Norm matters: efficient and accurate normalization schemes in deep networks

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

Spotlight, NeurIPS 2018
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 ($\sqrt{\sum x^2}$), numerically unstable.
Batch-norm Leads to norm invariance

The key observation:

• Given input $x$, weight vector $w$, its direction $\hat{\mathbf{w}} = \frac{w}{\|w\|}$

• Batch-norm is norm invariant: $BN(\|w\|\hat{\mathbf{w}}x) = BN(\hat{\mathbf{w}}x)$

• Weight norm only affects effective learning rate, e.g. in SGD:

$$\Delta \hat{\mathbf{w}} = \frac{\eta}{\|w\|^2} (I - \hat{\mathbf{w}}\hat{\mathbf{w}}^\top) \nabla L(\hat{\mathbf{w}}) + O(\eta^2)$$
Weight decay before BN is redundant

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

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

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

Weight normalization, for a channel $i$:

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

Bounded Weight Normalization:

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

$\rho$ - 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:
  $$\|x\|_1 = \sum_i |x_i| \quad \|x\|_\infty = \max_i \{|x_i|\}$$

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:
With a few more tricks...

- Can now train ResNet18 ImageNet with bottleneck operations in int8:

**Come visit our poster**

“Scalable Methods for 8-bit Training of Neural Networks”

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

NeurIPS 2018
Thank you for your time!