Bio-inspired learnable divisive normalization for ANNs

Vijay Veerabadran[*], **Ritik Raina**[*], **Virginia R. de Sa**[*+]
[*]Department of Cognitive Science, [+] Halıcıoğlu Data Science Institute,
University of California, San Diego
La Jolla, CA 92093
vveeraba@ucsd.edu

Abstract

In this work we introduce DivNormEI, a novel bio-inspired convolutional network that performs *divisive normalization*, a canonical cortical computation, along with lateral inhibition and excitation that is tailored for integration into modern Artificial Neural Networks (ANNs). DivNormEI, an extension of prior computational models of divisive normalization in the primate primary visual cortex, is implemented as a modular layer that can be integrated in a straightforward manner into most commonly used modern ANNs. DivNormEI normalizes incoming activations via learned non-linear within-feature shunting inhibition along with across-feature linear lateral inhibition and excitation. In this work, we show how the integration of DivNormEI within a task-driven self-supervised encoder-decoder architecture encourages the emergence of the well-known contrast-invariant tuning property found to be exhibited by simple cells in the primate primary visual cortex. Additionally, the integration of DivNormEI into an ANN (VGG-9 network) trained to perform image classification on ImageNet-100 improves both sample efficiency and top-1 accuracy on a held-out validation set. We believe our findings from the bio-inspired DivNormEI model that simultaneously explains properties found in primate V1 neurons and outperforms the competing baseline architecture on large-scale object recognition will promote further investigation of this crucial cortical computation in the context of modern machine learning tasks and ANNs.

1 Introduction

Past computational models of vision have served the important coupled goals of understanding biological vision and progressing towards creating machines with powerful visual capability. Hubel and Wiesel's seminal work characterizing receptive fields in the cat striate cortex (Hubel & Wiesel, 1968) inspired Fukushima's Neocognitron (Fukushima, 1980) (a hierarchical extension of this building block) and later the LeNet (LeCun et al., 1998) model that combined convolutional neural networks with gradient-based learning, and is the predecessor to most of today's modern ANNs that achieve tremendous success in the field of computer vision.

Following this line of brain-inspired architectures for computer vision, in this work we introduce DivNormEI, a novel computational model of divisive normalization and lateral interactions that is an extension of prior computational neuroscience models of normalization and horizontal connections (Blakeslee & McCourt, 1999; Robinson et al.; Schwartz & Simoncelli, 2001; Grossberg & Raizada, 2000; Li, 1998). Divisive normalization has been extensively studied in the fields of neuroscience and perception; these studies have highlighted the importance of this canonical computation for several traits of biological vision such as nonlinear response properties, efficient coding, invariance with respect to specific stimulus dimensions and redundancy reduction. Here, we combine divisive

3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference, Virtual.

normalization with linear lateral interactions to design a single circuit that we call DivNormEI. Different from prior models that were fit to explain behavioral/physiological data or used a small stimulus set, here we explore training DivNormEI's parameters in a task-driven fashion by optimizing performance on large-scale supervised object recognition and by optimizing self-supervised objective functions with gradient-based learning. We demonstrate the emergence of the ubiquitous contrast invariant tuning property in a self-supervised ANN equipped with DivNormEI. Additionally, we report our observation of improved large-scale object recognition accuracy on the ImageNet-100 dataset by virtue of the DivNormEI block. Comparing the tuning properties of our model's simple-cell equivalent neurons pre- and post-normalization, we observe the crucial role played by lateral connections and normalization to emergence of contrast invariance. We observed that a VGG-16 network pretrained on the large-scale ImageNet dataset did not show this property of contrast invariant tuning, further highlighting the specific role of normalization and lateral connections in developing this particular invariance. We hypothesize that the hierarchical incorporation of our DivNormEI (out of the scope of this paper) will promote development of invariance to more stimulus dimensions that shall be advantageous for high-level vision tasks such as image segmentation and object detection.

2 Methods

In this work, we extend prior computational models of divisive normalization and lateral interactions, and implement them within the framework of modern ANNs. We develop a learnable version of this nonlinear normalization computation along with linear excitatory and inhibitory lateral connections in a single module we call DivNormEI explained as follows.

2.1 DivNormEI layer: Learnable divisive normalization with lateral connections

We begin by defining a typical instantiation of divisive normalization that has been widely studied by prior art wherein, divisive normalization computes the ratio of an individual neuron's response to the summed activity of other neurons in its neighborhood. The following equation summarizes this well-studied formulation of a neuron i's normalization corresponding to an input stimulus drive x:

$$y_i(x) = \frac{y_i(x)}{\Sigma_j \mathbf{w}_j * y_j(x) + \sigma^2} \tag{1}$$

where j represents neighboring neurons of i and $\mathbf{w}, \sigma$ are learnable free parameters.

We now define our proposed learnable divisive normalization layer with lateral connections called DivNormEI. Let Y be an intermediate feature map – in response to stimulus drive X – that is being normalized by DivNormEI s.t. $Y \in \mathbb{R}^{\mathrm{h,w,c}}$. c is the number of features present in Y (e.g. if Y is the conv1 feature map in a ResNet-50 network, $c = 64$) while h, and w are the spatial dimensions of Y. In our model, each neuron in feature map k at spatial location i, j receive three kinds of activity modulation from neighboring neurons with i, j at their center: (i) divisive (or shunting) modulation that performs learned upscaling / downscaling (similar to gain control), (ii) linear excitatory modulation that positively influences activity with a learned additive operation and (iii) linear inhibitory modulation that negatively influences activity with a subtractive operation.

Once the intermediate feature map Y arrives into the DivNormEI block, the first modulatory influence described above, i.e., divisive normalization of incoming features is performed as follows:

$$\tilde{Y}_{i,j,k}(x) = \frac{Y^2_{i,j,k}(x)}{\Sigma_{m,n} \mathbf{w}^{div}_{m,n,k} Y^2_{m,n,k}(x) + \sigma^2} \tag{2}$$

Per the above equation, divisive normalization of neurons in feature map k is performed using the weighted summed activity of neighboring neurons (indexed by m, n) with a learnable 2-D depthwise-convolutional weight matrix $\mathbf{w}^{div}_k \in \mathbb{R}^{\mathrm{h_{div},w_{div}}}$ that is unique to each feature map. h_{div}, w_{div} represent the spatial extent of divisive normalization. Neurons in feature map k are normalized only by their spatial neighbors in the same feature map k. This particular design choice ensures that activations in neighboring spatial locations have reduced redundancy after divisive normalization.

The divisively normalized output $\tilde{Y}_{i,j,k}$ from Eqn. 2 is then modulated by two opposing lateral interactions: additive excitation and subtractive inhibition. Two learnable weight matrices $\mathbf{w}^{exc} \in \mathbb{R}^{\mathrm{h_e,w_e,c,c}}$ and $\mathbf{w}^{inh} \in \mathbb{R}^{\mathrm{h_i,w_i,c,c}}$ compute the weighted summed activity of neighboring neurons

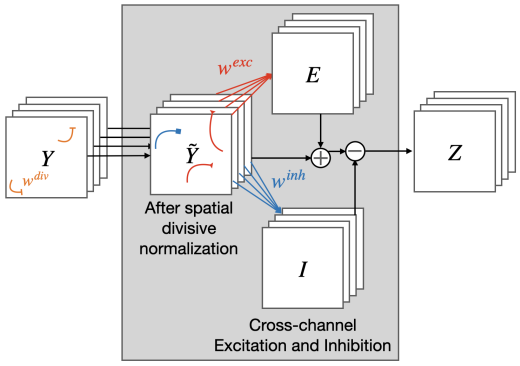

Figure 1: Architecture diagram for DivNormEI described above with spatial divisive normalization and cross-channel linear excitation and inhibition

from all feature channels for excitation and inhibition as follows (h_e, w_e and h_i, w_i denote the spatial extent of excitatory and inhibitory lateral connections):

$$E_{i,j,k}(x) = \Sigma_{m,n,o}\mathbf{w}^{exc}_{m,n,o,k} * \tilde{Y}_{m,n,o}(x) \tag{3}$$

$$I_{i,j,k}(x) = \Sigma_{m,n,o}\mathbf{w}^{inh}_{m,n,o,k} * \tilde{Y}_{m,n,o}(x) \tag{4}$$

Eqn. 3 and 4 are implemented as 2D convolutions with kernels w^{exc} and w^{inh} respectively. Linear excitation and inhibition E and I are integrated with $\tilde{Y}$ as follows to produce the normalized output:

$$\tilde{Z}_{i,j,k}(x) = \tilde{Y}_{i,j,k} + E_{i,j,k}(x) - I_{i,j,k}(x) \tag{5}$$

$$Z_{i,j,k}(x) = \gamma(BN(\tilde{Z}_{i,j,k}(x))) \tag{6}$$

In the above Eqn. 6, BN corresponds to batch normalization (Ioffe & Szegedy, 2015) and γ represents the ReLU nonlinearity. In our experiments, we initialize $\mathbf{w}_{div}$ to be a set of c 2-D Gaussian kernels to build strong local divisive inhibition that prevent redundancies with gradually decreasing inhibition from far-off neurons. However, it is to be noted that these parameters are also trained with backpropagation along with $\mathbf{w}^{exc}$ and $\mathbf{w}^{inh}$ that are initialized randomly. Unless specified otherwise, all lateral connection weights $\mathbf{w}^*$ are maintained non-negative at each step of training. In subsequent sections, we discuss our experiments training ANNs with DivNormEI using self-supervised and supervised objective functions. In all our experiments, we set $h_{div} = w_{div} = 5$, $h_e = w_e = 9$ and $h_i = w_i = 7$ based on hyperparameter optimization w.r.t classification accuracy on a custom dev set containing a proportion of ImageNet-100 train images.

3 Experiments

3.1 Experiment 1: In-silico electrophysiology with task-driven ANN

Berkeley Segmentation Dataset 500. In this experiment, we explored the self-supervised task of image super-resolution on natural images from the Berkeley Segmentation Dataset (Arbelaez et al., 2010), referred from here onward as BSDS500. For training super-resolution models in this experiment, we sample random crops of size 48x48 pixels from the training images of BSDS500 (original image size is 321x481 pixels) and use them as the high-resolution ground truth images. Corresponding low-resolution input images are obtained by down-sampling the ground truth images by a factor of 4 to size 12x12 pixels.

Encoder-decoder architecture for super-resolution. We implemented an encoder-decoder architecture for super-resolution. The encoder contains a fixed convolution layer initialized with a Gabor filter bank and a DivNormEI layer. The Gabor filter bank contains square filters of size 15px and 21px. At each filter size, we design filters selective to 4 orientations ($\theta = 0, \pi/4, \pi/2, 3\pi/4$), 2 spatial frequencies (2 cycles per degree, 3 cycles per degree) and 4 phase values ($\phi = 0, \pi/2, \pi, 3\pi/2$). The

encoder's output thus contains 64 feature maps. The decoder consists of 3 layers that upsample the encoder's output; first two layers are instantiated with transposed convolution (Zeiler & Fergus, 2014) with 64 filters each followed by batch normalization and ReLU nonlinearity. The last decoder layer is a 1x1 convolution with $tanh$ nonlinearity that produces the final output in image space.

Lateral connections encourage data-driven emergence of contrast invariant tuning Simple cells in primary visual cortex of cats and primates maintain contrast-invariant orientation tuning, i.e., the orientation selective response of neurons remains roughly steady despite varying input stimulus contrast (Troyer et al., 1998; Nowak & Barone, 2009). In this paper, we studied whether lateral connections and data-driven self-supervised learning can contribute to the emergence of this property. To study this hypothesis, we generated 100 sinusoidal grating stimuli that correspond to 25 grating orientations obtained at uniform intervals between 0 and π at 4 contrast levels (shown in Fig. 2.E). For each of the 64 feature maps in our encoder, we computed the neural tuning curves in response to the above 100 stimuli. Using stimuli at each contrast level, we computed the average of these tuning curves after ordering them such that each neuron's response to its preferred orientation stimulus was at the center of its tuning curve.

The average tuning curves **before normalization** are shown as a function of stimulus contrast in Fig.2.A, wherein the orientation selectivity decreases with decreasing stimulus contrast . This is similar to the average tuning curves of an ImageNet pretrained VGG-16 that we show in Fig. 2.B, i.e., both the self-supervised pre-normalization encoder neurons and ImageNet-pretrained VGG-16 neurons behave similarly and show a lack of contrast invariance.

On the other hand, the average tuning curve (of the same neurons in Fig.2.A) **after normalization using DivNormEI** as shown in Fig.2.C post-normalization is significantly more invariant to stimulus contrast (orientation selectivity and tuning curve variance is consistent across contrast levels). This post-normalization behavior shown in Fig. 2.C is similar to the reference behavior of cat primary visual cortical neurons we show in Fig. 2.B obtained from (Busse et al., 2009).

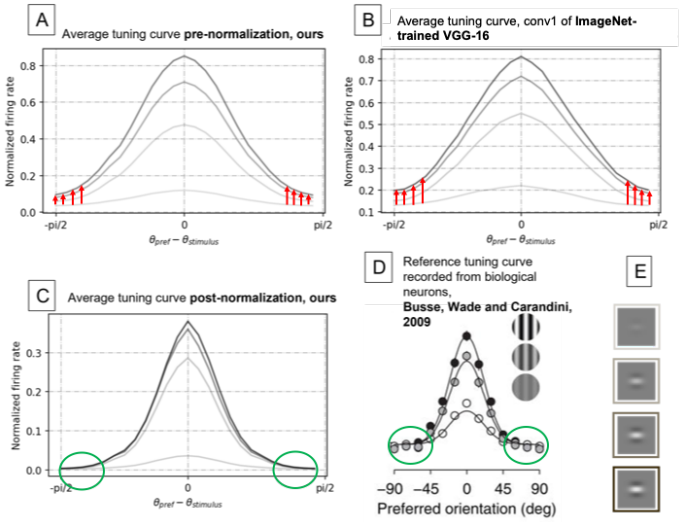

Figure 2: (A) Average tuning curve of neurons in our self-supervised encoder's output before application of DivNormEI layer. (B) Average tuning curve of the conv1 neurons of the VGG-16 model (Simonyan & Zisserman, 2014). (C) Average tuning curve of our self-supervised encoder's output after DivNormEI layer. (D) Reference of the contrast invariant tuning property in cat V1 simple cells. (Busse et al., 2009) (E) Example sinusoidal grating stimulus at four contrast levels.

3.2 Experiment 2: DivNormEI improves object recognition accuracy on ImageNet-100

In this experiment, we evaluated the utility of our proposed DivNormEI model for the computer vision task of object recognition on the ImageNet-100 dataset (a subset of the ImageNet dataset with 100 randomly sampled classes which are also present in the validation set, standardized by Tian et al. (2020)). For fast prototyping and GPU memory constraints, we created a custom shallower variant

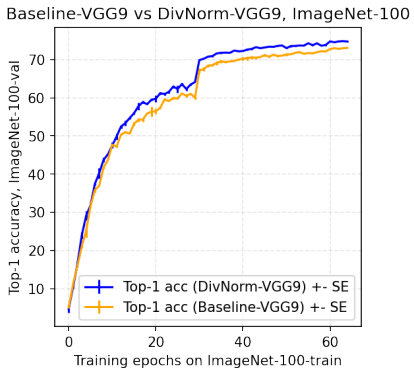

Figure 3: Image classification performance of Baseline-VGG9 and Divnorm-VGG9, error bars computed over 2 random seeds. The steep accuracy increase at epoch 30 coincides with learning rate decay for both architectures

of the VGG architecture with 9 layers of processing that we used in this experiment. We compared the following two architectures on ImageNet-100 classification: (a) Baseline-VGG9 – a 9-layer ANN with 3x3 convolution layers, max pooling and fully connected layers for classification, and (b) DivNorm-VGG9 – a 9-layer ANN with the same architecture as Baseline-VGG9, with the addition of an end-to-end trainable DivNormEI block at the output of the first convolution layer (output of DivNormEI block sent to subsequent 8 layers for classification). We trained two initializations of each of these models wherein the first pair of Baseline-VGG9 and DivNorm-VGG9 were initialized with the same weights, same as the second pair of Baseline-VGG9 and DivNorm-VGG9 models. **We observed that DivNorm-VGG9 models outperformed Baseline-VGG9 models on the ImageNet-100 validation set.** DivNorm-VGG9 had better sample efficiency than Baseline-VGG9, and was more accurate in classification by 1.8% (Top-1 validation accuracy of Baseline-VGG9: 73.3%, DivNorm-VGG9: **75.12%**) . This observation suggests that DivNormEI is also relevant to improving the discriminative power of modern ANNs and can be integrated into further computer vision solutions like image segmentation and detection that rely on object semantics and discriminability.

4 Related work

Computational models of divisive normalization and lateral connections have been studied previously by the neuroscience and vision science communities. Of these models, we find Schwartz & Simoncelli (2001)'s model to be most relevant to our divisive normalization computation, where the authors developed a simple yet neurally plausible circuit for divisive normalization optimized to maximize independence of filter responses. Our model is also related to Robinson et al.'s model that performs orientation- and spatial-frequency dependent normalization of filter responses with model parameters fit to best explain a suite of brightness illusions. Our work is also related to prior computational models of lateral interactions in the primary visual cortex such as Li (1998); Grossberg & Raizada (2000). Our proposed model is an extension of these sophisticated yet small-scale models (in terms of size and stimulus exposure) to integration within large-scale gradient-trained ANNs.

We find our work to be related to Ballé et al. (2015) where a deep network integrated with learnable Generalized Divisive Normalization modules (albeit without lateral connections that are present in our model) is trained to perform the task of image density modeling and compression. Our proposed work is also related to Burg et al. (2021) where the authors develop an end-to-end trainable model of divisive normalization similar to ours. Key differences between our work and the above work are: (i) our model performs linear lateral inhibition and excitation on top of learnable divisive normalization, (ii) parameters of our model are estimated with computer vision tasks such as super-resolution and image recognition, whereas Burg et al. (2021) train their model to predict V1 neuronal responses recorded from macaque primary visual cortex.

5 Discussion

We developed DivNormEI, a novel computational model of divisive normalization and lateral interactions – both canonical computations that are ubiquitously found in biological visual neurons associated with diverse functions such as contrast normalization, redundancy reduction, and nonlinear neuronal response properties. We conducted two experiments to address (1) the emergent biological similarity from data-driven training of our model and (2) its utility in modern ANNs trained on computer vision problems. We computed the orientation tuning curves of neurons post normalization by DivNormEI and observed their response to be invariant to input stimulus contrast. This property of contrast invariant tuning is similar to that of primary visual cortical neurons. We also tested the specific role of divisive normalization in developing contrast invariant tuning; i.e., an ImageNet-pretrained VGG-16 model exposed to a million natural images still does not possess this property. We also compared two pairs of identical convolutional architectures with the difference that one network among each pair contained a DivNormEI layer after its first convolution layer on large-scale object recognition from images in the ImageNet-100 dataset. We observed that the architectures with DivNormEI blocks possessed higher sample efficiency and classification accuracy compared to their *identical* baseline architectures without DivNormEI. We find this superior performance of DivNorm-architectures suggestive of the role of DivNormEI (and similar brain-inspired computations) in improving performance on computer vision tasks like image segmentation, where object discriminability is key. The studies in this paper were limited to specific forms of lateral interaction and by application to smaller-sized deep networks due to time-limited computational constraints. However, we believe that our promising initial findings encourage further investigation of the role and implementation of divisive normalization and other relevant lateral and recurrent cortical computations in modern ANN architectures.

6 Acknowledgments

This work was supported by the Sony Research Award program from Sony Research, the Kavli Symposium Inspired Proposals award from the Kavli Institute of Brain and Mind and support from the Department of Cognitive Science at UC San Diego.

References

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Density modeling of images using a generalized normalization transformation. *arXiv preprint arXiv:1511.06281*, 2015.

Barbara Blakeslee and Mark E McCourt. A multiscale spatial filtering account of the white effect, simultaneous brightness contrast and grating induction. *Vision research*, 39(26):4361–4377, 1999.

Max F Burg, Santiago A Cadena, George H Denfield, Edgar Y Walker, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6):e1009028, 2021.

Laura Busse, Alex R Wade, and Matteo Carandini. Representation of concurrent stimuli by population activity in visual cortex. *Neuron*, 64(6):931–942, 2009.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

Stephen Grossberg and Rajeev DS Raizada. Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision research*, 40(10-12):1413–1432, 2000.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Zhaoping Li. A neural model of contour integration in the primary visual cortex. *Neural computation*, 10(4):903–940, 1998.

Lionel G Nowak and Pascal Barone. Contrast adaptation contributes to contrast-invariance of orientation tuning of primate v1 cells. *PLoS one*, 4(3):e4781, 2009.

Alan Robinson, Paul Hammon, and Virginia de Sa. A neurally plausible model of lightness illusions combining spatial filtering and local response normalization.

Odelia Schwartz and Eero P Simoncelli. Natural signal statistics and sensory gain control. *Nature neuroscience*, 4(8):819–825, 2001.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Todd W Troyer, Anton E Krukowski, Nicholas J Priebe, and Kenneth D Miller. Contrast-invariant orientation tuning in cat visual cortex: thalamocortical input tuning and correlation-based intracortical connectivity. *Journal of Neuroscience*, 18(15):5908–5927, 1998.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.