

# **Examine the Capability of Multiple Instance Learning Designed for Predicting Tumor Purity with MNIST Dataset**

OUYANG Hui

## **Introduction**

Tumor purity, by definition, is the percentage of cancer cells within the tumor. As one of the quantitative indicators of tumor content of the samples, tumor purity needs to be estimated accurately to be used in cancer research. A novel multiple instance learning (MIL) model is designed to predict tumor purity from H&E stained histopathology slides, and it has been demonstrated that this MIL model performance well by obtaining high accuracy when comparing to the genomic tumor purity values and efficiently giving the spatial information (M.U. Oner, J. Chen, E. Revkov et al., 2021). This report discusses the performance of MIL model by using the MNIST data set. For the purpose of this report, each bag consists of 100 images with only digit 0 and digit 7 in it will be labeled with  $x$ , where  $x$  is the fraction of digit 0, and the objective is to predict the label of each bag (i.e. the fraction of digit 0 in the bag).

## **MIL model to predict tumor purity**

The basic concept of MIL is to arrange data in bags, where each bag contains a set of instances, and there is one single label  $Y$  per bag. In the case of predicting tumor purity, an instance is a patch of the both top and bottom slides of a tumor sample, and each bag is a set of the patches from the same sample. And the label of each bag is the tumor purity of that sample inferred from genomic sequencing data as ground truth. Then data are split into training, validation, and test sets, and slides from the same patient should be in the same set.

The MIL model to predict tumor purity involves three stages:

- Firstly, ResNet18 model is used feature extractor module extracts a feature vector for each patch inside the bag.
- Then the MIL pooling filter, namely 'distribution' pooling, summarizes extracted features into a bag-level representation by estimating marginal feature distributions.
- Finally, at bag-level representation transformation stage, three-layer multi-layer-perceptron is applied to predict the sample-level tumor purity.

Compared with genomics methods and pathologists' readings, this model is cost-effective and have ability to provide information about the spatial organization of the tumor microenvironment (M.U. Oner, J. Chen, E. Revkov et al., 2021).

## Apply MIL model on MNIST

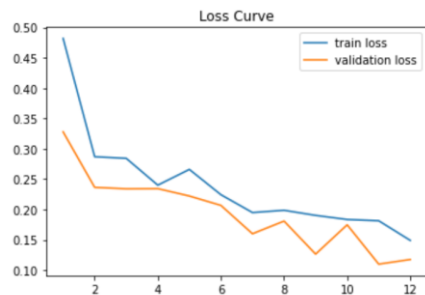
To migrate the MIL model to MNIST case in an analogy way, each bag is consisting of total of 100 images of digit 0 and digit 7 rather than patches of tumor slides. And the label of a bag is the fraction of 0s in this bag instead of tumor purity. Finally, the objective is to predict the fraction of 0s instead of predicting the tumor purity.

To prepare the data set, images of digit 0 and digit 7 are selected from MNIST data set. Each bag is implemented by adding the images of 0s with a randomly generated integer number between 0 and 100 and then adding the remaining 7s. And the number of 0s in it divided by 100 acts as the ground-truth label of each bag.

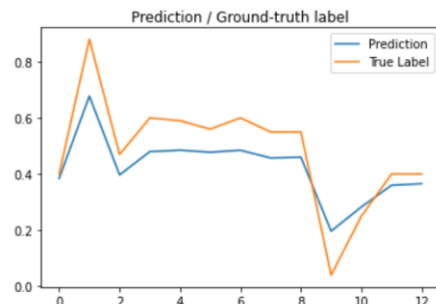
For the purpose of this report, the ratio of training, validation, and test sets are 70%, 20%, and 10% respectively. When constructing the MIL model, the source code of feature extractor module MIL pooling filter and the bag-level representation transformation module refers to [GitHub - onermustafaumit/SRTPMs: Spatially Resolved Tumor Purity Maps \(SRTPMs\)](#).

## Results

By plotting the loss curve it can be seen that train loss and validation loss both decreases with the increase of the epoch, which shows that the model was successfully get trained. However, it can be seen that more epochs should be applied. But due to the time limitation, only 12 epochs were used.



The graph below compares the prediction value with true label value, which shows that the MIL generally got good accuracy in predicting the label.



## Reviews and Conclusion

Overall the MIL model shows a good performance on the customized MNIST dataset. However, there are improvement can be down to reach higher performance. For example, as the distribution-based pooling filter designed for patches of the tumor slides shows advantages against point estimate-based counterparts, but it may not be the most suitable algorithms for MNIST dataset. Also, at the first time of model being trained, I found the loss value became nan after few epochs, and after changing the learning rate, it worked well. This suggested the impacts of learning rate to the performance of model, and more training processes should be down to find a proper learning rate when time is available.

In conclusion, this experiment shows the good capability of MIL model in dealing with different types of data. By the method of analogy, it demonstrates a way to migrate a model from the original field to another field, which allows the techniques of machine learning to produce more value.

## References

M.U. Oner, J. Chen, E. Revkov et al. Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study. bioRxiv (2021), <https://doi.org/10.1101/2021.07.08.451443>