Sparse coding

From Scholarpedia

Peter Foldiak and Dominik Endres (2008), Scholarpedia, 3(1):2984. revision #37405 [link to/cite this article]

Curator: Dr. Peter Foldiak, University of St Andrews, U.K.
Curator: Dr. Dominik Endres, University of St Andrews, UK

Mammalian brains consist of billions of neurons, each capable of independent electrical activity. Information in the brain is represented by the pattern of activation of this large neural population, forming a neural code. The neural code defines what pattern of neural activity corresponds to each represented information item. In the sensory system, such items may indicate the presence of a stimulus object or the value of some stimulus parameter, assuming that each time this item is represented the neural activity pattern will be the same or at least similar.

One important and relatively simple property of this code is the fraction of neurons that are strongly active at any one time. For a set of N binary neurons (which can either be 'active' or 'inactive'), the average (i.e., expected value) of this fraction across all information items is the sparseness of the code.

This average fraction can vary from close to 0 to about 1/2. Average fractions above 1/2 can always be decreased below 1/2 without loss of information by replacing each active neuron with an inactive one, and vice versa. **Sparse coding** is the representation of items by the strong activation of a relatively small set of neurons. For each stimulus, this is a different subset of all available neurons.

Contents

Local codes

At one extreme of low average activity fraction are local codes, in which each item is represented by a separate neuron or a small set of neurons. This way there is no overlap between the representations of any two items in the sense that no neuron takes part in the representation of more than one item. An analogy might be the coding of characters on a computer keyboard (without the Shift and Control keys), where each key encodes a single character. Note that locality of coding does not necessarily imply that only one neuron encodes an item, it only says that the neurons are highly selective, corresponding to single significant items of the environment (e.g. "grandmother cell").

This scheme has the advantage that it is simple and is also easy to decode. In a neural network, it is possible to make associations between a locally encoded item and any output by Hebbian strengthening of the synaptic connections between their neural representations in a single trial.

Local coding avoids the risk of unwanted interference between separately stored memories. It is also often possible to represent multiple locally coded items simply by the combined activity of the corresponding neurons without ambiguity.

However, there are many disadvantages to local coding: its representational capacity (here and in the following: the number of distinct representable information items) is low; that is, N binary neural units can only represent a maximum of N different items. This is a very small number compared to the number of stimuli we need to be able to discriminate even if N is comparable to the number (billions) of neurons in the human brain. Representational capacity also limits memory capacity, as no different outputs can be assigned to states that are indistinguishable. Another drawback of local coding is that as each representation is totally distinct from all others, associations linked to one item do not generalize to any other items when this would be beneficial. Generalization is an essential and widely observed behavior (e.g., McLaren and MacKintosh (2002)); if we have good discrimination, we are unlikely to ever encounter situations that are indistinguishable from each other. Therefore we need to apply our experiences from different situations having relevant aspects.

The lack of interaction between locally encoded items may be both an advantage, when it avoids unwanted interference, and a disadvantage, when such interactions are necessary for useful generalization.

Dense distributed codes

The other extreme (average activity ratio ≈ 0.5) corresponds to dense, or "holographic" coding. Here, an information item is represented by the combination of activities of all neurons. For N binary neurons, that implies a representational capacity of 2^N. Given the billions of neurons in a human brain, 2^N is beyond astronomical.

As the number of neurons in the brain (or even just in a single cortical area, such as primary visual cortex) is substantially higher than the number of receptor cells (e.g. in the retina), the representational capacity of a dense code in the brain is much greater than what we can experience in a lifetime (the factor of the number of moments in a lifetime adds the requirement of only about 40 extra neurons). Therefore the greatest part of this capacity is redundant.

Decoding a dense representation can be difficult as the whole neural population must be taken into account. For populations larger than a few thousand neurons, this goes beyond the capabilities of single neurons and may require a multilayer network. Dense encoding also implies that only one pattern can be represented at a given time. Superimposing more than one pattern makes the code hard to decode and may even introduce ambiguities (see Binding Problem) and interference. Dense codes also limit the number of memories that can be stored in an associative memory by simple, local learning rules (e.g. Hebbian learning). The mapping between such a pattern and an output is usually complex (not linearly separable) and cannot be learned easily and quickly by biologically plausible mechanisms. On the plus side, dense codes may facilitate good generalization performance and the high redundancy may lead to high fault tolerance to the failures of some of the units or connections.

Sparse codes

Sparse codes (small average activity ratio) are a favourable compromise between dense and local codes (Földiák 2002). The small representational capacity of local codes can be remedied with a modest fraction of active units per pattern because representational capacity grows exponentially with average activity ratio (for small average activity ratios). Thus, distinct items are much less likely to interfere when represented simultaneously. Furthermore, it is much more likely that a single layer network can learn to generate a target output if the input has a sparse representation (Willshaw and Dayan 1990). This is due to the higher proportion of mappings being implementable by a linear discriminant functions ('linear separability', see Perceptron). Learning in single layer networks is therefore simpler, faster and substantially more plausible in terms of biological implementation. By controlling sparseness, the amount of redundancy necessary for fault tolerance can be chosen. With the right choice of code, a relatively small amount of redundancy can lead to highly fault-tolerant decoding. For instance, the failure of a small number of units may not make the representation undecodable. Moreover, a sparse distributed code can represent values at higher accuracy than a local code. Such distributed coding is often referred to as coarse coding.

A code with a given average sparseness can contain codewords of varying sparseness. If the goal is to maximize sparseness while keeping representational capacity high, a sensible strategy is to assign sparse codewords to high probability items and more distributed codewords to lower probability items. This implements codes with with low energy expenditure, since neural activity consumes energy. Alternatively, if one only stored

the identities of the active units, the resulting code would have a short average description length (Cover and Thomas 2006). Several aspects of perceptual learning could be explained by the predicted increase of sparseness of the encoding of items with increased probability.

Measures of sparseness

Measures of sparseness are based on various measures of peakedness of the distributions of activities. Consider a set of N units indexed by i $(i = 1, \ldots, N)$ having activities x_i^k for the k^{th} $(k = 1, \ldots, K)$ activity pattern.

Sparseness can be calculated separately for each individual activity pattern k calculated across the set of units - also known as "population sparseness" (Willmore & Tohurst, 2001). Often these values are then averaged across all patterns to get an average sparseness.

Alternatively, a measure of selectivity can be calculated separately for each single neuron i across all patterns - also known as "lifetime sparseness" (Willmore & Tohurst, 2001). These values can be averaged across the population of units to give an average selectivity for the population.

The average sparseness and the average selectivity, however, are not necessarily equal, except when the measure is a linear functional of the activity distribution across the units.

Examples of sparseness measures:

Average activity ratio

If we consider neurons to be either active or inactive (binary), the simplest measure of sparseness is the average fraction of active neurons. Binary activity values can derive e.g. from observing spiking neurons for short time intervals (short enough so that at most 1 spike falls into any of these intervals), or by thresholding their spike counts in longer time intervals. In this case, average selectivity and average sparseness are equal (i.e. the total number of active units across the population and patterns divided by NK).

Monotonic (sub-)linear functions of the neural activity

For continuous-valued neurons, a penalty or cost function S can be defined (Olshausen and Field 1996, Harpur and Prager 2000) which has low values for sparse codes (it is the opposite of sparseness). Minimizing functions containing these terms are also useful for learning sparse codes (see Learning sparse codes section).

It is common to choose simply $S(|x_1^k|, ..., |x_N^k|) = \sum_i S(|x_i^k|)$, which can be interpreted as the negative logarithm of a prior distribution over independent code elements. Independence is desirable, because it helps minimize redundancy and maximizes the information capacity of the code (see Information Theory in Neuroscience and Independent Component Analysis for details). This type of cost function constrains selectivity of the units (i.e. the shape of the marginal distributions of the code elements) which by itself is not enough to guarantee sparseness. A simple counterexample would be a code of perfectly correlated code elements, i.e. $|x_1^k| = ... = |x_N^k|$ for all k. Thus, it is necessary to add another requirement. A natural one is representational accuracy, i.e. it should be possible to reconstruct the encoded information items from the code. If $S(|x_i^k|)$ is chosen so that the individual code elements are driven towards low entropies, then good reconstruction across all k will only be possible if the code elements are (near) independent a priori (Olshausen and Field 1996, Harpur and Prager 2000). In that case, the joint distribution of the code elements is given by the product of the marginals, and penalty functions of the above type can be adequate measures of code sparseness.

A code comprised of the activity of continuous-valued neurons is usually called sparse if the information in a given activity pattern is conveyed by a few large x_i^k, while most of the rest is near zero. Thus $S(|x_i^k|)$ must increase with $|x_i^k|$ to keep most values small. But to allow a few large $|x_i^k|$ values, $S(|x_i^k|)$ must be sublinear, or at most linear, i.e.

$$S(|x_i^k| + |x_j^k|) \leq S(|x_i^k|) + S(|x_i^k|)$$

Some examples of this type of sparseness penalty:

- $S(x) \propto |x|^a,\ 0 < a < 1$ is sublinear everywhere (Harpur and Prager 2000).
- The special case $S(x) \propto |x|$. In practice, this penalty will promote sparseness (Harpur and Prager 2000, Endres 2006) and it also corresponds to an exponential prior. Exponential distributions have been observed in neurons during natural stimulation, and they also optimise the information per spike ratio for non-negative responses.
- $S(x) \propto \log(1 + x^2)$ (Olshausen and Field 1996). While not sublinear for small $|x|$, its region of superlinearity can be shrunk arbitrarily by scaling x .

Kurtosis

Field (1994) used kurtosis as a measure of sparseness. Kurtosis excess is defined as

$$K_k = \frac{\langle (x_i^k - \mu_k)^4 \rangle_i}{\sigma_k^2} - 3$$

where $\mu_k = \langle x_i^k \rangle_i$ and $\sigma_k^2 = \langle (x_i^k - \mu_k)^2 \rangle_i$ are the expected value and variance of x_i^k . The expectation is taken across i, i.e. across all units in the neural population. The average kurtosis is then $K = \langle K_k \rangle_k$. Positive excess kurtosis values ("leptocurtic distributions") indicate that the distribution is "more peaked" than the normal distribution with a heavier tail. For neurons, this peak is expected to be near 0 activity, suggesting that a large fraction of neurons in the population are inactive or almost inactive, and therefore in certain circumstances this may be a good measure of sparseness. Kurtosis has only been studied for response distributions of model neurons where negative responses are allowed. It is unclear whether kurtosis is actually a sensible measure for realistic, non-negative response distributions.

Coefficient of variation

If the x_i^k are non-negative, another possible choice is the coefficient of variation

$$C_k = \frac{\sigma_k}{\mu_k}$$

or its closely related variant introduced by Treves and Rolls (Rolls and Tovee 1995)

$$T_k = \frac{\mu_k^2}{\langle (x_i^k)^2 \rangle} = \frac{1}{C_k^2 + 1}.$$

The latter is a measure of 'breadth of tuning' or 'density' where high values correspond to low sparseness. Often these measures are used to measure the degree of selectivity in neurons (sometimes somewhat confusingly called 'lifetime sparseness' or 'single-unit sparseness') rather than sparseness (sometimes also referred to as 'population sparseness').

Average selectivity and average sparseness, however, are related, as highly selective units tend to produce sparse codes. In the case of statistically independent units, these values will be equal.

Sparse codes have distributions concentrated near zero with a heavy tail, corresponding to significant responses of only a small number of neurons for a given stimulus, but those few neurons respond strongly. The measure T_k is 0.5 for an exponential distribution. Distributions that are more peaked at zero with a heavier tail have values less than 0.5, whereas distributions that are more evenly spread out have values greater than 0.5.

Linear vs. non-linear transformations for sparse coding

If a sparse code and a target output can be associated by a linear function, then it will also be possible to linearly associate any invertible linear transform of the code with the same target output. However, the sparseness of the code, as measured by any of the above functions, might be greatly affected by this transformation. One might therefore argue that an alternative definition of sparseness should be used here. According to this view, the sparseness of a code is the highest achievable sparseness value (using one of the former measures) by any invertible linear transform. This type of sparseness could therefore only be achieved by nonlinear transformations.

Sparse coding in the brain

One of the striking facts about single-cell physiological recording from sensory cortical areas is the difficulty of finding stimuli that effectively activate the neurons. This is true even at lower levels of sensory processing, such as in primary visual cortex, where the underlying stimulus parameters determining neural selectivity are often assumed to be known. This problem becomes much more severe in higher sensory areas, where the neural selectivity is more complex, and where even the relevant parameters are hard to define. These difficulties reflect the narrow tuning of cortical neurons which, when viewed on the population level, suggest a general sparse coding strategy. The selectivities determined experimentally also seem to match the probabilistically salient features of natural stimulation. This is what would be expected if the neural code conformed to information theoretic principles (Cover and Thomas 2006).

Field (1987, 1994) applied this idea to simple cells in primary visual cortex suggesting that basis functions limited both in space and frequency domains (such as Gabor functions) maximize sparseness when applied to natural images. Olshausen and Field (2004) give examples of sparse coding in other brain regions.

Neurons with complex selectivities in higher-level visual cortex (e.g. neurons selective to faces), could be interpreted as forming sparser codewords predicted for high probability items (e.g. faces) by the minimum description length principle.

Sparse coding is also relevant to the amount of energy the brain needs to use to sustain its function. The total number of action potentials generated in a brain area is inversely related to the sparseness of the code, therefore the total energy consumption decreases with increasing sparseness. Neural activity distributions observed experimentally can be approximated by exponential distributions, which has the property of maximizing information transmission for a given mean level of activity (bits/spike) (Baddeley et al. 1998).

Learning sparse codes

A number of algorithms have been able to find sparse encodings in neural network models. Földiák (1990) showed that anti-Hebbian lateral connections within a layer of nonlinear artificial neurons and Hebbian forward connections between these layers, combined with a local threshold control mechanism can learn a sparse code with low information loss, representing the inputs by the combination of uncorrelated components. Olshausen and Field (1996) and Harpur and Prager (2000) defined an explicit objective function that combined the requirement of high sparseness and low reconstruction error. The minimization of this function on natural images leads to a set of basis functions that resemble the localized receptive fields of simple cells in primary visual cortex. Figure 1 shows a 2 times overcomplete sparse neural basis learned by minimizing such an objective function (Endres 2006). The code provides sparseness and good reconstruction (see Figure 2).

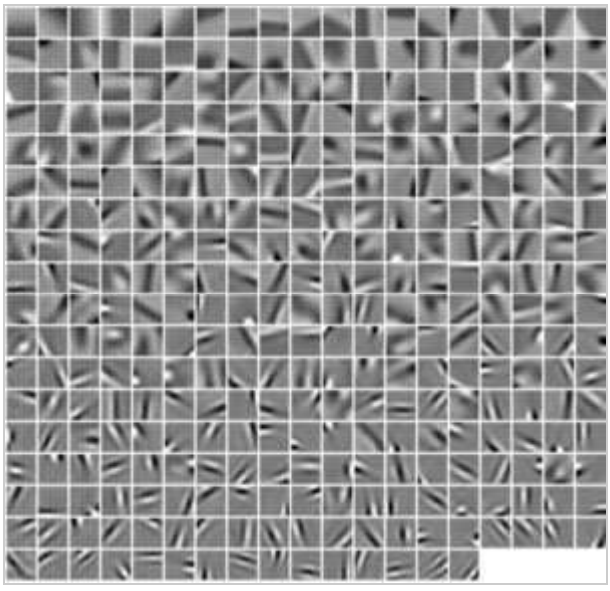

Figure 1: Sparsely encoding basis functions learned from natural images

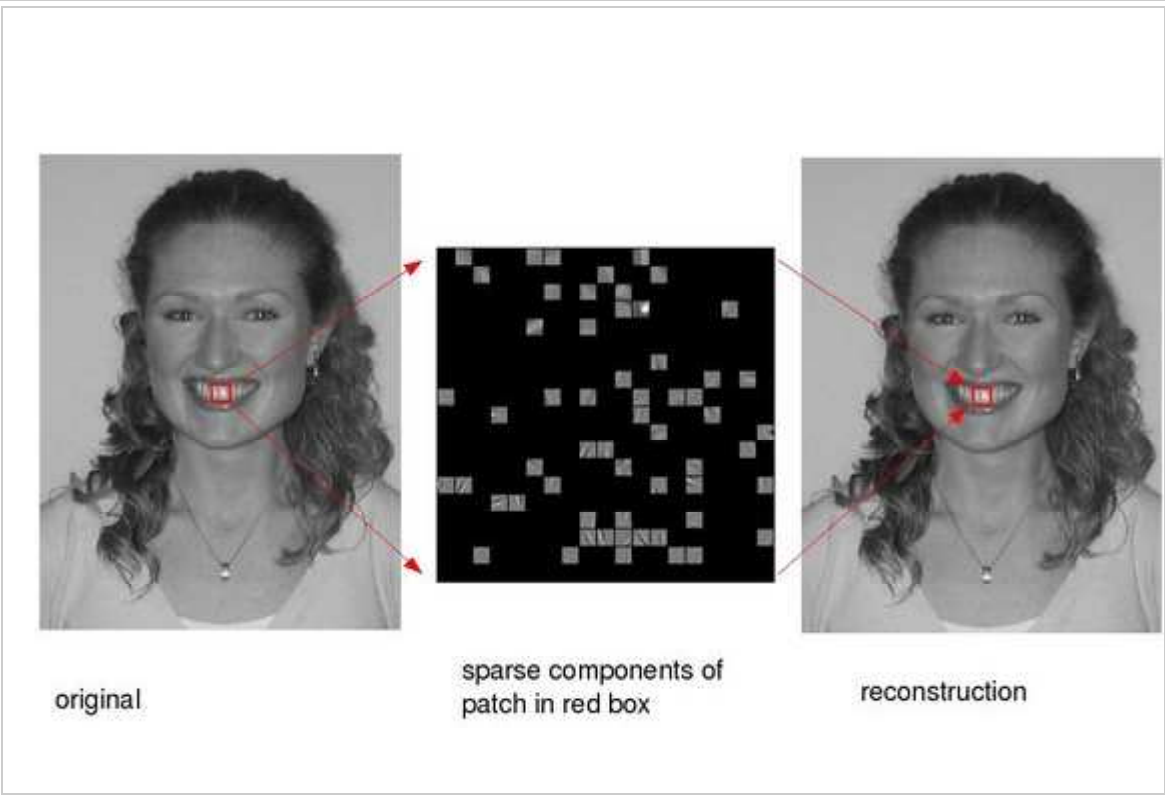

Figure 2: A natural image (left), encoded and reconstructed (right) with the basis functions from fig. 1. The sparse decomposition is exemplified on the image patch within the red square: most basis functions are not used in the encoding of a given patch, which is indicated by black areas in the basis function picture.

Alternatively, one can translate sparseness measures into prior probability distributions and reconstruction measures into likelihoods. Bayesian learning will then result in more principled way of trading off sparseness and reconstruction error Olshausen and Field (1997).

Factorial and explicit coding

In the case of distributed codes, codewords could be assigned randomly to the items that need to be encoded. Even with sparse codes, we could pick a random codeword from a set of low activity ratio codewords. Such a code, however, would not guarantee sensible generalization, as items that are subjectively similar may not get similar codewords. An alternative goal is to assign different parts of the codewords to different aspects of the items in a way that these aspects give useful generalizations. These aspects are then decodable locally, by considering only some parts of the code. This is especially important, as cortical neurons receive inputs from only a tiny fraction of all other neurons. A probabilistic view of this goal is that the probability of the combination of the different aspects (the whole encoded item) should be the product of the probabilities of the individual aspects of the item. This implies (by the definition of statistical independence) a representation that is factorial. Statistical independence (or near-independence) of the components is therefore a necessary (though not sufficient) condition for an "explicit code". Sparse coding is therefore closely related to Independent Component Analysis (Bell and Sejnowski 1997, Harpur and Prager 2000), and in fact under certain conditions they are formally equivalent (see Olshausen and Field (1997), Appendix A).

For instance, when encoding a scene consisting of separate and unrelated objects, an explicit code would consist of the superposition of the codes for the individual objects. In this case we can talk about the sparseness of the encoding of the individual components as well, not just that of the whole scene. We can consider the encoding of component items to be local or sparse, while the encoding of the whole, composite item would be more densely distributed. Sparseness itself, therefore, is not an ultimate goal of encoding. An item, such as a visual stimulus containing millions of pixels, will have a relatively small number of causes (e.g. parameters describing the objects in the scene and their positions). This sparse causal structure of the environment could be the reason why the heuristic of looking for sparse components is useful for finding the small number of causes of a given stimulus from a large number of potential causes.

Explicit coding is important as the correct generalizations tend to depend on only some of these components (e.g. presence of a 'tiger' component), while many other components (e.g. colour of leaves) may be irrelevant to the correct generalization (e.g. running away). An explicit code therefore reflects the similarity of the items in the

similarity of the codewords. Note that this "similarity" is not a property of the code itself; rather it depends on an interpretation of the code with respect to the world (and could be considered semantic).

Explicit codes also make the task of decoding components much easier, as only the relevant parts of the representation needs to be considered. Even more interestingly, the overlaps of the codewords could reflect the relationships between the encoded items implicitly.

References

- R Baddeley, L F Abbott, M C Booth, F Sengpiel, T Freeman, E A Wakeman, E T Rolls, 1998, Responses of neurons in primary and inferior temporal visual cotrices to natural scenes, Proc R Soc Lond B Biol Sci, 264:1775-1783.
- A J Bell, T J Sejnowski, 1997, The 'independent components' of natural scenes are edge filters, Vision Research, 37:3327-3338.
- T M Cover, J A Thomas, 2006, Elements of Information Theory, 22nd edition, Wiley-Interscience, ISBN 0471241954
- D M Endres, 2006, Bayesian and Information Theoretic Tools for Neuroscience (http://hdl.handle.net/10023/162) , PhD dissertation, University of St Andrews.
- D J Field, 1987, Relations between the statistics of natural images and the response properties of cortical cells, J Opt Soc Am A, 4(12):2379-2394.
- D J Field, 1994, What is the goal of sensory coding?, Neural Computation, 6:559-601.
- P Földiák, 1990, Forming sparse representations by local anti-Hebbian learning (http://www.st-andrews.ac.uk/%7Epf2/FoldiakSparseBC90.pdf) , Biological Cybernetics, 64:165-170.
- I P L McLaren, N J MacKintosh, 2002, Associative learning and elemental representation: II. Generalization and discrimination, Animal Learning & Behavior, 30(3):177-200.
- B A Olshausen, D J Field, 1996, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature, 381: 607-609.
- B A Olshausen, D J Field, 1997, Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research, 37:3311-3325.
- E T Rolls, M J Tovee, 1995, Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex, Journal of Neurophysiology, 73(2):713-726.
- B Willmore, D Tolhurst, 2001, Characterising the sparseness of neural codes, Network: Comput. Neural Syst. 12:255-270.
- D Willshaw, P Dayan P, 1990, Optimal plasticity from matrix memories: what goes up must come down, Neural Computation 2:85-93.

Internal references

- Valentino Braitenberg (2007) Brain. Scholarpedia, 2(11):2918.
- Howard Eichenbaum (2008) Memory. Scholarpedia, 3(3):1747.
- Erkki Oja (2008) Oja learning rule. Scholarpedia, 3(3):3612.
- John Dowling (2007) Retina. Scholarpedia, 2(12):3487.
- Ernst Niebur (2007) Saliency map. Scholarpedia, 2(8):2675.

Recommended reading

- P Földiák, 2002, Sparse coding in the primate cortex (http://www.st-andrews.ac.uk/%7Epf2/FoldiakSparseHBTNN2e02.pdf) , in The Handbook of Brain Theory and Neural Networks, Second Edition, pp 1064-1068, ed. Michael A. Arbib, MIT Press, ISBN 0-262-01197-2.
- G Harpur, R Prager, 2000, Experiments with low-entropy neural networks, in R Baddeley, P Hancock, P Földiák (eds.) Information Theory and the Brain, Cambridge University Press, pp 84-100, ISBN 0-521-63197-1.
- B A Olshausen, D J Field, 2004, Sparse coding of sensory inputs, Current Opinion in Neurobiology, 14:481-487.

External Links

- Peter Földiák's website (http://www.st-andrews.ac.uk/~pf2/)
- Dominik Endres' website (http://www.st-andrews.ac.uk/~dme2/)

See also

- Neuronal code, Receptive field, Vision, Independent component analysis, Saliency map,

Peter Foldiak, Dominik Endres (2008) Sparse coding. Scholarpedia, 3(1):2984, (go to the first approved version)
Created: 21 January 2007, reviewed: 28 January 2008, accepted: 28 January 2008
Invited by: Dr. Eugene M. Izhikevich, Editor-in-Chief of Scholarpedia, the free peer reviewed encyclopedia
Action editor: Dr. Eugene M. Izhikevich, Editor-in-Chief of Scholarpedia, the free peer reviewed encyclopedia

Categories: Computational Neuroscience | Neural Networks | Unsupervised Learning

---

- ISSN 1941-6016