Subash Chebolu

Chebolu.1@osu.edu

Predicting Breast Cancer using Wisconsin Breast Cancer Database

Introduction/Background

One of the biggest areas of medical research in the world right now is cancer. This project aims to use data about certain breast cancer instances to determine whether the instance is malignant or benign. With the predictive capabilities of machine learning in the modern era fueled by the massive amount of compute through distributed systems and GPU's, it might be possible to create a highly accurate predictor of tumor types, benign or malignant, through features within data collected from prior instances of breast cancer and ground truth.

Problem Statement

The project is aimed to provide a machine based second opinion on whether a breast tumor is malignant or benign. This could help to increase the accuracy of determining whether a tumor is related to possible breast cancer which could help with earlier detection. The implementation is reliant on leveraging compute from distributed systems and powerful GPU's to generate a deep learning model that can hypothesize on features given about an instance of suspicious breast tumors.

Approach to be Used

The project will use a dataset from Kaggle, the OSC compute resources, and a mix and match of deep learning models that have already been created. The dataset is from the Wisconsin Breast Cancer Database and hosted on University of California Irvine's machine learning repository and on Kaggle. The dataset contains 30 quantitative features on different instances of breast tumors and a ground truth binary classification feature representing if the tumor was benign or malignant (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data). Furthermore, The OSC compute resources will also be used to train a distributed deep learning model on the TensorFlow/Keras platform. Lastly, the deep learning models to be use have yet to be determined although the starting point will rely on the inbuilt models that Keras and TensorFlow already have. The model will be changed based on literature and trial and error throughout the project to find the best accuracy possible.

Work Plan

With the due date of 12/12/19 in mind, the project will have about 2 months to be completed. The first 2 weeks will be used for data preprocessing/cleansing and setting up the OSC environment I require (such as installing new dependencies: TensorFlow, Keras, etc.). The next month will be used for experimenting and testing better and better accuracies and results. The last 2 weeks will be used to sum up the process in a project report with best obtained results.

Expected Outcome

The outcome is to create an accurate machine learning model that can detect whether a tumor is malignant or benign given certain features of the tumor.