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 step 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. Devising dynamic knobs to extend the achievable accuracy-complexity tradeoffs via run-time adaptation.

---

**Figure 2. Prune and Quantize on CIFAR-10 dataset**

- 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).

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

<table>
<thead>
<tr>
<th>Mem. (KB)</th>
<th>Optimal</th>
<th>Optimal ARM</th>
<th>Bit-width</th>
<th>ARM Bit-width</th>
<th>ARM Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>15</td>
<td>83.10</td>
<td>8</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>7</td>
<td>82.64</td>
<td>8</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>7</td>
<td>81.99</td>
<td>8</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>6</td>
<td>81.49</td>
<td>8</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>6</td>
<td>80.42</td>
<td>8</td>
<td>1.57</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>5</td>
<td>78.17</td>
<td>8</td>
<td>6.53</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>5</td>
<td>71.85</td>
<td>8</td>
<td>17.17</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3. Multilevel Classification with ConvNets**

- 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).

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

- 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