref:

1. https://www.who.int/news-room/fact-sheets/detail/malaria
2. https://en.wikipedia.org/wiki/Malaria
3. https://www.who.int/malaria/en/
4. https://www.cdc.gov/malaria/diagnosis_treatment/diagnostic_tools.html#tabs-2-1

Problem Statement

1. In some non-tropical regions , malaria became rare (average 1700 cases in ths U.S per year), the laboratorians does not perform this test regularly.
2. In some regions which are lack of medical resources (also suffered the most), well-trained Medical personnel who are able to justify test result, also hard to find.

Base on previous statements, we can implement machine learning on the task classifying the microscopy images ,make the justification much more easier for medical personnel who may not so familiared with this task.

Datasets and Inputs

The datasets comes from https://ceb.nlm.nih.gov/repositories/malaria-datasets/. The cell images of Giemsa-stained thin blood smear slides from 150 P. falciparum-infected and 50 healthy patients were collected and photographed at Chittagong Medical College Hospital, Bangladesh.

The dataset contains a total of 27,558 cell images with equal instances of parasitized and uninfected cells.

ref:

1. https://ceb.nlm.nih.gov/repositories/malaria-datasets/

Solution Statement

We can train a CNN like model for the task , which should have acceptable accuracy/precision/recall. Also, considering we may need to run this model for making predictions on some terminal devices(like mobile phones), the model should by lightweight.

Benchmark Model

After surveyed several report, the reported accuracy of microscopy method is around 93~95% , the model should be comparable to that (Although the data source is different in these cases). Also , compare this from scratch model with other models trained by transfer learning.

ref:

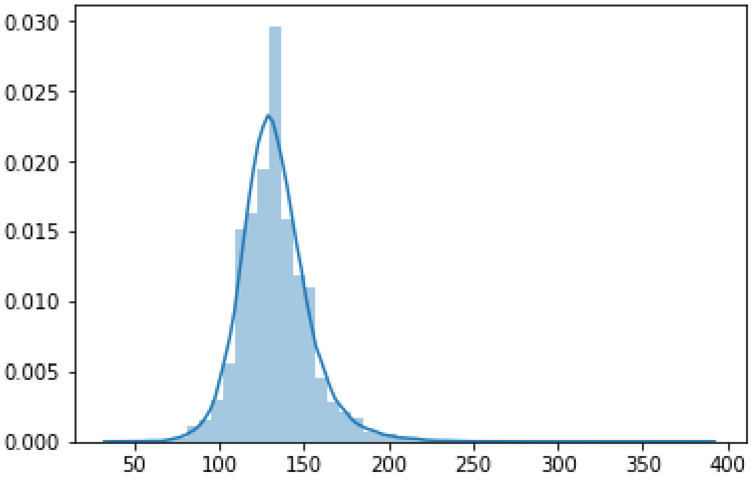
1. https://www.hindawi.com/journals/jpr/2019/1417967/
2. https://www.ncbi.nlm.nih.gov/pubmed/19407111

Evaluation Metrics

Since our data source is categorical balanced (equal instances for parasitized/uninfected) , accuracy should be ok for the model evaluation . We can also put precision/recall into consideration too. (we care more about recall than precision in disease detection task.)

Project Design

1. Split data into train/validation/test split in ratio 0.8/0.1/0.1.
2. Create image augmentation object for preprocessing and resizing , since the distribution of image sizing is from 50~200, resizing to 64*64 for computational effectiveness (also because resize image to bigger one will not increase the "information" contained).



3. Build CNN by using separable convolution layer (lightweight) with "selu" activations (self normalizations) and Global average pooling layer + dense layer for output.
4. Train and select the model with best validation accuracy, and compare with some transfer learning models.
5. Implement Grad-cam for CNN visualization and model check.

-