Norm Loss: An efficient yet effective regularization method for deep neural networks

Theodoros Georgiou†, Sebastian Schmitt‡, Wei Chen†, Thomas Bäck† and Michael Lew†

†Leiden Institute of Advanced Computer Science

‡Honda Research Institute Europe GmbH

Introduction

CNNs can suffer from diverse issues, such as:

- Exploding, vanishing gradients
- Scaling-based weight space symmetry
- Covariant-shift

We propose a **weight soft-regularization method based on the Oblique manifold**.

Related Work

Weight regularization

- Weight decay
- Weight normalization
- Weight orthogonalization

Activation normalization

- Batch normalization
- Group normalization
- Kalman normalization

Preliminaries

Given weight vector:

$$\mathbf{W} \in \mathbb{R}^{n \times p}, \tag{3}$$

p is the dimensionality of each weight vector of a filter, while n is the number of filters.

The Oblique manifold defines:

$$ddiag(\mathbf{W}\mathbf{W}^T) = I \tag{4}$$

Proposed method, the Norm Loss (nl)

$$L_{nl} = \sum_{c_o=1}^{C_o} \left(1 - \sqrt{\sum_{c_i=1}^{C_i}\sum_{i=1}^{F_h}\sum_{j=1}^{F_w} w_{ijc_ic_o}^2}\right)^2 \tag{1}$$

where F_w, F_h, C_i, C_o are the filter (or weight vector) width, height, number of input and number of output channels respectively. The loss is penalizing the weight vector of each neuron if its Euclidean norm is different from one. The final loss function then becomes:

$$L_{total} = L_{target} + \lambda_{nl} \cdot L_{nl} \tag{2}$$

Effect: Slowly steer the weight vectors to unit norm.

Connection to WD

Weight decay update rule:

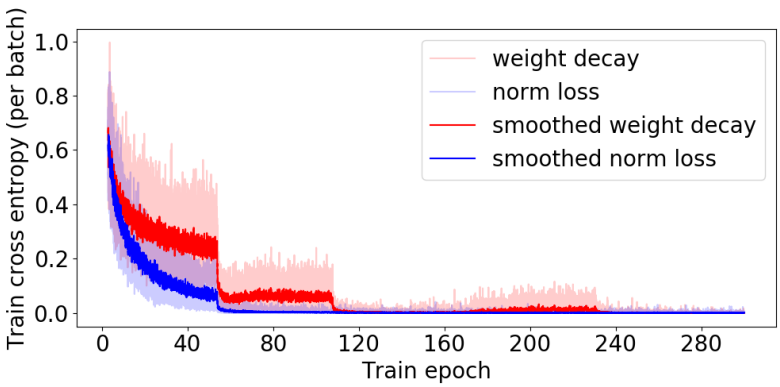
$$\frac{\partial L_{wd}}{\partial w_{ijc_ic_o}} = 2w_{ijc_ic_o} \tag{5}$$

Norm Loss update rule:

$$\frac{\partial L_{nl}}{\partial w_{ijc_ic_o}} = 2w_{ijc_ic_o}\left(1 - \frac{1}{\|w_{c_o}\|}\right) \tag{6}$$

The Norm Loss can be seen as an extension of the weight decay where the weight decay factor and its sign are regulated during training by the norm of the weight vector.

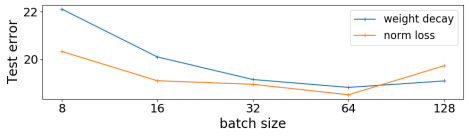
Cross entropy evolution (CIFAR-10)



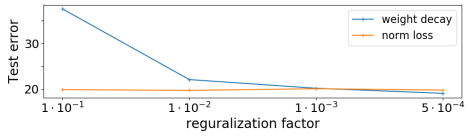
ImageNet (ResNet50)

regul.	Top-1 error	Top-5 error
wd (repr)	25.29	7.86
nl (Ours)	24.34	7.44
wd	23.85	-
ONI	23.30	-

Varying batch size



Varying reg. factor



CIFAR-100

model	regul.	error
ResNet110	wd (repr)	27.9 (28.398)
ResNet110	WN	- (28.38)
ResNet110	PBWN	- (27.03)
ResNet110	nl (Ours)	26.24 (**26.526**)
WRN-28-10	wd (repr)	18.85 (19.138)
WRN-28-10	wd	- (18.85)
WRN-28-10	OLM	- (18.76)
WRN-28-10	OLM-L1	- (**18.61**)
WRN-28-10	nl (Ours)	18.57 (18.648)

CIFAR-10

model	regul.	error
ResNet110	wd (repr)	6.32 (6.568)
ResNet110	wd	6.43 (6.61)
ResNet110	WN	- (7.56)
ResNet110	PBWN	- (6.27)
ResNet110	ONI	- (6.56)
ResNet110	nl (Ours)	5.9 (**5.996**)
WRN-28-10	wd (repr)	3.9 (3.966)
WRN-28-10	wd	- (3.89)
WRN-28-10	OLM	- (**3.73**)
WRN-28-10	nl (Ours)	4.47 (4.662)

Conclusions

Norm Loss:

- Comparable and sometimes better performance to the state of the art on popular architectures and benchmarks
- Lower computational complexity than most weight regularization methods
- High convergence speed
- Less sensitive to hyper parameters such as batch size and regularization factor