Recap of MIL model in Tumor Purity Prediction

Abstract

Tumor Purity, which refers to the proportion of cancer cells in a tumor issue, has been found to be critical in tumor formation and therapeutic response. Nevertheless, traditional way of estimating tumor purity is fulfilled by manually counting percentage of tumor nuclei over a region of interest in the slide and has been proved to be time-consuming and tedious. Tumor purity can also be inferred from different types of genomic data in different methods, yet these methods fail to distinguish low tumor content samples and do not provide spatial information of the location of the cancer cells. In this recap, we will talk about an MIL (Multiple Instance Learning) model in predicting tumor purity of therapeutic slides and helping to classify them as benign or malignant. After the model is summarized, we will simulate the process using a simpler collection of data sets and see if the accuracy requirements are kindly met.

Introduction

This deep multiple instance model has done the following investigations which might be found helpful in future clinic therapies. This model is designed to reduce the overall workload of clinics and avoid false negative counts in low tumor density examples. In addition, the model successfully provides spatial information of the slides and demonstrates a better prediction for slides with spatial information like from both top and bottom perspective. Subsequently, this MIL model has some knowledge after training and can make a reliable prediction over cancerous vs. normal tissues.

Discussion

I would like to put forward the major points that will be covered in this simulation process. First, we will try to represent patients' slides with 0s and 7s as indicators of malignant or benign in the data preparation process. Second, low tumor purity

samples are not readily available in the data sets applied in MIL training process, but samples of low-density zeros are yet accessible in MNIST training data sets by manually shifting digit images of 0s and 7s. This could provide us some insights about whether the model is functioning appropriately in a low-density data set. Last but obviously not the least, absolute mean errors will continue to be compared according to the accuracy requirements of the model whereas the comparison will be refined to original MIL model instead of other genetic models.

Methods

MIL Model

This model aims to predict a label for a bag consisting of **N** cropped instances from the sample slides. In order to fulfill this task, this model incorporates three modules which are feature extractor module, MIL pooling filter, and bag-level representation transformation module. In the setup process, the first step is to extract **J** features from each instance and formulate a feature vector. In the second step, an MIL pooling filter is applied and aggregates the feature vectors into a bag-level representation as the deep neural model input. The last step tries to assign a label for that input and tells us some information about the samples.

Our Model

An analogous version of the algorithm can be implemented without taking complex therapy slides into consideration. In this recap revision, we reuse the data sets from MNIST and perform regression on images with label 0(malignant) and 7(benign) and packet 100 random training images with either label 0 or 7 from MNIST into our bag. Given any bag input **X**, our objective is to predict a label of tumor purity as our output. In order to apply loss function minimization process in the MIL model predefined in the essay, a similar definition of tumor purity is followed as below: the percentage of 0s in the given bag. In other words, the likelihood of being 'cancerous' increases as the number of 0s increases in the training bag. Prior to feature extraction process, a ground-truth label is attached to the training bag by summing up all the 0s in that bag. Afterwards, the original training process is reused and repeated until convergence of

the neural network.

Results

Technology and methods in this model design may need to be more sophisticated as it does not generate expected outputs. I have been struggling with these for a couple of days so here is my apology for this.

Limitation

This model can normally be used as a replication for the original MIL model. However, it is not based on originally correlated data sets where instances of one unique bag comes from the same patient and this limitation will probably prevent the model from working well over other data sets, but it turns out not to be of too much hurt for a simplified reproduction model.

Another disadvantage lies in the spatial variation property originating from the essay and playing an important role in clinic therapies will never be discovered in my design. We believe this area requires a lot of dive-in work to reduce pathologist's workload but is still beyond the scope of this report.

Region-of-interest examination is also reserved as a future research interest to conduct mean-absolute-error analysis over high density and low-density areas. Nevertheless, the output of the model is dependent on the sum of 0 images in the input so it is highly probable this investigation will turn out to be tedious.

Lastly, I fail to build a sophisticated sufficiently model to support the training and prediction process which is under the requirements of this essay. It is now put on my schedule for my future experience in machine learning.