IRENA

Invariant Representations Extraction in Neural Architectures

NeuroTHIx

Project name: **IRENA** (**I**nvariant **R**epresentations **E**xtraction in **N**eural **A**rchitectures)
Team: **NeuroTHIx**
Team members: **Du Xiaorui, Erdem Yavuzhan, Cristian Axenie**

Approach

In our proposed solution we aim at a combination of Neuroscience principles, Abstract thinking and Prototyping, towards a solution aiming at bringing the efficiency and robustness of biological intelligence to technical systems that solve real-world problems. We start from the proposed Neuroscientific Research Challenges and propose a novel model and system capable of learning invariant representation.

Preamble

In this project we aim at building an unsupervised learning system that is based on and inspired by our biological intelligence for the problem of learning invariant representations. Mammalian visual systems are characterized by their ability to recognize stimuli invariant to various transformations. With our proposed model, we investigate the hypothesis that this ability is achieved by the temporal encoding of visual stimuli, and why not, other sensory stimuli. By using a model of a multisensory cortically inspired network, we show that this encoding is invariant to several transformations and robust with respect to stimulus variability. Furthermore, we show that the proposed model provides a rapid encoding and computation, in accordance with recent physiological results [9].

Elucidating the mechanisms of invariant pattern recognition is an active field of research in neuroscience [1–5]. However, very little is known about the underlying algorithms and mechanisms. A number of models have been proposed which aim to reproduce capabilities of the biological visual system, such as invariance to shifts in position, rotation, and scaling [6–8]. Starting from these premises, in order to build our approach, we reviewed functional organization in sensory cortical regions - how the cortex represents the world.

We consider four interrelated aspects of cortical organization and computation: (1) the set of receptive fields of individual cortical sensory neurons - the data representations, (2) how the interaction between cortical neurons reflects the similarity of their receptive fields - the data composition, (3) the spatial distribution of receptive-field properties - the data hierarchization and (4) how the spatial distributions of different receptive-field properties interact with one another - the cue-integration and correlation learning.

Driven by these principles and the study in [9] which shows how the neurophysiology data are generally well explained by the theory of input-driven self-organization, we explore a cortically-inspired model of cortical maps offering a parsimonious account for a wide range of map-related phenomena, that has the potential to explain phenomena related to the formation of hierarchical invariant representations. We decided to go deeper and address the problem of learning underlying relations behind such transformations by using biologically plausible representation (i.e. population codes) and computation (i.e. competition, cooperation, correlation).

Introducing the core model and system

Using cortical maps as neural substrate for distributed representations of sensory streams, our system is able to learn its connectivity (i.e., structure) from the long-term evolution of sensory observations. This process mimics a typical development process where self-construction (connectivity learning), self-organization, and correlation extraction ensure a refined and stable representation and processing substrate, as also shown in [10]. Following these principles, we propose a model based on Self-Organizing Maps (SOM) [11] and Hebbian Learning (HL) [12] as main ingredients for extracting underlying correlations in sensory data, the basis for subsequently extracting invariant representations. Our vision is that biological systems process visual input using a distributed representation, with different areas encoding different aspects of the visual interpretation. While current engineering habits tempt us to think of this processing in terms of a pipelined sequence of filters and other feed-forward processing stages, cortical anatomy suggests quite a different architecture, using strong recurrent connectivity between visual areas.

Starting point and motivation

Interestingly enough we started from the challenges that Merck set up in the Neuroscience track.

For motivating the use of SOM we started from the challenge quote: *"[...] the 'location neurons' behave similar to the vector quantization algorithm [...], i.e. the 'location neurons' would realize the same mechanism which is also conjectured to lead to invariant representations! "*.
In order to address the "specialization" of the neurons we used SOM, which is extending the basic process in Vector Quantization (i.e. competition, Winner-Take-All (WTA)) by joining it with cooperation in updating the neural weights (i.e. Soft-WTA). This will allow for a topological representation of the input space in the latent space of representation, such that close inputs are mapped closed in the latent space.
Moreover, SOM are responsible for extracting the statistics of the incoming data and encoding sensory samples in a distributed activity pattern. This activity pattern is generated such that the neuron closest to the input sample, in terms of its preferred value, will be strongly activated. Activation decays as a function of distance between input and preferred value. Using the SOM distributed representation, the model learns the boundaries of the input data, such that, after relaxation, the SOM provide a topology preserving representation of the input space. We extend the basic SOM, introduced in [11], in such a way that each neuron not only specializes

in representing a certain (preferred) value in the input space, but also learns its own sensitivity (i.e., tuning curve shape). Using this mechanism, the model optimally allocates resources (i.e., neurons): a higher amount to areas in the input space which need a finer representation; and a lower amount for those areas that don't. This feature emerging from the model is consistent with recent work on optimal sensory encoding in neural populations [13]. This claims that, in order to maximize the information extracted from the sensory streams, the prior distribution of sensory data must be embedded in the neural representation.

For motivating the use of HL we started from the challenge quote: *"[...] interestingly, this 'OR' operation is exactly what one needs when creating equivalency classes, so this hints at another connection to our conjectured algorithm for invariant representations [...]"*.
In order to address the underlying transformations / mathematical operations performed in neural circuitry, we employed a biologically plausible model that links the field of equivalence classes to accounts of Hebbian learning and categorization [14], namely HL. The second component of our model is the unsupervised Hebbian linkage, more precisely a covariance learning rule akin to the ones introduced in [12]: a fully connected matrix of synaptic connections between neurons in each input SOM, such that the projections propagate from presynaptic units to postsynaptic units in the network. Using an all-to-all connectivity pattern, each SOM unit activation is projected through the Hebbian matrix. The Hebbian learning process is responsible for extracting the co-activation pattern between the input layers (i.e., SOMs) and for eventually encoding the learned relation between the sensors. The effective correlation pattern encoded in a matrix, imposes constraints upon possible sensory values. Moreover, after the network converges, the learned sensory dependency will make sure that values are "pulled" towards the correct (i.e., learned) corresponding values, will neglect outliers, and will allow inferring missing sensory quantities.

Core model

In the following subsection we give a brief overview on the underlying mechanisms in our model. This is also present in the developed demo code (Python code).

In order to give an intuition on the inner workings of the aforementioned mechanisms, we start with a simple bimodal scenario, depicted in Figure 1b, in which the correlation among two sensors is represented by a simple nonlinear relation, e.g., power-law, as depicted in Figure 1a. Here sensory data can be n-dimensional, yet for simplicity we look at time series.

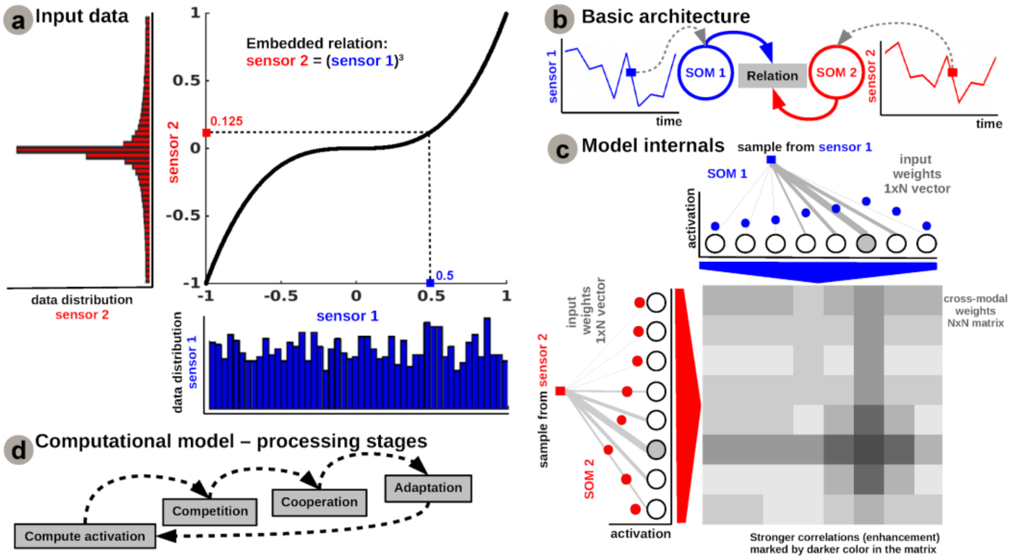

Figure 1. Model architecture. (a) Input data resembling a nonlinear relation and its distribution; (b) Basic architecture; (c) Model internal structure; (d) Processing stages.

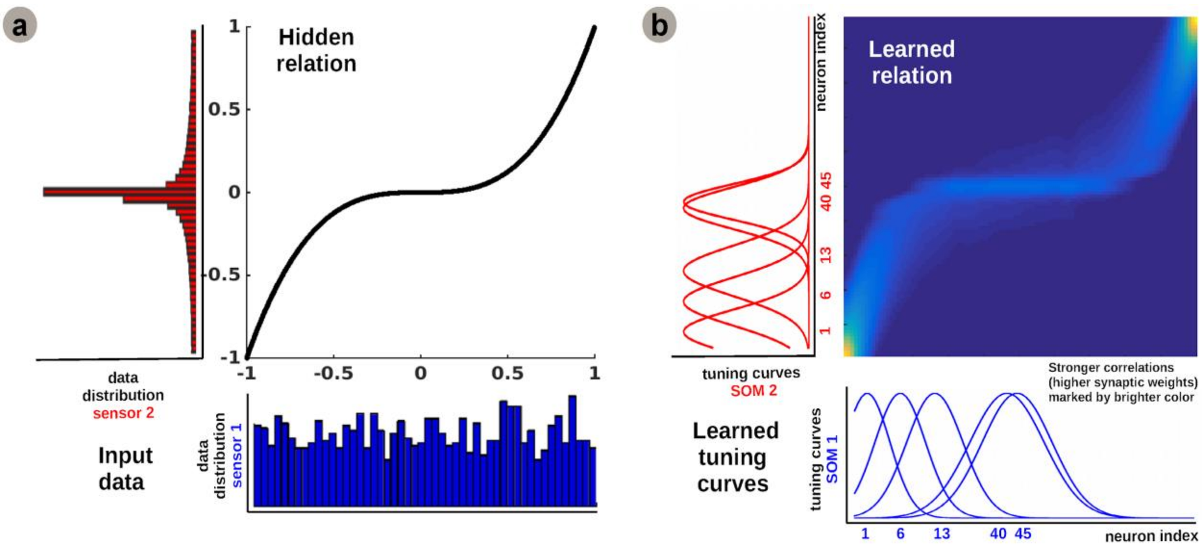

Figure 2. Extracted sensory relation and data statistics using the proposed model: (a) Input data statistics and hidden relation; (b) Learned preferred values and underlying relation.

Recall that, using these mechanisms, the network optimally allocates resources (i.e., neurons): a higher amount to areas in the input space which need a finer representation; and a lower amount for those areas that don't. The unsupervised Hebbian learning process is responsible for extracting the co-activation pattern between the input layers (i.e., SOMs), as shown in Figure 1c, and for eventually encoding the learned relation between the sensors, as shown in Figure 2b. The central panel of Figure 2b demonstrates that connections between uncorrelated (or weakly correlated) neurons in each population are suppressed (i.e., darker color-lower value) while correlated neurons' connections are enhanced (i.e., brighter color-higher value).

Code of the basic implementation is attached with the final submission of the project.
NeuroTHIx_IRENA_codebase_final.zip

A unified framework for invariant representations

In the following section we provide the perspective and extension ideas for in case we pass the qualifiers and will engage with Merck to bring this ideas to life, in a real-world problem of learning invariant representations in visual scenes.

One of the key differences between biological and engineered visual systems is that engineered solutions traditionally use a feed-forward sequence of processing stages and filters, whereas biology (e.g. in primates) uses strong recurrent connectivity between different brain areas which process different types of information in parallel. Understanding this biological style of visual processing could help with the long term technological goal of matching or surpassing the visual capabilities of biological systems. Some key architectural properties that currently are largely unique to biological vision systems include the strong recurrent connectivity between cortical areas, the ability of seemingly weak input to dominate the activity of a multi-area system, and the ability of a distributed representation to arrive at a coherent interpretation of weak or noisy input. We considered such principles in our model, yet there are still some interesting aspects to develop on top of our initial system.

In order to demonstrate that we are planning to design a system having these properties, which analyzes visual input with a network of recurrently connected SOM (i.e. through HL matrix), as shown in the basic model using SOM and HL. Each map represents a different aspect of the visual interpretation, such as light intensity or optic flow, as shown in Figure 3.

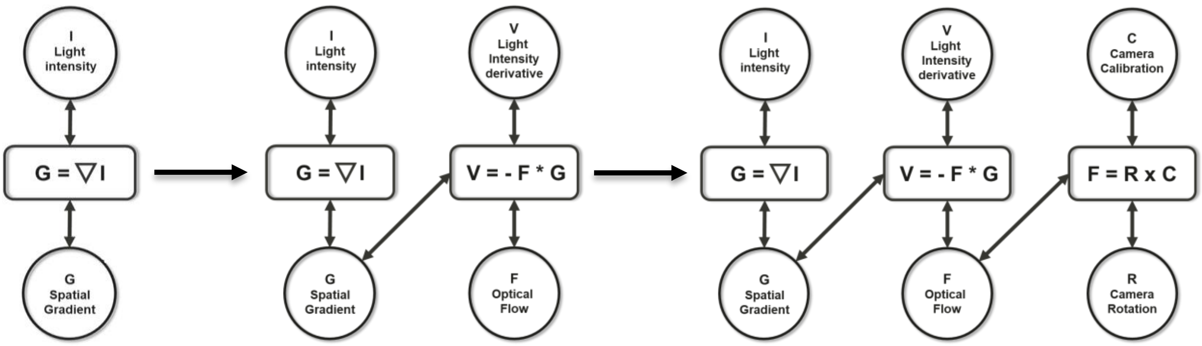

Figure 3: Learning invariant representations in the visual scene: Sample network architectures in which each circle is a sensory quantity and each square is a learnt relation using our model employing SOM and HL. Incrementally build a complex visual scene understanding from coupling pairs of sensors.

The network could contain, for example, optic flow map F, a light intensity map I, a spatial intensity gradient map G, a temporal intensity derivative map V, a camera calibration map C, and a single (non-mapped) estimate of the three-dimensional rotation R of the camera, which we assume to be fixed at a single point but free to rotate, like an eye in its socket.
As one can see, the relationships between these SOMs, as shown mathematically in the rectangles in Fig. 1, are (i) that G should be the gradient of I, (ii) that spatial variation G in brightness should match time variation V according to the local optic flow F, and (iii) the optic flow F at each point should correspond to the camera motion R with respect to the direction C that the pixel is aimed in. Together, these relations simply express the idea that the input should be explainable in terms of some kind of camera rotation in front of some kind of image.

In our view, such a problem boils down to actually learning invariant representation, such as those relating the SOMs for each sensor through HL. We think that we only have available the input derivative V for our network. Given the input V, it is not possible to simply solve for F and G. The only constraint on the vectors F and G at any pixel is that their scalar product should be -V. This is a very weak constraint, eliminating only one out of the four degrees of freedom in F and G at each pixel. Even if either F or G is known, the other is still under-constrained, being limited only to a line of possibilities. When trying to solve for F, this is known as the aperture problem. The other constraints, namely that the optic flow F at each pixel should be consistent with some overall camera rotation R, and that G should be the gradient of some map I (i.e., a conservative vector field), clearly do not constrain R or I at all. Thus all of the constraints in the system are weak constraints, and it is a priori not clear that the system will be able to find a correct interpretation of the input.

We believe that developing our model to implement such a network is a step towards learning invariant representations. More precisely, this small visual system is simple enough that it is easy to implement and verify for correctness, yet complex enough to solve non-trivial problems such as inferring the grayscale image or the optical flow, as shown in Figure 4. As we might not know what are the mathematical relations describing quantities in our visual scene, we think that our model can learn such mathematical relations directly from the data in an unsupervised manner using SOM and HL.

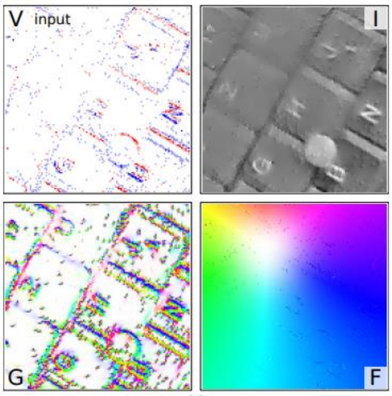

Figure 4: Sample network data after learning the underlying relations

Implementing the unified IRENA for visual scene understanding

In the context of proving that IRENA can cope with high-dimensional, rich and structured input as found in visual scenes we performed a series of exploratory experiments. The sensory quantities used in the experiments are the ones depicted in the network in Figure 3. We, hence, consider: light intensity (I), light intensity derivative (V), spatial gradient (G), and optical flow (F).

<u>Training the system</u>

In order to train the system we generated sensory data corresponding to the quantities in the network depicted in Figure 3. For this we used live images / video frames and applied traditional algorithms from OpenCV (e.g. Sobel, gradient) to generate our sensory quantities.

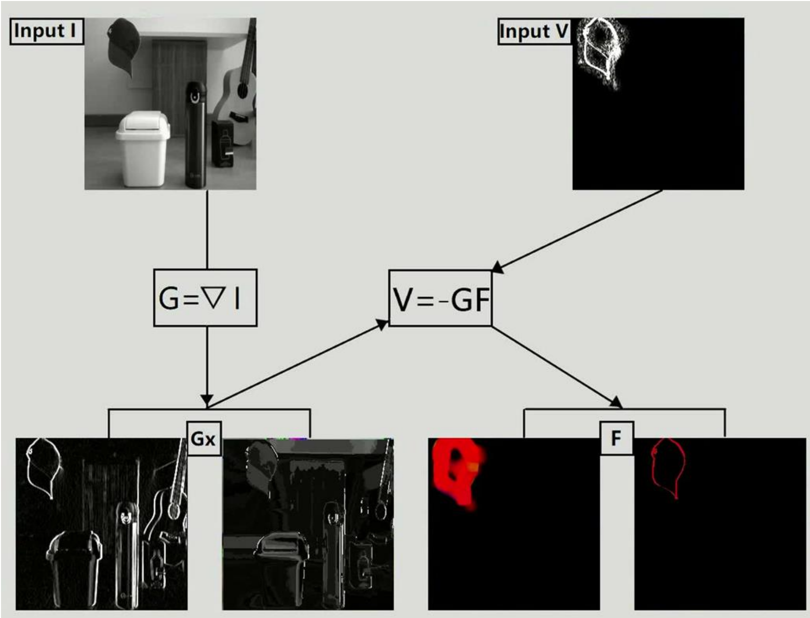

Figure 5: Learning the visual scene: given sensory data the system learns the underlying relations unsupervisedly

Figure 5 shows the training process of the system. We remind the reader that the network is the one depicted in Figure 3 (center) and we feed each sensory SOM with the respective sensory data we generated for a naturalistic visual scene. The network is fed with all sensory quantities simultaneously, and the network extracts the underlying (hidden, unknown) relations among the sensory data streams. The system has no prior information about he hidden relations among the sensory data describing the visual scene.

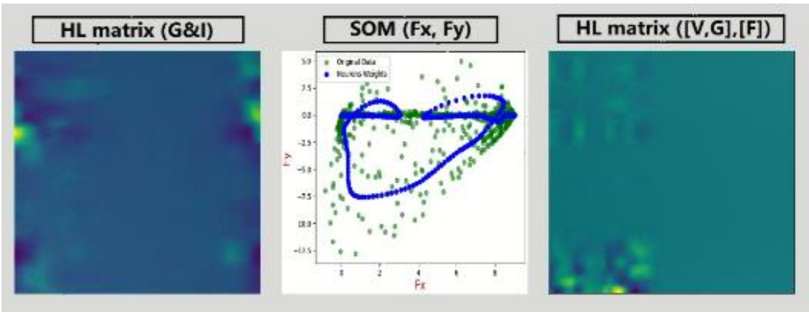

Figure 6: Learning the visual scene: The SOM and HL weights dynamics during training

We probed the weights of the input SOM networks corresponding to the sensory inputs and we can see that the SOM are learning the underlying distribution of the data in the input frames (see the optic flow SOM, Fx and Fy plot in Figure 6). Moreover, the correlation learning among the sensors (e.g. Hebbian linkage between spatial gradient and light intensity – Figure 6 left panel) is offering a hint about the highly nonlinear relations among the sensory quantities.

<u>Inference</u>

After training the system can exploit the learned relations to infer missing or noisy sensory quantities describing the scene. **This is a core feature of IRENA, the capability to learn underlying invariant sensory relations in order to infer missing ones, de-noise**

eventually corrupter sensory streams and fuse the sensory data into a globally consistent representation of the visual scene.
Such inference capabilities are described in Figure 7, where given the input I and V the system exploits the relations between spatial gradient, light intensity and light intensity derivative to infer optic flow.

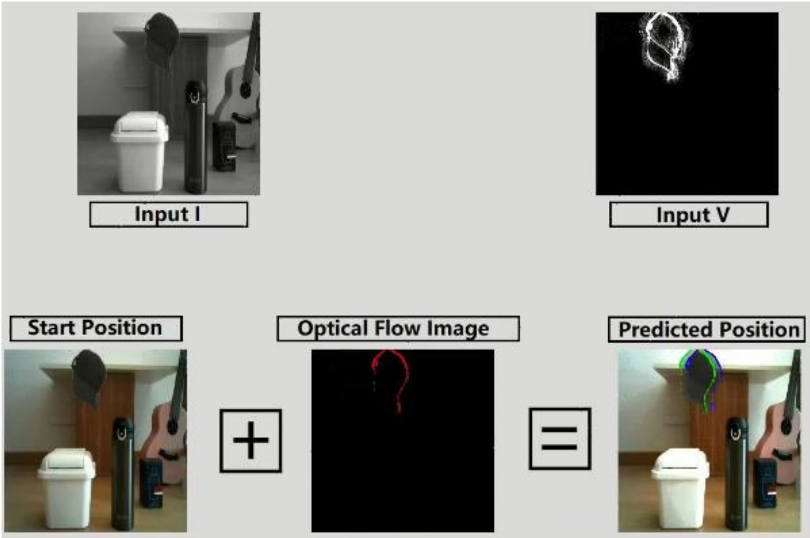

Figure 7: Learning the visual scene: given sensory data in only some modalities (I and V) the system infers the optic flow autonomously

IRENA offers a computational layer for extracting sensory relations for rich visual scenes, withy learning, inference, denoising and sensor fusion capabilities. The system is also capable, through its underlying unsupervised learning capabilities, to embed semantics and perform scene understanding.

<u>Semantics capabilities of IRENA</u>

IRENA's is capable, through its computational mechanisms, of object classification. The SOMs are used to learn the distribution of the input data, and then use SOMs as the input of HL to learn the correlation between SOMs. Actually, the essence of SOM is also basically able to cluster and subsequently classify the input data. Here, we give two example to show that our model actually include the object classification. At the moment such high level computation was not added to the visual scene understanding. This can be easily added thorough the same mechanism, exploiting the underlying relations, for example the shape of the object to track through optical flow and priors about the shape (i.e. apply Gabor filtering for certain object / texture orientations). **Semantics for clustering and classification come for free in IRENA and work directly at the input level offering a complex understanding of the input space.**

Experiment on object color clustering for subsequent classification:

1- Designed a 2-D SOM (size = (20, 30)) network.
2- Generated input vectors encoding RGB colors (e.g. 'black', 'blue', 'darkblue', 'skyblue', 'greyblue', 'lilac', 'green',red', 'cyan', 'violet', 'yellow', 'white','darkgrey', 'mediumgrey', 'lightgrey')

3- After 500 epochs training, we can get the final SOM picture below. We can see the neurons are representing the input space distribution (RGB space) well. It proves that our model have the ability of clustering by preserving the input space topology.

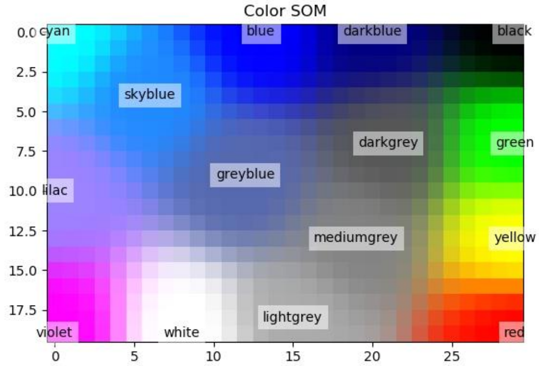

Figure 8: Semantics in IRENA: SOM clustering capabilities

This experiment demonstrates the capabilities of the SOM to cluster the input data in the same substrate used later in IRENA's correlation learning.

Experiment on MNIST classification:

1- Design a 2-D SOM (size = (30, 30)) network.

2-Download MNIST data, and then choose 100 numbers (0-9) as the input of the SOM.

3-After 200 epochs training, we can get the distribution of final SOM like in Figure 9.

4-Input 10 test data points. We can see the corresponding locations are activated. Each of the plotted MNIST test digits is well fitted into the target locations. Although digit 2 and 0 are not completely split, it seems to be resolved if the number of training iteration increases.

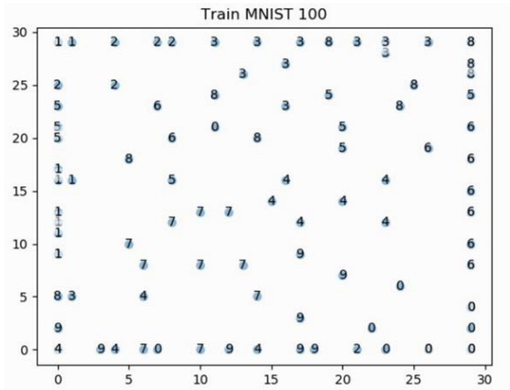

Figure 9: Semantics in IRENA: Training SOM for clustering and classification in MNIST

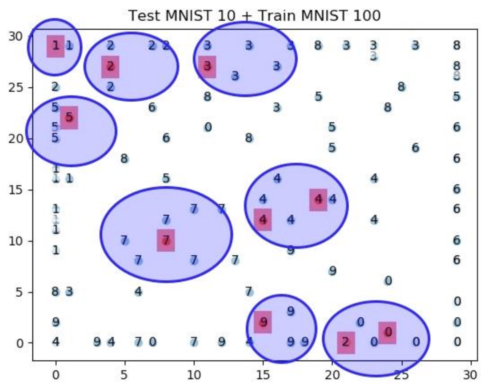

Figure 10: Semantics in IRENA: SOM capabilities in clustering and classification in MNIST

Figure 10 depicts the capabilities of the SOM to classify while still preserving input space topology (i.e. 2, 3, 5 are close to each other, whereas 4, 7, 9 are closer in the latent representation space). **This experiment shows that the SOM can classify the highly variable input data (i.e. handwritten text) exploiting the underlying topology preserving and the competition and cooperation between the neurons. This is a feature emerging without other costs from the computational model we employ.**

Code for the extended (semantics extraction) visual scene interpretation is available in NeuroTHIx_IRENA_codebase_final.zip

Conclusion

Our model provides an answer to the list of questions of the challenge:

• *Is the mechanism which was described above biologically plausible? Can it be connected to anatomical details of the brain?*

Yes, the model uses neural circuits implementing competition and cooperation in topology preserving populations of neurons. Furthermore, is using Hebbian learning as a means to learn correlations among the activation patterns of neurons in the population code. Such circuitry is describing various cortical areas, as we described also in the Preamble section.

• *Can you extend the model towards one which is (1) biologically plausible, (2) shows a relation to the creation of invariant representations (in the spirit of the universal cortical algorithm) and (3) does not only lead to location neurons on a hexagonal grid but to true grid neurons?*

In the basic formulation and implementation we provided, we show that the selected neurally inspired mechanisms have the potential to provide a system for representation and computation to learn invariant representations in low-dimensional spaces - as shown in Section “Introducing the core model and system”.
Our extended experiments demonstrated that the unified framework for invariant representations can be applied to learning visual scenes representations, include semantics for object classification and clustering and can be further extended to full 3D visual scenes - vision emphasized in the Section “Unified framework for invariant representations”.

Bibliography

[1] Fujita, Ichiro, et al. "Columns for visual features of objects in monkey inferotemporal cortex." Nature 360.6402 (1992): 343.

[2] Wyss, Reto, Peter König, and Paul FMJ Verschure. "Invariant representations of visual patterns in a temporal population code." Proceedings of the National Academy of Sciences 100.1 (2003): 324-329.

[3] Logothetis, Nikos K., and David L. Sheinberg. "Visual object recognition." Annual review of neuroscience 19.1 (1996): 577-621.

[4] Tanaka, Keiji. "Mechanisms of visual object recognition: monkey and human studies." Current opinion in neurobiology 7.4 (1997): 523-529.

[5] Rolls, Edmund T. "Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition." Vision: The Approach of Biophysics and Neurosciences. 2001. 366-395.

[6] Perrett, David I., and Mike W. Oram. "Neurophysiology of shape processing." Image and Vision Computing 11.6 (1993): 317-333.

[7] Wallis, Guy, and Edmund T. Rolls. "Invariant face and object recognition in the visual system." Progress in neurobiology 51.2 (1997): 167-194.

[8] Riesenhuber, Maximilian, and Tomaso Poggio. "Hierarchical models of object recognition in cortex." Nature neuroscience 2.11 (1999): 1019.

[9] Bednar, James A., and Stuart P. Wilson. "Cortical maps." The Neuroscientist 22.6 (2016): 604-617.

[10] Westermann, G.; Mareschal, D.; Johnson, M.H.; Sirois, S.; Spratling, M.W.; Thomas, M. Neuroconstructivism. Dev. Sci. 2007, 10, 75–83.

[11] Kohonen, T. Self-Organizing Maps; Wiley: Hoboken, NJ, USA, 2001.

[12] Chen, Z.; Haykin, S.; Eggermont, J.J.; Becker, S. Correlative Learning: A Basis for Brain and Adaptive Systems; Wiley: Hoboken, NJ, USA, 2007.

[13] Ganguli, D.; Simoncelli, E.P. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural Comput.* **2014**, *26*, 2103–2134.

[14] Tovar, Ángel E., and Gert Westermann. "A Neurocomputational Approach to Trained and Transitive Relations in Equivalence Classes." Frontiers in psychology 8 (2017): 1848.