Data sets in Machine/Deep learning

- In deep learning, there are many benchmark data sets
- Researchers and practitioners use the data sets to evaluate and to develop their goal
- Here, we introduce some of the famous data sets in different domains, including image, text, speech, and time series

- **Some sources from (with great appreciation and acknowledgements)**
  - https://machinelearningmastery.com/

Image type data sets:

- **MNIST**
  - http://yann.lecun.com/exdb/mnist/
  - Purpose: Object classification
  - The MNIST database of handwritten digits, available from the above link, has a training set of 60,000 examples, and a test set of 10,000 examples (The examples are 28×28 gray-scale 2-D image of digits 0 t0 10).

- **Fashion-MNIST**
  - https://github.com/zalandoresearch/fashion-mnist
  - Purpose: Object classification
  - A training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 gray-scale image, associated with a label from 10 classes (T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot).

Image type data sets:

- **The Quick, Draw!**
  - ❑https://github.com/googlecreativelab/quickdraw-dataset
  - ❑Purpose: Object classification
  - ❑The Quick Draw Dataset is a collection of 50 million drawings across 345 categories, contributed by players of the game Quick, Draw!. The drawings were captured as timestamped vectors, tagged with metadata including what the player was asked to draw and in which country the player was located.
- **CIFAR**
- https://www.cs.toronto.edu/~kriz/cifar.html
  - Purpose: Object classification
  - **The CIFAR-10 dataset**
    - The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.
  - **The CIFAR-10 dataset**
    - 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.

Image type data sets:

- **The Street View House Numbers (SVHN) Dataset**
  - ❑http://ufldl.stanford.edu/housenumbers/
  - ❑Purpose: Object classification
  - ❑It can be seen as similar in flavor to MNIST (e.g., the images are of small cropped digits), but incorporates an order of magnitude more labeled data (over 600,000 digit images) and comes from a significantly harder, unsolved, real world problem (recognizing digits and numbers in natural scene images). SVHN is obtained from house numbers in Google Street View images.
  - ❑ Typical use: MNIST-like 32-by-32 images centered around a single character (many of the images do contain some distractors at the sides).
  - ❑73257 digits for training, 26032 digits for testing (531131 additional).
  - ❑10 classes corresponding to the 10 classes.

- **ImageNet**
  - http://www.image-net.org/
  - Purpose: Object classification, object localization, and object detection
  - It is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images.
  - Extremely huge data sets (14,197,122 images)

Image type data sets:

- **Common Object in Context (COCO)**
  - ❑http://cocodataset.org/#home
  - ❑Purpose: object detection, object segmentation, and captioning
  - ❑123,287 images, 886,284 instances

- **Large-scale CelebFaces Attributes (CelebA)**
  - http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
  - ❑Purpose: face attribute recognition, face detection, landmark localization, and face editing & synthesis.
  - ❑A large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter. CelebA has large diversities, large quantities, and rich annotations, including
    - ➢ **10,177** number of **identities**
    - ➢ **202,599** number of **face images**
    - ➢ **5 landmark locations**, **40 binary attributes** annotations per image

Text type data sets in Natural Language Processing:

- **Reuters-21578 Text Categorization Collection**
  - http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
  - Purpose: Text classification and sentiment analysis
  - It is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories.

- **Reuters Corpora (RCV1, RCV2, TRC2)**
  - https://trec.nist.gov/data/reuters/reuters.html
  - Purpose: Text classification and sentiment analysis
  - This corpus, known as "Reuters Corpus, Volume 1" or RCV1, is significantly larger than the older, well-known Reuters-21578 collection heavily used in the text classification community.

Text type data sets in Natural Language Processing:

- **Large Movie Review Dataset**
  - http://ai.stanford.edu/~amaas/data/sentiment/
  - Purpose: Binary sentiment classification
  - It consists a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided.

- **News Group Movie Review**
  - ❑http://www.cs.cornell.edu/people/pabo/movie-review-data/
  - ❑Purpose: Sentiment Classification
  - ❑A collection of movie reviews from the website imdb.com and their positive or negative sentiment

Text type data sets in Natural Language Processing:

- **Project Gutenberg**
  - https://www.gutenberg.org/
  - Purpose: Language modeling (developing a statistical model for predicting the next word in a sentence or next letter in a word given whatever has come before).
  - A large collection of free books that can be retrieved in plain text for a variety of languages

- **Brown University Standard Corpus of Present-Day American English**
  - https://en.wikipedia.org/wiki/Brown_Corpus
  - Purpose: Language modeling
  - A large sample of English words

Text type data sets in Natural Language Processing:

- **The PASCAL Object Recognition Database Collection**
  - ❑http://host.robots.ox.ac.uk/pascal/VOC/databases.html
  - ❑Purpose: Image captioning
  - ❑The dataset has 20 classes, including aeroplane, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, train, TV.

- **Aligned Hansards of the 36th Parliament of Canada**
- https://www.isi.edu/natural-language/download/hansard/
  - ❑Purpose: Machine translation
  - ❑Pairs of sentences in English and French
  - ❑1.3 million pairs of aligned text chunks (sentences or smaller fragments) from the official records (*Hansards*) of the 36th Canadian Parliament

Text type data sets in Natural Language Processing:

- **Stanford Question Answering Dataset (SQuAD)**
  - https://rajpurkar.github.io/SQuAD-explorer/
  - Purpose: Question Answering
  - It is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

- **Deepmind Question Answering Corpus**
  - https://github.com/deepmind/rc-data
  - Purpose: Question Answering
  - Question answering about news articles from the Daily Mail. It contains a script to generate question/answer pairs using CNN and Daily Mail articles downloaded from the Wayback Machine.

Speech type data sets

- **TIMIT Acoustic-Phonetic Continuous Speech Corpus**
  - https://catalog.ldc.upenn.edu/LDC93S1
  - Purpose: Speech recognition (transforming audio of a spoken language into human readable text)
  - Not free, but listed because of its wide use. Spoken American English and associated transcription. It contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance.

- **LibriSpeech ASR corpus**
  - http://www.openslr.org/12/
  - Purpose: Speech recognition
  - Large collection of English audiobooks taken from LibriVox.
  - 1000 hours corpus of read English speech

Time series type data sets

- **Monthly Sunspot Dataset**
  - Purpose: Prediction (Univariate time series)
  - https://raw.githubusercontent.com/jbrownlee/Datasets/master/monthly-sunspots.csv
  - This dataset describes a monthly count of the number of observed sunspots for just over 230 years (1749-1983). The units are a count and there are 2,820 observations. The source of the dataset is credited to Andrews & Herzberg (1985).

- **Daily Female Births Dataset**
  - https://raw.githubusercontent.com/jbrownlee/Datasets/master/daily-total-female-births.csv
  - Purpose: Prediction (Univariate time series)
  - This dataset describes the number of daily female births in California in 1959. The units are a count and there are 365 observations. The source of the dataset is credited to Newton (1988).

Time series type data sets

- **EEG Eye State Data Set**
  - ❑http://archive.ics.uci.edu/ml/datasets/EEG+Eye+State
  - ❑Purpose: Classification predictive modeling (Multivariate time series)
  - ❑This dataset describes EEG data for an individual and whether their eyes were open or closed. There are a total of 14,980 observations and 15 input variables. The class value of ‘1’ indicates the eye-closed and ‘0’ the eye-open state.

- **Ozone Level Detection Dataset**
  - ❑http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection
  - ❑Purpose: Classification predictive modeling (Multivariate time series)
  - ❑This dataset describes 6 years of ground ozone concentration observations and the objective is to predict whether it is an “ozone day” or not.
  - ❑The dataset contains 2,536 observations and 73 attributes.

Source of machine/deep learning data sets

- There are still many data sets freely available
- Useful resources in the web:
  - UCI machine learning repository
    - http://archive.ics.uci.edu/ml/index.php
  - NLTK corpora
    - http://www.nltk.org/nltk_data/
  - Stanford NLP collection
    - https://nlp.stanford.edu/links/statnlp.html#Corpora
  - Torchvision image data sets:
    - https://pytorch.org/docs/stable/torchvision/datasets.html#cifar
  - Kaggle data sets
    - https://www.kaggle.com/datasets?utm_medium=paid&utm_source=google.com+search&utm_campaign=datasets&&gclid=Cj0KCQjwv8nqBRDGARIsAHfR9wBNRGQCjxnoypMTh4q7TI9OA3NtBmp9fyyaD6IGquQuaSgxc5wgvFYaAswCEALw_wcB