Report for Question 1

Introduction

Cancer

Cancer is a group of diseases, which cells of the body begin to grow out of control. This can happen when gene mutations occur, which result in errors either in division or apoptosis processes. The abnormal growths of cells would form tumours. Tumours may develop in any parts of the body, while some regions can have higher possibility than others. The most common type of cancer is Breast Cancer, which usually happen in women. In 2020, 2.3 million females were diagnosed with breast cancer, with 685,000 deaths globally [1]. Other common cancers are prostate cancer, gastric cancer, colon cancer, and etc. Cancers are often treated with surgery, radiation therapy, and/or chemotherapy. In general, earlier starts the treatment, higher possibility of cured, since in advanced stage, cancer cells can transfer to other parts of the body, which may make the situation much more complicated. In this case, treatment options for advanced stage cancers can be very limited, leading to a relatively low 5-year survival rate, for example, it can be less than 20% for gastric cancers [2].

Causes of cancers are very diverse, such as radiation, chemicals, viruses, and other factors, which ultimately leads to tumour-associated mutations, however, in many cases, the causes are not quite clear. In this case, the high heterogeneity of cancers requires the diagnosis system to be highly precise. However, the currently available histopathological systems do not have satisfying results, leading to inefficient treatment plans [3].

Genomic analysis has great benefits on study cancer type and particular patients. It involves the manipulation of large datasets and the application of complex methods. Tumour purity refers to the percentage of cancerous cell within the tumour, which plays an important role in affecting the quality of genomic analysis [4]. Tumour tissue usually have complex cellular components, including both cancerous and noncancerous tissues. Among multiple types of noncancerous cells, stromal and immune cells have been found to be essential for carcinogenesis, malignancy progression and treatment resistance [5]–[7]. However, there is limited knowledge about cell characteristics under purity levels.

Machine learning

Machine learning is a field of computer science that applies algorithms to datasets so that can be used for pattern recognition, forecasting, classification, and prediction. Machine learning techniques can be used in situations where there are large amounts of data and multiple variables, but no clear pattern can be observed and no clue for predicting the outcome, especially, where human bias can largely affect the result. For example, in fraud detection or credit scoring system, machine learning can be used for quick and accurate analysis of numerous data without human intervention.

Common machine learning techniques are neural networks, decision trees, regression analysis, clustering analysis, Bayesian networks, and etc., which are very useful indicator for future behaviour based on complex variables.

Machine learning methods have become increasingly popular in the research community and have emerged as a promising data analysis alternative. They provide potential for large improvements in fast and high-resolution analysis. In the case of tumour purity prediction tasks, pathological labelled ground truth can be used to measure the accuracy of predicted output from the network. The difference between the ground truth and prediction is called the “Loss”. By minimizing the loss through repeatedly updating parameters, features can be extracted and learned, and prediction is expected to be increasingly accurate. This process is known as “Training”. When the loss is within defined tolerance, the optimal parameters are thus found and “Training” is done, so that when new raw input to the network, it can quickly predict the higher resolution purity.

Method

The model

In this report, a novel multiple instance learning (MIL) model [4], [8] was used for regression on digit 0 and digit 7 from the MNIST dataset [9], which has 12188 images for training, and 2008 images for validation, with image size of 28 × 28. The input for this model was separated into multiple batches, where each batch consists of 100 images with a random fraction of digit 0 and rest of digit 7. In this case, the train set had 121 batches, while the test set had 20 batches (dropped out the last 88 cases in train set and 8 cases in validation set).

The model consists of three main blocks. First, a ResNet18 [10] model was used as the feature extractor module is used for extracts a feature vector for each image inside the batch. Next, a distribution pooling filter is used as the MIL pooling filter, which estimated marginal feature distributions so that summarized extracted features into a batch-level representation. In this experiment, the standard deviation was set as 0.05 and the estimated marginal distribution were uniformly binned into 21 bins. Finally, the batch-level representation transformation module with a three-layer multi-layer perceptron can predict the digit image to be 0 or 7.

Training

Initially, the neural network model was set to be trained using ADAM optimizer with a learning rate of 0.00001 and L2 regularization on the weights with a weight decay of 0.0005. During the training process, a bag is directly represented by a batch, so that for 100 instances in a bag, the batch size was set as 100. For each instance inside a bag, 28 features are extracted. The loss criterion was L1 loss which minimize sum of all absolute error. Then, an unseen validation set is used for evaluation.

Result

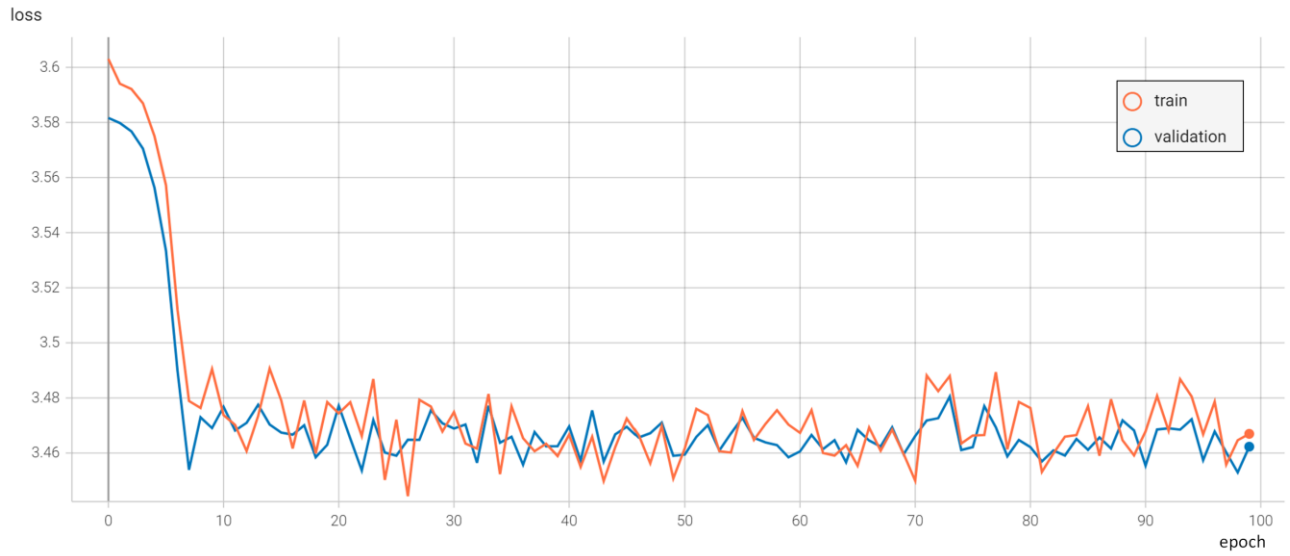

Figure 1. Loss (per bag) changes across 100 epochs for train set (orange) and validation set (blue)

According to the figure above, both train and validation loss were decreased dramatically in first 10 epochs, showing the network was working, while they became more stable, which have minor fluctuation around 3.47, in later epochs which might indicate they had converged.

Reference

[1] World Health Organization, "Breast cancer," Mar. 26, 2021. https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed Jan. 30, 2022).

[2] H. Katai *et al.*, "Five-year survival analysis of surgically resected gastric cancer cases in Japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese Gastric Cancer Association (2001–2007)," *Gastric Cancer*, vol. 21, no. 1, pp. 144–154, Jan. 2018, doi: 10.1007/S10120-017-0716-7/TABLES/6.

[3] S. Lou *et al.*, "Comprehensive Characterization of Tumor Purity and Its Clinical Implications in Gastric Cancer," *Frontiers in Cell and Developmental Biology*, vol. 0, p. 3843, Jan. 2022, doi: 10.3389/FCELL.2021.782529.

[4] M. Umit Oner *et al.*, "Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study," *bioRxiv*, p. 2021.07.08.451443, Jul. 2021, doi: 10.1101/2021.07.08.451443.

[5] M. R. Junttila and F. J. de Sauvage, "Influence of tumour micro-environment heterogeneity on therapeutic response," *Nature 2013 501:7467*, vol. 501, no. 7467, pp. 346–354, Sep. 2013, doi: 10.1038/nature12626.

[6] D. Zeng *et al.*, "Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures," *Cancer Immunology Research*, vol. 7, no. 5, pp. 737–750, May 2019, doi: 10.1158/2326-6066.CIR-18-0436.

[7] B. Zhang, Q. Wu, B. Li, D. Wang, L. Wang, and Y. L. Zhou, "M6A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer," *Molecular Cancer*, vol. 19, no. 1, pp. 1–21, Mar. 2020, doi: 10.1186/S12943-020-01170-0/FIGURES/6.

[8] M. U. Oner, "onermustafaumit/SRTPMs: Spatially Resolved Tumor Purity Maps (SRTPMs) [Source Code]." Dec. 16, 2021. doi: 10.5281/zendo.5606981.

[9] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012, doi: 10.1109/MSP.2012.2211477.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.