

Machine Learning Engineer Nanodegree

Capstone Proposal

Malaria Cells Detection

Daniel Valcarce

May 14th, 2019

Domain Background

According to the *World malaria report* [1](#), in 2017, an estimated 219 million cases of malaria occurred worldwide (95% confidence interval [CI]: 203–262 million), compared with 239 million cases in 2010 (95% CI: 219–285 million) and 217 million cases in 2016 (95% CI: 200–259 million).

In 2017, there were an estimated 435 000 deaths from malaria globally, compared with 451000 estimated deaths in 2016, and 607000 in 2010. Nearly 80% of global malaria deaths in 2017 were concentrated in 17 countries in the WHO African Region and India; 7 of these countries accounted for 53% of all global malaria deaths.

Effective surveillance of malaria cases and deaths is essential for identifying the areas or population groups that are most affected by malaria, and for targeting resources for maximum impact. A strong surveillance system requires high levels of access to care and case detection, and complete reporting of health information by all sectors, whether public or private.

Problem Statement

In order to reduce the burden for microscopists in resource-constrained regions and improve diagnostic accuracy, the *National Library of Medicine* [2](#) provide us with a dataset [3](#) with which I will create a model to detect malaria parasite in thin blood smear images.

Datasets and Inputs

A detailed description about how the images had been obtained and labeled can be found in the dataset page [4](#).

The dataset contains a total of 27560 images separated in two folders, Parasitized folder and Uninfected folder with 13780 images each.

The image sizes varies between 49x58 and 349x241.

Solution Statement

The solution proposed is to create a CNN image classifier to know when an image of a cell is parasitized or uninfected, applying the progressive resizing technique [5](#), consisting in a progressive, sequential resizing of all images while training, from smaller to bigger sizes.

The Accuracy is a good metric to evaluate the model due to the dataset is well balanced for both classes, having the same amount of images for each, parasitized and uninfected, but other metrics like F1-score or Specificity could be interesting to use as well.

Benchmark Model

In the research article "*Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images*" [6](#), Dr Rajaraman S and his team shown a table [7](#) with the performance metrics results obtained during their experiments, comparing their custom model against other well-known models like AlexNet, VGG-16, ResNet-50, Xception and DenseNet-121.

For this project I will use the same metrics and then compare the results against the values of their custom model.

Evaluation Metrics

In the article cited before the metrics used are:

- Accuracy
- AUC
- Sensitivity
- Specificity
- F1-score
- MCC

I will use the same evaluation metrics in order to compare if the model performs better or worse than theirs.

Project Design

Because the dataset is clean and well balanced, I can skip this step.

The data will be split in three groups, Training with the 75% of the data (20670 images), Validation with 20% (5512 images) and Test with the 5% (1378 images).

The first step will be to create a basic CNN model with the default values for the hyperparameters. The images for this model will be resized to a small value and finally the model will be trained and validate using cross-validation.

The next step will be to create a new CNN model and use transfer learning to get the weights from the first model. The images for this model will be bigger than the previous one. Finally, I will train and validate using cross-validation as before.

The main idea is to repeat these steps a few more times, increasing the size of the images in each step.

Each step will be evaluated with the metrics mentioned in the section above, so in this way I could tuning a little bit each model before moving to the next one.

At this moment is not sure how many steps I will use, nor the final architecture or the hyperparameters of the models.

After finishing the design and tuning, the model will be used against the Test data to evaluate how well or worst performs compared with the benchmark model.

References

1. <https://www.who.int/malaria/media/world-malaria-report-2018/en/>
2. <https://www.nlm.nih.gov/>
3. https://ceb.nlm.nih.gov/proj/malaria/cell_images.zip (zip file ~345 MB)
4. <https://ceb.nlm.nih.gov/repositories/malaria-datasets/>
5. <https://towardsdatascience.com/boost-your-cnn-image-classifier-performance-with-progressive-resizing-in-keras-a7d96da06e20>
6. <https://peerj.com/articles/4568/>
7. <https://peerj.com/articles/4568/#table-2>