frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 12 August 2015
doi: 10.3389/fncom.2015.00104

Diversity priors for learning early visual features

*Hanchen Xiong *, Antonio J. Rodríguez-Sánchez, Sandor Szedmak and Justus Piater*

Intelligent and Interactive Systems Group, Institute of Computer Science, University of Innsbruck, Innsbruck, Austria

This paper investigates how utilizing diversity priors can discover early visual features that resemble their biological counterparts. The study is mainly motivated by the sparsity and selectivity of activations of visual neurons in area V1. Most previous work on computational modeling emphasizes selectivity or sparsity independently. However, we argue that selectivity and sparsity are just two epiphenomena of the diversity of receptive fields, which has been rarely exploited in learning. In this paper, to verify our hypothesis, restricted Boltzmann machines (RBMs) are employed to learn early visual features by modeling the statistics of natural images. Considering RBMs as neural networks, the receptive fields of neurons are formed by the inter-weights between hidden and visible nodes. Due to the conditional independence in RBMs, there is no mechanism to coordinate the activations of individual neurons or the whole population. A diversity prior is introduced in this paper for training RBMs. We find that the diversity prior indeed can assure simultaneously sparsity and selectivity of neuron activations. The learned receptive fields yield a high degree of biological similarity in comparison to physiological data. Also, corresponding visual features display a good generative capability in image reconstruction.

Keywords: restricted Boltzmann machine, diversity prior, V1 simple cell, inhibition, Markov networks

OPEN ACCESS

Edited by:
Judith Peters, Netherlands Institute for Neuroscience - Royal Netherlands Academy of Arts and Sciences, Netherlands

Reviewed by:
Joel Zylberberg, University of Washington, USA
Rajat Mani Thomas, Netherlands Institute for Neuroscience - Royal Netherlands Academy of Arts and Sciences, Netherlands

****Correspondence:***
Hanchen Xiong, Institute of Computer Science, University of Innsbruck, Technikerstrasse 21a, A-6020 Innsbruck, Austria
hanchen.xiong@uibk.ac.at

Received: *30 March 2015*
Accepted: *27 July 2015*
Published: *12 August 2015*

Citation:
Xiong H, Rodríguez-Sánchez AJ, Szedmak S and Piater J (2015) Diversity priors for learning early visual features. Front. Comput. Neurosci. 9:104. doi: 10.3389/fncom.2015.00104

1. Introduction

Much has been advanced in the knowledge of the brain in the last century since the foundation of modern neuroanatomy by Ramón y Cajal (Ramón y Cajal, 1888, 1904; Jones, 2007). The work of HUBEL and WIESEL (1959) was the first breakthrough in the understanding of simple cells in area V1 of the visual cortex. V1 simple cells perform an early stage processing of the visual input from the retina and the lateral geniculate nucleus (LGN). One important property of V1 simple cells is that their receptive fields are *selective* in terms of location, orientation, and frequency, which can be modeled by Gabor filters. Another characteristic on V1 simple cells is that their activation pattern—when analyzed as a population—is sparse (Field, 1994). Selectivity (also referred to as "lifetime sparseness" by Willmore and Tolhurst, 2001) is related to a neuron having a response only to a small number of different (although similar) stimuli and providing a much lower response to other (usually very different) stimuli. Sparsity (or "population sparseness" by Willmore and Tolhurst, 2001) is a term expressing that the fraction of neurons from a population that is activated by a certain stimulus should be relatively small. Selectivity and sparsity would be due to a redundancy-reduction mechanism, where the visual cortex has evolved to encode visual information as efficiently as possible (Barlow, 1989). This *sparse coding* would then enhance coding efficiency, and when tested, leads in fact to Gabor-like representations (Olshausen and Field, 1996). Although sparse coding has been very successful at generating receptive fields similar to

those of simple cells, sparsity does not necessarily imply selectivity (Willmore and Tolhurst, 2001). In addition to this, recent multi-unit neurophysiological recordings found that just maximizing sparsity does not correlate with visual experience, suggesting that coding efficiency is also due to lateral, recurrent and feedback connections for the purpose of resolving ambiguities (Berkes et al., 2009). In order to show the (lack of) relationship between sparsity and selectivity, we illustrate these concepts in **Figure 1A**. Each row (red) in this figure represents how one neuron selectively responds to different visual stimuli while each column (blue) describes how many neurons are activated by one stimulus. Although selectivity and sparsity can be related at their average values, they are not necessarily correlated: Selective neurons do not ensure sparse neuron coding (**Figure 1C**); similarly, sparsely activated neurons are not necessarily narrowly selective (**Figure 1D**).

Another hypothesis on how to achieve coding efficiency is dependence minimization, which can be achieved applying *independent component analysis* (ICA) (Hyvärinen and Oja, 2000). ICA is a dimensionality reduction methodology widely used in signal processing for decomposing a compound signal into their components (or so-called bases) that are as independent as possible. In ICA, independence maximization is achieved by pursuing extrema of the kurtosis (a measure of function "peakedness") of each components' distribution. Applying ICA on natural images has also produced receptive fields like those of V1 simple cells (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998). Be either ICA or sparse coding, in the end, they are two successful learning strategies that can learn primary visual cortex-like receptive fields (Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998). Another successful learning strategy at emulating the hierarchical architecture of the brain is *deep learning* (Bengio, 2009; LeCun et al., 2015), which is usually constructed with a stack of restricted Boltzmann machines (RBMs). RBMs have recently attracted increasing attention due to its successes in learning representations (Hinton, 2002; Hinton and Salakhutdinov, 2006). In RBMs, there is no connection among hidden units (**Figure 2**), which makes inference and learning of RBMs quite easy and fast. That means that given some visible data, all hidden units are conditionally independent from each other (see Section 2.2). Even so, RBMs provide a nonlinear coding of natural images, which goes beyond sparse coding or ICA. However, the capability of RBMs is still limited when learning receptive fields similar to those of V1 simple cells. When RBMs are trained on natural images, many learned features can be rather

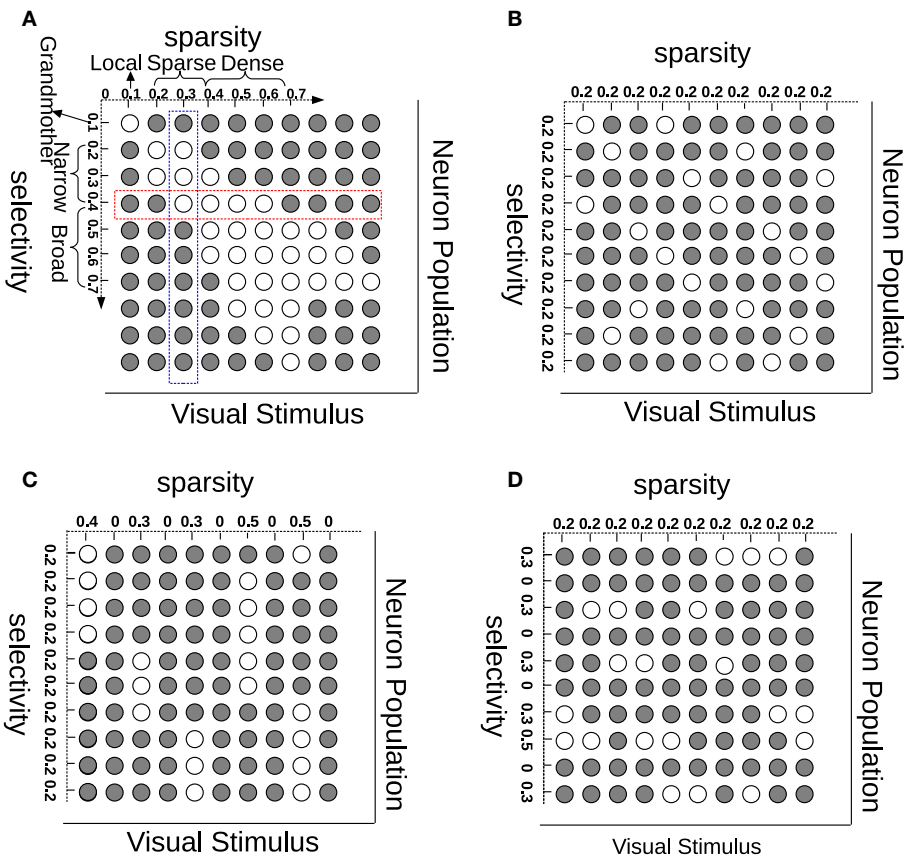

FIGURE 1 | Understanding sparsity and selectivity. White circles indicate activations while gray circles denote inactivations. **(A)** Explaining the concepts of sparsity and selectivity; **(B)** An example of good sparsity and good selectivity; **(C)** An example of good selectivity but bad sparsity; **(D)** An example of good sparsity but bad selectivity. See text for further description.

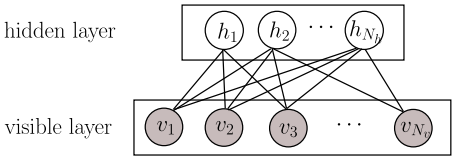

FIGURE 2 | A graphical model of a restricted Boltzmann machine (RBM). Gray circles represent observed variables while empty circles are hidden variables.

distributed, unlocalized and repeated, which is far from the (selective and sparse) nature of the learning task. Prior work has exploited different strategies to adapt RBMs toward learning selective or sparsely-activated neurons (Lee et al., 2007; Goh et al., 2010; Luo et al., 2011) on visual inputs. Meanwhile, most of those works focus on either one property, thus not ensuring sparsity and selectivity simultaneously in the resulting emulated neurons, which as mentioned before may be suboptimal for coding efficiency.

Empirically, neither sparse coding nor ICA can yield both, good selectivity and sparsity simultaneously (Willmore and Tolhurst, 2001). In this paper, we propose a novel hypothesis to interpret the selectivity and sparsity of neuron activations through the *diversity of neurons' receptive fields*. Based on the analysis exposed above, we can see that the effect of sparsity is to better differentiate neurons, while the goal of selectivity is to avoid "over-tolerant" neurons, thus both aimed at reducing ambiguities. We propose that, in order to reach bot—high degrees of neural population sparsity and individual neuronal selectivity—we need one condition: diverse receptive fields. To the best of our knowledge, the diversity of receptive fields (features) has rarely been exploited to guide learning, even though it has been achieved unintentionally in several existing models. By contrast to conventional models, we use diversity as a starting point instead of as a result. An earlier pioneering work focusing on the importance of diversity in neural coding was presented by Padmanabhan and Urban (2010).

We argue that selectivity and sparsity of neurons' activations can be seen as two epiphenomena of the diversity of receptive fields. To verify this hypothesis, we impose a *diversity prior* on the inter-weights within the RBMs when learning simple neurons' receptive fields from natural images. This prior will introduce a bias over the inter-weights toward higher degrees of *sum similarity minimization*. The prior indirectly coordinates neurons' activations by diversifying the inter-weights within the RBMs, which would mimic the effect of inhibition. It is worth noting that the prior is only employed in the learning phase, yet its implicit effect on coordinating neurons' activations will remain after learning. In this sense, the diversity prior is in line with the influence of inhibitory interneurons (King et al., 2013) (see Section 2.3 for more details). It should be finally noted that we do not consider an RBM (even if trained with diversity priors) as a full biologically-plausible model of V1 simple cells, since we are not considering many other aspects and properties of simple cells, e.g., contrast normalization, contrast adaption, etc. The purpose of our study is to verify and advocate for using *diversity* as a new principle in order to guide the learning of more similar primary visual cortex cell receptive fields.

2. Materials and Methods

In this section, we describe our basic experimental setup, which includes the construction of visual stimulus data, the restricted Boltzmann machine (RBM), and the proposed prior for training. For the RBM, a brief introduction of the model and its probabilistic properties is provided in Section 2.2. Readers are referred to Hinton (2002) for a more detailed and deeper study.

2.1. Images

The benchmark database from Olshausen and Field (1996)[1] was used in this paper. This database consists of 10 natural images, which were preprocessed with a pseudo-whitening filter, which flattens the spectrum of natural images by rescaling Fourier coefficients. This step is commonly applied (Olshausen and Field, 1996; Willmore and Tolhurst, 2001), and to some extent is similar to retinal processing. Alternatively, a similar preprocessing function is the log transform, which is more often used in ICA (van Hateren and van der Schaaf, 1998). Then, 100,000 small patches (size 14×14) were extracted from random positions of the 10 whitened images. Furthermore, a sigmoid function was applied to the pixel intensities to fit their values into the range [0, 1]. In addition, the patches with variances smaller than 0.1 were filtered out in order to accelerate training.

2.2. Restricted Boltzmann Machines

The restricted Boltzmann machine (RBM) is a two-layer, bipartite Markov network, which is a "restricted version" of the Boltzmann machine with only inter-connections between a hidden layer and a visible layer. RBMs have been recently rather popular in constructing deep neural networks (DNNs) (Hinton and Salakhutdinov, 2006). A graphical model of an RBM is presented in **Figure 2**. Input data is binary and N_v dimensional; they are fed into N_v units in the visible layer $\mathbf{v}$. The N_h units in the hidden layer $\mathbf{h}$ are stochastic binary variables, i.e., $\mathbf{v} \in \{0, 1\}^{N_v}$, $\mathbf{h} \in \{0, 1\}^{N_h}$. The joint probability of $\{\mathbf{v}, \mathbf{h}\}$ is:

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathbf{Z}} \exp(-E(\mathbf{v}, \mathbf{h})) \qquad E(\mathbf{v}, \mathbf{h}) = -\mathbf{v}^\top \mathbf{W}\mathbf{h} - \mathbf{h}^\top \mathbf{b} - \mathbf{v}^\top \mathbf{c} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{N_v \times N_h}$ is the matrix of symmetric weights, $\mathbf{b} \in \mathbb{R}^{N_h \times 1}$ and $\mathbf{c} \in \mathbb{R}^{N_v \times 1}$ are biases for hidden units and visible units, respectively. $\mathbf{Z} = \sum_{\mathbf{v},\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$ is the partition function for normalization. In our experiment, to fit the size of small image patches, N_v is equivalent to 196, and N_h is 200, i.e., 200 hidden units. Because of the restricted connections in RBMs, hidden units h_j are conditionally independent of each other given the visible data $\mathbf{v}$,

$$p(\mathbf{h}|\mathbf{v}) = \prod_j p(h_j|\mathbf{v}) \qquad p(h_j = 1|\mathbf{v}) = \mathcal{S}(\mathbf{v}^\top \mathbf{W}_{\cdot j} + b_j) \tag{2}$$

[1] Available on http://redwood.berkeley.edu/bruno/sparsenet/.

and similarly, visible units v_i are conditionally independent of each other given $\mathbf{h}$.

$$p(\mathbf{v}|\mathbf{h}) = \prod_i p(v_i|\mathbf{h}) \qquad p(v_i = 1|\mathbf{h}) = \mathcal{S}(\mathbf{W}_{i\cdot}\mathbf{h} + c_i) \quad (3)$$

where $\mathbf{W}_{i\cdot}$ and $\mathbf{W}_{\cdot j}$ denote the ith row and jth column of matrix $\mathbf{W}$, b_j and c_i are the jth and ith entry of vector $\mathbf{b}$ and $\mathbf{c}$, respectively. $\mathcal{S}(\cdot)$ is the logistic function $\mathcal{S}(x) = \frac{1}{1+\exp(-x)}$. Given training data $\mathcal{D} = \{\mathbf{v}^{(l)}\}_{l=1}^{L}$, an RBM can be learned by maximizing the average log-likelihood of $\mathcal{D}$:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \mathcal{L}(\mathcal{D}) = \arg\max_{\mathbf{W}} \frac{1}{L}\sum_{l=1}^{L}\left(\log \sum_{\mathbf{h}} p(\mathbf{v}^{(l)}, \mathbf{h})\right) \quad (4)$$

Since the log-likelihood is concave with respect to $\mathbf{W}$, $\mathbf{b}$, $\mathbf{c}$ (Koller and Friedman, 2009, Chapter 20), based on Equation (1), *gradient ascent* can be applied on Equation (4) by computing the gradient of $\mathcal{L}(\mathcal{D})$ with respect to $\mathbf{W}$, $\mathbf{b}$, $\mathbf{c}$ as:

$$\nabla_{\mathbf{W}}\mathcal{L}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbb{E}_{\mathbf{v}^{(l)}\in\mathcal{D},\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}^{(l)})}(\mathbf{v}^{(l)}\mathbf{h}^\top) - \mathbb{E}_{\mathbf{v},\mathbf{h}\sim p(\mathbf{v},\mathbf{h})}(\mathbf{v}\mathbf{h}^\top)\right] \quad (5)$$

$$\nabla_{\mathbf{b}}\mathcal{L}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}^{(l)})}(\mathbf{h}) - \mathbb{E}_{\mathbf{h}\sim p(\mathbf{v},\mathbf{h})}(\mathbf{h})\right] \quad (6)$$

$$\nabla_{\mathbf{c}}\mathcal{L}(\rfloor) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbb{E}_{\mathbf{v}^{(l)}\in\mathcal{D}}(\mathbf{v}^{(l)}) - \mathbb{E}_{\mathbf{v}\sim p(\mathbf{v},\mathbf{h})}(\mathbf{v})\right] \quad (7)$$

where $\mathbb{E}_p(\cdot)$ denotes the expected values with respect to p. Obviously, the first terms in Equations (5–7) are easy to compute with $\mathbf{v}^{(l)}$ from $\mathcal{D}$ and $\mathbf{h}$ inferred using Equation (2). However, the sampling $\mathbf{v}, \mathbf{h} \sim p(\mathbf{v}, \mathbf{h})$ in the second term of Equation (5) makes learning practically infeasible because it requires a large number of Markov chain Monte Carlo (MCMC) iterations to reach equilibrium. Fortunately, we can compute an efficient approximation to the exact gradient: contrastive divergence (CD), which works well in practice (Hinton and Salakhutdinov, 2006). By using CD_k, only a small number of k steps are run in block Gibbs sampling (usually $k = 1$), and Equation (5) can finally be approximated as

$$\nabla_{\mathbf{W}}\hat{\mathcal{L}}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbf{v}^{(l)}p(\mathbf{h}^{(l)+}|\mathbf{v}^{(l)})^\top - p(\mathbf{v}^{(l)-}|\mathbf{h}^{(l)+})\, p(\mathbf{h}^{(l)-}|\mathbf{v}^{(l)-})^\top\right] \quad (8)$$

$$\nabla_{\mathbf{b}}\hat{\mathcal{L}}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[p(\mathbf{h}^{(l)+}|\mathbf{v}^{(l)}) - p(\mathbf{h}^{(l)-}|\mathbf{v}^{(l)-})\right] \quad (9)$$

$$\nabla_{\mathbf{c}}\hat{\mathcal{L}}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbf{v}^{(l)} - p(\mathbf{v}^{(l)-}|\mathbf{h}^{(l)+})\right] \quad (10)$$

where $\mathbf{h}^{(l)+}$ denotes the inferred hidden vector from the lth observed data point $\mathbf{v}^{(l)}$ (using Equation 2), and $\mathbf{v}^{(l)-}$, $\mathbf{h}^{(l)-}$ are vectors after one-step block Gibbs sampling (using Equations 2, 3 and again Equation 2).

2.3. Imposing a Diversity Prior

In RBMs, columns of $\mathbf{W}$ are basis images, with which $\mathbf{v}$ can be reconstructed from $\mathbf{h}$. To some extent, they can also represent neurons' receptive fields. To this end, a natural choice of biasing parameters is to diversify the columns of $\mathbf{W}$ as much as possible. The way in which we approach diversification is minimizing *square cosine similarities* among columns of $\mathbf{W}$:

$$\arg\min_{\mathbf{W}} \sum_{j=1}^{N_h}\sum_{k\neq j}^{N_h}\left\|\frac{\mathbf{W}_{\cdot,j}^\top\mathbf{W}_{\cdot,k}}{||\mathbf{W}_{\cdot,j}||||\mathbf{W}_{\cdot,k}||}\right\|^2 \quad (11)$$

Note that the denominator in Equation (11) is necessary, because eliminating it will generate many "dead" neurons. This repulsive design among $\mathbf{W}_{\cdot,j}$ was also employed in the local competition algorithm (LCA) (Rozell et al., 2008). Zylberberg et al. (2011) also found that inhibition between two neurons are proportional to the similarity (measured by the vector dot product) between their receptive fields. Here, in order to gain a more clear understanding on how the diversity prior can replicate the effect of neural inhibition, an illustrating example is presented in **Figure 3**. In particular, for computing the gradient with respect to $\mathbf{W}$, Equation (8) needs to infer the activations of the hidden units. The prior, which can bias the columns of $\mathbf{W}$ toward a more diverse population will indirectly coordinate the activations by suppressing the emergence of similar receptive fields, and therefore leads to a similar effect neural inhibition has during learning. Also, the effect from the prior will remain after learning with the learned diverse $\mathbf{W}$. An extreme case is that the activation probabilities of neurons are exclusive to each other. Sparsity and selectivity are expected to be enhanced simultaneously by using this diversity-induced bias (Equation 11) (**Figure 1B**). We can define the prior probability distribution over parameters $p(\mathbf{W})$ as

$$p(\mathbf{W}) \propto \exp\left(-\lambda\cdot\sum_{j=1}^{N_h}\sum_{k\neq j}^{N_h}\left\|\frac{\mathbf{W}_{\cdot,j}^\top\mathbf{W}_{\cdot,k}}{||\mathbf{W}_{\cdot,j}||||\mathbf{W}_{\cdot,k}||}\right\|^2\right). \quad (12)$$

Then, the parameters can be estimated via maximum a posteriori (MAP):

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} p(\mathbf{W}|\mathcal{D}) = \arg\max_{\mathbf{W}} p(\mathbf{W})\prod_{l=1}^{L}\sum_{\mathbf{h}} p(\mathbf{v}^{(l)}, \mathbf{h}|\mathbf{W}) \quad (13)$$

In our previous work (Xiong et al., 2014), we used absolute cosine similarities, of which the derivative cannot be analytically computed and therefore we had to resort to MCMC-based simulated annealing to conduct MAP. However, here by using the square cosine similarity, Equation (13) can be converted to a constrained concave optimization:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \mathcal{L}(\mathcal{D}) - \lambda\sum_{j=1}^{N_h}\sum_{k\neq j}^{N_h}(\mathbf{W}_{\cdot,j}^\top\mathbf{W}_{\cdot,k})^2 \quad (14)$$

$$s.t. \qquad \forall j \in [1, N_h],\ ||\mathbf{W}_{\cdot,j}|| = 1$$

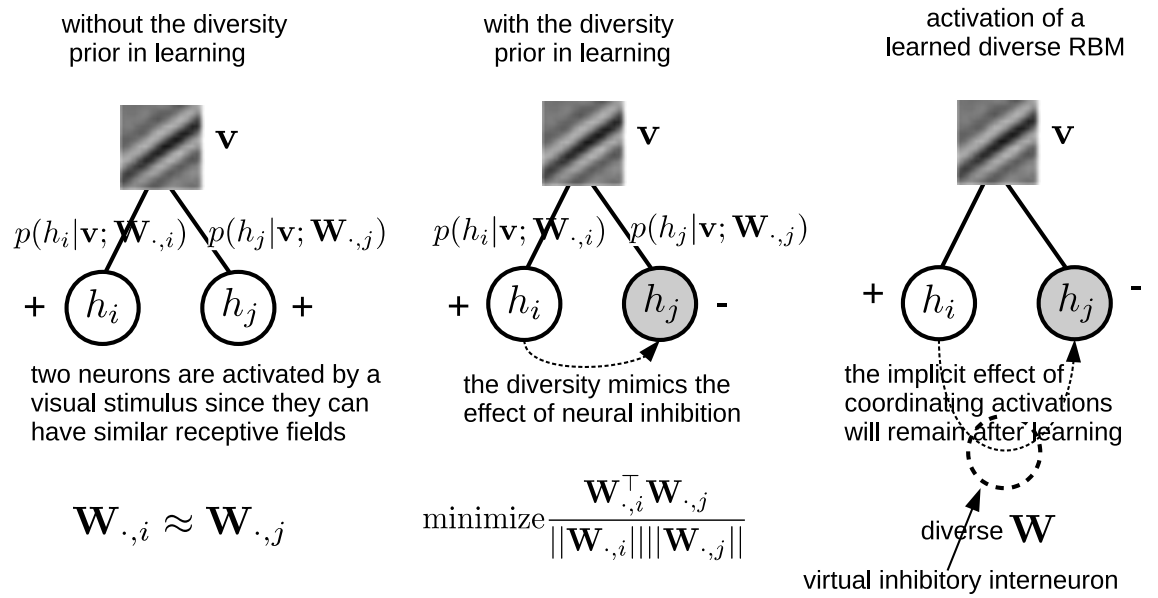

FIGURE 3 | Left, Middle: An illustrating example shows how the diversity prior would mimic the effect of inhibition among neurons' activations during learning. Empty circles denote activated neurons while gray circles are inactivated ones. **Right**: although the diversity prior is only employed in the learning phase, its implicit effect on coordinating neurons' activations will remain after learning. Intuitively, it can be considered as if there would exist virtual inhibitory interneurons which are induced by the diversity prior.

In this paper, since the above optimization problem is concave with respect to $\mathbf{W}$, we employed gradient ascent to solve it (see the Appendix for details), and derived an iterative update of $\mathbf{W}$ as

$$\mathbf{W}_{\cdot,j}^{t+1} = \mathbf{W}_{\cdot,j}^{t} + \nabla_{\mathbf{W}}\hat{\mathcal{L}}(\mathcal{D}) - 2\lambda\left(\sum_{k\neq j}^{N_h}(\mathbf{W}_{\cdot,k}\otimes\mathbf{W}_{\cdot,k}) + C\frac{||\mathbf{W}_{\cdot,j}||-1}{||\mathbf{W}_{\cdot,j}||}\mathbf{I}_{N_v}\right)\mathbf{W}_{\cdot,j} \quad (15)$$

where $\otimes$ denotes the outer product between vectors, and $\mathbf{I}_{N_v}$ is a $N_v \times N_v$ identity matrix. In Equation (15) we can see that the iterative update of $\mathbf{W}$ is composed of two parts, where the first is the gradient of the log-likelihood while the second is the gradient of the log prior.

3. Results

In this section the learned receptive fields are shown, with which we measure the selectivity and sparsity of neurons' activations. We also compare the learned receptive fields with physiological data. Finally, we test the learned receptive fields in an image reconstruction experiment. The training dataset, the code of learning RBM, the learned diverse RBM and other materials used in our experiments are available at: https://iis.uibk.ac.at/public/xiong/resources.html#Diverse_RBM. Following Hinton (2002), we conducted training on mini-batches at one epoch. In all 400 epochs were run and it takes around 18 h with our Matlab code on an Intel core i7 laptop.

3.1. Basis Images

In **Figure 4**, a subset of basis images (i.e., columns of $\mathbf{W}$) of RBMs trained with the diversity prior are shown. They look quite similar to the receptive fields of simple cells in macaque monkey V1 (Zylberberg et al., 2011, **Figure 3**). Rigorously speaking, basis images cannot be directly considered as receptive fields since they are internal connections or representations instead of response characteristics. The receptive fields of ICA are usually estimated as the inverse of the weight matrix (van Hateren and van der Schaaf, 1998), while in sparse coding reverse correlation is used for receptive fields (Olshausen and Field, 1996). Here, we employed a reverse correlation method similar to Hosoya (2012) who also developed a probabilistic model. For each hidden unit, its receptive field is estimated as

$$RF = \sum_{s=1}^{S} p(h_j = 1|\mathbf{v}_s)\mathbf{v}_s, \quad (16)$$

where $p(h_j = 1|\mathbf{v}_s)$ is computed as in Equation (2), while $\{\mathbf{v}_s\}_{s=1}^{S}$ is a set of visual stimuli which are randomly selected in the training database. This is a little different from the procedure by Hosoya (2012), since they generate synthetic $\mathbf{v}_s$ from a Gaussian distribution. Meanwhile, we arrived at a finding similar to Hosoya (2012). By linearly fitting the *RF* of each unit to its corresponding basis images, we found that our basis images are almost identical to their corresponding receptive fields.

3.2. Selectivity and Sparsity

There exist several ways to measure selectivity and sparsity, out of which *kurtosis* and *Treves-Rolls sparseness* are popularly used

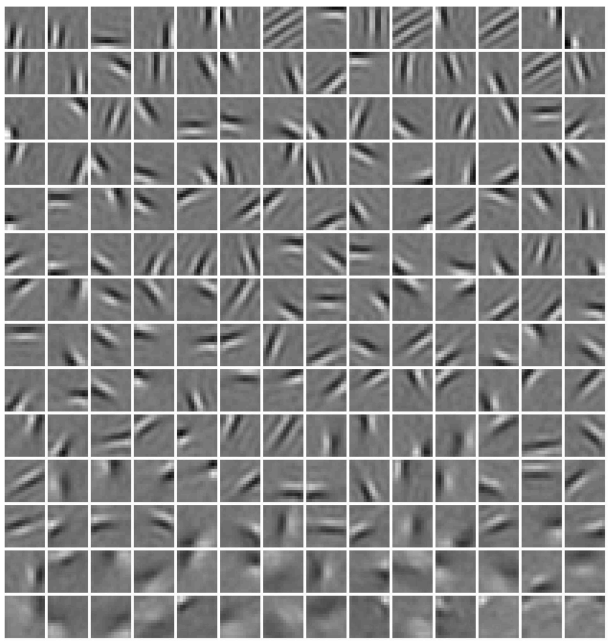

FIGURE 4 | Basis images (columns of W) learned in our model. They can be also considered as receptive fields since we found that they are almost identical.

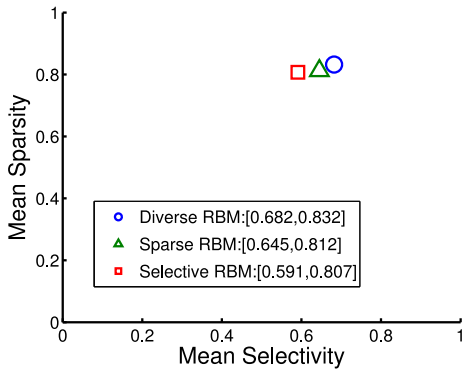

FIGURE 5 | Mean sparsity and mean selectivity of neurons' activations in diverse RBM, sparse RBM and selective RBM, respectively.

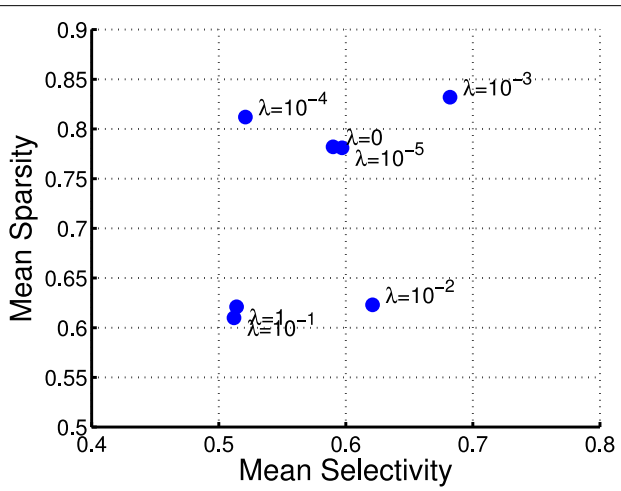

FIGURE 6 | Selectivities and sparsities when using different λ-values in the diverse RBM.

(Willmore and Tolhurst, 2001). Willmore and Tolhurst (2001) empirically proved that there exists a high correlation between these two measures. In other words, there would be no difference in using these two measures to quantify neurons' activations. Here, we use Treves-Rolls sparseness.

For a neuron, its selectivity is computed across all L input visual stimuli:

$$selectivity = 1 - \frac{(\sum_{l=1}^{L} r_l/L)^2}{(\sum_{l=1}^{L} r_l^2/L)} \tag{17}$$

where r_l is the activation probability of the neuron given the lth stimulus, computed as in Equation (2).

The sparsity of population activations by one stimulus is computed across all N_h neurons:

$$sparsity = 1 - \frac{(\sum_{j=1}^{N_h} r_j/N_h)^2}{(\sum_{j=1}^{N_h} r_j^2/N_h)} \tag{18}$$

where r_j denotes the activation probability of the jth neuron by the stimulus. We computed the mean selectivity of all 200 neurons and the mean sparsity on all training small patches. The results are plotted in **Figure 5**. Two relevant models (selective RBM and sparse RBM, see Section 4.1) were tested as well for comparison. It can be seen that using the diversity prior in learning can result in comparable selectivity and sparsity as using selectivity prior or sparse prior. Meanwhile, the diversity prior should be preferred since it generates a much smaller number of "dead" neurons (see Section 4.1). In our experiment, $\lambda = 10^{-3}$ was used to obtain the above result. To check how λ value affects sparsity and selectivity, in **Figure 6** a plot with several λ is presented. When λ is small, e.g., $0, 10^{-5}, 10^{-4}$, the effect of the diversity prior is weak or totally removed and both selectivity and sparsity decrease (**Figure 6**). The receptive fields of a diverse RBM trained with $\lambda = 10^{-5}$ are shown in **Figure 7A**. If we use a big value of λ, e.g., $10^{-2}, 10^{-1}, 1$, the iterative update of $\mathbf{W}$ Equation (15) is greatly dominated by the prior part, and therefore the fitness to the training data $\mathcal{D}$ deteriorates. It can be seen that selectivity and sparsity also decrease (even to a larger degree) using relatively large λs (**Figure 6**). The receptive fields learned with $\lambda = 1$ are displayed in **Figure 7B**.

3.3. Comparison with Biological Data

To better compare our receptive fields against physiological results (Ringach, 2002), we first fitted our receptive fields to Gabor filters:

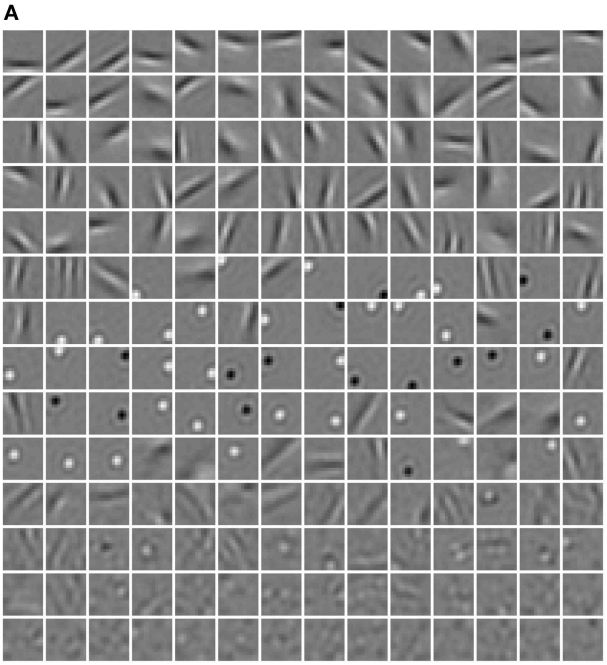

FIGURE 7 | The receptive fields learned using (A) $\lambda = 10^{-5}$, (B) $\lambda = 1$.

$$G(x, y; x_0, y_0, A, \sigma_x, \sigma_y, \theta, f, \phi) = A\cos(2\pi f x' + \phi) \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right)$$

$$x' = (x - x_0)\cos\theta + (y - y_0)\sin\theta$$

$$y' = -(x - x_0)\sin\theta + (y - y_0)\cos\theta \quad (19)$$

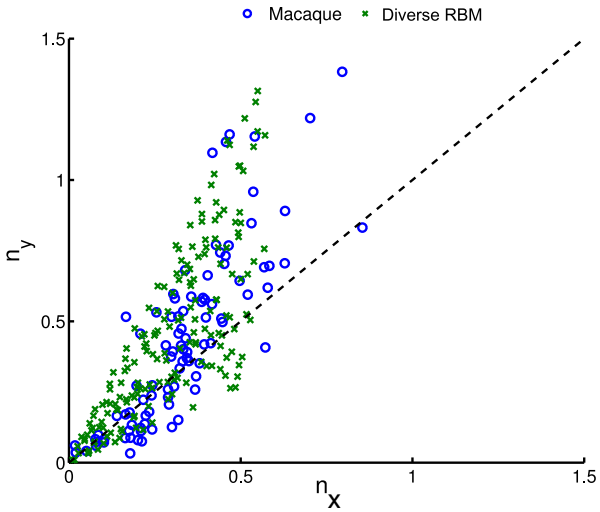

FIGURE 8 | After fitting receptive fields with Gabor filters, we pooled their shape profiles (n_x, n_y), for comparison to physiological data of macaque monkeys (Ringach, 2002).

whose parameters are the center position (x_0, y_0), amplitude A, size (σ_x, σ_y), orientation θ, spatial frequency f and phase ϕ. The fitting is done via the *Nelder-Mead Simplex* method, and therefore is not very reliable. Similar to Hosoya (2012) and Zylberberg and DeWeese (2013), we conducted quality control by filtering out some receptive fields which were poorly fitted. First, we compared our receptive fields with those of macaque monkey V1 cells[2] (Ringach, 2002) in units of the sinusoidal wavelength: $(n_x, n_y) = (\sigma_x f, \sigma_y f)$. In **Figure 8**, we pooled (n_x, n_y) of our receptive fields as well as the data from Zylberberg and DeWeese (2013). We found that they don't deviate very much although they slightly differ from each other. We also checked the statistics of aspect ratios within receptive fields: $\frac{n_y}{n_x}$. In **Figure 9** two histograms are displayed, which are global distributions of aspect ratios from our receptive fields and from the macaque monkey V1 cells, respectively. We can see that they are also quite close.

3.4. Image Reconstruction

Reconstruction using RBMs is quite straightforward. First, small, non-overlapping patches (size 14×14) were extracted from a preprocessed image. For each small patch $\mathbf{v}$, the activation probability of each neuron $p(h_j|\mathbf{v})$ can be computed as in Equation (2). Then, instead of using binary states of h_j, $p(h_j|\mathbf{v})$ is used for recovering $\mathbf{v}$ by using Equation (3). It is worth noting that although RBMs are probabilistic models, we use the value of $p(v_i|\mathbf{h})$ to recover the intensity of each pixel and thus the reconstruction is deterministic.

Out of the 10 images in the original database, 8 were used for training and the remaining 2 were used for testing the image reconstruction. The two test images were whitened and sigmoid-mapped using the same preprocessing procedure as the training images. They are shown in the left panel of **Figure 10**. In the right

[2]Data are available at: http://www.ringachlab.net/lab/Data.html.

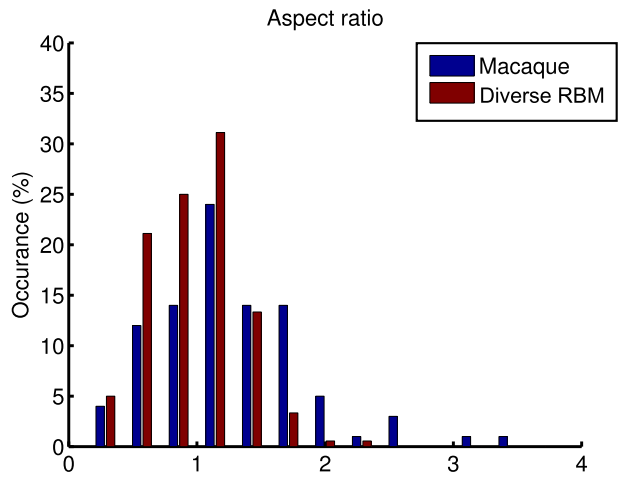

FIGURE 9 | A comparison of the histograms of aspect ratios (n_x/n_y) within macaque monkey V1 neurons and the learned diverse RBM neurons.

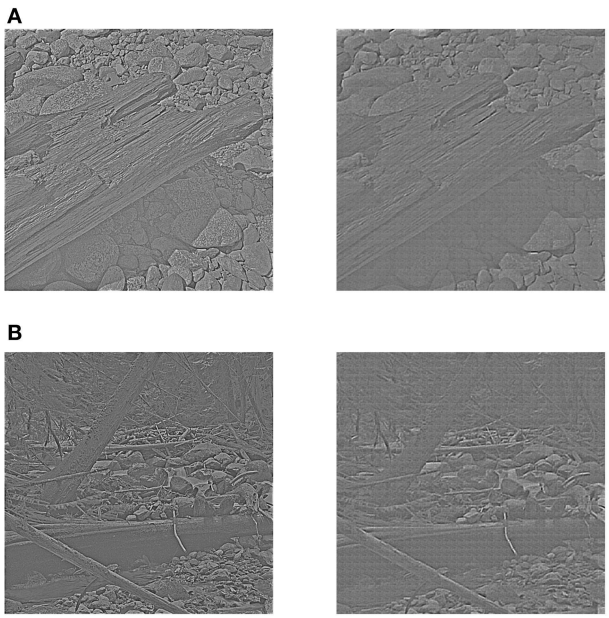

FIGURE 10 | Reconstruction using receptive fields of the learned diverse RBM. The left panel of **(A,B)** shows two test images after preprocessing, while the right panel of **(A,B)** shows two corresponding reconstructions.

panel of **Figure 10** the reconstructions of the two test images are presented. It can be seen that the reconstructions look very good in qualitative terms.

4. Discussion

4.1. Sparsity and Selectivity Prior on RBM

There are previous studies that learn simple cell receptive fields through the use of RBMs, either enforcing sparsity or selectivity. One recent example of the former is the sparse group restricted Boltzmann machine (SGRBM) (Luo et al., 2011), an RBM trained with the CD algorithm plus an $l1/l2$ norm regularization on the activations of the neuron population. At each iteration, given a visual stimulus, and after computing the activation probabilities of the whole neuron set, SGRBM attempts to minimize the $l1/l2$ norm of the set of activation probabilities. Although $l1/l2$ norm regularization can ensure sparsity, it can also lead to many "dead" (never responding) and "potential over-tolerant" (always responding) neurons (see **Figure 1D**). In the case of the latter, a study that enforces selectivity is the one from Lee et al. (2007) which uses a selectivity-induced regularization that suppresses the average activation probability of each neuron to all training stimuli.

One limitation of this strategy, as argued by Goh et al. (2010), is that decreasing average activation probabilities cannot guarantee selectivity. Instead, it will result in many similar neurons with uniformly low activation probabilities to all types of visual stimuli, which are prone to be "dead" as well. Following this line of thought and in order to prove the validity of our diverse RBM, two additional RBMs were trained using the CD algorithm with sparse regularization (sparse RBM) (Luo et al., 2011) and the CD algorithm with selectivity regularization (selective RBM) (Lee et al., 2007). For both of them, 200 hidden neurons were learned and their receptive fields are presented in **Figure 11**. We can see that the neurons' receptive fields learned in sparse RBM and selective RBM look similar to those of our RBM trained with a diversity prior. However, both sparse CD and selective CD led to many useless, "dead" neurons. We estimated the rough number of "dead" neurons by counting the number of neurons whose maximal activation probabilities to all training stimuli is smaller than 0.1, and the results are shown in **Figure 12**. Furthermore, we also computed the mean selectivity and the mean sparsity of neurons in sparse RBM and selective RBM in the same way as we did for the diverse RBM; their results are also shown in **Figure 5**.

4.2. The Equivalent to a Diversity Prior in Biological Systems

Knowing about how neuron receptive field properties arise is of great importance in visual neuroscience in order to hypothesize the circuits and connections that give rise to those properties. On one hand, one of the characteristics of simple cells in V1 is selectivity to oriented stimuli. These can be obtained through placing some constraint in learning from natural images. An example is the influential work by Olshausen and Field (1996). A set of coefficients is then formed such that they have a cost associated to them depending on how the activity is distributed. The aim is to increase sparsity, meaning lower cost. This approach leads to V1-like simple-cell receptive fields through the learning of a set of weights that correspond to the connections of the input layer with simple neurons in area V1.

On the other hand, inhibition seems to play a central role in the shaping of simple-cell receptive fields. We can consider three types of inhibitory inputs: feedforward, lateral and feedback (also known as recurrent). Feedforward inhibition is regarded as the main source of orientation selectivity in simple cells

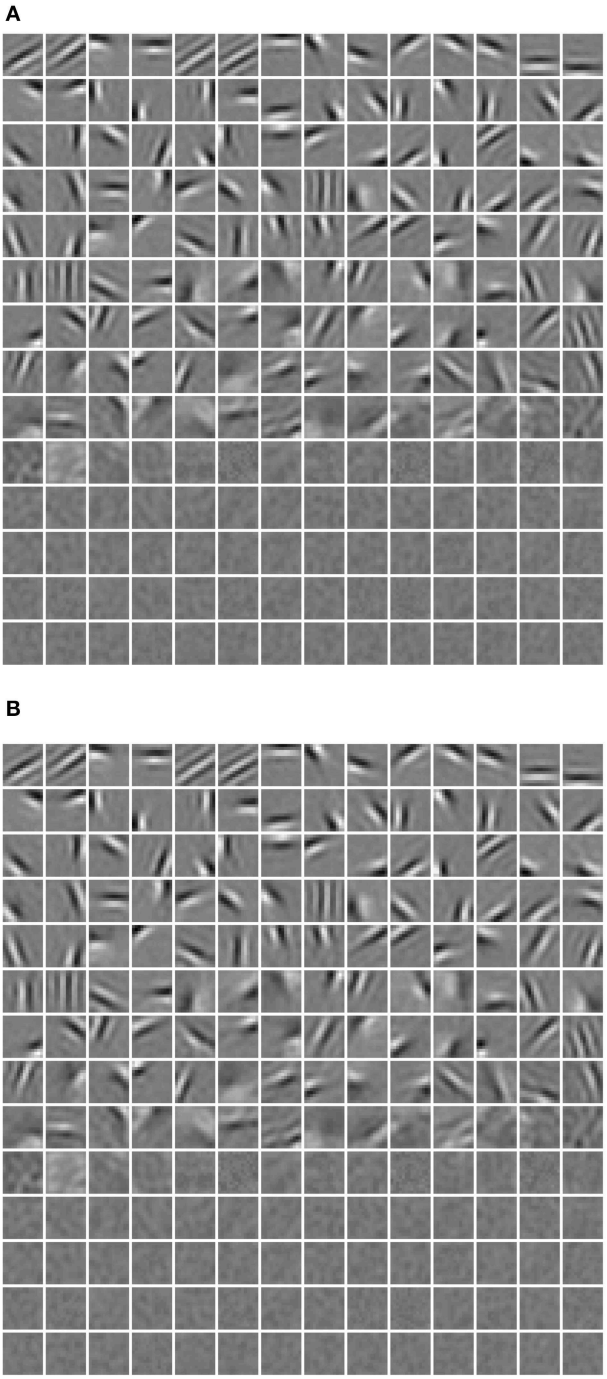

FIGURE 11 | The receptive fields learned in the learned (A) sparse RBM, (B) selective RBM.

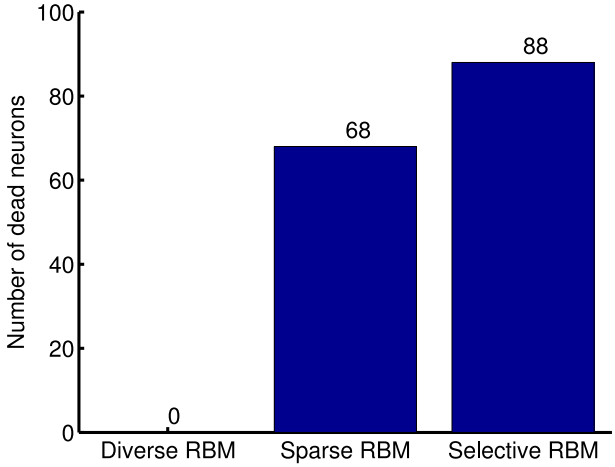

FIGURE 12 | Number of dead neurons in the learned diverse RBM, sparse RBM and selective RBM, respectively.

by some researchers (Heggelund, 1981; Celebrini et al., 1993; Ferster and Miller, 2000) and has been modeled by others, e.g., (Azzopardi et al., 2014). The classical role of feedback connections was the enhancement of receptive-field responses to top-down modulations (Ito and Gilbert, 1999; Treue, 2003), which have been successfully modeled for attention (Rodriguez-Sanchez et al., 2007) and contour integration (Neumann and Sepp, 1999; Tschechne and Neumann, 2014). But other studies are in support of feedback connections as the source of simple-cell selectivity through recurrent connections, most recently from Angelucci and Bressloff (2006). The appearance of orientation selectivity this way has also been proposed in models of recurrent inhibition, e.g., (Sabatini, 1996; Carandini and Ringach, 1997). Finally, even though there is an alive discussion regarding if orientation selection is achieved through feedforward or recurrent connections, it is interesting to note that none of them rule out that lateral inhibition can at least be partially blamed for this selectivity, e.g., (Celebrini et al., 1993; Angelucci and Bressloff, 2006). Lateral connections have in fact being made explicit into recent sparse coding models (Garrigues and Olshausen, 2008; King et al., 2013).

The common ground of all the aforementioned works is that inhibition is fundamental to the selectivity properties of simple cells, irrespective of where that inhibition comes from. Inhibition is also linked to the appearance of sparse sensory coding (Vinje and Gallant, 2000; Haider et al., 2010). We can conclude then, that inhibition would generate RF diversity, since as we have shown in this work (**Figure 1**), imposing diversity generates both selective and sparse neural populations. By explicitly favoring diversity in our model, we would be mimicking the effect that inhibition should have on feature learning in a biological system.

5. Conclusion

We test a recent new concept, that of diversity (Padmanabhan and Urban, 2010; O'Donnell and Nolan, 2011), by applying diversification on the columns of **W** when using a RBM to learn receptive fields. This diversification has the implication of providing a set of neurons that is at the same time sparse and selective, which, as mentioned in the introduction, is not always the case for sparse models. Imposing diversity is thus a more general condition to achieve both, sparsity and selectivity.

References

Angelucci, A., and Bressloff, P. C. (2006). Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. *Progr. Brain Res.* 154, 93–120. doi: 10.1016/S0079-6123(06)54005-1

Azzopardi, G., Rodríguez-Sánchez, A., Piater, J., and Petkov, N. (2014). A push-pull corf model of a simple cell with antiphase inhibition improves snr and contour detection. *PLoS ONE* 9:e98424. doi: 10.1371/journal.pone.0098424

Barlow, H. B. (1989). Unsupervised learning. *Neural Comput.* 1, 295–311.

Bell, A., and Sejnowski, T. (1997). The "independent components" of natural scenes are edge filters. *Vis. Res.* 37, 3327–3338.

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Berkes, P., White, B., and Fiser, J. (2009). "No evidence for active sparsification in the visual cortex," in *Neural Information Processing Systems (NIPS)* (Montreal, QC), 108–116.

Carandini, M., and Ringach, D. L. (1997). Predictions of a recurrent model of orientation selectivity. *Vis. Res.* 37, 3061–3071.

Celebrini, S., Thorpe, S., Trotter, Y., and Imbert, M. (1993). Dynamics of orientation coding in area v1 of the awake primate. *Vis. Neurosci.* 10, 811–825.

Ferster, D., and Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annu. Rev. Neurosci.* 23, 441–471. doi: 10.1146/annurev.neuro.23.1.441

Field, D. (1994). What is the goal of sensory coding? *Neural Comput.* 6, 559–601.

Garrigues, P., and Olshausen, B. A. (2008). "Learning horizontal connections in a sparse coding model of natural images," in *Neural Information Processing Systems (NIPS)* (Vancouver, BC), 505–512.

Goh, H., Thome, N., and Cord, M. (2010). "Biasing restricted Boltzmann machines to manipulate latent selectivity and sparsity," in *Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning* (Vancouver, BC).

Haider, B., Krause, M. R., Duque, A., Yu, Y., Touryan, J., Mazer, J. A., et al. (2010). Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron* 65, 107–121. doi: 10.1016/j.neuron.2009.12.005

Heggelund, P. (1981). Receptive field organization of simple cells in cat striate cortex. *Exp. Brain Res.* 42, 89–98.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 1771–1800. doi: 10.1162/089976602760128018

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hosoya, H. (2012). Multinomial bayesian learning for modeling classical and nonclassical receptive field properties. *Neural Comput.* 24, 2119–2150. doi: 10.1162/NECO_a_00310

HUBEL, D., and WIESEL, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw.* 13, 411–430. doi: 10.1016/S0893-6080(00)00026-5

Ito, M., and Gilbert, C. D. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron* 22, 593–604.

Jones, E. (2007). Neuroanatomy: cajal and after cajal. *Brain Res. Rev.* 55, 248–255. doi: 10.1016/j.brainresrev.2007.06.001

King, P. D., Zylberberg, J., and DeWeese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1. *J. Neurosci.* 33, 5475–5485. doi: 10.1523/JNEUROSCI.4188-12.2013

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, H., Ekanadham, C., and Ng, A. Y. (2007). "Sparse deep belief net model for visual area v2," in *Neural Information Processing Systems (NIPS)*, eds J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Vancouver, BC: Curran Associates, Inc.), 873–880.

Luo, H., Shen, R., Niu, C., and Ullrich, C. (2011). "Sparse group restricted boltzmann machines," in *AAAI Conference on Artificial Intelligence (AAAI)* (Granada).

Neumann, H., and Sepp, W. (1999). Recurrent v1–v2 interaction in early visual boundary processing. *Biol. Cybern.* 81, 425–444.

O'Donnell, C., and Nolan, M. F. (2011). Tuning of synaptic responses: an organizing principle for optimization of neural circuits. *Trends Neurosci.* 34, 51–60. doi: 10.1016/j.tins.2010.10.003

Olshausen, B., and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609.

Padmanabhan, K., and Urban, N. N. (2010). Intrinsic biophysical diversity decorrelates neuronal firing while increasing information content. *Nat. Neurosci.* 13, 1276–1282. doi: 10.1038/nn.2630

Ramón y Cajal, S. (1888). Sobre las fibras nerviosas de la capa molecular del cerebelo. *Rev. Trim. Histol. Norm. Patol.* 1, 33–49.

Ramón y Cajal, S. (1904). Variaciones morfologicas, normales y patologicas del reticulo neurofibrilar. *Trab. Lab. Investig. Biol. Madrid* 3, 9–15.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* 88, 455–463.

Rodriguez-Sanchez, A. J., Simine, E., and Tsotsos, J. K. (2007). Attention and visual search. *Int. J. Neural Syst.* 17, 275–288. doi: 10.1142/S0129065707001135

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.* 20, 2526–2563. doi: 10.1162/neco.2008.03-07-486

Sabatini, S. P. (1996). Recurrent inhibition and clustered connectivity as a basis for gabor-like receptive fields in the visual cortex. *Biol. Cybern.* 74, 189–202.

Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Curr. Opin. Neurobiol.* 13, 428–432. doi: 10.1016/S0959-4388(03)00105-3

Tschechne, S., and Neumann, H. (2014). Hierarchical representation of shapes in visual cortex from localized features to figural shape segregation. *Front. Comput. Neurosci.* 8:93. doi: 10.3389/fncom.2014.00093

van Hateren, J. H., and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. B. Biol. Sci.* 265, 359–366.

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Willmore, B., and Tolhurst, D. (2001). Characterising the sparseness of neural codes. *Netw. Comput. Neural Syst.* 12, 255–270. doi: 10.1080/713663277

Xiong, H., Szedmak, S., Rodríguez-Sánchez, A., and Piater, J. (2014). "Towards sparsity and selectivity: bayesian learning of restricted boltzmann machine for early visual features," in *24th International Conference on Artificial Neural Networks (ICANN14)* (Hamburg).

Zylberberg, J., and DeWeese, M. R. (2013). Sparse coding models can exhibit decreasing sparseness while learning sparse codes for natural images. *PLoS Comput. Biol.* 9:e1003182. doi: 10.1371/journal.pcbi.1003182

Zylberberg, J., Murphy, J. T., and DeWeese, M. R. (2011). A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS Comput. Biol.* 7:e1002250. doi: 10.1371/journal.pcbi.1002250

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix

A1. MAP optimization with the Sum Similarity Minimization Prior

The optimization problem (Equation 14) can be rewritten as

$$\max_{\mathbf{W}} \mathcal{L}(\mathcal{D}) - \lambda \underbrace{\left[\sum_{j=1}^{N_h} \sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,j}^{\top} \mathbf{W}_{\cdot,k})^2 + C \sum_{j=1}^{N_h} (||\mathbf{W}_{\cdot,j}|| - 1)^2 \right]}_{O}, \tag{A1}$$

where C is an extra parameter that is set relatively large to guarantee the satisfaction of the constraints in Equation (14). In our experiment, C is equivalent to 10^4. In this way, the constrained optimization problem is converted to an unconstrained one. It was already shown that the gradient ascent can used to maximize $\mathcal{L}(\mathcal{D})$. It is easy to see that O is also convex with respect to $\mathbf{W}$; therefore, the same gradient ascent can be also applied on $-\lambda O$. The gradient of O with respect to $\mathbf{W}$ can be computed as

$$\frac{\partial O}{\partial \mathbf{W}_{\cdot,j}} = 2 \sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,j}^{\top} \mathbf{W}_{\cdot,k}) \mathbf{W}_{\cdot,k} + 2C(||\mathbf{W}_{\cdot,j}|| - 1) \frac{\mathbf{W}_{\cdot,j}}{||\mathbf{W}_{\cdot,j}||} \tag{A2}$$

$$= 2 \sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,k} \otimes \mathbf{W}_{\cdot,k}) \mathbf{W}_{\cdot,j} + 2C(||\mathbf{W}_{\cdot,j}|| - 1) \frac{\mathbf{W}_{\cdot,j}}{||\mathbf{W}_{\cdot,j}||} \tag{A3}$$

$$= 2 \left(\sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,k} \otimes \mathbf{W}_{\cdot,k}) + C \frac{||\mathbf{W}_{\cdot,j}|| - 1}{||\mathbf{W}_{\cdot,j}||} \mathbf{I}_{N_v} \right) \mathbf{W}_{\cdot,j}, \tag{A4}$$

where $\otimes$ denotes the outer product between vectors and $\mathbf{I}_{N_v}$ is a $N_v \times N_v$ identity matrix.