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

Theodoros Georgiou.

Universiteit Leiden

Honda Research Institute EU

# Introduction



CNNs can suffer from diverse issues, such as:

- exploding, vanishing gradients
- scaling-based weight space symmetry
- covariant-shift

# Existing approaches



Weight regularization

- weight decay
- weight normalization
- weight orthogonalization
- ...

# Existing approaches



Weight regularization

- weight decay
- weight normalization
- weight orthogonalization
- ...

Activation normalization

- Batch normalization
- Group normalization
- Kalman normalization
- ...

# Oblique manifold



Given weight vector:

$$\mathbf{W} \in \mathbb{R}^{n \times p}, \tag{1}$$

$p$ is the dimensionality of each weight vector of a filter, while $n$ is the number of filters. The Oblique manifold defines:

$$ddiag(\mathbf{W}\mathbf{W}^T) = I \tag{2}$$

# Proposed method - Norm Loss



$$L_{nl} = \sum_{c_o=1}^{C_o} \left( 1 - \sqrt{\sum_{c_i=1}^{C_i} \sum_{i=1}^{F_h} \sum_{j=1}^{F_w} w_{ijc_i c_o}^2} \right)^2 \tag{3}$$

The loss is penalizing the weight vector of each neuron if its Euclidean norm is different from one.

$$L_{total} = L_{target} + \lambda_{nl} \cdot L_{nl} \tag{4}$$

Effect: Slowly steer the weight vectors to unit norm.

# Connection to weight decay



Weight decay update rule:

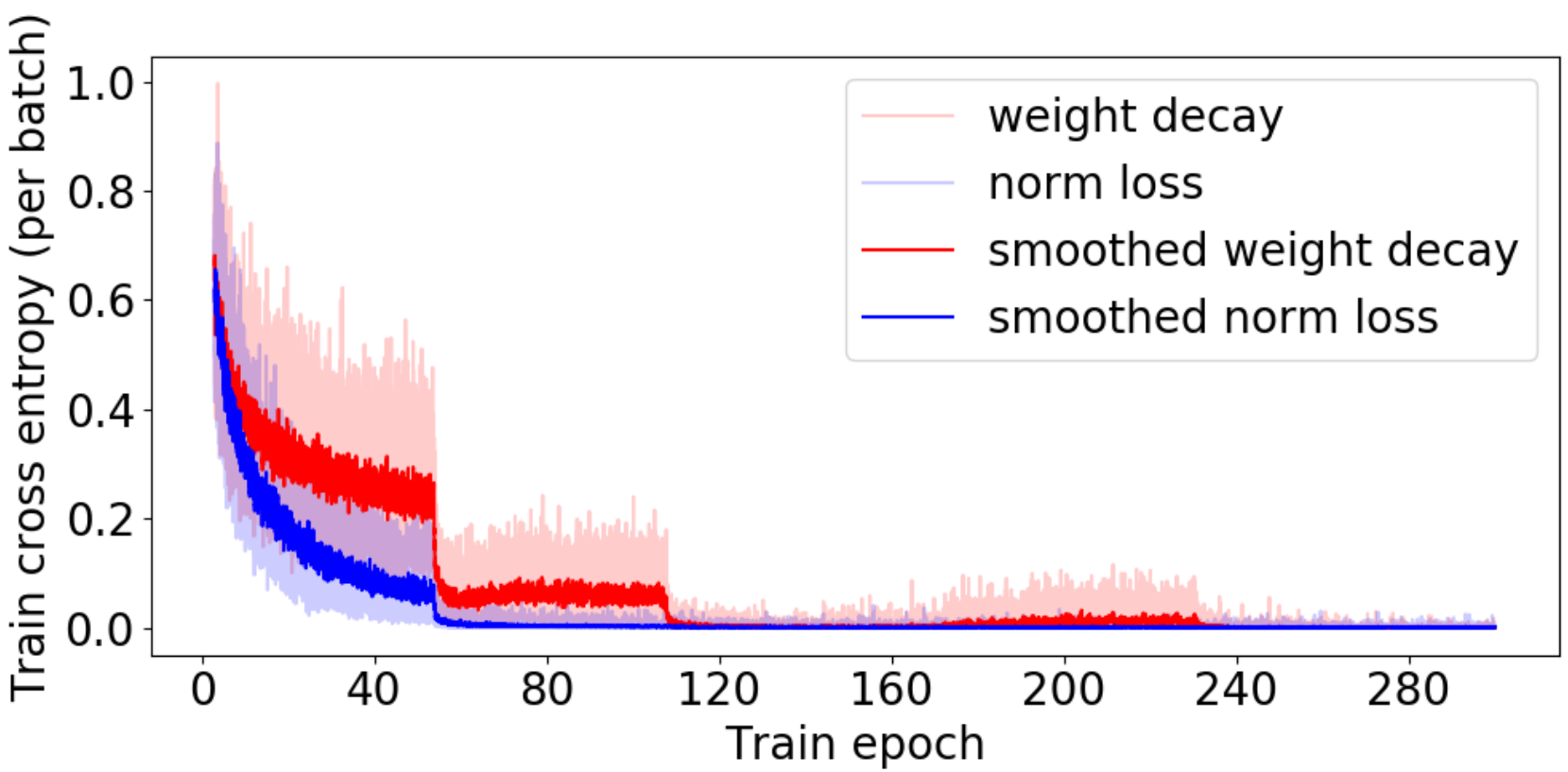
$$\frac{\partial L_{wd}}{\partial w_{ijc_i c_o}} = 2w_{ijc_i c_o} \tag{5}$$

Norm Loss update rule:

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

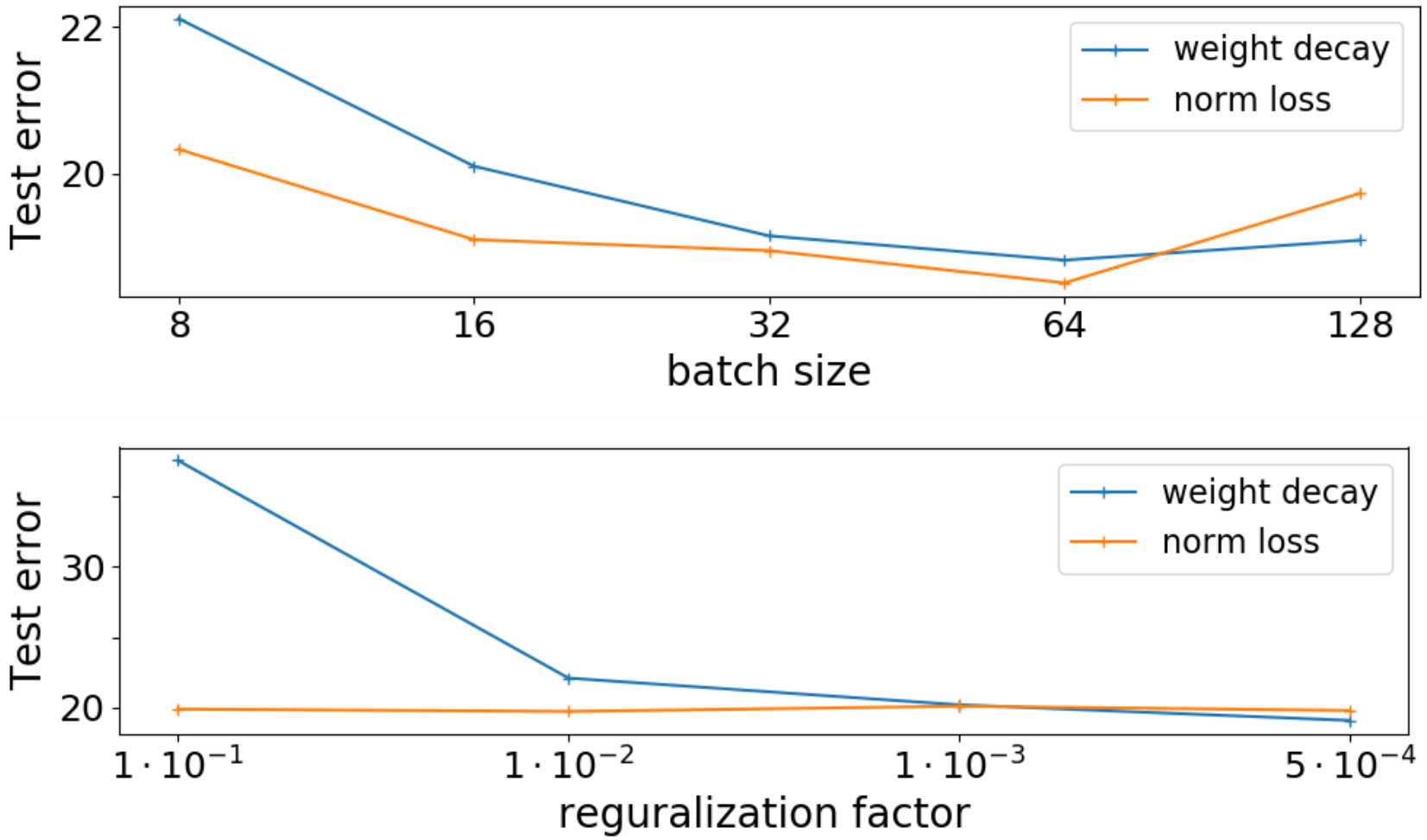
# Cross entropy during training (CIFAR-10)





# Hyper parameter robustness





# Results (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) |

# Results (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) |

# 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

Thank you!