CS-594 AML
Deep Learning Project Proposal

(Erik Osland, Latika Bhurani, Rashmi Arora, Shoaib Khan)

Introduction

This project will be to diagnose the aggressiveness of a form of breast cancer called invasive ductal carcinoma (IDC). The aggressiveness of the disease is possible to determine by inspecting images of the tumor. This project's primary objective is to accurately discriminate between images that display IDC and those that do not. Several neural network architecture/solver combinations will be explored. The result of each experiment will be summaries with a confusion matrix and F1 score. The project will be later extended to check the difference optimizers' performance on different datasets.

Data

The data set[1,2] represents 162 images that were converted into 277,524 individual 50x50 patches. Each patch is labeled 1 for the presence of IDC or 0 for its absence. Roughly 39% of the examples are IDC positive, where the remainder are negative. Analyzing an individual patch is enough to determine if IDC is present in the patch. In total there are 1.53 GB of data available for analysis.

Figure 1: Images without IDC (left) and image with IDC (right).
The entire dataset consists of similar images.

Tasks

With almost 280,000 individual images at roughly 1.5 GB of data, there will be a substantial amount of preprocessing work. Null pixels will need to be imputed and color values normalized to range of 0 to 1 to improve training performance. After the preprocessing steps the data will be run through several neural network architectures and evaluated against several different solvers.

NN Techniques:

Several deep learning techniques have been shown to have promise in image recognition and medical image diagnostics. We plan to employ following techniques:

- FeedForward Neural Networks - for its simplicity
- Convolutional Neural Networks - as images are locally correlated
- Kohonen Neural Networks - for its successful use in similar medical image diagnostics[3]

Optimization Techniques:

In addition to evaluating a variety of neural network architectures we plan on exploring a variety of different solvers. The most common solver in neural networks is SGD, however, in [5] it is shown that the solver selection can play an important role in the final accuracy of the model. We plan to explore the techniques as mentioned below:

- Stochastic Gradient Descent
- Heavy Ball
- AdaGrad
- ADAM
- RMSProp

The result of this exploration would be to show testing and training accuracy with respect to the epoch for the selected data set and model as well as the time efficiency for each.

References:

1. https://www.kaggle.com/paultimothymooney/breast-histopathology-images
2. http://gleason.case.edu/webdata/jpi-dl-tutorial/IDC_regular_ps50_idx5.zip
3. https://www.ncbi.nlm.nih.gov/pubmed/9481716
4. http://spie.org/Publications/Proceedings/Paper/10.1117/12.2043872
5. http://papers.nips.cc/paper/7003-the-marginal-value-of-adaptive-gradient-methods-in-machine-learning.pdf