Data Preparation for Machine Learning

Improve your ML models by curating your vision data

SBB CFF FFS

AI Retailer Systems

FORTUNE 500

arm

curbFlow

Frontify

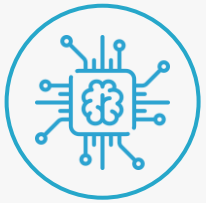
How does it work?

LIGHTLY

Get the most out of images, text, and voice
by filtering your data with Artificial Intelligence

Customized

We provide you with a suite of different filters and modifiable parameters



Unsupervised

Our algorithms work on raw and unlabeled computer vision data

Fast

Our highly optimized embeddings and filters allow you to filter 100’000 images in a few hours

Academic Dataset Studies

In the following slides, we show benchmarks of our filtering solution on well-known academic datasets for various computer vision tasks such as classification or segmentation.

Note that those datasets have been manually filtered by original authors before publishing.

Academic Dataset: CIFAR10

Dataset Description

Task:

- Image Classification

Training Set:

- 50'000 images
- 10 classes

Test Set:

- 10'000 images

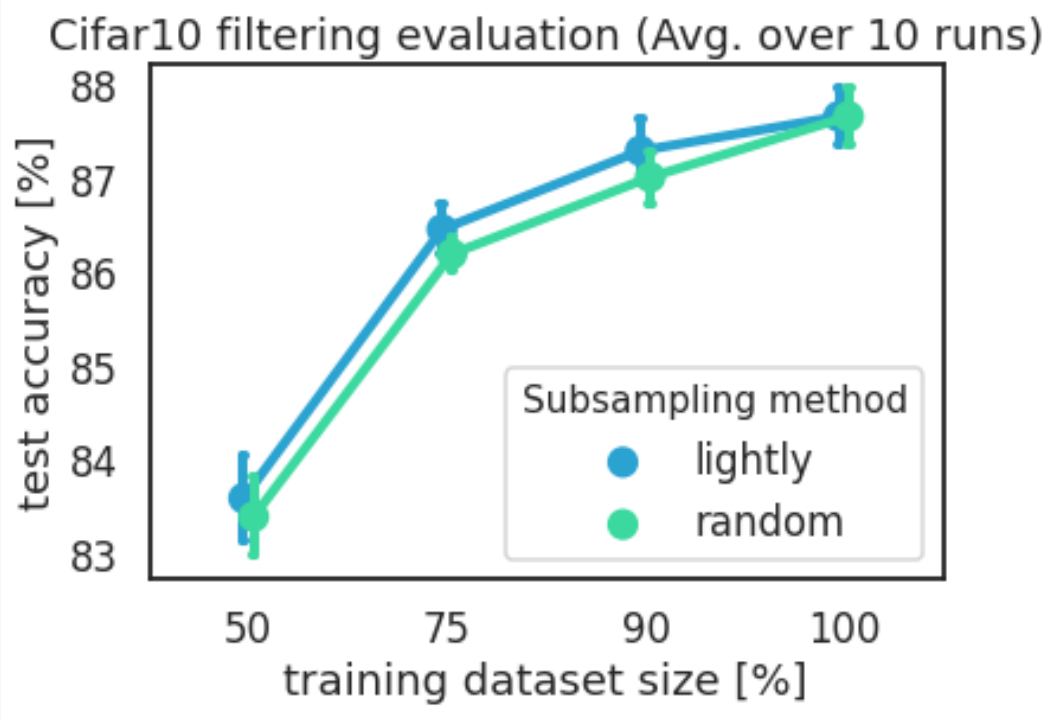
Evaluation Method

Experiment:

- Resnet34
- Train for 100 epochs
- SGD, wd=5e-4
- lr=0.1, decay by 10 at epochs 60 and 80

Results Using Lightly

- We report the best test accuracy (mean + std) over several runs with different random seeds



Academic Dataset: CamVid

Dataset Description

Task:

- Semantic Segmentation

Training Set:

- 367 images
- 11 classes

Test Set:

- 101 images

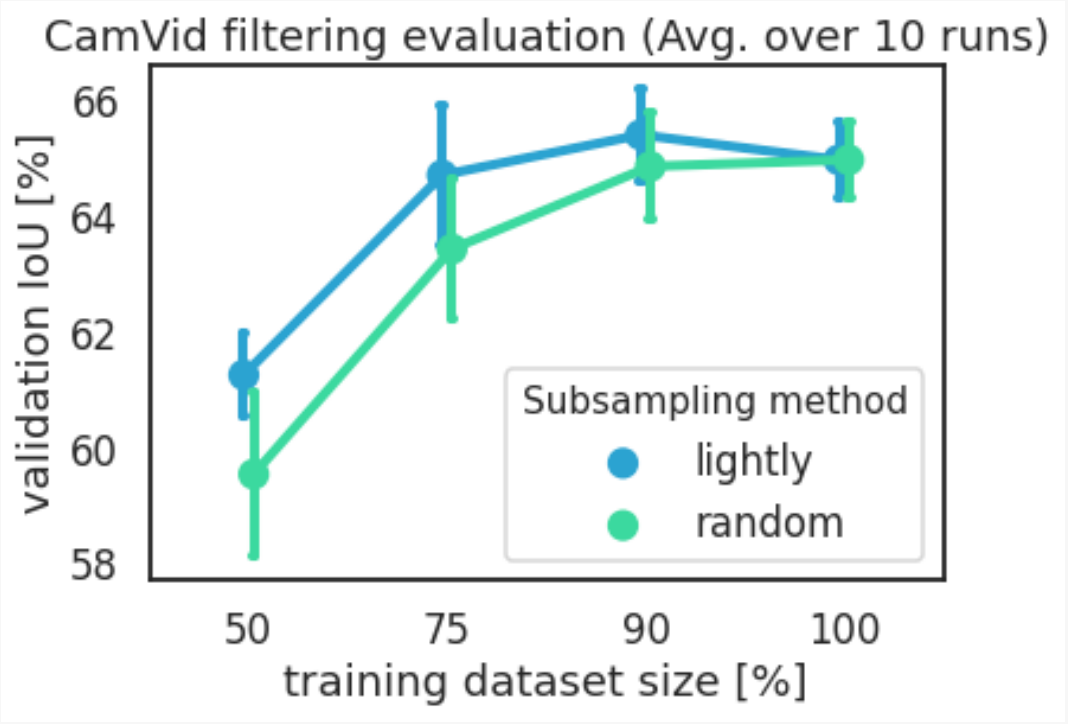
Evaluation Method

Experiment:

- E-NET
- Train for 300 epochs
- Code from: https://github.com/davidtvs/PyTorch-ENet

Results Using Lightly

- We report the best validation IoU averaged (mean + std) over several runs with different random seeds



Academic Dataset: Cityscapes

Dataset Description

Task:

- Semantic Segmentation

Training Set:

- 2975 images
- 19 classes

Test Set:

- 500 images

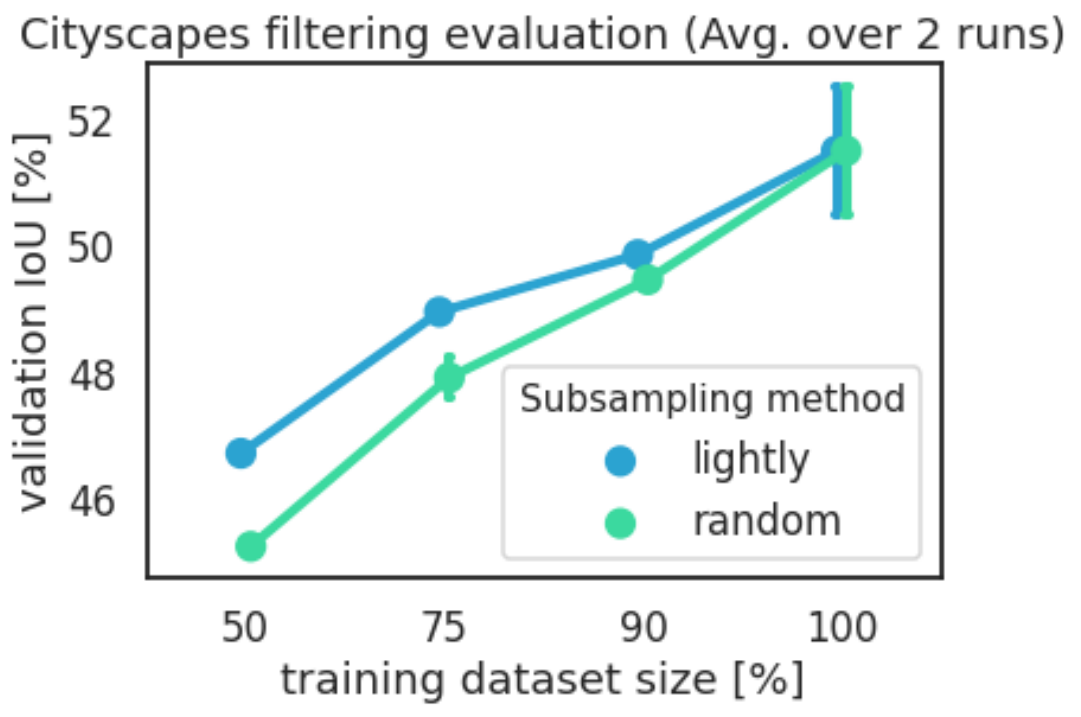
Evaluation Method

Experiment:

- E-NET
- Train for 300 epochs
- Code from: https://github.com/davidtvs/PyTorch-ENet

Results Using Lightly

- We report the best validation IoU averaged (mean + std) over several runs with different random seeds



Get in Contact with us

Matthias

MSc. HEC Paris | B.A. HSG, VUS Harvard

LinkedIn: /in/matthiasheller/

E-mail: matthias@whattolabel.com

Igor Susmelj

MSc. ETH Zurich | BSc. ETH Zurich

LinkedIn: /in/igorsusmelj/

E-mail: igor@whattolabel.com

Build better ML based-products, accelerate your product development, and save up to 50% of your data related costs