FAST.AI DATASETS

In machine learning and deep learning we can’t do anything without data. So the people that create datasets for us to train our models are the (often under-appreciated) heros. Some of the most useful and important datasets are those that become important “academic baselines”; that is, datasets that are widely studied by researchers and used to compare algorithmic changes. Some of these become household names (at least, among households that train models!), such as *MNIST*, *CIFAR 10*, and *Imagenet*.

At fast.ai we (and our students) owe a debt of gratitude to those kind folks who have made datasets available for the research community. We’ve teamed up with AWS to try to give back a little: we’ve made some of the most important of these datasets available in a single place, using standard formats, on reliable and fast infrastructure (see below for a full list and links). If you use any of these datasets in your research, please give back by citing the original paper (we’ve provided the appropriate citation link below for each), and if you use them as part of a commercial or educational project, consider adding a note of thanks and a link to the dataset.

We use these datasets in our teaching, because they provide great examples of the kind of data that students are likely to encounter, and the academic literature has many examples of model results using these datasets which students can compare their work to. In addition, we also use datasets from Kaggle Competitions, because the public leaderboards on Kaggle allow students to test their models against the best in the world (the Kaggle datasets are not listed here).

For each dataset below, click the ‘source’ link to see the dataset license and details from the creator, the ‘cite’ link for the paper for citations, and the ‘download’ link to access to dataset from AWS Open Datasets.

IMAGE CLASSIFICATION

Source	Citation	Download	Description
MNIST	LeCun et al., 1998a	download	Classic dataset of small (28x28) handwritten grayscale digits, developed in the 1990s for testing the most sophisticated models of the day; today, often used as a basic “hello world” for introducing deep learning. This fast.ai datasets version uses a standard PNG format instead of the special binary format of the original, so you can use the regular data pipelines in most libraries; if you want to use just a single input channel like the original, simply pick a single slice from the channels axis.
CIFAR10	Krizhevsky, 2009	download	60000 32x32 colour images in 10 classes, with 6000 images per class (50000 training images and 10000 test images). Very widely used today for testing performance of new algorithms. This fast.ai datasets version uses a standard PNG format instead of the platform-specific binary formats of the original, so you can use the regular data pipelines in most libraries.

Source	Citation	Download	Description
CIFAR100	Krizhevsky, 2009	download	This dataset is just like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a “fine” label (the class to which it belongs) and a “coarse” label (the superclass to which it belongs).
Caltech-UCSD Birds-200-2011	Lin et al. 2015	download	An image dataset with photos of 200 bird species (mostly North American); it can also be used for localization. Number of categories: 200; Number of images: 11,788; Annotations per image: 15 Part Locations, 312 Binary Attributes, 1 Bounding Box
Caltech 101	L. Fei-Fei et al., 2004	download	Pictures of objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. The size of each image is roughly 300 x 200 pixels. Can also be used for localization.
Oxford-IIIT Pet	O. M. Parkhi et al., 2012	download	A 37 category pet dataset with roughly 200 images for each class. The images have a large variations in scale, pose and lighting. Can also be used for localization.
Oxford 102 Flowers	Nilsback, M-E. and Zisserman, A., 2008	download	A 102 category dataset consisting of 102 flower categories, commonly occuring in the United Kingdom. Each class consists of 40 to 258 images. The images have large scale, pose and light variations.
Food-101	Bossard, Lukas et al., 2014	download	101 food categories, with 101,000 images; 250 test images and 750 training images per class. The training images were not cleaned. All images were rescaled to have a maximum side length of 512 pixels.
Stanford cars	Jonathan Krause et al., 2013	download	16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split. Classes are typically at the level of Make, Model, Year.

NLP

Source	Citation	Download	Description
IMDb Large Movie Review Dataset	Andrew L. Maas et al., 2011	download	A dataset for binary sentiment classification containing 25,000 highly polarized movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well.

Source	Citation	Download	Description
Wikitext-103	Stephen Merity et al., 2016	download	A collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia. Widely used for language modeling, including the pretrained models used in the fastai library and ULMFiT algorithm.
Wikitext-2	Stephen Merity et al., 2016	download	A subset of Wikitext-103; useful for testing language model training on smaller datasets.
WMT 2015 French/English parallel texts	Callison-Burch et al., 2009	download	French/English parallel texts for training translation models. Over 20 million sentences in French and English. Dataset created by Chris Callison-Burch, who crawled millions of web pages and then used a set of simple heuristics to transform French URLs onto English URLs, and assumed that these documents are translations of each other.
AG News	Xiang Zhang et al., 2015	download	496,835 categorized news articles from >2000 news sources from the 4 largest classes from AG's corpus of news articles, using only the title and description fields. The number of training samples for each class is 30,000 and testing 1900.
Amazon reviews - Full	Xiang Zhang et al., 2015	download	34,686,770 Amazon reviews from 6,643,669 users on 2,441,053 products, from the Stanford Network Analysis Project (SNAP). This full dataset contains 600,000 training samples and 130,000 testing samples in each class.
Amazon reviews - Polarity	Xiang Zhang et al., 2015	download	34,686,770 Amazon reviews from 6,643,669 users on 2,441,053 products, from the Stanford Network Analysis Project (SNAP). This subset contains 1,800,000 training samples and 200,000 testing samples in each polarity sentiment.
DBPedia ontology	Xiang Zhang et al., 2015	download	40,000 training samples and 5,000 testing samples from 14 nonoverlapping classes from DBpedia 2014.
Sogou news	Xiang Zhang et al., 2015	download	2,909,551 news articles from the SogouCA and SogouCS news corpora, in 5 categories. The number of training samples selected for each class is 90,000 and testing 12,000. Note that the Chinese characters have been converted to Pinyin.

Source	Citation	Download	Description
Yahoo! Answers	Xiang Zhang et al., 2015	download	The 10 largest main categories from the Yahoo! Answers Comprehensive Questions and Answers version 1.0 dataset. Each class contains 140,000 training samples and 5,000 testing samples.
Yelp reviews - Full	Xiang Zhang et al., 2015	download	1,569,264 samples from the Yelp Dataset Challenge 2015. This full dataset has 130,000 training samples and 10,000 testing samples in each star.
Yelp reviews - Polarity	Xiang Zhang et al., 2015	download	1,569,264 samples from the Yelp Dataset Challenge 2015. This subset has 280,000 training samples and 19,000 test samples in each polarity.

IMAGE LOCALIZATION

Source	Citation	Download	Description
Camvid: Motion-based Segmentation and Recognition Dataset	Brostow et al., 2008	download	Segmentation dataset with per-pixel semantic segmentation of over 700 images, each inspected and confirmed by a second person for accuracy.
PASCAL Visual Object Classes (VOC)	Everingham, M et al., 2010	download	Standardised image data sets for object class recognition - both 2007 and 2012 versions are provided here. The 2012 version has 20 classes. The train/val data has 11,530 images containing 27,450 ROI annotated objects and 6,929 segmentations.

COCO

Probably the most widely used dataset today for object localization is COCO: Common Objects in Context. Provided here are all the files from the 2017 version, along with an additional *subset* dataset created by fast.ai. Details of each COCO dataset is available from the COCO dataset page. The fast.ai subset contains all images that contain one of five selected categories, restricting objects to just those five categories; the categories are: chair couch tv remote book vase.

- fast.ai subset
- Train images
- Val images
- Test images
- Unlabeled images
- Testing Image info
- Unlabeled Image info

- Train/Val annotations
- Stuff Train/Val annotations
- Panoptic Train/Val annotations