Tensorial Neural Networks: Generalization of Neural Networks and Application to Model Compression

Furong Huang, University of Maryland

We propose tensorial neural networks (TNNs), a generalization of existing neural networks by extending tensor operations on low order operands to those on high order operands. The problem of parameter learning is challenging, as it corresponds to hierarchical nonlinear tensor decomposition. We propose to solve the learning problem using stochastic gradient descent and drive the nontrivial backpropagation rules using the generalized tensor algebra we defined. Our proposed TNNs has three advantages over existing neural networks: (1) TNNs naturally apply to high order input object and thus preserve the multi-dimensional structure in the input, as there is no need to flatten the data. (2) TNNs interpret the design of existing neural network architectures. (3) Mapping a neural network to TNNs with the same expressive power results in a TNN of fewer number of parameters. TNN based compression of neural network improves existing low-rank approximation based compression methods as TNNs exploit two other types of invariant structures, periodicity and modulation, in addition to the low rankness. Experiments on LeNet-5 (MNIST), ResNet-32 (CIFAR10) and ResNet-50 (ImageNet) demonstrate that our TNN based compression outperforms (5% test accuracy improvement universally on CIFAR10) the state-of-the-art low-rank approximation based compression methods under the same compression rate, besides achieving orders of magnitude faster convergence rates due to the efficiency of TNNs.