Training Strategies for Solving Data imbalance in Nuclear Classification

1st Ye Zhang
Shenzhen Graduate School
Harbin Institute of Technology University
Shenzhen, China
zhangye_zoe@163.com

2nd Zijie Fang
International Graduate School
Tsinghua University
Shenzhen, China
vison307@gmail.com

3rd Zhifan Lin
Shenzhen Graduate School
Harbin Institute of Technology University
Shenzhen, China
200111528@stu.hit.edu.cn

4th YiFeng Wang
Shenzhen Graduate School
Harbin Institute of Technology University
Shenzhen, China
wangyifeng@stu.hit.edu.cn

5th Yongbing Zhang*
Shenzhen Graduate School
Harbin Institute of Technology University
Shenzhen, China
ybzhang08@hit.edu.cn

***Abstract*—Nuclei segmentation and classification are prerequisite procedure exploring immune micro environment (IME), which plays an important role in tumor diagnosis and treatment. However, in practice, we may encounter imbalance problem. Aiming at distribution imbalance of dataset and category imbalance of nuclei, we develop data synthesis and augmentation methods for segmentation and classification task and these methods have been proven to improve model performance in test set.**

***Index Terms*—nuclei segmentation, nuclei classification**

I. Introduction

Nuclei segmentation and classification are prerequisite procedure exploring immune micro environment (IME), which plays an important role in tumor diagnosis and treatment. Hence, accurate nuclei segmentation and classification are of great significance for patient diagnosis and prognosis.

The colon nuclei identification and counting (CoNIC) challenge aims at developing algorithms that perform segmentation, classification and counting of 6 different types of nuclei within the current largest known publicly available nuclei-level dataset in CPath, containing around half a million labelled nuclei.

Nowadays, segmentation networks based on U-Net are emerging one after another, such as, Hover-Net [1], Triple U-Net [2], U-Net-FCN [3], Micro-Net [4] and so on.

However, we usually develop algorithms using identified distribution training, validation and test sets. Meantime, each dataset is used to validate model respectively. This is totally different from the practical scenario. For that we summarize the main challenges as several points:

- The segmentation result affects the classification performance especially the boundary district;
- The number of nucleus of different types are imbalanced. When training and test data comes from different population distribution, which will cause severe deviation;
- The training dataset have inconsistent staining distributions, which increases the training difficulty.

Aiming at last two points, we use new data synthesis and augmentation methods, which can solve imbalanced type problem (P2). Meantime, we add weight for type loss (TP branch of HoverNet), which can overcome imbalanced data distributions (P3).

*Corresponding author: ybzhang08@hit.edu.cn

II. Data and Methods

A. Data Description

The match data comes from the Lizard [5], which is collected from six sources: Glas, CRAG, CoNSeP, DigestPath, PanNuke and TCGA. The dataset contain 6 different types nuclei: epithelial, lymphocyte, plasma, eosinophil, neutrophil or connective tissue and the dataset can be downloaded from https://warwick.ac.uk/fac/cross_fac/tia/data/lizard/.

In the stage of algorithm development, we can only use first five datasets to split training and validation set. In test stage, the TCGA is used to testing.

B. Data Synthesis

In given dataset, epithelial nuclei is more numerous, while eosinophil and neutrophil nuclei are less numerous. There is no doubt, this will cause classification boundary to shift. In order to alleviate the category imbalance, we use real images to synthetic new images. In detail, we screen eosinophil and neutrophil from the original images and paste them on the real background. In the process, we add shape augmentation into screened images. The instance of synthetic images is shown as follows:

Using the method, we synthesize some images containing fewer types of nuclei, which are used as training inputs along with the original images.

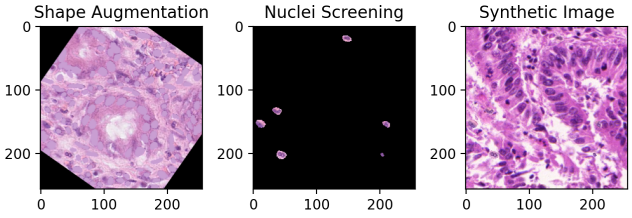

Fig. 1. An instance for data synthesis.

C. Data Augmentation

Data augmentation have achieved outstanding performance whatever feature augmentation or shape augmentation. However, these methods are mostly used in natural image tasks. For pathological images, how to select appropriate augmentation methods to adapt the differences in staining is particularly important. For examples, the images from CRAG present pale pink and the images from DigestPath present deep purple.

We convert the original images color domain to a wider color domain, which can reduce the sensitivity of trained net for color. And the gamut distribution before and after the conversion can be shown as follows:

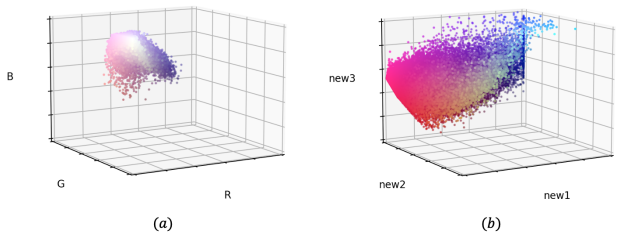

Fig. 2. Color gamut: (a) before conversion; (b) after conversion.

D. Methods

Besides above data preprocessing methods, we keep baseline model HoverNet as our training net. The HoverNet consists of three independent decoder branches: NP, HV and TP and these three branches can fulfill segmentation and classification tasks.

E. Weighted Loss

We use weighted category loss to calculate cross entropy loss (CE) and dice loss (DICE). These two naive losses can be formulated as:

$$CE = -\frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{K} X_{i,k}(I) log Y_{i,k}(I) \tag{1}$$

$$DICE = 1 - \frac{2 \times \sum_{i=1}^{N}(Y_i(I) \times X_i(I)) + \epsilon}{\sum_{i=1}^{N}(Y_i(I) + X_i(I)) + \epsilon} \tag{2}$$

Following above, the weighted loss functions are defined as:

$$WCE = -\frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{K} \omega_{i,k} X_{i,k}(I) log Y_{i,k}(I) \tag{3}$$

$$WDICE = 1 - \omega_i \times \frac{\sum_{i=1}^{N}(Y_i(I) \times X_i(I)) + \epsilon}{\sum_{i=1}^{N} Y_i(I) + \sum_{i=1}^{N} X_i(I) + \epsilon} \tag{4}$$

III. Results

In this part, we list top-3, baseline and our model scores. From the table we can see that although our model is better than baseline, there are still a big gap compared with other methods. We're also looking forward to learning from others in future.

Model	mPQ+	PQ	PQ+
Top-1	0.459	0.661	0.658
Top-3	0.457	0.630	0.651
Top-3	0.452	0.659	0.653
ours	0.362	0.577	0.572
baseline	0.296	0.550	0.544

IV. Conclusions

At last, we have some conclusions about the competition. First, we are very happy to have the opportunity to participate in the competition and finish the competition seriously. Second, in the process of the algorithm development, we try to adjust model but the results are depressing. Finally, we hope to have the opportunity to learn from the experience of other participates in this competition.

References

[1] Graham, Simon, et al. "CoNIC: Colon Nuclei Identification and Counting Challenge 2022." arXiv preprint arXiv:2111.14485 (2021).
[2] Zhao, B., et al. "Triple U-net: Hematoxylin-aware Nuclei Segmentation with Progressive Dense Feature Aggregation." Medical Image Analysis 65(2020):101786.
[3] P. Naylor, M. Laé, F. Reyal and T. Walter, "Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map," in IEEE Transactions on Medical Imaging, vol. 38, no. 2, pp. 448-459, Feb. 2019, doi: 10.1109/TMI.2018.2865709.
[4] Raza, Shan E. Ahmed, et al. "Micro-Net: A unified model for segmentation of various objects in microscopy images." Medical image analysis 52 (2019): 160-173.
[5] Graham, Simon, et al. "Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.