NLP mastering

Skyline System Ltd.

Week 1

August 3, 2023

Text classification tasks

The team will start with the implementation of text classification task in PyTorch Lightening (https://www.pytorchlightning.ai/index.html).

For the completion of this task, following steps are required to be done:

1. Dataset downloading
2. Preprocessing
3. Splitting dataset for training, validation, and testing
4. Model training and evaluation for text classification
5. Selecting performance metrics for text classification (accuracy, F1, precision, and recall etc.)
6. Analyzing model predictions
7. Visualizing model predictions
8. Performing error analysis

Sample datasets to start with

In this task, you will work on any two datasets.

1. https://archive.ics.uci.edu/dataset/380/youtube+spam+collection
2. https://www.kaggle.com/wcukierski/enron-email-dataset
3. https://www.kaggle.com/rtatman/fraudulent-email-corpus

Choose one dataset for the initial coding of the entire task. You'll use the second dataset to ensure that your code is modular enough as with minimal effort you can reuse the same code for the second dataset.

Structuring your deep learning project

It is necessary to structure your deep learning project so that it is usable and understandable by millions of users in the open-source community. This leads to determining how to write organized, modularized, and extensible code. Like other software, deep learning code also has a project structure, a documentation, and design principles such as object-oriented programming. For further necessary reading, please go through: https://theaisummer.com/best-practices-deep-learning-code/.

Writing Quality Code

We need you to ensure that your written code is a quality code, you are requested to make you code modular enough so that I's Readable, Extendable and Deployable. Following is a good read in this regard: How to Write Good Quality Machine Learning Code by Ujwal Tiwar at Medium

https://medium.com/@ujwalkaka/how-to-write-good-quality-machine-learning-code-6e2f0131e46

Finally in the code quality regards, please keep in mind that **we will adopt coding practices suitable for open source,** e.g., python virtual environment, testing, packaging etc. One suggestion regarding this includes formally adopt the use of **Github Co-pilot** too.

Please follow and explore further GitHub repositories of PyTorch and HuggingFace Transformers.

Preprints on ar*X*iv

On a related note, our goal should be to publish preprints on arXiv. Therefore, we must get used to rigorous evaluation and benchmarking. To get use too, I would suggest reading the latest preprints in the domains of NLP and deep learning so that at time we start writing those it’s not a new thing.