Tính năng bộ dữ liệu

Các tính năng xác định cấu trúc bên trong của tập dữ liệu. Nó được sử dụng để xác định cơ sở định dạng tuần tự hóa. Tuy nhiên, điều thú vị hơn với bạn là Tính năng chứa cấp độ cao thông tin về mọi thứ từ tên và loại cột cho đến ClassLabel. bạn có thể hãy coi Tính năng là xương sống của tập dữ liệu.

The Features format is simple: dict[column_name, column_type] . It is a dictionary of column cặp tên và loại cột. Kiểu cột cung cấp nhiều tùy chọn để mô tả loại dữ liệu bạn có.

Chúng ta hãy xem các tính năng của bộ dữ liệu MRPC từ điểm chuẩn GLUE:

```
>>> from datasets import load_dataset
>>> dataset = load_dataset('nyu-mll/glue', 'mrpc', split='train')
>>> dataset.features
{'idx': Value('int32'),
 'label': ClassLabel(names=['not_equivalent', 'equivalent']),
 'sentence1': Value('string'),
 'sentence2': Value('string'),
}
```

Tính năng Giá trị cho Bộ dữ liệu biết:

- Kiểu dữ liệu idx là int32 .
- Kiểu dữ liệu câu1 và câu2 là chuỗi.

Bộ dữ liệu hỗ trợ nhiều loại dữ liệu khác như bool , float32 và nhị phân chỉ để đặt tên một vài.

[!TIP]
Tham khảo Giá trị để biết danh sách đầy đủ các loại dữ liệu được hỗ trợ.

Tính năng ClassLabel thông báo cho Bộ dữ liệu cột nhãn chứa hai lớp. các các lớp được dán nhãn not_equivalent và tương đương. Nhãn được lưu dưới dạng số nguyên trong dataset. When you retrieve the labels, ClassLabel.int2str() and ClassLabel.str2int() carries out việc chuyển đổi từ giá trị số nguyên sang tên nhãn và ngược lại.

Nếu kiểu dữ liệu của bạn chứa danh sách các đối tượng thì bạn muốn sử dụng tính năng Danh sách. Hãy nhớ
Tập dữ liệu SQuAD?

```
>>> from datasets import load_dataset
>>> dataset = load_dataset('rajpurkar/squad', split='train')
>>> dataset.features
{'id': Value('string'),
 'title': Value('string'),
 'context': Value('string'),
 'question': Value('string'),
 'answers': {'text': List(Value('string')),
  'answer_start': List(Value('int32'))}}
```

Trường câu trả lời được xây dựng bằng cách sử dụng mệnh đề của các tính năng vì và chứa hai trường con,
text và answer_start , tương ứng là danh sách chuỗi và int32 .

[!TIP]
Xem phần làm phẳng để tìm hiểu cách bạn có thể trích xuất các trường con lồng nhau thành trường con c
cột độc lập.

Kiểu tính năng mảng rất hữu ích để tạo các mảng có kích thước khác nhau. Bạn có thể tạo mảng với
hai chiều bằng Array2D và thậm chí cả mảng có năm chiều bằng Array5D.

```
>>> features = Features({'a': Array2D(shape=(1, 3), dtype='int32')})
```

Kiểu mảng cũng cho phép chiều thứ nhất của mảng là động. Điều này hữu ích cho
xử lý các chuỗi có độ dài thay đổi như câu mà không cần phải đệm hoặc cắt bớt
đầu vào thành một hình dạng đồng nhất.

```
>>> features = Features({'a': Array3D(shape=(None, 5, 2), dtype='int32')})
```

Tính năng âm thanh

Bộ dữ liệu âm thanh có một cột có loại Âm thanh, chứa ba trường quan trọng:

- mảng : dữ liệu âm thanh được giải mã được biểu diễn dưới dạng mảng 1 chiều.

- path : đường dẫn tới file âm thanh đã tải về.
- samples_rate : tốc độ lấy mẫu của dữ liệu âm thanh.

Khi bạn tải tập dữ liệu âm thanh và gọi cột âm thanh, tính năng Âm thanh sẽ tự động giải mã và lấy mẫu lại tệp âm thanh:

```
>>> from datasets import load_dataset, Audio

>>> dataset = load_dataset("PolyAI/minds14", "en-US", split="train")
>>> dataset[0]["audio"]
<datasets.features._torchcodec.AudioDecoder object at 0x11642b6a0>
```

[!WARNING]
Lập chỉ mục vào tập dữ liệu âm thanh bằng cách sử dụng chỉ mục hàng trước rồi đến cột âm thanh - dataset[0]["audio"] - to avoid decoding and resampling all the audio files in the dataset.
Mặt khác, đây có thể là một quá trình chậm và tốn thời gian nếu bạn có một tập dữ liệu lớn.

With decode=False , the Audio type simply gives you the path or the bytes of the audio file, mà không giải mã nó thành một đối tượng AudioDecoding torchcodec,

```
>>> dataset = load_dataset("PolyAI/minds14", "en-US", split="train").cast_column("audio", Audio(de
>>> dataset[0]
{'audio': {'bytes': None,
    'đường dẫn': '/root/.cache/huggingface/datasets/downloads/extracted/f14948e0e84be638dd7943ac365
  'english_transcription': 'Tôi muốn thiết lập một tài khoản chung với đối tác của mình',
  'ý định_lớp': 11,
  'lang_id': 4,
  'đường dẫn': '/root/.cache/huggingface/datasets/downloads/extracted/f14948e0e84be638dd7943ac3651
  'transcription': 'I would like to set up a joint account with my partner'}
```

Tính năng hình ảnh

Bộ dữ liệu hình ảnh có một cột có loại Hình ảnh, tải các đối tượng PIL.Image từ hình ảnh được lưu trữ dưới dạng byte:

Khi bạn tải tập dữ liệu hình ảnh và gọi cột hình ảnh, tính năng Hình ảnh sẽ tự động giải mã tập tin hình ảnh:

```
>>> from datasets import load_dataset, Image

>>> dataset = load_dataset("AI-Lab-Makerere/beans", split="train")
>>> dataset[0]["image"]
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x500 at 0x125506CF8>
```

[!WARNING]
Lập chỉ mục vào tập dữ liệu hình ảnh bằng cách sử dụng chỉ mục hàng trước rồi đến cột hình ảnh - dataset[0]["image"] - to avoid decoding all the image files in the dataset. Otherwise, this có thể là một quá trình chậm và tốn thời gian nếu bạn có một tập dữ liệu lớn.

With decode=False , the Image type simply gives you the path or the bytes of the image file, mà không giải mã nó thành PIL.Image ,

```
>>> dataset = load_dataset("AI-Lab-Makerere/beans", split="train").cast_column("image", Image(deco
>>> dataset[0]["image"]
{'bytes': None,
 'đường dẫn': '/Users/tên người dùng/.cache/huggingface/datasets/downloads/extracted/772e7c1fba622c
```

Tùy thuộc vào tập dữ liệu, bạn có thể nhận được đường dẫn đến hình ảnh được tải xuống cục bộ hoặc nội dung của hình ảnh dưới dạng byte nếu tập dữ liệu không được tạo từ các tệp riêng lẻ.

Bạn cũng có thể xác định tập dữ liệu hình ảnh từ các mảng có nhiều mảng:

```
>>> ds = Dataset.from_dict({"i": [np.zeros(shape=(16, 16, 3), dtype=np.uint8)]}, features=Features
```

And in this case the numpy arrays are encoded into PNG (or TIFF if the pixels values precision is important).

Đối với các mảng đa kênh như RGB hoặc RGBA, chỉ hỗ trợ uint8. Nếu bạn sử dụng kích thước lớn hơn chính xác, bạn sẽ nhận được cảnh báo và mảng bị giảm xuống uint8.
Đối với hình ảnh có thang độ xám, bạn có thể sử dụng độ chính xác số nguyên hoặc số float mà bạn muốn m tương thích với Gối. Cảnh báo được hiển thị nếu số nguyên hoặc độ chính xác nổi của hình ảnh của bạn quá và trong trường hợp này, mảng bị hạ cấp: một mảng int64 được hạ cấp xuống int32 và float64 mảng được hạ cấp xuống float32.