Introduction:

Tumor purity describes the number of tumor cells in one certain cancer tissue. Estimation of tumor purity brings huge insight for diagnosis. Previously, the qualification methods were mainly based on molecular analyses [1] and wide genome analysis [2]. However, these methods all have severe limitations. For instance, percent tumor nuclei estimation may have huge inter-observer variability, causing biased and false-negative results [3]. Methods such as genomic tumor purity cannot offer the spatial location of the cancer cell, hindering the accurate diagnosis and therapeutic response.

This article, 'Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study' designs one novel machine learning structure, outperforming the traditional methods above. It applied deep multiple instance learning (DMIL) combined with ResNet to predict the tumor purity. Using the novel filter–distribution pooling, its model achieves outstanding performance than the normal methods such as max-pooling and average pooling. 100 bags for each sample are created before training which can decrease the false-negative rate and bring insight into the spatial location of tumor cells. This new method is tested within 8 cohorts such as BRCA, GBM, LGG, LUAD, LUSC, OV, PRAD, and UCEC, gaining lower mean absolute error than the percent tumor nuclei estimate method. Based on its result, they discover that the top and bottom slides of the tumor sample have a significant difference in tumor purity (Wilcoxon signed-rank test, $p<0.05$). Figuring out that using both top and bottom slides of the sample may be better than just taking one slide. Additionally, the author indicates that possible reasons for why pathologists may gain high percent tumor nuclei estimations are unsuitable region size and inappropriate selection.

Experiment:

This article, it generates 100 bags from the top and bottom slides of the sample at first. Using ResNet to extract feature map from each bag, then, MIL pooling filter transforms the feature maps to bag-level representation. Taking the average of 100 predictions as the result of the sample. The MIL neural network is trained with the Adam optimizer and MAE as loss function (Formula.1).

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

Formula.1: The formula of MAE

Based on the structure above, I design one simple MIL learning structure in the MNIST dataset. At first, the sample of digit 1 and digit 7 from MNIST are extracted. Taking 13007 samples as training dataset while 2163 samples as testing dataset. To mix each sample, I used NumPy to concatenate samples by the first axis. Each bag will be a long figure, containing 100 samples. Each epoch will divide 20% samples as validation datasets. If the MAE in the validation dataset doesn't increase within 30 steps, the training will be stopped. This model has the same hyperparameters and loss function of the article with an early stopping callback. Using CNN as the base model to make a prediction of the proportion of digit 1 in the bag. The CNN contains five convolution layers and three max-pooling layers. DNN with tanh as activating function is taken as the output layer.

Results:

The result indicates that CNN model converges quickly. After almost 5 epochs, loss function keeps consistent (figure.1). Model achieves MAE=0.0278 in training dataset and MAE=0.0367 in testing dataset. Applying this model on the test dataset, model achieves MAE=0.0242.

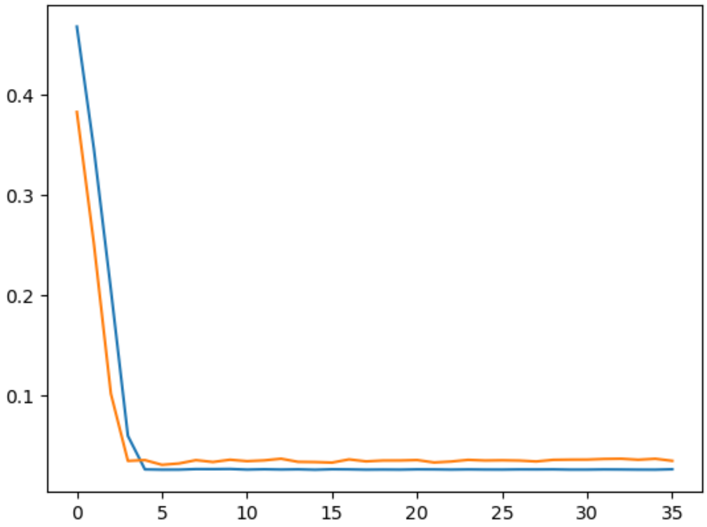

Figure. 1: The history of training

Reference

[1] Benelli, Matteo; Romagnoli, Dario; Demichelis, Francesca (2018). Tumor purity quantification by clonal DNA methylation signatures. Bioinformatics.

[2] Luo, Zhihui; Fan, Xinping; Su, Yao; Huang, Yu S (2018). Accurity: Accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants. Bioinformatics.

[3] Smits, A. J. et al (2014). The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. Modern Pathology 27, 168–174.

[4] Junttila, M. R. & de Sauvage, F. J (2013). Influence of tumour micro-environment hetero-geneity on therapeutic response. Nature 501, 346.