Best of both, Structured and Unstructured Sparsity in Neural Networks

BOSCH
Invented for life

Christoph Schulte
cschulte@techfak.uni-bielefeld.de
Bosch Sicherheitssysteme GmbH
Bielefeld University

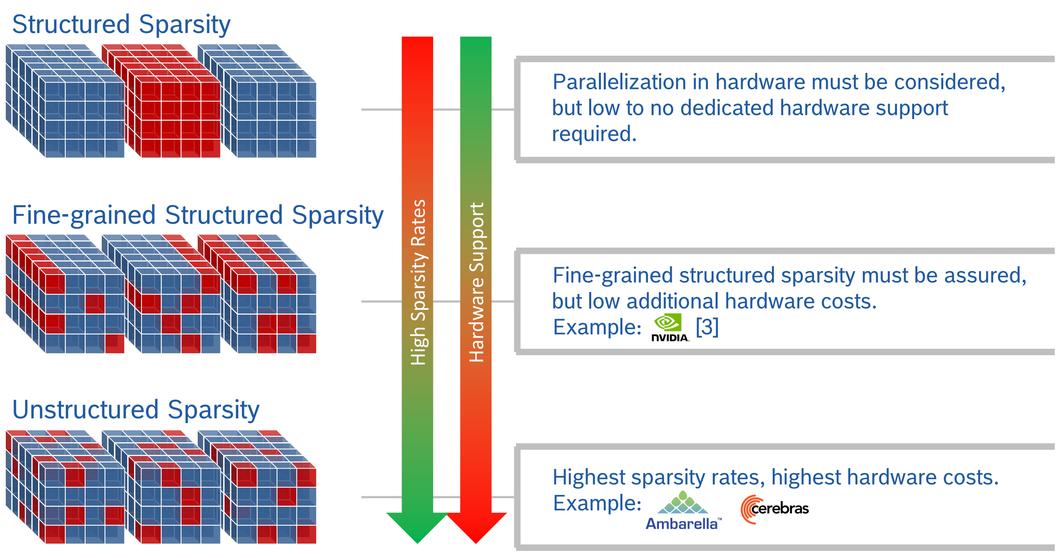
Sven Wagner
sven.wagner@de.bosch.com
Bosch Sicherheitssysteme GmbH

Armin Runge
armin.runge@us.bosch.com
Bosch Security Systems LLC

Dimitrios Bariamis
dimitrios.bariamis@de.bosch.com
Bosch Sicherheitssysteme GmbH

Barbara Hammer
bhammer@techfak.uni-bielefeld.de
Bielefeld University

Motivation



Is there an overhead associated with very high unstructured sparsity rates?

Definition of Sparsity

Different sparsity rate definitions exist:

$$sparsity = \frac{\theta_{zero}}{\theta}$$

- ➕ Reflects model size
- ➕ Simple to use
- ⚠️ Does not allow latency or throughput estimation

$$sparsity_{op} = \frac{\sum_l^L (\theta_{l,zero} \sum_i^{O_l} w_i h_i)}{\sum_l^L (\theta_l \sum_i^{O_l} w_i h_i)}$$

- ➕ Reflects computational complexity
- ➕ Allows latency and throughput estimation if system is compute-bound
- ⚠️ No consideration of memory accesses and memory bandwidth, i.e., inaccurate if system is memory-bound
- ⚠️ No consideration of potential overheads (e.g., sparsity encoding)

Sparsity rate that reflects number of operations is used

Structured and Unstructured Sparsity

Neural network architecture

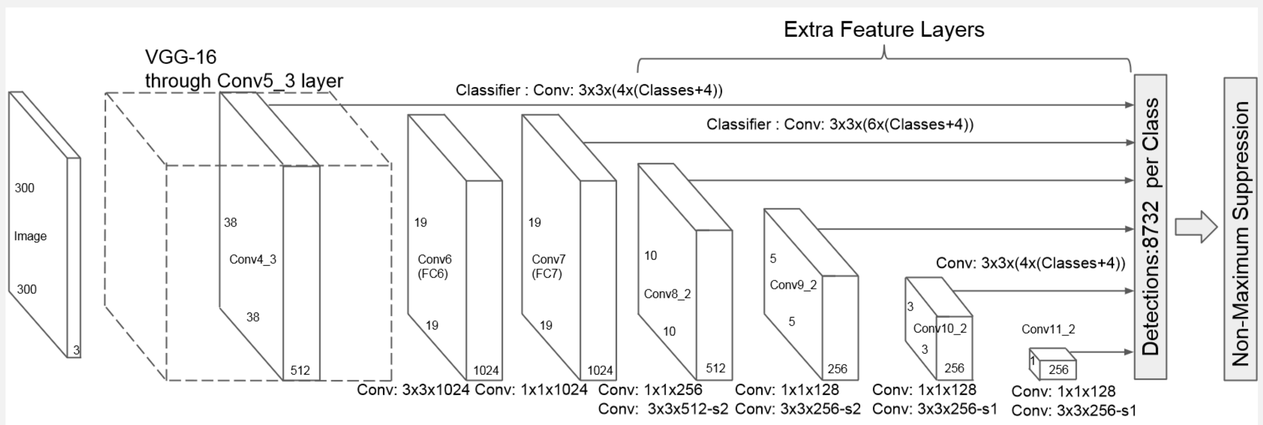


Figure 1: **S**ingle **S**hot multibox **D**etector (SSD) architecture with VGG feature extractor [1].

Unstructured pruning measurements

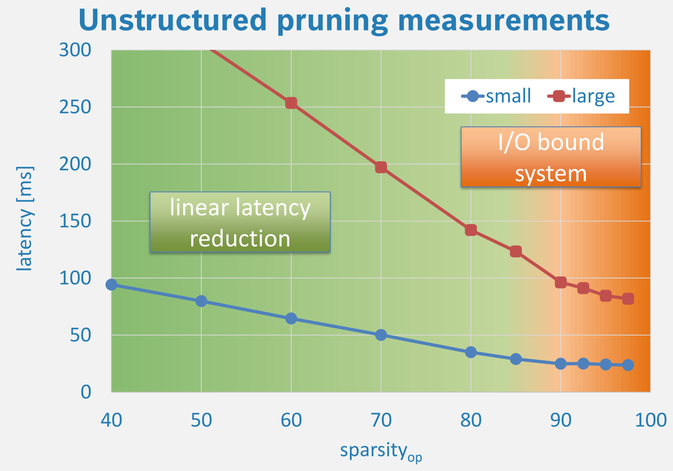


Figure 2: Sparsity rate vs. latency using unstructured pruning.

Almost linear latency reduction up to 85% $sparsity_{op}$.
Limited latency reduction for higher sparsity rates.

Experimental Settings

- **Evaluated models:**
  - Two SSD object detection (small: 5.4M parameters; large: 22M parameters) models with 7 classes and 1280x720 input resolution (see Figure 1).
- **Pruning method:** Iterative, global, L1-norm-based pruning approach
  - Small model pruned to 70% $sparsity_{op}$ without significant accuracy degradation.
- Models are **quantized to 8 bit** (activations & coefficients) using the Ambarella tools
- Latencies are measured on the **Ambarella CV22** [2].

Detailed comparison of unstructured pruned models

Model	small	large
$GMACs_{base}$	41.14	163.76
$sparsity_{op}$	80.0%	95.0%
$GMACs_{pruned}$	8.23	8.19
$latency$ [ms]	35.03	84.51
mem_p [MB]	3.08	3.41
mem_a [MB]	12.72	21.08

Limited latency reduction, because system is I/O bound.

Combining different sparsity patterns

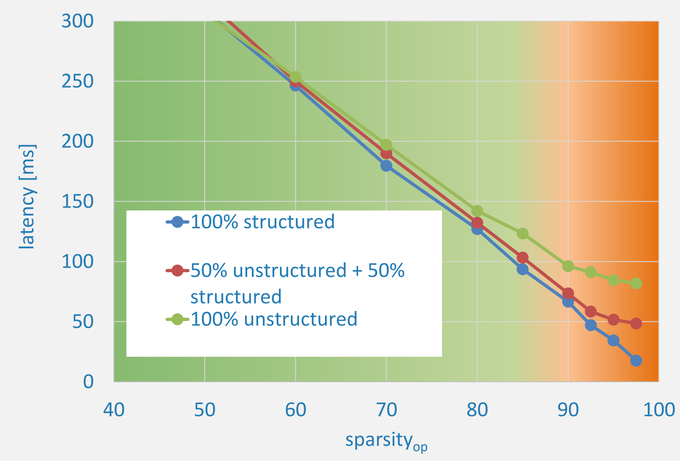


Figure 3: Sparsity rate vs. latency for large SSD models with unstructured sparsity, structured sparsity and combined sparsity.

Ratio of structured and unstructured sparsity a trade-off between latency and accuracy in high-sparsity regimes

Conclusion

- **Sparsity rate definition** matters; pruned weights contribute differently to number of required operations and latency.
- Linear speedup can be observed for **low to medium unstructured sparsity** rates on **Ambarella CV22**, a dedicated AI accelerators with coefficient sparsity support
- **High** unstructured sparsity rates lead to an **overhead**, making the baseline model a crucial factor
- **Structured sparsity** modifies the model architecture and, hence, is not affected by this overhead
- A **combination** of both, unstructured and structured sparsity can be used to achieve low latency, high throughout enabled by high sparsity rates

References

[1] Liu, Wei, et al. "**Ssd: Single shot multibox detector**." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016.

[2] Ambarella International LP. 2022. "**CV22S Computer Vision SoC for IP Cameras.**" https://www.ambarella.com/wp-content/uploads/Ambarella_CV22S_Product_Brief.pdf (2023/04/13).

[3] Nvidia Corporation. "**Accelerating Inference with Sparsity Using the NVIDIA Ampere Architecture and NVIDIA TensorRT.**" https://developer.nvidia.com/blog/accelerating-inference-with-sparsity-using-ampere-and-tensorrt/ (2021/06/20)

Link to the paper:
https://doi.org/10.1145/3578356.3592583