Normalizing the Normalizers: Comparing and Extending Network Normalization Schemes

Mengye Ren[†], Renjie Liao[†], Raquel Urtasun[†], Fabian H. Sinz[‡], Richard S. Zemel[†*]

[†]University of Toronto [‡]Baylor College of Medicine [*]Canadian Institute for Advanced Research (CIFAR)

Summary

- Batch normalization (BN) requires a set of independent samples.
- BN is nontrivial in RNNs, and performs poorly with small batch.
- We generalize the formulations of: batch normalization (BN); layer normalization (LN); and divisive normalization (DN).
- DN is believed to be a canonical computation of the brain.
- We found two modifications of standard DN, a smoothing term and L1 regularization, provide significant benefit in both CNNs and RNNs.

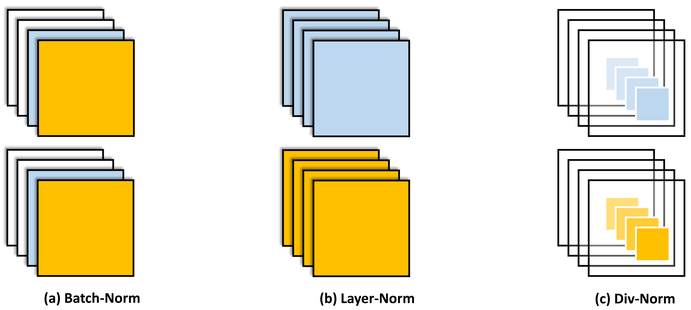
General Form of Normalization

- BN, LN, and DN can each be expressed in the following formulation:

$$z_{n,j} = \sum_i w_{i,j} x_{n,i} + b_j$$
$$v_{n,j} = z_{n,j} - \mathbb{E}_{\mathcal{A}_{n,j}}[z]$$
$$\tilde{z}_{n,j} = \frac{v_{n,j}}{\sqrt{\sigma^2 + \mathbb{E}_{\mathcal{B}_{n,j}}[v^2]}}$$

	Range	Normalizer Bias
BN	$\mathcal{A}_{n,j} = \{z_{m,j} : m \in [1,N], j \in [1,H] \times [1,W]\}$ $\mathcal{B}_{n,j} = \{v_{m,j} : m \in [1,N], j \in [1,H] \times [1,W]\}$	$\sigma = 0$
LN	$\mathcal{A}_{n,j} = \{z_{n,i} : i \in [1,L]\}$ $\mathcal{B}_{n,j} = \{v_{n,i} : i \in [1,L]\}$	$\sigma = 0$
DN	$\mathcal{A}_{n,j} = \{z_{n,i} : d(i,j) \leq R_{\mathcal{A}}\}$ $\mathcal{B}_{n,j} = \{v_{n,i} : d(i,j) \leq R_{\mathcal{B}}\}$	$\sigma \geq 0$

Table: Different choices of the summation and suppression fields $\mathcal{A}$ and $\mathcal{B}$.



(a) Batch-Norm (b) Layer-Norm (c) Div-Norm

Impact of Normalizer Bias

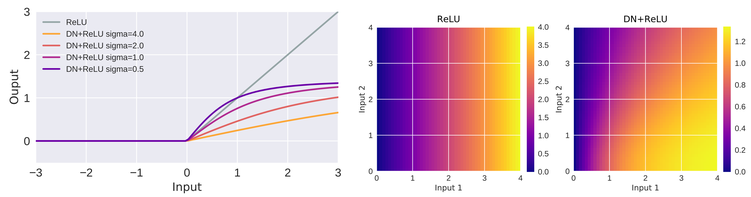


Figure: DN followed by ReLU can be viewed as a new activation function.

L1 Regularizer

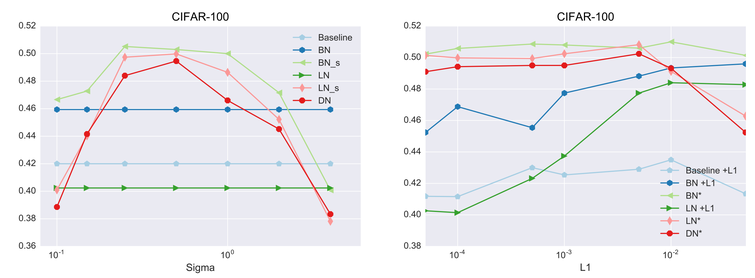
- We introduce the L1 regularizer on top of the centered activation to help decorrelate the filter response:

$$\mathcal{L}_{L1} = \alpha \frac{1}{NL} \sum_{n,j} |v_{n,j}|$$

CNN Image Classification

- We tested BN, LN, DN in a small CNN on object classification.
- Both smoother and L1 regularizer contribute to better performance.
- BN*, LN*, and DN* denote models with L1 regularizer.

Model	CIFAR-10 Acc.	CIFAR-100 Acc.
Baseline	0.7565	0.4409
BN	0.7807	0.4814
LN	0.7211	0.4249
BN*	**0.8179**	**0.5156**
LN*	0.8091	0.4957
DN*	0.8122	0.5066



Impact of L1 Regularizer

- We measure pairwise correlation (Corr) and mutual information (MI) in the first layer of a trained CNN.
- L1 regularizer makes pre-activations more independent.

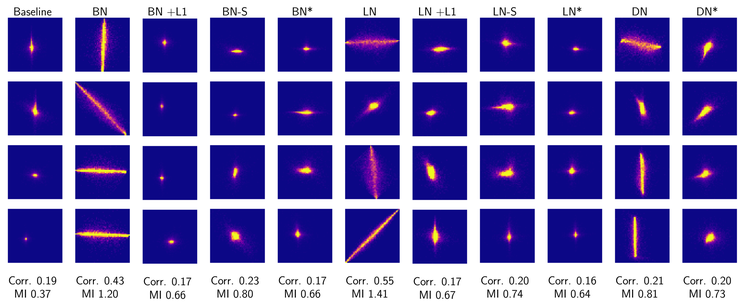


Figure: First layer CNN pre-normalized activation joint histogram.

RNN Language Modeling

- We tested BN, LN, DN in RNNs on the task of language modeling.
- We report the perplexity on Penn Treebank (PTB) test set (lower is better).

Model	LSTM	TanH RNN	ReLU RNN
Baseline	115.720	149.357	147.630
BN	123.245	148.052	164.977
LN	119.247	154.324	149.128
BN*	116.920	129.155	138.947
LN*	**101.725**	129.823	**116.609**
DN*	102.238	**123.652**	117.868

CNN Image Super-Resolution

- We also tested DN on low-level image processing CNNs.
- BN shows damaging effect on contrast and color.
- DN outperforms Super Resolution CNN (SRCNN) baseline.
- We report peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) on x3 and x4 (zoom-in ratio) benchmarks.

Table: Average test results on BSD200 Dataset.

Model	PSNR (x3)	SSIM (x3)	PSNR (x4)	SSIM (x4)
Bicubic	27.19	0.7636	25.92	0.6952
A+	27.05	0.7945	25.51	0.7171
SRCNN	28.42	0.8100	26.87	0.7378
BN	21.89	0.7553	21.53	0.6741
DN*	**28.44**	**0.8110**	**26.96**	**0.7428**

(a) Bicubic (b) SRCNN (c) BN (d) DN*