Can We Gain More from Orthogonality Regularizations in Training Deep CNNs?

Nitin Bansal Xiaohan Chen Zhangyang Wang

Department of Computer Science and Engineering, Texas A&M University

Overview

- We develop novel orthogonality regularizations on training deep CNNs, by borrowing ideas and tools from sparse optimization.
- These plug-and-play regularizations can be conveniently incorporated into training almost any CNN without extra hassle.
- The proposed regularizations can consistently improve the performances of baseline deep networks on CIFAR-10/100, ImageNet and SVHN datasets, based on intensive empirical experiments, as well as accelerate/stabilize the training curves.
- The proposed orthogonal regularizations outperform existing competitors.

Preliminaries

Goal We aim to regularize the (overcomplete or undercomplete) CNN weights to be "close" to orthogonal ones, for improving both training stability and final accuracy.

Notation The weight in one fully-connected layer is denoted as $W \in \mathbb{R}^{m \times n}$. For convolutional layer $C \in \mathbb{R}^{S \times H \times C \times M}$, we reshape C into $W' \in \mathbb{R}^{m' \times n'}$ where $m' = S \times H \times C$ and $n' = M$ to reduce it to the form of fully-connected layer.

Mutual Coherence The mutual coherence of a weight W is defined as

$$\mu_W = \max_{i \neq j} \frac{|\langle w_i, w_j \rangle|}{||w_i|| \cdot ||w_j||}, \tag{1}$$

where w_i denotes the i-th column of W, $i = 1, 2, ..., n$. In order for W to have orthogonal or near-orthogonal columns, μ_W should be as low as possible (zero if $m \geq n$).

Restricted Isometry Property We rewrite the Restricted Isometry Property condition of W as:

$$\delta_W = \sup_{z \in \mathbb{R}^n, z \neq 0} \left| \frac{||Wz||^2}{||z||^2} - 1 \right|, \tag{2}$$

where z is k-sparse. Note that δ_W reduces to the spectral norm of $W^TW - I$, denoted as $\sigma(W^TW - I)$, if we let $k = n$.

Orthogonality Regularization

Soft Orthogonality Regularization (SO) SO simply minimizes the distance from the Gram matrix of W to the identity matrix:

$$\text{(SO)} \qquad \lambda ||W^TW - I||_F^2, \tag{3}$$

Double Soft Orthogonality Regularization (DSO) DSO tries to regularize better when W is overcomplete, by appending another term to (3).

$$\text{(DSO)} \qquad \lambda(||W^TW - I||_F^2 + ||WW^T - I||_F^2). \tag{4}$$

Mutual Coherence Regularization (MC) We suppress μ_W to enforce orthogonality. Assuming columns of W are normalized to unit vectors (*what if not?*), we propose the following MC regularization based on (1):

$$\text{(MC)} \qquad \lambda ||W^TW - I||_\infty, \tag{5}$$

Spectral Restricted Isometry Property Regularization (SRIP) We suppress σ_W to enforce orthogonality, and propose the following SRIP regularization based on (2):

$$\text{(SRIP)} \qquad \lambda \cdot \sigma(W^TW - I). \tag{6}$$

Power Methods for Efficient SRIP Implementation To avoid the computationally expensive EVD, we approximate the computation of spectral norm using the truncated power iteration method. Starting with a randomly initialized $v \in \mathbb{R}^n$, we iteratively perform the following procedure a small number of times (2 times by default) :

$$u \leftarrow (W^TW - I)v, v \leftarrow (W^TW - I)u, \sigma(W^TW - I) \leftarrow \frac{||v||}{||u||}. \tag{7}$$

Links

arXiv preprint:

Source Codes:

Experimental Results

- We perform our experiments on several most popular state-of-the-art models: ResNet(including several different variants), Wide ResNet and ResNext. Datasets include CIFAR-10, CIFAR-100, SVHN and ImageNet.
- All results endorse the advantages of orthogonality regularization in improving the final accuracies: evident, stable, reproducible, and sometimes with a large margin. SRIP is the best among all, and incurs negligible extra computational load.

Table 1: Top-1 error rate comparison by ResNet 110, Wide ResNet 28-10 and ResNext 29-8-64 on CIFAR-10 and CIFAR-100. * indicates results by us running the provided original model.

Model	Regularizer	CIFAR-10	CIFAR-100
ResNet-110	None	7.04*	25.42*
	SO	6.78	**25.01**
	DSO	7.04	25.83
	MC	6.97	25.43
	SRIP	**6.55**	25.14
Wide ResNet 28-10	None	4.16*	20.50*
	SO	3.76	18.56
	DSO	3.86	18.21
	MC	3.68	18.90
	SRIP	**3.60**	**18.19**
ResNext 29-8-64	None	3.70*	18.53*
	SO	3.58	17.59
	DSO	3.85	19.78
	MC	3.65	17.62
	SRIP	**3.48**	**16.99**

Table 2: Top-5 error rate comparison on ImageNet.

Model	Regularizer	ImageNet
ResNet 34	None	9.84
	OMDSM	9.68
	SRIP	**8.32**
Pre-Resnet 34	None	9.79
	OMDSM	9.45
	SRIP	**8.79**
ResNet 50	None	7.02
	SRIP	**6.87**

Table 3: Top-1 error rate on SVHN using Wide ResNet 16-8.

Regularizer	ImageNet
None	1.63
SRIP	**1.56**

Effects on the Training Process

We carefully inspect the training curves (in term of validation accuracies w.r.t epoch numbers) of different methods on CIFAR-10 and CIFAR-100, with ResNet-110 curves shown here. Top: CIFAR-10; Bottom: CIFAR-100.