

Machine Learning Engineer Nanodegree

Capstone Proposal

Eddy Shyu May 21, 2017

Proposal

Domain Background

I've chosen an image classification problem that identifies whether a cervix is one of 3 classes. This is important for cervical cancer screening, because the cervix type determines whether a particular cancer treatment will be effective, or whether a different procedure is required. My mom had cervical cancer before so this is an interesting topic for me.

Problem Statement

The problem is to take a cervix image and determine whether it is one of 3 types. The types are based on whether the "transformation zone" is located mostly outside of the cervix (ectocervical, type 1), partially inside and outside of the cervix (type 2), or mostly outside of the cervix (ectocervical, type 3). Cervical cancer tends to begin in the transformation zone. For type 2 and 3, the transformation zone, and hence cancer cells, may not be visible from the outside. Treatment that normally works for type 1 cervixes are not completely effective for type 2 and type 3. Type 2 and 3 cervixes will require different medical procedures in order to completely remove cancerous tissues.

The measure of success will be the final total loss on the test data (cross-entropy), which measures whether each image was correctly classified as one of the three types.

Datasets and Inputs

The data are jpg images of cervixes, and their labels as type 1, 2 or 3. These are available from the [Kaggle site](#). In addition to the training and test sets, there are also more images that may be useful for training or for validation, but may be of lower image quality or pictures of the same cervix from the original training set. I will try to use some images from the additional data sets to see how they affect performance.

Solution Statement

I will pre-process the data and feed it into a convolutional neural network. I will use transfer learning, so I will use ResNet as the base architecture, and replace the output layer with my own output layer. Then I will train the model on the training set data in order to minimize the loss. The loss is the cross-entropy loss, which is lower when the predictions for each test set matches the actual type of the cervix.

Benchmark Model

The benchmark will be the final test images and labels when released by Kaggle. My model's predictions can be compared to the correct labels of the test set using the average cross entropy loss.

Evaluation Metrics

The model can be evaluated by how well its predictions align with the actual labels of the final test data set. This is the average cross entropy, which is the actual label (one-hot encoded) times the log of the predicted probability. Kaggle refers to this as the logloss.

Project Design

Exploratory Data Analysis

- I will use EDA to better understand the distribution of the data (whether there are a similar number of samples per cervix type). I will also visually examine images to understand their differences, and to see what extraneous objects (such as medical equipment, lenses) might affect training.

Data Pre-processing

- I will process images to remove those that are too different from the rest in their respective cervix type, or are too blurry to provide meaningful information.
- I will look for ways to crop out parts of the image that are unrelated to the cervix, such as medical equipment.
- I also want to try some other traditional image processing approaches, such as edge detection, to see if this improves the final results.
- Since the transformation zone tends to be a darker red, and the rest of the cervix is a light pink, I may keep the three colors. I may look into different color representations others than rgb.
- I will also equalize the histogram of each image, to remove differences in lighting.
- I will generate additional images using rotation and translation.
- I will also normalize to range between -1 and 1, and center the data around 0.

Model Design

- I will start with the ResNet architecture and replace the output layer, to train just the output layer's weights while leaving the weights of the other layers fixed.
- I will look into training earlier layers, to see if that helps performance.