Tensor Methods for Neural Network Compression

Lokesh Veeramacheneni

Fraunhofer SCAI

M.Sc Moritz Wölter

Prof. Dr. Jochen Garcke

Table of contents

What is a Tensor?

- Tensor is a d-dimensional array

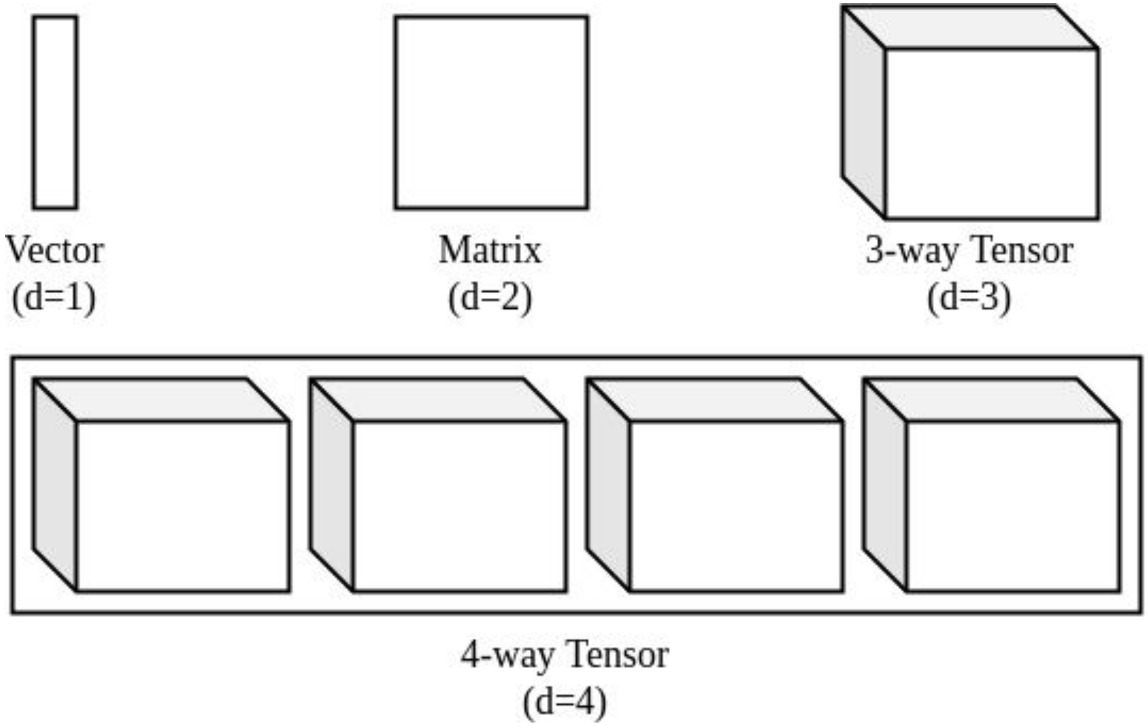
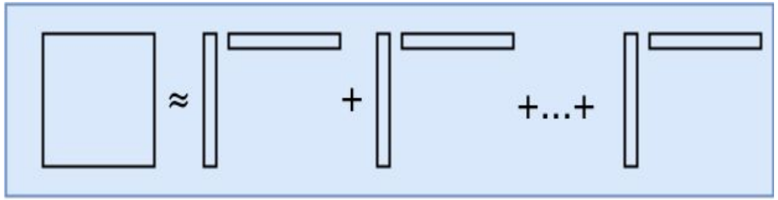


Figure 1: Representation of tensor upto four dimensions [1]

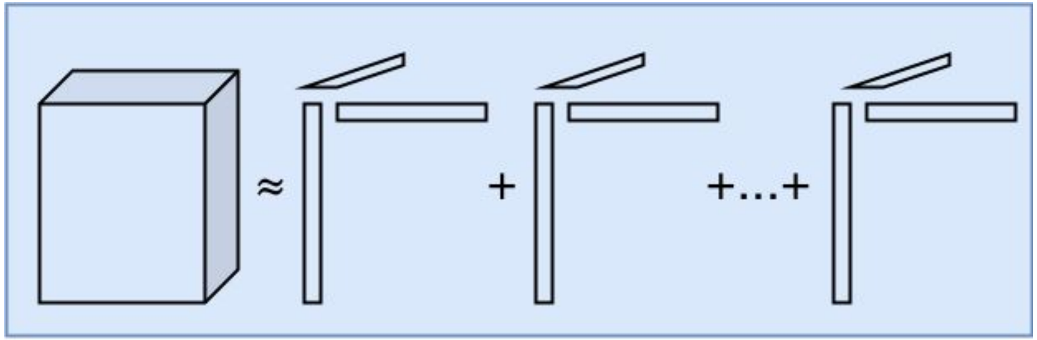
Candecomp/PARAFAC (CP) decomposition

- CP-decomposition can be viewed as matrix SVD generalized to tensors
- Unlike SVD, no orthogonality constraints are required
- It is defined as sum of d-dimensional outer products

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r,$$



Singular Value Decomposition



CP Decomposition

CP Decomposition

- Computed using Alternating Least Squares(ALS) method
- In 3-way decomposition, A, B and C are optimized sequentially [1]
- In ALS, we minimize the cost function ||X-M||

$$\min_{\hat{\mathbf{A}}} \|\mathbf{X}_{(1)} - \hat{\mathbf{A}}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}}\|_F,$$

$$\hat{\mathbf{A}} = \mathbf{X}_{(1)} \left[(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}}\right]^{\dagger}.$$

Equation representing the optimization of variables in ALS method [1]

CP Decomposition

- Input tensor shape (2 x 3 x 3)
- Approximation by various ranks

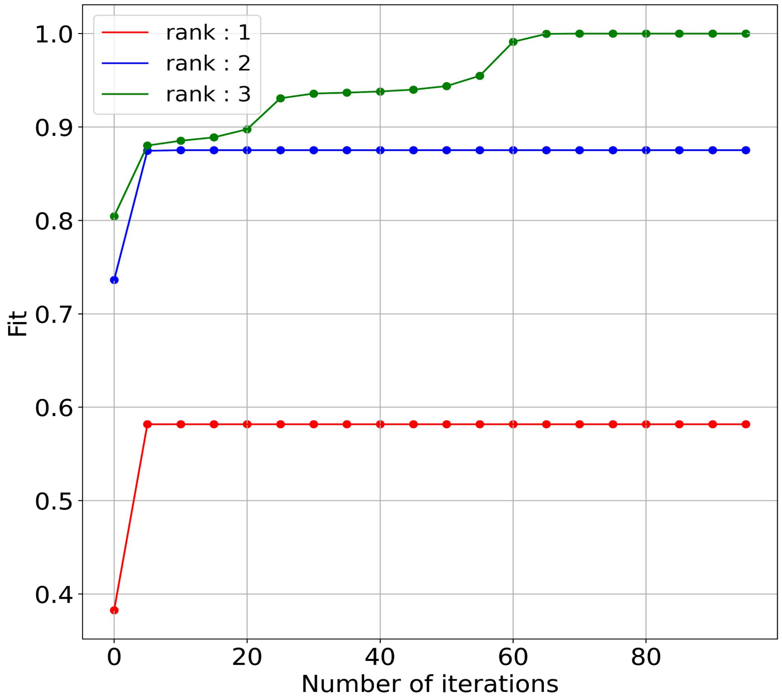


Figure 2: Convergence of ALS method with different ranks

CP Decomposition

- Unit tested with tensorly implementation
- Parameters
  - Rank : 2
  - Maximum iterations : 100

```
[[[ 0.0944,  0.5825, -0.9335],
  [ 0.2137,  0.3705,  0.1285],
  [ 1.5734, -0.7498, -1.0312]],

 [[ 0.4535, -0.6247, -0.9851],
  [ 0.5228,  0.5664,  0.4380],
  [-0.4044, -1.4392, -0.5378]]]
```

Input Tensor

```
[[[ 0.0988,  0.5762, -0.9241],
  [ 0.2404,  0.3301,  0.1888],
  [ 1.5858, -0.7666, -1.0045]],

 [[ 0.4423, -0.6093, -1.0089],
  [ 0.4388,  0.6829,  0.2543],
  [-0.4394, -1.3917, -0.6140]]]
```

Reconstruction from ALS method

```
[[[ 0.0984,  0.5765, -0.9252],
  [ 0.2398,  0.3303,  0.1848],
  [ 1.5857, -0.7668, -1.0050]],

 [[ 0.4426, -0.6091, -1.0080],
  [ 0.4388,  0.6840,  0.2571],
  [-0.4398, -1.3908, -0.6143]]]
```

Reconstruction from tensorly

AlexNet

- Convolutional Neural Network (CNN)
- First five convolutional layers
- Last three fully connected layers act as classifier

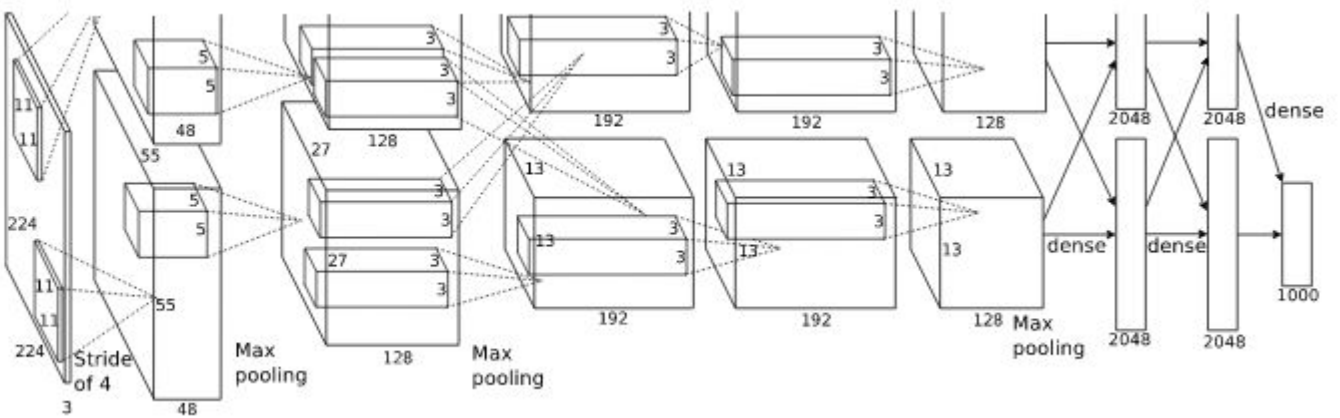


Figure 3: AlexNet architecture [2]

AlexNet

- Convolutional Neural Network (CNN)
- First five convolutional layers
- Last three fully connected layers act as classifier

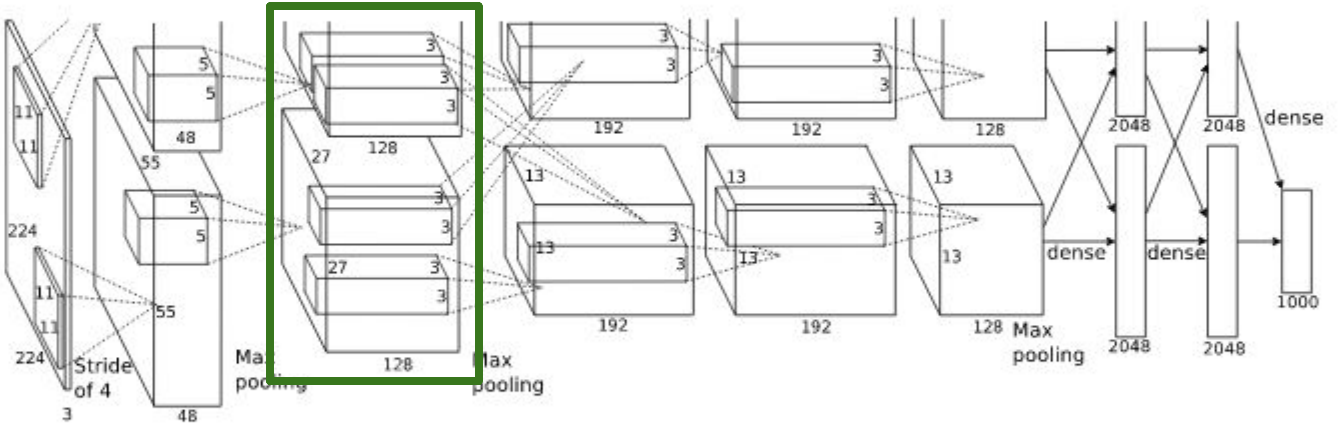


Figure 3: AlexNet architecture [2]

AlexNet Results

- Classification task on CIFAR 10 dataset
- Fine tuning over 10 epochs
- Learning rate : 0.002
- Rank : 45
- Maximum iterations : 100

	Without CP Decomposition	With CP Decomposition
Classification accuracy	84%	65%

References

[1] Kolda, Tamara G., and Brett W. Bader. "Tensor decompositions and applications." *Society for Industrial and Applied Mathematics (SIAM) review* 51, no. 3 (2009): 455-500.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, pp. 1097-1105. 2012.