Open access • Journal Article • DOI:10.3109/0954898X.2013.859323

A modular attractor associative memory with patchy connectivity and weight pruning. — Source link

Cristina Meli, Anders Lansner

Institutions: Royal Institute of Technology

Published on: 19 Nov 2013 - Network: Computation In Neural Systems (Taylor & Francis)

Topics: Attractor network, Content-addressable memory, Learning rule, Bcpnn and Hebbian theory

Department of Computational Biology
TRITA-CSC-CB 2013:01•ISRN KTH/CSC/CB--13/01--SE•ISSN 1653-5707

A modular attractor associative memory with patchy connectivity and weight pruning

Cristina Meli and Anders Lansner

Computational Biology (CB),
School of Computer Science and Communication (CSC)
Royal Institute of Technology (KTH)
S-100 44 STOCKHOLM, Sweden

A modular attractor associative memory with patchy connectivity and weight pruning

Cristina Meli and Anders Lansner

TRITA-CSC-CB 2013:01

Abstract

An important research topic in neuroscience is the study of mechanisms underlying memory and the estimation of the information capacity of the biological system. In this report we investigate the performance of a modular attractor network with recurrent connections similar to the cortical long- range connections extending in the horizontal direction. We considered a single learning rule, the BCPNN, which implements a kind of Hebbian learning and we trained the network with sparse random patterns. The storage capacity was measured experimentally for networks of size between 500 and 46K units with a constant activity level, gradually diluting the connectivity. We show that the storage capacity of the modular network is comparable with the theoretical values estimated for simple associative memories and furthermore we introduce a new technique to reduce the connectivity, which enhances the storage capacity up to the asymptotic value.

Keywords: Storage Capacity; BCPNN; Neural Network; Sparse Coding; Associative Memory; Hebbian Learning;

A modular attractor associative memory with patchy connectivity and weight pruning

Cristina Meli and Anders Lansner

Abstract

An important research topic in neuroscience is the study of mechanisms underlying memory and the estimation of the information capacity of the biological system. In this report we investigate the performance of a modular attractor network with recurrent connections similar to the cortical long- range connections extending in the horizontal direction. We considered a single learning rule, the BCPNN, which implements a kind of Hebbian learning and we trained the network with sparse random patterns. The storage capacity was measured experimentally for networks of size between 500 and 46K units with a constant activity level, gradually diluting the connectivity. We show that the storage capacity of the modular network is comparable with the theoretical values estimated for simple associative memories and furthermore we introduce a new technique to reduce the connectivity, which enhances the storage capacity up to the asymptotic value.

Introduction

The basic mechanism of information processing in the cortex is supposed to be a kind of association between concepts (Palm 1982, Rolls & Treves 1998). Information acquired through the senses is stored by creating an internal representation corresponding to a pattern of elevated neuronal activities. Later the brain is able to remember and recognize a pattern stored previously even if just a partial or distorted external stimulus is provided. This primary process has been extensively investigated giving rise to a specific type of artificial network, the Neural Associative Memory. Most of associative memory models are based on Hebb's theory of cell assemblies. Hebb hypothesized that synaptic connections increase their strength when two neurons are simultaneously active and when groups of neurons tend to be active at the same time and repeatedly, a kind of association can be assumed thus forming a cell-assembly (Hebb, 1949). These concepts have been formalized slightly different in the Willshaw-Palm and Little-Hopefield models (Willshaw et al., 1969, Palm 1980, Hopefield 1982), followed by more realistic attractor memory models including spiking neurons and conductance-based synapses (Fransen & Lansner 1995, Compte et al. 2000). The process of association can be realized through two different mechanisms: hetero-association which connects two different patterns and auto-association which connects one pattern to itself. A key issue concerns the amount of information that a system is able to store and recall correctly. The storage capacity of a network can be expressed in terms of number of patterns (pattern capacity) stored in the whole system or number of bits (information capacity) stored in the whole system or per synapse. Optimal values for the storage capacity have been estimated through the analysis of each specific model, using different methods. Without a global optimization procedure, the upper values obtained in the limit of large networks can be considerably lower. The main results have recently been summarized by Palm (2013).

Asymptotic bounds are typically computed for full connectivity and depend upon the specific learning rule, the task performed by the network and the activity level. The connectivity and the retrieval procedure can also affect the asymptotic value, but on a smaller scale (Bosh & Kurfess 1998, Schwenker, Sommer & Palm 1996). The storage capacity obtained for auto-associative memories with a local learning rule is half of the corresponding value achieved with hetero-association since the memory matrix is symmetric and contains roughly half of the information of an arbitrary matrix of the same size (Palm, 1992). In this report we investigate pattern completion of sparse Hebbian auto-associative memories, in different operating conditions. For such networks, the asymptotic completion capacity obtained with one step retrieval is, $\ln 2/4 \cong 0.17$ for binary storage and $1/8 \ln 2 \cong 0.18$ for additive storage (Willshaw et al. 1969, Palm 1980, Palm 1988, Palm 1991). Higher capacities have been obtained with iterative retrieval and a pattern recognition task, namely $\ln 2/2 \cong 0.34$ for the binary Hebb rule and $1/4 \ln 2 \cong 0.36$ for the additive Hebb rule (Palm & Sommer, 1992). An extensive study on the storage capacity of large auto-associative networks has been presented by Treves and Rolls (1991). Through the analysis of a wide class of models, with different Hebbian learning rules and architectures, they showed that the storage capacity is mostly affected by the sparseness of the neuronal representation and by the number of synapses per neuron rather than by the total number of units and connections in the system. They showed also that in the limit of sparse coding, the total number of patterns that can be stored in an auto-associative network with diluted connectivity is roughly proportional to the number of connections per neuron. . In the first section we briefly introduce attractor networks and the specific model we use, the Bayesian Confidence Propagation Neural Network (BCPNN). The second section outlines the new features we have introduced in the model, in the third section we present the details of the method and finally the results of the numerical simulations, followed by the discussion.

Attractor Networks and neocortex

Attractor networks are currently considered, both for their architecture and dynamical properties, one of the best abstract model of the cerebral cortex (Palm 1981, Hopefield 1982, Amit 1989, Rolls & Treves 1998, Lansner 2009). This type of network can generically be described as a system of N interacting units which converges over time towards a stable state belonging to some subspace. The time evolution can vary depending on whether the trajectories in the state space are strongly or weakly affected by the initial states. Different kind of equilibrium states or attractors can be reached such as fixed point, cyclic, line or chaotic attractors. In the simplest form of these neural networks, stationary fixed point attractors are identified with memories. When the network is in an attractor state, each unit persists in the same state over time, but activity is terminated by neural adaptation and synaptic depression. Attractor networks are thus a formalization of Hebbian learning, implemented in different ways through several learning rules, accounting for the complex phenomena leading to the changes in the synaptic efficacies. For this type of network, a key parameter is the size of the basins of attraction of the stored memories. The set of points in the states space from which the system evolves to an attractor, defines its basin of attraction. Similar patterns, corresponding to states lying in the same basin will be recalled as a single memory. Spurious stationary states can arise from the overlapping of patterns, leading to a loss of information since they can disturb the retrieval process. The idea that different parts of the brain act as an attractor network has diffused rapidly, supported by the experimental evidence of states similar to attractors found in cortical slices (Cossart et al. 2003, Shu et al. 2003). The architecture of this type of networks bears some resemblance to the structure of

cerebral cortex, especially its layers 2/3, characterized by many lateral recurrent connections between a huge number of similar units. The cortex is the largest single structure in the mammalian brain and it is assumed to be involved in high level functions such as memory, attention, sensory perception, thought and language. It consists of about $2 \cdot 10^{10}$ neurons in humans (Pakkenberg & Gundersen 1997), covering the outer part of the whole cerebrum like a thin sheet, about 3mm thick in humans (Hofman 1985). The basic elements are mostly excitatory pyramidal neurons (75-80%) and inhibitory interneurons (20-25%) organized in a modular structure (Braitenberg & Schüz 1998). This modularity reflects the way neurons are connected to each other. Both locally and over long distances, connectivity in the cortex is mostly recurrent and it appears to be locally denser in the vertical direction, giving rise to columnar structures. The average overall connectivity is very sparse, a total of about 10^{14} synapses in the human cortex gives a connectivity density on the order of 10^{-6} . The smallest observed module, the minicolumn, comprises some 100 neurons (Mountcastle 1997, Buxhoeveden & Casanova 2002, Amirikian & Georgopoulos 2003), while the hypercolumn is a larger module formed by a group of minicolumns, 100 minicolumns on average (Hubel & Wiesel 1977, Amirikian & Georgopoulos 2003). The minicolumn is sometimes considered the smallest repetitive building block of the cortex (Peters & Yilmaz 1993, Buxhoeveden & Casanova 2002). Although with some differences in the size and structure, the module is repeated throughout the volume. Evolutionary studies have shown that the increase in the volume has occurred through an increase in the number of minicolumns (Rakic 1995). In the horizontal direction cortex is organized in layers. These differ somewhat in thickness and composition of neurons and connections depending on location. In cortex, six main layers can be observed through the entire volume, with some exceptions as in those regions where layer IV is missing. The cortex is further partitioned into distinct functional areas: sensory, motor and association cortex, with a further subdivision within the single areas.

BCPNN

The BCPNN model implements a kind of Bayesian-Hebbian learning, based on probabilistic considerations and a derivation from Bayes rule (Lansner & Ekberg 1989, Lansner & Holst 1996). During the learning process, connections between simultaneously active units are increased in strength while those between anti-correlated units are weakened and may even become inhibitory, leading to the formation of new attractors, i.e. memories. The basic idea of the model is to express weights and their biases in terms of the probability of each unit being active and the probability of two units being active at the same time. The state of the network at a certain stage is thus conditioned by the previous states through Bayesian probabilities. The memory matrix is formed by a local learning rule:

$$w_{ij} = \log\left(\frac{P_{ij}}{P_i P_j}\right) \; ; \quad \beta_j = \log(P_j) \tag{1}$$

the weights w_{ij} and the biases β_j are computed from the probability P_i of the presynaptic unit being active, the probability P_j of the postsynaptic unit being active and the probability P_{ij} of the pre- and postsynaptic units being active at the same time. The retrieval is realized through an iterative process which updates units' activities according to biases and weights values.

The relaxation process ends when a stationary state is reached (i.e. an attractor). For each unit the support S_j is computed by a weighted sum of the inputs o_i added to the bias term (2), the potential m_j is computed from the support value through an ordinary differential equation (3):

$$S_j = \beta_j + \sum_{i=1}^{N} o_i w_{ij} \tag{2}$$

$$\tau_m \frac{dm_j}{dt} = S_j - m_j \tag{3}$$

and finally the output o_j of the post-synaptic unit is computed from the potential m_j through an exponential activation function. A complete derivation of equations 1-3 can be found in Lansner & Holst (1996) both for the single layer Bayesian network and multilayered network, i.e. a network with a modular structure. One simple way to estimate the probabilities P_i and P_{ij} is by counting the occurrences of activation (C_i, C_j) and co-activation (C_{ij}) of the units during the learning process:

$$P_i \cong \frac{C_i}{C} \;\; ; \;\; P_j \cong \frac{C_j}{C} \;\; ; \;\; P_{ij} \cong \frac{C_{ij}}{C} \tag{4}$$

where C is the total number of inputs. After the training, the weights and the biases are set to:

$$w_{ij} = \log\left(\frac{C_{ij}C}{C_i C_j}\right) \;\; ; \;\; \beta_j = \log\left(\frac{C_j}{C}\right) \tag{5}$$

In order to apply the same equations in the modular case, correlated units are grouped to form independent modules. The single module or "hypercolumn" represents a discrete coded attribute and the sum of its inner units' activities should sum to 1. A schematic view of the artificial unit is represented in Fig. 1.

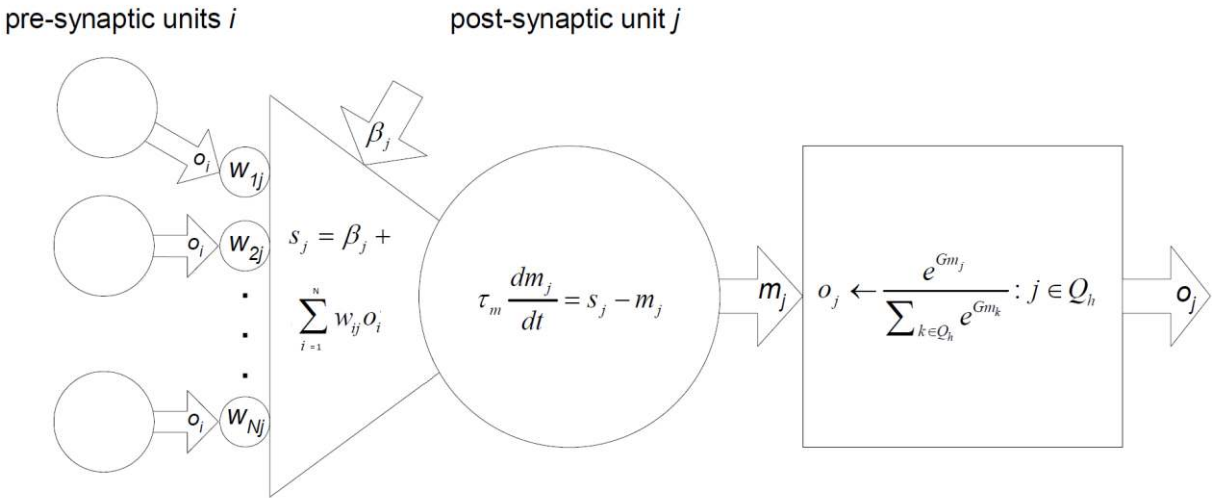

Figure 1.On the left, a schematic model of a single processing unit in a BCPNN is represented. The box on the right is relative to the whole module and Q_h *represents the set of units pertaining to it. The output of each unit in the hypercolumn is computed by passing its potential through a softmax function.*

Goals

The aim of this study is to investigate and improve the performance, in terms of storage capacity, of large recurrent networks up to about 50K units. We focused our analysis on the structure and the connectivity pattern of the network, introducing new features in the abstract model. First, we considered a modular network to represent the columnar structure observed in the cortex with long-range connections between hypercolumns, neglecting the local connections within each module. To flexibly regulate the activity level in the network we used silent hypercolumns, as previously proposed (Johansson & Lansner 2006). Second, we introduced a new technique to obtain a diluted network based on non-random removal of connections between single presynaptic units and clusters of postsynaptic units, resulting in "patchy connectivity". We want to verify if and how a change in the connectivity pattern can affect the storage capacity and also to investigate if it is possible to get closer to biology by introducing a criterion to prune connections, based on their effective contribution to the performance, hence obtaining a higher storage capacity. A theoretical estimation of the storage capacity for the BCPNN is beyond the scope of this paper. Here we present the results acquired through simulations and we discuss in a qualitative way the factors which possibly enhance the storage capacity.

Network Model

Our network has been implemented based mainly on the model presented by Johansson and Lansner (2006). They investigated the performance of a modular attractor network using different learning rules and introducing the concept of silent hypercolumns to vary the

activity level. In our model we used a single learning rule, the BCPNN, that was the best performing and we introduced new features related to the connectivity pattern, maintaining the same architecture and the same technique to regulate the activity level. We have previously also studied memory retrieval dynamics in a biophysically detailed spiking version of the model, with a similar hypercolumnar structure and minicolumns comprising different types of neurons and synapses (Lundqvist et al. 2006, Djurfeldt et al. 2008). In the following the key aspects of the model are reported, highlighting the differences with the previous study (Johansson & Lansner, 2006). First we describe the structure of the network and patchy connectivity, then we explain how the activity can be regulated using silent hypercolumns and finally we present a new technique to obtain the diluted weight matrix by weight pruning.

Patchy Connectivity

Long-range corticocortical and callosal connections, show mostly a point-to-surface mapping which differs in complexity for the extent and number of their terminal regions, both in the tangential and vertical directions. Axons spreading in the tangential direction over long distances, often form terminal clusters covering a region comparable to the size of a single or a group of columns (Goldman & Nauta 1977, DeFelipe et al. 1986, Gilbert & Wiesel 1989, Bosking et al. 1997). The branching geometry can be quite different, first in the number of branches originating from the single axon and hence in the number of terminal clusters and second in the structure of arborisation. Some axons show a kind of parallel architecture that can be related to simultaneous activation of different columns closely located, some others have a serial architecture with branches generating from a single trunk at larger distances from the parent cell body; sometimes both architectures are present in the same axon. The distribution of connections in the vertical direction is also variable: an individual cluster may terminate in one specific layer or involve more layers, and branches pertaining to the same axon can have different distributions. A detailed description of morphology is reported by Houzel, Milleret and Innocenti (1994). A definite relation between morphology and functional properties is not easy to establish, but similarities in the organization have been found for connections that are supposed to have similar functions (Gilbert & Wiesel 1969, Martin & Whitteridge 1984, Boyd & Matsubara 1991). Reproducing this complex structure with a single artificial system is complicated and requires several steps. We attempted a first step in this direction, introducing a modular structure and changing the connectivity pattern. In our model, identical binary units are grouped in modules of constant size and are connected to each other through recurrent, long-range patchy connections, with specialized local treatment of the connections within the module. Each module can be identified with a single hypercolumn and each unit with a single minicolumn. In the previous model (Johansson & Lansner, 2006), connections were implemented only between single units pertaining to different modules, while with “patchy connectivity” each presynaptic unit is connected to a group of terminal clusters through single connections and all minicolumn units pertaining to a module share the same input connections, though with individual weights. Thus, in our abstract model each processing unit represents a group of biological neurons and in the same way each connection between two units represents a group of synapses, even though sometimes we generically refer to it as a synapse in the following. A schematic view of patchy connectivity is shown in Fig. 2. It is important to verify if this change in the connectivity pattern, can affect the dynamical properties and the storage capacity of the system. Long-range connections have been previously examined in relation to the wiring cost of biological networks. In order to ensure compactness and low energy

consumption, i.e. the amount of wiring inside the cortex volume, long-range connections extending in the horizontal direction are supposed to be wired in a much more efficient way than randomly. Voges et al. (2009) investigated different spatial configurations and have shown that the optimal wiring is achieved with patchy-networks, characterized by a single long-range connection to each terminal cluster.

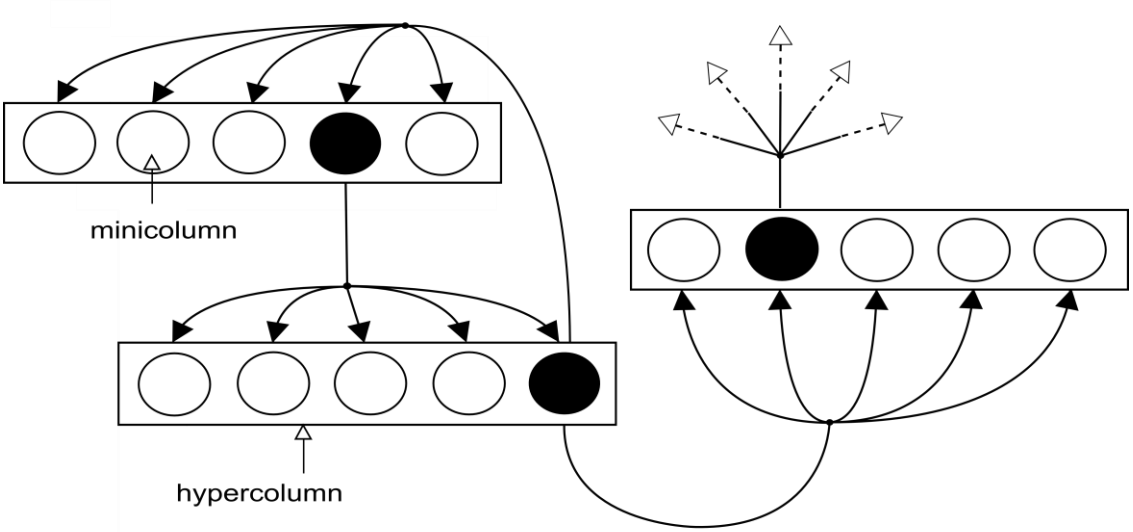

Figure 2.Schematic view of patchy connectivity in a network with hypercolumns. Each unit is connected to a number of terminal clusters and all the units in a module receive input from the same set of presynaptic units. Connections can be recurrent between hypercolumns and between single units pertaining to different modules, but all local connections within a hypercolumn are removed.

Silent Hypercolumns

In order to estimate the storage capacity of the brain, it is important to have an activity level comparable to the activity measured in the biological system. Many experiments in different cortical areas showed that only a small fraction of neurons are active in response to an external stimulus (Lennie 2003, Waydo et al. 2006). In a network with hypercolumns, a pattern of activity can be represented by a binary vector of size N, composed by H sub-vectors of size U. The local activity is normalized, provided that the sum of non-normalized activity values exceeds 1. The global activity in the network can be regulated through the introduction of silent hypercolumns. In our model, a silent hypercolumn can be seen as representing an attribute which is not relevant for and does not contribute to represent a particular stimulus. This interpretation follows from experimental studies on the activation patterns in the visual cortex (Cheng et al. 2001, Tsunoda et al. 2001), which have shown that single stimuli activate patches of neurons, each one covering an area corresponding to a hypercolumn and that the set of active hypercolumns, changes in response to different stimuli. Thus with silent hypercolumns it is possible to regulate the activity level more freely

than just by varying the module size (U). Moreover the number of units in the module should be consistent with the number of minicolumns in the biological network and therefore once the optimal size is reached, the usage of silent hypercolumns is preferred to further increase the sparseness of activity. Previous studies on attractor networks tested with uniformly distributed random patterns have shown that the storage capacity increases when sparse coding conditions are applied (Fulvi Mari & Treves 1998, Fulvi Mari 2004). However, the information content per pattern and hence the total amount of information that the network is able to store, decreases. In a modular network with the constraint of unary activity in each module, the information content per random binary pattern is equal to $I_p = H\ log_2(U)$ bits, where H is the number of hypercolumns and U the number of units in each module. If S hypercolumns are randomly set as silent, it becomes instead:

$$I_p = log_2(U^{H-S}\binom{H}{H-S}) \tag{6}$$

The storage capacity expressed in bits per connection can then be computed by:

$$M_{bits} = \frac{I_p P_{stored}}{N_s} \tag{7}$$

where P_{stored} is the total number of stored patterns and N_s is the total number of connections in the network, that can be approximated by :

$$N_s = cN^2 - HU^2 \tag{8}$$

where c is the connectivity density, N the total number of units and the last term accounts for not having connections within hypercolumns, which makes a negligible difference for large networks. In our implementation, a fraction of the hypercolumns is set automatically silent during pattern generation and in that case all the units in the module are OFF, as shown in Fig. 3 for a single pattern.

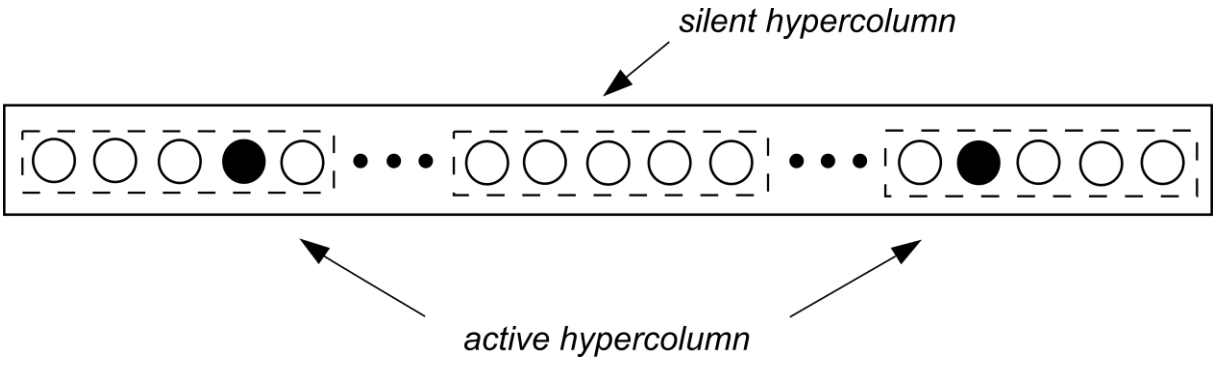

Figure 3.Schematic representation of a modular pattern with 5 units per hypercolumn

Weight Pruning

The total number of synapses in the adult brain is about constant over time, however alterations in the structure of synaptic connectivity are observed in the developing brain and also later related to learning and memory encoding. The formation of new synapses or their elimination is an activity-dependent process and is referred to as "structural plasticity" (Bailey & Kandel 1993, Muller et al. 2002, Lamprecht & LeDoux 2004). In a similar way, the connectivity pattern in an artificial network can be achieved through an activity-dependent weight pruning, aiming to remove the least useful connections. A standard technique to obtain a diluted weight matrix is random removal of connections. In Johansson & Lansner (2006) this technique was applied to remove single connections between two units and blocks of connections between two modules. We implemented a random and a non-random method removing single patchy connections, i.e. the connections between a presynaptic unit i and all the postsynaptic units j = 1, …, U in a hypercolumn. A schematic representation of the different methods is shown in figure 4. The non-random method has been developed in the attempt to realize more realistic connectivity patterns. The idea here is to estimate the overall impact of each presynaptic unit on the postsynaptic hypercolumn through a local score, e.g. a measure of the value of a patchy connection. In this way it is possible to operate a selection by removing the patches with a lowest value. Both with the random and non-random removal we obtain an asymmetric weight matrix, which however still allows for stable fixed point attractors in the operating conditions employed here.

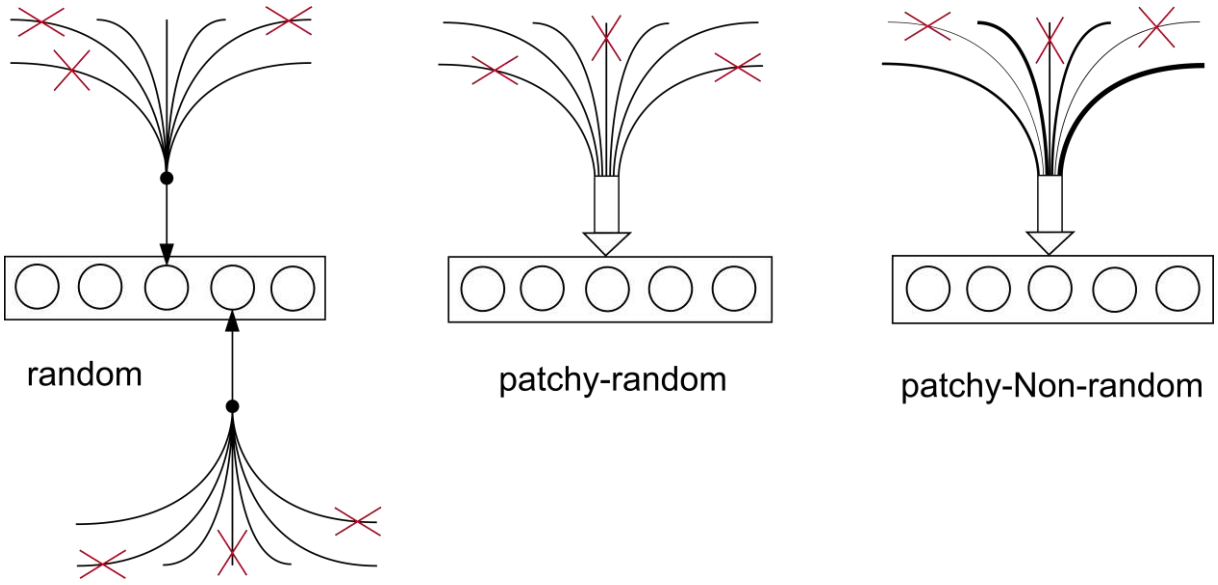

Figure 4.Schematic of standard random, patchy random and patchy non-random removal. With the standard random connections are removed independently for each unit, while with the patchy random and non-random the same set of connections are removed for each hypercolumn. For the non-random case, the score of patchy connections is represented by a different line thickness.

Methods

The storage capacity has been measured under different conditions for networks of increasing size, having the structure described above. We tested the effect of patchy connectivity and weight pruning in the very simple case of a non-spiking network trained with the BCPNN learning rule. We started to investigate small systems of 50 hypercolumns with 10 units per module, in order to compare with the systems investigated in the previous study (Johansson & Lansner 2006). The procedure adopted, is standard for testing the performance of an associative memory. In the following we describe in detail the learning and retrieval phases, and we explain how the storage capacity can be measured in this context.

Learning & Retrieval

In a non-spiking model, patterns can be stored by "one-shot learning", i.e. after a single presentation, by clamping the network in an attractor state automatically through the input, without dynamics. The training set is generated with a fixed fraction of silent hypercolumns and by random activation of a single unit in each active hypercolumn. After the whole training set has been encoded, weights and biases are computed according to (5). The diluted weight matrix is then obtained by removing a fixed percentage of patchy connections. The procedure is executed independently for each hypercolumn, through a random selection or by measuring the score of each patchy connection. In our implementation, the average strength of the patchy connections is evaluated by summing the absolute values of the weights:

$$M_1 = \sum_{j=1}^{U} |w_{ij}| \tag{9}$$

we also considered a weighted sum :

$$M_2 = \sum_{j=1}^{U} |P_{ij}\, w_{ij}| \tag{10}$$

and a number of different statistical measures based on the mutual information. However, the above measure (9) gave the best results and we did not test the others on the larger networks. The performance of the system was measured through pattern completion. The test patterns consisted of a subset of the training set, randomly chosen. A certain number of hypercolumns were selected to perform the retrieval without external input. The missing portion of the test pattern was restored based on the activities of the connected units according to (2). Each pattern was presented repeatedly until a stable state was reached, or a fixed number of iterations had been executed. If the final state perfectly matches the original stored pattern, the retrieval is considered as correct. The stability could then be tested through the spontaneous evolution of the network, after a single presentation of the input, by comparing the final state reached in a fixed number of iterations with the original stored pattern.

Estimation of the Storage Capacity

The storage capacity was measured by computing the percentage of patterns correctly retrieved. As described above, each single pattern must be identical to the original stored pattern and if the network is able to restore all the test patterns correctly, then the whole retrieval is considered 100% successful. The number of training patterns was increased until the percentage fell below 95%, which is our reference point. The curve can be steep or gentle depending on the operating conditions, as shown in Fig. 5. Each curve represents the average response of the system tested for a fixed set of parameters. In our implementation, the training set and the final configuration of the connectivity pattern are not preserved at the end of the simulation, so we average over many runs. Each point of a curve was obtained by repeating the same experiment 10 times. The pattern capacity is the crossing point at 95% and it is determined through a linear interpolation. We considered networks of size comprised between 500 and 46K units, with a constant activity level fixed to 0.1 and 0.04 and we decreased the connectivity to 3%, reaching the diluted regime[1] beyond 11K units. Thus the largest networks approached the operating conditions of the diluted networks studied by Treves and Rolls (1991), and we can compare our results with the pattern capacity predicted according to their formula (11), although it has been derived for networks without a modular structure:

$$P_{stored} \approx \frac{C^{RC}}{a\ln(\frac{1}{a})}k \qquad (11)$$

where C^{RC} is the number of synapses per unit, a the sparseness of the representation and k is a factor that gathers the characteristics of the network and is between 0.2-0.3 (see also Rolls (2012)).

[1] According to Treves and Rolls the diluted regime is reached when the ratio C^{RC}/N between the number of synapses per unit and the total number of unit in the network is in the range 0.1-0.01.

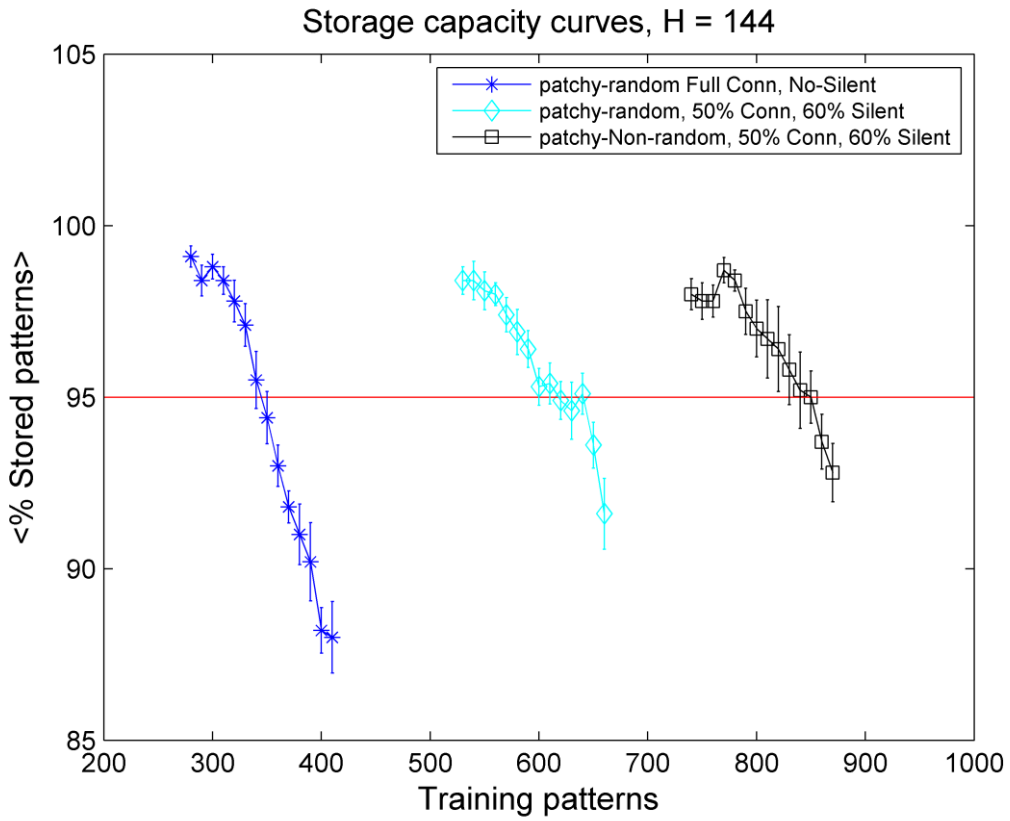

Figure 5: Crossing point for a network of 144 hypercolumns with full connectivity and without silent H (first on the left), with half connectivity and 60% silent H, using patchy random removal (in the middle) and non-random removal (on the right).

Implementation

In a non-spiking model, patterns can be stored by "one-shot learning", i.e. after a single presentation, by clamping the network in an attractor state automatically through the input, without dynamics. The model has been implemented in a parallel environment using the MPI library and all the simulations have been executed on a Cray machine (XE6 system). For a modular network, the simplest way to implement parallelism is to assign a module or a group of modules to a single CPU. The training and retrieval phases described above are executed locally by each CPU for the hypercolumns it is in charge of and all the activities and other data are exchanged through collective communications. The collective scheme is suitable for the training process since we have a fully connected system, while during the testing a considerable amount of unutilized data is exchanged. On the other hand, a single collective communication can be faster than a series of point to point communications when the size of the system is not too large and furthermore in this particular case the retrieval is performed on a subset of the training set. Thus we maintained the same scheme in both phases.

Results

In the following we present our results on the storage capacity, starting with small systems of 500 units up to large systems of 46K units. First we tested the effect of decreasing connectivity using random and non-random methods, for systems with a constant number of units in each hypercolumn (U=10). In the same way we tested the effect of increasing the number of silent hypercolumns, hence the sparseness of the representation, for different levels of connectivity. The same experiments for systems without silent hypercolumns are also presented, where the sparse coding limit is reached by changing the structure of the network, i.e. by fixing the size (N = HU) and increasing the number of units in each module. The scaling properties have been investigated for systems having about 1400 synapses per unit, the same as in most of the preliminary tests on small systems, gradually decreasing the connectivity to 3%.

4.2. Systems with sparse Connectivity

In Figs. 6 and 7 the storage capacity is plotted as a function of the connectivity for a system of 500 and 1440 units respectively. Each curve is related to a specific method and to a different percentage of silent hypercolumns, from zero to 60% with 20% increments. On the top of the figures, we compare the patchy random with the standard random method, on the bottom the patchy random with the non-random method. The systems have been tested with 50% occlusion, i.e. only half of the hypercolumns in the network receive input during the retrieval. From Figs. 6 and 7 it is evident that the larger system is stable in the interval investigated, while the smaller system shows a drop in the performance when the connectivity level is below 50%. This happens because the amount of active input connections to units drops below a critical value and the recurrent network can no longer complete the attractor states. However, the overall response of the two systems is quite similar. The storage capacity is higher when the sparseness of the representation increases. When the connectivity is reduced, the pattern capacity decreases and the information capacity shows a gentle raise until 20-30% of the connectivity for the larger system and until around 40-50% for the system of 50 hypercolumns. More interesting is the comparison between the different methods. The patchy random and the standard random removal, show a similar response for both systems while a visible difference is observed when the non-random method is used. The storage capacity is higher when connections are removed through a selection and there is a maximal difference at about 40% connectivity for the 144 hypercolumn case. However, the effect of the non-random pruning method tends to disappear when the connectivity is too low, showing a worse performance compared to the patchy random for the system with a higher percentage of silent hypercolumns. The curves in figure 6, representing the standard random method, reproduce earlier results for the BCPNN learning rule (Johansson & Lansner 2006).

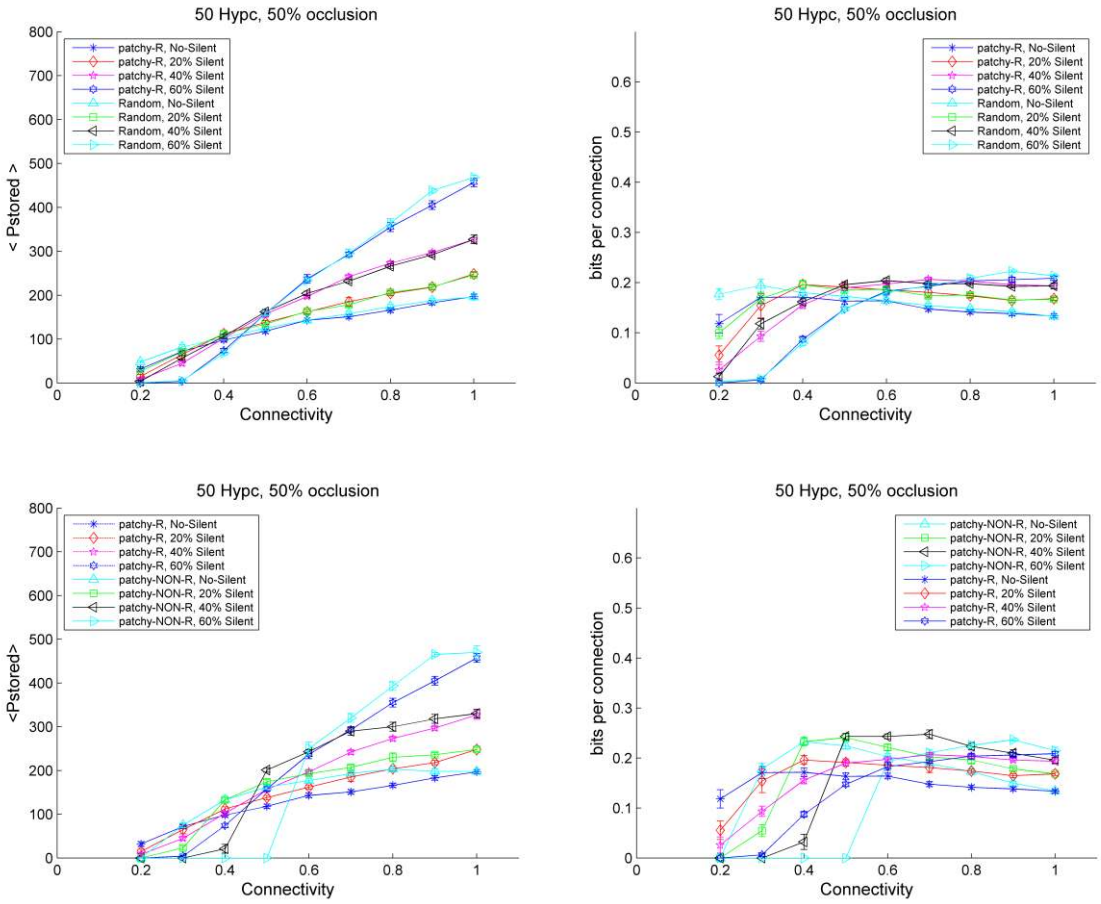

Figure 6: The storage capacity in terms of patterns (on the left side) and bits per connections (on the right) for a network of 50 hypercolumns. The patchy random method is compared with the standard random (on the top) and with the non-random (on the bottom).

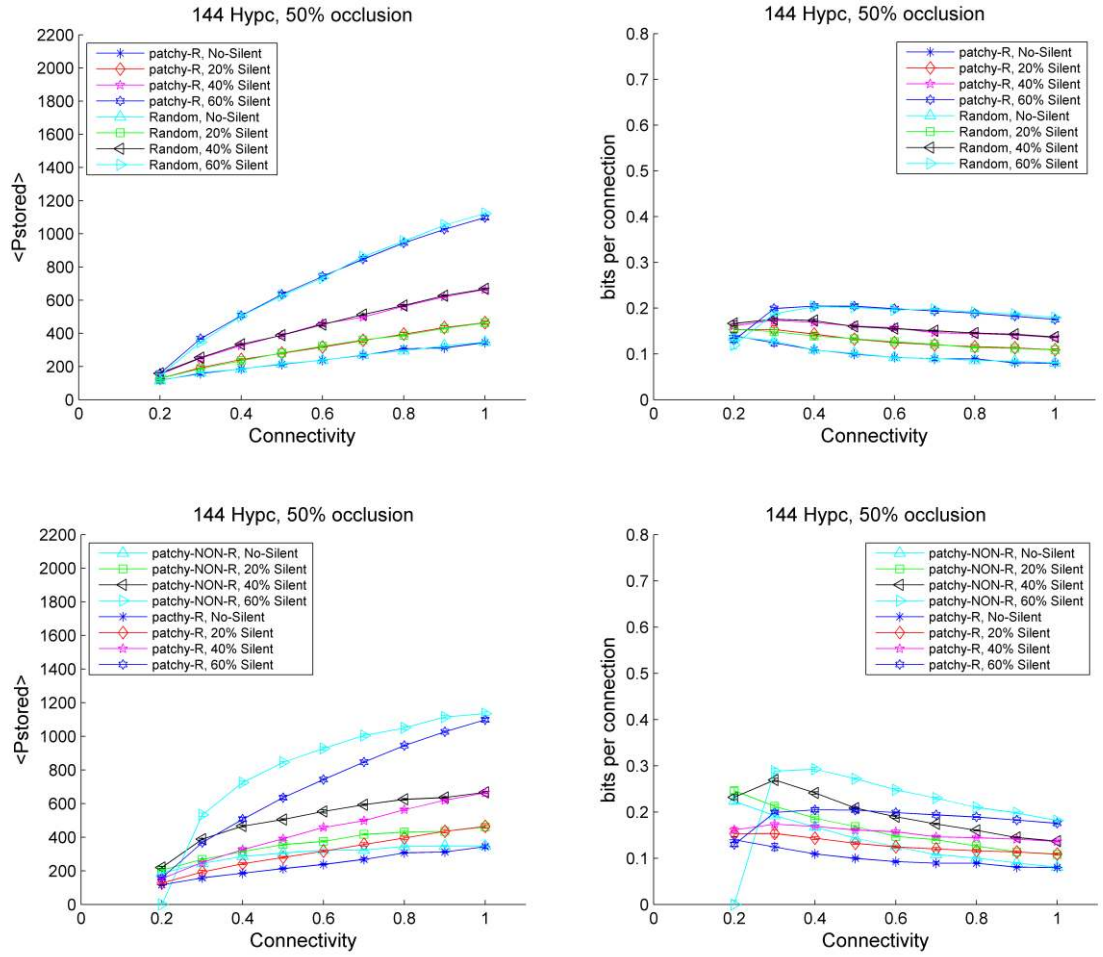

Figure 7: The storage capacity in terms of patterns and bits per connection for a network of 144 hypercolumns. In this case the system is more stable and it is possible to compare the different methods till the lowest level of connectivity.

Systems with sparse activity

In Fig. 8 the storage capacity is plotted as a function of the fraction of silent hypercolumns in the network, for three different systems of size 1440, 2880 and 5760 units respectively. We performed two tests by setting the connectivity in order to have about the same number of synapses per unit in the three systems, around 1400 and around 700 synapses per unit. We assigned exactly the same task to each system, keeping constant the number of hypercolumns to be filled in, equal to 43, which corresponds to 30% occlusion in the smallest system. First, we notice that the response of the three systems is about the same, when the random removal is applied (Fig. 8, top) while with the non-random method we observe three different responses with the largest system performing better (Fig. 8, bottom). This effect will be further discussed in the next section. As expected, the systems having a higher number of synapses per unit are able to store more patterns and the difference becomes larger as the number of silent hypercolumns increases (Fig. 8, left), while the systems having less synapses per unit can store more bits per connection (Fig. 8, right), until the total number of synapses, N_s , is the leading term in eq. (7). When the representation is too sparse, the storage capacity collapses and this happens first for the systems with less synapses per unit, when the fraction of silent hypercolumns is around 0.80. In terms of bits per connection the maximum is slightly

shifted, at around 0.70 for the systems with 700 synapses per unit and at around 0.85 for the systems with 1400 synapses per unit, due to the decreasing of the information content per pattern (eq. 6). A similar trend is observed for the non-random method (Fig. 8, bottom), with a steeper drop in the performance for the systems having less synapses per unit. As expected, the storage capacity is considerably higher for the non-random compared to the patchy random (full-connectivity curves) and the difference is larger as the sparseness increases, up to about 40% more. In a network with a modular structure, the activity level can be regulated by varying the number of units inside each module. In Fig. 9 the storage capacity is plotted as a function of the sparseness a, for two systems of the same size but different structure. One network consists of 288 hypercolumns with 10 units (S-system) the other network consists of 2880 units with variable number of hypercolumns (U-system). The size of the module has been gradually increased from 10 to 60 units, in order to compare with the same level of sparseness achieved while increasing the percentage of silent hypercolumns from 50% to 83%. The information content per pattern for the system with silent hypercolumns, S-system, is higher compared to the system without silent, i.e. the U-system:

$$I_p^U = \mathrm{H} \log_2 U \tag{12}$$

$$I_p^S = (\hat{H} - S) \log_2 \hat{U} + \log_2 \binom{\hat{H}}{\hat{H} - S} \tag{13}$$

$\hat{H}$ represents the constant number of hypercolumns and $\hat{U}$ their constant size in the S-system, while H and U are variable in the U-system with :

$$HU = \hat{H}\hat{U} = N \quad ; \quad H = \hat{H} - S \tag{14}$$

I_p^U can be rewritten in terms of the S-system parameters and by separating the logarithm we obtain:

$$I_p^U = \left(\hat{H} - S\right) \log_2 \hat{U} + \log_2 \left(\left(\frac{\hat{H}}{\hat{H} - S}\right)^{\hat{H} - S}\right) \tag{15}$$

which is always less than I_p^S . For a fixed activity level, the percentage of silent hypercolumns in the test portion is variable since they are set randomly, therefore the task performed by the two systems can be slightly different. We assigned a similar task to the networks by setting the number of test hypercolumns to a higher value for the U-system. For increasing values of the sparseness we kept constant the number of test hypercolumns in the S-system, while we

decreased the number of test hypercolumns in the U-system, since the information content per module increases. The connectivity was set to 50% for both systems, using random and non-random removal. Fig. 9 shows that the U-system performs better in terms of pattern capacity when the patchy random removal is applied (blue and red curves, left), while with the non-random method the difference to the S-system is reduced and the trend is reversed for sparseness above the value 0.03 (green and cyan curves, left). In terms of information capacity, we observe a similar response for the patchy random (blue and red curves, right), while with the non-random method the S-system performs better and the difference to the U-system increases with the sparseness (green and cyan curves, right). This effect suggests that the non-random removal is less powerful in the U-system when the size of the module is comparable with the number of hypercolumns in the network. To summarize, we do not observe any important effect that prevent the use of silent hypercolumns and we believe that for larger networks, a combination of the two methods should be applied to reach the required level of sparseness.

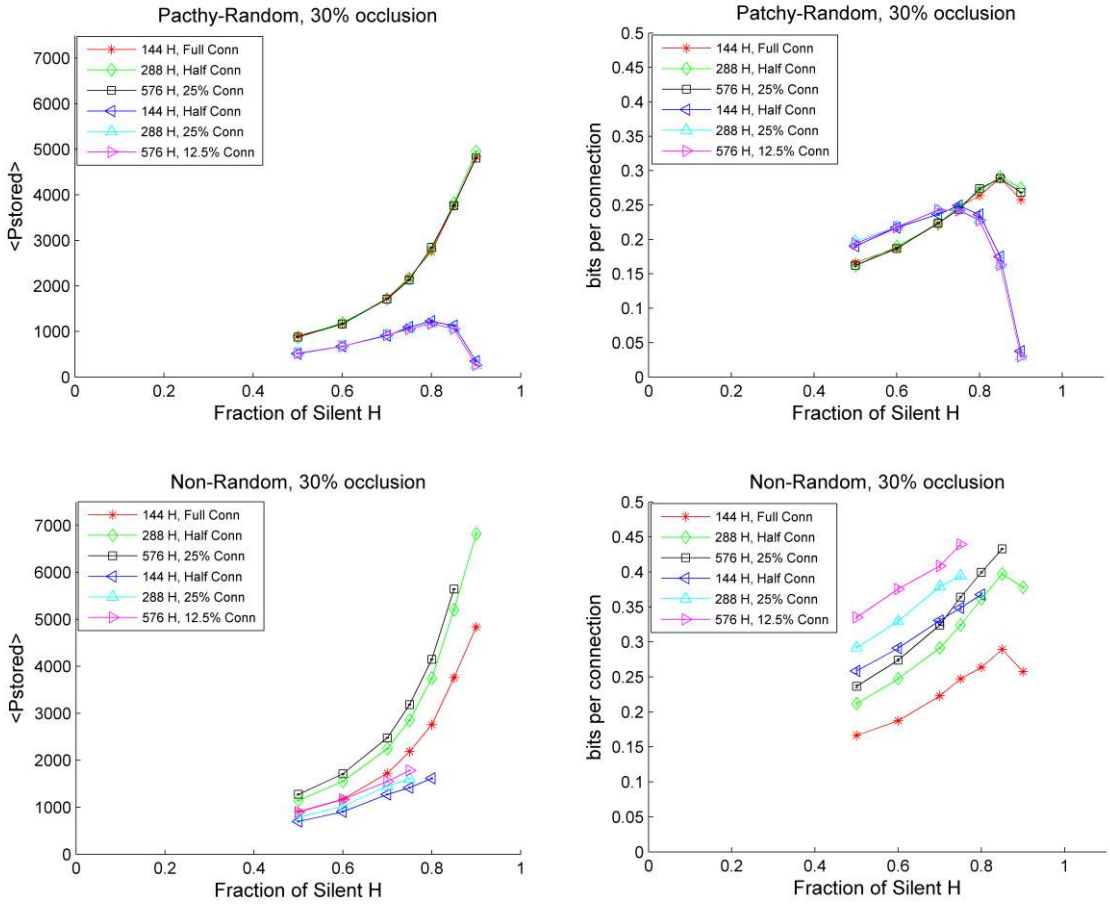

Figure 8. The storage capacity in terms of patterns (left) and bits per connection (right) for three networks of 144, 288 and 576 hypercolumns. The connection matrix is diluted proportionally to the size, using patchy-random (top) and non-random method (bottom).

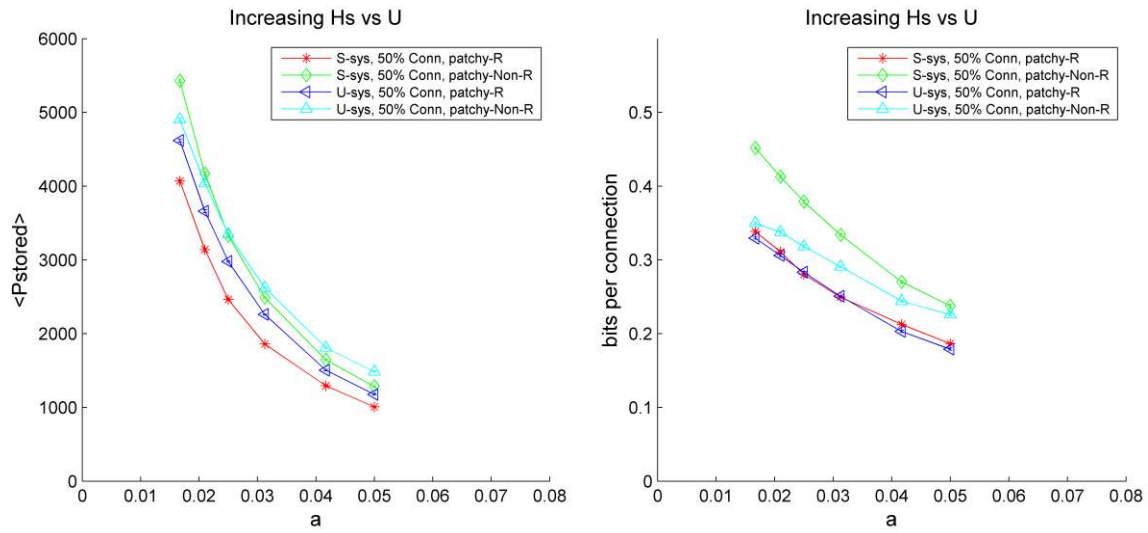

Figure 9.The storage capacity of a network of 2880 units with variable module size (U-system), is compared with the storage capacity of a network having the same number of units and a fixed size module (S-system), for increasing values of the sparseness. The connectivity is diluted to 50%, using both patchy random and non-random removal.

Scaling to large Networks

We present here the effect of non-random removal for large networks with diluted connectivity, using silent hypercolumns to lower the activity level. In Fig. 10 the storage capacity is plotted for networks of increasing size, starting from a fully connected system of 144 hypercolumns up to a system of 4608 hypercolumns with a weight matrix sparseness around 3%. The structure of the network and the number of synapses per unit are fixed to 10 units per module and 1430 synapses respectively. The systems have been tested using patchy random and non-random methods first without then with 60% silent hypercolumns. As in the previous tests, each network performs the same task with the number of hypercolumns to be filled in set to 43. First we remark that sparseness is crucial to enhance the storage capacity: we obtained about double the value with silent hypercolumns both for the pattern capacity and the information capacity. Second, we obtained a considerable increase of the storage capacity with the non-random removal. In accordance with the results presented in Treves & Rolls (1991), systems having the same number of synapses per unit, can store roughly the same number of patterns, as shown in Fig. 10 for the patchy random method. When the sparse matrix is achieved through a selection, we observe an increase of the storage capacity with the size of the network and this means that the measure we used for the pruning (9) is effective at least for this range of parameters. The difference can be explained comparing the connectivity patterns in the two systems after the removal. In a fully connected network, trained with sparse random patterns, only part of the synapses are involved in the learning process. With the non-random method we try to identify and preserve these connections, in order to exploit the whole information stored during learning. When connections are removed randomly, part of this information is lost and this explains the difference in performance. For larger networks, the connectivity is further reduced and the efficiency of the configuration does not change since the number of synapses per unit is the same, hence we obtain a constant storage capacity. The corresponding increase observed with the non-random method, suggests that the distribution of patchy connections according to the measure (9), improves the efficiency of the configuration in this interval. The weakest connections are removed first from each input

set, preventing the loss of information which occurs with a random selection. Therefore, we expect a higher storage capacity comprised between the value achieved with the random method and the storage capacity of the fully connected network. This effect should persist until the number of synapses per unit is not too small compared to the size of the network. The information capacity achieved with the patchy random method is about 0.08 bit per synapse for the systems without silent hypercolumns, i.e. an activity level equal to 0.1 and about 0.19 bit per synapse when the sparseness is increased to 0.04. The maximum information capacity reached with the non-random method is about 0.19 bit per synapse for the system without silent and about 0.35 bit per synapse for the system with 60% silent, approaching the asymptotic values for a standard auto-associative memory presented in Willshaw et al. (1969), Palm (1991) and Palm & Sommer (1991). In fact, in the context of a binary associative memory, which has maximum information capacity when the connection matrix is filled to 50%, if we can avoid storing non-used connections we would about double the information capacity. This is what we approach here with our non-random selection. The maximum pattern capacity we achieved with the non-random method, with the highest dilution (3%) and sparseness (a=0.04), namely about 2140 patterns, is a bit lower compared to the value predicted by formula (12), which falls in the range 2220-3330.

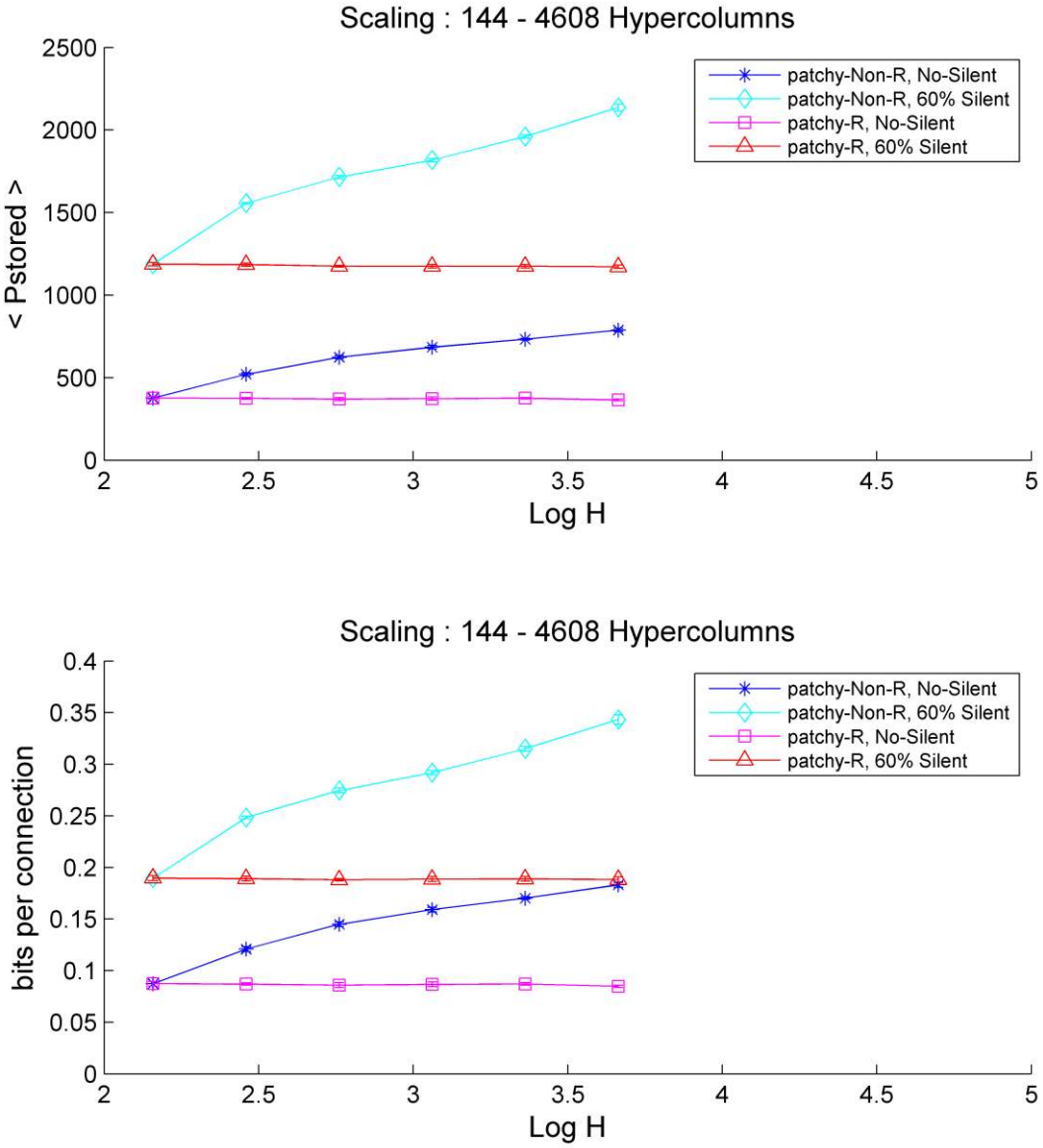

Figure 10.The effect of non-random removal on large networks, up to 46080 units, with sparse connectivity up to 3%. The storage capacity in terms of patterns (top) and bits per connection (bottom) is plotted for systems without silent hypercolumns and systems with 60% silent.

Discussion

The main subject addressed in this paper is the assessment of the storage capacity of an attractor network, after imposing specific constraints on the connectivity and activity level. We developed an abstract model reproducing, with some approximations and on a smaller scale, the modular structure and the sparse connectivity of the mammalian cortical network, as well as the low level of activity estimated for the cortex. Our network represents an extension of a previous model (Johansson & Lansner, 2006), characterized by a similar modular

structure but without patchy connectivity. The results of the simulations showed that the pattern capacity of a modular network with patchy connectivity is lower, but yet comparable with the values achieved with standard associative memories and that higher values can be obtained increasing the sparseness of the representation, with a corresponding information capacity close to the asymptotic value for standard associative memories. We proposed a local method to prune the connectivity of the network, which further improves the storage capacity. With the random removal of patchy connections, we obtained a constant storage capacity for networks of increasing size having the same number of synapses per unit, in agreement with the theoretical results presented by Treves and Rolls (1991). With the non-random removal, the storage capacity increased and seemed to depend upon the size of the network. We speculate that this effect could disappear for larger networks, resulting in a constant storage capacity but higher compared with the random removal. In principle, this method can be applied to any kind of associative network, adopting proper measures to evaluate the value of connections. The results achieved in this study are important for the development of more realistic high-capacity large-scale networks with sparse connectivity. First, we have found that introducing patchy connectivity and silent hypercolumns in the abstract model, does not imply a drastic reduction for the storage capacity. The performance of a modular network with patchy connectivity is equivalent to that of a modular network with random connections between units pertaining to different modules. Furthermore, with silent hypercolumns it is possible to reproduce a more realistic activity pattern, characterized by a small fraction of active hypercolumns. A second point made in this paper, is related to the diluted connectivity of the biological network. The connectivity pattern obtained with the standard techniques is rather unrealistic since part of the information acquired during learning is lost and part of the connections in the final configuration, do not bring any information. With the non-random removal of patchy connections, we improved the performance of the modular network and obtained a connectivity pattern with most of connections actively involved in the encoding. However, the proper selection is much more difficult if the network connectivity is diluted from the start as in the brain. The method we propose could then be implemented over time by pruning patches (terminal clusters) that have a low value according to some local scoring, and sprout a terminal cluster somewhere else which may be more useful and thus having a higher score and better chance to survive. An abundance of connections during early childhood would also help such a structural plasticity process. Our model is partly incomplete regarding both the architecture of the network and the new method we proposed for the removal of connections. We considered mostly modules with ten units, while in the biological network we observe more like 100 minicolumns in each hypercolumn (Hubel & Wiesel 1977, Amirikian & Georgopoulos 2003). Moreover, the scaling properties have been investigated for systems having about 1400 synapses per unit, while in the biological network it is in the order of 104 (Rolls & Treves 1998, Rolls 2008). We used a simple measure to evaluate the strength of the single patchy connection, that allowed obtain a higher storage capacity for the range of parameters investigated. However we are not able to predict at this stage if this measure would be effective for similar networks operating in different conditions. For instance, an interesting aspect that could be further developed is the analysis of the storage capacity using non-random patterns. In this context, it would be interesting to test if our simple measure (eqn. 9) is still efficient. More in depth theoretical analysis of the structural plasticity processes proposed here would give valuable additional insights and better implementations. The next step would be to apply the non-random removal to the long-range connections in a spiking network model (Lundqvist et al., 2006, Djurfeldt et al., 2008). It would also be interesting to analyse different aspects related to the geometry of the network. We could consider a more complex architecture with modules having different size and introduce a spatial reference in the network and new physical variables like the transmission

velocity to simulate more specific events. In conclusion, this study has shown that the storage capacity of large-scale attractor memory networks with patchy connectivity is comparable to and scales as efficiently as do traditional random sparsely connected networks. They are, however, more biologically plausible and much more efficient when it comes to wiring of the network. We have further shown that a biologically plausible method to prune network connectivity based on a local measure is able to enhance significantly the storage capacity. It remains to extend this to a strategy that works also for networks that are sparsely connected from the very beginning.

References

Amirikian B., Georgopoulos A.P. (2003) Modular organization of directionally tuned cells in the motor cortex : Is there a short range order? Proc. Natl. Acad. Sci., 100, 12474-12479

Amit D. (1989) Modelling brain function: The world of attractor neural networks. New York: Cambridge University Press.

Bailey C.H., Kandel E.R. (1993) Structural Changes Accompanying Memory Storage. Ann. Review of Physiology, Vol.55, 397-426

Bosch H., Kurfess F.J. (1998) Information storage capacity of incompletely connected associative memories. Neural Networks, 11(5): 869-876.

Bosking W.H., Zhang Y., Schofield B., FitzPatrick D. (1997) Orientation Selectivity and the Arrangement of Horizontal Connections in Tree Shrew Striate Cortex. J. NeuroSci., 17, 2112-2127

Boyd G. and Matsubara J. (1991) Intrinsic connections in cat visual cortex: a combined anterograde and retrograde tracing study. Brain Res., 560, 207-215.

Braitenberg V., Schüz A. (1998) Cortex : Statistics and geometry of neuronal connectivity. New York: Springer Verlag

Buxhoeveden D.P., Casanova M. F.(2002) The minicolumn hypothesis in neuroscience. Brain, 125 (5), 935-951

Cheng, K., Waggoner R.A. and Tanaka K. (2001) Human Ocular Dominance Columns as Revealed by High-Field Functional Magnetic Resonance Imaging. Neuron 32(2): 359-374

Compte A., Brunel N., Goldman-Rakic P.S., Wang X-J (2000) Synaptic mechanisms and network dynamics underlying visuospatial working memory in a cortical network model. Cerebral Cortex 10: 910–923.

Cossart R., Aronov D., Yuste R., Attractor dynamics of network Up states in the neocortex. NATURE 423 (6937) (2003) 283-288

DeFelipe J., Conley M., Jones E.G. (1986) Long-range focal collateralization of axons arising from corticocortical cells in monkey sensory-motor cortex. J. NeuroSci., 6, 3749-3766

Djurfeldt M., Lundqvist M., Johansson C., Rehn M., Ekeberg Ö and Lansner A. (2008) Brain-Scale simulation of the neocortex on the IBM Blue Gene/L supercomputer. IBM J.RES. & DEV. VOL. 52 NO. 1/2

Fransén E., Lansner A. (1995) Low spiking rates in a population of mutually exciting pyramidal cells. Network: Computation in Neural Systems 6: 271–288.

Fulvi Mari C., Treves A. (1998) Modeling neocortical areas with a modular neural network. BioSystems, 48, 47-55

Fulvi Mari C. (2004) Extremely Dilute Modular Neuronal Networks: Neocortical Memory Retrieval Dynamics. J. Comp: NeuroSci. 17, 57-59

Gilbert C.D., Wiesel T.N. (1989) Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. J. Neurosci., 9, 2432-2442

Goldman P.S., Nauta W.J.H. (1977) Columnar distribution of corticocortical fibers in the frontal association, limbic and motor cortex of the developing rhesus monkey. Brain Res., 122, 393-413

Hebb D.O. (1949) The Organization of Behaviour. John Wiley

Hofman M. A. (1985) Size and shape of the cerebral cortex in mammals: I. the cortical surface. Brain behavior and Evolution, 27, 28-40

Hopefield JJ (1982) Neural networks and physical systems with emergent collective computational properties. PNAS 81: 3088-3092.

Houzel J.C., Milleret C., Innocenti G. (1994) Morphology of Callosal Axons Interconnecting Areas 17 and 18 of the Cat. European Journal of Neuroscience, Vol.6, 898-917

Hubel D. H., Wiesel T.N. (1977) Functional architecture of macaque monkey visual cortex. Proceedings of the Royal Society of London B, 198, 1-59

Johansson C., Lansner A. (2006) On the Storage Capacity of an Abstract Cortical Model with Silent Hypercolumns. Tech Rep KTH, TRITA-NA-P0501

Lamprecht R., LeDoux J. (2004) Structural Plasticity and Memory. NATURE REVIEWS NEUROSCIENCE VOL 5, 45-54

Lansner A. (2009) Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. TINS-673

Lansner A., Ekberg Ö. (1989) A one layer feedback artificial Neural network with bayesian Learning Rule. International Journal of Neural Systems, vol. 1, No 1, 77-87

Lansner A. and Holst A. (1996) A higher order Bayesian neural network with spiking units. International Journal of Neural Systems, Vol. 7, No. 2 115-128

Lennie P. (2003) The Cost of Cortical Computation. Current Biology 13, 6: 493-497

Lundqvist M., Rehn M., Djurfeldt M. and Lansner A. (2006) Attractor dynamics in a modular network model of neocortex. Network: Computation in Neural Systems 17(3): 253-276

Martin K.A.C. and Whitteridge D. (1984) Form, function and intracortical projections of spiny neurons in the striate visual cortex of the cat. J. Physiol., 353, 463-504.

Mountcastle V.B. (1997) The columnar organization of the cortex. Brain, 120, 701-722

Muller D., Nikonenko I., Jourdain P. and Alberi S. (2002) LTP, Memory and Structural Plasticity. Current Molecular Medicine Vol.2, 605-611(7).

Pakkenberg B.,Gundersen H.J.G. (1997) Neocortical Neuron Number in Humans : Effect of Sex and Age. The journal Comparative Neurology, 384, 312-320

Palm G. (1980) On Associative Memory. Biol. Cybernetics 36, 19-31

Palm G. (1981) Towards a theory of cell assemblies. Biol. Cybernetics 39(3), 181-194

Palm G. (1982) Neural Assemblies. An alternative approach to artificial intelligence. Berlin, Heidelberg, New York: Springer.

Palm G. (1988) On the asymptotic information storage capacity of neural networks. Neural Computers (271-280). New York: Springer-Verlag.

Palm G. (1991) Memory capacities of local rules for synaptic modification. A comparative review. Concepts in Neuroscience, 2, 97-128.

Palm G., Sommer F.T. (1992) Information capacity in recurrent McCulloch-Pitts networks with sparsely coded memory states. Network 3, 177-186

Palm G. (1992) On the Information Storage Capacity of Local Learning Rules. Neural Computation 4, 703-711

Palm G. (2013) Neural associative memories and sparse coding. Neural Networks 37, 165-171

Peters A., Yilmaz E. (1993) Neuronal organization in area 17 of cat visual cortex. Cerebral Cortex, 3(1), 49-68

Rakic P. (1995) A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. Trends in Neurosci. 18, 383-388

Rolls E.T. (2008) Memory, attention and decision-making. A Unifying Computational Neuroscience Approach. Oxford: Oxford UniversityPress.

Rolls E. T. (2012) Advantages of dilution in the connectivity of attractor networks in the brain. Biologically Inspired Cognitive Architectures 1, 44-54

Rolls E. T., Treves A. (1998) Neural Networks and Brain Function, Oxford University Press, New York.

Schwenker F., Sommer F.T., Palm G. (1996) Iterative retrieval of Sparsely Coded Associative Memory Patterns. Neural Networks, 9(3): 445-455.

Shu Y., Hasenstaub A., Mccormick D.A. (2003) Turning on and off recurrent balanced cortical activity, NATURE 423 (6937) 288-293

Treves A., Rolls E. T. (1991) What determines the capacity of autoassociative memories in the brain? Network 2 371-397.

Tsunoda K., Yamane Y., Nishizaki M. and Tanifuji M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nature NeuroSci. 4(8): 832-838.

Voges N., Guijarro C., Aertsen A. and Rotter S. (2010) Models of cortical networks with long-range patchy projections, J Comput Neurosci 28 : 137-154

Waydo S., Kraskov A., Quiroga Q. R., Fried I. and Kock C., (2006) Sparse Representation in the Human Medial Temporal Lobe. Journal of Neuroscience, 26(40): 10232-10234.

Willshaw D.J., Buneman O.P. and Longuet-Higgins H.C. (1969) Non-holographic associative memory, Nature, 222, 960-962