Neural Networks 144 (2021) 210–228

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue

Performance of biologically grounded models of the early visual system on standard object recognition tasks

Michael Teichmann[1], René Larisch[1], Fred H. Hamker[*]

Chemnitz University of Technology, Str. der Nationen, 62, 09111, Chemnitz, Germany

ARTICLE INFO

Article history:
Available online 20 August 2021

Keywords:
Brain-inspired neural networks
Spiking neural networks
Hebbian learning
Spike timing-dependent plasticity
Object recognition

ABSTRACT

Computational neuroscience models of vision and neural network models for object recognition are often framed by different research agendas. Computational neuroscience mainly aims at replicating experimental data, while (artificial) neural networks target high performance on classification tasks. However, we propose that models of vision should be validated on object recognition tasks. At some point, mechanisms of realistic neuro-computational models of the visual cortex have to convince in object recognition as well.

In order to foster this idea, we report the recognition accuracy for two different neuro-computational models of the visual cortex on several object recognition datasets. The models were trained using unsupervised Hebbian learning rules on natural scene inputs for the emergence of receptive fields comparable to their biological counterpart. We assume that the emerged receptive fields result in a general codebook of features, which should be applicable to a variety of visual scenes.

We report the performances on datasets with different levels of difficulty, ranging from the simple MNIST to the more complex CIFAR-10 or ETH-80. We found that both networks show good results on simple digit recognition, comparable with previously published biologically plausible models. We also observed that our deeper layer neurons provide for naturalistic datasets a better recognition codebook. As for most datasets, recognition results of biologically grounded models are not available yet, our results provide a broad basis of performance values to compare methodologically similar models.

1. Introduction

The human visual system can identify objects in a visual scenery fast and with high accuracy. To explain the functions and process of its high performance in object recognition, different biologically motivated neural networks in the area of computational neuroscience have been developed. In recent years, neural networks using different unsupervised Hebbian learning rules have been able to reproduce a wide range of phenomena observed in the visual cortex, such as the emergence of simple-cell receptive fields (Brito & Gerstner, 2016; Carlson et al., 2013; Clopath et al., 2010; Gupta & Garg, 2011; Harpur & Prager, 1996; King et al., 2013; Larisch et al., 2020; Miconi et al., 2016; Wiltschut & Hamker, 2009; Zylberberg et al., 2011) or complex-cell receptive fields (Masquelier et al., 2007; Rolls, 2012; Spratling, 2005; Teichmann et al., 2012), the influence of inhibition and lateral connectivity on orientation selectivity (Banitt et al., 2007; Larisch et al., 2020; Miconi et al., 2016; Palmer & Miller, 2007; Sadeh et al., 2015) and the representation efficiency of the neural code (King et al., 2013; Larisch et al., 2020).

Despite these successes, such studies are often limited to explore only a specific phenomenon while neglecting its influence on the ability to recognize objects. Hence, many of these models are by design unable to perform object recognition at a useful level (e.g. Banitt et al., 2007; Buchs & Senn, 2002; Clopath et al., 2010; Gupta & Garg, 2011; Palmer & Miller, 2007; Sadeh et al., 2015), which raises concerns to what extend the described mechanisms match their biological counterpart. A biologically realistic model of the visual cortex must at some point also be accurate in recognizing objects to count as a convincing replica of the brain. For this reason, object recognition performance should be considered as an additional metric for evaluating models of the visual cortex, e.g. V1 simple-cell learning, to demonstrate the quality of the presented methods in this important aspect of coding.

In recent years, an increasing number of biologically motivated neural networks have been tested on established datasets for object recognition, as on hand written digits (MNIST) (e.g. Diehl & Cook, 2015; Illing et al., 2019; Kheradpisheh et al., 2018; Larisch et al., 2018; Spratling, 2017; Tavanaei & Maida, 2017), but only

* Corresponding author.
E-mail addresses: michael.teichmann@informatik.tu-chemnitz.de (M. Teichmann), rene.larisch@informatik.tu-chemnitz.de (R. Larisch), fred.hamker@informatik.tu-chemnitz.de (F.H. Hamker).

[1] Equal contribution.

https://doi.org/10.1016/j.neunet.2021.08.009
0893-6080/

a few on more difficult datasets like CIFAR-10 (e.g. Illing et al., 2019; Panda & Roy, 2016). This lack of comparable biologically grounded models on datasets besides of MNIST, makes the interpretation of the obtained recognition accuracy values difficult. Often, the performance of biologically grounded models is compared to those of deep neural networks, which originate from the field of computer vision. Despite the fact that these deep neural networks have some biological inspiration, they are trained often in a supervised manner (Lecun et al., 1998), e.g., by means of the backpropagation algorithm, which is considered not being the type of learning occurring in the brain (Bengio et al., 2016; Whittington & Bogacz, 2019). Further, deep networks are mainly designed to succeed in the object recognition task while the visual system likely solves multiple tasks at the same time (Kandel & Schwartz, 1995). As a consequence, these deep neural networks are poor references for approaches which aim to model the visual system with a higher degree of realism. This lack elucidates the need of a broader basis of biologically grounded models for comparison. Although the focus of this manuscript is on biologically motivated unsupervised learning approaches, it has to be mentioned that other biologically motivated approaches, such as spike-based reinforcement learning (e.g. Mozafari et al., 2019, 2018), are worth further investigation.

To improve the interpretability of the performance of biological grounded network models and to encourage other researchers to report their model performances, we evaluate the recognition accuracy for two models of the visual cortex on several object recognition datasets. Both neural network models employ biologically inspired Hebbian learning rules but differ in some fundamental aspects. One is rate-based and implements Hebbian synaptic plasticity, intrinsic plasticity, and structural plasticity. It implements the recurrent connectivity between the layers 4 and 2/3 of the visual areas V1 and V2 and includes feedback from V2 to V1. Further, it uses a trace learning approach to obtain complex-cell properties. The second network is spike-based and stacks two similar layers with recurrent connections within these layers, but not between layers, and utilizes two biologically motivated spike timing-dependent plasticity (STDP) rules. Both networks have been trained on natural scenes and not on the evaluation datasets used in the present study. This is done to account for the development of receptive fields comparable to those found in macaque monkey (Olshausen & Field, 1996). Thus, the results can be understood as a transfer of the learning from the natural scene domain to the specific evaluation dataset domains. This is an uncommon procedure in the machine learning domain, but it is mandatory for one of the main research goals for computational neuroscience models: gaining insights in the processing that takes place in the regarded brain areas as typically explored by electrophysiological recordings. Due to this, we have not changed our model training, or any other parameters, to improve our results in object recognition. Instead, we assume that an appropriate learning from naturalistic input should lead to a general codebook of features, which should be applicable on any visual scene (cf. Serre et al., 2007), at least at the level of low- and mid-complex features.

To address several issues on object recognition and cover different levels of difficulty, we have chosen a variety of datasets to evaluate the model performances. These span from digit or character recognition, to complex images of real world objects in different resolutions. This provides a larger base for comparison for other researchers and to comment on the suitability of the evaluation datasets to assess the representation efficiency of the model neurons. The presented evaluation datasets are the well-established and basic MNIST dataset (Lecun et al., 1998); the larger, and characters including, balanced set of extended MNIST (EMNIST Cohen et al., 2017); and the real world street view house number dataset SVHN (Netzer et al., 2011). Moreover, the well-established object recognition dataset CIFAR-10 (Krizhevsky, 2009), with its tiny resolution real world images; the larger resolution ETH-80 (Leibe & Schiele, 2003), where photographs of unseen toy figures have to be recognized from different viewpoints; and the face/motorbike subset from Caltech 101 (Li Fei-Fei et al., 2004), which contains real world photographs in larger resolution and has been recently used for evaluation by a related publication (Kheradpisheh et al., 2018).

2. Methods

In the following, we describe the investigated rate-based and spike-based model, giving an overview of the network architectures, training protocols and the preprocessing of the object recognition datasets. A more detailed description of the models can be found in Appendix. Further details can be found for the rate-based model in Teichmann (2018) and for the spike-based model in Larisch et al. (2020).

2.1. Rate-based model

The model has been developed to resemble the processing of early parts of the ventral visual stream, which in particular includes the development of receptive fields comparable to those in macaque monkeys, i.e. simple-cell, complex-cell, and V2-like receptive fields. We achieve this by the combination of three important cortical plasticity mechanisms. (1) Hebbian and anti-Hebbian synaptic plasticity to learn the synapse strengths of excitatory and inhibitory neurons, including trace learning (see Teichmann et al., 2012) to learn invariant representations. (2) Intrinsic plasticity to regulate the neural responses and stabilize the learning in deeper layers (Teichmann & Hamker, 2015). (3) Structural plasticity to modify the connectivity and to overcome the initial bias resulting from the wiring of connections by the designer (Teichmann, 2018). An exhaustive description of the model mechanisms can be found in Teichmann (2018). The model presented here differs in the use of a larger input layer, which causes different numbers of neurons also in the subsequent layers, slightly modified homeostatic mechanisms and a slightly modified inhibitory learning rule. The network has been simulated with the neurosimulator ANNarchy (v4), which can simulate rate, as well as spike-based neural network models on parallel hardware[2] (Vitay et al., 2015).

In contrast to the previous implementation (Teichmann, 2018), we increased the input layer size to $32 \times 32 \times 2$. The first two dimensions (x,y) refer to the image geometry and the third dimension refers to the responses of two different input neuron types, namely the on- and off-center cells, which have been found in the retina and the succeeding thalamus nucleus.

The firing of the input layer neurons is determined through a preprocessing step on the input images, with which we obtain the on- and off-center cell firing rates from the image. This is done by image whitening, proposed by Olshausen and Field (1997), which utilizes a filter in the frequency domain. In the spatial domain this filter can be understood as an on- or off-center filter. The method resembles the processing step of neurons in the lateral geniculate nucleus (LGN) of the thalamus, which is the major input source to layer-4 of the primary visual cortex (V1), the first cortical area processing visual stimuli. The resulting values from the preprocessing determine the firing rates of the corresponding input neurons.

The network consists of four layers, containing a population of excitatory and inhibitory neurons. Note, that here the term

[2] https://bitbucket.org/ANNarchy.

layer refers to a typical joint excitatory and inhibitory layer in the cortex. Other researchers count each neuron population as separate layer (e.g. Saunders et al., 2019), following that scheme we use a nine layer network. The network implements a rich neocortical-like connectivity between the neurons. The populations have been connected in a retinotopic fashion, which means that initially neighboring neurons are connected and more distant neurons are not connected, i.e. the neurons are locally connected. Note, that structural plasticity changes the connectivity based on the weight development during learning Examples for the change of the initial connections to the one after learning for the first layer receptive fields are shown in S1c. We based the decision to connect neurons from two populations on anatomical data. The connectivity on the population level is illustrated in Fig. 1a. The major connection paths of the model are the feedforward connections to excitatory and inhibitory neurons of a layer emanating from the excitatory neurons of the previous layer. Inhibitory and excitatory neurons within a layer are recurrently connected, as well as neurons within an inhibitory population. Further, the second layer of V2 (V2-L2/3) is recurrently connected to the second layer of V1 (V1-L2/3). This feedback connection is intended to investigate (in future simulations) the influence of feedback on the processing and should allow in future model versions the modulation of the neuronal activity by cognitive feedback, e.g. evoked through visual attention. The network organization and mechanisms have not been selected or tuned to improve the object recognition performance in the following tasks. The amount of neurons and the geometry of the layers can be found in Table 1a.

Due to the rich network structure, the most important circuits motifs are implemented, where each motif contributes differently to the network processing. The particular contribution of the single motifs is subject of ongoing research. Thus, we give just a brief overview. The feedforward path leads to the development of Gabor-like orientation selective neurons, the so called simple-cells, in layer 4 of V1 and determines the hierarchical structure of the layers, where the neurons in the subsequent layers become selective to increasingly complex visual features (DiCarlo & Cox, 2007; Kobatake & Tanaka, 1994). As presented in earlier studies, the lateral inhibitory path leads to decorrelated neuronal activity in the targeted neurons (Földiák, 1990; Larisch et al., 2020; Teichmann, 2018; Wiltschut & Hamker, 2009), which causes an improvement in the diversity of their orientation preference (Földiák, 1990; Larisch et al., 2020) and increases the robustness against perturbations in the input (Kermani Kolankeh et al., 2015; Larisch et al., 2018). The excitatory feedback path modulates population activity in earlier areas (Maunsell, 2015), which is important for different cortical mechanisms such as visual attention (Beuth, 2019).

As mentioned, the network employs a variety of different plasticity mechanisms. Most important for learning receptive fields are the excitatory and inhibitory plasticity. The excitatory plasticity is implemented by a covariance learning rule combined with an Oja normalization (L2-normalization) (see Teichmann, 2018). The inhibitory plasticity follows an anti-Hebbian learning rule, which tunes the inhibitory weights with respect to the correlation between the neurons. This rule originates in the rule proposed by Teichmann (2018) and Wiltschut and Hamker (2009) but additionally corrects the baseline of the weights to achieve zero weights for uncorrelated neurons. The implementation is similar to King et al. (2013), who developed the rule independently.

To achieve the development of realistic receptive fields and response properties, the network has been trained on randomly chosen patches taken from natural scenes. For examples of the first layer receptive fields see S1b. The natural scenes (Fig. 1c) have been taken from Olshausen and Field (1996). Olshausen and Field (1996) and others demonstrated that receptive fields comparable to macaque monkeys can be obtained when training a network on this kind of natural scene images. For training, we presented 200,000 patches to the network. The presentation of a randomly selected patch followed a protocol mimicking fixational eye movements. This means, a patch is selected and the next nine patches are drawn from the neighborhood of its predecessor (for further details please see Teichmann (2018)). Each patch is presented for 100 ms, a bit shorter than fixational eye movements to save computation time. This presentation protocol implements a time series of slightly changed images, which is required to enable the network to exploit temporal correlations for learning invariant representations, which has been demonstrated to be beneficial for learning complex-cell like properties (Teichmann et al., 2012).

2.2. Spike-based model

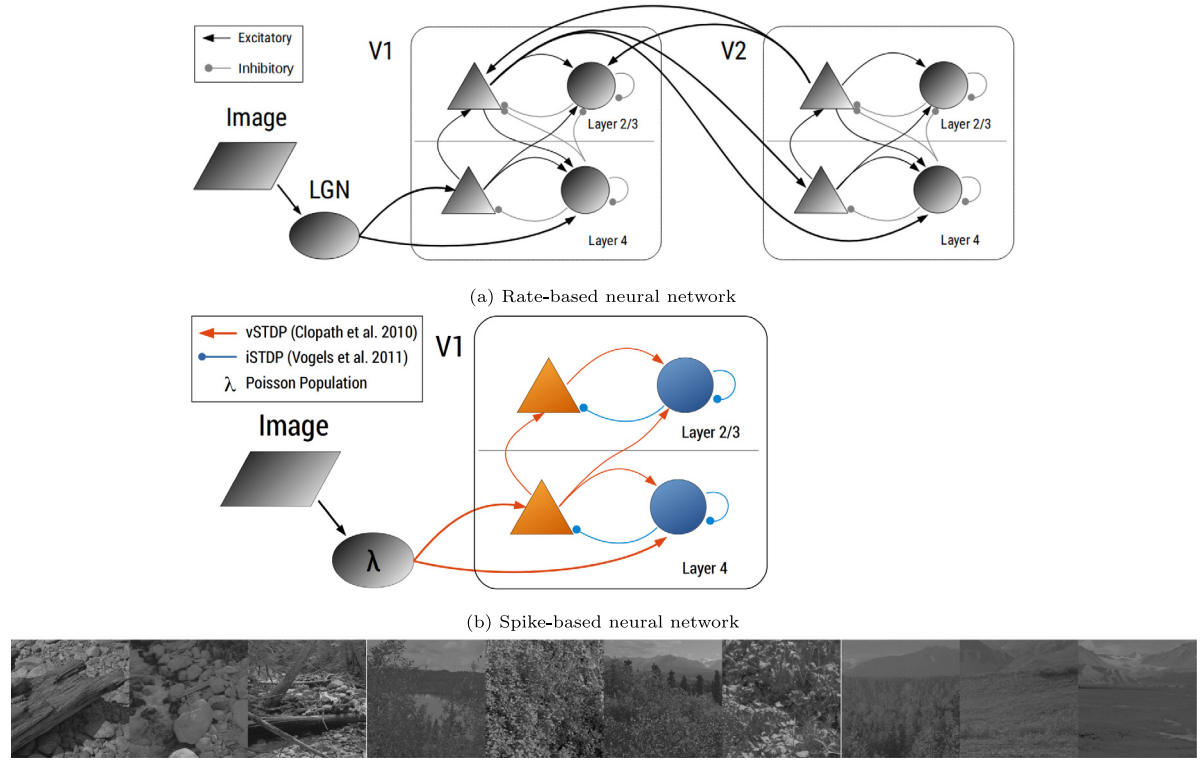
This model has been developed to investigate the interplay between excitatory and inhibitory plasticity by using detailed, phenomenological, STDP learning rules (Larisch et al., 2020). The self-organization by these plasticity rules leads to the development of an efficient stimulus encoding and phenomena like contrast-invariant orientation tuning of the neurons through emerging inhibitory gain-control. Similar to the rate-based model, this model simulates the cortical processing via differential equations over time and has been simulated with the neurosimulator ANNarchy (v4).

The spike-based model consists of two layers (five populations) which in turn consist of an excitatory and inhibitory neuron population. The model extends the previously proposed single layer architecture of Larisch et al. (2018) by an additional layer of excitatory and inhibitory neurons (Fig. 1b).

The input layer of the network consists of $18 \times 18 \times 2$ spiking neurons, where the first two dimensions represent the input geometry (x,y) and the third dimension refers to the responses of two different input neuron types, namely the on- and off-center neurons. The firing of the input layer neurons has been determined similarly to the rate-based model by applying whitening on the input images to obtain on- and off-center cell responses. From the obtained firing rates Poisson distributed spike events are generated, representing thalamic input from the lateral geniculate nucleus (LGN).

The input layer is connected to all neurons of the excitatory and inhibitory population of the first layer. Thus, inhibitory neurons receive feedforward excitation. Excitatory and inhibitory connections within a layer are recurrently connected, respecting Dale's law which states that excitatory neurons form only excitatory synapses and inhibitory neurons form only inhibitory synapses. Additionally, inhibitory neurons have connections to other inhibitory neurons of the same population. A layer consists of four times more excitatory neurons than inhibitory neurons, following findings in the visual cortex. The second layer uses the same connection scheme. The number of neurons in each population is given in Table 1b. Neurons are connected to all neurons of the connected populations. Due to a very similar network structure to the rate-based network, similar circuit motifs are implemented in the spiking-based network, except of the excitatory feedback path and the not implemented layers.

The model employs a set of phenomenologically grounded spike timing-dependent plasticity (STDP) rules. All excitatory synapses follow the voltage-based triplet STDP rule proposed by Clopath et al. (2010), for a detailed reimplementation and code see Larisch (2019). This rule has been demonstrated to reproduce phenomena like triplet STDP, different spiking burst experiments, and different voltage clamp experiments. It has also been used to



(a) Rate-based neural network

(b) Spike-based neural network

(c) Natural scene images

Fig. 1. Architectures of the proposed neural networks. (a) Areas, layers, and neuron populations of the rate-based neural network. The arrows indicate whether a connection between a neuron population exists or not. Triangles indicate an excitatory and circles an inhibitory neuron population. (b) Spike-based neural network architecture. Red triangles indicate excitatory populations and blue circles inhibitory populations. Red arrows indicate excitatory connections and blue arrows indicate inhibitory connections. (c) The natural scene images (Olshausen & Field, 1996) used for network training. Each image has a resolution of 512 × 512 pixels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Geometry and number of neurons in each layer of (a) the rate-based neural network and (b) the spike-based neural network.

(a) Rate-based neural network				(b) Spike-based neural network			
Layer	Type	Geometry	Neurons	Layer	Type	Geometry	Neurons
LGN	Input	32 × 32 × 2	2048	LGN	Input	18 × 18 × 2	648
V1 - L4	Exc	23 × 23 × 4	2116	V1 - L4	Exc	–	324
V1 - L4	Inh	23 × 23 × 1	529	V1 - L4	Inh	–	81
V1 - L2/3	Exc	17 × 17 × 7	2023	V1 - L2/3	Exc	–	324
V1 - L2/3	Inh	17 × 17 × 2	578	V1 - L2/3	Inh	–	81
V2 - L4	Exc	13 × 13 × 10	1690				
V2 - L4	Inh	13 × 13 × 3	507				
V2 - L2/3	Exc	10 × 10 × 16	1600				
V2 - L2/3	Inh	10 × 10 × 5	500				

demonstrate how the connectivity between neurons is influenced by the spike repetition frequency and spike order, as well as the development of V1 simple-cell receptive fields by presenting natural scenes. The inhibitory synapses use the symmetric STDP rule from Vogels et al. (2011) to implement a competition between the neurons. This causes decorrelated activity patterns and the development of an improved representation efficiency (Larisch et al., 2020).

Again, to achieve realistic receptive fields and response properties, the network has been trained on randomly chosen patches from natural scenes taken from Olshausen and Field (1996). Examples of the first layer receptive fields presented in S1a. The input values obtained from natural scene images are normalized by applying whitening, and setting the maximum firing rate of all scenes to 100 Hz. We presented 400,000 randomly selected natural scene patches to train the first layer. Each patch was presented for 125 ms. After that, the plasticities of the first layer are turned off and the second layer is trained on another 400,000 randomly selected natural scene patches.

2.3. Data preprocessing and performance measurement

After learning, all plasticity mechanisms of the networks were deactivated. Importantly, both neural networks have, at this time, never seen any of the images from the evaluation datasets. They developed their receptive fields only from natural scenes. We processed the images from the evaluation datasets by whitening every image similarly to the natural scene images. Further, we rescaled the resulting values so that the input evoked a sufficient network activity.

For image presentation in the test phase after learning, we reset the network activity for each patch (including the membrane potential) to zero (rate-based) or to the resting potential (spike-based), so that the neuron responses are not influenced

by previous presentations. To cope with different image sizes, we split larger images into multiple tiles having the size of our input population (see below). We recorded, for each population, the neural responses on the image tiles and concatenated these to one vector which now represents the population response on this image. The size of this response vector is thus determined by the number of neurons in the regarded population times the number of tiles into which the image was split. We used the response vectors, obtained on the training set of the regarded evaluation dataset, to train one linear support-vector machine (SVM) for each population and each evaluation dataset. The use of a linear readout is motivated by the theory about the visual system, which states that each stage of information processing untangles the visual representation and so simplifies a linear readout (DiCarlo & Cox, 2007). Using the trained SVM, we predicted the classes for the response vectors of the different neuron populations, obtained from the test images of the evaluation dataset, and calculated the accuracy score (fraction of correct predictions). The datasets ETH-80 and the face/motorbike subset of Caltech 101 have no dedicated training and test data. Thus, we randomly split the images into training and test images (cf. Kheradpisheh et al., 2018), which causes larger deviations in the problem difficulty. Due to this, we repeated the training and testing of the SVM 20 times per model instance. To get insights into how reliable our models achieve their performance, we trained 5 independent instances of our models and repeated the complete performance evaluation for each. We report the average accuracy over all model instances and classifier trainings together with its standard deviation.

For the spike-based model, we used the liblinear SVM implementation provided by the Python package scipylearn (version 0.19), with a linear kernel, a C-parameter of one (the default value), and the squared hinge loss with L2 penalty. For the rate-based model, we used the Matlab interface to liblinear (multicore version 2.11–1), with the default C-parameter of one, L2-regularized, and with L2-loss in dual form. The described SVM configuration is used for all datasets.

2.3.1. Tiling and recording for the rate-based model

Cropping or resizing are common ways to map differently sized input images on the fixed sized input layer of a neural network. As another approach, we split larger images into tiles. The networks considered here have been trained unsupervised, this means no classification layer is part of the networks, and the neurons only provide features encoding their input information. The rate-based model presented here has an input size of 32×32. The datasets having a smaller resolution, such as MNIST (28×28), are presented centered and padded with zeros to the network. The larger datasets, ETH-80 and the face/motorbike subset of Caltech 101, can be tiled into a set of 32×32 pieces covering the full image, so that we obtain multiple vectors of neuron responses encoding each tile.

We tiled ETH-80 (128×128) into 49 (7×7) patches with 16 pixel (50%) overlap between the tiles. The face/motorbike subset of Caltech 101 (160×242) has been tiled into 135 (9×15) patches, with again 16 pixel overlap. Where the 15th patch in y-dimension covers just 18 image pixels, all other pixels have been padded with zeros. To measure the firing rates of the neurons, we recorded the rate 42 ms after the input presentation starting in the LGN. All subsequent populations have been recorded considering the delay on the shortest path to the neurons, allowing all neurons to have the same time to converge. This means, we recorded the neurons in V1-L4 at 44 ms, V1-L2/3 at 46 ms, V2-L4 at 49 ms, and V2-L2/3 at 51 ms.

2.3.2. Tiling and recording for the spike-based model

The spike-based model has a limited input size of 18×18. Because of this, we also had to tile the images with a smaller resolution, i.e. MNIST, EMNIST, SVHN and CIFAR-10, into four overlapping patches. Due to this, the MNIST and EMNIST datasets were divided with an overlap of 8 pixels, and samples from the SVHN and CIFAR-10 datasets with an overlap of 4 pixels. The larger images from the ETH-80 and face/motorbike subset have been tiled as follows: Images of the ETH-80 dataset have been resized to 126×126 pixels, divisible by the input size of 18 pixels, to create 49 (7×7) non-overlapping patches from each image. For the face/motorbike subset of Caltech 101, with its asymmetrical resolution, we tiled each image into 392 (14×28) overlapping patches with an overlap of 10 pixels along the y-axis and 8 pixels along the x-axis. We cropped a 2 pixel wide margin on the right side (x-axis) of the images.

The brightness of every pixel (after whitening) is transferred into a spike train by a Poisson process, causing a random spike pattern with an average spike rate relative to the original brightness value. To measure a reliable firing rate of the neurons, we presented every patch for 125 milliseconds and saved the average firing rate for each excitatory and inhibitory neuron.

3. Results

We evaluated our two models on several common object recognition datasets, to determine the quality of their information representation. We selected datasets of different difficulty and realism to allow for a better comparison of the obtained performances. We will report the accuracy and its standard deviation for each neuron population separately (1) to allow other researchers to compare models of different depth and (2) to identify the suitability of the different representation levels for the regarded object recognition tasks. The scope of this paper is biologically motivated unsupervised approaches. However, for comparison we report accuracy values of a few other well-known methods.

3.1. MNIST

The first dataset we used for evaluation is MNIST[3] (Lecun et al., 1998). It is one of the most common small resolution datasets for object recognition. It contains 60,000 handwritten digits dedicated for training and 10,000 digits for testing (Fig. 2a). Each image has a resolution of 28×28 and is gray scaled. The images have black background, so that object recognition systems are not distracted by unrelated information. Recognition accuracies come close to 100 percent and differ just by a few percent between different approaches.

For this dataset we obtained the highest mean accuracy values when using the neural responses from the first layer excitatory neurons for classification. We achieved 98.04% and 97.95% for the rate- and spike-based model respectively (Table 2). The excitatory populations in deeper model layers slightly degrade in their classification performance. The inhibitory populations showed the same effect, but on lower values. We also measured for the rate-based model the accuracy on raw pixel values after applying whitening, i.e. the model input. Here we achieved 93.19% accuracy. We did not measure the accuracy of the spike-based model's input layer, as it is just a conversion of the whitened input into Poisson distributed spike trains.

Since MNIST is one of the most used benchmarks, an exhaustive number of models for comparison exists. As we focus on biologically grounded unsupervised approaches, we will report

[3] http://yann.lecun.com/exdb/mnist.

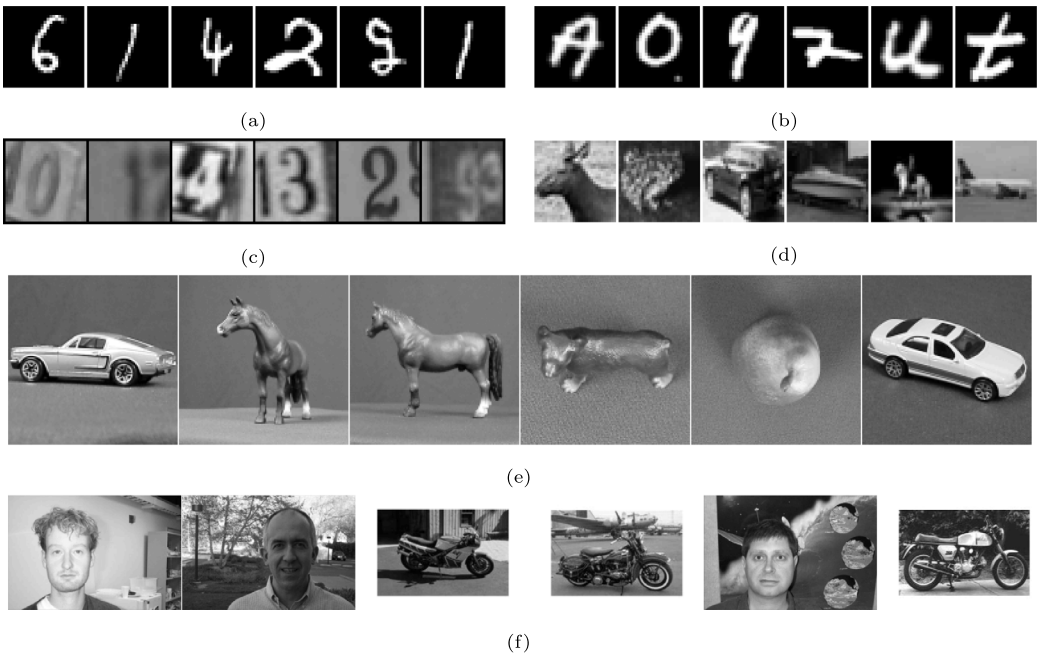


Fig. 2. Overview on the different object recognition datasets. Six example images for each of the datasets: (a) MNIST (Lecun et al., 1998), (b) Balanced EMNIST (Cohen et al., 2017), (c) SVHN (Netzer et al., 2011), (d) CIFAR-10 (Krizhevsky, 2009), (e) ETH-80 (Leibe & Schiele, 2003), and (f) the face/motorbike subset from Caltech 101 (Li Fei-Fei et al., 2004).

Table 2
Accuracy values on the MNIST dataset. (a) Accuracy values for previously published models. Ranked by their achieved accuracy. (b) Accuracy values for the proposed rate-based neural network and (c) the proposed spike-based neural network.

(a) Previously published models

Model	Learning type	Learning rule	Acc. in %
NMFSC (Kermani Kolankeh et al., 2015)	Unsupervised	NMF with sparseness constr.	91.54
ICA (Kermani Kolankeh et al., 2015)	Unsupervised	ICA	92.02
Rate-based NN (Kermani Kolankeh et al., 2015)	Unsupervised	Hebbian	92.5
Spiking RBM (Neftci et al., 2014)	Supervised	Contrastive divergence	92.6
SNN (Querlioz et al., 2013)	Unsupervised	Rectangular STDP	93.5
RBM autoencoder + SNN (Eliasmith et al., 2012)	Unsupervised	RBM/DBN + optimizing MSE	94
SNN (Diehl & Cook, 2015)	Unsupervised	Exponential STDP	95.0
SNN with STDP and R-STDP (Mozafari et al., 2019)	Reinforcement	Reward-modulated STDP	97.2
SNN (Larisch et al., 2018)	Unsupervised	Phenomenological STDP	98.08
Convolutional SNN (Tavanaei & Maida, 2017)	Unsupervised	Rectangular STDP	98.36
Deep Conv. SNN (Kheradpisheh et al., 2018)	Unsupervised	Rectangular STDP	98.4
l-RG + SNN (Illing et al., 2019)	–	Gabor-filter + supervised STDP	98.6
l-ICA (Illing et al., 2019)	Unsupervised	ICA	98.8
l-RG (Illing et al., 2019)	–	Gabor-filter	98.9
Spiking autoencoder (Panda & Roy, 2016)	Unsupervised	Backpropagation	99.05
LeNet-5 (Lecun et al., 1998)	Supervised	Backpropagation	99.1

(b) Rate-based neural network

Layer	Neuron type	Acc. ± STD in %
LGN	–	93.19 ± 0.02
V1 - L4	Exc	$\mathbf{98.04} \pm 0.08$
V1 - L4	Inh	95.75 ± 0.21
V1 - L2/3	Exc	97.52 ± 0.09
V1 - L2/3	Inh	95.51 ± 0.23
V2 - L4	Exc	96.80 ± 0.26
V2 - L4	Inh	94.10 ± 0.53
V2 - L2/3	Exc	95.20 ± 0.51
V2 - L2/3	Inh	93.08 ± 0.63

(c) Spike-based neural network

Layer	Neuron type	Acc. ± STD in %
V1 - L4	Exc	$\mathbf{97.95} \pm 0.11$
V1 - L4	Inh	95.79 ± 0.27
V1 - L2/3	Exc	97.01 ± 0.09
V1 - L2/3	Inh	92.08 ± 0.87

just a few other methods to better assess the values. A recent review of a broader range of algorithm classes can be found in Illing et al. (2019), and with focus on spiking neural networks (supervised and unsupervised) in Tavanaei et al. (2019).

Both of our networks show a good performance in comparison to related previously published unsupervised networks (Diehl & Cook, 2015; Kermani Kolankeh et al., 2015; Kheradpisheh et al., 2018; Querlioz et al., 2013; Tavanaei & Maida, 2017), while being outperformed by other backpropagation-based approaches (Lecun et al., 1998; Panda & Roy, 2016). We outperform the single-layer spiking neural network (SNN) of Diehl and Cook (2015), which implements an excitatory (6400 neurons) and inhibitory neuron population and uses simplified STDP as learning mechanism. Similarly to our spike-based network, the inputs have been converted to Poisson distributed spike trains, adding some degree of randomness to the system. In contrast to our work, it

has been trained directly on the MNIST data and the neurons received complete digits as input. This caused that the model neurons developed digit like receptive fields, similar to an earlier rate-based neural network of our group, which achieved 92.5% accuracy (288 neurons) (Kermani Kolankeh et al., 2015). The here introduced models developed in their first layer (excitatory as well as inhibitory neurons) Gabor-like receptive fields, so that the neurons behave like edge detectors. This is due to the training on natural scene images and the locality of the connections of the individual neurons (rate-based 10×10 initially, spike-based 18×18). Models utilizing localized filters on this dataset have been found superior in comparison to models working with digit like receptive fields (Illing et al., 2019). The single layer convolutional spiking neural network of Tavanaei and Maida (2017) and the three layer convolutional SNN (on–off center DoG plus two layers with rectangular STDP) of Kheradpisheh et al. (2018) have been again trained on the MNIST data. The first achieved an accuracy value of 98.36% and the latter 98.4% accuracy. Both networks perform better than our proposed networks. Illing et al. (2019) reported even higher accuracy values of 98.8% (l-ICA), when using localized receptive fields and finding the weights with independent component analysis (ICA), and 98.9% when using random Gabor-filters (l-RG) and 5000 units. In both networks, the readout from the hidden to the classification layer was learned through a supervised backpropagation algorithm. When converting the l-RG into a SNN, a spike implementation of backpropagation has been used, for which the authors reported a degraded performance of 98.6%. End-to-end trained networks with backpropagation, as used by Lecun et al. (1998), can reach performances of 99.1%, and above. They are clearly rendered superior to our approaches in terms of recognition performance, as all their network features are tuned to the recognition task. Spiking autoencoders (Panda & Roy, 2016) can achieve a rivaling result to supervised methods with a performance of 99.05%. Internally they also use backpropagation as learning principle, however, their receptive fields are tuned to achieve a minimal image reconstruction error. For the sake of completeness we also add our previous implementation of the spike-based model (single layer) (Larisch et al., 2018), which achieved 98.08% accuracy with a very similar configuration.

3.2. Balanced EMNIST

The extended MNIST (EMNIST) dataset[4] (Cohen et al., 2017) has been proposed as an enhanced version of MNIST. Just like MNIST, it is based on the same larger NIST dataset. It contains various subsets which, besides digits, also include handwritten letters. Here, we used the "Balanced" subset of EMNIST (Fig. 2b), which provides, in comparison to MNIST, a larger amount of examples (112,800 training examples and 18,800 for testing). Including letters, so that 47 classes have to be recognized. Each class contains the same, balanced, amount of training and testing images. Like MNIST, the images have a resolution of 28×28 pixels and black background. In contrast to the near perfect accuracy on MNIST, Cohen et al. (2017) reported on Balanced EMNIST just 78.02% recognition accuracy with an OPIUM-based classifier, a supervised method to learn with extreme learning machines. Thus, potential algorithms should show, in contrast to MNIST, more than a few percent difference in their performance, which should allow for a better comparison of different algorithms.

As well as for the MNIST dataset, our networks achieved their highest accuracy values in the excitatory population of V1-layer 4 (83.82% and 83.17%) and showed a performance decrease in deeper layers (Table 3). Likewise, the inhibitory populations reach lower values than the excitatory populations of the same layer. Despite that this dataset appears more valuable than MNIST to compare algorithms, there is a lack of performance values of other biologically inspired algorithms. This lack is might be caused by the novelty of the dataset.

3.3. SVHN

The images of MNIST and EMNIST are somewhat an idealized object recognition problem, since the images are clutter free and have no distracting background or other objects. A more realistic setup within the same data domain is the Street View House Numbers (SVHN) dataset (Netzer et al., 2011). This dataset[5] contains photographs of house numbers in Google Street View images (Fig. 2c). Two dataset versions are available: the original street view images in variable resolution, with bounding boxes for each digit, and a MNIST-like version, where the images are cropped and centered on the digits. We use the MNIST-like version. Its images have background and noise. Further, the target digit in the center is often surrounded by other distracting digits, which makes the recognition problem even more difficult. This might allow complex biologically grounded approaches to demonstrate advantages of their complexity, i.e. noise reduction, figure ground segregation, or attention mechanisms.

The images of the MNIST-like version have a slightly larger resolution of 32×32 pixels and are in RGB color, so we converted the images to gray scale. The task is the same as in MNIST: recognizing digits from 0 to 9 (10 classes). The dataset contains 73,257 images for training and 26,032 images for testing. A larger set of additional training data is also available, but not used.

The first excitatory population of our networks achieved the highest accuracy (75.4% and 76.26%). Deeper layers and the inhibitory populations show a drop in performance (Table 4), down to raw input accuracy, which renders the neural code and the receptive fields of these layers less suited for this dataset. Nevertheless, we outperformed simple methods like binary image features (63.3%) (Netzer et al., 2011). More sophisticated methods like k-Means (90.6%) or stacked sparse autoencoder (89.7%) clearly outperform our networks. These algorithms have also been applied on the gray scaled images, but, in contrast to us, used the additional images for training. The supervised trained DenseNet (Huang et al., 2018) achieved on colored images an even higher accuracy of 98.41.

3.4. CIFAR-10

Real world object recognition is an even more complex problem than digit or character recognition. Objects can have enormous variations, e.g. different views or different instances of the same class. One of the most common datasets for real world object recognition with small resolution is CIFAR-10[6] (Krizhevsky, 2009). It contains 60,000 RGB colored images (50,000 for training and 10,000 for testing), with a resolution of 32×32 pixels, from 10 different object categories (Fig. 2d). While the resolution is similar to MNIST, the images are photographs from real world objects in a natural environment, obtained via a web search and downscaled to its tiny resolution. The object classes are: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Again, we converted all images to gray scale. Our models achieved the highest accuracy in the first excitatory population (46.54% and 45.08%) (Table 5). In deeper layers the accuracy drops to the accuracy achieved on the input. The accuracy values of

[4] https://www.westernsydney.edu.au/bens/home/reproducible_research/emnist.

[5] http://ufldl.stanford.edu/housenumbers/.

[6] https://www.cs.toronto.edu/~kriz/cifar.html.

Table 3
Accuracy values on the Balanced EMNIST dataset. (a) Accuracy values for previous published models. (b) Accuracy values for the proposed rate-based neural network and (c) the proposed spike-based neural network.

(a) Previously published models

Model	Learning type	Learning rule	Accuracy in %
OPIUM-based (Cohen et al., 2017)	Supervised	Extreme learning machine	78.02

(b) Rate-based neural network

Layer	Neuron type	Acc. ± STD in %
LGN	–	70.02 ± 0.03
V1 - L4	Exc	**83.82** ± 0.09
V1 - L4	Inh	78.69 ± 0.43
V1 - L2/3	Exc	83.42 ± 0.16
V1 - L2/3	Inh	78.41 ± 0.51
V2 - L4	Exc	82.26 ± 0.45
V2 - L4	Inh	76.58 ± 0.94
V2 - L2/3	Exc	79.69 ± 0.63
V2 - L2/3	Inh	74.16 ± 0.97

(c) Spike-based neural network

Layer	Neuron type	Acc. ± STD in %
V1 - L4	Exc	**83.17** ± 0.15
V1 - L4	Inh	77.33 ± 0.28
V1 - L2/3	Exc	81.01 ± 0.18
V1 - L2/3	Inh	71.65 ± 2.93

Table 4
Accuracy values on the SVHN dataset. (a) Accuracy values for previously published models. (b) Accuracy values for the proposed rate-based neural network and (c) the proposed spike-based neural network.

(a) Previously published models

Model	Learning type	Learning rule	Accuracy in %
Binary features (Netzer et al., 2011)	–	–	63.3
HOG (Netzer et al., 2011)	–	–	85.0
Stacked sparse autoencoder (Netzer et al., 2011)	Unsupervised	Backpropagation	89.7
k-Means (Netzer et al., 2011)	Unsupervised	Clustering	90.6
DenseNet (Huang et al., 2018)	Supervised	Backpropagation	98.41

(b) Rate-based neural network

Layer	Neuron type	Acc. ± STD in %
LGN	–	40.78 ± 0.02
V1 - L4	Exc	**75.40** ± 0.55
V1 - L4	Inh	52.68 ± 1.17
V1 - L2/3	Exc	63.78 ± 0.83
V1 - L2/3	Inh	46.45 ± 1.52
V2 - L4	Exc	55.39 ± 0.92
V2 - L4	Inh	39.51 ± 0.62
V2 - L2/3	Exc	43.70 ± 0.75
V2 - L2/3	Inh	36.51 ± 0.51

(c) Spike-based neural network

Layer	Neuron type	Acc. ± STD in %
V1 - L4	Exc	**76.26** ± 0.31
V1 - L4	Inh	62.94 ± 1.91
V1 - L2/3	Exc	69.29 ± 1.11
V1 - L2/3	Inh	42.97 ± 2.03

Table 5
Accuracy values on the CIFAR-10 dataset. (a) Accuracy values for previously published models. (b) Accuracy values for the proposed rate-based neural network and (c) the proposed spike-based neural network.

(a) Previously published models

Model (all on colored images)	Learning type	Learning rule	Accuracy in %
Perceptron (Illing et al., 2019)	Supervised	Backpropagation	41.1
2-layer NN (Krizhevsky, 2009)	Supervised	Backpropagation	49.78
l-RG (Illing et al., 2019)	–	Gabor-filters	52.0
l-ICA (Illing et al., 2019)	Unsupervised	ICA	53.9
SIFT (Bo et al., 2010)	–	SIFT features	65.6
Spiking autoencoder (Panda & Roy, 2016)	Unsupervised	Spiking backprop.	75.42

(b) Rate-based neural network

Layer	Neuron type	Acc. ± STD in %
LGN	–	34.25 ± 0
V1 - L4	Exc	**46.54** ± 0.29
V1 - L4	Inh	39.11 ± 0.67
V1 - L2/3	Exc	43.22 ± 0.85
V1 - L2/3	Inh	37.52 ± 1.13
V2 - L4	Exc	40.81 ± 0.71
V2 - L4	Inh	35.77 ± 1.18
V2 - L2/3	Exc	37.39 ± 1.03
V2 - L2/3	Inh	34.10 ± 0.92

(c) Spike-based neural network

Layer	Neuron type	Acc. ± STD in %
V1 - L4	Exc	**45.08** ± 0.50
V1 - L4	Inh	38.92 ± 0.68
V1 - L2/3	Exc	41.23 ± 0.98
V1 - L2/3	Inh	31.92 ± 0.31

the inhibitory populations are lower than the respective excitatory population. On the colored images, a simple perceptron classifier can achieve 41.1% accuracy (Illing et al., 2019). A more sophisticated classifier, a simple 2-layer neuronal network with 1000 hidden neurons, trained with backpropagation, can achieve 49.78% accuracy (Krizhevsky, 2009). As baseline, a linear SVM on the gray scaled and whitened input images performs at 34.05% accuracy. On colored images a better performance of 53.9% can be achieved with independent component analysis (l-ICA) (Illing et al., 2019) or of 52% with random Gabor-filters (l-RG) (Illing

et al., 2019). SIFT features, a classical method from computer vision, outperformed these approaches with 65.6% recognition accuracy (Bo et al., 2010). Even better performed the tuned features of a spiking autoencoder with 75.42% (Panda & Roy, 2016), again on colored images. Other approaches using backpropagation, not listed, achieve accuracy values from 80% to above 95%. A survey of a few such networks implemented as spiking convolutional neural network can be found in Tavanaei et al. (2019).

Since we are using a transfer learning approach where both neural networks have been trained on natural scenes (consisting of landscape scenes), we suspect that our performance in deeper layers is degraded by the differences between the complex features learned from these natural scenes compared to those that are diagnostic for CIFAR-10 objects. In addition, a lack of color processing in our models makes a comparison difficult to those models trained on datasets where color is helpful. The performance on our gray scaled and whitened input is already degraded by about 7% in comparison to a perceptron classifier on the colored images. A difference similar as between our results and the results obtained with ICA (l-ICA) and random Gabor-filters (l-RG) (46.54% and 45.08% to 53.9% and 52%).

3.5. ETH-80

ETH-80[7] (Leibe & Schiele, 2003) was proposed as benchmark to measure the performance in tasks with high object variability and view-point variations (Kheradpisheh et al., 2018). The dataset contains images from eight different object classes photographed from 41 different view points. For each object class the dataset provides 10 different object instances (Fig. 2e). The original dataset images have RGB color and a resolution between 400 × 400 and 700 × 700 pixels. We used a version where all images have been cropped and resized to a resolution of 128 × 128 pixels and we converted the images to gray scale. The testing conditions aim to measure how well the algorithms generalize to new unseen objects, when trained on a few example objects. Therefore, the classifier is trained on all views of five randomly chosen instances and tested on the remaining five unseen instances.

In contrast to the trend in the other datasets, the deeper layers of the rate-based network showed improved accuracies (80.64% in V1-layer 2/3 or 80.26% in V2-layer 2/3) in comparison to the population of excitatory neurons in V1-layer 4 (78.44%) (Table 6). The spike-based model has again its best accuracy in its first excitatory population (79.31%), the excitatory neurons in the second layer show a slight decrease in their performance (78.66%).

Under the same conditions, Kheradpisheh et al. (2018) achieved with their deep convolutional spiking neural network an accuracy of 82.8%. The authors also reported results for the classical HMAX (69%), which we clearly outperform, a pre-trained AlexNet (79.5%), which has a comparable performance to our spike-based network and was exceeded by our rate-based network, and a fine-tuned AlexNet, which shows, as expected, a superior performance (96.2%). Under a different training and testing condition, Mozafari et al. (2018) achieved an accuracy of 89.5% with a reward-STDP approach. They used nine instances of each object class for the training set and set the remaining one instance in the test set.

[7] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/object-recognition-and-scene-understanding/analyzing-appearance-and-contour-based-methods-for-object-categorization/.

3.6. Caltech 101 - face and motorbike subset

A dataset allowing the comparison of the recognition performance on real world objects in a sufficiently high resolution is the final stage of our comparison. The Caltech 101 dataset[8] (Li Fei-Fei et al., 2004) provides a measure of how well algorithms generalize to unseen object instances. It consists of 101 different object classes and a background class. However, many classes contain only a few examples. Thus, we evaluated our models on the face and motorbike classes, having the most examples (435 faces, 798 motorbikes). The resolution of the dataset images are variable, thus, we rescaled all images to 160 × 242 pixels and converted all images to gray scale values (Fig. 2f). We trained the classifier on 200 randomly selected images from each class and used the remaining images as testset. This preprocessing and the training procedure is chosen similar to the one of Kheradpisheh et al. (2018) and of Mozafari et al. (2018) to allow comparison. The rate-based network achieved its best accuracy of 98.08% (Table 7) with the inhibitory population in V2-layer 4 and the spike-based network achieved 97.12% with the inhibitory population in V1-layer 4. However, the next layer excitatory population obtained a similar performance. As for the ETH-80 dataset it appears that deeper layers achieved good or even better accuracies than the first excitatory population. Also the smaller inhibitory populations show the tendency to similar performance. However, when comparing the performance of the different layers to the network input (96.39%) the improvements appear minor. This also relativizes the performance of the deep convolutional SNN reported by Kheradpisheh et al. (2018), achieving 99.1% accuracy.

3.7. Input representation

So far we applied linear separability as a measure for the information representation of the neural code. However, it could be questioned in how far the presented networks develop a neural code that discriminates well the regarded evaluation datasets by having different responses for different inputs. A method to analyze how similar or dissimilar both networks represent input samples from the same or different categories is the representational dissimilarity matrix (RDM) (Kriegeskorte et al., 2008), where the difference between the response vectors for different stimuli is quantified. We also used the t-SNE method (van der Maaten & Hinton, 2008) to map the differences in the 2D domain. A detailed description of the usage of both methods can be found in the supplementary material. We performed both analyses exemplarily on the response vectors of each population for the MNIST, balanced EMNIST, and CIFAR-10 dataset, see S2 to S19 in the supplementary material. Since the results of both networks are very similar, we report only the results of the rate-based network below.

The best performing population of the rate-based network (V1-layer 4 excitatory) shows higher dissimilarity values, in the RDM, between response vectors from different classes for the MNIST (S2a), as well as the EMNIST dataset (S6a). This indicates that samples of different classes are encoded differently in the neural representation. This is supported by the visualization via t-SNE, which shows a good separation between the response vectors (S3a and S7a). In comparison, the neural representation of the inhibitory population are more similar for inputs from different classes (RDM: S2b and 6b; t-SNE: 3b and S7b), a trend we had also observed for the accuracy values. The response dissimilarity values decrease with layer depth, suggesting a more similar representation in the deeper layers (RDM: S2, S4, S6, S8; t-SNE: S3, S5, S7, S9), again similar to the trend of the accuracy

[8] http://www.vision.caltech.edu/Image_Datasets/Caltech101.

Table 6
Accuracy values on the ETH-80 dataset. (a) Accuracy values for previously published models. (b) Accuracy values for the proposed rate-based neural network and (c) the proposed spike-based neural network.

(a) Previously published models

Model	Learning type	Learning rule	Accuracy in %
HMAX (Kheradpisheh et al., 2018)	Unsupervised	Random sampling	69.0
Pre-trained AlexNet (Kheradpisheh et al., 2018)	Supervised	Backpropagation	79.5
Deep Conv. SNN (Kheradpisheh et al., 2018)	Unsupervised	Rectangular STDP	82.8
Fine-tuned AlexNet (Kheradpisheh et al., 2018)	Supervised	Backpropagation	96.2

(b) Rate-based neural network

Layer	Neuron type	Acc. ± STD in %
LGN	–	73.38 ± 2.02
V1 - L4	Exc	78.44 ± 2.50
V1 - L4	Inh	79.50 ± 2.42
V1 - L2/3	Exc	**80.64** ± 2.20
V1 - L2/3	Inh	79.75 ± 2.22
V2 - L4	Exc	79.90 ± 2.15
V2 - L4	Inh	79.58 ± 2.11
V2 - L2/3	Exc	80.26 ± 1.92
V2 - L2/3	Inh	79.59 ± 2.08

(c) Spike-based neural network

Layer	Neuron type	Acc. ± STD in %
V1 - L4	Exc	**79.31** ± 2.13
V1 - L4	Inh	70.52 ± 2.26
V1 - L2/3	Exc	78.66 ± 2.81
V1 - L2/3	Inh	74.76 ± 1.83

Table 7
Accuracy values on the faces and motorbike subset from Caltech 101. (a) Accuracy values for previously published models. (b) Accuracy values for the proposed rate-based neural network and (c) the proposed spike-based neural network.

(a) Previously published models

Model	Learning type	Learning rule	Accuracy in %
Deep SNN with R-STDP (Mozafari et al., 2018)	Reinforcement	Reward-modulated STDP	98.9
Deep Conv. SNN (Kheradpisheh et al., 2018)	Unsupervised	Rectangular STDP	99.1

(b) Rate-based neural network

Layer	Neuron type	Acc. ± STD in %
LGN	–	96.39 ± 1.04
V1 - L4	Exc	97.34 ± 0.75
V1 - L4	Inh	97.85 ± 0.59
V1 - L2/3	Exc	97.87 ± 0.55
V1 - L2/3	Inh	97.93 ± 0.50
V2 - L4	Exc	97.93 ± 0.44
V2 - L4	Inh	**98.08** ± 0.44
V2 - L2/3	Exc	98.00 ± 0.42
V2 - L2/3	Inh	97.88 ± 0.50

(c) Spike-based neural network

Layer	Neuron type	Acc. ± STD in %
V1 - L4	Exc	96.81 ± 0.77
V1 - L4	Inh	**97.12** ± 0.75
V1 - L2/3	Exc	97.10 ± 0.77
V1 - L2/3	Inh	96.74 ± 0.69

values. In contrast to the well separable character datasets, on CIFAR-10 the t-SNE visualization did not show a good linear separability of the response vectors (S11 and S13), which could be expected given the accuracy values. This is supported by the RDM, where the dissimilarity values appear similarly distributed regardless of the class (S10 and S12). This indicates that little class specific information is present in the developed neural code of the network.

4. Discussion

Since object recognition is one of the fundamental tasks the human visual system has to solve, a realistic model of it should also convince in this task. Presently however, biologically grounded models of the visual system, aiming to simulate and to explain the cortical processing with a high degree of biological detail, rarely proved their performance in object recognition. Albeit it would be important to measure the efficiency of the achieved representation, particularly for newly proposed plasticity mechanisms. The lack of data from comparable algorithms complicates the evaluation. To encourage other researchers to measure and report their results, we evaluated the accuracy values of two neuronal network models of the early visual system on several common object recognition datasets.

Both models implement the interplay of inhibitory and excitatory neurons to allow insights in receptive field formation and neural coding. Although both models share similarities in their basic circuit and their plasticity mechanisms, they use fundamentally different concepts to achieve their results. The rate-based model utilizes three different important plasticity mechanisms and implements a complex recurrent circuity of the first two visual areas. The spike-based model uses a simpler and shallower form of this recurrent circuit, capturing the main feedforward and lateral inhibitory pathways. It focuses on plausible STDP learning mechanisms which have been demonstrated to account for many phenomena of cortical plasticity.

Both models have been trained on patches of natural scene images, with the aim to develop receptive fields comparable to those found in the visual cortex. We found it an important aspect to measure the representation efficiency of their emerging neuronal code to prove the quality of the utilized model mechanisms for learning and receptive field emergence. An underused method to evaluate the quality of receptive field formation is object recognition, a main objective of the modeled brain circuit. We assume that the emerging neural representation from the training on natural scenes provides a general codebook of image features for different object recognition tasks (cf. Serre et al., 2007). However, the constraint to train the model on natural scenes means that our models have never seen any images from the evaluation datasets during training and have not been optimized or adjusted on the object recognition tasks. Thus, the model structures are not adapted to the specific requirements of the considered object recognition tasks, for example the size of the input, color processing, differences in the frequency domain of the data to those of the natural scenes, or tuning to other highly task specific aspects.

To obtain results for the differently sized datasets, we tiled them into pieces and used the aggregated responses as scene descriptor. We presented all results for a gray scaled conversion of the evaluation datasets, which allows others to compare their results without the need to implement workarounds for color processing, which might lack biological plausibility (as processing the RGB channels separately Illing et al., 2019).

To allow a comparison to a broader range of results, we selected a range of common datasets. The range spanned from characters to objects and from idealized recognition tasks to real world tasks with distracting elements. This allows us to discover actual limits of our approaches and to report values allowing the comparison with future approaches.

Our networks achieved good accuracy values on MNIST, which turn out to be excellent in comparison to other approaches when restricting these to the same amount of neurons (cf. Diehl & Cook, 2015; Illing et al., 2019). However, it has to be questioned if MNIST, with its low differences of the accuracies between different approaches, is a suitable dataset for the evaluation of advances of biologically grounded models for visual processing. A presumably better dataset is the balanced set of EMNIST. While it is structurally similar to MNIST, it requires the recognition of far more classes due to the inclusion of letters. Our networks performed on this dataset with about $83 - 84\%$ accuracy, better than the presently available supervised approach (78%) which provides the baseline performance for this dataset (Cohen et al., 2017).

However, on these quite restricted datasets the strengths of biologically grounded deep recurrent networks, with their complex neuronal interactions, seem to play no role for the recognition performance. Our first layer neurons with their Gabor-like receptive fields achieved the best performance here. Illing et al. (2019) have shown for MNIST that with an increased amount of neurons a pool of random Gabor functions can come close to the near perfect accuracies of standard convolutional neural networks.

To further increase the difficulty of the recognition task by distracting elements, we selected the SVHN dataset, which includes distracting noise and objects. We expected this setup would potentially favor models incorporating advanced mechanisms, providing noise reduction or attention. Albeit we have shown in earlier publications that our network types are more robust to noise (Kermani Kolankeh et al., 2015; Larisch et al., 2018), we reached just a fair performance. The accuracy of our models was higher than the one of simple approaches, like binary features, but has been outperformed by standard computer vision approaches, like HOG (Netzer et al., 2011). As well as for EMNIST, no results of closely related algorithms have been available for comparison. Similar to MNIST and EMNIST, we observed on SVHN that the first layer of our networks were most appropriate for this task.

To evaluate the recognition performance on realistic objects, we tested our models on the very common CIFAR-10 dataset. We assumed that on these more naturalistic input images our learned representations provide improved performance values. Particularly when using responses of deeper layers which should have learned to represent more complex features. Somewhat surprisingly, we observed poor results in comparison to existing computer vision approaches or common deep neural networks. Deeper layer representations dropped in their performance, rendering them less suitable for this dataset. The differences in spatial resolution and content between our natural scene dataset (landscapes), used for training our networks, and the tiny resolution objects of CIFAR-10 might be part of the reason. Another part is the missing color processing, which degraded our general performance by about 7%, indicated by the difference in the performance on our gray scaled and whitened network input (linear SVM classifier) to the raw RGB images (perceptron classifier Illing et al., 2019). The implementation of chromatic processing would require different cell types at the LGN level, conversion into LMS color space, and a more complex routing through the neuron layers. Interestingly, a shallow network, using a population of random Gabor functions applied to color images, is able to achieve only about $5.5 - 7\%$ better performance (Illing et al., 2019), a similar amount to the degradation when using gray scale images and our first layer responses.

With ETH-80 we provided results for a somewhat idealized recognition task without distracting elements, but with larger resolution real world objects. This task was intended to measure object variability and view-point invariance of object recognition systems. Which is an important aspect for the processing in the visual cortex, particularly to measure the performance of the so called complex-cells. Our rate-based model was designed to account for the emergence of complex-cell properties in its second (V1-layer 2/3) and in its last layer (V2-layer 2/3). For these layers we observed an increased recognition accuracy over their predecessors, with the overall best accuracy in V1-layer 2/3. Contrarily, the spike-based model does not employ a dedicated mechanism for learning invariant representations, so it achieved its best performance in its first layer. Our models perform on ETH-80 at the level of a pre-trained AlexNet and clearly above the classical HMAX (Kheradpisheh et al., 2018), but slightly below a recently published deep convolutional spiking neural network trained with unsupervised STDP (Kheradpisheh et al., 2018).

Our final dataset was the classic Caltech 101. It contains higher resolution real world objects so that higher level representations should play a role. However, this dataset is unbalanced and we took just the two most frequent classes for evaluation, which lead to a simplistic binary recognition task. Nevertheless, the realistic objects provide a basis to evaluate the performance of the neural code in deeper network layers. Hence, we observed for the rate-based model our best performance in the deeper inhibitory population of V2-layer 4 and the second best in the even deeper excitatory population of V2-layer 2/3. The spike-based model had its best performance in the inhibitory population of the first layer, but a similar performance in the excitatory population of its second layer. However, the recognition results of all layers have been within about one standard deviation of the range of variation. As for MNIST, the recognition accuracies have been close to perfect, so that model advances could hardly be assessed with this dataset. Interestingly, throughout all datasets both presented networks performed approximately similar, despite their different architectures and learning rules.

Many biologically motivated models have been designed to explain the behavior of a localized subgroup of neurons, such as our presented spike-based model which analyzes and codes features with a limited spatial extend and full lateral interactions within the field of view of the neural population. In contrast, the presented rate-based model processes a larger fraction of the image and the neurons have spatially limited forward and lateral connectivity. This aspect might determine their capabilities for object recognition and play a crucial role to select a suitable set of algorithms to compare with. The models from Diehl and Cook (2015) and Kermani Kolankeh et al. (2015) also use unsupervised learning rules in a network with inhibitory connections to recognize images from the MNIST dataset. Both models have been trained by presenting the full digits to the neurons. As a consequence, the learned receptive fields represent full digits. Diehl and Cook (2015) reported an accuracy of 95.0%, using 6400 neurons, and Kermani Kolankeh et al. (2015) reported an accuracy of 92.5%, with 288 neurons. Similarly, all other listed approaches which have been applied on full digits achieve results of 95% and below (cf. Illing et al., 2019).

In contrast to that, our presented networks learn Gabor-like receptive fields in their first layer, as found in simple-cells in layer

4 of the primary visual cortex, due to the statistics of natural scenes (Bell & Sejnowski, 1997). By nature, Gabor-filters are localized filters. We achieved top accuracy values on MNIST of 98.04%, with 2116 neurons (excitatory neurons in V1-layer 4), for the presented rate-base model and 97.95% for the spike-based model, with 324 (excitatory neurons in V1-layer 4). Other approaches using localized features achieve comparable recognition accuracy values above 95% (Illing et al., 2019; Kheradpisheh et al., 2018; Tavanaei & Maida, 2017). Consequently, this aspect should be taken into account when comparing recognition rates.

For the neuron populations in deeper layers we observed a decreasing recognition accuracy. This could be expected for the character recognition tasks, namely MNIST, EMNIST, and SVHN. The domain shift from natural scenes (landscapes) to characters impairs the ability of deeper layer representations to distinguish the different classes. The deeper layer representations (area V2) learned from natural scenes are presumed to be more selective for naturalistic content than the representations in V1. A presumption that was found in macaque monkeys as well as in humans (Freeman et al., 2013) and which we also found for the rate-based model in previous experiments (Teichmann, 2018). Nevertheless, we expected that the learned representations in deeper layers are of use for more naturalistic datasets, such as CIFAR-10, ETH-80, and Caltech 101. While we found no overall decrease in the performance on the ETH-80 and Caltech 101 dataset, the accuracies on CIFAR-10 dropped. We suspect this is caused by the differences in resolution and content to our natural scene training set, which contains nothing similar to the CIFAR-10 object categories. The use of different data domains for training and testing the neural networks is caused by the different research goals of computational neuroscience in comparison to computer vision. Computational neurosciene aims to create brain-like systems, which requires a stimulation protocol which is somehow realistic (cf. Olshausen & Field, 1996), whereas computer vision aims to solve tasks of use. Because of that, the data required to train a system which promises to develop visual cortex like information processing has to be, in an advanced case, stereoscopic natural scene videos containing not too many man made objects, a requirement not common for standard object recognition datasets. Another reason for the poor performance in comparison to other approaches is the lack of color processing and the enormous effort to implement biologically plausible networks accounting for it (cf. Sincich & Horton, 2005). This issue complicates the comparison to state of the art methods which could be easily applied on colored images. Consequently, there is a need for comparable results on tasks which match the different levels of circuit complexity, such as object recognition on gray scaled images or even more complex tasks such as stereoscopic object recognition.

Despite all limitations to obtain competitive results with biologically grounded neural networks, we argue that it is important to use measures for the information processing on relevant tasks to assess the relevance of the neural representations for behavior. But not only the tasks have to be chosen with respect to the capabilities of the network, biologically grounded neural networks have to evolve as well. The brain utilizes several mechanisms to ensure that the visual cortex learns behavior relevant representations. Namely these are attention mechanisms (Maunsell, 2015), providing spatial- and feature-based attention, and active vision where entities of interest are focused (Eckmann et al., 2020), as well as perceptual learning (Dosher & Lu, 2017). Also reward based learning via the basal ganglia guides the cortex to memorize relevant stimuli (Villagrasa et al., 2018). As well as the emotional system in terms of the amygdala is suspected to guide visual learning (Liebold et al., 2015). All these signals innervate higher visual areas such as V4 or IT from which the information is signaled via feedback connections downwards the cortex and the thalamus to LGN. Due to this, the complex network circuitry of the here presented models is a direct consequence of the goal to cover the vast complexity of the brain circuity. The presented rate-based model already implements parts of the cortical feedback pathway, however, without the use of additional information to enforce the relevance of the presented stimuli. Further circuits motifs we implemented are namely the excitatory feedforward path, the inhibitory feedforward and lateral path. The implementation of the different circuit motifs, allows to investigate different aspects and functions of cortical processing, which was not scope of this manuscript. How the single motifs contribute to the ability to recognize object is subject of intense research (e.g. Kar & DiCarlo, 2021), and would be interesting to analyze in future works.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding

This work has been supported by a grant from the European Social Fund (ESF) and the Free State of Saxony, Germany and in part by German Research Foundation (DFG, 416228727) - SFB 1410 Hybrid Societies.

Appendix A. Learning in the rate-based neural network

Architecture

The model is comprised by two areas mimicking the areas V1 and V2 of the macaque visual cortex. Each of the areas consists of the two most important layers for the feedforward pathway, using the anatomical names layer 4 and layer 2/3. Each layer contains an excitatory and inhibitory neuron population. The model receives input from the LGN area, named after the thalamic nucleus providing the main source of input to the visual cortex. For an overview of the layer sizes see Table 1a.

Following Dale's law, our neurons form only connections of one type. This means, excitatory neurons form excitatory connections and inhibitory ones form inhibitory connections. The connectivity between the neurons follows a few main motives. Lateral: excitatory and inhibitory neurons within a layer are reciprocally connected and inhibitory neurons form connections to the other inhibitory neurons in its population. Forward: excitatory neurons of each layer project to the excitatory and inhibitory neurons of the next layer, i.e. layer 4 neurons project to layer 2/3 and these project to layer 4 of the next area and so on. Recurrent: within each area there are several additional recurrent connections, which are numerously found in the cortical tissue (Potjans & Diesmann, 2014). Recurrent feedback: layer 2/3 neurons have direct connections to layer 2/3 neurons of the preceding area (cortico-cortical). Inter area connections are not target specific and project to excitatory as well as inhibitory neurons (Isaacson & Scanziani, 2011).

We derived this connectivity structure following anatomical studies (Anderson & Martin, 2009; Douglas & Martin, 2004; Thomson & Bannister, 2003). These studies also served as basis for Potjans and Diesmann (2014), who developed a statistical model informed by the connectivity. We use the same major

Table 8
Overview on the initial connection windows (receptive field sizes) for each neuron type in each layer of the rate-based neural network. The target layer indicates the postsynaptic layer and the source layer the presynaptic layer. The synapse type is identical with the neuron type of the source layer. The offset is used to shift the connection window by the given amount to compensate different layer sizes.

Target layer	Neuron type	Source layer	Neuron type	Init RF size	Offset
V1-L4	Excitatory	LGN	Excitatory	10 × 10 × 2	0
		V1-L4	Inhibitory	11 × 11 × 1	−5
	Inhibitory	LGN	Excitatory	10 × 10 × 2	0
		V1-L4	Excitatory	11 × 11 × 4	−5
		V1-L2/3	Excitatory	7 × 7 × 7	−6
		V1-L4	Inhibitory	11 × 11 × 1	−5
V1-L2/3	Excitatory	V1-L4	Excitatory	7 × 7 × 4	0
		V2-L2/3	Excitatory	8 × 8 × 16	−7
		V1-L4	Inhibitory	7 × 7 × 1	0
		V1-L2/3	Inhibitory	9 × 9 × 2	−4
	Inhibitory	V1-L4	Excitatory	7 × 7 × 4	0
		V1-L2/3	Excitatory	9 × 9 × 2	−4
		V2-L2/3	Excitatory	8 × 8 × 16	−7
		V1-L4	Inhibitory	7 × 7 × 1	0
		V1-L2/3	Inhibitory	9 × 9 × 2	−4
V2-L4	Excitatory	V1-L2/3	Excitatory	5 × 5 × 7	0
		V2-L4	Inhibitory	7 × 7 × 3	−3
	Inhibitory	V1-L2/3	Excitatory	5 × 5 × 7	0
		V2-L4	Excitatory	7 × 7 × 10	−3
		V2-L2/3	Excitatory	4 × 4 × 16	−3
		V2-L4	Inhibitory	7 × 7 × 3	−3
V2-L2/3	Excitatory	V2-L4	Excitatory	4 × 4 × 10	0
		V2-L4	Inhibitory	4 × 4 × 3	0
		V2-L2/3	Inhibitory	5 × 5 × 5	−2
	Inhibitory	V2-L4	Excitatory	4 × 4 × 10	0
		V2-L2/3	Excitatory	5 × 5 × 16	−2
		V2-L4	Inhibitory	4 × 4 × 3	0
		V2-L2/3	Inhibitory	5 × 5 × 5	−2

motive for the existence of connections on the population level, but based our individual connections on theoretical assumptions. The simplest assumption is the retinotopic organization of the visual cortex. This means, neighboring neurons receive their input from neighboring afferent neurons. The connection windows are organized in rectangular patches and are of limited size, so that each neuron sees just a fraction of the total input to a layer. The second assumption is locality. This means, that neurons are only connected to a local area of surrounding neurons. For instance inhibitory neurons receive lateral input from neighboring excitatory neurons, which are likely to share similarities in their receptive field. The connections within an inhibitory population follow the same principle, inhibitory neurons are connected to their spatial neighbors. Note, this principle implements a retinotopic organization into the model, so that the x-y coordinates in our neuronal grid imply a neighborhood relation. For an overview of the initial sizes of the neural connectivity patches see Table 8.

This connectivity scheme has functional implications. The neurons develop spatially localized receptive fields, or technically spoken, feature detectors. The lateral locality induces that neighboring neurons are initially connected and the inhibitory circuity can provide decorrelation to neighboring excitatory neurons. This in turn causes that each neuron is enforced to develop a linearly independent receptive field, also inhibitory neurons are enforced to act independently from their neighbors.

Structural plasticity

We developed a structural plasticity mechanism based on two fundamental principles. Locality: new synapses can only be formed in proximity of existing synapses. Experience dependency: the formation probability is higher in the surrounding of strong synapses and weak synapses are more likely to be removed. Hence, we call the structural plasticity model experience-dependent spatial growth model. We based our assumptions on various neuroscientific findings of which a detailed description can be found in Teichmann (2018).

We implement synapse formation, as well as removal, as probabilistic processes. To determine the probability p_j^{form} that a potential, non-formed, synapses j become a synapse, we cumulate the synaptic weight strengths w_i for the connections emerging from the neighboring (in distance d) afferent neurons of the regarded neuron. This sum is normalized by the size of the neighborhood $|B(j,d)|$ and the maximum weight strength. Finally this is multiplied by a scaling factor c_s to obtain the formation probability for this potential synapse (Eq. (1)).

$$p_j^{form} = c_s \cdot \frac{1}{|B(j,d)|} \cdot \sum_{i \in B(j,d)} \frac{w_i}{\max_k(w_k)} \tag{1}$$

The probability for synapse removal p_j^{remove} depends only on the weight strength (Eq. (2)). Weak synapses are likely to be removed, whereas strong ones are subjected to be very stable. Note, that weight values in our network are positive defined, so the minimum weight strength is zero. The probability for removing a weight follows a logistic function, with its maximum at the parameter ρ. The parameter w^{half} indicates the relative weight strength where the logistic function is at the half of its maximum value. We adjust the function in the way that at $2 \cdot w^{half}$ the logistic function is decreased to 10% of its maximum value (Eq. (3)).

$$p_j^{remove} = \frac{\rho}{1 + e^{\left(\frac{w_i}{\max_k(w_k)} - w^{half}\right)/\Delta}} \tag{2}$$

$$\Delta = \frac{w^{half}}{\ln\left(\frac{1}{0.1} - 1\right)} \tag{3}$$

Intrinsic plasticity

Intrinsic plasticity is thought to stabilize the operating point of a neuron. This is important to ensure that all neurons participate

equally in the encoding of stimuli. We found intrinsic plasticity in our previous studies of superior importance for achieving adequate information encoding in multi-layer networks, by enforcing a similar operating point in all neurons of a population (Teichmann & Hamker, 2015). In general it can be said that information encoding is optimal when the neural responses are independent, i.e. no information is duplicated (aimed by inhibition) and all information in the input leads to a response (Simoncelli & Olshausen, 2001). This helps the brain not to waste resources on unresponsive or hyperactive neurons.

From our previous work (Teichmann et al., 2012; Wiltschut & Hamker, 2009), we know that our neural circuit with recurrent excitation and inhibition leads to sparse activity and single neurons follow an exponential firing statistic without enforcing it. Thus, we decided to regulate with intrinsic plasticity the mean and the variance of each single neuron, without forcing a particular distribution. To achieve a similar operating point in all neurons of a population, we use a parameterized rectified linear activation function for each neuron (Eq. (4)). With a threshold θ and a slope a parameter. Thereby m_j denotes the membrane potential and r_j the firing rate of a neuron. w_{ij} is an excitatory weights from the presynaptic neuron i to the postsynaptic neuron j and c_{kj} denotes an inhibitory weight from the presynaptic neuron k. $(.)^+$ denotes top half rectification.

$$\tau_m \frac{dm_j}{dt} = a_j \cdot \left(\sum_i w_{ij} r_i - \sum_k c_{kj} r_k - \theta_j \right) - m_j \tag{4}$$

$$r_j = m_j^+ \tag{5}$$

The parameters for threshold and slope are adapted to enforce the same mean and variance within a population. The threshold is increased if a neuron is more active than the population average, used as target value θ_{target}, and decreased if it is below (Eq. (6)).

$$\tau_\theta \frac{d\theta_j}{dt} = (r_j - \theta_{target}) \delta(\theta_j) \tag{6}$$

The slope is adapted based on the squared neuron activity (Eq. (7)), using the average over the squared population activity a_{target}.

$$\tau_a \frac{da_j}{dt} = (a_{target} - r_j^2) - \delta(a_j - 1) \tag{7}$$

Both terms use a small additional component $\delta(x)$, which implements a constant drift to the origin of the parameters, which are zero for the threshold and one for the slope (Eq. (8)). Where ϵ denotes the drift strength.

$$\delta(x) = \epsilon \cdot sgn(x) \tag{8}$$

Synaptic plasticity

The synaptic plasticity mechanisms in our model consist of three main mechanisms:

- Hebbian learning, with rapid homeostasis through Oja normalization, for all excitatory weights,
- anti-Hebbian learning for all inhibitory weights,
- a long term homeostasis, penalizing high activities, to keep the weight vectors in a meaningful range.

The Hebbian plasticity for the excitatory weights uses a covariance term to determine the weight changes (Sejnowski, 1977; Teichmann et al., 2012; Wiltschut & Hamker, 2009) (Eq. (9)). Long term potentiation (LTP) occurs for presynaptic activities above a certain threshold θ and long term depression (LTD) for activities below. When the threshold is chosen as the average activity of the presynaptic neuron than the weight change will be similar to the covariance of pre- and postsynaptic activity. We chose the value of the threshold as the population average of the presynaptic activity, which should be close to the unknown true average activity of the presynaptic neuron. The weight is bound through an online normalization to the L_2 norm of the weight vector, called Oja normalization (Oja, 1982). The length of the weight vector is controlled by the parameter α. This implements a rapid form of homeostasis (synaptic scaling) which takes place with the speed of the plasticity.

$$\tau_{Learn,j} \frac{dw_{ij}}{dt} = (r_i - \theta) \cdot Ca_j - \alpha_j \left(Ca_j\right)^2 w_{ij} \tag{9}$$

with

$$w_{ij} = (w_{ij})^+$$

Instead of using directly the postsynaptic activity, we compute the calcium level Ca_j as surrogate (Eq. (10)). We use a temporal trace over the postsynaptic activity to define the calcium level (Teichmann et al., 2012), which allows us to define a time window determining which past activities are incorporated in the weight change. Long time windows implement trace learning (Földiák, 1991; Teichmann et al., 2012) (several 100 ms), whereas short time windows are similar to the direct use of the activity (below 100 ms). Only for the excitatory connections from layer 4 neurons to layer 2/3 neurons a long time window (500 ms) is used, all other connections have a short time window (10 ms). As shown in our previous work (Teichmann et al., 2012) this leads to the development of complex-cell like receptive field properties in the layer 2/3 neurons.

$$\tau_{Ca} \frac{dCa_j}{dt} = r_j - Ca_j \tag{10}$$

Part of the calcium dependent weight change for excitatory synapses is the calcium dependent learning rate (Eq. (11)). In biological recordings a higher calcium level caused a higher alteration of the synaptic efficiency (Shouval et al., 2002). Similar to this findings we implemented an exponentially decreasing time constant with increasing calcium level.

$$\tau_{Learn,j} = 5000 + 30\,000 \cdot e^{-15 \cdot Ca_j} \tag{11}$$

As already mentioned, the Oja term serves a form of rapid homeostasis. However, it was found that experimental studies and theoretical largely differ in the speed of the homeostasis (Zenke & Gerstner, 2017; Zenke et al., 2017). We made also the controller variable α of the weight vector length of each neuron adaptive (Eq. (12)) (Teichmann et al., 2012), to allow a long term adaptation of the weight vector length to the desired working range. We penalize strong neuronal activities and currents by increasing α via the helper function H (Eq. (13)) and relax it with a low constant value ϵ_α. We intermingled this with the threshold parameter of the intrinsic plasticity. This long term homeostasis allows the neurons to find their working range by themselves and omits pathologically high firing rates, while giving room to increase the weight vector length to reach higher activity levels.

$$\tau_\alpha \frac{d\alpha_j}{dt} = \alpha_j \cdot \left(H_j - \epsilon_\alpha (1 - \delta\theta_j)\right) \tag{12}$$

$$\tau_H \frac{dH_j}{dt} = \frac{1}{2} \left(r_j - \gamma_r\right)^+ \cdot \sum_i w_{ij} r_i + \frac{1}{10} \left(\sum_i w_{ij} r_i - \gamma_c \right)^+ - K - H_j \tag{13}$$

with

$$\alpha_j = (\alpha_j)^+$$

and

$$H_j = (H_j)^+$$

Table 9
Parameter overview for the rate-based neural network.

Parameter	Description	Value
Activation function		
τ	Time constant membrane potential	10 ms
τ_{Ca}	Layer 4, time constant calcium trace	10 ms
τ_{Ca}	Layer 2/3, time constant calcium trace	500 ms
Synaptic plasticity		
τ_α	Time constant alpha regulation	50 000 ms
τ_H	Time constant helper function	100 ms
γ_r	Threshold on the firing rate	0.6
γ_c	Threshold on the excitatory current	1.3
ϵ_α	Constant decay	0.015
K	Constant decay	0.005
δ	Scaling factor intrinsic plasticity	10
Intrinsic plasticity		
τ_θ	Time constant threshold	10 000 ms
τ_a	Time constant slope	10 000 ms
ϵ	Drift strength	0.01
Structural plasticity		
d	Distance	3
c_s	Build factor	0.1
ρ	Removal factor	0.4
w^{half}	Excitatory removal function	0.01
w^{half}	Inhibitory removal function	0.15

For learning the inhibitory connections we employ anti-Hebbian learning, which decorrelates the neural responses and allows the network to learn the independent components of the input (Falconbridge et al., 2006; Földiák, 1990). Unlike the excitatory connections, we use no calcium dependent learning. The weight develops relative to the covariance of the pre- and postsynaptic activity (Eq. (14)) (King et al., 2013; Teichmann et al., 2012; Wiltschut & Hamker, 2009). The anti-Hebbian product of pre- and postsynaptic activity is normalized by the product of the presynaptic activity r_k and the temporal average over the postsynaptic activity $\langle r_j \rangle$. Again we intermingled the term with the intrinsic plasticity, through lowering postsynaptic average by the threshold from the intrinsic plasticity. This implements a weight change relative to the covariance (Földiák, 1990; Vogels et al., 2011) and balances the inhibition without explicitly setting a target activity (cf. Vogels et al., 2011). We extended this first part of the equation by multiplying the right term with the inhibitory weight c_{kj}, which causes that the weights develop relative to the covariance of the activities (King et al., 2013; Teichmann et al., 2012; Wiltschut & Hamker, 2009). Thus, connected neurons with highly correlating activities should develop stronger inhibitory weights than weak correlated neurons. An overview of the parameters can be found in Table 9.

$$\tau_c \frac{dc_{kj}}{dt} = r_k r_j - r_k(\langle r_j \rangle - \theta_j)^+ \cdot (1 + c_{kj}) \tag{14}$$
$$c_{kj} = (c_{kj})^+$$

Appendix B. Learning in the spike-based neural network

Architecture

The proposed spike-based neural network extends our previously published networks (Larisch et al., 2020, 2018) by a second neural layer. The input layer consists of 648 neurons and has a geometric shape of 18 × 18 × 2, where the last dimension addresses the input neuron type of on- and off-center neurons. It receives voltage pulses generated by a Poisson process representing the image pixels after the applied whitening procedure. The input layer projects to all neurons in the two populations of the first layer (V1-L4). The first layer consisting of an excitatory population with 324 neurons and an inhibitory population with 81 neurons, satisfying the 4 : 1 ratio of excitatory to inhibitory neurons of the primary visual cortex. The inhibitory and excitatory population are recurrently connected and the inhibitory neurons have connections to the other inhibitory neurons of their population (self-inhibition is excluded). The second layer (V1-L2/3) consists of the same amount of excitatory and inhibitory neurons (Table 1) and the same connectivity structure as the first layer. It receives its input from the excitatory population of the first layer (Fig. 1b). Following a brief description of the neuron model and the learning rules is given.

Neuron model of LGN

The number of pulses received by the input layer is determined by a Poisson generation process, which uses the input pixel values after applying the whitening procedure to determine the pulse density. We use a very simple neuron model (Eq. (15)) where the membrane potential (u) is directly dependent on the input current from the Poisson process ($I_{Poisson}$). Each input pulse increases the membrane potential directly above the spiking threshold and causes a spike of the corresponding input neuron. Additionally, we compute a spike trace variable, required to determine the weight change for the subsequent excitatory connections (Eq. (16)). The spike trace ($\bar{x}$) is increased at each spike (with $X_i = 1$ if the neuron spikes and $X_i = 0$ otherwise) and decays exponentially over time.

$$u = I_{Poisson} \tag{15}$$

$$\tau_x \frac{d\bar{x}_i}{dt} = -\bar{x}_i + X_i \tag{16}$$

Neuron model of the other layers

In all excitatory and inhibitory populations the adaptive exponential integrate-and-fire neuron model (Clopath et al., 2010) is used. The membrane potential (u) is defined as follows (Eq. (17)), with the membrane capacitance (C), the leak conductance (g_L), the slope factor (Δ_T), and the resting potential (E_L).

$$C\frac{du}{dt} = -g_L(u - E_L) + g_L\Delta_T e^{\frac{u-V_T}{\Delta_T}} - w_{ad} + z + I_{exc} - I_{inh} \tag{17}$$

The change of the membrane potential is mainly affected by the excitatory (I_{exc}) and inhibitory input currents (I_{inh}). The currents decay with time and increase with incoming spikes (Eq. (18)).

$$\begin{aligned} \tau_{I_{exc}} \frac{dI_{exc}}{dt} &= -I_{exc} + w_i^{exc} \sum_{i \in Exc} \delta(t - t_i') \\ \tau_{I_{inh}} \frac{dI_{inh}}{dt} &= -I_{inh} + w_j^{inh} \sum_{j \in Inh} \delta(t - t_j') \end{aligned} \tag{18}$$

If the membrane potential exceeds the adaptive spiking threshold (V_T) the neuron spikes. This causes that the spiking threshold is set to a high value and decays back to a lower value ($V_{T_{rest}}$) (Eq. (19)).

$$\tau_{V_T} \frac{dV_T}{dt} = -(V_T - V_{T_{rest}}) \tag{19}$$

Two other dynamical variables influence the membrane potential: the hyperpolarizing adaptation current (w_{ad}) and the afterpotential (z). If the neuron spikes, w_{ad} is increased by the

Table 10
Parameters for the neuron model of the spike-based neural network.

Global parameter values			
Parameter (values from Clopath et al. (2010))	Value	Parameter	Value
C, membrane capacitance	281 pF	τ_z, spike current time constant	40 ms
g_L, leak conductance	30 nS	τ_{V_T}, spike threshold time const.	50 ms
E_L, resting potential	−70.6 mV	τ_x, spike trace time constant	15 ms
Δ_T, slope factor	2 mV	τ_{wad}, adaption time constant	144 ms
$V_{T_{rest}}$, spike threshold at rest	−50.4 mV	I_{sp}, spike current after spike	400 pA
$V_{T_{max}}$, spike threshold after spike	−30.4 mV	a, subthreshold adaptation	4 nS
w^e_{min}, min. excitatory weight	0.0	b, spike-triggered adaption	0.805 pA
τ_-, time constant for $\bar{u}_-$	10.0 ms	τ_+, time constant for $\bar{u}_+$	7.0 ms
θ_-, plasticity threshold(LTD)	−70.6 mV	θ_+, plasticity threshold (LTP)	−45.3 mV
Parameter (added)	Value	Parameter	Value
$\tau_{I_{exc}}$, excitatory input time const.	1.0 ms	$\tau_{I_{inh}}$, inhibitory input time const.	10.0 ms

value b and decays exponentially (Eq. (20)). z is increased by the amount of I_{sp} and also decays exponentially (Eq. (21)).

$$\tau_{w_{ad}} \frac{dw_{ad}}{dt} = a(u - E_L) - w_{ad} \tag{20}$$

$$\tau_z \frac{dz}{dt} = -z \tag{21}$$

For a correct functioning of the voltage-depending learning rule, the membrane potential has to be fixed at a constant high value for 2 ms after a spike is generated, subsequently it is reset to the resting potential (E_L). An overview of the parameters is given in Table 10.

Spike timing-dependent plasticity

The plasticity of all excitatory connections follows the voltage-based triplet STDP learning rule from Clopath et al. (2010). A more detailed description of the voltage-based learning rule and the adaptive exponential integrate-and-fire neuron model, together with code, can be found in Larisch (2019).

The weight development for a connection from the presynaptic neuron i is given by Eq. (22).

$$\frac{dw_i}{dt} = A_{LTP}\bar{x}_i(u - \theta_-)^+(\bar{u}_+ - \theta_-)^+ - A_{LTD}\frac{\bar{\bar{u}}}{u_{ref}}X_i(\bar{u}_- - \theta_-)^+ \tag{22}$$

where X_i is the spike counter of the presynaptic neuron (which is 1, if the corresponding neuron has spiked, and 0 otherwise). $\bar{x}_i$ is the presynaptic spike trace (as presented in Eq. (16)). u is the membrane potential of the postsynaptic neuron. $\bar{u}_-$ and $\bar{u}_+$ are two running averages over the membrane potential. Both averages are defined similarly (Eq. (23) for $\bar{u}_+$), but differ in the time constant τ_+ and τ_-.

$$\tau_+ \frac{d\bar{u}_+}{dt} = -\bar{u}_+ + u, \tag{23}$$

The relative strength between LTD and LTP is controlled via a homeostatic mechanism (Eq. (24)), which depends on the quadratic difference between the postsynaptic membrane potential and the resting potential ($\bar{\bar{u}}$) in relation to a constant reference value u_{ref}.

$$\tau_{\bar{\bar{u}}} \frac{d\bar{\bar{u}}}{dt} = [(u - E_L)^+]^2 - \bar{\bar{u}} \tag{24}$$

The parameters A_{LTP} and A_{LTD} are the learning rates for the LTP and respective LTD term, θ_- and θ_+ are threshold parameters.

Table 11
Parameters for the excitatory synapses of the spike-based neural network. E1 refers to the excitatory population and I1 to the inhibitory population of V1-layer 4. E2 refers to the excitatory population and I2 to the inhibitory population of V1-layer 2/3.

Projection-specific parameters			
Parameter (custom values)	$LGN \rightarrow E1$	$LGN \rightarrow I1$	$E1 \rightarrow I1$
$\tau_{\bar{\bar{u}}}$	750 ms	750 ms	750 ms
w^e_{max}	5.0	3.0	0.7
w_{init} (bounds of random uniform distribution)	[0.025, 1.0]	[0.0, 0.1]	[0.0, 0.1]
A_{LTP}	10.8×10^{-5}	5.4×10^{-5}	4.5×10^{-7}
A_{LTD}	8.4×10^{-5}	4.2×10^{-5}	3.6×10^{-7}
$\bar{\bar{u}}_{ref}$	60.0 mV^2	55.0 mV^2	55.0 mV^2
Parameter (custom values)	$E1 \rightarrow E2$	$E1 \rightarrow I2$	$E2 \rightarrow I2$
$\tau_{\bar{\bar{u}}}$	750 ms	750 ms	750 ms
w^e_{max}	3.25	1.0	1.0
w_{init} (bounds of random uniform distribution)	[0.2, 1.25]	[0.01, 0.1]	[0.01, 0.5]
A_{LTP}	8.75×10^{-5}	6.0×10^{-5}	7.2×10^{-6}
A_{LTD}	7.5×10^{-5}	4.8×10^{-5}	5.6×10^{-6}
$\bar{\bar{u}}_{ref}$	50.0 mV^2	45.0 mV^2	50.0 mV^2

As done in Clopath et al. (2010), we equalize the norm of the weights projecting from the Off-LGN population and On-LGN population for each neuron every 20 seconds during the training. The weights are limited by a hard upper (w^e_{max}) and lower (w^e_{min}) bound.

The development of the inhibitory connections follows the phenomenologically motivated, symmetric inhibitory STDP rule (iSTDP) from Vogels et al. (2011) (Eq. (25)).

$$w(t+dt) = \begin{cases} w(t) + \eta(\bar{x}_{post} - \rho) & \text{if } t = t_{pre} \text{ (presynaptic spike)} \\ w(t) + \eta\bar{x}_{pre} & \text{if } t = t_{post} \text{ (postsynaptic spike)} \end{cases} \tag{25}$$

The learning rate is set by the parameter η. ρ controls the balance between LTD or LTP. Moreover, ρ defines a setpoint for the postsynaptic firing rate by up or down regulating the inhibitory strength (Vogels et al., 2011). The post- and presynaptic spike traces ($\bar{x}_{pre}$ and $\bar{x}_{post}$) are defined similar to Eq. (16). In consequence, temporal near post- and presynaptic spikes cause an increase of the inhibitory weights, whereas isolated spikes decrease the weights. Similar to the excitatory weights, the inhibitory weights are limited by a lower (w^i_{min}) and an upper (w^i_{max}) bound. An overview of the parameters is given in Table 10. The specific values for the different excitatory projections can be found in Table 11 and the one for the inhibitory projections in Table 12.

Table 12
Parameters for the inhibitory synapses of the spike-based neural network. E1 refers to the excitatory population and I1 to the inhibitory population of V1-layer 4. E2 refers to the excitatory population and I2 to the inhibitory population of V1-layer 2/3.

Projection-specific parameters				
Parameter	$I1 \rightarrow E1$	$I1 \rightarrow I1$	$I2 \rightarrow E2$	$I2 \rightarrow I2$
τ_{post}	10.0 ms	10.0 ms	10.0 ms	10.0 ms
τ_{pre}	10.0 ms	10.0 ms	10.0 ms	10.0 ms
w^i (initial)	0.0	0.0	[0.1, 0.3]	[0.05, 0.13]
w^i_{min}	0.0	0.0	0.0	0.0
w^i_{max}	0.5	0.5	0.5	0.5
η	4.0×10^{-6}	4.0×10^{-6}	8.0×10^{-6}	4.0×10^{-6}
ρ	35×10^{-2}	45×10^{-2}	50×10^{-2}	60×10^{-2}

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2021.08.009.

References

Anderson, J. C., & Martin, K. A. C. (2009). The synaptic connections between cortical areas V1 and V2 in macaque monkey. *Journal of Neuroscience*, *29*(36), 11283–11293. http://dx.doi.org/10.1523/JNEUROSCI.5757-08.2009, Retrieved from https://www.jneurosci.org/content/29/36/11283.

Banitt, Y., Martin, K. A. C., & Segev, I. (2007). A biologically realistic model of contrast invariant orientation tuning by thalamocortical synaptic depression. *Journal of Neuroscience*, *27*(38), 10230–10239. http://dx.doi.org/10.1523/JNEUROSCI.1640-07.2007, Retrieved from https://www.jneurosci.org/content/27/38/10230.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338. http://dx.doi.org/10.1016/S0042-6989(97)00121-1, Retrieved from http://www.sciencedirect.com/science/article/pii/S0042698997001211.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2016). Towards biologically plausible deep learning. arXiv Preprint. Retrieved from https://arxiv.org/abs/1502.04156.

Beuth, F. (2019). *Visual attention in primates and for machines - neuronal mechanisms* (Dissertation), Technische Universität Chemnitz, Retrieved from https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-356553.

Bo, L., Ren, X., & Fox, D. (2010). Kernel descriptors for visual recognition. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 244–252). Curran Associates, Inc., Retrieved from http://papers.nips.cc/paper/4147-kernel-descriptors-for-visual-recognition.pdf.

Brito, C. S. N., & Gerstner, W. (2016). Nonlinear hebbian learning as a unifying principle in receptive field formation. *PLoS Computational Biology*, *12*(9), 1–24. http://dx.doi.org/10.1371/journal.pcbi.1005070, Retrieved from https://doi.org/10.1371/journal.pcbi.1005070.

Buchs, N., & Senn, W. (2002). Spike-based synaptic plasticity and the emergence of direction selective simple cells: Simulation results. *Journal of Computational Neuroscience*, *13*(3), 167–186. http://dx.doi.org/10.1023/A:1020210230751, Retrieved from https://doi.org/10.1023/A:1020210230751.

Carlson, K. D., Richert, M., Dutt, N., & Krichmar, J. L. (2013). Biologically plausible models of homeostasis and STDP: Stability and learning in spiking neural networks. In *The 2013 international joint conference on neural networks (IJCNN)* (pp. 1–8). http://dx.doi.org/10.1109/IJCNN.2013.6706961.

Clopath, C., Büsing, L., Vasilaki, E., & Gerstner, W. (2010). Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nature Neuroscience*, *13*(3), 344–352. http://dx.doi.org/10.1038/nn.2479, Retrieved from http://www.nature.com/doifinder/10.1038/nn.2479.

Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)* (pp. 2921–2926). http://dx.doi.org/10.1109/IJCNN.2017.7966217.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. http://dx.doi.org/10.1016/j.tics.2007.06.010, Retrieved from http://www.sciencedirect.com/science/article/pii/S1364661307001593.

Diehl, P., & Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, *9*, 99. http://dx.doi.org/10.3389/fncom.2015.00099, Retrieved from https://www.frontiersin.org/article/10.3389/fncom.2015.00099.

Dosher, B., & Lu, Z.-L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science*, *3*(1), 343–363. http://dx.doi.org/10.1146/annurev-vision-102016-061249, Retrieved from https://doi.org/10.1146/annurev-vision-102016-061249 (PMID: 28723311).

Douglas, R. J., & Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, *27*(1), 419–451. http://dx.doi.org/10.1146/annurev.neuro.27.070203.144152, Retrieved from https://doi.org/10.1146/annurev.neuro.27.070203.144152 (PMID: 15217339).

Eckmann, S., Klimmasch, L., Shi, B. E., & Triesch, J. (2020). Active efficient coding explains the development of binocular vision and its failure in amblyopia. *Proceedings of the National Academy of Sciences*, *117*(11), 6156–6162. http://dx.doi.org/10.1073/pnas.1908100117, Retrieved from https://www.pnas.org/content/117/11/6156.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205. http://dx.doi.org/10.1126/science.1225266, Retrieved from https://science.sciencemag.org/content/338/6111/1202.

Falconbridge, M. S., Stamps, R. L., & Badcock, D. R. (2006). A simple hebbian/anti-hebbian network learns the sparse, independent components of natural images. *Neural Computation*, *18*(2), 415–429. http://dx.doi.org/10.1162/089976606775093891, Retrieved from https://doi.org/10.1162/089976606775093891.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, *64*(2), 165–170. http://dx.doi.org/10.1007/BF02331346, Retrieved from https://doi.org/10.1007/BF02331346.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200. http://dx.doi.org/10.1162/neco.1991.3.2.194, Retrieved from https://doi.org/10.1162/neco.1991.3.2.194 (PMID: 31167302).

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981. http://dx.doi.org/10.1038/nn.3402, Retrieved from https://doi.org/10.1038/nn.3402.

Gupta, A., & Garg, A. (2011). Development of receptive field structure of simple cell using spike timing dependent plasticity (STDP). *International Journal of Computer Applications(IJCA) Special Issue on Electronics, Information and Communication Engineering ICEICE(4)*, *4*, 13–18.

Harpur, G. F., & Prager, R. W. (1996). Development of low entropy coding in a recurrent network. *Network. Computation in Neural Systems*, *7*(2), 277–284. http://dx.doi.org/10.1088/0954-898X_7_2_007, Retrieved from https://doi.org/10.1088/0954-898X_7_2_007 (PMID: 16754387).

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely connected convolutional networks. arXiv Preprint. Retrieved from https://arxiv.org/abs/1608.06993.

Illing, B., Gerstner, W., & Brea, J. (2019). Biologically plausible deep learning — But how far can we go with shallow networks? *Neural Networks*, *118*, 90–101. http://dx.doi.org/10.1016/j.neunet.2019.06.001, Retrieved from http://www.sciencedirect.com/science/article/pii/S0893608019301741.

Isaacson, J. S., & Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron*, *72*(2), 231–243. http://dx.doi.org/10.1016/j.neuron.2011.09.027, Retrieved from http://www.sciencedirect.com/science/article/pii/S0896627311008798.

Kandel, E. R., & Schwartz, J. H. (1995). *Essentials of neural science and behavior*. Norwalk, CT: Appleton & Lange.

Kar, K., & DiCarlo, J. J. (2021). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, *109*(1), 164–176.e5. http://dx.doi.org/10.1016/j.neuron.2020.09.035, Retrieved from https://www.sciencedirect.com/science/article/pii/S0896627320307595.

Kermani Kolankeh, A., Teichmann, M., & Hamker, F. H. (2015). Competition improves robustness against loss of information. *Frontiers in Computational Neuroscience*, *9*, 35. http://dx.doi.org/10.3389/fncom.2015.00035, Retrieved from https://www.frontiersin.org/article/10.3389/fncom.2015.00035.

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., & Masquelier, T. (2018). STDP-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, *99*, 56–67. http://dx.doi.org/10.1016/j.neunet.2017.12.005, Retrieved from http://www.sciencedirect.com/science/article/pii/S0893608017302903.

King, P. D., Zylberberg, J., & DeWeese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *The Journal of Neuroscience*, *33*(13), 5475–5485. http://dx.doi.org/10.1523/JNEUROSCI.4188-12.2013, Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23536063.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*(3), 856–867. http://dx.doi.org/10.1152/jn.1994.71.3.856, Retrieved from https://doi.org/10.1152/jn.1994.71.3.856 (PMID: 8201425).

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. http://dx.doi.org/10.3389/neuro.06.004.2008, Retrieved from https://www.frontiersin.org/article/10.3389/neuro.06.004.2008.

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images (Chapter 3)*. University of Toronto.

Larisch, R. (2019). [Re] Connectivity reflects coding a model of voltage-based STDP with homeostasis. *ReScience C*, *5*(3), http://dx.doi.org/10.5281/zenodo.3538217, Retrieved from https://doi.org/10.5281/zenodo.3538217.

Larisch, R., Gönner, L., Teichmann, M., & Hamker, F. H. (2020). Sensory coding and contrast invariance emerge from the control of plastic inhibition over excitatory connectivity. http://dx.doi.org/10.1101/2020.04.07.029157, BioRxiv Preprint. Retrieved from https://www.biorxiv.org/content/early/2020/05/12/2020.04.07.029157.

Larisch, R., Teichmann, M., & Hamker, F. H. (2018) A neural spiking approach compared to deep feedforward networks on stepwise pixel erasement. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, & I. Maglogiannis (Eds.), *Artificial neural networks and machine learning – ICANN 2018* (pp. 253–262). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-01418-6_25.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. http://dx.doi.org/10.1109/5.726791.

Leibe, B., & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization. In *2003 IEEE computer society conference on computer vision and pattern recognition, 2003. Proceedings. Vol. 2* (pp. II–409). http://dx.doi.org/10.1109/CVPR.2003.1211497.

Li Fei-Fei, Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop* (p. 178). http://dx.doi.org/10.1109/CVPR.2004.383.

Liebold, B., Richter, R., Teichmann, M., Hamker, F. H., & Ohler, P. (2015). Human capacities for emotion recognition and their implications for computer vision. *I-Com*, *14*(2), 126–137. http://dx.doi.org/10.1515/icom-2015-0032, Retrieved from https://doi.org/10.1515/icom-2015-0032.

Masquelier, T., Serre, T., Thorpe, S., & Poggio, T. (2007). Learning complex cell invariance from natural videos: A plausibility proof. Retrieved from https://hdl.handle.net/1721.1/39833.

Maunsell, J. H. (2015). Neuronal mechanisms of visual attention. *Annual Review of Vision Science*, *1*(1), 373–391. http://dx.doi.org/10.1146/annurev-vision-082114-035431, Retrieved from https://doi.org/10.1146/annurev-vision-082114-035431 (PMID: 28532368).

Miconi, T., McKinstry, J. L., & Edelman, G. M. (2016). Spontaneous emergence of fast attractor dynamics in a model of developing primary visual cortex. *Nature Communications*, *7*(1), 13208–13218. http://dx.doi.org/10.1038/ncomms13208, Retrieved from https://doi.org/10.1038/ncomms13208.

Mozafari, M., Ganjtabesh, M., Nowzari-Dalini, A., Thorpe, S. J., & Masquelier, T. (2019). Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks. *Pattern Recognition*, *94*, 87–95. http://dx.doi.org/10.1016/j.patcog.2019.05.015, Retrieved from https://www.sciencedirect.com/science/article/pii/S0031320319301906.

Mozafari, M., Kheradpisheh, S. R., Masquelier, T., Nowzari-Dalini, A., & Ganjtabesh, M. (2018). First-spike-based visual categorization using reward-modulated STDP. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(12), 6178–6190. http://dx.doi.org/10.1109/TNNLS.2018.2826721.

Neftci, E., Das, S., Pedroni, B., Kreutz-Delgado, K., & Cauwenberghs, G. (2014). Event-driven contrastive divergence for spiking neuromorphic systems. *Frontiers in Neuroscience*, *7*, 272. http://dx.doi.org/10.3389/fnins.2013.00272, Retrieved from https://www.frontiersin.org/article/10.3389/fnins.2013.00272.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning 2011*. Retrieved from http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*, 267–273. http://dx.doi.org/10.1007/BF00275687, Retrieved from https://doi.org/10.1007/BF00275687.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. http://dx.doi.org/10.1038/381607a0, Retrieved from http://dx.doi.org/10.1038/381607a0.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325. http://dx.doi.org/10.1016/S0042-6989(97)00169-7, Retrieved from http://www.sciencedirect.com/science/article/pii/S0042698997001697.

Palmer, S. E., & Miller, K. D. (2007). Effects of inhibitory gain and conductance fluctuations in a simple model for contrast-invariant orientation tuning in cat V1. *Journal of Neurophysiology*, *98*(1), 63–78. http://dx.doi.org/10.1152/jn.00152.2007, Retrieved from https://doi.org/10.1152/jn.00152.2007 (PMID: 17507506).

Panda, P., & Roy, K. (2016). Unsupervised regenerative learning of hierarchical features in spiking deep networks for object recognition. In *2016 international joint conference on neural networks (IJCNN)* (pp. 299–306). http://dx.doi.org/10.1109/IJCNN.2016.7727212.

Potjans, T. C., & Diesmann, M. (2014). The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network model. *Cerebral Cortex*, *24*(3), 785–806. http://dx.doi.org/10.1093/cercor/bhs358, Retrieved from https://doi.org/10.1093/cercor/bhs358.

Querlioz, D., Bichler, O., Dollfus, P., & Gamrat, C. (2013). Immunity to device variations in a spiking neural network with memristive nanodevices. *IEEE Transactions on Nanotechnology*, *12*(3), 288–295. http://dx.doi.org/10.1109/TNANO.2013.2250995.

Rolls, E. T. (2012). Invariant visual object and face recognition: Neural and computational bases, and a model, VisNet. *Frontiers in Computational Neuroscience*, *6*, 35. http://dx.doi.org/10.3389/fncom.2012.00035, Retrieved from https://www.frontiersin.org/article/10.3389/fncom.2012.00035.

Sadeh, S., Clopath, C., & Rotter, S. (2015). Processing of feature selectivity in cortical networks with specific connectivity. *PLOS ONE*, *10*(6), 1–20. http://dx.doi.org/10.1371/journal.pone.0127547, Retrieved from https://doi.org/10.1371/journal.pone.0127547.

Saunders, D. J., Patel, D., Hazan, H., Siegelmann, H. T., & Kozma, R. (2019). Locally connected spiking neural networks for unsupervised feature learning. *Neural Networks*, *119*, 332–340. http://dx.doi.org/10.1016/j.neunet.2019.08.016, Retrieved from http://www.sciencedirect.com/science/article/pii/S0893608019302333.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology*, *4*(4), 303–321. http://dx.doi.org/10.1007/BF00275079, Retrieved from https://doi.org/10.1007/BF00275079.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426. http://dx.doi.org/10.1109/TPAMI.2007.56.

Shouval, H. Z., Bear, M. F., & Cooper, L. N. (2002). A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proceedings of the National Academy of Sciences*, *99*(16), 10831–10836. http://dx.doi.org/10.1073/pnas.152343099, Retrieved from https://www.pnas.org/content/99/16/10831.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216. http://dx.doi.org/10.1146/annurev.neuro.24.1.1193, Retrieved from https://doi.org/10.1146/annurev.neuro.24.1.1193 (PMID: 11520932).

Sincich, L. C., & Horton, J. C. (2005). The circuitry of V1 and V2: Integration of color, form, and motion. *Annual Review of Neuroscience*, *28*(1), 303–326. http://dx.doi.org/10.1146/annurev.neuro.28.061604.135731, Retrieved from https://doi.org/10.1146/annurev.neuro.28.061604.135731 (PMID: 16022598).

Spratling, M. W. (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 753–761. http://dx.doi.org/10.1109/TPAMI.2005.105.

Spratling, M. W. (2017). A hierarchical predictive coding model of object recognition in natural images. *Cognitive Computation*, *9*(2), 151–167. http://dx.doi.org/10.1007/s12559-016-9445-1, Retrieved from https://doi.org/10.1007/s12559-016-9445-1.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2019). Deep learning in spiking neural networks. *Neural Networks*, *111*, 47–63. http://dx.doi.org/10.1016/j.neunet.2018.12.002, Retrieved from http://www.sciencedirect.com/science/article/pii/S0893608018303332.

Tavanaei, A., & Maida, A. S. (2017). Multi-layer unsupervised learning in a spiking convolutional neural network. In *2017 international joint conference on neural networks (IJCNN)* (pp. 2023–2030). http://dx.doi.org/10.1109/IJCNN.2017.7966099.

Teichmann, M. (2018). *A plastic multilayer network of the early visual system inspired by the neocortical circuit* (Dissertation), Technische Universität Chemnitz, Retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa2-318327. ISBN 978-3-96100-065-4.

Teichmann, M., & Hamker, F. (2015). Intrinsic plasticity: A simple mechanism to stabilize hebbian learning in multilayer neural networks. In T. Villmann & F.-M. Schleif (Eds.), *Proc workshop new challenges in neural computation - NC2 2015,machine learning reports* (pp. 103–111). ISSN 1865-3960. http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_03_2015.pdf.

Teichmann, M., Wiltschut, J., & Hamker, F. (2012). Learning invariance from natural images inspired by observations in the primary visual cortex. *Neural Computation*, *24*(5), 1271–1296. http://dx.doi.org/10.1162/NECO_a_00268, Retrieved from https://doi.org/10.1162/NECO_a_00268 (PMID: 22295987).

Thomson, A. M., & Bannister, A. P. (2003). Interlaminar connections in the neocortex. *Cerebral Cortex*, *13*(1), 5–14. http://dx.doi.org/10.1093/cercor/13.1.5, Retrieved from https://doi.org/10.1093/cercor/13.1.5.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605, Retrieved from http://jmlr.org/papers/v9/vandermaaten08a.html.

Villagrasa, F., Baladron, J., Vitay, J., Schroll, H., Antzoulatos, E. G., Miller, E. K., & Hamker, F. H. (2018). On the role of cortex-basal ganglia interactions for category learning: A neurocomputational approach. *Journal of Neuroscience*, *38*(44), 9551–9562. http://dx.doi.org/10.1523/JNEUROSCI.0874-18.2018, Retrieved from https://www.jneurosci.org/content/38/44/9551.

Vitay, J., Dinkelbach, H., & Hamker, F. (2015). ANNarchy: a code generation approach to neural simulations on parallel hardware. *Frontiers in Neuroinformatics*, *9*, 19. http://dx.doi.org/10.3389/fninf.2015.00019, Retrieved from https://www.frontiersin.org/article/10.3389/fninf.2015.00019.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., & Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, *334*(6062), 1569–1573. http://dx.doi.org/10.1126/science.1211095, Retrieved from http://science.sciencemag.org/content/334/6062/1569.abstract.

Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, *23*(3), 235–250. http://dx.doi.org/10.1016/j.tics.2018.12.005, Retrieved from http://www.sciencedirect.com/science/article/pii/S1364661319300129.

Wiltschut, J., & Hamker, F. H. (2009). Efficient coding correlates with spatial frequency tuning in a model of V1 receptive field organization. *Visual Neuroscience*, *26*(1), 21–34. http://dx.doi.org/10.1017/S0952523808080966.

Zenke, F., & Gerstner, W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *372*(1715), Article 20160259. http://dx.doi.org/10.1098/rstb.2016.0259, Retrieved from https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2016.0259.

Zenke, F., Gerstner, W., & Ganguli, S. (2017). The temporal paradox of Hebbian learning and homeostatic plasticity. *Current Opinion in Neurobiology*, *43*, 166–176. http://dx.doi.org/10.1016/j.conb.2017.03.015, Retrieved from http://www.sciencedirect.com/science/article/pii/S0959438817300910 (Neurobiology of Learning and Plasticity).

Zylberberg, J., Murphy, J. T., & 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 Computational Biology*, *7*(10), 1–12. http://dx.doi.org/10.1371/journal.pcbi.1002250, Retrieved from https://doi.org/10.1371/journal.pcbi.1002250.