Lung Cancer Detection

Sandy Weng

Abstract

The goal of this project was to use neural network models to classify three types of microscopic lung tissues. I worked with data provided by Kaggle.

Design

This project aims to build a CNN model to detect lung cancer. Classifying these images accurately could help doctors diagnose patients in early stages of the illness which can improve the patients’ chances of recovery.

Data

The dataset contains 15,000 images with 5000 images for each class - normal, lung adenocarcinoma, and lung squamous cell carcinoma. Normal lung tissues under the microscope look like fibroids while lung tissues that have small cell lung cancer look like tumor masses. A subset of 400 and 6000 images were used for KNN baseline model and transfer learning, respectively.

Algorithms

Models

To start off my project I used KNN on 2 classes. Then I proceeded to implement a CNN from scratch and then VGG16 with more data. For multi-class, I used VGG16 and InceptionV3 on a smaller dataset before settling on InceptionV3 because it yielded the highest accuracy score.

Model Evaluation and Selection

The entire training dataset of 15,000 was split into 80/20 train vs. test sets. All the models were evaluated on the accuracy and loss metrics. The accuracy score for validation sets were used to determine which model was performing the best. The full dataset was used in an InceptionV3 model because it had the best validation score of 0.99 on the smaller dataset.

Final InceptionV3 scores:

- Accuracy: 0.973
- Loss: 0.06

Tools

- Numpy for data manipulation
- Scikit-learn and Keras for modeling
- Plotly and Seaborn for plotting

Communication

- Slides
- Visuals