





The Neocortex-Inspired Locally Recurrent Neural Network (NILRNN) as a Model of the Primary Visual Cortex

Franz A. Van-Horenbeke^(✉)  and Angelika Peer 

Human-Centered Technologies and Machine Intelligence Lab,
Faculty of Science and Technology, Free University of Bozen-Bolzano, Bolzano, Italy
{fvanhorenbeke,angelika.peer}@unibz.it
<https://hct.projects.unibz.it/>

Abstract. Our recently introduced Neocortex-Inspired Locally Recurrent Neural Network is a machine learning system that is able to learn feature extraction functions from sequential data in an unsupervised way. While it was designed with the main purpose of feature learning, its structure and desired functioning is highly inspired by models of the feedforward circuits in the neocortex. In this work, we study the behavior of our system when it takes shifting images as input, and we compare it with known behavior of the primary visual cortex. The results show that some of the best-known emerging properties in the primary visual cortex, such as the emergence of simple and complex cells as well as orientation maps, also occur in our system, indicating that also their behaviors can be considered analogous. This validates our system as a potential model of the primary visual cortex that may contribute to further understanding of its functioning. In addition, considering that most areas in the neocortex show similarities in terms of structure and operation, future studies of our system over inputs other than images may also bring new insights about other neocortical areas.

Keywords: Brain-inspired machine learning · Biologically inspired neural networks · Cognitive architectures · Feature learning · Unsupervised learning · Models of the visual cortex

1 Introduction

Within the field of cognitive neuroscience, computational models of different regions of the brain have been extensively used when aiming to understand how these systems work and how the different forms of cognition emerge (e.g., models of the basal ganglia [5], models of the hippocampus [3], etc.). One region that has been intensively studied and that has inspired the development of a series of computational models is the primary visual cortex [14], which is the earliest area in the neocortex processing the incoming visual information. The main

advantage of computational models over more descriptive or conceptual models is that they allow us to simulate the system and test different hypotheses. Indeed, they provide full control over the parameters of the system, which allows us to understand how they affect its behavior, and they give the possibility to measure any state variable. Still, not only the models that have been designed with this purpose have been useful in the advancement of our knowledge about the functioning of the different regions of the brain: There are also machine learning systems that, while designed with a complete different purpose, have still drawn inspiration from the structure of regions of the brain. This way, they show analogies with those regions in terms of behavior, and have brought insights about their functioning. A typical example is that of convolutional neural networks (CNNs): These networks have not been designed with the purpose of modeling our visual system, and therefore their design puts performance before being analogous to our brain. Still, they were inspired by the architecture and connections of our visual cortex. This, added to the fact that, due to their great success, they have been extensively and deeply studied, has put them as relevant sources of understanding our visual system [11].

In this sense, our recently proposed Neocortex-Inspired Locally Recurrent Neural Network (NILRNN) [22] is also a machine learning system that mainly seeks performance, but which is strongly inspired by our neocortex. In this study, we compare the behavior of the NILRNN with that of the primary visual cortex, as it is one of the best-known regions of the neocortex. The results show that both systems have analogous behaviors upon all the visual cortex properties for which the NILRNN was evaluated. This, added to the fact that the neocortex seems to be quite homogeneous along its areas in terms of structure and functioning [17], suggests that the use of our system in different applications may bring new insights about the operation of not only the visual cortex, but also other areas of the neocortex.

This article is organized as follows: Sect. 2 introduces concepts about the neocortex and the primary visual cortex, as well as about related computational systems. Section 3 describes the NILRNN architecture. Section 4 presents the results obtained regarding the visual cortex properties for which NILRNN was evaluated. Finally, Sect. 5 discusses on the results obtained and on possible implications.

2 Background

The neocortex is a thin layered region of the brain that is involved in high-level cognitive functions such as sensory perception, rational thought, voluntary motor control or language [13]. It is organized in general in a six-layered structure [18], and it is divided into areas that perform different functions [16]. For example, the primary visual cortex is the first area in the neocortex processing the input visual information. It gets the visual input from the thalamus, processes it and forwards it to the next areas in the visual cortex [21]. Still, the neocortex seems to be quite uniform along most of its areas in terms of structure and operation,

and therefore it appears to have a common underlying algorithm along those different areas [17].

The primary visual cortex is one of the better-known areas in the neocortex, and many computational models of it exist. These models usually focus on layers 2, 3 and 4 of the neocortex, and on the feedforward connections, which are the ones in charge of bringing the visual information through the different processing areas to the higher abstraction areas. Neurons in these layers of the primary visual cortex are sensitive to small regions of the input stimuli known as receptive fields, and are typically classified into simple and complex cells: Simple cells, mainly found in layer 4 (L4), tend to fire after edges in their receptive field with a particular orientation and position, while complex cells, mainly found in layers 2 and 3 (L2/3), tend to fire after edges with a particular orientation, but independently of their position (i.e., small shifts in the input affects little their response) [6]. This behavior is typically studied using as visual stimuli sine gratings as those shown in Fig. 1, for which simple cells tend to respond to a specific orientation and phase, while complex cells respond to a specific orientation but are more phase-invariant. In addition, neurons in layers 2, 3 and 4 with similar receptive fields and orientation preferences are found to be located close to each other, forming smooth ordered maps [10]. However, such order does not seem to exist in terms of phase [12].



Fig. 1. Examples of sine gratings of different orientations, spatial frequencies and phases.

In computational models of the primary visual cortex, the behavior of simple cells is typically achieved through Hebbian-like learning techniques, which model how neurons in our brain learn [4]. These learning rules applied over small regions of input images lead the modeled neurons to learn edge patterns of a particular orientation and phase. Regarding complex cells (mainly in L2/3), their expected behavior is often achieved by pooling simple cells (mainly in L4) of similar orientations but different phases, achieving this way a strong response to that orientation in a more phase-invariant way [9]. Considering that neurons in the primary visual cortex with similar orientations but different phases tend to be close to each other, models that satisfy such property can achieve the desired complex cell behavior by just pooling the neurons in a localized region of L4. Antolik et al. [1] proposed a model able to achieve such orientation order and phase disorder in a biologically plausible way by introducing lateral and feedback connections that allow neurons in L4 to contribute to the firing of their neighbors with certain time delay. This makes nearby neurons respond to input patterns that tend to occur close in time, but not to the same input. This way, if the model gets as input shifting images (mimicking the input to our visual system),

nearby neurons with similar receptive fields will learn edges with similar orientations but shifted in space, leading to the desired order. Another advantage of this model is that it does not explicitly rely on properties of the input images, and therefore may be also valid for other areas of the neocortex processing different types of data (as most areas have a similar structure).

Computational models of the visual cortex have inspired several machine learning systems, being CNNs a well-known successful example that has also contributed to its understanding [11]. CNNs, however, do not rely on learning a set of patterns that show orientation order but phase disorder to then pool nearby neurons together, but they are explicitly designed (i.e., hardwired) to pool neurons detecting the same pattern at slightly shifted positions of the input image. A key factor to their success seems to be that those shifted versions of the same pattern contribute essentially with the same information to the overall meaning of the input, and by grouping them, the network is losing little relevant information while simplifying the representation. However, this idea is not applicable in general to domains other than vision (e.g., shifting the elements of a generic feature vector may completely change its meaning), and neither seems to correspond to anything occurring in other regions of the neocortex. Our recently proposed NILRNN [22], on the other hand, is indeed designed to achieve orientation order and phase disorder when having shifting images as input. To do so, it relies on the same principle as that of the model by Antolik et al. [1], i.e., it pushes nearby neurons to learn patterns typically occurring close in time, which are then pooled together. This way, it can be argued that NILRNN works because input patterns that tend to occur close in time also have in general a very similar meaning, and can therefore be grouped together. This approach has the added benefit that it applies to almost any domain that deals with sequential data, and is a mechanism that may be also occurring in neocortical regions other than the visual cortex. In fact, considering that, due to such mechanism, the activity in the pooling neurons varies slower in time than the input, this approach is also consistent with the neocortex-related slowness principle, which states that the environment changes in a slower timescale than the sensory input we get from it, and therefore, good representations of the environment should also change in such slower timescale [23]. This makes NILRNN a more accurate model of the visual cortex in terms of structure as well as a potential model of other areas of the neocortex. Still, as we mentioned in Sect. 1, NILRNN was designed as a feature learning system for machine learning applications rather than as a model of the neocortex. In this regard, NILRNN has already shown its effectiveness outperforming other feature learning systems in classification tasks over sequential data domains such as speech recognition or action recognition [22].

3 Materials and Methods

NILRNN is an unsupervised feature learning neural network for sequential data. The NILRNN **feature extraction system**, shown in Fig. 2, consists of three

layers: the input layer, the recurrent layer (analogous to L4) and the max pooling layer (analogous to L2/3). Neurons in the recurrent layer are arranged in two dimensions, with neurons close to each other being connected through recurrent connections. This allows neurons to contribute to the firing of their nearby neurons in the next timestep, which, during the training phase, pushes them to learn input patterns that tend to occur successively in time. This way, a form of self-organization mechanism emerges, with a global order appearing due to the local interactions among the components. Since we will work with images as input, the input to this layer will be also partially connected, following a connection pattern that mimics the one observed in the primary visual cortex: Neurons in the recurrent layer are connected to a region of the input (i.e., their receptive field, see Sect. 2), in a way that, when moving along the neurons in the recurrent layer in both directions, the corresponding receptive fields also shift smoothly in both directions, similar to the connection patterns of convolutional layers. This way, neighbor neurons have same or very similar receptive fields. Neurons in the recurrent layer make use of sigmoid activation functions, since their desired behavior consists of just working as detectors of specific patterns in the input. Regarding the max pooling layer, it has a similar input connection pattern, with each neuron pooling neurons from a region of the recurrent layer. All these connection kernels have an approximately circular shape.

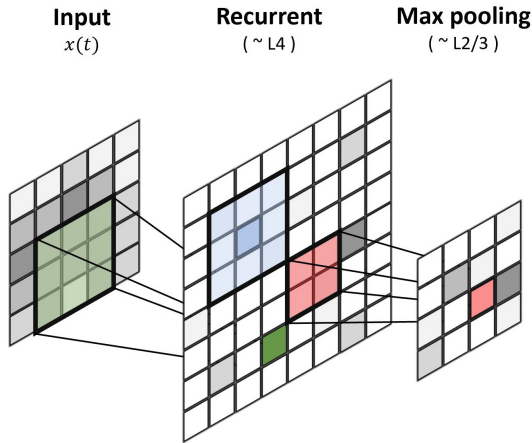


Fig. 2. NILRNN feature extraction system architecture with partially-connected input. The green cells represent input feedforward connections. The blue cells represent local recurrent connections. The red cells represent max pooling connections. (Color figure online)

This feature extraction system can of course be trained as part of a larger neural network and in a supervised way, but in this study we are interested in analyzing its behavior when trained in an unsupervised way, similar to what

occurs in the brain. The NILRNN **unsupervised learning system** (see Fig. 3) relies on self-supervised learning techniques similar to those used in autoencoders [7] (Hebbian learning methods are avoided because they typically require complementary mechanisms to lead to the desired results, making the system more complex and harder to design [15]). This way, since the neural network is recurrent, it is trained to reconstruct the input (same as autoencoders) as well as to predict the following inputs. On the other hand, since the max pooling layer does not need to learn any weights nor contributes in a positive way to the desired learning of weights in the recurrent layer, it is not included in the self-supervised learning system. The output layer is formed of several channels with the size and shape of the input layer corresponding to each of the predictions at the different timesteps. Each of these channels are connected to the recurrent layer following a pattern that is symmetric to that defining the connections between the input layer and the recurrent layer. Neurons in the output layer also make use of sigmoid activation functions, which means that the input to the network needs to be in the range $(0, 1)$. Finally, the recurrent layer is designed to have a sparse activity by adding a sparsity term to the cost function, similar to how is done for sparse autoencoders [7]. Sparsity, which consists of allowing only a small percentage of the neurons to be active at a given time, is a behavior that has been observed in the neocortex [8], and it is often very appropriate to represent the observations of the real world because these observations can usually be described through the presence of a limited number of features out of a considerably larger number of possible features (e.g., the presence of certain objects, their location, etc.), besides showing other advantages. This way, the cost function is given by the following equation:

$$J(W, b) = J_{error} + \lambda \cdot J_{regularization} + \beta \cdot J_{sparse} \quad (1)$$

where W and b represent all the variable weights and bias units, J_{error} is the squared-error cost term, $J_{regularization}$ is the L_2 regularization term, J_{sparse} is the sparsity term based on the KL-divergence and applied only to the recurrent layer, and λ and β are cost function weights.

4 Results

To evaluate how the NILRNN has an analogous behavior to that of the neocortex, we have used a network that takes as input an image patch of size 16×16 , and has a recurrent layer formed of neurons with a receptive field of size 69 pixels and connected in a recurrent way to 29 neurons of the same layer. Their specific receptive field is defined by a stride of 0.33 neurons (i.e., every three neurons, the receptive field shifts one pixel), leading to a layer size of 46×46 . The max pooling layer is defined by a kernel size of 21 neurons and by a stride of 1 neuron, having also a size of 46×46 . Regarding the self-supervised learning system, its output is formed of three channels (i.e., it has a size of $16 \times 16 \times 3$, reconstructing the current and next two inputs). The cost function is characterized by weights $\lambda = 1.5 \cdot 10^{-6}$ and $\beta = 0.15$, and by a desired sparsity parameter of 0.04.

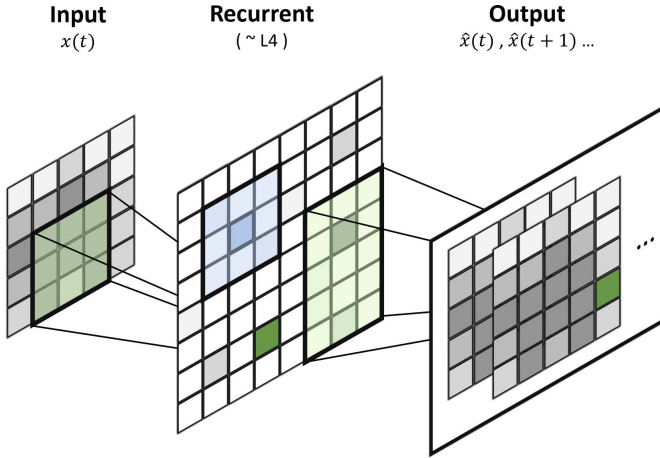


Fig. 3. NILRNN self-supervised feature learning system architecture with partially-connected input and output. The green cells represent input and output feedforward connections. The blue cells represent local recurrent connections. (Color figure online)

For the training, we have used truncated backpropagation through time with a truncation horizon of 4 timesteps, Adam optimization with a stepsize of $2.5 \cdot 10^{-3}$, and a batch size of 1000 samples, and we have trained the system on 400.000 batches. Note that all these hyperparameters have been set manually, and therefore better results may be obtained using other values. Regarding the training input, it consists of sequences of image patches of size 16×16 obtained by moving laterally a 16×16 window along the image at random velocities and directions, and which are taken from whitened natural images [19] normalized to the interval $[0.1, 0.9]$.

Once the system has been trained, we have set images of drifting sinusoidal gratings with different phases, orientations and spatial frequencies as those shown in Fig. 1 as input, and we have analyzed the responses of the neurons in both the recurrent and max pooling layers, similar to how has been done when studying the behavior of the primary visual cortex [2] or of models of it [1]. Figure 4 shows the resultant weights from the training at the input feedforward connections, once normalized (i.e., the input patterns that the neurons at the recurrent layer have learnt to detect). This figure shows that, as expected, the neurons learn to detect edges in the input, with neighbor neurons tending to detect edges with similar orientations but different phases. This can be also observed in the **orientation** and **phase maps** of the recurrent and max pooling layers shown in Fig. 5. These maps are obtained by finding, for each neuron, the orientation and phase of the input pattern that draws the maximum response, for any spatial frequency. As can be seen in this figure, the orientation map for the max pooling layer looks similar and has similar characteristics to those typically obtained from the primary visual cortex (i.e., it

has homogeneous regions appearing periodically, pinwheels where many different orientations meet, etc.) [2]. The orientation map for the recurrent layer has similar properties, but with the regions being more scattered, which is something that also occurs in other models of the primary visual cortex, and that, to the best of our knowledge, does not contradict any experimental evidence [1]. In addition, and as expected, regions at similar positions of the two layers have similar orientation preferences. Regarding the phase maps, that of the recurrent layer does not appear to have any order, which is consistent with experimental evidence. On the other hand, some homogeneous regions appear at the phase map of the max pooling layer, but these regions are in general smaller than those at the orientation map, and have the shape of the max pooling kernel (which is a 21 neurons kernel with the shape of a 5×5 square without its 4 corners). This way, they seem to appear simply because the maximum value for a neuron in the max pooling layer corresponds to the maximum of all the maximum values of the neurons it is pooling, so neurons in the recurrent layer having high maximum values will lead most neurons pooling them to have those same maximum values for the same input patterns (i.e., same phases). Other than that, it does not seem to exist any phase order.

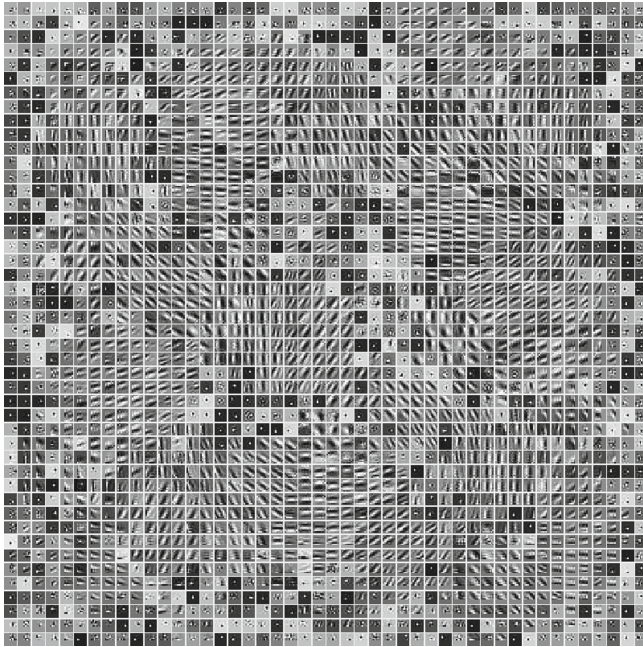


Fig. 4. Normalized weights at the input feedforward connections of an NILRNN trained with shifting images as input.

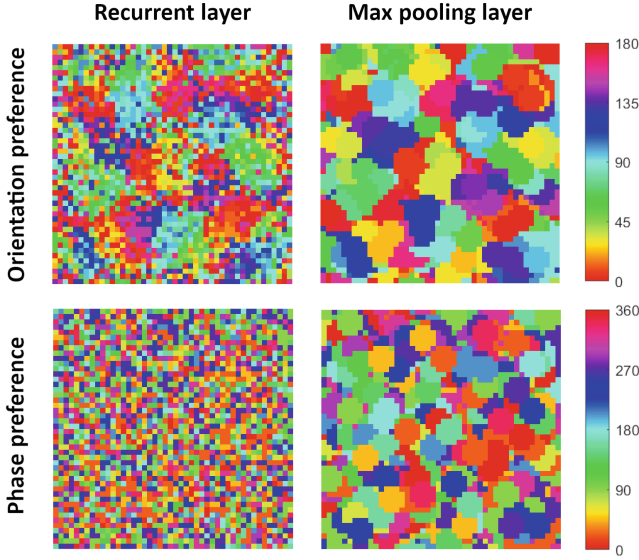


Fig. 5. Orientation and phase maps at the recurrent and max pooling layers of an NILRNN trained with shifting images as input.

In order to evaluate whether the neurons in both layers behave as simple or complex cells, we have calculated their **modulation ratios** [20]. The modulation ratio of a neuron is calculated as the ratio between the first harmonic and the average of the response to a drifting sine of the spatial frequency and orientation for which the maximum response is obtained. Neurons with a more simple-like behavior respond in a strong way to a particular phase with respect to the others, and therefore have a higher modulation ratio, while those with a more complex-like behavior respond more homogeneously, and therefore have a lower modulation ratio. Typically, neurons are classified as complex cells if their modulation ratio is below 1, and as simple cells otherwise. Figure 6 shows the histograms with the modulation ratios of all the neurons of each layer of the NILRNN, as well as a typical distribution obtained when taking a sample of neurons from the primary visual cortex across layers. In this figure we can see how, as expected, most neurons in the recurrent layer behave as simple cells, while most neurons in the max pooling layer behave as complex cells, analogously to what has been observed in the primary visual cortex (see Sect. 2). On the other hand, if we consider the neurons of both layers of our system altogether, the shape of the resultant histogram would be very similar to that obtained for the neurons in the primary visual cortex (except for the relative height of the two peaks, which depends on the size of the sample of neurons at each layer, and therefore should not be considered as a relevant difference).

Finally, Fig. 7 shows the **orientation tuning curves** and **phase responses** of some representative neurons from the inside of the orientation-wise homogeneous regions of both layers of the NILRNN. The orientation tuning curves show

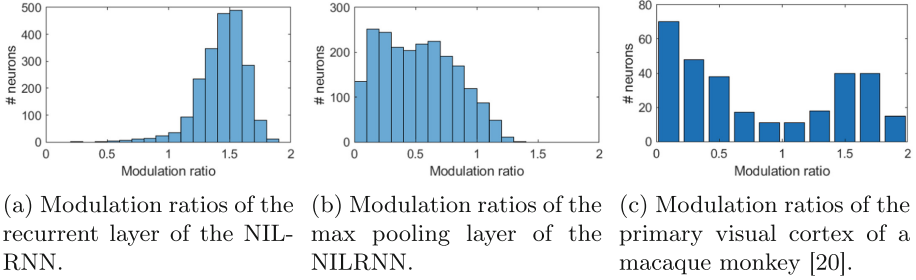


Fig. 6. Modulation ratios of the neurons of each layer of an NILRNN trained with shifting images as input, and of a sample of neurons from several layers of the primary visual cortex of a macaque monkey.

the maximum response value of each neuron as a function of the orientation for any phase and spatial frequency. The phase responses show the response obtained as a function of the phase for the orientation and spatial frequency that give the strongest response. The orientation tuning curves show that all these neurons are indeed finely tuned to a narrow band of spatial frequencies. As expected, and analogously to what has been observed in the primary visual cortex [20], the spatial frequency bands of the neurons in the max pooling layer are broader than those of the neurons in the recurrent layer. Regarding the phase responses, the figure shows how the neurons in the recurrent layer are also finely tuned to a narrow band of phases, while the neurons in the max pooling layer are much more phase-invariant, which corresponds to the behavior of simple and complex cells, respectively, as commented in Sect. 2.

5 Discussion

NILRNN is a neocortex-inspired artificial neural network for the unsupervised learning of features in sequential data. It is strongly inspired by computational models of the primary visual cortex, and it relies on brain-inspired machine learning mechanisms and principles such as sparsity, slowness or self-organization. The results presented in this study show that its behavior is in different ways analogous to that of the primary visual cortex. This means that it can function to some extent as a model of the primary visual cortex, and contribute to obtaining new insights about its principles of functioning, similar to what has occurred with CNNs. In fact, NILRNN is analogous to the primary visual cortex to a larger extent than CNNs both in terms of structure and of emerging behavior, allowing it for example to develop orientation maps. Furthermore, NILRNN is based on ideas that do not only apply to the visual cortex, but which may also apply to other areas of the neocortex (i.e., it relies on grouping together input patterns that tend to occur close in time, and not on grouping together spatially shifted versions of the same pattern). This allows NILRNN to be applicable to domains

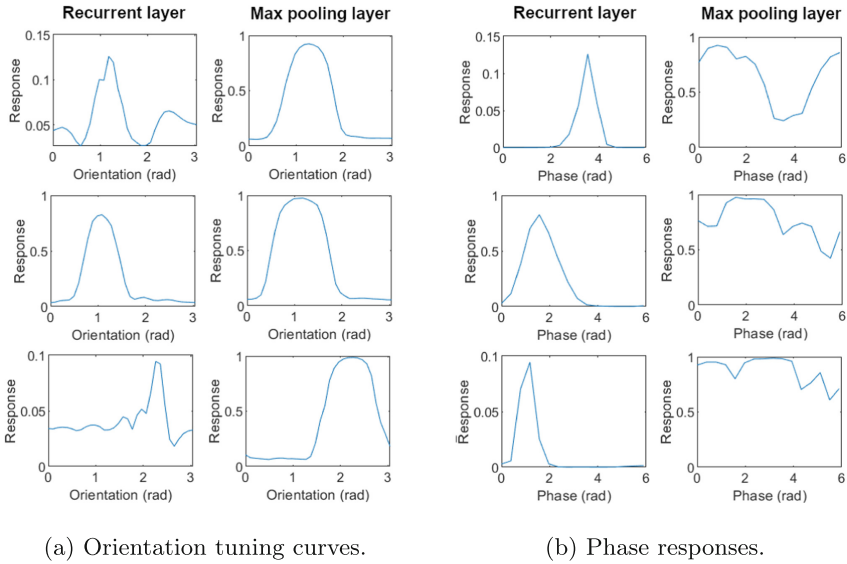


Fig. 7. Orientation tuning curves and phase responses of three representative neurons of each layer of an NILRNN trained with shifting images as input. The positions of the three neurons in both layers are (28,10) (top), (23,38) (middle) and (21,13) (bottom).

with very different properties from those of computer vision, as well as to possibly serve as a model of different areas of the neocortex, and, thus, contribute to the advancement of our knowledge about those areas and the neocortex in general.

Acknowledgements. This research was supported by the Euregio project OLIVER (Open-Ended Learning for Interactive Robots) with grant agreement IPN86, funded by the EGTC Europaregion Tirol-Südtirol-Trentino within the framework of the third call for projects in the field of basic research.

References

1. Antolik, J., Bednar, J.A.: Development of maps of simple and complex cells in the primary visual cortex. *Front. Comput. Neurosci.* **5**, 17 (2011)
2. Blasdel, G.G.: Orientation selectivity, preference, and continuity in monkey striate cortex. *J. Neurosci.* **12**(8), 3139–3161 (1992)
3. Burgess, N.: Computational models of the spatial and mnemonic functions of the hippocampus. In: Andersen, P., Morris, R., Amaral, D., Bliss, T., O’Keefe, J. (eds.) *The Hippocampus Book*, pp. 715–750. Oxford University Press (2007)
4. Choe, Y.: Hebbian learning. In: Jaeger, D., Jung, R. (eds.) *Encyclopedia of Computational Neuroscience*, pp. 1305–1309. Springer, New York (2015)
5. Cohen, M.X., Frank, M.J.: Neurocomputational models of basal ganglia function in learning, memory and choice. *Behav. Brain Res.* **199**(1), 141–156 (2009)

6. Gilbert, C.D.: Laminar differences in receptive field properties of cells in cat primary visual cortex. *J. Physiol.* **268**(2), 391–421 (1977)
7. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016). <http://www.deeplearningbook.org>
8. Graham, D.J., Field, D.J.: Sparse coding in the neocortex. *Evol. Nervous Syst.* **3**, 181–187 (2006)
9. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**(1), 106–154 (1962)
10. Hubel, D.H., Wiesel, T.N.: Sequence regularity and geometry of orientation columns in the monkey striate cortex. *J. Compar. Neurol.* **158**(3), 267–293 (1974)
11. Lindsay, G.W.: Convolutional neural networks as a model of the visual system: past, present, and future. *J. Cogn. Neurosci.* **33**(10), 2017–2031 (2021)
12. Liu, Z., Gaska, J.P., Jacobson, L.D., Pollen, D.A.: Interneuronal interaction between members of quadrature phase and anti-phase pairs in the cat's visual cortex. *Vision. Res.* **32**(7), 1193–1198 (1992)
13. Lukatela, K., Swadlow, H.A.: Neocortex. *The corsini encyclopedia of psychology*, pp. 1–2 (2010)
14. Martinez, L.M., Alonso, J.M.: Complex receptive fields in primary visual cortex. *Neuroscientist* **9**(5), 317–331 (2003)
15. McClelland, J.L.: How far can you go with hebbian learning, and when does it lead you astray. *Processes of change in brain and cognitive development: attention and performance xxi*, vol. 21, pp. 33–69 (2006)
16. Mesulam, M.M.: From sensation to cognition. *Brain J. Neurol.* **121**(6), 1013–1052 (1998)
17. Mountcastle, V.B.: The columnar organization of the neocortex. *Brain J. Neurol.* **120**(4), 701–722 (1997)
18. Narayanan, R.T., Udvary, D., Oberlaender, M.: Cell type-specific structural organization of the six layers in rat barrel cortex. *Front. Neuroanat.* **11**, 91 (2017)
19. Ng, A.: *Deep learning and unsupervised feature learning handouts* (2011). <https://web.stanford.edu/class/cs294a/handouts.html>
20. Ringach, D.L., Shapley, R.M., Hawken, M.J.: Orientation selectivity in macaque v1: diversity and laminar dependence. *J. Neurosci.* **22**(13), 5639–5651 (2002)
21. Tong, F.: Primary visual cortex and visual awareness. *Nat. Rev. Neurosci.* **4**(3), 219–229 (2003)
22. Van-Horenbeke, F.A., Peer, A.: Nilrnn: a neocortex-inspired autoencoder-like locally recurrent neural network for unsupervised feature learning in sequential data (2022). (manuscript in preparation)
23. Wiskott, L.: Slow feature analysis: a theoretical analysis of optimal free responses. *Neural Comput.* **15**(9), 2147–2177 (2003)