Exam Report

Yongchao Qiao

In this 7-day competition, I got a good evaluation score on the test set, but I think the best model I got is still not enough to meet the demands for industrial purpose since the classification score is not that high. However, now I will introduce everything I did to get the existing best model for me and other possible directions of improvement if given more time. Generally, there are three main directions I followed to get this best model, they are data augmentation, ensemble models and matched evaluation methods. Then I will explain them in detail but follow an order in processing the competition.

I. Fundamental preparation

This part is about initial processing for the original data and predict file that will be submitted. Actually, it contains my work on day1 and day2. First, I uploaded the train.zip to the cloud and read images as well as category names. Then I saved images and targets into different tuples group by the file suffix, like png or txt. Next, I still write a loop to check whether the images and targets are matched correctly. Since these pictures are not in one same size, I resized them in one same size as (32, 32, 3), fortunately this size was verified as the most suitable one for my model which helped me train the model quickly and precisely. Then I reshaped all the image arrays with size (n_images, 3072) to get the x_total and replace the name strings with corresponding category numbers, like “red blood cell” with “0”, to get the target_total set. Then I get train set and validation set from these two total sets. Next, I built the MLP network, trained the model and check my predict file. In the process of tuning the parameter, each time I adjusted the value of only one parameter to detect the best value of this parameter.

The second part is the work from day2. Since the dataset is very imbalanced and not that big, I used oversampling to balance and increase observations of the data. For this image dataset, I chose six methods like rotation, horizontal flip, vertical flip, shear, feature-wise_center and feature-wise_std_normalization to do the augmentation since these operations were reasonable for the cell images. Also, there are two options to augment the data: a) Augment the whole dataset which means use all the information of the original dataset; b) Augment only the training set of day1, which means only use the information of the training set of day1. The reason why I used the training set of day1 is that the evaluation score on validation data of day1 is only 0.02 higher than that of the blackboard, and I thought that these two test sets might have a similar distribution. Therefore, the score on day1’s validation set can be an indicator for me to evaluate the model.

II. Feedback and analysis

This is the main part that I used to find my best model. During the competition, I focused on my evaluation score for each submission on the blackboard, since it could tell me the direction of improvement regarding my future models. On day1, I split the whole original dataset into training and validation set to train the model while on day4, I used whole dataset as the training set and the validation set of day1 still as the validation set. On day3 I split the augmented dataset with whole information of the original set into training set and validation set to train the model while on day5 I used the augmented dataset with only information of day1’s training set as the training set and the validation of day1 as the validation set. As the table 1 shows below, in each column, the evaluation score of using whole information is always higher than that of only training set information. In each row, the evaluation score of using augmented dataset is always higher than that of original dataset. Then one highest score can be found which is from the augmented dataset

using whole information of the original dataset. After this analysis, I thought I could get a better model only from the augmented dataset using whole information of the original dataset.

Table 1: Comparison of four days' submissions

	Original dataset	Augmented dataset
Only training set information	0.771066 (Submission of Day1)	0.791168 (Submission of Day5)
Whole information	0.772244 (Submission of Day4)	0.793945 (Submission of Day3)

III. Ensemble models

The ensemble model is an important tool for me to improve my evaluation score on the blackboard. Actually, I used this tool in the submission of day5 and I got an ensemble model from six single models. However, the score is still lower than that of day3. So it indicates that the whole information of the original dataset is really important. Then I used this tool in the submission of day6 and I got my best evaluation score in this competition. That is, on day6, I used the augmented dataset with whole information of the original dataset to train the model and combine one better model I got with the best model on day3 to get the ensemble model. Finally, I got my best model in this competition with the evaluation score as 0.804228.

IV. Evaluation methods

After reading the introduction of this competition, I found that the final evaluation method was the mean of macro-averaged F1-score and the Cohen's Kappa score. So I used the same evaluation method to test my model after each training. This really helped me narrow the distance between the score I got on my own validation set and that of the blackboard.

V. Other ideas to make possible improvement

The first idea is that since the evaluation method is the mean of macro-averaged F1-score and the Cohen's Kappa score, I tried to define a metrics function and add it in the model.compile step. However, I failed since it ran successfully in the training procedure but not in the prediction procedure. Therefore, if this problem can be resolved, the model can still be improved. Then the second one is that if I can find a method to narrow the difference in one same category and boost the difference between different categories, the model can also be improved, since I have scanned cell images in different categories and found that there is much difference even though in one same category as well as little difference between different categories.