Machine Learning Engineer Nanodegree

Capstone Proposal

Liyuan Liu
April 29th, 2017

Proposal

In my capstone project, I will solve a problem posted on the Kaggle. It is a Cervical Cancer Screening problem. The goal is to find out an effective cancer treatment. Since I have just learned deep learning, I want to try the convolutional neutral network for this problem.

Domain Background

Cervical cancer is so easy to prevent if caught in its pre-cancerous stage that every woman should have access to effective, life-saving treatment no matter where they live. Today, women worldwide in low-resource settings are benefiting from programs where cancer is identified and treated in a single visit. However, due in part to lacking expertise in the field, one of the greatest challenges of these cervical cancer screen and treat programs is determining the appropriate method of treatment which can vary depending on patients' physiological differences.

Especially in rural parts of the world, many women at high risk for cervical cancer are receiving treatment that will not work for them due to the position of their cervix. This is a tragedy: health providers are able to identify high risk patients, but may not have the skills to reliably discern which treatment which will prevent cancer in these women. Even worse, applying the wrong treatment has a high cost. A treatment which works effectively for one woman may obscure future cancerous growth in another woman, greatly increasing health risks.And one thing to note is that even though health providers may not have skills to discern which treatment that will prevent cancer in these women, the workflow to make treatment decision would improve if we can

identify the type of cervix--different location of transformation zone. For more details you can refer to this document

In Schiffman et al, they pointed out that there are four major steps in cervical cancer development and the first step is infection of metaplastic epithelium at the cervical transformation zone. Thus if we can identify different location of transformation zone, it would help to decide which treatment to use further.

Problem Statement

According to the background of cervix cancer, the problem here is to classify the type of the cervix which is the same as the location of transformation zone.

There are 3 types of cervix. You can refer to this document for more details. In terms of machine learning, this is a classification problem. And the input are images of cervix.

Datasets and Inputs

We can find the datasets here. Kaggle provides us with both training set and test set. The inputs are images of cervix. The datasets also include labels which record the type of cervix. There are 3 types of cervix, Type_1, Type_2 and Type_3.

- In the training set includes 1481 images in total. There are 250 images of type_1, 781 images of type_2 and 450 images of type_3.
- I will randomly choose $1/3$ from each type of images for test set. So there will be 493 images in test set and 988 images in the training set.
- There are 3 different sizes of images, that is 2448×3264, 3096×4128 and 3264×2448. The images are colors with just one layer for each image.

Solution Statement

This is a supervised learning problem. The solutions are probabilities of each type of cervix.
The metric is log loss,

$$logloss = \sum_{i=1}^{N} \sum_{j=1}^{3} y_{i,j} log(p_{i,j})$$

where N is the sample size.

Benchmark Model

I have searched for a long time to find out the literature about the classification of

cervix types. But most of them focused on the cervical cancer diagnosing. Nothing is related to classification of cervix types. I can find the logloss which the kaggler achieved. But I don't know which model they used.

So after I preprocess the data, I will use some models such as multi-layer perceptron to gain some benchmark results.
Now we can assume a naive benchmark model where the probability of each type for each image is 1/3.

Evaluation Metrics

As I have described above, the metrics for both benchmark model and solution model would be logloss, the formula is ,
$logloss = \sum_{i=1}^{N} \sum_{j=1}^{3} y_{i,j} log(p_{i,j})$
where N is the sample size, $y_{i,j} = 1$ if the image i belongs to $type\_j$ and $p_{i,j}$ is the probability derived from the model where image i belongs to $type\_j$

Project Design

Data Preprocess

The input are images of cervix with some characteristics.

- First the image are of different size. So I will resize the images to the same size, say, 100 × 100.
- Second, the images are colors. I will attempt to convert the RGB image to grayscale image.
- Third, the images have very high resolution, say, 3096 × 4218, so we can balance the training time and space against the information loss by reducing the size. I will try to reduce the size to, for example, 100 × 100
- Forth, the labels are 1, 2, 3, I will first take a one-hot encode to convert the single numbers to vectors

The Solution Model

I know that the convolutional neural network is a great tool for image classification.

In this model, I will use convolutional layers and fully-connected layers. In tuning the model, I'll try different sizes of layers and number of layers.

Compare with benchmark model

After the layers are constructed, I'll input the images and then get the output. And then to see whether my solution model is better than the benchmark model.

Fine tune the model

From the output, I can get the feedback of the model. Then try different sizes of layers and different number of layers to improve the performance of the model.