Feature 2

Course: Large-scale AI engineering FS25

Author: Maja Gwóźdź

Date: May 16, 2025

We present a simple pretokeniser that works with the files from Assignment 2. For the test, we have used the `unsloth/Mistral-Nemo-Base-2407-bnb-4bit` tokenizer and the `train_data.parquet` dataset from the assignment. For the sequence length, we chose 200, for batch size 4096, and the `flush-rows` parameter was set to 2500. The experiment took approximately 8 minutes. All CLI arguments are described in the `pretokenize.py` file. To accelerate the process, we flush after accumulating 25000 chunks (this is the default value). We also add the slightly modified `dataset.py` and `train.py` files to the submission but those are not necessary for testing. To do a minimal test, run the `pretokenize.py` file with the default parameters (adjust the paths), and then run `check.py` to see if the experiment was successful. We also attach some minimal `sbatch` files that were used for the experiment (again, please adjust the paths accordingly).

Below are log excerpts from the experiments. The first one is from running the `pretokenize.py` file, and the second one is from the `check.py` run.

```
Tokenising: 192it [07:20,  2.29s/it]t]2,  2.39s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:20,  2.30s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:21,  2.30s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:22,  2.31s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:23,  2.31s/it]
```

```
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:25,  2.32s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:23,  2.31s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:24,  2.32s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:25,  2.32s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:27,  2.33s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:28,  2.33s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:32,  2.36s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:31,  2.35s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:33,  2.36s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:32,  2.36s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet
Tokenising: 192it [07:36,  2.38s/it]
Wrote 4,558,148 rows to /iopsstor/scratch/cscs/mgwozdz/datasets/train_data_tok.parquet

Input shape: torch.Size([32, 200])
Labels shape: torch.Size([32, 200])
Ignored tokens in loss: 128 out of 6400 (2.00%)
all shapes & masking OK
```