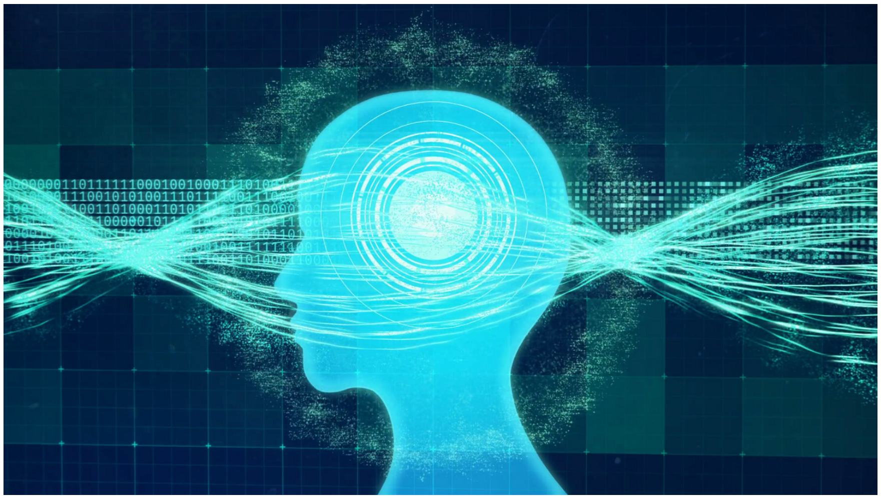
Jesus Rodriguez Follow
Chief Scientist, Managing Partner at Invector Labs, Angel Investor, Columnist at CIO.com, Board Member of Several Software Companies
Jul 25 · 7 min read

Understanding Memory in Deep Learning Systems: The Neuroscience, Psychology and Technology Perspectives



Today I decided to assemble several articles form last year in a longer essay that explains the foundations of memory in deep learning systems.

Memory modeling is an active area of research in the deep learning space. In recent years, techniques such as Neural Turing Machines(NTM) have made significant progress setting up the foundation for building human-like memory structures in deep learning systems. In the past, I've written extensibly about the role of memory in artificial intelligence(AI) so I am not planning to bore you restating the same points. Instead, I would like to approach the subject from a different angle and attempt to answer three fundamental questions that we should have in mind when thinking about memory in deep learning models:

a) What makes memory such a complex subject in deep learning systems?

b) Where can we draw inspiration about memory architectures?

c)What are the main techniques used to represent memories in deep learning models?

In order to effectively answer the first two questions, we should look at both the biological and psychological theories of memory. That should take us to the two schools of thought that have influenced our knowledge about memory the most: neuroscience and cognitive psychology. Following that same trend of thought we are going to structure this essay in three main parts. The first part will explain the neuroscience theory of memory. The second part will approach memory from the perspective of cognitive psychology while the final segment will focus on how deep learning is drawing inspiration from those disciplines to incorporate memory into neural networks. So let's start in the place where memories are created: the human brain.

The Neuroscience Theory of Memory

Understanding how memories are created and, sometimes, destroyed as well as the differences between long and short terms memory have been an important area of neuroscience research in the last decade. One of the iconic subjects that inspired that level of research was been known as patient HM.

Henry Gustav Molaison(HM) suffered an accident at the age of nine that caused him to experience convulsions regularly for the following years. In 1952, at the age of twenty-five, HM underwent a surgery to relieve his symptoms. The procedure was considered initially successful until the doctors discovered that they have accidentally cut part of HM's hippocampus. as a result, HM was unable to retain new memories.

The idea of living without new memories is the analog of always living in the present. Trust me, I am not talking about the mindfulness way but in the way that you can't relate to a recent event in the past or envision an event in the future. Patient HM went about his day only retaining information for a few minutes, greeting the same people and asking the same questions over and over again. The HM case was pivotal to help neuroscientists understand how memories are created, stored and recalled.

The modern neuroscience theory of memory involves three fundamental areas of the brain: the thalamus, the prefrontal; cortex and the hippocampus. The thalamus can b considered a router that processes sensory information(vision, touch, speech) and relays is to the sensory lobes of the brain for evaluation. The evaluated information eventually reaches the prefrontal cortex where it enters our consciousness forming short term memories. The information is also sent to the hippocampus which distributes different fragments to various cortices forming long term memories. One of the biggest challenges neuroscience today is to understand how those scattered fragments of memories can be reassembled into cohesive memory experiences. This is known in neuroscience as the "binding problem".

The Binding Problem

Considered one of the most puzzling aspects of the neuroscience theory of memory, the binding problem challenges the concept of recreating memories from other sensory information. Take the experience of going to a concert with your loved one. Memories about the event will be broken down and stored across different regions of the brain. However, it will only take one experience such as listening to a melody of the same band or seeing your wife dancing to recall the entire memory of the concept. How is this possible?

One theory that solves the binding problem states that memory fragments are linked by electromagnetic vibrations that are constantly flowing through the brain. There vibrations create a temporal(not spatial) link between memory fragments allowing them to activate together as a cohesive memory.

The neuroscience theory of memory give us the foundation to understand some of the main components of an intelligent memory architecture. However, human memory is not only a by-product of the components of the brain but it is also deeply influenced by contextual circumstances. That will be the subject of the next post.

The Cognitive Psychology Theory of Memory

The "binding problem" of the neuroscience theory of memory explains how scattered memory fragments can be recalled into cohesive memories. It turns out that, in order to explain the binding problem, we need to expand beyond the architecture of our brain and evaluate all sorts of psychological contextual elements that deeply influence how memories are recalled. One of the main theories in cognitive psychology that tries to explain the associative nature of memory is known as the Priming Effect.

Associative Memory and the Priming Effect

Like all good theories in cognitive psychology, let's try to explain the Priming Effect in the context of experiments. Think about the first thing that comes to mind when you hear the word DINNER. Was it wine?(for me it was), dessert?, maybe a Saturday night date? As you can see, something as simple as a word can evoke a mixed set of emotions and even other related words. We are effectively recalling associated memories.

One of the most remarkable results of the prior experiments is to notice how fast you were able to retrieve those related words or memories. That happens because associated memories are part of what Economics Nobel Prize winner Daniel Kanehman calls System 1; they happen quickly and they produce a series of related emotional and physical responses. In psychology, that type of phenomenon is known as Associatively Coherent.

Going back to our word game; the fact that the word DINNER evokes the idea of WINE or DESSERT is known as a priming effect in the sense that *"dinner primes dessert".* Priming has an important role explaining how memory works. The priming effect does not only applies to words but also to emotions, physical reactions, instincts and other cognitive phenomenon's. In the context of memory, the priming effect tells us that memories are not only recalled by associated ideas but by "primed ideas".

The Availability Heuristic

Another important element of the cognitive psychology theory of memory covers how we recall the frequency of events. For instance, if I ask you *"how many concerts have you attended in the last decade? "* you are likely to overestimate the number if the answer feels fluent or you have recently attended a concern. Otherwise, if you don't enjoy your last concert experience, the number might be too low. This cognitive process is known as the Availability Heuristic and explains how our memories are deeply influenced by the rapid availability of an answer.

By now we have an idea of how we can think about memory in the context of the brain(neuroscience) and our social settings(cognitive psychology). How are those theories imitated in deep learning algorithms?

Memory and Deep Learning

From the neuroscience and cognitive psychology theories of memories we know that any artificial memory system should have a specific set of characteristics to resemble human memory.

a)Partition a memory into segments that describe different areas of knowledge

b)Reassemble disparate segments into cohesive information structures

c)Retrieve data based on contextual and not directly related information as well as external data references

No discipline in computer science can benefit more from a human-like memory system than deep learning. Since its early days, there have been efforts in the deep learning space to model systems that simulate some of the key characteristics of human memory.

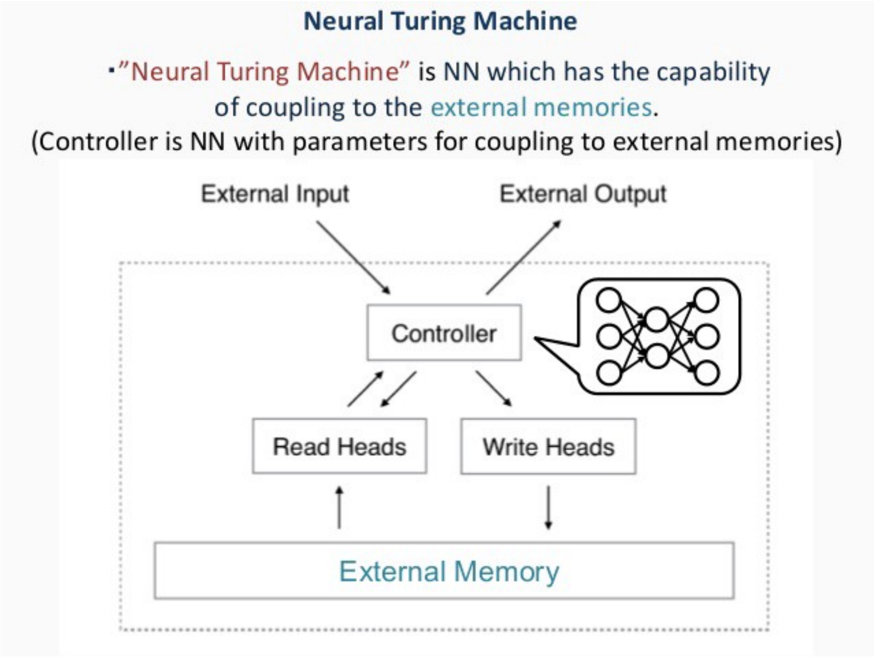
Deep Learning and Explicit Memory

In order to understand the relevance of memory in deep learning models, we should differentiate between the concepts of implicit and explicit knowledge. Implicit knowledge is typically subconscious and, consequently, hard to explain. We can find examples of implicit knowledge in areas such as speech and vision analysis such as recognizing a monkey in a picture or the tone and mood in a spoken sentence. Contrasting with that model, explicit knowledge is easily modeled declaratively. For instance, understanding that a monkey is a kind of animal or that certain adjectives are offensive are classic examples of explicit knowledge. We know that deep learning algorithms have made incredible progress representing implicit knowledge byyt they still struggle modeling and “memorizing” explicit knowledge.

What makes explicit knowledge so difficult in the context of deep learning algorithms? If you think about the traditional architecture of neural networks with millions of interconnected nodes, we will realize that they lack the equivalent of a working memory system that can store fragments of inferred pieces of knowledge and their relationships so that it can be easily acceded from different layers in the network. Recently, new deep learning techniques have been created to address this limitation.

Neural Turing Machines

The rapid evolution of deep learning algorithms has triggered the need for memory systems that can resemble the characteristics of human memory when processing explicit knowledge. One of the most popular techniques in the memory modeling space is known as Neural Turing Machines(NTM) and was introduced by DeepMind in 2014.



Neural Turing Machine

▪”Neural Turing Machine” is NN which has the capability of coupling to the external memories.
(Controller is NN with parameters for coupling to external memories)

NTM works by expanding a deep neural network with memory cells that can store complete vectors. One of the greatest innovations of NTM is that it uses heuristics to read and write information. For instance, NTM implements a mechanism known as content-based addressing that can retrieve vectors based on input patterns. This is similar to the way humans recall memories based on ctextual experiences. Additionally, NTM includes mechanics for increasing the prominence of memory cells based on how often they are recalled.

NTM is not the only techniques that enables memory capabilities in deep learning systems but is certainly one of the most popular. Imitating the biological and psychological functions of human memory is not an easy endeavor and has become one of the most important areas of research in the deep learning space.