Compression of Deep Neural Networks

Abstract:

In spite of their state-of-the-art performance across a wide spectrum of problems, deep neural networks (DNN) have the drawback that they require extensive computational resources for training as well as execution at runtime. This not only results in high energy consumption, but also in slow runtimes and high memory requirements, which greatly limits the adoption of DNNs in industrial applications and their deployment into resource constrained devices, e.g., smartphones, embedded systems or IoT devices. This talk will present DeepCABAC, a recently proposed compression algorithm for deep neural networks that is inspired from quantization and coding techniques used in the field of video coding. Experimental results show that DeepCABAC consistently attains higher compression rates than previously proposed coding techniques for neural network compression. For instance, it is able to compress the VGG16 ImageNet model by x63.6 with no loss of accuracy, thus being able to represent the entire network with merely 8.7MB. Furthermore, the talk will discuss a new data structure, the Compressed Entropy Row (CER) format, for efficiently representing the weight matrices of neural networks. The CER format can be regarded as a generalization of sparse data structures and is shown to be more energy and time efficient under practically relevant assumptions. Moreover, CER allows inference in the compressed domain, i.e. without decompressing the weight matrix, which is especially relevant for devices with limited memory capacity. Experimental results show that CER is a powerful lossless compression scheme for weight matrices, which attains up to x15 compression ratios for state-of-the-art DNN models.