

Compression of convolutional neural networks

Csanád Sándor

Faculty of Mathematics and Computer Science, Babeş-Bolyai University

Robert Bosch SRL

`scsanad@cs.ubbcluj.ro`

Convolutional neural networks (CNNs) are state-of-the-art methods in many computer vision problems such as in image categorization, object detection or image segmentation. However, due to their high memory and computation needs it is hard to use them on different mobile and embedded devices. To tackle this problem pruning can be applied to reduce the network size (by this also reduce the number of multiplications). Most methods prune networks in an unstructured way and store the remaining parameters in sparse matrix format. However, sparse matrix multiplications are less efficient in most deep learning frameworks and storing the matrix indices uses additional memory. Moreover, most pruning methods remove parameters based on their statistics and not considering the loss of the network.

Our method is based on the compressor-critic framework of *DeepIoT* [1]. It uses a compressor network to find redundancies between network parameters and prunes whole filters by setting high dropout probabilities for them. To learn redundancies, the compressor gets as input the weights or feature maps of the target network and prunes filters by considering also the expected loss of the network.

References

- [1] Shuochao Yao, Yiran Zhao, Aston Zhang, Lu Su, and Tarek Abdelzaher. 2017. DeepIoT: Compressing Deep Neural Network Structures for Sensing Systems with a Compressor-Critic Framework. In Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems (SenSys '17), Rasit Eskicioglu (Ed.). ACM, New York, NY, USA, Article 4, 14 pages. DOI: <https://doi.org/10.1145/3131672.3131675>