Abstract

Compression of convolutional neural networks using tensor Decomposition methods

The doctoral dissertation is focused on compressing convolutional neural networks (CNNs). Due to the fact that CNNs weights are mathematically represented as a 4-th-order tensor, it is possible to compress them by using tensor decomposition methods. In such a way, the original convolution is approximated with a pipeline of smaller operations. This dissertation presents novel CNNs compression methods based on various tensor decompositions. The first proposed approach is based on the hierarchical Tucker-2 (HT2) decomposition, which can be seen as a generalization of the Tucker-2 decomposition. In the HT-2 approach, the original convolutional layer is replaced with a sequence of four smaller convolutional layers, two of which are pointwise convolutions and the other two are 1D horizontal and vertical convolutions. The next approach is based on direct tensor-train (TT) decomposition, which by decomposed circularly permuted original weight tensor allows efficient compression of CNNs. In the further part of this work, CNNs compression based on reduced storage direct tensor ring (TR) decomposition was proposed. TR is a generalization of the TT decomposition, and its unique circular permutation-invariant property allows for finding the best low-storage TR decomposition at a given prescribed error, which is particularly suited to the neural network compression problem. In the reduced storage TR approach, the original convolution is approximated with a sequence of two contractions and two 1D convolutions. The last proposed approach describes a general framework for nested compression of CNNs, where the decomposed factors are further decomposed after the fine-tuning stage, which implies larger compression. The experiments carried out in this dissertation confirm the efficiency and applicability of the proposed approaches, which were published in reputable and well-known academic journals and conferences.

Słowa kluczowe: splotowe sieci neuronowe, kompresja sieci neuronowej, sieci tensorowe, modele dekompozycji tensorów.