DATE 21

Optimization Tools for ConvNets on the Edge

Valentino Peluso
Politecnico di Torino
10129 - Italy

PhD Advisors:
Enrico Macii
Andrea Calimera

Edge ConvNets

- The proliferation of low-power devices fueled the Internet-of-Things (IoT), enabling ubiquitous sensors to sample and transmit data over the internet.
- Thanks to the recent breakthroughs in Artificial Intelligence (AI), **Convolutional Neural Networks** (ConvNets) in particular, computers took a further steps towards human intelligence.
- Embedding AI in IoT end-nodes is the premise of a new paradigm—**Artificial Intelligence of Things** (AIoT)—where sensors will evolve from passive data collectors to active intelligent devices able to infer the meaning of data locally.
- This shift will improve **efficiency**, **scalability**, and **security**.

Challenges

- Memory and computational requirements of ConvNets (Fig. 1).
- **Multi-objective** optimization: memory, energy, and power, besides accuracy.
- High diversity in hardware and use-cases.

Contributions

- Develop **cross-layer optimizations** for software-to-silicon mapping of ConvNets, with vertical strategies spanning from hardware to software.
- Offer a collection of methods for the optimization of ConvNets, addressing different design goals: memory, energy, and power.
- Devise dynamic knobs to extend the achievable accuracy-complexity tradeoffs via **run-time adaptation**.

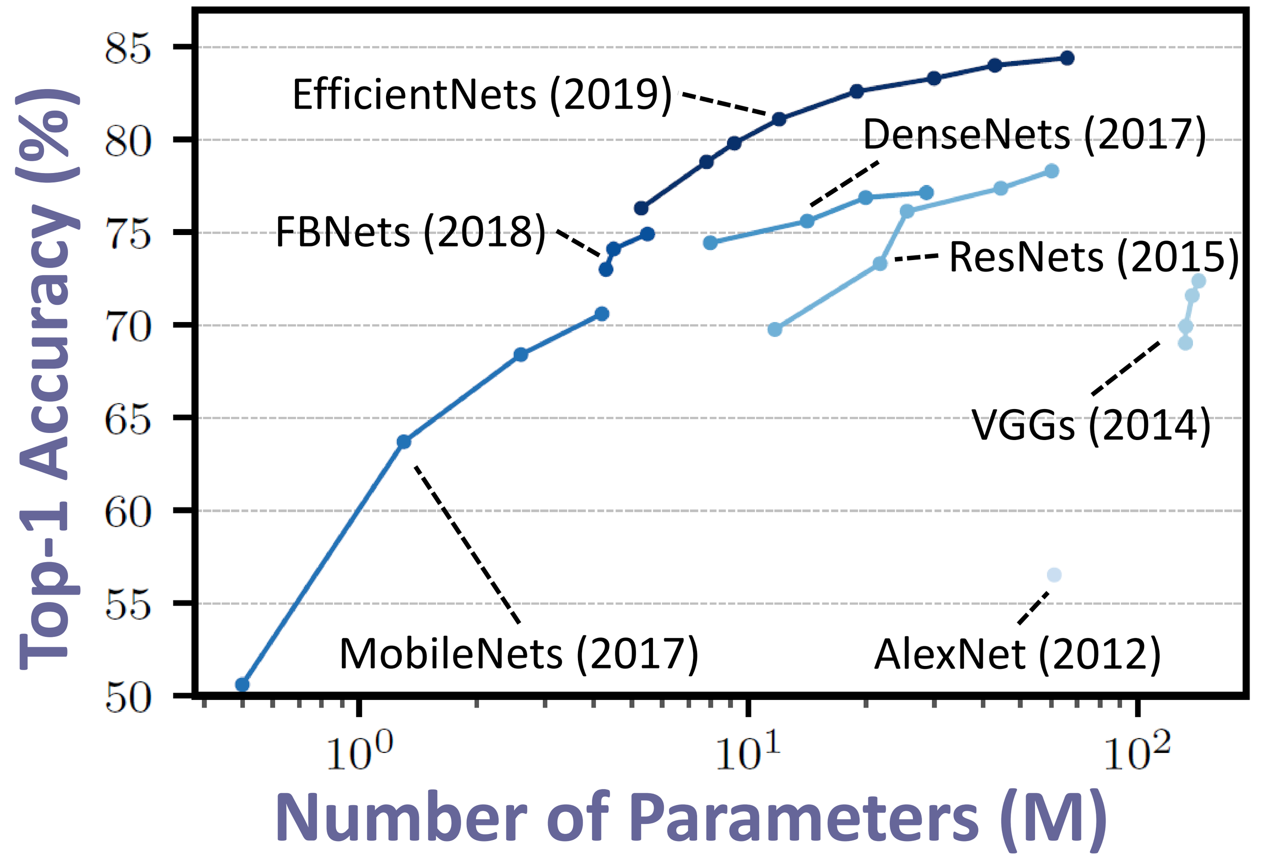

Figure 1. Top-1 Accuracy vs. Number of Parameters of modern ConvNets on the ImageNet dataset.

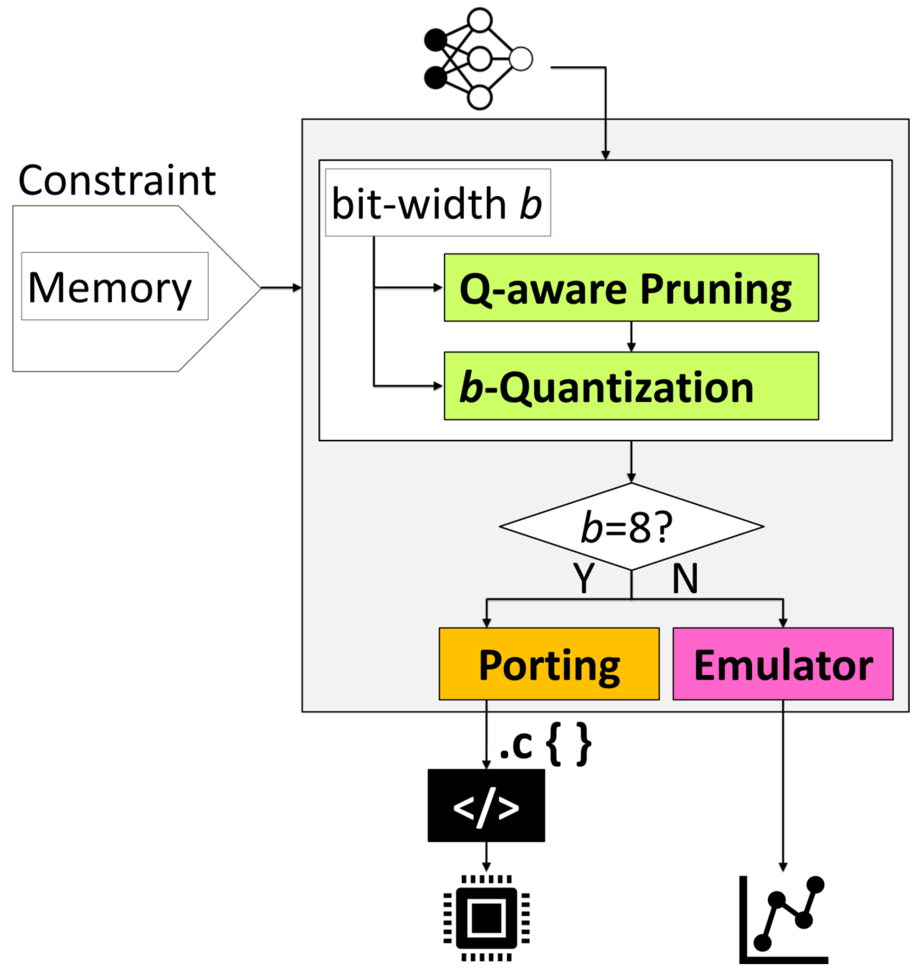

Figure 2. Prune and Quantize optimization flow.

Mem. (KB)	Optimal Bit-width	Optimal Top-1	ARM Bit-width	ARM Loss
245	15	83.10	8	0.25
115	7	82.64	8	0.20
98	7	81.99	8	0.59
82	6	81.49	8	0.70
66	6	80.42	8	1.57
49	5	78.17	8	6.53
33	5	71.85	8	17.17

Table 1. Results of Prune and Quantize on CIFAR-10 dataset

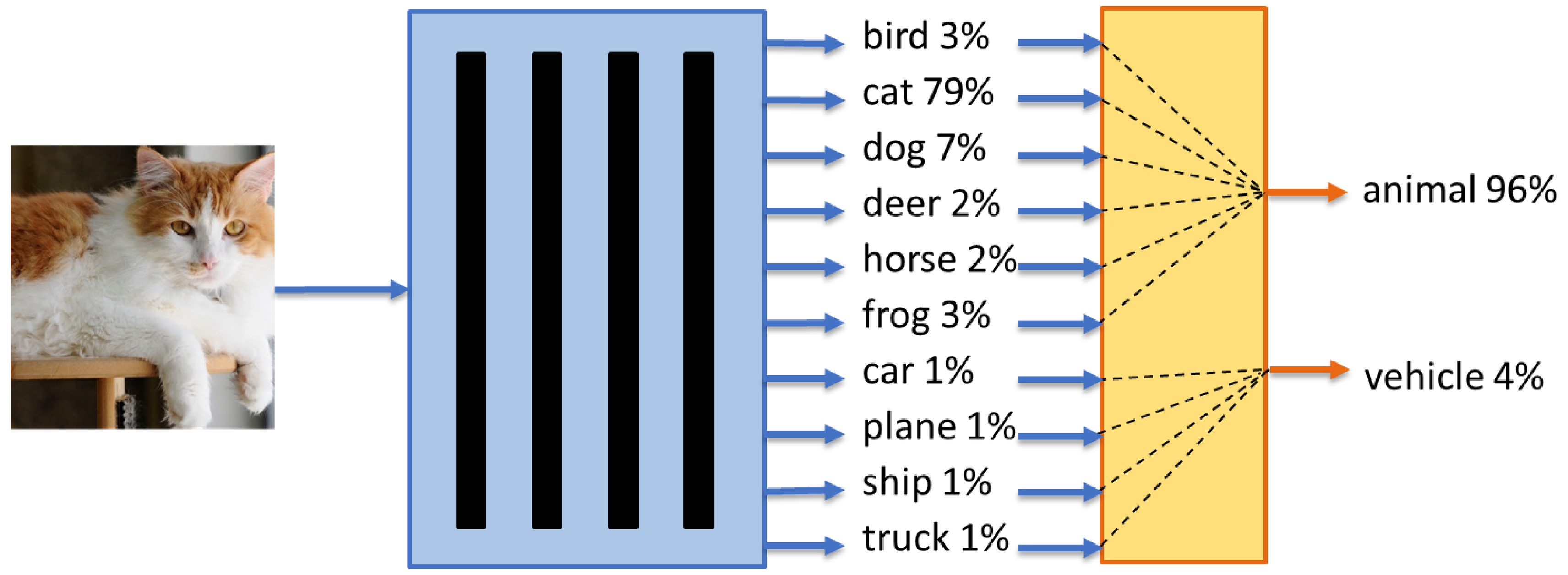

Figure 3. Multilevel Classification with ConvNets

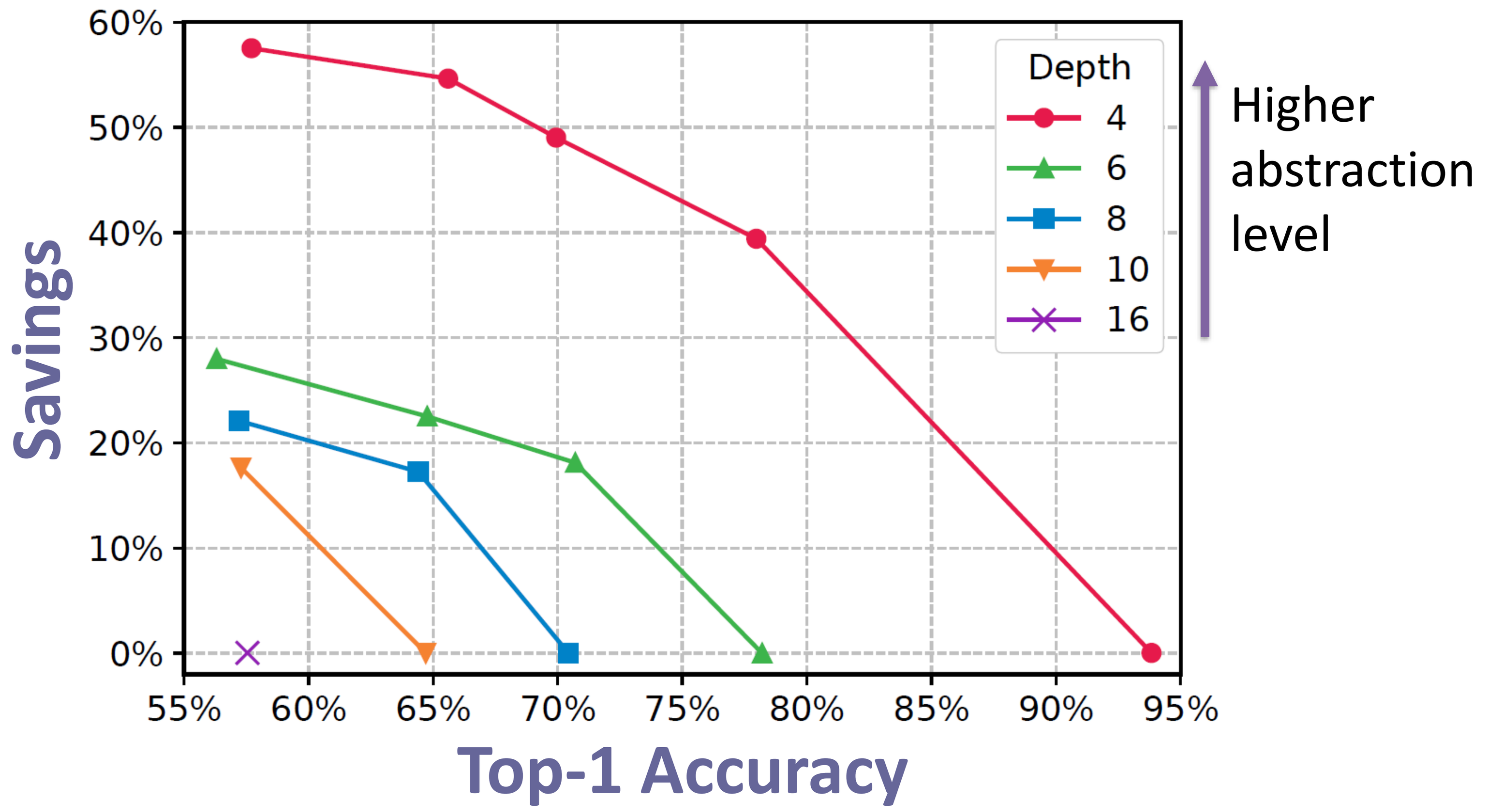

Figure 4. Scalable-Effort Classification with SqueezeNet on the ImageNet dataset.

1 Prune and Quantize — Cortex-M MCUs

- **Motivation:** Identify the best combination of pruning and quantization for memory-constrained applications.
- **Challenge:** General-purpose cores (ARM) have a limited instruction-set (minimum bit-width *b* is 8-bit).
- **Goal:** Assess the optimality of hardware-compliant solutions.
- **Results:** 3x compression with < 1% accuracy loss compared to arbitrary bit-width (Table 1).

2 Scalable-Effort ConvNets — ASICs/DSPs

- **Motivation:** State-of-art ConvNets are trained as static classifiers that expend equal efforts no matter the surrounding context and level of accuracy required.
- **Goal:** Design **Adaptive ConvNets** able to move in the abstraction-accuracy-energy space
- **Methods:** **Multilevel classification** (Fig. 3) and run-time **per-layer precision scaling**.
- **Results:** Achieves better trade-offs than static ConvNets with up to 58% energy savings or 36% higher accuracy (Fig. 4)

3 Voltage-Scaled ConvNets — Cortex-A CPUs

- **Motivation:** Modern embedded System-on-Chips have limited thermal design power, which prevents the execution of intensive workloads (like ConvNets) for long runtime at maximum voltage.
- **Goal:** Assess **thermal and power reliability** of embedded ConvNets under reactive and proactive DVFS policies.
- **Results:** On MobileNets for ImageNet classification, proactive DVFS achieves up to 17% faster processing than reactive DVFS.

References

1. M. Grimaldi, V. Peluso and A. Calimera, “Optimality Assessment of Memory-Bounded ConvNets Deployed on Resource-Constrained RISC Cores,” IEEE Access, 2019
2. V. Peluso and A. Calimera, “Scalable-effort ConvNets for Multilevel Classification,” Proc. ICCAD 2018.
3. V. Peluso, R.G. Rizzo, and A. Calimera, “Performance Profiling of Embedded ConvNets under Thermal-Aware DVFS,” Electronics, 2019