Quản lý bộ đệm

Khi bạn tải xuống tập dữ liệu từ Ôm mặt, dữ liệu sẽ được lưu trữ cục bộ trên thiết bị của bạn.
máy tính.
Các tập tin từ Ôm mặt được lưu trữ như bình thường trong bộ đệm ôm mặt_hub, tại
~/.cache/huggingface/hub theo mặc định.
Xem tài liệu về bộ đệm Hub để biết thêm chi tiết và cách thay đổi vị trí của nó.

Bộ đệm Hub cho phép Bộ dữ liệu tránh tải xuống lại các tệp dữ liệu từ Ôm mặt
mỗi khi bạn sử dụng chúng.

Datasets also has its own cache to store datasets converted in Arrow format (the format
used by Dataset objects).

Hướng dẫn này tập trung vào Bộ đệm bộ dữ liệu và sẽ chỉ cho bạn cách:

- Thay đổi thư mục bộ đệm.
- Kiểm soát cách tải tập dữ liệu từ bộ đệm.
- Dọn dẹp các tập tin bộ nhớ đệm trong thư mục.
- Bật hoặc tắt bộ nhớ đệm.

Thư mục bộ đệm

Thư mục bộ đệm Datasets mặc định là ~/.cache/huggingface/datasets . Thay đổi bộ đệm
location bằng cách đặt biến môi trường shell, HF_HOME sang thư mục khác:

```
$ export HF_HOME="/path/to/another/directory/datasets"
```

Ngoài ra, bạn có thể đặt biến môi trường HF_DATASETS_CACHE để chỉ kiểm soát
thư mục bộ đệm dành riêng cho bộ dữ liệu:

```
$ export HF_DATASETS_CACHE="/path/to/datasets_cache"
```

This only applies to files written by the datasets library (e.g., Arrow files and indices).
It does not affect files downloaded from the Hugging Face Hub (such as models, tokenizers, or

raw dataset sources), which are located in ~/.cache/huggingface/hub by default and controlled
riêng biệt thông qua biến HF_HUB_CACHE:

```
$ export HF_HUB_CACHE="/path/to/hub_cache"
```

Nếu bạn muốn di chuyển tất cả bộ nhớ đệm Ôm Mặt — bao gồm các tập dữ liệu và các bản tải xuống trung
— thay vào đó hãy sử dụng biến HF_HOME:

```
$ export HF_HOME="/path/to/cache_root"
```

Điều này dẫn đến:

- bộ nhớ đệm của bộ dữ liệu /path/to/cache_root/datasets
- bộ đệm trung tâm /path/to/cache_root/hub

Những sự khác biệt này đặc biệt hữu ích khi làm việc trong môi trường chia sẻ hoặc được nối mạng.
file systems (e.g., NFS).
Xem vấn đề #7480 để thảo luận về cách người dùng gặp phải các vị trí bộ đệm không mong muốn khi
HF_HUB_CACHE không được đặt cùng với HF_DATASETS_CACHE.

Khi tải tập dữ liệu, bạn cũng có tùy chọn thay đổi nơi lưu trữ dữ liệu.
Thay đổi tham số cache_dir thành đường dẫn bạn muốn:

```
>>> from datasets import load_dataset
>>> dataset = load_dataset('username/dataset', cache_dir="/path/to/another/directory/datasets")
```

Chế độ tải xuống

After you download a dataset, control how it is loaded by load_dataset() with the
tham số download_mode. Theo mặc định, Bộ dữ liệu sẽ sử dụng lại tập dữ liệu nếu nó tồn tại. Nhưng nếu bạ
cần tập dữ liệu gốc mà không áp dụng bất kỳ chức năng xử lý nào, hãy tải xuống lại các tệp dưới dạng
hiển thị dưới đây:

```
>>> from datasets import load_dataset
>>> dataset = load_dataset('rajpurkar/squad', download_mode='force_redownload')
```

Tham khảo DownloadMode để biết danh sách đầy đủ các chế độ tải xuống.

Tệp bộ nhớ đệm

Clean up the Arrow cache files in the directory with Dataset.cleanup_cache_files():

```
# Trả về số lượng file cache đã bị xóa
>>> dataset.cleanup_cache_files()
2
```

Bật hoặc tắt bộ nhớ đệm

Nếu bạn đang sử dụng tệp được lưu trong bộ nhớ đệm cục bộ, nó sẽ tự động tải lại tập dữ liệu với bất kỳ tệp
các biến đổi bạn đã áp dụng cho tập dữ liệu. Vô hiệu hóa hành vi này bằng cách đặt đối số
load_from_cache_file=False in Dataset.map():

```
>>> updated_dataset = small_dataset.map(add_prefix, load_from_cache_file=False)
```

Trong ví dụ trên, Bộ dữ liệu sẽ thực thi hàm add_prefix trên toàn bộ
tập dữ liệu một lần nữa thay vì tải tập dữ liệu từ trạng thái trước đó.

Disable caching on a global scale with disable_caching():

```
>>> from datasets import disable_caching
>>> disable_caching()
```

Khi bạn tắt bộ nhớ đệm, Bộ dữ liệu sẽ không tải lại các tệp đã lưu trong bộ nhớ đệm khi áp dụng nữa
chuyển đổi thành tập dữ liệu. Bất kỳ biến đổi nào bạn áp dụng trên tập dữ liệu của mình sẽ cần phải được áp

> [!TIP]
> Nếu bạn muốn sử dụng lại tập dữ liệu từ đầu, hãy thử đặt tham số download_mode trong
> load_dataset() instead.

Cải thiện hiệu suất

Việc tắt bộ đệm và sao chép tập dữ liệu vào bộ nhớ sẽ tăng tốc hoạt động của tập dữ liệu. Có hai tùy chọn để sao chép tập dữ liệu vào bộ nhớ:

1. Set datasets.config.IN_MEMORY_MAX_SIZE to a nonzero value (in bytes) that fits in your Bộ nhớ RAM.
2. Đặt biến môi trường HF_DATASETS_IN_MEMORY_MAX_SIZE thành giá trị khác 0. Ghi chú rằng phương pháp đầu tiên được ưu tiên cao hơn.