MULTIPLE INSTANCE LEARNING IN MNIST DATA

ZHOU Yihang

Macau University of Science and Technology

An applicant to the program of Msc of Biomedical Data Science at Nanyang Technological University

Application number: C2213722

1. INTRODUCTION

There are some jobs that require extensive experience and a lot of time in medical research as well as in day-to-day treatment. However, due to the shortage of medical resources, many articles based on machine learning have begun to appear in recent years. For example, Hemdan et al. [1] proposed COVIDX-Net to identify whether you have COVID-19, to relieve the pressure of radiologists to read X-rays. Similarly, there are applications such as reading and 3D reconstruction of cryo-lenses[2]. In this project, the work is based on the work that presented by Oner et al.[3] which want to predict the tumor purity by applicating the multiple instance learning (MIL) to the digital histopathology slides. The multiple instance learning is selected because counting the tumor nuclei is tedious and time-consuming for the expert of phathologist. Under this reason, the labels are not enough for supervised learning and the MIL is a good choice.

However, this project is to simulating the process that predicting the tumor purity by replace the tumor slides dataset with MNIST[4]. The MNIST dataset is a well-known dataset which usually introduced to the beginners.

And in this project, we first mix the images of 0 and 7 with specific ratio. And do the regression on the ratio by applying the multiple instance learning. Images of 0 and 7 represent the normal slides and tumor slides. We can simply simulate the paper presented by [3] in this ways.

As a result, we can accurate predict the ratio and the maximum error is 0.038, mean error is 0.023.

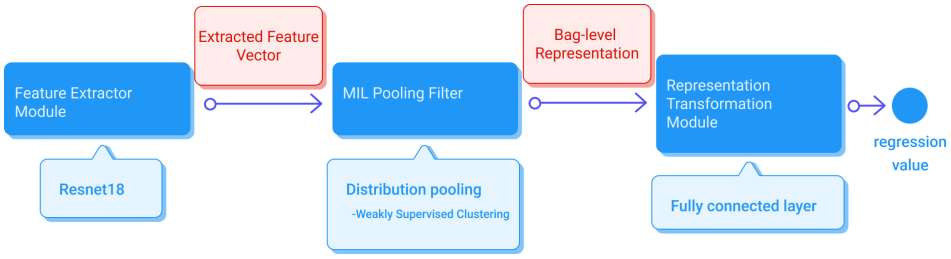

Fig. 1. Pipeline of the model

2. METHODS

In this section ,the overall pipeline of the model is described first. And the necessary traning hyperparameters is proved then. To evaluate the model, we test the performance of the tranined model. The regression result is close to the ground truth and the error is no larger than 0.5 which means the model is stable and accuate.

2.1. Data Processing

Understanding data is the key to bioinformatics. Whatever the models or algorithm is, it cannot gain a good performance without a suitable data pre-processing. In this project, we need to mix the images of zero and seven and then randomly choose a ratio. So we filtered the corresponding images and calculating the number of each number with specific random selected ratio. Then generate the random index for each class. By repeating many times, the train dataset and test dataset can be obtained.

2.2. Pipeline of model

The structure of the model is highly similar to the one described in the paper[3]. But a convolution layer is inserted into the begin of te model because of the channel of the MNIST images is one but the image of tumor slides is different. Figure 1 illustrates the pipeline of the model. The feature extractor is the well-known ResNet which present in 2016. This module extract the feature of each patch so that patches of every bag can be described as a extracted feature vector. The MIL pooling filter is a weakly supervised clustering method whcih present by Oner et al.[5] in 2019 which is distinguish form the traditional mean pooling or average pooling and well-supported in mathematical. And the transformation layer is to calcualting the final regression result. It is composed of several fully-connected layers and the shape of the layer's output is an float.

2.3. Hyperparameters

Some of the hyperparameter configurations made the loss drop slowly so that Adam optimizer is adopted finally. Table 1 illustrates the detailed hyperparameter of the model.

Learning rate	Train Dataset Size	Test Dataset Size	Batch Size	Optimizer	Epoch
3e-4	10000	1000	10	Adam	40

Table 1. Detailed Information of hyper-parameters

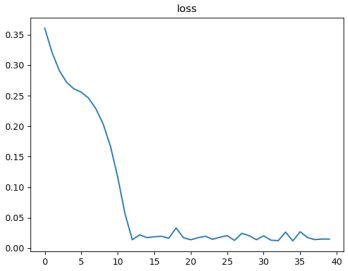

Fig. 2. Loss value during training

3. CONCLUSION AND RESULT

In this project, the MIL performed well in this project but there still are some space to improve. The train dataset is not big enough to fit the model well so that the loss didn't drop in the later epoch as the Figure 2 shows. This project illustrated a simple version of the paper which devoted to develop a model that can recognize the tumor slides. The internal mechanism can be showed in this simple example.

4. REFERENCES

[1] Ezz El-Din Hemdan, Marwa A Shouman, and Mohamed Esmail Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv preprint arXiv:2003.11055*, 2020.

[2] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger, "Exploring generative atomic models in cryo-em reconstruction," *arXiv preprint arXiv:2107.01331*, 2021.

[3] Mustafa Umit Oner, Jianbin Chen, Egor Revkov, Anne James, Seow Ye Heng, Arife Neslihan Kaya, Jacob Josiah Santiago Alvarez, Angela Takano, Xin Min Cheng, Tony Kiat Hon Lim, Daniel Shao Weng Tan, Weiwei Zhai, Anders Jacobsen Skanderup, Wing-Kin Sung, and Hwee Kuan Lee, "Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study," *bioRxiv*, 2021.

[4] Li Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

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