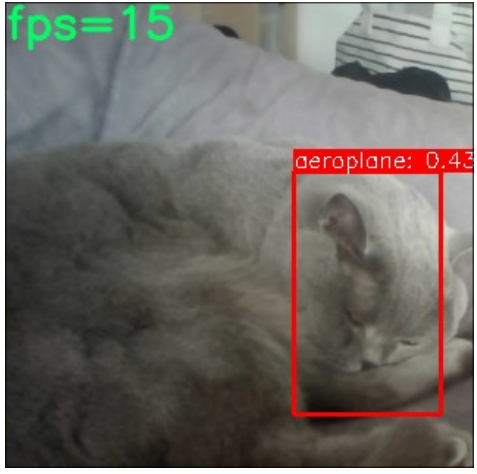
Karlsruhe Institute of Technology

Embedded Machine Learning Lab

Max Schik, Darius Schefer | March 20, 2024

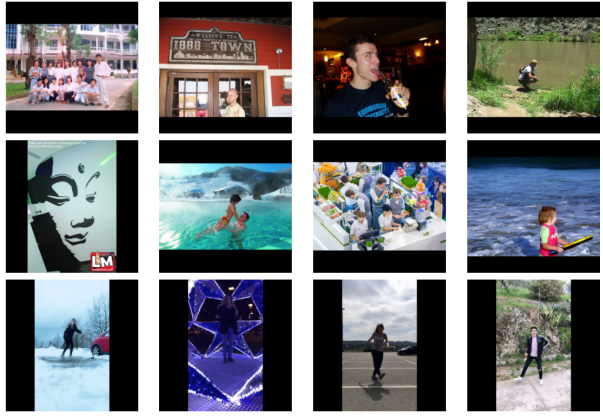
Contents

Cat



Data

- More data
  - Human Dataset [a] (17,300 images)
  - Tiktok Dancing [b] (2615 images)
- Data augmentation
  - Albumentations[c] library
  - Rotation, flipping, contrast



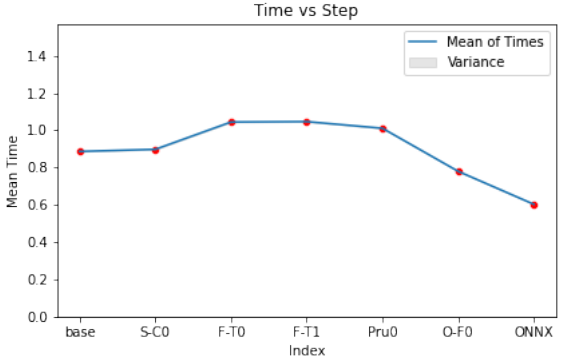
[a] https://www.kaggle.com/datasets/fareselmenshawii/human-dataset

[b] https://www.kaggle.com/datasets/tapakah68/segmentation-full-body-tiktok-dancing-dataset

[c] https://albumentations.ai/

Adaptation and optimization

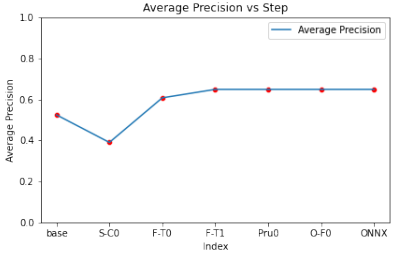
- Person-only-detection, fine-tuning
- Iterative pruning
- Batch norm optimization
- Inference

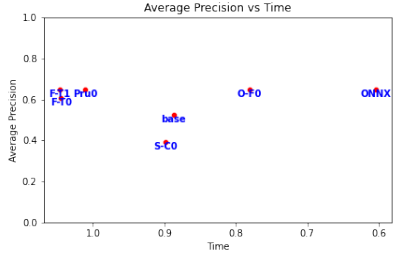


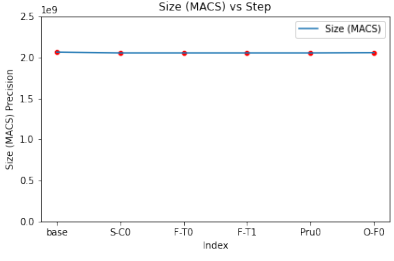
Inference

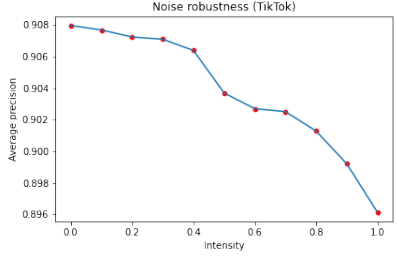
- ONNX
- TensorRT
  - TensorRT supports only a subset of ONNX spec
  - `ReflectivePad` → `ConstantPad`
  - Simplification of graph with `onnx-simplifier`[a]

[a] `https://github.com/daquexian/onnx-simplifier`


Average Precision vs Step
Average Precision
Average Precision
Index
base
S-C0
F-T0
F-T1
Pru0
O-F0
ONNX


Average Precision vs Time
Average Precision
Time
F-T1
Pru0
F-T0
O-F0
ONNX
base
S-C0


Size (MACS) vs Step
1e9
Size (MACS)
Size (MACS) Precision
Index
base
S-C0
F-T0
F-T1
Pru0
O-F0


Noise robustness (TikTok)
Average precision
Intensity

Detections


aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00
aeroplane: 1.00

Integration and pipeline

- Configuration
  - File containing steps and parameters
  - Reproducibility
  - Version control
- Camera loop
  - TensorRT for faster inference
  - Avg. 20fps

```
--- !experiment
start_weights_path: "./weights/voc_pretrained.pt"
augment: True

steps:
- !strip_classes
    finetune: true
    finetune_epochs: 15
- !pruning
    target_acc: 0.3
    prune_ratio: 0.05
    batch_size: 64
    num_train_epochs: 10
    num_eval_batches: 10
- !operator_fusion {}

```

Live Demo