Norm matters: efficient and accurate normalization schemes in deep networks

Elad Hoffer*, Ron Banner*, Itay Golan*, Daniel Soudry

Spotlight , NeurIPS 2018

*Equal contribution

Batch normalization

Shortcomings:

- Assumes independence between samples (problem when modeling time-series, RL, GANs, metric-learning etc.)
- Why it works? Interaction with other regularization
- Significant computational and memory impact, with data-bound operations –up to 25% of computation time in current models (Gitman, 17')
- Requires high-precision operations ($\sqrt{\sum_i x_i^2}$) , numerically unstable.

Batch-norm Leads to norm invariance

The key observation:

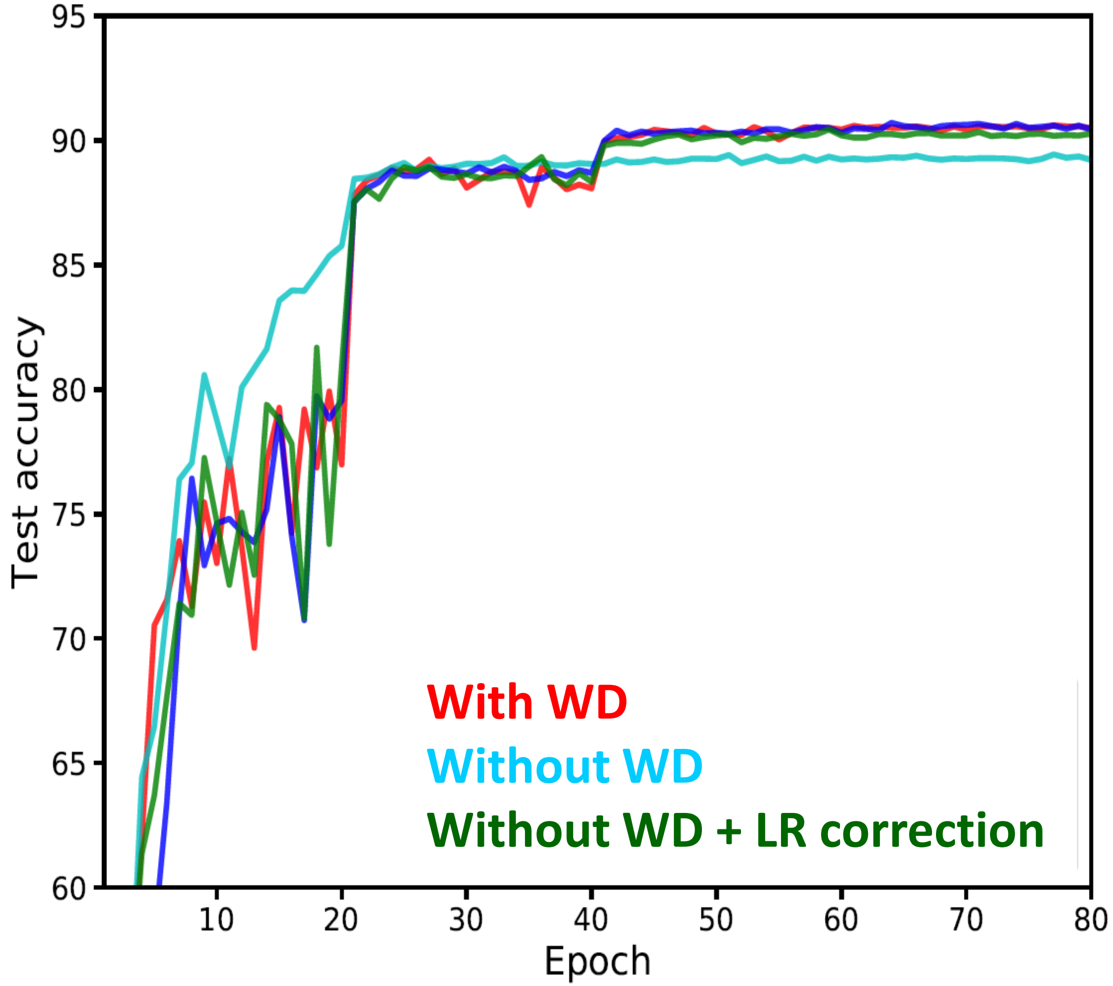
- Given input x, weight vector w, its direction $\hat{w} = \frac{w}{\|w\|}$
- Batch-norm is norm invariant: $BN(\|w\|\hat{w}x) = BN(\hat{w}x)$
- Weight norm only affects effective learning rate, e.g. in SGD:

$$\Delta \hat{w} = \frac{\eta}{||w||^2}(I - \hat{w}\hat{w}^\top)\nabla L(\hat{w}) + O(\eta^2)$$

Weight decay before BN is redundant

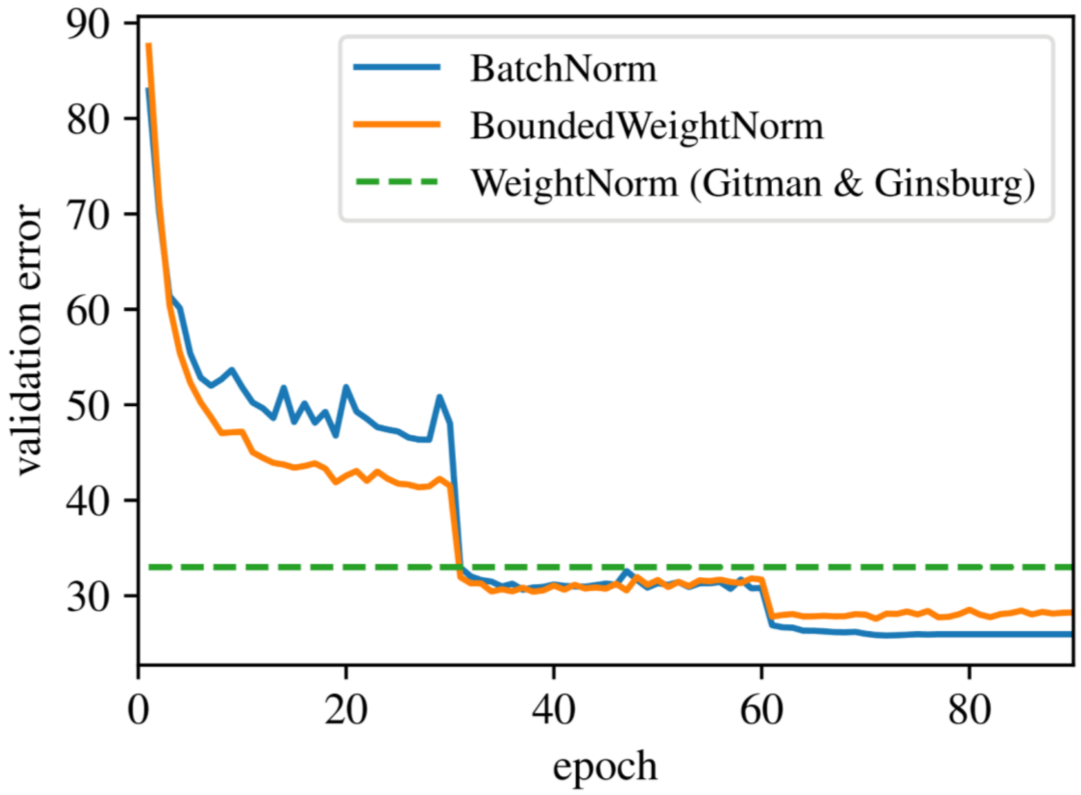
- Weight-decay equivalent to learning-rate scaling
- Can be mimicked by

$$\hat{\eta}_{\text{Correction}} = \eta \frac{\|w\|_2^2}{\|w_{[\text{WD on}]}\|_2^2}$$



Improving weight-norm

This can help to make weight-norm work for large-scale models



Resnet 50, ImageNet

Weight normalization, for a channel i:

$$w_i = g_i \frac{v_i}{\|v_i\|}$$

Bounded Weight Normalization:

$$w_i = \rho \frac{v_i}{\|v_i\|}$$

ρ - constant determined from chosen initialization

Replacing Batch-norm – switching norms

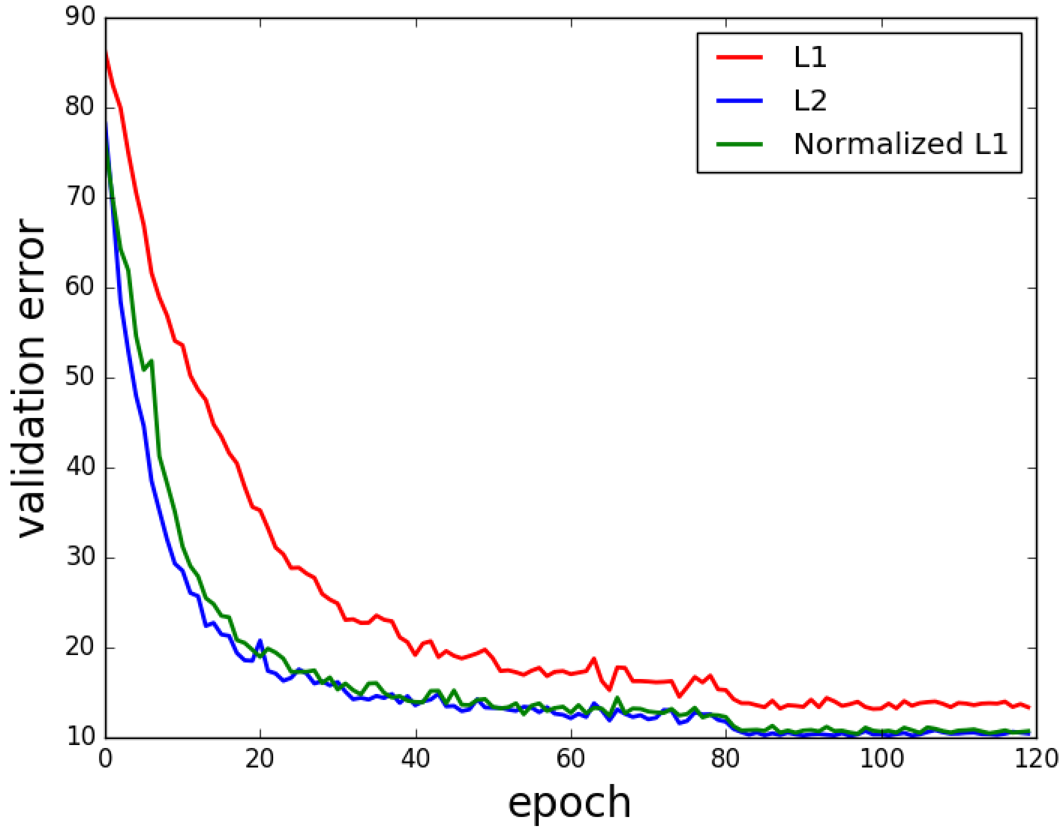
- Batch-normalization – just scaled L^2 normalization: $$\hat{x}_i = \frac{x_i - \langle x \rangle}{\frac{1}{\sqrt{n}}\|x-\langle x \rangle\|_2}$$
- More numerically stable norms:

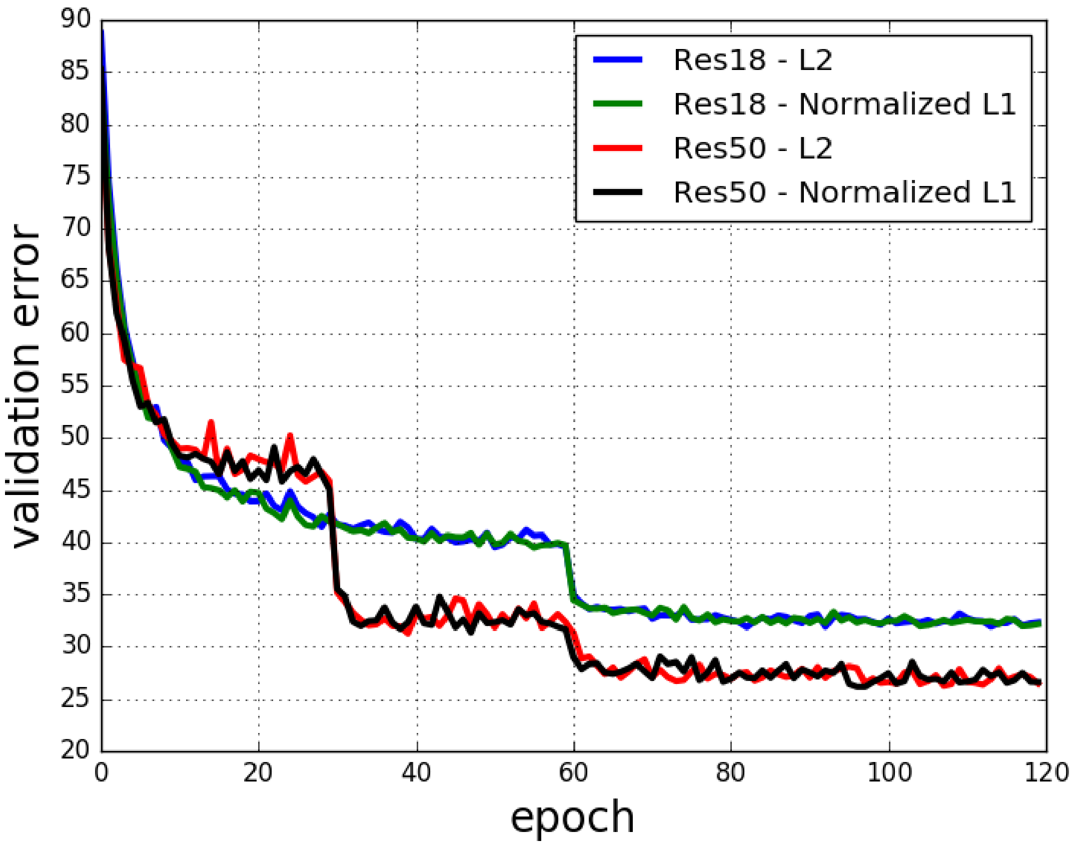
$$\|x\|_1 = \sum_i |x_i| \qquad \|x\|_\infty = \max_i\{|x_i|\}$$

We use additional scaling constants so that the norm will behave similarly to L^2, by assuming that neural input is Gaussian, e.g.:

$$\frac{1}{\sqrt{n}} E\|x-\langle x \rangle\|_2 = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} E\|x-\langle x \rangle\|_1$$

L^1 Batch-norm (Imagenet, Resnet)

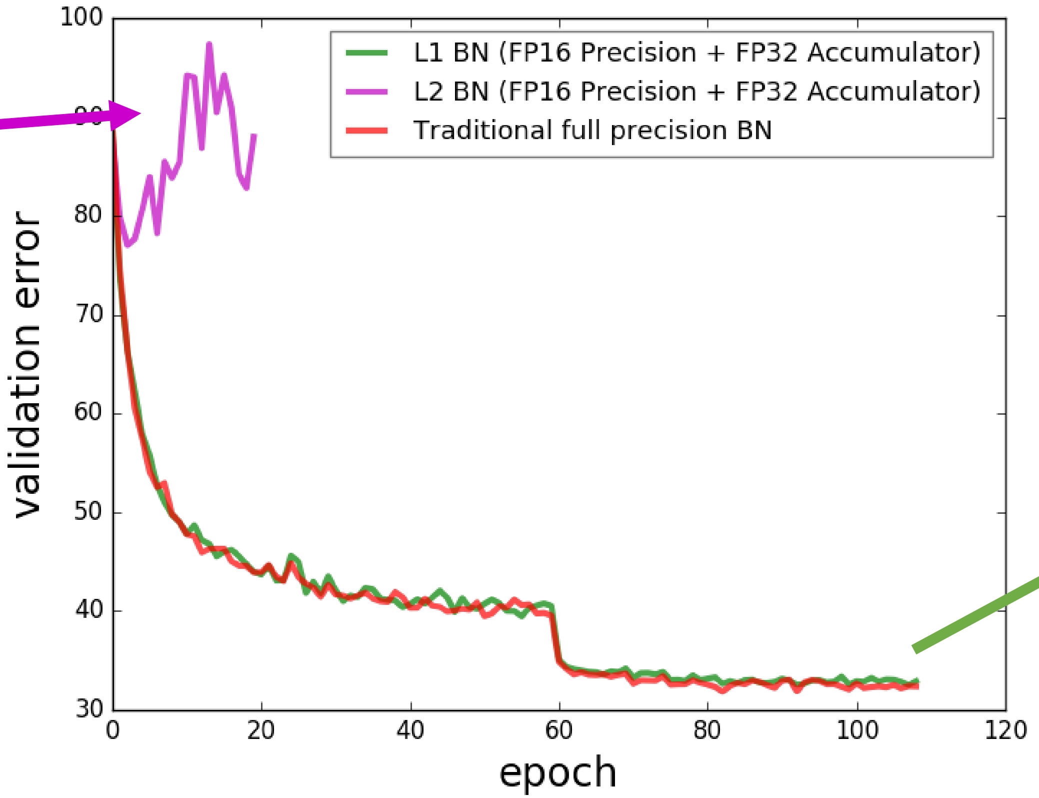




Low precision batch-norm

- L^1 batch-norm alleviates low-precision difficulties of batch-norm.
- Can now train using Batch-Norm on ResNet50 without issues on FP16:

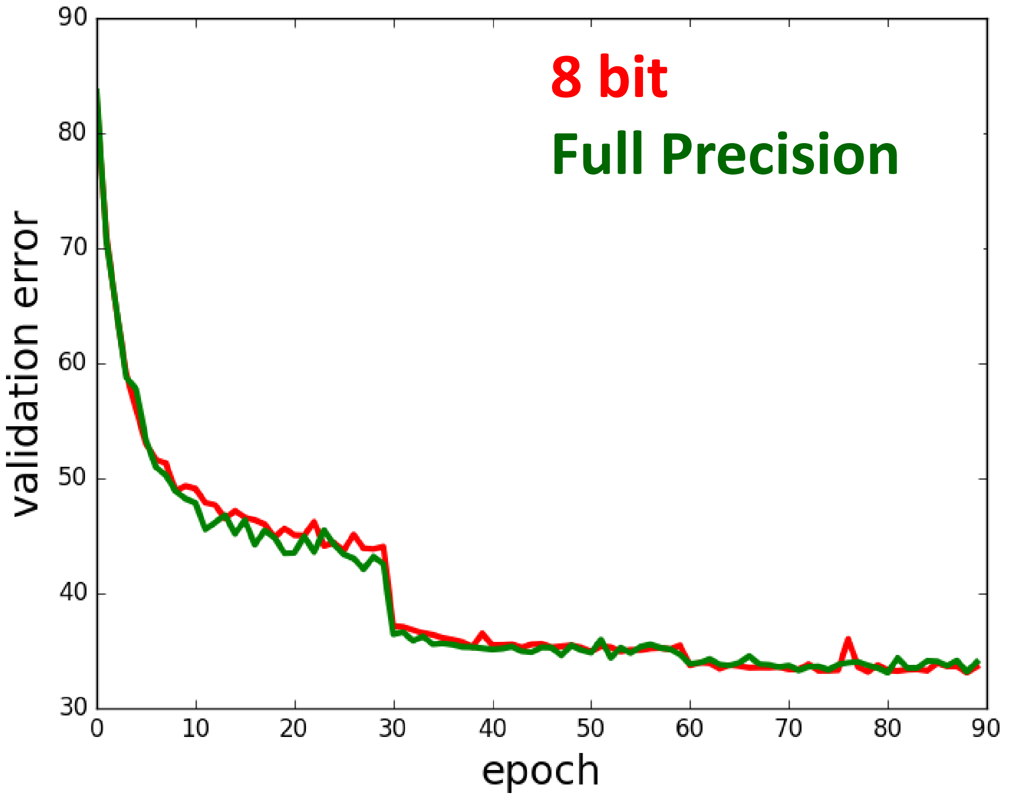
Regular BN in FP16 **fails**



L1 BN in FP16 **works** **as well as L2 in FP32**

With a few more tricks…

- Can now train ResNet18 ImageNet with bottleneck operations in **Int8**:



<u>Also at NeurIPS 2018</u>

"Scalable Methods for 8-bit Training of Neural Networks"

**Ron Banner, *Itay Hubara, *Elad Hoffer, Daniel Soudry*

Thank you for your time!
Come visit us at poster #27