Modelling Emergent Phenomena in Associative Memory with Neural Cell Assemblies

Kailash Nadh, BSc.

November, 2011

School of Engineering and Information Sciences
Middlesex University

Submitted in partial fulfilment of the requirements
of Middlesex University for the degree of Doctor of Philosophy

Acknowledgements

I thank my supervisors, Dr. Chris Huyck, Dr. Roman Belavkin, and Dr. Peter Passmore, a very versatile group, for the exploratory freedom they granted me with their slightly atypical and excellent approach to supervision, and the valuable feedback and insights they have offered throughout. I would like to mention the regular and long meetings with Dr. Chris that would often end in massive digressions—discussions that he describes as "hand waving"—that have heavily influenced this thesis.

I thank my friends who have been supportive of my work and interests—those who are of the opinion that artificial intelligence will pave way to humankind's doom, and those who are of the opinion that a technological singularity is impending. I also thank my family for their continued support.

I would like to mention EPSRC grant GR/R13975/01 that partially supported this work, and the scholarship provided by Middlesex University.

I dedicate this thesis to my supervisor Dr. Chris Huyck, for introducing me to the PhD programme, for providing excellent guidance, and for being an exceptional mentor.

Abstract

The fundamental mammalian behaviours of perception, recognition, categorisation, generalisation, and many other psychological phenomena are intrinsically bound to the basic cognitive process of memory formation and association. While many mechanisms for creating artificial associative memory exist, Hebbian Cell Assemblies(CAs) offer a neurobiologically plausible means of doing so.

This thesis is an exploratory study of the dynamics of CAs and CA based associative memory. It looks into how complex phenomena emerge from associative memories by modelling and simulating the cognitive processes of formation, association, and retrieval of memories as CAs. From these elemental processes, higher order behaviour are obtained, namely, emergent context sensitivity, spatial cognitive mapping (modelled in an embodied virtual agent), and emergence of novel behaviour (explored via an autonomous game playing agent). The thesis also confirms that CAs are capable of performing real world tasks, with a natural language processing model capable of resolving with better accuracy than many symbolic machine learning models, the prepositional phrase attachment ambiguity, a common syntactic and semantic ambiguity.

The results from the models are novel and manifold, hinting at a unified model of associative memory. They suggest that neurobiologically inspired models—in particular, the CA model—may be better at performing certain AI tasks than other traditional computational models. They also suggest the possibility of a CA based associative memory model that may be able to account for many higher order processes, and demonstrate how CAs can be used to model tasks in AI that resemble processes in the brain.

Contents

List of Tables

List of Figures

List of Publications

1. Nadh, K. (2011). A Neuro-computational Approach to Prepositional Phrase Attachment Ambiguity Resolution. Unpublished manuscript.
2. Huyck, C. R., Belavkin, R., Jamshed, F., Nadh, K., Passmore, P., Byrne, E., and Diaper, D. CABot3: A simulated neural games agent. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, Barcelona, Spain.
3. Nadh, K. (2010). Learnt novelty with overlapping cell assemblies. In *Proceedings of the York Doctoral Symposium on Computing*, number 3, pages 47-54, York, UK. University of York.
4. Nadh, K. and Huyck, C. R. (2010). A pong playing agent modelled with massively overlapping cell assemblies. *Neurocomputing*, 73(16-18):2928-2934.
5. Huyck, C. R. and Nadh, K. (2009). Multi-associative memory in fLIF cell assemblies. In *A. Howes, D. Peebles, R. C. E., editor, 9th International Conference on Cognitive Modelling*, pages 81-87, Manchester, UK.
6. Nadh, K. and Huyck, C. R. (2009). Prepositional phrase attachment ambiguity resolution using semantic hierarchies. In *The Ninth IASTED International Conference on Artificial Intelligence and Applications*, Innsbruck, Austria.

Chapter 1

Introduction

Trying to comprehend the emergence of intelligence in biological beings is a humbling task. The field of artificial intelligence (AI) is a modern version of the age old human quest for understanding the nature of intelligence, and thus selves. What makes the mystery of intelligence even more confounding is that it emerges from seemingly fundamental processes such as memory formation and association. It is known that many higher cognitive phenomena arise from the interplay of human memories that are encoded as large distributed networks of associations [Haxby et al., 2001; LaBar et al., 1999; Squire, 1992]. For the same reason, understanding and including associative memory mechanisms in artificial models of intelligence is important. This thesis is the result of such an exploration.

Theories from different disciplines give varying accounts for human associative memory. Many such accounts hold varying views for the same set of processes. Symbolism and connectionism are two such common views. Symbolic models postulate that cognition evolves from symbols that can be stored, retrieved, and transformed based on specific sets of rules [Newell, 1980]. On the other hand, connectionist models propose a different view inspired by the nervous system, where cognition is considered to emerge from dynamic networks of interconnected nodes based on biological brain cells or neurons [Feldman and Ballard, 1982]. While symbolic systems facilitate information processing, on their own, they are weak tools for understanding intelligence, as they incorrectly assume that humans are mere symbol processors. Humans perceive and manipulate symbols, but the underlying processes are complex and not well understood. If humans were mere symbol processors, for instance, the word *apple* would encompass no abstract "meaning" and would simply represent a symbol. Here, a symbol is considered to be a discrete unit that can be used to denote a set of objects. While symbolic systems are useful tools for studying high level cognitive behaviour, they do not necessarily provide insights into the nature of biological intelligence [Bechtel and Abrahamsen, 2002; Smolensky, 1987].

Connectionist models are large parallel networks of nodes inspired by biological networks of neurons in the brain. Such systems can perform a wide variety of tasks, from financial predictions [Fadlalla and Lin, 2001] to analyses of protein folding mechanisms [Rost, 2001]. In connectionist models, information processing is done by the interaction of networks of individual nodes resembling biological networks of neurons. Unlike symbolic systems, they allow modelling of cognitive processes with fine and often neurobiologically realistic accounts. Such neurobiologically inspired architectures are also good at performing many tasks in AI [Fransen and Lansner, 1998; Huyck, 2001; Knoblauch et al., 2007; Levy and Horn, 1999; Valiant, 2005; Wennekers, 2007; Wennekers and Palm, 2000; Wickelgren, 1999].

The biological brain is capable of information reuse at very low levels. Information learnt in one context can be automatically used in another. Perceived stimuli such as audio, visual, and tactile information have complex abstract representations in the brain that allow an individual to not merely perceive, but to "understand" them. Such processes emerge from the interaction of a vast interconnected network of biological neurons. Still, many computational connectionist models attempt to keep neurobiological influences to a minimum. For instance, backpropagation networks [Hecht-Nielsen, 1992] for supervised learning are not physiologically possible in biological neurons. It is also known that neurons do not exist in a fully connected network in the brain where every neuron connects to every other neuron, but in distributed circuits [Schüz, 1998], unlike many computational models that rely on fully connected artificial neural networks (ANN)

[Hopfield, 1982; Kohonen, 1982; Kosko, 1988; Willshaw et al., 1969]. While such models may be good for solving computational problems, they may not necessarily aid in understanding characteristics of intelligence.

On the other hand, computational models of Hebbian Cell Assemblies (CA) [Hebb, 1949]—specialised circuits of neurons—provide a neurobiologically and psychologically realistic framework for modelling associative memory, and possibly, intelligence. Hebb [1949] considered CAs to be the neural basis of the fundamental process of associative memory from which higher cognitive phenomena emerge. This thesis explores many interesting properties of CAs such as their formation that accounts for long term memories, reverberative behaviour that accounts for short term memories, synaptic and overlapping associations that create complex associative memories, competitive behaviour that supports categorisation and other higher processes (discussed in Section 3.2) In the process, the work described in this thesis demonstrates that the CA model is capable of performing many different kinds of tasks ranging from machine learning tasks to psychologically plausible cognitive tasks. It also provides some insight into how certain low level neural mechanisms in the brain may underpin higher mental processes. While there are advanced systems—for example, ACT-R, the symbolic cognitive architecture [Anderson, 1998]—that can model complex cognitive phenomena, this thesis focuses on CA based associative memory and investigates how it may give rise to higher order processes.

Associative memory is a fundamental cognitive process. In the brain, memories are not encoded as discrete entities, but in large distributed networks [Anderson and Bower, 1980]. Different ideas, concepts, and associations between them are gradually acquired and integrated into a vast memory system. These concepts and associations are critical to complex mental processes. Like many cognitive processes, associative memory has a strong neural basis.

It is thought that concepts, ideas, and associations between them are encoded in the brain as neural CAs. CAs exhibit dynamics that provide a unified account for long term and short term memories that traditional models compartmentalising the two processes cannot. Associative memory in itself has a wide range of properties. Concepts can have one to one, one to many, and many to many relationships. Concepts can also vary in their nature, for instance, derivative abstract concepts and discrete tokens. Associations can also exist in different forms such as hierarchical and contextual associations. They can also be of different types such as, abstract *is a* and *has a* relationships to higher level semantic and spatial associations. How associations emerge from CAs, and in turn, how certain fundamental higher order cognitive phenomena such as categorisation and generalisation arise, are explored in this thesis.

A series of models that explore different aspects of CA based associative memory, capable of performing a variety of tasks are developed. The objective is to gain insight into the fundamental neural mechanisms underlying associative memory, and to evaluate the viability of neurobiologically faithful computational CAs for modelling complex processes. The exploratory nature of this thesis is briefly described as a preamble in the next section.

1.1 Preamble

An inductive method is adopted in this thesis, where the cognitive processes of memory formation, association, and retrieval in CAs and higher order phenomena emerging from them are explored via computational simulations. While there is a great deal of neurobiological data supporting these processes, their precise dynamics are not well understood. For the same reason, the approach is highly exploratory, somewhat arbitrary, and slightly unconventional.

The principal problem considered in this thesis is the emergence of “intelligent” processes from associative memories. Such processes are numerous, indiscrete, and hence, poorly understood. So, a few such processes—while considerably different from one another—are modelled in a reasonably neurobiologically faithful framework for the purpose of exploring and construing their nature. The models each demonstrate, the emergence of context sensitivity from an associative memory; spatial cognitive mapping with an associative memory; emergence of novel behaviour from generalisation in an associative memory; and natural language disambiguation with a large associative memory of semantic data. Even though these models are unrelated to each other, they all focus on, and are based on increasingly complex CA based associative memories. The first two models associative memories with synaptic associations, while the latter two study overlapping associative memories. These two important types of associative memories on which all the work described in this is based on, are discussed in Section 3.2.

While it is not practical to delve into expansive details of the processes modelled, they are overviewed in some detail, and their models are described in detail. Some of these models have also been published elsewhere [Huyck and Nadh, 2009; Nadh and Huyck, 2010]. The key contributions of the work described in this thesis is briefly summarised in the next section, and is discussed in detail in Chapter 6.

1.2 Contributions

The findings from the models described in this thesis and their contributions to the understanding of the nature of CA based associative memories is briefly summarised below.

Context sensitivity can be implicit in the natural dynamics of an associative memory of CAs. The behaviour of a CA, and the associative memory it is a part of, can vary drastically based on the activity of CAs associated with it. The model, described in Section 5.1, demonstrates a low level neural mechanism underlying context.

The sequential activation of CAs in an associative memory representing various elements in an environment supports wayfinding, a form of spatial cognitive mapping. The model, described in Section 5.2, shows an embodied agent in a virtual 3D world performing a spatial cognitive task with such an associative memory.

Generalisation can implicitly emerge from the inherent characteristics of an associative memory of continuous, overlapping CAs. The model shows a virtual agent learning to play a simple game of *Pong* autonomously or via human supervision. The model is described in Section 5.3.

A large associative memory of overlapping CAs encoding semantic data is able to perform a natural language disambiguation task with an accuracy that is on par with statistical machine learning models that perform the same task. The model makes use of complex semantic associations that are formed in its associative memory to disambiguate sentences from a large English corpus that are affected by the prepositional phrase attachment ambiguity. The model is described in detail in Section 5.4.

The results from these four models are novel and multifold, and they highlight the benefits of the strong neurobiological underpinnings of the computational CA model. More specifically, the findings demonstrate that the CA model is capable of performing a variety of different tasks ranging from language disambiguation to spatial cognitive mapping. Furthermore, the models show how complex phenomena can emerge from the fundamental CA associative memory processes, hinting at the possibility of a unified model of associative memory, and in part, AI.

1.3 Structure

This section briefly summarises the rest of the chapters, outlining the thesis.

Chapter 2 provides background material and important concepts that support the work described in this thesis, namely, human associative memory and its neurobiology; ANNs in general; and some related computational models of associative memory.

Chapter 3 discusses Hebbian CAs in detail. It covers their neurobiology, and various dynamics exhibited by them. The chapter also discusses how CAs provide a unified account for the fundamental processes of short term and long term memory, and how they are central to higher order cognitive phenomena such as perception, recognition, and recollection.

Chapter 4 describes in detail the neural network architecture with which the models described in this thesis have been developed. It discusses the biologically inspired *Fatiguing Leaky Integrate and Fire* (FLIF) neurons that the architecture is based on. The chapter also describes the learning algorithm used in the architecture, and the processes by which CAs emerge in the network.

Chapter 5 describes the models developed and their findings, highlighting different capacities of CAs, namely emergent context sensitivity in CAs; spatial cognitive mapping with CAs embodied in a virtual agent; emergence of novel behaviour in CAs in a self learning game playing agent; and prepositional phrase attachment ambiguity resolution with CAs, a natural language disambiguation task.

Chapter 6 discusses the outcomes and implications of the work described, and briefly discusses prospects for future work.

CHAPTER 2

Background and related work

Many connectionist models, especially ANNs, have at least loose resemblances with human associative memory. As mentioned in the introduction, connectionist models unlike symbolic models, are neurobiologically inspired. Still, disparities are prevalent between different models as discussed in the following sections. This chapter overviews ANNs in general, various concepts of associative memory, and reviews a selected few computational models of associative memory. It provides elementary background information leading to CAs and CA based associative memory that are the focus of this thesis.

2.1 Artificial neural networks

ANNs are mathematical or computational implementations inspired by the biological brain that attempt to perform information processing like the brain itself. Their architectures are generally comprised of networks of independent functional nodes resembling biological neurons. Most ANNs are described by the connection topology of their nodes, the characteristics of individual nodes, and rules that drive "learning" in them. The advent of such models can be traced back to the infludential work of Lapicque [Brunel and van Rossum, 2008] from more than a century ago. He proposed the integrate and fire neuron model which is still extensively used in computational neural models. Later, the influential work by McCulloch and Pitts [1943], followed by Hebb [1949], Minsky [1954], Widrow and Hoff [1960], Rosenblatt [1962], and others paved the way to modern ANNs.

The McCulloch and Pitts model was based on a network of simple thresholded binary switches that received input from other switches. These switches became active if inputs exceeded a critical threshold and remained quiescent otherwise. These networks of switches were primitive abstractions of biological neurons that paved the way to modern ANNs.

Rosenblatt's *perceptron*—a slightly more sophisticated model, and one of the first practical ANNs—learnt from examples and performed simple pattern recognition and classification tasks. These eventually lead to the modern and widely used ANNs such as self organizing maps [Kohonen, 1982], the Hopfield network [Hopfield, 1982], and multilayer perceptrons [Rumelhart et al., 1986].

The model neuron in most modern ANNs is similar to the McCulloch and Pitts neuron, a fundamental thresholded unit capable of receiving and transmitting signals. Such a simple neuron in a network receives inputs from other connected neurons. If the sum of the inputs exceeds the neuron's threshold, it becomes active or "fires" by setting its output to one, and zero otherwise. These two states enable the neurons to perform simple binary classifications of linearly separable problems [Krogh, 2008]. The firing behaviour in such neurons is an abstraction of the firing behaviour of biological neurons [Bevan and Wilson, 1999] by which they transmit electric potential via their outgoing connections. More complex ANNs take into account the properties of biological neurons such as dynamic thresholds and spiking behaviour [Izhikevich, 2003; Maass and Bishop, 2001]. Many such ANNs are based on the Hodgkin-Huxley neuron model, the landmark work of Hodgkin and Huxley [1952]. Their model mathematically describes the chemical processes in a biological neuron that account for its elementary properties. The FLIF ANN used in this thesis (Chapter 4) encompasses many characteristics of biological neurons such as integrate and fire behaviour, leaking of potential and fatiguing.

ANNs have been applied to a variety of computational problems, with most applications concerning their inherent capacity for pattern recognition [Bishop, 1995]. This important feature of ANNs is commonly used to discover implicit underlying patterns in poorly understood data, such as weather forecasting [Kuligowski and Barros, 1998] and financial prediction [Fadlalla and Lin, 2001]. Similarly, ANNs are able to work with noisy and partial information unlike discrete symbolic systems. Such properties of ANNs are what are generally classed as "intelligent". While ANNs are neurobiologically inspired, most models are not faithful implementations of biological systems. There are also large disparities between different models due to the lack of a unified framework and definitive knowledge of the dynamics of biological neural systems [Basheer and Hajmeer, 2000]. While such ANNs are versatile, they are limited in their abilities to perform complex, and ultimately, human like tasks such as complex decision making. Nonetheless, ANNs are becoming increasingly sophisticated, with many models having neurobiologically realistic characteristics being used for modelling cognitive phenomena, for example, complex associative memories and cognitive mapping [Botvinick and Plaut, 2006; Fransen and Lansner, 1998; Garagnani et al., 2009; Samsonovich and Mcnaughton, 1997]. Ultimately, most ANNs are intricate tools for problem solving, and some, the means of exploring the nature of intelligence.

2.2 Associative memory

Human associative memory is a remarkably complex process that underpins many cognitive phenomena. It continues to develop throughout an individual's life, where new concepts are acquired, learnt, associated, recalled, and purged. Such a memory is not a disjoint islands of events, experiences, and images, but a complex network of vast collections of memories. This network is woven together by the continuous modification and organisation of personal experiences and factual knowledge that an individual acquires. Functionally, human associative memory is a *heteroassociative* memory, as it is a large collection of concepts with relationships of varying degrees of complexity, as opposed to an *autoassociative* memory that does not account for associations between different representations.

Any given concept in an associative memory can be associated with many other concepts, and the retrieval and modification of concepts may depend on criteria such as content and context. There may also be a wide range of types of associations between concepts. For instance, associations between factual concepts in the human associative memory may vary by strength of physical association and extent of feature overlaps [Tulving and Markowitsch, 1998]. The ubiquitous involvement of associative memory in many important aspects of conscious intelligent behaviour in humans, from perception and recognition to more complex decision making, may be considered a testament to the critical role it plays in the emergence of intelligence. Hence, it is only reasonable to assume that to create human like AI, computational models should have a strong basis in associative memory.

Associative learning, and thus, the mechanism of learnt association was notably demonstrated by the Russian physiologist Ivan Pavlov [1927] with his famous conditioning experiment, popularly known as the Pavlovian conditioning experiment. In the experiment, Pavlov presented his dogs with a ringing bell, and food shortly thereafter, repeatedly over many days. This repeated co-presentation of food and the ringing bell caused the dogs to associate the two stimuli. As a result, the dogs started salivating at the sound of the ringing bell, even when food was absent. It was proposed that if associative memory can be induced in animals, it is only reasonable to assume that similar processes may reside in the human brain.

Modelling associative memory in artificial neural networks is an important step in simulating intelligence [Pershin and Di Ventra, 2010], but models capable of simulating these processes are not common and usually incomplete. A model inspired by Pavlov's conditioning experiment is described in Chapter 5. It demonstrates how the cognitive process of context sensitivity can implicitly emerge from the associative learning of CAs.

In humans, associative memory is considered to have two major components, procedural memory and declarative memory [Tulving, 1985]. While these two systems overlap considerably, the classification is based on their observed nature. There is a large body of evidence that supports this division [Cavaco et al., 2004; Cohen et al., 1997; Molinari et al., 1997; Ullman, 2004; Ullman et al., 1997]. Even though these are independent systems, they interact with each other in carrying out various tasks. Procedural memory or *how to* memory is composed of implicit memories that do not require conscious recollection. For instance, the acquired skill of riding a bicycle or a similar skill learnt by repetitive practice manifests the action in realtime without conscious recollection. Such memories are non-propositional by nature [Tulving, 1985]. It is usually referred to as *implicit memory* for the same reason, as the learning process and the learnt knowledge are mostly unconscious. Aspects of Pavlovian conditioning, mentioned earlier, may be a part of procedural memory [Poldrack and Packard, 2003]. On the other hand, declarative memory concerns learning, organising, and using facts and events. The brain regions corresponding to declarative memory are well studied, and it is better understood than procedural memory [Eichenbaum, 2000; Squire, 1992].

A real world event experienced and learnt by an individual has many details such as spatial, audio, visual, tactile, and linguistic information. The recollection of such an event from memory involves pulling together a number of details from different regions of the brain concerning their perceptions, and compiling into a coherent representation that can be consciously interpreted [Cohen et al., 1997]. This is facilitated by the interaction between different brain regions. This highly distributed and interconnected nature of memories that encompasses various relationships between encoded concepts make them highly relational in nature [Cohen et al., 1997]. Activation of certain memories in such a network may cause activation of other related memories based on many given factors such as context. This dynamic nature makes the memory system highly flexible, scalable, and capable of rapidly adapting to novel scenarios. CAs encompass such dynamics by nature, and hence, are considered to be the neural basis of associative memory. The neurobiology of CAs and their various properties are discussed in detail in Chapter 3.

The two subsystems of declarative memory—episodic memory and semantic memory–are attributed to episodic and factual knowledge respectively [Tulving and Donaldson, 1972]. Most computational AI systems borrow characteristics from declarative memory, as attempts of modelling intelligence are usually based on systems capable of acquiring and manipulating factual knowledge. The models described in this thesis (Chapter 5) mainly encode semantic memories. Many machine learning systems that do not resemble biological associative memory are also based on similar concepts [Kolen and Pollack, 1991; Stainslaw Jankowski, 1996]. Hence, understanding declarative memory, especially semantic memory is important to modelling associative memory. The following subsections discuss episodic and semantic memories further.

2.2.1 Episodic memory

Episodic memory is the comprehensive autobiographical catalogue of personal experiences and events. Tulving and Thomson [1973] described it as “concerned with storage and retrieval of temporally dated, spatially located, and personally experienced events or episodes, and temporal-spatial relations among such events”. It is highly personal and intrinsically bound to an individual’s “self” and its perception of time. While episodic memory and semantic memory interact in complex ways, items in episodic memory are not well remembered compared to the factual knowledge in semantic memory. This is because stimuli processed semantically are better remembered than those processed merely perceptually [Kapur et al., 1994]. This may be why episodic memories are often not as intricate as semantic memories and are prone to interference [Tulving and Donaldson, 1972].

Episodic memories, being personal in nature, also have strong emotional bindings. Formation and consolidation of such memories are affected by emotional states, and recalling of episodic memories may evoke emotions. For instance, Brown and Kulik's [1977] so called *flashbulb memories* are unusually vivid circumstantial memories—detailed snapshots of moments of time—whose formation is triggered by heightened levels of emotion and states of consequentiality. A canonical example cited by many is the vivid recollection of the moment of learning of the assassination of President Kennedy. While Brown and Kulik [1977] proposed a specialised mechanism responsible for creating such vivid memories, it is known that such memories are a characteristic of episodic memory [Davidson and Glisky, 2002; McCloskey et al., 1988]. It is such characteristics that differentiate episodic memory from semantic memory which is more general purpose in comparison. Due to this nature of episodic memory, in AI, it is seldom applied to general tasks and often limited to computational models of embodied intelligence [Dodd and Gutierrez, 2005; Ramamaurthy et al., 2004; Rickel and Johnson, 2000].

2.2.2 Semantic memory

Semantic memory encodes information pertaining to facts, concepts, meanings, and the relationships between them. It is a vast encyclopaedia of conceptual knowledge whose constituents range from abstract concepts to symbolic information such as names of people and places. It facilitates the rapid acquisition, organisation, and retrieval of such information in real time. The role of semantic memory in human intelligence is ubiquitous, as such factual information is extensively used in higher cognitive process such as decision making and language. It is critical to intelligence and manifests the fundamentals of an individual's interaction with the environment. For instance, subjects with damage to the brain regions associated with semantic memory are unable to do basic tasks such as object recognition and naming, eventually leading to severe problems in language comprehension [Hodges et al., 1992].

There is evidence that different categories of perceived objects elicit similar activity in the brain regions that correspond to the perception of those objects. These regions have also been observed to be consistent across different subjects. Objects with specific feature sets such as man made tools associated with movement in space may have their features encoded in different regions that concern visual, linguistic, and motor movements [Le Clec'H et al., 2000]. Such regions have been found to elicit similar activity when subjects do naming tasks, write about the objects, or simply think about them. This suggests that these processes involve retrieval of learnt information of specific attributes and features of different objects that are distributed and co-operative in nature [Kreiman et al., 2000; Martin and Chao, 2001; Vandenberghe et al., 1996].

Such distributed organisation of knowledge resembles semantic networks, a high level knowledge representation scheme. A semantic network is a graph, where concepts are represented as vertices and relationships between them as edges [Quillian, 1967]. This resemblance is interesting, as semantic memory in its simplest form may be considered a semantic network. Figure 2.1 shows an example of a simple semantic network with a small group of interlinked concepts, where the links specify their relationships. What constitutes the encyclopaedic knowledge of the environment in an individual is the vast underlying network of abstract concepts, instances of concepts, categories, and relationships between them—all associated coherently and recollected at will.

Semantic memory is an immensely complex system with a number of modalities. While its high level behaviour is well studied—mostly through visual and semantic recognition, association, and recollection tasks [Federmeier and Kutas, 2001; Moss et al., 1995; Warburton et al., 1996]—its low level characteris-

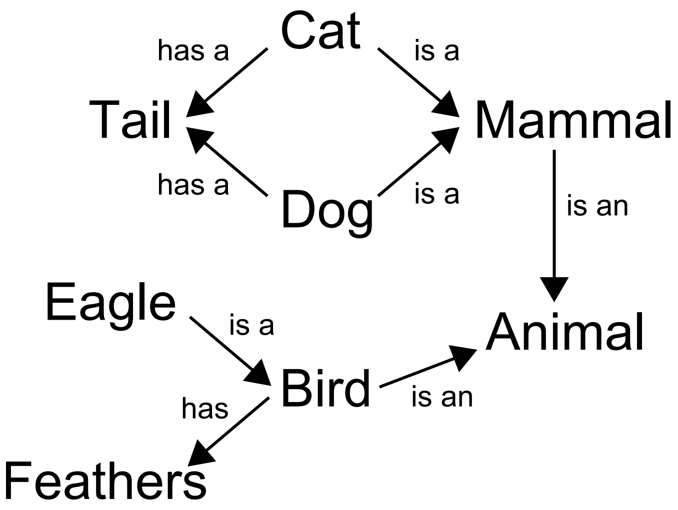


Figure 2.1: A simple semantic network

tics are not well understood and warrant further research. Given a large number of examples of an object, an abstract representation of it gradually emerges implicitly. This representation may be categorical in nature. For example, given a number of apples of different colours, shapes, and sizes, an abstract concept that represents an apple emerges that then aids in recognising and classifying apples with novel characteristics. This emergent behaviour seems to be an inherent property of associative memory. There is evidence that there is a dissociation in the brain regions that represent such abstract knowledge and information of specific instances [Beauregard et al., 1997; Kiehl et al., 1999; Mummery et al., 1998; Vinson et al., 2003]. For instance, the visual memory of a particular object may be neuroanatomically different from the abstract category it belongs to. Such characteristics are modelled in CAs in a simulation (described in Section **??**), where emergent semantic relationships in a large associative memory are used for natural language processing. While it is difficult to quantify these associations, the model provides some insight into their nature.

2.3 A review of related computational associative memory models

The number of ANNs applied to a wide range of problems is staggering. However, general purpose models are few, as many are localised to specific domains. While it is difficult to compare all such systems, this section discusses in some detail, four widely used ANN models of general associative memory. These ANNs are discussed in contrast to the computational CA based ANN used in this thesis (Chapter 4). Some mathematical details of learning in these ANNs are presented so as to highlight the similarities and differences with the ANN used in this thesis. While these ANNs are powerful tools capable of modelling associative memories of varying complexities, the work described in this thesis demonstrates that the CA model with its strong neurobiological underpinnings provides a range of dynamics that are able to account not just for complex associative memories, but higher processes emerging from them.

2.3.1 The Hopfield network

The Hopfield network [Hopfield, 1982] is a widely used autoassociative ANN. As an autoassociative network [Bishop, 1995], it is able to recall a stored pattern upon being presented with the same pattern or a partial version of it. It is a fully connected network of simple neurons—binary threshold units—with symmetric bidirectional connectivity, that serves as a content addressable memory. The model neuron in the Hopfield network is a simple unit that can have one of the two states, firing or inactive. The state s_i of a neuron i is the only value it can take and is usually either 1 or 0, or, 1 or -1. Neurons in the network connect to every other neuron but themselves with symmetric connection strengths or “weights”. As a result, the Hopfield network does not have distinct input or output neurons and every neuron acts as an input and output. Figure 2.2 shows an example network with two neurons i and j, where W is the weight matrix which is the collection of all individual connection weights between neurons, and $w_{ij} = w_{ji}$.

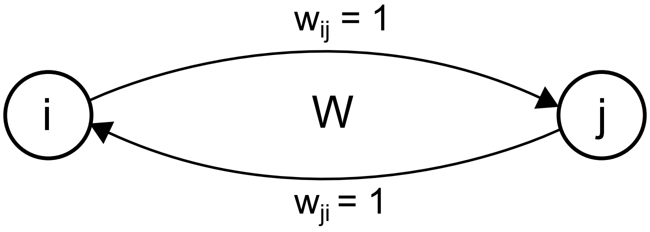


Figure 2.2: An example Hopfield network with 2 neurons

The neurons in the network usually update their states synchronously and each neuron retains its state until it is updated. Equation 2.1 shows the state s_i of a neuron i, where w_{ij} is the connection weight of the neuron i from neuron j, s_j is the state of the neuron j, and θ_i is its threshold. By the rule, patterns presented to the network eventually converge to a global "stable state". This resembles the formation of Hebbian CAs (Section 3.1). Such stable states are called *attractors* (further discussed in Section 3.4 in the context of CAs).

$$s_i = \begin{cases} 1 & \sum_j w_{ij} s_j \geq \theta_i \\ 0 & \sum_j w_{ij} s_j < \theta_i \end{cases} \tag{2.1}$$

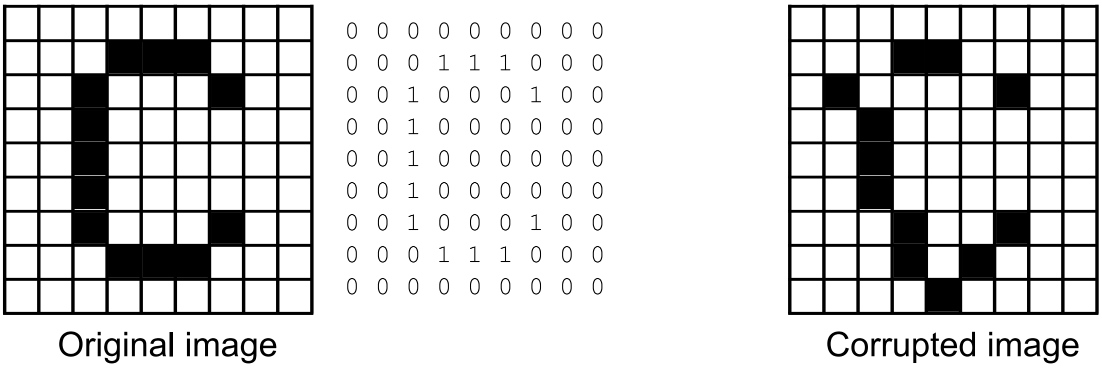
At any instant, the network has a collective state that can be represented as a vector s. In the example in Figure 2.2, the network can have two stable states depending on the two possible values the weight matrix can have. If $w_{ij} = 1$, $s = (1, 1)$ or $s = (0, 0)$. If $w_{ij} = 0$, $s = (1, 0)$ or $(0, 1)$. The network is trained by presenting it with a pattern, where the values of neurons are set externally to the corresponding pattern. After many iterations, the network converges to a stable state, where the pattern is stored in the network's weight matrix. Subsequently, when presented with the same pattern or a variation of it, the network should settle to the previously attained stable state, performing a pattern recollection.

This is the fundamental process that facilitates pattern recognition in the Hopfield network.

Every network state in the Hopfield network has an "energy" associated to it. The network behaves in such a way that with each iteration, it tries to minimise its energy so as to converge to a local minima. When the network is presented with a pattern, and it converges to such a state of minimal energy, it is considered to have attained a stable state. The energy E of the network is given by Equation 2.2.

$$E = -\frac{1}{2}\sum_{ij} w_{ij} s_i s_j \tag{2.2}$$

Hopfield networks have been used for numerous tasks ranging from classic computer science problems such as the theoretical *Travelling salesman problem* [Hopfield and Tank, 1985] to image processing [Paik and Katsaggelos, 1992]. They are widely used for image processing tasks, especially, optical character recognition (OCR). For example, a 9 × 9 black and white pixel image of the letter C can be applied directly as a pattern to a 9 × 9 Hopfield network, where every white pixel can be considered to have the state 0, and every black pixel, 1. During training, the image, essentially a large pattern, converges to a stable state in the network. If the same image or a partial version of it is then presented to the network, it should converge to the stable state attained earlier or one close to it, effectively recognising the original form of the image. Figure 2.3 illustrates this example, where the original image and a corrupt version of it that may recall the original image, are shown. Every cell represents a single pixel in the image and also a neuron in the network. Black cells represent firing neurons (1) and white cells represent inactive neurons (0). The pattern corresponding to the original image that is presented to the network is also shown.



```
0 0 0 0 0 0 0 0 0
0 0 0 1 1 1 0 0 0
0 0 1 0 0 0 1 0 0
0 0 1 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0
0 0 1 0 0 0 1 0 0
0 0 0 1 1 1 0 0 0
0 0 0 0 0 0 0 0 0
```

Figure 2.3: A visualisation of optical character recognition in a Hopfield network

2.3.2 Bidirectional associative memory

Bidirectional associative memory (BAM) [Kosko, 1988] is a recurrent heteroassociative ANN. It contains two “layers” of neurons X and Y where the connectivity of the neurons are bidirectional. The neurons are simple binary units that generate the output value of $+1$ if they fire, and -1 if they are inactive. The two layers in the network transmit information back and forth and act as either input or output layers depending on the direction of information propagation. That is, if an external pattern is presented to X, X acts as the input layer and Y as the output layer, and vice versa.

As a heteroassociative memory, the network is able to associate two different patterns with each other, as opposed to an autoassociative memory like the Hopfield network that can only perform single pattern recalls. That is, the network encodes pattern pairs that after learning, when presented with one pattern of a pair, results in the recollection of the other. In addition, the network can also act as an autoassociative memory. Figure 2.4 illustrates an example BAM network. The network is fully connected, and X and Y are the two layers with m and n neurons respectively. The number of neurons in the two layers need not be equal. Like in the Hopfield network, the connections are bidirectional—$w_{x_i y_j} = w_{y_j x_i}$.

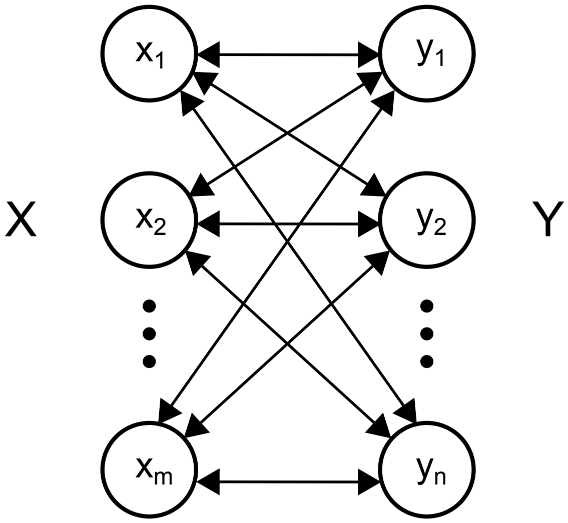


Figure 2.4: An example BAM network

When a pattern is presented to X, it sends computed information to the output layer Y. The output layer then transmits the results of computations from it back to X. This process is repeated until the network attains a stable state when information transmitted back and forth stops changing. Like the Hopfield network, patterns are stored in the network in a weight matrix W. To encode and associate a pair of patterns in the network, they are applied to X and Y respectively. If there are N pairs of patterns, and x_i and y_i are such a pair, the weight matrix encoding them is shown in Equation 2.3. Here, x and y are row vectors, and t denotes their transposition.

$$W = \sum_{i=1}^{N} {x_i}^t y_i \tag{2.3}$$

After training, any pattern x_i applied to the input layer X, should produce the corresponding y_i in the output layer Y, or vice versa. When a novel pattern is presented, the network outputs the closest output pattern. Thus, the BAM network is capable of storing patterns with simple one to one associations.

2.3.3 Self organizing maps

Kohonen's [1982] self organizing maps (SOM) are feedforward ANNs that do unsupervised learning. The SOM network consists of an input layer and a *Kohonen layer* or output layer. Unlike the Hopfield Network or BAM, the output in the SOM network is derived from a single neuron in the output layer based on a "winner takes all approach", where neurons compete for activation. The output neuron is determined by the internal state of the network and does not need to be specified externally—hence, unsupervised learning. The activations of the neurons are continuous and are usually between 0 and 1. The network functions by organising its state into specific topologies, given large input patterns. When presented with a novel input pattern, the neuron closest to its topology in the state space fires. The network adapts to this in such a way that its weights are proportional to the "closeness" with the winning neuron, encoding the input pattern. This self organisation bears resemblances to the formation of CAs via Hebbian learning.

Figure 2.5 shows an example SOM network with an input layer X with two neurons and an output layer Y with four neurons. The input neurons are fully connected to the output neurons, but the output neurons do not have connections among themselves. Information flows in a single direction in the network, from the input layer to the output layer. If an input pattern P, for example $P = (0, 1)$, is applied to the input neurons, the output layer may organise itself in such a way that neuron y_1 fires. Similarly, for an input $P = (1, 0)$, the neuron y_3 may fire. By the dynamics of the network, every time these inputs are presented, the weights of the neurons in the output layer are adjusted so that they move closer to the input patterns, subsequently resulting in the firing of a single neuron closest to the pattern.

Given P, training the network involves setting arbitrary weights to all neurons, where w_{ij} represents the connection weight between neurons i and j. The winning neuron is then determined based on Equation 2.4, where d_{ij} represents

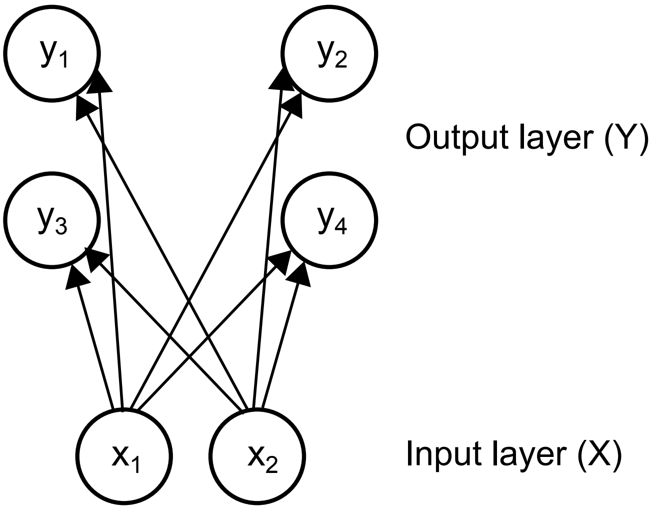


Figure 2.5: An example SOM network

the neuron with the minimum distance from the input pattern.

$$d_{ij} = \sqrt{\sum_{i \in Y} (P_i - w_{ij})^2} \tag{2.4}$$

The neurons in the neighbourhood of the winning neuron are adjusted closer to the input pattern based on Equation 2.5, where η is a dynamic learning constant that degrades (approaches 0) over time. The neighbourhood may be the radius from the winning neuron determined based on the topography of the network. These steps are repeated until the network is considered to have attained convergence depending on the properties of the input pattern.

$$w_{ij_t} = w_{ij_{t-1}} + \eta(P_i - w_{ij_{t-1}}) \tag{2.5}$$

SOM networks are generally applied to machine learning tasks in AI. They are extensively used for data mining and classification of complex datasets [Vesanto and Alhoniemi, 2000] due to their ability to organise large multi-dimensional data into smaller spaces. Like the Hopfield network, they are also used in image processing. If a network is trained with an input pattern representing the pixels of an image of the letter C, a single neuron in the output layer will fire, representing C. In the future, a partial input representing the same letter may cause the same neuron to fire based on similarity, performing letter recognition. The network may be trained with patterns representing all 26 letters in the English alphabet, where 26 output neurons will represent each of the letters.

Kohonen's SOM network is inspired by self organising processes thought to be in the brain [Kohonen, 1990]. CA formation may be considered such a process. However, standard SOMs are not suitable for modelling complex associative memories. They are inherently unable to do heteroassociation, a primary characteristic of complex associative memories. Processing in the SOM network is dependent on its fully connected topology that is unlike the topology in the brain [Schüz, 1998]. Like the Hopfield network, the simplistic neurons in the SOM network have some properties of the biological neuron that influence learning in the brain. The ANN described in this thesis encompasses many such properties of the biological neuron. Chapter 4 discusses these in detail.

2.3.4 Adaptive resonance theory network

The adpative resonance theory (ART) network, developed by Grossberg [1976] and Carpenter [1987], is an ANN model based on Grossberg's ART theory of human cognitive information processing. The ART model describes multiple ANNs with varying capabilities such as ART2 [Carpenter and Grossberg, 1987] that supports learning of continuous patterns, and ART3 [Carpenter and Grossberg, 1990] that encompasses basic aspects of neurotransmitters found in the brain. ART is different from the models discussed in previous sections, as it is inspired by ideas of cognition. It supports certain cognitive processes such as context sensitivity.

The basic ART network does unsupervised learning and consists of two layers of neurons, an input layer or the comparison layer and an output layer or the recognition layer. In addition to the two layers, there is a *reset* mechanism.

Neurons in the two layers are fully connected with each other and thus have strong feedback links. The back and forth connections between the layers are different from each other—connections from the input layer are known as *bottom up* links, and connections from the output layer to the input layer are known as *top down* links. This is inspired by cognitive processes in the brain thought to be of a similar nature [Grossberg, 1976].

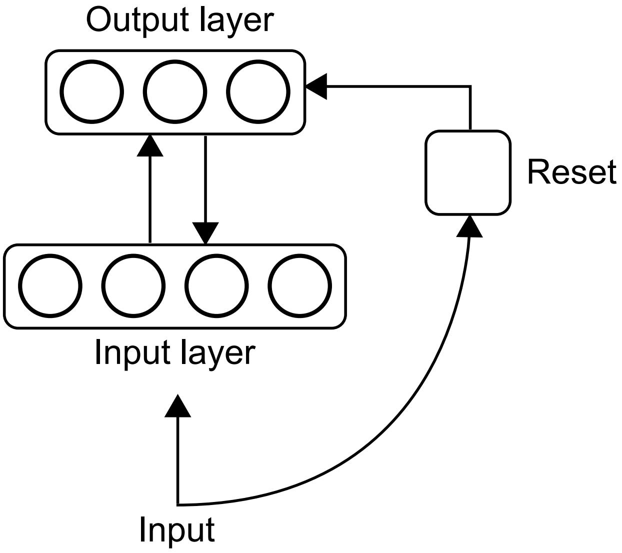


Figure 2.6: Basic architecture of the ART network

Figure 2.6 shows the basic architecture of the ART network. The reset mechanism is a symbolic module that heavily influences the system. In the network, the input layer takes an input pattern and propagates the resultant activity to the output layer. This activity is stored in the weight matrix as the long term memory representation of the input pattern. The activation a_i of a neuron i in the output layer with a connection from neuron j in the input layer, provided a pattern P to the input layer, is given in Equation 2.6, where w_{ij} is the connection weight and P_i is a component of the input pattern.

$$a_i = \sum w_{ij} P_i \tag{2.6}$$

Subsequently, like the SOM network, the output is derived from a single neuron in the output layer, decided based on a winner takes all strategy. Activation of the winning neuron in the output layer is then fed back to the input layer. The reset module compares this activation with the original input against a *vigilance threshold* (ρ)—a value between 0 and 1 that dictates how general or fine grained categorical representations should be in the output layer. If the comparison exceeds ρ ($r > \rho$), the input pattern is considered to have been categorised. This is known as the *vigilance test.* The value of r is given by $\frac{a}{\sum P_i}$. If $r < \rho$, the neurons in the output layer are searched to find the closest match to the input. If no such neuron is found, a new categorical representation for the input is formed in the output layer. The network is considered to "resonate" when it learns a categorical representation for an input pattern on passing the vigilance test.

The self organising behaviour in the ART network is similar to that of the SOM network and CAs. The network learns from inputs and autonomously categorises them based on their hidden characteristics. Hence, it is extensively used for categorisation tasks [Gan and Lua, 1992; Song et al., 1999].

Sequitur

The previous sections briefly discussed a few widely used computational models of associative memory. While they are versatile tools used for many AI tasks, they are not good at accounting for many dynamics of complex associative memories. The "neurons" in these models are rudimentary abstractions that bear little resemblance to the biological neuron. Also, the fully connected networks in these models are unlike the connectivity of neural networks in the brain [Kangas et al., 1989; Schüz, 1998]. Due to their topology dependent learning mechanisms, they are unable to make use of connection topologies thought to be in the brain [Koenig et al., 2005] that may underlie processes supporting memory. For instance, while the brain is thought to encode associative memories via synapses and physical overlapping of memories (discussed in Section 3.2.4), these models are only capable of simple synaptic associations.

Being an autoassociative memory, the Hopfield network is unable to represent associations between different patterns, unlike human associative memory that encodes many complex associations and relationships between memories. The standard SOM and ART networks are distinct in their way of encoding memories, but do not account for heteroassociative memories. While BAM is heteroassociative, it is limited in its ability to encode multi-associative memories, as it can only encode pattern pairs. Even though the SOM and ART networks are inspired by the self organising processes thought to be in the brain [Kohonen, 1990] such as the formation of CAs, they are limited in their ability to form complex associative memories. The ART network, although influenced by cognitive processes, lacks neurobiological plausibility. For instance, learning in the network is driven entirely by the symbolic reset mechanism. Extensions of some of these models have been introduced so as to address certain shortcomings [Hagiwara, 1990], but they do not necessarily unify diverse dynamics.

It is important to consider these shortcomings that arise from the lack of neurobiological bases. While these models are excellent tools for various tasks, it is reasonable to assume that a model that can offer a unified account for many of the observed properties of associative memory systems may be better for many AI tasks. Properties of biological neurons such as fatiguing and decaying may aid in the graceful "forgetting" of memories, an important characteristics of human associative memory and learning. A unified model of associative memory that encompasses such characteristics may help bring together various facets of the human associative memory that constitute intelligence. As is shown in this thesis, the FLIF ANN used for modelling CA based associative memory (Chapter 4) offers dynamics stemming from neurobiologically inspired properties such as decay and fatiguing of neurons that are able to account for many such properties. The different associative memory models developed using the ANN (Chapter 5) support this prospect. It is not the goal of this thesis to discourage other models, but to explore the CA model of associative memory as a strong candidate for modelling in AI.

2.4 Chapter summary

This chapter provided the background material and important concepts that support the work described in this thesis. It briefly discussed ANNs in general, and provided an overview of various facets of human associative memory and its neurobiology. It also reviewed four related computational models of associative memory models in detail, in comparison to the CA model and the FLIF ANN the work described in this thesis is based on. The next chapter examines Hebbian CAs in detail, discussing their neurobiology, general characteristics, and the mechanisms by which they form associative memories.

CHAPTER 3

Hebbian cell assemblies

Hebb [1949] proposed the idea of CAs—reverberating circuits of neurons in the brain that are considered to be the neural basis of concepts—in his seminal book, *The Organization of Behavior*. It served as an impetus to the interest in neuropsychology in the second half of the twentieth century that continues today, greatly influencing the modern field of cognitive neuropsychology. The hypothesis also inspired many computational models of AI, especially the idea of attractor neural networks [Amit, 1989]. Attractor neural networks are powerful tools for modelling a wide range of tasks [Belavkin and Huyck, 2010; Byrne and Huyck, 2010; Fransen and Lansner, 1998; Garagnani et al., 2009; Huyck, 2001, 2008; Huyck and Nadh, 2009; Knoblauch et al., 2007; Wennekers, 2007, 2009; Wennekers and Palm, 2000; Wickelgren, 1999]. The ANN described in this thesis (Chapter 4) is such a system.

The CA hypothesis was radically different from previous accounts of information processing in the brain, as it suggested that not individual neurons, but certain higher processes emerging from the interplay of large networks of neurons give rise to higher order psychological phenomena such as memory [Elman, 1991; Milner et al., 1998]. Computational CA models are based on this principle and may provide added benefits for modelling neuropsychological tasks.

Though Hebb merely hypothesised the existence of CAs, there is now extensive research suggesting their presence and their role in giving rise to higher order cognitive phenomena [Engel et al., 1991; Funahashi, 2001; Fuster, 1999; Harris, 2005; Hauk et al., 2004; Kelso et al., 1986; Maurer et al., 2003; Molnár et al., 2008; O'Neill et al., 2008; Pasupathy and Connor, 2002; Peyrache et al., 2010; Plenz and Thiagarajan, 2007; Pulvermuller, 1999; Sakurai, 1998; Spatz, 1996].

This chapter provides a detailed overview of CAs and their neurobiology—how they inherently support associative memory, and some of their characteristics that shed light on the underlying mechanisms that cause the emergence of higher order phenomena. The background information discussed in this chapter is meant to provide an understanding of the fundamentals on which the work described in this thesis is based.

3.1 Neurobiology of CAs

In the brain, objects, ideas, and abstract concepts are represented not by individual neurons, but by groups of simultaneously active neurons with high mutual synaptic strengths [Wennekers and Palm, 2000]. CAs are central to associative memory as they, by nature, represent stimuli encoded in groups of neurons. The CA theory postulates that CAs are formed by a Hebbian learning rule, whereby modifications in the synaptic transmission efficacy are driven by the correlations in the firing activity of pre-synaptic and post-synaptic neurons [Gerstner and Kistler, 2002]. In Hebb's words:

> "When an axon of cell A is near enough to excite cell B or repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." [Hebb, 1949]

That is, if two neurons A and B co-fire, where A is the pre-synaptic neuron transmitting a signal to the post-synaptic neuron B, the strength of the transmission synapses or the mutual synaptic strength increases. This is known as the Hebbian learning rule.

When neurons are quiescent, the inside and outside of their cell bodies are in a state of electrical equilibrium. They accumulate potential transmitted by pre-synaptic neurons. When this accumulated potential crosses a critical threshold, they fire. This transient event is known as an action potential. Here, when neuron A fires, it undergoes a change in polarity that is propagated via its axonal terminals and is passed to the connected post-synaptic neuron B, via a release of specialised chemicals known as neurotransmitters. The behaviour of B is determined by how its receptors react to the neurotransmitters. The resultant behaviour may be excitatory or inhibitory, where excitation increases the likelihood and inhibition reduces the likelihood of B firing. This is illustrated in Figure 3.1, where the pre-synaptic neuron A is shown to be sending signals to the post-synaptic neuron B.

Such changes are encoded in the synapses of the neurons so that the likelihood of B responding to signals coming from A is higher in the future. This property of synapses to change their strength based on neural activity is termed *synaptic plasticity* [Abbott and Nelson, 2000]. The more two neurons co-fire, the higher their mutual synaptic strength grows and the larger the said likelihood becomes. Thus, stimuli repeatedly exciting a large number of neurons can increase their mutual synaptic strength, assembling them into a CA. Such a CA may respond to the original stimuli or variations of it in the future, and stay active even after external stimuli disappears. This reverberative behaviour (discussed in detail in Section 3.2) is one of the fundamental dynamics of the CA. If a variation of the stimuli excites certain neurons of the CA, the activity may spread to other neurons and activate the whole CA due to the high mutual synaptic strength of its constituent neurons.

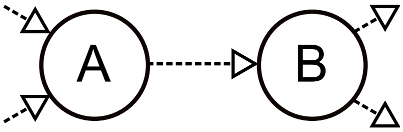


Figure 3.1: Co-firing pre-synaptic and post-synaptic neurons

As mentioned in the beginning of this chapter, there is extensive evidence for CAs in the mammalian brain. Hebb's postulate that not individual neurons, but activity of large groups of neurons are the basis of mental processes, has been experimentally observed. From these observations, it is evident that CAs are central to the fundamental processes of associative memory. Two such recent studies are briefly discussed in the following subsections.

3.1.1 Hebbian learning in the rat hippocampus

The hippocampus is a region of the mammalian brain thought to be extensively involved in memory processes such as formation and association [Eichenbaum, 2004; Hartley et al., 2007; Hassabis et al., 2009; O'Reilly and Rudy, 2001; Shohamy and Wagner, 2008; Squire, 1992; Tulving and Markowitsch, 1998]. Recently, O'Neill et al. [2008] studied the firing patterns of specialised neurons in the rat hippocampus that are responsible for encoding spatial and navigational information, aiming to better understand the underlying mechanisms of reactivation of neurons encoding such information. The study revealed that these neurons fire recurrently during sleep, even after external stimuli have long ceased. The rate of co-firing of these groups of neurons increased based on the number of times they were active together when not asleep, for instance, neurons representing similar places that co-fired during an exploration task. In contrast, neurons representing dissimilar places experienced reduced co-firing rates. These are indicative of the Hebbian learning mechanism that drives the formation of CAs. The reactivation behaviour of neurons are thought to govern the consolidation of memories [Carr et al., 2011], further indicating CA formation through repeated co-firing of neurons.

3.1.2 Human cortical CAs

Molnár et al. [2008] analysed slices of cortical regions of the human brain in order to study cortical neural activity. The data they recorded was the first of its kind for human brains. They found that individual action potentials are capable of triggering activity in groups of neurons. The activity can propagate further and excite large networks of neurons. The initial potential was achieved by applying external stimulation to a particular neuron. Such synaptic strength possessed by a neuron is accounted for by the Hebbian learning mechanism, where co-firing increases synaptic efficacy of neurons. The initial action potentials triggered long lasting activity across large networks of neurons that persisted even after the removal of the external stimuli. This network of synaptically connected neurons spanning different brain regions affirms Hebb's idea that not individual neurons, but groups of neurons are responsible for information processing. The researchers suggested that such CAs may be the building blocks of higher order cognitive phenomena as proposed by Hebb.

3.2 Dynamics of CAs

CAs exhibit several interesting properties that provide insight into different aspects of information processing in the brain. These properties provide a viable account for the fundamental processes of memory formation and association. Computational CA models that can emulate these properties are able to provide a unified mechanism for simulating neurobiologically realistic long term and short term memories unlike traditional separatist models. This section overviews the basic properties of CAs, namely, how formation of CAs account for long term memories; how their reverberative behaviour accounts for short term memories; how synaptic associations between CAs form associative memories; how overlapping CAs can account for another type of association; and the competitive nature of CAs.

3.2.1 Formation and long term memory

As discussed in the previous sections, CAs are formed when mutual synaptic strength of neurons with inter-synaptic connections increase as a result of repeated co-firing. An external stimulus may excite a group of neurons leading to the formation of a CA that is bound to that particular stimulus. If the stimulus is the sensory information of seeing an object, for instance, an apple, the CA formed subsequently is the neural representation of the apple. This process of physical synaptic modification resulting in the formation of CAs drive long term memory formation [Buzsáki and Draguhn, 2004; Girardeau et al., 2009]. This apple CA may be activated in the future when encountering the same apple, or a part of it. It may also be activated by the sight of different apples with similar features, as excitation of a few neurons of a CA may lead to the activation of the CA as a whole, performing completion effects [Lansner, 2009]. While in reality, there are many complex processes involved in long term memory formation, this is a simplified account of the underlying neural mechanisms.

3.2.2 Reverberation and short term memory

Reverberation in terms of CAs is the process of activity cascading through a CA even after an external stimulus that triggered it ceases. This persistent firing behaviour of neurons in the CA may be triggered by an external stimulus, or by activity from a connected CA. The reverberation may persist until the recurrent energy of the CA dissipates due to fatiguing neurons, or activation from connected CAs cease. Such transient activation of CAs accounts for short term memory, where a concept represented by a CA can be considered to be active in the working memory for a short period of time. This implies that CAs are a part of the working memory only when they are active. When a CA is transiently active, it may enter a labile state where further modifications are possible, for instance, association with a different CA [Shirvalkar, 2009]. So, CAs once formed are subject to changes in the future such as association, weakening, or even fractionation into smaller CAs, influenced by various factors.

3.2.3 Association via synapses

CAs can be associated with each other via the same mechanism that drives their formation. Repeated co-firing of neurons may assemble them into a CA. If a group of co-firing neurons with synaptic connectivity belong to different CAs, an increase in their mutual synaptic strengths will drive them to association, in turn, associating the CAs they are a part of. Two such CAs with synaptic association may activate each other in the future when one of them becomes active. This mechanism may underlie the association of non similar, but co-occurring concepts. For example, while an aeroplane may not share any superficial or abstract similarities with the sky, the thought of one often elicits the thought of the other. This may be due to the repetitive co-occurrence of the two concepts, where the CA representing the aeroplane is co-active with the CA representing the sky. This behaviour is consistent with aspects of Pavlovian conditioning discussed earlier, where stimuli are associated via repeated co-presentation.

There is another type of CA association where concepts sharing a large number of abstract features may be encoded by a common subset of neurons in CAs representing them. This association, association via neural overlaps, is discussed in the next section.

3.2.4 Association via neural overlaps

A neuron is not limited to a single CA and may belong to multiple CAs. So, there may be physical regions of neural overlap between CAs [Sakurai, 1996, 1999]. Such neuronal overlaps may be governed by factors such as similarity of initial external stimuli. There is evidence that neurons in a CA represent different aspects or features of the information that the CA encodes [Osan et al., 2011]. This implies that similar stimuli may be bound to CAs that have physical overlaps, where neurons in the overlapping regions encode similarities.

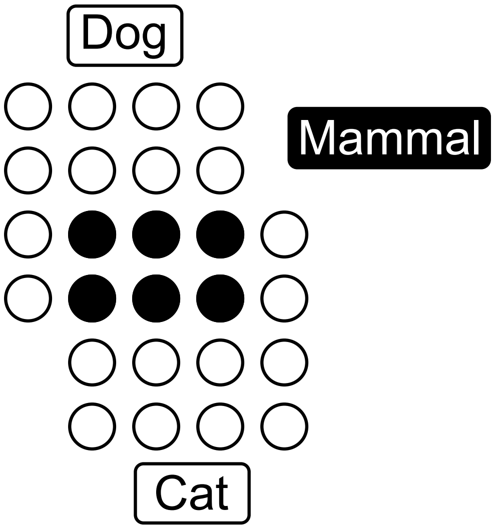


Figure 3.2: An example of neural overlap in CAs

Figure 3.2 shows an example of two physically overlapping CAs. The *dog* CA and the *cat* CA are composed of 16 neurons each, of which six are shared between the two. These shared neurons represent the abstract features of mammals shared by dogs and cats. While these shared neurons may be a part of the *dog* and *cat* CAs, they may represent a separate *mammal* CA. The behaviour of this *mammal* CA may depend on context. When external stimuli activates the *dog* CA, activation may spread to the *mammal* and *cat* CAs. While this is an extreme simplification of biological CAs that may have tens of thousands of neurons each, and may share thousands of neurons with other CAs, it serves as an illustrative example. A model shows this behaviour in overlapping CAs [Huyck, 2007]. Such behaviour may be the basis of “train of thoughts”, where thinking of a dog leads to thoughts of different kinds of dogs, and other related animals, as it is known that human memories are distributed in such ways [Levy and Horn, 1999].

Complex cognitive processes are represented by large scale overlapping neural circuits in the brain [LaBar et al., 1999]. There is evidence that brain regions involved in vastly different processes such as linguistics and motor movements have information encoded across them in a distributed manner [Kable et al., 2002]. In human subjects, the brain regions activated in a passive reading task involving action words such as *pick* and *kick* were found to physically overlap with regions concerning fingers and feet respectively [Hauk et al., 2004].

Sharing of neurons across CAs provide increased capacities for encoding information as opposed to encoding them in discrete neurons. That is, overlapped representations allow encoding of more information than the available number of neurons. Such distributed representations also increase the resilience of encoded memories. While it may seem that extensive overlaps may reduce the discriminability of the concepts they encode, that is not necessarily the case [Levy and Horn, 1999; Palm, 1991; Valiant, 2005; Yoshizawa et al., 1993]. Moreover, spreading of activation across overlapping CAs enable passing of information from CA to CA. Hence, overlapped encoding of CAs is an important mechanism underlying associative memory.

3.2.5 Inter-CA competition

CAs compete for activation via their constituent inhibitory neurons. When two CAs are active, mutually inhibitory connections between them send inhibitory signals trying to suppress each other's activation. This behaviour supports classification of ambiguous stimulus, where multiple CAs activated by the same stimulus compete, with the CA better incorporating the stimulus becoming more active. For instance, when sensory stimuli representing objects with spatial and temporal equivalence are encountered, active regions in the brain corresponding to them compete and one representation wins, as two objects cannot be usually perceived to occupy the exact same space at any given time [Keysers and Perrett, 2002]. Similarly, it has also been observed that models of CAs in the cortical regions concerning hearing compete for activation when activated with stimuli representing vowels throughout the period of stimulation [Hoshino and Zheng, 2004]. Competition also limits the number of CAs that may be active in a system with capacity constraints, such as the working memory [Kaplan et al., 1991]. In computational CA models, competition between CAs via inhibitory connections reduce the probability of cascades of neural activity causing undesirable CAs to activate [Connolly and Reilly, 2005].

3.3 CAs and associative memory

CAs by nature are memory states. As discussed in the previous sections, formation of CAs is considered the basis of long term memory. A CA formed by an external stimulus may be the mental representation of it. This external stimulus or variations of it can activate the CA which may reverberate even after the stimulus ceases. This reverberation is considered the basis of short term memory. Reverberating CAs may develop associations with other CAs. This may be based on the similarity of the stimuli or concepts representing them, contiguity, or depending on how often they are co-active [Wickelgren, 1999]. Such associations may develop via the co-activation of CAs, where associations are encoded as strengthened synaptic efficacies via Hebbian learning. Associations between CAs may also be represented by physical neural overlaps as discussed in Section 3.2.4, where overlapping regions represent shared features. Chapter 5 describes models that explore synaptic and overlapping associations in CAs.

3.4 CAs as attractor states

The idea of attractor networks was inspired by an important early work of Hopfield [1982]. They are recurrent networks of neurons, in which patterns of activity tend to settle to “attractor states” over time [Amit, 1989]. When presented with an external stimulus, activity elicited in such a network may settle to a particular state that then becomes the internal representation of it. When the same stimulus or its variants are presented to the network in the future, activity in the network may converge to the previously formed attractor state—that is, become “attracted” to it. For this reason, attractor networks may also serve as content addressable associative memory systems, as it is possible to retrieve previously learnt states.

A neural network can encode attractor states by stimulus driven synaptic modification of groups of neurons via Hebbian learning [Amit and Brunel, 1995]. Such groups of neurons with modified synaptic efficacies bound to a particular stimulus may fire in response to the same stimulus or variants of it in the future. Since this is essentially the process of CA formation (discussed in Section 3.1), CAs can be considered attractor states, where CAs formed by external stimuli are their mental representations. There are studies suggesting the presence of attractor dynamics in the brain [Cossart et al., 2003; Rolls, 2007; Wills et al., 2005].

Attractor networks have been used for a wide range of tasks [Byrne and Huyck, 2010; Fransen and Lansner, 1998; Garagnani et al., 2009; Huyck, 2007, 2008; Knoblauch et al., 2007; Wennekers, 2007, 2009; Wennekers and Palm, 2000; Wickelgren, 1999]. The ANN used in this thesis (discussed in the next chapter) is an attractor network.

3.5 Chapter summary

This chapter discussed Hebbian CAs in detail. It delved into their neurobiology and physiology, various dynamics exhibited by them, and the mechanisms by which they form different kinds of associative memory. It also briefly reviewed two recent neurobiological studies on CAs in mammalian brains. The next chapter discusses the FLIF ANN with which CAs are modelled in the various simulations described in this thesis.

CHAPTER 4

The FLIF neural network

The previous chapter discussed Hebbian CAs and their properties. Computational CA models encompassing such neurobiologically inspired characteristics are able to model complex processes. An ANN based on *Fatiguing Leaky Integrate and Fire* (FLIF) neurons that can simulate CAs via the mechanisms described in the previous chapter, is detailed here. It is based on the *Connection, Association, and Networking Technology* (CANT) neural network framework written in the Java programming language [Huyck, 2001]. The architecture is able to simulate a large number of neurons in real time, with as many as several hundred thousands in a single network. All associative memory models developed in this thesis are modelled with this ANN. These models are described in detail in the next chapter.

4.1 The FLIF neuron

The FLIF neuron model is an extension of the widely used Leaky Integrate and Fire (LIF) neuron model [Gerstner, 2002; Maass and Bishop, 2001; Tal and Schwartz, 1997]. They encompass many characteristics of biological neurons [Huyck, 2007]. Like the biological neuron, the FLIF neuron integrates incoming potential from pre-synaptic neurons. On surpassing a critical threshold, it fires with a burst transmitting the accumulated potentials via its synapses to post-synaptic neurons. At every instant the neuron is integrating potential, while below the threshold, some potential leaks away. This leaking behaviour is based on how potential leaks through membranes of biological neurons [Churchland and Sejnowski, 1992]. Similarly, every time the neuron fires, it experiences fatigue, the “tiredness” felt by the neuron. Fatigue affects the firing threshold of the neuron in such a way that it becomes relatively difficult for it to fire in the near future. As a result, repeatedly firing neurons, for instance in a reverberating CA, die out, eventually enabling transient short term activation of CAs. It has been demonstrated that CAs modelled with FLIF neurons are Turing complete [Byrne and Huyck, 2010], implying that it is possible to model complex computational processes in a purely neural system.

4.2 Network properties

The parameters governing the four characteristics of the FLIF neuron (fatiguing, leaking, integration, and firing), the connectivity of neurons, and other properties of the FLIF ANN are discussed in this section.

4.2.1 Topology

In the network, neurons follow a connection topology inspired by the biological distanced biased connectivity. The neurons are laid out in a $N = rows \times columns$ grid where neurons on the edges are near the neurons on the opposite edges, making the network toroidal. It is possible to partition the network into smaller toroidal networks or *subnets* for modularity. Each neuron connects to neighbouring neurons within a fixed city block distance (d), where $d = 4$, as it has been observed to work well across many simulations [Huyck, 2000, 2007; Huyck and Nadh, 2009]. In addition to the local connections, a connection is made to another random area in the subnet via a long distance axon [Churchland and Sejnowski, 1992]. Hence, the closer the two neurons are, the higher their likelihood of being interconnected. This is illustrated in Figure 4.1. The figure shows the partial connectivity of a single neuron in a hypothetical 5×5 network where its local connections are within d, and a long distance connection to a neuron at the bottom of the network. The neuron also connects to the neuron at the opposite edge of the network, as on a torus, it is in the immediate vicinity.

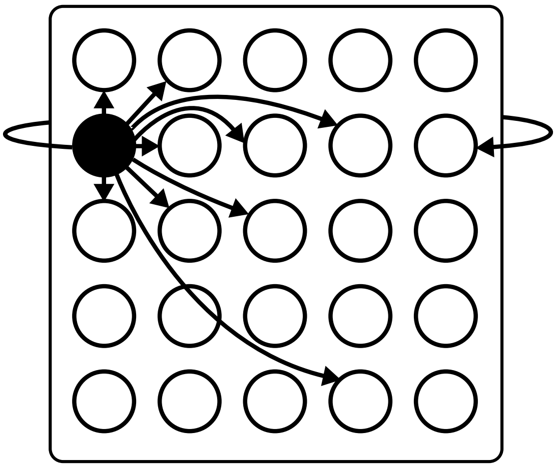


Figure 4.1: Connectivity of a single neuron in a hypothetical 5x5 FLIF neural network

The connectivity rule is shown in Equation 4.1. There exists a connection between neuron i and j only if $c_{ij} = 1$, where d is the distance, κ is the connection probability (a constant), and r is a random number between 0 and 1.

$$
\begin{aligned}
D &= \left(\frac{1}{d * \kappa}\right) \\
c_{ij} &= \begin{cases} 1 & r < D \\ 0 & r > D \end{cases}
\end{aligned}
\tag{4.1}
$$

Thus, neurons in the network are sparsely connected. Even though there are a staggeringly large number of neurons—an estimated 100 billion—in the human brain, on average, a neuron synapses with only about 1000 other neurons [Koenig et al., 2005]. The connection topology in the neural network is an extreme but faithful simplification of this. It adheres to the connection topology in the brain, where neurons reside in distributed circuits and not in fully connected networks [Schüz, 1998]. This is in contrast to the fully connected ANNs previously discussed in Section 2.3.

4.2.2 Excitatory and inhibitory neurons

The network has two different types of neurons, excitatory and inhibitory neurons. Excitatory neurons transmit positive potentials upon firing to post-synaptic neurons which increases the likelihood of them firing. When an inhibitory neuron fires, it transmits negative potential to post-synaptic neurons, reducing their accumulated potential and decreasing the likelihood of them firing. This is based on the observation that, generally, neurons in the brain show either excitatory or inhibitory behaviour [Eccles, 1986]. The excitatory and inhibitory nature of a neuron is determined by how its receptors respond to chemical neurotransmitters coming from a pre-synaptic neuron [Vicario-Abejn et al., 1998]. While the network does not account for neurotransmitters, their behaviour is abstracted into the excitatory and inhibitory classification of neurons.

In the network, synaptic weight or synaptic efficacy w_{ij} of the synapses of a neuron i connected to j has the polarity p, so that every time i fires, the potential propagated through its synapses is affected by p. For all synapses of i, $p = 1$ if i is excitatory and $p = -1$ if i is inhibitory.

The percentage of inhibitory neurons in the network is determined by a constant ι, where $0 < \iota < 1$. During the initialisation of the network, i is inhibitory if $r < \iota$, where r is a random number between 0 and 1, and r is different for every neuron in the network. Typically, the higher the value of ι, the larger the number of inhibitory neurons in the network.

4.2.3 Decay and fatigue

As discussed in Section 4.1, neurons in the network leak some integrated potential when they are not firing. The network computes this and all other algorithms at discrete time steps or cycles, where each time step t is equal to 10 milliseconds in simulated time. This enables the network to gracefully ignore the chemical synaptic delay in biological neurons [Bennett and Zukin, 2004]. Decay is modelled in the network by the decay constant δ based on Equation 4.2, where A_{i_t} is the total activation of a neuron i at every time t it is not firing.

$$A_{i_t} = \frac{A_{i_{t-1}}}{\delta} \tag{4.2}$$

In contrast, a neuron fatigues when it fires. Fatigue affects the firing threshold of the neuron in such a way that it becomes more difficult for it to fire in the immediate future. When a neuron has not fired in a cycle, its fatigue level reduces so that it can fire again based on a recovery mechanism. This is shown in Equation 4.3, where θ_{i_t} is the firing threshold i at time t, θ is the threshold constant, and $F_{i_{t-1}}$ is its fatigue level at $t{-}1$. The state of the neuron is represented by σ_{i_t}, where $\sigma_{i_t} = 1$ if it fires at time t, and $\sigma_{i_t} = 0$ otherwise.

$$\theta_{i_t} = \theta + F_{i_{t-1}} \tag{4.3}$$

Equation 4.4 shows how the fatigue level F_{i_t} changes based on the neuron's fatigue level $F_{i_{t-1}}$ at time $t-1$, where f is the fatigue rate constant and f_r is the fatigue recovery constant. Also, F_{i_t} is always ≥ 0.

$$F_{i_t} = \begin{cases} F_{i_{t-1}} + f & \sigma_{i_t} = 1 \\ F_{i_{t-1}} - f_r & \sigma_{i_t} = 0 \end{cases} \tag{4.4}$$

The fatigue mechanism makes it so that active CAs eventually die out. This dynamic threshold reduces the likelihood of learnt CAs in a network remaining active indefinitely. Such a scenario may lead to undesirable effects such as simulated epilepsy, where too many neurons in the network are active, rendering it dysfunctional [Connolly and Reilly, 2005]. Besides fatigue, inhibitory neurons also aid in reducing this likelihood.

4.2.4 Activation

Neurons in the network have a critical threshold. When the integrated potential of a neuron exceeds its threshold, it fires, propagating its potential via its synapses to post-synaptic neurons. In biological neurons, threshold potentials are governed by chemical transmissions [Lodish et al., 1999]. Activation of neurons in the network is based on Equation 4.5, where A_{i_t} is the activation of a neuron i at time t when the neuron is not firing.

$$A_{i_t} = \frac{A_{i_{t-1}}}{\delta} + \sum_{j \in V_i} w_{ij} \tag{4.5}$$

The current total activation is the sum of remnant activation from the last time step $t-1$ (modified by the decay constant δ) and the output of neuron $j \in V_i$, where V_i is the set of all firing pre-synaptic neurons connected to i. This is weighted by the connection strength w_{ij} between i and j. The neuron fires at $t+1$ if the accumulated activation A_{i_t} exceeds its threshold θ_{i_t+1}. Upon firing, $A_{i_t} = 0$, when the neuron sends its activation to all connected post-synaptic neurons, losing all potential in the process. Fatiguing causes θ to be dynamic as described in Section 4.2.3.

4.2.5 Learning

Learning in the network is based on a correlatory Hebbian learning rule [Huyck, 2004] that is derived from Hebb's original postulate. If a pre-synaptic and post-synaptic neuron fire together, their synaptic weights get adjusted so as to increase their mutual synaptic efficacy. On the other hand, if only the pre-synaptic neuron fires, it experiences reduced synaptic efficacy. This is known as anti-Hebbian learning. These two rules together act on two neurons i and j in such a way that their mutual synaptic weights are adjusted based on the frequency of them firing together, improving the likelihood of the post-synaptic neuron j firing in response to the pre-synaptic neuron i in the future. In addition, a compensatory rule modifies the two learning rules so as to make learning in the network balanced. Equation 4.6 and 4.7 show the Hebbian and anti-Hebbian learning rules respectively, where w_{ij} is the total synaptic strength of the synapse from neuron i to j, χ is the axonal median constant affecting the synaptic strength, and λ is the learning rate.

$$\Delta^{+} w_{ij} = \mu_{ij}^{+} \left(1 - \frac{w_{ij}}{\chi}\right) \lambda \tag{4.6}$$

$$\Delta^{-} w_{ij} = \mu_{ij}^{-} \left(-\lambda w_{ij}\right) \tag{4.7}$$

Equation 4.8 shows the strengthening compensatory modifier μ_{ij}^{+} and Equation 4.9 shows the weakening compensatory modifier μ_{ij}^{-}. C is an arbitrary constant ($C = 1.3$) that has been observed to work well across different simulations [Huyck, 2000, 2007; Huyck and Nadh, 2009], w_i is the total connection strength of i, and B is the saturation base which is the desired total connection strength from a neuron.

$$\mu_{ij}^{+} = C^{(B-w_i)} \tag{4.8}$$

$$\mu_{ij}^{-} = C^{(w_i-B)} \tag{4.9}$$

The strengthening and weakening compensatory rules modify the Hebbian and anti-Hebbian learning rules respectively. They try to retain the overall synaptic connection strength of i close to B by increasing it if it is too low and decreasing it if it is too high. This keeps neurons with low synaptic strengths, for instance in a CA, from getting isolated from neurons with high mutual synaptic strengths. Similarly, it keeps neurons with high mutual synaptic strengths from becoming too dominant by counterbalancing their strengths. This ensures balanced activity distribution across CAs formed in the network. Such a mechanism is neurobiologically plausible, as biological neurons have constrained capacities for storing and transmitting potential [Huyck, 2007].

4.3 Interaction of CAs in the network

CAs by nature are reverberating neural circuits. Such circuits form in the neural network via unsupervised learning. That is, neurons self-organise into CAs based on how stimuli activate them without any explicit instructions on how to achieve a target state. This assembly process is driven by the previously described Hebbian learning rules. While the CAs that form in the network have a faithful resemblance to biological CAs, they are extreme simplifications. For instance, while an average neuron in the brain may have several thousand connections to other neurons, a neuron in the network may only have a few hundred connections or fewer based on the configuration.

When two connected neurons in the network co-fire, their synaptic weights get strengthened, increasing the likelihood of the pre-synaptic neuron activating the post-synaptic neuron in the future. While two such neurons may be a CA, a CA may also have tens of hundreds of neurons. When an external stimulus causes a few neurons of a CA to fire, the high synaptic strengths may cause other neurons in the CA to fire and in turn activate the whole CA. This activity may cascade around the CA and cause it to reverberate even after the external stimulus ceases. In the network, this transient reverberation may die out due to fatiguing neurons or because of inhibitory neurons. This is necessary, as neurons should not enter a biologically implausible state of perpetual activity. Moreover, neurons may need to participate in different CAs for performing different tasks. A visualisation of reverberation of a CA in a 40 × 40 network is shown in Figure 4.2.

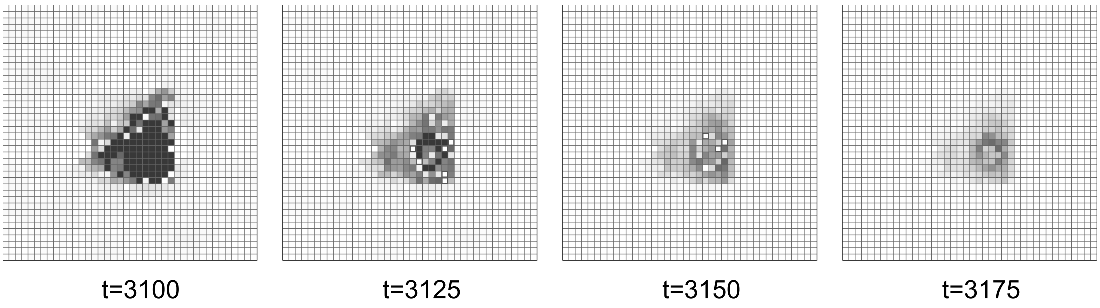


Figure 4.2: Visualisation of activity in a reverberating CA

Each cell in the grid is a neuron whose activation level is indicated by the strength of its shade of grey, where grey and white cells are non firing neurons, and black cells are firing neurons. The network was presented with a stimulus for 3000 cycles, where 50 random neurons of a total of 200 neurons in the centre of the network were externally stimulated. These 50 random neurons changed every cycle. External stimulation involves manually setting the activation levels of neurons to values exceeding their firing thresholds so they fire immediately. At the end of 3000 cycles, the external stimulus was removed. The CA formed as a result of the co-firing of these neurons can be seen to have the most activation at $t = 3100$. The CA reverberates for roughly 200 cycles, with its activation level peaking at $t = 3100$, and waning thereafter with each subsequent step. The overall activation of the CA is shown in Figure 4.3.

Equation 4.10 gives the activation level of a CA at time t, where a_{i_t} is the activation of neuron $i \in V_t$, V_t being the set of all neurons firing at t in the reverberating CA.

$$A_t = \frac{1}{|V_t|} \sum_{i \in V_t} a_{i_t} \tag{4.10}$$

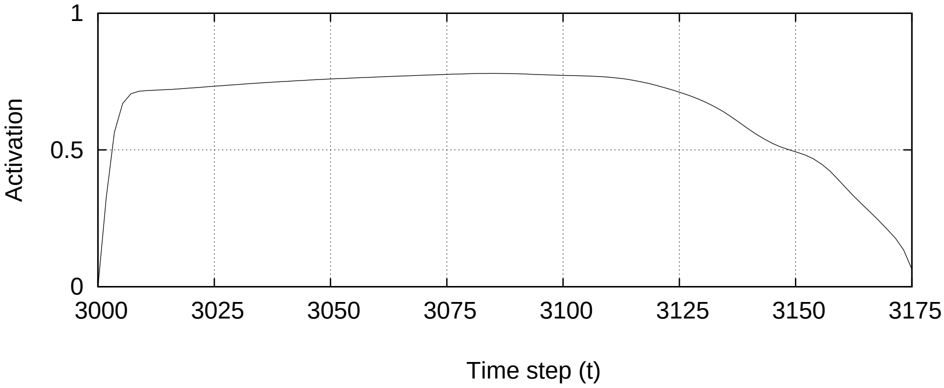


Figure 4.3: Activity of a reverberating CA in the FLIF neural network

When a CA is active, it may in turn activate other CAs that have synaptic connections or share neurons with it. This is a basic associative memory mechanism in CAs. Since neurons in the network have distance biased connectivity, in larger networks, CAs physically distant from each may still share neurons. In some cases, the initially active neurons of a CA may not be strong enough to activate the whole CA, and the partial reverberation may die out quickly. Such partial activation of a CA may be a pseudo-stable state, where the activity represents a part of the information the CA encodes. For example, partial activation in a CA encoding a complex shape may represent a specific feature of the shape, such as a curve.

As discussed in Section 3.2.5, CAs compete with each other for activation via mutual inhibitory connections. Figure 4.4 shows the activation levels of two competing CAs A and B in a 20×20 network (normalised to the range $[0, 1]$). These were recorded over four different random network configurations to show the uniform competitive behaviour under different configurations. The two CAs were allowed to form over 3000 cycles by externally stimulating the first 200 and the last 200 neurons alternatingly. As a result, neurons in the two halves of the network assembled into competing CAs. At $t = 3000$, training was stopped and twenty random neurons of each CA were externally stimulated for 100 cycles. These neurons propagated activity to other neurons, activating the whole CAs. They continued to reverberate for roughly 100 cycles after the stimulus was removed. While reverberating, both CAs competed for activation by trying to inhibit each other, gradually dying out from neuronal fatigue and inter-CA inhibition.

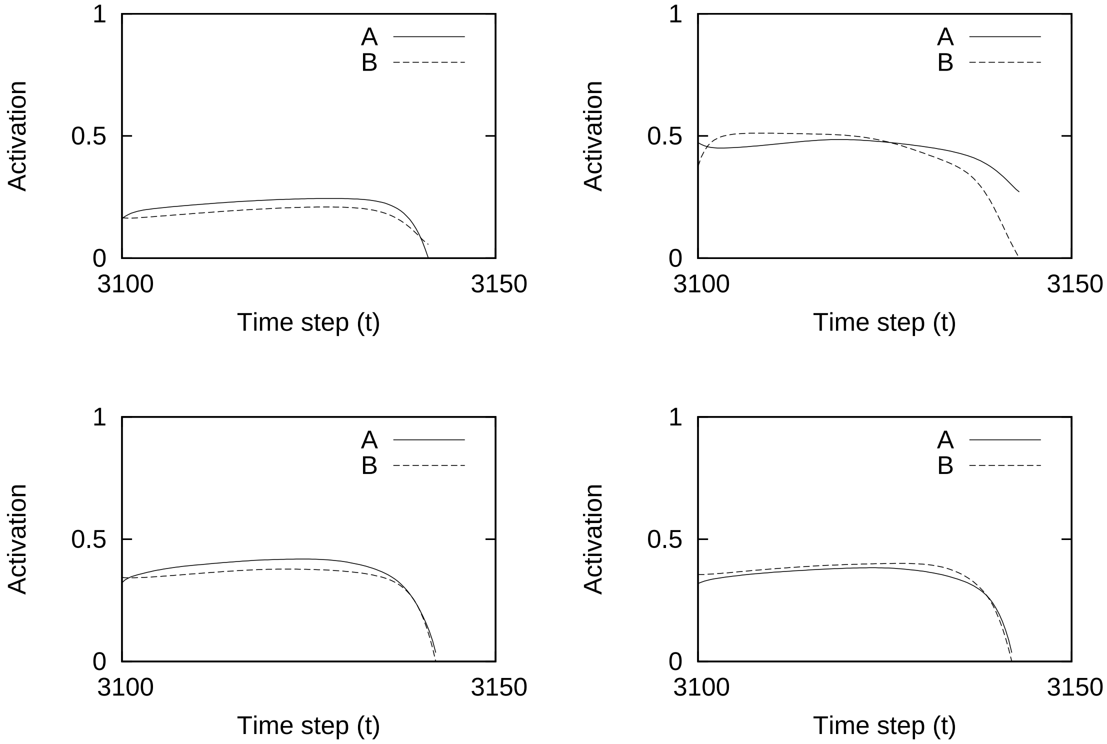


Figure 4.4: Activity of two competing CAs

4.4 Chapter summary

This chapter discussed in detail, the FLIF ANN used for modelling CAs in simulations described in this thesis. It discussed various properties of the FLIF neuron along with its mathematical and computational details, and described the learning algorithms by which the ANN models CAs. It described the topology of the ANN, and discussed the formation and interaction of CAs within the network. The next chapter presents in detail, the various associative memory models developed in this thesis.

CHAPTER 5

CAs at work: models and tasks

CAs are considered to be the neural basis of many mental processes. As such, computational models of CAs offer dynamics that other systems do not. The previous chapter described such a computational model—an ANN framework based on FLIF neurons capable of modelling relatively neurobiologically faithful CAs. Prior chapters discussed various background concepts leading to the work described in this chapter, from the basics of human associative memory to the neurobiology and behavioural characteristics of CAs. This chapter describes a series of associative memory models developed in the FLIF ANN. The models explore various properties of CA based associative memory and the higher order phenomena emerging from them.

Each model described in this chapter explores increasingly complex forms of CA based associative memory, namely, emergent context sensitivity from an associative memory of CAs with synaptic associations; spatial cognitive mapping supported by the sequential activation of CAs in an associative memory of CAs with synaptic associations; novel behaviour from emergent generalisation in an associative memory of continuous overlapping CAs; natural language disambiguation supported by emergent relationships of semantic data in a large associative memory of overlapping CAs. In that respect, they are quite distinct from one another. However, processes they model are unified to a certain extent, as they stem from the same set of mechanisms.

It seems worthwhile to mention that in the course of this thesis, many other CA based models were developed—A CA based content addressable memory based on the classic *Jets and Sharks* model [Rumelhart and McClelland, 1982] that explored learnt prototypes [Huyck and Nadh, 2009]; a simple image recognition tool based on physiologically realistic rudimentary retinal cells, and a simplified version of line and angle detection in the visual cortex region of the brain; and several simple categorisation tasks based on many to many associations from synaptic and overlapped encoding of CAs. These simple models aided the development of the models described in this chapter, which are summarised below.

Emergent context sensitivity in an associative memory

Inspired by Pavlov's [1927] conditioning experiment, the model shows how context sensitivity can be inherent in an associative memory. It demonstrates that learnt CAs representative of different external stimuli can behave differently in different contexts. In the model, a hypothetical dog reacts differently to *food* depending on the context—it *salivates* when *hungry* and *lies down* when *not hungry.* Even though these concepts are learnt in the same manner, the interplay of CAs representing them gives rise to context effects.

Spatial cognitive mapping with a sequential associative memory in an embodied agent in a virtual environment

The model shows an embodied agent in a virtual 3D environment capable of simple spatial cognitive mapping. The CA based agent explores the environment while learning features such as rooms, doors, and door colours as sequential associations via visual cues. The agent can then navigate to a particular room based on the learnt landmarks. The model shows that complex phenomena such as spatial cognitive mapping can be accounted for by the dynamics of CAs in an associative memory.

Novel behaviour from a massively overlapping associative memory in a game playing agent

The model is a virtual agent capable of learning and playing the classic game of *Pong*. The agent learns the game either by “watching” a human play, or autonomously from environmental feedback. Game moves are encoded in a massively overlapping associative memory. It is then able to perform novel moves in the game based on generalisation effects that emerge from its associative memory.

Natural language disambiguation with an associative memory of semantic hierarchies

A natural language processing model that disambiguates prepositional phrase (PP) attachment ambiguity in the English language is described. The model uses a large overlapping associative memory of semantic hierarchies of sentences extracted from a corpus to disambiguate novel sentences. It attains a resolution accuracy on par with many known machine learning systems that perform the same task.

5.1 Emergent context sensitivity in an associative memory

The model was conceived as a result of research into types, tokens, and the nature of associations that may be present in an associative memory. Types are abstract memories where as tokens are instantial episodic memories. While there are many computational models of associative memory, such properties are usually ignored. Early versions of the model were attempts at emulating explicit types of associations, but it turned out to be difficult to find low level neurobiological and psychological bases for the process. After many fruitless iterations, further research indicated that context is a process that affects memory association and retrieval extensively, causing memories to behave differently under different circumstances [Maratos et al., 2001; Nieuwenhuis et al., 2005]. While there is extensive research studying high level effects of context, the low level neural mechanisms underpinning the process are not well understood. Subsequently, context modelled as CAs was attempted.

The model was devised to contain elements of Pavlovian conditioning to test if context could be modelled via conditioned learning of associative memories, like their high level behaviour suggests. While emergent effects were not expected, the resultant model saw the implicit emergence of context from simple associative learning. This is an interesting prospect, as without the need for explicitly modelling context, it emerged as a by product of CA formation and association, processes with strong neurobiological backgrounds. The psychological process of context is briefly discussed in the following subsection (Section 5.1.1), and the model in the subsequent subsections.

5.1.1 Context

Context is the implicit knowledge available in a situation that facilitates wider awareness and perception of the situation, beyond what is explicitly available. This is evident in language. For example, in human conversations, communication is based on implicit situational information, eliminating the need to explicitly describe various facets of a topic. Symbols in the environment can also have different semantics in different contexts. Thus, context reduces the need for redundancy and facilitates efficient information exchange and processing. Context is also a low level process, where activity of neural circuits in the brain vary under different conditions. For instance, activity of reward sensitive regions in the human brain vary based on the context in which they are activated [Nieuwenhuis et al., 2005]. Memory retrieval is also heavily influenced by context, for example, influence of emotional contexts on episodic memory retrieval [Maratos et al., 2001]. Similarly, a concept may have different semantics in different contexts. For instance, neural representations of words in the brain activate differently depending on the context they are perceived [Federmeier et al., 2000]. That is, the activity of a CA may be influenced by the activity of associated CAs depending on different factors. There is evidence suggesting that neural representations of context may be conjunctive [O'Reilly and Rudy, 2001; Rudy and O'Reilly, 2001]. How such context dependent neural behaviour in turn produces high level context sensitivity remains to be understood well.

The model described in this section, inspired by the Pavlovian conditioning experiment [Pavlov, 1927], explores how context sensitive behaviour can emerge from the underlying characteristics of CAs in forming associative memories.

5.1.2 The model

The model represents a hypothetical dog that is conditioned similarly to Pavlov's experiment. It drives the formation of a simple associative memory with five CAs representing an object, two states, and two actions—*food*, *hungry*, *not hungry*, *salivate*, and *lie down* respectively. The model learns these concepts individually, and then the associations between them. It then sees differential activation behaviour of actions for the same object depending on the active state or context. That is, while *food* is equally associated to all other CAs, activating *food* in the context of one of the two state CAs, and in the absence of a context, produces completely different behaviour in the model. The specifics of the model and the simulation are described in the following sections.

5.1.3 Network properties

The model consists of a single 50×40 subnet named *Context*. Table 5.1 shows the network parameters (described in Chapter 4) for the model. Each of the five CAs to be learnt are designated 400 consecutive neurons in the network.

Table 5.1: Context model's network parameters

		Context
Learning rate	λ	0.05
Threshold	θ	3.5
Axonal median	χ	0.5
Fatigue	f	0.4
Fatigue recovery	f_r	0.8
Decay	δ	1.2
Saturation	B	18
Inhibitory neurons	ι	0.2
Neurons	N	50×40

5.1.4 Simulation

The 400 neurons designated to each of the five concepts—*food*, *hungry*, *not hungry*, *salivate*, and *lie down*—are externally stimulated for 300 cycles each consecutively. External stimulation involves choosing a random 50% of the neurons designated to the concepts every cycle, and setting their activation levels to values $> \theta$. Co-firing of these groups of neurons lead to the formation of CAs representing the concepts. When a concept is being learnt, its CA sends inhibition to other regions of the network via its inhibitory connections. The distance biased topology of the net ensures that the CA is able to affect a large area of the network. Thus, initially, all individual CAs inhibit each other.

After individual concepts have been learnt over 1500 cycles, two different scenarios resembling the behaviour of a hypothetical dog that *salivates* on being presented with *food* when *hungry*, and *lies down* when presented with the same *food* when *not hungry*, are learnt. This is done by choosing a random 50% of the neurons of the three CAs—the object CA, the action CA, and the state CA—simultaneously every cycle, and externally stimulating them for each of the scenarios in two stages. The stimulation for each stage lasts 300 cycles, and is as follows:

1. *food + hungry + salivate*
2. *food + not hungry + lie down*

During this two stage learning over 600 cycles, the co-activation of these CAs increases the mutual synaptic strength of inter-CA neurons, eventually associating them. As a result, when a CA is activated in the future, it may pass activation to other CAs it is associated with, in turn activating them. During co-activation, the three active CAs inhibit the two inactive CAs via mutual inhibitory connections. Since *food* is active in conjunction with all other CAs, it forms both excitatory and inhibitory associations with them.

After learning, the network is tested by externally stimulating the object CA in context of the state CAs for 100 cycles each—*food + hungry* and *food + not hungry.* The resultant activity in the network is discussed in the next section.

5.1.5 Results and discussion

When *food* and *hungry* are activated, they propagate positive activation via learnt excitatory connections to *salivate.* At the same time, they have some excitatory connections, but many learnt inhibitory connections to *not hungry* and *lie down.* As a result, *salivate* becomes fully active and reverberates, immediately suppressing any activity starting in *not hungry* and *lie down.* Similarly, *food* activated in context of *not hungry* fully activates *lie down*, that in turn suppresses any activity in *salivate* and *hungry.*

Figure 5.1 shows the state of the neurons designated to different concepts in the network before learning (A) and after learning (B). Before learning, physical synaptic connections between neurons in the network are distributed, following the distance biased topology (discussed in Section 4.2.1), and have random strengths as shown by solid arrows. However, after learning, while physical connections

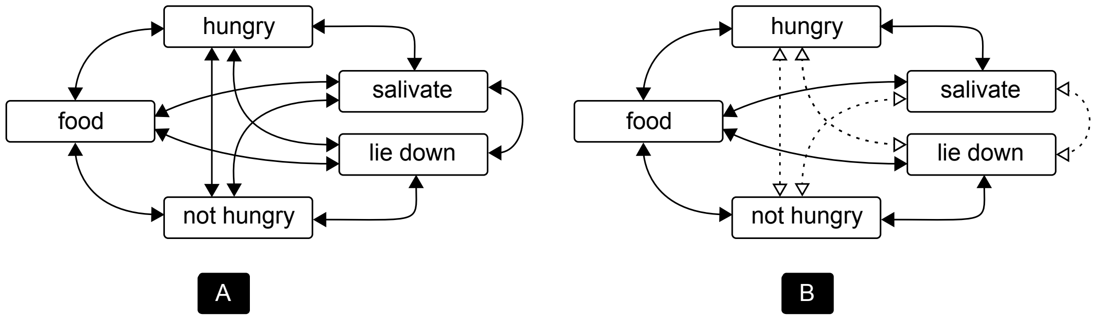


Figure 5.1: Initial and learnt states of CAs in the context network

remain the same, their synaptic efficacies change leading to formation of individual CAs and associations between them. The solid arrows in *B* show the strong excitatory associations and the dotted arrows show weak associations that collectively have an inhibitory effect. These associations are formed naturally via the dynamics of Hebbian learning.

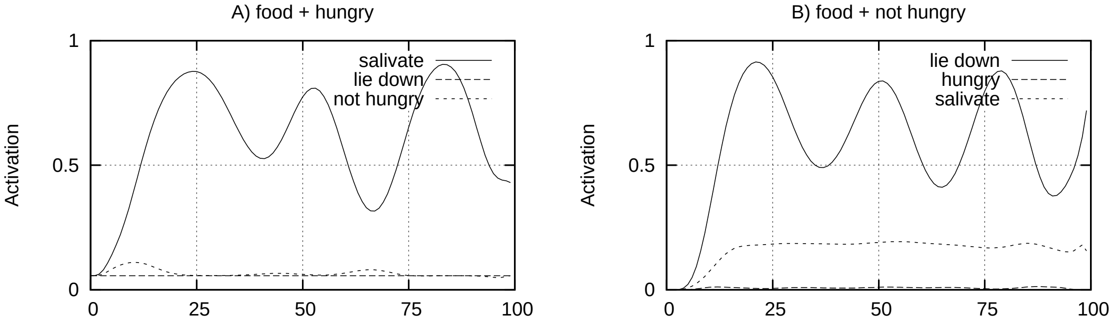


Figure 5.2: Activation levels of CAs when *food* is activated in different contexts

Figure 5.2 shows an example of the activation levels of reverberating CAs during the two tests. *A* shows *salivate* quickly rising and suppressing *lie down* and *not hungry* when a random 50% of the neurons of *food* and *hungry* are externally stimulated. Similarly, *B* shows *lie down*'s activation rising and *hungry* and *salivate* when *food* and *not hungry* are externally stimulated.

Figure 5.3 shows the activation level of CAs over 100 cycles when *food* alone is externally stimulated. Since *food* has excitatory associations with all other CAs, it propagates activation to them, causing them to become active and reverberate. These CAs oscillate, competing with each other for activation. That is, without an active context, when *food* alone is active, the model behaves in an entirely different way.

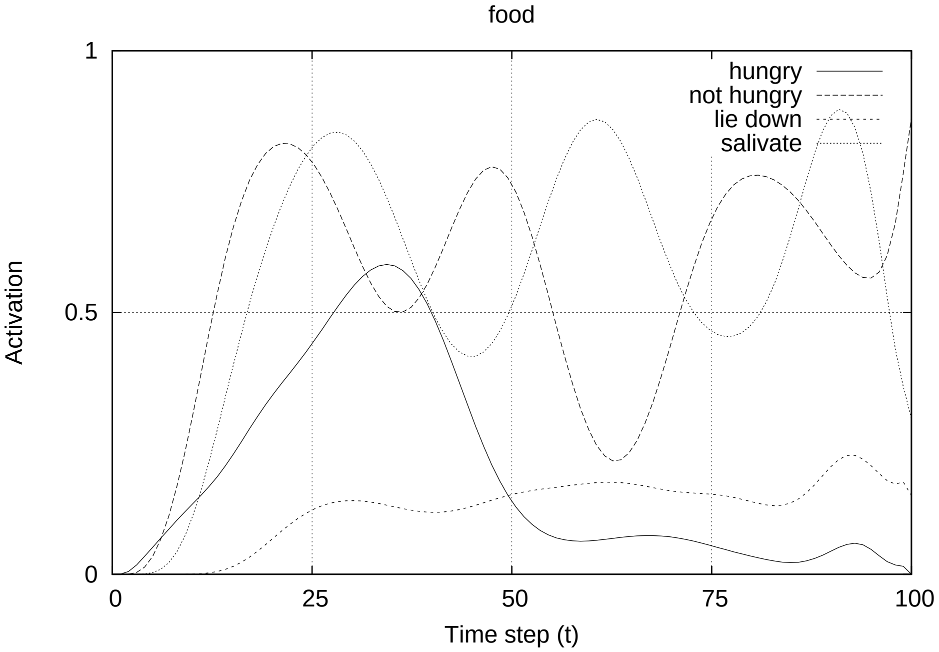


Figure 5.3: Activation levels of CAs when *food* alone is activated

In essence, the behaviour of the model differs radically under different contexts and when no context is present. This behaviour is inherent in the underlying dynamics of CA formation and association. Without explicitly encoding context, it emerges implicitly from the associations between concepts. Similar to context in the real world, learnt concepts in the associative memory act as context under different conditions. The results from this model are novel, as it shows how a simple CA based associative memory may implicitly give rise to context effects without the need of any explicit external mechanisms. It shows how a group of concepts with many to many associations can implicitly develop contextuality, a fundamental process underpinning many higher order phenomena.

While the model is an extreme simplification of the actual neurobiology of context sensitivity, the activation behaviour shown by the CAs in the model resemble the variable activity of certain brain regions influenced by context [Nieuwenhuis et al., 2005; Wagner et al., 1998]. Context also plays an important role in efficient memory retrieval by helping differentiate between different items in memory. Without context, humans often find stimuli ambiguous, for example, the semantics of certain phrases. However, provided context, disambiguation is faster [Gennari et al., 2007]. The activity in the model accounts for this behaviour. When *food* alone is active, it activates multiple CAs it is associated with, but when a context CA is active in conjunction, the subsequently active CA is more relevant. The behaviour of the model reveals some interesting dynamics of context that may be considered for further research on context effects in associative memories.

5.2 Spatial cognitive mapping with a sequential associative memory in an embodied agent in a virtual environment

The spatial cognitive mapping model is an evolved version of a much simpler earlier model that involved an agent in a virtual environment with two rooms. The impetus of the original model was the idea of sequential memories in human associative memories. A task was required for the purpose, and spatial cognitive mapping was chosen, as aspects of it are considered to involve sequential memories, and more importantly, CAs [Dragoi and Buzaski, 2006; Tanila et al., 1997]. Spatial cognitive mapping is a term for many processes [Foo et al., 2005; Nardini et al., 2008; Spiers and Maguire, 2008], of which, one is landmark based navigation. Landmark based navigation is thought to involve sequential activation of episodic memories concerning landmarks in an environment [Buzsáki, 2005; Dragoi and Buzaski, 2006; Pastalkova et al., 2008]. The simulation was devised in such a way that both sequential memories and landmark based navigation could be combined into a single model.

The agent and the virtual environment was inspired by the *CABot* project—a fully neural embodied agent capable of complex vision, natural language parsing, goal setting, and spatial navigation [Huyck et al., 2011]. A version of the spatial cognitive mapping model described here has been incorporated into the CABot3 agent [Huyck et al., 2011].

For creating the 3D environment, *IRRLICHT* [Gebhardt et al., 2009], a cross-platform 3D engine was used. The behaviour of the completed model was found to resemble certain low level neurobiological processes involved in the brain, providing some insight into their nature.

5.2.1 Spatial cognitive mapping

Spatial cognitive mapping is the psychological process by which an individual acquires, stores, and recalls information pertaining to surroundings such as physical features, relative routes, and landmarks for the purpose of interacting with the environment [Downs and Stea, 1973]. It is generally used in path finding and environmental navigation. Many biological beings—mammals, birds, and insects—use some form of this process [Bennett, 1996] to facilitate fundamental interactions crucial for survival in their environments. The ability of many animals to explore distant territories and find their way back to their nests are examples of this process.

The idea of *cognitive maps* was introduced by Tolman [1948] in his landmark paper *Cognitive Maps in Rats and Men* that describes how rats in a maze form mental maps of the maze for efficient navigation. Later, O'Keefe and Nadel's [1978] work inspired a current of neurobiological and behavioural psychology research on cognitive mapping. However, there is considerable speculation and disagreement on the neural mechanisms underpinning it [Bennett, 1996; Cruse and Wehner, 2011; Mackintosh, 2002; Sturz et al., 2006; Tversky, 1993; Wang and Spelke, 2002]. Whether spatial cognitive mapping involves the formation of explicit integrated spatial "maps" of environments is not clear. There is evidence suggesting that it may involve multiple strategies such as path integration and rather simple associative memories of environmental landmarks [Foo et al., 2005; Nardini et al., 2008; Spiers and Maguire, 2008].

The involvement of the hippocampus region of the brain and its specialised *place cells* or *grid cells* that encode spatial locations that facilitate spatial cognitive mapping is well studied [Doeller et al., 2010; Dragoi and Buzaski, 2006; Leutgeb et al., 2005; Tanila et al., 1997]. However, spatial navigation may not necessarily need complex map like representations, but may be based on simple associative learning of environmental landmarks [Foo et al., 2005]. For example, an individual new to a city may not initially memorise routes, but may navigate by relying on easily recognisable landmarks like buildings. There is evidence suggesting that sequential activation of CAs in brain regions involved in navigation plays an important role in this process [Buzsáki, 2005; Dragoi and Buzaski, 2006].

This section describes an embodied agent in a virtual environment that demonstrates an extreme simplification of landmark based navigation behaviour. The agent learns landmarks in a 3D environment and associations between them as a simple CA based sequential associative memory. It is then able to perform basic navigation by recognising previously learnt landmarks, resembling similar behaviour in humans [Foo et al., 2005]. The resemblances of the model with neurobiological and psychological processes are discussed in Section 5.2.6. The model serves as a proof of concept for the phenomenon and explores how such a complex process may be supported by a relatively simple associative memory.

5.2.2 The model

The agent exists in a virtual 3D environment composed of four interconnected rooms. The rooms each have a colour coded door—yellow, blue, red, and green respectively. Figure 5.4 shows a snapshot of the virtual environment. The agent has a primitive visual system capable of detecting doors, colours, and open spaces. It explores the environment by moving from room to room, learning in the process, doors and their colours, association between doors and rooms, and association between different rooms. That is, $(room - door[colour])_m - room_n$ episodes. Once learnt, the agent is able to navigate to a particular room by "recollecting" the colour of the target door relative to its current location. This is enabled by the sequential activation of CAs in its associative memory of learnt landmarks. The agent is only able to move in straight lines and make left turns in the environment

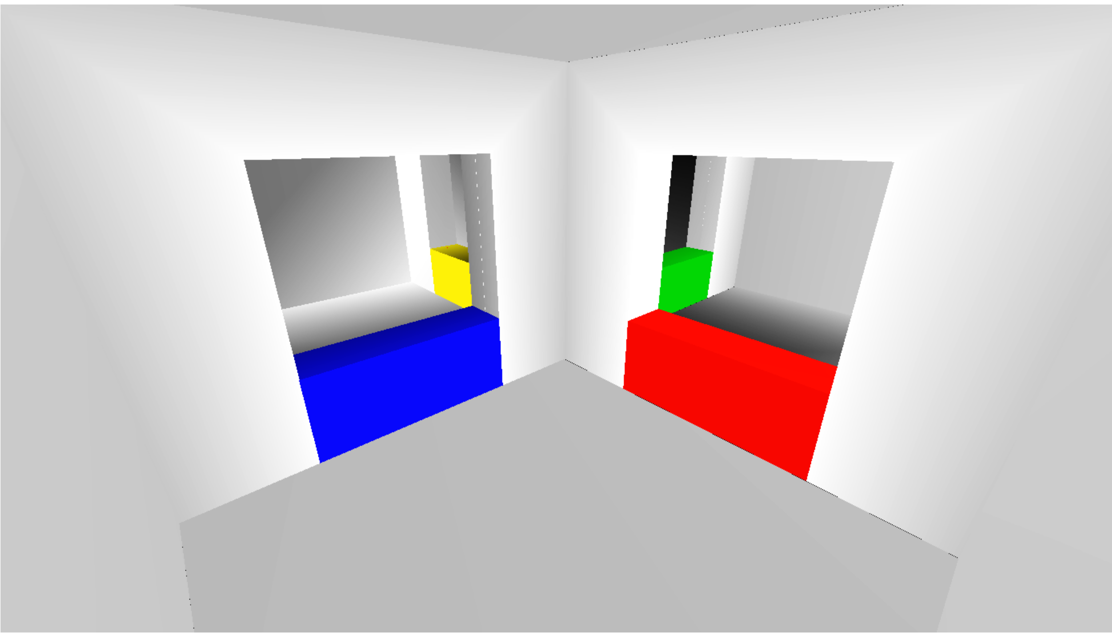


Figure 5.4: Spatial cognitive mapping agent's 3D virtual environment

for the sake of simplicity, as the model focuses on the associative memory aspect of spatial cognitive mapping and not properties of physical navigation.

5.2.3 Network properties

The model has a large network that is divided into smaller subnets for modularity—*Colour* that encodes colours recognised by the agent; *Object* that encodes the landmarks recognised by the agent; *Door* that encodes the doors encountered by the agent; *Room1* and *Room2* that encode rooms the agent visits; *Episode* that encodes the associations between doors and rooms as different episodes; and *Goal* that encodes the target room in a navigation task.

Blocks of 200 neurons are designated for CAs to be learnt representing rooms, doors, and episodes in their corresponding subnets. That is, a block for each of the four colours in *Colour*, four doors in *Door*, four rooms each in *Room1* and *Room2*, and four episodes in *Episode*. While these blocks of neurons are not CAs initially, they form CAs via co-firing during the exploration task. *Object* has two blocks representing door and empty space. Apart from these subnets that are part of the agent's associative memory, there is a *Vision* subnet with neurons that act as visual receptors for the agent's vision. *Vision* is not considered to be a part of the agent's learning system. The network parameters of the subnets are given in Table 5.2.

Figure 5.5 shows the connection topology of the subnets. Arrows represent random, low weight inter-subnet connections. Double headed arrows represent to and fro connections between subnets. In such cases, two neurons in different subnets may have multiple connections with each other. Details of intra-subnet connections are listed below.

Every excitatory neuron in *Colour* connects to 50 random neurons in the block representing door in *Object* with an initial synaptic weight of 0.5. They also connect to 5 random neurons in *Door* with an initial synaptic weight of 0.001

Every excitatory neuron in *Object* connects to 10 random neurons in *Door* with a synaptic weight of 0.001

Every excitatory neuron in *Door* connects to 15 random neurons in *Colour* with a synaptic weight of 0.001. They also connect to 5 random neurons in *Episode* and *Goal* with a synaptic weight of 0.001

Every excitatory neuron in *Room1* connects to 5 random neurons in *Episode* and *Room2* with a synaptic weight of 0.001

Every excitatory neuron in *Room2* connects to 5 random neurons in *Episode* and *Room1* with a synaptic weight of 0.001

Every excitatory neuron in *Episode* connects to 5 random neurons in *Room1*, *Room2*, and *Door* with a synaptic weight of 0.001

Every excitatory neuron in *Goal* connects to 5 random neurons in *Door* with a synaptic weight of 0.001

5.2.4 Vision

The agent is capable of primitive vision with which it “sees” limited aspects of the environment. The visual system is made of a 100 × 80 grid (*Vision*) of binary *on-off* receptors that respond to the four colours the agent recognises—yellow, blue, red, and green. *Vision* receives visual stimuli from the environment in the form of a 100 × 80 grid of pixels. Each pixel in this grid is mapped to a receptor in the corresponding position in *Vision*. If a pixel has one of the four recognised colours, its receptor switches on, otherwise, it is off. The environment also sends external stimulation to *Colour*, activating the corresponding colour CA. This is done by a symbolic module that reads the colour of individual pixels in the pixel grid,

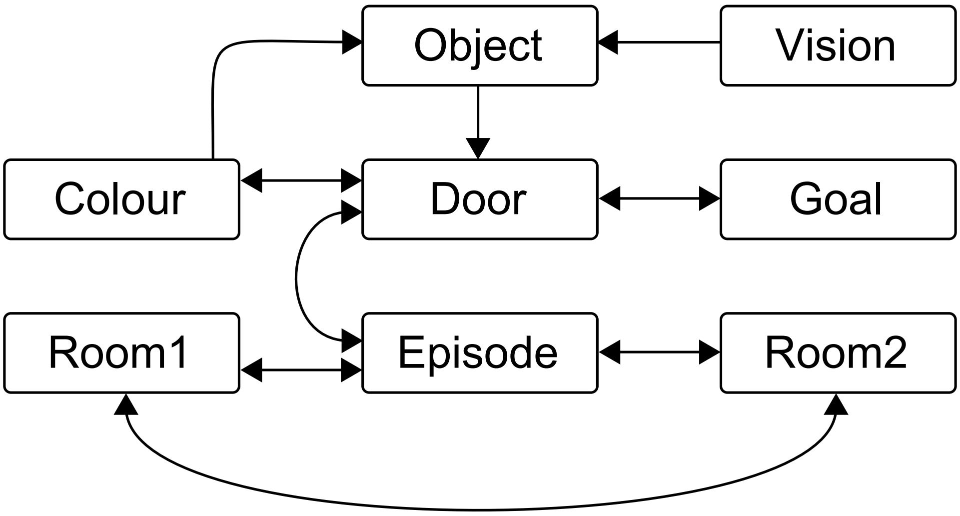


Figure 5.5: Spatial cognitive mapping model's network structure

and externally excites the neurons in *Color* corresponding to the most prominent colour. Thus, the agent is only able to recognise a single colour at a time—the most prominent colour in its visual field.

When a receptor is on, it sends out positive activation. Every receptor in the top 20 and bottom 20 rows of *Vision* connects to 30 random neurons of the 200 neurons representing door in *Object*, with a synaptic weight of .008. When the agent has a door at a distance in its view, a few receptors are on, not strong enough to cause the neurons in *Object* to fire. As it approaches the door, the image grows larger, triggering more receptors. When the agent is face to face with the door, a large number of receptors are on that excite neurons in *Object* signalling the presence of the door. Figure 5.6 illustrates how receptors in *Vision* may appear when the agent is far away from a door and when the agent is up close to a door, where black cells represent receptors that are on. Learning in the agent depends on visual feedback from the environment as discussed in the following section.

Table 5.2: Spatial cognitive mapping model's network parameters

		Colour, Object, Door, Room1, Room2, Episode, Goal	Vision
Learning rate	λ	0.1	0.0
Threshold	θ	4.5	4.5
Axonal median	χ	0.5	NA
Fatigue	f	0.3	0.0
Fatigue recovery	f_r	0.8	0.0
Decay	δ	1.2	0.0
Saturation	B	31	NA
Inhibitory neurons	ι	0.3	0.0
Neurons	N	40×20	80×100

5.2.5 Simulation

The simulation consists of two tasks, exploration and navigation. During exploration, the agent explores the environment by moving through all the rooms exactly once and returning to the starting room, learning landmarks in the process. During navigation, the agent goes to a randomly chosen room using the "knowledge" it has acquired during exploration. In both tasks, the path the agent takes is fixed. It only moves forward in straight lines and makes left turns. Figure 5.7 shows a bird's eye view of the environment and the path taken by the agent.

A symbolic module controls the agent and drives the learning processes. CAs representing different colours in *Colour* are learnt prior to the tasks via the external stimulation of a random 50% of the neurons designated to the colours every cycle, for 200 consecutive cycles. During exploration, the agent learns various features from the external stimulation of the designated blocks leading to CA formation. At the start of the exploration, the agent moves forward from whichever room it is in. *Object* signals the presence of a door when the agent encounters one,

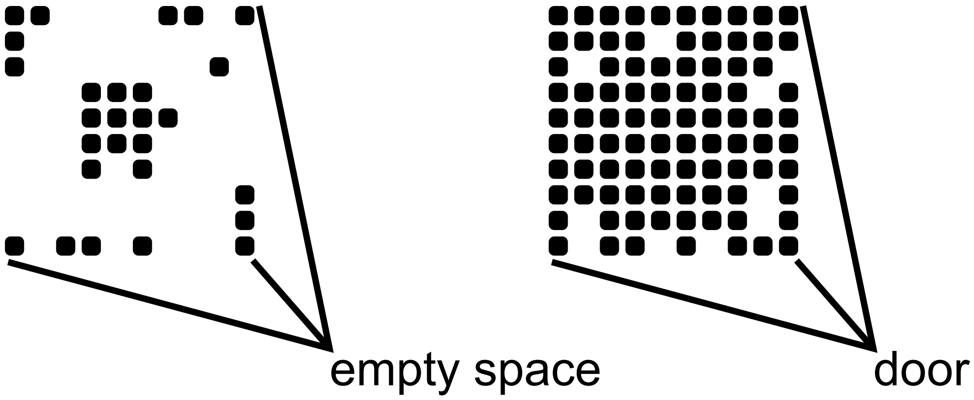


Figure 5.6: Spatial cognitive mapping agent's vision

triggering a five step learning process:

1. The current room is encoded in *Room1* and *Room2* via the external stimulation of neurons designated to the room for 300 cycles. For instance, if the current room is the first room, the first 200 neurons in *Room1* and *Room2* each
2. The door in the visual field is encoded in *Door* via the external stimulation of the neurons designated to it, similar to step 1
3. The newly learnt door CA in *Door* is associated with its colour CA representing its colour in *Colour*. While the colour CA in *Colour* is already active from environmental stimuli, the door CA in *Door* is externally stimulated for 300 cycles so that it remains co-active with the colour CA
4. The agent moves forward a few steps and when the door disappears from view, it assumes it is the next room. This new room is learnt in *Room1* and *Room2* as in step 1
5. The passage from the previous room to the present and the door connecting the two rooms is learnt in *Episode*. This is done via the co-activation of CAs

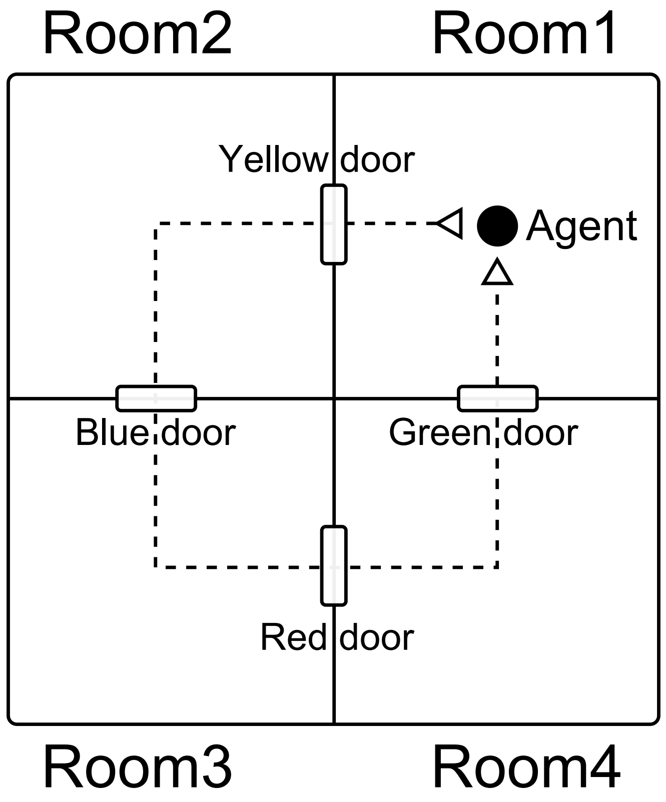


Figure 5.7: Bird's eye view of the spatial cognitive mapping agent's virtual environment

representing the previous room in *Room1*, the connecting door in *Door*, and the current room in *Room2* by external stimulating them for 300 cycles

Figure 5.8 shows an example of step 5, where individual CAs representing various features have been learnt, and the agent is learning the episode $(room_1 - door[yellow]_1) - room_2$, via the external stimulation of a random 50% of the neurons designated to them. The blackened areas represent CAs that have formed, and the dotted arrows represent associations between CAs learnt as a result of co-activation caused by the external stimulation.

Exploration ends once the agent has gone through all rooms and has returned to the starting room. By then, it has learnt all rooms, doors, and the

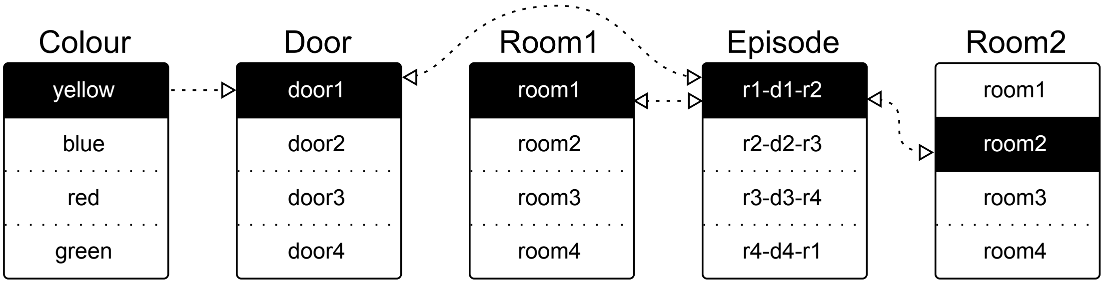


Figure 5.8: Learning in the spatial cognitive mapping model's subnets

associations between them. Since the agent's knowledge is encoded as episodes of sequences, it is possible to query its memory. For instance, after exploration, externally activating a colour CA in *Colour* activates the door CA representing the door with that particular colour in *Door*, which in turn activates the associated episode CA in *Episode*. The episode CA further propagates activation to *Room1* and *Room2*, activating the room CAs representing the rooms connected by the door. Since the agent only moves in one direction, these are rooms before and after the particular door. Similarly, activating a room CA in *Room1* activates the succeeding room CA in *Room2* via the episode CA in *Episode* that links them. The episode CA will also activate the door CA connecting the two rooms in *Door*, which subsequently activates the colour CA in *Colour*. Thus, with this simple sequential associative memory, the agent is able to serially "recollect" features of the environment, given a single landmark.

After exploration, the agent is tested with a navigation task by making it go to a randomly chosen room. This is achieved by setting a goal in *Goal*. The CA representing the target room is externally activated in *Room2*. This CA activates the episode CA in *Episode* it is associated to which subsequently activates the associated room in *Room1* and the connecting door in *Door*. The active door CA is the door that leads to the target room, that is, the goal. The goal is memorised by externally exciting 200 neurons in *Goal* for 300 cycles, so that it remains co-active with the door CA in *Door*. Thus, the goal CA is associated to the target door. That is, if the target door CA in *Door* is active in the future, it will propagate activation to *Goal* activating the goal CA, signalling attainment of the goal.

With the goal in memory, the agent moves forward, looking for the target door. On approaching a door, *Vision* signals *Colour* with the dominant colour, and the colour CA subsequently activates the door CA representing the door with that colour in *Door*. Here, the agent has recognised a previously learnt feature from an environmental cue. If the door in view is the target door, the active door CA in *Door* activates the goal CA in *Goal*, indicating the achievement of the goal. The navigation test ends with the agent moving forward until the door disappears from view, entering the target room.

5.2.6 Results and discussion

The model was tested on 30 different network configurations, where the connectivity, initial synaptic weights, and excitatory and inhibitory neurons in the network varied based on the rules described in Section 4.2. The agent correctly completed the navigation task every time. While this is not a definitive measure of the agent's abilities, it shows that landmark based spatial cognitive mapping is possible with a simple associative memory. Another mechanism involved in spatial cognitive mapping—a method of encoding an individual's relative spatial location with specialised cells in the hippocampus region of the brain—is shown in the model described in Section 5.3.

While the model demonstrates an extremely simplified version of landmark based spatial cognitive mapping in humans, it has reasonable neurobiological basis. Landmark based spatial navigation is considered to involve episodic and semantic memories [Burgess et al., 2002; Buzsáki, 2005]. The agent's associative memory and the sequential activation of CAs in it resemble the sequential activation of hippocampal CAs involved in navigation [Buzsáki, 2005; Dragoi and Buzaski, 2006; Pastalkova et al., 2008].

In the agent's associative memory, *Episode* binds different landmarks together. Every CA in *Episode* associates two consecutive rooms in *Room1* and *Room2*, the door connecting them in *Door*, and the colour of the door in *Colour* via the door itself. Every CA in *Episode* can be considered an "episode" encoding segments of the course of the agent's exploration. Although they lack temporal information, it is possible to recall a specific episode sequentially. This behaviour is consistent with one proposed mechanism of encoding episodic memories [Fortin et al., 2002]. For example, during the navigation task, when the agent is instructed to go to a randomly chosen room, it recalls the episode of having passed the room, subsequently recalling other features such as the connecting door and its colour. So, *Episode* and other subnets act as a semantic memory, as they together encode factual information about the agent's environment. Similarly, the goal CA in *Goal* that becomes active when the agent encounters a target door during navigation, resembles the behaviour of a subset of neurons in the rat hippocampus that fire when it is in the vicinity of a target goal or landmark in a navigation task [Gothard et al., 1996].

Activating a CA in any of the subnets in the agent's associative memory activates the associated CAs in other subnets via the corresponding episode CA in *Episode*. Figure 5.9 shows an example of such a case. The activity shown in the figure was recorded after the agent learnt the first episode, $(room_1 - door[yellow]_1) - room_2$. A random 50% of the neurons of the CA representing the first room in *Room1* were externally stimulated for ten cycles, activating the room CA representing the first room in *Room1*. Subsequently, CAs in *Episode, Room2, Door*, and *Colour* became active sequentially, reverberating for many cycles.

As discussed in Section 3.2.5, CAs compete with each other for activation. This is why, for instance, despite being in the same subnet, door CAs representing

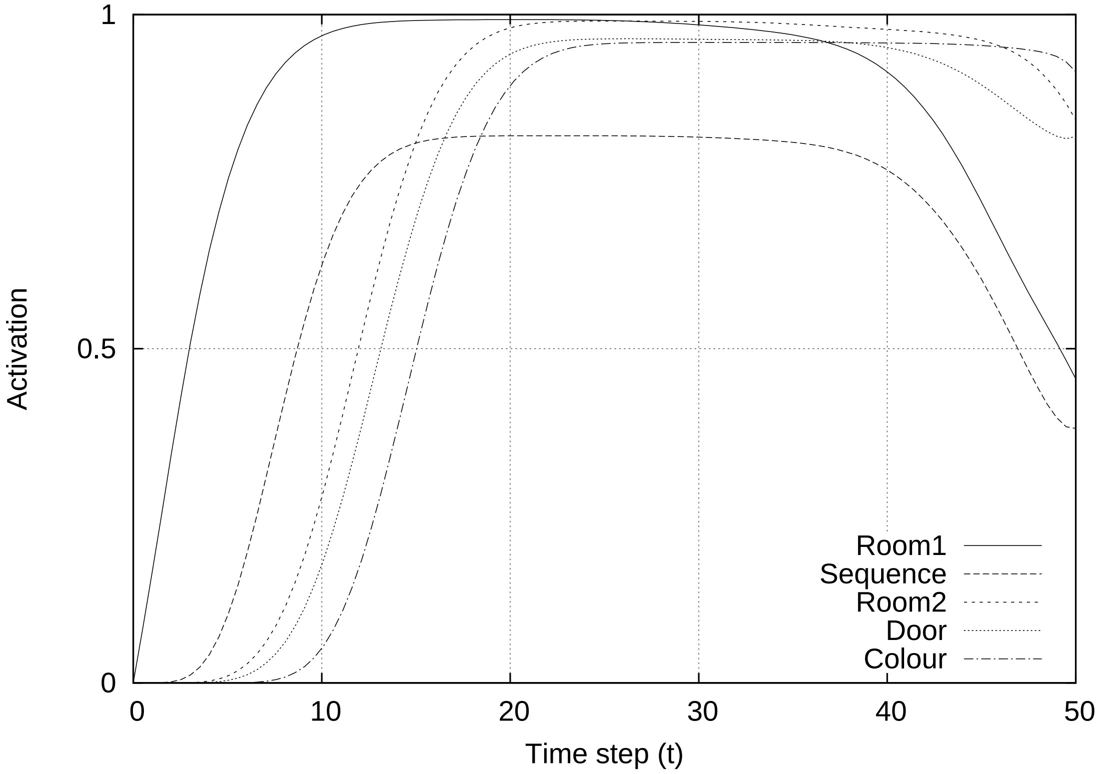


Figure 5.9: Activation of CAs in different subnets in the spatial cognitive mapping model

two different doors do not usually activate each other. While the CAs can activate each other, during learning, they are learnt in isolation. Without association via co-activation or physical overlap of neurons, their inhibitory connections become stronger. In the model, CAs are designated discrete groups of neurons in different subnets so as to drive learning and association in particular ways to ensure the desired behaviour. However, this approach is biased and purely methodological. If the agent was moved to an environment composed of ten rooms with doors of recurring colours, it would fail to navigate properly. Since door colours are the only unique environmental cues the agent relies on, to encode features and create a sequential "map" of its environment, multiple doors with the same colour would activate multiple door CAs in *Door*, and in turn, multiple episode CAs in *Episode*, and room CAs in *Room1* and *Room2*. In such a scenario, the agent would fail to distinguish between different doors and fail to identify its present location or a target location correctly. Still, learning concepts and their associations in the model are neurobiologically plausible.

In conclusion, a better design driven entirely by environmental cues and a scalable network for memorising a large number of landmarks may yield an agent with better navigational capabilities. Such a model is currently under development, and is briefly discussed in Section 6.2. The model has also been applied to a more sophisticated, fully neural agent with complex vision, shape and object recognition, goal planning and natural language parsing [Huyck et al., 2011]. In this system, the agent's environment has distinctive landmarks such as pyramids and stalactites, and the spatial cognitive mapping model works in conjunction with these features. This shows the adaptability and the viability of the model in embodied virtual agents. Furthermore, it serves as a proof of concept exploring some of the processes involved in spatial cognitive mapping. While the subnets in the model are unlikely to correspond to the brain regions involved in spatial cognitive mapping, the low level neural behaviour of CAs is fairly consistent with their observed behaviour. More importantly, it demonstrates the versatility of the computational CA model.

5.3 Novel behaviour from a massively overlapping associative memory in a game playing agent

The previous models explored properties of associative memories formed via synaptic associations of CAs. This model was developed for exploring the dynamics of physically overlapping CAs (discussed in Section 3.2.4), as overlapped encoding is an important mechanism by which associative memories are formed. Further research indicated that overlapping CAs may support generalisation [Shohamy and Wagner, 2008]. An interactive agent was chosen for the model, as such a system would involve continuous, real time behaviour. This would enable testing if generalisation could affect behaviour in the model dynamically—an important characteristic of real world agents.

A minimalistic version of the classic arcade game of *Pong* was chosen for its simplicity. The game is an abstraction of the real world table tennis game that involves moving a paddle in the game environment so as to hit an approaching ball. In the model, the paddle's movements are locked to the Y-axis. Hence, the game dynamics could be faithfully reduced to learning gameplay as position variables. In terms of encoding the game moves on the neural network, the positions of the paddle relative to the ball are encoded as CAs. This enables the agent to recollect the position the paddle should be at a particular instant relative to the position of the ball in the environment, so as to repeat the previously encoded move, thus playing the game. Subsequently, it was learnt that the spatial positions of the paddle and the ball in the game could be represented in a way that resembles the behaviour of place cells [Doeller et al., 2010; Dragoi and Buzaski, 2006; Leutgeb et al., 2005; Tanila et al., 1997]—specialised cells in the brain that are thought to be involved in associative memory, and in particular, representing spatial positions.

To be able to play the game, the agent has to first learn the game moves. It was decided that instead of explicitly programming the moves into the agent, it would learn from the environment. For this purpose, two training modes—human training and self training—were devised in such a way that learning and game playing in the agent would be continuous processes, resembling biological agents. Whether massively overlapping CAs could be encoded in a small network, whether generalisation could be modelled in such a memory, and the constraints that may affect the process were unknown.

Initially in the model, formation of overlapping CAs were outside of the game constraints, where large CAs formed due to massive overlaps. These large CAs produced continuous recurrent activity, rendering the agent incapable of learning properly. Eventually, the network parameters were adjusted so that formation, overlapping, and inhibition were suitable for the game. Like the previously described models, these parameters were derived via manual parameter exploration. While the lack of a set of unified network parameters that work for many different models is a drawback, these models serve as proof of concepts for the complex processes that can be modelled with neurobiologically faithful CAs.

Although arrived at somewhat arbitrarily, the goals of the model are multifold: to model an associative memory of physically overlapping CAs, to explore generalisation effects in such overlapping memories, to see if the behaviour of hippocampal place cells can be emulated in CAs, and to model novel behaviour wholly from an associative memory in an interactive agent. The model is discussed in detail in the following sections.

5.3.1 Generalisation as a cognitive process

The ability to generalise from past events to adapt to novel situations is key to intelligence. For instance, generalisation is involved in drawing analogies between different objects [Gentner, 1983], acquisition of language [Goldberg, 2006], and the more abstract process of *binding* [Malsburg, 1995] which concerns how different concepts are segregated and bound for coherent interpretation. The role of generalisation is commonly exemplified in visual perception and recognition. In vision, stimuli produced by an object in the environment viewed from different angles or viewed under different illuminations can be considerably different. Nevertheless, the brain is not only able to recognise the particular object, but novel objects that share similarities. This may not be possible if recognition merely involves a direct comparison of encoded stimuli with environmental stimuli, as environmental stimuli are rarely the same. Instead, such processes are underpinned by generalisation and abstraction of previously encoded stimuli [Bunge et al., 2003; Tervaniemi et al., 1994]. In the case of vision, learnt features are thought to be distributed across overlapping regions [Tanaka, 1996]. While the high level behaviour of generalisation in various contexts is well studied, the low level neural mechanisms underpinning it are not well understood.

Section 3.2.4 described how neuron sharing between CAs—physical overlapping of CAs—is a method of encoding associative memories. Subsets of neurons in a CA encode different features of the concept represented by the CA [Osan et al., 2011] and thus, neural overlaps between CAs may represent features shared between them. The extent of such overlaps may depend on the similarity of features they encode [Wickelgren, 1999]. There is evidence for such large scale overlaps in the representation of memories in the brain [Haxby et al., 2001; LaBar et al., 1999]. Hence, overlapped encoding is thought to be a mechanism underlying generalisation [Shohamy and Wagner, 2008].

This section describes a model—a simulated agent that learns to play a single player version of the classic arcade game of Pong [Kent, 2002]—that encodes information as an associative memory of massively overlapping CAs, with which it is able to adapt to and play in previously unseen scenarios. The method of encoding is similar to that of hippocampal place cells. The results are novel and provide insights into how overlapped encoding of information in the brain may support generalisation, and how novel behaviour can be produced by processes that implicitly emerge from an associative memory.

5.3.2 The model

The model is a virtual agent that learns to play a game of Pong by “watching” a human play or autonomously from environmental feedback. Learning in the agent involves encoding game moves in an associative memory. The agent is then able to make previously unseen game moves by generalising from previously learnt game moves, where generalisation implicitly emerges from the overlapped encoding of CAs. Learning is driven by visual feedback from the game environment, perceived by the agent with a rudimentary visual system. Furthermore, learning in the model is continuous, where every perceived and played game move modifies the agent's memory.

The game environment consists of a vertically movable paddle and a ball in a rectangular enclosure of four walls. The paddle is locked to the Y-axis and confined to the wall on the left, while the ball is free to move around the environment. The ball bounces around the environment, rebounding and changing its direction when it touches a wall or the paddle. The goal of the game is to move the paddle along the Y-axis so as to hit the approaching ball and prevent it from crossing the paddle's Y-axis field, that is, going past the paddle and coming in contact with the wall behind it. When the human or the agent moves the paddle and hits the ball successfully, the move is considered a "hit", otherwise, a "miss". A simple score board keeps track of the hits and misses, and is used as a superficial measure of performance. Figure 5.10 illustrates the game environment, where the greyed paddle area is the Y-axis field to which the paddle's movements are limited, and the large white space is the ball's area.

The agent learns to play the game by encoding the game moves—the positions of the paddle relative to the positions of the ball. The representation of positions of the paddle and the ball resemble how place cells in the brain are thought to represent self-location. While the actual game of Pong usually involves taking into consideration both X and Y positions of the ball and a single axis of the paddle, for the sake of simplicity, the agent only perceives the Y-axis positions of the paddle and ball. The network properties of the model are described in the next section, and different aspects of the game are discussed in the sections

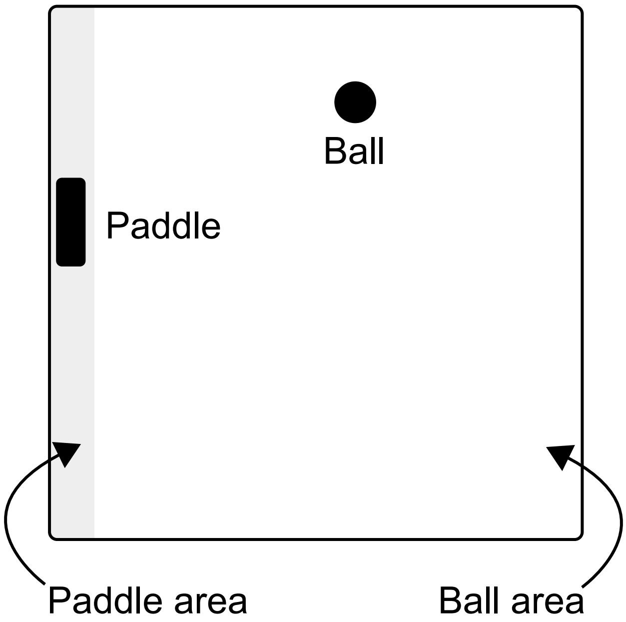


Figure 5.10: Pong game environment

thereafter.

5.3.3 Network properties

The model's network is divided into smaller subnets, namely, *Paddle* that encodes Y-axis positions of the paddle; *PaddleIn*, an intermediary subnet that relays visual signals to *Paddle*; *Ball* that encodes Y-axis positions of the ball; and *Control* that regulates aspects of the agent's vision. In addition, there is a *Vision* subnet made of visual receptors that is a part of the agent's simple visual system. *Paddle* and *Ball* together are the agent's associative memory, as learning only happens in and between them. The rest of the subnets are for signalling and control.

Figure 5.11 shows the connectivity of the subnets in the network. The black arrow heads represent low weight excitatory inter-subnet connections. The double headed arrow between *Paddle* and *Ball* represents to and fro connections, where two neurons in the subnets may have more than one connection with each other. The arrow between *Control* and *PaddleIn* represents random, all inhibitory connections. *Vision* connects to *PaddleIn* and *Ball* in a particular manner which is explained in Section 5.3.4. The details of these inter-subnet connections are given below.

Every excitatory and inhibitory neuron in *Paddle* connects to 20 random neurons in *Ball* with an initial synaptic weight of 0.001 and −0.001 respectively

Every excitatory neuron of each of the ten neurons in every row of *PaddleIn* connects to all the ten neurons in the corresponding row in *Paddle* with a synaptic weight of 6.0

Every excitatory and inhibitory neuron in *Ball* connects to 30 random neurons in *Paddle* with a synaptic weight of 0.001 and −0.001 respectively

Every inhibitory neuron in *Control* connects to 100 random neurons in *PaddleIn* with a synaptic weight of -8.0, for the purpose of disabling the agent's vision when necessary

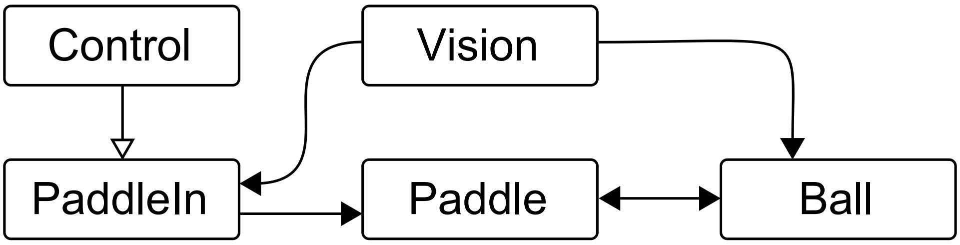


Figure 5.11: Pong model's network structure

Table 5.3 shows network parameters of the subnets. As in the previously described models, the parameters were determined via manual parameter exploration, and are set in such a way that learning between *Paddle* and *Ball* is gradual.

Table 5.3: Pong model's network parameters

		Vision	Control	PaddleIn	Paddle	Ball
Learning rate	λ	0.0	0.0	0.0	0.1	0.1
Threshold	θ	6.0	1.0	4.5	3.5	4.5
Axonal median	χ	NA	0.4	0.15	0.3	0.15
Fatigue	f	0.6	0.2	0.6	0.9	0.9
Fatigue recovery	f_r	0.8	0.8	0.8	0.3	0.3
Decay	δ	1.2	1.2	1.2	1.3	1.3
Saturation	B	NA	NA	NA	18	18
Inhibitory neurons	ι	0.2	1.0	0.1	0.2	0.2
Neurons	N	100×100	20 × 10	100 × 10	100 × 10	100 × 10

5.3.4 Vision

The agent is capable of rudimentary vision, and learns to play the game entirely visually. The visual system is similar to that of the previously described spatial cognitive mapping model (Section 5.2). It consists of the *Vision* subnet—a 100 × 100 grid of binary *on-off* receptors that respond to the colour black. *Vision* receives visual stimuli from the environment as a 100 × 100 grid of pixels. Each pixel in this grid is mapped to a receptor in the corresponding position in *Vision*. Since the paddle and the ball are the only coloured elements in the environment, pixels corresponding to their position switch on the receptors in *Vision*. That is, *Vision* is a one-to-one mapping of the elements in the environment, as seen by the allocentric agent.

When a receptor is on, it sends out positive activation. This is a value $> \theta$ of *PaddleIn* and *Ball* so that the neurons receiving the activation fire immediately. Every receptor of the ten receptors in the paddle area of *Vision*—the width of the Y-axis field of the paddle being ten pixels—connects randomly to the ten random neurons in the corresponding row in *PaddleIn* with a synaptic weight of

6.0. Similarly, each of the 90 remaining receptors in every row in the ball area of *Vision* connects randomly to the ten neurons in the corresponding row in *Ball* with a synaptic weight of 6.0. The ball moves constantly, the paddle occasionally, and *Vision* reflects their position in the environment in real time. The row-to-row connectivity between *Vision*, and *PaddleIn* and *Ball* makes it so that the activation propagated by the receptors in *Vision* excite neurons in *PaddleIn* and *Ball* reflecting the Y-axis positions of the paddle and ball respectively. However, not all neurons receiving activation fire at the same time due to fatigue. Unless inhibited by *Control*, *PaddleIn* subsequently relays the activation to *Paddle*. *Control* is a means of externally controlling the agent's vision, that blinds it of the paddle's position during testing. The purpose of the control mechanism is explained in the next section. In essence, neural firing in these subnets, and the CAs that form in *Paddle* and *Ball* are representative of the vertical positions of the paddle and ball respectively.

5.3.5 Simulation

In the simulation, the agent has two different training modes—human training and self training. After either of the training modes, the agent is left to play the game autonomously, and the gameplay is recorded. In both training scenarios, learning is driven entirely by visual stimuli. Initially, synaptic strengths of neurons in subnets representing the agent's associative memory, *Paddle* and *Ball*, are weak. As explained in the previous section, during training, firing neurons in *Paddle* and *Ball* reflect the Y-axis positions of the paddle and ball respectively. Prolonged co-firing of these neurons lead to the gradual formation of CAs in their corresponding subnets, and the inter-subnet association of CAs representing the paddle and ball in *Paddle* and *Ball* when the ball is close to the paddle, that is, when the Y-axis positions of the paddle and ball are close to each other.

Figure 5.12 shows an example of such an instance, where blackened areas in *PaddleIn*, *Paddle*, and *Ball* show neural firing from visual stimuli representing Y-axis positions of the paddle and ball. In *Paddle* and *Ball*, the co-firing of neurons leads to CA formation. The dashed arrows show how the visual stimuli excite neurons, and the dotted arrow connecting *Paddle* and *Ball* shows associative learning via co-activation of CAs.

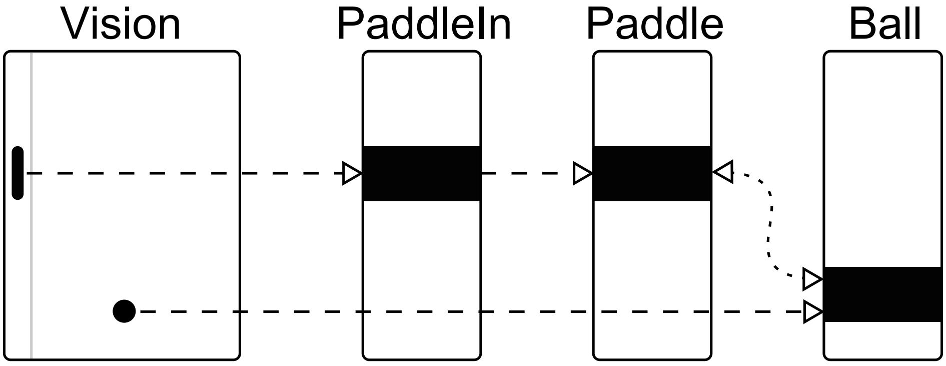


Figure 5.12: Learning in the Pong model's subnets

As the neural firing in *Paddle* and *Ball* shift based on the positions of the paddle and ball in the environment, depending on the extent of the movement, CAs representing them overlap. For instance, if the paddle is moved up one position, the CA learnt representing the current position will overlap with the antecedent CA that represents the previous position. Depending on the movement of the paddle and ball, a number of CAs representing their Y-axis positions form, overlapping with local CAs representing neighbouring positions. Thus, the formation of CAs representing the vertical positions of the paddle and ball involve the amalgamation of positions that are close to each other.

While position CAs in *Paddle* and *Ball* have strong associations with their neighbouring CAs via massive overlaps, inter-subnet associations between CAs depend on the position of the paddle in relation to the ball. During human training, if the human player constantly moves the paddle in such a way that it always avoids the ball by moving away from the approaching ball, the associations between the position CAs in *Paddle* and *Ball* will reflect these erratic moves. For example, the paddle positioned at the top when the ball is at the bottom. This is because associations between inter-subnet CAs in the model are driven entirely by co-activation which further depends on the vertical positions of the paddle and ball in the environment.

In addition, if there is no human input, the agent moves the paddle by itself depending on the activity in *Paddle*. Since *Paddle* is a one to one mapping of the paddle area in the environment, the activity of firing neurons in *Paddle* is averaged, and the paddle is moved to the Y-axis position in the environment corresponding to the most active region in *Paddle*. If there is no activity in *Paddle*, the paddle stays idle. However, if for instance, a group of neurons near the top of *Paddle* are firing due to recurring activity in a CA, the agent moves the paddle to the corresponding position in the environment. This is a cyclic process, as the paddle is moved depending on the activity in *Paddle*, and the position of the paddle in the environment in turn reflects neural activity in *Paddle* via *Vision*. When the agent learns gradually, the CAs formed as a result in *Paddle* may inhibit weaker activity, thus making the process dynamic.

When the paddle and the ball move away from a particular position, the neural activity in *Paddle* and *Ball* shift accordingly. As this happens, CAs of neurons receiving external stimulation from *Vision* representing the current position may suppress some of the previously active CAs. By the anti-Hebbian learning rule, synaptic strength between active CAs and inactive CAs are weakened. For example, excitatory neurons in *Paddle* will have stronger connections with neurons in the immediate vicinity than neurons at a distance. Similarly, neurons at a distance may receive inhibition from a group of firing inhibitory neurons. Activation from reverberating CAs representing the current position of the paddle may spread to overlapping regions forming larger CAs. That is, multiple Y-axis positions of the paddle close to each other may combine to form CAs that represent intermediary positions via spreading of activation. This process creates CAs representing new positions, or generalising from previously learnt positions.

As there is a considerable amount of randomness in the initial network topology in both training scenarios, and the game physics, the simulation was run 20 times in total, 10 times under each training scenario. Each trial lasted five minutes in simulated time (30,000 cycles). In the human training mode, a human played the game for a minute (6000 cycles) prior to testing. The two training modes and their outcomes are discussed in the subsections below, and the results and findings in Section 5.3.6.

Human training

In the human training mode, the paddle is controlled via keyboard input by a human player using the *up* and *down* keys. As the game progresses, the player moves the paddle in accordance with the ball with the goal of hitting it. Since there is two way learning between *Paddle* and *Ball*, recurrent activation from *Vision* drives the formation and associations of CAs in these subnets. This essentially encodes the player's action of moving the paddle in relation to the ball, and hitting the ball by aligning the paddle to it. As explained in the previous section, CAs in both subnets are continuous, overlapping with antecedent and precedent CAs, representing the continuity of the movement of the paddle and the ball.

CA formation and association in the subnets depend on the player's behaviour. Consistent behaviour reinforces certain CAs and associations between them while weakening the associations between certain inactive CAs. Similarly, if the player's behaviour is erratic, the encoded moves may reflect this behaviour. Thus, the agent observes the player play and learns from the player's moves. Since learning is highly dependent on the nature of the moves, to some extent, the agent encompasses the gameplay behaviour of the human player.

Human training lasts a minute in simulated time, after which, the agent is tested for six minutes. *Control* inhibits *PaddleIn* so that the visual stimuli from *Vision* representing the position of the paddle in the environment is not relayed to *Paddle*. That is, the agent is blinded to the paddle's position in the environment so that it has to recollect the paddle's position in relation to the ball from its learnt internal states. As the ball moves around in the environment, it continues to elicit neural activity in *Ball*. This activates CAs learnt in *Ball*, that in turn activate CAs in *Paddle* to which they are associated. Activation of CAs in *Paddle* makes the agent move the paddle to the corresponding position in the environment. As CAs in *Paddle* and *Ball* are continuous, overlapping, and have inter-subnet associations, *Paddle* receives constant activation from CAs in *Ball* that in turn translates to the agent moving the paddle in the environment. This causes the agent to move the paddle in accordance to the ball in real time, hitting it, depending on the accuracy of the moves learnt during training. Since a reward mechanism is absent, it is possible for erratic training to reinforce flawed game moves.

Figure 5.13 shows the number of hits and misses by the agent over ten trials. The number of misses are small compared to the number of hits as a result of supervised training by the human player. Moreover, since the number of moves possible in the training time of two minutes is small, the agent performs well in the novel situations it encounters in the following six minutes of testing, supported by emergent generalisation.

Self training

In the self training mode, the agent acquires game moves on its own. Initially, as there is no human intervention or pre-existing memory, the agent remains quiescent, not moving the paddle. For the agent to learn the game moves, it has to perceive the paddle's position relative to the ball. So, a simple vision control mechanism regulates the vision so that the agent only perceives the position of the

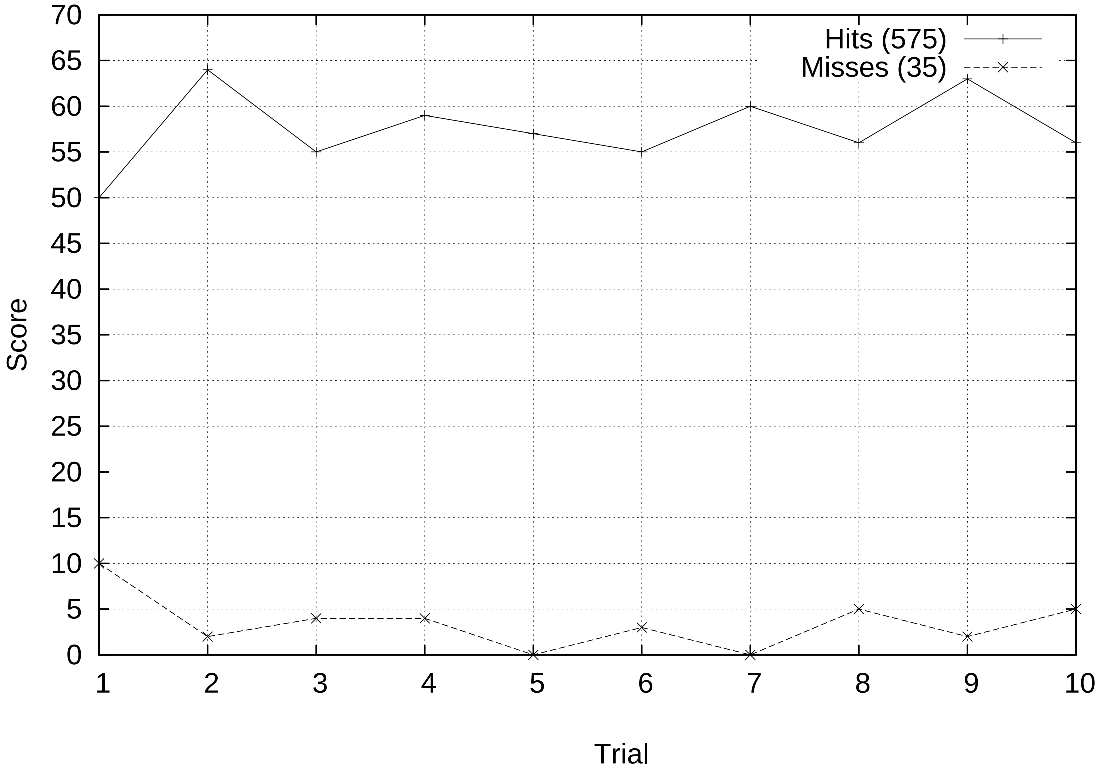


Figure 5.13: Pong agent's game score in human trained gameplay

paddle when the ball approaches it. For this, *Control* inhibits *PaddleIn* so that visual stimuli from *Vision* representing the paddle is not relayed to *Paddle*.

When the moving ball crosses the paddle's Y-axis field (Figure 5.10), due to the connectivity of *Vision* to *Paddle* via *PaddleIn*, *Vision* sends neural activation to *Paddle*, perceiving the ball in place of the paddle. This sudden activity in *Paddle* makes the agent move the paddle to the corresponding position in the environment. Then, inhibition from *Control* is released, enabling the agent to visually perceive the position of the paddle in relation to the position of the ball. This co-firing of neurons in *Paddle* and *Ball* lasts for 50 cycles before *Control* shuts of visual perception again. This is a symbolic mechanism, where control is externally excited so that its inhibitory neurons inhibit *PaddleIn*, preventing the visual stimuli from being relayed to *Paddle*. Since the ball is continuously moving, it then bounces off the point of impact on the wall on the paddle's side, where the paddle currently is. This process repeats every time the ball crosses the paddle area, causing the paddle to move to the position of impact, triggering learning via co-firing of neurons in *Paddle* and *Ball*, leading to the eventual formation of CAs. Gradually, the agent learns enough positions so that the approaching ball elicits activity in learnt CAs before crossing the paddle area, making the agent move the paddle to the particular position before the approaching ball, hitting it. Thus, with a simple mechanism of relative learning, the agent acquires the game moves autonomously.

There is no separate test in the self training mode, as learning and playing are continuous. The agent learns and plays at the same time, improving with time. Figure 5.14 shows the score distribution of the agent across ten trials. Unlike the small number of misses in the human training mode, the percentage of misses compared to the hits are significant in each of the trials. This is because learning in self training mode is driven entirely by visual environmental feedback triggered by the misses when the ball crosses the paddle's Y-axis field and hits the wall.

Figure 5.15 shows the progression of scores over an example self training trial, where the dots indicate when hits and misses occurred. Initially, the number of misses are greater than the number of hits, but the agent learns from the misses and the number of successful hits quickly surpass the number of misses. This growth is drastic, as the agent only needs to learn a few moves, from which generalisation emerges that further enables the agent to cope with novel situations.

5.3.6 Discussion

The previous section described the two different types of training in the model and their results. Overlapping associations between positions of the paddle and

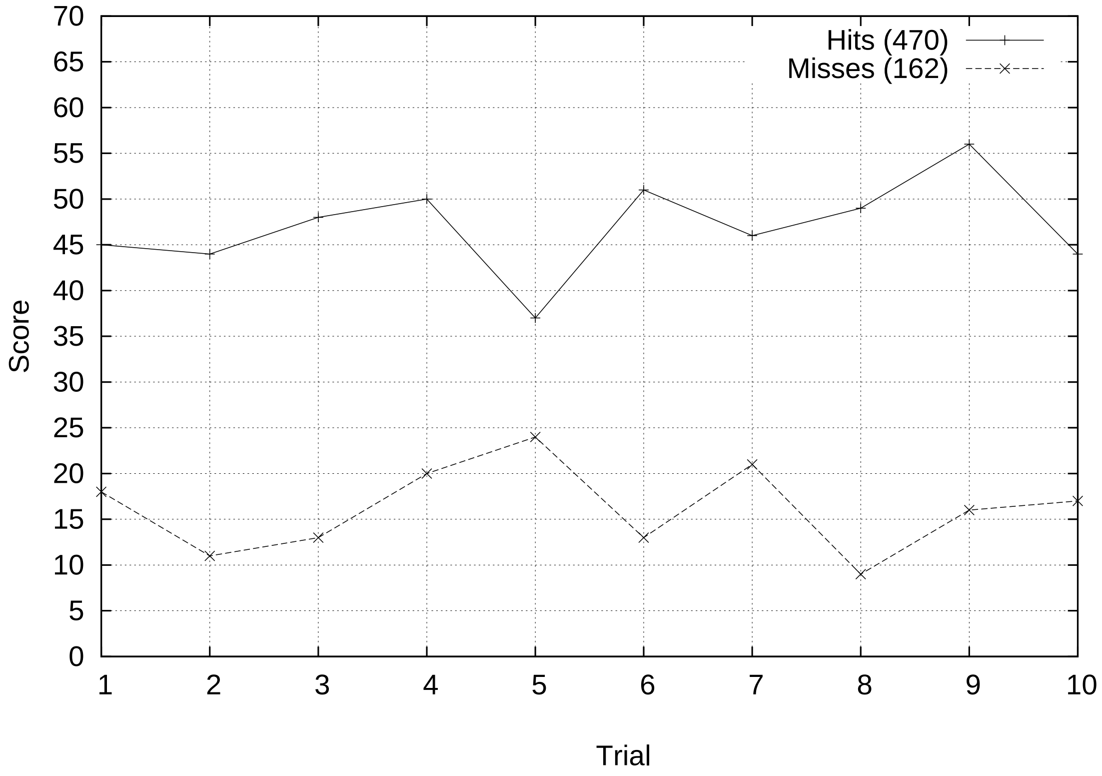


Figure 5.14: Pong agent's game score in self trained gameplay

ball learnt by the agent are dynamic, and change over time. The agent, driven by visual stimuli, is able to play the game reasonably well. In both training scenarios, the number of successful hits are significantly higher than the number of misses (Figure 5.13 and Figure 5.14).

Overlapping CAs formed in *Paddle* and *Ball* are continuous. Figure 5.16 shows the visualisation of *Paddle* at different instances after an example human training session. Each panel in the figure is a full representation of the state of *Paddle*, where colour gradients indicate activation levels of neurons in the subnet relative to its state. The states were recorded when at least 50 neurons fired in *Paddle* due to internal activation from associated CAs in *Ball*, at least 10 cycles apart, capturing the neural activation associated with continuous movement. It

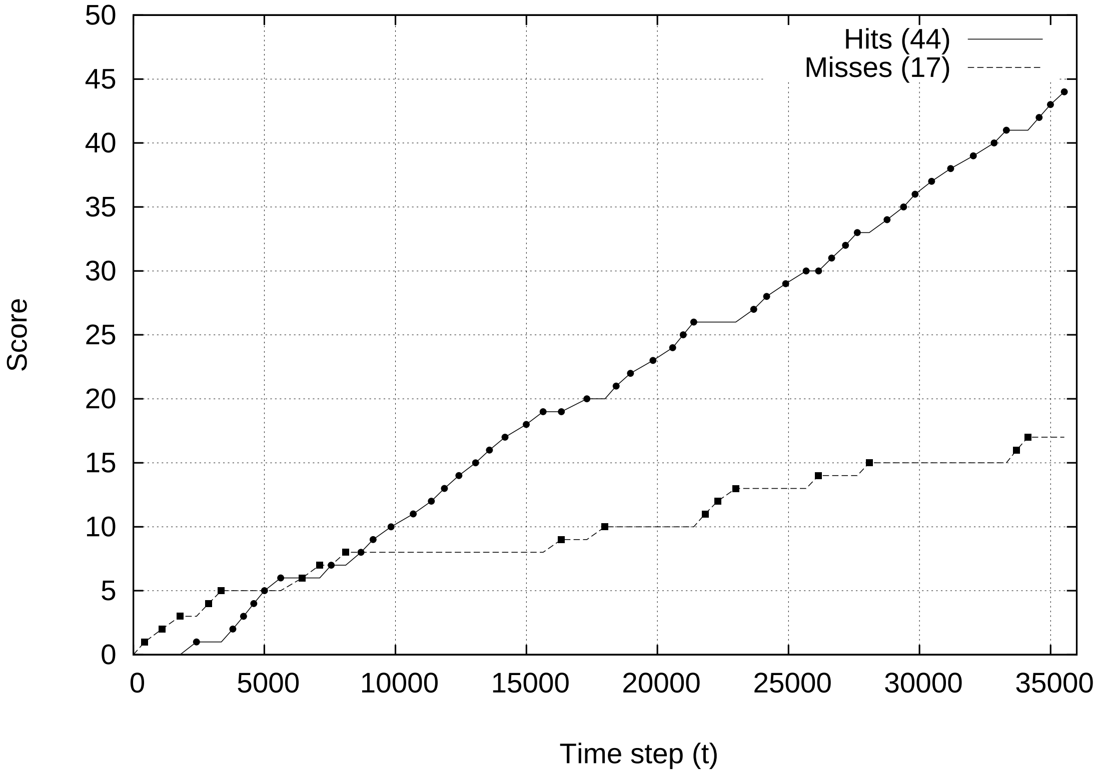


Figure 5.15: Progression of the Pong agent's game score during self trained gameplay

can be seen that after $t = 6061$, this criteria was met only at $t = 6239$. This may have been due to the ball moving away to a certain position in the environment that was not learnt previously, before coming back to a previously learnt position close to that as $t = 6061$. Since the activation is internal, invoked by *Ball*, they are CAs. The figure shows activation spreading and overlapping with neighbouring regions, indicating physical overlaps between CAs. Due to their varying levels of overlap and highly dynamic nature, it is difficult to identify distinct CAs or approximate their physical boundaries in the subnet. These CAs represent learnt positions of the paddle in the environment in relation to the ball. The spreading of activation from CA to CA gives rise to generalisation effects. For instance, if

A, B, and C respectively are three immediate consecutive positions, and the agent has learnt the positions A and C, the spreading of activation between A and C as the paddle moves in relation to the ball may cause them to overlap, giving rise to an intermediate CA representing B in the process.

Figure 5.17 is a similar visualisation where the recorded states are at least 200 cycles apart. Unlike Figure 5.16 where movements are successive and activation is local, the figure shows the ball at arbitrary positions in the environment. The regions of activation are spread out and overlapping, where the most active neurons may represent the current position of the ball. For instance, at $t = 7774$ and $t = 7974$, the activation is spread over a large region, encompassing multiple positions. This indicates physical overlap between multiple CAs, where new CAs may have formed. If there was a position that was not learnt during training, this overlap may cause neighbouring positions to be "generalised" into a previously unseen position, that is, result in the formation of intermediate CAs. In the game, the agent moves the paddle depending on such activity in *Paddle*.

As mentioned in the introduction, the behaviour of neurons in *Paddle* resemble hippocampal place cell CAs, whereby activation of CAs in the subnet reflect the position of the paddle in the environment. It has been shown that the activation patterns of place cells in rats represent their allocentric positions in the environment, where activity shifts in relation to their movements [Dragoi and Buzaski, 2006; Leutgeb et al., 2005]. In such cases, activity is localised to a region representing a minimal, scaled down version of the perceived environment. This is similar to the shifting of CA activations in *Paddle* (shown in figures 5.16 and 5.17), where *Paddle* represents the environment, and the region of activation corresponds to the vertical position of the paddle in the environment. In the model, *Paddle* only accounts for the size of the environment and Y-axis positions, where as place cells are thought to account for various factors such as size of the environment, orientation, trajectory, and multi-dimensional spatial positions. In addition, the size of *Paddle* is fixed in the model, unlike place cells that seem to dynamically account for environments of varying dimensions.

Paddle does not represent the agent's self-position, as the agent is external to the game environment. However, it is reasonable to consider *Paddle* as the agent's place cell mechanism, as there is evidence suggesting that place cells in humans encode their spatial positions even when interacting within a virtual environment [Doeller et al., 2010]. There is also evidence suggesting that place cell activations may be continuous and overlapping, and learning and recollection in them may also be continuous where new information may be added to already learnt representations, resulting in the emergence of intermediate representations [Leutgeb et al., 2005]. The model shows this behaviour, whereby continuous learning results in the merging of new representations of positions with existing representations, causing intermediate positions to emerge, that then aid the agent in gameplay.

In conclusion, the model provides insights into the dynamics of overlapped encoding of associative memories. The results support the view that generalisation may emerge from the integration and overlap of a new representation with existing representations [Shohamy and Wagner, 2008]. It unifies overlapped encoding of CAs, emergence of generalisation from such overlapping memories, and some behavioural aspects of hippocampal place cells—processes that have strong neurobiological and neuropsychological underpinnings—into a single system. With an interactive game playing agent that learns and plays the game well, the model demonstrates the aforementioned processes and their proposed nature. Moreover, the agent producing novel behaviour is an interesting model of AI. While it is not a definitive representation of the processes, it is intriguing nonetheless, and like the other experimental models, serves as an impetus for further research.

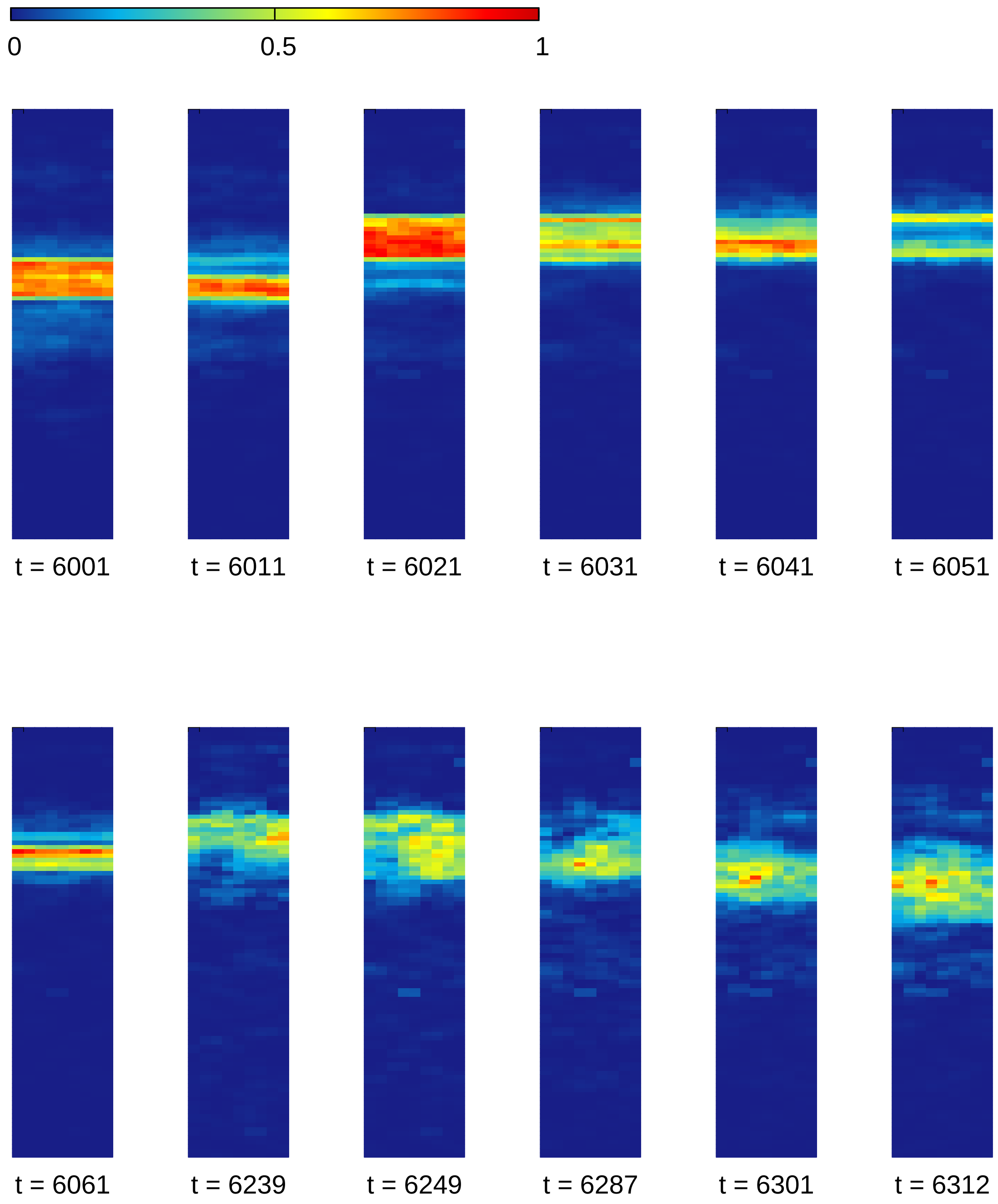


Figure 5.16: Visualisation of activity in *Paddle* during continuous movement of the ball

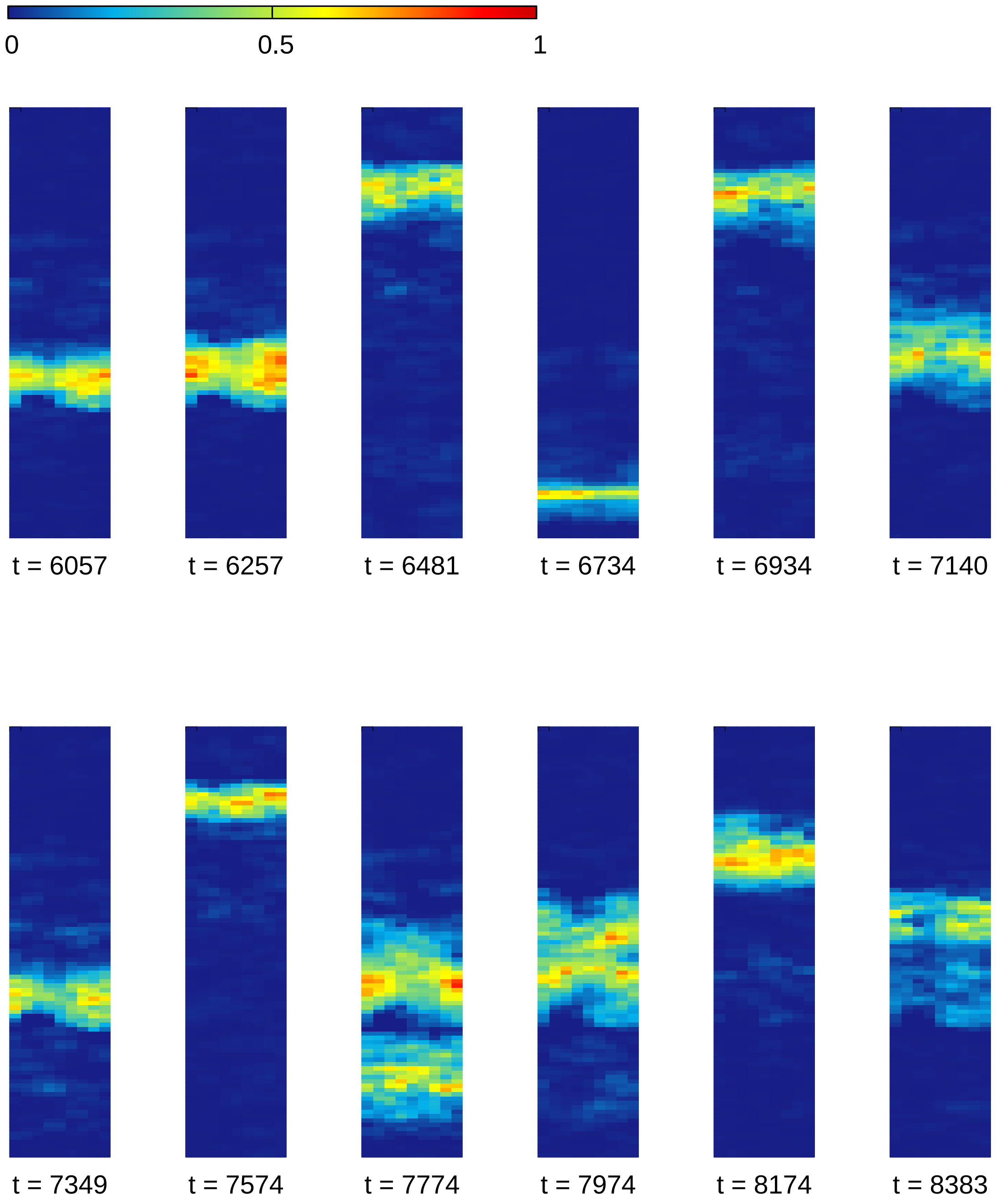


Figure 5.17: Visualisation of activity in *Paddle* during arbitrary movement of the ball

5.4 Natural language disambiguation with an associative memory of semantic hierarchies

The models described in the previous sections concerned neurobiological and psychological processes. The model described in this section is the CA based implementation of a machine learning task in natural language processing—disambiguation of prepositional phrase (PP) attachment ambiguity [Hindle and Rooth, 1993] in the English language. The model is based on prior work in machine learning [Nadh and Huyck, 2009]. It was conceived to test the computational capabilities of neurobiologically plausible CAs, and to evaluate how a neural approach may work for certain machine learning tasks, in this case, natural language processing.

The network in the model is relatively large with over 300,000 neurons, as it has to accommodate a large amount of real world data. As a result, computational limitations restricted development. For instance, the number of outgoing inter-subnet synapses per neuron had to be restricted to 90 to cope with memory constraints. The model also required about eight hours on average to complete a full simulation on a reasonably powerful personal computer. Derivation of suitable network parameters for the model was made difficult by this long time window. As a result, unlike the previous models, the development ended up spanning many months. Eventually, a relatively good set of parameters was derived and the model was able to resolve the PP attachment ambiguity better than many known machine learning models, suggesting that a neural approach may benefit many similar machine learning tasks. More importantly, the results show the viability of CAs as a general purpose tool for computational problems. The PP attachment ambiguity in general, and the model are described in the following subsections.

5.4.1 PP attachment ambiguity

Semantic and syntactic ambiguities are important challenges in natural language processing. Such ambiguities make parsing of text into symbolic representations difficult. The PP attachment ambiguity is such a problem. In the English language, PP attachment ambiguities arise when a PP follows a verb phrase (VP) and a noun phrase (NP) [Hindle and Rooth, 1993]. A canonical example is the sentence "*I saw the girl with the telescope*" (Example 1). Here, the phrases *saw* (VP), *the girl* (NP), and *with the telescope* (PP) may be combined in two ways, resulting in two different semantic interpretations as listed below. These two cases are also illustrated in Figure 5.18.

A. The PP can attach to the VP, resulting in the interpretation meaning *the girl* was seen through *the telescope*. In this attachment, *the telescope* is the instrument of the verb *saw*—(VP *saw* (NP *the girl*) (PP *with the telescope*)).

B. The PP can attach to the NP resulting in the semantic interpretation *the girl* possessing *the telescope* was seen—(VP *saw* (NP *the girl* (PP *with the telescope*))).

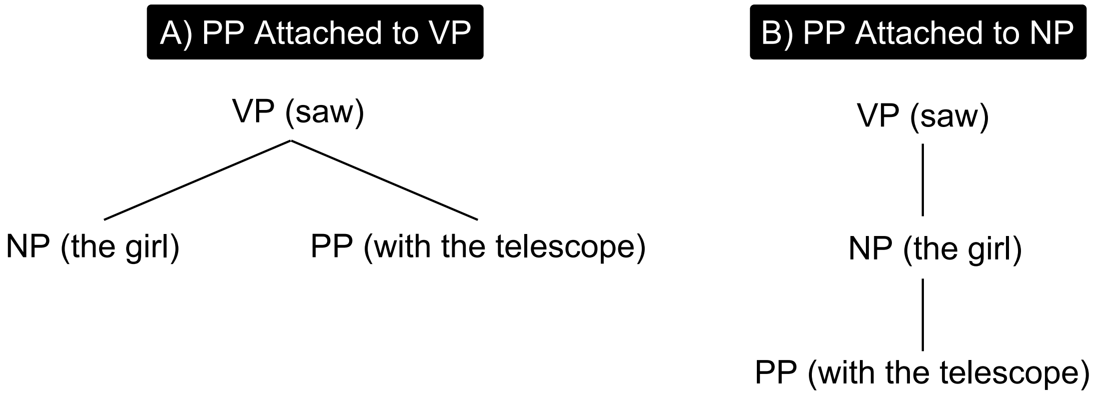


Figure 5.18: An example of PP attachment ambiguity

An example for a case where a human annotator may consider PP attaching to the NP "correct" is the sentence "*I saw the girl with the flowers*". Here, it makes sense for the PP to attach to the NP meaning that *the girl* possessing *the flowers* was seen, as unlike in Example 1, flowers are not typically used as instruments of seeing. Additionally, attachment decisions can be influenced by factors like context, making it impossible to disambiguate all cases of the ambiguity from sentences on their own.

Many machine learning models attempt to resolve the PP attachment ambiguity using quadruples of headwords of ambiguous sentences—(v n_1 p n_2), where v is the head verb, n_1 is the head noun of the NP, p is the preposition, and n_2 is the head noun of the PP. In Example 1, the quadruple is *saw girl with telescope*. While such quadruples may aid in efficient disambiguation, even a single misattachment decision may result in multiple parsing errors [Lin, 1998].

The quadruples of ambiguous sentences are usually obtained from standard corpora, as is the data used in the model described. Previously seen quadruples can be resolved relatively well, but sparseness of real world data in general [Atterer and Schütze, 2007] reduces the possibility of frequent occurrences of highly similar quadruples. So, many machine learning models adopt methods of generalising from training data. The model described in this section is the neural CA based implementation of prior work [Nadh and Huyck, 2009] that used a method of measuring semantic similarities between quadruples for disambiguation. Both methods work on the assumption that semantically similar sentences should have similar PP attachments. The higher the similarity, the higher the probability of having similar attachments. In the previous work, semantic similarity between quadruples was calculated by an algorithm that counted the frequency of common terms describing individual words in a quadruple in semantic trees obtained from a lexical dictionary.

The model described here achieved an average disambiguation accuracy of 84.56% (standard deviation of 2.85) over six trials, and the highest accuracy of 88.33% in a single trial. The results are on par with the original machine learning model, and other known methods of PP attachment ambiguity disambiguation. The results suggest that neurobiologically plausible CAs are powerful tools for performing computational tasks usually limited to machine learning systems.

5.4.2 The model

In the model, ambiguous quadruples extracted from a large training set—the *Penn Treebank* (PTB) corpus [Marcus et al., 1993]—and sense hierarchies of words in the quadruples from the semantic dictionary *WordNet* [Miller, 1990] are encoded in a network as CAs, along with their correct attachment decisions. Over time, large overlapping CAs representing semantic relationships between training quadruples emerge, that are then used to classify novel quadruples in a test set. The PTB and WordNet datasets are further discussed in Section 5.4.3.

As mentioned earlier, the derivation of a suitable set of parameters for the network was complicated by the computational constraints imposed by the size of the network. The behaviour of the ANN at such a scale had never been tested. For deriving parameters, the training set itself was divided into a smaller training set and a test set (50% – 50%). The model was trained and tested on these two smaller test sets repeatedly for the purpose of parameters derivation so as to avoid biasing results from the actual test. The parameters had to be adjusted to enable gradual learning in the large network over prolonged periods of time, as the dataset is large, requiring prolonged learning.

5.4.3 Datasets

The model learns from examples with known attachment decisions from a training set. These are extracted from the *Wallstreet Journal* subset of the PTB corpus. The PTB is an English corpus annotated with part-of-speech [Charniak, 1997] and syntactic structures by lexicographers [Marcus et al., 1993]. Sentences in the corpus are represented in a standard parenthesised tree structure as shown in the example in Figure 5.19.

```
(S
  (NP (PRP I))
  (VP (VBD saw)
    (NP (DT the) (NN girl))
    (PP (IN with)
      (NP (DT the) (NN telescope))))
  (. .)
)
```

Figure 5.19: Annotated form of the sentence *I saw the girl with the telescope*

```
Verb attachment                            Noun attachment
(VP (*)                                    (VP (*)
  (NP *)                                     (NP *
  (PP-* (*)                                    ( PP-* (*)
    (NP (*))                                     (NP (*)) )
  )                                           )
)                                           )
```

Figure 5.20: Annotated tree structure of sentences with PP attachment ambiguity

A simple recursive decent parsing algorithm creates quadruples from sentences extracted from the corpus of the forms shown in Figure 5.20, where * is a wildcard representing one or more tagged items. For the one verb and two nouns in a quadruple, word sense hierarchies are obtained from WordNet. Word sense hierarchies are lexical trees of sequences of hypernyms where each subordinate sequence or *synset* is a set of synonyms of the superordinate word. The model trains on the word sense hierarchies along with their parent quadruples. Figure 5.21 shows an example sense hierarchy of the noun *telescope*. In the example, sequences are generic terms for the words in their immediate superordinate position, for instance, *scientific instrument* ← *instrument*. However, words may have multiple sense hierarchies, for instance, the word *bat* may refer to the animal, or a wooden club used in certain games. In such cases, only the first sense hierarchy is used and the rest is discarded, as WordNet orders entries based on the frequency of their occurrence across multiple corpora [Lee et al., 2000]. While this may yield false positives in some cases, the probability of the first sense being correct remains high due to the frequency heuristic.

```
telescope, scope
 => magnifier
  => scientific instrument
   => instrument
    => device
     => instrumentality, instrumentation
      => artifact, artefact
       => whole, unit
        => object, physical object
         => physical entity
          => entity
```

Figure 5.21: Word sense hierarchy of the noun *telescope*

The quadruples thus extracted from the PTB corpus are divided into a training set and a test set as shown in Table 5.5. Table 5.4 shows some example quadruples from the dataset.

Table 5.4: Example quadruples from the PTB

v	n_1	p	n_2	attachment
closing	port	for	time	VP
filed	brief	in	appeal	VP
admit	victory	in	congress	NP
regards	year	as	period	VP
terminated	negotiations	for	purchase	NP

Table 5.5: Some statistics of the dataset extracted from PTB

Total sentences processed	49,208
Sentences with PP attachment ambiguity	7810
Quadruples in the training set	4810
Quadruples in the test set	3000
Training quadruples with verb attachments	3847
Training quadruples with noun attachments	963

5.4.4 Network properties

The model has a large network partitioned into four input subnets *Verb, Noun1, Preposition*, and *Noun2* that encode the four individual components of quadruples, v, n_1, p, and n_2 respectively. Two attachment subnets, *VerbAttach* and *NounAttach* encode the attachment decisions of training quadruples. Figure 5.22 shows the gross connectivity of the subnets. The arrows represent random, low weight excitatory inter-subnet connections. Every neuron in the four input subnets connects to 90 random neurons in the corresponding attachment subnet, totalling to over 18 million inter-subnet synaptic connections, hence the aforementioned computational constraints.

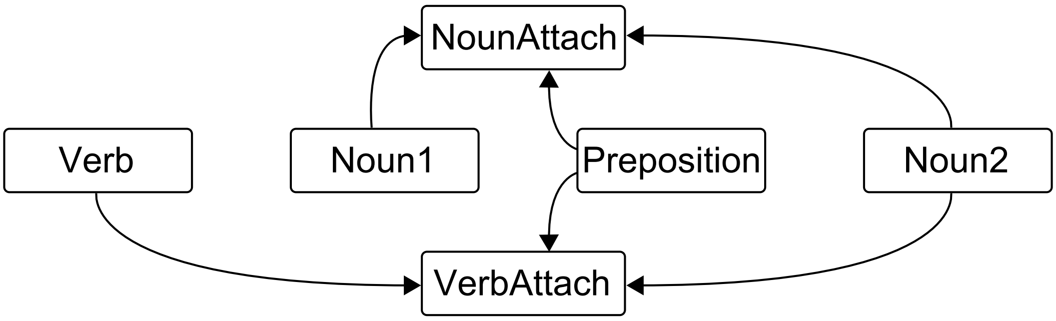


Figure 5.22: PP attachment disambiguation model's network structure

Table 5.6 shows the network parameters for the subnets. The number of neurons in each subnet is derived from the number of words in the full training set. Each word in a quadruple, v, n_1, p, and n_2, and all words in their sense hierarchies are designated 20 neurons each in their corresponding input subnets. The size of the two attachment subnets is determined based on the maximum number of attachment decisions of either attachments in the training set. As 79.9% of the training quadruples have verb attachments, the maximum size of both subnets are based on this figure, where the attachment decisions of training quadruples are assigned 10 neurons each. Hence, a large number of neurons in *NounAttach* never participate in any activity, as only 20.1% of the training quadruples have noun attachments. Still, both subnets have the same number of neurons in order to reduce learning bias. If *NounAttach* had fewer neurons, its connection density would be significantly higher than that of *VerbAttach*, which would result in the activation of a large number of its neurons, affecting the model's behaviour. How neurons are designated to words in the training set is discussed in the next section.

The specifics of inter-subnets connections are listed below.

Every excitatory neuron in *Verb* connects to 90 random neurons in *VerbAttach* with a synaptic weight of .08

Every excitatory neuron in *Noun1* connects to 90 random neurons in *NounAttach* with a synaptic weight of .08

Every excitatory neuron in *Preposition* connects to 90 random neurons in *VerbAttach* and *NounAttach* with a synaptic weight of .08

Every excitatory neuron in *Noun2* connects to 90 random neurons in *VerbAttach* and *NounAttach* with a synaptic weight of .08

Table 5.6: PP attachment ambiguity disambiguation model's network parameters

		Verb	Noun1	Prep	Noun2	VA, NA
Learning rate	λ	0.2	0.2	0.2	0.2	0.1
Threshold	θ	4.0	4.0	4.0	4.0	3.5
Axonal median	χ	0.5	0.5	0.5	0.5	0.5
Fatigue	f	0.6	0.6	0.6	0.6	0.6
Fatigue recovery	f_r	0.8	0.8	0.8	0.8	0.8
Decay	δ	1.2	1.2	1.2	1.2	1.2
Saturation	B	20	20	20	20	20
Inhibitory neurons	ι	0.2	0.2	0.2	0.2	0.25
Neurons	N	2666×20	4746×20	54×20	5076×20	3847×10

5.4.5 Simulation

Initially, every v, n_1, and n_2 of all the training quadruples, and words from their individual sense hierarchies obtained from WordNet are designated 20 neurons each in *Verb*, *Noun1*, and *Noun2* respectively. Similarly, each unique preposition p is designated 20 neurons in *Preposition.* The subnets each only have a single instance of a word even if it occurs multiple times across different quadruples. However, the words in *Noun1* and *Noun2* are not common, as they represent the collection of every unique n_1 and n_2 distinct from each other, respectively. The designation of neurons is serial, where the first 20 neurons in *Verb* represents the first v, the second 20 represents the first word in its sense hierarchy, and so on for every word in its sense hierarchy and every v. Thus, the subnets represent a "bag of words" including every word and its sense hierarchy of all quadruples in the training set. This serial designation of neurons does not encompass the relationships between words in sense hierarchies. For example, initially, there is no physical overlap between the neurons that represent *telescope* and *magnifier* until the associations are acquired via gradual learning during training.

The simulation consists of two modes, a training mode and a test mode. In the training mode, every training quadruple is presented to the network for 100 cycles each. Neurons designated to each of the words v, n_1, and n_2, words in their sense hierarchies, and the p of a quadruple are externally stimulated in their corresponding subnets. This involves setting the activation levels of the designated neurons to values $> \theta$ of their corresponding subnets. Similarly, a set of ten neurons representing the attachment decision of the quadruple is externally stimulated in *VerbAttach* or *NounAttach*, depending on the attachment. The order of this set is assigned serially. For example, for the first quadruple in the training set with a verb attachment, the first ten neurons in *VerbAttach* are stimulated, for the second, the next ten and so on for every quadruple in the training set.

The co-firing of neurons in the four input subnets and the corresponding attachment subnet gradually result in the formation of CAs representing the training quadruple across them. Initially, in the input subnets, CAs representing individual words are formed. Via gradual learning, these CAs overlap with CAs representing similar words in other quadruples via shared words in their sense hierarchies. For instance, CAs representing *boy* and *girl* in *Noun1* may share a large number of neurons, as their sense hierarchies share many common words—for example (*girl, individual, person, someone* ...) $\cap$ (*boy, individual, person, someone* ...).

Learning in the model sees the emergence of CAs with associations via overlap within subnets and associations via synapse across subnets. Quadruples with similar words have overlapping representations in the input subnets. The CAs and their associations may change as new quadruples are learnt. If a training quadruple with a verb attachment is semantically similar to a previously learnt quadruple with the same attachment, its constituent words in the four input subnets may activate the previous quadruple's CA in *VerbAttach* during external stimulation while learning. This may result in the two CAs in *VerbAttach* forming associations. Similarly, quadruples with semantic similarities may have CAs in the attachment subnets with associations varying based on their similarities. As a result, CAs in *VerbAttach* and *NounAttach* may become active in response to activation in the four input subnets representing input quadruples, before the ten corresponding neurons corresponding to the attachment decision are externally stimulated.

The model classifies novel inputs based on this behaviour, where attachment decisions of novel inputs are decided based on the activity elicited by them in the two attachment subnets. If a novel quadruple presented to the input subnets excites more neurons in *VerbAttach* than *NounAttach*, it is considered to have a verb attachment and vice versa. When a novel input activates CAs in either of the attachment subnets, it is the result of associations between previously learnt quadruples with semantic similarities. That is, if many quadruples similar to a novel quadruple have verb attachments, the model assumes that the novel quadruple has the same attachment.

After training, the model is tested by presenting 3000 novel quadruples from the test set. The four words v, n_1, p, and n_2 of each quadruple are presented to the corresponding input subnet by externally stimulating their corresponding neurons in the subnet for 100 cycles each. The number of neurons firing during this period in *VerbAttach* and *NounAttach* due to the activation propagated by the inputs are recorded, and the attachment decision is attributed to the subnet with the largest number of active neurons. The results are discussed in the next section.

5.4.6 Results and discussion

The model attained an average resolution accuracy of 84.56% (standard deviation of 2.85) over six trials. A particular trial yielded the highest accuracy of 88.33%, which is better than the result from the prior work the model is based on [Nadh and Huyck, 2009]. However, the statistical significance of this figure could not be verified due to the aforementioned computational and time constraints, and hence, it could be an outlier. All test quadruples disambiguated incorrectly have noun attachments, as *VerbAttach* has the largest number of active neurons in most cases. As repeated co-activation of CAs increase the strength of their associations, the common verbs and nouns in the large number of quadruples with verb attachments result in reinforcement of CAs representing them, inhibiting weak or sparse activity in their subnets. Since repeating words in the four input subnets have the same set of neurons designated to them, CAs representing repeating verbs and nouns have stronger learnt connections to *VerbAttach* due to the sheer volume (79.1%) of quadruples with verb attachments. The results are presented in Table 5.7.

The CAs learnt across different subnets are large, and overlap with other CAs extensively and change dynamically over the course of learning. During training, a quadruple presented later may be similar to a quadruple learnt earlier. In such a case, the CAs of the newly learnt quadruple may overlap with the CAs

Table 5.7: PP attachment disambiguation model's results

Attachment	Quadruples in the test set	Correct predictions	Accuracy
Verb	2416	2416	100%
Noun	584	234	40.06%
Total	2650	3000	88.33%

of the earlier quadruple, affecting other quadruples that may have been associated to it in the process. Thus, the effect of CAs of different quadruples on one another is highly dynamic. Table 5.8 shows the Pearson's correlation coefficients of three different *VerbAttach* states during training. Pearson's correlation is a measure of linear dependence between two variables. Here, it shows the similarity between different network states, where states are represented by the neurons firing at particular instances. If two states have a high similarity coefficient, it means that many of the same neurons fired at both instances, indicating the similarity of different CAs that were active at those instances. Since these are training states, external stimulation contributes substantially to the activity in the subnet. The three states shown in the table are at cycles $t = 2250, t = 2750$, and $t = 5150$ while learning the quadruples *put stock on list*, *put touch on compromise*, and *enter venture in april* respectively. These example quadruples have verb attachments and were selected to illustrate how CAs representing different quadruples may differ.

Table 5.8: Pearson's correlation coefficient of different states of *VerbAttach*

put stock on list	*put touch on compromise*	0.52
put touch on compromise	*enter venture in april*	0.11
enter venture in april	*put stock on list*	0.1

The higher similarity coefficient of *put stock on list* and *put touch on compromise* is indicative of the underlying semantics. Both quadruples roughly translate to the abstract form *put something on something.* On the other hand, *enter venture in april* is semantically different compared to the other two quadruples, and hence the low similarity coefficients. Interestingly, *put touch on compromise* has a higher similarity to *enter venture in april* than *put stock on list.* This may be due to a number of reasons, from the learning time window to similarities between previously learnt quadruples. The similarities may also change as learning progresses and new quadruples are introduced. This is how the semantic similarity dynamic of the quadruples, encoded across the network, enable classification of novel quadruples. While this is not a definitive measure of the learning dynamics, it is an indicator of the superficial behaviour of the model.

In the real world, context influences attachment decisions. However, like many models [Ratnaparkhi et al., 1994; Stetina and Nagao, 1997; Toutanova et al., 2004; Zhao and Lin, 2004], the model described works in the null context. Even though no system working in the null context can ever resolve all instances of PP attachment ambiguity, null context models continue to produce good results. For instance, Ratnaparkhi et al.'s [1994] maximum entropy model used lexical information within verb phrases obtained from the PTB WSJ corpus and no external semantic knowledge achieved an accuracy of 81.6%. Stetina and Nagao's [1997] decision tree and semantic dictionary method for word sense disambiguation attained an accuracy of 88.1%. A Markov chain random walk model that used WordNet synsets resolved the PP attachment ambiguity with an accuracy of 87.5% [Toutanova et al., 2004]. Nakov and Hearst [2005] used the world wide web as a training set for disambiguation achieving an accuracy of 83.82%. These results are presented in Table 5.9, even though it is difficult to conduct a fair comparison of different disambiguation systems due to the lack of a standard dataset.

The model, like many models, relies on external semantics for disambigua-

Table 5.9: Comparison of results from the PP disambiguation model with prior work

Model	Result
Maximum Entropy Model [Ratnaparkhi et al., 1994]	81.6%
Decision trees and WordNet [Stetina and Nagao, 1997]	88.1%
Nearest neighbour method [Zhao and Lin, 2004]	86.5%
Markov chain random walk model [Toutanova et al., 2004]	87.5%
Disambiguation with a semantic dictionary [Nakov and Hearst, 2005]	83.8%
Semantic hierarchies for lattice construction [Nadh and Huyck, 2009]	88.1%
Semantic hierarchies as overlapping CAs	
Average	84.56%
Highest	88.33%

tion. It combines this data—semantic hierarchies from WordNet—with examples, and uses it as a large semantic dictionary. It represents semantic data as CAs in a large network, unlike symbolic semantic systems used in other machine learning models such as the prior work this model is based on [Nadh and Huyck, 2009]. Groups of neurons represent individual words in the semantic hierarchies. While this representation flattens the hierarchies losing the hierarchical relationships in the process, some forms of the relationships are gradually acquired via Hebbian learning. Due to certain characteristics of CAs—spreading of activation, overlapping, and inhibition—as seen in previous models, these relationships may not be as discrete or accurate as in a symbolic model. Still, the model is able to perform the task well with its neurobiologically plausible CA based associative memory. Curating the dataset manually, for example, reducing jargon and acronyms in quadruples for which there are no semantic hierarchies in WordNet, may produce better results. However, the goal of the model is to explore large CA based associative memories and not specifically scrutinise the natural language processing task at hand. The results suggest that such a neural approach may be better for tasks in AI usually restricted to symbolic machine learning systems—at least, some natural language processing tasks—hinting at interesting prospects, and demonstrating the versatility of the CA model.

5.5 Chapter summary

This chapter described the four models developed in the thesis, namely, emergent context sensitivity in CAs; spatial cognitive mapping with CAs embodied in a virtual agent; emergence of novel behaviour in CAs in a self learning game playing agent; and prepositional phrase attachment ambiguity resolution with CAs, a natural language disambiguation task. It presented in detail, the models, the neurobiological concepts relating to them, and their findings, highlighting different capacities of CAs. The next chapter summarises the findings from the models, their novel contributions, and briefly discusses the prospects for future work.

Chapter 6

Discussion and conclusion

This thesis has presented a body of work aimed at exploring and construing the nature of associative memory and complex processes that emerge from it—processes that constitute intelligence. The different experimental models described in the thesis explored different aspects of associative memory. Moreover, all models are based on the CA, a construct with strong neurobiological underpinnings. Together, the models point at a unified CA based associative memory model with many interesting capabilities. In addition, many of the models described resemble observed neurobiological and psychological processes in the brain. The work thus highlights the benefits of a neurobiological approach to modelling in AI, in particular, demonstrating the CA model to be a powerful candidate.

The results from the models are novel, and make contributions to the understanding of associative memory, the nature of CAs, and some aspects of AI in general. The outcomes, with an emphasis on the possible novel contributions, are summarised in the following section.

6.1 Summary of findings

Context sensitivity can emerge from an associative memory, wholly from the underlying characteristics of CA formation and association. The model (Section 5.1) showed how memory retrieval is affected by context, where the activation of CAs representing different concepts is influenced by CAs representing different contexts. This behaviour implicitly emerges from the learnt associations in the associative memory. The behaviour of the model is similar to some observations in the brain thought to underlie context sensitivity.

A form of **spatial cognitive mapping**—landmark based navigation—is possible with an associative memory of CAs representing various features in an environment, and the associations between them. The model encodes episodic and semantic memories, and relies on their sequential activation to facilitate the navigation of an agent in an environment. The sequential activation of memories in the model resembles processes in the brain thought to be involved in landmark based spatial navigation. The model demonstrates how a process as complex as spatial navigation can be implemented with a relatively simple CA based associative memory. Having been incorporated into a more advanced agent, it serves as a prototype for complex spatial navigation in more advanced systems.

Novel behaviour can emerge from the **generalisation** of memories in an associative memory of overlapping CAs. Generalisation, a fundamental process constituting intelligence, is shown to implicitly emerge from an associative memory of overlapping CAs, resembling processes in the brain. In particular, the similarities of the model with hippocampal place cells thought to underlie spatial navigation and associative memory are demonstrated. Based on these processes, an agent learns to play a game of Pong by observing human play, or on its own, subsequently producing novel behaviour. The model serves as a prototype for advanced virtual and mechanical agents that may need to evolve novelty, a fundamental characteristic of intelligence.

A large overlapping memory of CAs encoding **real world data** can act as a powerful **categoriser** and perform a machine learning task. Using the associations that emerge from the overlapped encoding of large sets of hierarchical semantic data, the natural language processing model disambiguates sentences with PP attachment ambiguity with an accuracy on par with many known machine learning models. The model demonstrates that a neurobiological approach may be beneficial for certain tasks in AI that are usually limited to machine learning.

The models explored the dynamics of neurobiologically plausible associative memories with synaptic connections and overlapped encoding of CAs. The models, with their novel outcomes, have shown that the characteristics of CAs can inherently give rise to complex phenomena resembling similar low level processes in the brain, hinting at a unified model of associative memory. Moreover, the models, in particular the natural language disambiguation model, demonstrate that the CA model may be a powerful general purpose tool for tasks in AI.

6.2 Final remarks and future work

Like many works, the research presented in this thesis raises interesting questions and prospects, warranting further investigation. There are many characteristics of associative memory yet to be understood. For instance, while the models demonstrated effects of local inhibition among CAs, the brain is thought to have global inhibitory mechanisms [Garagnani et al., 2009; Rolls, 2007] that affect memory, such as large scale context effects. Besides, the models do not take into account the dynamic states a CA can have. This is an important aspect, as variable activation of CAs may underpin higher processes. If a CA is very active at an instant, but not as active at another instant, the difference in activation level may have significant effect on how the CA influences an associative memory.

In the ANN used, a cycle of neural activity is considered to be 10 milliseconds in simulated time. However, the models ignore the temporal properties of firing neurons. In the brain, precise timing of neural firing and similar temporal properties are important, and are thought to be a mechanism of encoding memories [Dragoi and Buzaski, 2006; Engel et al., 1991; Kaplan et al., 1991; Wennekers and Palm, 2000]. These are characteristics that further investigations into associative memories may need to consider. In addition, while the resemblances of the models with low level neurobiological and psychological processes in the brain have been highlighted, they need to be evaluated against real world data.

The sections detailing the models each presented a table of network parameters to highlight how the parameters differ across models, and how the parameter values were adjusted to obtain desired behaviour. Many of these parameters such as fatigue and decay are neurobiologically inspired. It is important to consider how a set of optimal parameters can be obtained that can unify the differences across these models and future models.

A better neural framework to succeed the ANN (described in Chapter 4), that is more computationally efficient and may consider some of the above mentioned characteristics, is being planned. In addition, a model that combines the dynamics of overlapped encoding of CAs and hippocampal place cells for performing an advanced cognitive mapping task—a virtual agent in a complex maze—is being developed.

Ultimately, this thesis is a small contribution to the tremendous task of comprehending, and one day, creating intelligence. Like many endeavours in the field of AI, it is a part of continuing work, to which, the magnificence of intelligence is the impetus.

Bibliography

Abbott, L. and Nelson, S. (2000). Synaptic plastiticy: taming the beast. *Nature Neuroscience*, 3:1178–1183.

Amit, D. J. (1989). *Modelling Brain Function: The world of attractor neural networks*. Cambridge University Press, Cambridge, United Kingdom.

Amit, D. J. and Brunel, N. (1995). Learning internal representations in an attractor neural network with analogue neurons. *NETWORK*, 6(3):359–388.

Anderson, J. (1998). *The atomic components of thought.* Lawrence Erlbaum Associates, Mahwah, N.J.

Anderson, J. R. and Bower, G. H. (1980). *Human Associative Memory: A Brief Edition*. Lawrence Erlbaum Associates, USA.

Atterer, M. and Schütze, H. (2007). Prepositional phrase attachment without oracles. *Computational Linguistics*, 33(4):469–476.

Basheer, I. A. and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, 43:3–31.

Beauregard, M., Chertkow, H., Bub, D., Murtha, S., Dixon, R., and Evans, A. (1997). The neural substrate for concrete, abstract, and emotional word lexica a positron emission tomography study. *Journal of Cognitive Neuroscience*, 9(4):441–461.

Bechtel, W. and Abrahamsen, A. (2002). *Connectionism and the mind.* Blackwell, Oxford.

Belavkin, R. V. and Huyck, C. R. (2010). Conflict resolution and learning probability matching in a neural cell-assembly architecture. *Cognitive Systems Research*, 12(2):93–101.

Bennett, A. (1996). Do animals have cognitive maps? *The Journal of Experimental Biology*, 199(1):219–224.

Bennett, M. V. and Zukin, R. S. (2004). Electrical coupling and neuronal synchronization in the mammalian brain. *Neuron*, 41(4):495–511.

Bevan, M. and Wilson, C. (1999). Mechanisms underlying spontaneous oscillation and rhythmic firing in rat subthalamic neurons. *Neuroscience*, 19:7617–7628.

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press.

Botvinick, M. M. and Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113:201–233.

Brown, R. and Kulik, J. (1977). Flashbulb memories. *Cognition*, 5(1):73–99.

Brunel, N. and van Rossum, M. C. W. (2008). Lapicque's 1907 paper: from frogs to integrate-and-fire. *Biological Cybernetics*, 97:337–339.

Bunge, S. A., Kahn, I., Wallis, J. D., Miller, E. K., and Wagner, A. D. (2003). Neural circuits subserving the retrieval and maintenance of abstract rules. *Journal of Neurophysiology*, 90(5):3419–3428.

Burgess, N., Maguire, E. A., and O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, 35(4):625–641.

Buzsáki, G. (2005). Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus*, 15(7):827–840.

Buzsáki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science (New York, N.Y.)*, 304(5679):1926–1929.

Byrne, E. and Huyck, C. (2010). Processing with cell assemblies. *Neurocomputing*, 74(1–3):76–83.

Carpenter, G. A. and Grossberg, S. (1987). ART 2: self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26(23):4919–4930.

Carpenter, G. A. and Grossberg, S. (1990). ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, 3(2):129–152.

Carr, M. F., Jadhav, S. P., and Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, 14(2):147–153.

Cavaco, S., Anderson, S. W., Allen, J. S., Castro-Caldas, A., and Damasio, H. (2004). The scope of preserved procedural memory in amnesia. *Brain*, 127(8):1853–67.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI Magazine*, 18:33–44.

Churchland, P. and Sejnowski, T. (1992). *The Computational Brain*. MIT Press, USA.

Cohen, N. J., Poldrack, R. A., and Eichenbaum, H. (1997). Memory for items and memory for relations in the procedural/declarative memory framework. *Memory*, 5(1/2):131–178.

Connolly, C. and Reilly, R. (2005). A proposed model of repetition blindness. In Cangelosi, A., Bugmann, G., and Borisyuk, R., editors, *Modelling language, cognition and action: Proceedings of the 9th Neural Compuation and Psychology Workshop*, pages 279–288. World Scientific.

Cossart, R., Aronov, D., and Yuste, R. (2003). Attractor dynamics of network up states in the neocortex. *Nature*, 423(6936):283–288.

Cruse, H. and Wehner, R. (2011). No need for a cognitive map: Decentralized memory for insect navigation. *PLoS Computational Biology*, 7(3):e1002009.

Davidson, P. S. R. and Glisky, E. L. (2002). Is flashbulb memory a special instance of source memory? Evidence from older adults. *Memory*, 10(2):99–111.

Dodd, W. and Gutierrez, R. (2005). The role of episodic memory and emotion in a cognitive robot. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 692–697.

Doeller, C. F., Barry, C., and Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, 463(7281):1476–4687.

Downs, R. M. and Stea, D. (1973). *Cognitive maps and spatial behaviour. Process and products*, pages 8–26. Aldine, Chicago.

Dragoi, G. and Buzaski, G. (2006). Temporal encoding of place sequences in hippocampal cell assemblies. *Neuron*, 50:145–157.

Eccles, J. C. (1986). Chemical transmission and Dale's principle. *Progress in Brain Research*, 68:3–13.

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, 1(1):41–50.

Eichenbaum, H. (2004). Hippocampuscognitive processes and neural representations that underlie declarative memory. *Neuron*, 44(1):109–120.

Elman, J. L. (1991). Distributed Representations, Simple Recurrent Networks, And Grammatical Structure. *Machine Learning*, 7(2–3):195–225.

Engel, A. K., Knig, P., and Singer, W. (1991). Direct physiological evidence for scene segmentation by temporal coding. *Proceedings of the National Academy of Sciences of the United States of America*, 88(20):9136–9140.

Fadlalla, A. and Lin, C.-H. (2001). An analysis of the applications of neural networks in finance. *Interfaces*, 31(4):112–122.

Federmeier, K. D. and Kutas, M. (2001). Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):202–24.

Federmeier, K. D., Segal, J. B., Lombrozo, T., and Kutas, M. (2000). Brain responses to nouns, verbs and class-ambiguous words in context. *Brain*, 123(12):2552–2566.

Feldman, J. and Ballard, D. (1982). Connectionist models and their properties. *Cognitive Science*, 6(3):205–254.

Foo, P., Warren, W. H., Duchon, A., and Tarr, M. J. (2005). Do humans integrate routes into a cognitive map? map versus landmark-based navigation of novel shortcuts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2):195–215.

Fortin, N. J., Agster, K. L., and Eichenbaum, H. B. (2002). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience*, 5(5):458–462.

Fransen, E. and Lansner, A. (1998). A model of cortical associative memory based on a horizontal network of connected columns. *Network: Computation in Neural Systems*, 9(2):235–264.

Funahashi, S. (2001). Neuronal mechanisms of executive control by the prefrontal cortex. *Neuroscience Research*, 39:147–165.

Fuster, J. M. (1999). *Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate.* MIT Press.

Gan, K. and Lua, K. (1992). Chinese character classification using an adaptive resonance network. *Pattern Recognition*, 25(8):877–882.

Garagnani, M., Wennekers, T., and Pulvermüller, F. (2009). Recruitment and consolidation of cell assemblies for words by way of Hebbian learning and competition in a multi-layer neural network. *Cognitive Computation*, 1(2):160–176.

Gebhardt, N. et al. (2009). Irrlich engine - a free open source 3D engine. http://irrlicht.sourceforge.net.

Gennari, S. P., MacDonald, M. C., Postle, B. R., and Seidenberg, M. S. (2007). Context-dependent interpretation of words: Evidence for interactive neural processes. *NeuroImage*, 35(3):1278–1286.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Gerstner, W. (2002). Integrate-and-fire neurons and networks. *The handbook of brain theory and neural networks*, 2:577–581.

Gerstner, W. and Kistler, W. K. (2002). Mathematical formulations of Hebbian learning. *Biological Cybernetics*, 87(5–6):404–415.

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsaki, G., and Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience*, 12(10):1222–1223.

Goldberg, A. (2006). *Constructions at Work*. Oxford University Press, Oxford, Oxfordshire.

Gothard, K. M., Skaggs, W. E., Moore, K. M., and Mcnaughton, B. L. (1996). Binding of hippocampal CA1 neural activity to multiple reference frames in a landmark-based navigation task. *The Journal of Neuroscience*, 16(2):823–35.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: II. feedback, expectation, olfaction, illusions. *Biological Cybernetics*, 23(4):187–202.

Hagiwara, M. (1990). Multi-dimensional associative memory. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 3–6.

Harris, K. (2005). Neural signatures of cell assembly organization. *Nature Reviews Neuroscience*, 6:399–407.

Hartley, T., Bird, C. M., Chan, D., Cipolotti, L., Husain, M., Vargha-Khadem, F., and Burgess, N. (2007). The hippocampus is required for short-term topographical memory in humans. *Hippocampus*, 17(1):34–48.

Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Moyneus, P., and Maguire, E. (2009). Decoding neuronal ensembles in the human hippocampus. *Current Biology*, 19:546–554.

Hauk, O., Johnsrude, I., and Pulvermller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

Hebb, D. O. (1949). *The organization of behavior.* Wiley, New York.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception (Vol. 2): computation, learning, architectures*, pages 65–93, Orlando, USA. Harcourt Brace & Co.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19:1:103–120.

Hodges, J. R., Patterson, K., Oxbury, S., and Funnell, E. (1992). Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain : A journal of neurology*, 115 (Pt 6):1783–1806.

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544.

Hopfield, J. and Tank, D. (1985). Neural computation of decisions in optimization problems. *Biological Cybernetics*, 52(3):141–152.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Hoshino, O. and Zheng, M. (2004). Dynamic neuronal information processing of vowel sounds in auditory cortex. In Rajapakse, J. and Wang, L., editors, *Neural Information Processing: Research and Development*, pages 19–38. Springer-Verlag, Berlin.

Huyck, C. (2000). A practical system for human-like parsing. In Horn, W., editor, *Proceedings of the 14th European Conference on Artificial Intelligence.*

Huyck, C. R. (2001). Cell assemblies as an intermediate level of cognition. *Emergent Neural Computational Architectures Based on Neuroscience*, 2036/2001:383–397.

Huyck, C. R. (2004). Overlapping cell assemblies from correlators. *Neurocomputing*, 56:435–9.

Huyck, C. R. (2007). Creating hierarchical categories using cell assemblies. In *Connection Science*, volume 19:1, pages 1–24.

Huyck, C. R. (2008). CABot1: A videogame agent implemented in fLIF neurons. In *Cybernetic Intelligent Systems, 7th IEEE International Conference*, pages 1–6, London.

Huyck, C. R., Belavkin, R., Jamshed, F., Nadh, K., Passmore, P., Byrne, E., and Diaper, D. (2011). CABot3: A simulated neural games agent. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, Barcelona, Spain.

Huyck, C. R. and Nadh, K. (2009). Multi-associative memory in fLIF cell assemblies. In A. Howes, D. Peebles, R. C. E., editor, *9th International Conference on Cognitive Modelling*, pages 81–87, Manchester, UK.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, 14(6):1569–1572.

Kable, J. W., Lease-spellmeyer, J., and Chatterjee, A. (2002). Neural substrates of action event knowledge. *Journal of Cognitive Neuroscience*, 14(5):795–805.

Kangas, J., Kohonen, T., Laaksonen, J., Simula, O., and Ventä, O. (1989). Variants of self-organizing maps. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 517–522. IEEE Service Center.

Kaplan, S., Sonntag, M., and Chown, E. (1991). Tracing recurrent activity in cognitive elements (TRACE): A model of temporal dynamics in a cell assembly. *Connection Science*, 3:179–206.

Kapur, S., Craik, F., Tulving, E., Wilson, A., Houle, S., and Brown, G. (1994). Neuroanatomical correlates of encoding in episodic memory: Levels of processing effect. *Proceedings of the National Academy of Sciences*, 91(6):2008–2011.

Kelso, S. R., Ganong, A. H., and Brown, T. H. (1986). Hebbian synapses in hippocampus. *Proceedings of the National Academy of Sciences*, 83(14):5326–5330.

Kent, S. L. (2002). *The ultimate history of video games.* Prima Life, Roseville, California.

Keysers, C. and Perrett, D. I. (2002). Visual masking and RSVP reveal neural competition. *Trends in Cognitive Sciences*, 6(3):120–125.

Kiehl, K. A., Liddle, P. F., Smith, A. M., Mendrek, A., Bruce, and Hare, R. D. (1999). Neural pathways involved in the processing of concrete and abstract words. *Human Brain Mapping*, 7(4):225–233.

Knoblauch, A., Kupper, R., Gewaltig, M.-O., Körner, U., and Körner, E. (2007). A cell assembly based model for the cortical microcircuitry. *Neurocomputing*, 27:649–677.

Koenig, T., Studer, D., Hubl, D., Melie, L., and Strik, W. (2005). Brain connectivity at different time-scales measured with EEG. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1015–1024.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Kohonen, T. (1990). The self-organizing map. In *Proceedings of the IEEE*, volume 78, pages 1464–1480.

Kolen, J. F. and Pollack, J. F. (1991). Multiassociative memory. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pages 7–10.

Kosko, B. (1988). Bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):49–60.

Kreiman, G., Koch, C., and Fried, I. (2000). Imagery neurons in the human brain. *Nature*, 408(6810):357–361.

Krogh, A. (2008). What are artificial neural networks? *Nature biotechnology*, 26(2):195–197.

Kuligowski, R. J. and Barros, A. P. (1998). Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather and Forecasting*, 13(4):1194–1204.

LaBar, K. S., Gitelman, D. R., Parrish, T. B., and Mesulam, M. (1999). Neuroanatomic overlap of working memory and spatial attention networks: A functional MRI comparison within subjects. *Neuroimage*, 10:695–704.

Lansner, A. (2009). Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. *Trends in Neuroscience*, 32:3:178–186.

Le Clec'H, G., Dehaene, S., Cohen, L., Mehler, J., Dupoux, E., Poline, J. B., Lehericy, S., Moortele, P. F. V. D., and Bihan, D. L. (2000). Distinct cortical areas for names of numbers and body parts independent of language and input modality. *Neuroimage*, 12:381–391.

Lee, C., Lee, G., and Yun, S. J. (2000). Automatic WordNet mapping using word sense disambiguation. In *Proceedings of the 2000 Joint SIGDAT conference on*

Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13, EMNLP '00, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leutgeb, J. K., Leutgeb, S., Treves, A., Meyer, R., Barnes, C. A., Mcnaughton, B. L., Moser, M. B., and Moser, E. I. (2005). Progressive transformation of hippocampal neuronal representations in "morphed" environments. *Neuron*, 48(2):345–358.

Levy, N. and Horn, D. (1999). Associative memory in a multi-modular network. *Neural Computation*, 11:1717–1737.

Lin, D. (1998). Dependency-based evaluation of MINIPAR. In *Proceeds of the Workshop on the Evaluation of Parsing Systems*, Granada.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H., and Matsudaira, P. (1999). *Molecular Cell Biology (Lodish, Molecular Cell Biology).* W. H. Freeman, 4th edition.

Maass, W. and Bishop, C. M. (2001). *Pulsed Neural Networks.* MIT Press.

Mackintosh, N. (2002). Do not ask whether they have a cognitive map, but how they find their way about. *Psicologia*, 23:165–185.

Malsburg, C. v. d. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5(4):520–526.

Maratos, E. J., Dolan, R. J., Morris, J. S., Henson, R. N. A., and Rugg, M. D. (2001). Neural activity associated with episodic memory for emotional context. *Neuropsychologia*, 39(9):910–920.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19:2:313–330.

Martin, A. and Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, 11(2):194–201.

Maurer, A. P., Cowen, S. L., Burke, S. N., Barnes, C. A., and McNaughton, B. L. (2003). Organization of hippocampal cell assemblies based on theta phase precession. *Hippocampus*, 16(9):785–794.

McCloskey, M., Wible, C. G., and Cohen, N. J. (1988). Is there a special flashbulb-memory mechanism? *Journal of Experimental Psychology: General*, 117(2):171–181.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysic*, 5:115–133.

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Milner, B., Squire, L. R., and Kandel, E. R. (1998). Cognitive neuroscience and the study of memory. *Neuron*, 20(3):445–468.

Minsky, M. L. (1954). *Theory of Neural-Analog Reinforcement Systems and its Application to the Brain-Model Problem.* PhD thesis, Princeton University.

Molinari, M., Leggio, M. G., Solida, A., Ciorra, R., Misciagna, S., Silveri, M. C., and Petrosini, L. (1997). Cerebellum and procedural learning: evidence from focal cerebellar lesions. *Brain*, 120(10):1753–1762.

Molnár, G., Oláh, S., Komlósi, G., Füle, M., Szabadics, J., Varga, C., Barz, P., and Tamás, G. (2008). Complex events initiated by individual spikes in the human cerebral cortex. *PLoS Biology*, 6(9):222.

Moss, H. E., Ostrin, R. K., Tyler, L. K., and Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4):863–883.

Mummery, C. J., Patterson, K., Hodges, J. R., and Price, C. J. (1998). Functional neuroanatomy of the semantic system: Divisible by what? *Journal of Cognitive Neuroscience*, 10(6):766–777.

Nadh, K. and Huyck, C. R. (2009). Prepositional phrase attachment ambiguity resolution using semantic hierarchies. In *The Ninth IASTED International Conference on Artificial Intelligence and Applications*, Innsbruck, Austria.

Nadh, K. and Huyck, C. R. (2010). A pong playing agent modelled with massively overlapping cell assemblies. *Neurocomputing*, 73(16–18):2928–2934.

Nakov, P. and Hearst, M. (2005). Using the web as an implicit training set: Application to structural ambiguity resolution. In *Proceedings of HLT-NAACL*.

Nardini, M., Jones, P., Bedford, R., and Braddick, O. (2008). Development of cue integration in human navigation. *Current biology : CB*, 18(9):689–693.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4(2):135–183.

Nieuwenhuis, S., Heslenfeld, D. J., von Geusau, N. J. A., Mars, R. B., Holroyd, C. B., and Yeung, N. (2005). Activity in human reward-sensitive brain areas is strongly context dependent. *NeuroImage*, 25(4):1302–1309.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map.* Oxford University Press, New York.

O'Neill, J., Senior, T. J., Allen, K., Huxter, J. R., and Csicsvari, J. (2008). Reactivation of experience-dependent cell assembly patterns in the hippocampus. *Nature Neuroscience*, 11(2):209–215.

O'Reilly, R. C. and Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, 108(2):311–345.

Osan, R., Chen, G., Feng, R., and Tsien, J. Z. (2011). Differential consolidation and pattern reverberations within episodic cell assemblies in the mouse hippocampus. *PLoS ONE*, 6(2):e16507.

Paik, J. K. and Katsaggelos, A. K. (1992). Image restoration using a modified hopfield network. *IEEE Transactions on Image Processing*, 1(1):49–63.

Palm, G. (1991). Memory capacities of local rules for synaptic modification. a comparative review. *Concepts in Neuroscience*, 2.

Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, 321(5894):1322–1327.

Pasupathy, A. and Connor, C. (2002). Population coding of shape in area V4. *Nature Neuroscience*, 5:12:1332–1338.

Pavlov, I. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Translated by Anrep GV*. London Oxford University Press.

Pershin, Y. V. and Di Ventra, M. (2010). Experimental demonstration of associative memory with memristive neural networks. *Neural Netw.*, 23(7):881–886.

Peyrache, A., Benchenane, K., Khamassi, M., Wiener, S., and Battaglia, F. (2010). Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution. *Journal of Computational Neuroscience*, 29(1):309–325.

Plenz, D. and Thiagarajan, T. C. (2007). The organizing principles of neuronal avalanches: cell assemblies in the cortex? *Trends in Neurosciences*, 30(3):101–110.

Poldrack, R. A. and Packard, M. G. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia*, 41(3):245–251.

Pulvermuller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, 22:253–336.

Quillian, M. (1967). Word concepts: A theory of simulation of some basic semantic capabilities. *Behavioral Science*, 12:410–30.

Ramamaurthy, U., D'Mello, S., and Franklin, S. (2004). Modified sparse distributed memory as transient episodic memory for cognitive software agents. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 6, pages 5858–5863.

Ratnaparkhi, A., Reynar, J., and Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 250–255.

Rickel, J. and Johnson, W. L. (2000). Task-oriented collaboration with embodied agents in virtual worlds. In *Embodied conversational agents*, pages 95–122. MIT Press, Cambridge, MA, USA.

Rolls, E. T. (2007). An attractor network in the hippocampus: Theory and neurophysiology. *Learning & Memory*, 14(11):714–731.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan Book.

Rost, B. (2001). Review: Protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134:204–218.

Rudy, J. W. and O'Reilly, R. C. (2001). Conjunctive representations, the hippocampus, and contextual fear conditioning. *Cognitive, Affective, & Behavioral Neuroscience*, 1(1):66–82.

Rumelhart, D. and McClelland, J. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement and some tests and extensions of the model. *Psychological Review*, 89:1:60–94.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). *Learning internal representations by error propagation*, pages 318–362. MIT Press, Cambridge, MA, USA.

Sakurai, Y. (1996). Hippocampal and neocortical cell assemblies encode memory processes for different types of stimuli in the rat. *The Journal of Neuroscience*, 16(8):2809–2828.

Sakurai, Y. (1998). The search for cell assemblies in the working brain. *Behavioural Brain Research*, 91:1–13.

Sakurai, Y. (1999). How do cell assemblies encode information in the brain? *Neuroscience & Biobehavioral Reviews*, 23(6):785 – 796.

Samsonovich, A. and Mcnaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *The Journal of Neuroscience*, 17(15):5900–5920.

Schüz, A. (1998). Neuroanatomy in a computational perspective. In Arbib, M. A., editor, *The handbook of brain theory and neural networks*, pages 622–626. MIT Press, Cambridge, MA, USA.

Shirvalkar, P. R. (May 2009). Hippocampal neural assemblies and conscious remembering. *Journal of Neurophysiology*, 101(5):2197–2200.

Shohamy, D. and Wagner, A. D. (2008). Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron*, 60(2):378–389.

Smolensky, P. (1987). Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, 1(2):95–109.

Song, X.-H., Hopke, P. K., Fergenson, D. P., and Prather, K. A. (1999). Classification of single particles analyzed by atofms using an artificial neural network, art-2a. *Analytical Chemistry*, 71(4):860–865.

Spatz, H.-C. (1996). Hebb's concept of synaptic plasticity of neuronal cell assemblies. *Behavioural Brain Research*, 78:3–7.

Spiers, H. J. and Maguire, E. A. (2008). The dynamic nature of cognition during wayfinding. *Journal of Environmental Psychology*, 28(3):232–249.

Squire, L. R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological review*, 99(2):195–231.

Stainslaw Jankowski, Andrzej Lozowski, J. M. Z. (1996). Complex-valued multistate neural associative memory. *IEEE Transactions on Neural Networks*, 7(6):1491–1496.

Stetina, J. and Nagao, M. (1997). Corpus based PP attachment ambiguity resolution with a semantic dictionary. In *Proceedings of Workshop on Very Large Corpora*, pages 66–80.

Sturz, B. R., Bodily, K. D., and Katz, J. S. (2006). Evidence against integration of spatial maps in humans. *Animal Cognirion*, 9(3):207–17.

Tal, D. and Schwartz, E. L. (1997). Computing with the leaky integrate-and-fire neuron: logarithmic computation and multiplication. *Neural Computation*, 9(2):305–318.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19(1):109–139.

Tanila, H., Shapiro, M. L., and Eichenbaum, H. (1997). Discordance of spatial representation in ensembles of hippocampal place cells. *Hippocampus*, 7(6):613–623.

Tervaniemi, M., Maury, S., and Näätänen, R. (1994). Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. *Neuroreport*, 5(7):844–846.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4):189–208.

Toutanova, K., Y., A. N., and Manning, C. (2004). Learning random walk models for inducing word dependency distributions. In *Proceedings of International Conference on Machine Learning*.

Tulving, E. (1985). How many memory systems are there? *American Psychologist*, 40(4):385–398.

Tulving, E. and Donaldson, W. (1972). Episodic and semantic memory. In *Organization of memory*, pages 381–402. Academic Press.

Tulving, E. and Markowitsch, H. J. (1998). Episodic and declarative memory: Role of the hippocampus. *Hippocampus*, 8(3):198–204.

Tulving, E. and Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):359–380.

Tversky, B. (1993). Cognitive maps, cognitive collages, and spatial mental models. In Frank, A. U. and Campari, I., editors, *Spatial Information Theory: A Theoretical Basis for GIS*, pages 14–24, Berlin. Springer Verlag.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition*, 92(1–2):231–270.

Ullman, M. T., Corkin, S., Coppola, M., and Hickok, G. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, 9:266–276.

Valiant, L. G. (2005). Memorization and association on a realistic neural model. *Neural Computation*, 17:527–5557.

Vandenberghe, R., Price, C., Wise, R., Josephs, O., and Frackowiak, R. S. J. (1996). Functional anatomy of a common semantic system for words and pictures. *Nature*, 383 (6597):254–256.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600.

Vicario-Abejn, C., Collin, C., McKay, R. D. G., and Segal, M. (1998). Neurotrophins induce formation of functional excitatory and inhibitory synapses between cultured hippocampal neurons. *The Journal of Neuroscience*, 18(18):7256–7271.

Vinson, D. P., Vigliocco, G., Cappa, S., and Siri, S. (2003). The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain and Language*, 86(3):347–365.

Wagner, A. D., Desmond, J. E., Glover, G. H., and Gabrieli, J. D. E. (1998). Prefrontal cortex and recognition memory: fMRI evidence for context-dependent retrieval processes. *Brain*, 121:1985–2002.

Wang, R. F. and Spelke, E. S. (2002). Human spatial representation: insights from animals. *Trends in Cognitive Sciences*, 6(9):376–382.

Warburton, E., Wise, R. J., Price, C. J., Weiller, C., Hadar, U., Ramsay, S., and Frackowiak, R. S. (1996). Noun and verb retrieval by normal subjects. Studies with PET. *Brain*, 119 (Part 1):159–79.

Wennekers, T. (2007). A cell assembly model for complex behaviour. *Neurocomputing*, 70(10–12):1988–1992.

Wennekers, T. (2009). On the natural hierarchical composition of cliques in cell assemblies. *Cognitive Computation*, 1(2):128–138.

Wennekers, T. and Palm, G. (2000). Cell assemblies, associative memory and temporal structure in brain signals. In *R.Miller, editor, Time and the Brain. Conceptual advances in Brain Research, Vol. 2*, pages 251–274. Harwood Academic Publishers.

Wickelgren, W. A. (1999). Webs, cell assemblies, and chunking in neural nets. *Canadian Journal of Experimental Psychology*, 53(1):118–131.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record, Part 4*, pages 96–104, New York. IRE.

Wills, T. J., Lever, C., Cacucci, F., Burgess, N., and O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876.

Willshaw, D., Buneman, O., and Longuet-Higgins, H. (1969). Non-holographic associative memory. *Nature*, 222:960–962.

Yoshizawa, S., Morita, M., and Amari, S.-I. (1993). Capacity of associative memory using a nonmonotonic neuron model. *Neural Networks*, 6(2):167–176.

Zhao, S. and Lin, D. (2004). A nearest-neighbor method for resolving PP-attachment ambiguity. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 428–434.

Index

Appendix

Following are the URLs for downloading various models described in Chapter 5.

Emergent context sensitivity in an associative memory (Section 5.1)
— http://nadh.in/research/files/model_context.zip

Spatial cognitive mapping with a sequential associative memory in an embodied agent in a virtual environment (Section 5.2)
— http://nadh.in/research/files/model_cogmap.zip

Novel behaviour from a massively overlapping associative memory in a game playing agent (Section 5.3)
— http://nadh.in/research/files/model_pong.zip

Natural language disambiguation with an associative memory of semantic hierarchies (Section 5.4)
— http://nadh.in/research/files/model_ppca.zip