

SP2171 DISCOVERING SCIENCE AY 2003/2004

Introduction to Human and AI Learning

by

Abel Yang (U031876N)

Hu Yi (U032048J)

Wee Wei Wen (U032150L)

Mentors:

Kamalesh Basu

Huegesh Marimuthu

March 22, 2004

Contents

1	Overview	1
2	Human Learning	3
2.1	The Concept of Abstraction	3
2.2	Case Study: The learning of abstract concepts in babies	3
2.3	1st and 2nd Order Generalizations	5
2.4	Inherent Knowledge/Abilities	6
2.5	Possible Processing Pathways	7
3	Modelling Human Memory	9
3.1	The Theory of Memory	9
3.2	The Modal Model of Memory	11
3.3	Case Study: Learning with developmental changes in neural networks	16
3.4	Conclusion and Hypothesis	18
4	Introduction to Artificial Intelligence	19
4.1	Classification of AI	19
4.2	Proposed Model	23
5	Conclusion and Future Work	29

Abstract

Artificial intelligence attempts to model aspects of the decision making process in living organisms, and to some extent the human learning and decision making process. In this study, we focus on the human learning process and its means of learning and memory. By comparing the natural model to existing models of artificial intelligence, we show that both generalization and association is necessary in learning systems. We also propose a model for an artificial learning system based on the models we have reviewed.

Chapter 1

Overview

A.I. An abbreviation for Artificial Intelligence, this term has become a very familiar one to many people. To date, many intelligent systems have been developed for a variety of uses, from playing strategic games such as Chess and Warcraft, to assisting with scientific studies, to visual recognition and language processing.

The human brain is the most powerful intelligent system currently known to mankind in terms of learning capability, higher reasoning and in the understanding of abstract concepts. It is a system found in every human being and has been in existence for thousands of years, aiding each new generation of humanity to finally result in the modern civilization it is today.

Given the power of the human brain, what we propose to do is to attempt to create an AI model based on the human one in the quest for a better AI. We intend to focus on various aspects of the human ‘A.I.’ and also examine some of the other A.I. models currently in use today.

In our first chapter, we will examine the human learning process, focusing on two important areas: The understanding of abstract concepts and the abilities which are inherent in humans from birth. We will also touch on the possibility of there being more than one processing pathway in the human brain, each used when needed. The theories proposed here will aid in the construction of the human A.I. model.

In our second chapter, we will touch on the subject of memory. Memory is essential to learning, largely because without it, anything learnt can not be remembered and as such

is not really learnt. First we will cover the case of a man named 'H.M.', whose memory capabilities have been somewhat compromised. Following that, we shall examine some models pertaining to memory.

In the third chapter, we shall examine some of the A.I. models currently in existence, reviewing their processes and the theories behind them. We shall then use this knowledge to propose a human A.I. model.

Chapter 2

Human Learning

2.1 The Concept of Abstraction

An important part of the human learning process is the learning of abstract concepts. Abstraction is the formation of a generalized idea or theory developed from specific concrete examples. In human beings however, abstract concepts can be complex and whose examples can seemingly have nothing in common. For example, a bird, a worm and a monkey are all animals, but are very, very different. Thus, understanding the process by which humans form abstract ideas would be very important to understand the human learning process. To this end, we have chosen to examine the learning of abstract ideas in children (babies, 18 months onwards), because it lays the foundation for subsequent abstract concepts learnt later in life.

2.2 Case Study: The learning of abstract concepts in babies

The experiments we mentioned in the case study are referenced from “*The emergence of abstract ideas: evidence from networks and babies*”. [2]

The first issue pertains to the labelling (the naming) of objects. There is a strong correlation between labelling and more abstract thought, such that children are more likely to form taxonomic rather than thematic groupings and to make inferences based on deeper and functional properties when the objects are named than when they are not. It is interesting to note that the label need not be a word the child knows.

Further experiments have come to support these findings. When children were presented with a named object and asked to find one that had the same name from a different set of objects, they consistently showed different response based on whether the object was made of a solid or non-solid material. It should first be noted that the experiment was quite highly controlled, that the only two possible criteria for the grouping of the objects presented was the object's material and the object's shape. All the objects and names involved in the experiment are novel, thus rather than children's knowledge about specific categories, the task was designed to measure the children's generalized expectations about how novel names are attached to categories and about how categories are to be formed. It was found that with solid objects, the children chose to consistently group them by their shape. This is in contrast to the non-solid objects, where the children chose to consistently group them by the material they were made of, as opposed to shape. It should also be noted that this was found in children 18 to 24 months onwards; very young children of generally less than 18 months were not as consistent in their choice and grouping of the different objects.

Similar results were found in the case of word distinction. It was observed that when a sound, like a tone, was used to label a word, children of age 18-24 months did not take the tone to be a label for an object. They did however, accept word labels for the object and had no trouble in selecting the right one. Once again, all objects and word labels used in the experiment were novel. Children under 18 months were also tested and interestingly, they did accept the tone as a label and actually performed better in selecting the correct objects for that section of the test.

In a test on slightly older children of age 30-36 months, the children were exposed to additional sets of objects besides the two sets in the first test, the same shape and same material sets. The additional sets were cross-solidity sets; objects of the same shape but of a different solidity material from the named object. It should be noted that the material was the same, only of a different solidity. In the case of the non-solid cross solidity test, two different sets of objects were used, some in a non-constructed shape typical of non-solid objects and another set in a constructed shape, more typical of solid objects. The results were similar to the first test: The children still selected objects of the same shape for the solid named object and objects of the same material for the non-solid named object. Some children did make selections when it came to the cross-solidity tests, but overall, much

fewer children made selections. Also, more children made selections when the non-solid test object was in a non-constructed shape as compared to when the non-solid test object was in a constructed shape.

In tests conducted on the source of the label for the objects, there were a few interesting findings. Children from the original age group of 18-24 months were tested and it was found that children did not take a word as a label when played a recording of the word. They did however, take it to be a label when the source was a spoken word. They produced a similar response when a non-word sound was used. A second test using animal sounds to label animals was also conducted. In this test, sounds, some animal in origin and some artificially synthesized, were used as labels for animals. It was found that the children accepted animal sounds as labels and did not accept the artificial sounds as labels. It was also found that in this case, the source of the sound did not make a difference.

The case study's conclusion was merely that "*abstraction may often be nothing more than the result of specific learning of specific instances - not a separate process, not a separate kind of knowing - by the natural and very ordinary process of generalization by similarity*". However, based on the results of these experiments, a few more conclusions can be drawn.

2.3 1st and 2nd Order Generalizations

These results prove a few things. Firstly, this supports the theory that children learn the distinctions between objects and substances as they learn their first object and substance terms. [2] This theory is that children will first make specific associations between names and specific individual objects. The process of learning is very slow, an interesting parallel to the AI learning system of the same nature. This is shown by the fact that children younger than 18 months were less consistent in their grouping of objects, the younger, the less consistent. Thus, what the experiment shows is that children are able to make a first order generalization based on the properties of specific objects (that balls are ball shaped, for example).

While the learning process for this initial step may be clear, the second order generalizations is not. An example of a second order generalization might be that "since balls are ball

shaped and cups are cup-shaped, ___ are ___ shaped”, as given in the journal article. This is evidenced by the experiments, where children have made the second order generalization that solid objects are labelled by their shapes and non-solids are labelled by their material. Likewise, the experiments on sounds are evidence of similar second order generalizations, like words being labels when spoken with the mouth as a source. Subsequently, third order generalizations and so on can be made based on currently know generalizations and so on.

While this can be accepted as logical, it must be noted that the training process plays a significant part. As noted from the younger children, the training process must be fairly specific and have little room for error. What is meant by this is that the training process must be designed to pertain to the specific variable which is to be trained. In the minds of the younger children, solidity was not yet a distinction criteria. They had not yet made the generalizations that solids were referred to by shape and that non-solids were referred to by material. Thus, they were not as consistent in their groupings of the objects. Likewise for sounds. This indicated that a child might think that a ball is a ball shaped object which rolls, until he is presented with a series of non-ball shaped objects which also roll and is told that they are not balls.

One comment on the experiments is that little detail is given on the groups of children involved. It is unlikely that the same infants were tested at different ages; it is more likely that different groups of children at different ages were tested. Perhaps the experiments might be more accurate if the same groups were tested.

2.4 Inherent Knowledge/Abilities

What is very important but not mentioned at all in the case studies and other journal articles of the same nature is the fact that these experiments demonstrate that children have some inherent abilities and knowledge.

First and foremost, the most obvious inherent ability is that the children are able to make generalizations based on similar criteria. While this might seem simple and obvious, largely because the entire human race is born with this ability, it becomes a very important point in the creation of a model of the human learning process. An AI, for example, created based on this model, must be programmed to be able to observe similarities based on knowledge

and form the initial first order generalizations based on them. Without this ability, the whole model fails.

Also, since the children are too young to understand explanations, the experiments would seem to indicate that children have an inherent rudimentary understanding of the concepts of shape, material, etc. in the sense that the shape of an object is the object's physical shape, while material is what the object is made of. Likewise, they already classify the source of a sound as a possible distinction criteria.

Another inherent ability is their associative ability. Children have to be able to associate label with object in order for any of the experiments to have worked. Once again, this basic ability to associate a thing with another thing and to know that one thing indicated another thing has to be inherent for the labelling process to work in the first place, since, as mentioned above, the children are far too young to understand explanations.

2.5 How many processing pathways? A study of Monkey vs Human response time

A series of visual processing experiments were conducted on a series of human and monkey subjects [3]. These experiments were conducted under extremely tight time constraints (stimuli for 30 ms, response required in less than 1s) and were based on flashing the subjects a picture and requiring a response to a given question for the series of pictures, like "is there a human/monkey in the picture". Yes or no responses were required. The pictures also varied in degrees of sharpness and colour(i.e. colour or black and white). It was noticed that the monkeys have a generally faster reaction time(250-300ms) than the humans(350-450ms), although they had roughly equal success rates(90% and 94% respectively). While this can be explained by the fact that humans have larger brains and that within the cortical areas the conduction velocity of the axons can be relatively slow, it is the other finding which is significant.

While colour was found to have no effect, the more important observation is that the subjects continued to perform significantly above chance levels until the sharpness of the picture was at a contrast of 3.1% of normal levels. What is interesting here is that at contrast levels of twice to three times of that, it is already very difficult to actually comprehend what

is in the picture, even without a time limit. This, coupled with the fact that the human subjects could not give an explanation as to why they chose the pictures they did for the lower contrast levels, nor could they identify what the picture was, provides a very interesting insight to how the human mind might possibly work. The reason given was that humans might share some of the more basic, coarse and faster processes with monkeys, inherited from their common ancestor and when faced with great time constraints, they are forced to rely on them.

What this would mean for a possible AI model is that, to be similar to a human, it should have not one, but two processing pathways: one for higher reasoning at the cost of speed when time is not a constraint, and one for making swift decisions when time is a constraint. This second processing pathway should also be inherent to a certain extent and also preferably have no need for modification.

Another thought concerning this second pathway is that it could possibly be the reason for “gut feelings” and “intuition” in humans. Since the humans involved in the study demonstrated the ability to correctly detect and assess the pictures even though their conscious minds could not, it is very much likely that what is called “intuition” might just be the perception of stimuli on a subconscious level, which triggers biologically appropriate responses, in spite of the conscious mind being unable to locate the reason for those responses. This might possibly be the key to giving an AI the equivalent of “intuition”.

Chapter 3

Modelling Human Memory

3.1 The Theory of Memory, the Hippocampus and H.M.

The hippocampus is a structure located in the medial temporal lobe of the human brain. It is a key factor in the formation of long-term memories and is currently credited with being the area of the brain in which short-term memories are converted into long-term memories. As this formation of long term memories is important to the human learning process, we shall thus examine the hippocampus, the results of removing the hippocampus and their implications on learning and memory storage.

Like most parts of the human brain, the exact mechanism of the processes of the hippocampus is currently unknown. However, it has been possible to determine the function the hippocampus has in the long-term memory formation pathway by examining the effects that the removal of the hippocampus has on the human being. As such experiments are currently unethical and illegal, the existence of a man dubbed H.M. [1] is fortunate. In 1953, H.M.'s hippocampus was removed by doctors as a cure for his epilepsy. It is interesting to note that it did work. However, shortly after the operation, it was noticed that H.M. was having memory troubles and he was brought in for observation. A recent review paper by Suzanne Corkin [1] provides us with the findings of H.M.'s case from that time till now. The following observations are all referenced from this paper, which makes references to the various papers concerning H.M. So far, observations and experiments on H.M. have provided us with the following information regarding the function and process of the hippocampus:

The most significant observation was H.M.'s inability to form new long-term memories. He is still able to remember things for very short periods of time, but unable to recall pieces of information after a longer period. He had a general inability to remember facts, names, images and other pieces of information associated with declarative long-term memories. This proves that the hippocampus to be uninvolved in the formation of short-term memories and that its removal affects the long-term memory formation process.

The next observation is that all of H.M.'s long-term memories of events prior to the operation remained intact. Although he was unable to recall any of the events after his operation, he can still remember events in his life prior to the operation. This indicates that the hippocampus is not the area in which long-term memories are stored. This would indicate that it is indeed the area of the brain responsible for the conversion of short-term memories to long-term memories. It also indicates that the hippocampus is not required for the access of those long-term memories either.

It is observed as well, that H.M.'s other mental functions are all in working condition and not destroyed or severely impaired by the removal of his hippocampus. His motor processes are unaffected, as are his sensory processes and emotional processes. Thus, this indicated that the hippocampus has but the one specific function as mentioned above.

It should be noted that H.M. is capable of remembering motor skills learnt after the operation. This fact indicates that unlike declarative long term memories, the formation of procedural long-term memories do not take place in the hippocampus.

All this implies that the human brain has separate processes for the formation of memories of motor skills and of declarative data. As memory is a significant part of the learning process, this would imply that, biologically at the very least, the process of learning a motor skill and a piece of declarative data are different. Another implication of this is that the human brain functions as several entities much like machines on a factory assembly chain; each area has its own function(s) and interacts or passes on information to other areas to create a whole mental process. At least in the case of the hippocampus, this also shows there to be no back-up system; this area of the brain is specialised and unique to its function.

At this point it should be mentioned that the storage site of long term memories is currently unknown. One possibility is that memories are not stored in any specific area of the brain, but instead broken up as fragments and coded to a particular assembly code.

Studies have shown that if a particular area of the brain involved in sensory processing is damaged, the person is unable to recall that aspect of the memory. This ties in with the theory, that once assembled, the memory is re-processed by all the areas of the brain involved in the creation of that memory. All this gives us input for a possible AI model of the human learning process in terms of information storage, processing and recall.

3.2 The Modal Model of Memory by Atkinson & Shiffrin

In the field of scientific research on human memory, there has yet to emerge a dominant theory which explains all the mechanisms. One of the most influential models was the modal model proposed by Atkinson and Shiffrin in 1968.

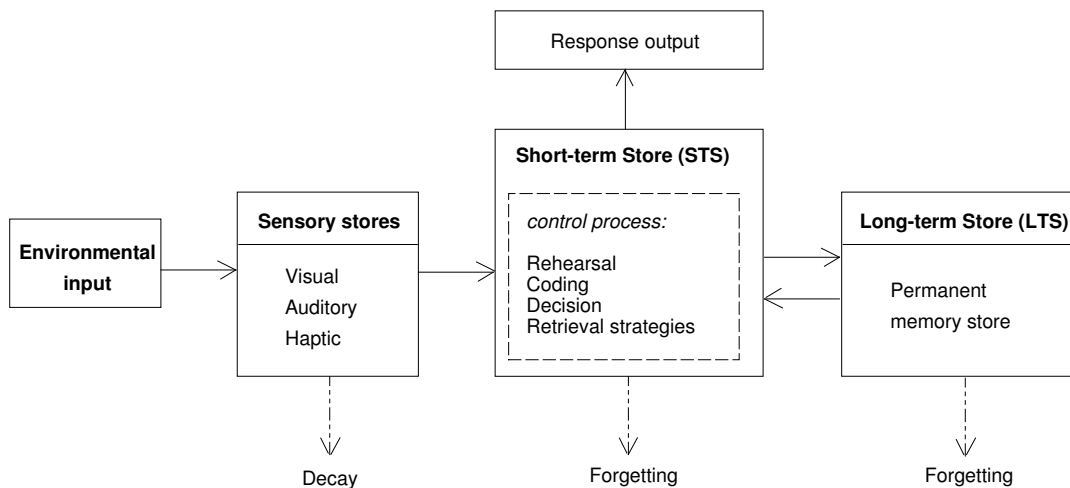


Figure 3.1: The model of human memory proposed by Atkinson & Shiffrin(1968).

In this model, information is assumed to flow from the environment through sensory stores, which are parts of the perceptual system, into a short-term store. Atkinson and Shiffrin also proposed in their publication that the short-term store has limited capacity; and the longer an item resides in this store, the greater the probability of its transfer to long-term memory.

Short-term Memory

Short-term memory is the temporary memory store accessed after recent exposure to a stimulus to be recalled. Short-term store is probably the most commonly accessed portion in this model. It is where the input is analyzed and the response output is generated. It also controls important processes such as rehearsal, coding, decision and retrieval strategies.

In Atkinson and Shiffrin's theory proposed together with their model, the probability of a memory item being transformed into a long-term memory increases as the time it resides in the short-term store increases. This hypothesis was later commonly rejected as evidence suggested that merely holding a memory item in short-term memory store did not guarantee transferring to long-term memory. Instead, the processing which the item underwent plays a more important part [4]. In the case of word learning, observing the characteristics (shape, font, upper or lower case) helps little; acoustical training by reading it aloud would help slightly more; by far the most important processing is the judgment about its meaning and the relation between it and a pre-existing concept or experience. Based on more studies on the role of short-term memory store on learning, Baddeley & Hitch proposed another more complicated model for the short-term store [7], in which they rename the short-term memory as working memory.

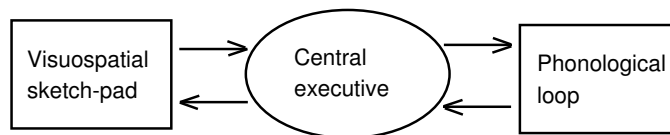


Figure 3.2: The model of working memory proposed by Baddeley & Hitch(1974).

In this new model, working memory is subdivided into three components and different functions of working memory are better explained. The central executive has more functions than mere storage. It allows information input to interact with the long-term memory, and then encode them into a form that can be stored in long-term store. The two subsidiary systems, the visuospatial sketchpad and the phonological loop, hold memory traces and engage in the rehearsal of information received. Repeating information input into these two subsidiary systems enable repeating executions, hence after going through the central

executive for a number of times, the information is better learnt and stored in the long-term store.

Long-term Memory

Long-term memory is defined as the permanent memory store accessed after a considerable period between the presentation of a stimulus and its recall.

Long-term memory is not a single entity but is composed of several separate systems. The two major categories are declarative memory and non-declarative (procedural) memory, which is distinct from each other with respect to the kind of information processing involved. Declarative memory refers to the aspect of memory that stores facts and events. Non-declarative memory is a memory of skills and procedures. The following is the fractionation of memory proposed by Squire in 1991 [8].

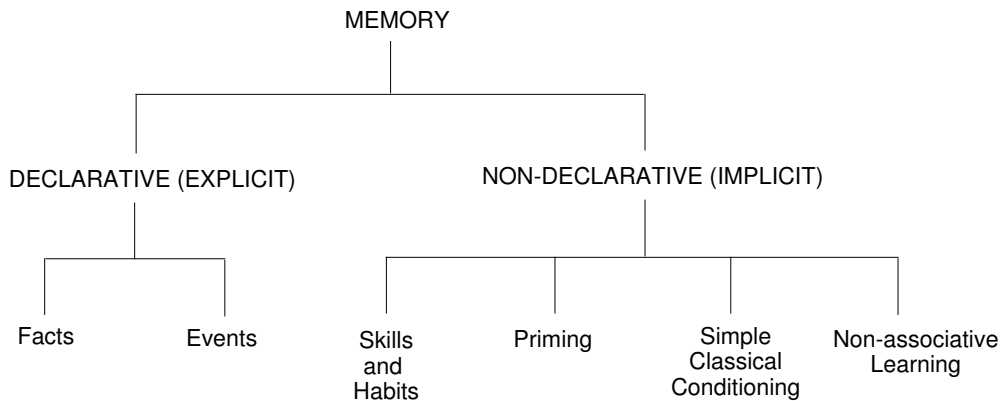


Figure 3.3: A taxonomy of long-term memory proposed by Squire(1991).

From this taxonomy of memory, we can easily identify the distinctions between declarative and procedural memory. Declarative memory can be formed after a single trial of a particular event; whereas procedural memory, in general, is acquired through a series of trials.

Evidence supporting such a two-component system comes from the study of retrograde amnesia. Patients lose memories that were acquired before the onset of amnesia, most commonly due to the damage of hippocampus and temporal lobes. In human amnesia,

such as the patient H.M. mentioned earlier on, it has been commonly observed that patients lose recent memories much easier than remote memories. Thus there seems to be a relationship between the integrity of the hippocampus and the storage of “new” memory. Plausibly declarative memory depends mainly on the hippocampus and related structures, while procedural memory is governed by other brain mechanisms and brain circuits.

As it is difficult to measure the extent of amnesia in humans, experiments were carried out on animals to test this hypothesis. In one study, monkeys learned 100 discrimination problems, 20 each in the 1st, 5th, 9th, 13th and 15th week [6]. In the 17th week, some monkeys’ hippocampus formations were removed bilaterally. In one-trial performance tests conducted after surgery, the monkeys with lesions were impaired on the discriminations that were learned 2 or 4 weeks before surgery. By contrast, they remember the discriminations learned longer before surgery as well as normal monkeys. Moreover, the monkeys with lesions remembered the discriminations learned long before surgery significantly better than the discriminations learned just prior to surgery. These results indicate that hippocampal damage selectively affects a class of memories that are used for only a short time and spares a class of memories that are destined to endure. Indeed, early memories are preserved despite damages in hippocampus because they become independent of the hippocampus as time passes after learning.

Through further studies on declarative memory, Tulving [4] proposed two subcategories of the declarative memory, namely episodic and semantic memory.

Episodic memory refers to autobiographical memory for events that have a particular temporal and spatial context. It is indeed what most people regard as memory. The nervous feelings you had when you delivered your first public speech and details about the happenings during a particular gathering belongs to episodic memory since they are both related to a specific time and place.

Semantic memory refers to factual memory about the world, including factual memory that derives from particular events. A chemistry student does not have to recollect any events at a particular time and place before he can recall the bonding between hydrogen and oxygen in water molecules. Thus the memory about chemical bonding is semantic. An example of a piece of semantic memory derived from particular events would be the best looking girl you saw in your prom night.

Although both episodic memory and semantic memory are impaired in a patient with memory defects, in some studies it has been shown that semantic memory is relatively preserved. Swiss psychiatrist Claparede [4], on a morning in 1911, pricked a Korsakoff's amnesic patient with a pin hidden in his hand. On the next morning, when the doctor asked for handshaking, the patient withdrew her hand without knowing why she did so. Thus semantic memory seems to have a different neural basis than episodic memories. Furthermore, the similarity between the common concepts of memory and episodic memory, and the well-observed link between space/time-related memory and the frontal lobes, suggest that the frontal lobes are more important for episodic memory than for semantic memory.

The following chart, adapted from website of McGraw-Hill Companies, shows the comparison between them.

Characteristic	Episodic Memory	Semantic Memory
Source	Sensation	Comprehension
Units	Events	Facts,Ideas
Organization	Temporal	Conceptual
Reference	Self	Universe
Registration	Experiential	Symbolic
Temporal	Present	Absent
Vulnerability	More chance of disruption	Less chance of disruption
Access	Deliberate	Automatic
Queries	Time? Place?	What?
Reports	Remember	Know
Development	Later in life	Early in life
Amnesia	Affected	Unaffected

Table 3.1: A Comparison between Episodic Memory and Semantic Memory. Adapted from <http://www.dushkin.com/connectext/psy/ch07/table7.mhtml>

Forgetting

Forgetting, a process that occurs in both short-term store and long-term store, marks a crucial difference between the ways in which memory is stored in human brain and computers. Any form of memory stored in any part of the brain is subjected to decay at all times; whereas computer memory can be stored for much longer and can be retrieved at any time in the exactly the same form as it was installed. People never have perfect memory. In contrast, most of the studies about machine learning assume perfect memory.

Through studies on the functions of rapid eye movement (REM) sleep [9] in both humans and other mammals, it has been suggested that REM serves as a memory consolidation mechanism, during which some memories are discarded to prevent overloading of the biological neural networks.

3.3 Case Study: Learning with developmental changes in neural networks

When artificial neural networks are designed to simulate human learning, one important difference between a human and a learning machine is often overlooked. For humans, especially in the case of small children, learning and physical growth occurs at the same time; in contrast, most of the time, the learning machine has been assumed to possess fixed memory capacity. In 1993, Jeffery L. Elman developed his own argument about training neural networks – the importance to start small [10].

In his experiment, a simple recurrent¹ network was designed to learn human language. The training data consisted of 50,000 sentences from a few grammatical categories and with various lengths. The network was trained to predict the following word based on the words given at the beginning of the sentence. The performance of the machine is measured by the degree to which the network's predictions match the conditional probability distributions of the training data.

When the network was first trained without any specific strategy, its learning result was disappointing just as expected. In contrast, the efficiency greatly improved when certain

¹Recurrent means the internal states are fed back at every time step to provide an additional input.

learning strategy was applied.

Incremental Input

The training data was arranged in order of increasing complexity and the network was trained with the simplest input first.

The experiment was divided into five phases. In each phase, the network was trained with 10,000 sentences. The percentage of complex sentences included in each phase was 0, 25, 50, 75 and 100 respectively.

The performance turned out to be very good as the error rate was as low as 0.177.

Such pleasing result demonstrates good analogy with children's language learning. Children cannot learn languages with all the complexity at once. One has to begin with phrases of the simplest structure, and then move on until the complexity of adult language is reached. Besides language learning, other forms of human learning also occurs in an incremental manner. How babies grasp abstract concept, as discussed earlier in our report, is one good example.

However, researchers also noticed a significant dis-analogy. As the learning environment in this experiment was carefully designed to ensure a strictly increasing complexity, it is not a good model for the learning environment to small children. Unlike the network, children are exposed, wittingly or unwittingly, to all aspects of adult language right from the moment they start to learn. Hence, another strategy was invented to provide a better simulation.

Incremental Memory

Since children have a more or less constant environment for language learning, in this experiment, there is no special treatment on the training data before it is passed to the neural network. At the same time, to enable the learning mechanism to encounter only simple data at the beginning, its memory capacity was set low at first.

Due to the limitation in its memory capacity, the neural network could only process short sentences and would discard the longer ones. Throughout the training process, the neural network were tested periodically to decide whether it was qualified to increase its

memory. In this way, the neural network have to learn the simpler sentences before they were able to process the more complicated ones. Though training input was kept constant, the learning data being processed still organized in an incremental manner. The final result turned out to be as good as the experiment in which incremental input was used.

Despite the pleasing outcome, the learning curve of the neural network seemed to be very deep at the beginning. It took much longer for the neural network to master the first level of sentences than any subsequent level. This drawback may be attributed to the small proportion of simplest sentences present in the training data thus a longer time is required for obtaining sufficient data. Another possible explanation will be that there are much more possible rules for generation when the structure of input is simple [10].

3.4 Conclusion and Hypothesis

As we have discussed above, children learned best at the very time when their memory and attention span were increasing; when a network is treated with raw training data, it produces great result when its memory capacity increases as learning progresses. However, knowing about such facts is insufficient to create a good neural network. In our SP2172 project, we have to examine how to create a neural network with adjustable memory capacity. The long period required for the network to start to “learn” could be another problem that we have to face.

Chapter 4

Introduction to Artificial Intelligence

4.1 Classification of AI

Artificial intelligence systems may be classified by the means learning is achieved. To this end we have three different types of artificial intelligence systems that mimic either or both of the cognitive and associative aspects of the learning process.

A cognitive system attempts to generalize its input into an abstract form to simplify processing. Of interest here is the technique used in classifying a set of stimuli into a specific general category for further processing. Such techniques include predicate mapping [11], semantic mapping [11] or frames [12] among others.

In general, a cognitive system receives stimuli as input, performs a transformation on the stimuli to simplify its characteristics and passes the simplified data on to be processed. The transformation on the stimuli performs some abstraction on the stimuli, such that the stimuli can be mapped to a more abstract form or category. In either case, such cognitive systems require some degree of associative processing to match a set of stimuli to a particular generalized object. However, once stimuli are correctly placed in its categories, the system can not only perform categorization but can also make a number of basic inferences from information encoded within the categories.

This brings us to the definition of grounded abstraction defined by Zucker [11] which

is a mapping from a perception P_a to a perception P_g such that P_g is simpler than P_a . Such mapping applies to both predicate and semantic mapping. In terms of processing, the abstract representation of the stimuli should be as simple as possible, such that processing is reduced. For learning, a highly simplified set of stimuli may be insufficient to form a new transformation, but an unprocessed set of stimuli may be too large to be processed as a whole. This requires the system or teacher to decide on a set of details that allows the frame to be stored for further use in both processing stimuli and learning. Without prior knowledge of future stimuli, the system can only guess what is necessary to be stored, effectively requiring a significant amount of teacher interaction or a large number of trials to create a suitable abstraction that is optimal for both learning and processing.

Frames as described by Minsky [12] are a general case of mapping a set of input stimuli to a framework describing the stimuli. Simplification occurs in subsequent learning, as a frame is modified by new stimuli to keep the common terminals and discard the terminals that lie outside the new stimuli. A mature system can therefore match stimuli to a frame with minimal error. This allows the matching process to act as a transformation from the stimuli to a simplified form suitable for processing.

The method of frames and simplification suffer from one drawback. In the case of a large set of data such as speech, image or a large block of text, the system needs a mapping to extract the attributes of the stimuli that can be passed to the transformation to be simplified. When the set of stimuli is small, for example pressure receptors on the fingertips of a robot, the stimuli can easily be characterised and simplified. However when the set is large, for example an image, of 10,000 pixels, the stimuli must go through an additional simplification process that will allow it to recognise the difference between two possibilities, for example the letter A and the letter B. This may necessarily involve some form of associative system that will process the data prior to sending it to the cognitive system, like a neural net or a set of pattern matching rules.

An associative system primarily associates its stimuli with a specific response. Such systems can range from simple rule-based systems to a neural net. Artificial intelligence systems based on an associative system can be very efficient, but suffer from a lack of generalization ability.

Rule based systems process stimuli by mapping a stimulus vector to an output vector.

Such means of processing severely limits the capabilities of the system but given a large enough set of stimuli and rules, i.e. the dimensionality of the stimulus vector is large, the system may be sufficiently complex for some applications. A main drawback of such rule based systems is that the system does not have a capacity for learning and memory and needs to be hard coded with the rules that map stimuli to output.

A neural net system attempts to mimic the function of neurons arranged in a network pattern, akin to the structure of the brain. By considering the neuron as a black box that processes a number of input stimuli to arrive at an output, and linking these artificial neurons in a network that mimics the dendritic connections of the brain, the neural net can be said to be an approximation of a natural brain.

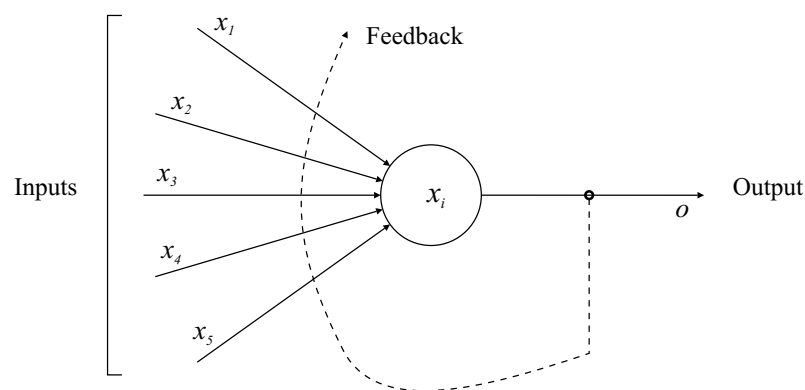


Figure 4.1: An artificial neuron simulates a natural neuron by receiving stimuli from multiple sources and returning a single output based on the summation of the input values. The neuron may have a feedback path depending on the method of learning in the system.

From the model of an artificial neuron in Fig. 4.1, processing takes place by comparing the weights assigned to each of the inputs to the neurons, and returning an appropriate response based on a function of the inputs. While this may seem simple, it is by employing a large number of neurons that the system derives its processing abilities. Of interest will be the connections between the neurons and the means of assigning weights to each neuron in the process of learning.

Forms of learning for neural nets include error correction learning, memory based learn-

ing and Hebbian learning among others [13]. In error correction learning, the neural net obtains feedback through a channel dedicated to feedback. This channel does not connect to any neurons in an ordinary manner; rather it gives each neuron a positive or negative feedback based on the response that the neural net returned, effectively correcting its decision with each feedback signal. A shortcoming of this model is that the feedback needs to be given by a separate channel, therefore requiring additional stimulus receptors to sufficiently derive feedback from the environment.

In memory based learning, the neural net stores a set of data that pertains to the response of the neuron. This method of learning allows the set of responses to be modelled as a vector equation, and is applied in the support vector machine, hence the name. In this model, stimuli are mapped to a vector that is then compared to a set of training vectors. The nearest training vector to the stimulus vector is used to generate output. Subsequently, if feedback is available, the stimulus vector may be added to the set of training vectors with the appropriate responses. A support vector machine employs a similar technique, but instead of maintaining a complete set of training vectors it maintains a set of support vectors that define a decision plane in the vector space of the neurons. A response to a stimulus vector will depend on the location of the stimulus vector in the vector space relative to the decision plane.

The support vector machine is highly suited for pattern classification and can provide good generalization performance on pattern classification problems [13]. As the support vector machine considers the input stimuli as a vector and assigns a response according to the position of the vector, it does not closely model the natural brain, and is better suited for a large number of stimuli with a small number of outputs.

For the Hebbian learning model, we introduce the Hebbian synapse [14], which states that if two neurons on each side of a synapse are activated simultaneously, then the strength of the synapse is increased. This model has been further extended to propose that if two neurons on each side of a synapse are activated asynchronously, the strength of the synapse decreases [15]. Such a model has also been shown to play a role in the functioning of the brain [16] and there is strong physiological evidence for Hebbian learning in the hippocampus [13].

The Hebbian synapse is a model very suited for learning, and allows feedback to be

received from channels other than a specialized feedback channel in the case of error correction learning. A possible network that can take advantage of Hebbian learning is shown in Fig. 4.2a and Fig. 4.2b.

While an associative network is highly efficient, in many cases it lacks the ability to make extensive generalizations, and make generalizations from generalizations. Between humans and an associative system, humans could perform generalizations better than an associative neural net [17]. Over a cognitive system, an associative neural net is easier to implement as the stimuli to the system is broken down in a consistent manner which is then passed to the neurons. In contrast, a cognitive system requires input to be simplified to a sufficient level so as to be easy to process, but not simplified too far that information is lost. To overcome these shortcomings, a hybrid model has been proposed and in some instances deployed. These hybrid models use the output of both associative and cognitive system to perform pattern matching with a higher degree of abstraction.

While an associative system has been shown to be limited in its ability to generalize, we present this hypothesis that the limitations of such a system results from its limited number of neurons rather than an inadequacy in the design. The computational abilities of the natural brain can be expressed in terms of a neural net, and generally consists of nothing more than neurons. While the neuron density and synaptic density may be extremely large with respect to current limits in computing power, it is ultimately a neural net, albeit extremely complex. Therefore it might be possible that apart from the design limitations of the neural net, the capacity of the network limits its ability to form associations.

4.2 Proposed Model

We therefore propose the following model that could mimic the natural learning process. Such a model, as shown in Fig. 4.3, implements a Hebbian learning system on the neuron level and receives its input from a set of receptors. These receptors do not need to connect to all the nodes in the input layer, but a subset of the nodes. Each node also does not need to be connected to all the other nodes in the following layer, and need not be connected to all the nodes in the preceding layer. However, the number of connections to each node should not be too low. Each node may be a simple neuron or may contain a neural net,

depending on the need for complexity.

The nested structure of the system allows for further expansion should additional neurons be needed, while at the same time simplifying the computational requirements of the network. This allows the network to ‘grow’ new neurons should there be a need for additional capacity.

Each neuron will not only take into account the value of the input stimulus but also the time in which the input arrives. By allowing the active state of the neuron to decay over time as opposed to immediately turning off, the network can respond to and learn from stimuli that range over time. Interaction between each neuron and consequently each sub-network is constrained to a discrete time step in order to simplify computation, although such a restriction should ideally not be imposed.

The state of a neuron can therefore be described by the following equation:

$$s_t = d(s_{t-1}) + \sum_{i=1}^k (n_{i_t} \cdot w_{i_t})$$

where s_t is the state of the neuron at time t , $d(x)$ is a decay function, n_i , where $i = 1 \dots k$ are the synapse values for synapses $1 \dots k$ at time t , and w_{i_t} are the synapse weights for synapses $1 \dots k$ at time t . The synapse weights can be described by the following equation:

$$w_{s_{t+1}} = n_{s_t} \cdot a \ln(bs_t)$$

where $w_{s_{t+1}}$ is the synapse weight for synapse s at time $t + 1$, n_{s_t} is the synapse value for synapse s at time t , s_t is the neuron state at time t , and a and b are constants.

Ideally an implementation will maintain a number of spare neurons such that the system may encode new stimuli that it encounters. These spare neurons can be created by replacing an existing neuron with a sub-network, allowing the network to be extensible. As a natural brain does not have new neurons being created, such behaviour does not mimic any known neural models. It is however advantageous to have such behaviour as it reduces the computing power needed to process stimuli.

As the network is extensible, there may be a number of spare neurons that can respond to new stimuli. If these neurons are not reinforced by subsequent stimuli, then the synaptic weights will eventually be reduced from a lack of synchronous activity. If the neurons are

activated from repeated stimuli, the synaptic weights increase and the synapses corresponding to the common set of attributes for a class of objects will be dominant. A neuron that has had all its synaptic weights reduced to zero would be treated as a spare neuron. Given a sufficiently large number of synapses, this model can mimic a form of generalization.

While the number of input and output interfaces is fixed, the system is by no means constrained. As the neurons are sensitive to time-dependent stimuli, it can receive stimuli that are multiplexed and transmitted through a single channel such as speech or text strings. Likewise, it can also provide time dependent output by inserting a delay in the processing of each neuron and allowing the output of a layer to propagate to the same layer.

Qualities of the Proposed Network

This network has the following qualities:

- The network is extensible

By replacing an existing neuron with a neural net, the number of neurons in the network can be easily increased without the need for extensive changes in the structure of the network.

- The network can be compartmentalized

As each sub-network is connected to the main network through a small set of synapses, the network can be compartmentalized such that each sub-network can be processed in a distributed manner.

- The network can make a generalization from similar stimuli

Given that the network implements a Hebbian learning model, similar attributes of different stimuli provides synchronous stimuli for neurons that respond to the attributes, hence forming a general representation of the object.

- The network is time-dependent

By allowing the value of each neuron to decay and inserting a delay of time t in each processing stage, the network can be made time-dependent. The time of arrival of each stimulus signal will make a difference as to which neurons are activated and

which synapses are strengthened. As the decay of the neuron state with respect to time need not be linear, the current state of the network is not only dependent on previous stimuli but also on how long ago the stimuli occurred.

- The network can associate causes with effects

As the network is time-dependent, it can associate current stimulus with a prior one, allowing it to associate causes with effects to the extent of the maximum decay time of the neurons.

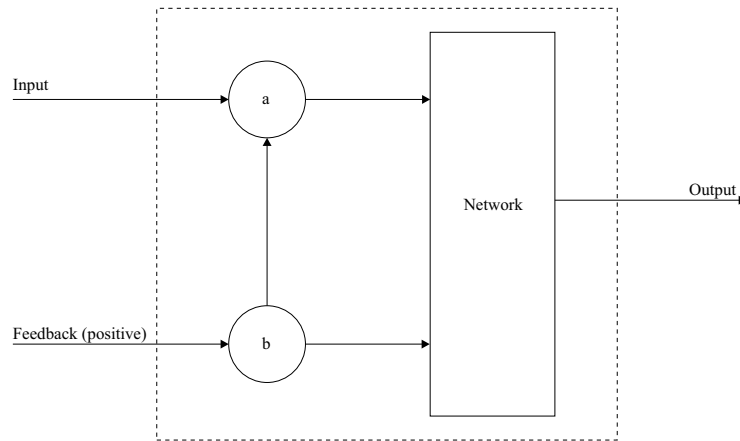


Figure 4.2a: A system that can take advantage of the hebbian learning system has stimuli and feedback processed by different neurons. When the feedback neuron fires, the input neuron is still active from the input stimulus. As the feedback is positive, both sides of the synapse are still active, thus the strength of the connecting synapse is increased.

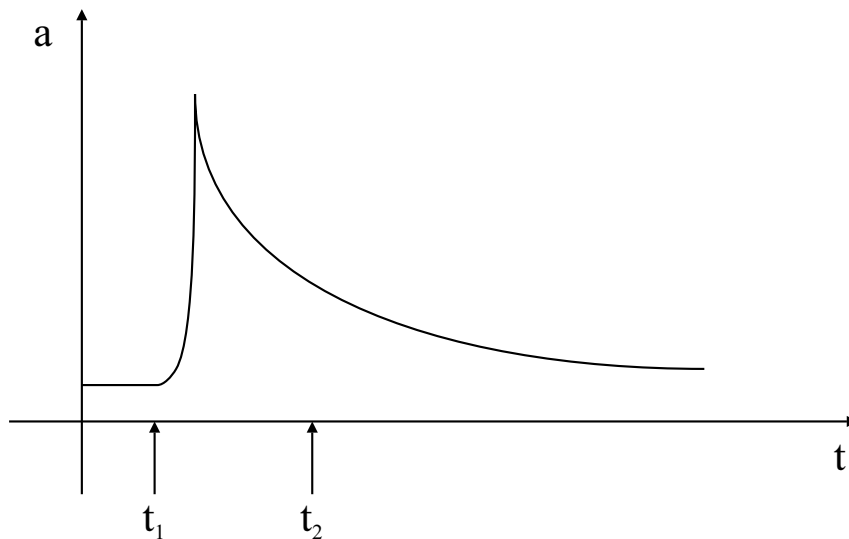


Figure 4.2b: When a neuron is activated at time t_1 , it may stay active for a period of time such that a feedback neuron can further stimulate it by activating it at time t_2 to reinforce the strength of the feedback synapse.

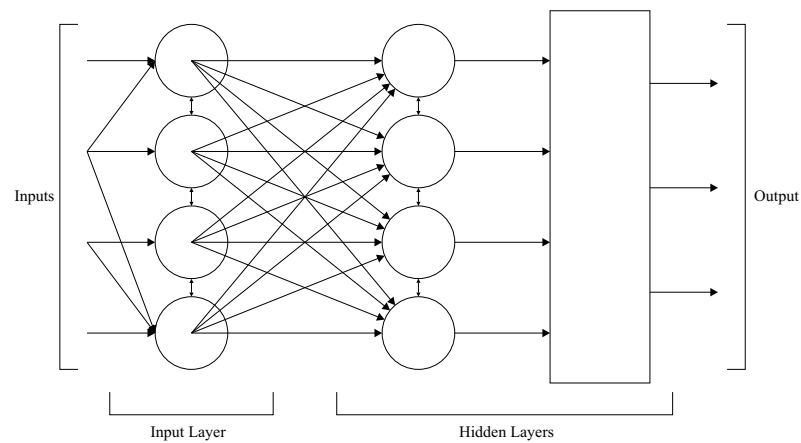


Figure 4.3: Each neuron is connected to other neurons in the same layer and the subsequent layer, and neurons in the input layer does not receive a complete set of the stimuli. There is no explicit feedback pattern, instead feedback is obtained through the same means as the original stimuli. Each neuron in the input may be a black box that represents a neural net or may be a neuron.

Chapter 5

Conclusion and Future Work

In this study we have reviewed the formation of memory and generalizations in both natural and artificial systems. The formation of memory in natural systems depends heavily on the hippocampus for memory related to events, while memory related to motor skills and basic responses seem to be stored without a need for the hippocampus. Natural systems can also form generalizations given a set of stimuli, and that such generalizations are not commonly formed in artificial systems.

Artificial systems form memory by storing prior training data as part of the neural net weights or as part of a set of rules that are applied to a set of stimuli. Artificial systems also happen to learn better when memory is limited at the beginning and then progressively increased, such that it learns simpler stimuli before learning complex stimuli. This behaviour mimics the behaviour of natural systems, and could be due to the neural net itself.

We have also reviewed different types of artificial systems, from cognitive systems to associative systems. Cognitive systems learn by making generalizations while artificial systems learn by associating stimuli with previous stimuli.

For further study, we present the following hypotheses:

A larger number of neurons will improve the generalization ability of a neural network. Currently neural nets can perform generalization albeit to a level that cannot match the performance of a cognitive system or a real human. This could be attributed to a limited number of neurons used.

Conversely, a limited number of neurons will restrict the learning abilities of a network,

therefore forcing it to learn only simpler stimuli, thus building a base for more complex stimuli to be more efficiently generalized.

We have also proposed the following model for further study:

A neural net that can be nested, i.e. each node in the network can be a simple neuron or a sub-network. This allows the system to be extensible, and therefore avoid the problem of insufficient neurons while at the same time minimising the computing power needed to process a set of stimuli. This model may be tested by examining its responses in a virtual world similar to an online role-playing game.

Acknowledgement

We are grateful to Kamalesh Basu, Huegesh Marimuthu, Looi Shiang Yong, Marcelle Leng, Tang Yong Han and Wong Leo E for their many interesting and stimulating discussions about the content discussed in this report. We also thank everyone who has helped us in one way or another.

Bibliography

- [1] Corkin, Suzanne (2002). “*What’s new with the amnesic patient H.M.?*”. *Nature Review Neuroscience* 3, 153-160.
- [2] Eliana Colunga & Linda B. Smith (2003). “*The emergence of abstract ideas: evidence from networks and babies*”. *Philosophical Transactions: Biological Sciences* 358, 1205-1214.
- [3] Michèle Fabre-Thorpe (2003). “*Visual categorization: accessing abstraction in non-human primates*”. *Philosophical Transactions: Biological Sciences* 358, 1215-1223.
- [4] Alan D. Braddeley (2002). “*The Psychology of Memory*”. *Handbook of Memory Disorders*, John Wiley & Sons, Ltd.
- [5] Atkinson, R.C. & Shiffrin, R.M.(1968). “*Human memory: a proposed system and its control processes*”. In K.W. Spence, “*The Psychology of Learning and Motivation: Advances in Research and Theory*”, Vol.2, 89-195. New York: Academic Press. Quoted by [4].
- [6] Harold J. Morowitz (1994). “*The Mind, the Brain, and Complex Adaptive Systems*”. Addison-Wesley, 1994.
- [7] Alan D. Braddeley (1986). “*Working Memory*”. Oxford: Clarendon Press. Quoted by [4].
- [8] Squire, L.R. & S. Zola-Morgan (1991). “*The Medial Temporal Lobe Memory System*”. *Science* 253(1991): 1380-1386.

- [9] J.A. Horne (2000). “*REM – by default?*” *Neuroscience and Biobehavior Review* 24, 777-797.
- [10] Jeffery L. Elman (1993). “*Learning and Development in Neural Networks: The Importance of Starting Small*”. *Cognition* 48, 71-99.
- [11] J. Zucker (2003). “*A grounded theory of abstraction in artificial intelligence,* ”. *Philosophical Transactions: Biological Sciences* 358, 1293-1309.
- [12] M. Minsky (1974). “*A framework for representing knowledge*”. MIT A.I. Lab, Memo 306.
- [13] S. Haykin (2003). “*Neural Networks, a comprehensive foundation*”. Prentice Hall, 2003
- [14] D. O. Hebb (1949). “*The organization of behavior: A neuropsychological theory*”. New York: Wiley, Quoted by [\[13\]](#).
- [15] G. S. Stent (1973). “*A Physiological mechanism for Hebb’s postulate of learning*”. *Proc. Natl. Acad. Sci. USA* 70, 997-1001.
- [16] A. Zador, C. Koch, T. H. Brown (1990). “*Biophysical model of a Hebbian synapse*”. *Proc. Natl. Acad. Sci. USA* 87, 6718-6722.
- [17] R. Spiegel, I. P. L. McLaren (2003). “*Abstract and associatively based representations in human sequence learning*”. *Philosophical Transactions: Biological Sciences* 358, 1277-1283.