Improving the storage capacity of neocortical associative networks by structural plasticity and hippocampal training.

Andreas Knoblauch, Marc-Oliver Gewaltig, Rüdiger Kupper, Ursula Körner, Edgar Körner

2007

Preprint:

This is an accepted article published in Proceedings of the 4th HRI Global Workshop on Advances in Computational Intelligence. The final authenticated version is available online at: https://doi.org/[DOI not available]

Improving the storage capacity of neocortical associative networks by structural plasticity and hippocampal training

Andreas Knoblauch, Marc-Oliver Gewaltig, Rüdiger Kupper, Ursula Körner, Edgar Körner

***Abstract*— In this work we analyze how structural processes and synaptic consolidation during hippocampal training can improve the performance of neocortical associative networks by emulating full (or increased) synaptic connectivity. In our model the hippocampus can store a set of activity patterns by one-shot learning. Then the hippocampus trains the neocortex by repeatedly replaying the patterns in a sequence. The synapses of the neocortical network are consolidated depending on Hebbian learning. In each time step a fraction of the unconsolidated synapses are removed and replaced by the same number of new synapses at random locations thereby maintaining total connectivity. We show that this procedure can massively increase the synaptic capacity of a cortical macrocolumn (factor 10-20 or even up to factor 200 for pattern capacity). In a second step we analyze the model with respect to the time (or number of repetitions) necessary to increase effective connectivity from base level to a desired level. The analysis shows that acceptable training time requires a certain fraction of unconsolidated synapses to keep the network plastic.**

I. Introduction

The hippocampal formation plays a crucial role in organizing cortical long-term memory. It is believed that the hippocampus is able of fast (one-shot) learning of new episodic information followed by extensive time periods where corresponding neocortical representations are trained and "compressed" [1]. Here, compression usually refers to processes such as chunking spatially and temporally distributed activity patterns. We take the complementary approach and optimize the synaptic network by structural plasticity, e.g., replacing unused synapses, thereby making full use of the potential connectivity [2].

We apply the frameworks of structural plasticity and hippocampus-induced learning to the training of neocortical associative networks [3]. Associative networks such as the Hopfield or Willshaw model are at the heart of many cortex theories and have been analyzed for a long time with respect to information storage capacity and plausible retrieval strategies [3], [4]. For example, it is well known that a completely connected network can store about 0.7 bits per synapse. However, for incompletely connected networks the capacity per synapse can be massively reduced or even vanish, depending on the retrieval algorithm [4].

A. Knoblauch is with the Honda Research Institute Europe, email: andreas.knoblauch@honda-ri.de

M.-O. Gewaltig is with the Honda Research Institute Europe, email: marc-oliver.gewaltig@honda-ri.de

R. Kupper is with the Honda Research Institute Europe, email: ruediger.kupper@honda-ri.de

U. Körner is with the Honda Research Institute Europe, email: ursula.koerner@honda-ri.de

E. Körner is with the Honda Research Institute Europe, email: edgar.koerner@honda-ri.de

In this work we analyze how structural processes and synaptic consolidation [5] during hippocampal training can improve the performance of neocortical associative networks by emulating full (or increased) synaptic connectivity. In our model the hippocampus can store a set of activity patterns by one-shot learning. Then the hippocampus trains the neocortex by repeatedly replaying the patterns in a sequence. The synapses of the neocortical network are consolidated depending on Hebbian learning. In each time step a fraction of the unconsolidated synapses are removed and replaced by the same number of new synapses at random locations thereby maintaining total connectivity. We show that this procedure can massively increase the synaptic capacity of a cortical macrocolumn (factor 10-20 or even up to factor 200 for pattern capacity). In a second step we analyze the model with respect to the time (or number of repetitions) necessary to increase effective connectivity from base level to a desired level. The analysis shows that acceptable training time requires a certain fraction of unconsolidated synapses to keep the network plastic.

II. How much information can a synapse store?

A. Synaptic (state) plasticity

Here we develop upper bounds on the amount of information a single synapse can store. We can think of two components that contribute to the total synaptic storage capacity. First, information can be stored in the state properties of a given synapse. The state of a synapse may include synaptic strength (i.e., the amplitude of the postsynaptic potential), synaptic delay (i.e., the time difference between the presynaptic spike and the postsynaptic potential), but also the postsynaptic composition of receptors. From information theory [6], [7] we then know that the capacity $C_{\rm state}$ of a synapse due to plasticity of its state cannot exceed the following bound,

$$C_{\rm state} \leq \log_2(\#\text{synaptic states}) \, . \qquad (1)$$

This suggests that we can potentially store unlimited information per synapse, for example if the synaptic weight is real-valued. However, real synapses are subject to noise as well as a number of adaptation mechanisms which suggest that real synapses may have a rather small number of functionally distinctive states (perhaps on the order of 10 or even binary [8]). Similarly, technical implementations on digital computers have typically only a limited numerical precision for representing a synaptic state (e.g., 8 bits for weights between 0 and 255). Furthermore, numerical experiments for various network models show that increasing numerical

precision does not necessarily lead to an increase in performance [9]. Another well-known fact is that associative networks with binary synapses can store almost the same or even more information per synapse (0.69 bits per synapse [10], [11]) than comparable models with gradual synapses (0.14 bps [12] or at most 0.72 bps [13]). In all these cases the bound of eq. 1 cannot be achieved because of algorithmic limitations of the network implementations. Thus, it seems reasonable to assume that the state capacity is bounded by a relatively small value (of perhaps 0.1-1bps)

B. Structural (location) plasticity

Structural plasticity provides a complementary way to store information in a neural network. This includes processes such as the generation of new synapses and the elimination of unused (silent) synapses as well as dendritic and axonal growth and remodeling which is now considered a regular physiological feature of adult brains [14], [15], [16]. Structural plasticity appears to be the solution of the brain for the impossibility to establish a fully connected network as commonly employed by artifical neural networks. Actually, the number or density of functional (i.e., non-silent) synapses seems to be the very bottleneck of the brain, both for anatomical and metabolic reasons [17]. In particular, the number of synapses per cortical volume is remarkably constant over different species [18].

Think of a new synapse that is generated by a neuron. The amount of storable (location) information is determined by the number of potential target locations, i.e., by all the target cell dendrites that the neuron's axon can reach. Thus, similarly as before, the capacity $C_{\rm loc}$ of a synapse due to structural plasticity plasticity cannot exceed the bound

$$C_{\rm loc} \quad \leq \quad \log_2(\#\text{potential locations}) \; . \qquad (2)$$

Again the precise location of a synapse on a dendrite may be real-valued and thus our bound unlimited. However, it is usually assumed that the number of functionally relevant target locations on a dendritic tree is rather small. For example, one could divide a neuron into a relatively small number (10-1000) of iso-potential compartments [19], or even more conservatively, simply count the number of potential target cells [2], [3].

Thus the decisive question is how many cells can be targeted by a typical cortical neuron? The typical view is that a cortical neuron is connected to about 10 percent of its neighbors within a cubic-millimeter of cortex containing about 100000 potential target cells [20], [21]. This would correspond to an upper bound $C_{\rm loc} \leq \log_2 100000 \approx 17$bps. From the 100000 cells in 1mm3 cortex only about 30 percent are potential targets for *fast* structural processes (minute time scale) such as dendritic spine growth and retraction [2], [3] which corresponds to a bound $C_{\rm state} \leq \log_2 30000 \approx 15$bps. Clearly the final answer to our question depends on the considered time scale of structural plasticity. For example, if we allow a larger time window of days or even months the much slower structural processes involving dendritic and axonal growth may become important. Then potential targets may lie in the whole sphere containing the neuron's axonal tree, i.e., in a much larger range than 1mm^3. Thus, it seems reasonable to assume that the location capacity is bounded by perhaps 15-20bps.

The surprising result is that the total synaptic capacity may actually be dominated by structural plasticity with capacities more than one order of magnitude larger than obtained for the more commonly investigated synaptic plasticity. However, this conclusion assumes that the upper bound eq. 2 could actually be reached by a neural network realization. This assumption is not obviously true as shown by the negative example for the state capacity mentioned above: The Hopfield model is actually algorithmically unable to exploit the high potential capacity of gradual-valued synapses. In contrast, we will show here that even simpler associative networks with binary synapses are actually able to exploit the potential of location capacity,i.e., for a target population of n neurons we can indeed achieve $C \approx \log_2 n$.

III. How much information can a synaptic network store?

A. Neural associative networks

Here we fix our ideas about the storage capacity of structural plasticity to a concrete model of neural associative memory. Neural associative memory networks and cell assemblies play a prominent role in virtually any brain theory, in particular concerning neocortex and hippocampus (e.g., [10], [22], [23], [24], [12], [25], [18], [26], [27], [28], [29]). The main assumptions are that functionally relevant entities (e.g., objects) are represented by distributed activity patterns, that these activity patterns can be stored in the local recurrent (auto-associative) connections by Hebbian learning, thereby constituting local cell assemblies (defined as the active neurons of an activity pattern), and that cell assemblies at different cortical locations are linked by hetero-associative Hebbian learned synaptic connections.

Formally speaking, *associative memories* are systems that contain information about a finite set of associations between pairs of address and content patterns $\{(\mathbf{u}^\mu \rightarrow \mathbf{v}^\mu) : \mu = 1, ..., M\}$. A typically noisy address pattern $\tilde{\mathbf{u}}$ can be used to retrieve an associated content pattern $\hat{\mathbf{v}}$ which ideally equals the content pattern $\mathbf{v}^\mu$ associated to the most similar address pattern $\mathbf{u}^\mu$. This is a variant of the *Best Match Problem* in [30], and efficient solutions of have widespread applications, e.g. for cluster analysis, speech and object recognition, or information retrieval in large databases [31], [32], [33], [34], [35].

In *neural implementations* of associative memory the information about the associations is stored in the synaptic connectivity of one or more neuron populations. From the technical perspective, neural implementations can be advantageous over hash-tables or simple look-up-tables if the number of patterns is large, if parallel implementation is possible, or if fault-tolerance is required, i.e. if the address patterns $\tilde{\mathbf{u}}^\mu$ may differ in unpredictable ways from the

original patterns $\mathbf{u}^\mu$ used for storing the associations (but see [36]).

B. The Willshaw model

An attractive model of neural associative memory both for biological modeling and technical applications is the so-called Willshaw or Steinbuch model with binary neurons and synapses [10], [37], [11], [13] illustrated in Fig. 1. Each address pattern $\mathbf{u}^\mu$ is a binary vector of length m containing k one-entries and $m-k$ zero-entries. Similarly, each target pattern $\mathbf{v}^\mu$ is a binary vector of length n containing l one-entries and $n-l$ zero-entries. Typically, the patterns are sparse, i.e., $k \ll m$ and $l \ll n$.

The M pattern pairs are stored *hetero-associatively* in a binary *memory matrix* $\mathbf{A} \in \{0,1\}^{n \times m}$, where

$$A_{ij} \quad = \quad \min\left(1, \sum_{\mu=1}^{M} u_i^\mu \cdot v_j^\mu\right) \in \{0,1\}\,. \qquad (3)$$

The *neural interpretation* is that of two neuron populations, an address population u consisting of m neurons and a target population v consisting of n neurons. The patterns $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ describe the activity states of the two populations at time μ, and A_{ij} is the strength of the Hebbian learned synaptic connection from neuron u_i to neuron v_j. Note that for the auto-associative case $u = v$ (i.e., if address and target populations are identical), the network can be interpreted as an undirected graph with $m = n$ nodes and edge matrix $\mathbf{A}$ where patterns correspond to cliques of $k = l$ nodes.

An important variable for estimating the performance of the Willshaw model is the matrix or memory load p_1 which is the fraction of one-entries in the memory matrix $\mathbf{A}$. Obviously the memory load p_1 is an increasing function of the pattern number M and can be computed as follows: The probability that a given synapse is *not* set by the association of one pattern pair is $1 - kl/mn$, therefore after learning M pattern associations,

$$p_1 \quad = \quad 1 - \left(1 - \frac{kl}{mn}\right)^M \approx 1 - e^{-Mkl/mn}, \qquad (4)$$

$$M \quad = \quad \frac{\ln(1-p_1)}{\ln(1-kl/mn)} \approx -\frac{mn}{kl}\ln(1-p_1), \qquad (5)$$

where the approximation is valid for $kl \ll mn$.

After learning, the stored information can be retrieved applying an address pattern $\tilde{\mathbf{u}}$. Vector-matrix-multiplication yields the neural potentials $\mathbf{x} = \tilde{\mathbf{u}} \cdot \mathbf{A}$ of the target population, and imposing a threshold Θ gives the (one-step) retrieval result $\hat{\mathbf{v}}$,

$$\hat{v}_j \quad = \quad \begin{cases} 1\,, & x_j = (\sum_{i=1}^m \tilde{u}_i A_{ij}) \geq \Theta \\ 0\,, & \text{otherwise} \end{cases}\,. \qquad (6)$$

Choosing $\Theta = \sum_{i=1}^m \tilde{u}_i$ will be referred to as the *Willshaw threshold* and ensures that the retrieval result $\hat{v}^\mu$ includes all active units of the original content pattern v^μ, plus possibly add-noise (i.e., false one-entries). More exactly, if the address pattern $\tilde{u}$ contains a fraction λk of the one-entries of u^μ, then a zero-entry in $\mathbf{v}^\mu$ will become a "false" one in $\hat{\mathbf{v}}$ with the error probability

$$p_{01} \quad = \quad \sum_{s=0}^{\lambda k} (-1)^s \binom{\lambda k}{s} \left(1 - \frac{l}{n}\left(1 - \frac{\binom{m-k}{s}}{\binom{m}{s}}\right)\right)^{M-1} \qquad (7)$$

$$\approx \quad {p_1}^{\lambda k}. \qquad (8)$$

For the exact formula see [3]. In the following we use the binomial approximation eq. 8 which assumes independently generated one-entries in the memory matrix. Although this is obviously not true, the approximation is very useful for our analysis and still sufficiently good for many parameters, for example sufficiently sparse patterns with $k = O(n^2/\log n)$ (see [38]).

C. The storage capacity of the Willshaw model

Obviously, if we store more and more patterns the memory load approaches $p_1 \to 1$ and the error probability p_{01} becomes unacceptable. In order to compute the maximal possible memory load $p_{1\epsilon}$ and the maximal number of storable patterns M_ϵ we bound the error probability p_{01} by $p_{01\epsilon}$,

$$p_{01} \leq p_{01\epsilon} := \frac{\epsilon l}{n-l}, \qquad (9)$$

where we call $\epsilon > 0$ the *fidelity parameter*. For example, $\epsilon = 0.01$ means that the add-errors in the retrieval result $\hat{v}$ is at most one percent of the content pattern activity l. Or equivalently, the expected Hamming distance between $\hat{v}$ and v^μ is at most ϵl. With the approximation eq. 8 we obtain the maximal memory load $p_{1\epsilon}$ and the corresponding pattern capacity M_ϵ,

$$p_{1\epsilon} \quad \approx \quad \left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda \cdot k}} \qquad \left(\Leftrightarrow k \approx \frac{\mathrm{ld}\frac{\epsilon l}{n-l}}{\lambda \mathrm{ld} p_{1\epsilon}}\right), \qquad (10)$$

$$M_\epsilon \quad \approx \quad -\lambda^2 (\mathrm{ld} p_{1\epsilon})^2 \ln(1-p_{1\epsilon}) \frac{k}{l} \frac{mn}{(\mathrm{ld}\frac{n-l}{\epsilon \cdot l})^2}. \qquad (11)$$

If ϵ is sufficiently small the totally stored information is $MnI(l/n) \approx -Ml\mathrm{ld}(l/n)$, where $I(p) := -p\mathrm{ld}(p) - (1-p)\mathrm{ld}(1-p)$ is the Shannon information of a binary random variable with probability p. Dividing by the number of synapses mn we obtain the normalized network storage capacity C_ϵ in bits per synapse,

$$C_\epsilon \quad \approx \quad \lambda \mathrm{ld} p_{1\epsilon} \ln(1-p_{1\epsilon}) \eta \qquad (12)$$

where $\eta \approx (1 + \frac{\ln \epsilon}{\ln(l/n)})^{-1} \to 1$ for $l/n \to 0$. Thus, for $p_{1\epsilon} = 0.5$ and logarithmic $k \sim \log n$ (eq. 10) we obtain the well-known storage capacity $C = \ln 2 \approx 0.69$ bps of the Willshaw model.

D. Diluted Willshaw networks

The analysis so far is valid for fully connected networks, i.e., where each pair of a address neuron and content neuron is actually connected by a binary synapse. The analysis can be generalized to diluted networks with connectivity $p < 1$, i.e., where only a fraction pmn of the mn potential synapses

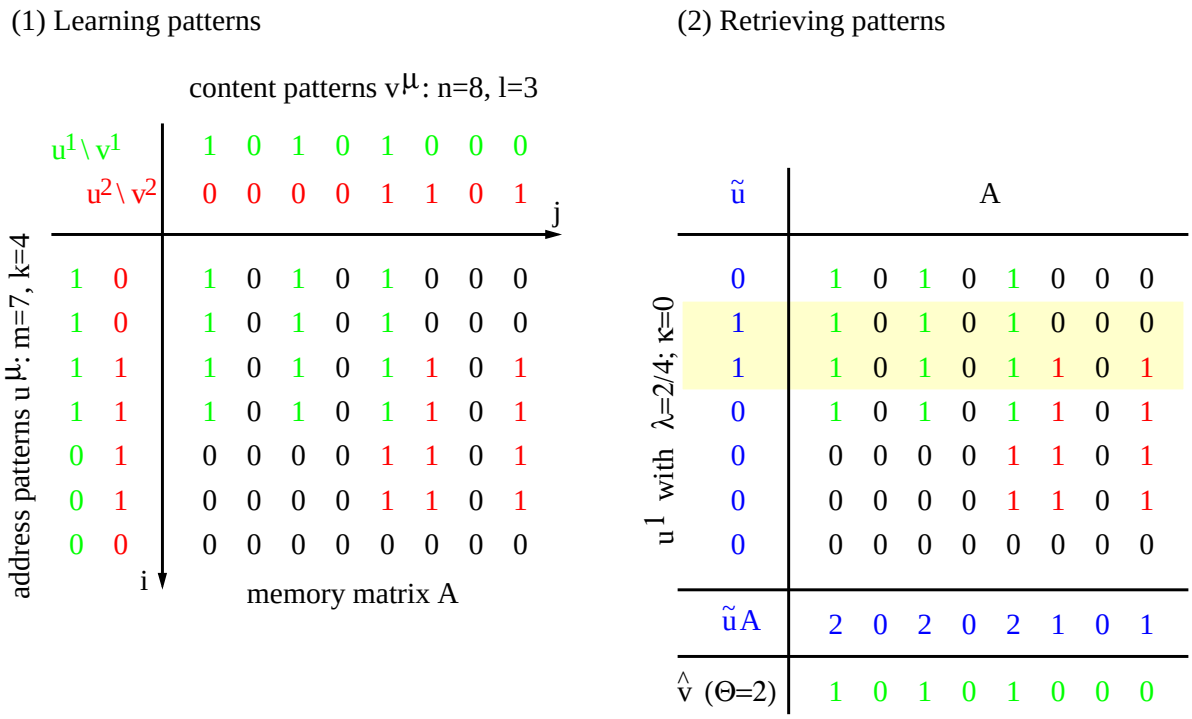

Fig. 1. Working principle of the binary Willshaw model (for hetero-association). **Left:** During the learning phase M associations between address patterns $\mathbf{u}^\mu$ and content patterns $\mathbf{v}^\mu$ are stored in the binary memory matrix $\mathbf{A}$ representing binary synaptic weights of the connection from address population u to content population v. Initially all synapses are inactive. During learning of pattern associations, the synapses are activated according to Hebbian coincidence learning (eq. 3). **Right:** For retrieval an address pattern $\tilde{\mathbf{u}}$ is propagated through the network. Vector-matrix-multiplication yields the membrane potentials $\mathbf{x} = \tilde{\mathbf{u}}\mathbf{A}$. To obtain the retrieval result $\hat{\mathbf{v}}$ (here equal to $\mathbf{v}^1$) a threshold Θ is applied. For pattern part retrieval, $\tilde{\mathbf{u}} \subseteq \mathbf{u}^\mu$ we can simply choose the Willshaw threshold $\Theta = |\tilde{\mathbf{u}}|$ yielding a superset $\hat{\mathbf{v}} \supseteq \mathbf{v}^\mu$ of the original pattern (i.e., no missing one-entries).

is actually realized [3]. The resulting formulae for pattern capacity M_ϵ and network storage capacity C_ϵ are

$$M_\epsilon \approx -\lambda^2 (\mathrm{ld} p_{1\epsilon})^2 \ln(\frac{1-p_{1\epsilon}}{p}) \frac{k}{l} \frac{mn}{(\mathrm{ld}\frac{n}{\epsilon \cdot l})^2} \quad (13)$$

$$C_\epsilon \approx \lambda \frac{\mathrm{ld} p_{1\epsilon} \ln \frac{1-p_{1\epsilon}}{p}}{p} \eta \quad (14)$$

where $p_{1\epsilon} \geq p_{1,\min} := 1 - p$ is as in eq. 10. Proofs can be found in [3]. We require the formulae just to discuss the possible impact of structural plasticity and hippocampal learning by comparing the capacity of diluted to fully connected networks (see discussion).

E. Willshaw model with structural plasticity

As discussed in section II-B, structural plasticity includes the elimination and regeneration of synapses where the total number of synapses may be kept approximately constant [19], [2], [14], [15], [16]. One potential function of structural plasticity could thus be the "compression" of diluted neural networks by emulation of full connectivity, because, with time, any possible potential synapse will actually be created. For example, a diluted synaptic network with connectivity $0 < p \leq 1$ endowed with structural plasticity could emulate the fully connected Willshaw network if the number $p_1 mn$ of *non-silent* synapses is not larger than the number pmn of synapses that are realized at a time. However, this would require additional mechanisms such as consolidation of the useful synapses [5] and arbitrary repetition or replay of the learning signal [1]. In section IV we will present a simple model of structural plasticity, synaptic consolidation, and hippocampal training that is capable of emulating full connectivity in neocortical networks.

For the moment let us assume that emulation of full connectivity is actually possible. Then we can estimate the consequences for the storage capacity of the Willshaw model: If emulation of full connectivity in a diluted Willshaw network with connectivity $p = p_{1\epsilon}$ is possible then the resulting network is equivalent to the fully connected Willshaw model where all the useless silent (0-)synapses have been eliminated. Thus, the network has only $p_{1\epsilon} mn$ synapses and the capacity increases by a factor $1/p_{1\epsilon}$. Similarly, in an equivalent inhibitory implementation of the Willshaw model [3], [39] (with essentially inverted memory matrix), the capacity would increase by factor $1/(1-p_{1\epsilon})$. This motivates the definition of the synaptic capacity

$$C_\epsilon^S := \frac{C_\epsilon}{\min(p_{1\epsilon}, 1-p_{1\epsilon})} \quad (15)$$

$$\approx \lambda \frac{\mathrm{ld} p_{1\epsilon} \ln(1-p_{1\epsilon})}{\min(p_{1\epsilon}, 1-p_{1\epsilon})} \eta \quad (16)$$

which gives an upper bound for the storage capacity per synapse in the Willshaw model with structural plasticity.

Figure 2 compares the asymptotic network capacity C to the synaptic capacity C^S. While C is limited by $\ln 2 \approx 0.69$ bps the synaptic capacity C^S can become arbitrary large for sparse and dense connectivity with $p_{1\epsilon} \to 0$ or $p_{1\epsilon} \to 1$.

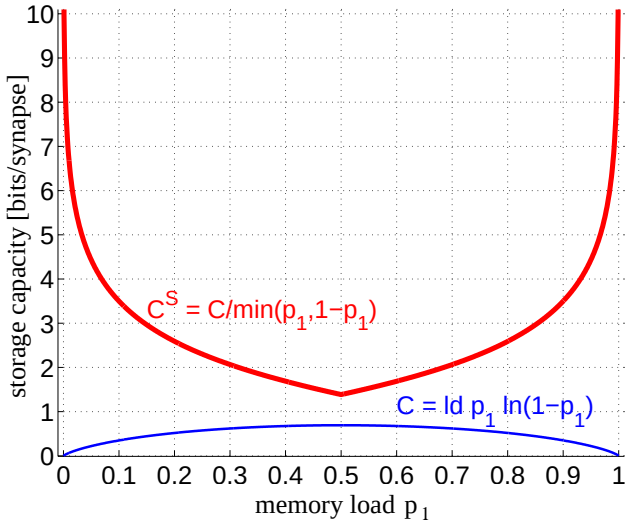

Fig. 2. Asymptotic network capacity C and synaptic capacity C^S as functions of the the memory load $p_{1\epsilon}$. We have $C \leq \ln 2 \approx 0.69$ bits per synapse for $p_{1\epsilon} = 0.5$ (balanced potentiation) where C^S is minimal. In contrast, C^S can become arbitrary large for sparse potentiation ($p_{1\epsilon} \to 0$) and dense potentiation ($p_{1\epsilon} \to 1$).

How fast grows C^S with the network size? For *sparse potentiation* ($p_{1\epsilon} \to 0$) we obtain with eq. 10

$$C^S_\epsilon \approx -\mathrm{ld} p_{1\epsilon} \approx \frac{\mathrm{ld}\frac{n-l}{\epsilon l}}{\lambda k} \approx \frac{\mathrm{ld} n}{\lambda k} \qquad (17)$$

Thus, for $\lambda k = 1$ it is actually possible to achieve the upper bound eq. 2.

Similarly, for *dense potentiation* ($p_{1\epsilon} \to 1$) we obtain with eq. 10

$$p_{1\epsilon} \approx (\frac{\epsilon l}{n-l})^{1/\lambda k} = e^{\frac{\ln(\epsilon l/(n-l))}{\lambda k}} \approx 1 - \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k}, \qquad (18)$$

$$C^S_\epsilon \approx -\mathrm{ld}(1-p_{1\epsilon}) \approx \mathrm{ld} k \ . \qquad (19)$$

Thus, we can reach the upper bound eq. 2 also for dense potentiation, for example for $k = n^d$ and $d \to 1$. In the regime of dense potentiation we must be more careful about the convergence of the binomial approximation of the error probability eq. 8: Previous analyses suggested that this approximation becomes exact only for $k = O(\log n)$ or $k = O(k^{1/3})$ and would massively overestimate performance for larger k [11], [40]. However, a recent technical report showed that eq. 8 becomes exact at least for $k = O(n/\log^3 n)$ [38] which includes $k = n^d$ for any $d < 1$.

In summary, the Willshaw model with structural plasticity can reach the bound $C^S \leq \log_2 n$ (eq. 2) both for sparsely and densely potentiated networks. The resulting capacities (in bits per synapse) for realistic network sizes are at least one order of magnitude larger than previously estimated for networks having only synaptic plasticity (see discussion).

IV. Model of structural plasticity and cortico-hippocampal interplay

Here we discuss more closely the possible interplay between a low-capacity one-shot memory system (e.g., the hippocampus) and a high-capacity learning system requiring extended training (e.g., the neocortex). We identify both systems with simple associative memories as introduced before. First we describe rather informally our ideas how the hippocampus may implement one-shot learning and train the high-capacity networks of neocortex. Then we propose a simple model of structural plasticity which can easily be analyzed. Finally we apply this model to the Willshaw associative memory model and investigate the time requirements for generating high capacity compressed memory networks.

A. Cortico-hippocampal interplay

Fig. 3 illustrates our ideas how a hippocampal low capacity one-shot learning system (HC) may interact with a neocortical high capacity associative learning system (A) in order to boost the storage capacities of neural networks in the brain. According to this scenario, neural activity in a cell population U causes associated activity patterns in population V. This may happen through purely internal processes (i.e., neural activity cascading through multiple processing stages as indicated by the lower pathway in Fig. 3) but also through external processes (e.g., activity in U may cause interactions with the environment which can be sensed by V). Now the goal is to learn the causal relations between U and V in a direct way by modifying the synaptic weights and structure of the high-capacity associative connection A from U to V. For example, A could be identified with the memory matrix of the Willshaw associative memory model (see eq. 3; see Fig. 1).

The basic idea is that the U→V associations are stored temporarily by one-shot learning in the connections from and to HC and that HC is able to reactivate the activity patterns of U and V in a sequence. This enables the high capacity synapses connecting U to V to slowly learn the associations by structural plasticity processes. It is straightforward to assume that one-shot Hebbian learning of the connections from and to HC is accomplished similarly to the diluted Willshaw model (see section III-D; see also [3]). However, we will not discuss here how HC stores and replays sequences of activity patterns (e.g., see [1], [26], [41], [42] for some ideas). On the other hand, the slow structural plasticity processes in A could effectively implement the memory compression as discussed in section III-E by pruning and regenerating synapses, thereby maintaining a constantly low connectivity level (of perhaps 10 percent within 1mm^3, [20], [21]). This is possible since HC can repeat the activity patterns to be stored and thereby providing a teacher signal for U and V.

B. Simple model of structural plasticity

Our model assumes simple binary synapses as in the Willshaw model. Since for diluted networks, a synapse can be either realized or not realized, each synapse can have one

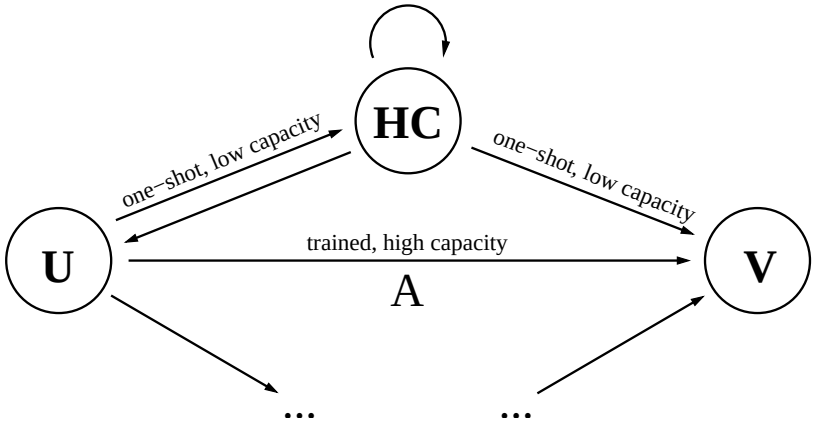

Fig. 3. Hypothesized interplay between a hippocampal low capacity one-shot learning system (HC) with a neocortical high capacity learning system (A). Each arrow corresponds to a synaptic connection between two neuron populations. We assume that neural activity in population U causes activity patterns in V (through internal and/or external processes, illustrated by the lower pathway). These associations are stored temporarily by one-shot learning in the connections from and to HC. HC is able to reactivate the activity patterns of U and V in a sequence. By repeated reactivation of corresponding U/V activity patterns the information can be transferred into the high-capacity "short-cut" connection A (which could be interpreted, for example, as the memory matrix of the Willshaw model, eq. 3). This is mainly achieved by slow learning mechanisms such as ongoing structural plasticity.

of three states. Potential synapses not yet realized are called p-synapses (cf., [2]), a real synapse with weight zero a 0-synapse, and a real consolidated synapse with weight one a 1-synapse [5]. The basic learning algorithm of the hippocampal learning system HC can be described as follows:

1) At the beginning, set all real synapses to 0-synapses.
2) Prune or erase a fraction $p_{\rm e}$ of the remaining 0-synapses.
3) Replace the pruned synapses by the same number of newly generated synapses located at random positions.
4) Present all pattern associations to be learned in a sequence. For each pair of activity patterns employ binary clipped Hebbian learning on the real synapses (cf. eq. 3), i.e., a 0-synapse changes to a 1-synapse if both presynaptic cell (in S) and postsynaptic cell (in R) are active.
5) GOTO step 2

Let us assume that there are N potential synapses (including real synapses) where a fraction p of the potential synapses is actually realized. We refer to p as the connectivity. For example, for the Willshaw model there are $N = mn$ potential synapses and pmn real synapses, where m is the size of population S and n is the size of population R. Obviously, our algorithm maintains a constant connectivity level p and a constant number of real synapses for the synaptic connection from S to R. The learning task is then to learn a specific binary connection matrix A containing $p_1 mn$ 1-synapses, where p_1 should be smaller than p. For example, A could be identified with the memory matrix of the Willshaw model (eq. 3). The learning rule of step 2 in our algorithm should provide the full information for constructing A. For example, for the Willshaw model all M patterns to be stored should be presented in the sequence of step 2.

An interesting question is how much time is necessary for learning A? We will determine the probability $p^{(T)}$ that a given 1-synapse of A is realized after T loops of our algorithm. During each loop the probability that the considered 1-synapse is realized is $1-(1-1/N)^g$, where g is the number of newly generated synapses. For $g \ll N$, this probability can simply be approximated by g/N. Since in the first loop $g = pN$ we have obviously $p^{(1)} = p$. For $g^{(s)} \ll N$ being the number of newly generated synapses in loop s we can thus write

$$p^{(T)} = 1 - \prod_{s=1}^{T} (1 - \frac{g^{(s)}}{N}) \quad (20)$$

with

$$g^{(1)} = pN \quad (21)$$

$$g^{(s)} = p_e {p_0}^{(s)} N \quad \text{for } s > 1 \quad (22)$$

where ${p_0}^{(s)} N$ is the number of real 0-synapses at the beginning of loop s which is the difference between all real synapses and the real 1-synapses at the beginning of loop s,

$${p_0}^{(s)} N \approx pN - p^{(s-1)} p_1 N = (p - p_1 p^{(s-1)}) N \quad (23)$$

Note that this is only an approximation since we use ${p_0}^{(s)}$ and $g^{(s)}$ as random variables while defining $p^{(s-1)}$ in eq. 20 as a probability (i.e., eq. 23 is exact only for the *expectation value* $E({p_0}^{(s)})$). Nevertheless, we can insert eq. 23 in eq. 22 and eq. 20 and obtain finally

$$p^{(1)} = p \quad (24)$$

$$p^{(T)} \approx 1 - (1-p) \prod_{s=2}^{T} (1 - p_e (p - p_1 p^{(s-1)})) \quad (25)$$

$$= 1 - (1-p) \prod_{s=1}^{T-1} (1 - p_e (p - p_1 p^{(s)})) \ . \quad (26)$$

The analysis becomes particularly simple if we assume $p_1 \ll p$, i.e., if the 1-synapses to be learned are only a small fraction

of all real synapses. Then we have ${p_0}^{(s)} \approx p$, $g^{(s)} \approx p_e pN$ for $s > 1$, and therefore

$$p^{(T)} \approx 1-(1-p)(1-p_e p)^{T-1} \quad (27)$$
$$\approx 1-(1-p)e^{-p_e p(T-1)} , \quad (28)$$

where the approximation is valid for large T, i.e., many replays of the pattern associations. We may be interested in time T_β where $p^{(T_\beta)} = \beta$, with typically $\beta \approx 1$. Resolving for T_β yields

$$T_\beta = 1+\frac{\ln\frac{1-\beta}{1-p}}{\ln(1-p_e p)} \quad (29)$$
$$\approx 1+\frac{\ln(1-p)-\ln(-\ln\beta)}{p_e p} , \quad (30)$$

where the approximation is valid for $\beta \to 1$, since then $\ln(1-(1-\beta) \approx -(1-\beta)$, and small $p_e \ll 1$, i.e., when structural changes are slow compared to the repetition period of the learning system HC.

C. Application to the Willshaw model

Figure 4 shows results from simulations and the analysis of the model for structural plasticity and cortico-hippocampal interplay applied to the Willshaw associative memory model. Here the activity patterns in areas S and R are assumed to be binary vectors of sizes m and n each containing k and l one-entries, respectively. Then the synaptic connection A from U to V in Fig. 3 is realized by the binary memory matrix eq. 3 of the diluted Willshaw model (see section III-D; see also [3]) Thus, the connection realizes a fraction p of the mn potential synapses, and initially (i.e., before learning) all real synapses are 0-synapses. Now the learning goal is to make A identical to the memory matrix of the completely connected Willshaw model (eq. 3) by a combination of Hebbian learning and structural changes of the network.

After the first "learning-shot" the M pattern associations are stored in A and in the connections between U,V, and HC resulting in diluted memory matrices as described in sections III-B,III-D. Then the hippocampal learning system HC sequentially replays the M pattern associations T times while the synapses in A continue Hebbian learning. In each replay epoch a fraction p_e of the remaining 0-synapses are replaced by the same number of randomly chosen potential p-synapses, and Hebbian learning can change some of the newly generated 0-synapses to 1-synapses. If the number $p_1 mn$ of 1-synapses necessary to realize the complete memory matrix eq. 3 is not larger than the total number pmn of real synapses then the learning goal (constructing the complete memory matrix) can be achieved in finite time. This process is illustrated in Fig. 4. The simulations show that the approximation eq. 26 is quite good for many parameter sets, and that approximation eq. 28 becomes exact if the network contains many more silent synapses than 1-synapses required for the memory matrix ($p_1/p \to 0$).

In the following we compute how much time is necessary to have a fraction γ of the M pattern association to be stored to be completely connected. For one pattern association we need the generation of kl 1-synapses if the cell assemblies in population U have size k and the cell assemblies in population V have size l. Thus, we require $\beta^{kl} = \gamma$ or $\beta = \gamma^{1/kl}$. This is accomplished after time

$$T_\gamma = 1+\frac{\ln\frac{1-\gamma^{1/kl}}{1-p}}{\ln(1-p_e p)} \quad (31)$$
$$\approx 1+\frac{\ln(1-p)-\ln\left(-\frac{\ln\gamma}{kl}\right)}{p_e p} \sim \log(kl) , \quad (32)$$

if we again assume $p_1 \ll p$. For the approximation we assume additionally $p_e \ll 1$ and $\gamma, \beta \approx 1$. Thus, there appears to be only a weak dependence on assembly size k. However, loading the memory matrix to the high-fidelity limit $p_{1\epsilon}$ (see eq. 10) will cause difficulties for large k. Then T_γ increases not only due to increasing k, but possibly also due to increasing $p_{1\epsilon}$: for $p_{1\epsilon} \to p$, or equivalently $k \to \mathrm{ld}(\epsilon l/n)/(\lambda \mathrm{ld} p)$, eq. 32 is no longer valid and we obtain $T_\gamma \to \infty$ (see Fig. 5). Thus, in order to keep the network plastic we have to require $p > p_{1\epsilon}$. A reasonable choice could be $p = 2p_{1\epsilon}$. This would decrease the synaptic capacity C^S by factor 2 and thus imply $C(0.5) = C^S(0.5) = \ln 2$ for "incompressible" networks with $p_{1\epsilon} = 0.5$.

In this section we have hypothesized how a hippocampal learning system HC may train neocortical networks in order to increase their storage capacities by several orders of magnitudes and, by means of structural plasticity, effectively implement a memory compression by emulating full connectivity $p = 1$ (cf. section III-E). The two components that contribute in increasing the storage capacity of incompletely connected networks are the following:

1) *Structural plasticity* can serve to transform an incompletely connected network into an effectively completely connected network. That means, in the incompletely connected network without structural plasticity the connectivity within a cell assemblies remains incomplete (connectivity level p). In contrast, in the model including structural plasticity and training by a hippocampus-like structure as proposed in this section, the cell assemblies become completely connected with time (connectivity $\beta \approx 1$ after time T_β). The effect of this is, that a much larger number of pattern associations can be stored in a network with given size and connections (replace $p \ll 1$ by $\beta \approx 1$ in eq. 13).
2) *Pruning* or erasing functional irrelevant 0-synapses can directly increase the storage capacity per connection (Fig. 2). However, there is always a trade-off between plasticity and storage capacity: In a network where most silent synapses have been pruned, C^S can be very large (see eq. 16), but learning additional patterns will require much more time. It is reasonable to assume that with increasing lifetime more and more 0-synapses are pruned and therefore it becomes increasingly difficult to learn new information. In principle, a smaller number of real 0-synapses could be compensated by faster structural changes with larger p_e.

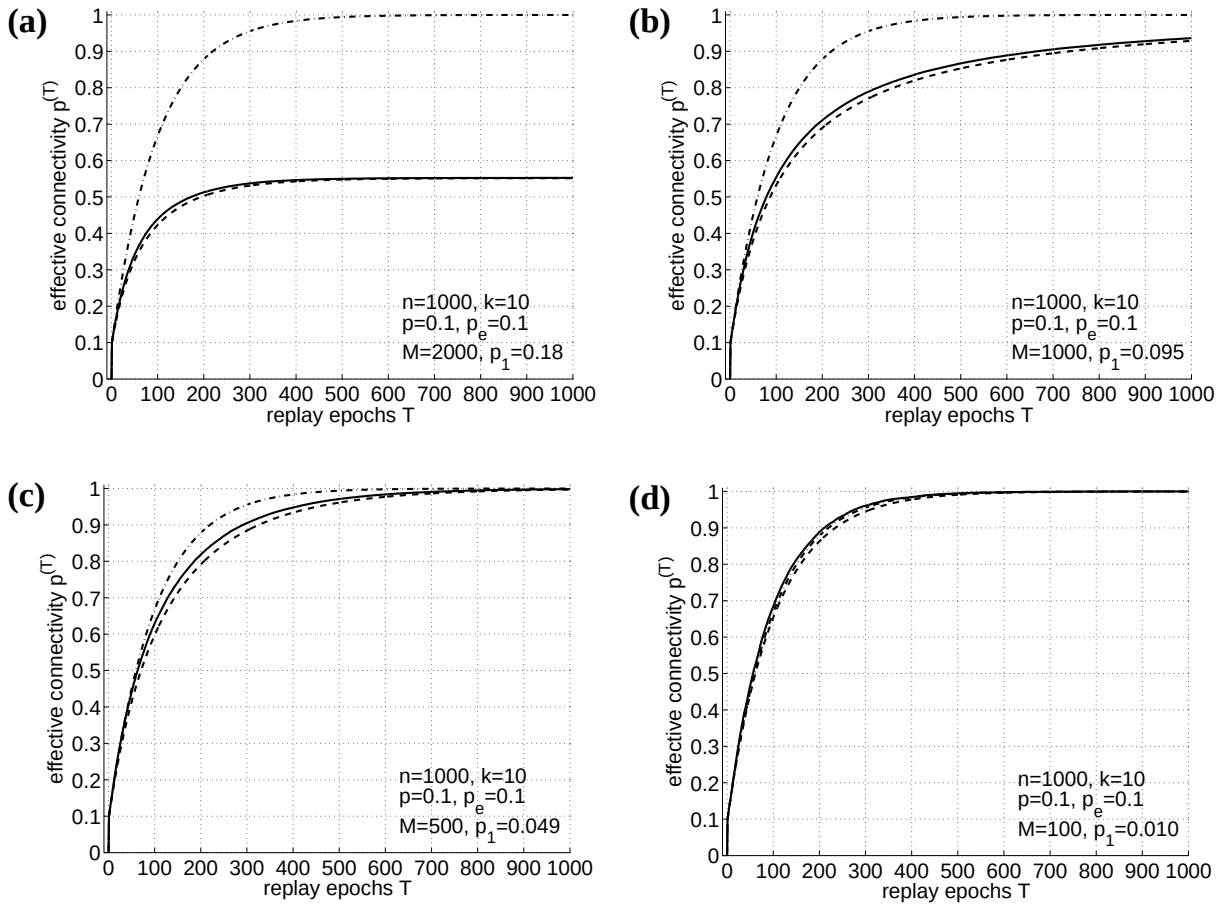

Fig. 4. Simulation of the structural plasticity model applied to the Willshaw associative memory model (n=1000 neurons per population, assembly size $k = 10$, connectivity $p = 0.1$, and erasing fraction $p_e = 0.1$). **a** : The effective connectivity $p^{(T)}$ (the fraction of realized 1-synapses) as a function of number of replay epochs T for $M = 2000$ stored associations which corresponds to a memory load of $p_1 = 0.18$ in the fully connected Willshaw model. The curves correspond to a simulation (solid) and to the theoretical values according to eq. 26 (dashed) and eq. 28 (dash-dotted). $p^{(T)} \to 1$ is not possible since $p_1 > p$. Approximation eq. 28 is bad. **b** : Same as (a) but for $M = 1000$ and $p_1 = 0.095$. For $p_1 \leq p$ full effective connectivity $p^{(T)} \to 1$ is possible, but the convergence is slow since $p_1 \approx p$. **c** : Same as (a) but for $M = 500$ and $p_1 = 0.049$. **d** : Same as (a) but for $M = 100$ and $p_1 = 0.010$. For $p_1 \ll p$ full effective connectivity $p^{(T)} \to 1$ is achieved quite fast and even approximation eq. 28 becomes good.

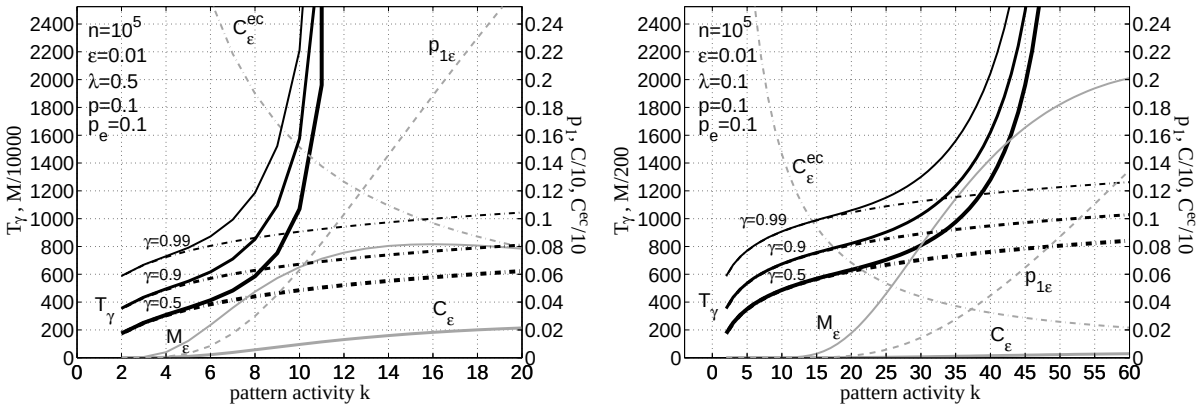

Fig. 5. The plot illustrates the dependence of T_γ (black) on the cell assembly size k. T_γ is number of replay epochs that is necessary for having a fraction γ of the M stored associations completely learned. The plot shows T_γ as computed by eq. 26 and $p^{(T_\beta)} = \beta$ using $\beta = \gamma^{1/k^2}$ (solid) and the approximation eq. 32 (dash-dotted) for $n = 10^5$ neurons per population, high-fidelity parameter $\epsilon = 0.01$, network connectivity $p = 0.1$, and structural plasticity where in each replay epoch a fraction $p_e = 0.1$ of the silent synapses are replaced. **Left** panel corresponds to fault tolerance with $\lambda = 0.5$, **right** panel to $\lambda = 0.1$. T_γ is relatively small and eq. 32 approximates well for small k, but T_γ diverges for $p_1 \to p = 0.1$. The gray curves are high-fidelity matrix load $p_{1\epsilon}$ (dashed), pattern capacity M_ϵ (solid medium), storage capacity C_ϵ (solid thick), and compression capacity C_ϵ^{ec} (dash-dotted).

V. Discussion

In this report we have investigated the information storage capacity of structural plasticity [19], [2] in combination with synaptic consolidation [5] and hippocampal training of cortical networks [1]. In section II we have computed simple upper bounds for the amount of information (in bits per synapse) that can be stored in cortical networks by synaptic plasticity and structural plasticity, respectively. Surprisingly, the upper bounds suggest that the total synaptic storage capacity of cortical networks may actually be dominated by structural plasticity, which has gained only little attention so far.

Then we showed in section III that the upper bound for structural plasticity (eq. 2) can actually be reached by simple models of synaptic networks. In particular, we showed for the Willshaw network model [10], [11] that in certain parameter regimes the storage capacity $C^S \sim \log n$ (in bits per synapse) can grow logarithmic with the neuron number if structural changes during learning are possible, for example the elimination of useless or silent synapses. In contrast, the storage capacity of static networks is limited by $C \leq 0.72$ bps [13].

Formally speaking, our analysis of the Willshaw model implies two separate phases of operation. In the first phase, information is stored by Hebbian learning in a fully connected static network. Structural plasticity occurs then in the second phase where the useless synapses are eliminated and thus the synaptic network gets compressed (see also [43], [36], [3]). In particular the assumption of fully connected networks during learning is unrealistic, both for biology and for VLSI implementations. Therefore, we developed in section IV a more realistic model of cortico-hippocampal interactions where Hebbian learning and structural plasticity occur simultaneously thereby maintaining a constant low level p of connectivity. In this model, the hippocampal learning system is capable of one-shot learning in diluted networks, but has only a relatively low storage capacity. By sequentially replaying the pattern information acquired by one-shot learning, the hippocampal system can train neocortical networks thereby boosting the storage capacity of neocortex. The hippocampal replay would be accompanied by ongoing structural changes such as pruning and regeneration of silent synapses. This procedure enables the limited number of synapses to "find their right places" where they can contribute most effectively to information storage. This way, even sparsely connected cortical networks can develop with time the same functional connectivity as fully connected networks.

Thus, our results imply the hypothesis that one function of structural plasticity may be to compensate for an incomplete network connectivity (cf. [19], [2]). This emulation of complete connectivity can massively increase the storage capacity of a typical cortical macrocolumn (corresponding to about 1mm^3 cortex containing about $n = 100000$ neurons each having about 10000 synapses [20], [21]). By learning in the incomplete static network ($p = 0.1$) the macrocolumn can store at most $M \leq 46040$ patterns or $C \leq 0.17$ bits per synapse, where the patterns should have a relatively high activity ($k = 340$ or $k = 493$, see eqs. 13,14; see also [3]). In comparison, emulating complete connectivity by structural plasticity and pruning functionally irrelevant silent synapses (section III-E) can boost the capacity to $M \geq 8958305$ associations or $C^S \geq 3.14$ bits per synapse, where the patterns should be very sparse (e.g., $k = 17$ or even $k \leq 5$; see table 3 in [3]). Thus, a neocortical network trained by the hippocampus and capable of structural changes as suggested by our model could store about factor 200 more pattern associations and still about factor 20 more information per physically realized synapse.

References

[1] G. Buzsaki, "The hippocampo-neocortical dialogue." *Cerebral Cortex*, vol. 6, pp. 81–92, 1996.

[2] A. Stepanyants, P. Hof, and D. Chklovskii, "Geometry and structural plasticity of synaptic connectivity." *Neuron*, vol. 34, pp. 275–288, 2002.

[3] A. Knoblauch, "On compressing the memory structures of binary neural associative networks," Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany, Internal Report HRI-EU 06-02, April 2006.

[4] B. Graham and D. Willshaw, "Improving recall from an associative memory." *Biological Cybernetics*, vol. 72, pp. 337–346, 1995.

[5] S. Fusi, P. Drew, and L. Abbott, "Cascade models of synaptically stored memories." *Neuron*, vol. 45, pp. 599–611, 2005.

[6] C. Shannon and W. Weaver, *The mathematical theory of communication.* Urbana/Chicago: University of Illinois Press, 1949.

[7] T. Cover and J. Thomas, *Elements of information theory.* New York: Wiley, 1991.

[8] C. Petersen, R. Malenka, R. Nicoll, and J. Hopfield, "All-or-none potentiation at CA3-CA1 synapses." *Proceedings of the National Academy of Science, USA*, vol. 95, pp. 4732–4737, 1998.

[9] U. Vollmer and A. Strey, "Experimental study on the precision requirements of rbf, rprop and bptt training." in *Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN)*, 1999, pp. 239–244.

[10] D. Willshaw, O. Buneman, and H. Longuet-Higgins, "Non-holographic associative memory." *Nature*, vol. 222, pp. 960–962, 1969.

[11] G. Palm, "On associative memories." *Biological Cybernetics*, vol. 36, pp. 19–31, 1980.

[12] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the National Academy of Science, USA*, vol. 79, pp. 2554–2558, 1982.

[13] G. Palm, "Memory capacities of local rules for synaptic modification. A comparative review." *Concepts in Neuroscience*, vol. 2, pp. 97–128, 1991.

[14] C. Woolley, "Structural plasticity of dendrites." in *Dendrites.*, G. Stuart, N. Spruston, and M. Häusser, Eds. Oxford, UK: Oxford University Press, 1999, pp. 339–364.

[15] S. Witte, H. Stier, and H. Cline, "In vivo observations of timecourse and distribution of morphological dynamics in Xenopus retinotectal axon arbors." *Journal of Neurobiology*, vol. 31, pp. 219–234, 1996.

[16] F. Engert and T. Bonhoeffer, "Dendritic spine changes associated with hippocampal long-term synaptic plasticity." *Nature*, vol. 399, pp. 66–70, 1999.

[17] P. Lennie, "The cost of cortical computation." *Current Biology*, vol. 13, pp. 493–497, 2003.

[18] M. Abeles, *Corticonics: Neural circuits of the cerebral cortex.* Cambridge UK: Cambridge University Press, 1991.

[19] P. Poirazi and B. Mel, "Impact of active dendrites and structural plasticity on the memory capacity of neural tissue." *Neuron*, vol. 29, pp. 779–796, 2001.

[20] V. Braitenberg and A. Schüz, *Anatomy of the cortex. Statistics and geometry.* Berlin: Springer-Verlag, 1991.

[21] B. Hellwig, "A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex." *Biological Cybernetics*, vol. 82, pp. 111–121, 2000.

[22] D. Hebb, *The organization of behavior. A neuropsychological theory.* New York: Wiley, 1949.
[23] V. Braitenberg, "Cell assemblies in the cerebral cortex." in *Lecture notes in biomathematics (21). Theoretical approaches to complex systems.*, R. Heim and G. Palm, Eds. Berlin Heidelberg New York: Springer-Verlag, 1978, pp. 171–188.
[24] G. Palm, *Neural Assemblies. An Alternative Approach to Artificial Intelligence.* Berlin: Springer, 1982.
[25] ——, "Cell assemblies as a guideline for brain research." *Concepts in Neuroscience*, vol. 1, pp. 133–148, 1990.
[26] E. Rolls, "A theory of hippocampal function in memory." *Hippocampus*, vol. 6, pp. 601–620, 1996.
[27] R. Fay, U. Kaufmann, A. Knoblauch, H. Markert, and G. Palm, "Integrating object recognition, visual attention, language and action processing on a robot using a neurobiologically motivated associative architecture." in *Proceedings of the NeuroRobotics GI Workshop at 27th German Conference of Artificial Intelligence*. University of Ulm, 2004.
[28] R. Hecht-Nielsen, *Confabulation theory.* Heidelberg: Springer-Verlag, 2007.
[29] A. Knoblauch, R. Kupper, M.-O. Gewaltig, U. Körner, and E. Körner, "A cell assembly based model for the cortical microcircuitry." *Neurocomputing*, vol. 70(10-12), pp. 1838–1842, 2007.
[30] M. Minsky and S. Papert, *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press, 1969.
[31] T. Kohonen, *Associative memory: a system theoretic approach.* Berlin: Springer, 1977.
[32] R. Prager and F. Fallside, "The modified Kanerva model for automatic speech recognition." *Computer Speech and Language*, vol. 3, pp. 61–81, 1989.
[33] D. Greene, M. Parnas, and F. Yao, "Multi-index hashing for information retrieval." *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, pp. 722–731, 1994.
[34] X. Mu, M. Artiklar, P. Watta, and M. Hassoun, "An RCE-based associative memory with application to human face recognition." *Neural Processing Letters*, vol. 23, pp. 257–271, 2006.
[35] M. Rehn and F. Sommer, "Storing and restoring visual input with collaborative rank coding and associative memory." *Neurocomputing*, vol. 69, pp. 1219–1223, 2006.
[36] A. Knoblauch, "Neural associative memory for brain modeling and information retrieval." *Information Processing Letters*, vol. 95, pp. 537–544, 2005.
[37] K. Steinbuch, "Die Lernmatrix." *Kybernetik*, vol. 1, pp. 36–45, 1961.
[38] A. Knoblauch, "Asymptotic conditions for high-capacity neural associative networks." Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany, Internal Report HRI-EU 07-02, February 2007.
[39] ——, "On the computational benefits of inhibitory neural associative networks." Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany, Internal Report HRI-EU 07-05, May 2007.
[40] G. Palm and F. Sommer, "Associative data storage and retrieval in neural nets." in *Models of Neural Networks III*, E. Domany, J. van Hemmen, and K. Schulten, Eds. New York: Springer-Verlag, 1996, pp. 79–118.
[41] E. Körner, M.-O. Gewaltig, U. Körner, A. Richter, and T. Rodemann, "A model of computation in neocortical architecture." *Neural Networks*, vol. 12, pp. 989–1005, 1999.
[42] R. O'Reilly and K. Norman, "Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework." *Trends in Cognitive Sciences*, vol. 6(12), pp. 505–510, 2002.
[43] A. Knoblauch, "Optimal matrix compression yields storage capacity 1 for binary Willshaw associative memory." in *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003.*, ser. LNCS 2714, O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, Eds. Berlin: Springer Verlag, 2003, pp. 325–332.