Chapter 5 Datasets Library

Wednesday, August 2, 2023 10:19 AM

5.2 Dataset ~~से~~ HF Hub ~~से~~ ~~को~~ ~~लाना~~

load_dataset

flexible

Single path

list of paths

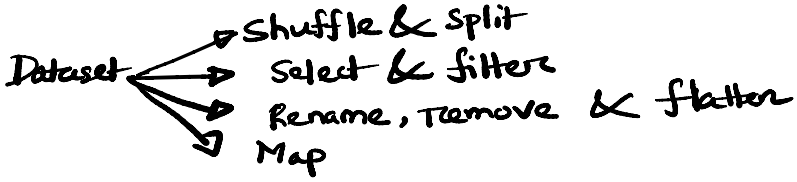
dict. maps split names to file path.

Automatic decompression of i/p files.

GZIP, ZIP, TAR

5.3 Slice & dice a dataset

Dataset won't perfectly prepare for training



5.4 Big data:-

Millions of doc

GB of size

HF Datasets library

frees memory management problems

memory-mapped files.

GPT-J corpus → The pile 1.2 TB

T-5 u → C4 ≈ 3 TB

Datasets library uses

Apache arrow Traditional ROW/column format

Streaming API. Iterable dataset.

Arrow's Memory mapped format

enables access to bigger than RAM datasets

multiple processes to work with same dataset W/o moving/copying

Pile

English text corpus.

Pile

English text corpus.

EleutherAI

Diverse range of topics

14.3 GB

2 columns:- meta, text

num_rows:- 15518009

RAM usage:-

psutil.Process().memory_info().rss

rss = resident set size

Dataset size

name_of_dataset["train"].dataset_size

Size ≈ 20 GB

কিন্তু RAM used only 5600 MB

Wes Kinney's famous rule of thumb for pandas library

5X/10X RAM লাগে dataset size এর থেকে।

Datasets library treat each datasets as a memory-mapped file

mapping betn RAM & file storage w/o fully load into memory

done by apache arrow memory format & pyarrow lib.

Parquet:- compress করে file এ store করে রাখে data

80% Compression Ratio.

যদি dataset এর size আরো বড়ো হয়, disk এর চেয়েও বেশি

তখন Streaming = True argument এর দ্বারা dataset এর element

কে access করা হয় on the fly, w/o need to download whole dataset.

Access first element

next(iter(name_of_dataset))

elements can be processed by

IterableDataset.map()

Shuffle করার জন্য

Shuffle করার জন্য

```
sh_ds = name_of_dataset.shuffle(buffer_size=10,000, seed=42)
    next(iter(sh_ds))
```

Select করার জন্য

```
d = dataset.take(5)
list(d)
```

Skip করার জন্য .skip()

একাধিক dataset কে combine করার জন্য :-

```
interleave_datasets()

Combined_dataset = interleave_datasets([dataset_1, dataset_2])
list(islice(combined_dataset, 2))
```

5.5 Creating your own dataset :-

Creating a corpus of Github issues :-

to track bugs/features in Github repo.

use cases :-

time to take to close open issues/pull requests.

training a multilabel classifier to tag metadata based on issue's description

Creating semantic search engines to find issues match user query.

Getting data :-

all issues of HF datasets in repo's issues tab

to download all repo's issues, use Github REST API :-

5.6 Semantic search with FAISS :-

Sentence ⟶ Transformer Encoders ⟶ embed. vectors.

cosine sim. for measuring how close two embeddings are :-

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

FAISS (Facebook AI similarity Search)

FAISS (Facebook AI similarity Search)

Sim. search based on embed.

Convert long passages into single embed.

" Question " "

Compare 2 embed. using FAISS

5.6 Use embedding for semantic search:—

rep. each token as embed. vector

pool individual embed. for vector rep. of whole sentences.

Loading & preparing