Wu, Bijiao; Wang, Dingheng; Zhao, Guangshe; Deng, Lei; Li, Guoqi

Hybrid tensor decomposition in neural network compression. (English) Zbl 1475.68325

Summary: Deep neural networks (DNNs) have enabled impressive breakthroughs in various artificial intelligence (AI) applications recently due to its capability of learning high-level features from big data. However, the current demand of DNNs for computational resources especially the storage consumption is growing due to that the increasing sizes of models are being required for more and more complicated applications. To address this problem, several tensor decomposition methods including tensor-train (TT) and tensor-ring (TR) have been applied to compress DNNs and shown considerable compression effectiveness.

In this work, we introduce the hierarchical Tucker (HT), a classical but rarely-used tensor decomposition method, to investigate its capability in neural network compression. We convert the weight matrices and convolutional kernels to both HT and TT formats for comparative study, since the latter is the most widely used decomposition method and the variant of HT. We further theoretically and experimentally discover that the HT format has better performance on compressing weight matrices, while the TT format is more suited for compressing convolutional kernels. Based on this phenomenon we propose a strategy of hybrid tensor decomposition by combining TT and HT together to compress convolutional and fully connected parts separately and attain better accuracy than only using the TT or HT format on convolutional neural networks (CNNs). Our work illuminates the prospects of hybrid tensor decomposition for neural network compression.

MSC:
68T07 Artificial neural networks and deep learning
15A23 Factorization of matrices
15A69 Multilinear algebra, tensor calculus

Keywords:
neural network compression; hybrid tensor decomposition; hierarchical Tucker; tensor-train; balanced structure

Software:
LOBPCG; htucker; ImageNet; MNIST; CIFAR

Full Text: DOI

References:


[18] Hinton, G.; Vinyals, O.; Dean, J., Distilling the knowledge in a neural network, (NIPS deep learning workshop (2014))

[19] Hitchcock, F. L., The expression of a tensor or a polyadic as a sum of products, Studies in Applied Mathematics, 6, 1, 164-189 (1927) - Zbl 53.0095.01


[26] Krizhevsky, A., Learning multiple layers of features from tiny images (2009), Computer Science Department, University of Toronto, (Master's thesis)


[31] Liu, J.; Luo, J.; Shah, M., Recognizing realistic actions from videos, (IEEE conference on computer vision and pattern recognition (2009))


of the international conference on learning representation.


[44] Vanhoucke, V.; Mao, M., Improving the speed of neural networks on CPUs, (NIPS deep learning and unsupervised feature learning workshop (2011))


[50] Zhao, Q.; Sugiyama, M.; Cichocki, A., Learning efficient tensor representations with ring-structured networks, (IEEE international conference on acoustics, speech and signal processing (2019)), 8608-8612


This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.