

Improved Willshaw Networks with Local Inhibition

Philippe Tigréat, Vincent Gripon, Pierre-Henri Horrein

Electronics Department, Telecom Bretagne

Brest, France

Email: {philippe.tigreat, vincent.gripon, ph.horrein}@telecom-bretagne.eu

Abstract—Willshaw networks are a type of associative memories with a storing mechanism characterized by a strong redundancy. Namely, all the subparts of a message get connected to one another. We introduce an additional specificity, by imposing the constraint of a minimal space separating every two elements of a message. This approach results from biological observations, knowing that in some brain regions, a neuron receiving a stronger stimulation can inhibit its neighbors within a given radius. We experiment with different values of the inhibition radius introduced, and we study its impact on the error rate in the retrieval of stored messages. We show that this added constraint can result in significantly better performance of the Willshaw network.

Keywords—Willshaw Networks; Clique-Based Neural Networks; Lateral Inhibition.

I. INTRODUCTION

Associative memories are a type of computer memories that are part of the broader category of content-addressable memories. Where addressable memories associate an address with a piece of data, associative memories have the characteristic of associating patterns to one another. Among this group, we distinguish between hetero-associative memories, and auto-associative memories. An hetero-associative memory will associate together patterns in pairs. For instance, if the pattern p_1 was associated with pattern p_2 , the request p_1 will bring the response p_2 . Auto-associative memories follow a different principle, as they will associate a pattern with itself. The main application of these memories is pattern completion, where a request made of a subpart of a stored message will get as response the completed pattern. It is today widely accepted that the working principle of the brain can often be likened to the operation of an associative memory.

The prominent model for associative memories was introduced by John Hopfield [1]. Hopfield networks are associative memories made of a set of N neurons that are fully interconnected. The training of these networks, given n binary vectors x^μ of length N , consists in modifying the weight matrix W according to the formula:

$$w_{ij} = \frac{1}{n} \sum_{\mu=1}^n x_i^\mu x_j^\mu, \quad (1)$$

where element w_{ij} at the crossing between line i and column j of W is the real-valued connection weight from neuron i to neuron j .

As connections are reciprocal and not oriented, we have :

$$w_{ij} = w_{ji} \quad \forall i, j \in \llbracket 1, N \rrbracket \quad (2)$$

for any indices i and j in the list of neurons, which makes W symmetrical.

The binary values considered for the stored messages are usually -1 and 1, but can be adapted to work with other binary alphabets. The Hopfield model has a limited efficiency, in particular it doesn't allow a storage of more than $0.14N$ messages [2]. The limits of the model can be explained by the facts that each entry of the matrix is modified at every time step of the storing procedure, and that the changes are made in both directions and can therefore cancel each other out. This overfitted characteristics of associative memories is very different from that observed in learning applications. Indeed, an overfitted learning system recognizes only the training samples and fails at generalizing to novel inputs. To the contrary, an overfitted storing system recognizes everything and does not discriminate anymore between stored and nonstored data.

Willshaw networks [3] are another model of associative memories in which information is carried by the existence or absence of connections. Its material is made of a set of N neurons and N^2 potential connections between them. A message is then a fixed size subset of the N neurons, and can be represented by a sparse vector of length N with ones at these neurons' positions and zeros everywhere else. The connection weights are binary, and the active units in a message get fully interconnected as soon as it is memorized, thus forming a clique. Figure 1 gives an example of such a network. The performances of Willshaw networks are way superior to those of Hopfield memories, given that stored messages are sparse (i.e., they contain a small proportion of nonzero elements). Further theoretical and numerical comparison between Hopfield and Willshaw networks can be found in [4]–[7].

Recently, a novel type of associative memories was proposed by Gripon et al. [8], called Gripon-Berrou Neural Networks (GBNNs) or clique-based neural networks. These associative memories make use of powerful yet simple error correcting codes. These networks consider input messages to be nonbinary, and more precisely to be words in a finite alphabet of size l . This specific structure allows the separation of nodes into different clusters, each being constituted of the same number l of nodes. Connections between nodes inside a given cluster are forbidden, only the connections between nodes in two different clusters are allowed. There again, this model brings a significantly improved performance as compared to the former state-of-the-art of associative memories, namely Willshaw networks [9]–[11]. For instance, with 2048 nodes and 10000 stored messages of order 4 and 2-erasures queries, a Willshaw network will have an error rate close to 80%, while a clique-based neural network will only make 20% of wrong retrievals.

In both Hopfield and Willshaw models, the number of

messages the network can store and retrieve successfully is linearly proportional to the number of nodes, with a greater proportionality constant for Willshaw networks [5]. In clique-based neural networks however, storage capacity grows quadratically as a function of the number of units.

One of the objectives of the present work is to explain the performance improvement brought by the separation of the network into clusters. We therefore study a network that can be considered as an intermediate between the Willshaw and Gripon-Berrou models. More precisely, our proposed model adds a locally exclusive rule for nodes to be active in the network.

We focus here on the phenomenon observed in biological neural networks, called lateral inhibition [12]. It can also be referred to as surround suppression [13]. This translates in the inhibition exerted by some neurons on their close neighbors when these have an activity inferior to their own. We consider that the Willshaw model is not totally biologically plausible, as it does not feature this phenomenon of inhibition of close neighbors. We propose a model of Willshaw network that is improved in terms of plausibility, by the introduction of local inhibition that results in the prohibition of short-range connections. We show that this modification brings a performance improvement in the retrieval of stored messages.

Section II introduces Willshaw networks and biological considerations related to our model. Section III details modifications in our implementation as compared to the classic Willshaw model, including the constraint applied on the space between connected neurons. Section IV presents the results we obtain, and gives some theoretical explanations.

II. WILLSHAW NETWORKS AND BIOLOGICAL CONSIDERATIONS

Willshaw networks are models of associative memories constituted of a given number of neurons. A stored message, or memory, is a combination of nodes taken in this set. The storage of this information element corresponds to the creation of connections with unitary weights between every two neurons in this message. The graphical pattern thus formed is termed "clique". The storing process of n binary vectors x^μ of length N is equivalent to the modification of elements of the network's connection matrix W , according to the formula:

$$w_{ij} = \max_{\mu} x_i^{\mu} x_j^{\mu} \quad (3)$$

Note that here, the max operator is performed coefficient-wise. Equivalently, the connection weight between nodes i and j is equal to 1 if, and only if, those two nodes are used in a same message among the n stored messages.

The network's density d is defined as the expected ratio of the number of 1s in the matrix W to the number of 1s it would contain if every possible message was stored. In the case of uniform i.i.d. messages, all containing exactly c active nodes, binomial arguments quickly lead to the formula:

$$d = 1 - \left(1 - \frac{\binom{c}{2}}{\binom{N}{2}}\right)^n \quad (4)$$

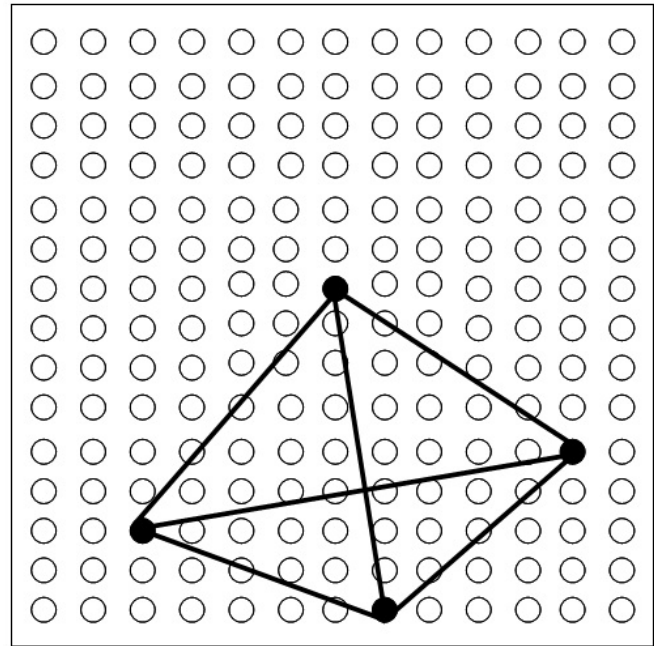


Figure 1. Willshaw network. Black nodes represent a message, and the connections between them are the means of its storage in the network.

The efficiency of an associative memory is defined as the ratio of the maximal amount of information carried by the messages it is capable of storing then retrieving with high probability, over the total information quantity represented by its set of connection weights. For a Willshaw network with N nodes, the number of potential connections or binary resource is

$$Q = \frac{N(N-1)}{2} [\text{bits}]. \quad (5)$$

After M messages have been stored in the network, the amount of information it contains is

$$B = M \left(\log_2 \left(\binom{N}{c} \right) \right) [\text{bits}]. \quad (6)$$

Hence the efficiency of a Willshaw network is

$$\eta = \frac{2M \left(\log_2 \left(\binom{N}{c} \right) \right)}{N(N-1)}. \quad (7)$$

The maximal attainable efficiency is $\ln(2)$ [14].

The stimulation of a Willshaw network with an input request can be performed as the product of the sparse input vector by the network's connection matrix. The resulting vector then contains the output scores of the network's neurons. The score of a neuron is thus the sum of unitary stimulations it receives from the request elements it is connected to. Neurons must then be selected based on their score.

Figure 2 defines a procedure that can be used for the recovery of a complete message from a subpart of its content. The Global Winner-Takes-All step consists in discarding all active neurons with a score below the maximum.

We aim to modify classic Willshaw networks in a way that is

Data: Subpart x of a stored message
Result: Set of nodes z active after treatment
 $z = x$
Repeat
 $y = Wz$
 $z = \text{GlobalWinnerTakesAll}(y)$
while (convergence not reached
and max. nb. of iterations not reached)
Return z

Figure 2. Message retrieval procedure in a classic Willshaw network

relevant in regard to biological observations. Emphasis is put on lateral inhibition, a phenomenon that has been observed in several areas of the brain. It is notably present in sensory channels. For vision, it operates at the level of retinal cells and allows an increase in contrast and sharpness of signals relayed to the upper parts of the visual cortex [13] [17]. In the primary somatosensory area of the parietal cortex, neurons receive influx coming from overlapping receptive fields. The Winner-Takes-All operation resulting from the action of inhibitory lateral connections allows to localize precisely tactile stimuli, despite the redundancy present in the received information [18]. The same scheme of redundancy among sensory channels, and filtering via lateral inhibition, is present in the auditory system [12]. WTA is observed in the inferior colliculus and in upper levels of the auditory processing channel.

III. PREVENTING CONNECTIONS BETWEEN NEIGHBOR NEURONS

Classic Willshaw networks have no topology. Their material is constituted with a list of neurons each having an index as sole referent. There is neither a notion of spatial position in these networks, nor, a fortiori, of spatial distance. We get closer here to a real neural network, by arranging them on a two-dimensional map. In the model we propose, the respective positions of two neurons impact the possibility for them to get connected together. The considered network is composed of a number N of nodes evenly distributed along a square grid, of side $S = \sqrt{N}$. Stored messages are of constant order, meaning they are all constituted of the same number of neurons. We forbid connections between nearby neurons. To this end, we apply a threshold σ on the spatial length of a connection. Stored messages must necessarily be conform to this constraint. Each message is formed in a random manner, units are chosen iteratively. Each new element of the message is picked from the positions left available after the removal of the neighbors of the formerly selected nodes, as indicated on Figure 3. One can consider the introduced constraint as applied on the network’s material, as the weights of a predetermined set of short-range connections will be enforced to stay null all along the network’s life. During the formation of a message, it is practical to pick neurons to satisfy this constraint in a sequential manner, with a local inhibition applied on a neuron’s neighborhood from the moment it is selected until the message generation is complete.

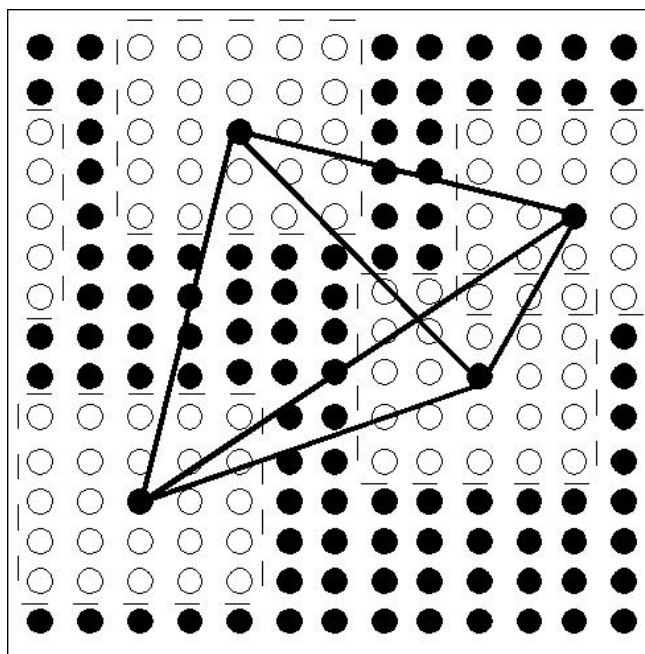


Figure 3. Willshaw network with a constraint on local connections.

A link can be drawn between this approach and Kohonen Self-Organizing Maps [15], where closeby neurons encode more similar information. Long-range distance therefore separates information elements that are different in nature, whereas shorter-range distance depicts a difference in degree. Local competition is particularly relevant in this scheme.

During retrieval, the network is stimulated iteratively with a request that will most often change from one iteration to the next. Each node of the request will first stimulate every other element it is connected to. Scores are initialized at zero at the start of every iteration, and each stimulation is a unitary increment to the score of the receiver unit. For the first iteration, after the stimulation we apply a global Winner-Takes-All rule, which consists in excluding from the research scope all units that do not achieve the maximal score observed in the network. We know indeed that the neurons from the searched message will all have the maximum possible score, equal to the number of elements in the request. Once non-maximum elements are put to zero, we only pay interest in the remaining neurons during the rest of the retrieval process. Moreover, for every iteration after the first one, neurons in the new request are the only ones that can receive stimulation as the algorithm proceeds to only discard neurons from then on.

Thereafter, we can keep using the global Winner-Takes-All principle iteratively, but other algorithms such as Global Winners-Take-All (GWsTA) or Global Losers-Kicked-Out (GLsKO) [16] are more efficient in discriminating the right nodes from the spurious ones that can appear during retrieval.

Global Winners-Take-All relies on the calculation of a threshold score to select winner neurons. This threshold is chosen such that neurons with an activity above it are in number at least as large as the order of stored messages.

Data: Subpart x of a stored message
Result: Set of nodes z active after treatment
Phase I
 $y = Wx$
 $z = \text{GlobalWinnerTakesAll}(y)$
Phase II
Repeat
 $y = Wz$
 $a = \text{active nodes in } y$
 $m = \text{nodes in } a \text{ with minimal score}$
 $z = a - m$
while (convergence not reached
and max. nb. of iterations not reached)
Return z

Figure 4. Message retrieval procedure in a Willshaw network with lateral inhibition

Global Losers-Kicked-Out consists in putting off, at each iteration, all the units that do not have the highest score, or a subgroup sampled randomly in this ensemble.

These two algorithmic techniques allow to get rid of an important proportion of false-positives. In the clique-based GBNN, clusters play a similar role.

The iterative nature of the process means that a message retrieved as output from the network is typically reinjected in it until input and output no longer differ. A limited number of iterations is applied in the case where the network would not converge to a stable solution, an observable case in which it can oscillate between two states.

In addition to these two stopping criteria that are the maximum number of iterations and convergence, comes a third one which is the identification of a clique. Indeed, if we observe that the units still active after an iteration are in number equal to the order of stored messages, and that they all have the same score, this means it is a stored message. This ensemble is then retained as the response given by the network for the current request.

Figure 4 shows the message retrieval procedure used in the results we present. Phase II uses Global Losers-Kicked-Out.

We experiment the storage of messages of order c in the connection matrix of the network. Messages are formed with the constraint of a minimal space between connected nodes. Two units in a message must be spaced apart at a distance superior to a minimum σ . In order to ease computations and avoid edge effects, we choose to use the L_1 distance, even though we believe this method should work using any distance. This way, when picking a node x for a message, all nodes located in a square grid centered on x , of side $2\sigma+1$, are excluded from the possible choices for the elements of the message remaining to be filled. Moreover, this distance is applied in a cyclic way, meaning a node located on the right edge of the grid will be considered a direct neighbor of the element located at the crossing between the same line and the left edge of the grid. All four corners of the grid will also be neighbors to one another. As a result, the spatial distance we consider can be written:

$$d(A, B) = \min(\text{abs}(x_a - x_b), S - \text{abs}(x_a - x_b), \text{abs}(y_a - y_b), S - \text{abs}(y_a - y_b)), \quad (8)$$

where abs denotes the absolute value function.

We call the network so described a torus. During retrieval, only a sample from the nodes of the complete message are stimulated, the inputs are subparts of stored messages. Units that are close to elements of an input will not reach the maximum score in the network, and will therefore be ruled out after the first Global Winner-Takes-All operation. During the second phase of the algorithm, nodes in the vicinity of input neurons will also be more likely to reach a low score if they are activated, and to be discarded. Hence, the local inhibition used initially during the creation of messages impacts the retrieval process as well.

We pay interest in the network's ability to return the exact memory associated to a request. Hence every difference, even marked by a single unit, between the expected pattern and the network's output is counted as an error.

We measure the performance of the network as the ratio of the number of successfully retrieved messages over the total number of requests.

Various parameters can impact this performance, albeit to different degrees:

- the length S of the grid's side
- the number M of stored messages
- the minimal space σ between two elements of a message
- the order c of stored messages
- the number of erasures c_e applied on stored messages to obtain the corresponding request messages

Let's note that the constraint applied on the length of connections reduces the number of messages one can form for a given network. It thus lowers the information quantity carried by single messages. Hence, there is a tradeoff on the individual quantity of information of the messages and the performance of the retrieval algorithm.

The behavior of this network is interesting in relation to Willshaw networks and clique-based neural networks, as it is close to a classic Willshaw network and displays the added feature of prohibited connections as observed in GBNNs. One could talk about sliding-window clustering here.

IV. RESULTS

For every configuration of the network, messages and requests we test, we store a set of thousands of messages in the network. These messages are generated randomly following the local inhibition pattern described in section III. We then request it with the full set of queries associated to stored messages.

For each network size, we observe that there is an optimal value of the minimal distance σ , that lowers the most significantly the error rate, as compared to the corresponding Willshaw network without constraint on local connections. For a given minimal distance, the reduction in error rate depends on the number of stored messages, with an optimal number of messages which is a function of the network size. For cliques of order 4 and 2 erasures, the maximal reachable improvement

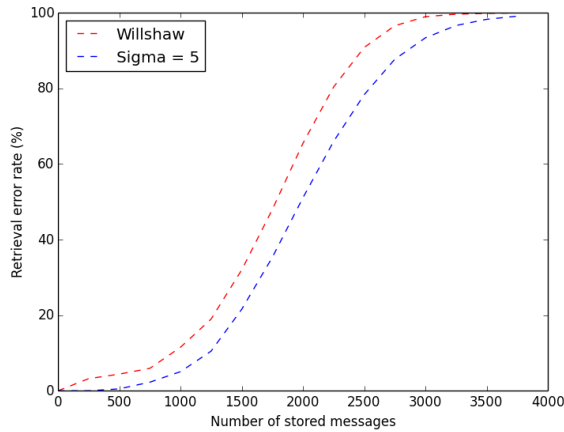


Figure 5. Evolution of the retrieval error rate with and without constraint $\sigma = 5$ in a network of side length 20 with 400 neurons, stored messages of order 6 and 1 erasure applied to form corresponding requests, with a maximum of 5 iterations.

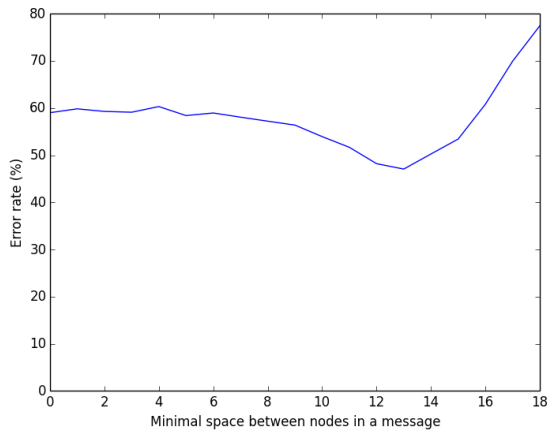


Figure 6. Minimal connection distance effect on performance in a network of side length 20 with 400 neurons, stored messages of order 4 and 2 erasures applied to form corresponding requests. The case where minimal spacing $\sigma = 0$ corresponds to a classic Willshaw network.

is close to 15%, and seems to be the same for all network sizes. In this configuration, the minimal distance bringing the best performance is approximately the third of the network side.

The evolution of the retrieval error rate as a function of the number of stored messages is slower with the appropriate constraint on connections than for a classic Willshaw network, as can be seen on Figure 5.

For a constant number of stored messages, the graph of the error rate as a function of σ is characterized by a progressive decay down to a minimum, followed by a rapid growth for upper values of σ , as shown on Figure 6.

This can be explained by two phenomena. On one part, the prohibition of a growing part of the possible connections gradually decreases the probability of a "false message", characterized by the intrusion of a spurious node in the output. The existence of a node that is connected to all elements in

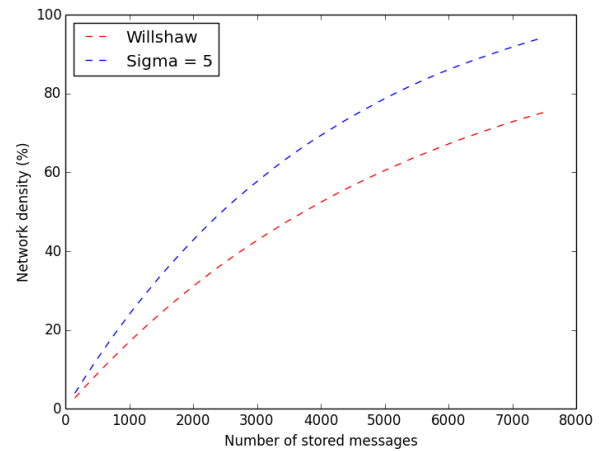


Figure 7. Evolution of the density with and without constraint in a network of side length 20 with 400 neurons, stored messages of order 6.

a request yet is not part of the corresponding message will potentially cause an error. In fact, forbidding some connections has the effect of reducing the number of concurrent nodes susceptible to cause errors. We can estimate the mean number of concurrent nodes remaining after the choice of k neurons of a message:

$$N_{competitors} = N \left(1 - \left(\frac{(2\sigma + 1)^2}{N} \right) \right)^k$$

The corresponding number of nodes blocked by the constraint on connections is, on average:

$$N_{blocked} = N \left(1 - \left(1 - \left(\frac{(2\sigma + 1)^2}{N} \right) \right)^k \right)$$

This explains the decay phase in error rate observed for the first values of σ . Let's note that it comes with a decrease in the diversity of messages, namely the total number of different messages that can be stored in the network. Following this decay, the decrease in the number of concurrent nodes has another effect: the reuse of connections by different messages becomes more frequent as the choice for possible connections gets reduced. This comes to counteract the former phenomenon and raises the error rate.

The network density grows faster as messages are stored in the network, than for a classic Willshaw network, as shown on Figure 7. This is because of the decrease in number of possible connections due to the spacing constraint.

Besides, we observe that the maximal improvement in performance, for given values of c and c_e , varies little as a function of the network size. This can be explained by the fact that the minimal distance giving the best performance is approximately proportional to the side of the network. Consequently, the proportion of neurons in the network that cannot be connected to the $c - c_e$ neurons in the request remains more or less the same for different network sizes,

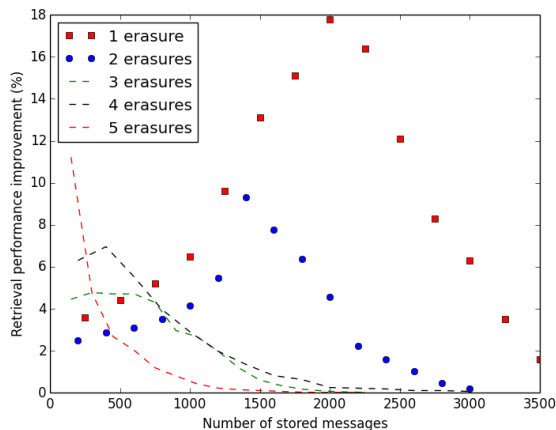


Figure 8. Maximal improvement obtained over a classic Willshaw network of side length 20 with messages of order 6.

with the optimal minimal distance.

The benefits brought by the constraint on connections seems stronger for smaller numbers of erasures. For erasures of about half the units of the messages, the maximum gain will be lower, yet for a high amount of erasures the performance can be noticeably enhanced by the added constraint. Moreover, the graph of the best performance improvement as a function of the number of stored messages has a shape that varies depending on the number of erasures. The performance improvement over a classic Willshaw network also depends on the number of messages stored in the network. It reaches a peak for a certain number of stored messages, and then decays when additional messages get stored. The higher the number of erasures, the earlier this peak is reached during the storage. For lower numbers of erasures, average performance gain increases more slowly at first but decays faster after the maximum is reached, as illustrated by Figure 8.

The greatest performance improvement is most often observed for a Willshaw network and a number of stored messages giving a performance between 40 and 60%. The performance gain is then often close to 15%.

V. CONCLUSION

We introduced a modified version of Willshaw neural networks that has interesting properties regarding storage capacity and retrieval performance. By prohibiting certain types of connections in the network, we observe that the retrieval ability can be enhanced, and that the value of the threshold on inter-neuron connection spacing has a direct impact on performance. This is relevant with observations on clique-based neural networks, in that it shows constraining connections in a Willshaw network modifies its capacity in a way that depends on the nature of the applied constraint. It is a step forward in understanding why the use of clusters in GBNNs brings significantly higher capacity as compared to Willshaw networks. To some extent, it also emulates biological observations of lateral inhibition in the brain and sensory channels, as we prevent neighbor neurons from connecting and

therefore let them compete for activity. This makes sense with a framework in which close-by neurons encode patterns that differ only in degree and where only one unit that resonates most with input stimuli must activate. Future work might involve a deeper theoretical analysis of this result, and a further attempt to explain the gain in performance brought by clustering the pool of neurons in GBNNs. It may also involve experimenting with other constraints on connections based on the relative locations of neurons.

ACKNOWLEDGEMENTS

This work was supported by the European Research Council under Grant ERC-AdG2011 290901 NEUCOD.

The authors would like to thank NVIDIA for providing us with a free graphics card allowing to speed up computations for the experiments performed during this work.

REFERENCES

- [1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, 1982, pp. 2554–2558.
- [2] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 33, no. 4, 1987, pp. 461–482.
- [3] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, "Non-holographic associative memory," *Nature*, 1969, pp. 960–962.
- [4] H. C. Longuet-Higgins, D. J. Willshaw, and O. P. Buneman, "Theories of associative recall," *Quarterly reviews of biophysics*, 3(02), 1970, pp. 223–244.
- [5] J. D. Keeler, "Comparison between kanerva's sdm and hopfield-type neural networks," *Cognitive Science*, vol. 12, no. 3, 1988, pp. 299–329.
- [6] D. Willshaw and P. Dayan, "Optimal plasticity from matrix memories: What goes up must come down," *Neural Computation*, vol. 2, no. 1, 1990, pp. 85–93.
- [7] A. Knoblauch, G. Palm, and F. T. Sommer "Memory capacities for synaptic and structural plasticity," *Neural Computation*, vol. 22, no. 2, 2010, pp. 289–341.
- [8] V. Gripon and C. Berrou, "Sparse neural networks with large learning diversity," *Neural Networks, IEEE Transactions on*, vol. 22, no. 7, 2011, pp. 1087–1096.
- [9] —, "A simple and efficient way to store many messages using neural cliques," *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2011 IEEE Symposium on*, 2011, pp. 1–5.
- [10] —, "Nearly-optimal associative memories based on distributed constant weight codes," *Information Theory and Applications Workshop (ITA), 2012*, pp. 269–273.
- [11] B. K. Aliabadi, C. Berrou, and V. Gripon, "Storing sparse messages in networks of neural cliques," *Proceedings of the national academy of sciences*, vol. 25, no. 5, 2014, pp. 461–482.
- [12] S. A. Shamma, "Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, 1985, pp. 1622–1632.
- [13] H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster "Inhibitory stabilization of the cortical network underlies visual surround suppression," *Neuron*, vol. 62, no. 4, 2009, pp. 578–592.
- [14] G. Palm, "Neural associative memories and sparse coding," *Neural Networks*, vol. 37, 1987, pp. 165–171.
- [15] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, 1998, pp. 1–6.
- [16] A. Aboudib, V. Gripon, and X. Jiang "A study of retrieval algorithms of sparse messages in networks of neural cliques," *arXiv preprint arXiv:1308.4506*, 2013.
- [17] J. R. Cavanaugh, W. Bair, and J. A. Movshon, "Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons," *Journal of neurophysiology*, vol. 88, no. 5, 2002, pp. 2530–2546.
- [18] M. A. Heller and E. Gentaz, "Psychology of touch and blindness," *Psychology Press*, 2013.