UNIVERSITY OF CALIFORNIA, MERCED

A Review of Attractor Neural Networks and Their Use in Cognitive Science

A Thesis submitted in partial satisfaction of the requirements for the degree of Master of Science

in

Cognitive and Information Sciences

by

Makenzy Lee Gilbert

Committee in charge:

Professor Jeffrey Yoshimi, Chair
Professor David Noelle
Professor Colin Holbrook

The Thesis of Makenzy Lee Gilbert is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

Jeffrey Yoshimi - Chair

_____

David Noelle

_____

Colin Holbrook

University of California, Merced 2024

Table of Contents

List of Figures

Acknowledgements

I want to express my gratitude to my parents who with their encouragement and support have given me the opportunity to succeed. I am forever grateful for everything you have done.

I would also like to acknowledge the contributions of my committee members, Dr. David Noelle and Dr. Colin Holbrook, your assistance and insights have been crucial in shaping this thesis and academic development.

Additionally, I express my deepest appreciation to my committee chair, Dr. Jeffrey Yoshimi, who has been instrumental in my academic journey. Without your initial enthusiasm and passion in my undergraduate studies, I would never have pursued higher education, let alone in this specific domain. I am fortunate to have your support, and your continued mentorship is invaluable.

Lastly, I would like to thank myself for my resilience. This thesis stands as a testament to my unwavering commitment to my goals and my determination to make the best out of any challenge I face.

Abstract

A Review of Attractor Neural Networks and Their Use in Cognitive Science

by Makenzy Lee Gilbert for the partial satisfaction of the requirements for the degree of Masters of Science in Cognitive and Information Sciences, University of California, Merced, 2024

Dr. Jeffrey Yoshimi, Chair

This literature review explores the role of attractor neural networks (ANNs) in modeling psychological processes in artificial and biological systems. By synthesizing research from dynamical systems theory, psychology, and computational neuroscience, the review provides an overview of the current understanding of ANN function in memory formation, memory reinforcement, retrieval, and forgetting. Key mathematical foundations of ANNs, including dynamical systems theory and energy functions, are discussed to explain the behavior and stability of these networks. The review also examines empirical applications of ANNs in cognitive processes such as semantic memory and episodic recall, as well as highlighting the hippocampus' role in pattern separation and completion. The review addresses challenges like catastrophic forgetting and noise effects on memory retrieval. By identifying gaps between theoretical models and empirical findings, it highlights the interdisciplinary nature of ANN research and suggests future areas for exploration.

Introduction

Attractor neural networks are used to study memory processes in both artificial and biological systems. This literature review aims to synthesize the fragmented research on attractor neural networks, specifically their role in memory-related processing. Integrating findings from multiple domains, including dynamical systems theory, psychology, and computational neuroscience, this review will provide a comprehensive view of the present understanding of how attractor networks function and their relevance to memory formation, reinforcement, retrieval, and forgetting. Understanding memory through attractor neural networks is important for theoretical and applied sciences. In theoretical neuroscience, these models help describe the core principles of neural dynamics and memory encoding. In applied contexts, insights from attractor networks can inform the development of artificial systems and therapeutic strategies for memory-related disorders. Reviewing where the literature currently stands is crucial, highlighting the interdisciplinary nature of this research and detecting gaps between the theoretical models and empirical findings to identify areas where further exploration is needed.

Neural networks are computational models that are inspired by the human brain.These models contain layers of connected “neurons” or nodes that transmit information. These networks are designed to process information; each connection between the nodes has a weight that is updated as the network “learns”. This allows a model to iteratively improve performance over time.

An attractor neural network is a specific type of neural network, In an ANN, certain patterns of neural activity become stable states called “attractors.” When one of these networks receives partial input of a previously stored pattern, such as a picture missing parts, the network can complete the pattern by settling into one of these attractor states. This capacity to recall entire patterns from fragmented input makes ANN models useful for studying memory processes in cognitive science. By examining how these networks learn, retrieve, and forget memories, researchers have gained insights into the fundamental principles of memory dynamics and its neural realization.

1. History

Perhaps the best-known early study of recurrent attractor networks is the work of John Hopfield in the 1980s. He developed simple recurrent networks and studied how they could be used to model associative memory. These have since become known as “Hopfield Networks”. The networks demonstrate how neural networks store and retrieve information through the system-wide interactions of binary neurons.

Hopfield used mathematical tools from dynamical systems theory and physics (in particular thermodynamics) to study these networks. He described memories as stable states or attractors the system naturally evolves towards. The attractors are minima in an energy landscape.

Interactive Activation and Competition (IAC) networks like attractor networks, use recurrent connections to stabilize activation patterns, but they emphasize the competitive interactions between units. Within the same layer, units can inhibit each other, which helps to resolve ambiguities and select the most appropriate representation. This competitive dynamic is not seen in standard attractor networks, which primarily focus on achieving stable states (Rumelhart, McClelland, & PDP Research Group, 1986).

Paul Smolensky and the PDP research group's work in the PDP volumes lays out a fundamental example of how attractor neural networks can be applied in cognitive tasks. Smolensky's room schema model describes how an ANN can recall layouts of specific rooms. The model uses memory representations in the form of attractors, allowing for the filling-in of missing information much like humans can visualize or remember an entire room from a few details or in dim lighting (Rumelhart, Smolensky, McClelland, & Hinton, 1986).

Around the same time, Ackley, Hinton, and Sejnowski introduced Boltzmann machines, a different class of networks similar to Hopfield networks, but with stochastic elements (Ackley et al, 1985). Around this time Smolensky developed similar ideas, in particular what he called the harmonium (Smolensky, 1986). This work laid the foundation for later developments, such as Restricted Boltzmann Machines (RBMs) (Hinton & Salakhutdinov, 2006), which separate the network into a visible and hidden layer. These networks were ultimately more popular because they could be trained using fast learning algorithms (Hinton, 2006). In 1989, Daniel J. Amit built on these ideas in a book-length manuscript (Amit, 1989), clarifying the mathematics and unpacking the connection between recurrent networks and memory encoding and retrieval. Subsequent researchers like McRae (McRae et. al, 1997), explored the application of these networks to semantic memory and other cognitive functions. For instance, McRae's work demonstrated how correlated features within concepts help the network settle on the correct meaning faster, which is crucial for tasks like semantic priming and feature verification.

The Seidenberg "triangle model" is a framework that describes how a collection of layers in a larger model cooperate to process information. It has three main layers that represent orthography (spelling), phonology (sound), and semantics (meaning). The original model is connectionist and emphasizes how these three types of representations interact through learned connections (Seidenberg and McClelland, 1989). It was one of the first models to explain how the brain processes written and spoken language using parallel distributed processing. The multiple hidden layers connect orthography, phonology, and semantics. When certain limitations arose in the original model, recurrent connections were added that produced attractor dynamics in some layers (Plaut et. al, 1996; Monaghan, Chang, & Welbourne, 2017). These attractors work within this architecture by ensuring that each representation (spelling, sound, meaning) can activate the correct patterns in the other representations. See Figure 4 for a diagram of the model and its interacting parts. For example, seeing the orthographic representation of a word can lead to the correct phonological and semantic representations, due to the stabilizing influence of the attractor networks.

Researchers in cognitive and computational neuroscience (CCN) have continued to explore and validate these attractor models in the context of psychological phenomena. This includes studies on the role of the hippocampus in pattern separation and completion, which are essential for episodic memory and associative recall.

In applied contexts, attractor networks offer insight into disorders and mental health. Often the stability of these attractor states provides an understanding of conditions such as schizophrenia, a mental disorder that includes delusions, hallucinations, and disorganized thinking. Studies indicate that shallow attractors cause disorganized thought. At the same time, psychedelic research suggests that psychedelic intervention may improve mental health through the adaptive influences of entropy that allow the brain to explore more attractors in a landscape to reduce dysfunctional thinking patterns. These contrasting views call for further study on attractor neural networks in memory and cognition (Musa et al., 2022; Hipólito et al., 2023).

2. Mathematical and Computational Foundations

Examining the formal mathematical basis of ANNs including dynamical systems theory (DST) is crucial for understanding how these networks can be used to model memory. DST is the study of systems that evolve over time under deterministic rules. Continuous time systems are represented by differential equations, whereas discrete time systems are represented by difference equations. Both types of equations can be implemented on a computer. Differential equations are numerically integrated and both types of systems involve repeatedly iterating an update function. This means solutions are not exact, but approximated at discrete points.

DST involves studying the evolution of points within a state space over time. A state space represents all possible states of a system, each point representing a potential state of the system(Yoshimi et al., 2023). For an n-dimensional continuous system, the state space is R^n. Each dimension corresponds to one variable necessary to describe the state of the system. The equations can be quite complicated and nonlinear which can lead to complex chaotic behaviors. Given a dynamical system, we can take any initial state of the system and define trajectories from that state. Using the system we can fill the state space with trajectories, resulting in a phase portrait. The phase portrait gives us an immediate sense of the dynamics of a system. The nice thing about them is they give us a visual way to understand dynamics that are otherwise somewhat hidden in the abstract equations.

An important kind of behavior in a system is an attractor, a state or set of states that trajectories tend to lead toward. An attractor represents a set of states that a system will naturally progress towards. Attractors can be stable points, limit cycles, or chaotic strange attractors. The key characteristic of an attractor is the ability to metaphorically draw and pull on nearby trajectories. This is evident in Figure 1, where the fixed point attractors in red represent stable states. Each attractor has its own basin of attraction. The two basins in this example are separated by the green basin boundary. This means that on either side of the boundary, an initial condition will follow a trajectory towards the corresponding attractor.

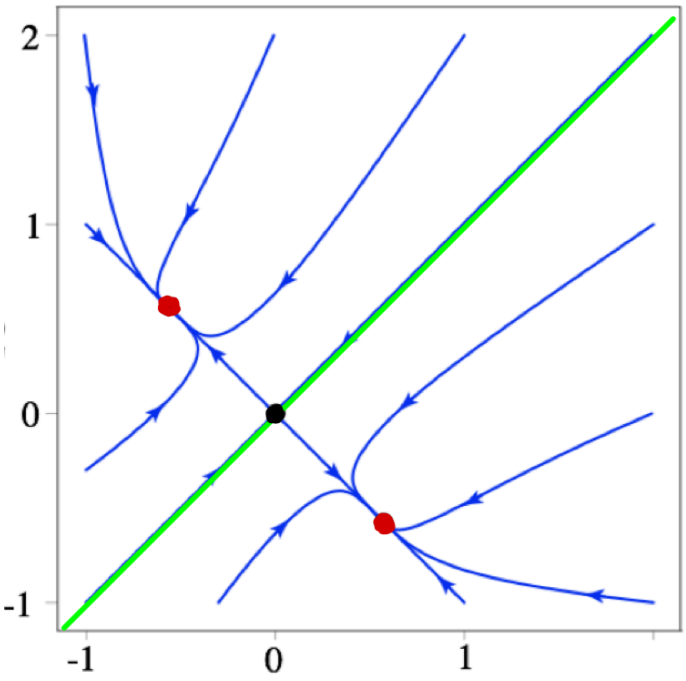


Figure 1. Phase portrait of a two-node recurrent network. Two fixed-point attractors and their respective basins of attraction. The attractors in red are stable states which the system moves towards, given the dynamics. Each attractor has its own basin of attraction, depicted by the surrounding trajectories that converge towards these fixed points. The green line is the basin boundary and the triangular area above it is the basin for the attractor in the upper left and the triangular area below it is the basin for the attractor on the lower right (Yoshimi et al., 2023).

In the context of neural networks, state variables usually correspond to activations of nodes. The total patterns of activation across a set of nodes is the state of the system at a time. This state can be thought of as a point in the state space, which is here an activation space, the set of all possible activation patterns of the network. In certain attractor networks like a Hopfield network, attractors are patterns of activation that the network will tend to settle into over time. Parameters are another kind of variable in a dynamical system, that is fixed while the system runs. In a recurrent neural network, weights are often treated as parameters, which stay fixed while studying the attractors and orbits in a state space, but which are allowed to vary during learning. When there is a change to a system's parameters, you can see changes in the total set of trajectories within a state space. Changes that introduce new attractors are called bifurcations (Yoshimi et al., 2023).

Daniel J. Amit's work in 1992 studied how recurrent networks are easily interpreted as dynamical systems where neural activity patterns progress towards stable states known as attractors. As with Hinton and Sejnowski, Amit also introduced stochastic elements into these network models, explaining how noise and random perturbations influence the dynamics of memory storage and retrieval. This is crucial for understanding real-world neural behavior, where noise is always present.

Applied to neural networks, each state is a pattern of neural activity and each attractor corresponds with a stable pattern of activity within the network. These stable patterns can represent memories or cognitive states. The overall network dynamics allow convergence to these attractor states, which enables them to complete partial or noisy information. Because of this, attractor networks are capable of recall and recognition. These concepts as they are instantiated in attractor networks are discussed further below.

Often it is difficult to study a dynamical system and to find its attractors directly, and so an indirect method is used. Functions are defined on the state space that can be used to find stable attracting points. One example of such a function is the Lyapunov function. For a system described by a set of differential equations, where *x* is a state variable that varies with time *t*, the Lyapunov function $V(x)$ must satisfy these conditions:

1. $V(x) > 0 \; for \; all \; x \neq 0 \; and \; V(0)=0$
2. The derivative of $V(x)$ along the trajectories of the system denoted as $\frac{dV}{dt}$ must be less than or equal to zero, $\frac{dV}{dt} \leq 0$ (Perelson, Oster, & Katchalsky, 1976).

Convergence properties for Hopfield networks have been proven, which show that under certain conditions, the trajectory of a dynamical system will converge to an attractor. This means that regardless of the initial state, the system will eventually reach a stable state. This reinforces the use of Lyapunov functions in predicting the long-term behavior of these systems (Bruck, 1990).

One prominent type of Lyapunov function is an energy function. Lyapunov functions like energy functions are valuable because they facilitate certain mathematical analyses but they also have a nice visualization. They can be visualized as graphs above the state space of a dynamical

system, what are often called energy surfaces or energy landscapes. The attractors can be thought of as “low points” in these energy landscapes.

This kind of picture motivates a standard metaphor, the ball and landscape metaphor, where we think of an initial state of a dynamical system (a pattern of activity in a neural network) as a ball that is placed on the surface and allowed to roll to a minimum which corresponds to an attractor. Figure 2 illustrates the ball and landscape metaphor, depicting the concept of attractors and basins of attraction within a two-dimensional coordinate space. The initial state of a dynamical system, represented as a ball, is placed on the surface and allowed to roll towards a minimum, corresponding to an attractor. The energy landscape shown above the state space helps us visualize how initial states progress towards attractor states. A standard confusion is between the attractors and basins themselves in the actual state space (bottom of figure) and the hills and valleys of the energy function V, which are “above” the state space and help us study it.

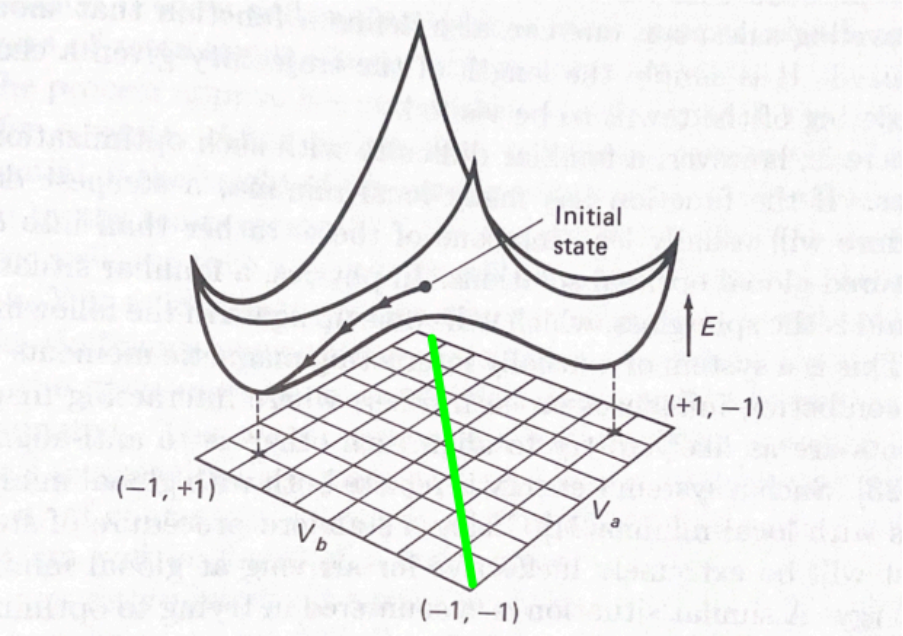


Figure 2. Amit's work emphasizes that the basin of attraction for a specific attractor includes the region in the state space where initial conditions lead to that attractor. The two crosses on the grid represent attractors, the green line representing the basin boundary separating two basins, and are distinct from the hills and valleys of the energy function, which is situated "above" the state space to aid in our understanding (Amit, 1989 p.89, Figure 2.11).

Again, a basin of attraction for a specific attractor is the region in the state space from which initial conditions will lead to that attractor, as in Figure 2. The efficiency of memory retrieval depends on the dynamics of these basins. Amit’s work discusses how the structure of these basins affects memory retrieval. A larger basin means that the network can withstand more noise and still retrieve the correct memory. He describes how the volume of the solution space in J-space (representing the strengths of connections in a network) is related to the basin of attraction for each memory pattern. As this volume increases, the more robust the memory retrieval process becomes. He further explains that spurious states, which are unintended stable

attractors that occur due to the nonlinearity of the dynamics, can affect stability but are typically less significant when the basins are large and the signal-to-noise ratio is favorable (Amit, 1989).

As a network attempts to store more patterns, it nears "memory saturation." At this point, the network's ability to store and retrieve memories becomes strained. Amit describes that as the number of stored patterns increases, the landscape of basins becomes more crowded. Additionally, more spurious become entangled with the intended memory states' basins and this makes for unreliable retrieval (Amit, 1989).

In a two-node network, the basin of attraction will be a boundary within the state space, this carves up where the boundary in the space for which initial conditions will evolve towards that attractor. In an energy landscape, the basin of attraction is represented differently. It is a mathematical visualization of the underlying dynamics, usually depicted as a surface where valleys correspond to attractors. This visualization helps us understand how the system behaves, showing where stable states are to be. However, it is important to note that this surface is just a tool to help us see the patterns; it doesn't represent the exact states of the attractors themselves. In continuous systems, evolution happens smoothly and basins are more complex than the distinct boundaries seen in discrete systems. Both systems benefit from using tools such as energy landscapes.

To gain a better understanding of these concepts, including attractors and energy landscapes we turn to the Hopfield model and describe it in more detail. Hopfield networks are a specific type of recurrent neural network (RNN). They have binary state neurons and symmetrical weights. The network nodes can be in one of two states: +1, or -1. The network is updated according to this equation:

$$s_i(t + 1) = sign\,(\Sigma_j w_{ij} s_j(t))$$

where *sign*(x) is defined as:

+1 if $x > 0$
-1 if $x < 0$
0 if $x = 0$

The dynamics of the network are visualized by an energy function E that decreases over time, which effectively describes the network's evolution towards a stable state.

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j - \sum_i \theta_i s_i$$

The Hopfield energy function E acts as a Lyapunov function because it either decreases or stays the same as the network updates. This ensures the system is stable and evolving towards a local minimum of the energy landscape. In a ball and landscape analogy, we can think of the height of the ball at the lowest point in a valley being the Lyapunov function $V(x)$. If the ball is at the bottom, the function is zero, which means this is a stable state. As the ball is pushed away from the lowest point, the height increases, thus moving away from a stable state.

Additionally, Hopfield networks parallel the Ising model from statistical mechanics, which describes ferromagnetism in materials. By viewing neural states as analogous to spin states and synaptic weights as interaction strengths, this comparison helps understand a

network's energy landscape and the stability of multiple attractors (Amit, 1989). This analogy offers valuable insights into their operation and stability. In an example system with three elements, each particle (analogous to a neuron) has a spin that can be either up (+1+1+1) or down (−1−1−1). The interaction between spins is described by the Hamiltonian (the energy function for this system):

$$H = -\sum_{\langle i,j \rangle} J_{ij} s_i s_j$$

where J_{ij} represents the interaction strength between spins i and j

Two concepts related to this model are entropy and free energy. Entropy measures the level of randomness in a system. In the present context, higher entropy indicates a larger number of possible neural configurations. Entropy is related to the probability distribution of network states. For a given probability distribution $\rho(t)$ of network states at time t, entropy *S* is defined as:

$$S = -\sum_I \rho_I(t) \ln \rho_I(t)$$

Entropy affects the stability of attractor states in memory. Higher entropy states tend to allow the network to be more adaptable to new information, while lower entropy states tend to maintain the stability of existing memory states (Hopfield, 1982; Amit, 1989).

Free energy is the combination of the entropy and the energy in the system at a time. Free energy *F* is defined as:

$$F = E - TS$$

As the network learns, *F* is reduced over time. This means that the system's energy *E* decreases or the product of temperature *T* and entropy *S* is controlled and minimized. The network gradually becomes more organized and efficient during learning. This is due to the network's adjustments to the weights reducing free energy, which in turn lowers the entropy in a controlled way (Amit, 1989).

Thus far the focus has been on attractors in a state space and energy functions used to identify and study those attractors. Recall that these are the state variables of a dynamical system. We now consider the parameters; the weights. They are updated using a learning algorithm. In the case of Hopfield networks, this learning algorithm is usually some version of the Hebb rule. Introduced by Donald Hebb in 1949, this principle proposes that the synaptic connection between two neurons is strengthened when they are activated at the same time (Yoshimi et al., 2023). The Hebbian learning rule can be expressed as:

$$\Delta w_{ij} = \eta \cdot x_i \cdot y_j$$

Where:

Δw_{ij} is the change in weight between one neuron *i* and another neuron *j*, η is the learning rate, or small constant parameter that determines how fast learning occurs, x_i is the activation of

the presynaptic neuron i, and y_j is the activation of the postsynaptic neuron j. The principle behind this rule is that repeated coactivation of neurons leads to stronger connections between them. Intuitively if the weight between neuron *I* and *J* is initially small, and if *I* is consistently active when neuron *J* is also active the weight between them will then increase. This means that in the future, activation of neuron *I* is more likely to lead to activation of neuron *J*.

For RNNs, the application of the Hebb rule often results in the introduction of a new fixed point attractor to the network. Whatever the activation pattern when the rule is applied will tend to become a new attractor. In the ball and landscape metaphor, it is as if a new valley is added wherever the ball happens to be. Thus, Hebbian learning can adjust weights to reinforce certain patterns of activity, which effectively "burns in" attractors that correspond to learned memories. Recall that in Hopfield networks weights are symmetric. What this does from a dynamical systems standpoint is prevent cycles from occurring; the network only learns fixed point attractors (Hopfield, 1982).

Concepts that would be especially useful when considering attractors as models of memory would be the "strength" of an attractor, where intuitively a strong attractor (relative to a set of attractors) is one with a larger basin of attraction than others in the set and where states are pulled to the attractor more rapidly (the energy landscape is deeper). Then the idea would be that with repeated learning of a pattern, it would get "stronger", "larger", and "deeper." Surprisingly there has been very little work on making these intuitive ideas precise, though the intuitive ideas are sometimes invoked (Kaneko, 1998; Zemel & Mozer 2001; Graves, Wayne, & Danihelka 2014; Deng et al., 2020,). As noted in the conclusion this is something I would like to work on in the future.

Understanding the emergence and implications of 'spurious' attractors is important for the scope of attractor neural network models. The literature on neural networks and dynamical systems is ambiguous regarding the definition of a spurious attractor. The term "spurious attractor" lacks a consistent definition and is frequently used in a relative sense depending on the context in which it is discussed. As described by Amit, they can be mistakenly identified as valid due to the proximity of actual stored patterns (Amit, 1989). For others, they are merely nuisances or byproducts of the training process that do not negatively impact the system's performance. In certain contexts, spurious attractors are considered to be states that need to be explicitly removed to ensure the proper functioning of the network (Frolov et. al., 2010; Robbins & McCallum, 2004).

Given this lack of consistency, the following definition can be proposed: When a network is trained to have a specific attractor A, another attractor A′ that appears as part of the training process is considered spurious. In this case, a spurious attractor A′ is defined in relation to an intended attractor A. A spurious attractor is thus a relative concept, dependent on the intended attractor as well as the specific goals of the network training. Its precise meaning and implications vary across different studies and applications and are worth interpreting in further study.

Simulated annealing is a probabilistic method used to find the best solution to a problem by mirroring the process of heating and cooling metal. In metalwork, high temperatures are applied. The high-energy state allows atoms to move more freely, and then it is slowly cooled to make the metal stronger and remove defects. In simulated annealing, the 'temperature' parameter starts high, allowing the system to escape local minima or spurious attractors. This process allows the system to explore a wide range of possible states with high randomness. Just like higher temperature in physical systems makes atoms move more freely, higher noise in neural

networks allows the system to explore a wider range of states. As the randomness, noise, or 'temperature' parameter in the system is slowly reduced, the exploration is slowly limited and allows for the settling into an approximated global minimum energy configuration, this is close to the best possible solution. This is similar to the metal achieving a low-energy, defect-free configuration, actual annealing through physical heating and cooling, and simulated annealing through the controlled introduction and reduction of noise (Branke et al., 2008; Du et al., 2019).

Noise plays a role in the dynamics of these unintended spurious attractors. As noted by Amit the presence of noise can metaphorically "flatten" the attractor landscape and this makes transitions between states easier. This leads to a higher probability of the system settling into a spurious attractor. These unintended local minima can be dealt with in this context through simulated annealing which allows the system to explore a wider range of states through the introduction of noise. Simulated annealing helps understand how noise can control both the stability and change of these attractors. The probability of transitioning from one state to another is given by the equation:

$$\Pr(\Delta x)=\frac{exp[-\beta\Delta E(x)]}{exp[-\beta\Delta E(x)+exp[\beta\Delta E(x)]]}$$

Where: Δx represents a proposed step in the space of *x*, ΔE(x) is the change in energy associated with the step. β is a parameter that represents an inverse temperature, controlling the level of noise introduced (Amit, 1989).

Another complementary approach to understanding attractors and dynamics in RNNs (and their relevance to psychology) is in terms of what Smolensky (1986) called harmony landscapes. Roughly speaking, a harmony landscape is an inverted Energy landscape, where the dynamics lead towards peaks of maximum harmony rather than towards troughs of minimum energy. In his famous room schema example (discussed in more detail shortly), a given partial description of a room will "fill in" missing information by evolving towards the peak of such a landscape. In neural networks, this harmony is a measure of the network's internal consistency or stability. The network's states evolve to maximize the harmony, in the same way that energy-minimizing works in the Hopfield network (Smolensky, 1986).

In harmony landscapes, each potential state of a network is represented by a point within a multidimensional space. Harmony then of a given state is the height at that point, the higher harmony states correspond to the opposite in an energy landscape. By adding a negative sign to the energy function you can transform the landscape from one seeking to minimize energy to one where we aim to maximize harmony. The network's dynamics can be visualized as a movement through the landscape towards points of maximum harmony, which correspond to stable attractors. Learning, as described by Smolensky in "The Proper Treatment of Connectionism" (1993), involves the gradual adjustment of connection strengths or parameters, which results in the adaptation of old and creation of new concepts, categories, and schemata through shifting these harmony landscapes.

Smolensky gives the example of a room schema attractor model to illustrate how ANN models can be applied to cognitive tasks such as associative memory and spatial memory. The model was designed to encode and recall layouts of specific rooms, by stabilizing the memory representation in the form of attractors. This allows the model to fill in missing or degraded information and recover entire representations from partial cues which mimics the way humans can visualize entire rooms from just a few remembered details or a dimly lit room (Rumelhart, Smolensky, McClelland, & Hinton, 1986).

The room schema model works through a network of interconnected units that mirror the associative aspects of different room layouts. Features like a bed or room size of specific rooms are encoded in these connections, influencing how room information is processed and recalled. Learning in this model follows Hebbian-like rules, strengthening connections between frequently co-activated units, thus reinforcing common room patterns or schemas. The integration of Smolensky's harmony landscapes and the room schema model into the broader framework of attractor neural networks enhances the understanding of memory (Smolensky, 1986).

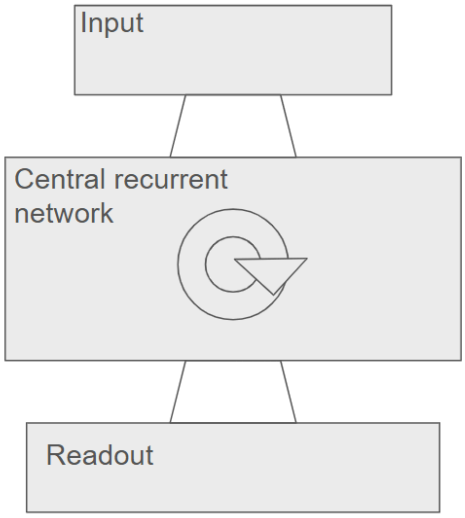


Figure 3. This diagram shows a recurrent attractor network integrated into a larger system. The system consists of an input layer that feeds into a central recurrent network (RNN) and a readout layer for processing outputs.

Recurrent attractor networks are frequently used as part of larger systems that contain other neural networks feeding into the RNN and reading out from them. It is helpful to differentiate these components. As shown in Figure 3: one simple layer of nodes serves as input to the RNN, allowing us to study what kind of inputs the system can deal with. In the room schema case, different room patterns are sent in. In an image classification network, raw pixel patterns are inputs. This setup is sufficient when the interest is in how the RNN classifies inputs. However, there are cases where the attractor states of the RNN are “read out” and used in other networks, for example, to control a process or to be further classified as in a multi-layer Restricted Boltzmann Machine or deep belief net.

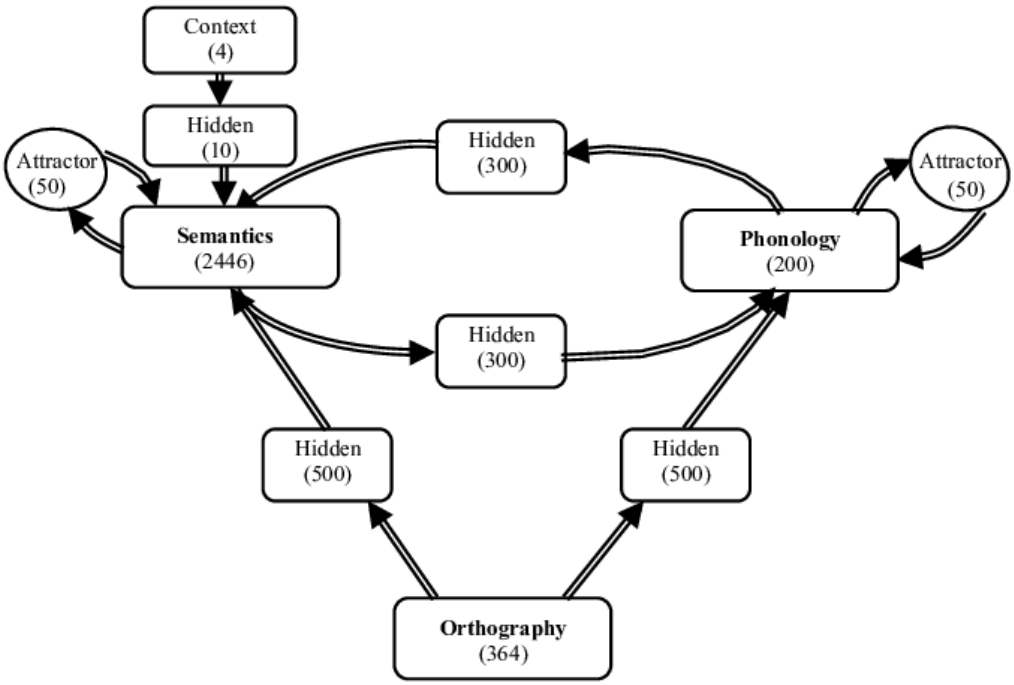


Figure 4: The triangle model for reading (Seidenberg & McClelland, 1989).

This modular approach aligns well with the components of the triangle model originally proposed by Seidenberg and McClelland in 1989. In the triangle model, the orthographic, phonological, and semantic representations can be seen as different types of inputs to the recurrent attractor networks. For example, a simple layer of nodes representing orthographic inputs (the written words) are inputs. The system then processes and classifies various word forms. Similarly, phonological inputs (sound) and semantic inputs (meaning) are fed into their respective attractor networks. This setup is analogous to how room patterns or raw pixel patterns are fed into an RNN in other contexts, such as image classification networks. In Figure 4 the attractor networks aid in stabilizing these representations as they are fed back into the semantics or phonology modules. These attractor states within these networks can be read out and utilized by other neural networks for further processing.

3. Attractor Network Models of Memory

Memory retrieval is a crucial cognitive function:the recall of stored information within neural architectures. In psychology, memory encompasses the processes of encoding, storage, and retrieval of information. Attractor neural networks, those with recurrent connections, are essential for modeling these retrieval processes. In these networks, memory is modeled as stored patterns of neural activation, encoded in the weights and retrieved through network dynamics. The following sections explore specific applications of attractor neural networks in memory retrieval and the mechanisms underlying these processes.

Attractor neural networks can recall entire memories from incomplete or noisy inputs. When a fragment of a stored pattern is introduced to the network, Hebbian learning strengthens

the connections necessary to activate the full attractor. This capability for pattern completion stems directly from the synaptic reinforcement dictated by Hebbian learning.

Retrieval, in psychological terms, is the process of accessing and bringing information to conscious awareness. From a modeling standpoint, this process is the reconstruction of a stored pattern from partial inputs to an original memory pattern. It is not essential for retrieval to be conscious. For example, implicit memory is an unconscious process, where past experiences influence thoughts and perceptions without awareness. Retrieval manifests itself through better performance on tasks or changes due to prior exposure, such as effortlessly retrieving your favorite song lyrics from years ago. Retrieval should be understood as a broad term that subsumes both recognition and recall, which will bebe differentiated below.

A cue, psychologically, is a piece of information that triggers the retrieval of a certain memory. In neural networks, this is implemented as a fragment of a stored pattern which initiates the retrieval process. During the retrieval phase, learning is typically turned off, and the network relies on fixed weights to reconstruct the full memory. Once a cue is presented, activation spreads through these connections, enabling the network to retrieve the complete stored pattern without further modifications to the weights. This can be thought of as giving a hint or providing parts of a memory to be retrieved such as the first couple words of a song to be remembered, leading to retrieving whole lines of the song. Cues in Figure 5 are fragments of images, where the full memory is the full image, that is, the entire memory to be retrieved, For example, a cue here might be a dog ear and in that case the final attracting point to which the trajectory leads is the full retrieval of the dog image.

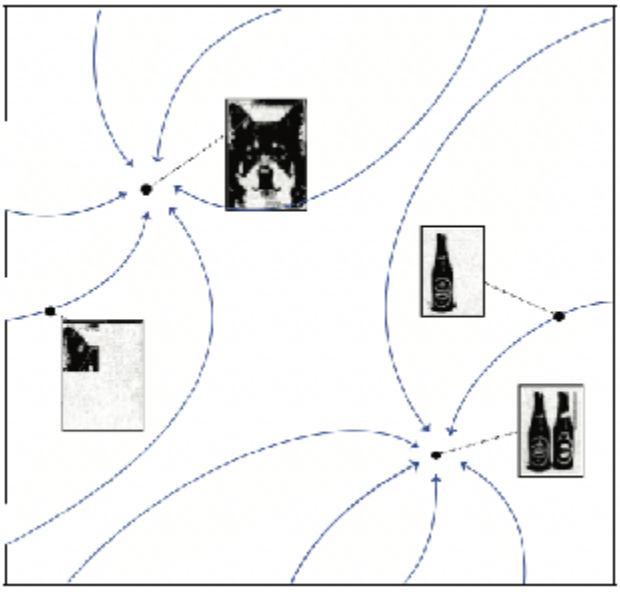


Figure 5. Diagram of cues and the retrieval process in a neural network. Fragmented images represent initial cues that initiate the retrieval process. As activation spreads the cues evolve to complete images that correspond to the attracting fixed points of the network (Yoshimi et al., 2023).

Priming, in psychology, is the phenomenon where exposure to one stimulus influences the future response to subsequent stimuli. This describes the effect of previous experience on

memory and perception. In neural networks, priming occurs when prior activation of nodes (or partial activation of a trained pattern)leads to faster and more accurate pattern completion, moving the initial condition closer to the basin of attraction, hence the network is “primed” to retrieve a pattern.

Recognition is the process of identifying whether information has been encountered before or determining if the current piece of information matches something stored in memory when presented with it. Amit describes recognition as a process of identifying whether the network's current state is within the basin of attraction of a known pattern. This involves less dynamic transition compared to recall and is more resilient to noise (Amit, 1989). Recognition operates when a stimulus directly places the network state near or in the basin of the attractor allowing for immediate identification without an actual total network evolution. An example of this could be asking someone “Is Jason at the event?”. This requires recognizing whether the specific person is present or not, a single dimension question because it specifically asks about Jason.

Interactive Activation and Competition (IAC) networks do well in recognition tasks due to their ability to handle multiple dimensions effectively. In these networks, each dimension, such as the features associated with "Jason" (appearance, actions, etc.), competes within the network. The competition lets the network quickly converge to the suitable representation; these dimensions in IAC networks have explicit representations, which leads to a straightforward recognition process.

Free recall involves retrieving a memory without direct or specific cues. An example would be asking someone “Who was at the event?” requiring a free recall, the question poses initial conditions that are not strongly biased towards any particular basin of attraction, so the network navigates multiple attractors associated with persons who attended the event in this instance. Searching through your memory to bring back a piece of information purely from your internal retrieval cues. The exact nature of this difference, especially in the context of RNNs, is not entirely clear, differentiating recognition from recall involves an additional process of being given abstract information and selecting the right attractor.

In the context of an attractor network, the network must navigate its energy landscape to find the attractor that represents the desired memory. Free recall involves a more complex and dynamic process of locating the correct attractor from a potentially vast and undifferentiated state space. For example, if the network has learned several memory patterns, a free recall task would require the network to converge on one of these patterns without additional input. In a Hopfield network, free recall is modeled by starting the network in a random state and letting it evolve towards an attractor. The depth and width of the basin of attraction determine the ease of retrieving the memory state.

Cued recall is the retrieval of information from memory with a specific cue. The cue could be stimuli such as words or phrases. This cue helps access the stored information faster. In ANN terms cued recall can be thought of as placing the network weakly into the basin of an attractor. As an example asking “Who was at the event that presented a talk and was wearing blue?” In this way multiple dimensions of information each *narrow the search space* to recall from memory. The network then uses this partial memory to navigate towards the associated attractor. Amit touches on this idea in his work, but the definitions are unclear. By clearly carving this boundary we can understand how these networks can be used to better model retrieval mechanisms and how specific dynamics such as cues and priming affect these processes.

Behavioral experiments have shown that memories formed through repeated exposure are strengthened and easily retrieved. For example, studies on the list length effect demonstrate that longer or repeated presentation of items leads to deeper attractors, making those items easier to recall. Ruppin and Yeshurun (1991) show that successful recall and recognition of an item decreases as the length of a list of learned items increases. As the memory load increases, the width of the specific memory's basin of attraction is reduced. The simulations showed that as the load for memory increased, a longer amount of iterations were needed. Additionally, maintenance rehearsal is the process of repeatedly practicing items. This rehearsal improves recall and recognition by deepening the basins of attraction. In the same paper, items presented for longer periods or rehearsed at more intervals reduce the effects of memory load by effectively lowering the network's 'temperature', which increases memory capacity and reduces recognition failures (Ruppin & Yeshurun, 1991).

Empirical validation of attractor neural networks with human data is described in a study conducted by McRae, de Sa, and Seidenberg (1997). They developed a Hopfield model to explore word meaning and validated its predictions with human data. The model itself is made up of input units for word forms and output units for binary features, trained with the Hebbian learning rule for storing feature covariations. The binary features are characteristics of words that can have one of two possible values, like "yes" or "no," "true" or "false," or "0" or "1." In the context of the Hopfield model used in the study, these represented specific properties of words, such as "Is it a living thing?" In simulations of one of their experiments, which involved priming, the convergence time was dependent on how *similar* the initial and final states were. This experiment highlighted the importance of individual features, basically showing how important each individual feature of the words was for convergence. By showing that convergence time is influenced by individual features, the study shows how each feature of a word contributes to the overall process of word recognition and meaning (McRae et. al, 1997).

In another experiment they focused on feature verification, that is, convergence time relative to the strength of correlated features. Both of these simulations aligned with the human trials. The outcomes showed that though there are limitations in modeling tasks where extensive reasoning and integration is needed, these attractor networks could account for patterns of performance in speed-related memory phenomena such as priming (McRae et. al, 1997).

McRae's model can be linked to the structure in Figure 3. The input units correspond to the features, the main network consists of interconnected units where each unit represents a semantic feature, and the output units represent the final activation pattern after the network settles. This setup mirrors the layered architecture seen in Seidenberg's model, where different types of representations (orthographic, phonological, semantic) interact through learned connections.

In a recent study, Pereira-Obilinovic, Aljadeff, and Brunel (2023) propose a new model that overcomes key limitations of traditional attractor networks by incorporating forgetting and by allowing weights to learn during retrieval (most traditional models clamp the weights during retrieval) . Continuous learning allows the network to incorporate new information over time, while continuous forgetting helps to remove less relevant information. This model shows the balance needed for memory retention and retrieval in neural networks. The study reveals that memory retrieval dynamics in neural networks are influenced by the age of the memory. Recent memories are retrieved as fixed-point attractors, with stable neural activity. Older memories become chaotic attractors, with heterogeneity and fluctuations. This allows the network to balance the storage and retrieval of various memories and adapts to their age and relevance.

Fixed-point attractors quickly converge to a consistent neural activity pattern when a related cue is presented, representing recent memories. On the other hand, chaotic attractors exhibit high variability and chaotic dynamics, which reflects the natural decay and increased interference over time for older memories (Pereira-Obilinovic, Aljadeff, & Brunel, 2023).

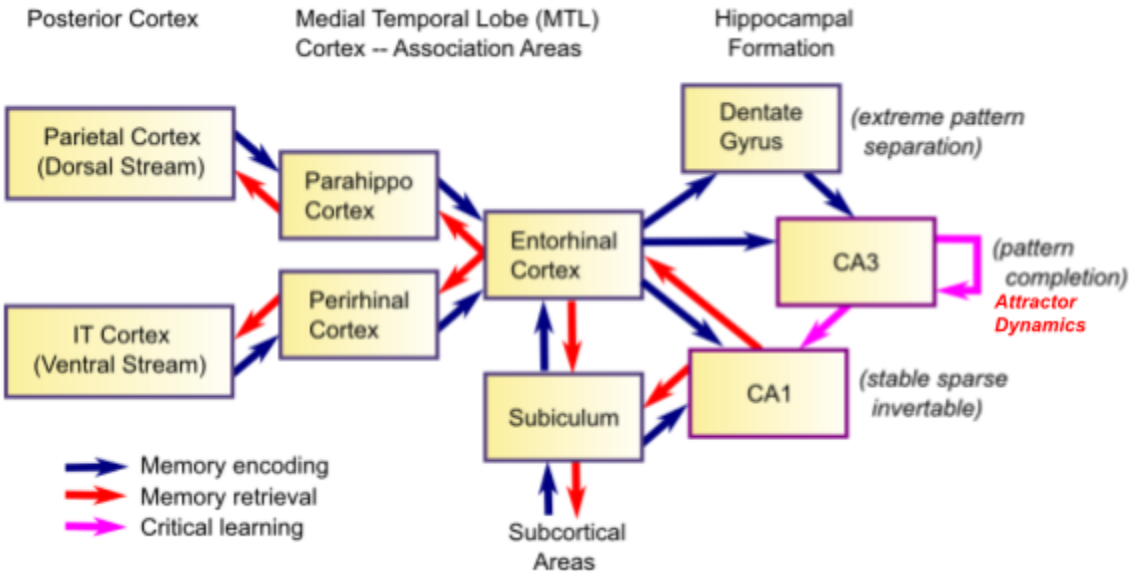
A challenge in neural network models is catastrophic forgetting, where new memories overwrite older ones. The study addresses this with continuous learning and forgetting. Online Hebbian learning updates synaptic connections based on current neural activity, allowing adaptation to new information. A forgetting mechanism, implemented through a decay term in the synaptic weights, ensures older memories fade, creating space for new ones. Using dynamic mean field theory (DMFT), the study shows that the optimal forgetting timescale maximizes the number of retrievable memories, allowing efficient retention and recall over time. Pereira-Obilinovic's work gives insights into memory retrieval in attractor neural networks. Their model addresses age-dependent dynamics and balances continuous learning with forgetting. By integrating fixed-point and chaotic attractors, they provide a deeper understanding of memory storage and retrieval, their model overcomes significant limitations of classical approaches (Pereira-Obilinovic, Aljadeff, & Brunel, 2023).

4. Attractor Models in Computational Cognitive Neuroscience

Building on the memory applications discussed in the last section, I will discuss attractor neural networks in the computational cognitive neuroscience (CCN) literature. The views of connectionist models and CCN models differ slightly. Connectionist modeling usually refers to how information is processed and distributed across networks of neurons through weighted connections between simplified "nodes", which are highly approximated model neurons. Computational cognitive neuroscience takes a more biologically plausible positioning. Whereas the principles of connectionism are integrated, so are the specific brain regions, their functional roles, and their physiological interactions.

The hippocampus is the brain region involved in episodic memory. The process of memory encoding begins with high-level sensory input being activated in the entorhinal cortex (EC) region of the hippocampus. This propagates to the dentate gyrus (DG) and to the CA3 region where dense recurrent connections are considered the main 'memory trace'. This is where the most prominent attractor dynamics arise. See Figure 6. In parallel, the EC sends activation to CA1 which can replay the same pattern back to the EC as an autoencoder would, reversing and recreating the original pattern (O'Reilly et. al, 2012).

These activity patterns trigger learning in CA3 and between CA3 and CA1, strengthening the memory and when recall is initiated CA3 reactivates CA1 and EC, then the cortex and effectively returns the original neural activity pattern and the memory.



[1]

Figure 6. This diagram maps the flow of memory encoding and retrieval in the brain. The hippocampus integrates information from the dorsal and ventral pathways through the entorhinal cortex. The hippocampal formation, including the dentate gyrus (DG), CA3, and CA1 regions, processes and stores episodic memories, which can later be recalled and reactivated through these pathways. This is the essential feedback loop for stabilizing and reactivating memory patterns in the brain (O'Reilly et. al, 2012, Figure 8.2).

Within the CCN view, the "Complementary Learning Systems" (CLS) framework was originally proposed by McClelland and O'Reilly (O'Reilly et al., 2014), describing how the hippocampus and the neocortex are equally crucial for memory formation and retrieval. The hippocampus is specifically involved in the formation and recall of episodic memories through pattern separation. The neocortex is the largest portion of the cerebral cortex, involved in high-level functions such as sensory perception. It plays a large role in processing sensory information and higher cognitive functions. A visual of how these regions are integrated is shown in Figure 7. The neocortex generalizes information slowly and forms semantic memory through pattern completion in this CLS framework (O'Reilly et al., 2014). Attractor networks are important for stabilizing memory states within these regions. In the prefrontal cortex, attractor networks are involved not only in semantic memory but also in working memory and task management. Within the CA3 region in the hippocampus; this area specifically aids in memory retrieval from partial cues and is crucial for attractor dynamics due to having dense recurrent connections in this region (Rennó-Costa, Lisman, & Verschure, 2014).

[1] In the brain when there is recurrent processing it is often thought of as supporting attractor dynamics.

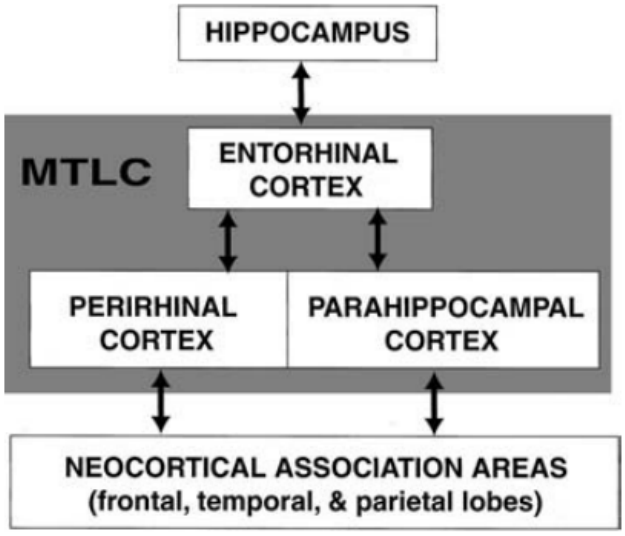


Figure 7. A representation of the neocortex, medial temporal lobe cortex (MTLC), and hippocampus. The MTLC acts as a bridge between the two important areas of the complementary learning systems approach. This area is at the top of the cortical processing hierarchy and is responsible for integrating refined outputs from the cortical modules and relaying them to the hippocampus, as well as receiving feedback from the hippocampus, and sending it back to the specialized modules (Norman & O'Reilly, 2003, Figure 1).

O'Reilly's work also considers the issue of catastrophic forgetting. This phenomenon occurs when new learning disrupts former knowledge. The CLS combats this by using each of the primary regions' functional roles, the hippocampus encoding new information without any interference with prior memory, and the neocortex slowly integrates the new information learned, creating stable representations in long-term memory (O'Reilly et al., 2014).

5. Other Uses of Attractor Models in The Cognitive Sciences

In applied contexts, attractor neural networks can inform research into memory disorders as well as mental health applications. The stability of attractor states influences how memories are retrieved, stored, and processed. Cognitive maps are mental representations for organizing and categorizing experiences. For example, cognitive maps are thought to aid humans and animals in navigating through environments (Peer et al., 2021). In schizophrenia research, hypotheses have been proposed linking the condition to changes in attractor landscapes and cognitive maps. The hypothesis is that shallow attractors contribute to thought disorders, specifically dysfunctional thoughts or speech. The shallow basin of attraction acts to decrease the threshold for state switching, which leads to inappropriate associations between memories that have no relation and disorganized thought patterns (Rolls, 2010; Musa et al., 2022).

As previously mentioned, the CA3 region of the hippocampus is important for the formation of stable cognitive maps. In schizophrenic patients, disordered synaptic plasticity and decreased inhibition may result in an unstable attractor landscape, therefore affecting pattern separation and excess pattern completion (Musa et al., 2022). This disparity leads to abnormal thoughts and behaviors in patients. Understanding thought disorders through the mechanics of

shallow attractors gives insights into therapeutic applications such as target behavioral intervention or pharmacological intervention aimed towards restoring stability for patients with schizophrenia.Another proposal is that excessive noise flattens the attractor landscape, making state switches easy and leading to the disorganized thought patterns characteristic of the disorder.

One particularly interesting thread that remains unresolved is the difference between positive and negative disentrenchment. This refers to a landscape characterized by shallow attractors, which allows for state transitions to occur frequently and with ease. In schizophrenia studies, disentrenchment is considered negative as it leads to the disordered thoughts of schizophrenics. Conversely, in the use of psychedelics, disentrenchment is positive because it helps patients overcome entrenched thought patterns that mark rumination or cyclical behaviors. Exploring this as well as implications across various domains of cognition is something I intend to pursue in my future dissertation research.

Some have taken this to be a hint that could be used in a positive way. The idea is that in the absence of this noise, entrenched attractors result in rumination and stuck states which are alleviated by the destabilizing properties of psychedelics. These substances increase entropy and promote better flexibility in thought patterns. This flexibility and increase in entropy allow the system to explore more states in a state space, which can effectively push a system away from rumination or maladaptive patterns of thinking. In this way, the brain can access a variety of attractors or stable states, promoting resilient behaviors (Hipólito et al., 2023).This “entropic brain hypothesis” proposes that mental health is marked by the brain's capacity to navigate multiple stable states in a state space, while illness then is compared to a system's rigidity in certain states (Hipólito et al., 2023). Psychedelics help to increase the formation of new attractor states, promoting mental flexibility.

This contrasts with the shallow cognitive map theory, which suggests that shallow attractors contribute to the disorganized thoughts of schizophrenics. In essence, while the entropic brain hypothesis focuses on increasing the number of stable states to improve mental health, the shallow cognitive map theory highlights the negative impact of having too few deep attractors. This discrepancy shows room for further exploration of the balance between a deepened attractor landscape and shallow ones and what implications this has for memory formation and retrieval.

6. Conclusion

Deepening the understanding of attractor neural networks’ role in processes like memory retrieval, pattern recognition, and learning will help connect theoretical models with real-world findings. An interdisciplinary approach including dynamical systems theory, computational cognitive neuroscience, and psychology can create more biologically realistic models of neural functioning. Further evidence and interpretation are needed to showcase these models' utility as tools for modeling aspects of memory.

Despite the extensive research on attractor networks since the 1980s, gaps remain. Despite the many existing studies, there is still more to be done studying the psychological relevance of attractor networks. In addition to further empirical work, some basic foundational work is also needed. The concept of attractor strength, while intuitive, has not been precisely formalized. “Strong” and “deep” attractors, or “large” basins make intuitive sense, but they have not been precisely formulated. Similarly for how learning relates to “deepening” of attractors and how this facilitates better retrieval. Though there has been some work (Heino et al., 2023), more research is needed, and this is part of what I intend to pursue in my future work.

Spurious attractors present an issue in ANN models. These often unintended states can interfere with memory. However there is no consensus on the definition. I proposed one definition above, but more work clarifying their definition, formation, and overall dynamics is important for future study.

The distinction between recall and recognition processes in ANNs is inadequately defined. Recall involves reconstructing a memory from partial cues, requiring complex navigation of the attractor landscape. Recognition, on the other hand, is a simpler process, identifying whether a presented stimulus matches a stored pattern. Future work should focus on clarifying these mechanisms in the ANN context.

One unresolved area in applied contexts is the distinction between positive and negative disentrenchment in the context of shallow map theory versus deep attractors. Shallow attractors contribute to disordered thoughts seen in schizophrenia, while deep attractors can lead to rumination and entrenched mental states. I plan to explore this distinction and its implications across various cognitive domains in my future dissertation research.

To conclude, while attractor neural networks are an important tool for advancing understanding of cognitive processes such as memory, further interdisciplinary research and interpretation is needed to address the existing gaps and refine the terminology across disciplines.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*(1), 147–169. https://doi.org/10.1207/s15516709cog0901_7

Amit, D. J., & Amit, D. J. (1989). Modeling brain function: The world of attractor neural networks. Cambridge university press.

Branke, J., Meisel, S., & Schmidt, C. (2008). Simulated annealing in the presence of noise. *Journal of Heuristics, 14*, 627-654.

Bruck, J. (1990). On the convergence properties of the Hopfield model. *Proceedings of the IEEE, 78*(10), 1579–1585. https://doi.org/10.1109/5.58341

Deng, H., Hua, Y., Song, T., Xue, Z., Ma, R., Robertson, N., & Guan, H. (2020, April). Reinforcing Neural Network Stability with Attractor Dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 3765-3772).

Du, K. L., Swamy, M. N. S., Du, K. L., & Swamy, M. N. S. (2019). Hopfield networks, simulated annealing, and chaotic neural networks. In *Neural Networks and Statistical Learning* (pp. 173-200).

Fehérvári, J. G., Balogh, Z., Török, T. N., & Halbritter, A. (2024). Noise tailoring, noise annealing, and external perturbation injection strategies in memristive Hopfield neural networks. *APL Machine Learning, 2*(1), 016107.

Frolov, A. A., Husek, D., Muraviev, I. P., & Polyakov, P. Y. (2010). Origin and elimination of two global spurious attractors in Hopfield-like neural network performing Boolean factor analysis. *Neurocomputing, 73*(7-9), 1394-1404.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Heino, M. T., Proverbio, D., Marchand, G., Resnicow, K., & Hankonen, N. (2023). Attractor landscapes: A unifying conceptual model for understanding behaviour change across scales of observation. *Health Psychology Review*, *17*(4), 655-672.

Hipólito, I., Mago, J., Rosas, F. E., & Carhart-Harris, R. (2023). Pattern breaking: A complex systems approach to psychedelic medicine. *Neuroscience of Consciousness, 2023*(1), niad017. https://doi.org/10.1093/nc/niad017

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507. https://doi.org/10.1126/science.1127647

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.

Kaneko, K. (1998). On the strength of attractors in a high-dimensional system: Milnor attractor network, robust global attraction, and noise-induced selection. *Physica D: Nonlinear Phenomena*, *124*(4), 322-344.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99.

Monaghan, P., Chang, Y. N., & Welbourne, S. R. (2017). Different processes for reading words learned before and after onset of literacy. In *CogSci*.

Musa, A., Khan, S., Mujahid, M., et al. (2022). The shallow cognitive map hypothesis: A hippocampal framework for thought disorder in schizophrenia. *Schizophrenia, 8*, 34. https://doi.org/10.1038/s41537-022-00247-7

Norman, K. A., & O'Reilly, R. C. (n.d.). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *University of Colorado at Boulder*.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2012). *Computational cognitive neuroscience*. Wiki Book, 4th Edition (2020). https://CompCogNeuro.org

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive Science, 38*(6), 1229-1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2021). Structuring knowledge with cognitive maps and cognitive graphs. Trends in cognitive sciences, 25(1), 37-54.

Pereira-Obilinovic, U., Aljadeff, J., & Brunel, N. (2023). *Phys. Rev. X, 13*, 011009. Published 27 January 2023.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. In *Connectionist Psychology* (pp. 367-454). Psychology Press.

Rennó-Costa, C., Lisman, J. E., & Verschure, P. F. (2014). A signature of attractor dynamics in the CA3 region of the hippocampus. PLoS computational biology, 10(5), e1003641.

Robins, A. V., & McCallum, S. J. (2004). A robust method for distinguishing between learned and spurious attractors. *Neural Networks, 17*(3), 313-326.

Rolls, E. T. (2010). Attractor networks. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 119-134.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models* (Vol. 2, pp. 7-57). MIT Press.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations. The MIT press.

Ruppin, E., & Yeshurun, Y. (1991). Recall and recognition in an attractor neural network model of memory retrieval. Connection Science, 3(4), 381-400.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological review, 96(4), 523.

Smolensky, P. (1986). Chapter 6: Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations* (pp. 194–281). MIT Press. ISBN 0-262-68053-X

Smolensky, P. (1993). On the proper treatment of connectionism. In *Readings in philosophy and cognitive science*. https://doi.org/10.7551/mitpress/5782.003.0044

Smolensky, P., & Legendre, G. (2006). The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1. MIT press.

Yoshimi, J., Hotton, S., Tosi, Z., Gordon, C., & Noelle, D. C. (2023). Neural networks in cognitive science.

Zemel, R. S., & Mozer, M. C. (2001). Localist attractor networks. Neural Computation, 13(5), 1045-1064.