Lưu trữ đám mây

Bộ dữ liệu ôm mặt

Trung tâm tập dữ liệu Ôm Mặt là nơi tập hợp ngày càng nhiều các tập dữ liệu trải rộng trên nhiều loại của các miền và nhiệm vụ.

Không chỉ là nơi lưu trữ đám mây: Dataset Hub là một nền tảng cung cấp phiên bản dữ liệu nhờ git, cũng như Trình xem tập dữ liệu để khám phá dữ liệu, khiến nó trở thành một nơi tuyệt vời để lưu trữ Bộ dữ liệu sẵn sàng cho AI.

Hướng dẫn này chỉ ra cách nhập dữ liệu từ bộ lưu trữ đám mây khác bằng hệ thống tệp triển khai từ fsspec .

Nhập dữ liệu từ bộ lưu trữ đám mây

Hầu hết các nhà cung cấp lưu trữ đám mây đều có triển khai Hệ thống tệp fsspec, rất hữu ích cho nhập dữ liệu từ bất kỳ nhà cung cấp đám mây nào có cùng mã.
Điều này đặc biệt hữu ích khi xuất bản bộ dữ liệu về Ôm Mặt.

Hãy xem bảng sau để biết một số ví dụ về các nhà cung cấp dịch vụ lưu trữ đám mây được hỗ trợ:

Nhà cung cấp lưu trữ	Triển khai hệ thống tập tin
Amazon S3	s3fs
Bộ nhớ đám mây của Google	gcsfs
Azure Blob/DataLake	adlf
Lưu trữ đám mây của Oracle	ocif

Hướng dẫn này sẽ chỉ cho bạn cách nhập tệp dữ liệu từ bất kỳ bộ lưu trữ đám mây nào và lưu tập dữ liệu trê Ôm Mặt.

Giả sử chúng tôi muốn xuất bản tập dữ liệu về Ôm mặt từ tệp Parquet từ đám mây kho.

Trước tiên, hãy khởi tạo hệ thống tệp lưu trữ đám mây của bạn và liệt kê các tệp bạn muốn nhập:

```
>>> import fsspec
>>> fs = fsspec.filesystem("...")# s3 / gcs / abfs / adl / oci / ...
>>> data_dir = "path/to/my/data/"
>>> pattern = "*.parquet"
>>> data_files = fs.glob(data_dir + pattern)
["path/to/my/data/0001.parquet", "path/to/my/data/0001.parquet", ...]
```

Sau đó, bạn có thể tạo tập dữ liệu trên Ôm mặt và nhập các tệp dữ liệu, ví dụ:

```
>>> from huggingface_hub import create_repo, upload_file
>>> from tqdm.auto import tqdm
>>> destination_dataset = "username/my-dataset"
>>> create_repo(destination_dataset, repo_type="dataset")
>>> for data_file in tqdm(fs.glob(data_dir + pattern)):
...with fs.open(data_file) as fileobj:
...path_in_repo = data_file[len(data_dir):]
...upload_file(
...path_or_fileobj=fileobj,
...path_in_repo=path_in_repo,
...repo_id=destination_dataset,
...repo_type="dataset",
...)
```

Hãy xem tài liệu về ôm mặt_hub về các tệp tải lên tại đây nếu bạn đang tìm kiếm thêm tùy chọn tải lên.

Cuối cùng, bây giờ bạn có thể tải tập dữ liệu bằng cách sử dụng Bộ dữ liệu:

```
>>> from datasets import load_dataset
>>> ds = load_dataset("username/my-dataset")
```