Open access • Book Chapter • DOI:10.1007/978-3-319-48506-5_21

A Modular Network Architecture Resolving Memory Interference Through Inhibition

— Source link

Randa Kassab, Randa Kassab, Frédéric Alexandre, Frédéric Alexandre

Institutions: University of Bordeaux, French Institute for Research in Computer Science and Automation

Published on: 12 Nov 2015 - International Joint Conference on Computational Intelligence

Topics: Content-addressable memory

A Modular Network Architecture Resolving Memory Interference through Inhibition

Randa Kassab, Frédéric Alexandre

A Modular Network Architecture Resolving Memory Interference through Inhibition

Randa Kassab and Frédéric Alexandre

Institut des Maladies Neurodégénératives, Université de Bordeaux, CNRS, UMR 5293, Bordeaux, France
LaBRI, Université de Bordeaux, Bordeaux INP, CNRS, UMR 5800, Talence, France
Inria Bordeaux Sud-Ouest, 200 Avenue de la Vieille Tour, 33405 Talence, France
{randa.kassab,frederic.alexandre}@inria.fr

Abstract. In real learning paradigms like pavlovian conditioning, several modes of learning are associated, including generalization from cues and integration of specific cases in context. Associative memories have been shown to be interesting neuronal models to learn quickly specific cases but they are hardly used in realistic applications because of their limited storage capacities resulting in interferences when too many examples are considered. Inspired by biological considerations, we propose a modular model of associative memory including mechanisms to manipulate properly multimodal inputs and to detect and manage interferences. This paper reports experiments that demonstrate the good behavior of the model in a wide series of simulations and discusses its impact both in machine learning and in biological modeling.

Keywords: Associative memory, interference, inhibition, biological systems

1 Introduction

In the domain of machine learning, models of neural networks are classified along their architecture and their mode of learning [6], specifically corresponding to supervised and unsupervised modes. In contrast, in the domain of cognitive science, a natural learning paradigm considered in a realistic behavioral and ecological environment often associates several neuronal architectures and learning modes. This is for example the case with pavlovian conditioning that has been shown to require learning a variety of invariants and to modify the neuronal circuitry in several brain regions including the amygdala, hippocampus and cortex [11]. Consequently, in addition to developing efficient models of neural networks designed for their specific characteristics, there is also a need for a more systemic view of learning, considered at the global cognitive level.

Such an approach was already proposed twenty years ago in [12] arguing that the brain exploits complementary learning systems, with a slow and procedural learning in the cortex, able to extract structures and regularities in the data and to generalize, compared with a quick learning in the hippocampus able to retain the specifics of one's life experiences. This paper, with a very strong impact in both cognitive and machine learning communities, proposes that these systems might be respectively implemented with classical neural models of pattern matching like the multilayer perceptron for the slow learning and models of associative memory for the quick learning.

As an illustration, these models can be contrasted with the property of generalization. Generalization is often reported as a desirable property of artificial neural networks. This phenomenon occurs if, when a network is presented with an example it has never seen before, it is able to interpolate a satisfactory response from the combination of close previously learned examples. Such a response can be judged satisfactory not only because from a limited learning phase the network behaves well in a wider domain but also because in some sense learning went beyond specific cases and was able to extract some general structures or regularities in the example space. In some cases, however, this property might be considered a flaw. This is the case for example when there is no useful topography in the example space or when the goal is to learn some arbitrary association. Consider for example learning to associate a phone number with a name: there is nothing to learn from the euclidean distance between two such numbers and you can in no way discover an association if it was not instructed to you before. This contrasts the cases of learning a general rule from a set of examples, as it is for example studied with layered architectures like the multilayer perceptron, versus learning by heart specific cases like in associative memories.

Neural models of associative memories have been proposed with recurrent networks like the Hopfield model [7] and the Willshaw model [19]. Based on classical connectionist characteristics (like units with non linear activation functions and hebbian learning), the recurrent architecture of these networks indicates that learning is mainly focused on the inner characteristics of an example to be memorized and not on the elaboration of abstract representations in intermediate layers. Nevertheless, some problems can appear if too close examples are learned. In such a case, the network might elaborate an answer from the combination of several learned examples; what would be called generalization in other circumstances is called here interference.

As a consequence, models of associative memories are generally used as content addressable memories, where few prototypes are stored as stable states of the network and noisy or incomplete patterns are presented as inputs and reconstructed to the closest stored example. Beyond this use as an autoassociative memory (where initial input and final result have the same dimension), the adaptation to heteroassociative memory is straightforward: just virtually split the recurrent network in two sets of neurons A and B. The recurrent connectivity includes connections within A and within B (seen as two autoassociative memories) and between A and B (heteroassociative memory between the two sets of different dimension A and B). As configurations of A+B are learned as prototypes, proposing an incomplete pattern A (B neurons being set to 0) will result in the reconstruction of A+B, yielding the answer B. The main acknowledged weakness of these models is about their limited capacity of storage and the associated risk of catastrophic interference when this capacity is exceeded or when too close prototypes are stored [4, 9]. The best solution to this problem is to require a sparse coding, which intrinsically also limits the maximum number of stored prototypes. An associated strategy is to orthogonalize the inputs and project their encoding in higher dimensions, which results in larger weight matrices to manipulate [13]. In both cases, this might prevent associative memories from being applied to large scale realistic problems and can accordingly explain why they didn't have the same expansion in the machine learning community than layered networks. It is consequently highly desirable to develop scalable models of associative memory.

In previous work, we have proposed a modular network model of associative memory [8] grounded on biological data [1, 17]. These data report heterogeneities in the hippocampal structure that might support the coexistence of autoassociative and heteroassociative networks in this region. Specifically, the hippocampus is a neuronal structure known to be involved in episodic memory [18], corresponding to the storage of specific episodes including their context and their emotional or motivational significance. For example, the hippocampus is involved in contextual learning of pavlovian conditioning [3], linking neutral stimuli and their context to biologically significant events (reward and punishment). Though primarily oriented toward biological modeling, we have also explained in [8] the interest of such a segregation from an information processing point of view (cf. the concluding section for a summary). In addition, we have also postulated an additional mechanism for the association of autoassociative memories, that might result in a more robust system, particularly more resistant to interference. The goal of this paper is to evaluate more precisely the performances of this mechanism from an information processing point of view.

In the next section, we will present this model together with its formalism based on the associative memory initially proposed by Willshaw [19]. Then we will report the experiments that were conducted to evaluate its resistance to interference and the associated results. We will conclude by explaining the interest of such a mechanism both in neuroscience and in information processing domains.

2 Multiple Associative-Memory Model

The model is made up of two autoassociative networks that are heteroassociatively linked through a layer of intermediate cells (Fig. 1). The goal is to associate two multi-element patterns in such a way that when at least some elements of the first pattern are presented both patterns can be recalled as a whole. In the hippocampus, these two patterns are considered to represent two important dimensions of episodic memories: 1) The perceptual dimension arises from the integration of different kinds of signals coming from the perception of the outer world: exteroception. 2) The emotional dimension reflects the perception of internal cues of different valences related to pain and pleasure: interoception.

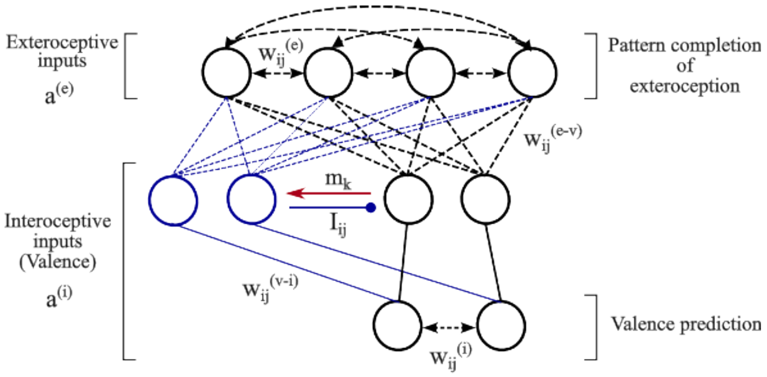

Fig. 1. The architecture of the hippocampal model. Black lines denote the basic circuit of the model while blue lines denote changes in circuitry mediated by one group of associated cells (blue) following the detection of valence-overload interference (red arrow). Autoassociative and heteroassociative connectivities between hippocampal cells are denoted respectively by bidirectional dashed lines and simple dashed lines without arrows. Inhibitory connections between valence cells are denoted by lines ended with circles. Stable non-plastic connections, both excitatory and inhibitory, are denoted by solid lines.

Then, the two autoassociative networks considered in the model receive and store independently two types of input patterns, $a^{(e)}$ and $a^{(i)}$. The layer of intermediate cells is organized into a small number of ordered groups of valence cells that receive valence-related information from the same interoceptive pathways as the interoceptive autoassociative network. The cells in the first group can be directly activated by interoceptive inputs to the model and can therefore be thought of as the primary valence cells. Interoceptive inputs on the cells in the other groups, which are termed associated cells, are conditional, that is, they can not evoke postsynaptic activity within associated cells unless a concomitant signal, m_k, related to the activity pattern of a precedent group is applied.

The valence cells belonging to the same group of intermediate cells are not interconnected. By contrast, inhibitory connections, I_{ij}, exist between cells belonging to different groups. The inhibitory connections are not plastic. They are prewired such that an inhibitory connection from cell i to cell j exists ($I_{ij} = 1$) if the two cells belong respectively to different groups, k and l, and l precedes k ($l < k$). Thus, each group of associated cells, once activated, silences excitable cells in its preceding groups including the primary group of valence cells. This means that at most valence cells in one group can be active at a time.

The formation of extero-interoceptive associations is done at the level of heteroassociative links, $w_{ij}^{(e-v)}$, between the exteroceptive autoassociative network and the groups of intermediate valence cells. These latter provide direct excitatory input to the interoceptive autoassociative network through non-plastic connections, $w_{ij}^{(v-i)}$. These connections are prewired only between valence cells that are sensitive to the same kind of valence.

The classical binary version of the Willshaw network [19] is chosen as the basis for the implementation of both auto- and heteroassociative memory functions in the model. The neurons are simple McCulloch-Pitts binary threshold units and learning begins with all the synaptic weights set to zero. Synaptic plasticity is achieved according to a clipped version of Hebbian learning: a single coincidence of presynaptic and postsynaptic activity changes the synaptic weight w_{ij} from 0 to 1, while further co-activations do not induce further changes. The recall process is done by presenting a cue pattern $\tilde{x}$ and counting the dendritic sum for each cell j ($s_j = \sum_{i=1}^{n} w_{ij}\tilde{x}_i$) in one-time step. The output cells that have a dendritic sum equal to or higher than the number of active inputs are activated. The quality of a recalled pattern can be assessed according to its Hamming distance (HD) from the originally stored pattern (i.e. the number of elements that differ between the two patterns. For example, if x=(0 1 1 1 0) and y=(1 1 0 1 0) then HD(x,y)=2).

Similarly to cholinergic models of the hippocampus [5, 14], our model operates in transition between two modes, storage and recall, depending on a hyperparameter ACh. This mechanism is inspired from biological data describing mode switching under the dynamic regulation of the levels of acetylcholine (ACh) released from septal cholinergic projections to the hippocampus. During recall, a retrieval cue, $a^{(e)}$, is applied to the exteroceptive autoassociative network. The pattern of activity obtained at the output, $\hat{a}^{(e)}$, drives retrieval in the heteroassociative network. An intermediate valence cell, l, can fire only if the dendritic sum of its excitatory inputs exceeds the threshold value and if it does not receive inhibitory inputs from other valence cells that have already fired. The activity of the intermediate valence cells, $\tilde{a}^{(i)}$, triggers recall in the interoceptive autoassociative network yielding the valence prediction by the model, $\hat{a}^{(i)}$.

Just after delivery of the interoceptive information, two novelty-detection processes take place to compare the retrieved patterns to the actual patterns from extero- and interoception. The novelty condition occurs when the Hamming distance between two patterns exceeds pre-specified thresholds ($\mathrm{HD}^{(e)} > e$ or $\mathrm{HD}^{(i)} > v$). Novelty induces ACh dynamics that favor learning of new inputs, otherwise the model settles in recall mode.

During learning, excitatory intrinsic synaptic transmission along the recurrent connections is removed and activity in the model is purely driven by afferent extero- and interoceptive inputs, $a^{(e)}$ and $a^{(i)}$. In the model, two kinds of interference can occur due to a saturation, or overload of learning. The first kind of interference occurs within the autoassociative memories when too many or too close inputs are stored. It is called pattern overload and can be much mitigated using sparse patterns and low memory load conditions. The second kind of interference is called valence overload and is more likely to occur when elements making up the stored patterns become simultaneously associated to different valences. Consider for example learning AB+, AC- and BD-, where A, B, C and D are exteroceptive patterns and + and - are interoceptive valences. Since

A and B are simultaneously associated to + and - valences, the recall of AB would probably generate an interference (both responses produced). The model deals with valence-overload interference by monitoring activity of intermediate valence cells, $y^{(v)}$. If any activity is observed among intermediate valence cells ($\sum_i y_i^{(v)} > 0$) in response to exteroceptive inputs a matching process takes place to determine whether this activity matches interoceptive valence-specific inputs. A mismatch ($\mathrm{HD}^{(v)} > v$) signals a potential interference to a successive group of associated valence cells that become able to respond to valence-related inputs and rapidly silence valence cells that were active in preceding groups.

3 Experiments

The validity of the proposed model is examined through a series of numerical experiments (cf. [8] for the description of other numerical experiments with this model). The simulated model is configured with 150 cells in the exteroceptive autoassociative network and 3 cells in the interoceptive autoassociative network. The intermediate valence cells are organized into 5 groups of 3 cells each.

Inputs are provided to the model as two independent patterns of activity. The exteroceptive inputs are generated as random 150-element binary patterns with 6 elements being active (set to 1). The interoceptive inputs are modeled by 3 binary cells to differentiate positive, negative and neutral valence states. One of these cells switches to its active state according to whether a pleasant (100), unpleasant (010), or neutral (001) stimulus is present.

The performance is evaluated by comparing the output patterns recalled by the model against the original representation of the input patterns that were presented to the model as new information to be stored. Specifically, two kinds of recall errors are considered when evaluating simulation results. Pattern completion errors which reflect the Hamming distance between the learned and retrieved activation for exteroceptive patterns, and valence prediction errors which reflect the Hamming distance between the correct and predicted valence. In both cases, errors are scored when Hamming distance is greater than zero.

Two types of simulations are set out to test the model for its ability to rapidly link exteroceptive patterns and their emotional valences while avoiding valence overload interference. The first set of simulations examines the effect of the number of stored patterns on the accuracy of valence prediction. The model is tested under full-cue and partial-cue recall conditions. The number of stored patterns is kept low enough that under full-cue conditions almost no pattern overload occurs at the level of autoassociative memories. This is important to ensure that any prediction errors might be detected arise directly from valence overload at the level of heteroassociative links between exteroceptive and interoceptive patterns. The second set of simulations focuses on how to quantify the ability of associated valence units to orthogonalize conflicting associations arising from a change in previously learned valence values.

In all of the simulations, the performance of the proposed model, also called the full model, is compared with that of a reduced model with the groups of associated cells removed. A third model with a single autoassociative memory in which both exteroceptive and interoceptive information are merged into a single pattern is also considered to further delineate benefits of the proposed architecture under partial-cue conditions. All results are averaged over 10 simulation runs and are displayed throughout the figures as mean ± standard error of the mean. The novelty-detection thresholds, e and v, are set to zero for all the simulations.

4 Results

4.1 Storage capacity

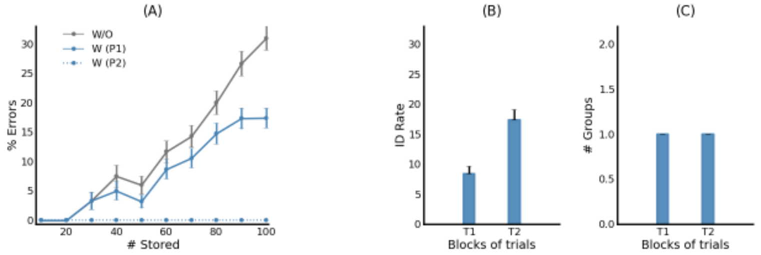

Fig. 2. Influence of the number of stored patterns on the accuracy of valence prediction. (A) Percentage of prediction errors of the model without associated cells (W/O) and with associated cells after one block (W (P1)) and two blocks (W (P2)) of training trials. (B) Rates of interference detection during the first (P1) and second (P2) training trials. (C) Number of groups of associated cells needed to resolve interference detected during training trials T1 and T2.

The first set of simulations is run by varying the number of training patterns and observing how valence prediction is affected with and without the groups of associated cells included in the model (Fig. 2). Training patterns are presented randomly into blocks of N trials with N varying from 10 to 100 in steps of 10. At the different values of N, the full and reduced models were able to recognize exteroceptive patterns with pattern completion errors less than 0.3%. However, there was a noticeable difference between the two models in terms of valence prediction.

As illustrated in Fig. 2A, following the first presentation of training patterns, both models perform perfectly up to N=20, after which point valence prediction errors begin to occur more frequently with increasing size of the blocks of training trials. But as expected, adding the associated cells decreases valence prediction errors at each value of N. For instance, at N=100, the percentage of prediction errors is about 32% for the reduced model but falls to about 20% for the full model. This reduction results from the identification of about 7% of the stored associations as interfering associations (Fig. 2B). Interference effect is accordingly reduced through the recruitment of one group of associated cells (Fig. 2C). During the second presentation of training patterns, the full model detects all the interfering associations that remain and orthogonalizes them using the same group of associated cells (Fig. 2C). Therefore, the performance of valence prediction differs significantly between the two models after the second presentation of training patterns: the reduced model continues to commit the same prediction errors while the proposed model performs with no errors at all.

4.2 Pattern completion

In the above simulations, it is pertinent to emphasize that no differences were observed between the autoassociative model and the heteroassociative model with the groups of associated cells removed. This is of no surprise because in both cases valence prediction is initiated by a complete set of exteroceptive cues. Under partial-cue conditions, the proposed model as well as its reduced version are expected to take advantage of the fact that pattern completion of exteroceptive cues is performed prior to valence prediction. To test this premise, the three models are trained in the same manner as in the previous simulations except that recall is triggered by partial versions of the original trained patterns. Specifically, the block size is set to 100 training patterns and the model is cued with partial versions with either 1, 2 or 3 of the 6 active inputs turned off.

As shown in Fig. 3A and B, all the models perform similarly and reasonably well in terms of pattern completion of exteroceptive cues. The accuracy of valence prediction of the autoassociative model is much worse than that of the heteroassociative models and monotonously drops as the number of deleted elements increases (Fig. 3C). On the contrary, the heteroassociative models are much less sensitive to the percentage of deleted elements. The accuracy of valence prediction with the 1/6 partial-cue condition is the same as that obtained with the full-cue condition (Fig. 3D and E). This is because exteroceptive patterns are almost perfectly reconstructed as shown in Fig. 3B. The removal of two or three of the six active cues causes a proportional decrease in the accuracy of pattern completion of exteroceptive patterns. Consequently, the improvement in valence prediction by the proposed model is less pronounced but still highly significant as compared to the reduced model. For all the percentages of removal simulated, the model makes use of one group of associated cells to tackle valence-overload interference (Fig. 3G).

4.3 Discrimination

Here we investigate the functional significance of the groups of associated cells using numerical simulations with reversal learning tasks. The task in the first set of simulations involves two phases. In the first phase the model is presented repeatedly with 50 training patterns [e.g. A+, B-, C (neutral), etc.] over 4 blocks of trials and the percentage of prediction errors made at the beginning of each trial is measured and displayed in Fig. 4A. This is a simple discrimination learning

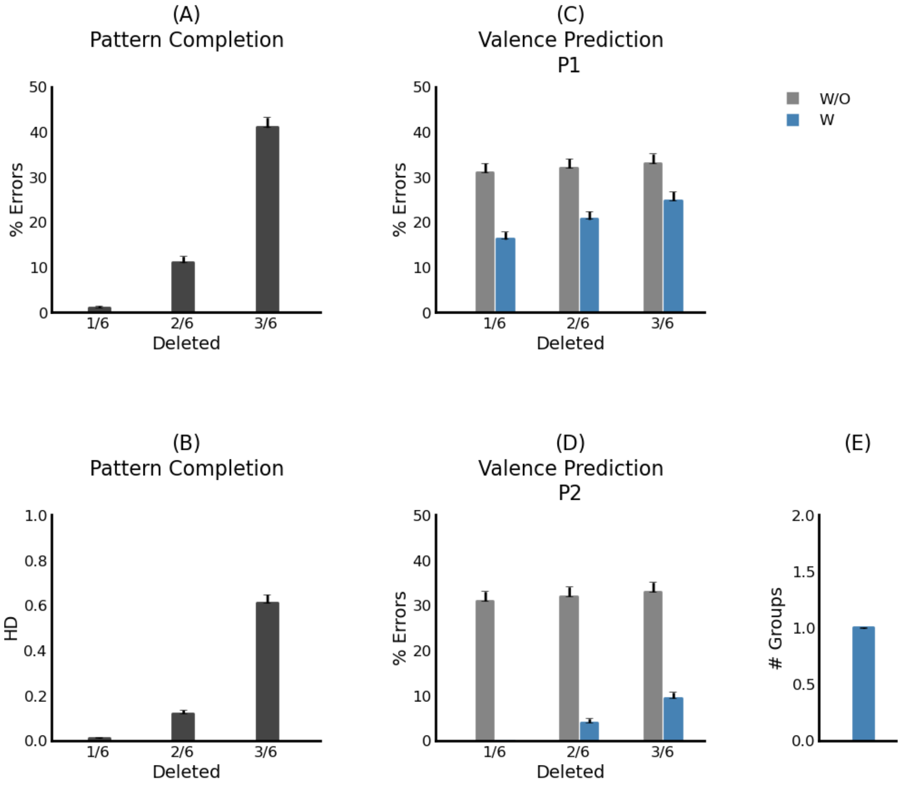

Fig. 3. Performance of the proposed model after training on 100 input patterns. The model is tested using partial cues in which 1, 2, or 3 out of 6 active elements in the original inputs are turned off. (A) Pattern completion performance, defined as the percentage of retrieved patterns that differ at least by one element from the originally stored patterns. (B) Pattern completion performance, defined in terms of Hamming distance between the stored and retrieved patterns. (C, D) Valence prediction performance of the proposed model with (w) and without (w/o) associated cells after one and two blocks of training trials. (E) Maximal number of groups of associated cells needed to resolve interference detected under all simulation conditions (one and two blocks of training trials P1 and P2, and for 1/6, 2/6 and 3/6 partial-cue conditions).

problem similar to those tested in the previous simulations. Thus as was observed before, valence-overload interference occurs at the early stages of learning and exhibits the recruitment of one group of associated cells to tackle it. When the groups of associated cells are removed the reduced model shows impaired performance that persists over the repeated trials. In the second phase, emotional valences of the training patterns are randomly changed to other value with a probability of 50% [e.g. A-, B (neutral), C (neutral), etc.]. As shown in Fig. 4A the proposed model quickly learns to reverse its behavior as all the emotionally changed patterns are detected and learned on the first training trials after reversal. On the other hand, the reduced model fails to acquire the new associations since the old ones have not been unlearned.

4.4 Reversal learning

Further analysis of the model behavior is based on a cue-context reversal learning task similar to that established by [10] to investigate reversal learning in patients with mild amnesic cognitive impairment. To simulate this task, three groups of 4 exteroceptive patterns each are formed such that one of the 6 active elements is used to encode the presence of a sensory cue and the others to encode contextual cues. No overlap is allowed between cells encoding for different cues or contexts (cf. Table 1).

Table 1. The experimental design of the task of [10]. Note. A–H refer to eight cue shapes, 1–8, eight contexts, + and – indicate respectively positive and negative valences.

Training patterns			**Task**	
Group 1	Group 2	Group 3	Phase 1	Phase 2
(original)	(cue reversal)	(context reversal)	(acquisition)	(retention & reversal)
A1+	E1–	A5–	Group1	Group1
B2+	F2–	B6–		Group2
C3–	G3+	C7+		Group3
D4–	H4+	D8+		

In the first phase of acquisition, the model is repeatedly presented with the training patterns in the first group and valence prediction is evaluated over four blocks of training trials. Fig. 5 shows that both full and reduced models make correct valence prediction after a single exposure to the training patterns. Then, the reversal phase is immediately followed by exposing the models to new training patterns from the second and third groups, in addition to the old ones. The training patterns are also presented repeatedly four times in random order. The results show that, in the first block of trials, valence prediction errors are made for both new and old patterns. This reflects the fact that heteroassociative connections are irrelevantly strengthened between the original patterns and valences

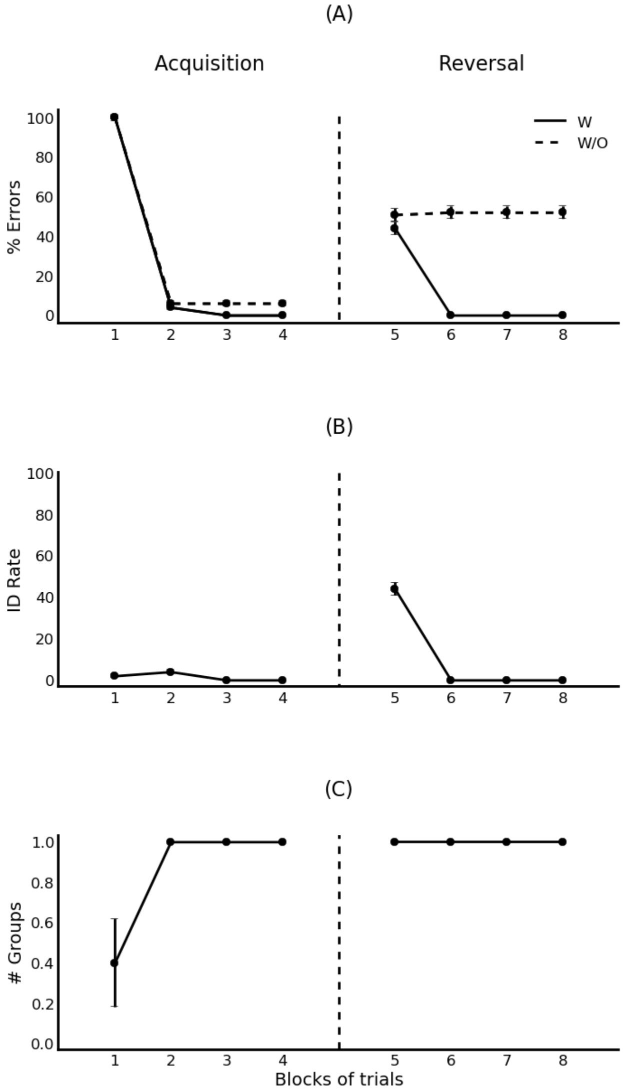

Fig. 4. Discrimination reversal learning. (A) Percentage of prediction errors of the model with (w) and without (w/o) associated cells. (B) Rates of interference detection over each block of trials. (C) Number of groups of associated cells needed to resolve interference across the different blocks of trials.

of new patterns. When interference is detected, one group of valence-associated cells is recruited and prediction errors fall to zero rapidly on the third block of trials after reversal. In contrast, the number of prediction errors the reduced model makes is still the same as the blocks progress for the same reason stated above.

5 Discussion

The primary goal of this paper was to propose a new framework to consider associative memories, particularly to make them efficient even in adverse conditions. Indeed, associative memories have powerful properties for learning by heart specific patterns and recalling them from partial information. They can learn quickly and recall patterns as they were initially presented, without modification nor generalization. It has been shown [12] that such properties are highly desirable for specific classes of cognitive functions and a class of models able to emulate them should consequently be considered carefully. Nevertheless, associative memories are little exploited in classical machine learning because they suffer from limited storage capacities, particularly when patterns to be stored are close, resulting in interferences and catastrophic forgetting [6].

It has also been shown [6] that associative memories can be simply used in auto-association for pattern retrieval but also in heteroassociation between two different classes of inputs. In the model presented here, we exploit this heteroassociative view to propose a modular network. From an information processing point of view, we explain in that paper that a heteroassociation between two data spaces of different size leads to more robust retrieval than a simple autoassociation with a flat vector concatenating both kinds of information because the evaluation of the Hamming distance between stored and actual patterns would consider in this latter case that one error in any dimension yields the same penalty, which is obviously not the case. This is confirmed in the paper, considering comparison of performances between similar autoassociative and heteroassociative models.

A modular view of associative memories is also exploited to implement another powerful property of our model, for managing interferences, using another set of units called associated cells. When an association is learned between a high dimensional data space and a smaller space representing labels (valences in the present case), one central problem is about the association of close patterns with different labels or of different combinations of patterns with different labels. This classical problem has been termed configural learning [2]. With a fully automatic algorithm to insert associated cells between the heteroassociative modules, we have proposed in the present model a mechanism able to detect interference at the heteroassociative level and to trigger new learning accordingly. The experiments reported here, particularly comparing performances of reduced and full heteroassociative models, show that our model is very efficient at performing such a learning. In addition, this learning process is very quick, which preserves another important specificity of episodic learning.

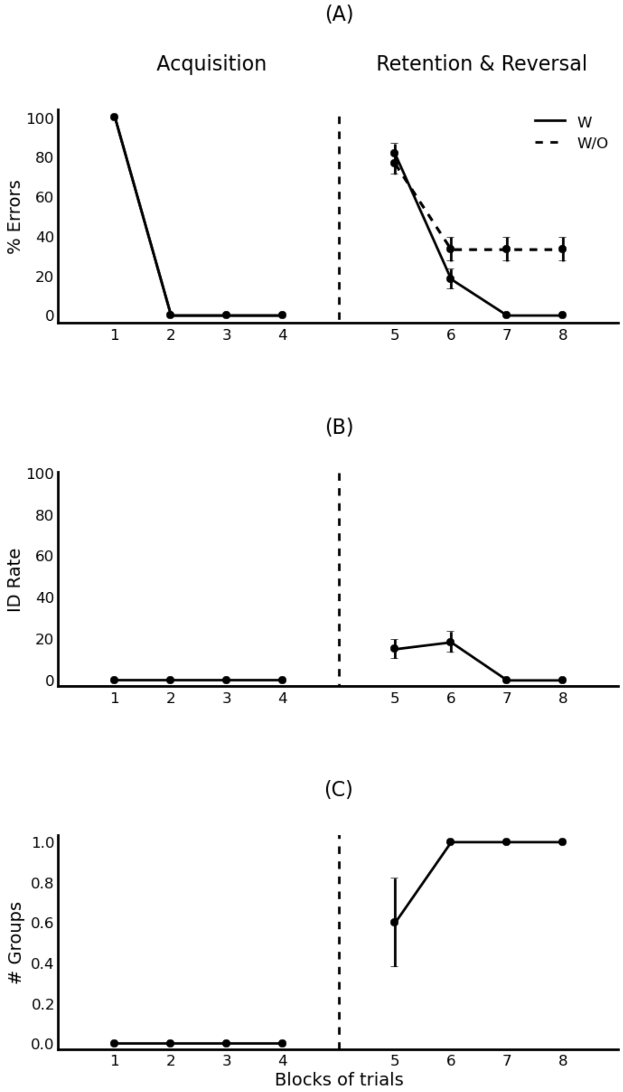

Fig. 5. Cue-context reversal learning. (A) Percentage of prediction errors of the model with (w) and without (w/o) associated cells. (B) Rates of interference detection over each block of trials. (C) Number of groups of associated cells needed to resolve interference across the different blocks of trials.

In this paper, we also propose to relate the very good properties of modular heteroassociative memories to two different frameworks. In the framework of brain modeling, the model has been primarily built as a biologically informed model of the hippocampus [8]. In addition to proposing some evidences for the implementation of a modular network in this cerebral structure, we also propose that heteroassociation could take place between exteroception and interoception, corresponding to different kinds of hippocampl inputs. In further studies, this could be extended to other classes of hippocampal inputs, particularly related to the frontal cortex.

The focus was set here on interoception and exteroception because this study was related to other studies in the team [3] related to pavlovian conditioning. This learning paradigm is very interesting because it is an excellent basis for a systemic view of learning in the brain, with adaptive processing involving (at least) the amygdala, the hippocampus and the cortex [11]. Extending the duality between procedural learning in the cortex and specific cases learning in the hippocampus[12], we explain in [3] that the amygdala is designed to learn pavlovian associations from cues extracted by both structures with their own way of learning and also report, in accordance to other authors [15], a synergy between the three modes of learning, where an event in one learning module (an error of prediction, the occurrence or the storage of a specific case) can trigger or modify learning in another learning module. Considering the importance of such a distributed learning principle, better understanding its details deserves additional work.

Bio-inspiration was also a strong motivation for this work because, in addition to classical evaluation of performances, one of the experiments we made was also designed to reproduce behavioral and cognitive data in the medical domain for amnesic impairments [10]. Related medical data strongly suggest the central role of the hippocampus in this memory process, giving additional interest to the complementary learning system hypothesis [16]. The cognitive framework initiated in [12] postulates how procedural learning in the cortex, slowly learning and able of generalization, might be instructed by specific cases learned quickly in the hippocampus avoiding interferences. Adding emotional aspects with the dissociation between interoceptive and exteroceptive cues, extends this framework of mnemonic synergy in the brain, proposed for medical purposes.

In the framework of machine learning, we have also presented this work as a new model of associative memory and its main results have been described here mainly in the framework of information processing. This is also the reason why we use simple binary units in the hippocampal model, even if more complex functioning rules might be expected in the framework of a biologically inspired model: Even if more complex units might be considered in future works, particularly to fit with more precise biological data, the main goal of the present work was to settle the main computational principles of our modular model. Beyond the case for pavlovian conditioning with interoceptive and exteroceptive cues, we believe that it is not rare in the information processing domain to cope with such associations between data of different dimensions, as it is the case for example with labeled data (high-dimensional data associated with a symbolic label). In this case, we claim that combining autoassociation and heteroassociation as proposed here results in more robustness in the retrieval phase. More generally and beyond associative memories, heteroassociation using intermediate cells to reduce ambiguities is a general class of approaches in machine learning and the criteria proposed here to avoid interference and keep associations simple could be extended to other models of machine learning in further works. This could illustrate other cases where a primary biological inspiration yields efficient learning principles.

References

1. P. Andersen. *The Hippocampus Book*. Oxford Neuroscience Series. Oxford University Press, USA, 2007.
2. Catalin V. Buhusi and Nestor A. Schmajuk. Attention, configuration, and hippocampal function. *Hippocampus*, 6(6):621–642, January 1996.
3. Maxime Carrere and Frederic Alexandre. A pavlovian model of the amygdala and its influence within the medial temporal lobe. *Frontiers in Systems Neuroscience*, 9(41), 2015.
4. B. Graham and D. Willshaw. Capacity and information efficiency of the associative net. *Network: Computation in Neural Systems*, 8(1):35–54, 1997.
5. Michael E. Hasselmo, Bradley P. Wyble, and Gene V. Wallenstein. Encoding and retrieval of episodic memories: Role of cholinergic and gabaergic modulation in the hippocampus. *Hippocampus*, 6(6):693–708, 1996.
6. J. Hertz, A. Krogh, and R. Palmer. *Introduction to the theory of neural computation*. Addison Wesley, 1991.
7. J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences, USA*, pages 2554–2558, 1982.
8. Randa Kassab and Frederic Alexandre. Integration of exteroceptive and interoceptive information within the hippocampus: a computational study. *Frontiers in Systems Neuroscience*, 9(87), 2015.
9. A. Knoblauch, G. Palm, and F. T. Sommer. Memory capacities for synaptic and structural plasticity. *Neural Computation*, 22(2):289–341, 2010.
10. E. Levy-Gigi, O. Kelemen, M. A. Gluck, and S. Kéri. Impaired context reversal learning, but not cue reversal learning, in patients with amnestic mild cognitive impairment. *Neuropsychologia*, 49(12):3320–6, 2011.
11. Stephen Maren. Building and Burying Fear Memories in the Brain. *The Neuroscientist*, 11(1):89–99, February 2005.
12. J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–457, 1995.
13. B.L. McNaughton and L. Nadel. Hebb-marr networks and the neurobiological representation of action in space. In *Neuroscience and Connectionist Theory*, pages 1–63. Hillsdale, NJ: L. Erlbaum, 1990.
14. M. Meeter, J. M. Murre, and L. M. Talamini. Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus*, 14(6):722–41, 2004.
15. Ahmed A. Moustafa, Mark W. Gilbertson, Scott P. Orr, Mohammad M. Herzallah, Richard J. Servatius, and Catherine E. Myers. A model of amygdala–hippocampal–prefrontal interaction in fear conditioning and extinction in animals. *Brain and Cognition*, 81(1):29–43, February 2013.
16. Randall C. O'Reilly, Rajan Bhattacharyya, Michael D. Howard, and Nicholas Ketz. Complementary Learning Systems. *Cognitive Science*, December 2011.
17. T. Samura, M. Hattori, and S. Ishizaki. Sequence disambiguation and pattern completion by cooperation between autoassociative and heteroassociative memories of functionally divided hippocampal CA3. *Neurocomputing*, 71(16–18):3176–183, 2008.

18. Endel Tulving. Episodic and semantic memory. *Organization of Memory. Academic Press.*, 1972.
19. D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222(5197):960–962, 1969.