Question 1

Introduction

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.

Tumor purity is estimated by two main approaches: percent tumor nuclei estimation and genomic tumor purity inference. However, both pathologists' slide reading approach and genomics methods have limitations. On the one hand, counting tumor nuclei is tedious and time-consuming. There also exists inter-observer variability between pathologists' estimates. On the other hand, while genomic tumor purity is accepted as the golden standard, it does not apply to the low tumor content samples. Besides, it does not provide spatial information of the locations of the cancer cells.

To overcome these challenges, A new deep multi-instance learning (MIL) model was developed by Oner et al [1] to predict tumor purity from H&E stained digital histopathology slides. They used data from ten different TCGA cohorts and a local Singapore cohort. In the study, the MIL model consisting of three modules: feature extractor module, MIL pooling filter, and bag-level representation transformation module. They used neural networks to implement the feature extractor module and the bag-level representation transformation module to parameterize the learning process fully. The predictions were consistent with genomic tumor purity values, and they outperformed pathologists' percent tumor nuclei estimates in the TCGA cohorts.

Regression Task

1. Datasets

Based on the background above, I implement a simpler version of the method using

the MNIST data set for regression on digit 0 and digit 7. To be more specific, I process the MNIST datasets and divide it into train and test dataset, which just keep the digit 0 and digit 7. To get a random patch(bag) containing two digits per iteration, I generate a random decimal x in the range 0 to 1, which represents the proportion of the digit 0 in a patch(bag).

Besides , in order to get a stable training processing, I uniform the pixels into the distributed between 0 to 1. Because the images are simple, I do not perform a data augmentation on the dataset.

2. Model and Training Details

According to the requirement, I use the source code given in the paper[1] and modify its output channel into one. As for optimizer, I used SGD with learning rate 0.01 firstly but found that the loss can not converge normally, even with a cosine learning rate scheduler. So I turned to ADAM. After several experiments, I find that choosing 5e-4 as the learning rate with 200 epochs results in a best performance

3. Results and Conclusions

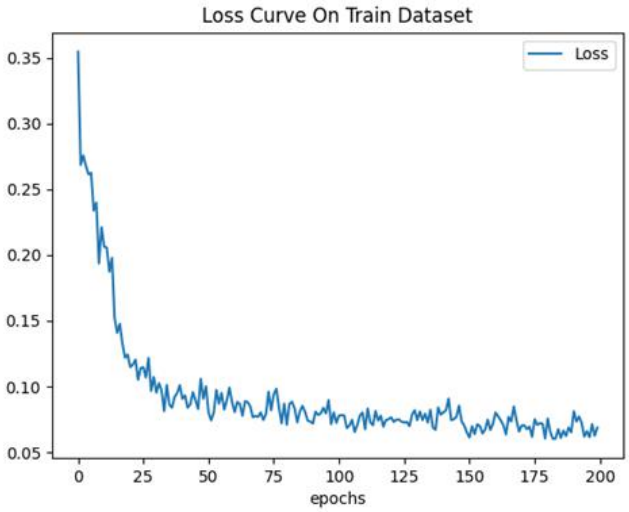

Figure 1

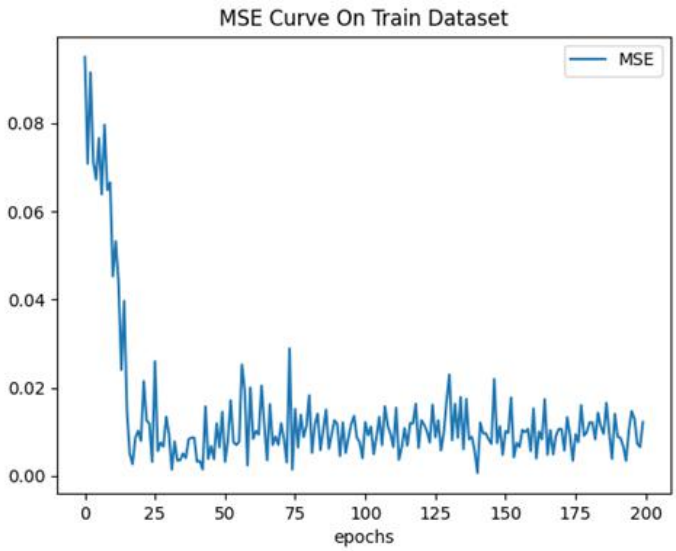

Figure 2

According to the Loss curve in Figure 1, this approach yields good results on the MNIST datasets. We can see the value of loss tends to stabilize in the final stage. Besides, I also use Mean Square Error (MSE) to evaluate this method on the test dataset, the curve in Figure 2 also proves the excellent performance of MIL.

References

[1] Oner M U, Chen J, Revkov E, et al. Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study[J]. bioRxiv,2021.