Project Proposal

Motivation: What problem are you tackling? Is this an application or a theoretical result?

Cancer is one of the leading causes of death globally, and Lung Cancer is one of the most common types in humans. One way to reduce the risk of Lung Cancer is through early detection from CT scans at early stages. Currently doctors are overworked and on occasion make mistakes such as catching subtle things in a CT scan. By creating a model to detect these subtle changes we can increase the chance of cancer being caught early and help doctors avoid this mistake. The tool would be used as a supplemental aid for doctors already reviewing CT scans. This tool would be used by hospitals if it helps their doctors become more accurate when diagnosing this form of cancer.

Method: What machine learning techniques are you planning to apply or improve upon?

For this project, we plan to use a multi-layered Convolutional Neural Network (CNN) followed by a Linear Regression Classifier to build our model. We intend to adopt a similar architecture to VGG16, which comprises 3 Convolution Layers (Conv + ReLU) and 1 Fully Connected Layer to extract features from each CT scan of the lung, capturing all abnormalities. We will then feed these extracted features into the Linear Regression Classifier to generate predictions.

Intended experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

For this project, we will use the Radiology CT Images dataset from Nation Lung Screening Trial. This dataset contains three annual screenings with single-view posteroanterior from 26,722 participants, with each annual screening containing 156 CT scans of the chest. Given the substantial size of the dataset, originally 11.14 TB, we have opted to take a sample of 1000 participants from the dataset. We will be mainly focusing on the latest annual screening for each of the participants, retaining only the relevant images and dropping the rest. Also there's a csv file from the datasets that contains the clinic data, which labels the official result of each participant as either containing Lung Cancer or not. Thus, we have to assign the label to each image associated with the participant. After we have all the data ready, we will split it into training sets and testing sets and pass the training sets to the model we have predefined earlier and start training the model to correctly predict the label of each CT scan.

We can then use the testing sets to evaluate our model using the Binary Cross-Entropy Loss to measure the disparity between the predicted probability distribution and true labels. We will also use metrics such as Accuracy, Precision, Recall, and F1 Score from the Scikit-learn library to assess the performance of our model. In addition, we will utilize a Confusion Matrix to provide a more detailed breakdown of the model's performance with four values True Positive (TP), True Negative (TN), False Positive (FP), and FN (False Negative).

Reference

1. Vani Rajasekar, M.P. Vaishnnave, S. Premkumar, Velliangiri Sarveshwaran, V. Rangaraaj, et al. “Lung Cancer Disease Prediction with CT Scan and Histopathological Images Feature Analysis Using Deep Learning Techniques.” *Results in Engineering*, Elsevier, 18 Apr. 2023, www.sciencedirect.com/science/article/pii/S2590123023002384.

2. “NLST.” *The Cancer Imaging Archive (TCIA)*, 10 Jan. 2024, www.cancerimagingarchive.net/collection/nlst/.