MiniProject 3: Multi-label Classification of Image Data

COMP 551 (001/002), Fall 2021, McGill University

1 Key points

- Lead TA's **Nishanth and Safa**
- This project is **due on November 25th at 11:59pm**.
- Like previous projects, you should form a group of three to work on this problem. All the rules for forming a group apply. The grading is same for all the team members in a group.
- For this project you will need to participate in a Kaggle competition. In addition, you will have to submit the following on MyCourses to be graded:
  1. **code.zip**: Your data processing, classification and evaluation code (.py and .ipynb files).
  2. **writeup.pdf**: Your two-page report of the project.
- We recommend using **Overleaf** for writing your report and **Google colab** for coding and running the experiments. The latter also gives access to the required computational resources. Both platforms enable remote collaborations. You may also choose to use the Google cloud credits provided to you to run the experiments.
- This project also includes a Kaggle competition. **It is important that you participate in it and follow the competition rules given below**:
  1. **Do not** register on Kaggle unless you have finalized your group in MyCourses.
  2. Create a group on Kaggle, where the **group name should be the same as the registered group in MyCourses**.
  3. Only one Kaggle account is allowed per individual. Students must form a team before uploading their solution. Any violations will be noted and penalized appropriately.
  4. Privately sharing code or data outside the team is not allowed. It is okay to share small parts of code on Kaggle, if it is made available to all the participants in the class.
  5. Participants are not allowed to use data from outside. They should use the data provided to train their classifier. But you are allowed to perform any operations on the given data.
  6. Team mergers are not allowed in this competition.
  7. Each team may submit a maximum of 2 solutions per day.
  8. You may select up to 2 submissions as a final solution for private leaderboard ranking (the best one will be selected).

2 Problem Statement

In this mini-project, you will develop models to classify the image data. You will use the combo MNIST dataset (https://www.kaggle.com/c/comp-551-fall-2021/data) provided to you for this problem. Combo MNIST dataset has 60,000 images for training your classifier. Out of the 60,000 training images, 30,000 images have labels associated with them. The remaining 30,000 images do not have labels. You can get creative to use them to train your classifier. Sample images are shown in Figure 1 to give you an idea of the image dataset. The dataset also comes with 15,000 test images whose predictions has to be uploaded to Kaggle to get a score. Roughly 40% these 15,000 images are used to show the rankings in the public leaderboard, while the remaining 60% form private leaderboard which is hidden from everyone. The final rankings are based on the private leaderboard. Please read the evaluation page on Kaggle (https://www.kaggle.com/c/comp-551-fall-2021/overview/evaluation) carefully for formatting your submission file.

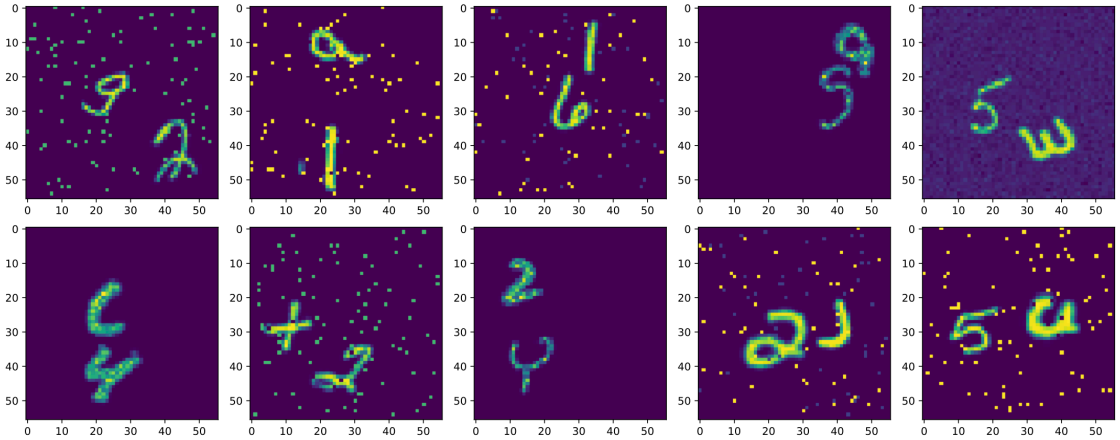

Figure 1: Sample images from combo MNIST dataset

Each image consists of two characters: one letter from the English alphabet (26 letters in total) and one digit from numbers 0-9 (10 digits in total). These characters can be of different size and may appear in any part of the image with different orientations. Additionally, these images may also be affected by noise. Your task is to train a model that can correctly identify the characters in the image. You are free to use any Python libraries you like to extract features, pre-process the data, evaluate your model, and to tune the hyper-parameters, etc.

3 Evaluation

The evaluation is divided into two parts: (i) report writing and (ii) Kaggle competition.

3.1 Report writing

i **Writing (1.5 points):** the overall presentation quality of the report.

ii **Experimentation (1.5 points):** Discussion on various experiments you tried, justifications for choosing the model architecture, hyperparameters, etc.

iii **Creativity (1.5 points):** Originality of your work, how you use the additional unlabelled data, uniqueness of your ideas, etc.

3.2 Kaggle competition

i **Baseline 1 (1 point):** You will receive this grade if your model's performance is above baseline 1 (above 51% accuracy on the private set).

ii **Baseline 2 (1 point):** your model's performance is above baseline 2 (above 70% accuracy on the private set).

iii **Baseline 3 (1 point):** your model's performance is above baseline 3 (above 87% accuracy on the private set).

iv **Top 50% (1 point):** your final standing is in the top 50% in the private leaderboard.

v **Top 10% (1 point):** your final standing is in the top 10% in the private leaderboard.

Finally, **half a point (.5)** is allocated for the active participation in the competition. This includes: making submissions regularly, initiating and taking part in the discussions, and sharing knowledge in terms of research papers or its summaries, blogs, or answering your peers' questions.