INFR10086 Machine Learning (MLG) Semester 2, 2023/4

Tutorial 5: Data sets

In this tutorial, we will get to “feel” the scale of machine learning by looking a few commonly used data sets.

ImageNet

ImageNet[1] is one of the large image data sets. It is well known for the rise of deep learning, and still how researchers test their computer vision models. The most used subset (on Kaggle[2]) contains about 1.2 million images.

Discussion. The entire ImageNet on Kaggle is 156 G. If you network’s bandwidth is 25 M per second. How long does it take just to download the data set?

Since it will take too long to download tht entire data set and most likely we don’t have a spare 156 G of disk space, we will look at a few sample files. Download

```
https://homepages.inf.ed.ac.uk/htang2/mlg2023/tutorial-5/imagenet-samples.tar.gz
```

Untar the file and browse the directory a little. Below is how the data set is structured.

```
imagenet-samples/
|-- ILSVRC
|   `-- Data
|       `-- CLS-LOC
|           `-- train
|               |-- n02017213
|               |   `-- n02017213_7894.JPEG
|               |-- n02091134
|               |   `-- n02091134_12759.JPEG
|               `-- n04447861
|                   `-- n04447861_2895.JPEG
|-- LOC_synset_mapping.txt
`-- LOC_train_solution.csv
```

[1] https://image-net.org

[2] https://www.kaggle.com/c/imagenet-object-localization-challenge/overview/description

Discussion.

- What do you see in `LOC_synset_mapping.txt`? What do you think this file is for?
- What do you see in `LOC_train_solution.csv`? What do you think this file is for?
- What do you see in the three JPEG files?
- Could you find the entry `n02017213_7894` in `LOC_synset_mapping.txt` and `LOC_train_solution.csv`? What about `n02091134_12759` and `n04447861_2895`?

Discussion.

- What do you think ImageNet is for?
- Suppose you want to build a model that takes an image and outputs the result that you see in `LOC_train_solution.csv`. Is this a regression task, a classification task, or something else?

LibriSpeech

LibriSpeech[3] is a data set consisting of 1000 hours of read English speech. Similar to ImageNet, it is one of the large speech data set, and is continued to be used by the speech community for nearly a decade. The entire data set is about 60 G, and will take again take a while to download.

Instead, download

`https://homepages.inf.ed.ac.uk/htang2/mlg2023/tutorial-5/librispeech-samples.tar.gz`

Untar the file and browse the directory a little. Below is how the data set is structured.

```
librispeech-samples/
|-- BOOKS.TXT
|-- CHAPTERS.TXT
|-- SPEAKERS.TXT
`-- train-clean-100
    |-- 103
    |   `-- 1240
    |       |-- 103-1240-0000.flac
    |       |-- 103-1240-0001.flac
    |       |-- 103-1240-0002.flac
    |       `-- 103-1240.trans.txt
    `-- 909
        `-- 131041
            |-- 909-131041-0000.flac
            |-- 909-131041-0001.flac
            |-- 909-131041-0002.flac
```

[3] `https://www.openslr.org/12/`

```
            `-- 909-131041.trans.txt
```

Discussion.

- What are in the first few lines of `103-1240.trans.txt`?
- What do you hear in `103-1240-0000.flac` and `103-1240-0001.flac`?

Discussion.

- What do you think LibriSpeech is for?
- Suppose you want to build a model that takes an audio recording and outputs the result that you see in `103-1240.trans.txt`. Is this a regression task, a classification task, or something else?

`enwik9`

The `enwik9` is one text data set of decent size. It contains exactly 1 billion bytes, and the zip version is about 308 M. It is also simple to use—it is just a single file. The link below contains a few samples from `enwik9`.

`https://homepages.inf.ed.ac.uk/htang2/mlg2023/tutorial-5/enwik9-samples`

Discussion.

- What are in the first few lines of `enwik9-samples`?
- What do you think `enwik9` is for?
- Is the task you have in mind a regression task, a classification task, or something else?
- How should we evaluate the model and declare success?

Europarl

The Europarl[4] data set is a text data set extracted from the proceedings of European Parliament. The data set is about 1.5 G in size. Again, we won't be able to look at the entire data set. The link below contains a few samples from Europarl.

`https://homepages.inf.ed.ac.uk/htang2/mlg2023/tutorial-5/europarl-v7-samples.fr-en.en`
`https://homepages.inf.ed.ac.uk/htang2/mlg2023/tutorial-5/europarl-v7-samples.fr-en.fr`

[4] `https://www.statmt.org/europarl/`

Discussion.

- What are in the first few lines of `europarl-v7-samples.fr-en.en`?
- What are in the first few lines of `europarl-v7-samples.fr-en.fr`?
- What do you think Europarl is for?
- Suppose you want to build a model that takes in `europarl-v7-samples.fr-en.en` and outputs the result in `europarl-v7-samples.fr-en.fr`. Is this a regression task, a classification task, or something else?
- How should we evaluate the model and declare success?