“**Compressing Deep CNNs using Basis Representation and Spectral Fine-tuning and Sparsity Based Learning**”

Abstract:

We propose an efficient and straightforward method for compressing deep convolutional neural network (CNNs) that uses basis filters to represent the convolutional layers, and optimizes the performance of the compressed network directly in the basis space. Specifically, any spatial convolution layer of the CNN can be replaced by two successive convolution layers: the first is a set of three-dimensional orthonormal basis filters, followed by a layer of one-dimensional filters that represents the original spatial filters in the basis space. While other methods have used low rank analysis to approximate the filters (or their activations), we jointly fine-tune both the basis and the filter representation to directly mitigate any performance loss due to the truncation. We refer to this as spectral fine-tuning since the basis filters and the weights of linear combination are the spectral decomposition of the equivalent spatial filters.

We also employ a minimum L1 norm criteria to auto-compress the network during the training process. We demonstrate the generality of the proposed approach by applying it to several well known deep CNN architectures and data sets for image classification and object detection. We also present the results of direct comparisons with other state-of-the art compression techniques, and show that our approach yields competitive results and performs better in many cases.