Life-long learning: consolidation of novel events into dynamic memory representations

Emilia I. Barakova, Tino Lourens and Yoko Yamaguchi

RIKEN-BSI, DEI Lab, 2-1, Hirosawa, Wako-shi, Saitama, 351-0198, Japan,
Honda RI, Japan Co., Ltd, 8-1, Honcho, Wako-shi, Saitama, 351-0114, Japan.

Abstract. Life-long learning paradigm accentuates on the continuity of the on-line process of integrating novel information into the existing representational structures, and recategorization or update of these structures. This paper brings up the hypothesis, that memory consolidation is a biological mechanism that resembles the features of life-long learning paradigm. A global model for memory consolidation is proposed on a functional level, after reviewing the empirical studies on the hippocampal formation and neocortex. Instead of considering memory as storage, the proposed model reconsiders the memory process as recategorization. Distinct experiences that share a common element can be consolidated in the memory in a way such that they are substrata for a new solution. The model is applied to an autobiographical robot.

1 Motivation

Life-long learning paradigm accentuates on the continuity and recategorisation in the learning process. The learned novel fact, event or skill is not independent from the background of the learner. In contrast, the knowledge obtained from previous tasks alleviates acquiring of new information by constraining the possible hypothesis space for those tasks. Furthermore, it is a basis for emergence of novel solutions derived by translating the experience from different domains. The process by which the novelty finds its place in the complete system of previously learned structures, updates these structures, or in other words causes reconceptualization.

While living organisms naturally store and use creatively information, experienced even once in a life-span, it is extremely hard to simulate similar phenomena on artificial agents[1][20]. The major difficulty in simulating life-long learning systems stems from the complexity of the integration and categorization of continuously changing sensory and motor information within the existing representational structures and concepts. The newly acquired information in its turn changes the representations and causes reconceptualization. In addition, all the artificial learning systems suffer from catastrophic forgetting, a process by which the new knowledge dominates over or cancels the previously learned ones.

The neurophysiological mechanism, that resembles the main features of the long-life learning, the continuity, novelty integration and reconceptualization of the existing representational structures is the memory consolidation process. In this paper the consolidation process between the new coming sensory-motor signals and the forming representations and categories is modelled.

In neuroscience, the memory consolidation is understood as a process where declarative memories are gradually encoded into the neocortex. A wide array of neuropsychological, neuroanatomical, imaging, and neurophysiological data define the system crucial for declarative long-term memory and memory consolidation as: the hippocampal formation as one-shot memory encoder, and the neighboring cortical areas in the ventromedial temporal lobe as consolidating and integrating part and the associative neocortex as a long-term storage part, e.g. [2][9][16].

Miller [14] proposes a theory describing how the interaction of the hippocampus, the theta rhythm, and the isocortex allow the brain to represent contexts. He suggests that theta rhythm mediates and synchronizes this process.

In the recent work of Yamaguchi[22], that models the neural dynamics of the hippocampal network, is concluded, that theta phase precession provides an efficient on-line mechanism for memory storage of novel temporal sequences. This mechanism is used as starting point.

Memory consolidation is understood as a two-fold process. Initially, the integration of novel information into the existing representational structures or concepts is performed. On its turn, this integration updates the previously formed representations. Modelling this process is based on novelty encoding. Detailed modelling of the brain mechanisms ensuring the consolidation is not aimed, but rather adopting the main operational principles that will allow to achieve a life-long learning functionality.

Behaviourally-oriented memory functionality, as a part of a life-long learning system, is illustrated in a robotics framework. The ultimate goal is building an autobiographical robot. The current achievement is in learning distinct experiences, that share common elements (overlapping concepts). The test has to find out whether these experiences have been integrated into the memory in a way that enables emergence of solutions for novel problems. The proposed model for concept formation resembles the neocortical functionality and its interplay with the hippocampal formation in the following way: Reactivated concepts initiate a sequence of meaningful events, as previously formed and stored in another network, modelling the hippocampal functioning. The newly experienced events can change the topology of the concept-forming network. A further aim is to extend the experimental and thus the theoretical research to dynamical and different environments, so to create an autobiographical robot. As an autobiographical robot is understood a robot which can recollect its past and use it in an creative way.

2 Consolidation and complementary memory systems

There are two established views of what memory consolidation implies. The first view considers the consolidation as initiation of a cascade of cellular and molecular events by a distinctive experience, that forms a durable form of synaptic alteration [5]. The second view equalizes the consolidation to a post-processing of memory traces, during which the traces may be reactivated, analysed, and gradually incorporated into long-term event memory and general knowledge base [12]. We do not think that the two views are contradictory. The second view explains the recall dynamics of the stored experiences, while the first one describes the concepts of formation. Naturally, we will start with second one, since it has more direct relation to solving the life-long learning problem. Going to the detail of the memorization, the first view is naturally exploited.

The view of consolidation as a processes that alter the newly stored and still labile information to the more stable long term storage relies on the functionality of the cortico - hippocampal memory system. The episodic memory consolidation involves a recording process that transfers the episodic memory trace of an event from the hippocampus to the neocortex [11][12][16][14].

The brain areas, involved in memory consolidation include: the hippocampal formation (HF), the medial frontal areas (MTL), and the neocortical association areas (NAA). We will restrict our analysis to HF and NAA (Figure 1).

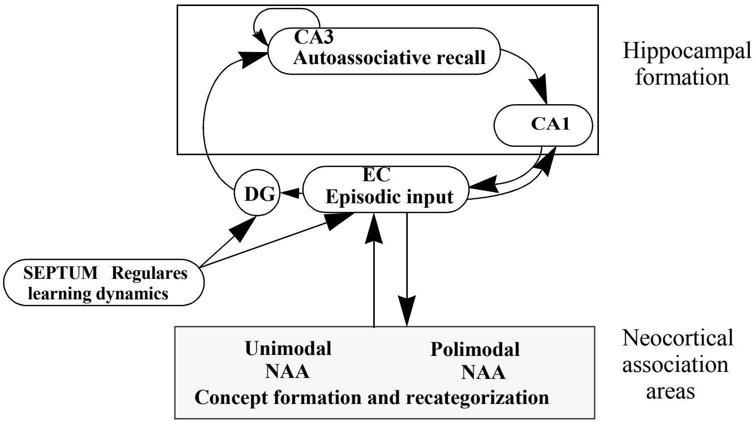
CA3
Autoassociative recall
Hippocampal formation
CA1
DG
EC
Episodic input
SEPTUM Regulares learning dynamics
Unimodal NAA
Polimodal NAA
Concept formation and recategorization
Neocortical association areas

Figure 1 Memory system involved in the consolidation process

There is a broad consensus that the neocortex is the final storage site for declarative long-term memories [7]. The representations are formed on the similarity principle. The use of overlapping representations allows the cortex to represent the shared structure of events, it therefore enables generalisation.

The hippocampus is associated with episodic memory encoding. In addition it directs the consolidation process by reactivation of the stored patterns in some cortical areas and allowing synapses in these areas to change. The functionality of the hippocampus by the memory formation and consolidation processes is mediated by the theta rhythm.

The theta rhythm is a major brain oscillation in the range of 4-10Hz. As a rhythmic activity it originates at the septum[18]. The theta bursts are sent from the medial septum and the diagonal band of the septum to all parts of the hippocampus [14]. The hippocampus then relays these rhythmic firing bursts to the cortex.

The studies on anterograde amnesia strongly suggest that theta activity of the hippocampus may be related to the encoding and/or retrieval of new information. Positive evidence came from studies which have documented that there is a preference for long-term potentiation (LTP) to occur in the hippocampal formation, and that theta activity induces or at least enhances LTP [4]. The fact that LTP is considered the most important electrophysiological correlate for encoding new information, underlines the potential importance of hippocampal theta for memory processes.

In trying to explain the possible functional significance of the theta rhythm in humans, we assume that synchronized bursts of a small set of hippocampal pyramidal cells induce theta activity in selected but distributed cortical regions which are relevant for performing a particular task.

There is a general consensus that memory recall involves reactivation of cortical patterns that characterize the original episodes. This reactivation can be instantiated by partial or noisy patterns. Kali and Dayan [8] define the computational role of memory system as cortical pattern completion. This view is a basis for their model of memory consolidation, performed by a Hopfield - type associative memory network. Obviously, their model of the consolidation process is restricted to the first view for consolidation, as listed in the beginning of this section. In general consensus with [8]our understanding of memory consolidation extends it in the following way:

- The central aspect of consolidation is an achievement of consensus between the two memory subsystems with different characteristics - the fast, on-line memory encoding by the hippocampus and the slowly changing representations of concepts in the neocortex.
- Memory consolidation is a twofold process of integration of novel information into the existing representational structures and updating the previously formed representations.
- The novel information is encoded by the theta rhythm. The so encoded information reaches the associative neocortex areas, where either update an older concept/representation or can form a new one.
- Once reactivated, the episode initiates a sequence of related events.

3 Computational Model

As a physiological process, memory consolidation allows for memories to become permanently encoded in the brain. Computationally it has to resolve the conflict between the two types of representations - hippocampal and neocortical. The hippocampus is specialized in on-line memorizing of specific events, while the neocortex learns slowly the regularities and concepts. The hippocampus assigns distinct representations to stimuli, thereby allowing rapid learning without suffering catastrophic forgetting. In contrast, the neocortex assigns similar representations to similar stimuli and uses overlapping representations. Therefore it can represent the shared structure of events and generalizes to previously experienced stimuli.

These arguments clarify that computationally the consolidation model has to deal with the trade-off of generalization and novelty encoding.

A biologically plausible model for hippocampal novelty encoding in one-shot learning is developed in [22]. According to it, the novel experience is encoded on the phase-locking principle by the theta rhythm and stored in the connections of the Hebbian network.

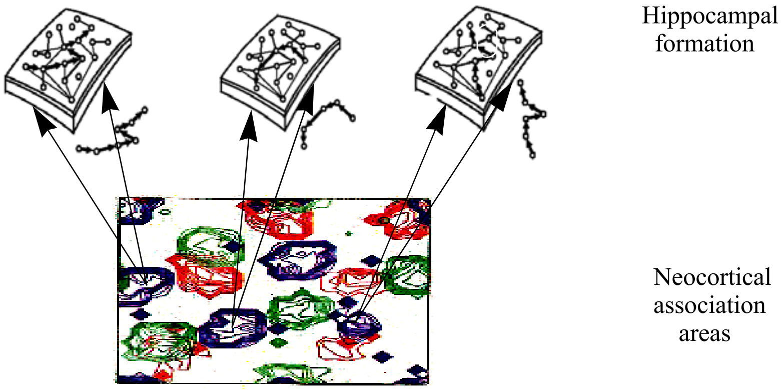

Figure 2 Schematic representation of computations in HF and NAA

The input information, characterizing a distinctive scenery from the environment is represented with a vector $I(t) = \{I_i(t)\}$. Its physical meaning in terms of the hippocampal model of Yamaguchi[22] is, that in an initially stable system of the hippocampal network is brought to oscillatory behaviour by introducing the positive valued input pattern $I(t)$ .Since the enthorinal cortex (EC) provides the main cortical input to the CA1 and CA3 hippocampal areas, this layer is considered as an anatomical counterpart of the neural layer, where the inputs are initially supplied to. In EC layer, the neural oscillator that models the functionality of the $i-th$ EC unit is presented as:

$$\dot{\phi}_i(t) = \omega_i(t) + \{\beta_o - I_i(t) - \cos\phi_0(t)\}\sin\phi_i \tag{1}$$

where $\omega_i(t)$ represents the native frequency and β_0 is a stabilization coefficient. Theta rhythm is modelled as an additional unit of the EC layer. Its oscillatory behaviour is described as:

$$\dot{\phi}_0(t) = \omega_0 . \tag{2}$$

A mechanism, that alleviates spatial coding of the information is called phase precession [15] and constitutes in gradual change in the native frequency $\omega_i(t)$. Because the firing phase advances with the successive cycles, the spatial displacement of any two place fields or in general memory events is reflected in the firing order of the cells within the theta cycle. Thus, the temporal firing order is expressed in a compressed form during any given theta cycle. Functionally, this serve for encoding of event sequences by making the neural activity pattern change rapidly within the effective time window of LTP (about a theta cycle). Without compression, the patterns, that become associated over this short time of retrieval of sequences will suffer interference [17].

The phase-modulated input patterns are further fed into a categorization network. The categorization network is accomplished by a Growing neural gas algorithm[6]. It combines competitive Hebbian learning [10] with a growing cell structure. The purpose of the self-organizing algorithm is dimensionality reduction for the input vector. In computational terms, the low-dimensional representation is equivalent of the concept formation. The competitive Hebbian learning adds to this the preservation of the topological structure of the input data [10]. In addition, the model is incremental. It starts with a small structure and grows until all the input space is represented, i.e. until all the concepts are represented. The method uses local computations therefore, it is prone to catastrophic forgetting. It adds new units any time a new concept is found.

Until now the two parts of the model, the one that approximates the hippocampal-like one-shot learning, and the one, that represents the concept formation as it is believed to happen in neocortex were described. The consolidation process, that features the interplay between the two structures takes place as schematically shown in Figure 2. The upper layer represents a number of stored sequences, found during different learning trials and encoded and stored by the hippocampal network. The lower level represents the concept formation. Some of the novel concepts, shown in red, are not distinctively formed yet. If a pattern is catheterized as belonging to a certain concept, that is close to a wanted solution, the categorization cause recall of a stored Hebbian hippocampal network, and initiates a sequence of related events.

4 Experimental evidence

There is not a well established definition for an autobiographical robot, although there are few researchers that have been using this term. Intuitively, we define as an autobiographical robot to possess authobiographical memory. The definition will be of practical significance if it makes explicit the borders of functionality of an autobiographical robot. So we clarify, that autobiographical robot is able to:

- Recall old episodes.
- The recalled episodes provoke a sequence of related memories.
- The old experiences can be used in a creative way i.e. as an emergent novel solution.

The first step of creating an autobiographical robot is a gathering of relevant experiences trough a single trial learning. The useful (according to a certain criterion) sequences of experience acquired during the robot life span are stored as the fixed Hebbian structures, that can reproduce the experienced event sequence. This is done in an agreement with the new findings, that by the recall process both the hippocampal and neocortical representations are reactivated[17].

According to the proposed cortico-hippocampal model, the novel perception/event) is first categorized as belonging to a particular existing category, or a new category is created. If catheterization to an existing cluster is made, this experience triggers a sequence of events, that have been following the same experience when it has happened in the past.

To be more specific, a new visual percept (the novel seen event or scenery) is categorized in the cortical structure. This novel event triggers a series of ideothetic information that has been experienced in the past, and represents an successful strategy for accomplishing a wished goal.

Figure 3 illustrates the concept formation maps in two different stages of life-long learning process. With the time of learning some close concepts has unified, and the others have diverged and separated. This shows the recategorization feature of our model, an important element of a life-long learning system. For simplicity the illustration is made for ideothetic stream of information.

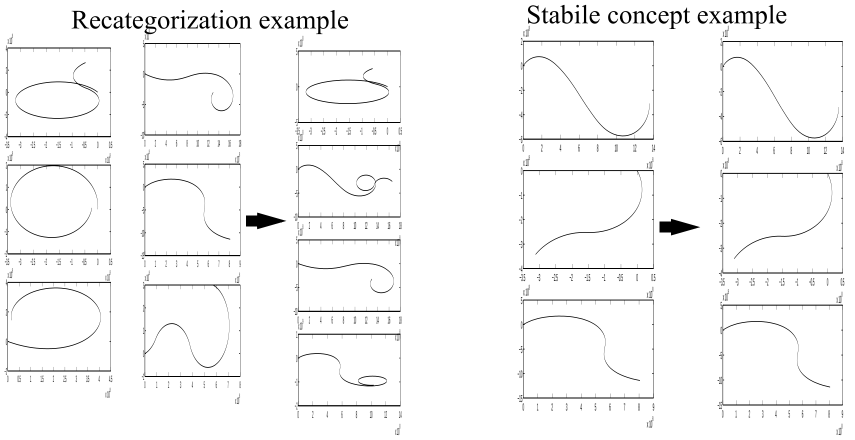

Figure 3 Two stages of concept formation by life-long learning

5 Discussion

Psychological studies[3][19] show that humans integrate and interpret new experiences on the basis of previous ones. Previous experiences are reconstructed with the actual body and context as the point of reference Conway[3]. This way the previously remembered event is recalled trough the prism of the present-day knowledge, understanding and mood. With this respect memory is reconceptualization process as opposed to a storage. The developed model clearly shows this property, and therefore is a good basis for developing of a life-long learning system.

The division of the memory structures to fast learning for novelty encoding and slow changing for representing the knowledge on conceptual level is a powerful mechanism to cope with the trade-off between generalization and inference.

The advantages of categorization of theta encoded signals will be shown in follow-up paper.

The proposed model, goes to a different degree of detail in the functionality of the two brain regions, mostly associated with consolidation process. The near future developments include creating an on-line version of the hippocampal encoding network and dynamical environment framework.

6 References

[1] Barakova E.I. and U. R. Zimmer, Dynamical Situation and Trajectory Discrimination of Raw Range Measurements, *AISTA*, Canberra, Feb. 2000.

[2] Cohen NJ and Eichenbaum H, *Memory, amnesia and the hippocampal system.* Cambridge, MA: MIT Press, 1995.

[3] Conway, M. A., & Bekerian, D. A. (1987). Organization in autobiographical memory. *Memory & Cognition*, **15,** 119~132.

[4] Pavlides C, Greenstein YJ, Grudman M, Winson J., Long-term potentiation in the dentate gyrus is induced preferentially on the positive phase of theta-rhythm. *Brain Res* 1988, **439**:383-387.

[5] Izquierdo I, Medina JH, Memory formation: the sequence of biochemical events in the hippocampus and its connection to activity in other brain structures. *Neurobiol. Learning and Memory* 1997, **68**:285-316.

[6] Frizke, B. (1995). A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems 7*. MIT Press, Cambridge MA.

[7] Fuster, J.M. *Memory in the Cerebral Cortex*. MIT Press, 1994.

[8] Kali S, Dayan P. 2000. Hippocampally-dependent consolidation in a hierarchical model of neocortex. *NIPS* 2000:24~30.

[9] Lavenex P. and David G. Amaral, Hippocampal-Neocortical Interaction: A Hierarchy of Associativity, *Hippocampus* 10:42-430 (2000).

[10] Martinetz, T., & Schulten, K. (1991). A neural-gas network learns topologies. In T. Kohonen at al. (Eds.), (pp. 397~402). Proc. of ICANN, Amsterdam, Netherlands.

[11] McClelland JL, McNaughton, B.L., O'Reilly, R.C. Why there are complementary learning systems in the hippocampus and neocortex. *Psychol. Rev.* 1995, 102, 419-457

[12] Marr D: Simple memory: a theory of archicortex. *Philos Trans R Soc Lond B* 1971, **262**:23-81.

[13] McClelland JL and Goddard, NH,1997. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6:654-665.

[14] Miller, R. Cortico-hippocampal interplay:Self-Organized, Phase-locked loops for Indexing Memory, Psychobiology, Vol. 17(2), 115-128, 1989.

[15] O'Keefe, J. and Recce M., 1993, Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*, 3:317-330.

[16] Squire, L. R. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99:195-231, 1992.

[17] Sutherland, G.R. and McNaughton, B. (2000) Memory trace reactivation in the hippocampal and neocortical neuronal ensembles. *Curr. Opin. Neurobiol.* 10, 180~186

[18] Swanson, L.V., Kohler, C. and Bjorklund A. The limbic region: I:The septohippocampal system. In A. Bjorklund, T. Hokfeld, and L.W. Swanson, edt. Handbook of chemical neuroanathomy, Vol. 5, pages 125-277. Elsevier, 1987.

[19] Rubin, D. C. (Ed.). (1996). Remembering our past: Studies in autobiographical memory. New York: Cambridge University Press.

[20] Thrun, S.B., and T. M. Mitchell. Integrating inductive neural network learning and explanation-based learning. In *Proceedings of IJCAI-93*, Chamberry, France, 1993.

[21] Wiener, S.I., V. Korshunov, R.Garcia, and A . Berthoz, 1995. Internal, sibstratial and landmark que control of hippocampal CA1 place cell activity. *European Journal of Neuroscience*, 7:2206-2219, 1995

[22] Yamaguchi, Y., 2003.A Theory of Hippocampal Memory Basedon Theta Phase Precession *Biological Cybernetics,* In press.