Additional Information about Task Sheet 1

There might be existing implementations of the topics we are offering, but the main objective of this course is for you to create the entire project from scratch. You are not permitted to rely on pre-existing implementations or simply copy and paste code from existing repositories.

In the following, you will discover further details regarding each section included in the initial task sheet.

Implementation

1. Data Preparation

In this milestone, one of your tasks is to develop a dataset class for each dataset you intend to use. If your research involves multiple datasets, you can begin with one and then incorporate additional datasets in subsequent milestones. If the data is not already present in the cluster, you can easily obtain it by downloading and saving it to the designated path (/home/space/datasets/<dataset>) within the cluster. Please check the following links to learn more about creating custom datasets in PyTorch.

- Writing Custom Datasets, DataLoaders and Transforms — PyTorch Tutorials 2.1.0+cu121 documentation
- https://pytorch.org/tutorials/beginner/basics/data_tutorial.html

Your custom dataset, created for each specific dataset, should resemble the code snippet provided below.

```
class CustomDataset(Dataset):
      def __init__(self, …):
            super.()__init__()
            [ … ]
      def __len__(self, …):
            [ … ]
      def __getitem__(self, …):
            [ … ]
```

The load_data function is supposed to create an instance of your custom dataset class by specifying the dataset name as an argument. As per the task sheet instructions, if the number of test/train samples is pre-defined, you can set n_train and n_test to None.

2. Data Visualization

Moreover, it is essential to develop data visualization tools.

Images: include sample images for each dataset to provide a visual representation of your data.
Text: incorporate sample text snippets and their corresponding labels in your report, and consider including a Word Cloud to enhance your data presentation.

***Note:* Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be

highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

3. Model Prototype

Additionally, your tasks include implementing the model, the associated loss function, and a basic training loop. In this milestone, there is no need to train the model or make it work perfectly.

***Note:* Make sure you can load the data and forward it through the untrained model.

Report

• Dataset overview

Your report should include the following:

1. Overview of datasets used, including general data information, data labels, number of data points per class, dataset visualization, and other relevant statistics.
2. Discussion on feature extraction methods, comparing classical techniques with modern ones using neural networks.
3. Consideration of data normalization and its necessity, along with any other data transformations employed.
4. Exploration of questions that can be asked about the data and how machine learning models can address these inquiries.

• Baseline method and evaluation

In your report, include:

1. Model Description: Briefly describe your model and its architecture (details about the model's architecture and training procedure are not needed).
2. Literature Review: Discuss prior approaches, their limitations, and how your model addresses them.
3. Evaluation Methods: Explain the methods used for assessing performance.
4. Class Differences: Examine variations among different classes in the dataset (in terms of how well they can be represented or classified by the baseline model).
5. Good Features Definition: Define what constitutes good features in machine learning.

• Discussion

Please check the task sheet. No further explanations are needed for this section.

The provided explanations serve as a supplementary resource to the first task sheet.

Important

1. Examination registration deadline: 21.11.2024 at 23:59
2. Submit your report on time.
3. Report <= 10 pages.
4. Provide a clean (`model class, dataset class, …`) and executable code on the cluster. Do not submit Jupyter Notebooks.