Deep Multiple Instance Learning Based Tumour Purity Map Obtaining Model on MNIST Dataset

Abstract

An accurate tumour purity estimation is crucial for pathology evaluation. Two approaches, percent tumor nuclei estimation and genomic tumor purity inference, are mainly utilized to obtain tumour purity. However, due to the inter-observer variability and loss of spatial information of tumor cells, both methods suffer from considerable limitations. As a consequence, a machine learning model is introduced to obtain tumor purity map under the framework of Multiple Instance Learning (MIL). With patches from image random cropping, it accepts bag of instances as input to estimate tumor purity value. Tumor purity maps can be later produced for downstream pathology tasks. In this report, we demonstrate the effectiveness of the method through experiments on MNIST dataset.

Introduction

Tumour purity is the proportion of cancer cells in the tumour tissue where multiple types of cells are included such as cancer cells, normal epithelial cells, stromal cells. Tumour purity is essential for downstream tasks as it provides a key attribute of the sample. For instance, if tumour purity is over-estimated, a false-positive report may be produced. Therefore, it is of significance to accurately estimate tumour purity.

Percent tumor nuclei estimation and genomic tumor purity inference are two primal approaches to estimate tumour purity value. The former requires pathologists to read H&E stained histopathology slides for percentage estimation by counting tumour nuclei, which can be time-consuming. Furthermore, due to the individual variability, the results may be inconsistent for different pathologists. The latter normally estimates tumour purity based on genomic data whose result has now been accepted as a golden standard. However, the spatial information of tumour microenvironment, which can be a significant factor in therapeutic response, is lost. Therefore, the above two approaches are limited.

In [1], a machine learning model under Multiple Instance Learning (MIL) that predicts the tumor purity from H&E stained histopathology is proposed, which both provides consistency caused by different individuals and spatial information which can later lead to a tumour purity map. The model accepts a bag of patches randomly cropped from the ROI (Region of Interests) of sample slide, and output tumour purity value for the sample. Each patch is called an "instance". During training process, the sample-level tumour purity value is used as label, so that no pixel-level annotations from pathologists are needed which largely saves time and energy.

The model is composed of three parts: feature extractor, pooling filter, and bag-level representation transformation. Feature extractor is a ResNet [4], focusing on high-order feature extraction. The feature map will be inputted into distribution pooling, where instance-level features are embedded into distribution space, rather than simply using max or mean pooling. The final module, a Multi-Layer Perceptron (MLP), will predict tumour purity based on bag-level representation.

In application scenarios, the model will accept bag of instances cropped from ROI on a sample slide, and the corresponding region will be represent based on the predicted tumour purity value, resulting in a tumour purity map.

Methodology

We utilised MNIST dataset to reproduce a simpler version of the above model. Each bag contains 100 images with x % of '0' and (1-x) % of '7', symbolising the tissue image from ROI. The '0' here represents tumour tissue, and '7' represents normal tissue. The label is the "tumour purity", i.e., sample-level label, whereas here we set it to 0.7 indicating that 70% of instances are '0'. The code has largely referred to [1] on https://github.com/onermustafaumit/SRTPMs.

The dataset is divided into training/validation/testing set. The input shape is (Batch, NumInstance, Width, Height), and the shapes of data are indicated in Table 1.

X_train / y_train	(70, 100, 28, 28)	(70,)
X_val / y_val	(13, 100, 28, 28)	(13,)
X_test / y_test	(15, 100, 28, 28)	(15,)

Table 1. Shape of Data

For parameter settings in model construction, we selected differently since only a simpler version model is reproduced. We set num_classes=1, since there is only one "sample", and all the images are cropped from these ROIs. We chose num_instances=100, as each bag contains 100 images (instances). We randomly selected num_features=28, num_bins=10, and sigma=0.1. For training hyper-parameters, we chose lr=3e-5 and num_epochs=15 due to the limited number of data. We also chose loss function as MSE, since it's the most common loss function for regression task.

Results

The training and validation loss curve is as indicated in Figure 1:

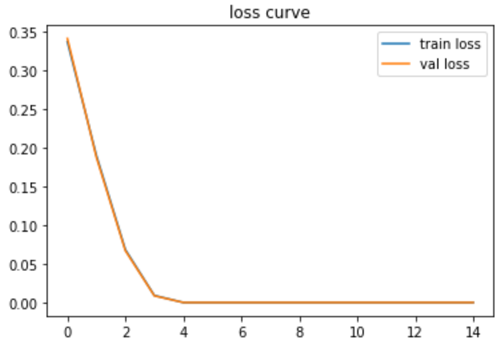


Fig 1. Loss curve

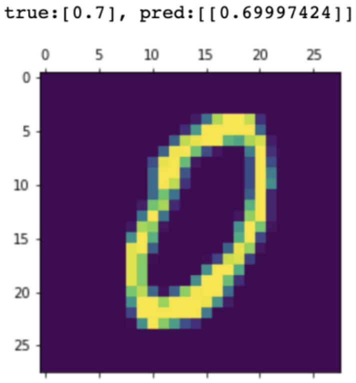
The testing result is attached below:

MAE: 0.00012845198313395182

MSE: 3.203978072482035e-08

RMSE: 0.00017899659417100749

The result is as expected, since we have only one “sample” with “tumour purity” of 0.7. In test set, we can visualize a certain instance and



```
true:[0.7], pred:[[0.69997424]]
```

Fig 2. A data point in test set

For an instance in test set, it is ‘0’ which symbolizes a patch of tumour tissue. We can see the model has predicted the bag level label as 0.7.

References

[1] Oner, Mustafa Umit, et al. "Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study." bioRxiv (2021).

[2] Oner, Mustafa Umit, et al. "Studying The Effect of MIL Pooling Filters on MIL Tasks." arXiv preprint arXiv:2006.01561 (2020).

[3] Oner, Mustafa Umit, Hwee Kuan Lee, and Wing-Kin Sung. "Weakly supervised clustering by exploiting unique class count." arXiv preprint arXiv:1906.07647 (2019).

[4] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.