Sparse Convolutional Neural Networks

Baoyuan Liu[1], Min Wang[1], Hassan Foroosh[1], Marshall Tappen[2], Marianna Pensky[3],
[1]Computational Imaging Lab, Computer Science, University of Central Florida, Orlando, FL, USA
[2]Amazon.com, Seattle, WA 98109
[3]Department of Mathematics, University of Central Florida, Orlando, FL, USA

Deep neural networks have achieved remarkable performance in both image classification[3] and object detection[1] problems at the cost of large amount of parameters and computational complexity. Results of ILSVRC competitions in recent years have demonstrated a strong correlation between the network size and the classification accuracy. The ILSRVR 2014 submission from VGG[4] builds a network with up to 16 convolutional layers that reduces the top-5 classification error to 7.4%, at the expense of approximately one month of network training with 4 high-end GPUs.

The structure of these networks makes it reasonable to conjecture that there exists significant redundancy in these huge networks. In this paper, we show that this redundancy makes it possible to significantly reduce the amount of computation required to process images by sparse decompositions over the convolutional kernels. Figure 1 shows the basic idea of our decomposition methods. Both inter-channel and intra-channel redundancy are exploit.

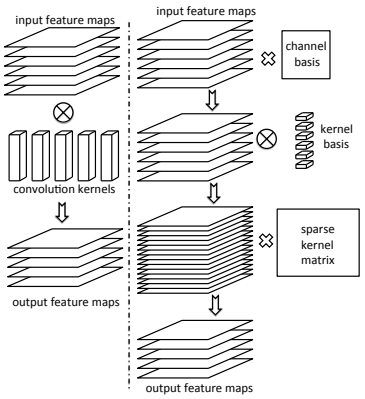
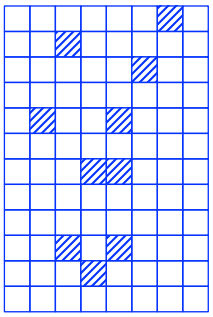


Figure 1: Overview of our sparse convolutional neural network. Left: the operation of convolution layer for classical CNN, which convolves large amount of convolutional kernels with the input feature maps. Right: our proposed SCNN model. We apply two-stage decompositions over the channels and the convolutional kernels, obtaining a remarkably(more than 90%) sparse kernel matrix, and converting the operation of convolutional layer to spare matrix multiplication

We first perform an initial decomposition based on reconstruction error of kernel weights, then fine-tune the network while imposing sparsity constraint. In the fine-tuning phase, we optimize the network training error, the sparsity of convolutional kernels as well as the number of convolutional basis simultaneously by minimizing a sparse group-lasso object function. Surprisingly high sparsity can be achieved in our model. As illustrated in the row for sparsity of Table 1, we are able to zero out more than 90% of the convolutional kernel parameters of the network in [3] with relatively small number of basis while keeping the loss of accuracy less than 1%.

In our Sparse Convolutional Neural Networks (SCNN) model, each sparse convolutional layer can be performed with a few convolution kernels followed by a sparse matrix multiplication. It could be assumed that the sparse matrix formulation naturally leads to highly efficient computation. However, computing sparse matrix multiplication can involve significant overhead that makes it difficult to actually achieve significant acceleration. Thus, we also propose an efficient sparse matrix multiplication algorithm. Based on the fact that the sparse convolutional kernels are fixed after training, we avoid the necessity of indirect and discontinuous memory access by

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.



(a) An example sparse matrix $\mathbf{B}$. The shadowed blocks represent non-zero elements and the blank blocks represent zero elements.

Input:
$\mathbf{A}$: 8×12 dense matrix
$\mathbf{B}$: 12×8 sparse matrix
Output:
$\mathbf{C} = \mathbf{A} \times \mathbf{B}$
Operations:
$\mathbf{c}_7 += \mathbf{a}_1 \times b_{1,7}$
$\mathbf{c}_3 += \mathbf{a}_2 \times b_{2,3}$
$\mathbf{c}_6 += \mathbf{a}_3 \times b_{3,6}$
$\mathbf{c}_2 += \mathbf{a}_5 \times b_{5,2}$
$\mathbf{c}_5 += \mathbf{a}_5 \times b_{5,5}$
$\mathbf{c}_4 += \mathbf{a}_7 \times b_{7,4}$
$\mathbf{c}_5 += \mathbf{a}_7 \times b_{7,5}$
$\mathbf{c}_3 += \mathbf{a}_{10} \times b_{10,3}$
$\mathbf{c}_5 += \mathbf{a}_{10} \times b_{10,5}$
$\mathbf{c}_4 += \mathbf{a}_{11} \times b_{11,4}$

(b) Generated Pseudo code for calculating $\mathbf{C} = \mathbf{A} \times \mathbf{B}$. $\mathbf{c_i}$ is the i^{th} column of $\mathbf{C}$ and $\mathbf{a_j}$ is the j^{th} column of $\mathbf{A}$. $b_{i,j}$ is the element of $\mathbf{B}$ at i^{th} column and j^{th} row

Figure 2: An example that illustrates how our algorithm generates code for calculating a dense matrix and a sparse matrix

layer	conv1	conv2	conv3	conv4	conv5
sparsity%	0.927	0.950	0.951	0.942	0.938
theoretical speedup	2.61	7.14	16.12	12.42	10.77
Actual speedup	2.47	4.52	6.88	5.18	3.92

Table 1: Sparsity, theoretical and actual speedup corresponding to each convolutional layer for our SCNN model. Results demonstrates that our highly sparse model could lead to remarkably acceleration for computation.

encoding the structure of the input sparse matrix into our program as the index of registers. Figure. 2 gives a simple example of how we generate codes from a sparse matrix. As illustrated in the last two rows of Table 1 ,our CPU-based implementation demonstrates much higher efficiency than off-the-shelf sparse matrix libraries and a significant speedup over the original dense network are realized. While convolutional network systems are dominated by GPU-based approaches, advances in CPU-based systems are useful because they can be deployed in commodity clusters that do not have specialized GPU nodes.

We further apply SCNN to object detection problem. We apply SCNN to accelerate the convolutional layers of SPP[2] model. To reduce the running time of fully connected layer, we propose to (1) sparsify the fully connected layer (2) adopt a cascade model composed of the last convolutional layer and the last fully connected layer.

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer Vision–ECCV 2014*, pages 346–361. Springer, 2014.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.