Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges.

> Cheng et al. Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges. IEEE Signal Processing Magazine 2018.

Parameter pruning and sharing

Quantization and binarization

- Methods: k-means, n-bits
- Drawbacks:
  - Lowered on large CNNs
  - Need Special bp method

Pruning and sharing

Designing the structural matrix

- Use special pattern matrix, for example:

$$R = \begin{bmatrix} r_1 & r_2 & \cdots & r_{n-1} & r_n \\ r_n & r_1 & \cdots & r_{n-2} & r_{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_3 & r_4 & \cdots & r_1 & r_2 \\ r_2 & r_3 & \cdots & r_n & r_1 \end{bmatrix}$$

The multiplication operation of this matrix can be accelerate by FFT.

- Drawbacks: Might bring **bias** to model

Low-rank factorization and sparsity

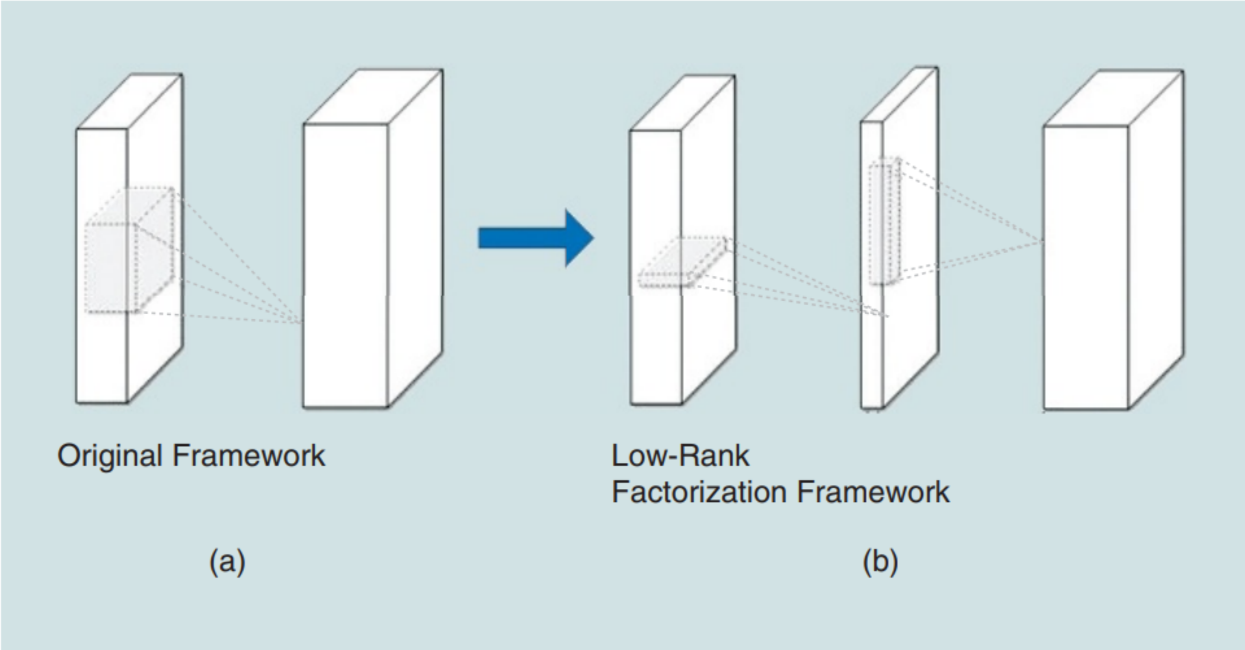

FIGURE 2. A typical framework of the low-rank regularization method. (a) is the original convolutional layer, and (b) is the low-rank constraint convolutional layer with rank-K.

- Canonical polyadic (CP) decomposition
- BN: Batch Normalization

Result

Model	TOP-5 Accuracy	Speedup	Compression Rate
AlexNet	80.03%	1	1
BN low-rank	80.56%	1.09	4.94
CP low-rank	79.66%	1.82	5
VGG-16	90.60%	1	1
BN low-rank	90.47%	1.53	2.72
CP low-rank	90.31%	2.05	2.75
GoogleNet	92.21%	1	1
BN low-rank	91.88%	1.08	2.79
CP low-rank	91.79%	1.20	2.84

Transferred/compact convolutional filters

Let T be transform matrix. Try to approximate T'.

$$T'\Phi(x) = \Phi(Tx)$$

KD

Using Other DNN to approximate **teacher** DNN.

- Monte Carlo Teacher