Deep Multiple Instance Learning-based Tumor Purity Prediction and Numerical Regression Task

Guo Peihong

Abstract

The proportion of cancer cells in the tumor tissue is referred to as tumor purity. In high-throughput genomic analysis, accurate tumor purity estimation is critical for accurate pathologic evaluation and sample selection to avoid normal cell contamination. A new deep multi-instance learning model was developed by Oner et al [1]. The model can predict tumor purity from H&E-stained histopathological slides such that the predictions are consistent with genomic tumor purity values. It also provides information about the spatial organization of the tumor microenvironment. It is unusual in that it represents each sample as a bag of patches cut from the top and bottom slides of the sample and uses the genomic tumor purity of the sample as the bag label, producing a stronger bag-level representation from the characteristics of the patches. Based on this idea, I tried to perform regression training on the numbers 0 and 7 using this network structure. The results show that this bag level approach yields very good results on the MNIST dataset.

Introduction

Tumor purity is estimated by two main approaches: percent tumor nuclei estimation and genomic tumor purity inference. The former offers a cellular level resolution and is extensively applicable. Counting tumor nuclei, on the other hand, is difficult and time-consuming. More crucially, pathologists' assessments are subject to inter-observer variability. Although the latter is recognized as the gold standard, it doesn't apply to samples with low tumor content. Furthermore, they do not provide spatial information about the cancer cells' positions. In other words, we lose knowledge of the tumor microenvironment's spatial arrangement. The MIL model of Oner et al. successfully predicted tumor purity from histopathological sections of different TCGA cohorts. This method allowed to find a significant difference in tumor purity between the top and bottom of the sample, indicating that tumor purity differs spatially. In addition, this study also obtained spatially resolved tumor purity maps and preferably used both slices of the sample to predict tumor purity.

This innovative approach takes a bag of patches cut from a sample's top and bottom slides as input and predicts the tumor purity of the sample as output. For each patch

inside the bag, the feature extractor module extracts a feature vector. By calculating marginal feature distributions, the MIL pooling filter, also known as 'distribution' pooling, aggregates collected features into a bag-level representation. Finally, the bag-level representation transformation module predicts tumor purity at the sample level. The ingenuity of this neural network lies in the combination of common network structures to perform a new level of tasks. The feature extractor module is a ResNet18[2] model, and the bag-level representation transformation module is a three-layer multilayer-perceptron.

Regression

1. Analysis and Methodology

I make an analogy between tumor purity prediction and numerical regression task. I parse out the MNIST datasets and make the datasets for digit 0 and digit 7: train0 and train7. The two datasets contain about 6000 images of 28*28 each, and we can analogize this training set to the top and bottom slides of a tumor sample. Next, I use the random module in python to take a random floating point number x between 0 and 1. This value is the ground truth of our regression task, which corresponds to genomic tumor purity. I follow this random number to make a bag containing 100 images. the percentage of the digit 0 in these 100 images is x. In my regression task, each image is a patch in the bag. The feature extraction for each patch in the bag is actually the feature extraction for each image in this newly composed bag. In this way, I succeed in corresponding the two tasks. I refer to the source code of [1] https://github.com/onermustafaumit/SRTPMs and train its network structure.

I set the patch_size to 100, the batch_size to 8, and the learning rate to 1e-4. But this learning rate is not the most ideal for training. So after several tests, I finally set the learning rate to 5e-4. 300 epochs were performed in total.

2. Result

Regarding the Loss function for this training, L1Loss was used. I output the Loss value for each Epoch, and some of the Loss values are shown in Figure **1**. Then the Loss curve were plotted, as shown in Figure **2**.

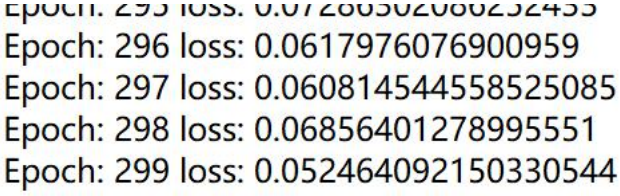

Fig1. some of the Loss values

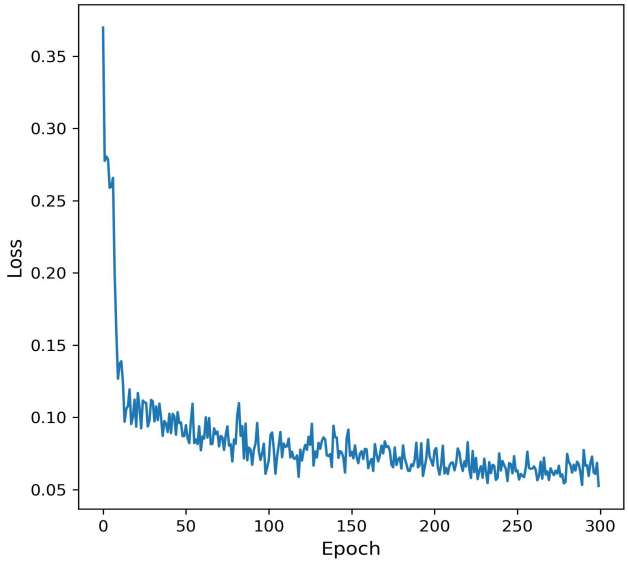

Fig2. Loss curve

We can see a good convergence effect from the Loss curve, which also proves the excellent performance of MIL.

Conclusion

Through this study, I deepened my understanding of biomedicine, especially in the area of tumor purity prediction, which also ignited my interest in exploring the world of biomedicine. The training of neural networks has enabled me to improve my ability in machine learning, especially deep learning. In this learning process, I felt the joy of problem solving.

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.

[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.