Integrating Flexible Normalization into Mid-Level Representations of Deep Convolutional Neural Networks

Luis Gonzalo Sanchez Giraldo, Odelia Schwartz

{`lgsanchez, odelia`}`@cs.miami.edu`

Computer Science Department

University of Miami

Coral Gables, FL 33146

Keywords: visual cortex, convolutional neural networks, divisive normalization

Abstract

Deep convolutional neural networks (CNNs) are becoming increasingly popular models to predict neural responses in visual cortex. However, contextual effects, which are prevalent in neural processing and in perception, are not explicitly handled by current

CNNs, including those used for neural prediction. In primary visual cortex, neural responses are modulated by stimuli spatially surrounding the classical receptive field in rich ways. These effects have been modeled with divisive normalization approaches, including flexible models, where spatial normalization is recruited only to the degree responses from center and surround locations are deemed statistically dependent. We propose a flexible normalization model applied to mid-level representations of deep CNNs as a tractable way to study contextual normalization mechanisms in mid-level cortical areas. This approach captures non-trivial spatial dependencies among mid-level features in CNNs, such as those present in textures and other visual stimuli, that arise from tiling high order features, geometrically. We expect that the proposed approach can make predictions about when spatial normalization might be recruited in mid-level cortical areas. We also expect this approach to be useful as part of the CNN toolkit, therefore going beyond more restrictive fixed forms of normalization.

1 Introduction

It has long been argued that an important step in understanding the information processing mechanisms in the brain is to understand the nature of the input stimuli (Attneave, 1954; Barlow, 1961). Visual processing of natural images is a paradigmatic example that has been studied extensively (Simoncelli and Olshausen, 2001; Zhaoping, 2006, 2014; Olshausen and Lewicki, 2014; Geisler, 2008; Hyvrinen et al., 2009). Structure in images can be captured in the form of statistical dependencies among the responses of filters acting on the image at different scales, orientations, and spatial locations (e.g.

Bell and Sejnowski, 1997; Olshausen and J.Field, 1997; Hyvrinen et al., 2009). These regularities often manifest in a nonlinear fashion (Simoncelli, 1997; Wegmann and Zetzsche, 1990; Zetzshe and Nuding, 2005; Golden et al., 2016). Therefore, it is natural to think that neural processing systems employ nonlinear operations to exploit dependencies, as they encode information about the input stimulus.

Both perception and neural responses are influenced by the spatial context, i.e., by stimuli that spatially surround a given point in space. Spatial contextual influences beyond the classical receptive field have been extensively documented for neurons in primary visual cortex (e.g. Levitt and Lund, 1997; Sceniak et al., 1999; Cavanaugh et al., 2002a,b). Models that are based on nonlinear statistical regularities across space in images have been able to capture some of these effects (Rao and Ballard, 1999; Schwartz and Simoncelli, 2001; Spratling, 2010; Karklin and Lewicki, 2009; Zhu and Rozell, 2013; Coen-Cagli et al., 2012; Lochmann et al., 2012).

Here, we focus on divisive normalization, a nonlinear computation that has been regarded as a canonical computation in the brain (Heeger, 1992; Carandini et al., 1997; Carandini and Heeger, 2012). From a coding perspective, divisive normalization acts as a transformation that reduces nonlinear dependencies among filter activation patterns in natural stimuli (e.g. Schwartz and Simoncelli, 2001). Different forms of divisive normalization have been considered in modeling spatial contextual interactions among cortical neurons. In its basic form, the divisive normalization operation is applied uniformly across the entire visual field. However, spatial context effects in primary visual cortex are better explained by a weighted normalization signal (e.g. Cavanaugh et al., 2002a,b; Schwartz and Simoncelli, 2001). Recently, more sophisticated models that

recruit normalization in a nonuniform fashion (Coen-Cagli et al., 2012) have shown better generalization at predicting responses of V1 neurons to natural images (Coen-Cagli et al., 2015). The rationale behind this form of *flexible normalization* (and related predictive coding models of (Spratling, 2010; Lochmann et al., 2012)) is that contextual redundancies vary with stimulus. In the flexible normalization model, divisive normalization is therefore only recruited at points where, according to the model, the pool of spatial context filter responses to an image are statistically dependent with the filter responses in a center location. This relates to highlighting salient information by segmentation in regions of the image in which the spatial homogeneity breaks down (Li, 1999).

As basic computational modules, it would be expected that non-linearities take place at different stages of the cortical processing hierarchy. However, studying these operations beyond the primary visual cortex level, for instance understanding when normalization is recruited for natural images, has been rather difficult. Learning models of surround divisive normalization in primary visual cortex has often relied on access to individual neural unit responses, which are then combined (e.g., in a weighted manner) to produce the modulation effect from the pool of units. In comparison to primary visual cortex, where different features such as orientation, spatial frequency and scale have a fairly well understood role in characterizing visual stimuli, the optimal stimulus space for intermediate cortical levels is less well understood (Poggio and Anselmi, 2016).

For instance, in V2, studies have previously shown selectivity to conjunctions of orientations (e.g., Ito and Komatsu, 2004), to figure-ground (e.g., Zhou et al., 2000; Zhaoping, 2005), and to texture stimuli (e.g., Freeman et al., 2013; Ziemba et al.,

2016) and texture boundaries (e.g. Schmid and Victor, 2014; Rowekamp and Sharpee, 2017). Some studies have characterized contextual surround modulation in area V2 (e.g. Shushruth et al., 2009; Ziemba et al., 2018) and V4 (e.g. Kim et al., 2019). The recent findings regarding texture sensitivity have also spurred surround experiments with naturalistic textures in V2 (Ziemba et al., 2018). We believe developing computational models of surround normalization offers a complementary route for hypothesizing what stimulus patterns might be relevant at intermediate levels.

In this work, we propose the use of deep CNNs to study how flexible normalization might work at intermediate level representations. CNNs have shown intriguing ability to predict neural responses beyond primary visual cortex (Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Cichy et al., 2016), including recent studies modeling neurophysiology data from area V2 (Laskar et al., 2018) and V4 (Pospisil et al., 2016, 2018). As we move up in the hierarchy, neural units at a given level combine the responses from early processing stages lending to a larger repertoire of possible stimuli acting at the higher level. In addition, CNNs have interestingly incorporated simplified forms of normalization (Jarrett et al., 2009; Krizhevsky et al., 2012; Ren et al., 2017). CNNs can therefore provide a tractable way to model representations that might be employed by intermediate levels of the visual processing hierarchy, such as secondary visual cortex (V2). Here, we integrate flexible normalization into the AlexNet CNN architecture (Krizhevsky et al., 2012), although our approach can be more broadly applied to other CNN and hierarchical architectures.

For intermediate level representations, we show that incorporating flexible normalization can capture non-trivial spatial dependencies among features such as those

present in textures, and more generally, geometric arrangements of features tiling the space. One instance of such geometric arrangements are texture boundaries where features detecting such boundaries are likely to align. Our focus here is on developing the framework for the CNN and demonstrating the learned statistics and spatial arrangements that result for intermediate layers of the CNN. We expect the proposed approach can make predictions about when spatial normalization might be recruited in intermediate areas, and therefore will be useful for interplay with future neuroscience experiments as well as become a standard component in the CNN toolkit.

1.1 Contributions of this work

Divisive normalization is ubiquitous in the brain (Carandini and Heeger, 2012), but contextual surround influences in visual cortex have mostly been studied in area V1. In primary visual cortex, models such as steerable pyramids or Gabor filters provide a good account for the single-unit selectivity of the neural receptive field, and are often used as a front end to more sophisticated normalization models. Such models learn statistical dependencies between units in center and surround locations and then apply an appropriate computation the reduces them. On the other hand, models of intermediate cortical neural unit selectivity of responses, such as visual cortical area V2, have been more elusive. Consequently, there is less understanding about what patterns of statistical dependencies emerge across space at these intermediate levels. Understanding such patterns of dependencies would be a basis for formulating models that reduce the dependencies and make predictions about when normalization is recruited in intermediate areas such as V2.

In this paper, we focus on characterizing the statistics and learning what statistical patterns of dependencies might emerge for intermediate cortical visual area model units. We then learn a model for reducing such dependencies. In the discussion, we elaborate on (i) predictions of our framework and future application to understanding V2 data beyond the classical receptive field; and (ii) the potential benefit and future application of testing the impact of such computations downstream.

We largely focus on the second convolutional layer of AlexNet. Second layer neural units combine V1-like features, captured by the first convolutional layer units, into bigger receptive fields. Since the network is constrained to capture information relevant to natural images, we expect the second layer will only learn such structures that are meaningful and not all possible combinations of simple features. We also examine and compare to other layers of AlexNet, namely the first and third convolutional layers.

The model of contextual interaction we propose to use is a normative model based on the premise that one of the purposes of normalization is to remove high order dependencies that cannot be removed by linear computations or point-wise non-linearities and that extend beyond the classical receptive field (This also means they extend beyond the reach of the max pooling layers). This class of model has been used to explain V1 contextual influences, but it has not been applied to higher order units (Coen-Cagli et al., 2009, 2012). Our results for the second layer offer potential predictions about when normalization might be recruited in areas beyond V1. This approach could be adapted to other hierarchical architectures and higher layers, and thus has more general applicability.

From a technical standpoint, models such as the mixture of Gaussian Scale Mixtures

(GSM)s and flexible normalization have been studied extensively for V1. Our main technical contribution is making these models applicable to units at intermediate stages where, unlike V1, units lack compact descriptors such as scale and orientation, and demonstrating the approach on CNN units. For V1 filters, models like (Coen-Cagli et al., 2009) enforce symmetry constraints on the covariance matrix to equalize the variances of surround units in opposing spatial locations, given the orientation structure of the receptive fields. The rationale was that for basic features such as oriented edges, symmetry would be expected: a vertical edge is equally likely on average to have a dependency with a vertical edge on the top or bottom, and to the left and right. Without this constraint, the model learned variances were sometimes skewed to one side. In intermediate areas, one would still expect certain symmetries, but we cannot assume the symmetry directions based on the orientation, since the filter structure in the CNN is not characterized by orientation. We found that learning proceeds well without having to incorporate symmetry constraints by modifying the model of (Coen-Cagli et al., 2009) as described in more detail in Section 4.

2 Normalization in Deep Neural Nets

Recently, new forms of normalization have been introduced to the deep neural networks tool set (Ioffe and Szegedy, 2015; Ba et al., 2016). The motivation for these computations is different from the divisive normalization models in neuroscience, which are based on observations of neural responses. Batch normalization (Ioffe and Szegedy, 2015) is a popular technique aimed at removing the covariate shift over time (i.e., in

batches) in each hidden layer unit, with the goal of accelerating training by maintaining global statistics of the layer activations. Layer normalization (Ba et al., 2016) on the other hand, employs averages across units in a given layer (and space in the case of convolutional networks) at every time step, introducing invariances in the network that benefit the speed of learning. Batch and layer normalization provide better conditioning of the signals and gradients that flow through the network, forward and backwards, and have been studied from this perspective.

Simple forms of divisive normalization that draw inspiration from neuroscience, such as those described in (Jarrett et al., 2009; Krizhevsky et al., 2012), have been used to improve the accuracy of deep neural network architectures for object recognition. However, the empirical evaluation of deeper architectures in (Simonyan and Zisserman, 2015) reached a different conclusion showing that the inclusion of local response normalization (LRN), where responses are normalized by the activity of other filters at the same spatial location, did not offer any significant gains in accuracy. One possible yet untested hypothesis for this case is that the increased depth may be able to account for some of the nonlinear behavior associated with LRN. Nevertheless, it is important to note that these empirical conclusions have only considered simple and fairly restrictive forms of normalization and measured their relevance solely in terms of classification accuracy. While accuracy with and without normalization can be the same for the standard benchmarks, other criteria such robustness to adversarial examples or other forms of noise could be used to evaluate the role of normalization.

Recent work that attempts at unifying the different forms of normalization discussed above has started to reconsider the importance of normalization for object recognition,

in the context of supervised deep networks (Ren et al., 2017). In their work, divisive normalization is defined as a localized operation in space and in features where normalization statistics are collected independently for each sample. Divisive normalization approaches arising from a generative model perspective have also been recently introduced (e.g. Balle et al., 2016). Other work on supervised networks inspired by primary visual cortex circuitry has proposed normalization as a way to learn a discriminant saliency map between a target and its null class (Han and Vasconcelos, 2010, 2014). Although these works extend beyond the simple normalization forms discussed in previous paragraphs, they are still limited to fixed normalization pools and to early stages of processing. None of these approaches have thus far considered the role of spatial dependencies and normalization in intermediate layers of CNNs to address questions in neuroscience.

Our work extends the class of flexible normalization models considered in (Coen-Cagli et al., 2012, 2015) which stem from a normative perspective where the division operation relates to (the inverse of) a generative model of natural stimuli. In previous work, flexible normalization models were learned for an oriented filter bank akin to primary visual cortical filters. Here, we develop a flexible normalization model that can be applied to convolution filters in deep CNNs. Our objective in this paper is to develop the methodology and to study the statistical properties and the structure of the dependencies that emerge in middle layers (specifically, we focus on the second convolutional layer of AlexNet). We expect our model to be useful in providing insights and plausible hypotheses about when normalization is recruited in visual cortical areas beyond primary visual cortex.

3   Background

We describe the Gaussian Scale Mixture and flexible normalization model, which serves as a background to our modeling.

3.1   Statistical model for divisive normalization

A characteristic of natural stimuli is that the coefficients obtained by localized linear decompositions such as wavelets or independent component analysis are highly non-Gaussian, generally depicting the presence of heavy tailed distributions (Field, 1987). In addition, these coefficients, even if linearly uncorrelated, still expose a form of dependency where the standard deviation of one coefficient can be predicted by the magnitudes of related coefficients across space, scale, and orientation (Simoncelli, 1997). In this sense, models that extend beyond linearity are needed to deal with nonlinear dependencies that arise in natural stimuli.

A conceptually simple yet powerful generative model that can capture this form of coupling is known as the Gaussian Scale Mixture (Andrews and Mallows, 1974; Wainwright and Simoncelli, 2000; Wainwright et al., 2001). In this class of models, the multiplicative coordination between filter activations is captured by incorporating a common mixer variable where local Gaussian variables are multiplied by this common mixer. Since in this generative model dependencies arise via multiplication, one can reduce the dependencies and estimate the local Gaussian variable via the inverse operation of division. The Gaussian variables may themselves be linearly correlated, which amounts to a weighted normalization.

Formally, a random vector X containing a set of m coupled activations is obtained

by multiplying an independent positive scalar random variable V (which we denote the mixer variable) with an m-dimensional Gaussian random vector G with zero mean and covariance $\mathbf{\Lambda}$, that is, $X = VG$. The random variable $X|V = v$ is a zero mean Gaussian random variable with covariance $\mathbf{\Lambda}v^2$, and X is distributed with pdf:

$$p_X(\mathbf{x}) = \int_0^\infty \frac{v^{-m}}{(2\pi)^{m/2}|\mathbf{\Lambda}|^{1/2}} \exp\left(-\frac{\mathbf{x}^\mathrm{T}\mathbf{\Lambda}^{-1}\mathbf{x}}{2v^2}\right)p_V(v)\mathrm{d}v. \tag{1}$$

For analytical tractability, we consider the case where the mixer V is a Rayleigh distributed random variable with pdf, $p_V(v) = \frac{v}{h^2}\exp\left(-\frac{v^2}{2h^2}\right)$, for $v \in [0,\infty)$, and scale parameter h. Integrating over v yields the following pdf:

$$p_X(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}|\mathbf{\Lambda}|^{1/2}h^m}a^{1-m/2}K_{m/2-1}(a), \tag{2}$$

where $K_\lambda(\cdot)$ is the modified Bessel function of the second kind, and

$$a^2 = \frac{\mathbf{x}^\mathrm{T}\mathbf{\Lambda}^{-1}\mathbf{x}}{h^2}. \tag{3}$$

To ease notation, we can let $\mathbf{\Lambda}$ absorb the scale parameter h. Reversing the above model to make inferences about G given X, results in an operation similar to divisive normalization. Given an instance $\mathbf{x}$ of X, we can compute the conditional expectation of the ith element of G as follows:

$$\mathbb{E}\left[g_i|\mathbf{x}\right] = \frac{x_i}{\sqrt{a}}\frac{K_{\frac{m-1}{2}}(a)}{K_{\frac{m}{2}-1}(a)}. \tag{4}$$

The divisive normalization is weighted, due to the term a, which incorporates the inverse of the covariance matrix in the computations of the normalization factor.

3.2 Flexible contextual normalization as a mixture of Gaussian scale mixtures

The Gaussian Scale Mixture model described above captures the coordination between filter activations (e.g., for receptive fields that lie in nearby spatial locations) through a single mixer variable. The normalization operation produced by performing inference on the GSM model is replicated across the entire image, which intrinsically assumes the statistics to be homogeneous across space. However, the statistical dependency between filter activations may vary depending on the particular visual image and set of filters, such as if the filters cover a single visual object or feature, or are spaced across the border of objects in a given image (Schwartz et al., 2006, 2009).

A more sophisticated model (Coen-Cagli et al., 2009) (see also Coen-Cagli et al., 2012), uses a two-component mixture of GSMs,

$$p_X(\mathbf{x}) = \Pi_{cs} p_X(\mathbf{x}|\mathbf{\Lambda}_{cs}) + (1 - \Pi_{cs}) p_{X_c}(\mathbf{x}_c|\mathbf{\Lambda}_c) p_{X_s}(\mathbf{x}_s|\mathbf{\Lambda}_s), \tag{5}$$

where $\mathbf{x}_c$ and $\mathbf{x}_s$ denote the set of responses from units with receptive fields in center and surround locations, and $\mathbf{\Lambda}_{cs}, \mathbf{\Lambda}_c, \mathbf{\Lambda}_s$ are the parameters that capture covariant structure of the neural responses, and Π_{cs} is the prior probability that center and surround are dependent. The subscript cs denotes parameters of the center-surround dependent component of the model, and c and s the respective center and surround parameters of the center-surround independent component. In this model, normalization is only recruited to the degree center and surround responses are deemed as statistically dependent. The first term of Eq.(5), $p_X(\mathbf{x}|\mathbf{\Lambda}_{cs})$, corresponds to center-surround dependent units. In the center-surround dependent component, the responses are coupled

13

linearly by the covariance $\mathbf{\Lambda}_{cs}$ and nonlinearly by the multiplicative mixer. The product $p_{X_c}(\mathbf{x}_c|\mathbf{\Lambda}_c)p_{X_s}(\mathbf{x}_s|\mathbf{\Lambda}_s)$ in the second term represents the case of statistical independence between the center group and the surround group.

Note that the covariance matrix (which provides a tuned or weighted normalization) is fixed for a given center-surround group. The covariance structure is learned for a given group of filters over an ensemble of images, and meant to capture the neural responses that may be co-active. We presume the covariance structure is fixed in the brain for any stimulus it encounters (akin to other weighted normalization models). The covariance matrix in the model is therefore not stimulus dependent for a given neuron.

4   Normalization in Deep Convolutional Networks

We next describe our approach for incorporating flexible normalization into convolutional layers of deep CNNs. We also explain how we modified the mixture of GSM model of (Coen-Cagli et al., 2009) to accomodate this.

4.1   Convolutional layers and Flexible Normalization

In their most basic form, convolutional neural networks are a particular instance of feed forward networks where the affine component of the transformation is restricted by local connectivity and weight sharing constraints. Convolutional layers of deep CNNs are arrangements of filters that uniformly process the input over the spatial dimensions. On 2-dimensional images, each of the CNN filters linearly transforms a collection of 2-dimensional arrays called input channels. For instance, RGB images are 2-dimensional
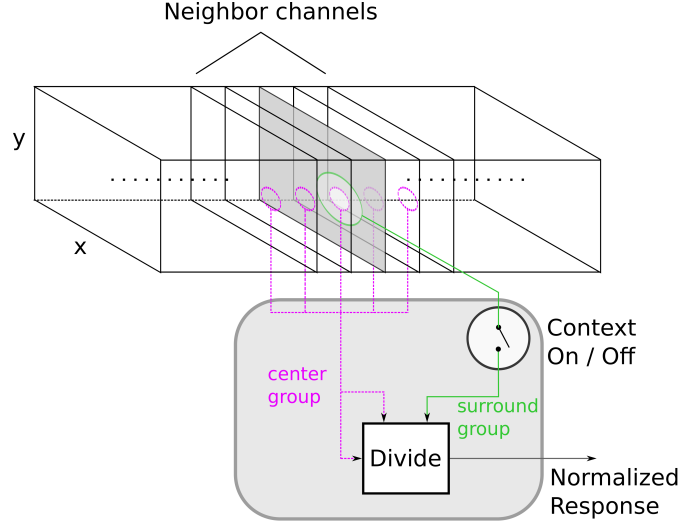
Figure 1: Schematic of flexible normalization on map computed by a convolutional layer of a deep CNN. As with flexible normalization, the surround normalization is gated and determined based on inference about statistical dependencies across space. To compute the normalized response of a filter k at location (x, y), the model uses responses from adjacent filters (channels) in the arrangement (akin to cross-orientation suppression in primary visual cortex) as the center group, and a set of responses from the same filter k at relative displacements from the (x, y) position to form the surround group (spatial context).

images with 3 channels. The output produced by each filter is a 2-dimensional array of responses called a map. Therefore, each convolutional layer produces as many output maps as filters. Let $I_{\mathrm{in}}(x, y, \ell)$ be the collection of 2-d input arrays, where x and y denote the spatial indexes and $\ell \in \mathcal{X}_{\mathrm{in}}$ the input channel index. A convolutional layer is a collection of 3-dimensional arrays $\{W_k(x, y, \ell)\}_{k \in \mathcal{C}_{\mathrm{out}}}$. The operation of convolution, which yields a map, is defined as:

$$I_{\mathrm{out}}(x, y, k) = \sum_{\ell} \sum_{x', y'} I_{\mathrm{in}}(x + x', y + y', \ell) W_k(x', y', \ell) \tag{6}$$

15

In addition to convolutions and point-wise nonlinearities, CNNs can include other nonlinear operations such as pooling and normalization whose outputs depend on the activities of groups of neural units. Here, we cascade the flexible normalization model with the output map of a convolution layer. Flexible normalization of the outputs of a convolutional layer is carried out at each spatial position and output channel of I_{out}. For channel k and spatial position (x, y), the normalization pool consists of two groups: (i) a group of activations at the same (x, y) position from spatially overlapping filters from neighboring channels to k, called the center group. We use the center group that is already part of the Alexnet CNN local normalization layer (akin to cross-orientation suppression in V1 (Heeger, 1992)); (ii) a set of responses from the same filter k at spatially shifted positions, called the surround group. According to the flexible normalization, the surround normalization is gated and determined based on inference about the statistical dependencies between center and surround activations.

Figure 1 depicts this arrangement of maps produced by the filters in a convolutional layer as a 3-dimensional array. For each map k, we compute the normalized response at each (x, y) location using the flexible normalization model introduced above.

4.2 Flexible normalization for convolutional layers

One of the main differences between our model and (Coen-Cagli et al., 2009) is that our model imposes statistical independence among surround responses in the center-surround independent component of the mixture. This is achieved by making:

$$p_{X_{\text{s}}}(\mathbf{x}_{\text{s}}|\mathbf{\Lambda}_{\text{s}}) = \prod_{\ell \in \mathcal{S}} p_{(X_{\text{s}})_\ell}\left((\mathbf{x}_{\text{s}})_\ell | (\mathbf{\Lambda}_{\text{s}})_\ell\right), \tag{7}$$

where $\mathcal{S}$ denotes the set of indexes of the surround units and $\mathbf{x}_s$ the vector of filter responses of the surround units. In other words, when the center units are independent from the surround units, the group of surround units do not share the same mixer. By having independent mixers in our model, we avoid making any assumptions about symmetries in the responses of the surround units. Symmetry constraints based on the orientation of the V1 model units were originally used in (Coen-Cagli et al., 2009) for learning the parameters of the model. It is important to bear in mind that for mid-level representations there is no clear intuition or explicit knowledge about the nature of the symmetries that may arise across space. A graphical model of the flexible normalization model proposed here is depicted in Figure 2.

4.3 Inference

Another key difference between our model and (Coen-Cagli et al., 2009) is the inference. In our model, we assume there exists a common underlying Gaussian variable $\hat{G}$ that generates both types of responses (center-surround dependent and center-surround independent). The coupling is therefore, a two-stage process. First, a latent response vector $\hat{G}$ is sampled from a Gaussian distribution with zero mean, and identity covariance. This response is then linearly mapped by one of two possible transformations depending on whether the response is center-surround dependent or independent. Subsequently, the multiplicative coupling is applied to the linearly transformed vector according to the type of response (dependent or independent). The main reason for the above choice is that, if we were only resolving the multiplicative couplings, the distribution of the inferred response would still be a mixture of Gaussians which cannot be
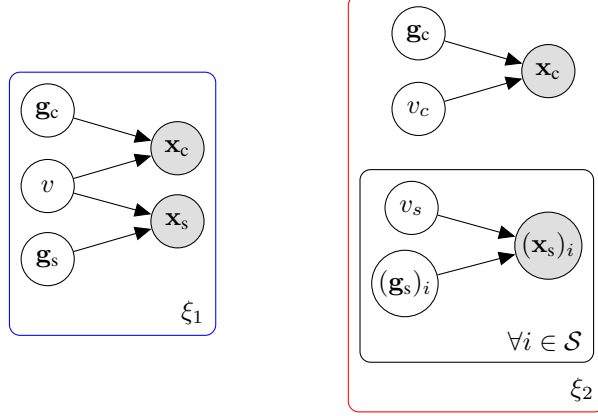
17

Figure 2: Flexible normalization model, based on a mixture of Gaussian Scale Mixtures. (left) Center-Surround dependent; (right) Center-Surround Independent. The model is similar to (Coen-Cagli et al., 2009, 2015), except that when center units are independent from surround units. We further impose independence of the surround unit activations. This removes the need to impose any symmetry constraints in learning the model parameters for higher CNN layers. $\mathbf{x}_c$ and $\mathbf{x}_s$ correspond to the filter responses of center and surround units, respectively. $\mathbf{g}_c$ and $\mathbf{g}_s$ are the Gaussian latent variables, v is the mixer in the center-surround dependent component ξ_1, v_c and v_s are the center and surround mixers for the center-surround independent component ξ_2, and $\mathcal{S}$ are the surround indexes. Note that for each surround unit filter response $(\mathbf{x}_s)_i$ there is an i.i.d draw of the mixer v_s.

decoupled by linear means.

Reversing the coupling by computing $\mathbb{E}\left[\hat{G}_i|\mathbf{x}\right]$ is also a two-stage process. First, posterior probabilities of $\mathbf{x}$ being center-surround dependent are obtained using Bayes rule, $p(\xi_1|\mathbf{x}) = \frac{p(\mathbf{x}|\xi_1)\Pi_{cs}}{p(\mathbf{x})}$. Then, conditional expectations $\mathbb{E}\left[G_i|\mathbf{x}, \mathbf{\Lambda}_{cs}\right]$ and $\mathbb{E}\left[G_i|\mathbf{x}_c, \mathbf{\Lambda}_c\right]$ are linearly mapped to a common space. Namely, we apply a linear transformation $\mathbf{Q}^\mathrm{T}$

to the center-surround independent component cs $\perp\!\!\!\perp$, such that:

$$\mathbf{Q}^{\mathrm{T}}\mathbf{\Lambda}_{\mathrm{cs}\perp\!\!\!\perp}\mathbf{Q} = \mathbf{Q}^{\mathrm{T}}\left[\begin{array}{c|c}\mathbf{\Lambda}_{\mathrm{c}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Lambda}_{\mathrm{s}}\end{array}\right]\mathbf{Q} = \mathbf{\Lambda}_{\mathrm{cs}}. \tag{8}$$

Inference in our flexible normalization model is given by:

$$\mathbb{E}\left[\hat{G}_i|\mathbf{c}\right] = p(\xi_1|\mathbf{x})\mathbb{E}\left[G|\mathbf{x}, \mathbf{\Lambda}_{\mathrm{cs}}\right] + (1 - p(\xi_1|\mathbf{x}))\left(\mathbf{Q}^{\mathrm{T}}\right)_{i,:}\mathbb{E}\left[G|\mathbf{x}, \mathbf{\Lambda}_{\mathrm{c}}, \mathbf{\Lambda}_{\mathrm{s}}\right], \tag{9}$$

where $\left(\mathbf{Q}^{\mathrm{T}}\right)_{i,:}$ denotes the ith row of $\mathbf{Q}^{\mathrm{T}}$. This inference can be followed by whitening

of the components of $\hat{G}$ yielding the desired identity covariance matrix, $\mathbf{I}$. However,

here, the relevant operation is the transformation that takes one covariance and makes

it equal to the other covariance matching the distributions of the center-surround de-

pendent and center-surround independent component after removing the multiplicative

couplings (Eq.(8)).

As we mentioned above, the covariance and therefore the whitening transformation

are presumed fixed. The flexible part of the model is gating this weighted surround nor-

malization (e.g., turning it on or off to the degree that center and surround are inferred

to be statistically dependent for a given stimulus). The inference about posterior prob-

abilities of center-surround dependence is therefore the part that is stimulus dependent.

Learning parameters of the model

In this work, our main purpose is to observe the effects of normalization in the responses

obtained at the outputs of a convolutional layer in a deep CNN. For this reason, we

apply the flexible normalization model to the responses of filters from a pre-trained

network that does not include flexible normalization [1]. The responses of a layer from

---

[1] In our work, as with the original AlexNet, filters were trained for object recognition

this pre-trained network are used to construct the set of center and surround units to be normalized. The parameters of the flexible normalization model, the prior Π_{cs} and covariances Λ_{cs}, Λ_c, and Λ_s, are then learned by Expectation Maximization (EM) fitting to the pre-trained CNN responses (Coen-Cagli et al., 2009)(see appendix for details).

5   Simulations

We integrate flexible normalization into the AlexNet architecture (Krizhevsky et al., 2012) pre-trained on the ImageNet ILSVRC2012 object recognition challenge. Since our main goal is to investigate what the effects of normalization are at the layer level rather than at the network level, we only learn the parameters of the divisive normalization model on top of the pre-trained filters. The divisive normalization is applied to the outputs of the convolutional layer. In particular, we integrate flexible normalization into the outputs of the second convolutional layer of AlexNet. In the Additional simulations in the Appendix, we also examine incorporating flexible normalization into the first and third layers of Alexnet.

We focus on the second layer for two reasons. First, it is comprised of combinations of V1-like units in the first layer, and so is likely to be more tractable in future studies comparing to neurophysiology studies in V2. Second, we found empirically that on average, as we move up from layer 1 to layer 3, the responses of units in AlexNet become less statistically dependent across space, suggesting that from an efficient coding perspective divisive normalization across space would have less influence as we move up the hierarchy.

In our model, the center neighborhoods are the same built-in neighborhoods that were induced by the local response normalization operation carried out in the original AlexNet architecture. The surround groups are obtained by taking activations from an approximately circular neighborhood with a radius of 4 strides apart[2], at every 45 degrees, which yields a total of 8 surround units. Figure 3a depicts the spatial arrangement of a center response and the positions of its surround responses.

5.1 Redundancy in activations of intermediate layers

As argued above, multiplicative couplings (high order correlations) between linear decomposition coefficients are common in natural images. As we show below, activations at intermediate layers such as the second convolutional layer of AlexNet, which we denote as "Conv2," display a significant amount of high order coupling.

Focusing on the Conv2 layer from AlexNet, we examine the structure of spatial dependencies within a unit. We show that even at spatial locations for which the filters have less than 20% overlap, the values of the activations of spatially shifted units expose high order correlations[3]. In Figure 3b, we display the energy correlations for the activations of a subset of units in the second convolutional layer (Conv2) of AlexNet. For each unit, we display the correlation of energies between the given unit and its spatial neighbors 4 strides apart in either the vertical or horizontal direction. Each one of the 3×3 tiles is the corresponding squared correlation for a particular unit. We see that not

---

[2]By stride, we mean the minimum spatial shift at which the convolution sum is evaluated.

[3]Correlations beyond first order include correlation of squares as a special case

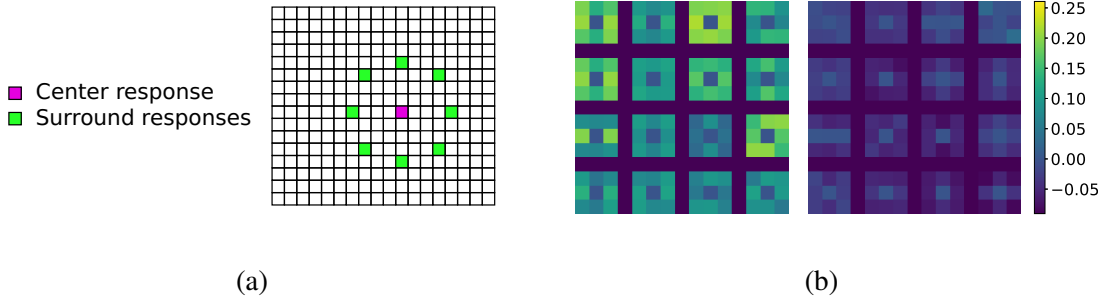(a)   (b)

Figure 3: Energy correlations in the second convolutional layer of AlexNet before and after flexible normalization. (a) Spatial distribution of center and surround activations in the normalization pool. (b) Correlation of energies between center and surround responses from a subset of 16 Conv2 units from AlexNet before (left) and after (right) flexible normalization. Each of the sixteen 3×3 tiles depicts the correlation between the center activation and each of the 8 surround units shown in (a). It is clear that normalization reduces the energy correlations.

only do these high order couplings remain for the outputs of the Conv2 layer, but also the regularities of how their values are distributed across space. For various units, it is clear that spatial shifts in particular directions have stronger couplings.

5.2   Dependency reduction by flexible normalization

To visually assess the effect of normalization on the correlation structure among units, we depict the joint conditional histograms of the unit activations after normalization and whitening. Previous studies with V1-like filters have shown that filter activations follow a bowtie-like structure that can be understood as a high order coupling (e.g. Schwartz and Simoncelli, 2001). In particular, the amplitude of one variable gives information about the variance (standard deviation) of the other variable. This dependency can be
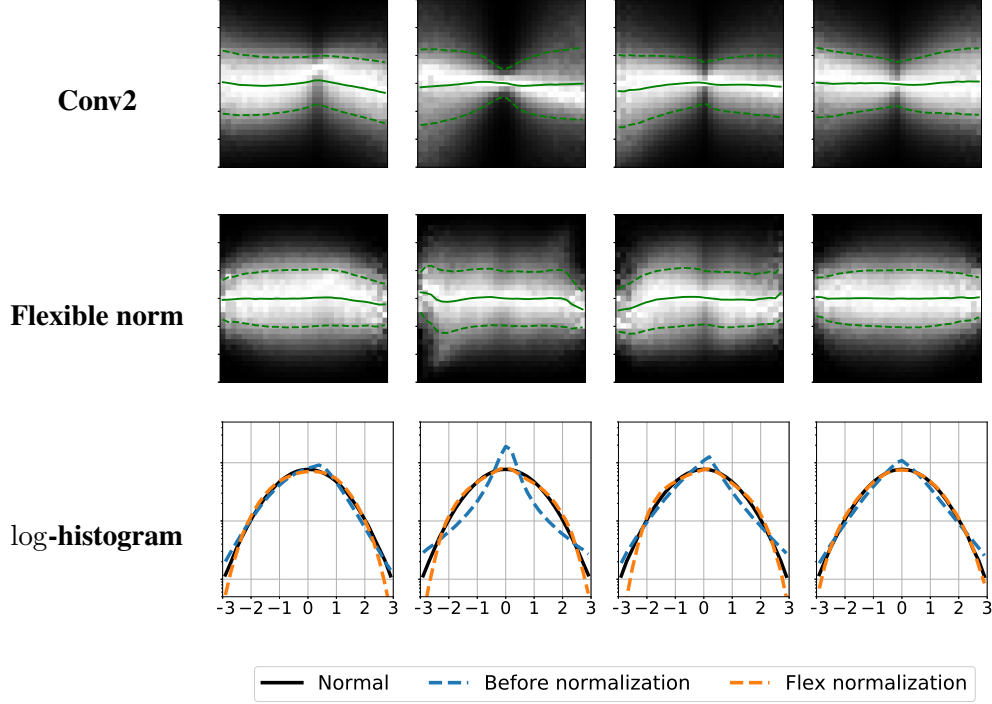
Figure 4: Marginal and joint conditional distributions of activations from example Conv2 units in AlexNet, before an after flexible normalization. The joint conditional distributions are a simple way to visually inspect dependencies. As we can see, Conv2 units at different spatial locations are nonlinearly coupled. Flexible normalization reduces these dependencies making the conditional distributions look closer to constant. In addition, the marginal log-histograms show that the normalized responses become closer to Gaussian, in agreement with the model assumptions.

reduced via divisive normalization from neighboring filter activations. Figure 4 shows the conditional histograms ($p(x_\text{s}|x_\text{c})$) for the same pair of center-surround unit activations before and after applying flexible divisive normalization. Along with the normalized conditional histograms, we show marginal log-histograms which give an idea of how normalization changes the marginal distributions from highly kurtotic to more Gaussian-like.

We further quantify the results for the flexible normalization model, and compare to a simpler baseline surround normalization model using a single GSM. At a population level, both normalization models, flexible and single GSM, consistently reduce mutual information between the center unit and spatial surround activations (2032 out of all 8×256 center-surround pairs). But the flexible normalization reduces mutual information beyond the level achieved by the control model (Figure 5). The distribution of the difference of the estimates of mutual information between single GSM minus flexible normalization is skewed to the right (1.8 skewness).

Also, computing the entropy of the activations before and after normalization shows consistent increase which is more pronounced in the flexible normalization model (Figure 6). Since the activations before and after normalization have been scaled to have unit variance, larger entropies correspond to random variables whose distributions are more similar to the Gaussian. We have also examined the expected likelihood of the test data. For more than 98% of the units, flexible normalization has higher likelihood compared to the single GSM model. Overall, the population quantities confirm that flexible normalization is better than the single GSM at capturing the Gaussian statistics and reducing the statistical dependencies.

5.3   Predicting homogeneity of stimuli based on mid-level features

The main idea of flexible normalization is that contextual modulation of neural responses should be present only when responses are deemed as statistically dependent. In the case of V1, co-linearity of stimuli in the preferred direction of the center unit would cause the flexible normalization model to invoke suppression (see also appendix
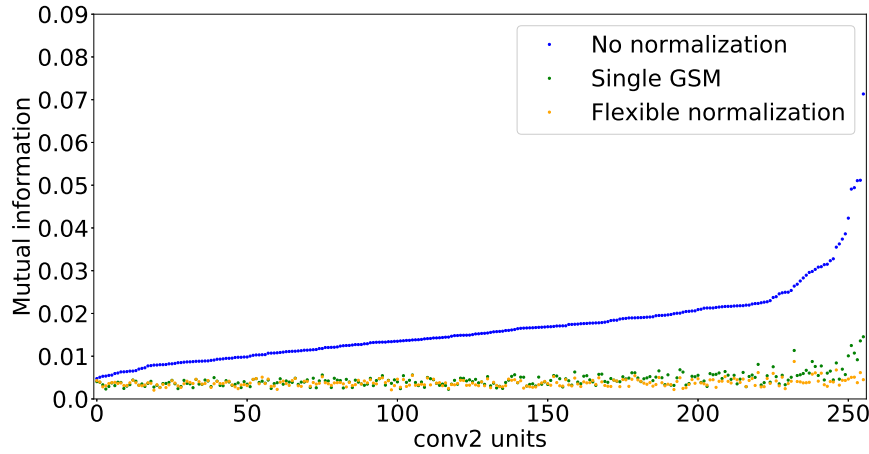
Figure 5: Population mutual information summary statistics for flexible normalization versus the control surround normalization model. Average mutual information between center and each of the 8 surround locations for all 256 channels in the second convolutional layer of AlexNet. The units are ordered with respect to the mutual information before normalization. A control normalization model where normalization is applied uniformly across the image is included for reference. Flexible normalization is able to reduce mutual information in cases where the fixed normalization model (control) cannot.

for Conv1 units). In other words, the model would infer high center-surround posterior probability from the stimuli.

For the case of mid-level features, we wanted to observe what structure in the stimuli would lead to center-surround dependence. Note that for intermediate level features, the notion of orientation is not as clear as in V1, where models may contain orientations in their filter parameterizations.

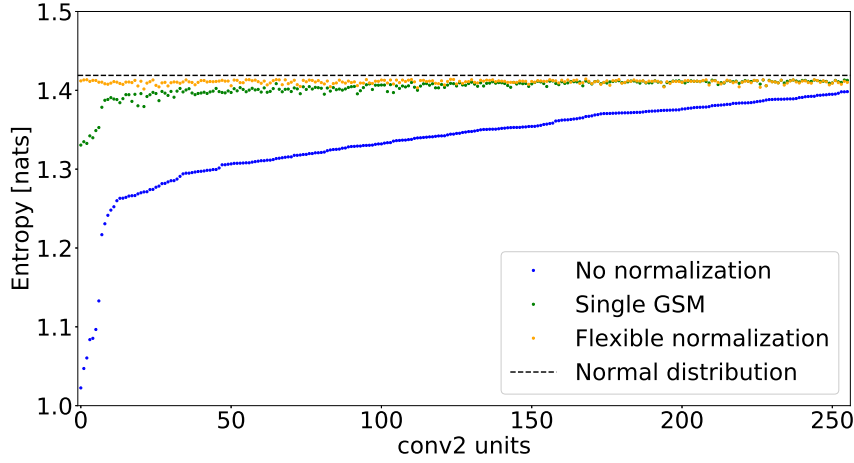Figure 7 shows some examples of image patches that cover the center-surround

Figure 6: Population entropy summary statistics for flexible normalization versus the control surround normalization model. Marginal entropies from standardized responses (zero mean and unit variance) before and after normalization for each unit in the second convolutional layer of AlexNet. Similar to mutual information, units are oredered based on their entropy values before normalization. Black dashed line indicates the theoretical upper bound which corresponds to the entropy of a unit variance Gaussian distributed random variable.

neighborhoods, for which the model finds a high posterior probability of Center-Surround dependence. Along with these images, a visualization of the receptive field of the second convolutional layer units is presented. In addition, The units depicted in Figure 7 are ordered based on the prior probability of center-surround dependence that is learned by our flexible normalization model. The top row of the figure corresponds to the lowest value of this prior probability among the units displayed in the figure.

As can be seen, for some of the Conv2 representations, the idea of co-linearity is present in the form of tiling of these mid-level features. The center-surround covariance

26

Figure 7: Tiling in Conv2 center-surround dependent units. Example units are ordered from lower learned prior of dependence (top unit on the left side table; .5193) to higher learned prior of dependence (bottom unit on the right side table; .9364). (First column) center-surround dependent covariances. Each black corresponds to the spatial location of the receptive fields. The line thickness between points depicts the strength of covariance between spatially shifted receptive fields. The size of the black circles depicts the variance relative to the center circle which has the same size in all units. (Second column) Conv2 units visualization with a method adapted from Zeiler and Fergus (2014). (Remaining columns) Image regions with high probability of being Center-Surround dependent and high activation values prior normalization according to our model. Note how regions can be seen as tiling the space with translations of the Conv2 unit receptive fields in directions with strong covariance.27
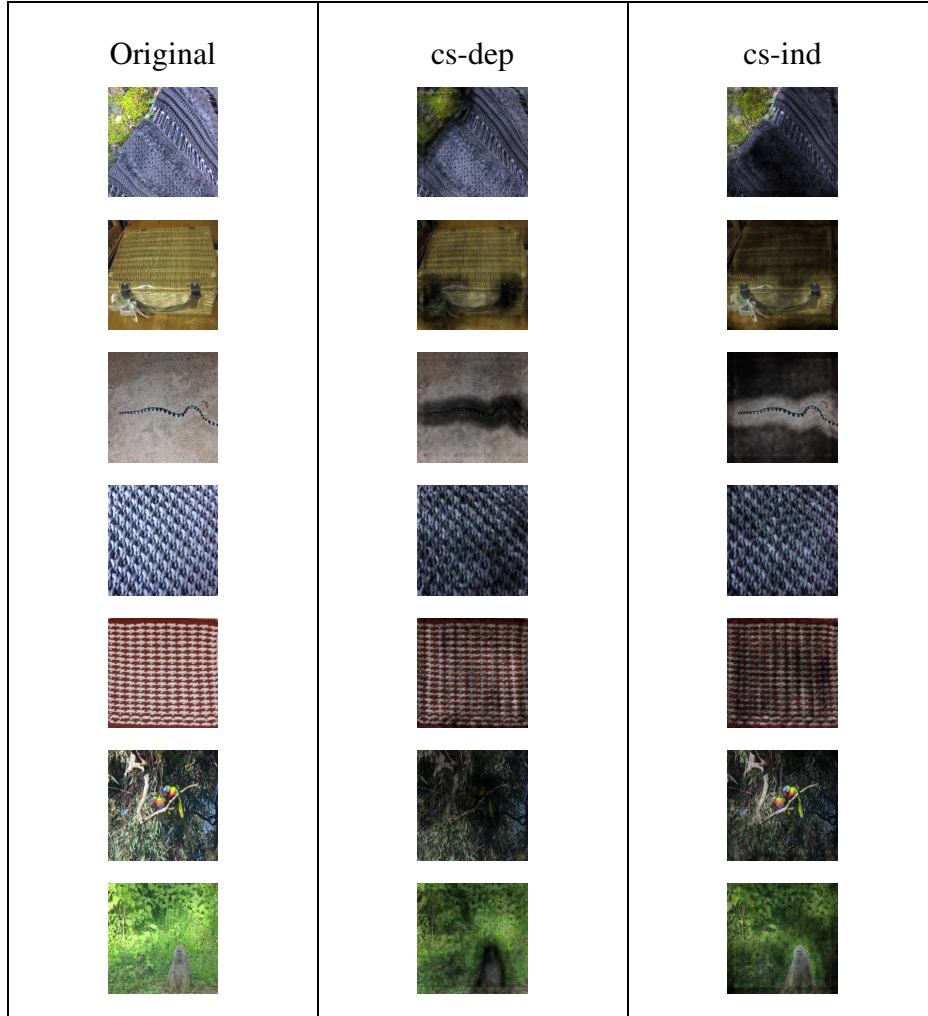
Figure 8: Inferred posterior distributions on ImageNet data. (left) Original images input to the CNN. For each image at each spatial location, we compute the geometric mean of posterior probabilities among all channels of the layer (Conv2). (middle) Areas that the model deemed as dependent while obscuring other areas. (right) Complementary display where high center-surround dependent areas are darker, instead.

also captures this property. By looking at the receptive fields of the Conv2 units (second column from the left in Figure 7) we can see that spatial arrangements of repetitions of these receptive fields seem more natural in certain configurations. For instance, the eighth row shows horizontal structure in the covariance matrix. The receptive field has

a horizontal structure, but also a texture boundary that includes vertical structure. The tiling according to the covariance matrix is along this main horizontal axis. Similarly, in the second row of the left table, vertical arrangements of translated versions of the receptive field give a continued pattern as appreciated in the corresponding patches (right column of the table). For other example units (e.g., the last 3 units which appear to capture texture-like structure), the covariance structure is more uniform across space.

We also looked at the spatial distribution of the posterior probabilities for entire images from ImageNet. For each channel in the convolutional layer, we obtained the posterior probabilities and computed the geometric mean across channels. Figure 8 shows the relation between the image content and the posterior probability by shading areas with high (middle column) and low (right column) posterior probability of being center-surround dependent. As mentioned earlier, a high posterior probability of center-surround dependent activations can be an indicator of the homogeneity of the region under consideration. As we can see, the posterior probabilities from Conv2 units capture this homogeneity at a more structured level compared to previous work on V1-like filters, where homogeneous regions correspond to elongated lines in the preferred orientations of the filters.

In particular, repeating this procedure for the first convolutional layer (Figure 15 in the Appendix), reveals that the first layer is not as effective as the second in capturing (and therefore suppressing) the background homogeneous texture structure prevalent in images, so as to highlight for instance the snake (third row) or birds (sixth row) in the scene.

29

5.4 Predicting surround modulation based on image homogeneity for intermediate level features

Previous work has considered predictions and experimental testing of the flexible normalization model for area V1 (Coen-Cagli et al., 2015). Here we consider predictions for the intermediate level features of the CNN. We focus on Conv2 units following the flexible normalization.

To characterize the influence of surround modulation, we presented each model unit with small natural images confined to the classical receptive field and a corresponding large image that extended beyond the classical receptive field. We obtained the modulation ratios by comparing the average responses of each unit, when presented with stimuli at apertures of 31 and 63 pixels, corresponding to the areas covered by a single unit and the center and surround neighborhood. Example images for both apertures are displayed in Figure 9a.

In the scatter plot of Figure 9b, each point corresponds to a single model unit. The scatter plot compares two types of stimuli, based on the model predictions. One set of images correspond to images deemed as homogeneous by the model posterior inference of center-surround dependence (top row of Figure 9a). The second set of images are those deemed heterogeneous by exposing low posterior probability in the component modeling center-surround dependencies (bottom row of Figure 9a). The vertical axis corresponds to the modulation ratios for the set of heterogeneous images and the horizontal axis for the homogeneous images. The model predicts more surround modulation for images that are deemed homogeneous than for images deemed heterogeneous.This mostly results in suppression by the surround, since most of the points lie above the

diagonal line. Note that some units do lie below the diagonal because in some cases the model would actually increase the energy of the response to match the expected variance of the Gaussian latent variable.

For this simulation, in addition to the posterior probabilities, images where selected based on the magnitude of the responses they elicited in the Conv2 units. Only images with high response levels and largest or smallest posterior probabilities were used. To compute the average response for each unit, a set of 40 homogeneous images and 80 heterogeneous images (40 where center energies are larger than surround energies and 40 for the opposite case) were selected for each unit, as explained above.

In a similar flavor to (Coen-Cagli et al., 2015), which focused on V1, an experiment to test the model predictions of surround modulation in V2 could use the same set of images (collected for all units in the CNN) to record responses from this area. Responses of each unit can be matched to CNN units by their level of correlation for the stimuli confined to the classical receptive field. Average responses for center and center-surround stimuli can then be compared between the CNN and the recorded units, by computing the modulation ratios of the experimental data to obtain a plot analogous to the one displayed in Figure 9b. Although we focus here on convolutional layer 2 of the CNN, one could expand the experimental comparisons to include other layers of the CNN, for instance layers 1 and 3.
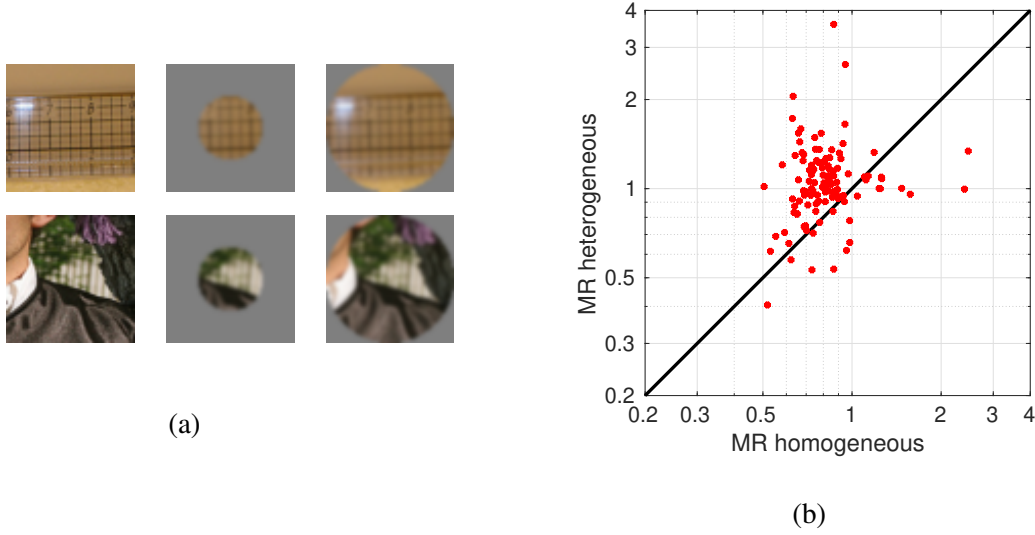
(a)

(b)

Figure 9: (9a) Example stimuli employed to compute modulation ratios. Top left image corresponds to an image patch that was deemed as homogeneous by the flexible normalization model and the bottom left to a heterogeneous patch. The center column contains the stimuli cropped at 31 pixel diameter for center aperture and at 63 pixel diameter for center-surround aperture. (9b) Modulation ratios obtained by comparing the average of the normalized responses of Conv2 units from AlexNet when presented with stimuli at 31 and 63 pixel apertures. The two apertures correspond to the areas covered by a single unit and the center-surround neighborhood, respectively. The scatter plot compares two types of stimuli. One set of images correspond to images deemed as homogeneous by the model posterior inference of center-surround dependence (vertical axis). The second set of images are those deemed heterogeneous by exposing low posterior probability that center and surround are dependent. In addition to the posterior probabilities, images where selected based on the magnitude of the responses they elicited in the Conv2 units. Only images with high response levels and largest or smallest posterior probabilities are used.

6 Discussion

V1 normalization has been studied extensively, including flexible models of divisive normalization by the spatial context. In modeling normalization of V1 units using a Gaussian Scale Mixture framework, previous work has shown that the learned covariance matrix (influencing the contextual normalization) has collinear structure (Coen-Cagli et al., 2012) (see also appendix). This reflects the oriented front end of V1 units and the collinear structure prevalent in images. Our work seeks to address richer receptive field structure beyond the level of V1-like units. We therefore build on a modified version of the model of (Coen-Cagli et al., 2012), and use intermediate level representations that are learned by convolutional networks trained for object recognition.

The second convolutional layer of AlexNet combines responses from different units from the first layer into more elaborate features. However, the statistical dependencies between such units have not been characterized, and consequently there has been little emphasis on deriving divisive normalization models from scene statistics as has been done extensively for V1-like model units. We therefore set out to study the statistics of the second layer of the CNN, and to examine the covariance structure that is learned. The covariance structure can be understood as a template of what spatial arrangements lead to high center-surround dependence for a given model unit.

First, we found that units in the Alexnet CNN had sparse marginal statistics and joint conditional statistical dependencies across space, similar to what has been observed for ICA and Gabor filters. Although this decreased in higher layers on average, the statistics in the second layer of AlexNet were still prominent. Further, we found in our simulations, that the learned covariances for Conv2 units included both collinear spatial

arrangements capturing structure such as tiling across space of texture boundaries; and covariances that were more uniform across all spatial directions capturing structure such as textures. Textures have also received attention in other studies in mid-level visual areas (Freeman et al., 2013; Rowekamp and Sharpee, 2017; Laskar et al., 2018; Ziemba et al., 2018) and in generative models such as (Portilla and Simoncelli, 2000; Gatys et al., 2015).

Furthermore, we found that from a technical perspective, adding independence among surround units in the center-surround independent component of the mixture Eq.(5) was crucial for learning the parameters of the model without imposing symmetry constraints. This is particularly necessary in the context of deep CNNs, and even more so at middle levels. Unlike Gabor filters or steerable pyramids, units of deep CNNs are not parametrized explicitly with respect to orientations, for instance. The same issue arises in cortical neurons, whereby higher neural areas beyond V1 combine orientations in a nonlinear manner, leading to rich structure that is not easily parameterized.

The model does a good job at reducing multiplicative dependencies, but it is not perfect. Currently, filters of the CNN are not learned jointly with flexible normalization. Another limitation of the model is that flexible normalization is incorporated in a single layer of the CNN. An important future direction which would require further technical development beyond the scope of this paper, is learning flexible normalization in multiple layers of the CNN (e.g., layers 1 and 2) simultaneously. Our approach can also be used with other classes of hierarchical architectures.

There are two important future applications of our approach. The first pertains

to understanding the benefits of the proposed normalization computation downstream. In Coen-Cagli and Schwartz (2013), the impact of V1 flexible normalization was examined downstream and shown to perform better on figure ground segregation than a model that always divided by the surround. What might be a benefit of flexible normalization for higher neural areas beyond V1? Figure 8 is suggestive in this regard, since for the second convolutional layer, background textures in example scenes are inferred to be more statistically dependent, highlighting objects in the scene (compare to Figure 15 in the Appendix for the first convolutional layer).

Flexible normalization has the potential to cope with different data sets and different amounts of clutter, by controlling the amount of selective suppression helpful for the given stimulus and task. This may be in line with the recent observation that V4 units respond more strongly to shapes than to textures, which (Kim et al., 2019) refer to as *de-texturization*. It may also have the potential to help in the texture bias that has been observed for texture-shape cue conflicts in CNNs relative to human subjects (Geirhos et al., 2019). An important future direction is to explore these questions for intermediate layers of the deep neural network.

A second important application is making predictions and testing against cortical neural data. Previous work on flexible normalization has tested predictions for natural images in area V1 (Coen-Cagli et al., 2015). Our approach can be used to make predictions about when normalization might be relevant in higher visual cortical areas to reduce redundancy. In particular, we expect our model can make useful predictions for testing normalization of neural responses to large stimuli that extend beyond the classical receptive field in intermediate cortical areas such as V2 (as explained in Section 5.4

and Figure 9).

Our modeling of surround dependencies in second layer units of Alexnet mostly revealed homogeneous patterns of dependencies for textures, and collinear patterns for texture boundaries. Suppression for extended homogeneous textures has been observed in V2 (Ziemba et al., 2018) and in V4 (Kim et al., 2019). However, the result of (Ziemba et al., 2018) regarding more suppression for spectrally matched noise than for textures may require an additional facilitatory mechanism as proposed in their paper. In addition, neurophysiology studies of V2 surround suppression have found qualitatively similar surround suppression for grating stimuli in V2 as for V1 (Shushruth et al., 2009). A direction for future work is to test the model against cortical data in area V2.

A more complete modeling account should include two stages of surround normalization (e.g., corresponding to V1 and V2) and consider what is inherited from area V1 and what is unique to area V2. Although we have emphasized model units based on the second layer of the CNN, our approach can more generally applied to other hierarchical models, such as those learned with unsupervised methods (e.g., Hosoya and Hyvrinen, 2015).

Acknowledgments

Appendix

6.1 Maximum likelihood estimation of covariance

Recall the pdf of our Gaussian scale mixture with Rayleigh mixer:

$$p_X(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} |\mathbf{\Lambda}|^{1/2} h^m} a^{1-m/2} K_{m/2-1}(a),\tag{10}$$

where $K_\lambda(\cdot)$ is the modified Bessel function of the second kind, and

$$a^2 = \frac{\mathbf{x}^{\mathrm{T}} \mathbf{\Lambda}^{-1} \mathbf{x}}{h^2}.\tag{11}$$

Eq.(10) can be employed to compute the likelihood function of the covariance matrix $\mathbf{\Lambda}$. Furthermore, notice that the scale parameter h can be simply dismissed by making part of the covariance[4]. If we take $\hat{\mathbf{\Lambda}}^2 = \mathbf{\Lambda} h^2$, Eq.(10) becomes:

$$p_X(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} |\hat{\mathbf{\Lambda}}|^{1/2}} \hat{a}^{1-m/2} K_{m/2-1}(\hat{a}),\tag{12}$$

where $\hat{a} = \mathbf{x}^{\mathrm{T}} \hat{\mathbf{\Lambda}}^{-1} \mathbf{x}$. From this point on, to simplify notation, we will refer to $\hat{\mathbf{\Lambda}}$ as $\mathbf{\Lambda}$. and $\hat{a}$ as simply a.

For an exemplar $\mathbf{x}_i$, the partial derivative of $\log$-likelihood function with respect to $\mathbf{\Lambda}^{-1}$ is given by:

$$\begin{aligned}\frac{\partial \log L(\mathbf{\Lambda}|\mathbf{x}_i)}{\partial \mathbf{\Lambda}^{-1}} = \frac{\partial \log (p_X(\mathbf{x}_i))}{\partial \mathbf{\Lambda}^{-1}} &= \frac{1}{p_X(\mathbf{x}_i)} \frac{\partial p_X(\mathbf{x}_i)}{\partial \mathbf{\Lambda}^{-1}} \\ &= \frac{\mathbf{\Lambda}}{2} - \frac{1}{2a} \frac{K_{m/2}(a)}{K_{m/2-1}(a)} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}.\end{aligned}\tag{13}$$

Based on Eq.(13), we propose the following iterative update rule for $\mathbf{\Lambda}$

$$\mathbf{\Lambda}_{\text{new}} \leftarrow \frac{1}{N} \sum_{i=1}^{N} g_m(a_i) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}},\tag{14}$$

---

[4]Adding scale back is straightforward.

where

$$g_m(a_i) = \frac{1}{a_i} \frac{K_{m/2}(a_i)}{K_{m/2-1}(a_i)}, \tag{15}$$

and $a_i = \sqrt{\mathbf{x}_i{}^{\mathrm{T}}\mathbf{\Lambda}_{\mathrm{old}}{}^{-1}\mathbf{x}_i}$.

Details about inference

Here, we work out the inference procedure for the full covariance GSM. Let $\backslash i$ denote the set of all indexes minus index i, and decompose the precision matrix $\mathbf{\Lambda}^{-1}$ into,

$$\mathbf{\Lambda}^{-1} = \begin{bmatrix} (\mathbf{\Lambda}^{-1})_{\backslash i, \backslash i} & (\mathbf{\Lambda}^{-1})_{\backslash i, i} \\[2mm] (\mathbf{\Lambda}^{-1})_{i, \backslash i} & (\mathbf{\Lambda}^{-1})_{i, i} \end{bmatrix}. \tag{16}$$

It can be shown that

$$\mathbb{E}\left[G_i|\mathbf{x}\right] = \frac{c_i}{\sqrt{a_{\backslash i}}} \left(\frac{a}{a_{\backslash i}}\right)^{\frac{m}{2}-1} \frac{|\mathbf{\Lambda}|^{\frac{1}{2}}}{\sigma_i|((\mathbf{\Lambda}^{-1})_{\backslash i, \backslash i})^{-1}|^{\frac{1}{2}}} \frac{K_{\frac{m-1}{2}}(a_{\backslash i})}{K_{\frac{m}{2}-1}(a)}, \tag{17}$$

where $\sigma_i{}^2 = (\mathbf{\Lambda})_{i,i}$, and

$$a_{\backslash i}{}^2 = \mathbf{x}_{\backslash i}{}^{\mathrm{T}}(\mathbf{\Lambda}^{-1})_{\backslash i, \backslash i}\mathbf{x}_{\backslash i} + 2x_i\mathbf{x}_{\backslash i}{}^{\mathrm{T}}(\mathbf{\Lambda}^{-1})_{\backslash i, i} + \tag{18}$$
$$+ x_i{}^2 \left[(\mathbf{\Lambda}^{-1})_{i, \backslash i}((\mathbf{\Lambda}^{-1})_{\backslash i, \backslash i})^{-1}(\mathbf{\Lambda}^{-1})_{\backslash i, i} + \sigma_i{}^{-2}\right].$$

Noticing that

$$(\mathbf{\Lambda}^{-1})_{i,i} = \sigma{i}^{-2} + (\mathbf{\Lambda}^{-1})_{i, \backslash i}((\mathbf{\Lambda}^{-1})_{\backslash i, \backslash i})^{-1}(\mathbf{\Lambda}^{-1})_{\backslash i, i}, \tag{19}$$

yields $a_{\backslash i} = a$. Furthermore, since $\sigma_i^2 = |\mathbf{\Lambda}||(\mathbf{\Lambda}^{-1})_{\backslash i, \backslash i}|$,

$$\mathbb{E}\left[G_i|\mathbf{x}\right] = \frac{x_i}{\sqrt{a}} \frac{K_{\frac{m-1}{2}}(a)}{K_{\frac{m}{2}-1}(a)}. \tag{20}$$

6.2 Mixture of Gaussian scale mixtures

The mixture model has the following general form:

$$p_X(\mathbf{x}) = \sum_{\alpha \in \mathcal{A}} \Pi_\alpha p_X(\mathbf{x}|\mathbf{\Lambda}_\alpha) \tag{21}$$

Parameter estimation for the above model Eq.(21) can be solved using expectation maximization (EM) algorithm. In particular, we use the conditional EM algorithm to update the parameters of each of the mixture components. For each partial **E-step**, we compute the posterior distributions over the assignment variable:

$$q(\alpha, \mathbf{x}_i) = \frac{\Pi_\alpha p_X(\mathbf{x}_i | \mathbf{\Lambda}_\alpha)}{\sum\limits_{\alpha' \in \mathcal{A}} \Pi_{\alpha'} p_X(\mathbf{x}_i | \mathbf{\Lambda}_{\alpha'})}, \text{ for all } \alpha \in \mathcal{A}. \tag{22}$$

In the partial **M-step**, we update all the mixture probabilities using Eq.(22),

$$\Pi_{\alpha'} \leftarrow \frac{1}{N} \sum_{i=1}^{N} q(\alpha', \mathbf{x}_i), \text{ for all } \alpha' \in \mathcal{A}, \tag{23}$$

and the corresponding covariance $\mathbf{\Lambda}_\alpha$ using a modified version of the fixed-point Eq.(14), as follows:

$$\mathbf{\Lambda}_\alpha \leftarrow \frac{\sum\limits_{i=1}^{N} q(\alpha, \mathbf{x}_i) g_m(\mathbf{x}_i | \mathbf{\Lambda}_\alpha) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}}{\sum\limits_{j=1}^{N} q(\alpha, \mathbf{x}_j)}. \tag{24}$$

where $g_m(\mathbf{x}_i | \mathbf{\Lambda}_\alpha) = g_m(\sqrt{\mathbf{x}_i^{\mathrm{T}} \mathbf{\Lambda}_\alpha^{-1} \mathbf{x}_i})$ from Eq.(15) We use a single fixed-point iteration per partial CEM iteration. The proposed fixed point update increases the likelihood at each iteration.

Two-GSM mixture model for flexible normalization

Gaussian scale mixture models have been used to explain non-linear dependencies among linear decompositions of natural stimuli such as images. In the simplest case, it is assumed that such dependencies carry over the entire stimuli. For example, in vision, commonly used approaches of local contrast normalization apply the same normalization scheme across the entire image. Spatial pools for normalization have been applied to explain responses to redundant stimuli. While this model is able to account

for suppressions of unit responses where spatial context is redundant, it can also lead to suppression in cases where context may not be redundant. A flexible normalization that suppresses responses only when the spatial context is deemed as redundant can be constructed as a mixture of GSMs. A simple version considers a component with full center-surround dependencies, a second component representing the center-surround independence results from the product of a center only and surround only GSM distributions:

$$p_X(\mathbf{x}) = \Pi_{\mathrm{cs}} p_X(\mathbf{x}|\mathbf{\Lambda}_{\mathrm{cs}}) + (1 - \Pi_{\mathrm{cs}}) p_{X_{\mathrm{c}}}(\mathbf{x}_{\mathrm{c}}|\mathbf{\Lambda}_{\mathrm{c}}) p_{X_{\mathrm{s}}}(\mathbf{x}_{\mathrm{s}}|\mathbf{\Lambda}_{\mathrm{s}}), \tag{25}$$

where $\mathbf{x}_{\mathrm{c}}$ and $\mathbf{x}_{\mathrm{s}}$ denote the sub-vectors of $\mathbf{x}$ containing the center and surround variables, respectively. The variants of the EM steps prsented in Eq.(22), Eq.(23), and Eq.(24) are discussed below. For each partial **E-step**

$$q(\mathrm{cs}, \mathbf{x}_i) = \frac{\Pi_{\mathrm{cs}} p_X(\mathbf{x}_i|\mathbf{\Lambda}_{\mathrm{cs}})}{Q(\mathbf{x}_i)}, \tag{26}$$

$$q(\mathrm{cs} \perp\!\!\!\perp, \mathbf{x}_i) = \frac{(1 - \Pi_{\mathrm{cs}}) p_{X_{\mathrm{x}}}(\mathbf{x}_{i,\mathrm{c}}|\mathbf{\Lambda}_{\mathrm{c}}) p_{X_{\mathrm{s}}}(\mathbf{x}_{i,\mathrm{s}}|\mathbf{\Lambda}_{\mathrm{s}})}{Q(\mathbf{x}_i)} = 1 - q(\mathrm{cs}, \mathbf{x}_i) \tag{27}$$

$$Q(\mathbf{x}_i) = \Pi_{\mathrm{cs}} p_X(\mathbf{x}_i|\mathbf{\Lambda}_{\mathrm{cs}}) + (1 - \Pi_{\mathrm{cs}}) p_{X_{\mathrm{c}}}(\mathbf{x}_{i,\mathrm{c}}|\mathbf{\Lambda}_{\mathrm{c}}) p_{X_{\mathrm{s}}}(\mathbf{x}_{i,\mathrm{s}}|\mathbf{\Lambda}_{\mathrm{s}}) \tag{28}$$

$$\tag{29}$$

Each partial **M-step** updates the center-surround dependent probability using Eq.(22),

$$\Pi_{\mathrm{cs}} \leftarrow \frac{1}{N} \sum_{i=1}^{N} q(\mathrm{cs}, \mathbf{x}_i). \tag{30}$$

Three partial **M-step** updates are required:

1. A center-surround dependent covariance $\mathbf{\Lambda}_{\mathrm{cs}}$ update,

$$\mathbf{\Lambda}_{\mathrm{c}} \leftarrow \frac{\sum_{i=1}^{N} q(\mathrm{cs}, \mathbf{x}_i) g_m(\mathbf{x}_i|\mathbf{\Lambda}_{\mathrm{cs}}) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}}{\sum_{j=1}^{N} q(\mathrm{cs}, \mathbf{x}_j)}, \tag{31}$$

2. a center-only covariance

$$\Lambda_c \leftarrow \frac{\sum\limits_{i=1}^{N}(1 - q(cs, \mathbf{x}_i))g_{m_c}(\mathbf{x}_{i,c}|\Lambda_c)\mathbf{x}_{i,c}\mathbf{x}_{i,c}^{\mathrm{T}}}{\sum\limits_{j=1}^{N}1 - q(cs, \mathbf{x}_j)}. \tag{32}$$

3. and a surround-only covariance

$$\Lambda_s \leftarrow \frac{\sum\limits_{i=1}^{N}(1 - q(cs, \mathbf{x}_i))g_{m_s}(\mathbf{x}_{i,s}|\Lambda_s)\mathbf{x}_{i,s}\mathbf{x}_{i,s}^{\mathrm{T}}}{\sum\limits_{j=1}^{N}1 - q(cs, \mathbf{x}_j)}. \tag{33}$$

Re-parameterization:

To simplify computations and directly enforce the nonnegative definiteness in our covariance estimation, we re-parametrize the likelihood function. Let us write,

$$\Lambda = \mathbf{B}^{\mathrm{T}}\mathbf{B}, \text{ and } \Lambda^{-1} = \mathbf{A}^{\mathrm{T}}\mathbf{A}. \tag{34}$$

Then, $\mathbf{B} = \mathbf{A}^{-\mathrm{T}}$ and

$$\frac{\partial \log L(\Lambda|\mathbf{x}_i)}{\partial \mathbf{A}} = \mathbf{B} - \frac{1}{a}\frac{K_{m/2}(a)}{K_{m/2-1}(a)}\mathbf{A}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}, \tag{35}$$

which yields the following fixed point update:

$$\mathbf{B}_{\text{new}} \leftarrow \frac{1}{N}\sum_{i=1}^{N}g_m(a_i)\mathbf{A}_{\text{old}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \tag{36}$$

A center-surround independent model with independent surround units

In this model the center-surround independent component has the extra property that requires surround units to be independent from each other. One consequence of this requirement is that the surround covariance Λ_s becomes a diagonal matrix. Note that diagonal covariance is a necessary but not sufficient condition for independence in this

case. The main feature for independence is that each one of the surround units has its own mixer (scaling rather than mixing) variable instead of a shared mixer as it is the case in the model previously discussed. If the mixer (scaling) variables are Rayleigh distributed, each surround unit ℓ in the center-surround independent component has a Laplace distribution:

$$f_\ell(x) = \frac{1}{2\sqrt{(\mathbf{\Lambda}_c)_{\ell,\ell}}} \exp\left(-\frac{|x|}{\sqrt{(\mathbf{\Lambda}_c)_{\ell,\ell}}}\right), \tag{37}$$

where $(\mathbf{\Lambda}_c)_{\ell,\ell}$ denotes the diagonal element of the surround covariance matrix $\mathbf{\Lambda}_c$. Note that this matrix has zero off-diagonal elements by definition. In this model,

$$p_{X_s}(\mathbf{x}_{i,s}|\mathbf{\Lambda}_s) = \prod_{\ell \in \mathcal{S}} f_\ell\left((\mathbf{x}_{i,s})_\ell\right) \tag{38}$$

In this modified version Eq.(33) becomes:

$$(\mathbf{\Lambda}_s)_{\ell,\ell} \leftarrow \frac{\sum_{i=1}^{N}(1 - q(\mathrm{cs}, \mathbf{x}_i))\left|(\mathbf{x}_{i,s})_\ell\right|}{\sum_{j=1}^{N} 1 - q(\mathrm{cs}, \mathbf{x}_j)}\sqrt{(\mathbf{\Lambda}_s)_{\ell,\ell}}. \tag{39}$$

The rest of the **EM** algorithm proceeds in the same way as described in Eq.(31) and Eq.(32).

Matching covariances for inference

Here, we describe how we obtain the transformation $\mathbf{Q}$ for equation (7) in the paper. Assuming that both matrices are full rank, we can write $\mathbf{\Lambda}_{cs} = \mathbf{A}^\mathrm{T}\mathbf{A}$ and $\mathbf{\Lambda}_{cs\perp} = \mathbf{B}^\mathrm{T}\mathbf{B}$. Furthermore, there exists a transformation $\mathbf{Q}$ such that:

$$\mathbf{Q}^\mathrm{T}\mathbf{\Lambda}_{cs\perp}\mathbf{Q} = \mathbf{\Lambda}_{cs}, \tag{40}$$

which is simply given by $\mathbf{Q} = \mathbf{B}^{-1}\mathbf{A}$.

6.3 Judging the effectiveness of normalization

As noted above, the Gaussian scale mixture introduces a multiplicative coupling between variables that cannot be removed by linear means. This coupling is captured by a simple dependency measure based on the energy of the variables. For zero mean, unit variance, and mutually independent G_i and G_j, define $X_i = C_i V$ and $X_j = G_j V$ where the mixer V is also independent of G_i and G_j. The covariance of X_i^2 and X_j^2 is given by,

$$
\begin{aligned}
\mathbb{E}\left[(X_i^2 - \mathbb{E}\left[X_i^2\right])(X_j^2 - \mathbb{E}\left[X_j^2\right])\right] &= \mathbb{E}\left[X_i^2 X_j^2\right] - \mathbb{E}\left[X_i^2\right] \mathbb{E}\left[X_j^2\right] \\
&= \mathbb{E}\left[G_i^2 V^2 G_j^2 V^2\right] - \mathbb{E}\left[G_i^2 V^2\right] \mathbb{E}\left[G_j^2 V^2\right] \\
&= \mathbb{E}\left[V^4\right] - \mathbb{E}\left[V^2\right]^2.
\end{aligned}
\tag{41}
$$

The strength of the coupling depends on the spread of V. A perfect inversion of the coupling, which would require explicit values of V, would make the expression Eq.(41) zero. Here, we use a related measure, the correlation between squared responses (Coates and Ng, 2011). This measure has been used to select groups of receptive fields that should be processed together in a subsequent layer of a deep network. The correlation between squared responses is computed in a two-step process. First, variables X are decorrelated by whitening using ZCA. For the pair whitened variables $(\tilde{X}_i, \tilde{X}_j)$, correlation of squared responses is given by:

$$
S(\tilde{X}_i, \tilde{X}_j) = \frac{\mathbb{E}\left[(\tilde{X}_i^2 \tilde{X}_j^2 - 1)\right]}{\sqrt{\mathbb{E}\left[(\tilde{X}_i^4 - 1)\right] \mathbb{E}\left[(\tilde{X}_j^4 - 1)\right]}}
\tag{42}
$$

6.4 Additional simulations

Flexible normalization on the first convolutional layer of AlexNet

We also trained our flexible normalization model on the responses of the first convolutional layer of AlexNet. The filters in this layer resemble the patterns that have been identified to elicit vigorous responses in V1 neurons. This is not the first time the flexible normalization model has been applied to filters modeling V1. For previous work, we refer the readers to Coen-Cagli et al. (2009, 2012, 2015). Nevertheless, to the best of our knowledge this is the first time the flexible normalization model has been applied to filters learned from data in a supervised learning task. In previous work, the orientation of the filters, which was known, was employed to restrict the model fitting by adding symmetry constraints to the covariance matrices of the model. As we explained in the main text, our modified model does not employ these symmetry constraints, but forces the surround variables to be fully independent, which translates into having a separate mixer variable for each one of them.

Covariance structure of the surround components of the first convolutional layer of AlexNet

Similarly to Figure 4a in the paper, we visualize the covariance structure of the surround covariance (Figure 10). As we can see, low frequency filters expose stronger correlation in their responses than the high frequency filters. Also the orientation of the filter is reflected in the covariance structure of the model, similar to the results obtained in (Coen-Cagli et al., 2009) for wavelet filters. The high frequency filters showed lower levels of correlation and weaker oriented patterns (see for instance lower right corner in
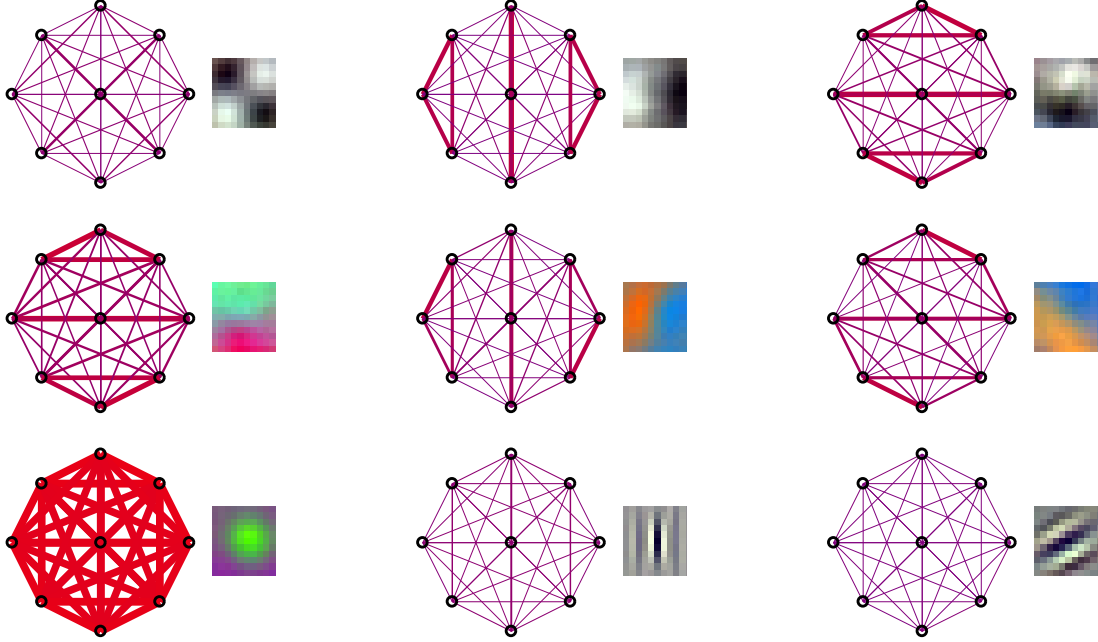
Figure 10: Covariance structure for different units in the first convolutional layer of AlexNet. We display the covariance structure of the surround pool along with the visualization of the corresponding filter. Thicker lines mean larger magnitude of the correlation. Line color linearly interpolates from blue for negative values to red for positive values

Figure 10).

High order correlations between surround components of the first convolutional layer of AlexNet

Here, we show the correlation of energies for the first layer units of AlexNet before and after normalization (Figure 11). We can see that the normalization procedure reduces the energy correlation significantly. In addition to the squared correlation, we also visualize the normalized conditional histograms before and after normalization, as well
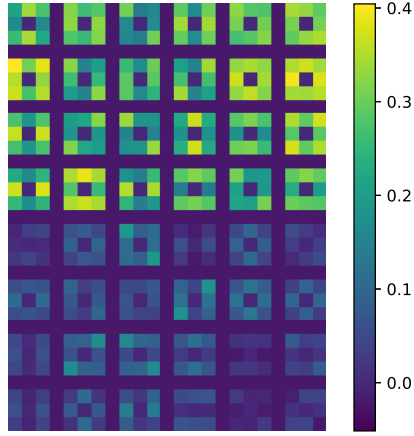
Figure 11: Correlation of energies between center and surround responses for a subset of units of the first convolutional layer of AlexNet. The upper half correspond to the correlation before normalization and the bottom half after flexible normalization.

as the marginal distributions of the center variable.

High order correlations between surround components of the third convolutional layer of AlexNet

We also incorporated flexible normalization into the third convolutional layer "Conv3" of Alexnet. We examined the mutual information and the entropy, in comparison to layer 2 and 1 (Figure 13). Interestingly, mutual information between center and surround responses reduced, and entropy increased, from Conv1 to Conv3. It is also interesting to note that this effect was obtained despite that the overlap between center and surround units was increased. Namely, in Conv1, center and surround responses were 6 strides apart which corresponded to 0% overlap. For Conv2, responses were chosen 4 strides apart resulting in 20% overlap, and for Conv3, 2 strides apart (roughly 33% over-
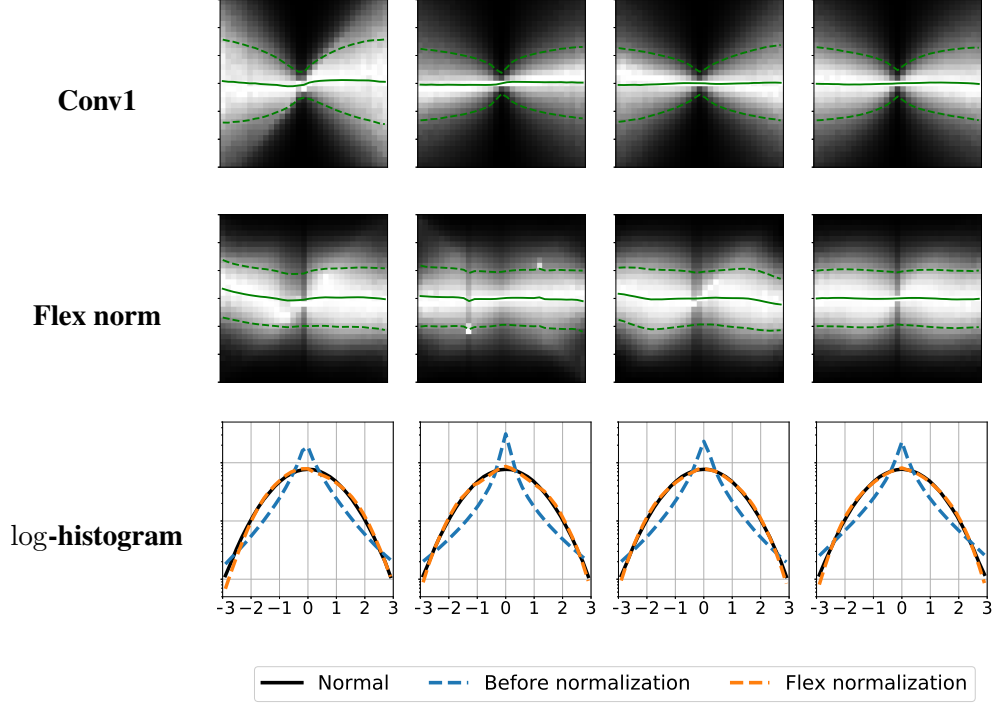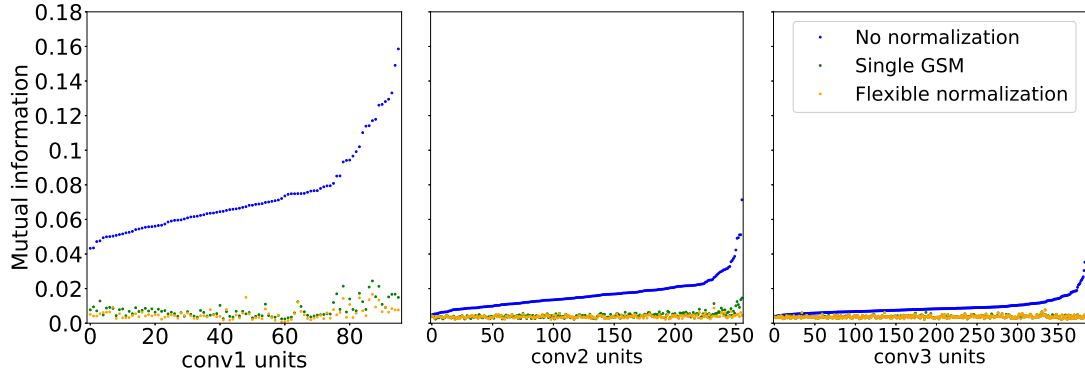
Figure 12: Normalized conditional histograms between center and surround responses from the first convolutional layer of AlexNet. The first two rows from the top are the conditional distributions before and after flexible normalization. The third row are the corresponding $\log$ histograms.
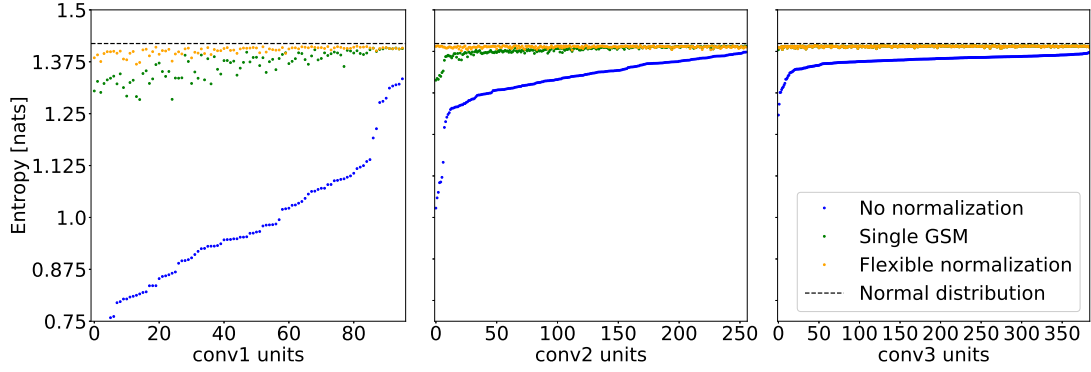
lap). The motivation behind increasing the overlap was that the prior probabilities of center-surround dependence would drop to the point of collapse in Conv3 when center and surround units were chosen to lie further apart.

Single GSM normalization of second layer units of AlexNet

In addition to flexible normalization, we looked at a simpler model which assumes the coupling between center and surround units remains the same across the entire image. This model is a particular case of the flexible normalization where $\Pi_{cs} = 1$. In this model, only the center-surround dependent covariance Λ_{cs} is of interest. As shown in

(a) Mutual Information

(b) Marginal Entropy

Figure 13: Population mutual information and marginal entropy summary statistics for flexible normalization versus the control surround normalization model for first, second and third convolutional layers of AlexNet.

more detail for the population statistics in the main text, the single GSM model reduces dependencies and makes the marginal distributions closer to Gaussian, but not as much as the mixture of GSMs model.
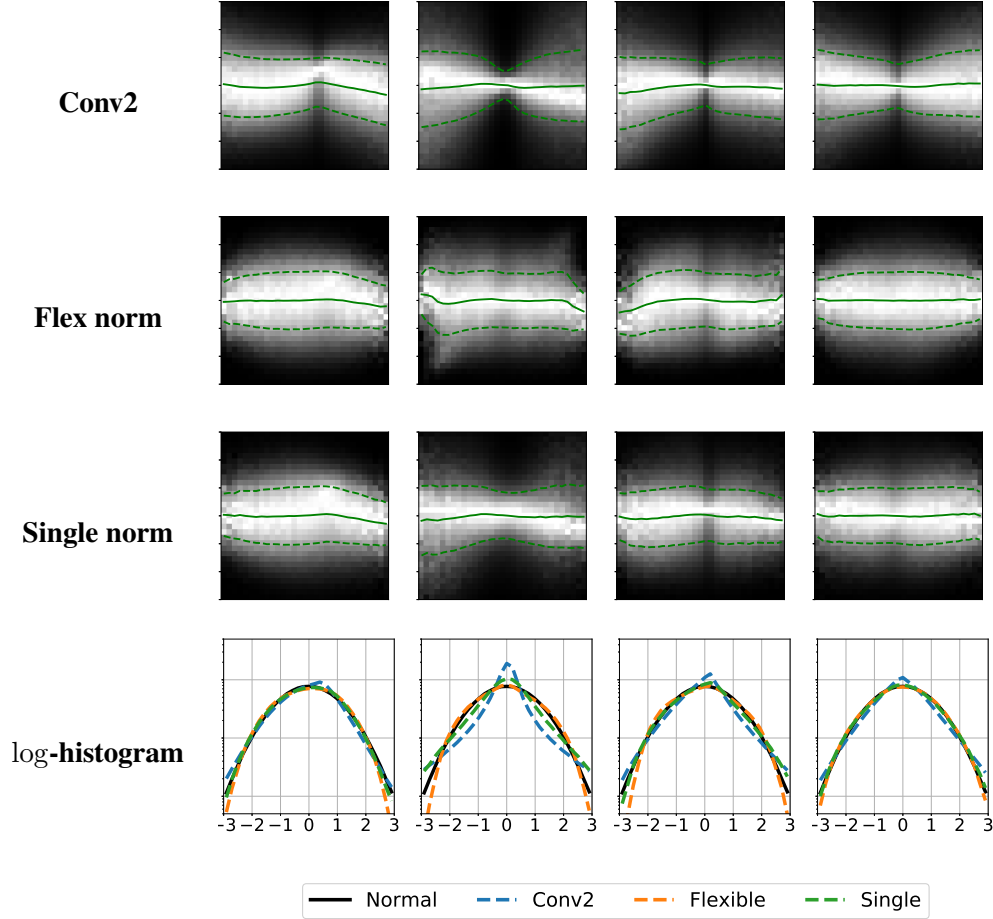
Figure 14: Comparison between flexible and single GSM normalized conditional histograms between center and surround responses from the second convolutional layer of AlexNet. First three rows from the top show the conditional distributions before and after flexible normalization and single GSM normalization. The fourth row shows the corresponding log-histograms.

Predicting homogeneity of stimuli based on the first convolutional layer features of AlexNet

We also mapped the inferred posterior probabilities of the center-surround dependent component for the first convolutional layer of AlexNet. In this case, the units predict less homogeneity in the images as compared to Conv2 units. Flexible normalization in

Conv2 picks up homogeneous structure that is not captured at the Conv1 level. For instance, the textures in the fourth and fifth rows in Figure 15 are consider heterogeneous by Conv1 units, but more homogeneous by Conv2 units. In addition, the foliage background in the image in the last row becomes more homogeneous in Conv2, highlighting one of the objects present in the scene.

References

Andrews, D. and Mallows, C. (1974). Scale mixtures of normal distributions. *J. Royal Stat. Soc.*, 36:99–102.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193.

Ba, L. J., Kiros, R., and Hinton, G. E. (2016). Layer normalization. *CoRR*, abs/1607.06450.

Balle, J., Laparra, V., and Simoncelli, E. P. (2016). Density modelling of images using a generalized normalization transformation. In *International Conference on Learning Representations*.

Barlow, H. B. (1961). *Possible Principles Underllying the Transformations of Sensory Messages*, chapter 13, pages 217–234. MIT press.

Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.
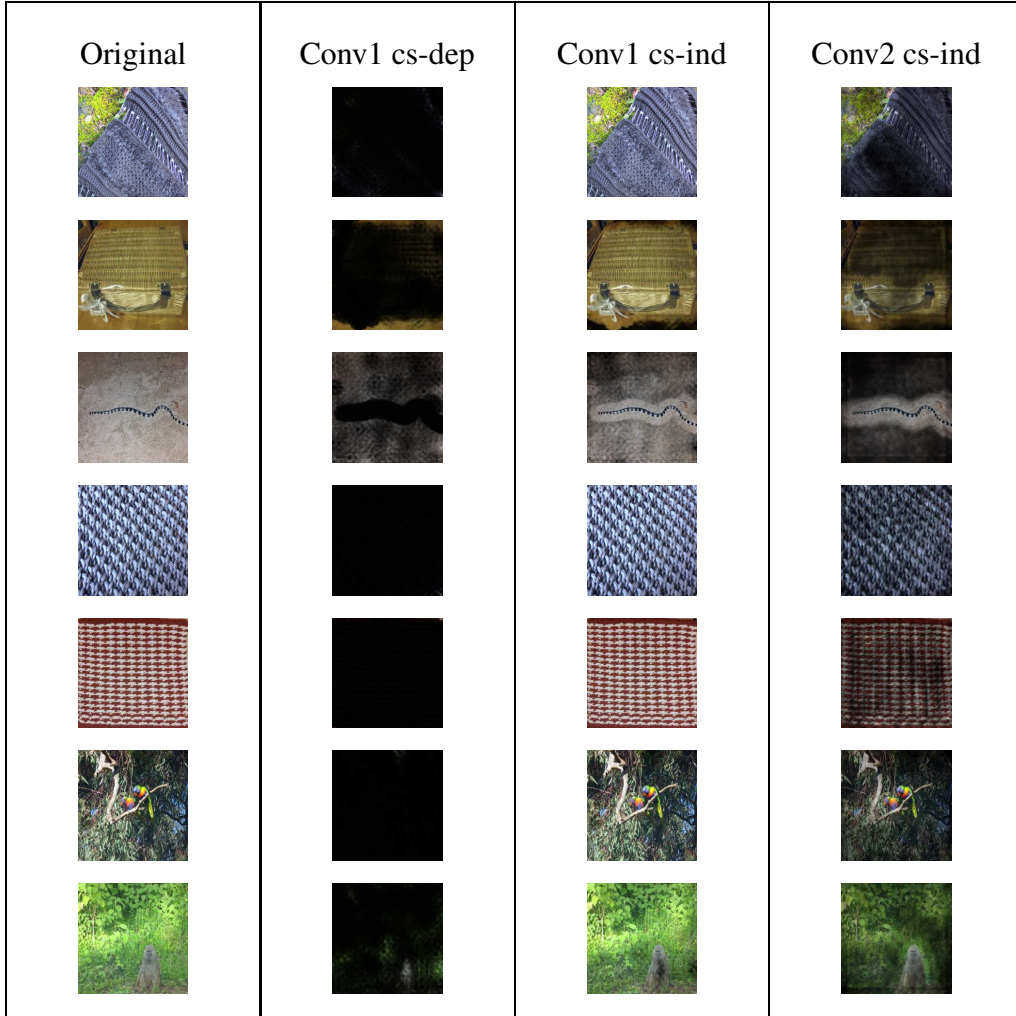
Figure 15: Inferred posterior distributions on ImageNet data. (left) Original images input to the CNN. For each image at each spatial location, we compute the geometric mean of posterior probabilities among all channels of the layer (Conv1). (second column from left) Areas that the model deemed as dependent while obscuring other areas. (third column from left) Complementary display where high center-surround dependent areas are darker, instead. (rightmost column) Conv2 inferred posterior, where high center-surround dependent areas are darker. Flexible normalization in Conv2 picks up homogeneous structure that is not captured at the Conv1 level.

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

Carandini, M., Heeger, D. J., and Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of Neuroscience*, 17(21):8621–8644.

Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2530–2546.

Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2547–2556.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. *arXiv preprint arXiv:1601.02970*.

Coates, A. and Ng, A. Y. (2011). Selecting receptive fields in deep networks. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2528–2536.

Coen-Cagli, R., Dayan, P., and Schwartz, O. (2009). Statistical models of linear and nonlinear contextual interactions in early visual processing. In *Advances in Neural Information Processing Systems 22*.

Coen-Cagli, R., Dayan, P., and Schwartz, O. (2012). Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Computational Biology*, 8(3).

Coen-Cagli, R., Kohn, A., and Schwartz, O. (2015). Flexible gating of contextual modulation during natural vision. *Nature Neuroscience*, 18:1648–1655.

Coen-Cagli, R. and Schwartz, O. (2013). The impact on mid-level vision of statistically optimal divisive normalization in v1. *Journal of Vision*, 13(8).

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 4(12):2379–2394.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2015). Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems 28*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59:167–192.

Golden, J. R., Vilankar, K. P., Wu, M. C., and Field, D. J. (2016). Conjectures regarding the nonlinear geometry of visual neurons. *Vision Research*, 120:74–92.

Han, S. and Vasconcelos, N. (2010). Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision research*, 50(22):2295–2307.

Han, S. and Vasconcelos, N. (2014). Object recognition with hierarchical discriminant saliency networks. *Frontiers in Computational Neuroscience*, 8:109.

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

Hosoya, H. and Hyvrinen, A. (2015). A hierarchical statisitical model of natural images explains tuning properties in v2. *The Journal of Neuroscience*, 35(29):10412–10428.

Hyvrinen, A., Hurri, J., , and Hoyer, P. O. (2009). *Natural Image Statistics: A probabilistic approach to early computational vision*. Springer.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org.

Ito, M. and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *Journal of Neuroscience*, 24(13):3313–3324.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *ICCV*, pages 2146–2153. IEEE.

Karklin, Y. and Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(1):83–87.

Kim, T., Bair, W., and Pasupathy, A. (2019). Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*.

Laskar, M. N. U., Giraldo, L. G. S., and Schwartz, O. (2018). Correspondence of deep neural networks and the brain for visual textures. *CoRR*, abs/1806.02888.

Levitt, J. B. and Lund, J. S. (1997). Contrast dependence of contextual effects in primate visual cortex. *Nature*, 387:73–76.

Li, Z. (1999). Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Network*, 10(2):187–212.

Lochmann, T., Ernst, U. A., and Deneve, S. (2012). Perceptual inference predicts contextual modulations of sensory responses. *Journal of Neuroscience*, 32(12):4179–4195.

Olshausen, B. A. and J.Field, D. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325.

Olshausen, B. A. and Lewicki, M. S. (2014). What natural scenes statistics can tell us about cortical representation. *The New Visual Neurosciences*, pages 1247–1262.

Poggio, T. and Anselmi, F. (2016). *Visual Cortex and Deep Networks: Learning Invariant Representations*. MIT Press.

Portilla, J. and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70.

Pospisil, D., Pasupathy, A., and Bair, W. (2016). Comparing the brain's representation of shape to that of a deep convolutional neural network. In *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (Formerly BIONETICS)*, BICT'15, pages 516–523, ICST, Brussels, Belgium, Belgium. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Pospisil, D. A., Pasupathy, A., and Bair, W. (2018). 'artiphysiology' reveals v4-like shape tuning in a deep network trained for image classification. *eLife*, 7:e38242.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a fucntional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Ren, M., Liao, R., Urtasun, R., Sinz, F. H., and Zemel, R. S. (2017). Normalizing the normalizers: Comparing and extending network normalization schemes. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR 2017.

Rowekamp, R. J. and Sharpee, T. O. (2017). Cross-orientation suppression in visual area v2. *Nature Communications*, 8.

Sceniak, M. P., Ringach, D. L., Hawken, M. J., and Shapley, R. (1999). Contrast's effect on spatial summation by macaque v1 neurons. *Nature Neuroscience*, 2(8):733–739.

Schmid, A. M. and Victor, J. D. (2014). Possible functions of contextual modulations and receptive field nonlinearities: pop-out and texture segmentation. *Vision Research*, 104:57–67.

Schwartz, O., Sejnowski, T., and Dayan, P. (2009). Perceptual organization in the tilt illusion. *Journal of Vision*, 9(4)(19):1–20.

Schwartz, O., Sennowski, T. J., and Dayan, P. (2006). Soft mixer assignment in a hierarchichal generative model of natural scene statistics. *Neural Computation*, 18(11):2680–2718.

Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825.

Shushruth, S., Ichida, J. M., Levitt, J. B., and Angelucci, A. (2009). Comparison of spatial summation properties of neurons in macaque v1 and v2. *Journal of Neurophysiology*, 102(4):2069–2083. PMID: 19657084.

Simoncelli, E. P. (1997). Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conference on Signals, Systems, and Computers*, pages 673–678. IEEE Computer Society.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics nad neural representation. *Annual Reviews Neuroscience*, 24:1193–1216.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR 2015.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area v1. *Journal of Neuroscience*, 30(9):3531–3543.

Wainwright, M. J. and Simoncelli, E. P. (2000). Scale mixtures of Gaussians and the statistics of natural images. In Solla, S. A., Leen, T. K., and Müller, K.-R., editors, *Adv. Neural Information Processing Systems*, volume 12, pages 855–861, Cambridge, MA. MIT Press.

Wainwright, M. J., Simoncelli, E. P., and Willsky, A. S. (2001). Random cascades on wavelet trees and their use in modeling and analyzing natural imagery. *Applied and Computational Harmonic Analysis*, 11(1):89–123. Special issue on wavelet applications.

Wegmann, B. and Zetzsche, C. (1990). Visual-system-based polar quantization of local amplitude and local phase of orientation filter outputs. In *Human Vision and Electronic Imaging: Models, Methods, and Applications*, volume 1249.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional net-works. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Zetzshe, C. and Nuding, U. (2005). Nonlinear and higher-order approaches to the en-coding of natural scenes. *Network*, 16(2-3):191–221.

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area v2. *Neuron*, 47(1):143 – 153.

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Network*, 17(4):301–334.

Zhaoping, L. (2014). *Understanding Vision: Theory, Models, and Data*. Oxford Uni-versity Press.

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, 20(17):6594–6611.

Zhu, M. and Rozell, C. J. (2013). Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. *PLOS Computational Biology*, 9(8).

Ziemba, C. M., Freeman, J., Movshon, J. A., and Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, 113(22):E3140–E3149.

Ziemba, C. M., Freeman, J., Simoncelli, E. P., and Movshon, J. A. (2018). Contextual modulation of sensitivity to naturalistic image structure in macaque v2. *Journal of Neurophysiology*.