Less is More: Towards Compact CNNs

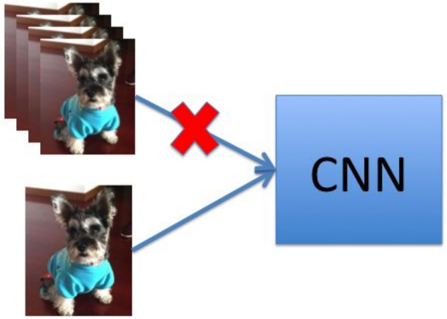
Hao Zhou[1], Jose M. Alvarez[2] and Fatih Porikli[2,3]

[1]University of Maryland, College Park, USA
[2]Data61/CSIRO, Canberra, Australia
[3]Australian National University, Canberra, Australia

Motivation

1. CNNs are very large (Millions of parameters)

2. Large memory footprint



Motivation

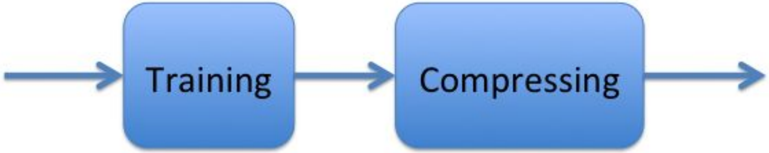
What we did

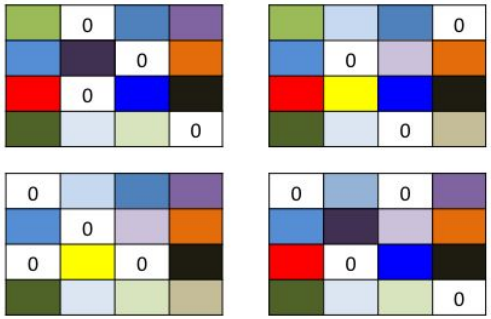
AlexNet: 60M → 14M

VGG: 133M → 74M

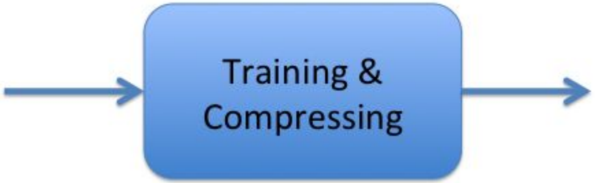
Contributions

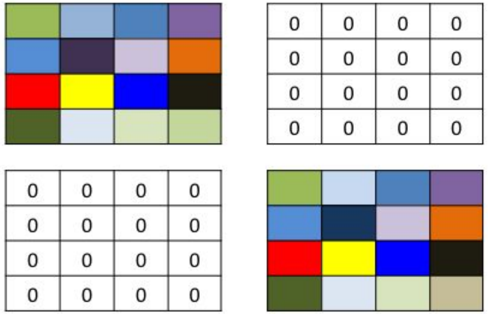
Others





Ours





Our method

Idea: adding sparse constraints to neurons.

$$\min_{\hat{\mathbf{W}}} \psi(\hat{\mathbf{W}}) + g(\hat{\mathbf{W}})$$

Loss for CNNs

Sparse Constraints

Our method

Idea: adding sparse constraints to neurons.

$$\min_{\hat{\mathbf{W}}} \underbrace{\psi(\hat{\mathbf{W}})}_{\text{Loss for CNNs}} + \underbrace{g(\hat{\mathbf{W}})}_{\text{Sparse Constraints}}$$

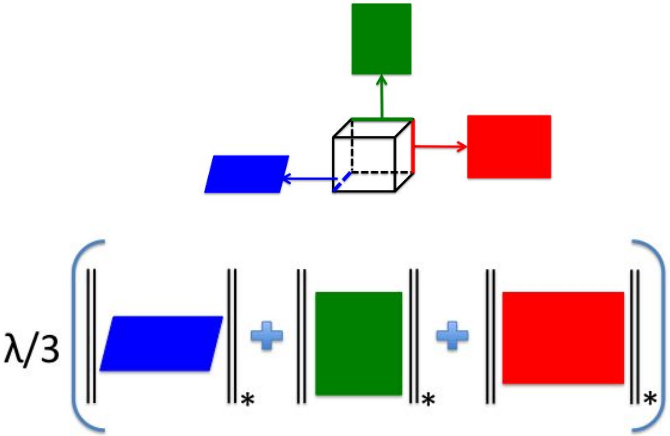
Forward-backward splitting:

$$\rightarrow \text{Forward: Backprop} \qquad \hat{W}^* \leftarrow \hat{W} - \tau \frac{\psi(\hat{W})}{\hat{W}}$$

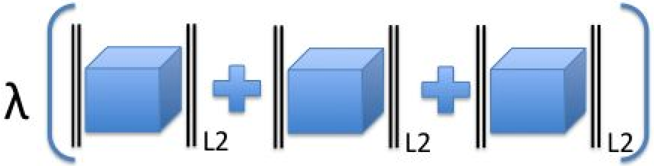
$$\rightarrow \text{Backward: Sparsity} \qquad \hat{W} \leftarrow \arg\min_{\hat{W}} g(\hat{W}) + \frac{1}{2\tau}||\hat{W} - \hat{W}^*||^2$$

Our method — sparse constraints

Sparse Constraints



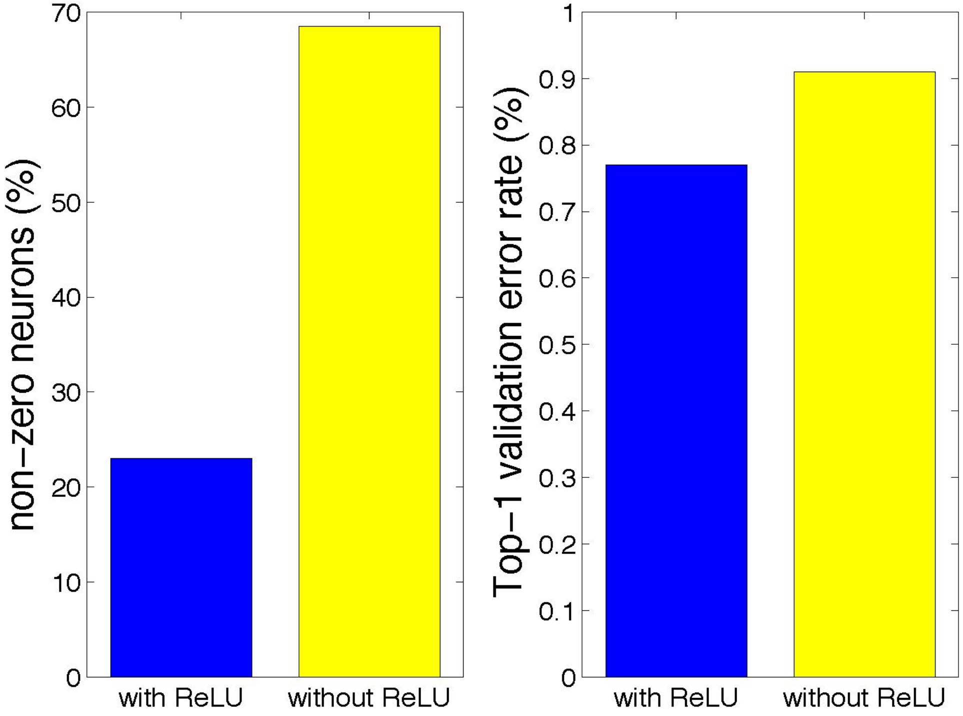
Tensor Low Rank



Group Sparsity

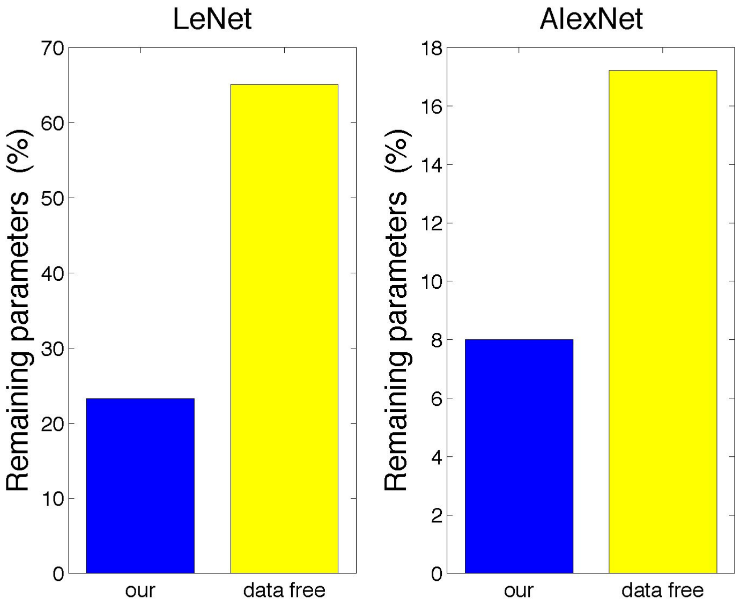
J.~Liu, P.~Musialski, P.~Wonka and J.~Ye: Tensor Completion for Estimating Missing Values in Visual Data. PAMI.(2013)

Experiments — ReLU

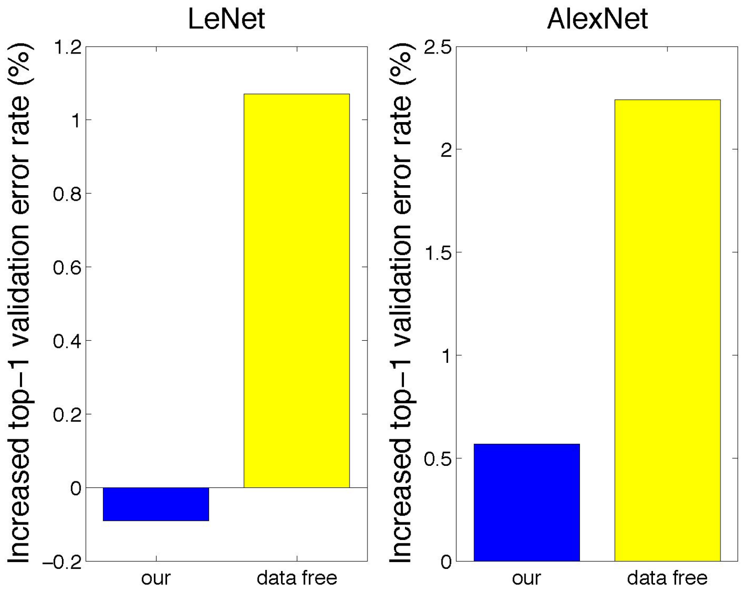


Conv2 on LeNet

Experiments



Non-zero parameters



Increased error rate

Srinivas, S., Babu, R.V.: Data-free parameter pruning for deep neural networks. In: BMVC.(2015)

Comments?
Questions?
Welcome to poster
#09