NORMALIZING THE NORMALIZERS: COMPARING AND EXTENDING NETWORK NORMALIZATION SCHEMES

Mengye Ren[*†], **Renjie Liao**[*†], **Raquel Urtasun**[†], **Fabian H. Sinz**[‡], **Richard S. Zemel**[†*]
[†]University of Toronto, Toronto ON, CANADA
[‡]Baylor College of Medicine, Houston TX, USA
[*]Canadian Institute for Advanced Research (CIFAR)
{mren, rjliao, urtasun}@cs.toronto.edu
fabian.sinz@epagoge.de, zemel@cs.toronto.edu

ABSTRACT

Normalization techniques have only recently begun to be exploited in supervised learning tasks. Batch normalization exploits mini-batch statistics to normalize the activations. This was shown to speed up training and result in better models. However its success has been very limited when dealing with recurrent neural networks. On the other hand, layer normalization normalizes the activations across all activities within a layer. This was shown to work well in the recurrent setting. In this paper we propose a unified view of normalization techniques, as forms of divisive normalization, which includes layer and batch normalization as special cases. Our second contribution is the finding that a small modification to these normalization schemes, in conjunction with a sparse regularizer on the activations, leads to significant benefits over standard normalization techniques. We demonstrate the effectiveness of our unified divisive normalization framework in the context of convolutional neural nets and recurrent neural networks, showing improvements over baselines in image classification, language modeling as well as super-resolution.

1 INTRODUCTION

Standard deep neural networks are difficult to train. Even with non-saturating activation functions such as ReLUs (Krizhevsky et al., 2012), gradient vanishing or explosion can still occur, since the Jacobian gets multiplied by the input activation of every layer. In AlexNet (Krizhevsky et al., 2012), for instance, the intermediate activations can differ by several orders of magnitude. Tuning hyperparameters governing weight initialization, learning rates, and various forms of regularization thus become crucial in optimizing performance.

In current neural networks, normalization abounds. One technique that has rapidly become a standard is batch normalization (BN) in which the activations are normalized by the mean and standard deviation of the training mini-batch (Ioffe & Szegedy, 2015). At inference time, the activations are normalized by the mean and standard deviation of the full dataset. A more recent variant, layer normalization (LN), utilizes the combined activities of all units within a layer as the normalizer (Ba et al., 2016). Both of these methods have been shown to ameliorate training difficulties caused by poor initialization, and help gradient flow in deeper models.

A less-explored form of normalization is divisive normalization (DN) (Heeger, 1992), in which a neuron's activity is normalized by its neighbors within a layer. This type of normalization is a well established canonical computation of the brain (Carandini & Heeger, 2012) and has been extensively studied in computational neuroscience and natural image modelling (see Section 2). However, with few exceptions (Jarrett et al., 2009; Krizhevsky et al., 2012) it has received little attention in conventional supervised deep learning.

Here, we provide a unifying view of the different normalization approaches by characterizing them as the same transformation but along different dimensions of a tensor, including normalization across

[*]indicates equal contribution

examples, layers in the network, filters in a layer, or instances of a filter response. We explore the effect of these varieties of normalizations in conjunction with regularization, on the prediction performance compared to baseline models. The paper thus provides the first study of divisive normalization in a range of neural network architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), and tasks such as image classification, language modeling and image super-resolution. We find that DN can achieve results on par with BN in CNN networks and out-performs it in RNNs and super-resolution, without having to store batch statistics. We show that casting LN as a form of DN by incorporating a smoothing parameter leads to significant gains, in both CNNs and RNNs. We also find advantages in performance and stability by being able to drive learning with higher learning rate in RNNs using DN. Finally, we demonstrate that adding an L1 regularizer on the activations before normalization is beneficial for all forms of normalization.

2 Related Work

In this section we first review related work on normalization, followed by a brief description of regularization in neural networks.

2.1 Normalization

Normalization of data prior to training has a long history in machine learning. For instance, local contrast normalization used to be a standard effective tool in vision problems (Pinto et al., 2008; Jarrett et al., 2009; Sermanet et al., 2012; Le, 2013). However, until recently, normalization was usually not part of the machine learning algorithm itself. Two notable exceptions are the original AlexNet by Krizhevsky et al. (2012) which includes a divisive normalization step over a subset of features after ReLU at each pixel location, and the work by Jarrett et al. (2009) who demonstrated that a combination of nonlinearities, normalization and pooling improves object recognition in two-stage networks.

Recently Ioffe & Szegedy (2015) demonstrated that standardizing the activations of the summed inputs of neurons over training batches can substantially decrease training time in deep neural networks. To avoid covariate shift, where the weight gradients in one layer are highly dependent on previous layer outputs, Batch Normalization (BN) rescales the summed inputs according to their variances under the distribution of the mini-batch data. Specifically, if $z_{j,n}$ denotes the activation of a neuron j on example n, and $B(n)$ denotes the mini-batch of examples that contains n, then BN computes an affine function of the activations standardized over each mini-batch:

$$\tilde{z}_{n,j} = \gamma \frac{z_{n,j} - \mathbb{E}[z_j]}{\sqrt{\frac{1}{|B(n)|}(z_{n,j} - \mathbb{E}[z_j])^2}} + \beta \qquad \mathbb{E}[z_j] = \frac{1}{|B(n)|} \sum_{m \in B(n)} z_{m,j}$$

However, training performance in Batch Normalization strongly depends on the quality of the aquired statistics and, therefore, the size of the mini-batch. Hence, Batch Normalization is harder to apply in cases for which the batch sizes are small, such as online learning or data parallelism. While classification networks can usually employ relatively larger mini-batches, other applications such as image segmentation with convolutional nets use smaller batches and suffer from degraded performance. Moreover, application to recurrent neural networks (RNNs) is not straightforward and leads to poor performance (Laurent et al., 2015).

Several approaches have been proposed to make Batch Normalization applicable to RNNs. Cooijmans et al. (2016) and Liao & Poggio (2016) collect separate batch statistics for each time step. However, neither of this techniques address the problem of small batch sizes and it is unclear how to generalize them to unseen time steps.

More recently, Ba et al. (2016) proposed Layer Normalization (LN), where the activations are normalized across all summed inputs within a layer instead of within a batch:

$$\tilde{z}_{n,j} = \gamma \frac{z_{n,j} - \mathbb{E}[z_n]}{\sqrt{\frac{1}{|L(j)|}(z_{n,j} - \mathbb{E}[z_n])^2}} + \beta \qquad \mathbb{E}[z_n] = \frac{1}{|L(j)|} \sum_{k \in L(j)} z_{n,k}$$

where $L(j)$ contains all of the units in the same layer as j. While promising results have been shown on RNN benchmarks, direct application of layer normalization to convolutional layers often leads to

a degradation of performance. The authors hypothesize that since the statistics in convolutional layers can vary quite a bit spatially, normalization with statistics from an entire layer might be suboptimal.

Ulyanov et al. (2016) proposed to normalize each example on spatial dimensions but not on channel dimension, and was shown to be effective on image style transfer applications (Gatys et al., 2016).

Liao et al. (2016a) proposed to accumulate the normalization statistics over the entire training phase, and showed that this can speed up training in recurrent and online learning without a deteriorating effect on the performance. Since gradients cannot be backpropagated through this normalization operation, the authors use running statistics of the gradients instead.

Exploring the normalization of weights instead of activations, Salimans & Kingma (2016) proposed a reparametrization of the weights into a scale independent representation and demonstrated that this can speed up training time.

Divisive Normalization (DN) on the other hand modulates the neural activity by the activity of a pool of neighboring neurons (Heeger, 1992; Bonds, 1989). DN is one of the most well studied and widely found transformations in real neural systems, and thus has been called a canonical computation of the brain (Carandini & Heeger, 2012). While the exact form of the transformation can differ, all formulations model the response of a neuron $\tilde{z}_j$ as a ratio between the acitivity in a summation field $\mathcal{A}_j$, and a norm-like function of the suppression field $\mathcal{B}_j$

$$\tilde{z}_j = \gamma \frac{\sum_{z_i \in \mathcal{A}_j} u_i z_i}{\left(\sigma^2 + \sum_{z_k \in \mathcal{B}_j} w_k z_k^p\right)^{\frac{1}{p}}}, \tag{1}$$

where $\{u_i\}$ are the summation weights and $\{w_k\}$ the suppression weights.

Previous theoretical studies have outlined several potential computational roles for divisive normalization such as sensitivity maximization (Carandini & Heeger, 2012), invariant coding (Olsen et al., 2010), density modelling (Ballé et al., 2016), image compression (Malo et al., 2006), distributed neural representations (Simoncelli & Heeger, 1998), stimulus decoding (Ringach, 2009; Froudarakis et al., 2014), winner-take-all mechanisms (Busse et al., 2009), attention (Reynolds & Heeger, 2009), redundancy reduction (Schwartz & Simoncelli, 2001; Sinz & Bethge, 2008; Lyu & Simoncelli, 2008; Sinz & Bethge, 2013), marginalization in neural probabilistic population codes (Beck et al., 2011), and contextual modulations in neural populations and perception (Coen-Cagli et al., 2015; Schwartz et al., 2009).

2.2 REGULARIZATION

Various regularization techniques have been applied to neural networks for the purpose of improving generalization and reduce overfitting. They can be roughly divided into two categories, depending on whether they regularize the weights or the activations.

Regularization on Weights: The most common regularizer on weights is weight decay which just amounts to using the L2 norm squared of the weight vector. An L1 regularizer (Goodfellow et al., 2016) on the weights can also be adopted to push the learned weights to become sparse. Scardapane et al. (2016) investigated mixed norms in order to promote group sparsity.

Regularization on Activations: Sparsity or group sparsity regularizers on the activations have shown to be effective in the past (Roz, 2008; Kavukcuoglu et al., 2009) and several regularizers have been proposed that act directly on the neural activations. Glorot et al. (2011) add a sparse regularizer on the activations after ReLU to encourage sparse representations. Dropout developed by Srivastava et al. (2014) applies random masks to the activations in order to discourage them to co-adapt. DeCov proposed by Cogswell et al. (2015) tries to minimize the off-diagonal terms of the sample covariance matrix of activations, thus encouraging the activations to be as decorrelated as possible. Liao et al. (2016b) utilize a clustering-based regularizer to encourage the representations to be compact.

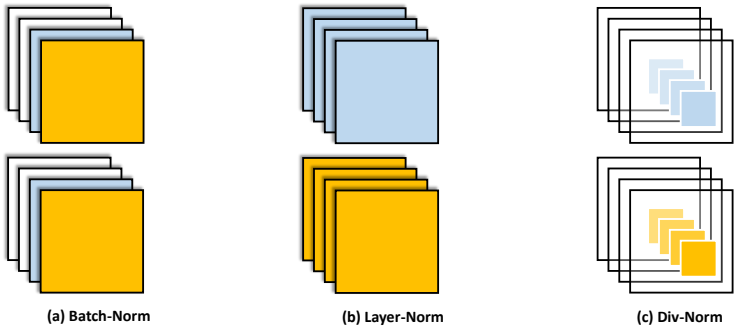


Figure 1: Illustration of different normalization schemes, in a CNN. Each $H \times W$-sized feature map is depicted as a rectangle; overlays depict instances in the set of C filters; and two examples from a mini-batch of size N are shown, one above the other. The colors show the summation/suppression fields of each scheme.

3 A Unified Framework for Normalizing Neural Nets

We first compare the three existing forms of normalization, and show that we can modify batch normalization (BN) and layer normalization (LN) in small ways to make them have a form that matches divisive normalization (DN). We present a general formulation of normalization, where existing normalizations involve alternative schemes of accumulating information. Finally, we propose a regularization term that can be optimized jointly with these normalization schemes to encourage decorrelation and/or improve generalization performance.

3.1 General Form of Normalization

Without loss of generality, we denote the hidden input activation of one arbitrary layer in a deep neural network as $\mathbf{z} \in \mathbb{R}^{N \times L}$. Here N is the mini-batch size. In the case of a CNN, $L = H \times W \times C$, where H, W are the height and width of the convolutional feature map and C is the number of filters. For an RNN or fully-connected layers of a neural net, L is the number of hidden units.

Different normalization methods gather statistics from different ranges of the tensor and then perform normalization. Consider the following general form:

$$z_{n,j} = \sum_i w_{i,j} x_{n,i} + b_j \tag{2}$$

$$v_{n,j} = z_{n,j} - \mathbb{E}_{\mathcal{A}_{n,j}}[z] \tag{3}$$

$$\tilde{z}_{n,j} = \frac{v_{n,j}}{\sqrt{\sigma^2 + \mathbb{E}_{\mathcal{B}_{n,j}}[v^2]}} \tag{4}$$

where $\mathcal{A}_j$ and $\mathcal{B}_j$ are subsets of z and v respectively. $\mathcal{A}$ and $\mathcal{B}$ in standard divisive normalization are referred to as summation and suppression fields (Carandini & Heeger, 2012). One can cast each normalization scheme into this general formulation, where the schemes vary based on how they define these two fields. These definitions are specified in Table 1. Optional parameters γ and β can be added in the form of $\gamma_j \tilde{z}_{n,j} + \beta_j$ to increase the degree of freedom.

Fig. 1 shows a visualization of the normalization field in a 4-D ConvNet tensor setting. Divisive normalization happens within a local spatial window of neurons across filter channels. Here we set $d(\cdot, \cdot)$ to be the spatial L_∞ distance.

3.2 New Model Components

Smoothing the Normalizers: One obvious way in which the normalization schemes differ is in terms of the information that they combine for normalizing the activations. A second more subtle but important difference between standard BN and LN as opposed to DN is the smoothing term σ, in the denominator of Eq. (1). This term allows some control of the bias of the variance estimation, effectively smoothing the estimate. This is beneficial because divisive normalization does not utilize information from the mini-batch as in BN, and combines information from a smaller field than LN. A

Model	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 1: Different choices of the summation and suppression fields $\mathcal{A}$ and $\mathcal{B}$, as well as the constant σ in the normalizer lead to known normalization schemes in neural networks. $d(i, j)$ denotes an arbitrary distance between two hidden units i and j, and R denotes the neighbourhood radius.

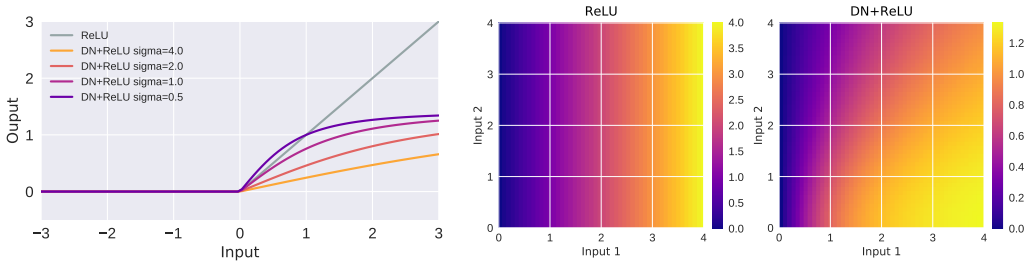


Figure 2: Divisive normalization followed by ReLU can be viewed as a new activation function. Left: Effect of varying σ in this activation function. Right: Two units affect each other's activation in the DN+ReLU formulation.

similar but different denominator bias term $\max(\sigma, c)$ appears in (Jarrett et al., 2009), which is active when the activation variance is small. However, the clipping function makes the transformation not invertible, losing scale information.

Moreover, if we take the nonlinear activation function after normalization into consideration, we find that σ will change the overall properties of the non-linearity. To illustrate this effect, we use a simple 1-layer network which consists of: two input units, one divisive normalization operator, followed by a ReLU activation function. If we fix one input unit to be 0.5, varying the other one with different values of σ produces different output curves (Fig. 2, left). These curves exhibit different non-linear properties compared to the standard ReLU. Allowing the other input unit to vary as well results in different activation functions of the first unit depending on the activity of the second (Fig. 2, right). This illustrates potential benefits of including this smoothing term σ, as it effectively modulates the rectified response to vary from a linear to a highly saturated response.

In this paper we propose modifications of the standard BN and LN which borrow this additive term σ in the denominator from DN. We study the effect of incorporating this smoother in the respective normalization schemes below.

L1 regularizer: Filter responses on lower layers in deep neural networks can be quite correlated which might impair the estimate of the variance in the normalizer. More independent representations help disentangle latent factors and boost the networks performance (Higgins et al., 2016). Empirically, we found that putting a sparse (L1) regularizer

$$\mathcal{L}_{L1} = \alpha \frac{1}{NL} \sum_{n,j} |v_{n,j}| \tag{5}$$

on the centered activations $v_{n,j}$ helps decorrelate the filter responses (Fig. 5). Here, N is the batch size and L is the number of hidden units, and $\mathcal{L}_{L1}$ is the regularization loss which is added to the training loss.

A possible explanation for this effect is that the L1 regularizer might have a similar effect as maximum likelihood estimation of an independent Laplace distribution. To see that, let $p_v(\mathbf{v}) \propto \exp(-\|\mathbf{v}\|_1)$ and $\mathbf{x} = W^{-1}\mathbf{v}$, with W a full rank invertible matrix. Under this model $p_x(\mathbf{x}) = p_v(W\mathbf{x}) |\det W|$.

Then, minimization of the L1 norm of the activations under the volume-conserving constraint $\det A = \text{const.}$ corresponds to maximum likelihood on that model, which would encourage decorrelated responses. We do not enforce such a constraint, and the filter matrix might even not be invertible. However, the supervised loss function of the network benefits from having diverse non-zero filters. This encourages the network to not collapse filters along the same direction or put them to zero, and might act as a relaxation of the volume-conserving constraint.

3.3 SUMMARY OF NEW MODELS

DN and DN*: We propose DN as a new local normalization scheme in neural networks. In convolutional layers, it operates on a local spatial window across filter channels, and in fully connected layers it operates on a slice of a hidden state vector. Additionally, DN* has a L1 regularizer on the pre-normalization centered activation ($v_{n,j}$).

BN-s and BN*: To compare with DN and DN*, we also propose modifications to original BN: we denote BN-s with σ^2 in the denominator's square root, and BN* with the L1 regularizer on top of BN-s.

LN-s and LN*: We apply the same changes as from BN to BN-s and BN*. In order to narrow the differences in the normalization schemes down to a few parameter choices, we additionally remove the affine transformation parameters γ and β from LN such that the difference between LN* and DN* is only the size of the normalization field. γ and β can really be seen as a separate layer and in practice we find that they do not improve the performance in the presence of σ^2.

4 EXPERIMENTS

We evaluate the normalization schemes on three different tasks:

- **CNN image classification:** We apply different normalizations on CNNs trained on the CIFAR-10/100 datasets for image recognition, each of which contains 50,000 training images and 10,000 test images. Each image is of size $32 \times 32 \times 3$ and has been labeled an object class out of 10 or 100 total number of classes.
- **RNN language modeling:** We apply different normalizations on RNNs trained on the Penn Treebank dataset for language modeling, containing 42,068 training sentences, 3,370 validation sentences, and 3,761 test sentences.
- **CNN image super-resolution:** We train a CNN on low resolution images and learn cascades of non-linear filters to smooth the upsampled images. We report performance of trained CNN on the standard Set 14 and Berkeley 200 dataset.

For each model, we perform a grid search of three or four choices of each hyperparameter including the smoothing constant σ, and L1 regularization constant α, and learning rate ϵ on the validation set.

4.1 CIFAR EXPERIMENTS

We used the standard CNN model provided in the Caffe library. The architecture is summarized in Table 2. We apply normalization before each ReLU function. We implement DN as a convolutional operator, fixing the local window size to 5×5, 3×3, 3×3 for the three convolutional layers in all the CIFAR experiments.

We set the learning rate to 1e-3 and momentum 0.9 for all experiments. The learning rate schedule is set to {5K, 30K, 50K} for the baseline model and to {30K, 50K, 80K} for all other models. At every stage we multiply the learning rate by 0.1. Weights are randomly initialized from a zero-mean normal distribution with standard deviation {1e-4, 1e-2, 1e-2} for the convolutional layers, and {1e-1, 1e-1} for fully connected layers. Input images are centered on the dataset image mean.

Table 3 summarizes the test performances of BN*, LN* and DN*, compared to the performance of a few baseline models and the standard batch and layer normalizations. We also add standard regularizers to the baseline model: L2 weight decay (WD) and dropout. Adding the smoothing constant and L1 regularization consistently improves the classification performance, especially for

Table 2: CIFAR CNN specification

Type	Size	Kernel	Stride
input	$32 \times 32 \times 3$	-	-
conv +relu	$32 \times 32 \times 32$	$5 \times 5 \times 3 \times 32$	1
max pool	$16 \times 16 \times 32$	3×3	2
conv +relu	$16 \times 16 \times 32$	$5 \times 5 \times 32 \times 32$	1
avg pool	$8 \times 8 \times 32$	3×3	2
conv +relu	$8 \times 8 \times 64$	$5 \times 5 \times 32 \times 64$	1
avg pool	$4 \times 4 \times 64$	3×3	2
fully conn. linear	64	-	-
fully conn. linear	10 or 100	-	-

Table 3: CIFAR-10/100 experiments

Model	CIFAR-10 Acc.	CIFAR-100 Acc.
Baseline	0.7565	0.4409
Baseline +WD +Dropout	0.7795	0.4179
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

the original LN. The modification of LN makes it now better than the original BN, and only slightly worse than BN*. DN* achieves comparable performance to BN* on both datasets, but only relying on a local neighborhood of hidden units.

ResNet Experiments. Residual networks (ResNet) (He et al., 2016), a type of CNN with residual connections between layers, achieve impressive performance on many image classification benchmarks. The original architecture uses BN by default. If we remove BN, the architecture is very difficult to train or converges to a poor solution. We first reproduced the original BN ResNet-32, obtaining 92.6% accuracy on CIFAR-10, and 69.8% on CIFAR-100. Our best DN model achieves 91.3% and 66.6%, respectively. While this performance is lower than the original BN-ResNet, there is certainly room to improve as we have not performed any hyperparameter optimization. Importantly, the beneficial effects of sigma (2.5% gain on CIFAR-100) and the L1 regularizer (0.5%) are still found, even in the presence of other regularization techniques such as data augmentation and weight decay in the training.

Since the number of sigma hyperparameters scales with the number of layers, we found that setting sigma as a learnable parameter for each layer helps the performance (1.3% gain on CIFAR-100). Note that training this parameter is not possible in the formulation by Jarrett et al. (2009). The learned sigma shows a clear trend: it tends to decrease with depth, and in the last convolution layer it approaches 0 (see Fig. 3).

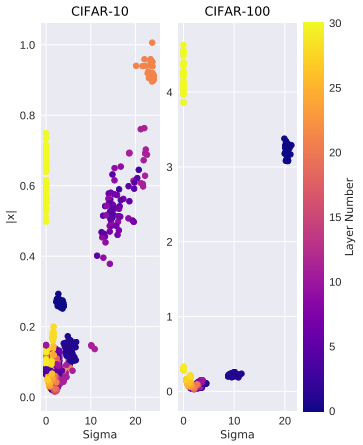


Figure 3: Input scale ($|x|$) vs. learned σ at each layer, color coded by the layer number in ResNet-32, trained on CIFAR-10 (left), and CIFAR-100 (right).

4.2 RNN EXPERIMENTS

To apply divisive normalization in fully connected layers of RNNs, we consider a local neighborhood in the hidden state vector $\mathbf{h}_{j-R:j+R}$, where R is the radius

Table 4: PTB Word-level language modeling experiments

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

of the neighborhood. Although the hidden states are randomly initialized, this structure will impose local competition among the neighbors.

$$v_j = z_j - \frac{1}{2R+1} \sum_{r=-R}^{R} z_{j+r} \tag{6}$$

$$\tilde{z}_j = \frac{v_j}{\sqrt{\sigma^2 + \frac{1}{2R+1} \sum_{r=-R}^{R} v_{j+r}^2}} \tag{7}$$

We follow Cooijmans et al. (2016)'s batch normalization implementation for RNNs: normalizers are separate for input transformation and hidden transformation. Let $BN(\cdot)$, $LN(\cdot)$, $DN(\cdot)$ be BatchNorm, LayerNorm and DivNorm, and g be either tanh or ReLU.

$$\mathbf{h}_{t+1} = g(W_x \mathbf{x}_t + W_h \mathbf{h}_{t-1} + b) \tag{8}$$

$$\mathbf{h}_{t+1}^{(BN)} = g(BN(W_x \mathbf{x}_t + b_x) + BN(W_h \mathbf{h}_{t-1}^{(BN)} + b_h)) \tag{9}$$

$$\mathbf{h}_{t+1}^{(LN)} = g(LN(W_x \mathbf{x}_t + W_h \mathbf{h}_{t-1}^{(LN)} + b)) \tag{10}$$

$$\mathbf{h}_{t+1}^{(DN)} = g(DN(W_x \mathbf{x}_t + W_h \mathbf{h}_{t-1}^{(DN)} + b)) \tag{11}$$

Note that in recurrent BN, the additional parameters γ and β are shared across timesteps whereas the moving averages of batch statistics are not shared. For the LSTM version, we followed the released implementation from the authors of layer normalization [1], and apply LN at the same places as BN and BN*, which is after the linear transformation of $W_x \mathbf{x}$ and $W_h \mathbf{h}$ individually. For LN* and DN, we modified the places of normalization to be at each non-linearity, instead of jointly with a concatenated vector for different non-linearity. We found that this modification improves the performance and makes the formulation clearer since normalization is always a combined operation with the activation function. We include details of the LSTM implementation in the Appendix.

The RNN model is provided by the Tensorflow library (Abadi et al., 2016) and the LSTM version was originally proposed in Zaremba et al. (2014). We used a two-layer stack-RNN of size 400 (vanilla RNN) or 200 (LSTM). R is set to 60 (vanilla RNN) and 30 (LSTM). We tried both tanh and ReLU as the activation function for the vanilla RNN. For unnormalized baselines and BN+ReLU, the initial learning rate is set to 0.1 and decays by half every epoch, starting at the 5th epoch for a maximum of 13 epochs. For the other normalized models, the initial learning rate is set to 1.0 while the schedule is kept the same. Standard stochastic gradient descent is used in all RNN experiments, with gradient clipping at 5.0.

Table 4 shows the test set perplexity for LSTM models and vanilla models. Perplexity is defined as $\text{ppl} = \exp(-\sum_x \log p(x))$. We find that BN and LN alone do not improve the final performance relative to the baseline, but similar to what we see in the CNN experiments, our modified versions BN* and LN* show significant improvements. BN* on RNN is outperformed by both LN* and DN. By applying our normalization, we can improve the vanilla RNN perplexity by 20%, comparable to an LSTM baseline with the same hidden dimension.

[1] `https://github.com/ryankiros/layer-norm`

Table 5: Average test results of PSNR and SSIM on Set14 Dataset.

Model	**PSNR** (x3)	**SSIM** (x3)	**PSNR** (x4)	**SSIM** (x4)
Bicubic	27.54	0.7733	26.01	0.7018
A+	29.13	0.8188	27.32	0.7491
SRCNN	29.35	0.8212	27.53	0.7512
BN	22.31	0.7530	21.40	0.6851
DN*	**29.38**	**0.8229**	**27.64**	**0.7562**

Table 6: Average test results of PSNR and SSIM 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**

4.3 Super Resolution Experiments

We also evaluate DN on the low-level computer vision problem of single image super-resolution. We adopt the SRCNN model of Dong et al. (2016) as the baseline which consists of 3 convolutional layers and 2 ReLUs. From bottom to top layers, the sizes of the filters are 9, 5, and 5 [2]. The number of filters are 64, 32, and 1, respectively. All the filters are initialized with zero-mean Gaussian and standard deviation 1e-3. Then we respectively apply batch normalization (BN) and our divisive normalization with L1 regularization (DN*) to the convolutional feature maps before ReLUs. We construct the training set in a similar manner as Dong et al. (2016) by randomly cropping 5 million patches (size 33×33) from a subset of the ImageNet dataset of Deng et al. (2009). We only train our model for 4 million iterations which is less than the one adopted by SRCNN, *i.e.*, 15 million, as the gain of PSNR and SSIM by spending that long time is marginal.

We report the average test results, utilizing the standard metrics PSNR and SSIM (Wang et al., 2004), on two standard test datasets Set14 (Zeyde et al., 2010) and BSD200 (Martin et al., 2001). We compare with two state-of-the-art single image super-resolution methods, A+ (Timofte et al., 2013) and SRCNN (Dong et al., 2016). All measures are computed on the Y channel of YCbCr color space. We also provide a visual comparison in Fig. 4.

As show in Tables 5 and 6 DN* outperforms the strong competitor SRCNN, while BN does not perform well on this task. The reason may be that BN applies the same statistics to all patches of one image which causes some overall intensity shift (see Figs. 4). From the visual comparisons, we can see that our method not only enhances the resolution but also removes artifacts, *e.g.*, the ringing effect in Fig. 4.

4.4 Ablation studies and discussion

Finally, we investigated the differential effects of the σ^2 term and the L1 regularizer on the performance. We ran ablation studies on CIFAR-10/100 as well as PTB experiments. The results are listed in Table 7.

We find that adding the smoothing term σ^2 and the L1 regularization consistently increases the performance of the models. In the convolutional networks, we find that L1 and σ both have similar effects on the performance. L1 seems to be slightly more important. In recurrent networks, σ^2 has a much more dramatic effect on the performance than the L1 regularizer.

Fig. 5 plots randomly sampled pairwise pre-normalization responses (after the linear transform) in the first layer at the same spatial location of the feature map, along with the average pair-wise

[2] We use the setting of the best model out of all three SRCNN candidates.

Figure 4: Comparisons at a magnification factor of 4.

correlation coefficient (Corr) and mutual information (MI). It is evident that both σ and L1 encourages independence of the learned linear filters.

There are several factors that could explain the improvement in performance. As mentioned above, adding the L1 regularizer on the activations encourages the filter responses to be less correlated. This can increase the robustness of the variance estimate in the normalizer and lead to an improved scaling of the responses to a good regime. Furthermore, adding the smoother to the denominator in the normalizer can be seen as implicitly injecting zero mean noise on the activations. While noise injection would not change the mean, it does add a term to the variance of the data, which is represented by σ^2. This term also makes the normalization equation invertible. While dividing by the standard deviation decreases the degrees of freedom in the data, the smoothed normalization equation is fully information preserving. Finally, DN type operations have been shown to decrease the redundancy of filter responses to natural images and sound (Schwartz & Simoncelli, 2001; Sinz & Bethge, 2008; Lyu & Simoncelli, 2008). In combination with the L1 regularizer this could lead to a more independent representation of the data and thereby increase the performance of the network.

5 Conclusions

We have proposed a unified view of normalization techniques which contains batch and layer normalization as special cases. We have shown that when combined with a sparse regularizer on the activations, our framework has significant benefits over standard normalization techniques. We have demonstrated this in the context of both convolutional neural nets as well as recurrent neural networks. In the future we plan to explore other regularization techniques such as group sparsity. We also plan to conduct a more in-depth analysis of the effects of normalization on the correlations of the learned representations.

Table 7: Comparison of standard batch and layer normalation (BN and LN) models, to those with only L1 regularizer (+L1), only the σ smoothing term (-s), and with both (*). We also compare divisive normalization with both (DN*), versus with only the smoothing term (DN).

Model	CIFAR-10	CIFAR-100	LSTM	Tanh RNN	ReLU RNN
Baseline	0.7565	0.4409	115.720	149.357	147.630
Baseline +L1	0.7839	0.4517	111.885	143.965	148.572
BN	0.7807	0.4814	123.245	148.052	164.977
BN +L1	0.8067	0.5100	123.736	152.777	166.658
BN-s	0.8017	0.5005	123.243	131.719	139.159
BN*	**0.8179**	**0.5156**	116.920	129.155	138.947
LN	0.7211	0.4249	119.247	154.324	149.128
LN +L1	0.7994	0.4990	116.964	152.100	147.937
LN-s	0.8083	0.4863	102.492	133.812	118.786
LN*	0.8091	0.4957	**101.725**	129.823	**116.609**
DN	0.8058	0.4892	103.714	132.143	118.789
DN*	0.8122	0.5066	102.238	**123.652**	117.868

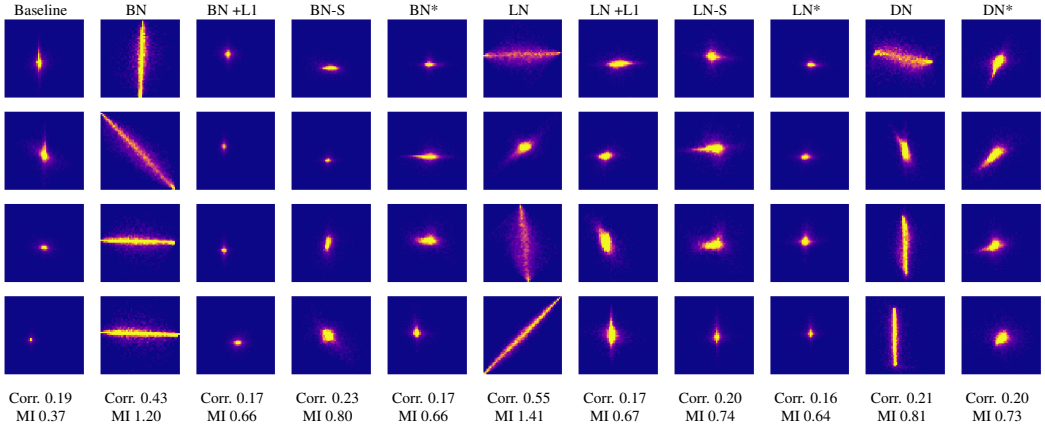


Figure 5: First layer CNN pre-normalized activation joint histogram

Acknowledgements RL is supported by Connaught International Scholarships. FS would like to thank Edgar Y. Walker, Shuang Li, Andreas Tolias and Alex Ecker for helpful discussions. Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

REFERENCES

Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20(10): 2526–63, 2008. ISSN 08997667. doi: 10.1162/neco.2008.03-07-486.

Abadi, Martín, Barham, Paul, Chen, Jianmin, Chen, Zhifeng, Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Irving, Geoffrey, Isard, Michael, Kudlur, Manjunath, Levenberg, Josh, Monga, Rajat, Moore, Sherry, Murray, Derek Gordon, Steiner, Benoit, Tucker, Paul A., Vasudevan, Vijay, Warden, Pete, Wicke, Martin, Yu, Yuan, and Zhang, Xiaoqiang. Tensorflow: A system for large-scale machine learning. *CoRR*, abs/1605.08695, 2016.

Ba, Jimmy Lei, Kiros, Jamie Ryan, and Hinton, Geoffrey E. Layer normalization. *CoRR*, abs/1607.06450, 2016.

Ballé, Johannes, Laparra, Valero, and Simoncelli, Eero P. Density modeling of images using a generalized normalization transformation. *ICLR*, 2016.

Beck, J. M., Latham, P. E., and Pouget, A. Marginalization in Neural Circuits with Divisive Normalization. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(43):15310–9, oct 2011. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1706-11.2011.

Bevilacqua, Marco, Roumy, Aline, Guillemot, Christine, and Morel, Marie-Line Alberi. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012.

Bonds, A. B. Role of Inhibition in the Specification of Orientation Selectivity of Cells in the Cat Striate Cortex. *Visual Neuroscience*, 2(01):41–55, 1989.

Busse, L., Wade, A. R., and Carandini, M. Representation of Concurrent Stimuli by Population Activity in Visual Cortex. *Neuron*, 64(6):931–942, dec 2009. ISSN 0896-6273. doi: 10.1016/j.neuron.2009.11.004.

Carandini, M. and Heeger, D. J. Normalization as a canonical neural computation. *Nature reviews. Neuroscience*, 13(1):51–62, nov 2012. ISSN 1471-0048. doi: 10.1038/nrn3136.

Coen-Cagli, R., Kohn, A., and Schwartz, O. Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, 18(11):1648–1655, 2015. ISSN 1097-6256. doi: 10.1038/nn.4128.

Cogswell, Michael, Ahmed, Faruk, Girshick, Ross, Zitnick, Larry, and Batra, Dhruv. Reducing overfitting in deep networks by decorrelating representations. *ICLR*, 2015.

Cooijmans, Tim, Ballas, Nicolas, Laurent, César, and Courville, Aaron. Recurrent batch normalization. *CoRR*, abs/1603.09025, 2016.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Dong, Chao, Loy, Chen Change, He, Kaiming, and Tang, Xiaoou. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2016.

Froudarakis, Emmanouil, Berens, Philipp, Ecker, Alexander S, Cotton, R James, Sinz, Fabian H, Yatsenko, Dimitri, Saggau, Peter, Bethge, Matthias, and Tolias, Andreas S. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature neuroscience*, 17(6):851–7, apr 2014. ISSN 1546-1726. doi: 10.1038/nn.3707.

Gatys, Leon A., Ecker, Alexander S., and Bethge, Matthias. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

Glorot, Xavier, Bordes, Antoine, and Bengio, Yoshua. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. Deep learning. Book in preparation for MIT Press, 2016.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *CVPR*, 2016.

Heeger, D. J. Normalization of cell responses in cat striate cortex. *Vis Neurosci*, 9(2):181–197, 1992. ISSN 09525238.

Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S., and Lerchner, A. Early Visual Concept Learning with Unsupervised Deep Learning. *CoRR*, abs/1606.05579, 2016.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., and LeCun, Y. What is the best multi-stage architecture for object recognition? *ICCV*, 2009.

Kavukcuoglu, K., Ranzato, M.'A., Fergus, R., and LeCun, Y. Learning invariant features through topographic filter maps. In *CVPR Workshops*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, 2012.

Laurent, César, Pereyra, Gabriel, Brakel, Philémon, Zhang, Ying, and Bengio, Yoshua. Batch normalized recurrent neural networks. *arXiv preprint arXiv:1510.01378*, 2015.

Le, Quoc V. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8595–8598. IEEE, 2013.

Liao, Q. and Poggio, T. Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. *CoRR*, abs/1604.03640, 2016.

Liao, Qianli, Kawaguchi, Kenji, and Poggio, Tomaso. Streaming Normalization: Towards Simpler and More Biologically-plausible Normalizations for Online and Recurrent Learning. *CoRR*, abs/1610.06160, 2016a.

Liao, Renjie, Schwing, Alexander, Zemel, Richard, and Urtasun, Raquel. Learning deep parsimonious representations. *NIPS*, 2016b.

Lyu, Siwei and Simoncelli, Eero P. Reducing statistical dependencies in natural signals using radial Gaussianization. *NIPS*, 2008.

Malo, J., Epifanio, I., Navarro, R., and Simoncelli, E. P. Nonlinear image representation for efficient perceptual coding. *TIP*, 15(1):68–80, 2006.

Martin, David, Fowlkes, Charless, Tal, Doron, and Malik, Jitendra. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

Olsen, S. R, Bhandawat, V., and Wilson, R. I. Divisive Normalization in Olfactory Population Codes. *Neuron*, 66(2):287–299, 2010. ISSN 10974199. doi: 10.1016/j.neuron.2010.04.009.

Pinto, N., Cox, D. D., and DiCarlo, J. J. Why is Real-World Visual Object Recognition Hard? *PLoS Comput Biol*, 4(1):e27, jan 2008. doi: 10.1371/journal.pcbi.0040027.

Reynolds, J. H. and Heeger, D. J. The normalization model of attention. *Neuron*, 61(2):168–85, jan 2009. ISSN 1097-4199. doi: 10.1016/j.neuron.2009.01.002.

Ringach, D. L. Population coding under normalization. *Vision Research*, 50(22):2223–2232, 2009. ISSN 18785646. doi: 10.1016/j.visres.2009.12.007.

Salimans, Tim and Kingma, Diederik P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *NIPS*, 2016.

Scardapane, S., Comminiello, D., Hussain, A., and Uncin, A. Group sparse regularization for deep neural networks. *CoRR*, abs/1607.00485, 2016.

Schwartz, O. and Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nat Neurosci*, 4 (8):819–825, 2001. ISSN 1097-6256. doi: 10.1038/90526.

Schwartz, O., J., Sejnowski T., and P., Dayan. Perceptual organization in the tilt illusion. *Journal of Vision*, 9(4):1–20, apr 2009. ISSN 1534-7362.

Sermanet, P., Chintala, S., and LeCun, Y. Convolutional neural networks applied to house numbers digit classification. *Proceedings of International Conference on Pattern Recognition ICPR12*, (Icpr):10–13, 2012. ISSN 1051-4651. doi: 10.0/Linux-x86_64.

Simoncelli, E. P. and Heeger, D. J. A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761, 1998.

Sinz, Fabian and Bethge, Matthias. Temporal Adaptation Enhances Efficient Contrast Gain Control on Natural Images. *PLoS Computational Biology*, 9(1):e1002889, jan 2013. ISSN 1553734X.

Sinz, Fabian H and Bethge, Matthias. The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction. In *NIPS*, 2008.

Srivastava, Nitish, Hinton, Geoffrey E, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

Timofte, Radu, De Smet, Vincent, and Van Gool, Luc. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, 2013.

Ulyanov, Dmitry, Vedaldi, Andrea, and Lempitsky, Victor S. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

Wang, Zhou, Bovik, Alan C, Sheikh, Hamid R, and Simoncelli, Eero P. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

Zaremba, Wojciech, Sutskever, Ilya, and Vinyals, Oriol. Recurrent neural network regularization. *CoRR*, abs/1409.2329, 2014.

Zeyde, Roman, Elad, Michael, and Protter, Matan. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pp. 711–730. Springer, 2010.

A EFFECT OF SIGMA AND L1 ON CIFAR-10/100 VALIDATION SET

We plot the effect of σ and L1 regularization on the validation performance in Figure 6. While sigma makes the most contributions to the improvement, L1 also provides much gain for the original version of LN and BN.

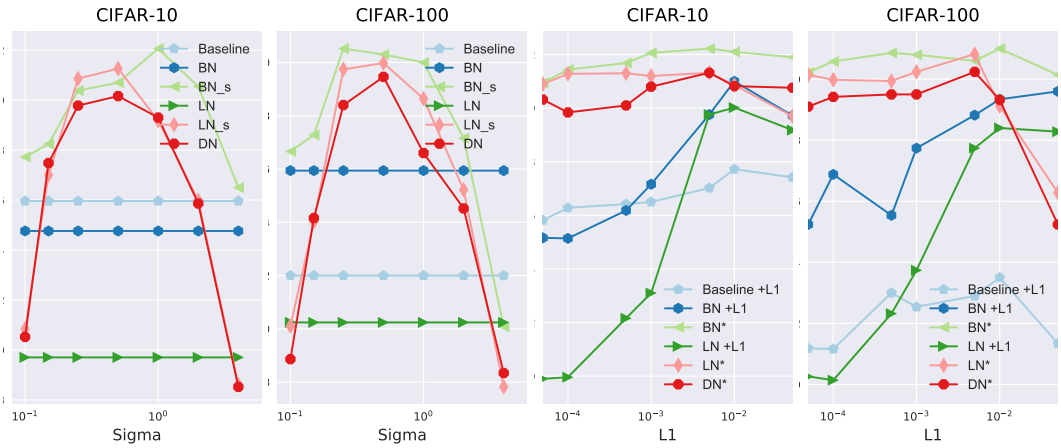


Figure 6: Validation accuracy on CIFAR-10/100 showing effect of sigma constant (a, b) and L1 regularization (c, d) on BN, LN, and DN

B LSTM IMPLEMENTATION DETAILS

In LSTM experiments, we found that have an individual normalizer for each non-linearity (sigmoid and tanh) helps the performance for both LN and DN. Eq. 12-14 are the standard LSTM equations, and let N be the normalizer function, our new normalizer is replacing the nonlinearity with Eq. 15-16. This modification can also be thought as combining normalization and activation as a single activation function.

This is different from the released implementation of LN and BN in LSTM, which separately normalized the concatenated vector $W_h\mathbf{h}_{t-1}$ and $W_x\mathbf{x}_t$. For all LN* and DN experiments we choose this new formulation, whereas LN experiments are consistent with the released version.

$$\begin{pmatrix} \mathbf{f}_t \\ \mathbf{i}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = W_h\mathbf{h}_{t-1} + W_x\mathbf{x}_t + \mathbf{b} \tag{12}$$

$$\mathbf{c}_t = \sigma(\mathbf{f}_t) \odot \mathbf{c}_{t-1} + \sigma(\mathbf{i}_t) \odot \tanh(\mathbf{g}_t) \tag{13}$$

$$\mathbf{h}_t = \sigma(\mathbf{o}_t) \odot \tanh(\mathbf{c}_t) \tag{14}$$

$$\bar{\sigma}(x) = \sigma(N(x)) \tag{15}$$

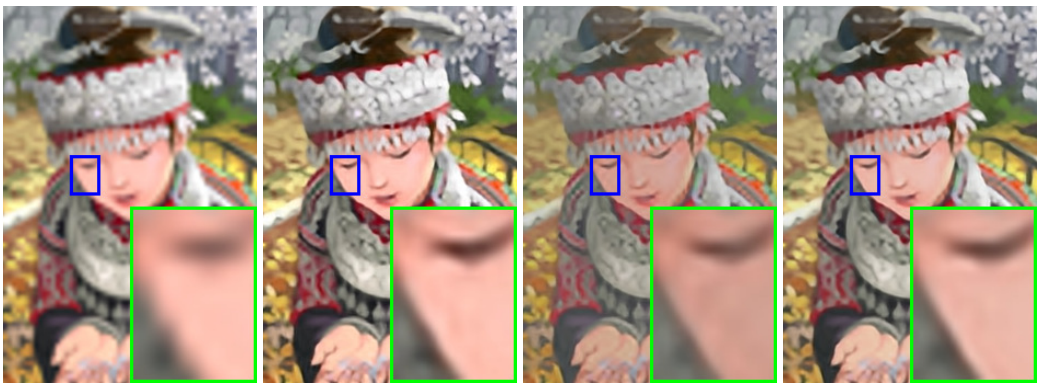
$$\overline{\tanh}(x) = \tanh(N(x)) \tag{16}$$

C MORE RESULTS ON IMAGE SUPER-RESOLUTION

We include results on another standard dataset Set5 Bevilacqua et al. (2012) in Table 8 and show more visual results in Fig. 7.

Table 8: Average test results of PSNR and SSIM on Set5 Dataset.

Model	**PSNR** (x3)	**SSIM** (x3)	**PSNR** (x4)	**SSIM** (x4)
Bicubic	30.41	0.8678	28.44	0.8097
A+	32.59	0.9088	30.28	0.8603
SRCNN	**32.83**	0.9087	30.52	0.8621
BN	22.85	0.8027	20.71	0.7623
DN*	**32.83**	**0.9106**	**30.62**	**0.8665**



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

Figure 7: Comparisons at a magnification factor of 4.