Machine Learning Engineer Nanodegree

Capstone Proposal

Amr Elbana
March 13st, 2022

1) domain background

Parasitism is the most frequent method of life; parasites make up more than half of all animal species. Parasites can be present in all animal species and can have a significant impact on human, domestic animal, and wildlife health.

Parasitology is the study of parasitism. It is an interdisciplinary field that includes morphology, taxonomy, biology, behavior, life cycles, pathogenesis, epidemiology, ecology, physiology, biochemistry, genetics, molecular biology, infection detection, immunology, and therapy.

2) Problem Statement

In both developing and developed countries, parasitic protozoa and helminth infections cause significant death, misery, and economic loss. The magnitude of the crisis confronting the international community cannot be overstated. Malaria infection, for example, is one of the most common and severe infections in impoverished nations, with 300-500 million clinical cases and 1-2 million fatalities per year, especially among children under the age of five. Malaria, on the other hand, has an insidious effect, reducing economic growth by 1.3 percent every year in Africa alone (p.a.). Animal productivity and food production are also threatened by parasites. Parasitic worms infect around 500 million big ruminants globally, resulting in annual economic losses of over $3 billion.

Despite many advances in parasite treatment and control, infections continue to spread due to a variety of factors, including urbanization (crowding); more intensive farming systems, greater animal translocation, further land, and marine development, inadequate effluent disposal, parasite drug resistance, and vector insecticide resistance.

The primarily investigated problem is malaria patients' classification via Convolutional Neural Network (CNN) and transfer Learning.

3) Datasets and Inputs

The dataset comprises segmented cells from the Malaria Screener research activity's thin blood smear slide photos. Researchers at the Lister Hill National Center for Biomedical Communications (LHNCBC), part of the National Library of Medicine (NLM), have developed a mobile application that runs on a standard Android smartphone attached to a conventional light microscope to reduce the burden on microscopists in resource-constrained areas and improve diagnostic accuracy. For each minuscule field of view, the smartphone's built-in camera captured photographs of slides. A professional slide reader manually annotated the photos.

a) The dataset contains parasitized and uninfected cells.
b) There are a total of 27,558 cell images with equal instances 2.
c) There are 13780 total Un-infected images.
d) There are 13780 total parasitized images.
e) In my project I will only use 1000 for each class as the project was developed locally and my pc is slow.
f) Project is linked with GitHub, and I commit every change to the GitHub repository.
g) The datasets were included in the project local directory after being downloaded via the provided links and unzipped locally.
h) The structure of the GitHub repository was made using CookieCutter. Package used to structure the repositories.
i) The dataset was used to create training, validation, and testing subsets for the CNN classifier.
j) The datasets are appropriate given the context of the problem.

4) Solution statement

For this project, I will use deep learning to reach the final solution. Deep learning is based on a neural network which is just an algorithm from machine learning algorithms. This algorithm did not require that the data be in a tabular form.

Convolutional neural networks are usually used for the image dataset as they show a great ability to extract the main features from the images.

The network tries to find the pattern of features that are found in the parasite cell and the uninfected cell and to learn how to differentiate between them with high accuracy.

As the model tries to classify the input images as parasite cells or uninfected ones. I will try to build a CNN from scratch and use a pre-trained model which is VGG-19 with batch normalization trained on ImageNet.

5) Benchmark

I will build one model and use one pre-trained model on the data and try to compare which one will give me the better result.

6) Evaluation metrics

Both models' performance was evaluated using the accuracy metric: the number of correct predictions divided by the total number of predictions. This. Accuracy is one of the widely used metrics for classification evaluation.

Also, the confusion Matrix is usually used to evaluate the classification algorithm It has 4 categories: True positives, True negatives, false positives, and false negatives. Using this matrix, we can calculate various useful metrics!

7) Project Design

I will include the design in step:

a) Import all libraries
b) Check that there are not any corrupted images.
c) Split the data into train, validation, test sets
d) Construct the image generator to resize and rescale the image
e) Make augmented files for the train images
f) Build the model from scratch and train the data.
g) import the pre-trained model and train the data.
h) Test the algorithm on sample images.
i) visualize the train and validation loss and accuracy
j) Check which one of the models has better accuracy
k) Build a web application