Cross-Layer Group Regularization for Deep Neural Network Pruning

Shuang Gao, Xin Liu

NVIDIA Corporation, 2788 San Tomas Expressway, Santa Clara, 95051, CA, USA

Introduction

- Applying sparsity regularization to achieve a compact model without loss in quality;
- Model pruning is an efficient method to remove redundant weights.

Previous Works

- Most existing methods only consider structural sparsity within an individual layer.

Our Solution

- Propose a Cross-Layer Group (CLG) regularization considering statistics from multiple layers, and align the weight sparsity across multiple layers.

Cross-Layer Group Regularization

- **Loss**: $L(X,W) = \sum_{i=1}^{N} E\left(y_i, f(x_i, W)\right) + \lambda R(W)$ **Reg Loss**: $R = \lambda \sum_{S \in G_M} R_S$ *with* $R_s = \sum_{i=0}^{k} \sqrt{p_i}\, \|W_i\|_2$

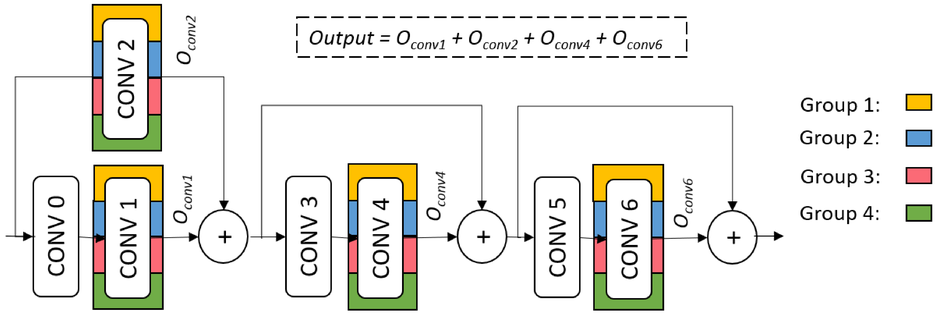

Figure 2: An example of layer group definition

Weight Sparsity Mismatch across Layers

- Multiple layers are connected by element-wise operations in residual blocks of ResNet [1].
- Propose to align weight sparsity across connected layers.

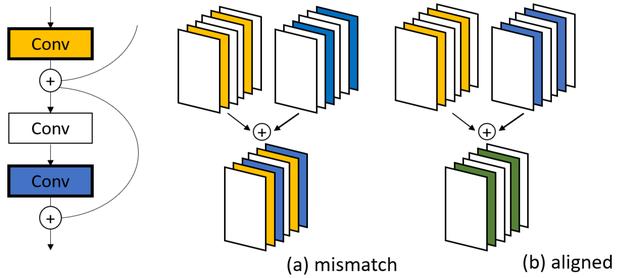

Figure 1: The proposed method reduces the sparsity mismatch among layers.

Sparsity Alignment Analysis

- The weights sparsity across multiple layers are well aligned with the proposed approach; whereas mismatched by L1 regularization.

Figure 3: Weights sparsity comparison between L1 regularization (top) and the proposed method (bottom)

Experimental Results

Layer Number	Method	Parameter Number	Accuracy	Pruned Ratio
56	ResNet v1	0.85M	93.0%	--
	Li's method [2]	0.73M	93.1%	13.7%
	ResNet v2[a]	1.67M	92.9%	--
	L1 regularization v2[a]	0.69M	93.6%	58.5%
	CLG-L1_A v2[a]	**0.44M**	**93.4%**	**74.0%**
	CLG-L1_B v2[a]	**0.24M**	**93.0%**	**85.7%**
110	ResNet v1	1.72M	93.5%	--
	Li's method [2]	1.16M	93.3%	32.4%
	ResNet v2[a]	3.32M	93.7%	--
	L1 regularization v2[a]	0.84M	93.5%	74.7%
	CLG-L1_A v2[a]	**0.46M**	**93.7%**	**86.1%**
	CLG-L1_B v2[a]	**0.32M**	**93.4%**	**90.4%**

[a] stands for v2 with bottleneck.

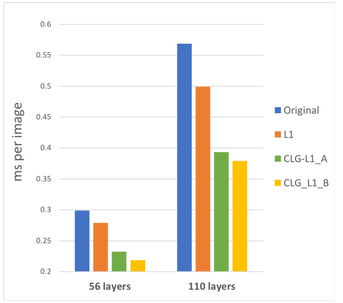

Figure 4: Runtime comparison

- >=74% less parameters, up to 50% speedup compared to original ResNet model;
- 37%-65% less parameters, up to 32% speedup compared to L1 regularization;
- 40% to 72% less parameters compared to Li's[2] pruning method.

[1] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision (ECCV), 2016.
[2] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. Pruning filters for efficient convnets. In International Conference on Learning Representations (ICLR), 2017.