Inference optimization with pruning and quantization

MSc Seminar

Jelena Bratulić

Supervisor: prof. Siniša Šegvić, PhD

25/05/2022

Fakultet
elektrotehnike i
računarstva

Table of contents

1. Pruning
2. Quantization
3. SwiftNet
4. Datasets
5. Results
6. Conclusion and future ideas

Pruning

- Performance optimization technique which removes unnecessary activations from the model
- Activations' importance is usually determined by L1 norm or some other metric
- Structured and unustructured pruning
- Lottery Ticket Hypothesis
    - creates submasks (winning tickets) from full model
    - masks are created by comparing the weights of fully trained model by its L1 norm
    - weights are pruned according to the previously calculated mask and the rest of the weights are trained from the same initialization as the fully trained model

Fakultet elektrotehnike i računarstva

Quantization

- Performance optimization technique that allows speeding up inference and decreasing memory requirements by performing computations and storing tensors at lower bitwidths (such as INT8 or FLOAT16) than floating-point precision.
- Quantization to INT8 requires additional step of calibration
  - ensures better approximation of lower and upper bounds for quantization process
  - achieved as multiple forward passes through calibration set (subset of validation or training set) and a statistics calculations from the activations' values
- TensorRT - NVIDIA's open-source optimization library for inference optimization

Fakultet elektrotehnike i računarstva

SwiftNet

- Architecture for efficient semantic segmentation introduced by our research group in 2019
- 3 main parts:
  - Recognition encoder (ResNet or MobileNet V2)
  - Upsampling decoder
  - Module for increasing the receptive field
    - spatial pyramid pooling
    - pyramid fusion
- Pyramidal fusion on multi-scale images - full resolution, half resolution and quarter of resolution

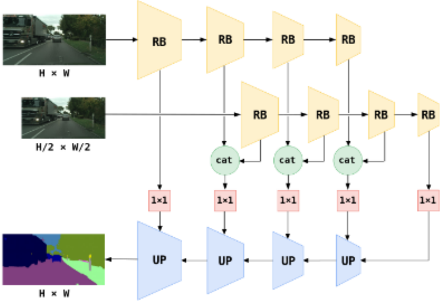


Figure: Architecture of two-level pyramidal SwiftNet

Datasets

Cityscapes

- High-resolution images (1024 x 2048) of outdoor scenes
- 2975 images for train, 500 images for val and 1525 images for test
- 19 classes (road, sidewalk, building, vegetation, car, ...)
- Training on 784x784 crops

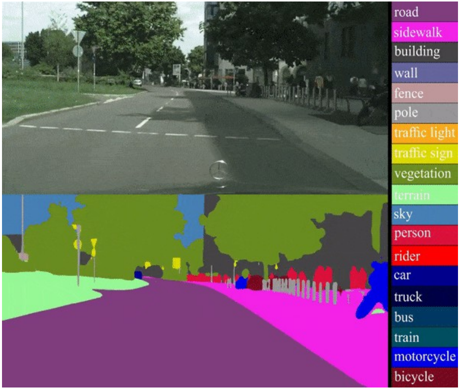


Figure: Frame from Cityscapes dataset.

Fakultet elektrotehnike i računarstva

Datasets

Ade20k

- indoor and outdoor scenes in various resolutions
- 20000 images for train, 2000 images for validation and 3000 images for test
- 150 classes (door, wall, building, person, vehicle, etc.)

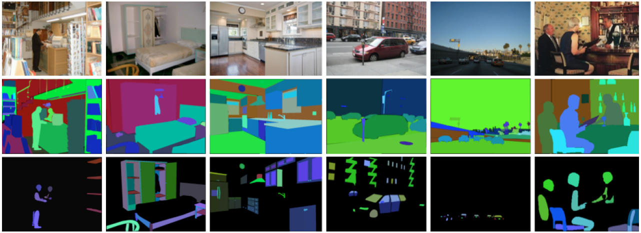
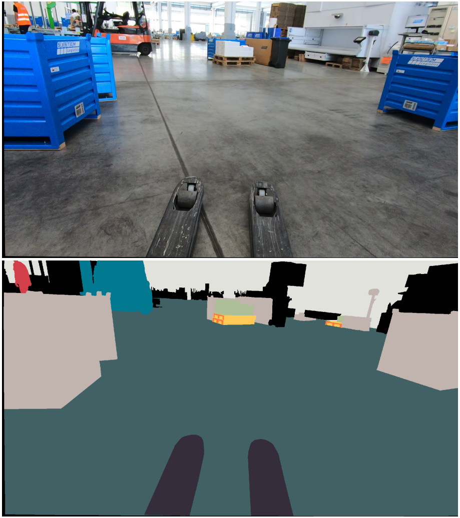


Figure: Different frames from Ade20k dataset.

Fakultet elektrotehnike i računarstva

Datasets

Romb Technologies



- In-house dataset acquired by **Romb Technologies Ltd**.
- High resolution (1920 x 1080) images of different warehouse scenarios

Results

Model	Number of parameters	GMACs	mIoU [%]	FPS
Full model	22 040 278	209.67	48.6	18.2
Pruned	21 004 589	143.13	49.5	23.4

Table: Performance on Romb data for iteratively pruned model on Ade20k and finetuned on Romb data. FPS were measured on Nvidia 1080 Ti.

Model	FP16		INT8	
	mIoU [%]	FPS	mIoU [%]	FPS
Full	48.6	16.4	48.6	30.4
Pruned	49.5	20.4	49.5	34.4

Table: Performance on Jetson AGX Xavier and old Romb data.

Fakultet elektrotehnike i računarstva

Results

Pruning - results on Ade20k

Model	Number of parameters	GMACs	mIoU [%]	FPS
Full	22 040 278	209.67	33.89	18.43
LTH	20 857 526	139.68	32.48	24.63
FT	20 857 526	139.68	32.05	24.63
RI	20 857 526	139.68	20.39	24.63

Table: Performance on Ade20 for model pruned by 50% in first 2 blocks (LTH - Lottery Ticket Hypothesis, FT - Fine tuning, RI - Random init). FPS were measured on Nvidia GTX 1080 Ti.

Results

Pruning - results on Ade20k

Model	Number of parameters	GMACs	mIoU [%]	FPS
Full	22 040 278	209.67	33.89	18.43
LTH	6 010 806	59.75	30.55	36.30
FT	6 010 806	59.75	31.49	36.30
RI	6 010 806	59.75	22.84	36.30

Table: Performance on Ade20k for model pruned by 50% in all 4 blocks (LTH - Lottery Ticket Hypothesis, FT - Fine tuning, RI - Random init). FPS were measured on Nvidia GTX 1080 Ti.

Fakultet elektrotehnike i računarstva

Results

Pruning - results on updated Romb data

Model	Number of parameters	GMACs	mIoU [%]
Full model from random	22 040 278	209.67	65.86
Full model from Ade20k	22 040 278	209.67	70.65
Pruned first 2 blocks	20 857 526	139.68	76.06
Pruned all blocks	6 010 806	59.75	75.54

Table: Performance on new Romb data

Results

Quantization - updated Romb data

Model	FP16		INT8	
	mIoU [%]	FPS	mIoU [%]	FPS
Full	70.65	16.4	70.60	30.4
Pruned 1st and 2nd block	76.1	23.0	75.6	37.9
Pruned all blocks	75.5	37.4	75.0	54.5

Table: Performance on Jetson AGX Xavier and updated data.

Fakultet elektrotehnike i računarstva

Conclusion and future ideas

- Benefits from pruning are dependent of dataset and model sparsity
    - fine tuning shows better performance for sparse models
- Unusuall behavior was noticed for pruned models with input feature maps of small resolution
- Quantization achieves great performance accelaration with minimal loss in accuracy
- Combining pruning with quantization achieves efficient models for real-time semantic segmentation on embedded system (AGX Jetson Xavier) with 30 FPS in INT8 for full model and 34 FPS in INT8 for pruned model
- Future ideas:
    - Combine Context-Aware pruning with Lottery Ticket Hypothesis

Fakultet elektrotehnike i računarstva

Appendix

Inference speed throught the pyramid

Level of pyramid	Model	Block 1	Block 2	Block 3	Block 4	Forward pass
0	Backbone	7.1	6.9	8.2	4.3	30.4
	Full	7.4	7.3	8.7	4.6	32.3
	Pruned	3.6	3.8	7.8	4.3	22.2
1	Backbone	1.9	2.0	2.5	1.8	9.2
	Full	2.0	2.1	2.7	1.9	9.9
	Pruned	1.0	1.3	2.4	1.8	7.4
2	Backbone	0.8	1.2	1.8	1.1	5.3
	Full	0.9	1.2	1.8	1.2	5.5
	Pruned	0.9	1.3	1.9	1.2	5.6

Table: Results for inference speed for different pyramid levels. Inference speed is measured in miliseconds (ms) and acheived on Nvidia GTX 1080 Ti.

Fakultet elektrotehnike i računarstva