Package 'torchdatasets'

June 20, 2024

Title Ready to Use Extra Datasets for Torch

Version 0.3.1

Description Provides datasets in a format that can be easily consumed by torch 'dataloaders'. Handles data downloading from multiple sources, caching and pre-processing so users can focus only on their model implementations.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.1

Imports torch (>= 0.5.0), fs, zip, pins, torchvision, stringr, withr, utils

Suggests testthat, readr, coro, tokenizers, R.matlab

URL https://mlverse.github.io/torchdatasets/, https://github.com/mlverse/torchdatasets

BugReports https://github.com/mlverse/torchdatasets/issues

NeedsCompilation no

Author Daniel Falbel [aut, cre], RStudio [cph]

Maintainer Daniel Falbel <daniel@rstudio.com>

Repository CRAN

Date/Publication 2024-06-20 12:40:01 UTC

Contents

`bank_marketing_dataset` *Bank marketing dataset*

Description

Prepares the Bank marketing dataset available on UCI Machine Learning repository here The data is available publicly for download, there is no need to authenticate. Please cite the data as Moro et al., 2014 S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Usage

```
bank_marketing_dataset(
  root,
  split = "train",
  indexes = NULL,
  download = FALSE,
  with_call_duration = FALSE
)
```

Arguments

`root`	path to the data location
`split`	string. 'train' or 'submission'
`indexes`	set of integers for subsampling (e.g. 1:41188)
`download`	whether to download or not
`with_call_duration`	whether the call duration should be included as a feature. Could lead to leakage. Default: FALSE.

Value

A torch dataset that can be consumed with `torch::dataloader()`.

Examples

```
if (torch::torch_is_installed() && FALSE) {
bank_mkt <- bank_marketing_dataset("./data", download = TRUE)
length(bank_mkt)
}
```

bird_species_dataset *Bird species dataset*

Description

Downloads and prepares the 450 bird species dataset found on Kaggle. The dataset description, license, etc can be found here.

Usage

```
bird_species_dataset(root, split = "train", download = FALSE, ...)
```

Arguments

root	path to the data location
split	train, test or valid
download	wether to download or not
...	other arguments passed to torchvision::image_folder_dataset().

Value

A torch::dataset() ready to be used with dataloaders.

Examples

```
if (torch::torch_is_installed() && FALSE) {
birds <- bird_species_dataset("./data", token = "path/to/kaggle.json",
                              download = TRUE)
length(birds)
}
```

cityscapes_pix2pix_dataset *Cityscapes Pix2Pix dataset*

Description

Downloads and prepares the cityscapes dataset that has been used in the pix2pix paper.

Usage

```
cityscapes_pix2pix_dataset(
  root,
  split = "train",
  download = FALSE,
  ...,
  transform = NULL,
  target_transform = NULL
)
```

Arguments

root	path to the data location
split	train, test or valid
download	wether to download or not
...	Currently unused.
transform	A function/transform that takes in an PIL image and returns a transformed version. E.g, transform_random_crop().
target_transform	A function/transform that takes in the target and transforms it.

Details

Find more information in the project website

---

dogs_vs_cats_dataset *Dog vs cats dataset*

Description

Prepares the dog vs cats dataset available in Kaggle here

Usage

```
dogs_vs_cats_dataset(
  root,
  split = "train",
  download = FALSE,
  ...,
  transform = NULL,
  target_transform = NULL
)
```

Arguments

root	path to the data location
split	string. 'train' or 'submission'
download	whether to download or not
...	Currently unused.
transform	function that takes a torch tensor representing an image and return another tensor, transformed.
target_transform	function that takes a scalar torch tensor and returns another tensor, transformed.

Value

A torch::dataset() ready to be used with dataloaders.

Examples

```
if (torch::torch_is_installed() && FALSE) {
dogs_cats <- dogs_vs_cats_dataset("./data", token = "path/to/kaggle.json",
                                  download = TRUE)
length(dogs_cats)
}
```

---

guess_the_correlation_dataset

Guess The Correlation dataset

Description

Prepares the Guess The Correlation dataset available on Kaggle here A copy of this dataset is hosted in a public Google Cloud bucket so you don't need to authenticate.

Usage

```
guess_the_correlation_dataset(
  root,
  split = "train",
  transform = NULL,
  target_transform = NULL,
  indexes = NULL,
  download = FALSE
)
```

Arguments

root	path to the data location
split	string. 'train' or 'submission'
transform	function that takes a torch tensor representing an image and return another tensor, transformed.
target_transform	function that takes a scalar torch tensor and returns another tensor, transformed.
indexes	set of integers for subsampling (e.g. 1:140000)
download	whether to download or not

Value

A torch dataset that can be consumed with `torch::dataloader()`.

Examples

```
if (torch::torch_is_installed() && FALSE) {
gtc <- guess_the_correlation_dataset("./data", download = TRUE)
length(gtc)
}
```

---

imdb_dataset *IMDB movie review sentiment classification dataset*

Description

The format of this dataset is meant to replicate that provided by Keras.

Usage

```
imdb_dataset(
  root,
  download = FALSE,
  split = "train",
  shuffle = (split == "train"),
  num_words = Inf,
  skip_top = 0,
  maxlen = Inf,
  start_char = 2,
  oov_char = 3,
  index_from = 4
)
```

Arguments

root	path to the data location
download	wether to download or not
split	train, test or valid
shuffle	whether to shuffle or not the dataset. TRUE if split=="train"
num_words	Words are ranked by how often they occur (in the training set), and only the num_words most frequent words are kept. Any less frequent word will appear as oov_char value in the sequence data. If Inf, all words are kept. Defaults to None, so all words are kept.
skip_top	skip the top N most frequently occurring words (which may not be informative). These words will appear as oov_char value in the dataset. Defaults to 0, so no words are skipped.
maxlen	int or Inf. Maximum sequence length. Any longer sequence will be truncated. Defaults to Inf, which means no truncation.
start_char	The start of a sequence will be marked with this character. Defaults to 2, because 1 is usually the padding character.
oov_char	int. The out-of-vocabulary character. Words that were cut out because of the num_words or skip_top limits will be replaced with this character.
index_from	int. Index actual words with this index and higher.

---

oxford_flowers102_dataset

102 Category Flower Dataset

Description

The Oxford Flower Dataset is a 102 category dataset, consisting of 102 flower categories. The flowers chosen to be flower commonly occuring in the United Kingdom. Each class consists of between 40 and 258 images. The details of the categories and the number of images for each class can be found on this category statistics page.

Usage

```
oxford_flowers102_dataset(
  root,
  split = "train",
  target_type = c("categories"),
  download = FALSE,
  ...,
  transform = NULL,
  target_transform = NULL
)
```

Arguments

root	path to the data location
split	train, test or valid
target_type	Currently only 'categories' is supported.
download	wether to download or not
...	Currently unused.
transform	A function/transform that takes in an PIL image and returns a transformed version. E.g, transform_random_crop().
target_transform	A function/transform that takes in the target and transforms it.

Details

The images have large scale, pose and light variations. In addition, there are categories that have large variations within the category and several very similar categories. The dataset is visualized using isomap with shape and colour features.

You can find more info in the dataset webpage.

Note

The official splits leaves far too many images in the test set. Depending on your work you might want to create different train/valid/test splits.

---

oxford_pet_dataset *Oxford Pet Dataset*

Description

The Oxford-IIIT Pet Dataset is a 37 category pet dataset with roughly 200 images for each class. The images have a large variations in scale, pose and lighting. All images have an associated ground truth annotation of species (cat or dog), breed, and pixel-level trimap segmentation.

Usage

```
oxford_pet_dataset(
  root,
  split = "train",
  target_type = c("trimap", "species", "breed"),
  download = FALSE,
  ...,
  transform = NULL,
  target_transform = NULL
)
```

Arguments

`root`	path to the data location
`split`	train, test or valid
`target_type`	The type of the target: • 'trimap': returns a mask array with one class per pixel. • 'species': returns the species id. 1 for cat and 2 for dog. • 'breed': returns the breed id. see `dataset$breed_classes`.
`download`	wether to download or not
`...`	Currently unused.
`transform`	A function/transform that takes in an PIL image and returns a transformed version. E.g, `transform_random_crop()`.
`target_transform`	A function/transform that takes in the target and transforms it.

Index