Neural Networks 25 (2012) 84–93

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Regarding the temporal requirements of a hierarchical Willshaw network

João Sacramento [a,*], Francisco Burnay [b], Andreas Wichert [a]

[a] INESC-ID Lisboa and Instituto Superior Técnico, Technical University of Lisbon, Av. Prof. Dr. Aníbal Cavaco Silva, 2744-016 Porto Salvo, Portugal
[b] Instituto de Plasmas e Fusão Nuclear, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 30 March 2011
Received in revised form 14 July 2011
Accepted 17 July 2011

Keywords:
Associative memory
Willshaw model
Retrieval time
Hierarchical neural network
Sparse coding

ABSTRACT

In a recent communication, Sacramento and Wichert (2011) proposed a hierarchical retrieval prescription for Willshaw-type associative networks. Through simulation it was shown that one could make use of low resolution descriptor patterns to decrease the total time requirements of recalling a learnt association. However, such a method introduced a dependence on a set of new parameters which define the structure of the hierarchy. In this work we compute the expected retrieval time for the random neural activity regime which maximises the capacity of the Willshaw model and we study the task of finding the optimal hierarchy parametrisation with respect to the derived temporal expectation. Still in regard to this performance measure, we investigate some asymptotic properties of the algorithm.

1. Introduction

In the strictest technical sense, an associative memory model is designed to solve a variation of the classical nearest neighbour determination problem. Instead of finding a solution for the original labelled classification task formulation (Cover & Hart, 1967; Fix & Hodges, 1951; Minsky & Papert, 1969), an associative memory is a system that stores information about a finite set of M associations of the form

$$S := \{(\mathbf{x}^\mu \mapsto \mathbf{y}^\mu) : \mu = 1, \ldots, M\}, \quad (1)$$

with most memory models assuming the patterns are binary vectors, i.e., $\mathbf{x} \in \{0, 1\}^m$ and $\mathbf{y} \in \{0, 1\}^n$. Given a possibly corrupt or incomplete pattern $\tilde{\mathbf{x}} \in \{0, 1\}^m$, the system should be able to find the best-matching (or rather, the 'nearest neighbour') $\mathbf{x}^\mu$ with respect to a desired similarity metric and then return a pattern $\hat{\mathbf{y}} \in \{0, 1\}^n$ ideally corresponding to the originally stored $\mathbf{y}^\mu$. Thus, the original association $(\mathbf{x}^\mu \mapsto \mathbf{y}^\mu)$ ought to be restored through a robust recall process.

Three different yet closely related tasks are usually identified with the above process: when $n = 1$ and $m \gg n$, the memory solves a binary classification problem over the labels 'known' and 'unknown' (learnt patterns being associated with the former) and is said to perform familiarity discrimination (Bogacz & Brown, 2003; Bogacz, Brown, & Giraud-Carrier, 2001; Greve, Sterratt, Donaldson, Willshaw, & van Rossum, 2009); when $m = n$ and $\forall \mu$: $\mathbf{x}^\mu = \mathbf{y}^\mu$ an autoassociative function is carried out and the memory is expected to perform pattern completion or correction; finally, the case of arbitrary m, n and $\mathbf{x}^\mu, \mathbf{y}^\mu$ is called heteroassociation. The latter is most easily comparable with a standard von Neumann computer memory.

The general quality of a neural associative memory implementation can be assessed with respect to several quantities. The most addressed in the literature is the storage capacity (Amit, Gutfreund, & Sompolinsky, 1985; Gardner, 1988; Knoblauch, Palm, & Sommer, 2010; Palm, 1980; Palm & Sommer, 1992; Willshaw, Buneman, & Longuet-Higgins, 1969), which is typically measured through the critical pattern capacity α_c (simply given by a normalisation of the number of patterns M over the number of content neurons n) or through the more general network capacity C, measured in bits per synapse (bps) and defined as the maximal mutual Shannon information (Cover & Thomas, 2006; Shannon, 1948) between stored and retrieved vectors $I(\mathbf{x}^1, \ldots, \mathbf{x}^M; \hat{\mathbf{y}}^1, \ldots, \hat{\mathbf{y}}^M)$ normalised over the count of synaptic contacts required by the network. The latter is usually preferable since it takes into account both the required resources and the information content of the patterns.

Another quantity of interest is the expected time necessary for the learning and retrieval processes to terminate, generally presented as a count of elementary operations and as a function of some of the parameters which define the memory task. The temporal requirements of an associative memory model deserve attention from both the technical and biophysical perspectives, partly determining the model's efficiency. From a purely algorithmic point of view, the pattern recognition community is currently facing the challenge of solving as quickly as possible the nearest neighbour determination problem in high dimensional space (large m, n) and for high pattern loads (large M), to cope with the

* Corresponding author. Tel.: +351 21 423 32 31; fax: +351 21 314 48 43.
E-mail addresses: joao.sacramento@ist.utl.pt (J. Sacramento), francisco.burnay@ist.utl.pt (F. Burnay), andreas.wichert@ist.utl.pt (A. Wichert).

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2011.07.005

increase in size of modern data sets. When the analogy with biological neural networks is to be taken in consideration, the temporal efficiency is equally important as it is linked to the energetic requirements of membrane potential determination (Knoblauch et al., 2010).

While attempting to solve the associative memory task using a neural computation approach a lot of effort has been placed in developing and studying recurrent networks (i.e., with feedback couplings), as for finite systems they can provide stronger error tolerance to pattern noise through an increase in the size of the basins of attraction (Amit et al., 1985; Derrida, Gardner, & Zippelius, 1987; Gardner, 1988; Gardner-Medwin, 1976; Golomb, Rubin, & Sompolinsky, 1990; Hopfield, 1982; Kosko, 1987; Palm & Sommer, 1992; Sommer & Palm, 1999), in exchange for additional iterations during the retrieval process.

If our aim is the least computational effort, it is known that under the sparse random coding regime (Barlow, 1972; Field, 1994; Olshausen & Field, 2004) the simpler Willshaw net achieves a high storage capacity (viz., $C = \log 2 \approx 0.7$ bits per synapse in the limit of $m, n \to \infty$) using a biologically plausible local learning rule of the Hebbian type and a parallel 'single-shot' retrieval prescription (Amari, 1989; Knoblauch et al., 2010; Nadal & Toulouse, 1990; Palm, 1980; Steinbuch, 1961; Willshaw et al., 1969). Besides allowing for high storage capacities, the coding restriction on the ℓ_0 pseudo-norm of the pattern vectors (or, equivalently, on the ℓ_1 norm since we assume the patterns are binary) imposed by the sparseness requirement also reduces the temporal complexity of learning and retrieval and seems to be in accordance with the signalling and maintenance energy budget of the mammalian brain (Laughlin & Sejnowski, 2003; Lennie, 2003; Levy & Baxter, 1996).

Due to the inherently parallel synchronous update mechanism, the temporal benefits of the single-shot retrieval procedure employed by the Willshaw model can only be fully exploited using specialised hardware constructs. Attempting to decrease the retrieval time on sequential computer implementations, a recent communication suggested the use of a hierarchical retrieval prescription in order to take advantage of the sparse structure of the stored patterns (Sacramento & Wichert, 2011).

However, the proposed model introduced a dependence on a new set of integer parameters which defined the hierarchy, and were obtained through exhaustive combinatorial search. It remained unclear whether this problem was tractable for high dimensional pattern spaces, and whether a heuristic approach could be derived in order to avoid the integer constrained optimisation. In this work we address these issues and compute refined time expectations for finite memories. En passant, we also show that asymptotically the hierarchical refinement procedure reduces the temporal complexity of the retrieval process when compared with the original single-layer network.

The rest of this paper is organised as follows. In Section 2, we review the network model presented in Sacramento and Wichert (2011) and derive exact expectations for the time requirements of learning and retrieval. Then, in Section 3, we analyse the optimisation task of determining the hierarchical configuration which minimises the time expectation we obtain. We show that even though the problem is difficult to solve analytically, the solution space grows with a polynomial of the pattern space dimension and can thus be tackled through enumeration. We also provide a heuristic method to solve the task and verify its validity empirically for several network configurations.

2. Model characterisation

In the first part of this work we will see how associative networks of the Willshaw type use a kind of plasticity (namely synaptic) and a local Hebbian learning rule to store and recall memory traces. After defining the equations which govern the learning and retrieval processes of the original single-layer model and its hierarchical variant, we will change focus to the statistical characterisation of their temporal requirements. Following the algorithm analysis tradition, we will adopt as our time measure a simple count of the number of necessary operations that either a sequential computer or a specialised hardware construct can perform in constant $O(1)$ time. Asymptotic comparisons using Bachmann–Landau notation can then be made, as well as finite numerical evaluations for particular cases. This approach has found widespread use across the literature, as it is mathematically tractable and abstract enough to establish a comparison between different models and implementation architectures.

2.1. Network equations for learning and retrieval

The original Willshaw model is a single-layer neural network comprising two populations of McCulloch–Pitts binary threshold neurons (McCulloch & Pitts, 1943): an address population of m neurons capable of establishing synaptic connections with n neurons which form the content population. We can then interpret the silent-firing (0–1) activity patterns of each set of neurons at a given synchronous time frame as our binary input (address) and output (content) vectors.

During the learning phase, we assume each pair $(\mathbf{x}^\mu, \mathbf{y}^\mu)$ from S is presented to the network independently at time μ. A Hebbian-type learning rule is applied and formation of synapses can occur as a consequence of new stimuli. Willshaw networks employ a particular non-additive clipped Hebb rule, where the synaptic strength factor is disregarded, i.e., we only care to check whether a synapse between any two neurons i and j is either present or not. Thus, the entire state of a network can be represented by a binary weight matrix $\mathbf{W} \in \{0, 1\}^{m\times n}$, where $W_{ij} = 0$ denotes an absent synapse from pre-synaptic neuron i to post-synaptic neuron j and $W_{ij} = 1$ a present one. After learning the M associations of S, the entries of the synaptic connectivity matrix are then given by

$$W_{ij} = \min\left(1, \sum_{\mu=1}^{M} x_i^\mu y_j^\mu\right), \tag{2}$$

which results in bidirectional synapses $\forall i, j,\ \ W_{ij} = W_{ji}$ for the autoassociative case of $m = n$ and $\forall \mu,\ \mathbf{x}^\mu = \mathbf{y}^\mu$.

Notice how this learning prescription leads to distributed storage, in the sense that each synaptic contact W_{ij} can store information about more than one pattern pair. It is also the simplest possible realisation of the hypothesis of Hebb (1949), as the synaptic update procedure is local and bounded. Note that due to the nonlinearity of the rule, S cannot in general be recovered from $\mathbf{W}$.

The retrieval process starts when the address population fires according to a certain cue pattern $\tilde{\mathbf{x}}$. The activity state $\hat{\mathbf{y}}$ of the content neuron population which will yield the output pattern of the network must then be updated. Each unit j computes (locally) its dendritic potential, corresponding to the sum of the incoming excitatory signals,

$$s_j = \sum_{i=1}^{m} W_{ij}\tilde{x}_i, \tag{3}$$

over which a nonlinear activation function is applied

$$\hat{y}_j = H[s_j - \Theta], \tag{4}$$

where H is the Heaviside step function. Note that the parameter Θ determines the highly nonlinear threshold operation which denoises the output and its choice will critically influence the quality of the recovered pattern $\hat{\mathbf{y}}$. Optimal threshold determination,