Compression of Convolutional Neural Networks by Filter Pruning

Csanád Sándor, Szabolcs Pável, Lehel Csató

Faculty of Mathematics and Informatics,
Babeş-Bolyai University, Cluj-Napoca, Romania
{csanad.sandor, szabolcs.pavel, lehel.csato}@cs.ubbcluj.ro

Convolutional neural networks 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 and the number of floating point operations.

Our pruning method iteratively removes filters from convolutional layers based on their importance values. As filter importance, we use the coefficients of a linear model trained on a dataset containing $(\mathbf{z}_i, s_i)$ pairs. Here $\mathbf{z}_i$ denotes which filters are active in the network and s_i defines its score, calculated from the loss when $\mathbf{z}_i$ is applied on the network. We present our experimental results on the ResNet architecture trained on the CIFAR-10 dataset.