Report

Common Measures of Tumor Purity

Tumor purity is the proportion of cancer cells in tumor tissue, which is essential for cancer research and clinical practice (Gong, Zhang, & Guo, 2020). There are primarily two approaches for estimating tumor purity in analyzing samples: tumor nucleus percentage estimation and genomic tumor purity inference (Oner et al., 2021). Tumor nucleus percentage is estimated by reading H&E-stained histopathology slides. This approach offers a spatial organization of the tumor microenvironment, i.e. the spatial distribution of non-cancerous cells present inside and around the tumor, which strongly influences the genomic analysis of tumor samples and may alter the biological interpretation of the results (Aran, Sirota, & Butte, 2015). Therefore, it is a good alternative. However, the method currently relies on manual observation and counting by pathologists, which is cumbersome and time-consuming and suffers from some errors. The alternative approach is to infer tumor purity from different types of genomic data, which is known as genomic tumor purity inference. Such tumor purity obtained from different types of genomic data is very accurate. However, they are not applicable to samples with low tumor content, and this method does not provide spatial organization of the tumor microenvironment. In a nutshell, both methods have distinct advantages and limitations.

Machine learning models for tumor purity prediction

In order to find a more fitting approach, based on the idea of visual intelligence, this paper combines the advantages of the above two approaches and builds a Multiple Instance Learning (MIL) model. The inputs to the model are histopathological slides, which guarantee that the spatial tissue information of the tumor microenvironment is preserved. And it also uses the genomic tumor purity as the “ground truth”, which is more accurate. The MIL model consists of three modules: a feature extraction module, a MIL pooling filter, and a packet-level representation

transformation module. ResNet18, as the feature extraction module, first partitions the packet into patches and feeds the features extracted from the patches into the MIL pooling layer. This MIL pooling operation differs from the common average pooling or maximum pooling in that it is distribution-based, i.e., after filtering by the pooling layer, the features of each individual patch are integrated into a holistic feature vector. Finally, the model performs the transformation from the integrated feature vector of the whole bag to the genomic tumor purity values through a three-layer multilayer perceptron. As a result, the machine learning model effectively predicts tumor purity while providing the spatial organization of the tumor microenvironment.

Implement a simpler version of the method

1. Data Processing

Digit 0 and digit 7 of the MNIST handwritten dataset are used as the data sources. And each bag consists of 100 images with a fraction x of digit 0 and 1-x of digit 7 as shown in Figure 1. Among these, the proportion of 0 in each bag is the label of the bag. In accordance with the above idea, 330 bag data were prepared and then 30 bag were randomly selected as test set data. In the remaining 300 data, in an 7:3 ratio, 70% as train set, 30% as validation set and then input into the MIL model for training.

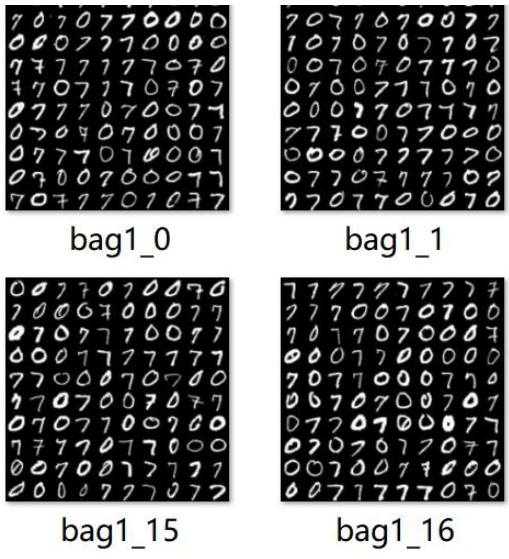

Figure 1. Images in the bag.

2. Model Training

Not finished.

References:

Aran, D., Sirota, M., & Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat Commun, 6*, 8971. doi:10.1038/ncomms9971

Gong, Z., Zhang, J., & Guo, W. (2020). Tumor purity as a prognosis and immunotherapy relevant feature in gastric cancer. *Cancer Medicine, 9*(23), 9052-9063. doi:https://doi.org/10.1002/cam4.3505

Oner, M. U., Chen, J., Revkov, E., James, A., Heng, S. Y., Kaya, A. N., . . . Lee, H. K. (2021). Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study. *bioRxiv*, 2021.2007.2008.451443. doi:10.1101/2021.07.08.451443