Tạo tập dữ liệu video

Hướng dẫn này sẽ chỉ cho bạn cách tạo tập dữ liệu video bằng VideoFolder và một số siêu dữ liệu.
Đây là giải pháp không cần mã để tạo nhanh tập dữ liệu video với hàng nghìn video.

> [!TIP]
> Bạn có thể kiểm soát quyền truy cập vào tập dữ liệu của mình bằng cách yêu cầu người dùng chia sẻ liên
> thông tin đầu tiên. Hãy xem hướng dẫn về bộ dữ liệu Gated để biết thêm thông tin về cách
> kích hoạt tính năng này trên Hub.

Thư mục Video

VideoFolder là trình tạo tập dữ liệu được thiết kế để tải nhanh tập dữ liệu video với một số
nghìn video mà không yêu cầu bạn phải viết bất kỳ mã nào.

> [!TIP]
> Hãy xem hệ thống phân cấp Mẫu phân chia để tìm hiểu thêm về cách VideoFolder tạo
> phân chia tập dữ liệu dựa trên cấu trúc kho lưu trữ tập dữ liệu của bạn.

VideoFolder tự động suy ra nhãn lớp của tập dữ liệu của bạn dựa trên tên thư mục.
Lưu trữ tập dữ liệu của bạn trong cấu trúc thư mục như:

```
thư mục/xe lửa/chó/golden_retriever.mp4
folder/train/dog/german_shepherd.mp4
folder/train/dog/chihuahua.mp4

thư mục/train/cat/maine_coon.mp4
thư mục/train/cat/bengal.mp4
folder/train/cat/birman.mp4
```

Nếu tập dữ liệu tuân theo cấu trúc VideoFolder thì bạn có thể tải nó trực tiếp bằng
load_dataset():

```
>>> from datasets import load_dataset

>>> dataset = load_dataset("path/to/folder")
```

This is equivalent to passing videofolder manually in load_dataset() and the directory in dữ liệu_dir :

```
>>> dataset = load_dataset("videofolder", data_dir="/path/to/folder")
```

Bạn cũng có thể sử dụng thư mục video để tải tập dữ liệu liên quan đến nhiều phần tách. Để làm như vậy, tậ thư mục nên có cấu trúc như sau:

```
thư mục/xe lửa/chó/golden_retriever.mp4
thư mục/train/cat/maine_coon.mp4
folder/test/dog/german_shepherd.mp4
thư mục/test/cat/bengal.mp4
```

> [!WARNING]
> Nếu tất cả các tập tin video được chứa trong một thư mục duy nhất hoặc nếu chúng không cùng cấp độ cấu trúc thư mục, cột nhãn sẽ không được thêm tự động. Nếu bạn cần thì hãy đặt drop_labels=False explicitly.

Nếu có thông tin bổ sung mà bạn muốn đưa vào về tập dữ liệu của mình, như chú thích văn bản hoặc các hộp giới hạn, hãy thêm nó dưới dạng tệp siêu dữ liệu.csv trong thư mục của bạn. Điều này cho phép bạn bộ dữ liệu cho các tác vụ thị giác máy tính khác nhau như chú thích văn bản hoặc phát hiện đối tượng. bạn c cũng sử dụng tệp JSONL siêu dữ liệu.jsonl hoặc tệp Parquet siêu dữ liệu.parquet .

```
thư mục/train/metadata.csv
thư mục/tàu/0001.mp4
thư mục/tàu/0002.mp4
thư mục/tàu/0003.mp4
```

Tệp siêu dữ liệu.csv của bạn phải có trường file_name hoặc *_file_name để liên kết các tệp video với siêu dữ liệu của họ:

```
tên_tệp,tính năng bổ sung
0001.mp4,Đây là giá trị đầu tiên của tính năng văn bản bạn đã thêm vào video của mình
0002.mp4,Đây là giá trị thứ hai của tính năng văn bản bạn đã thêm vào video của mình
0003.mp4,Đây là giá trị thứ ba của tính năng văn bản mà bạn đã thêm vào video của mình
```

hoặc sử dụng siêu dữ liệu.jsonl:

```
{"file_name": "0001.mp4", "additional_feature": "This is a first value of a text feature you added
{"file_name": "0002.mp4", "additional_feature": "This is a second value of a text feature you adde
{"file_name": "0003.mp4", "additional_feature": "This is a third value of a text feature you added
```

Ở đây file_name phải là tên của tệp video bên cạnh tệp siêu dữ liệu. Hơn
nói chung, nó phải là đường dẫn tương đối từ thư mục chứa siêu dữ liệu đến video
tài liệu.

Bạn có thể trỏ tới nhiều video ở mỗi hàng trong tập dữ liệu của mình, chẳng hạn nếu cả hai
đầu vào và đầu ra của bạn là video:

```
{"input_file_name": "0001.mp4", "output_file_name": "0001_output.mp4"}
{"input_file_name": "0002.mp4", "output_file_name": "0002_output.mp4"}
{"input_file_name": "0003.mp4", "output_file_name": "0003_output.mp4"}
```

Bạn cũng có thể xác định danh sách video. Trong trường hợp đó bạn cần đặt tên cho trường file_names hoặc
*_file_names . Đây là một ví dụ:

```
{"videos_file_names": ["0001_left.mp4", "0001_right.mp4"], "label": "moving_up"}
{"videos_file_names": ["0002_left.mp4", "0002_right.mp4"], "label": "moving_down"}
{"videos_file_names": ["0003_left.mp4", "0003_right.mp4"], "label": "moving_right"}
```

Chú thích video

Bộ dữ liệu phụ đề video có văn bản mô tả video. Một ví dụ về siêu dữ liệu.csv có thể trông
giống:

```
tên_tệp,văn bản
0001.mp4,Đây là chú chó tha mồi vàng đang chơi bóng
0002.mp4,Một chú chó chăn cừu Đức
0003.mp4,Một con chihuahua
```

Tải tập dữ liệu bằng VideoFolder và nó sẽ tạo một cột văn bản cho chú thích video:

```
>>> dataset = load_dataset("videofolder", data_dir="/path/to/folder", split="train")
>>> dataset[0]["text"]
"Đây là một chú chó tha mồi vàng đang chơi với một quả bóng"
```

Tải tập dữ liệu lên Hub

Ví dụ: khi bạn đã tạo tập dữ liệu, bạn có thể chia sẻ tập dữ liệu đó với bằng cách sử dụng ôm mặt_hub. Đảm bảo bạn đã cài đặt thư viện ôm mặt_hub và bạn đã đăng nhập vào Hugging Face account (see the Upload with Python tutorial for more details).

Tải tập dữ liệu của bạn lên với ômface_hub.HfApi.upload_folder :

```
từ ôm mặt_hub nhập HfApi
api = HfApi()

api.upload_folder(
    folder_path="/path/to/local/dataset",
    repo_id="username/my-cool-dataset",
    repo_type="dataset",
)
```

Bộ dữ liệu Web

Định dạng WebDataset dựa trên kho lưu trữ TAR và phù hợp với các bộ dữ liệu video lớn. Indeed you can group your videos in TAR archives (e.g. 1GB of videos per TAR archive) and có hàng ngàn kho lưu trữ TAR:

```
thư mục/train/00000.tar
thư mục/train/00001.tar
thư mục/train/00002.tar
None
```

Trong kho lưu trữ, mỗi ví dụ được tạo từ các tệp có chung tiền tố:

```
e39871fd9fd74f55.mp4
e39871fd9fd74f55.json
f18b91585c4d3f3e.mp4
f18b91585c4d3f3e.json
ede6e66b2fb59aab.mp4
ede6e66b2fb59aab.json
ed600d57fcee4f94.mp4
ed600d57fcee4f94.json
None
```

Ví dụ: bạn có thể đặt nhãn/chú thích/tính năng cho video của mình bằng cách sử dụng tệp JSON hoặc văn bả

Để biết thêm chi tiết về định dạng WebDataset và thư viện python, vui lòng kiểm tra
Tài liệu WebDataset.

Load your WebDataset and it will create on column per file suffix (here "mp4" and "json"):

```
>>> from datasets import load_dataset

>>> dataset = load_dataset("webdataset", data_dir="/path/to/folder", split="train")
>>> dataset[0]["json"]
{"bbox": [[302.0, 109.0, 73.0, 52.0]], "categories": [0]}
```