Machine Learning Engineer Nanodegree

Capstone Proposal

Moises Vargas June 25st, 2017

Proposal

Domain Background

Cervical cancer has become one of the most importance causes of death in womans by cancer at global level [1]. In the past years artificial intelligence has played important roll on detection of cervical cancer. Department of Biomedical Engineering of University of Malaya has a review article which study 103 journal paper between 2010 and 2014 which shows works on intelligent systems to cervical cancer using techniques like artificial neural networks, support vector machines, decision trees among others [2].

Cervical cancer can take up to three decades to develop, due this time it is crucial that cervical cancer can be prevented on time allowing patients to receive the appropriated method of treatment. Beyond of having the proper technology to detect cervical cancer it is needed to have similar technology to determine which kind of treatment is appropriate for every patient. Intel in partnering with MobileODT opened a challenge in June 2017 with the aim of identify which type of cervix a patient has, given a cervix image [3].

Problem Statement

Cervical cancer stages can be detected with current artificial intelligence methods, however there are different treatment depending on patient physiology. Detecting quickly which type of cervix a patient has, would lower the risk of apply the wrong treatment to a patient and will level up the survival possibilities of womans with cervical cancer. Developing an artificial intelligence method to detect what kind of cervix a patient has, base on a cervix image would potentially help on this problem.

Datasets and Inputs

The data that will be use it this in the solution of this proposal is being obtained from a competition from kaggle [3]. Data contain images of different type of cervix; There are 1438 examples of Type 1, 4346 of Type 2 and 2426 of Type 3, in total 8210 images. Images ranges in different sizes from 480 pixels of height and 640 pixels of width to 3264 pixels of height and 2448 pixels of width, the entire data set weights 33GB on disk in zip format.

All of cervix images of this data set are considering normal without cancer, the aim of the data is to discover models that can detect the cervix type rather than whether yes or not the cervix contains cancerous cells. This data set well suite the stated problem since the aim is to identify which type a cervix is based on the image cervix.

Solution Statement

This document propose to build a convolutional neural network (CNN), with specific well known architecture LeNet 5 [4]. This kind of neural network are suitable to identify patters on images upon perform convolutions on the image, this net will extract and weight features that best represent each type of cervix. This model can be measured by splitting the initial data in training, validation and test sets and measuring the accuracy of the model on these three datasets.

Benchmark Model

This document propose use the LeNet 5 architecture [4] as benchmark model, feed a resized version of the original data 200x200 pixels, with RGB channels and measure the training, validation and test sets to obtain the base line accuracy.

Evaluation Metrics

The accuracy metric proposed to use on both solution and benchmark model will measure how well the model is performing by counting the number of correct predictions. For instance for a data set the model will predict for each example which type of cervix this example belongs, if data set contains 100 examples and only 50 were correctly predicted the accuracy if defines as follow; 50/100 = 0.5 it traduce to the model performance is 50% of accuracy.

Project Design

In general the this project will have the following work flow:

1. Data exploration

In this section, 33GB of data will be extracted out from its zip format, data will be organized in three directories for its corresponding cervix type i.e Type_1, Type_2 and Type_3.

With a imaging processing tool, the images will be inspected to see in which format the images are, and identify possible corrupted images. If corrupted images are detected will be removed from the data set.

Explore and create basic counting of height and width of the images and find a good intermediate or reasonable image resizes. The data set if big 33GB and this step aims to normalize the image size and reduce the images weight on space disk.

Plot some random chosen images for every cervix type.

Count number of example per cervix type and perform and data augmentation or data reduction to obtain a data set with same number of example per cervix type.

2. Build benchmark model

Build the LeNet-5 and adapt input layer to feed the images with RGB channels, in this section training, validation and test sets will be measured and reported to compare with the solution model.

3. Build the solution model

Building the solution model base on LeNet-5 and apply modifications to the network to increase the evaluation metrics do be better than the benchmark model. This implies experimentation of different settings specifically to the Convolutional Neural Networks, for example convolutions has filters experimenting how by adding or removing filters, adding for convolution layers changing the patch size of the convolution. In addition experimentation with dropout layers and polling layers are considered in this section to improve model performance.

Experimentation using other color spaces for the images like YUV, HSV, HLS and gray should be considered as well as normalization of the images to be zero centered.

4. Report experiments and choose the best performed experiment model

After experimentation of the section 3, the best model with best accuracy on training, validation and test sets will be chosen as finally solution model.

References

1. OMICS International, "Artificial Intelligence Based Semi-automated Screening of Cervical Cancer Using a Primary Training Database", https://www.omicsonline.org/open-access/novel-benchmark-database-of-digitized-and-calibrated-cervical-cells-forartificial-intelligence-based-screening-of-cervical-cancer-ccoa-1000105.php?aid=68453
2. The Scientific World Journal, "Intelligent Screening Systems for Cervical Cancer", http://dx.doi.org/10.1155/2014/810368
3. Kaggle, "Intel & MobileODT Cervical Cancer Screening", https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening
4. Yann Lecun, "LeNet-5, convolutional neural networks", http://yann.lecun.com/exdb/lenet/