what is tumor purity

A tumor consists of a complex mixture of cells, such as cancer cells, normal epithelial cells, stromal cells, and infiltrating immune cells[1], and the percentage of cancer cells within the tumor is called **tumor purity**. An accurate tumor purity estimation is of great clinical importance.

tumor purity is now estimated by two main approaches

1. percent tumor nuclei estimation: a pathologist estimates tumor purity by reading H&E stained histopathology slides.

drawback: counting tumor nuclei is tedious and time-consuming, inter-observer variability between pathologists' estimates. Results are usually higher than genomic tumor purity.

2. genomic tumor purity inference: tumor purity inferred from different types of genomic data, which is referred to as genomic tumor purity, is usually used in genomics analysis. Nowadays, it is accepted as the golden standard.

drawback: do not apply to the low tumor content samples, do not provide spatial information of the locations of the cancer cells.

previous studies showed that percent tumor nuclei estimates by different pathologists are not only inconsistent but also different from genomic tumor purity values.

what and why is MIL

Supervised machine learning methods work by training a model over a labeled set of training examples and then deploying it for testing after performance evaluation[2], which require accurately labeled examples for training. Any noise or ambiguity in the labels can affect learning and, hence, the test performance of a classifier[3].

Multiple Instance Learning (MIL) is a weak supervision learning paradigm that allows modeling of machine learning problems in which labels are available only for groups of examples called bags. A positive bag may contain one or more positive examples but it is not known which examples in the bag are positive. All examples in a negative bag belong to the negative class.

In this situation, if a patient bring his slide and want to know weather there is cancer or not, we can use sample-level label as the bag label to predict his result. Because at this time, we only want to know whether it contains cancer cells, regardless of the specific situation of each cell. The latter requires doctors to separate and label each cell, which is particularly time-consuming and labor-consuming. Moreover, sample-level labels are also weak labels providing only aggregate information rather than pixel-level information. Yet, they can easily be collected from pathology reports, electronic health records, or different data modalities. Therefore, not only can MIL save our time and money, but also make labels more obtainable.

predict tumor purity from H&E stained histopathology slides

1. data pre-processing:

1) instances(patches): Top and bottom slides will be cropped from the sample to make instances. (size: 512×512 at first, 299× 299 after data augmentation)
2) bags: A bag was created by randomly sampling 200 patches (instances) from all available patches previously cropped over a sample's slides.
3) lables: The rest of a sample will be sequenced by ABSOLUTE[4] to obtain this bag's ground-truth label.

4) data sets: in each cohort, randomly segregate the data at the patient level into training, validation, and test sets.

2. feature extractor module

1) extract 128 features for each instance inside the bag.
2) use a ResNet18[5] model as the feature extractor module.

3. MIL pooling filter

1) summarize extracted features into a bag-level representation by estimating marginal feature distributions.
2) this pooling filter is based on distribution.[6]

4. representation transformation module

use a three-layer multilayer-perceptron as the bag-level representation transformation module.

5. model parameters

initialize the neural networks randomly and trained them end-to-end, optimizer: ADAM, learning rate: 0.0001, L2 regularization the weights with a weight decay: 0.0005, batch size :1, loss function:absolute error, employ early-stopping based on loss in the validation set.

6. model evaluation

1) create 100 bags for each sample in the test set and obtained tumor purity predictions from the trained model.
2) use the average of 100 predictions as the sample's tumor purity prediction during performance evaluation.

7. model results

1) obtained significant spearman's rank correlation between genomic tumor purity values and models' predictions in 8 cohorts(BRCA, GBM, LGG, LUAD, LUSC, OV, PRAD, UCEC).
2) correlation coefficients obtained from MIL predictions were significantly better than ones obtained from pathologists' estimates in all cohorts except LUSC and PRAD.
3) MIL predictions had lower mean-absolute-error and higher Spearman's correlation coefficient than pathologists' percent tumor nuclei estimates.
4) successfully predicted tumor purity from slides of ffpe sections using transfer learning with minimal training only in the first convolutional layer of the feature extractor module.
5) there is a variation in tumor purity between the top and bottom sections of a tumor sample. The degree of spatial variation in tumor purity is different for different cancer types.
6) predicting a sample's tumor purity using both the top and bottom slides together is better than using only one of them whenever possible.
7) region-of-interest selection is crucial in pathologists' percent tumor nuclei estimation, which may be the reason for their high percent tumor nuclei estimates.
8) this model learned discriminant features for cancerous vs. normal tissue histology without requiring exhaustive annotations from pathologists. And successfully classifies samples into tumor vs. normal.

8. reasons for prediction errors

1) data sets are not big enough.
2) samples with only one slide leads to higher prediction error.
3) predictions are based on morphology in H&E stained histopathology slides, while genomic tumor purity values were based on DNA data. Variations are inevitable between cellular level and molecular level.

building a similar prediction model based on MNIST

The MNIST database of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image.

I think there should be two models to complete the mission. Model I will be used to find out 7 and model II will be used to find out 0. Let's take the establishment process of model I as an example.

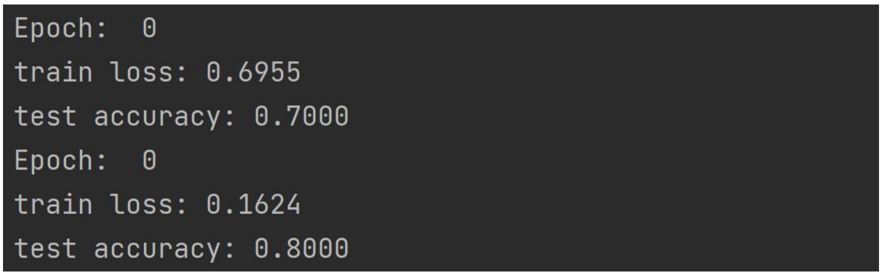
1. split the original MNIST datasets to bags

Assign the bag label as "1" if one of the instances is 7, else assign the bag label as "0". I was required to put 100 image into each bag randomly. However, I never conduct machine learning on pictures before, and I don't know how to transform pictures to matrix. Fortunately, I found open-source datasets on github (https://github.com/rosasalberto/mil). It provides MNIST bags. The dataset I used contains 4 bags and there are unequal quantities of instances in each bag. Bag will be positive if the instance '7' is contained in it.

2. model training and validation result

I never used neural network as learning method and only learned its theory in classes before. Therefore, I built a simple CNN model, with same parameters as the model in the given paper but without some complicated components like ResNet, to classify bags having 7.

It loads data and transform them to right format. Then training dataset will be sent into CNN to optimize parameters. And test validation will be conduct every 50 steps. To save time, epoch is only 0. We can see some of the results in picture1. Code is in the file named 'MIL-MNIST.py' for more information.



```
Epoch:  0
train loss: 0.6955
test accuracy: 0.7000
Epoch:  0
train loss: 0.1624
test accuracy: 0.8000
```

Figure 1: part of the results of my model

reference

[1]. Whiteside, T. The tumor microenvironment and its role in promoting tumor growth. Oncogene 27, 5904–5912 (2008).

[2]. S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, Supervised machine learning: A review of classification techniques. 2007.

[3]. D. F. Nettleton, A. Orriols-Puig, and A. Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," Artif. Intell. Rev., vol. 33, no. 4, pp. 275–306, Apr. 2010.

[4]. Carter, S. L. et al. Absolute quantifification of somatic DNA alterations in human cancer. Nature biotechnology 30, 413–421 (2012).

[5]. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition in Proceedings of the IEEE conference on computer vision and pattern recognition (2016), 770–778.

[6]. Oner, M. U., Kye-Jet, J. M. S., Lee, H. K. & Sung, W.-K. Studying The Effffect of MIL Pooling Filters on MIL Tasks. arXiv preprint arXiv:2006.01561 (2020).