Report for Q1

Summary of the paper

This paper applies Deep Multiple Instance Learning to obtain the tumor purity map. Tumor purity is the proportion of cancer cells in the tumor tissue. An accurate tumor purity estimation is crucial for accurate pathologic evaluation and for sample selection to minimize normal cell contamination in high throughput genomic analysis. They developed a novel deep multiple instance learning model predicting tumor purity from H&E stained digital histopathology slides. Their model successfully predicted tumor purity from slides of fresh-frozen sections in eight different TCGA cohorts and formalin-fixed paraffin-embedded sections in a local Singapore cohort. Their model can be utilized for high throughput sample selection for genomic analysis, which will help reduce pathologists' workload and decrease inter-observer variability.

Their MIL models successfully predicted tumor purity from histopathology slides in different TCGA cohorts. MIL model has a novel 'distribution' pooling filter that produces stronger bag-level representations from patches' features than standard pooling filters like max and mean pooling. Their analysis used data from ten different TCGA cohorts and a local Singapore cohort. The histopathology slides in each cohort were randomly segregated at the patient level into training, validation, and test sets. Then, they trained their MIL model on the training set, chose the best set of model weights based on validation set performance, and evaluated the best model on the held-out test set. Their MIL models learned discriminant features for cancerous vs. normal histology while being trained on the weak labels of genomic tumor purity values. By conducting clustering over these features, they successfully obtained cancerous vs. normal segmentation maps for H&E stained slides of the TCGA LUAD cohort.

They used Spearman's rank correlation coefficient, mean-absolute errors and AUC, ROC to evaluate the effectiveness of the MIL model. Comparing with the percent tumor nuclei estimates by pathologists, their MIL models' predictions gave a higher correlation and lower mean-absolute-error with genomic tumor purity values. MIL model learned discriminant features from genomic tumor purity values (which are

sample-level weak labels) without requiring pixel-level annotations from pathologists, which are expensive and tedious. It is pretty promising to explore weak labels further for new biomarkers in cancer studies. In other words, the question of 'How strong are the weak labels?' is still a valid research question to be explored in digital histopathology.

My understanding of MIL

The difference between MIL and other common machine learning methods of image processing is that MIL takes a picture as a bag and divides the picture into different regions as instance. Such a picture is a data set, and the manual labeling step is omitted. Other machine learning algorithms take a picture as a piece of data and need to manually mark the objects on the picture.

Weak labels (for example, the overall diagnosis of a CT film, rather than the specific area with problems) are more common in medical diagnosis. In an X-ray or CT image, all areas are normal, then the bag is healthy, and as long as one area is abnormal, then the bag is unhealthy. Very consistent with the definition of MIL framework. In the task of medical image processing for packet classification, MIL classifier uses the information of co-occurrence and instance structure, which is better than general supervised learning.

MY practice on the MNIST data set

Because of my limited level, I don't have the code to implement the MIL algorithm for the time being, so I used the relatively simple CNN algorithm to process this data set:

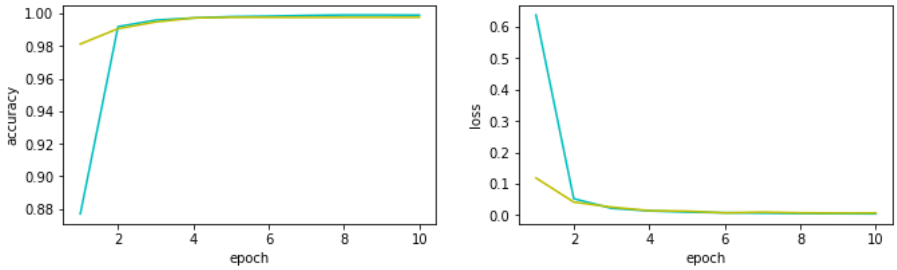
1.The experimental data set is directly imported from TensorFlow.Keras.datasets. Then, the data of 0 and 7 are separated, combined and normalized.

2.The build the model as follows: input layer, convolution layer, pool layer, convolution layer, pool layer, Flatten (filtering to full connection layer requires Flatten lamination), full connection layer, full connection layer and output layer.

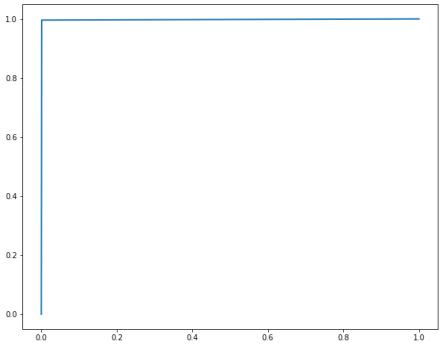
3.Training data.

4.Finally evaluating the model with accuracy and AUC value.

Draw the loss curve, acc curve of the training process:



Draw the ROC curve:



The AUC of the model is 0.9975.