NSYSU Practical and Innovative Analytics in Data Science - Spring 2023

Final Project

Goal

This course is designed to equip you with the necessary skills to undertake a practical, real-world project that encompasses the entire data science pipeline, including data preparation, organization, transformation, and result evaluation. The final project is intended to start you in these directions. This is an opportunity for you to jump into a dataset, understand its structure and then apply your data-cleaning skills. Specifically, you are expected to use data-centric AI techniques on a topic related to a classification dataset.

Dataset

The images are collected and modified from the Roman Number dataset. It is noted that we have manipulated both MNIST and Fashion MNIST datasets in the laboratory and homework. We will investigate whether machine learning or neural networks can help us recognize Roman number characters.

The original dataset was grouped into two folders – train and val, i.e., training and validation. Each folder has ten sub-folders – i to x. Each subfolder contained images of handwritten Roman numerals from 1 to 10. The training and validation dataset used in this final project is a variant of the above-mentioned dataset, which contains more than 4,300 characters from 10 different classes. **However, it is noted that the model (a modified ResNet50, which is a convolutional neural network) should be kept fixed**, and you were asked to modify the image data provided in any way they saw fit, subject to a maximum of 10,000 images. Your job is trying to perform data cleaning and figure out the best training and validation split.

The dataset is held on the Kaggle platform, which can be accessed from here. There is a *mapping.csv* that describes the mapping between ten classes and their corresponding label. In addition, the training set is organized into different directories and is fully labeled. Finally, we also provide a testing set containing 400 characters (about 10% of the final dataset) representing the real-world scenario for you to test.

Rules

You are allowed to use any external datasets, but the maximum of images submitted should be at most 10,000 images. You are also allowed to change the training and validation split. However, the model should be fixed in this project. In addition, do not upload your data to be publicly available during the final project.

The submission will be evaluated using the macro F1 score. The F1 is calculated as follows:

$$F_1 = 2 * \frac{precsion * recall}{precision + recall}$$

Where

$$\text{precsion} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

In "macro" F1 a separate F1 score is calculated for each class/label and then averaged.

Your prediction label should be encoded into numerical values according to the *mapping.csv*. You can test your results on the public leaderboard, which is also available on the Kaggle platform. However, your submission is restricted to less or equal to five times a day and it is a good practice that does not tune your results on the test set. The file should contain a header and have the following format:

Id, Expected
0000.jpg, 1
….

Id corresponds to the jpg filenames in test data and Expected corresponds to your prediction label. Follow the *sample_submission.csv* if you have trouble with formatting. It is noted that the macro F1 score shown on your Kaggle leaderboard is only about 10% of the test data. The final score will be shown after the competition ends, which will be the macro F1 score of the whole test data.

Grading Policy and Deliverables

- Deadline (6/04 11:59 pm)

The final score will be the summation of the **accuracy of your model on the whole test data (40%), the final presentation and the final report**. The former score will be calculated as follows:

$$\begin{cases} 0.285 * F1\ score & if\ macro\ F1\ score < 70\% \\ \min(40, 20 + (F1\ score - 70) * 1.33) & if\ macro\ F1\ score \geq 70\% \end{cases}$$

You are requested to hand in the code to reproduce your result. **Please include a link to your code to produce the dataset and your dataset, or upload the zip file via cyber university** for your final project. We will use your code and dataset to perform inference on the whole test set.

- Final Presentation (Scheduled at 6/5 (9:10~12:00)) (30%)

Each team is required to present their work to the class. The presentation time is 10 minutes, with an additional 3 minutes for Q & A. The grading score will be based on the clarity of the presentation, the relevance of the project to topics taught in this course, and the novelty of the work. Each team will also be given a grading list containing six characters from A+ to D (which will be translated into 6.5%~10% in the final grading). You need to provide a letter grade for other teams. The final score will be the summation of the grade from students (10%), TAs (10%) me (10%).

- Final Report (due 6/11 11:59 pm) (30%)

After the final, we will also post all the final writeups online so that you can read about each other's work. If you do not want your report to be posted online, please tell us a week before the submission deadline. Your report may contain the following sections:

- Abstract
- Introduction
- Dataset
- Methods
- Experiments and results
- Discussion
- Conclusion and future work
- Contributions
- Reference

Note that your results may not be positive, but you can still report what you have tried so far and have some discussion. The final project report can be at most **10 pages** long (including appendices and figures). The paper size is **standard A4** or 8.5 x 11 inches and the font size must be **greater than or equal to 10pt**. You are free to use single-column or two-column layouts and we will allow for extra pages containing only references. If someone else had advised or helped you on this project, your report must fully acknowledge their contributions. **Please include a section that describes what each team member worked on and contributed to the project.**

The report will be judged based on the clarity of the report, **the relevance of the techniques taught in this course, and the novelty of the procedure of the work (You can find existing solutions on the internet, but do not just follow the existing solution). The score will be higher if you use more automatic data cleaning and less manual labeling.**

Reference

- The original competition
- Introduction to Data-Centric AI
- Data augmentation
- Data cleaning
- Label issue
- More Data-Centric AI techniques
- Awesome production machine learning
- Awesome python data science