NSYSU Data Science Capstone Project - Spring 2022

Midterm Project

Goal

This course aims to prepare you to engage in a real-world project requiring you to apply skills from the entire data science pipeline: preparing, organizing, and transforming data, constructing a model, and evaluating results. The midterm 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 model-building skills. Specifically, you are expected to use deep learning to build a high accuracy model on a topic related to the open set classification problem.

Dataset

The images are collected and modified from Traditional Chinese Handwriting Dataset. It is noted that we have manipulated both MNIST and Fashion MNIST datasets in the laboratory and homework. We will investigate the possibility of whether machine learning or neural networks can help us to recognize handwritten traditional Chinese characters.

The original dataset was produced based on Tegaki, an open-source package. Total 13,065 different Chinese characters, with an average of 50 samples for each character, were generated. The training dataset used in this midterm project is a variant of the above-mentioned dataset, which contains more than 34,000 characters from 800 different classes that are common words in traditional Chinese. However, the testing scenario may exist words outside these 800 different classes. Your job is trying to correctly classify them into 800 different common word classes and reject the word if it does not belong to these 800 classes.

The dataset is held on the Kaggle platform, which can be accessed from here. There is a *mapping.csv* that describe the 800 common word 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 that contains 6,000 characters which represent the real-word scenario that may exist word outside the 800 common word classes.

Rules

You are allowed to use any external datasets, but you should report the dataset you used on the report. You are also allowed to use any pretrained models or any publicly available packages/code. However, do not upload your code/checkpoints to be publicly available during the midterm project. In addition, do not use closed source OCR software.

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*. In addition, you should return -1 if the word does not belong to the 800 common word classes. 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 than 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, 0
….

Both of the above types should be strings. 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 60% of the test data. The final score will be shown after the competition is over, which will be the macro F1 score of the whole test data.

Grading Policy and deliverables

- Deadline (4/19 11:59 pm)

The final score will be the summation of the **accuracy of your model on the whole test data (50%) and the midterm report (50%)**. The former score will be calculated as follows:

$$\begin{cases} 0.4375 * F1\ score & if\ macro\ F1\ score < 80\% \\ \min\left(50{,}35 + (F1\ score - 80)\right) & if\ macro\ F1\ score \geq 80\% \end{cases}$$

- Midterm Report (due 4/19 11:59 pm)

After the midterm, we will also post all the midterm 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 midterm 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**. In addition, you should use **English** to write the report. 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.**

You are requested to hand in the code to reproduce your result and the code should be written in **Python** unless you can't find the desired package in Python or reproducing the required algorithm in Python is tedious. **Please include a link to your code and your model checkpoints, or upload the zip file via cyber university** for your midterm project. We will use your code and checkpoints to perform inference on the whole test set.

The report will be judged based on the clarity of the report, the relevance of the techniques taught in this course, the novelty of the procedure and the correctness of your code and the technical quality of the work.

Reference

- Traditional Chinese Handwriting Dataset
- Tegaki
- CASIA Online and Offline Chinese Handwriting Database