Learning Divisive Normalization in Primary Visual Cortex

Max F. Günthner[1,*], Santiago A. Cadena[1-3], George H. Denfield[3,4], Edgar Y. Walker[3,4], Leon A. Gatys[1,2], Andreas S. Tolias[2-5,§], Matthias Bethge[1-3,§], Alexander S. Ecker[1-3,§]

[1] Institute for Theoretical Physics and Werner Reichardt Center for Integrative Neuroscience, University of Tübingen, Germany; [2] Bernstein Center for Computational Neuroscience, Tübingen, Germany; [3] Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, TX, USA; [4] Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA; [5] Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA; * max.guenthner@bethgelab.org; [§] these authors contributed equally.

Abstract

Divisive normalization (DN) has been suggested as a canonical computation implemented throughout the neocortex. In primary visual cortex (V1), DN was found to be crucial to explain nonlinear response properties of neurons when presented with superpositions of simple stimuli such as gratings. Based on such studies, it is currently assumed that neuronal responses to stimuli restricted to the neuron's classical receptive field (RF) are normalized by a non-specific pool of nearby neurons with similar RF locations. However, it is currently unknown how DN operates in V1 when processing natural inputs. Here, we investigated DN in monkey V1 under stimulation with natural images with an end-to-end trainable model that learns the pool of normalizing neurons and the magnitude of their contribution directly from the data. Taking advantage of our model's direct interpretable view of V1 computation, we found that oriented features were normalized preferentially by features with similar orientation preference rather than non-specifically. Our model's accuracy was competitive with state-of-the-art black-box models, suggesting that rectification, DN, and a combination of subunits resulting from DN are sufficient to account for V1 responses to localized stimuli. Thus, our work significantly advances our understanding of V1 function.

Keywords: primary visual cortex; natural stimulus; divisive normalization

Introduction

A crucial step towards understanding the visual system of the brain is to build models that predict neural responses to arbitrary stimuli with high accuracy (Carandini et al., 2005). The current state-of-the-art data-driven model in accurately predicting single-unit monkey V1 responses to natural stimuli is a three-layer black-box convolutional neural network (CNN) (Cadena et al., 2019). However, such multi-layer CNNs are difficult to interpret. For instance, we do not know what kind of nonlinear mapping the CNN approximates. A good candidate for such a nonlinearity is divisive normalization (Heeger, 1992) which was proposed to be a canonical neural computation throughout the visual pathway because it explains a wide variety of neurophysiological phenomena (Carandini & Heeger, 2012).

The basic idea of DN (Figure 1a) is that a unit's response

$$z_l = \frac{y_l}{\sigma_l + \sum_k p_{kl} \cdot y_k} \tag{1}$$

is given as its driving input activity y_l divisively normalized by a weighted sum over nearby units' driving inputs y_k (Carandini & Heeger, 2012; Heeger, 1992). In V1, the driving input is typically given as the half-wave rectified result of a linear filter applied to the stimulus. In the denominator, the semi-saturation constant σ_l defines how responses saturate with increasing driving input. The set of normalizing neurons k, as well as the according normalization weights p_{kl}, define which nearby neurons contribute, and with what strength, to the normalization of a given neuron l. While both are unknown, in this study we focus on neurons with the same RF location and localized stimuli covering the RF. In this setting, nonlinear effects such as cross-orientation inhibition have been described (Bonds, 1989; Busse, Wade, & Carandini, 2009; DeAngelis, Robson, Ohzawa, & Freeman, 1992; Heeger, 1992; Morrone, Burr, & Maffei, 1982).

To explain such localized normalization phenomena, it is currently assumed that the normalization weights for a given output unit are constant, $p_{kl} = p_l$ (Busse et al., 2009; Heeger, 1992). This leads to an orientation-nonspecific normalization uniformly pooling over all nearby neurons with similar RF location (Bonds, 1989; Busse et al., 2009; DeAngelis et al., 1992; Heeger, 1992; Morrone et al., 1982). However, those studies experimentally investigated the effect of DN while presenting simple stimuli such as gratings. It is not clear if these results generalize to responses to natural images, therefore limiting our understanding of visual processing in V1 (Carandini et al., 2005; Olshausen & Field, 2005).

It is currently unknown if DN is used in processing natural stimuli in V1 and if so, which types of neurons contribute to the normalization of a given unit and with what strength they do so. To answer this question, we propose an end-to-end trainable DN model to predict V1 spike counts from natural stimuli, learning all parameters directly from the data. By analyzing the learned normalization weights p_{kl}, we gain insights into how nearby neurons contribute to any given neuron's normalization pool.

Results

We investigate a dataset of 166 neurons recorded using multi-channel silicon probes in V1 of awake fixating monkeys,

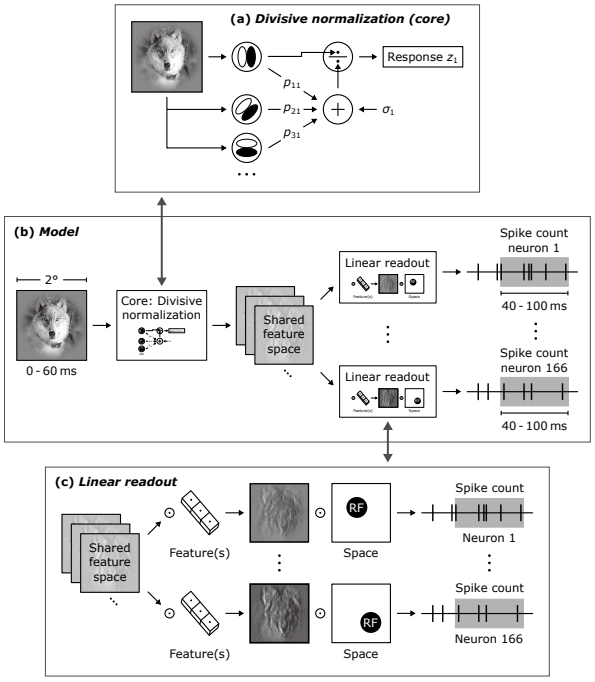

Figure 1: Model architecture. (a) Divisive normalization (DN) mechanism, model's core part (simplified). Visual input is passed through multiple linear filters. The normalized response is computed by dividing the driving input of one filter through a weighted sum of the driving inputs of all filters with normalization weights p_{kl} and the semi-saturation constant σ_l. (b) Data and model architecture. Our model predicts the spike counts for each of 166 neurons in a time window 40–100 ms after stimulus onset. The model is split into a core part, including 32 DN mechanisms (a), and a readout part (c). (c) Linear readout mapping the shared feature space to each neuron's spike count through an individual weighted sum over the entire feature space. Readout weights are factorized in feature weightings and a receptive field (RF) location mask.

who viewed a fast sequence of natural images and textures (Cadena et al., 2019). Images covered 2° of visual angle and were flashed for 60 ms without blanks in between (Figure 2). In this work, our goal is to predict the spike counts extracted in a time window of 40–100 ms after image onset (Figure 1b), accounting for typical response latencies in V1.

Learnable Divisive Normalization Model

We fit our DN model to all recorded neurons simultaneously (Figure 1). Fitting it to each neuron individually would be intractable because all normalizing subunits would have to be implicitly learned for each neuron separately, for which there is not enough data from individual neurons. Instead, when fitting the model to all neurons simultaneously, all neural responses are already provided which can be used by the normalization. In addition, it is sufficient to learn 32 features (indexed by l in

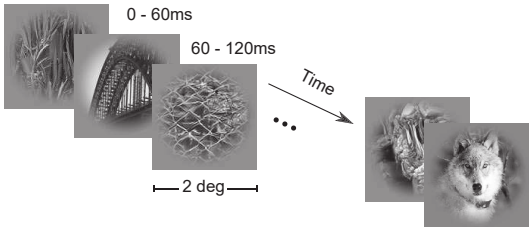

Figure 2: Natural stimuli covering two degrees visual angle, shown in fast sequence. Images were centered on RFs of recorded neurons. Adapted from Cadena et al. (2019).

Table 1: Accuracy (fraction of explainable variance explained, FEV) of different models. * Fitted to each neuron individually.

Model	**FEV (%)**
Linear-nonlinear Poisson (Cadena et al., 2019)	16.3*
Energy model	45.9
Nonspecific divisive normalization	47.8
Divisive normalization (ours)	**48.5**
Black-box 3-layer CNN (Cadena et al., 2019)	49.8

Equation 1) shared by all neurons, since many neurons in V1 perform similar computations. We achive both by leveraging the idea by Klindt, Ecker, Euler, and Bethge (2017) to split the model into two parts (Figure 1b).

In the first *core* part (Figure 1a), we learn our DN model. To match a more general formulation of DN (Carandini & Heeger, 2012), we compute the driving inputs $y_l = [\max(0, \mathrm{BN}(w_l * x))]^{n_l}$ by convolution with 32 kernels w_l of spatial size $13\,\mathrm{px} \times 13\,\mathrm{px}$. Stimuli x were downsampled by factor of two and cropped, keeping the central $46\,\mathrm{px} \times 46\,\mathrm{px}$ ($\approx 1.3^\circ$ visual field). Batch normalization without rescaling (BN) (Ioffe & Szegedy, 2015) leads to responses of unit variance and in the denominator of Equation 1 we use low-pass filtered inputs, $y_k \leftarrow \langle y_k \rangle$ (average pooling over $5\,\mathrm{px} \times 5\,\mathrm{px}$). When fitting our model, all parameters are learned, leading to a nonlinear feature-space shared by all neurons.

In the second *readout* part (Figure 1c), we map the learned feature maps z_l to the activity of individual neurons via a linear readout for each neuron, similar to previous work (Cadena et al., 2019; Klindt et al., 2017). To ensure that the readout does not model any complex computations nor contribute to normalization, we constrain the weights to be non-negative and factorize them into a location mask encoding a neuron's RF times a vector of feature weights. Additionally, we impose an L_1 sparseness prior.

Accuracy of DN Model is Competitive

Our DN model achieves an accuracy competitive to the data-driven state-of-the-art model of Cadena et al. (2019). We define the models' accuracies (Table 1) as fraction of explainable variance explained (FEV), which is the fraction of the stimulus-

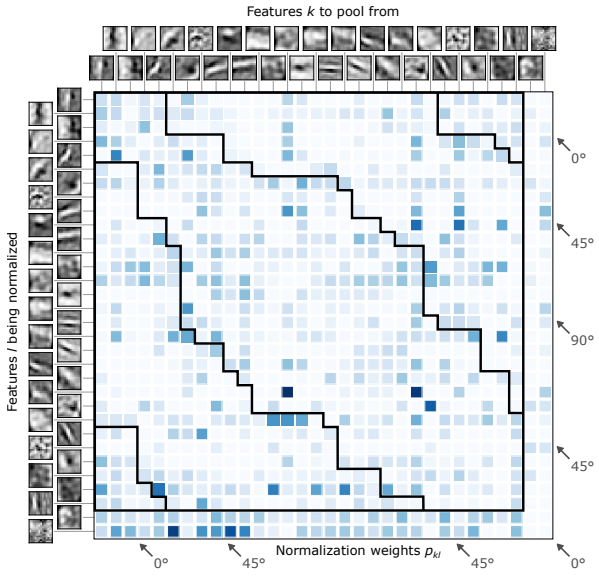

Figure 3: Normalization weights (white and blue squares) for features ordered by orientation. Last two features (bottom and right) show no orientation. Weights in diagonal directions (arrows) correspond to approximately constant orientation difference between features l being normalized and features k that can contribute to normalization. Diagonal lines mark 45° orientation difference boundary. Weights for similar orientation ($< 45^\circ$) are higher (darker color) compared to dissimilar orientation ($\geq 45^\circ$). Model with highest accuracy.

driven response that is explained by the model, ignoring unexplainable trial-to-trial variability in the response of the neurons. Thus, a perfect model would reach 100% FEV. Our DN model achieved an accuracy of 48.5% FEV, performing almost as well as the state-of-the-art black-box CNN, which reached a FEV of 49.8% (Cadena et al., 2019).

Removing the trainable DN module (Equation 1) from our full model, just keeping 32 channels of linear-nonlinear drives y_l directly followed by the readout, leads to an model that is able to approximate complex cells. Therefore, we refer to it as energy model (Adelson & Bergen, 1985), which reaches an accuracy of 45.9% FEV. The drop of 2.6 percentage points compared to our full model supports the hypothesis that DN is an important computational mechanism in V1 under stimulation with natural images. The linear-nonlinear Poisson model (Simoncelli, Paninski, Pillow, & Schwartz, 2004) is another classical standard model of V1. It was fit to each neuron in our dataset individually, reaching an even lower accuracy of 16.3% FEV (Cadena et al., 2019).

Normalization is Feature-Specific

We now investigate the structure of the learned normalization pool, i. e. the weights p_{kl} of the sum in the denominator of Equation 1: The higher the weight, the stronger feature l is normalized by feature k. For this analysis, we focus on

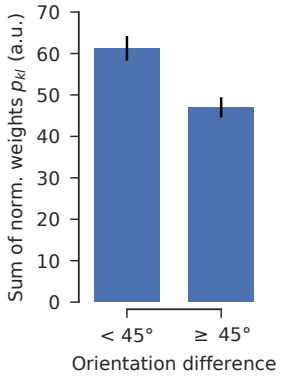

Figure 4: Sum of normalization weights averaged over models with FEV $> 48.0\%$ (38 models). For similarly oriented features ($< 45^\circ$) sum is 30.3% higher than for features with dissimilar orientation ($\geq 45^\circ$). Error bars show standard error across models. Difference is statistically significant (Wilcoxon signed rank test; $p < 1.1 \cdot 10^{-7}$, $N = 38$).

orientation-selective features. We find that weights are higher if the feature to be normalized and the normalizing feature exhibit similar orientation (Figure 3). In contrast, strongly differing orientations lead to lower weights.

To quantify this difference in contribution to normalization, we split the sum in Equation 1 into two parts: The contribution of features k with orientation similar to the driving feature l and that of features with dissimilar orientations. For each weight, we calculate the angle difference for its driving and normalizing feature by detecting the global maximum in the features' power spectrum. We assign the weights into a class of similar orientations (orientation difference $< 45^\circ$) and dissimilar orientation (orientation difference $\geq 45^\circ$). Subsequently, we sum up the weights in each of the two classes separately. We repeat this procedure for all models with an FEV greater than 48.0% (38 models) and average the sum of weights for similar and dissimilar oriented features across models. This analysis confirms our observation from above: The sum of weights accounting for normalization by similar orientations is 30.3% higher than the sum of weights accounting for dissimilar orientation (Figure 4). This result is statistically significant across models (Wilcoxon signed rank test; $p < 1.1 \cdot 10^{-7}$, $N = 38$). Hence, similarly oriented features contribute more strongly to the normalization of oriented features than dissimilarly oriented ones.

In a control experiment we fit a feature non-specific normalization model with constant normalization weights $p_{kl} = p_l$. This model's accuracy is 47.8% FEV, which is below that of our more general DN model in our main experiment, suggesting that indeed feature-specific normalization is necessary to account for V1 responses.

Discussion

Previous experimental work investigated suppressive phenomena within the RF only with simple stimuli, mainly consisting of a combination of driving and mask gratings. Some of them encountered weak orientation-specific phenomena in few cells, but all concluded that normalization is predominantly orientation-nonspecific (Bonds, 1989; Busse et al., 2009; DeAngelis et al., 1992; Heeger, 1992; Morrone et al., 1982). Thus, our findings do not stand entirely in contrast to previous experimental results, but we quantitatively refine them using a larger dataset of V1 responses to natural images: We find that oriented features are preferentially normalized by channels with similar orientation. The reason for the difference between our results and previous studies could be that we use natural stimuli, which have different image statistics compared to simple stimuli. Furthermore, most previous studies of DN were performed in cats, however, orientation-specific DN could be more specific to primate rather than cat visual cortex.

Our discovery of DN by similar orientations matches the implementation of the connectivity of neurons in mouse V1: Inhibitory parvalbumin-expressing interneurons strongly inhibit those excitatory pyramidal cells that share their visual selectivity (Znamenskiy et al., 2018). Furthermore, our empirical findings are consistent with a normative model: From the efficient coding hypothesis, Schwartz and Simoncelli (2001) derived an ecologically justified DN model which implies that normalization weights should not be uniform.

In conclusion, we developed a model consisting of one layer of subunits followed by a learned orientation-specific DN. We have no evidence that any additional computation might be missing to account for responses to localized natural stimuli in V1, given that this model performs with an accuracy competitive with state-of-the-art black-box models. Hence, our work significantly improves our understanding of V1 function and DN under conditions close to real-world visual stimulation.

Acknowledgements

We thank Fabian H. Sinz and David Klindt for valuable discussions. This work was supported by: German Federal Ministry of Education & Research (BMBF), Competence Center for Machine Learning (FKZ 01IS18039A); German Research Foundation (DFG) grant EC 479/1-1 (A.S.E.) & Collaborative Research Center (SFB 1233); National Eye Institute, National Institutes of Health, Award R01EY026927 (A.S.T.), DP1 EY023176 (A.S.T.), NIH-Pioneer Award DP1-OD008301 (A.S.T); International Max Planck Research School for Intelligent Systems (IMPRS-IS) (M.F.G. & S.A.C.); NEI/NIH Core Grant for Vision Research (EY-002520-37); NEI training grant T32EY00700140 (G.H.D.) & F30EY025510 (E.Y.W.); German National Academic Foundation (M.F.G. & L.A.G.); Intelligence Advanced Research Projects Activity (IARPA), Department of Interior/Interior Business Center (DoI/IBC) contract no. D16PC00003.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*, 284-299.

Bonds, A. B. (1989). Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience*, *2*, 41-55.

Busse, L., Wade, A. R., & Carandini, M. (2009). Representation of concurrent stimuli by population activity in visual cortex. *Neuron*, *64*, 931-942.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, *15*, e1006897.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., ... Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, *25*, 10577-10597.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*, 51-62.

DeAngelis, G. C., Robson, J. G., Ohzawa, I., & Freeman, R. D. (1992). Organization of suppression in receptive fields of neurons in cat visual cortex. *Journal of Neurophysiology*, *68*, 144-163.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181-197.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Klindt, D., Ecker, A. S., Euler, T., & Bethge, M. (2017). Neural system identification for large populations separating "what" and "where". In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3506–3516). Curran Associates, Inc.

Morrone, M. C., Burr, D. C., & Maffei, L. (1982). Functional implications of cross-orientation inhibition of cortical visual cells. I. Neurophysiological evidence. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *216*, 335-354.

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding V1? *Neural computation*, *17*, 1665–1699.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*, 819-825.

Simoncelli, E. P., Paninski, L., Pillow, J., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, *3*, 327–338.

Znamenskiy, P., Kim, M.-H., Muir, D. R., Iacaruso, M. F., Hofer, S. B., & Mrsic-Flogel, T. D. (2018). Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. *bioRxiv*, 294835.