

Contents



- Frameworks and APIs

- Pytorch
- Theano
- Tensorflow
- CNTK
- Keras
- Chainer
- Torch
- Caffe

- Datasets

- WordNet
- ImageNet
- MNIST
- ActivityNet
- COCO
- OpenImages
- Kaggle
- Reuters Corpora, ASU, ...

Machine Learning Frameworks and APIs

What gets the job done

PyTorch

Python framework for machine learning

Supports distributed training

Supports GPU

<http://pytorch.org/>

<https://github.com/pytorch/pytorch>



Theano

Python library for mathematical computing

Supports multidimensional arrays and symbolic differentiation

Supports GPU

<http://www.deeplearning.net/software/theano/>

<https://github.com/Theano/Theano>



theano

TensorFlow

End-to-end open-source machine learning platform

Originated at Google

Supports all platforms (including Google TPU)

<https://www.tensorflow.org>  **TensorFlow**

<https://github.com/tensorflow/tensorflow>

CNTK

Part of Microsoft Cognitive Toolkit

Supports feed-forward DNNs, CNNs, and RNNs/LSTMs

Implements SGD with backpropagation learning

Supports multiple GPUs

<https://cntk.ai>

<https://github.com/Microsoft/cntk>



Keras

High-level neural networks API written in Python

Can run on top of TensorFlow, CNTK, or Theano

Supports CNN and RNN models

Supports GPUs

<https://keras.io/>

<https://github.com/keras-team>



Chainer

A Powerful, Flexible, and Intuitive Framework for Neural Networks

Runs on multiple GPUs easily with very little coding

Supports feed-forward nets, convnets, recurrent nets, and recursive nets

<https://chainer.org>



Torch

Scientific computing framework with ML support

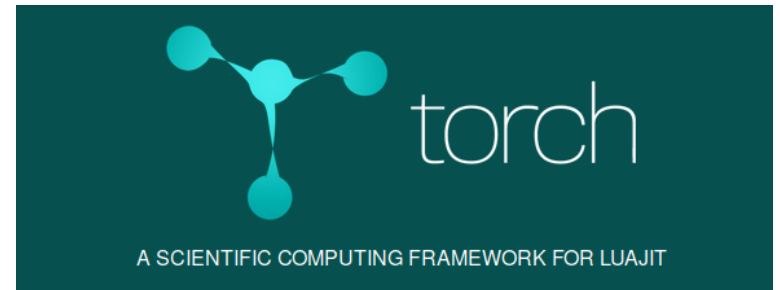
Extremely fast multidimensional linear algebra

Supports CNN, RNN, RL, and energy based models

Supports GPUs

<http://torch.ch/>

<https://github.com/torch/torch7>



Caffe

From Berkeley AI Research

Supports extremely fast CNN, RNN

Supports GPU

<http://caffe.berkeleyvision.org/>

<https://github.com/BVLC/caffe>



Caffe

Deep learning framework

Wikipedia

Long list of ML tools with feature comparisons

https://en.wikipedia.org/wiki/Comparison_of_deep-learning_software



Machine Learning Data Sets

Where is the data to train my AI

WordNet

Lexical database of English nouns, verbs, adjectives, and adverbs grouped into synsets

Synsets linked by conceptual-semantic and lexical relations

155,327 words, 175,979 synsets, 207,016 word-sense pairs

<https://wordnet.princeton.edu/>

WordNet

A Lexical Database for English

ImageNet

Image database organized according to the WordNet hierarchy (currently nouns only)

14,197,122 images, 21841 synsets indexed

One of the most widely used datasets for visual ML

<http://www.image-net.org/>



IMAGENET

MNIST

Database of images of handwritten digits

Size-normalized and centered on 28x28

60,000 images training set, 10,000 images test set

<http://yann.lecun.com/exdb/mnist/>

THE MNIST DATABASE
of handwritten digits

ActivityNet

Large collection of video snippets covering wide range of human activities in their daily life

200 classes, 100 videos per class, 1.54 activity instances per video, total 648 hours of video

<http://activity-net.org/>

<https://github.com/activitynet/ActivityNet>

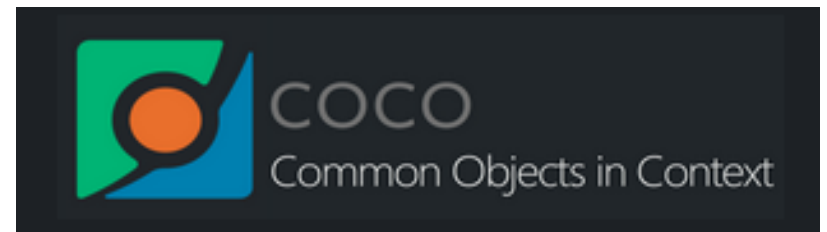


COCO

Large-scale object detection, segmentation, and captioning dataset

330K images (> 200K labeled), 1.5 million object instances, 80 object categories, 91 stuff categories, 5 captions per image, 250,000 people with keypoints

<http://cocodataset.org/>



OpenImages

15,851,536 boxes on 600 categories



Open Images

2,785,498 instance segmentations on 350 categories

36,464,560 image-level labels on 19,959 categories

391,073 relationship annotations of 329 relationships

478,000 crowdsourced images with 6,000+ categories

<https://storage.googleapis.com/openimages/web/index.html>

Kaggle

20,815 datasets

In a variety of real-life problem domains

Including many anonymized real-life datasets

<https://www.kaggle.com/datasets>



Reuters Corpora

RCV1 – 810,000 English language news stories

RCV2 – 487,000 news stories in thirteen languages

TRC2 – 1,800,370 news stories

<https://trec.nist.gov/data/reuters/reuters.html>



ASU Data Sets



Text-oriented datasets on tweets, blogs, Digg, Delicio.us, Last.fm, Youtube, ... with millions of entries

ASU Twitter dataset – 11,316,811 users, 85,331,846 connections

<http://socialcomputing.asu.edu/pages/datasets>



Wikipedia

Long list of datasets in every possible domain

[https://en.wikipedia.org/wiki/](https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)

[List_of_datasets_for_machine-learning_research](https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)



What Does One Need To Get Started in Machine Learning

Trust your rusty little computer

Less Is More

A regular personal computer

A reasonable multi-core CPU

A reasonable amount of RAM

A simple GPU

GNU Linux (Ubuntu, Debian, Fedora, ...)

Internet

That's all!