Karlsruhe Institute of Technology

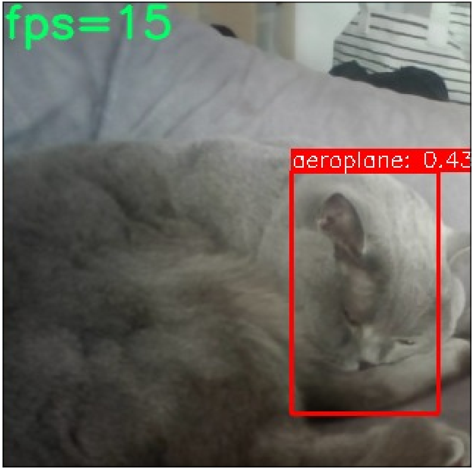
Embedded Machine Learning Lab

Max Schik, Darius Schefer | March 20, 2024

Contents

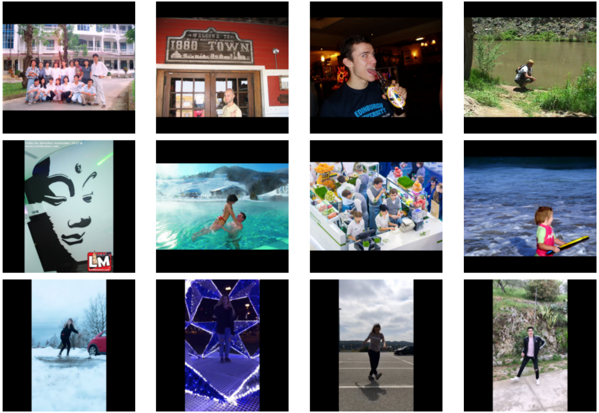
1. Data
2. Adaptation and optimization
3. Integration and pipeline
4. Live demo

Cat



Data

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



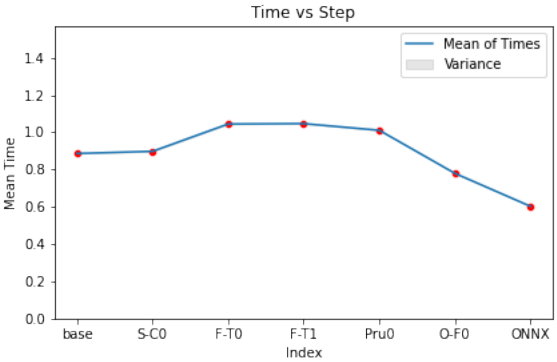
[a] https://www.kaggle.com/datasets/fareselmenshawii/human-dataset

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

[c] https://albumentations.ai/

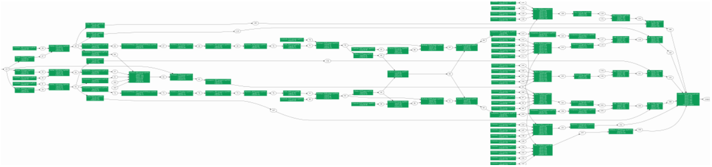
Adaptation and optimization

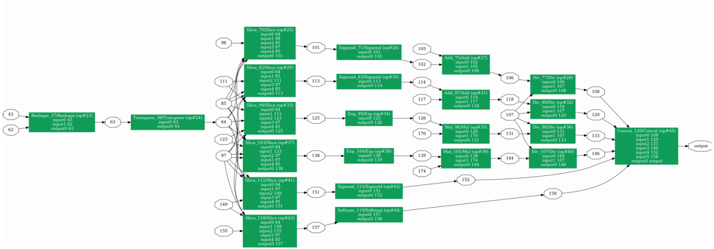
- Person-only-detection, fine-tuning
- Iterative pruning
- Batch norm optimization
- Inference



Inference

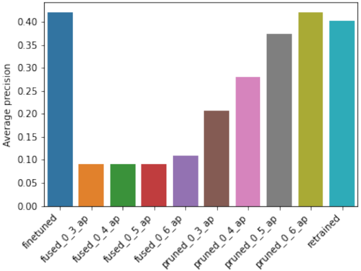
- ONNX
- TensorRT
  - TensorRT supports only a subset of ONNX spec
  - `ReflectivePad` → `ConstantPad`
  - Simplification of graph with `onnx-simplifier`[a]



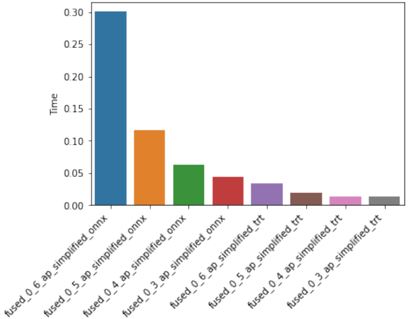


43
62
Reshape_37/Reshape (op#23) input0 43 input1 62 output0 63
63
Transpose_38/Transpose (op#24) input0 63 output0 64
98
111
85
64
123
97
149
155
Slice_70/Slice (op#25) input0 64 input1 98 input2 85 input3 97 input4 85 output0 101
Slice_82/Slice (op#29) input0 64 input1 85 input2 111 input3 97 input4 85 output0 113
Slice_94/Slice (op#33) input0 64 input1 111 input2 123 input3 97 input4 85 output0 125
Slice_103/Slice (op#37) input0 64 input1 123 input2 97 input3 97 input4 85 output0 138
Slice_112/Slice (op#41) input0 64 input1 97 input2 149 input3 97 input4 85 output0 151
Slice_118/Slice (op#43) input0 64 input1 149 input2 155 input3 97 input4 85 output0 157
101
113
125
138
151
157
Sigmoid_71/Sigmoid (op#26) input0 101 output0 102
Sigmoid_83/Sigmoid (op#30) input0 113 output0 114
Exp_95/Exp (op#34) input0 125 output0 126
Exp_104/Exp (op#38) input0 138 output0 139
Sigmoid_113/Sigmoid (op#42) input0 151 output0 152
Softmax_119/Softmax (op#44) input0 157 output0 158
105
102
114
117
126
170
139
174
Add_75/Add (op#27) input0 102 input1 105 output0 106
Add_87/Add (op#31) input0 114 input1 117 output0 118
Mul_96/Mul (op#35) input0 126 input1 170 output0 131
Mul_105/Mul (op#39) input0 139 input1 174 output0 144
106
118
107
131
144
152
158
Div_77/Div (op#28) input0 106 input1 107 output0 108
Div_89/Div (op#32) input0 118 input1 107 output0 120
Div_98/Div (op#36) input0 131 input1 107 output0 133
Div_107/Div (op#40) input0 144 input1 107 output0 146
108
120
133
146
Concat_120/Concat (op#45) input0 108 input1 120 input2 133 input3 146 input4 152 input5 158 output0 output
output

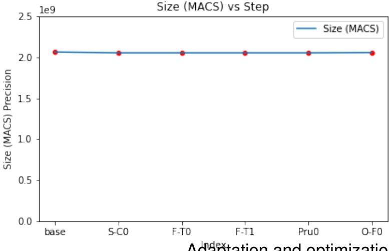
[a] https://github.com/daquexian/onnx-simplifier



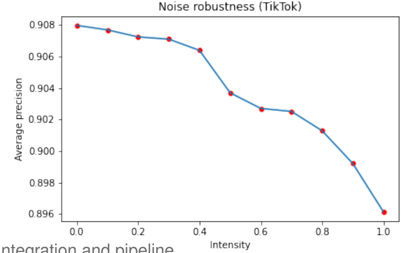
Average precision
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
finetuned
fused_0_3_ap
fused_0_4_ap
fused_0_5_ap
fused_0_6_ap
pruned_0_3_ap
pruned_0_4_ap
pruned_0_5_ap
pruned_0_6_ap
retrained



Time
0.00
0.05
0.10
0.15
0.20
0.25
0.30
fused_0_6_ap_simplified_onnx
fused_0_5_ap_simplified_onnx
fused_0_4_ap_simplified_onnx
fused_0_3_ap_simplified_onnx
fused_0_6_ap_simplified_trt
fused_0_5_ap_simplified_trt
fused_0_4_ap_simplified_trt
fused_0_3_ap_simplified_trt

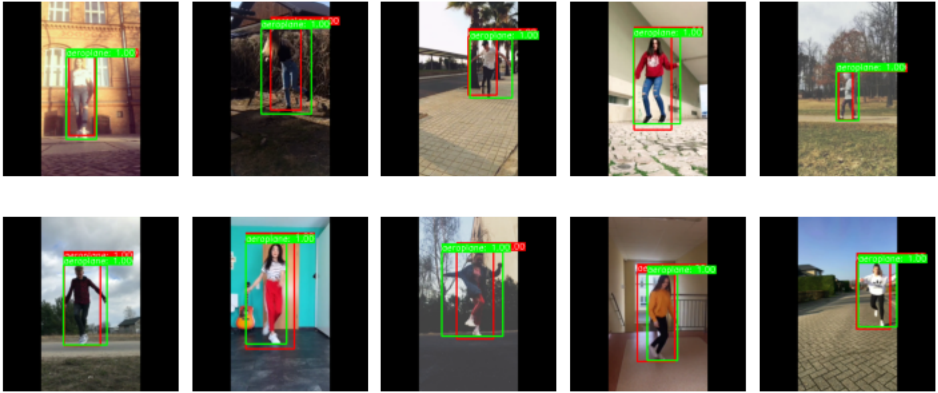


Size (MACS) vs Step
1e9
Size (MACS)
Size (MACS) Precision
0.0
0.5
1.0
1.5
2.0
2.5
base
S-C0
F-T0
F-T1
Pru0
O-F0
Index



Noise robustness (TikTok)
Average precision
0.896
0.898
0.900
0.902
0.904
0.906
0.908
0.0
0.2
0.4
0.6
0.8
1.0
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
aeroplane: 1.00

Integration and pipeline

- Configuration
  - File containing steps and parameters
  - Reproducibility
  - Version control

```
--- !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 {}

```

Pipeline

- jetcam
- inference using framework
  - separate thread for camera creates problem for tensorrt
- faster NMS by using torch implementation (30% faster than default implementation)

Live Demo