Datasets and Dataloaders

Rowel Atienza, PhD

University of the Philippines

github.com/roatienza

2022

"The temptation to form premature theories upon insufficient data is the bane of our profession." - *Sherlock Holmes*

Outline

Dataset (Collection, Labelling, Data Structure)

Dataloader (APIs)

Datasets

$$\mathcal{D} = \{\boldsymbol{x}, \boldsymbol{y}\},$$
$$\boldsymbol{x} \in \mathbb{R}^{M \times N \times \cdots}, \boldsymbol{y} \in \mathbb{R}^{S \times T \times \cdots}$$

Stores the samples and their corresponding labels

Data Sources

Audio, Speech

Image, Video

Sensors

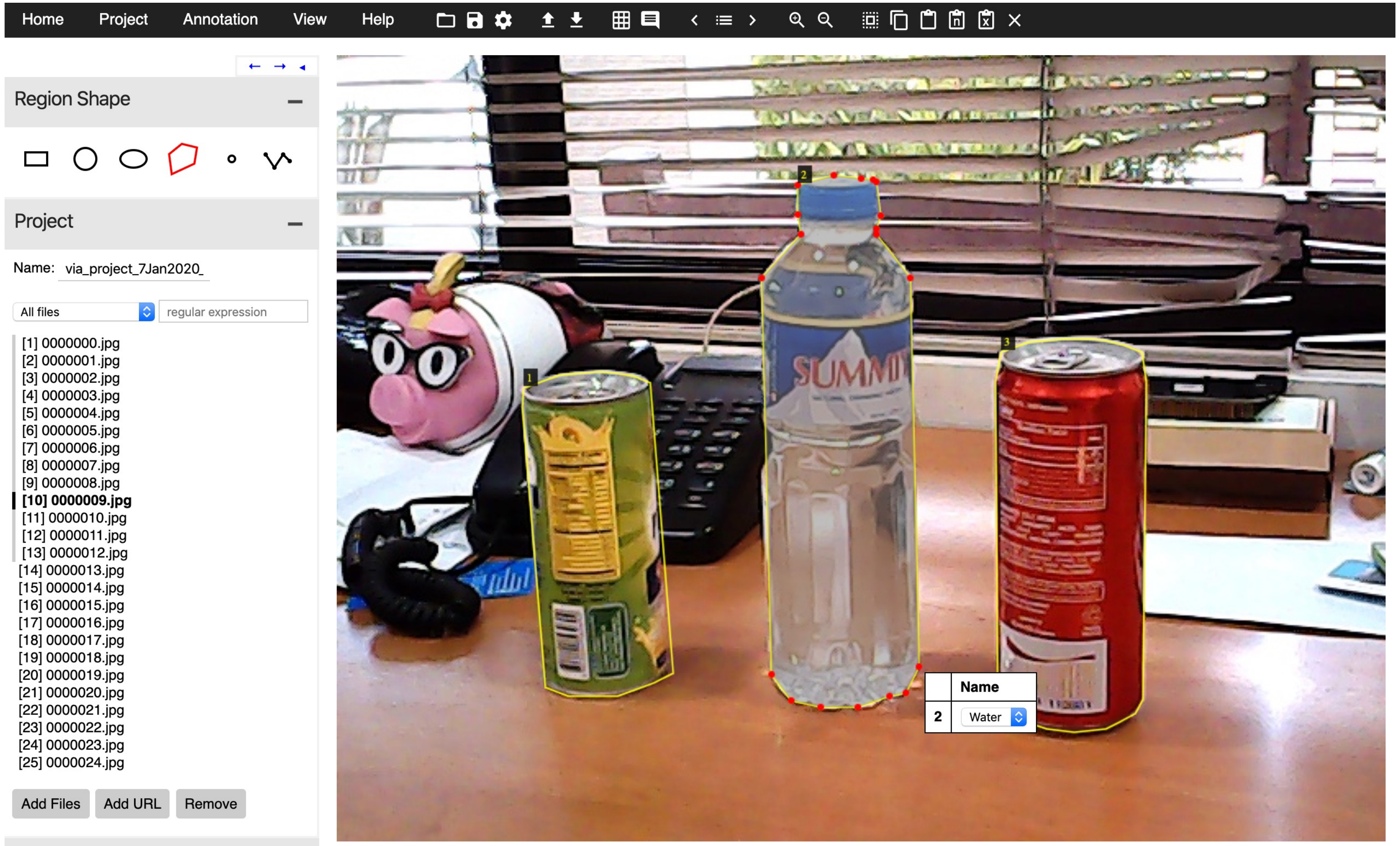
Internet

Business, Govt

Databases

Data Labelling or Annotation

VIA Annotation Tool



Name: via_project_7Jan2020_

All files | regular expression

[1] 0000000.jpg
[2] 0000001.jpg
[3] 0000002.jpg
[4] 0000003.jpg
[5] 0000004.jpg
[6] 0000005.jpg
[7] 0000006.jpg
[8] 0000007.jpg
[9] 0000008.jpg
[10] 0000009.jpg
[11] 0000010.jpg
[12] 0000011.jpg
[13] 0000012.jpg
[14] 0000013.jpg
[15] 0000014.jpg
[16] 0000015.jpg
[17] 0000016.jpg
[18] 0000017.jpg
[19] 0000018.jpg
[20] 0000019.jpg
[21] 0000020.jpg
[22] 0000021.jpg
[23] 0000022.jpg
[24] 0000023.jpg
[25] 0000024.jpg

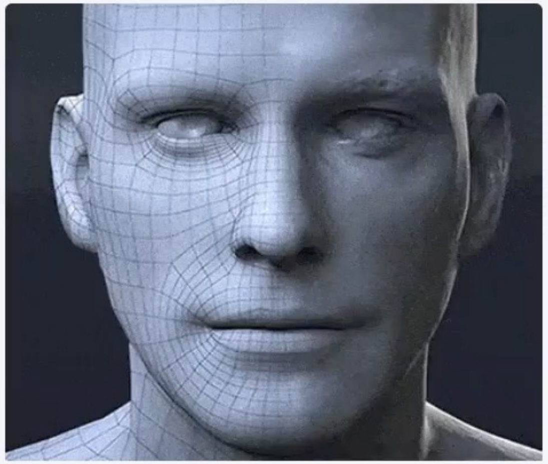
	Name
2	Water

https://www.robots.ox.ac.uk/~vgg/software/via/

COGITO

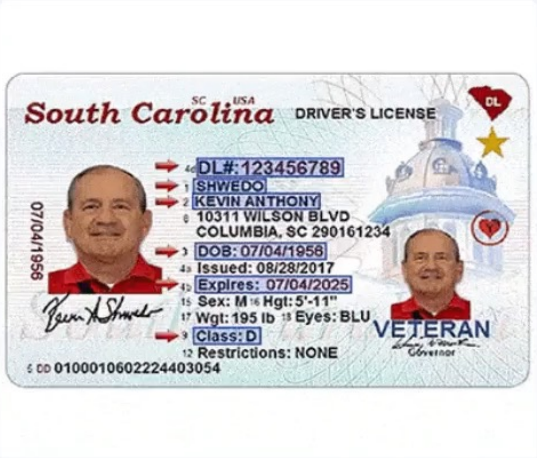
2D Bounding Boxes

Cuboid



Point & Landmark

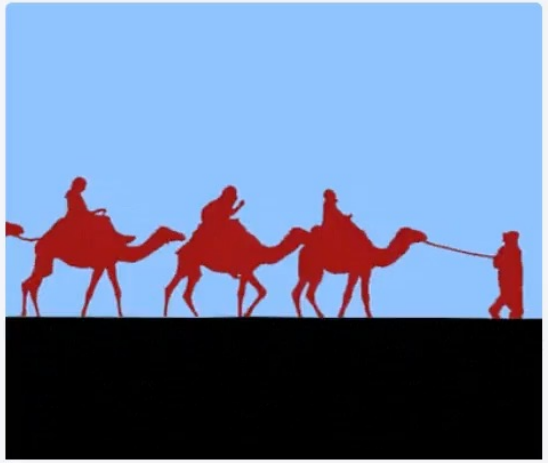
Lines & Splines



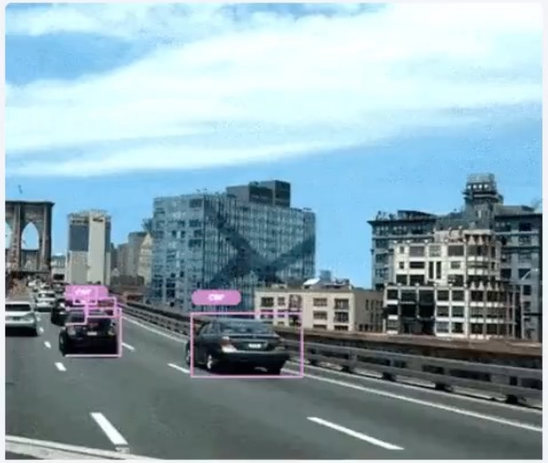
SC USA
South Carolina DRIVER'S LICENSE
DL
4d DL#:123456789
1 SHWEDO
2 KEVIN ANTHONY
8 10311 WILSON BLVD
COLUMBIA, SC 290161234
3 DOB: 07/04/1956
4a Issued: 08/28/2017
4b Expires: 07/04/2025
15 Sex: M 16 Hgt: 5'-11"
17 Wgt: 195 lb 18 Eyes: BLU
9 Class: D
12 Restrictions: NONE
5 DD 0100010602224403054
07/04/1956
VETERAN
Governor

Text Annotation

Polygons



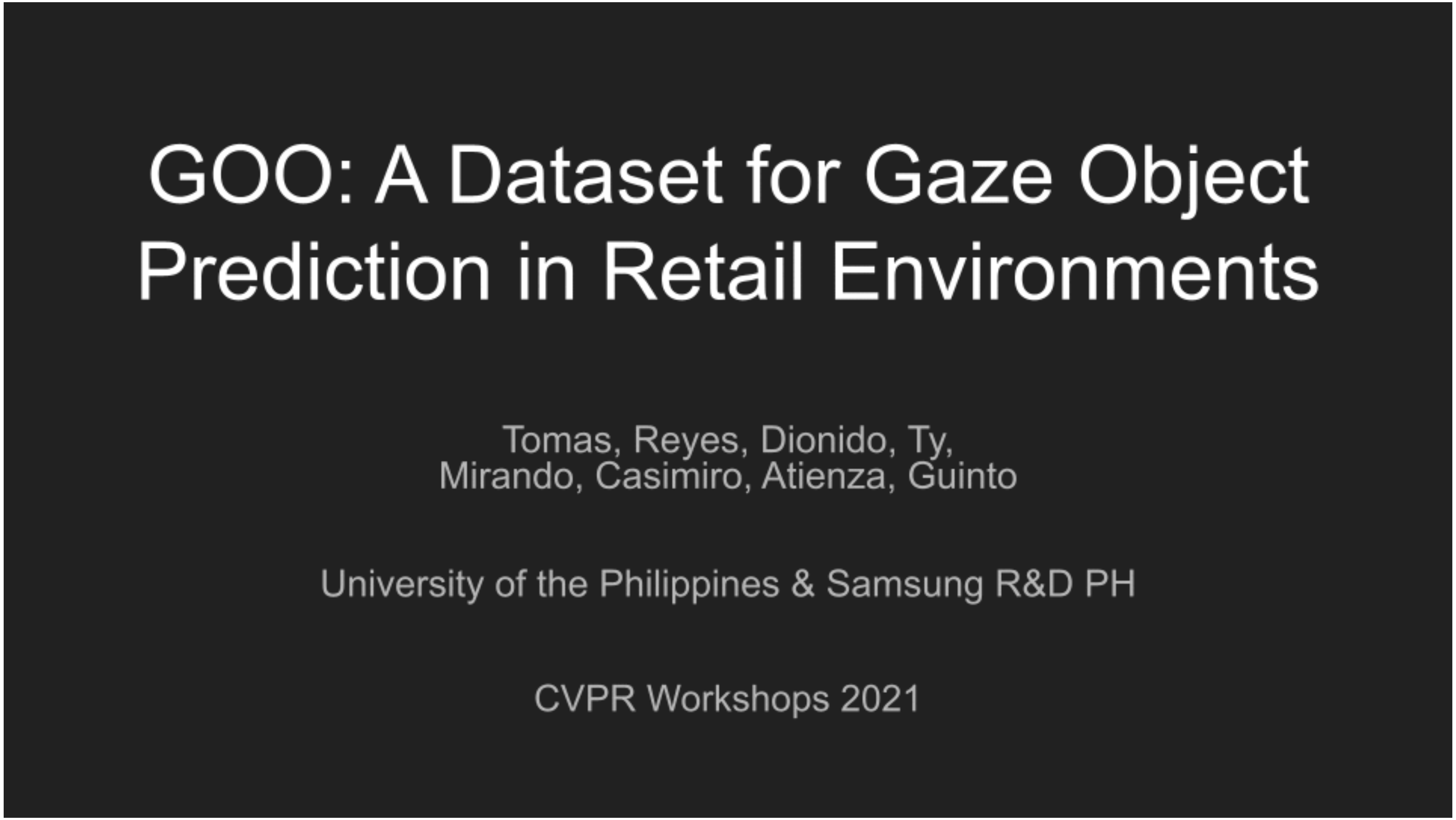
Semantic Segmentation



Video Annotation

Syntheic Datasets

https://github.com/upeee/GOO-GAZE2021



GOO: A Dataset for Gaze Object Prediction in Retail Environments

Tomas, Reyes, Dionido, Ty,
Mirando, Casimiro, Atienza, Guinto

University of the Philippines & Samsung R&D PH

CVPR Workshops 2021

Publicly Available Datasets

Paperswithcode https://paperswithcode.com/datasets

HuggingFace Datasets https://huggingface.co/datasets

Self-driving:

http://apolloscape.auto/index.html

https://waymo.com/intl/en_us/dataset-download-terms/

Issues on datasets

Sufficiency

Train/Validation/Test

Bias

Sufficiency

The target test performance is achievable given a capable model

Train/Validation/Test Split

70/10/20

70/20/10

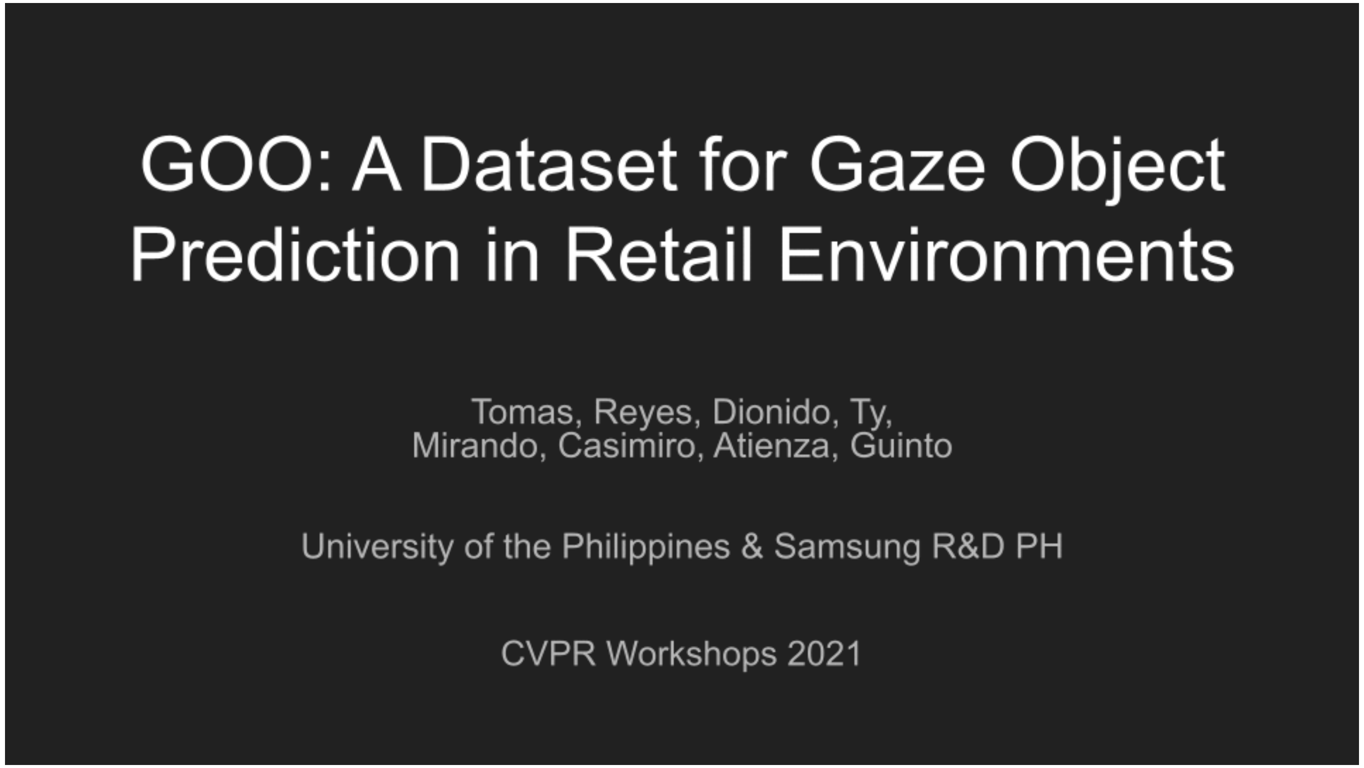
80/0/20

80/10/10

No hard rule but no duplication of samples in the splits!

Bias

Sampling, Exclusion, Racial, Measurement, Recall, Association, Observer



GOO: A Dataset for Gaze Object Prediction in Retail Environments

Tomas, Reyes, Dionido, Ty,
Mirando, Casimiro, Atienza, Guinto

University of the Philippines & Samsung R&D PH

CVPR Workshops 2021

https://www.telusinternational.com/articles/7-types-of-data-bias-in-machine-learning

PyTorch Datasets

https://pytorch.org/vision/stable/datasets.html

torch.utils.data.Dataset

Supports `__getitem__()` and `__len__()`

https://pytorch.org/docs/stable/data.html#torch.utils.data.Dataset

torchvision.datasets

https://pytorch.org/vision/stable/datasets.html

Common datasets for vision classification, detection, segmentation, action recognition, captioning

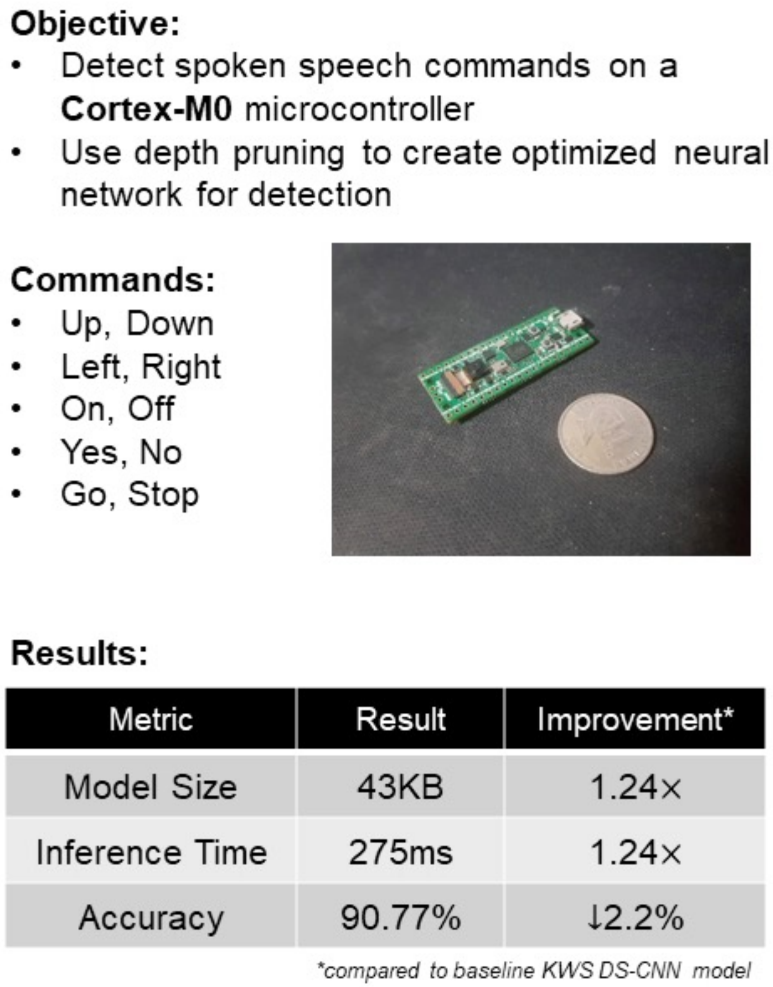
COCO 2017 Keypoint Detection Task

torchaudio.datasets

https://pytorch.org/audio/stable/datasets.html

Common datasets for automatic speech recognition (ASR), text to speech (TTS), keyword spotting (KWS), voice cloning, music genre recognition

Depth Pruning
For TinyML,
De Leon & Atienza
ICASSP 2022



Objective:

- Detect spoken speech commands on a **Cortex-M0** microcontroller
- Use depth pruning to create optimized neural network for detection

Commands:

- Up, Down
- Left, Right
- On, Off
- Yes, No
- Go, Stop

Results:

Metric	Result	Improvement*
Model Size	43KB	1.24×
Inference Time	275ms	1.24×
Accuracy	90.77%	↓2.2%

**compared to baseline KWS DS-CNN model*

torchtext.datasets

https://pytorch.org/text/stable/datasets.html

Common datasets for text classification, language modelling, machine translation, sequence tagging, question answer, unsupervised learning

Dataloaders

https://pytorch.org/tutorials/beginner/basics/data_tutorial.html

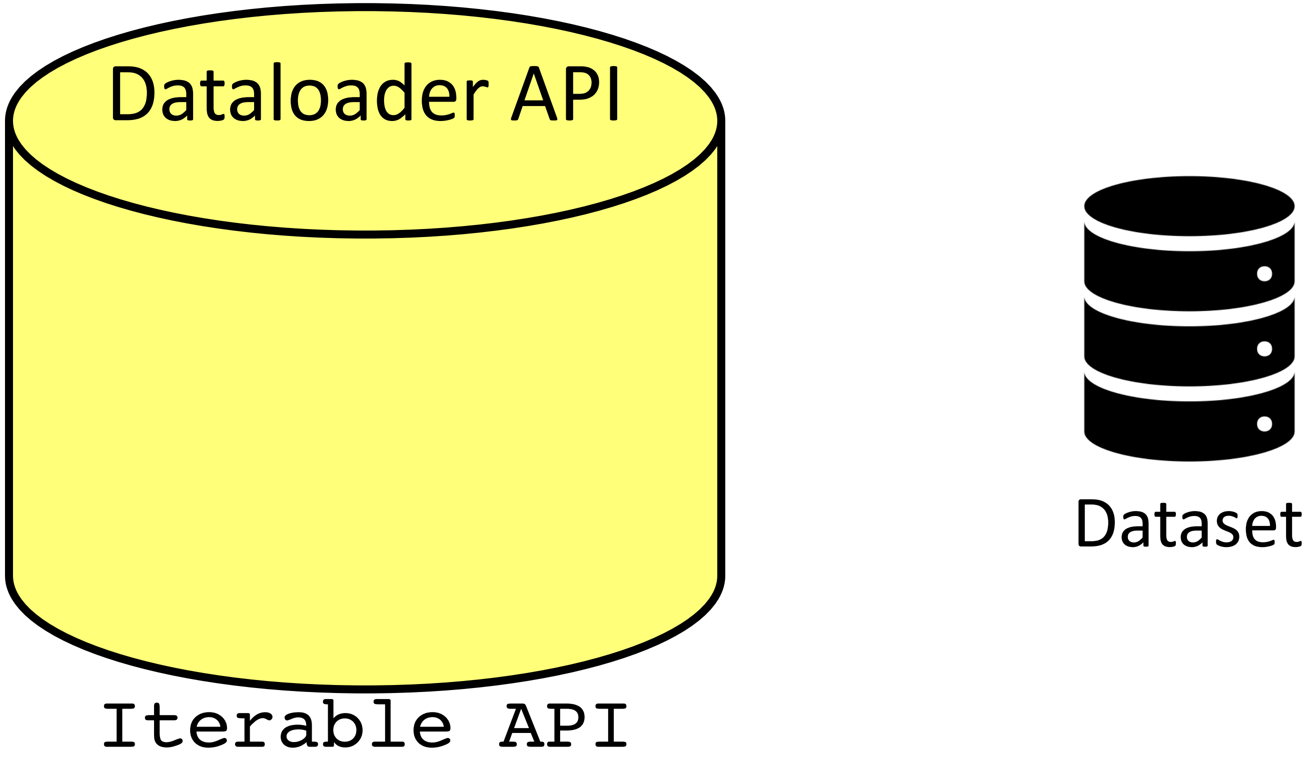
Why dataloaders

Mini-batch loading

Dataset shuffling

Multi-processing

Dataloader is Dataset Wrapper



Dataloader API

`Iterable API`

Dataset

torch.utils.data.DataLoader

```
DataLoader(dataset,
           batch_size=1,
           shuffle=False,
           num_workers=0,
           collate_fn=None,
           pin_memory=False,
            …)
```

Code demo is next

https://github.com/roatienza/Deep-Learning-Experiments/blob/master/versions/2022/datasets/python/dataloader_demo.ipynb