Chia sẻ tập dữ liệu bằng CLI

Tại Hugging Face, chúng tôi đang thực hiện sứ mệnh dân chủ hóa Machine Learning tốt và chúng tôi tin rằng về giá trị của nguồn mở. Đó là lý do tại sao chúng tôi thiết kế Bộ dữ liệu để bất kỳ ai cũng có thể chia sẻ tập dữ liệu với cộng đồng ML lớn hơn. Hiện tại có hàng ngàn bộ dữ liệu trong hơn 100 ngôn ngữ trong Hugging Face Hub và nhóm Hugging Face luôn chào đón những ngôn ngữ mới đóng góp!

Kho lưu trữ dữ liệu cung cấp các tính năng như:

- Lưu trữ dữ liệu miễn phí
- Lập phiên bản tập dữ liệu
- Lịch sử cam kết và những khác biệt
- Siêu dữ liệu cho khả năng khám phá
- Thẻ tập dữ liệu dành cho tài liệu, cấp phép, giới hạn, v.v.
- Trình xem tập dữ liệu

Hướng dẫn này sẽ chỉ cho bạn cách chia sẻ thư mục tập dữ liệu hoặc kho lưu trữ có thể dễ dàng được truy cập bởi bất cứ ai.

Thêm tập dữ liệu

Bạn có thể chia sẻ tập dữ liệu của mình với cộng đồng bằng kho lưu trữ tập dữ liệu trên Khuôn mặt ôm Trung tâm.
Nó cũng có thể là một tập dữ liệu riêng tư nếu bạn muốn kiểm soát ai có quyền truy cập vào nó.

Trong kho lưu trữ tập dữ liệu, bạn có thể lưu trữ tất cả các tệp dữ liệu của mình và định cấu hình tập dữ liệu tập tin nào đi đến phần chia nào.
Các định dạng sau được hỗ trợ: CSV, TSV, JSON, dòng JSON, văn bản, Parquet, Arrow, SQLite, Bộ dữ liệu Web.
Nhiều loại tệp nén cũng được hỗ trợ: GZ, BZ2, LZ4, LZMA hoặc ZSTD.
Ví dụ: tập dữ liệu của bạn có thể được tạo từ các tệp .json.gz.

Khi tải tập dữ liệu từ Hub, tất cả các tệp ở định dạng được hỗ trợ sẽ được tải, theo cấu trúc kho lưu trữ.

Để biết thêm thông tin về cách tải tập dữ liệu từ Hub, hãy xem phần tải tập dữ liệu từ hướng dẫn của Hub.

Tạo kho lưu trữ

Việc chia sẻ tập dữ liệu cộng đồng sẽ yêu cầu bạn tạo tài khoản trên hf.co nếu bạn chưa có một cái nữa.
Bạn có thể trực tiếp tạo kho lưu trữ dữ liệu mới từ tài khoản của mình trên Hugging Face Hub, nhưng hướng dẫn này sẽ chỉ cho bạn cách tải lên tập dữ liệu từ thiết bị đầu cuối.

1. Đảm bảo rằng bạn đang ở trong môi trường ảo nơi bạn đã cài đặt Bộ dữ liệu và chạy lệnh sau:

```
đăng nhập ôm mặt-cli
```

2. Đăng nhập bằng thông tin đăng nhập Hugging Face Hub của bạn và tạo kho lưu trữ dữ liệu mới:

```
repo ôm mặt-cli tạo tập dữ liệu my-cool-dataset --type
```

Thêm cờ -tổ chức để tạo kho lưu trữ trong một tổ chức cụ thể:

```
repo ôm mặt-cli tạo tập dữ liệu my-cool-dataset --type --tổ chức tên tổ chức của bạn
```

Chuẩn bị tập tin của bạn

Kiểm tra thư mục của bạn để đảm bảo các tệp duy nhất bạn đang tải lên là:

- Các file dữ liệu của tập dữ liệu
- Thẻ dữ liệu README.md

tải lên ôm mặt-cli

Sử dụng lệnh tải lên ôm mặt-cli để tải tệp trực tiếp lên Hub. Trong nội bộ, nó sử dụng các trình trợ giúp upload_file và upload_folder tương tự được mô tả trong hướng dẫn Tải lên. trong dụ bên dưới, chúng ta sẽ xem xét các trường hợp sử dụng phổ biến nhất. Để có danh sách đầy đủ có sẵn

tùy chọn, bạn có thể chạy:

```
>>> huggingface-cli upload --help
```

Để biết thêm thông tin chung về ôm mặt-cli, bạn có thể xem hướng dẫn CLI.

Tải lên toàn bộ thư mục

Cách sử dụng mặc định cho lệnh này là:

```
# Usage:  huggingface-cli upload [dataset_repo_id] [local_path] [path_in_repo] --repo-type dataset
```

Để tải lên thư mục hiện tại ở thư mục gốc của repo, hãy sử dụng:

```
>>> huggingface-cli upload my-cool-dataset . . --repo-type dataset
https://huggingface.co/datasets/Wauplin/my-cool-dataset/tree/main/
```

[!TIP]
Nếu repo chưa tồn tại, nó sẽ được tạo tự động.

Bạn cũng có thể tải lên một thư mục cụ thể:

```
>>> huggingface-cli upload my-cool-dataset ./data . --repo-type dataset
https://huggingface.co/datasetsWauplin/my-cool-dataset/tree/main/
```

Cuối cùng, bạn có thể tải thư mục lên một đích cụ thể trên repo:

```
>>> huggingface-cli upload my-cool-dataset ./path/to/curated/data /data/train --repo-type dataset
https://huggingface.co/datasetsWauplin/my-cool-dataset/tree/main/data/train
```

Tải lên một tập tin

Bạn cũng có thể tải lên một tệp bằng cách đặt local_path để trỏ đến một tệp trên máy của bạn. Nếu như trong trường hợp đó, path_in_repo là tùy chọn và sẽ mặc định là tên tệp cục bộ của bạn:

```
>>> huggingface-cli upload Wauplin/my-cool-dataset ./files/train.csv --repo-type dataset
https://huggingface.co/datasetsWauplin/my-cool-dataset/blob/main/train.csv
```

Nếu bạn muốn tải một tệp lên một thư mục cụ thể, hãy đặt path_in_repo tương ứng:

```
>>> huggingface-cli upload Wauplin/my-cool-dataset ./files/train.csv /data/train.csv --repo-type d
https://huggingface.co/datasetsWauplin/my-cool-dataset/blob/main/data/train.csv
```

Tải lên nhiều tập tin

Để tải lên nhiều tệp từ một thư mục cùng một lúc mà không tải lên toàn bộ thư mục, hãy sử dụng lệnh --include và --exclude các mẫu. Nó cũng có thể được kết hợp với tùy chọn --delete để xóa các tập tin trên repo trong khi tải lên những cái mới. Trong ví dụ bên dưới, chúng tôi đồng bộ hóa địa ph Dung lượng bằng cách xóa các tệp từ xa và tải lên tất cả các tệp CSV:

```
# Sync local Space with Hub (upload new CSV files, delete removed files)
>>> huggingface-cli upload Wauplin/my-cool-dataset --repo-type dataset --include="/data/*.csv" --d
None
```

Tải lên một tổ chức

Để tải nội dung lên kho lưu trữ thuộc sở hữu của một tổ chức thay vì kho lưu trữ cá nhân, bạn phải chỉ định rõ ràng nó trong repo_id :

```
>>> huggingface-cli upload MyCoolOrganization/my-cool-dataset . . --repo-type dataset
https://huggingface.co/datasetsMyCoolOrganization/my-cool-dataset/tree/main/
```

Tải lên một bản sửa đổi cụ thể

Theo mặc định, các tập tin được tải lên nhánh chính. Nếu bạn muốn tải tập tin lên chi nhánh khác hoặc tham chiếu, hãy sử dụng tùy chọn --revision:

```
# Tải tệp lên PR
ôm mặt-cli tải lên bigcode/the-stack . . --repo-type tập dữ liệu --revision refs/pr/104
None
```

Lưu ý: nếu bản sửa đổi không tồn tại và --create-pr không được đặt, một nhánh sẽ được tạo tự động từ nhánh chính.

Tải lên và tạo PR

Nếu bạn không có quyền chuyển sang repo, bạn phải mở PR và để tác giả biết về những thay đổi bạn muốn thực hiện. Điều này có thể được thực hiện bằng cách đặt --create-pr lựa chọn:

```
# Tạo một PR và tải các tập tin lên nó
>>> huggingface-cli upload bigcode/the-stack --repo-type dataset --revision refs/pr/104 --create-p
https://huggingface.co/datasets/bigcode/the-stack/blob/refs%2Fpr%2F104/
```

Tải lên đều đặn

Trong một số trường hợp, bạn có thể muốn đẩy các bản cập nhật thường xuyên lên kho lưu trữ. Ví dụ, điều r tập dữ liệu của bạn đang tăng lên theo thời gian và bạn muốn tải thư mục dữ liệu lên cứ sau 10 phút. Bạn có thể thực hiện việc này bằng tùy chọn --every:

```
# Tải lên nhật ký mới cứ sau 10 phút
huggingface-cli upload my-cool-dynamic-dataset data/ --every=10
```

Chỉ định một thông điệp cam kết

Sử dụng --commit-message và --commit-description để đặt thông báo tùy chỉnh và mô tả cho cam kết của bạn thay vì cam kết mặc định

```
>>> huggingface-cli upload Wauplin/my-cool-dataset ./data . --repo-type dataset --commit-message="
None
https://huggingface.co/datasetsWauplin/my-cool-dataset/tree/main
```

Chỉ định mã thông báo

To upload files, you must use a token. By default, the token saved locally (using huggingface-cli login) will be used. If you want to authenticate explicitly, use the --token lựa chọn:

```
>>> huggingface-cli upload Wauplin/my-cool-dataset ./data . --repo-type dataset --token=hf_****
None
https://huggingface.co/datasetsWauplin/my-cool-data/tree/main
```

Chế độ im lặng

Theo mặc định, lệnh tải lên ôm mặt-cli sẽ dài dòng. Nó sẽ in các chi tiết như thông báo cảnh báo, thông tin về các tệp đã tải lên và thanh tiến trình. Nếu bạn muốn silence all of this, use the --quiet option. Only the last line (i.e. the URL to the uploaded files) được in. Điều này có thể hữu ích nếu bạn muốn chuyển đầu ra sang lệnh khác trong tập lệnh.

```
>>> huggingface-cli upload Wauplin/my-cool-dataset ./data . --repo-type dataset --quiet
https://huggingface.co/datasets/Wauplin/my-cool-dataset/tree/main
```

Thưởng thức !

Xin chúc mừng, tập dữ liệu của bạn hiện đã được tải lên Hugging Face Hub nơi mọi người có thể tải nó trong một dòng mã!

```
dataset = load_dataset("Wauplin/my-cool-dataset")
```

Nếu tập dữ liệu của bạn được hỗ trợ thì nó cũng phải có Trình xem tập dữ liệu để mọi người khám phá nội dung tập dữ liệu.

Cuối cùng, đừng quên làm phong phú thẻ tập dữ liệu để ghi lại tập dữ liệu của bạn và biến nó thành tài liệu có thể khám phá được! Hãy xem hướng dẫn Tạo thẻ tập dữ liệu để tìm hiểu thêm.