Course: CSCI 1970 Individual Independent Study
Project Title: A literature review of model interpretability techniques and uncertainty quantification methods used in deep learning applications
Professor: Ritambhara Singh
Student: Alex Guo

Capstone Abstract: Recent advances in deep learning have led to an increasing interest in the applications of these new techniques to the medical field. However, deep learning models are often referred to as black boxes, and a major obstacle impeding the rapid adoption of these neural networks into the clinical setting is the trustworthiness of the models and questions of how and why a model performs the way it does. Throughout this independent study, I conducted a literature review of techniques used in the deep learning field to better understand how clinical deep learning models may be made more transparent, and eventually adopted on a large scale in a medical setting. First, I focused on interpretability techniques that are used to explain how a model came to a certain conclusion. I found common techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM), which utilizes backpropagation of a network's decision overlaid with an input image to generate a heat map of areas in a photo a convolutional neural network (CNN) model assigned high importance to in a classification. Another technique, Local Interpretable Model-Agnostic Explanations (LIME), generates local perturbations of an input image to detect the importance of various components of the image on the final decision. Finally, Testing with Concept Activation Vectors (TCAV) utilizes a second simple model to learn concepts which are then extrapolated to a vector representing a concept (CAV), which is computed against the weights of the original model in question to see how important each learned concept was in the final decision of the original model. These techniques provide ways to visualize and understand how CNN models come to the classification conclusions they do, which can be applied to improve trustworthiness and understandability in applications such as Opthalmology for automated retina scans and Radiology for automated MRI scans. Secondly, I explored uncertainty quantification methods, which are used to quantify the uncertainty within a model's final decision. Some popular techniques used here included Monte Carlo Dropout (MCDO), which involves using dropout layers computed during training on testing data to make predictions repeatedly for a given input, which allows the subsequent calculation of statistics

such as mean and variance of the final output. Another technique was the usage of deep ensembles, which requires the training of multiple versions of the same network in parallel with random initializations, and passing to each model the same input to allow for metrics such as mean and variance to be computed between the different model outputs. Techniques such as these allow for the detection of outputs a model is uncertain about and thus require further manual inspection by clinical experts, providing additional robustness against inaccuracies in any potentially automated clinical tool. In addition to my literature review, I additionally created PowerPoints to familiarize other lab members with an overview of this field of techniques, some of which can be found here:

Model Interpretability Overview

Uncertainty Quantification Overview

Interpretability/Uncertainty in Medical Applications