Surround Modulation: A Bio-inspired Connectivity Structure for Convolutional Neural Networks

Hosein Hasani
Department of Electrical Engineering
Sharif University of Technology
hasani.hosein@ee.sharif.edu

Mahdieh Soleymani Baghshah
Department of Computer Engineering
Sharif University of Technology
soleymani@sharif.edu

Hamid Aghajan
Department of Electrical Engineering
Sharif University of Technology
aghajan@ee.sharif.edu

Abstract

Numerous neurophysiological studies have revealed that a large number of the primary visual cortex neurons operate in a regime called surround modulation. Surround modulation has a substantial effect on various perceptual tasks, and it also plays a crucial role in the efficient neural coding of the visual cortex. Inspired by the notion of surround modulation, we designed new excitatory-inhibitory connections between a unit and its surrounding units in the convolutional neural network (CNN) to achieve a more biologically plausible network. Our experiments show that this simple mechanism can considerably improve both the performance and training speed of traditional CNNs in visual tasks. We further explore additional outcomes of the proposed structure. We first evaluate the model under several visual challenges, such as the presence of clutter or change in lighting conditions and show its superior generalization capability in handling these challenging situations. We then study possible changes in the statistics of neural activities such as sparsity and decorrelation and provide further insight into the underlying efficiencies of surround modulation. Experimental results show that importing surround modulation into the convolutional layers ensues various effects analogous to those derived by surround modulation in the visual cortex.

1 Introduction

The classical receptive field of a neuron is determined by the region of sensory space where stimuli elicit neural responses. In a sizable population of neurons in the visual cortex, the classical receptive field is surrounded by the nonclassical receptive field, through which the same stimulus can influence the neural response away from that of a mere classical receptive field response [20, 2]. This effect is often referred to as surround or contextual modulation and has been frequently reported in the mammalian visual cortex [20, 2, 29]. The strength of suppression depends on the disparity between the visual features of the stimuli in the classical and nonclassical receptive fields. Various visual features can induce surround suppression, such as spatial frequency, orientation, color, direction of motion, and luminance [2, 29, 59, 53]. Surround modulation has a maximum effect when the center and surround carry similar features. Thus, neurons become more sensitive to the contrast and less sensitive to constant features, and this may enhance the visual perception of objects.

The neural mechanism behind surround modulation is a matter of debate in the literature. However, in general, three types of connections have been identified to be involved in the modulation of

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

classical and nonclassical receptive fields [3]. Bottom-up feedforward connections from the lower areas affect the center of the receptive field, while intra-areal lateral connections mainly contribute to near surround inhibition, and inter-areal top-down feedback connections from higher cortical areas modulate wider areas including far surround.

Surround modulation is believed to be involved in numerous functions of mammalian visual systems. It plays a fundamental role in boundary detection, contour integration, perceptual grouping, figure-ground segmentation, and border ownership [46, 12, 15, 31, 50]. It is also incorporated in visual attention [25, 56], contrast gain control [18], and perception of depth and motion [2, 6, 5, 26]. In addition to these functions, surround modulation also has notable effects on the neural coding of the visual cortex. It increases sparsity, trial-to-trial reliability and temporal precision of spikes, and decreases temporal and spatial redundancies present in natural scenes by removing predictable components [48, 57, 58, 17].

Theoretical studies have shown that natural scenes can be represented by sparse codes [43, 44]. During natural vision, surround modulation forms a sparse representation in the visual cortex [57, 58, 17]. When image patches encompass both classical and nonclassical receptive fields, modulation becomes more suppressive and the mean spiking rate of individual neurons significantly decreases. During normal vision, wide-field stimulation makes neurons operate at optimum efficiency. Sparse coding increases neural selectivity, which means that each cortical neuron is generally silent but strongly responds to a limited set of visual patterns [45, 62]. Surround modulation strongly decorrelates responses across the population of neurons. This suggests that neurons carry more independent information and are unlikely to fire simultaneously [57].

Further electrophysiological studies have revealed that surround modulation, and hence sparse coding, dramatically increase the average information per spike, information transmission rate, and bandwidth efficiency [58]. On the other hand, dense codes are statistically redundant and metabolically inefficient because in this regime, each neuron responds to a broad set of stimuli, and thus the information is distributed across the neural population, rendering each spike less informative. Sparse coding also enhances the formation of synaptic connections, learning rate, and memory capacity [4, 62]. Together, these properties indicate that surround modulation is an essential mechanism for visual processing by offering informationally, computationally, and metabolically efficient neural codes.

There are historical links between the evolution of CNNs and neurophysiological findings. Hubel and Wiesel first described the concept of receptive fields in visual neurons and proposed a hierarchical structure for the visual cortex by introducing simple, complex, and hypercomplex cells [19, 21]. Motivated by these ideas, the first generations of CNNs were reported in the literature of artificial neural networks [13, 32, 33, 49]. CNNs consist of stacked convolutional layers. In each layer, features are extracted from the input to that layer through local convolution operations. Convolutional layers are followed by pooling operations, which make the extracted features invariant to the geometric transformations of the input [33, 49]. Across the layers, receptive fields become more extensive, and features become more abstract. Thus, a deep structure composed of simple convolution layers can achieve invariant object categorization [30]. Analogously, the visual cortex is composed of a hierarchical organization of distinct cortical areas [11]. Neurons in the early visual cortex have small receptive fields and capture low-level visual features. Along the ventral stream, receptive fields become larger, and neurons respond to more complex features. Finally, in the inferior temporal cortex, neurons respond to the semantic contents of the visual stimuli [22]. Recent studies have shown that this resemblance is not limited to the structure of these networks, and there are also strong correspondences between hierarchical features in the layers of CNNs and neuronal responses in the ventral visual stream with respect to the same stimuli. These correspondences exist even though the CNNs were not optimized to fit neural data and were merely optimized for object recognition tasks [7, 27, 61, 16].

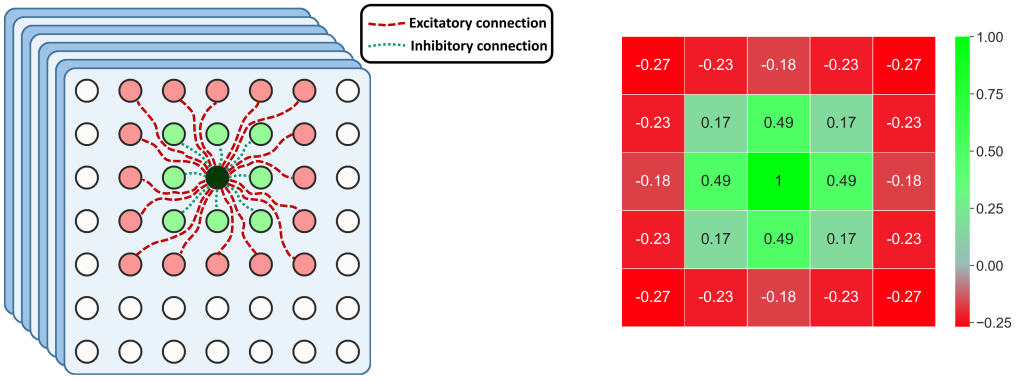
In this paper, motivated by the impressive properties of surround modulation in the biological visual system, we introduce a structure of connections for the standard CNNs to achieve architectures even more similar to the brain. We add local lateral connections to the activation maps of convolutional layers which imitate the function of surround modulation. We found that the proposed structure can outperform traditional CNNs in object recognition tasks while being also capable of speeding up the training procedure. Further evaluations on challenging tasks show that the incorporation of lateral connections can increase the generalization of networks on new domains with different visual

situations which were never seen during training. We also analyze the statistics of the feature space and compare the results with the effects of surround modulation in neural coding of the visual cortex.

Related Work Recently, seeking biologically inspired CNN architectures has drawn considerable attention. Most of the reported studies have focused on augmenting the structure of CNNs by adding recurrent and feedback connections to the convolution layers [41, 54, 36]. One possible explanation for the success of these models is that recurrence enlarges the effective receptive field and also spreads operations across time as well as the network structure. This allows these networks to process each piece of data several times, rendering them capable of reaching the performance of feedforward networks even with shallower models [37, 63, 38]. Linsley et al. [38] proposed a model based on gated recurrent units (GRUs) [8] with additional horizontal connections to solve the contour integration task. Spoerer et al. [54] added between-layer and within-layer recurrent connections to the CNN and evaluated the performance of their model under the presence of noise and clutter on an artificial digit dataset. Nayebi et al. [41] extended the idea by adding attributes such as gating and bypassing, long-range feedback links, and more considerations about the timing of neural dynamics. They also performed a thorough search among different architectures to select the most accurate one.

Our proposed method basically differs from the other biologically motivated models by not using recurrent connections with temporal dynamics. Instead, it employs the existing knowledge about the neural mechanism of surround modulation to incorporate a simple, hard-coded linear operation in the CNNs with no additional training parameters or fundamental changes in the structure of feedforward CNNs.

2 Surround Modulation in CNN



(a) (b)

Figure 1: A simple implementation of surround modulation through lateral connections. **(a)** During modulation, each unit excites near neighbors and inhibits far neighbors within a range based on the level of its own activity. **(b)** A 5×5 SM kernel with associated weights for excitatory connections (green) and inhibitory connections (red).

The Gaussian function has been used to derive a well-established model to describe the mechanism of center-surround modulation [9, 51, 52]. In this model, interactions within the classical receptive field are defined by an excitatory Gaussian, and interactions from the nonclassical receptive filed are defined by a broader inhibitory Gaussian. We employ this model to simulate the center-surround modulation in CNNs through lateral connections. We define a $(2k+1) \times (2k+1)$ kernel whose positive and negative elements respectively determine the excitatory and inhibitory connections between the unit in the location $(k+1, k+1)$ and its neighbors (Figure 1a). The weight associated with each neighbor is defined by the difference of Gaussian (DoG) function:

$$DoG_{\sigma_e,\sigma_i}[x,y] = \frac{1}{2\pi\sigma_e^2} exp(-\frac{(x-(k+1))^2+(y-(k+1))^2}{2\sigma_e^2}) - \frac{1}{2\pi\sigma_i^2} exp(-\frac{(x-(k+1))^2+(y-(k+1))^2}{2\sigma_i^2}). \quad (1)$$

where σ_e and σ_i are the standard deviations of the excitatory and inhibitory Gaussians, respectively. The surround modulation (SM) kernel is obtained by normalizing the DoG to the amplitude of its center:

$$SM[x, y] = \frac{DoG[x, y]}{DoG[k + 1, k + 1]}. \tag{2}$$

The SM kernel has been designed to increase the feature saliency by suppressing redundant and spatially constant responses within activation maps. Each unit in the activation map m of layer l is modulated by its spatial neighborhood with the corresponding weights in the SM kernel. One plausible approach is to linearly add the weighted responses of the surround units to the response of the center unit. To do this, it is sufficient to convolve each activation map by the SM kernel to obtain the surround modulated activation map:

$$act^{l,m}_{SM}[x, y] = (act^{l,m} * SM)[x, y]. \tag{3}$$

It is not necessary to strictly follow the exact DoG profile to achieve excitatory connections in the center and inhibitory connections in the surround. Especially for SM kernels with smaller sizes, providing a proper balance between the amounts of excitation and inhibition may entail subtle modifications of the kernel. For each specific task, finding the optimal design for the SM kernel and its most effective placement in the CNN may need performing hyperparameter search. However, the aim of this paper is not to explore all possible setups, but to present a simple, fixed setup which resembles the surround modulation phenomenon in the cortical nervous system. Hence, for all of our experiments we deploy a fixed SM kernel with a size of 5×5 ($k = 2$) obtained by setting $\sigma_e = 1.2$ and $\sigma_i = 1.4$ (Figure 1b). In our experiments, the SM kernel was added to activation maps of the first convolutional layer as such modulation is more common in the early visual cortex.

3 Experiments

To evaluate the performance of the proposed method, we set up three types of experiments. First, we apply the proposed model to a standard image classification task. Then, we evaluate the robustness of the model in different visual situations. Finally, we analyze the characteristics of neural activities in the presence of surround modulation and compare the results with those reported for the visual cortex.

In all experiments, baseline is a standard CNN which consists of multiple convolution layers with 3×3 kernels and stride of 1, multiple max-pooling layers with 2×2 kernels and stride of 2, and three fully-connected layers at the end of the network. The last layer predicts the probability of categories with a softmax activation function and the other fully-connected convolution layers have ReLU non-linearities. All implementations are done in Tensorflow [1], and during the training procedure, Adam optimizer [28] with a learning rate of 10^{-4} is used to minimize the cross entropy loss (see Supplementary Materials for more details). We train all of the networks from scratch by initializing trainable weights with Xavier initialization [14], and repeat each experiment 10 times.

3.1 Image Classification

We used the ImageNet dataset [10] for object recognition as it contains natural images with an acceptable resolution to test surround modulation. For further analysis, we composed a baseline dataset, hereby called baseline-ImageNet, by randomly choosing 100 categories. From each category, 500 instances were randomly chosen for training, 50 instances for validation, and 100 instances for the test set. All images were cropped around their centers and resized to 160×160 pixels. The baseline network contains 2.3M trainable parameters. It has seven convolutional layers, five pooling layers, and three fully-connected layers with Dropout [55] in each of them. We did not perform any data augmentation during training or any multi-cropping during the test. This was done to provide a standard and balanced dataset along with a standard CNN architecture, i.e. an impartial setting, to evaluate the exclusive effect of adding SM to the network.

We constructed the SM-CNN by adding the SM kernel to half of the activation maps in the first convolution layer of the baseline network. The SM-CNN has the same number of training parameters as the baseline network and requires less than 1% extra computations for performing surround

modulations. Figure 2 shows the validation loss and accuracy as a function of the number of training steps. Surround modulation increases both the training speed and accuracy of the baseline model ($p < 0.001$).

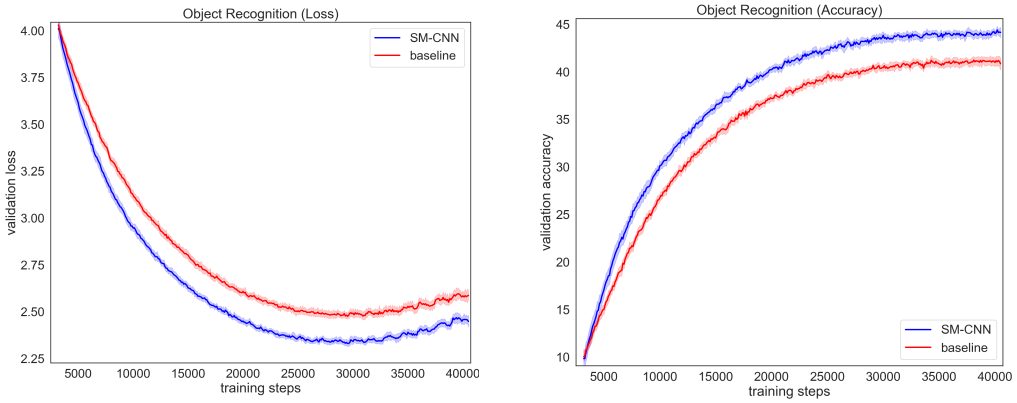


Figure 2: Baseline-ImageNet validation loss (left) and top-1 accuracy (right) of the baseline network and SM-CNN across training steps. The pale margins indicate confidence intervals in the trials.

To compare the performance of the models more precisely, we designed six additional networks as follows. Two control models were designed by adding an extra layer after the first convolutional layer of the baseline network with a kernel size of (5×5). The number of convolution kernels is the same as the number of activation maps in the first layer. The first control model ($\mathbf{C}^1_{\text{baseline}}$) has no activation function after the new layer, while the second one ($\mathbf{C}^2_{\text{baseline}}$) has a ReLU nonlinearity.

Four variants of the SM-CNN were also designed. In the first variant ($\mathbf{C}^1_{\text{randSM}}$), we replaced the SM kernel with a random kernel matching the first and second-order statistics of the SM kernel to analyze the advantages offered by the SM configuration. In the second variant ($\mathbf{C}^2_{\text{SM}}$), the SM kernel was applied to the input image as the preprocessing step. In the third variant ($\mathbf{C}^3_{\text{SM}}$), the SM kernel was added to all activation maps of the first layer. In the fourth variant ($\mathbf{C}^4_{\text{SM}}$), the SM kernel was applied after the first max-pooling layer, which is itself located after the first two layers of convolution. The top-1 accuracy of all the network variants is shown in Table 1. The main SM-CNN, as well as its third and fourth variants, outperform the traditional baseline variants ($p < 0.001$) even with fewer parameters and computations.

Table 1: Top-1 accuracy of the baseline and SM-CNN and their related control models for the baseline-ImageNet dataset.

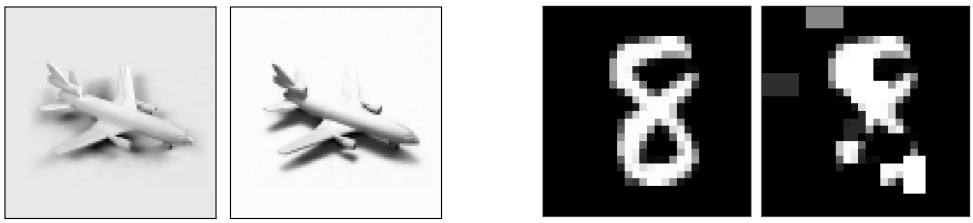
	baseline variants			SM-CNN variants				
	main	$\mathbf{C}^1_{\text{baseline}}$	$\mathbf{C}^2_{\text{baseline}}$	main	$\mathbf{C}^1_{\text{randSM}}$	$\mathbf{C}^2_{\text{SM}}$	$\mathbf{C}^3_{\text{SM}}$	$\mathbf{C}^4_{\text{SM}}$
ACC	$41.0_{\pm 0.6}$	$41.3_{\pm 0.6}$	$40.7_{\pm 0.5}$	$43.8_{\pm 0.7}$	$41.0_{\pm 0.9}$	$40.8_{\pm 0.4}$	$42.3_{\pm 0.6}$	$43.2_{\pm 0.4}$

The inferior performance of the baseline variants indicates that the surround modulation effect cannot be trivially replaced by a learnable convolution layer. The average performance of $\mathbf{C}^1_{\text{randSM}}$ is not higher than the baseline, indicating the effectiveness of the excitatory-inhibitory pattern in the proposed SM kernel. The performance levels of the third and fourth SM-CNN variants imply that surround modulation could be effective in various setups even when applied to the intermediate layers of the network, while the performance level of the second SM-CNN variant indicates that adding the SM kernel to the input of the network may not provide a notable improvement.

3.2 Robustness and Generalization

As discussed in the introductory remarks, surround modulation serves as a proxy function for various visual perception tasks. In this section, we examine the ability of the proposed method in handling challenging visual tasks. We train each model on the standard datasets and test them on domains with different situations as follows.

Illumination Here, we explored the generalization capability of the networks under drastic changes in illumination. To this end, we used the small NORB dataset [35], which contains images of 3D objects under six different lighting conditions (examples in Figure 3a). We trained the networks on training images from one lighting situation and tested them on test sets composed of unfamiliar lighting conditions. To assess whether the generalization capability offered by surround modulation can be obtained by other standard regularization methods, we examined three additional control models. In the first control model, we added L_2 regularization to all of the trainable weights of the baseline network. In the second control model, we added Dropout to the last three fully-connected layers of the baseline network. In the third control model, we added Batch Normalization [24] to all convolutional layers of the baseline network.



(a) (b)

Figure 3: Data samples for challenging image classification tasks. **(a)** An instance from the small NORB dataset with its different illumination variant. **(b)** An instance from the MNIST dataset with its artificially occluded variant.

The overall accuracies on familiar and unfamiliar test sets are reported in Table 2. Averaging across six conditions and ten trials shows that the baseline CNN lost about 36% of its accuracy in new lighting situations, whereas the SM-CNN lost just about 17%. The superior robustness of the SM-CNN over the standard regularization methods suggests that surround modulation can be employed as a regularization technique, although it was not directly designed for this purpose. Figure 4 shows the accuracies of familiar and unfamiliar test cases in one of the lighting conditions as a function of training steps.

Table 2: The overall test accuracy of the familiar and unfamiliar test sets for all six lighting conditions.

		light_0	light_1	light_2	light_3	light_4	light_5
familiar	baseline	$85.2_{\pm 1.6}$	$85.6_{\pm 1.4}$	$80.8_{\pm 1.1}$	$85.6_{\pm 1.2}$	$85.3_{\pm 1.1}$	$84.6_{\pm 1.8}$
	baseline_{L_2}	$86.8_{\pm 1.5}$	$86.5_{\pm 1.8}$	$81.2_{\pm 2.3}$	$85.8_{\pm 1.2}$	$85.6_{\pm 1.7}$	$86.4_{\pm 1.6}$
	$\text{baseline}_{\text{Dropout}}$	$88.0_{\pm 1.8}$	$88.4_{\pm 1.6}$	$82.7_{\pm 1.8}$	$84.7_{\pm 2.3}$	$\mathbf{87.4}_{\pm 1.9}$	$85.9_{\pm 1.8}$
	$\text{baseline}_{\text{BN}}$	$90.8_{\pm 0.8}$	$91.7_{\pm 1.6}$	$86.8_{\pm 2.0}$	$89.0_{\pm 1.6}$	$85.8_{\pm 0.8}$	$87.3_{\pm 1.1}$
	SM-CNN	$\mathbf{92.3}_{\pm 1.1}$	$\mathbf{93.0}_{\pm 0.8}$	$\mathbf{89.4}_{\pm 1.1}$	$\mathbf{91.7}_{\pm 0.9}$	$87.1_{\pm 1.3}$	$\mathbf{90.1}_{\pm 1.1}$
unfamiliar	baseline	$57.4_{\pm 1.5}$	$56.1_{\pm 2.4}$	$35.1_{\pm 2.1}$	$55.0_{\pm 1.6}$	$46.8_{\pm 2.1}$	$41.3_{\pm 2.0}$
	baseline_{L_2}	$59.8_{\pm 1.7}$	$56.7_{\pm 1.8}$	$35.5_{\pm 1.4}$	$56.0_{\pm 2.0}$	$46.9_{\pm 1.6}$	$42.5_{\pm 1.5}$
	$\text{baseline}_{\text{Dropout}}$	$66.8_{\pm 1.8}$	$67.4_{\pm 2.4}$	$36.8_{\pm 1.6}$	$62.2_{\pm 2.7}$	$49.1_{\pm 1.1}$	$48.9_{\pm 3.6}$
	$\text{baseline}_{\text{BN}}$	$65.1_{\pm 2.8}$	$57.2_{\pm 2.7}$	$31.6_{\pm 3.7}$	$64.8_{\pm 2.2}$	$43.1_{\pm 5.1}$	$50.8_{\pm 1.6}$
	SM-CNN	$\mathbf{80.8}_{\pm 1.2}$	$\mathbf{73.0}_{\pm 2.5}$	$\mathbf{61.4}_{\pm 2.5}$	$\mathbf{82.2}_{\pm 0.8}$	$\mathbf{65.8}_{\pm 1.3}$	$\mathbf{80.3}_{\pm 1.0}$

Occlusion A potential advantage of employing surround modulation by the biological visual system is increasing robustness to the presence of occlusion and clutter. Handling these conditions in the brain may be additionally facilitated by attention mechanisms through top-down feedback connections [47]. In our analysis, we investigate the robustness of the proposed surround modulation method in the presence of small clutter on the MNIST dataset [34].

We augmented the test images by adding from 8 to 15 random patches with different shapes and intensities to each image (one example is shown in Figure 3b). The length of each patch varied from 3 to 6 pixels. To estimate the generalization ability of the networks in challenging situations more accurately, we restricted the size of the training set to 1000 by randomly sampling 100 samples from each class. Networks were trained on the standard images with no additional augmentation but tested

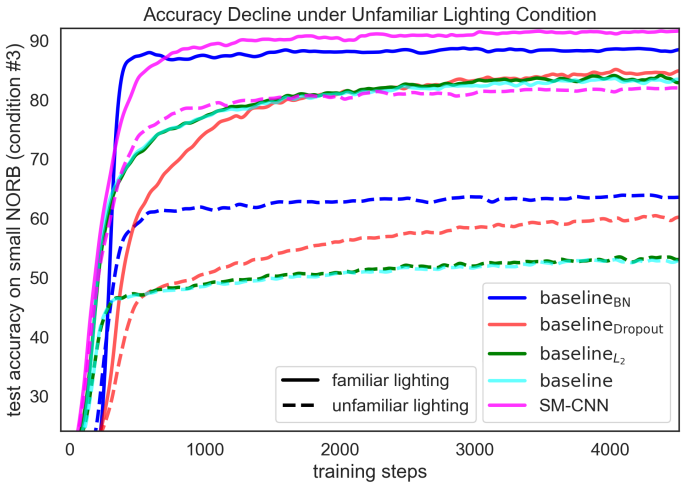


Figure 4: The accuracy of the familiar and unfamiliar test sets with training performed on the lighting condition 3 of the small NORB dataset.

on the standard and occluded images. As shown in Table 3, in the occluded scenario, the accuracy of the baseline method dropped by 37.8% while the accuracy of the SM-CNN dropped by 26.2%.

Table 3: Classification accuracy on the standard and occluded test images of the smaller version of the MNIST dataset.

	standard	occluded
baseline	$92.6_{\pm 0.3}$	$54.8_{\pm 1.1}$
SM-CNN	$93.2_{\pm 0.2}$	$67.0_{\pm 1.1}$

3.3 Neural Activity Characteristics

As stated in the introductory section, a particular effect of surround modulation is to increase the sparsity of the neural activities. This effect bodes well with the requirements of biological visual systems which need to operate under maximum information and energy efficiency constraints during natural vision. To assess whether our implementation of surround modulation is capable of inducing sparsity to artificial neural networks, we set up a task-driven approach. We optimized networks for object recognition on natural images and then analyzed the neural activities of the final networks and compared the results with observations reported in studies of the biological visual system. To this end, we deployed the networks which were trained on the baseline-ImageNet dataset and fed them with 5000 images from the test set. Then, we recorded the neural activity of the first and third convolution layers. In addition to the baseline network and the SM-CNN, we also analyzed the effect of the presence of Batch Normalization in the first convolution layer on neural activities. For each model, we trained six different networks from scratch to ensure that the statistics do not change in different trials.

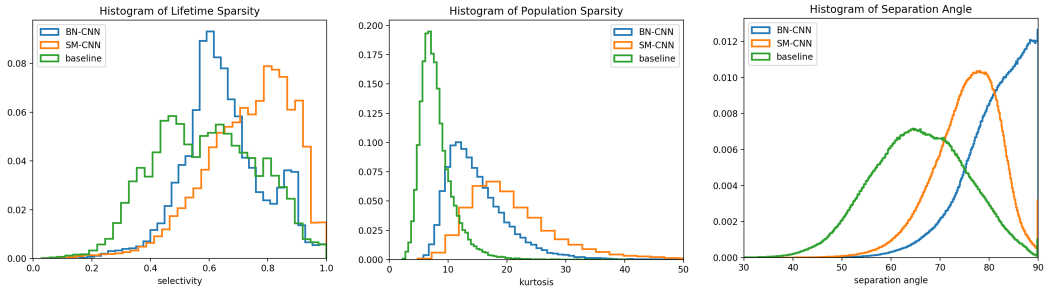
We report sparsity of the neural responses in two different ways: lifetime sparsity, which characterizes the sparsity of the response of each unit to the entire set of stimuli, and population sparsity, which characterizes the sparsity of the activity of the neural population in response to a stimulus [60]. We compute the population and lifetime sparsities with kurtosis [1] and the selectivity index, both of which are well-studied metrics in neuroscience [57, 58, 60, 45]. Figures 5a and 5b show the distributions of the selectivity index for the lifetime sparsity and the kurtosis index for the population sparsity of the third convolutional layer, respectively. We also estimated the sparsity of the overall activities by the Gini metric, which is a suitable sparsity measure especially for one-sided distributions [23] (see

[1] The distribution of the amplitudes of neural responses is one-sided, while kurtosis is suitable for two-sided distributions. So, similar to [57], we add to each distribution its mirrored version around the origin and estimate the kurtosis metric for the resulting two-sided zero-mean distribution.

Table 4: Average sparsity scores of neural activities from the first and third convolutional layers across six trials. In all of the sparsity measures, SM-CNN and baseline$_{\text{BN1}}$ have higher scores than the baseline network, indicating higher sparsity in their neural coding.

	First convolutional layer			Third convolutional layer		
	baseline	baseline$_{\text{BN1}}$	SM-CNN	baseline	baseline$_{\text{BN1}}$	SM-CNN
kurtosis	$5.8_{\pm 0.7}$	$22.0_{\pm 1.2}$	$22.6_{\pm 1.7}$	$11.6_{\pm 1.1}$	$17.2_{\pm 2.7}$	$28.2_{\pm 3.3}$
selectivity	$0.50_{\pm 0.02}$	$0.77_{\pm 0.01}$	$0.72_{\pm 0.02}$	$0.66_{\pm 0.03}$	$0.71_{\pm 0.07}$	$0.78_{\pm 0.01}$
Gini	$0.51_{\pm 0.02}$	$0.74_{\pm 0.01}$	$0.64_{\pm 0.01}$	$0.65_{\pm 0.03}$	$0.67_{\pm 0.09}$	$0.76_{\pm 0.01}$

Supplementary Materials for definitions). In all of the metrics, a higher value indicates higher sparsity. Table 4 shows the sparsity measures of the first and third layers of each model across the population and stimuli. As the results indicate, surround modulation, as well as CNN with Batch Normalization in its first layer (baseline$_{\text{BN1}}$), significantly increase the sparsity of the neural activities.



(a) (b) (c)

Figure 5: Distributions of sparsity and decorrelation for the third convolutional layer. **(a)** Distribution of lifetime sparsity based on selectivity index. **(b)** Distribution of population sparsity based on kurtosis index. **(c)** Distribution of decorrelation based on separation angles between random pairs of neurons.

To evaluate whether surround modulation increases independence between the units of a specific layer, we followed the same strategy employed in [57]. We randomly selected 1M pairs of neurons in each layer and computed the separation angles between their responses to 5000 stimuli (see Supplementary Materials). Neuron pairs that carry similar information would have low separation angles. Figure 5c shows the distribution of separation angles in the third convolutional layer of the models mentioned above. Surround modulation, as well as Batch Normalization, decrease the correlation between neural activities among the neural population, and this property also propagates to higher layers.

Together, these results indicate that our simple implementation of surround modulation in CNNs offers considerable efficiencies in neural coding of the early and intermediate layers. Surround modulation increases sparseness and decreases statistical redundancies, analogous to similar effects reported in studies of the biological visual system [57, 58, 17]. Our results also provide more insight about efficient neural coding in CNNs and suggest that sparse coding in these networks may offer similar properties as those considered for sensory cortices, such as improvement in the learning rate and prediction accuracy [4, 62].

4 Discussion

In the present study, we introduce a biologically plausible modification of CNNs by leveraging a well-known phenomenon in the biological visual system, the so-called surround modulation. The DoG function is a well-studied operator in computer vision, which is mainly used for low-level feature extraction like edge and blob detection [40, 39]. Inspired by the role of this function in modeling surround modulation, we implemented a simplified version of surround modulation by a linear operation that can be easily incorporated into the convolutional layers of CNNs. We found that the proposed method can significantly improve the performance of CNNs in standard image

classification tasks. While the resulting improvement is not trivial, it is also somewhat surprising, because the proposed method merely introduces some linear operations with no additional training parameters, and the increment in computations is also limited.

We investigated the impact of this simple form of surround modulation on improving the robustness of CNNs under difficult visual conditions. We also analyzed the potential implications of the proposed surround modulation structure in increasing sparsity and decorrelation in the neural activities of the CNNs. In most of the cases, the surround modulated CNN reaches the performance of the baseline in fewer optimization steps or with a smaller size of training data. Based on these experiments, it seems that surround modulation is capable of facilitating learning, an observation that is consistent with the neurophysiological studies of the brain [62]. A possible justification is that surround modulation increases feature saliency and decreases redundancies which may slow down training. We also analyzed neural activities in the presence of Batch Normalization which is also considered to improve the accuracy of the network and its training speed. Batch Normalization increases sparsity and decorrelation analogous to surround modulation. Thus, the success behind Batch Normalization presumably can be explained by efficient neural coding.

The potential gain offered by surround modulation on the training speed is interesting as it alleviates the need for numerous training steps on large datasets, which is a major drawback of traditional CNNs in comparison with the biological visual system. However, more meticulous studies are needed to explore the role of surround modulation in facilitating few-shot learning.

Our implementation of surround modulation is straightforward. It can be easily utilized in different CNN architectures and for different visual tasks. Semantic segmentation is one type of challenging tasks in which surround modulation may perform well. Surround modulation reduces sensitivity to constant textures and also gets involved in various functions associated with image segmentation such as border ownership, perceptual grouping, and contour integration [12, 50, 15, 31, 46]. However, in occlusion scenarios, one may need to employ top-down feedback links in order to incorporate more abstract features in the segmentation task [47].

In summary, this work introduces a new biologically-motivated connectivity structure for the CNNs resembling the structure of the visual cortex. As a result, more biologically plausible neural coding, better generalization performance, and more biologically plausible behavior in training speed are also achieved, even though the method was not explicitly designed for these gains. More work is needed to explore and compare various setups for implementing surround modulation. Here, we implemented surround modulation as near-surround lateral connections, but physiological studies have frequently reported the presence of extensive top-down feedback links [42], especially involved in far-surround modulation, which is not covered in the present study and is worth exploring.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] John Allman, Francis Miezin, and EveLynn McGuinness. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual review of neuroscience*, 8(1):407–430, 1985.

[3] Alessandra Angelucci and Paul C Bressloff. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. *Progress in brain research*, 154:93–120, 2006.

[4] Horace Barlow. Redundancy reduction revisited. *Network: computation in neural systems*, 12(3):241–253, 2001.

[5] Richard T Born and Roger BH Tootell. Segregation of global and local motion processing in primate middle temporal visual area. *Nature*, 357(6378):497, 1992.

[6] Giedrius T Buracas and Thomas D Albright. Contribution of area mt to perception of three-dimensional shape: a computational study. *Vision research*, 36(6):869–888, 1996.

[7] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[9] Gregory C DeAngelis, RALPH D Freeman, and IZUMI Ohzawa. Length and width tuning of neurons in the cat's primary visual cortex. *Journal of neurophysiology*, 71(1):347–374, 1994.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Daniel J Felleman and DC Essen Van. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[12] David J Field, Anthony Hayes, and Robert F Hess. Contour integration by the human visual system: evidence for a local "association field". *Vision research*, 33(2):173–193, 1993.

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

[14] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[15] Stephen Grossberg, Ennio Mingolla, and William D Ross. Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in neurosciences*, 20(3):106–111, 1997.

[16] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[17] Bilal Haider, Matthew R Krause, Alvaro Duque, Yuguo Yu, Jonathan Touryan, James A Mazer, and David A McCormick. Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron*, 65(1):107–121, 2010.

[18] David J Heeger. Normalization of cell responses in cat striate cortex. *Visual neuroscience*, 9(2):181–197, 1992.

[19] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[20] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*, 28(2):229–289, 1965.

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

[22] Chou P Hung, Gabriel Kreiman, Tomaso Poggio, and James J DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

[23] Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[25] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194, 2001.

[26] HE Jones, KL Grieve, W Wang, and AM to. Surround suppression in primate v1. *Journal of neurophysiology*, 86(4):2011–2028, 2001.

[27] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] James J Knierim and David C Van Essen. Neuronal responses to static texture patterns in area v1 of the alert macaque monkey. *Journal of Neurophysiology*, 67(4):961–980, 1992.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[31] V A_f Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *Journal of Neuroscience*, 15(2):1605–1615, 1995.

[32] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1 (4):541–551, 1989.

[33] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

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

[35] Yann LeCun, Fu Jie Huang, Leon Bottou, et al. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR (2)*, pages 97–104. Citeseer, 2004.

[36] Xin Li, Zequn Jie, Jiashi Feng, Changsong Liu, and Shuicheng Yan. Learning with rethinking: Recurrently improving convolutional neural networks through feedback. *Pattern Recognition*, 79:183–194, 2018.

[37] Qianli Liao and Tomaso Poggio. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*, 2016.

[38] Drew Linsley, Junkyung Kim, Vijay Veerabadran, Charles Windolf, and Thomas Serre. Learning long-range spatial dependencies with horizontal gated recurrent units. In *Advances in Neural Information Processing Systems*, pages 152–164, 2018.

[39] David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999.

[40] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980.

[41] Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel L Yamins. Task-driven convolutional recurrent models of the visual system. In *Advances in Neural Information Processing Systems*, pages 5290–5301, 2018.

[42] Lauri Nurminen, Sam Merlin, Maryam Bijanzadeh, Frederick Federer, and Alessandra Angelucci. Top-down feedback controls spatial summation and response amplitude in primate visual cortex. *Nature communications*, 9(1):2281, 2018.

[43] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996.

[44] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[45] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

[46] Esther Peterhans and Rüdiger von der Heydt. Subjective contours-bridging the gap between psychophysics and physiology. *Trends in neurosciences*, 14(3):112–119, 1991.

[47] Karim Rajaei, Yalda Mohsenzadeh, Reza Ebrahimpour, and Seyed-Mahdi Khaligh-Razavi. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *Biorxiv*, page 302034, 2019.

[48] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

[49] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019, 1999.

[50] Ko Sakai and Haruka Nishimura. Surrounding suppression and facilitation in the determination of border ownership. *Journal of Cognitive Neuroscience*, 18(4):562–579, 2006.

[51] Michael P Sceniak, Dario L Ringach, Michael J Hawken, and Robert Shapley. Contrast's effect on spatial summation by macaque v1 neurons. *Nature neuroscience*, 2(8):733, 1999.

[52] Michael P Sceniak, Michael J Hawken, and Robert Shapley. Visual spatial characterization of macaque v1 neurons. *Journal of neurophysiology*, 85(5):1873–1887, 2001.

[53] Zhi-Ming Shen, Wei-Feng Xu, and Chao-Yi Li. Cue-invariant detection of centre–surround discontinuity by v1 neurons in awake macaque monkey. *The Journal of physiology*, 583(2):581–592, 2007.

[54] Courtney J Spoerer, Patrick McClure, and Nikolaus Kriegeskorte. Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, 8:1551, 2017.

[55] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15 (1):1929–1958, 2014.

[56] Kristy A Sundberg, Jude F Mitchell, and John H Reynolds. Spatial attention modulates center-surround interactions in macaque visual area v4. *Neuron*, 61(6):952–963, 2009.

[57] William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

[58] William E Vinje and Jack L Gallant. Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1. *Journal of Neuroscience*, 22(7):2904–2915, 2002.

[59] Gary A Walker, Izumi Ohzawa, and Ralph D Freeman. Asymmetric suppression outside the classical receptive field of the visual cortex. *Journal of Neuroscience*, 19(23):10536–10553, 1999.

[60] Ben DB Willmore, James A Mazer, and Jack L Gallant. Sparse coding in striate and extrastriate visual cortex. *Journal of neurophysiology*, 105(6):2907–2919, 2011.

[61] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[62] Haishan Yao, Lei Shi, Feng Han, Hongfeng Gao, and Yang Dan. Rapid learning in cortical coding of visual scenes. *Nature neuroscience*, 10(6):772, 2007.

[63] Amir R Zamir, Te-Lin Wu, Lin Sun, William B Shen, Bertram E Shi, Jitendra Malik, and Silvio Savarese. Feedback networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1308–1317, 2017.