On the Combination of Tensor Decomposition and Quantization for CNN Compression

Clément Laroudie (clement.laroudie@cea.fr), Mohamed Ouerfelli (mohamed-oumar.ouerfelli@cea.fr)
CEA LIST/DIASI/SIALV/LVML

Motivation

- CNNs achieve strong predictive performance but require substantial computation, limiting their deployment on resource-constrained devices such as smartphones or edge AI systems.
- However, post training compression methods can reduce its footprint: Tensor decomposition (TD) and quantization (Q) are two widely used techniques for reducing the computational burden of CNNs.
- Although extensively studied individually, their joint behavior remains underexplored. We provide a preliminary empirical study.

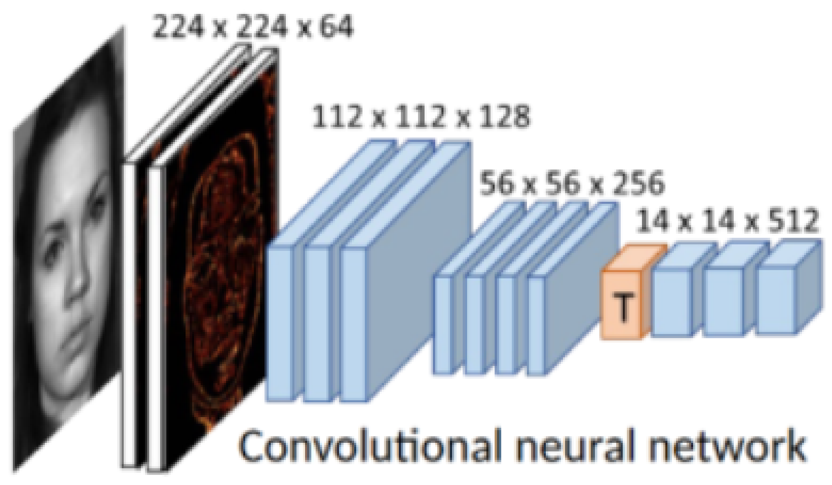

Figure 1: The layers of a CNN, where each feature map is a 3-way tensor.

- A convolution is parameterized by a kernel tensor $\mathcal{K} \in \mathbb{R}^{T\times S\times K_h\times K_w}$ mapping an input image $\mathcal{I} \in \mathbb{R}^{S\times H\times W}$ to an output image $\mathcal{Y} \in \mathbb{R}^{T\times H'\times W'}$.

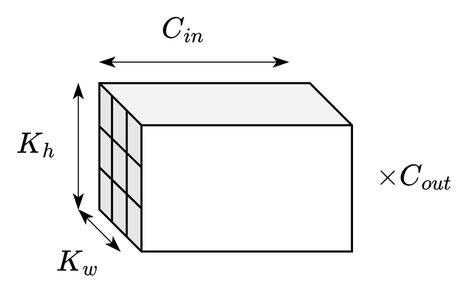

Figure 2: $\mathcal{K}$ is a 4-way tensor.

Methodology

Study of the two major decompositions:

- CP4 Decomposition [1]: Factorizing along the four modes (fig. 3).
- Tucker2 Decomposition [2]: Factorizing along the first two modes only (fig. 4).

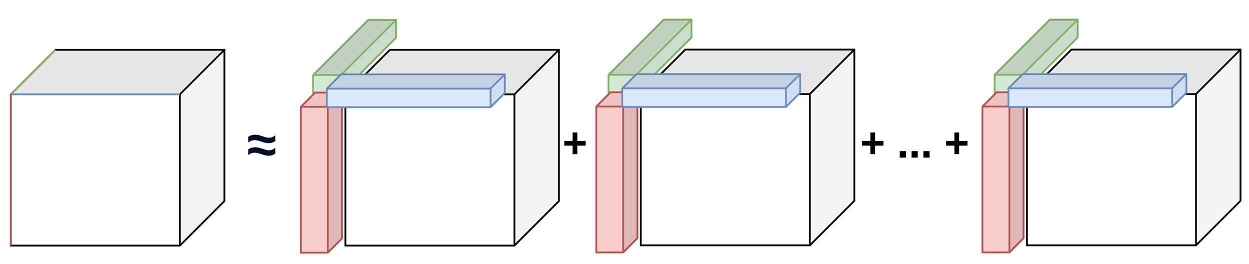

Figure 3: CP decomposition approximates a 3-way tensor T as a sum of rank-1 tensors – a generalization of matrix svd to tensors.

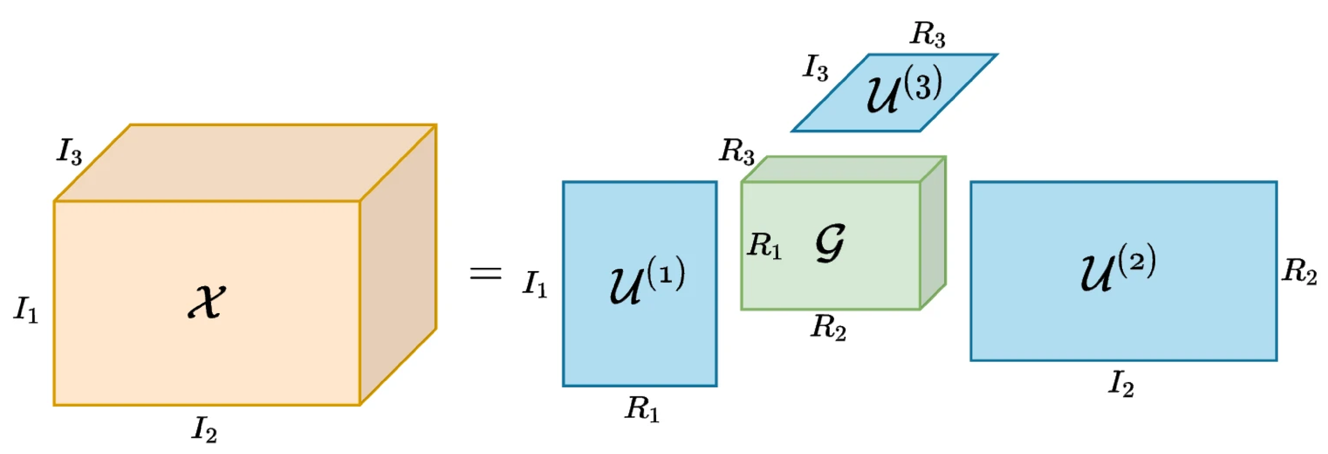

Figure 4: Tucker decomposition of a 3-way tensor $\mathcal{X}$ as a core tensor $\mathcal{G}$ and factors $\mathcal{U}^{(i)}$.

- **Models:** ResNet18/34/50, GoogLeNet, AlexNet.
- **Datasets:** CIFAR-10 and ImageNet
- **Rank Selection:** Both parameter ratio and automatic rank selection via VBMF [2].
- **Quantization:** FP16/INT8 with Pytorch and ONNX.

Preliminary Results

Tucker2

- Accuracy drops from TD+Q closely match the sum of the individual degradations, suggesting near-independence (fig. 5).
- The results also generalize with the Sigma Data-Aware Tucker Decomposition.
- **Practical pipeline:** Decompose layers with Tucker 2 (favoring later ones) to reach a desired ratio, then apply INT8 quantization to achieve an additional consistent reduction.

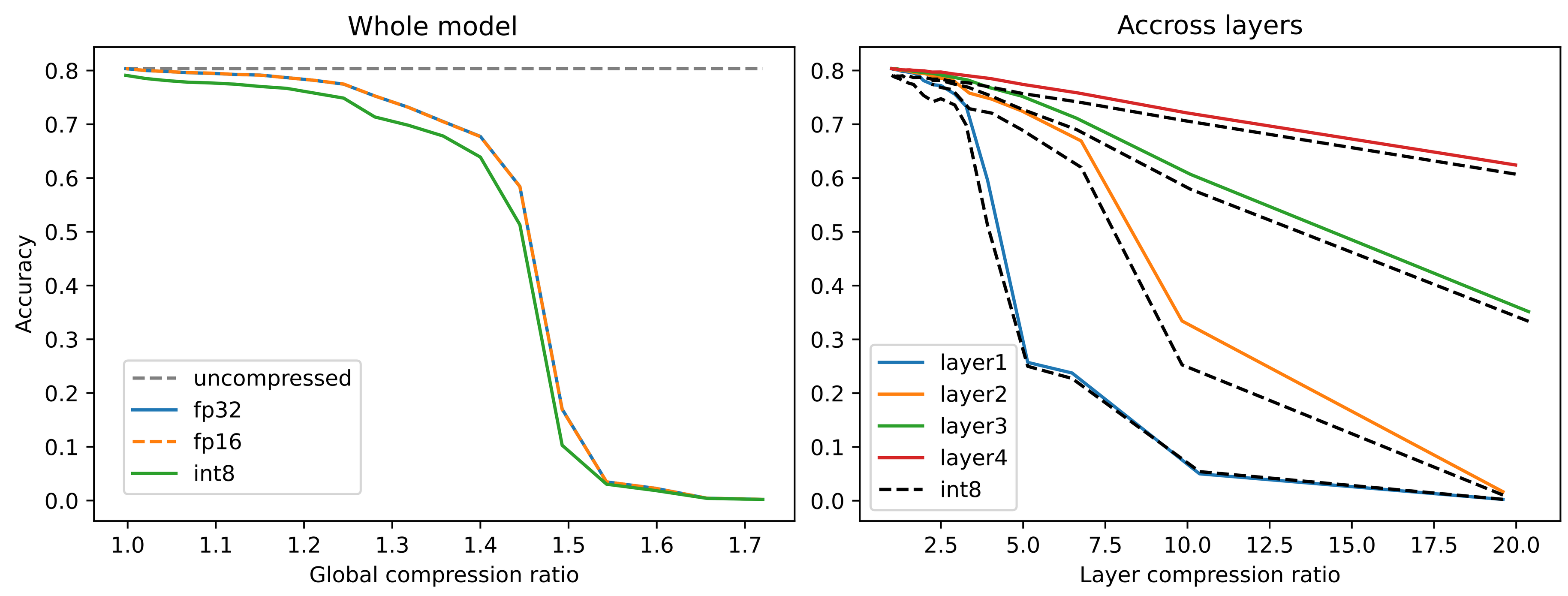

Figure 5: ResNet50 compression on ImageNet using Tucker-2 decomposition. Left: global compression with all convolutions compressed (except the first). Right: per-layer compression when only one layer is decomposed.

CP4

- CP + INT8 fails (fig. 6) due to exploding component ranges
- Corrective methods as EPC [3] (fig. 7) seem to mitigate this effect (not perfect yet).

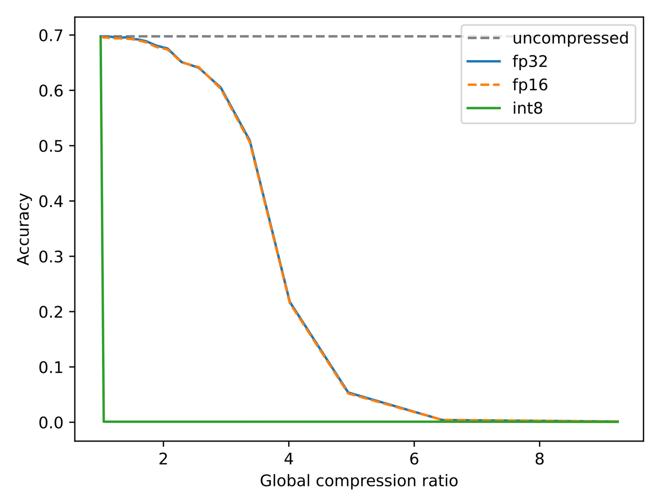

Figure 6: Resnet 18, global CP4 compression

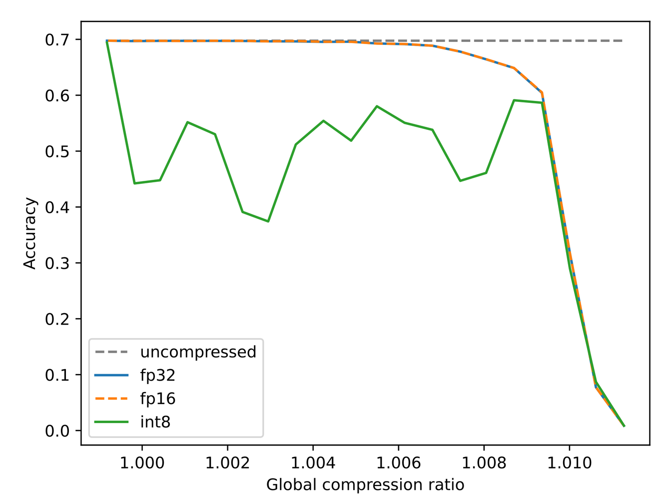

Figure 7: Resnet 18, first layer CP4 with EPC correction

Current & Future work

- **The CP irregularity:** We will continue investigating stable CP methods as EPC [3].
- **Lower precisions:** We need to investigate lower bit quantization (e.g. INT4).
- **Joint optimization:** We only evaluate post-training pipelines. Quantization-Aware Decomposition [4] and lightweight fine-tuning remain unexplored.
- **Limited to standard CNNs:** Extensions to transformers and attention models are planned.

References

[1] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*, 2014.

[2] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*, 2015.

[3] Anh-Huy Phan, Konstantin Sobolev, Konstantin Sozykin, Dmitry Ermilov, Julia Gusak, Petr Tichavský, Valeriy Glukhov, Ivan Oseledets, and Andrzej Cichocki. Stable low-rank tensor decomposition for compression of convolutional neural network. In *European Conference on Computer Vision*, pages 522–539. Springer, 2020.

[4] Daria Cherniuk, Stanislav Abukhovich, Anh-Huy Phan, Ivan Oseledets, Andrzej Cichocki, and Julia Gusak. Quantization aware factorization for deep neural network compression. *Journal of Artificial Intelligence Research*, 81:973–988, 2024.