Fall 2021 CS 747 Deep Learning
Assignment 4

Due date: November 19, 2021, 11:59:59 pm

Announcement Please note that this assignment requires using Pytorch, and it needs to run on GPUs. You are highly recommended to use Google Colab to finish this assignment, unless you have a local machine with powerful GPUs. Since the process of running one model is time-consuming, **please start the assignment early**.

1 Text Generation and Language classification with a RNN (100 points)

1.1 Part 1 (80 points)

Data Setup To download the data for the RNN tasks, go to the Assignment folder and run the `download_language` script provided:

```
$ cd CS747_Assignment4/
$ ./download_language.sh
```

The data for the generation task is the complete works of Shakespeare all concatenated together. The data we use for the language classification task is a set of translations of the Bible in 20 different Latin alphabet based languages (i.e., languages where converting from unicode to ASCII may be somewhat permissible).

Implementation This part of the assignment is split into two notebooks (`MP4_generation.ipynb` and `MP4_classification.ipynb`). The `MP4_generation.ipynb` will guide you through the implementation of your RNN model. You will use the same model framework you implement for classification portion in the `MP4_classification.ipynb` notebook.

Both of the RNN tasks in this assignment are not as computation heavy and can be trained in a short amount of time on GPU or CPU.

1.2 Part 2 (20 points)

Download or scrape your own dataset. Your data could be a book by your favorite author, a large codebase in your favorite programming language, or other text data you have scraped off the internet.

Design your own neural networks (it needs to be different from Part 1) and implement language generation task.

Submission Instructions

Please submit all your files on GMU Blackboard Portal before the due date.

- All of your code (python files and ipynb file) in a single ZIP file. The filename should be `netid_mp4_code.zip`.
- Your ipython notebooks with output cells converted to PDF format. The filenames should be `netid_mp4_output.pdf`.
- Submit an output Kaggle submission CSV file on a provided test subset for the RNN classification task to Kaggle competition. [1]
- A brief report in PDF format using this template[2], with name `netid_assignment4_report.pdf`.

[1] `https://www.kaggle.com/t/127d8de9a3314744aa909c0f244d2f3b`

[2] `https://docs.google.com/document/d/1H8Dh2z2SmKarOFD3_n66_Wtxi5Xz7-zpqlTQ7hEKuhE/edit?usp=sharing`