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. Before AlexNet showed the promise of convolutional neural networks in imag classification in 2012, various computer vision techniques were used in medical imaging. Since 2015, use of neural networks in medical imaging has gained prominence, convolutional neural networks in particular, and to a lesser extent, recurrent neural networks, autoencoders, restricted boltzmann machines. Within convolutional networks, the four most well-known architectures are AlexNet, VGG-19, GoogLeNet, and ResNet. ResNet is currently the best performing model. There are also multi-stream architectures, in which the original input channels (for instance, red, green and blue) can be included into later layers, and not just the input layer. In the case of medical imaging, the different channels can be different scales of the same image (zoomed in for more detail, zoomed out for more context). Another use of multi-stream is for 3D images, where the original data is divided into 2D slices that are fed as separate input streams.

When using transfer learning (using pre-trained architectures such as ResNet), we can either use it as a feature extractor, and replace the output layer. This does not require further network training. Another option is to fine-tune the existing network, by either replacing more of the later layers, or by training the network and updating the pre-trained weights based on the current data set.

Reference: A Survey on Deep Learning in Medical Image Analysis

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 cervices 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 cervices, 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.

1481 images are in the training set.

The image sizes vary; most are 3264 by 2448 (or 2448 by 3264), but others are larger (3096 by 4128). Some have more rows than columns, others have more columns than rows. When resizing to make the shapes standard, I may first try to rotate the images so that their longest side is the number of columns. That way, the distortion from converting a rectangle to a square will at least be the same. I may resize the images to be somewhere between 133 x 100, to 655 x 500 (to keep the aspect ratio 1 x 1.33).

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

My benchmark will be the Kaggle leaderboards; I want to get in the top 300 of the 661 participants. This requires a test loss of under 1.0.

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 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 bewteen -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.