Jian Cheng, Pei-song Wang, Gang Li, Qing-hao Hu, Han-qing Lu, 2018. Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology & Electronic Engineering*, 19(1): 64-77.
https://doi.org/10.1631/FITEE.1700789

Recent advances in efficient computation of deep convolutional neural networks

Key words: Deep neural networks; Acceleration; Compression; Hardware accelerator

Corresponding author: Jian CHENG
E-mail: jcheng@nlpr.ia.ac.cn
ORCID: http://orcid.org/0000-0003-1289-2758

Motivation

- Deep neural networks have evolved remarkably over the past few years and they are currently the fundamental tools of many intelligent systems.
- DNN-based methods are both computational-intensive and resource-consuming, which hinders the application of these methods on embedded systems like smart phones.
- Network acceleration and hardware implementation have become more and more important.
- We want to provide a comprehensive survey of network acceleration, compression, and accelerator design from both algorithm and hardware points of view.

Reviewed Methods

- Network pruning
- Low-rank approximation
- Network quantization
- Teacher-student networks
- Compact network design
- Hardware accelerators

Future Trends

1. Scalable (self-adaptive) compression.
2. Network acceleration for object detection and other computer vision tasks.
3. Hardware-software co-design.

Conclusions

- DNNs provide impressive performance while suffering from a huge computational complexity and a high energy expenditure.
- In this paper, we provide a survey of recent advances in efficient processing of DNNs from both the algorithm and hardware points of view.
- In addition, we pointed out a few topics that deserve further investigations in the future.