02-620 Machine Learning Project Proposal

Deepika Yeramosu | Neel Mehtani | Evan Trop
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
dyeramos@andrew.cmu.edu | nmehtani@andrew.cmu.edu | etrop@andrew.cmu.edu

1 Introduction

1.1 Problem

Metastatic cancer develops when cells spread from the primary tumor site through the blood of the lymph system to a secondary location in other organs or tissue. The problem we would like to address is to determine how we can apply machine learning image classification techniques to detect metastatic cancer in biopsy samples from patients who are suspected of having metastatic cancer.

1.2 Motivation

Current cancer diagnosis relies on a trained pathologist to detect tumor cells from patient biopsies under a microscope. This method is susceptible to inaccuracy and requires the pathologist to detect nuanced anomalies in the cell structure. An effective machine learning could be potentially more accurate and require significantly less resources in detecting metastatic cancer.

1.3 Methods

K-Nearest Neighbors: We will implement the K-Nearest neighbors algorithm first, so that we can determine how well a simplistic, non-parametric, and non-probabilistic model can correctly label an image from the PCAM test dataset. KNN is a model that is based on image similarity, so we can use the provided images and labels from the training dataset to label new test data. In this manner, the most common images will help influence the prediction of the new image based on the hyper parameter we use for k. We are considering approaches where we compare the difference in pixel values across the entire image or by comparing the differences of average pixel values based on certain subsections of the images provided to compare the likeness of a certain input.

SVM: We will implement the standard SVM classifier package from Scikit-Learn. Based on the input feature space that the images provide from their pixel values, we will attempt to learn a model which best fits the classification hyperplane that separates the PCAM image classes, either linearly or nonlinearly with the use of kernels.

R-CNN: We will implement a deep learning model architecture by utilizing a R-CNN model which will enable classification predictions for the PCAM images using neural networks. We may even compare how this form of deep learning algorithm compares to the YOLO method of image classification.

1.4 Dataset

We plan to use the PatchCamelyon dataset, which consists of 327,680 color images that are extracted from histopathologic scans of lymph nodes. From the dataset we will predict the presence of metastatic tissue in binary format (1 corresponding to the presence of metastatic tissue, 0 corresponding to the absence of metastatic tissue). The dataset comes pre-split into training, testing, and validation sets, all of which come with labels. All the data, both the images and the labels, come in HDF5 formatted files.

The dataset can be found here: https://www.kaggle.com/andrewmvd/metastatic-tissue-classification-patchcamelyon

References

[1] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling. "Rotation Equivariant CNNs for Digital Pathology". arXiv:1806.03962

[2] Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA: The Journal of the American Medical Association, 318(22), 2199–2210. doi:jama.2017.14585