NSYSU Practical and Innovative Analytics in Data Science - Fall 2024

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 aims to guide you in these areas. It provides an opportunity for you to delve into a dataset, understand its structure, and apply your data-cleaning skills. Specifically, you are expected to utilize data-centric AI techniques on a classification dataset-related topic.

Dataset

The images are collected and modified from the Roman Number dataset. It is noted that both the MNIST and Fashion MNIST datasets have been utilized in laboratory sessions and homework assignments. In this project, we will investigate whether machine learning or neural networks can help us recognize Roman numeral characters.

The original dataset is organized into two main folders: `train` and `val` (i.e., training and validation). Each of these folders contains ten subfolders labeled I to X, each containing images of handwritten Roman numerals from I to X. The training and validation dataset used in this final project is a variant of the aforementioned dataset, comprising about 4,400 characters across 10 different classes.

It is important to note that the model—a truncated version of ResNet50, which is a convolutional neural network—must remain fixed. You are required to modify the provided image dataset in any manner you deem appropriate, with the constraint of a maximum of 12,000 images. Your task is to perform data cleaning/preparation and determine the optimal training and validation split.

The dataset is hosted on the Kaggle platform and can also be accessed here. The training set is organized into different directories and is fully labeled. **Additionally, we provide a testing set containing 500 characters (approximately 10% of the final dataset) to simulate real-world scenarios for your evaluation.**

Rules

The maximum number of images submitted **should be at most 12,000**. You are also permitted to change the training and validation split. **However, the model must remain fixed for this project**. Additionally, 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 the "macro" F1 score, a separate F1 score is calculated for each class/label and then averaged.

Your prediction labels should be encoded into numerical values according to the *mapping.csv* file. You can test your results on the public leaderboard, which is also available on the Kaggle platform. However, submissions are restricted to a maximum of five times per day, and it is good practice not to tune your results based on the test set. The submission file should contain a header and follow the specified format:

```
Id, Expected
0000.png, 1
….
```

`Id` corresponds to the PNG filenames in the test data, and `Expected` corresponds to your predicted label. Refer to the *sample_submission.csv* file if you encounter formatting issues. It is important to note that the macro F1 score displayed on your Kaggle leaderboard is based on only about 10% of the test data. The final score, which will be the macro F1 score calculated over the entire test dataset, will be revealed after the competition concludes.

Grading Policy and Deliverables

- Deadline of dataset and code (12/15 11:59 pm)

The final score will be the sum of the **accuracy of your model on the entire test data (50%), the final presentation, and the final report**. The accuracy score will be calculated as follows:

$$\begin{cases} 0.357 * F1\ score & if\ macro\ F1\ score < 70\% \\ \min(50{,}25 + (F1\ score - 70) * 1.666) & if\ macro\ F1\ score \geq 70\% \end{cases}$$

You are required to submit the preprocessing code to reproduce your results. Please include a link to your **preprocessing code and dataset**, or upload a zip file via Cyber University for your final project. We will use your code and dataset to perform inference on the entire test set. Your submission should be a zip file with the following file structure, with your training and validation data split into different folders. You can rename the `sample_submission` folder as you like, but the name of your zip file should match your folder name. In this example, it should be called `sample_submission.zip`.

```
sample_submission/
│
├── train/
│   ├── i/
│   ├── ii/
│   ├── iii/
```

```
│   └── ...
│
└── val/
    ├── i/
    ├── ii/
    ├── iii/
    └── ...
```

- Final Presentation (Scheduled on 12/16 and 12/23 (9:10~12:00)) (25%)

Each team is required to present their work to the class. **The presentation time is 18 minutes, followed by an additional 5 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. **The final score will be the sum of the grades from the TAs (15%) and myself (10%).**

- Deadline of Final Report (due 12/29 11:59 pm) (25%)

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

- Abstract
- Introduction
- 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 include some discussion. The final project report can be at most 10 pages long (including appendices and figures). The paper size should be standard A4 or 8.5 x 11 inches, and the font size must be at least 10pt. You are free to use single-column or two-column layouts, and we will allow extra pages containing only references. **If someone else has 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 work's procedure (You can find existing solutions on the internet, but do not just follow them). The score will be higher if you use more automatic data cleaning and less manual labeling. **The final score will be the sum of the grades from the TAs (15%) and myself (10%).**

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