

Tensor Methods for Neural Network Compression

Lokesh Veeramacheneni

M.Sc Moritz Wölter

Prof. Dr. Jochen Garcke

Fraunhofer SCAI

Table of contents

1. Tensors
2. Tensor unfolding
3. Candecomp/PARAFAC decomposition
4. AlexNet architecture
5. CP decomposition on AlexNet
6. Results

What is a Tensor?

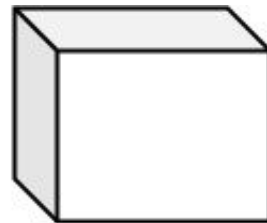
- Tensor is a d -dimensional array



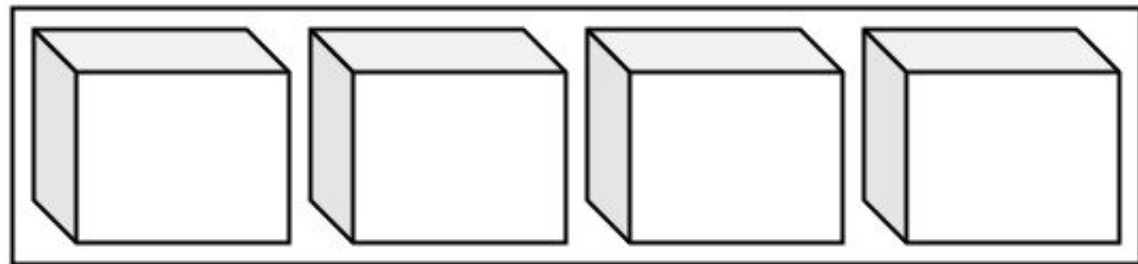
Vector
($d=1$)



Matrix
($d=2$)



3-way Tensor
($d=3$)

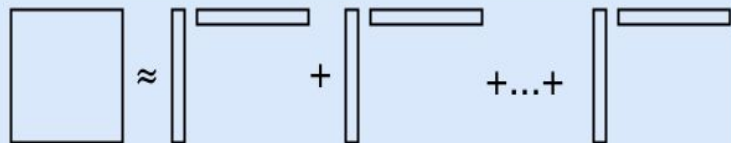


4-way Tensor
($d=4$)

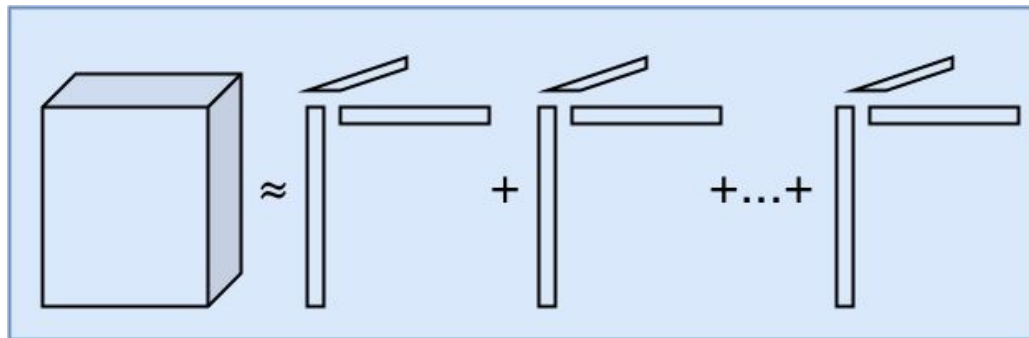
Candecomp/PARAFAC (CP) decomposition

- 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
- In ALS, we minimize the cost function $\|X-M\|$

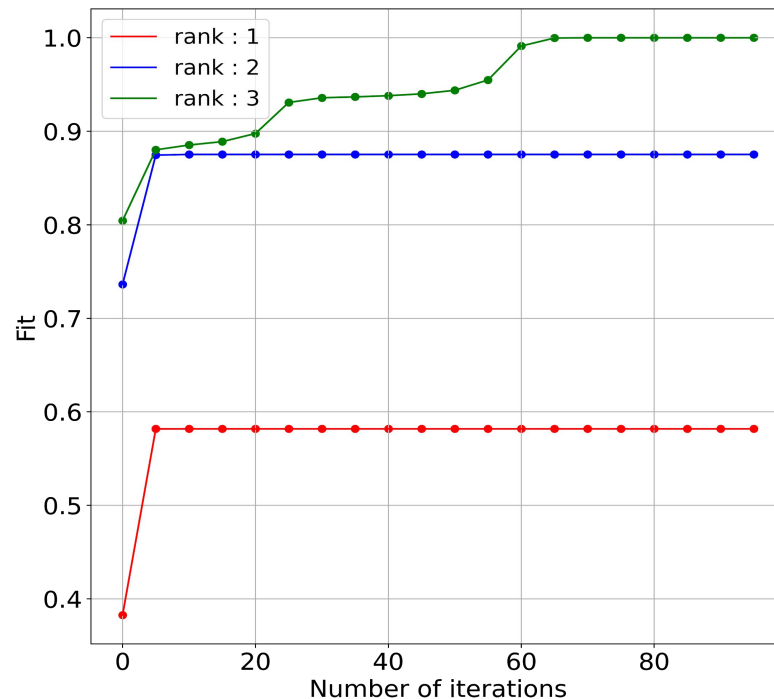
$$X_{(1)} \approx A(C \odot B)^T,$$

$$X_{(2)} \approx B(C \odot A)^T,$$

$$X_{(3)} \approx C(B \odot A)^T.$$

CP Decomposition

- Input tensor shape (2 x 3 x 3)
- Approximation by various 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]],
```

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

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

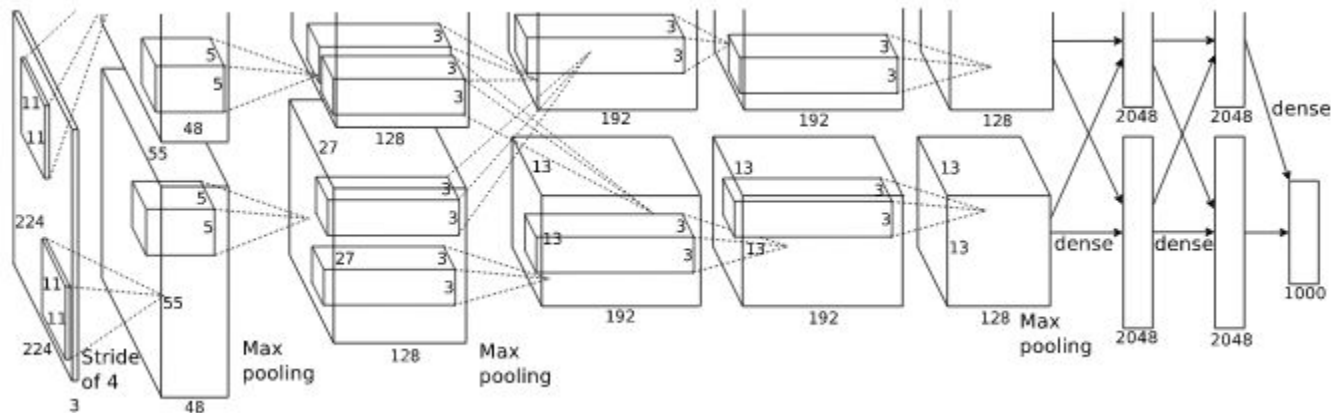
```
[[ 0.4535, -0.6247, -0.9851],  
 [ 0.5228,  0.5664,  0.4380],  
 [-0.4044, -1.4392, -0.5378]]]
```

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

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

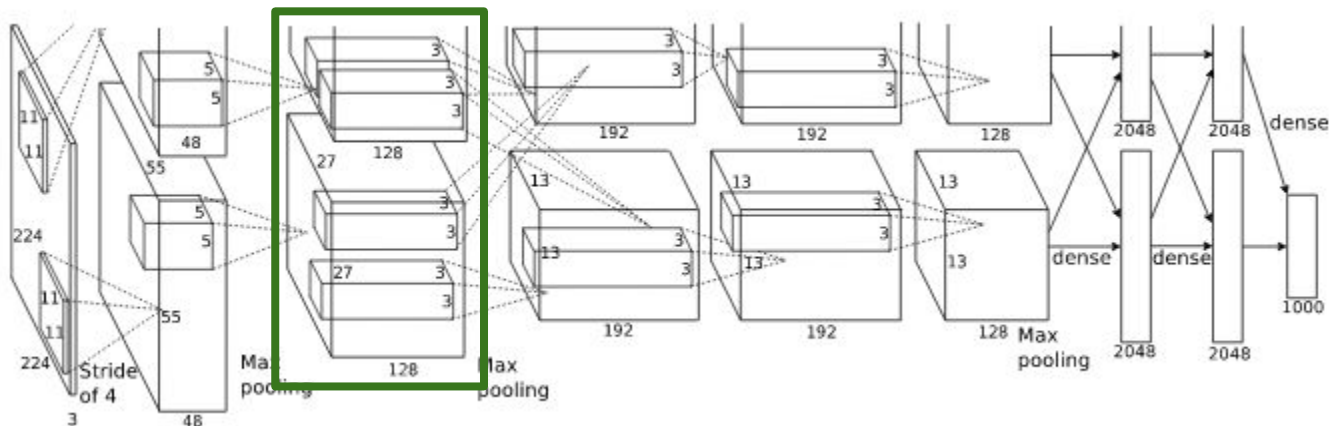
AlexNet

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



AlexNet

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



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
Accuracy	84%	65%