NTNU
Innovation and Creativity

Relevant Concepts of and a Framework for Conceptual Representations based on Connectionism

henning veflingstad

Master of Science in Informatics
Submission date: June 2007
Supervisor: Keith Downing, IDI
Co-supervisor: Sule Yildirim, ØSIR

Norwegian University of Science and Technology
Department of Computer and Information Science

The Project Description

It is accepted that it is difficult to formalize a theory of how humans represent knowledge [1]. However, there are theories on symbolic representations and also on sub-symbolic representations. The Symbol grounding problem is a well known problem of symbolic representations [2].

Early work on sub-symbolic knowledge representations started with the work of Rumelhart et al. [3]. Later, RAAMs were proposed by Pollack [4] to provide a connectionist representation of symbol structures. Elman showed that hierarchical categorization can emerge from a sequence of words fed into a neural network, namely SRN (Simple Recurrent Network) [5]. These categories represent lexical classes.

The work relevant to categorization is extended to work on evolution of language or origins of language recently [6]. However, different from the earlier work by Pollack and Elman, the main purpose of this work is not to propose means of representing symbolic structures by sub-symbolic (connectionist) means but to conceptualize objects of the real world. In [7], conceptualization or categorization of reality is equated to 'meaning'. The representation of meaning is achieved by representing each object in an agent's environment by employing a feature set. That is, objects are conceptualized by features and discriminated by discriminating features among conceptualizations.

This project requires a survey on the literature for conceptualization models including above stated ones, [8, 9, 10] and others as the first step. In [11], in Section 3 and Figure 1, a model of "concept to concept associations" is presented. However, a conceptualization model that is used to represent each concept has to be either borrowed from the existing literature or needs to be designed before the associations between concepts can be implemented. The later stage is to provide the associations between the concepts. The association models can be inspired from associative networks or a network that would represent Hebbian way of strengthening and weakening of weights between concepts.

References in Project Description

[1] The essence of neural networks, Robert Callan, ISBN 0-13-908732-X, 1999.

[2] Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42:pp. 335-346. Worskshop paper.

[3] Rumelhart et al: Learning Internal Representations by Error Propogation (from Rumelhart et al, Parallel Distributed Processing, 1986).

[4] J. Pollack, Recursive Distributed Representations (1990).

[5] Jeffrey L. Elman Finding Structure in Time (1990) Cognitive Science

[6] Origins of language

[7] Steels, L. (2004) The Evolution of Communication Systems by Adaptive Agents. In: Alonso, E., D. Kudenko and D. Kazakov (eds.) (2004) Adaptive Agents and Multi-Agent Systems. LNAI 2636. Springer Verlag, Berlin. p. 125-140.

[8] Cangelosi, A. (2004) The sensorimotor bases of linguistic structure: Experiments with grounded adaptive agents. In S. Schaal et al., editor, SAB04, pages 487--496. Los Angeles: Cambridge MA, MIT Press.

[9] Dorffner G., A Step Toward Sub-Symbolic Language Models without Linguistic Representations, in Reilly R., Sharkey N.(eds.): Connectionist Approaches to Language Processing (Vol. I), Hove: Lawrence Erlbaum, 1992.

[10] Michael Lewin, Concept Formation and Language Sharing: Combining Steels' Language Games with Simple Competitive Learning (2002)

[11] S. Yildirim, "Innate Planning Mechanisms", IJCAI'05, Modeling Action Selection Workshop, Extended Abstract.

Contents

List of Figures

List of Tables

Abstract

In this theses, the role of conceptual representations in higher level cognition is investigated. In the field of AI much work has been done on conceptual representations. I review important parts of this work before presenting a framework for utilizing conceptual representations in higher level cognitive function. It is proposed that algorithms are represented non-symbolically in the brain and that they employ non-symbolic concepts in their computation, the result being that more complex though arises. Three levels of cognition is proposed, with emphasis being on the conceptual level where the proposed non-symbolic algorithms reside and high level human thinking occurs. A case study on novelty detection in and between modalities is also presented.

acknowledgments

I would like to thank my supervisors Keith Downing and Sule Yildirim for all the help with this thesis.I especially want to thank Sule for putting so many hours and making my master theses a very exciting project.

1 Introduction

The motivation for this thesis was to investigate the role of conceptual representations in higher level cognitive behavior within the context of parallel distributed processing (PDP). Much work has been done on how to achieve conceptual representations in PDP networks. The literature on how to employ them in higher level cognitive behavior is more sparse. A literature study on conceptual representations from the fields of AI, psychology and neurophysiology was undertaken to gain insights as to the nature of conception. The relevant of which is presented in this thesis. The concept is a complex issue. It affords categorical inferences allowing us to construe the world. Using PDP one can glean aspects of the concept, how it arises and its processing. PDP was thus chosen as the methodology for this project, the objective being knowledge and an implementation employing concepts in its processing. Two case studies were performed: One investigating the role of algorithms in high level thought. The other, investigating how novelty can be used to validate associations. That is, how can an agent know that some patterns of activity are meaningful while others are just noise?

This thesis is structured as follows: In part 1 an overview of the relevant literature is presented. We will see how semantic task performance arises from hidden layer representations in PDP networks. This will make the concept more concrete, grounding our discussion in hidden layer representations. Then we will view the concept as arising from perceptual simulations within the theory of perceptual symbol systems (Barsalou, 1999). This will enlighten the means by which a perceptual theory of cognition can support

the complexities of thought. Next the issue of dimensionality of the representation is reviewed, making evident that stimuli must be reduced for generalization to occur. Afterwards, the issue of cognitive control is discussed, showing how conception may arise due to dopamine and prefrontal cortex interaction. Then, categorical perception is introduced and how categorization may influence perception is rewieved.

This comprises the foundation on which the my work is presented. Based on this I will in part 2 present a framework in which algorithms are represented. These algorithms are non-symbolic and they operate on non-symbolic concepts and if-then rules to achieve higher level thought. An implementation of a non-symbolic summation algorithm is also presented showing the feasibility of the approach. Based on the information previously reviewed I will also speculate as to where and how these algorithms are learned, but since this is a complex issue it is left out of the simulation. The next section discuss how novelty can allow an agent to know whether it has seen a perceived instance before without explicitly referencing memory. Then some shortcomings of the novelty model is identified and a solution is proposed. Before conclusions are presented, language's involvement in conception is discussed.

Part I
Overview of relevant litterature

2 Semantic cognition

In Rogers & McClelland (2006) a parallel distributed processing (PDP) approach is taken to semantic cognition. Semantic cognition is a complex phenomenon exhibiting behaviors still to be reconciled by a functional theory of cognition. They successfully show how PDP can account for many of these behaviors and thus add to the promise of PDP as a tool for investigating cognitive phenomena and implementing conceptual systems. I will in this section review some of the simulations they present. Specifically, we will see how semantic performance in PDP networks arise from similarity based generalization on patterns of activity in the hidden layers of neural networks. This will provide us with a powerful tool for understanding the nature of distributed representations in neural networks. Before we continue, a formal description of a semantic task is necessary. I will borrow this from Rogers & McClelland (2006)):

> "We define semantic tasks as those that require a person to produce or verify semantic information about an object, a depiction of an object, or a set of objects indicated verbally (e.g., by a word). By semantic information, we refer to information that has not previously been associated with the particular stimulus object itself (though it may well have been associated with other objects), and that is not available more or less directly from the perceptual input provided by the object or object depiction." (p.2)

Thus, verifying that a depicted object is a cat, or that it can purr, are clearly semantic tasks as long as the information has not previously been directly associated with the particular picture. However, verifying that two depicted cats are the same color is not a semantic task as the judgment can be made directly from the pictures (Rogers & McClelland, 2006)).

2.1 A PDP theory of semantic cognition

In PDP a set of simple processing units is interconnected with a set of weights. The units represent information by having a level of activity and convey this information by passing the activity to neighboring nodes scaled by the weight between them. The knowledge is thus encoded in the weights while the "state of the world" is encoded in the activity of the processing units. This is informed by, and is very similar to, neurons in the brain forming synapses with each other. Information in the network can be represented using either localist or distributed representations. When using localist representations, a single instance is represented as "owning" a unit—A unit is "turned on" when the instance is present—, while in distributed representations several units are used in representing an instance and each instance

is uniquely identified by a specific pattern of activation over these units. Distributed representations are particularly important to similarity based generalization as gradients of similarity can be learned. The networks learn by adjusting the weights between individual processing units in small increments according to an error measure. Sets of units are typically gathered in layers representing meaningful properties of the world while weights between these layers represents a learned mapping from one layer to the next. The maximum and minimum activity a unit can reach is constrained by an activation function transforming the sum of inputs a unit receives. In localist encodings a units is often "turned on" by setting the activity of its unit to this maximum, while it is "turned off" by setting it to its minimum. In this and later sections, a type of networks coined *feed forward networks* will be investigated. In figure 1 there is an illustration of a feed forward neural network. The picture is from Rogers & McClelland (2006) and is the semantic system our discussion will be centered around. It is treated as a simplified model of of experience with objects in the world and of spoken statements about these objects (Rogers & McClelland, 2006). The network has five layers of which two are input layers (item & relation), two are hidden layers(representation & hidden), and one is an output layer (attribute). The job of the network is to learn a set of weights able to map a pattern presented at the input layer(s) to an intended pattern at the output layer(s). Since the network has hidden layers it must also discover the necessary activation patterns across these layers to perform this task. The network thus learns an internal representation of the items presented to the network. The item, relation and attribute layers uses localist encoding. Units in the item layer represents an animal or plant and stands for an occurrence of the object itself, not its name. In the representation layer each unit in the representation layer represents the context in which the item is encountered, while the units in the attribute layer represents the predicted consequence of an encounter with the item in a given context. When the network has been trained sufficiently it will complete proposition such as *canary ISA with living thing, animal, bird and canary.*

There are some important issues to discuss before continuing. These concern , by what means does the network arrive at the internal representations, what issues arise from using localist encoding, and what kind of learning does our system do. The assumption here is that information from all modalities converge in a common semantic representational system. As we will see later, this results in an amodal representation. The use of localist encoding implies that the world has already been categorized before arriving to this semantic system, and as such, the network can be viewed as learning facts about concepts. That is, it learns facts about concepts and then generalizes to instances due to the semantic knowledge of some underlying system. Thus some underlying system has done a lot of the semantic task for us in that they have associated a perceived object with a category.

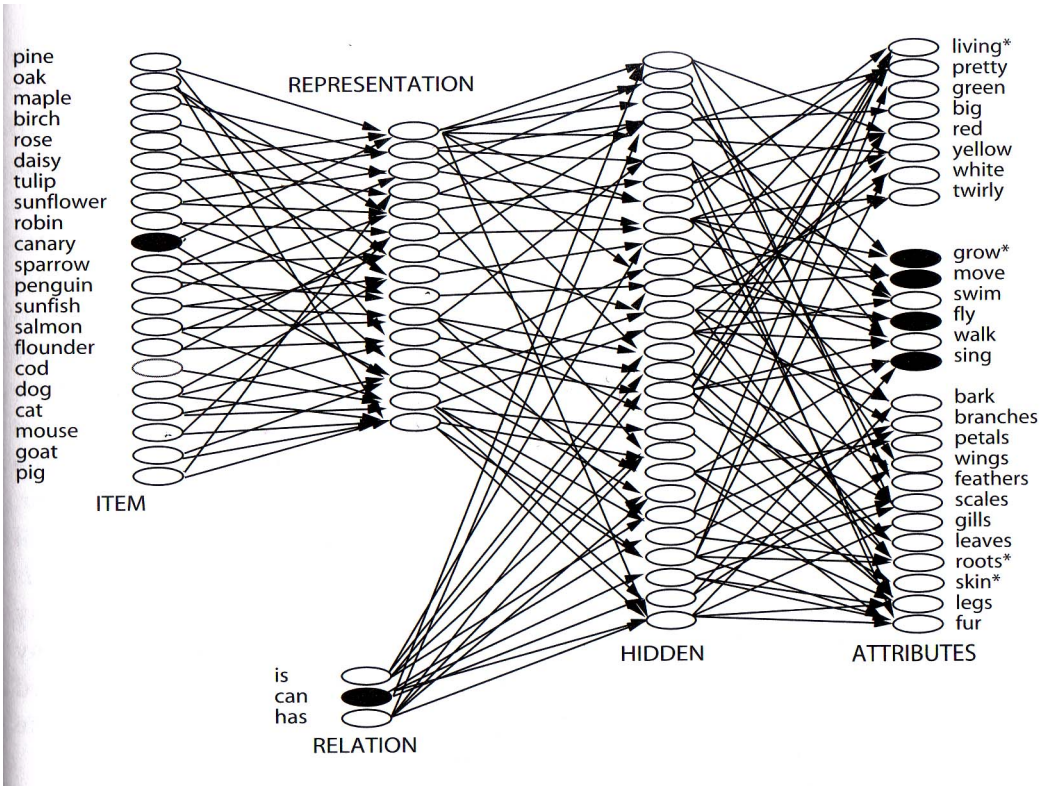

Figure 1: This figure shows the architecture of the network we will discuss here. There are however some units in the output layer that are not shown. These are the units naming the instances presented to the input layer and are the names at the input in the figure, the indermediate names fish, bird, flower and tree, and the superodinate names animal and plant. The context unit for probing names is not show either—ISA context. This picture was originally presented in (Rogers & McClelland, 2006).

It is reasonable to assume that the task of categorizing a perceived object such as a bird requires the knowledge that birds has wings, a beak and legs, and as such (at least some of) the knowledge present in our network has to be present in some underlying system. However, as stated in Rogers & McClelland, (2006) the inputs can be seen as a course discretization of the perceptual similarity relation among the perceived objects. Still this issue remains since the act of encoding encountered birds as similar would still require semantic knowledge of birds (an ostrich and a canary differ somewhat in perceptual similarity). The use of localist encoding also constrains the network in that novel instances can not be presented to the input. This however, is gracefully circumvented by a technique called *back propagation to representation* (Rogers & McClelland). Using this technique the representation for a novel instance in the representation layer is induced from its attributes and a context, and as such, the concept *pretty animal* can be represented.

The type of learning this system does is reminiscent to a person slowly learning from interacting with the environment over time. As the network interacts with its environment, the weights are adjusted in small increments using online learning. Fast learning such as a person being told that a perceived object is an animal is done using back propagation to representation, for then storing the derived representations in a fast learning system, a function attributed to the hippocampus.

The network presented will learn the weights necessary to map patterns at the input layers to the intended patterns at the output layer and in the course of this learning also develop internal representations of the instances. These internal representations are distributed, and as such, similarity is encoded as a continuous parameter (as opposed to the localist layers). Another important point is that the network discovers the hierarchical structure of the domain on its own based on similarity relations among the presented patterns. Thus one escapes the inherent perils of explicitly structuring the agents internal world (Brooks, 1991)

2.2 Progressive differentiation of internal representations

An interesting property of PDP networks is that their internal representation develop in a coarse to fine manner resembling the conceptual development of a child. Because of PDP's resemblance to the human brain it offers important insights as to why this happens in a child's brain. Understanding the process of progressive differentiation also gives us a better understanding of what properties in the world influences learning in PDP networks and thus why learning proceeds as it does. It turns out that coherently covarying properties in the patterns presented to the network guides the fine to coarse differentiation of the internal representations and thus what properties are learned first. This happens because the properties that covary the most with

each other will contribute to an error in the same direction while the more idiosyncratic properties will tend to cancel each other out. Another factor is that the error back propagates more strongly through weights representing meaningful mappings. For example, when learning starts all the weights are initialized to random values, and as a result all the patterns will derive the same internal representation. The property *living* in Figure 1. will thus produce an error on the same direction for all patterns (they are all living things) while the property *swim* will vary across the instances (only the fish can swim). Once *living* has been learned, the properties reliably separating the animals from the plants will be the ones with the most coherent error signal and thus guide learning. The internal representations now separate into animals and plants. As mentioned above there is another factor also influencing this fine to coarse differentiation: the errors back propagate more strongly through weights representing meaningful mappings. With reference to figure 1, the magnitude of the changes made to the weights feeding into the representation layer depends on how this will help reduce the error on the output nodes. This again depends on whether the weights feeding forward from the representation layer can make use of the induced changes. The errors will thus back propagate more strongly through the weights representing properties the network has learned or is currently focusing on.

It is worth to notice that although the frequency of a property in the patterns presented to the network does influence the ease of which it is learned, it is how the property covaries with other properties in the patterns that enforces the hierarchical structure of the internal representations. It should also be clear from this discussion that properties common to all patterns does not induce any hierarchical structure of the internal representations. As these properties are among the first to be learned, the internal representations are still very similar and the error from these properties will cause them to change in almost the same way.

In Rogers & McClelland (2006) they showed that by adding noise to the network weights the process ran in reverse with specific level categories being lost first, thus closely mirroring the effects seen in semantic dementia. The effect was attributed to specific level categories occupying a smaller representational space than intermediate and superordinate categories. Specific level categories are the categories for each instance (e.g. *mouse* and *canary*). intermediate level categories are the ones intermediate in the hierarchy (e.g. *tree* and *fish*), and superordinate categories are the top level ones (i.e. *animal* and *plant*). Intermediate categories will later be referred to as basic level categories also.

Thus, when noise was induced, representations of specific level categories ended up outside their representational space and thus induced an incorrect or more general naming response. PDP networks thus discovers the hierarchical structure intrinsic to the presented stimulus. This discovery proceeds in a coarse to fine manner throughout learning, while the process is reversed

when the network is lesioned.

2.3 Salience of coherently covarying properties

We have so far seen that the network structures its internal representations based on coherent covariation of properties in the patterns taught to the network. An interesting questions is how this influences generalizations from novel instances. Consider birds, they all have legs, a body, a head, and wings; However, the fact that birds have wings are more important to the concept birds then the fact that it has legs. It has been observed that at nine months of age, children are able to discriminate between perceptually similar objects from different categories at a superordinate level, but failing at this at a more intermediate level. Eleven month old children however succeed at discriminating at a more intermediate level (Rogers & McClelland, 2006). The ability to discriminate on the basis of semantic properties not directly available from the perceived stimulus seems to develop in a coarse to fine manner as the infant has more experience with the world. To investigate how the network in Figure 1 faired with respect to this task Rogers & McClelland (2006) trained the network on a set of plants and animals. The plants were divided into trees and flower; The animals into birds, fish, and mammals. During the course of training the network was never taught to activate any name units for the instances, so any hierarchical representation arising was the result of the network discovering the hierarchical nature of the domain based on its covarying properties (such as can fly, has skin, etc). They also trained a version of the network above employing distributed representations for the item layer. The patterns introduced to the item layer of this network were produced to respect perceptual similarities directly observable from a static environment. Thus all flower patterns were more similar to each other than they were to the tree patterns, but more similar to the tree patterns than they were to the fish patterns. Both networks discovered the same internal hierarchical representation. The distributed network thus represented all flowers (or birds, fish, etc) as equal in the sense that they clustered together in the representational space. The weights projecting from the item layer to the representation layer thus served the function of grouping together perceptual different items on the basis of a few abstract shared properties not present in the input (Rogers & McClelland). This shows how the output representations are responsible for shaping the internal representations in the network. It also makes clear a separation of function in the network: The weights from the item layer to the representation layer maps perceived instances in the world into conceptual representations while the weights projecting forward from the representation layer are responsible for categorical inferences. There is however an important point to mention: in the simulations presently discussed, all output patterns were, even within a category, slightly different from the other patterns—This was true for both

the distributed and localist version. The resultant conceptual representations for all patterns were thus slightly different from each other. A concept in this respect is a region of internal space occupied by representations from the same category, while each single representation is a conceptualization of a perceived object.

To test how the networks generalized from novel inputs they devised twenty novel patters: five trees, flowers, birds and fish. These patterns were made in such a way that four of the patterns (habituation patterns) within a group were similar to each other and different from any other pattern, while one pattern (test pattern) within each group was dissimilar from the others within the group but very similar to the habituation patterns from a cross category group (intermediate level) and a cross domain group (superordinate level). The cross category and cross domain similarities were based on more idiosyncratic perceptual similarities, while the test patterns' similarity to the within category habituation patterns were based on coherently covarying properties. A test pattern for bird might share *big* and *green* with the fish habituation patterns while sharing *wings* with all the bird habituation patterns. It was thus more similar to the fish with respect to appearance while conceptually being more similar to the birds.

The simulation recorded the internal representations from the novel stimulus during different stages of training. They presented the four habituation patterns for a category, recorded the activation patterns, and computed the centroid of these patterns. Then, the within category test item was presented and its distance to the centroid recorded. The same was done for the cross category and cross domain instance. The distance between the centroid and the different test items was interpreted as a measure of how novel the model found the test items with respect to the habituation patterns. From these distances, the probability for the model choosing an object from a contrasting domain was computed. The model chooses to explore the stimulus it finds more novel. Both simulations showed a coarse to fine differentiation of the internal representations during learning. The probability for choosing the contrasting stimulus started out with being near chance, but as training proceeded it first started to reliably choose the between category item from the contrasting global domain, and later the between category item from the contrasting intermediate domain.

The model had grouped the within category test items as more similar to the habituation patterns even though the between category test items were perceptually more similar. The model thus construed as similar those patterns sharing coherently covarying properties, while treating as dissimilar patterns sharing many more properties with the habituation patterns (These properties did not covary with the properties of the habituation patterns). It would seem that the model lends special weighting to the coherently covarying properties.

2.4 Typicality and frequency effects

In Rogers & McClelland (2006) typicality and frequency effects are investigated with respect to a phenomenon called basic level categories. Basic level categories are often intermediate categories—Somewhere between the superordinate and specific categories. They have been observed to be among the first children learn to name during early stages of lexical acquisition. However, during this stage they also show a tendency to overextend these categories to similar but semantically unrelated objects. Adults have been shown to prefer naming at the basic level for typical category members while more atypical members are named at a more specific level. For example, bird is a basic level category which adults usually prefer to use when naming a typical instance such as a canary, a penguin, however, is often named with its more specific name—Penguin. People are also faster at producing naming responses for basic level categories. With expertise basic level category effects tend to disappear with the experts preferring to name at a more specific level. In semantic dementia, basic level effects give away for naming at a more specific level as the disease proceeds.

Some of this seems to conflict with the coarse to fine differentiation of knowledge we have previously discussed. For example, how can basic level category labels be among the first to be acquired when the internal hierarchy first acquire the general categories? How can there be a preference for basic level categories during lexical acquisition, while superordinate categories being the ones most resistant to damage? Rogers & McClelland (2006) shows that typicality and frequency coupled with the properties previously discussed can account for all of this. To be able to manipulate the level at which the network in Figure 1. named instances they added three new ISA units to the relation layer. These were ISA-general, ISA-basic and ISA-specific. In the ISA-general context, the network was trained to respond with a general name while all other name units having zero as their target. The same was true for the other two contexts except for the required naming response being the one they imply. The frequency of pattern presentation during learning was manipulated such that basic level names were the target for an instance three times more often than specific or general names. The rest of the contexts were manipulated in such a way that they appeared almost as often for animal and plant items (the difference was insignificant). During training the network showed a preference for naming at the basic level with this naming unit being on an average the one most strongly activated. The difference between activation of the basic unit and the general- and specific unit was greatest during early stages of training and became attenuated as training proceeded. The basic level was, however, always the one showing the highest activation. When noise was introduced to the weights, performance degraded in a fashion similar to dementia with basic level responses giving away to general ones.

To investigate the importance of frequency on the results above, the simulation was run again with all naming responses being equally frequent. In early stages of training there were still, on an average, some basic level advantages. However, as training proceeded, the difference between the naming levels became negligible. At a closer inspection it turned out that the naming behavior of the network depended on the typicality of an item, with typical items exhibiting basic level advantages and atypical ones showing advantages for the specific level. That is, items sharing many properties with the prototypical element of their basic level category were the ones showing basic level advantages. Typicality is thus enough for the model to exhibit basic level advantages while naming frequency further facilitates this process.

In the introduction to this sub section it was mentioned that basic level advantages were attenuated with expertise—Experts prefer to name at a more specific level and are also faster at verifying properties at the specific level. Rogers & McClelland (2006) investigated the effects of expertise by making the network an expert on different basic level categories. This was achieved by making items from the domain of expertise appear more frequently than the other items in the training set. Two different levels of expertise were investigated. At one level, items in the expert domain was eight times more frequent than in the novice domain, at the other level they were sixteen times more frequent. Basic level names were three times more frequent than specific and general ones. The network showed a strong tendency to name at the basic level in the novice domain, while activation levels in the expert domain were almost equal for the specific and basic level names. Expertise thus attenuated the basic level advantages. Moreover, the mean Euclidean distance between the internal representations of the items in the expert domain, and between the items in the novice domain was recorded. The representations in the expert domain had a larger distance between them than did the ones in the novice domain. Expertise had thus differentiated the items in the expert domain more. In a simulation similar to the one above except that only a single item was the one most frequently presented, naming responses to this item was shown to occupy a larger semantic subspace than other items from the same category. That is, when the internal representation of the familiar items was slowly translated towards the representation for an item from the same category, the naming response for the familiar item was the one being the most active over the largest distance of space. Also, when queried for basic level names at different stages of training with different items, the network showed a tendency to overextend the name of the familiar item to semantically related items during early stages of training. During degradation, the familiar name was misapplied to other items more frequently than the name for more novel items. This effect became stronger with the amount of noise introduced.

The network thus exhibits behavior similar to that observed in human

subjects. It exhibits basic level advantages for highly typical items. As these items are named more frequently with their basic label this advantage becomes stronger. During degradation, the naming responses for specific and basic level names are the first to go while the general names are more persistent; familiar names are also overextended more frequently. Basic level advantages also become attenuated with expertise. What are the properties of the network responsible for this behavior? Since items being similar receive representations close in internal space, similarity based generalization causes their specific properties to interfere with each other during learning. The properties they share, however, does not interfere with each other and are thus easier to learn. As the frequency of presentation increases, the network receives more pressure to differentiate each of the frequent items. As a result, similarity based generalization does not cause as much interference and the specific properties are easier to learn. Since the main structuring force guiding the internal representations still is coherent covariation, the general categories occupy the largest region of internal space and is therefor more resilient to noise. Because frequency of presentation induces a greater separation of familiar items, the same effect causes them to be overextended during degradation.

2.5 Conclusions

In this section we have seen that semantic task performance can arise in PDP networks as a result of similarity based generalization of regions of internal representational space. These regions correspond to groupings in the hierarchical structure of the domain as dictated by the coherently covarying properties of the instances taught to the network.

3 Perceptual symbol systems

As opposed to classical amodal symbol systems such as frames and semantic nets, Barsalou (1999; 2003a; 2003b) presents a perceptual theory of cognition in which perceptual symbols gets integrated into simulators capable of simulating all aspects of a concept. This section will present this theory, Perceptual Symbol Systems (PSS), and look closer on some of the important aspects of it. Specifically we will see:

- how perceptual symbols gets extracted from experience and integrated into a simulator,
- how these simulators combine productively
- how they support abstract thought and propositional construal,

We will start with an account of how this theory is different from classical amodal symbol systems and recording systems, then move on to perceptual symbols, how they get integrated into simulators, and how these simulators can support higher level cognitive functions such as abstract thought and propositions. This section draws heavily on material from Barsalou (1999).

3.1 Classical symbol systems vs. PSS

Classical symbol systems relies on a process called transduction to derive a symbolic account of presented stimulus. Upon perceiving a chair, the transduction process may deliver a schematic representation of a chair which is then manipulated by the conceptual system according to its syntactic properties. These systems propose a modular account of cognition in which the conceptual system is its own module separate from perception, a module wherein all symbols are arbitrarily related to the states from which they came, thus being amodal. In contrast to this, PSS represents symbols as records of perceptual states in the modality they belong: the perceptual symbol for a perceived chair is represented in the vision modality, while the sound it makes when pushed along the floor is represented as a perceptual symbol in the auditory modality. As an agent perceives scene with a chair in it, a perceptual state will arise in the vision modality. Selective attention focuses on certain aspects of this state such as the chair and its parts (seat, legs, material it is made of, etc.) and stores records of them in long-term memory as a perceptual symbol. Later on, these symbols can be retrieved and reenacted in either off-line processing (e.g. thinking about a chair) or on-line processing (e.g. filling in missing parts of a partially occluded chair during perception). These perceptual symbols are modal and analogical. Why is best described in Barsalou's (1999) own words:

> "They are modal because they are represented in the same systems as the perceptual states that produced them. ...Because

> perceptual symbols are modal they are also analogical. The structure of a perceptual symbol corresponds, at least somewhat, to the perceptual state that produced it".

Since perceptual symbols are modal and analogical, they avoid the symbol grounding problem (Harnard, 1993) inherent to all classical symbolic theories. I will borrow from Harnard (1993) to explain the symbol grounding problem:

> "the real problem of symbol grounding is that the interpretation of the symbols, whether or not it is unique, is not intrinsic to the symbol system: It is projected onto it by the mind of the interpreter, whereas that is not true for the thoughts in my mind."

The analogical modal symbols thus avoid the symbol grounding problem because meaning is derived by their appearance and relations among the symbols are reflected by this appearance.

Another issue with these systems was put forward by Brooks (1991) :

> "...each animal species, and clearly each robot species with their own distinctly non-human sensor suites, will have their own different Merkwelt. Second, the Merkwelt we humans provide our programs is based on our own introspection. It is by no means clear that such a Merkwelt is anything like what we actually use internally–it could just as easily be an output coding for communication purposes (e.g., most humans go through life never realizing they have a large blind spot almost in the center of their visual fields). The first objection warns of the danger that reasoning strategies developed for the human-assumed Merkwelt may not be valid when real sensors and perception processing is used. The second objection says that even with human sensors and perception the Merkwelt may not be anything like that used by humans." (p 144)

It should be noted that some of this argument also applies to connectionism as well. Connectionism has its roots in knowledge about the human brain—Neurons and how they function. However, an artificial feed forward neural network is by no means biologically plausible. First of all, there are many types of neurons in the brain, they are interconnected in many different ways and they use many different types of neuro-transmitters to convey information. Second, back propagation, a commonly used training method, is also biologically implausible. Neural networks are simplifications of the neuronal architecture in the human brain, and when chosen to represent intelligence one assumes that intelligence can be represented with them. Also,

by selecting a specific learning method (e.g. back propagation) you implicitly constrain how the internal world of the agent is structured. However, as opposed to classical symbolic theories of cognition, connectionism does not explicitly structure the internal world of the agent, this world will, as we have seen, develop depending on the chosen interconnectivity, learning method, and the agents interaction with the world.

Regarding modality, Barsalou makes a claim about feed forward neural networks particularly important to this theses: He claims that that conceptual representations in feed forward neural networks are amodal. This would mean that implementing a modal conceptual system using feed forward neural networks is impossible—After all, the conceptual representations are amodal. This claim is based on two facts: First of all, hidden layer activity in feed forward neural networks are usually interpreted as conceptual representations. Second, before training the networks, their weights are initialized to small random values, thus optimizing training. If the weights were initialized to zero the networks could not learn. Now, Barsalou claims that because the hidden layer activity is often interpreted as conceptual representations, one has a modular architecture where the conceptual system is separated from the perceptual system; That is, the perceptual system resides on the input layer while the conceptual system on the hidden layer. The conceptual representations are arbitrarily related to the perceptual system because they reside in a different module and the weights between the modules are seeded with random numbers. That is, because the weights are randomly seeded at each trial, the resultant representations will vary from learning trial to learning trial, and within each trial their representation will be arbitrarily related to the perceptual system. However, he admits there to be certain invariants between learning trials (Barsalou, 2003a). In an earlier section we saw that the organization of the conceptual representations depend on the similarity structure of the instances presented to the network. That is, similar instances will be close in conceptual space wile dissimilar instances will be farther apart. Also, the amount of space "allocated" to an instance depends on its frequency of presentation, while how easy an instance is to learn depends on how similar it is to other instances. The information presented in that section would seem to suggest that even though the resultant conceptual representations are arbitrarily related to the training instances, their relation to each other depends on the similarity structure of the domain and stays invariant from trial to trial. Moreover, it is these invariants that support the conceptual representations performance in semantic cognition (e.g. basic level advantages in lexical acquisition). The overall error of the network on a given trial may vary depending on the initial weights (e.g. given a certain number of training epochs, the network may reproduce the required pattern of activity on the output units with differing level of precision), the performance of the network in semantic cognition, although dependent on the overall error, is invariant from trial to

trial given that the network reaches a certain level of precision.

This suggests that the initial weights, influencing the time it takes to reach a certain level of precision, does not influence the properties of the representations that justify their conceptual interpretation. These properties are invariant of the initial weights and depend on the similarity structure of the domain and frequency of presentation. With this said, it is still true that conceptual representations are arbitrarily related to the original perceptual states. In fact, this can not be disputed. However, with the previous discussion as a reference, the arbitrariness of these conceptual representations are not considered as an issue in this thesis as this does not seem to influence the behavior of the system of interest here.

There is one last point I would like to make with respect to the modal interpretation of symbols/hidden layer activity. Later in this thesis auto-encoders (Hinton, 1990) will be used to derive hidden layer representations in different modalities. In one of the simulations, one auto-encoder has an intended interpretation as the visual modality while another has an intended interpretation as the auditory modality. In each modality, the hidden layer representations develop with respect to similarities in the domain of this modality. When a representation is present at the hidden layer of the modality from where it came, it has a valid interpretation in that it will produce as valid response, as well as having relations to other representations in that modality reflecting what is learned about them. If I were to present this pattern to the other modality, this would not be true. With respect to this I pose the following question: Is it not the modal interpretation (by the system) of a symbol that justifies its modal interpretation by us. This question is not answered here, but it reflects underlying assumptions of simulations and discussions to come. That is, a pattern of activity will be given a modal interpretation by the system in which it has meaningful relationships to other patterns and has valid effects.

It is now time to continue our discussion about the important differences between classical symbol systems and PSS. Symbol grounding and the explicit structuring of the agents internal world was mentioned as problems inherent in symbolic approaches. Symbolic approaches, however has some very powerful properties; After all they have been intensely studied and used for many years with logic dating all the way back to the Aristotle (syllogisms). Symbolic systems has a clear semantics, they support productivity, and they lend themselves gracefully to computation. By having a clear semantics and supporting productivity, symbolic systems have the power to express almost anything. By productivity I am here referring to the ability to combine a finite set of symbols in an infinite number of ways. Symbolic systems do this effortlessly, while connectionist approaches has been rather lacking in this respect. However progress have been made:

Pollack (1990) showed with his recursive auto associative memory (RAAM) that connectionist approaches also support productivity. Hook (2005) has

proposed that microscopic loops between the basal ganglia, cerebellum and cerebral cortex might support recursion. If this is so, a connectionist approach to recursion should also be possible by mimicking this architecture. Associative chains can express chains of instances linked together. In Love (1999) asynchronous timing information is utilized to represent complex structures such as *John hit Ted*. BoltzCONS (Touretsky, 1990) can represent lisp-like lists. Even though connectionist approaches support productivity, it is often time consuming and awkward with respect to symbolic approaches. PSS, as we soon will see, supports this as effortlessly and gracefully as symbolic approaches. However, an implementation of PSS has, to my knowledge, yet to be.

It would seem that a perceptual theory of cognition escapes the problematic issues of symbolic theories while offering a promise of successfully implementing its more powerful features. With the previous discussion of semantic cognition as a reference, connectionism provides a powerful means to implement a perceptual theory of cognition. The next section will show that PSS offers a powerful description of how such a system must act to capture the complexities of thought.

3.2 Simulators

As records of related perceptual states get extracted by selective attention and stored in long-term memory as perceptual symbols, they become integrated into a simulator. This happens by storing the perceptual symbols in a frame. In the frame, spatial and content information is represented separately:

> "At one level, volumetric regions of an object is represented according to their spacial layout. At another level, the content of these subregions are represented as specializations." (Barsalou, 1999. p 590).

As more and more information accumulates in a subregion, this subregion constitutes a simulator of its own able to simulate different aspects of the region. Between specializations of the same subregion, inhibitory connections evolve, while excitatory connections evolve between specializations of different subregions belonging to the same instance. This way simulators for concepts such as chair, car, etc., develop. Event sequences also become organized into frames. When this happens the different subregions of the frame represents events being temporally separate. If an agent thinks about a specific chair, perceptual symbols for that chair is reenacted in their respective modality, a simulation has been run. A simulator thus is a concept, while a simulation a conceptualization. The chair can be simulated from different angles, with different materials and different colors. A simulator thus can simulate all aspects of a concept and also extend reality by adding

features to it never seen. An infinite number of simulations can be produced from a simulator. In addition to the perceptual symbols for the four sensory modalities (vision, audition, haptics and olfaction), perceptual symbols for introspective and proprioceptive sensations also become integrated into simulators.

Frames also support the notions of framing and background-dependent meaning. In framing a focal concepts is specified relative to a background concept and can not be specified independently of it. In background-dependent meaning, a focal concept changes as its background concept changes. A frame supports framing by organizing the background knowledge necessary to understand the focal concept. The entity frame for human organizes the background knowledge necessary for understanding foot. The perceptual symbols accessed when simulating a foot depend on the background frame and will vary depending on whether a human or dog is simulated in the background. Frames thus support background-dependent meaning.

Results from FMRI studies indicate that concepts are highly multimodal engaging modalities important when interacting with its respective members in the world. When people think about a hammer, activity can be seen in the motor areas of the brain. When they think about a bird, activity in the visual areas of the brain can be seen. However, the areas being active are typically not areas active in actual perception, they are immediately adjacent within the same modality-specific system (Barsalou, 1999). Brain damage can induce category-specific impairments affecting either living or non-living things. This has been attributed to a possible functional-visual separation of knowledge (Farah & McClelland, 1991). Also, when areas if the brain processing visual information is damaged, peoples performance in categorizing artifacts relying mostly on visual information degrades (e.g. birds). If motor areas of the brain is damaged, performance drops when categorizing tools. This support the contention that modalities important in processing an instance also are important in representing that instance. Important to the representation of a hammer is the perceptual symbols for how it feels to hold and how it is used. For birds, visual aspects is important. A simulator is therefor highly multimodal, simulating all modal aspects of a concept.

As stated in the last section, PSS supports productivity. Productivity is the ability to combine a finite set of symbols in an infinite number of ways. Barsalou (1999) proposes that productivity can be done in PSS by having simulators for relations among objects. A simulator for the relation *above* can be used to simulate "a lamp above a chair". This simulation can be combined with the relation *left of* to produce the simulation of "a lamp above a chair to the left of a table". By combining different simulators an infinite number of simulations can be produced.

When viewing an instance in a scene, information about this object is projected onto simulators in memory. If a simulators' frame contains a sim-

ulation of the instance or if it can produce a novel simulation fitting that instance well, this simulator will become active. This constitutes a type token mapping in which the perceived instance is a token mapped to its type (a simulator). This process can be construed as mapping the perceived instance onto the predicate ISA(chair, instance_a) and then producing the proposition *it is true that the perceived instance is a chair*. PSS thus supports propositions. The agent has also engaged in categorization by assigning a category (simulator) to the perceived instance. The simulator bound to the perceived instance contains a tremendous amount of information about it and therefor allows the agent to draw categorical inferences.

Barsalou (1999) also provides an account of how PSS might support abstract concepts such as truth and falsity. Important to an abstract concepts is a simulated background event sequence framing the concept. The abstract concept is a focal part of this sequence. Also important to the abstract concept is the perceptual symbols for introspective states. The concept of truth arises from validating event sequences against observed events. As an example an agent might be told that there is a cup on the table. As a result the agent will perform a simulation of this and try to establish its truth by mapping it to a perceived situation. Propositional construal is therefor important to abstract concepts. As an agent performs this sequence of events on a number of occasions, the introspective symbols arising from running the simulation, mapping it to a perceived scene and verifying its truth, becomes organized into a simulator for the concept of truth. In time the agent learns to simulate the experience of truth.

In PSS, language is also perceptual. As a person hear, speak, and read words, simulators for them develop. These simulators become connected with the simulators for their referents in the world. This way, language serves as a powerful means of building complex simulations and conveying them to other people.

3.3 Conclusions

In this section we have seen how a perceptual theory of cognition can be as expressive as classical symbolic theories while avoiding their inherent problems such as the symbol grounding problem. Another important problem with classical symbolic theories is that they are brittle and cumbersome. That is, they do not exhibit graceful degradation in the presence of error, and as the complexity of the modeled domain grows, so does the difficulty of modeling the domain. This section has also offered much information as to what a concept is: a concept arises by a skill for producing context-specific representations of a category (Barsalou et. al., 2003b). This implies that the concept itself is not represented but arises from the behavior of the system.

4 Dimensionality of representations

The dimensionality of the representational space is an important issue as the number of examples necessary for reaching a given level of performance grows exponentially with the underlying representation space (Edelman & Intrator, 1997). This is known as the curse of dimensionality. Thus, when implementing a perceptually grounded conceptual system one would need some way of reducing the dimensionality of the incoming stimulus, lest confusion prevail. Consider the task of human vision: *"the immediate successor of the retinal space in the processing hierarchy is, in primates, a million-dimensional space spanned by the activity if the individual axons in the optic nerve "*(Edelman & Intrator, 1997) . One could argue that all the information necessary to discriminate perceived objects are present in this million-dimensional space, and as such, no reduction is needed. There are, however, several obvious reasons why to prefer a reduction of the stimuli:

As noted above, as the number of dimensions goes down, the number of examples needed for reaching a given level of performance goes down. Thus, if a low dimensional representation (LDR) is extracted, subsequent cognitive processing will be better able to generalize. If the process responsible for reducing the dimensions is sensitive to informative dimensions and insensitive to uninformative dimensions, the resulting LDR will be more informative. For example, the process could compress highly correlated dimensions while possibly emphasizing dimensions being, for some reason, salient. Subsequent cognitive processing would thus, in addition to be relieved of dimensional pressure, be relieved of possibly distracting dimensions. In this section, a well known solution for unsupervised dimensionality reduction will be reviewed: auto-encoders.

4.1 Unsupervised dimensionality reduction

An auto-encoder (Hinton, 1989) is a feed forward neural network comprising three or more layers. The input and output layers has the same number of units while at least one of the internal layers has a number of units less than this. During training the auto-encoder is taught to reproduce at its output units, the same pattern as presented to its input units. Because of the internal layer, being narrower than the original stimulus, the network is forced to discover an internal representation with a lower dimensionality than the original. This dimensionality reductions is afforded by the compression of redundant features in the input (Hinton, 1989; Gluck & Meyers, 1993). That is, units correlated across the presented instances can be represented with fewer (possibly one) units at the internal layer. Another interesting property of these internal representations is that they are topology preserving (Edelman & Intrator, 1997; DeMers & Cottrell, 1993). With topology preserving I refer to the intrinsic relations among the stimuli being preserved to some

degree in the internal representations. Envision plotting a set of points on a rubber sheet and then crumbling the rubber sheet together—The resulting ball is our high dimensional space. Our internal representation unfolds this paper without tearing it—Thus being topology preserving—but may stretch the rubber in some places—Thus not being isometric. The auto-encoder being topology preserving has a very interesting implication: Stimuli being close in the original space will be close in the internal space. The likelihood of two stimuli being treated as similar (evoking the same response) is related to their distance in psychological space (Sehpard, 1987). Stimuli being similar will thus be conceived as being similar by the cognitive system operating on this LDR. In the discussion on semantic cognition we saw how typicality and frequency influenced the internal representations of neural networks. Typicality influenced how easy an item was to learn as well as what properties of it was emphasized by learning. Frequency of exposure moved an item further apart from its similar items, affectively allocating a larger piece of internal space to it and allowing its more individual properties to be emphasized. If the hidden layer units is our psychological space, then familiar items will be more distinctive than unfamiliar ones. Thus allowing cognition to better focus on these.

We will now return to the issue of dimensionality reduction, specifically the issue of non-linear mappings between the original and internal space. An auto-encoder only having one hidden layer can only learn linear mappings into the internal space. If an extra layer is added between the hidden and the input layer, the network can learn non-linear representations (DeMers & Cottrell, 1993), and thus possibly reducing the dimensionality even more. This might introduce strange consequences in similarity judgment though. For example, envision a string whose endpoints connect forming a circle. When reducing the dimensionality of this circle to one dimension, the endpoints will be disconnected and the string straightened. The endpoints, touching each other in the original space, will now be perceived as being on different sides of this one-dimensional space.

4.2 Conclusions

I conclude this section with noting that reducing the dimensionality of a high dimensional input stimuli is important as it influences cognition's ability to generalize. Auto-encoders provide a means to this reduction as well as preserving similarities among the stimuli. This similarity will however be distorted such that familiar objects are perceived as being more distinct than others. This reduction is done in an unsupervised way due to the network learning the identity-mapping. In a later section we will revisit properties of auto-encoders and discuss how they can explain some of the effects seen in categorical perception when extended with a supervised component. In fact, this extension makes it more similar to the network in figure 1., and

it is by virtue of this that it will be able to capture some effects seen in categorical perception.

5 Cognitive control

We have so far seen how conceptual representations can emerge in artificial neural networks. We have looked at different aspect of these representations such as the internal hierarchy arising from coherent covariation (Rogers & McClelland, 2006), the redundancy compression arising in auto encoders (Gluck & Meyers, 1993; Japkowicz, 1995). We have also looked at Barsalou's perceptual theory of cognition, perceptual symbol systems (Barsalou, 1999; Barsalou, 2003a; Barsalou, et. al., 2003b). However, the notion of control has yet to be mentioned. The human mind has the ability to represent a vast amount different goals, contingencies and abstract categories. For example, when told to jump each time someone claps their hands, a person is able execute this behavior even in the presence of distractions. This kind of performance is not captured very well in the sort of gradual learning we have thus far seen. A feed forward neural network would not be suitable as for each change of activity, the old state is erased. An Elman network (Elman, 1990) could represent the changing nature of behavior over time, but it does not seem likely that it could effortlessly represent this sort of abstract tasks "out of the box". In Barsalou's perceptual symbols system, a simulation could be run as a result of the command. When someone then claps their hands, the resultant auditory and visual states would bind to the simulation causing the proposition "someone clapped their hands" to be validated. A "jumping" simulation could then be run contingent on the validation of this proposition. Where are these decisions made? Since the goals and contingencies would need to be resistant to noise it seems unlikely they would reside in the different modalities in the brain as perceptual states—They would easily get erased as new perceptual states arise due to bottom up interference from the world. Also, neurons in these areas of the brain does not maintain their activity over long periods of time, a property important to representing goals. We will next turn our attention to the PFC for answers to this question.

5.1 Cognitive control through dopamine and prefrontal cortex interaction

To find some answers we will here take a closer look on a theory of Cognitive Control proposed by Miller & Cohen, 2001; Braver & Cohen, 2000). The prefrontal cortex (PFC) is a collection of neocortical areas receiving projection from virtually all cortical sensory systems, motor systems, and many subcortical structures (Miller & Cohen, 2001). This makes it a perfect place for behavior guiding rules to be seated (i.e., from here they can influence an abundant array of processing aspects). People with PFC damage seem to act on a whim, and are impaired in keeping with internal goals. They seem to have problems with adhering to newly learned rules, while old rules

and behaviors seem intact. These old rules, however, may be executed on a whim. The Stroop task is a task at which people with PFC damage show deficits. In this task a subject is shown a word for a color (e.g., Green) depicted in some color (e.g., green, red, etc). If the color the word names differ from its depicted color, the subject is faced with conflicting stimulus in which naming the written color is behavioral salient. This suggest they have difficulty adhering to the goal or rule of the task in the face of a stronger (more competitive) stimulus (Miller & Cohen 2001). *The task itself illustrates one of the fundamental aspects of cognitive control and goal directed behavior* (Miller & Cohen, 2001):

> "the ability to select a weaker, task-relevant response (or source of information) in the face of competition from an otherwise stronger, but task-irrelevant one."

How does the PFC select task-relevant responses in the face of stronger task-irrelevant ones? By building one the fundamental principle that processing in the brain is competitive, Miller & Cohen (2001) proposes the following function of the PFC in the service of cognitive control:

> "the active maintenance of patterns of activity that represent goals and the means to achieve them. They provide the bias signals throughout much of the rest if the brain, affecting not only visual processes but also other sensory modalities, as well as systems responsible for response execution, memory retrieval, emotional evaluation, etc. The aggregate effect of these bias signals is to guide the flow of neural activity along pathways that establish the proper mappings between inputs, internal states, and outputs needed to perform a given task."

Patterns of activity in the PFC can thus serve the function of selective attention by influencing the features of a scene to be attended through excitatory connections (Miller & Cohen, 2001). The neurons being influenced are then able to win over their competitors by mutual inhibition. It can serve as a place for representing goals and the context of a situation, and thus also serving the function of working memory. The PFC thus seems critical in tasks such as learning the association between a specific situation and a response. However, as this task is frequently executed, it gets "pushed down" and the behavior gets more automated.

Miller & Freedman et. al. (2002) performed two experiments to investigate the role the PFC in representing rules and categories. In the first, monkeys performed a delayed-match-to-category (DMC) task. To create the test images three species of cats and three breeds of dogs were morphed resulting in a large set of parametric blends of the prototype images. A specific morph belonged to the category it was the most of (i.e., it was a

cat if it was over 50% cat). This allowed the experiment to capture the sharp category boundaries often present in the real world. After training the monkeys performed at about 90% correct. They recorded activity in the lateral PFC and found many examples of neurons seeming to encode category membership.

In the second they trained monkeys to respond differently depending on whether a match or non-match rule was in effect. If the match rule was in effect and two subsequent pictures matched, the monkeys released a lever, if they did not match, the lever was released. The reverse was true for the non-match rule. Cues signifying the same rule were taken from different modalities while cues signifying different rules were taken from the same modalities. This allowed responses to the physical properties of the cue to be disambiguated from responses to the rule the cue signified. The most prevalent activity across the PFC was reported to encode the current rule. Rapid plasticity of the PFC neurons was also reported. That is, the neurons showed tuning to the specific task after very few learning trials.

The above evidence supports the view that the PFC is engaged in representing rules and categories. It also shows that the neurons in the PFC can exhibit plasticity, important in fast learning. The PFC is also known to support active maintenance in the face of interference (Miller & Cohen, 2001). It thus exhibits the main properties important for cognitive control. There is still one issue to be addressed: how does the system learn what to represent and when to update that representation? Braver & Cohen (2000) propose that *dopamine (DA) projections to the PFC serves to gate access of context representations into active memory through simple neuromodulatory effects on processing units in the PFC*. Midbrain DA neurons respond to an unpredicted reward, as learning proceeds DA responses migrate until they coincide with the stimulus predicting reward, and cease their activation to the now predicted reward. DA activity is also inhibited if the predicted reward fails to appear, if it appears earlier than expected DA activity is increased. (Braver & Cohen, 2000; Miller & Cohen, 2001) . DA activity thus seems to encode prediction error (Miller & Cohen, 2001). DA neurons can thus gate information that predicts reward into the PFC while also serving as a learning signal for strengthening the association between this information and the response of the DA neurons. This system thus learn what to gate into the PFC and when to update the representations inside the PFC. Braver & Cohen (2000) also presents an implementation of such a system successfully learning what to gate, to keep it memory in the face of distracting stimulus, and to give the right response at the right time, proving the validity of the approach. The weights intrinsic to the PFC and the weights from the stimulus to the PFC are however kept constant, thus leaving out the mechanism by which active maintenance arises.

5.2 Conclusions

The PFC thus is a perfect region of the brain in which to represent rules, goals and categories due to its interconnectivity and inherent properties. The representations in the PFC can be learned through interaction between the PFC and DA systems. These rules affects processing in other parts of the brain by exerting a biasing influence. As these rules are frequently executed they get "pushed down" and become more automated. The PFC thus seems to be important in both representing and achieving conceptual representations, as well as discovering the more abstract relationships among them. A closer look on this possible aspect of the PFC will be taken at a later point in the thesis.

6 Categorical perception

Categorical perception is the phenomenon that people tend to perform better in discrimination tasks when the stimuli comes from different categories. This phenomenon suggest that the way we categorize the world, in turn influences the way we perceive the world. In this section I will discuss ways in which cognition might influence perceptual learning through categorization. The reason for saying 'might' is that the observed effects of category learning on perception may be a result of changes in some higher level cognitive system (e.g., a higher sensitivity for learning categories in the PFC) and thus not a result of the perceptual system adjusting to facilitate category learning. This introduction to categorical influences on perceptual learning is synthesized from Goldstone (1994). Next, a short review of results of an experiment investigating this phenomenon is presented. This experiment is originally presented in Goldstone (1994). Then a model of hippocampal function proposed by Gluck & Meyer (1993) to explain stimulus conditioning in animals is presented and the ways in which this model might explain some of the results previously seen is discussed.

The first issue of perceptual learning we will discuss is that of preexposure. Preexposure is coined to describe the phenomenon that people are better able to perform discriminations on stimuli when they have previously been exposed to the stimuli. That is, by being exposed to a set of pictures before engaging a discrimination task, the subjects perform better at this task even though they received no feedback during preexposure as to how the stimuli should be discriminated. It would seem that the perceptual system becomes primed to the different stimuli.

The next issue relates to the process enabling people to become better at discriminating the stimuli. This is the issue of whether acquired equivalence of acquired distinctiveness is the process responsible. Acquired equivalence is the process by which sensitivity to the irrelevant dimensions of the stimuli is desensitized. For example, before category learning is engaged, the subjects discriminate all stimuli equally well, but after category learning, sensitivity to stimuli not straddling categories is lost. One account of evidence for this effect is that infants (2 months old) show sensitivity to differences between speech sounds they loose by the age of 10 months (Goldstone, 1994). Acquired distinctiveness is the opposite of this effect and thus denotes heightened sensitivity to the stimuli straddling category boundaries. This could happen by the perceptual system being sensitized to changes along a category relevant dimension at the category boundary. Acquired similarity could happen by the perceptual system being desensitized to changes along an irrelevant dimension and also by desensitization to changes inside a category along a relevant dimension. This brings us to the next issue: whether an entire dimension must be sensitized or if local regions can receive sensitizing.

If an entire dimension must be sensitized to an equal degree, than people receiving category training should afterward be equally able to discriminate stimuli within and between a category when than stimuli varies along the relevant dimension. However, even if an entire dimension must be sensitized, it is possible that the some regions becomes more sensitive to changes than others. For example, it could be that sensitivity is heightened more at the category boundary than within a category.

An issue remaining is that of integral dimensions. With integral dimensions, attending to one dimension without attending to the other dimension is rather difficult. With separable dimensions, this is easy. Integral dimensions, such as the brightness and saturation of color, has been argued to be psychologically fused (Goldstone, 1994). One would thus expect that changes in sensitivity to one integral dimension would also apply to the other. However, it could also be that pressure from categorization "defuses" these dimensions and only the relevant one become (de)sensitized. Also, since integral dimensions are easy to focus on at the same time, it could be that separate dimensions compete more with each other for attention then do integral ones.

To investigate the nature of categorization's influence on perceptual discrimination, specifically with respect to the issues mentioned above, Goldstone (1994) undertook a series of experiments. The specifics of the experiment is not important here so I will outline only the general points: Subjects were placed into groups, where each group categorized the stimuli according to different measures. There was one group of size-categorizers, one group of brightness-categorizers, one group that categorized based on both dimensions (thus having four categories). The categorizers first went though a training stage were they were presented with a stimuli and had to guess what category it belonged to and then received feedback as to whether the guess was correct or not. After training they went through a discrimination trial were they had to judge whether two subsequently presented stimuli were the same or different. There was also a control group who did not receive any category training before the discrimination trial. Size and brightness is separable dimensions, so an experiment differing only in the categorization relevant dimensions was also undertaken. The dimensions for this experiment was the integral dimensions brightness and saturation. By comparing how well the subjects fared at categorization within and between experiments, the nature of how category learning influenced perceptual discrimination was determined.

The results from the experiment with separable dimensions showed that acquired distinctiveness asserted its influence on the categorization relevant dimension. There was also local sensitization within the relevant dimension with sensitivity being highest at the category boundary. The effect on the irrelevant dimensions differed between the size-categorizers and brightness-categorizers, with the results showing a null effect and acquired equivalence

respectively. For the categorizers relying on both dimensions, acquired distinctiveness was found for both dimensions with local sensitization within the dimensions (sensitivity being the highest at the category boundary). However, acquired distinctiveness was lower when both dimensions were relevant than when only one was relevant. There was thus competition between the separable dimensions when they both were relevant.

The results from the experiment with integral dimensions were similar to the ones just presented. However, when only one dimension was relevant, acquired distinctiveness was also found for the other. Also, when both dimensions were relevant, no competition between the dimensions were found. These results confirm the idea that focusing on several dimensions is easy when they are integrated, and also that when attention is placed on one such dimension, attention on the other follows.

6.1 Hippocampal Mediation of Stimulus Representation

In Gluck & Meyer (1993) a computational model of hippocampal function in mediating stimulus representations are represented. The models learn what cues presented within a context predicts a specific outcome. The model incorporates two representational constraints assumed to reside in the hippocampus. These constraints are *redundancy compression* and *predictive differentiation*. The issue of redundancy compression has been discussed earlier, but for clarity I will repeat it here: redundancy compressions is the process by which correlated units in the input is compressed into a more compact representation (possibly a single unit). As earlier in this thesis, redundancy compressions is achieved here by employing an auto-encoder. Predictive differentiation is the process by which cues that predict a specific outcome causes the internal representations to separate themselves from other representations not containing the cue. This is achieved by augmenting the auto-encoder with an extra output unit denoting the predicted outcome. The role attributed to the hippocampus is to derive representation used for learning associations in other brain areas.

The process of predictive differentiation will result in acquired distinctiveness when the Euclidean distance between the stimulus representations in internal space is the measure of similarity. This measure of similarity is assumed in the rest of this section. When training this model on stimulus predicting different outcomes, the dimensions being predictive will acquire distinctiveness because of the internal representations being separated to accommodate similarity based generalization—They separate to avoid interfering with each other. Acquired distinctiveness would be highest at the category boundary because of the error signal changing drastically at this point forcing the two categories apart in representational space. As we have already seen, expertise with a category in PDP networks causes the representations belonging to this category to drift apart in representational space.

As such local sensitization within a category along the relevant dimension would also occur. It would also be natural to assume that the pressure to differentiate between category representations would be larger than the differentiation of within category differences, and therefor that local sensitization of within category items will be smaller than between categories.

The process of redundancy compressions is the same as acquired equivalence. When training this model on stimulus predicting different outcomes, the amount of acquired equivalence would depend on the nature of the stimuli. For example, consider the model previously being taught to categorize based on brightness. If it now was required to categorize based on size, brightness would not be a good criterion for separating the internal space and acquire equivalence.

Competition between dimensions when categorizing along two dimensions would be present until the internal space has separated enough to avoid interference (this assumes they were close enough to interfere with each other when learning started). During the early stages of training, the error signal from the predictive units will be more likely to interfere with each, both within and between instances. However, as the representations gain distinctiveness along the relevant dimensions, this effect will decrease.

With respect to integral dimensions it is not easy to see how this model can facilitate the result reported by Goldstone (1994). Within the context of stimulus conditioning using localist encoding the model supports this though. Gluck & Meyer (1993) show that if the model is first trained with stimulus A and B not predicting any outcome at all, these representations will become compressed with the background context. This compression of the two cues could be a way two stimuli become fused. If later, the model is trained with A predicting an out come (B is not presented during this training) then some of the predictions inferred from A will transfer to B. However, in the current context both dimensions are present during training and the training method would thus discover that B is not predictive of the outcome.

6.2 The effect of language on perceptual discrimination in artificial neural networks

In this section some results from (Cangelosi & Parisi, 2001) will be presented. How conceptual representations can influence language have previously been considered. We have seen how typicality and frequency of presentation can affect naming responses of PDP networks by the effect they have on the conceptual organization. The effects of language on the internal representations of a network will now be discussed . Specifically, how language has a beneficial effect on non-linguistic behavior. In Cangelosi & Parisi (2001) a set of agents were evolved using a genetic algorithm. Using a two segment arm, the agents were required to manipulate two different objects. If object A was

present, the default action was to push it away, for object B, the default was to pull it towards itself. The agent was controlled by a neural network receiving proprioceptive, retinal, and language input. At the output the network was required to produce motor/muscle signals to control the arm. Without going into detail on connectivity and training regime (the description of the different conditions for the simulations are simplified here), I will mention that the networks was evolved in three different conditions: no_language, late_language and early_language. In the no_language condition, the agent was required to perform the default action associated with the objects. In the late_language condition the agent received language input after being trained for a 1000 epochs in the no_language condition. The agent could receive as linguistic input either noun, default verb, opposite verb, noun and default verb, or noun and opposite verb. There was also an epoch without language input in the language conditions. The early_language condition was similar to the late_language condition except for that language was introduced at epoch zero.

Results showed that the late_language condition was the most successful, followed by the no_language condition. With respect to these results, Cangelosi & Parisi proposed that for language to have a beneficial effect, the cognitive behavior on which it will be grounded must be evolved before it is introduced. Interestingly, the beneficial effects of introducing language was also present when the network did not receive any linguistic input. The introduction of language had thus allowed the network to structure its internal world more efficiently. To investigate further the effects of language, the average within and between category distance for the objects were computed for the no_language, noun_only, verb_only, and noun+verb trials in the late_language condition. Between category distances were bigger for language trials than the no_language one. For the within category distance, the distance was smallest for the language conditions. Verb_only and verb+noun showed the strongest effects. Thus, with the introduction of language, the categories acquired similarity within, while they acquired distinctiveness between. Moreover, the verbs were the greatest contributors to this effect. Cangelosi & Parisi attributed this to the fact that nouns only covary with the objects while verbs covary with the action to be performed, thus further facilitating efficient internal representations.

6.3 Conclusion

We have reviewed results from Goldstone (1994) showing how category learning influences perception. For separable dimensions, the categories acquire distinctiveness along the relevant dimensions. Along the irrelevant dimensions, the results were inconclusive showing both a null effect and acquired similarity. When both dimensions were relevant for categorization, the dimensions competed with each other for distinctiveness. For integral di-

mensions the relevant dimensions and irrelevant dimensions acquired distinctiveness. When both dimensions were relevant for categorization, no competition was present. Along the relevant dimensions, there were also local sensitization with acquired distinctiveness also within a category. The within category distinctiveness was smaller than between category. PDP offers a way to understand some of these phenomena. The results for integral dimensions are however harder to explain. Cangelosi & Parisi (2001) showed how language can influence perceptual tasks. The introduction of language caused the internal representations to be further sculpted in the image of precision. The internal representations thus received within category similarity and between category distinction, suggesting that language is an important factor in conceptual representations. The beneficial effects of language was also present when the network did not receive any linguistic input.

Part II
Framework and case studies

7 Non-Symbolic Algorithms

This far we have reviewed how to achieve conceptual representations and what properties they exhibit. In Rogers & McClelland (2006) conceptual representations were achieved by employing distributed representations at the hidden layer of a neural network. These representations elicited behavior similar to human behavior in semantic tasks. In Barsalou (1999; 2003a; 2003b) a concept is represented as a simulator. An object in the world belongs to a simulator that can produce a simulation of it matching the perceived entity to some degree. These simulators are far more flexible than the hidden layer representations we have seen and support a wide range of complex behavior. In fact, simulators can develop for all perceptual states arising and integrate these states both spatially and temporally. As such, simulators can develop for any perceivable aspect of the world and any introspective aspect of the agent.

In this section a framework for the use of conceptual representations in high level cognitive function will be presented. That is, how PDP might be engaged in high level cognitive behavior such as algorithms, planning, and language production. This will be done within the framework of non-symbolic algorithms and non-symbolic concepts (Veflingstad & Yildirim, 2007). Within this framework it is assumed that algorithms exist in the human brain, that they are represented non-symbolically and that they operate on non-symbols. High level cognitive function can thus be represented without the use of any symbols by employing non-symbolic algorithms and concepts. It is further proposed that a thought simply requires the activation of concepts. To frame our discussion I will present three levels of cognition proposed in (Veflingstad & Yildirim, 2007):

1. Stimulus-Response level: This is the level where there is a direct functional mapping from the sensations of a situation to behavioral outcomes. For example, a robot might be wired up to avoid obstacles conforming to a particular pattern of activations across proximity sensors.
2. Conceptual Level: This is the level where there is formation of concepts in parallel to a functional mapping from the sensations of a situation to behavioral outcomes. For example, a robot might be wired up to avoid obstacles conforming to a particular pattern of activations across proximity sensors and in the meantime it forms the concept of obstacle. The obtained concepts are employed in high level cognitive tasks e.g. thinking, planning, decision-making.
3. The Language Level: This is the symbolic level and there is a functional mapping from the conceptual level to the symbolic language level. For example, a robot maps a concept

of obstacle to the word "obstacle" (p. 29)

Even though these levels are believed to be the major ones, their actual separation is vague. For example, it might be possible that concepts are formed at level 1 while level 2 is left with the function of using these concepts. The material we have reviewed this far also suggest that level 3 and 2 have a top down effect on the functioning of the lower levels. For example, the acquisition of concepts influence how the world is perceived (Goldstone, 1994), while the acquisition of language makes the conceptual representations more precise (Cangelosi & Parisi, 2001). I would like to stress that even though level 3 is termed symbolic its processing need not be symbolic as similarity based generalization can possibly rid the need for symbols. That is, whether or not the brain function symbolically is still a debate in the field of AI. Here it is proposed reasoning can be done without the use of any symbols. This, however, does not exclude the possibility of symbolic processing also being present in the brain.

The focus of this section will be on how an agent might utilize the concepts it forms during navigation in high level cognition. Since level 2 is attributed with this function it this level the discussion will be centered around and it is here the proposed non-symbolic algorithms reside. If the algorithm is not represented symbolically and it does not manipulate symbols, how is it represented and what does it manipulate? In the section on semantic cognition we saw how PDP can represent semantics. In fact, when the network in Figure 1 produces an output pattern as a result of a presented item and relation pair, it is actually completing a proposition. Completing a proposition can be seen as executing an algorithm. This would suggest that an algorithm can be represented within PDP. We have further seen that the semantic behavior arises from similarity based generalization suggesting that our algorithms should operate on distributed representations. It is thus proposed that non-symbolic algorithms operate on non-symbolic concepts, achieved by employing distributed representations such as the hidden layer representations we have seen earlier. I would like to stress that the distributed representation is not a concept but a specific conceptualization. The concept arises from the system through similarity based generalization. However, for simplicity distributed representations will be referred to as concepts here. This is justified because each step of the algorithm is assumed to operate on sets of patterns, treating them similarly. When I refer to, for example, the concept of *cold* it is this set I am referring to. A step in an algorithm often involve decision making, and, as such a step in an non-symbolic algorithm can be an if-then rule. An if then rule is essentially a classification, and is therefor represented with a feed forward neural network here. Also, an action is also a concept and might therefor also be a step in the algorithm. I will now describe how algorithms may be acquired using a bathing algorithm from Veflingstad & Yildirim (2007) as an example.

7.1 Representation and acquisition of algorithms

We will now consider the case of a child learning an algorithm through its experience with the world. Specifically, the child experiences cold lakewater and as a result of this does an evasive maneuver, stimulus-response in nature. The algorithm can be represented in english as:

1. sense the lakewater
2. if the lakewater is cold then get out of the lakewater

The first step of the algorithm consists of the concepts *sense* and *lakewater*. The second, involves an if then rule having the concepts *lakewater* and *cold* as its inputs and the concept *get_out_of* as a result. *sense* and *get_out_of* are action concepts. As the child senses the lakewater the concept *sense* is active. It is assumed that the child has previous experience with these things and thus has these concepts in its brain. During execution of step 1 *sense* and *lakewater* is active. These to concepts being simultaneous active corresponds to the thought "sensing lakewater". The execution of this step causes the actual sensing of the water at level 1. As a result, *lakewater* and *cold* becomes active. *sense* is now out of context and is assumed to have lost some or all of its activity. The concepts *lakewater* and *cold* being simultaneously active corresponds to the thought "the lakewater is cold". As a result of the lakewater being cold, the child feels physical discomfort and gets out of the water. This behavior happens at level 1 and as a result of this, *get_out_of* becomes active at level 2. When this happens for the first time, the child has no knowledge of this causality and can not predict this outcome. However, as a result of these concepts being simultaneously active, a causal link starts to form between *lakewater*, *cold* and *get_out_of*. This link is the if-then rule and captures the causality among the concepts. Once this link has formed, step 2 will be executed when the concepts in its precondition are active, allowing the child to plan ahead. This link can for example help forming thoughts such as "if the lakewater is cold, I should get out of it". Also, the resulting action *get_out_of* needs and argument so that the child knows what to get out of. A conceptual relevance link must must therefor be present. This is not a physically present link, but represents a construal of the situation. This could for example be that the child observes the sequence of which the concepts becomes activated and thus understands that it is the "cold lakewater"' that elicits the "get out of" action.

This demonstrates how algorithms can be acquired by experiencing the world. These algorithms algorithms can then be applied to produce more complex thought allowing the agent to reason about the world. However, what is the process allowing this learning to take place? Does cognition simply associate blindly everything that cooccurs? This would make learning

very hard, and make learning temporal associations impossible. I will propose in accordance with previously reviewed material that cognitive control (Braver & Cohen, 2000; Miller & Cohen, 2001; Miller & Freedman et. al., 2002) guides the process. Patterns of sustained activity in the PFC can assert selective attention honing sensation into informative grains. The PFC could thus be responsible for focusing on the relevant concepts during learning (as well as the relevant aspects of the concepts). Consider the algorithm above: for learning to be successful, irrelevant aspects of the environment must be excluded and the causality must be in the right direction. In accordance with the PFC function reviewed earlier it is here assumed that the relevant aspects are filtered out by excitation of the relevant concepts (and possibly inhibition of the irrelevant ones) and gating of the relevant information into the PFC. The causality of the situation is construed by DA neurons learning/knowing that the cold lakewater is a good predictor of the get_out_of action. In this example, it might be that prediction learning is unnecessary since the cue and result are simultaneously active, and that this learning is done by some other process construing the meaningful causal relationship. It might thus be the PFC which construes the conceptual relevance while also being the place where the algorithm is first represented. The algorithm will later be pushed down to more autonomous areas of the brain if the algorithm is executed repeatedly. This could happen by the child experiencing the same situation again or by the child exercising its new thought skill (executing the algorithm in the PFC).

To investigate the feasibility of this approach a non-symbolic summation algorithm was implemented. The aspect of cognitive control, being a complex issue, was left out. The simulation thus investigates the feasibility of learning long addition employing conceptual representations. This simulation has previously been presented in (Veflingstad & Yildirim, 2007). Because of its relevance it is repeated here.

7.2 Non-symbolic Summation

We have seen that the concepts that represent sensations can be associated with concepts that represent actions. We need to be able to simulate this association while representing the basic elements of algorithms using artificial neural networks. Towards this end a Non-Symbolic summation algorithm was implemented simulating an agent doing long addition. The simulation employed two neural networks. One, an auto-encoder learning the identity-mapping on 256 pixel grayscale images of the digits from zero to nine deriving a conceptual representation of the numbers seen (Figure 2). These images are not intended to represent the retinal activation patterns produced when perceiving a digit. The simulation assumes a nonverbal numerical representational system extending across different modalities. Evidence supporting this contention is presented in (Jordan & Brannon, 2006). Since the auto-

encoder is too simple a system to perform this task, the images are intended to represent all instances having a numerical interpretation known to the agent. The other network, henceforth called the summation network, was a four-layered feed-forward network taught to map the conceptual representations of two numbers and a carry to a conceptual representation of a number and a carry out (answer), thus carrying out the actual summation (Figure 3). To add two numbers the following steps was performed: Starting with the lowest order digits, one digit from each number was presented to the auto-encoder in turn. After each presentation, the conceptual representation derived on the conceptualization layer was copied to its corresponding position on the input neurons of the summation network. If these were the first digits to be added, carry in was set to the conceptualization of zero. If not, the conceptualization present at carry out was copied to carry in. Then activation was fed forward and the activation on the output neurons corresponding to the answer was copied to the conceptualization layer of the auto-encoder and decoder. This was the first digit in the answer. This process was carried out until there were no more digits to add and carry out was zero. The corresponding Symbolic expression of the summation algorithm can be given as follows:

1. Sense two digits from two numbers visually.
2. If there are no more digits to add and carry out is zero, display summation in a visual form.
3. Else, sum the two digits and the carry into a number and a carry.
4. Go to Step 1.

The auto-encoder was a five layered neural network, the middle layer being the one representing the concepts. The input and output layers had 256 neurons each. The two layers on each side of the middle layer had 60 neurons each, and the middle layer had 15 neurons. The network was trained using back-propagation for 10000 epochs with a learning rate of 0.01. At each epoch, the digits from zero to ten were presented. Weights were updated after each presentation of a digit. After presentation of a digit and before feeding activation forward, Gaussian noise with a mean of zero and a standard deviation of 0.2 was added to the input neurons. The summation network had 45 input neurons (3 * 15), 37 neurons in each of the two hidden layers, and 30 output neurons (15 * 2). It was trained using back-propagation for 10000 epochs with a learning rate of 0.01. At each epoch, every combination of two digits and a carry out of 200 possibilities were taught. Each digit having 10 possible values ranging from 0 to 9 leads to 100 possibilities for the digit pairs to be added. Having a carry value 0 or 1 increases the number of possibilities for digit pairs to 200. No Gaussian noise was added. Weights were updated after each combination. In both

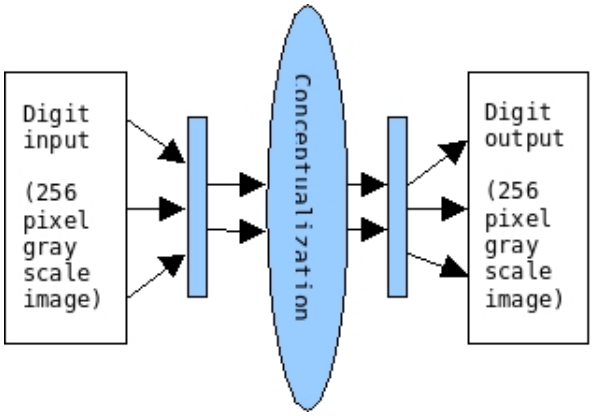

Figure 2: Auto-encoder

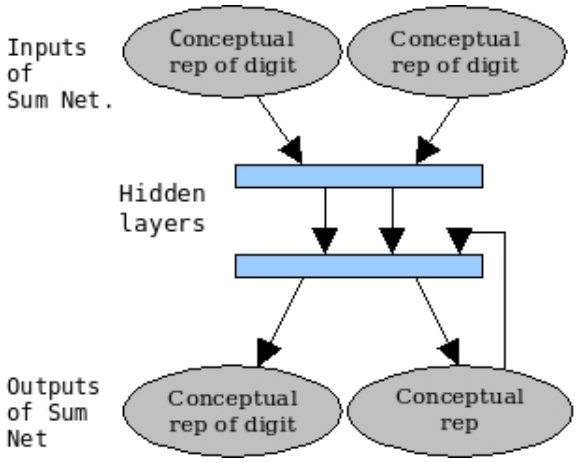

Figure 3: Summation network

networks, every hidden and output neuron also had a constant bias of one with a modifiable weight.

We ran the simulation ten times and averaged the results. After 10000 epochs the sum of squared errors was 1.83 and 3.16 in the auto-encoder and the summation network respectively. To test the performance of the summation network, training was turned off each 500^{th} epoch and 1000 summations trials were performed. At each summation trial two random numbers in the range 0-999 were selected and summed according to the procedure described above. The performance was recorded in the following manner: Failure to produce the correct carry or answer was an error. Each answer and carry produced was decoded in the auto-encoder. Then, for each image in the dataset, the sum of absolute differences with respect to the decoded image was computed. The image in the data set with the smallest sum of absolute differences was interpreted as the summation networks intended

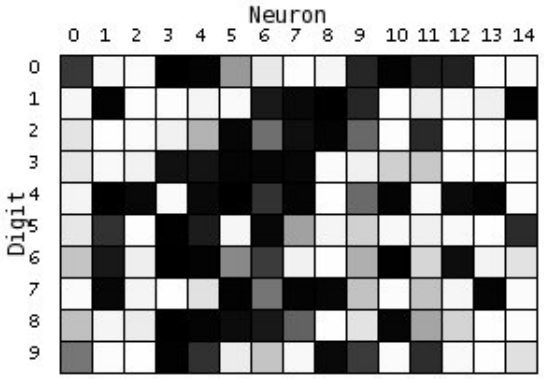

Figure 4: The activation pattern for the conceptualizations

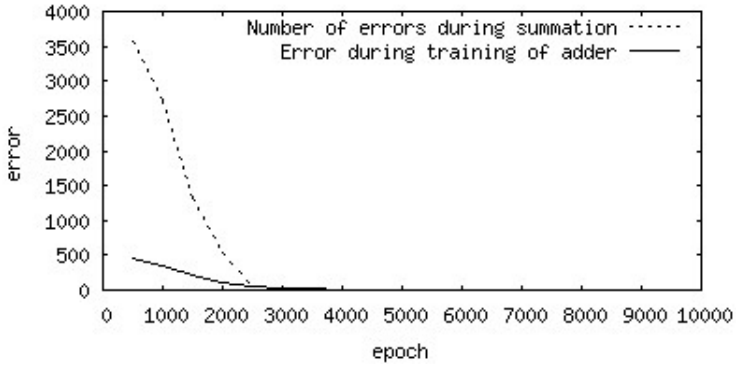

Figure 5: Errors during summation and training.

answer/carry. If this was incorrect an error was recorded. A reconstruction error was also computed as the average of the absolute differences computed above. After 4500 epochs the number of errors stayed below 3 and at epoch 10000 it dropped to zero. The reconstruction error after 10000 epochs was 3.11. Figure 4 shows the activation pattern for the conceptualizations.

This shows that the network had almost perfectly learned non-symbolic summation, the only "error" being an inability to produce an answer identical in appearance to the digits learned (an absolute difference of 3.11 is very close though). Figure 5 shows the change in squared sum and summation error over 10000 epochs, maximum error obtained being around 3500 when averaged for 10 runs.

7.3 Conclusion

In this section it was proposed that algorithms can exist in the human brain and that they are represented non-symbolically. It was shown how they can be acquired through observing the causality in the domain and by construing meaningful relationships among concepts. These algorithms can be represented by employing distributed representations in neural networks and they employ concepts in their processing. When an algorithm has been acquired it may help to produce complex thought and may thus be used in planning. It is thus proposed that algorithms are essential in reasoning. The PFC is a region of the brain suitable for acquiring these algorithms. As these algorithms are used frequently, they will be "pushed" down to more autonomous areas of the brain, thus releaving the need of attention.

8 Novelty detection in and between modalities

The ability for an autonomous intelligent agent to detect whether some perceived instance is novel is important to intelligence. Within every novel instance lies the promise of learning something new which may be of importance to the agent—the purpose for which it was created. Also, if the agent is endowed with a multi-modal conceptual system, then novelty can be measured across different modalities allowing the agent to know in what way it previously has experienced the perceived instance and thus what to expect from the world. Knowing in what ways it has not experienced the perceived instance allows it know what kind of information it can gain from investigating the instance. Novelty can also serve as an attentional factor causing an agent to explore instances perceived as novel. For example, after children have been habituated with a set of artifacts and then given a choice between a novel artifact and an artifact similar to the ones in the habituation set, they tend to choose the novel one (Rogers & McClelland, 2006). In Rogers & McClelland (2006) this was modeled by computing the Euclidean distance between the conceptual representation of the novel instance and the centroid of the conceptual representations of a set habituation instances. This measured the novelty of an item with respect to the current context. In NASA's OASIS project (Castañ̃o, R et. al) novelty detection is used to identify science opportunities. In-situ rover's are given the ability to classify whether a perceived rock is novel and use this information in prioritizing what rocks to sample. The OASIS project used a hybrid approach utilizing distance-based k-means clustering, probability-based and discriminative methods. However, one could also measure novelty by how well an agent is able to conceptualize the world. This approach would not require the maintenance of a context or any reference to memory. If such a novelty is available for conception it can explore the world with a preference for the unknown. Novelty is thus a natural way of endowing an agent with curiosity, the result being an agent "intent" on discovering the world. Here we will investigate such a novelty measure. The only information used to determine this measure is the information present in the network weights.

8.1 Novelty as reconstruction error

The approach taken here utilizes auto-encoders (Hinton, 1989) to derive compressed hidden layer representations of presented stimulus. The hidden layer representations gets compressed because a hidden layer narrower than the input/output layer is used, reducing the dimensionality of the presented stimulus. This has two important advantages:

- The dimensionality gets reduced by taking advantage of redundant information; That is, correlated features in the input stimulus. In

other words, redundancies gets compressed allowing further processing to easier focus on the more relevant properties of the stimulus.

- A low dimensional representation (LDR) avoids what is known as the curse of dimensionality: The number of examples necessary for reliable generalization grows exponentially with the number of dimensions (Edelman,, S. & Intrator, N., 1997) .

During training an auto-encoder it is taught to reproduce on its output units the same patterns as presented on its input units. Thus, after sufficient degree of training, a pattern presented to the network which reproduces a pattern on the output units with low error, will most likely be a pattern similar to one taught, while one with high error will be an unfamiliar one. A threshold function of the error signal can thus be used to judge novelty. In Elman (1988) this technique is used to recognize learned syllables in continuous speech. In Japkowicz (1995) it is applied to several other problems with success.(NOTE:Should mention some if the problems) In this article the technique is extended to include a link between two modalities (vision and audition) allowing the agent to both decide whether it has seen the presented instance before and if it has heard a word associated with the instance before. A simulation is implemented to investigate the following points:

- Will a link between modalities be able to function reliably throughout training? A fully conceptual system would have to learn, and thus update its internal representations, continuously. Moreover, small errors in conceptual representations would be expected to have a higher impact than errors in input representations—The redundancies are compressed. As a result, a link between modalities (working on conceptual representations) might be unable to function to an acceptable degree of precision.

- How resistant will this conceptual system be to noise?

- If our agent sees a familiar instance it does not have a word associated with, will the link infer a novel conceptual representation in the auditory modality?

- Is it possible to find a general procedure to derive the threshold function for deciding novelty.

The three first points above relates to the precision of the link between the modalities, the last one to a method for implementing an agent able to detect novelty.

8.2 The simulation

The purpose of this simulation is to investigate how well a link between two modalities fares in novelty detection. I wanted the modalities to be perceptually grounded while not having to deal with the complexities of implementing an auditory system able to deal with differing word lengths, stress and pronunciation. As a result, input to the to modalities was represented as pictures. Input to the vision modality was pictures of numbers while pictures of words (the words for the different numbers) was the input to the audition modality. As a result, the two modalities differ only in their number of input units (size of the pictures) and the appearance of their input pictures.

Two auto-encoding networks are employed in implementing the two modalities, vision and auditory (figure 6). Their function is to derive hidden layer conceptual representation of presented pictures, and to be noise resistant and able to generalize. The link between the modalities is implemented using a bidirectional feed forward network (figure 6). Its job is to learn a mapping in both directions between the conceptual representations. The link is made bidirectional because the network having its conceptual representation inferred has no way of deciding its novelty—The novelty measure is the difference between the input and output of the network. A bidirectional link allows the activity of the inferred concept to flow back to the original network; the difference between the original and resulting activity is the novelty measure. It is thus the link which is responsible for deriving the inter-modal novelty measure. Both the auto-encoders had one hidden layer consisting of 40 units, their number of input/output units depended on the size of the pictures they were trained on. The link had three layers in both directions, the hidden in both directions having 40 hidden units. The input to the vision modality was 256 pixel gray scale images of numbers. These numbers were separated into three sets: A training set used for training the network and testing performance, a set called the "similar set" used for testing performance, and a set called the "novel set" also used for testing performance. These sets are listed under "Numbers" in table 1. The input to the audition modality was 250 pixel gray scale images of words. As opposed to the vision modality, the audition modality had only one set of pictures, a training set consisting of the words for the numbers from zero to ten. This set is listed as "training set" under "Words" in table 1.

As is apparent from table 1 the size of the "Numbers" and "Words" training set is different. This was chosen to allow the simulation to know some numbers only by sight. When perceiving these numbers during testing, whether or not the link infers a novel representation in the audition modality can be investigated. The "Similar" and "Novel" sets were present for performance testing. The "Similar" set consisted of the numbers from zero to five in italic font thus only being slightly different than the learned

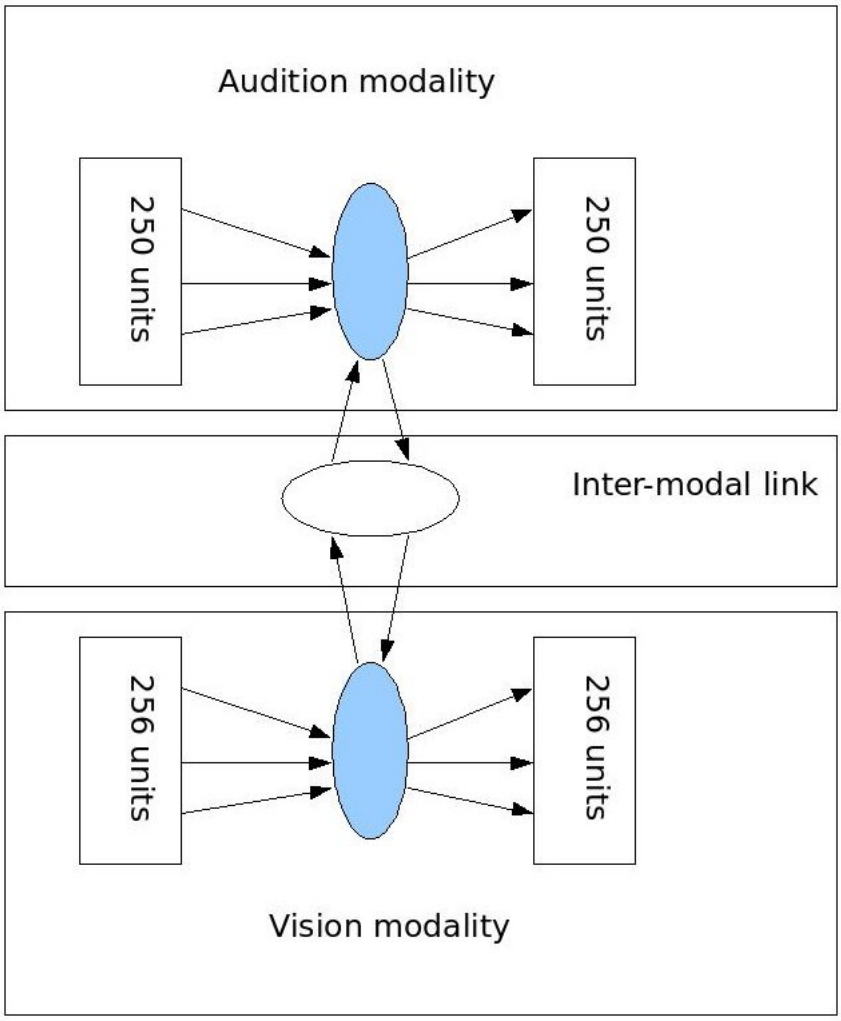

Figure 6: This figure shows the architecture of the networks. The ellipsis's marked blue has a conceptual interpretation.

Numbers			Words
Training set	Familiar set	Novel set	Training set
0	*0*	45	zero
1	*1*	46	one
2	*2*	47	two
3	*3*	48	three
4	*4*	49	four
5	*5*		five
6			six
7			seven
8			eight
9			nine
10			ten
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Table 1: The "training set" is the set used to train the vision modality. The familiar set is equal to the six first entries in the training set except the numbers are written in an italic font. The novel set is very different from anything the number-network is taught. The word set is the training set for the audition modality.

numbers. The Novel set was very different from anything taught to the modalities.

All the networks was trained using back propagation with a learning rate of 0.02 and online weight update. Also, all hidden and output layers had a constant bias of one with modifiable weights. Gaussian noise with a mean of zero and a standard deviation of 0.2 was added to the presented patterns in the vision modality during training. The networks were trained for 10,000 epochs, each one learning the patterns in the training sets. This was done according to the following recipe:

The numbers in the "Numbers" training set was presented to the vision modality in ascending order. If, at each presentation of a number there was a word at the corresponding index in the Words trainings set, then this word was presented to the audition modality and the link between the two networks was updated using the two conceptual representations as input and target (both representations functions as both target and input because the link is bidirectional). As an example, when the vision modality was presented with the number 1, the audition modality was presented with the word "one", then weights were updated in all networks. When the vision modality was presented with the number 20, only weights in the vision modality was updated.

I have yet to give a formal definition of how the novelty measure is computed. This is simply the sum of absolute differences between the input and the output of the network:

$$f(x) = sum(i = 1)^{n} |X_i - Y_i|] \tag{1}$$

In equation (1)X_i is the i^{th} input unit and Y_i is the i^{th} output unit. Every 100^{th} epoch, learning was turned off and the performance of the networks was tested in the following manner:

1. All the numbers in the training set was presented to the vision modality in ascending order. For each presented number, the novelty was computed according to equation (1) and recorded. Then the image in the Numbers training set the image on the output units was closest to was computed. The image it was closest to was the image in the "Numbers" training set it had the lowest sum of absolute differences with. If the image was closest to itself (it had associated correctly) activation was fed through the link, then from the hidden layer of the audition modality to its output layer. If the image associated wrong, an error was recorded. The novelty measure for the link was recorded using equation (1) with X being the original hidden layer pattern of the audition modality and Y being the pattern produced by the link. Also, the image in the "Words" training set the image on the output

units of the audition modality was closest to was computed in the same manner as above [1]. If this image had the same index as the image currently presented to the vision modality, the association was correct; if not, an error was recorded.

2. All the numbers in the Numbers similar set was presented to the vision modality in ascending order. Performance was recorded in the same manner as above (an association was deemed correct if the similar "image" with index one associated with the image at index one in the "Numbers" training set)

3. All the images in the "Numbers" novel set was presented to the vision modality. Only the novelty measure was recorded (No association error or novelty for the link was computed).

No noise was added during testing. Also, during phases 1 and 2 above, if the number presented to the vision modality associated wrong then no value was present for novelty and failure for that instance in the audition modality during the current epoch. To remedy this a novelty and failure of -1 was recorded. The simulation was run ten times and the results averaged with missing values treated specially: If at any epoch, a missing value existed for an instance, then the average for that instance during the current epoch was -1 for both the novelty measure and failure.

From the averaged result an upper bound on correctness was computed separately for each modality in the following manner:

The averaged results from each epoch was sorted in ascending order according to the novelty measure. Then the list was traversed until an entry with a failure larger than zero was found. If the entry before this one had a fail of zero, then that entries novelty measure was the upper bound on correctness. If it had a fail of minus one, the upper bound on correctness was zero. The results for the vision modality is shown in figure 7, the audition modality in figure 8.

As noted in the section above I also wanted to investigate how resistant to noise the link was. Therefor the simulation was run again in the same manner as above except that Gaussian noise with a mean of zero and a standard deviation of 0.2 was now added each time a picture was presented to the vision modality during performance testing. The results for the vision modality is shown in figure 9 and for the audition modality in figure 10.

8.3 Results and discussion

A simulation of intra-modal and inter-modal novelty detection was run. The purpose of the simulation was to investigate how well will the link would

[1] This was only computed if there was a number to word association for the current number.

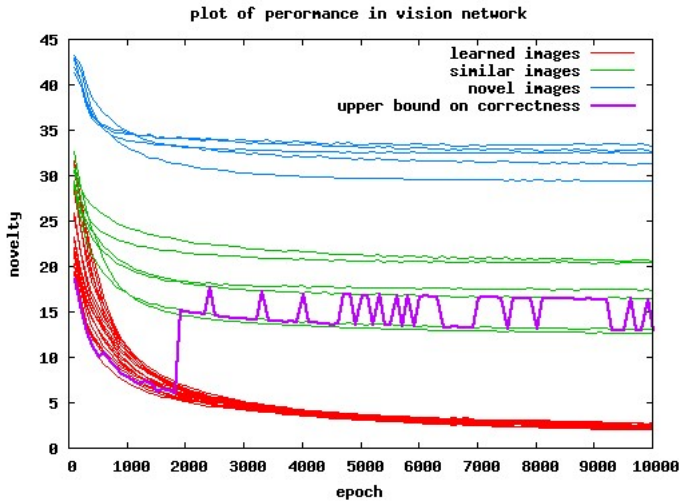

Figure 7: This plot shows the novelty measure for all the patterns presented to the vision modality. The upper bound on correctness is also plotted. It is clear from this graph that the network behaves correctly on patterns inducing much higher novelty than the patterns it was trained on (red).

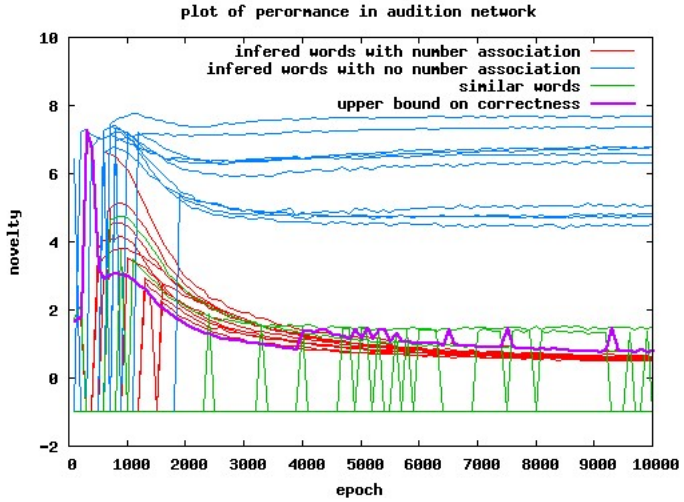

Figure 8: This plot shows the novelty for pictures presented to the audition modality. Note that in this plot, some pictures have missing values for certain epochs and as a result are plotted with a novelty of -1. From this plot one can clearly see a big separation of patterns that has word association (red) and patterns that does not (blue). However, The upper bound on correctness (purple) follow the red lines (the patterns the link and audition modality actually learned) much closer than in figure 7

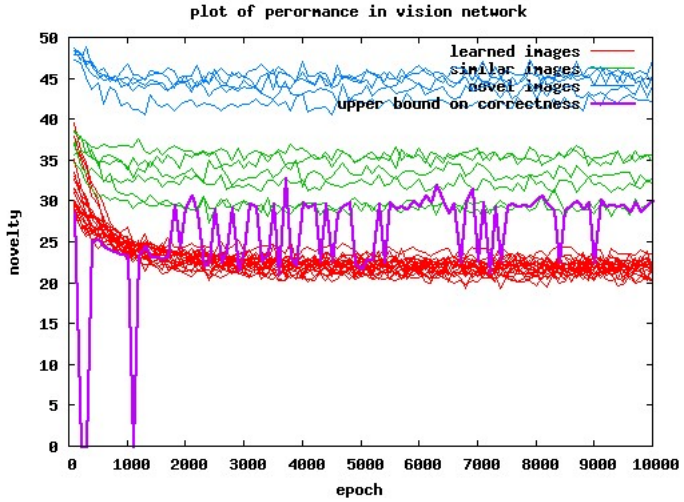

Figure 9: This plot shows the novelty measure for all pictures presented to the vision modality when Gaussian noise was added. The separation between the learned and novel pictures is not as pronounced as in figure 7. Also it is clear that at certain epochs the network actually behaves erroneous as the upper bound on correctness (purple) is lower than some of the learned words (red).

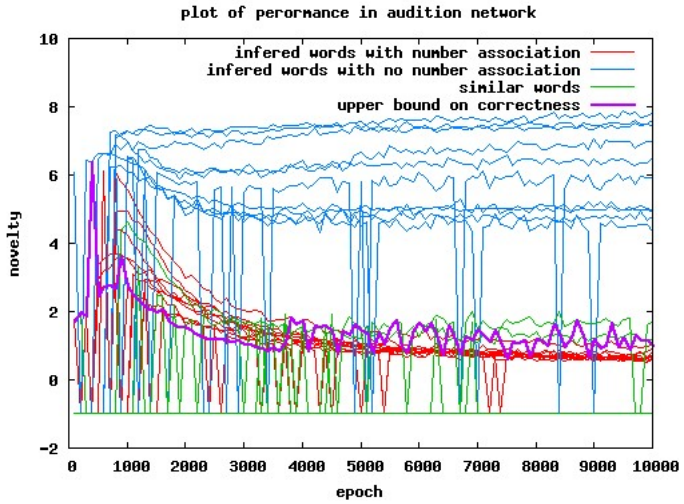

Figure 10: This plot shows the novelty measure for all pictures presented to the audition modality when Gaussian noise was added. Interestingly, this plot is more similar to its corresponding noiseless case than figure 9 with respect to increase in novelty. There is some erroneous behavior where the upper bound on correctness (purple) is lower than than the novelty measure for some of the learned images. Note that missing values are plotted with a novelty of -1.

function throughout training, how noise resistant the system would be, if the system would correctly handle pictures not having a word association and if it was possible to find a general procedure to derive a threshold function for deciding novelty. The simulation shows that the link, although behaving differently than the auto-encoder, functioned well throughout training. The main difference was that the link enforced a stricter judgment on novelty (the upper bound on correctness was much closer to the novelty measure of learned pictures) and that it needed almost twice as many epochs as the auto-encoder to reach a level where it behaved correctly. However, the link also enforced a much smaller separation of learned and similar pictures than did the auto-encoder, something which would in part be responsible for the upper bound being stricter. Also, differing from epoch to epoch, between two and three similar pictures associated correctly in the auto-encoder, while only between one and two in the link. This would also be a factor influencing the upper bound on correctness in the link. This would seem to suggest that the link generalizes to higher degree (but also with higher error), although the simulation was not detailed enough to suggest why the link had a higher error on associations for pictures in the similar set.

When noise was introduced into the system the separation between the different picture sets in the vision modality was much less pronounced and associations on the similar picture set in the vision modality had a higher error. Interestingly, the noise seems to influence the novelty measure of the learned and similar picture sets more than the novel picture set —The simulation does not offer any clues as to why this is so. The noise also causes the networks to associate learned pictures wrong, thus implying that an agent operating in a noise environment would from time to time either falsely judge learned images as novel, or associate learned images wrong (this would depend on the threshold function for judging novelty). This erroneous behavior is also present in the link, however, novelty measures seems some what unaffected of the noise. Although the simulation being to crude to offer any explanation of this behavior, it would be safe to assume that one reason for the novelty measures seeming unaffected by the noise would be that much of the noise is filtered away before activity reaches the hidden layer of the vision modality. It would seem that introduction of noise would affect both the selection of a threshold function and the ability of an agent to make correct associations.

On the issue of whether the link would be able to correctly induce novel conceptualizations in the audition modality when the conceptual representation present in the vision modality had no word association, the answer is yes. From the plot in figures 8 and 10 on can clearly see a separation between pictures with and without a word association. This behavior also seems to be very noise resistant.

As to the selection of a threshold function for deciding novelty, one answer is evident from the behavior shown in figures 7 to 10: The threshold

should be selected as a function deciding on the arithmetic mean of the novelty measure of the learned images multiplied with a function of two variables (possibly more). Moreover, the two variables would be how much training the network(s) has received (the current epoch) and how noisy the environment is. How the function of the current epoch and amount of noise added will behave is left unspecified here. This function would bound how strict the system would be in its judgment on novelty with respect to time. Function (2) below is the function for deriving the threshold, function (3) decides on this threshold.

$$\begin{aligned} &f(m, e, \sigma) = m * g(e, \sigma) \\ &e = \text{current epoch} \\ &m = \text{arithmetic mean of novelty measure} \\ &\sigma = \text{standard deviation of noise} \end{aligned} \tag{2}$$

$$\begin{aligned} &n(x) = \begin{cases} 0 & x < f(m, e, \sigma) \\ 1 & x \geq f(m, e, \sigma) \end{cases} \\ &x = \text{the novelty measure to decide on} \end{aligned} \tag{3}$$

The results presented here, although qualitative in nature, offers promising evidence supporting intra-modal and inter-modal novelty detection using the method outlined above. However, in the presence of noise, performance degraded and some erroneous behavior emerged. There are however a few questions relevant for further investigation left unanswered. It seems pertinent they are voiced here:

- How can the function $g(e, \sigma)$ be specified so it optimizes performance? It is highly probable that an agent operating in a real environment would encounter situations with differing amount of noise. If provided with a signal mirroring the amount of noise, one would want a function behaving optimally in all situations encountered.

- How will the method hold up in more complex domains? The simulation run here operated in a very simple environment. Will the method function well in a real-world scenario?

- How can this approach be extended to allow "many to many" relationships? There is one great flaw evident in this approach with respect to intra-modal novelty detection: It can only handle "one to one" relationships. The reason for this is that feed forward neural networks can only handle "many to one" relationships when the input and output units are interpreted as distributed patterns (i.e not using localist encoding or several clusters as input or output). The situation is even worse here since the link is bidirectional (i.e can only handle "one to one" relationships).

9 Novelty as an attentional guide

In the last section we saw that novelty can function as a continuous "type-checking" parameter. That is, using a novelty measure, the validity of an association can be assessed and compared with other associations. Using novelty, a system of PDP networks can thus validate associations as well as compare them to each other, guiding simulations along a path from probable to less probable associations. However, the links presented in the last simulation was unable to represent "many to one" associations. A wast amount of links would thus have to exist as each concept can include an immense number of conceptualizations, each possibly differing in associations to other conceptualizations in ways not relevant to the concept. Consider a car simulator: it has the ability to simulate a wast amount of different cars, each one possibly differing in the parts it is made up of. So this approach could prove to be untenable. However, if we have a link able to represent "many to many" associations this would no longer be a problem. The problem is now that it is not clear how to select among association that are more or less plausible. Using the network in figure 1 this is not a problem. In this network, the most plausible association will be the original one, while varying among associations can be seen as moving the pattern of activation on the hidden layer. Barsalou's (1999) solution of increasing excitatory connections between regions in a simulator and increasing inhibitory connections within regions, seems more natural. For example, in the previous simulation (7.2), inhibitory connections would develop within the hidden layers of the vision and audition modality, while excitatory connections would develop between. If this was a massively recurrent system, then upon thinking of a number, its most typical word association would appear on the hidden layer of the audition modality after activation has circulated for a period. To simulate another word association (if it exists) patterns of activity in the PFC could inhibit synaptic activity sustaining this pattern while exciting synaptic activity that do not, thus causing the pattern to settle on another attractor. This is in line with Barsalou (1999) proposing that selective attention focus on regions of simulations, thus guiding the simulation. The novelty of a perceived instance would now be the sum of errors on the output layer of the auto-encoder while the novelty of an association between concepts would be proportional to the amount of PFC influence. Novelty modeled this way supports both autonomous learning and autonomous exploration. That is, if the agent is modeled in such a way that it will investigate closer any object inferring novelty into the system, it would explore the world autonomously. The nature of this novelty tells the systems what type of information it lack and thus how the object should be investigated. DA neurons are known to encode novelty, so it is here assumed that the novelty in the system is reflected in DA activity in charge of gating information into the PFC. However, as the agent explores the environment with the goal of

reducing novelty, it further develop its abilities to conceptualize the world. This can happen as a result of perceptual similarity as well as discovering relationships among perceived objects. The discovery of relationships is here also attributed to DA neurons learning what properties predict predict reward, or the meaningful causal relationships among the properties. During, off-line processing, the guiding of a simulation along more or less probable associations can be seen as choosing among more or less novel associations where the novelty is proportional to the PFC's influence on processing. This is analogous to the amount of attention necessary for thinking a thought.

10 Simulation and language

We have previously seen how introduction of language into the conceptual system causes further differentiation of between category representations and increased similarity of within category representations (Cangelosi & Parisi, 2001). In Cangelosi & Riga (2006) they show how language can help acquisition of higher level behavior through grounding transfer. For example, an agent is first taught to execute the actions CLOSE_LEFT_ARM and CLOSE_RIGHT_ARM. After this behavior has been acquired the agent learns that "PULL [is] CLOSE_LEFT_ARM [and] CLOSE_RIGHT_ARM". This new behavior is the acquired by associating the new word PULL with the actions associated with the words CLOSE_LEFT_ARM and CLOSE_RIGHT_ARM. This process of grounding transfer can be seen as an implementation of Barsalou's (1999) symbol productivity mechanism (Cangelosi & Riga 2006). That is, the agent simulates the low order actions and then associates these simulations with the PULL action. Language thus seem to be important both in acquiring efficient conceptual representations and for acquiring more complex behavior. We will now consider how language might be important in imagination in the context of simulation.

Consider running a simulations of concepts that do not have perceptible referents in the world. For example, I might now tell you to imagine two new colors called A and B. These colors are distinct from each other and all other colors. It does not seem feasible that these concepts are represented within the color simulator because there does not exist any perceptual symbols for them in the vision modality. It would be much more sensible that by parsing in the sentence, the concepts A and B becomes integrated into a new simulator guided by the syntax of the sentence in same way as the grounding transfer above. They could thus possibly be represented by a simulation of their name combined with the construal of them being colors in the abstract relationship "distinct" to each other and all other colors. It is hard to see how such abstract thought can appear without the use of language in a perceptually grounded system. This suggest that language plays an important role in all aspects of concepts. That is, concepts can be

acquired by experiencing the world and discovering meaningful relationships in it, but language optimizes the conceptual representation, allows easier acquisition of new behavior and facilitates abstract thought. Or to put it in the context of perceptual simulation, language increases the skill for producing context-specific representations of a category.

Part III
Conclusions

11 Conclusions and discussion

In this thesis we have investigated what concepts are and how they may be represented. We have seen that conceptual representations can be achieved by employing distributed representations in a hidden layer of a neural network. A pattern of activity is in this respect a conceptualization while the concept(s) it belongs to is a region of space treated alike by similarity based generalization. That is, the conceptualization may still have its individual properties only attributed to itself, but the properties relevant to the concept are shared among the representations in that region of space. These regions of space are allocated as dictated by coherently covarying properties of the domain, and thus constitutes a hierarchical representation of it. In this hierarchical representation, the most general concepts occupy the largest amount of space, with their subordinate concepts distributed in clusters allocated inside this space. This hierarchic representation is discovered in a coarse to fine manner, mirroring the conceptual development of a child. Properties being highly typical for a concept are, however, easier to learn and may thus be acquired before properties of concepts superordinate to them, mirroring basic level advantages in lexical acquisition. These typical properties show a higher level of activation throughout training. Frequency of presentation also influences how easy a concept or pattern is to acquire. Frequency of presentation causes a higher pressure to differentiate the instance, thus allocating a larger amount of space to it. This in turn facilitates the learning of its individual properties, thus attenuating the basic level advantages. The properties that covary coherently in the domain becomes more salient than other properties. This allows concepts to be acquired based on especially informative properties, thus possibly overlooking perceptual similarity. When noise was introduced into the system, the hierarchy broke down in a fine to coarse manner. These effects are all due to similarity based generalization and the coarse to fine differentiation of conceptual distinctions, and support many findings in semantic cognition. PDP thus serve as a good starting place for achieving conceptual representations.

By viewing concepts as simulators (Barsalou, 1999; Barsalou, 2003a; Barsalou, 2003b), they are a skill to produce context-specific representations. This is also true of the hidden layer conceptual representations, although depending on whether the context is predictive. A simulator is comprised by a set of modality specific perceptual symbols extracted from perceptual states. Barsalou (1999) also offered valuable insights as to how simulators can support productivity and abstract thought.

We have also seen how categorization can influence perceptual discrimination (Goldstone, 1994). By acquiring categories, the category relevant boundaries acquire distinctiveness with emphasis on the category boundary. For separable dimensions, the irrelevant dimension may receive acquired acquired similarity, however, one null effect was also found in (Goldstone,

1994). For Integral dimensions, the irrelevant dimension also acquired distinctiveness. When two dimensions were relevant for categorization, the separable dimensions competed with each other, while the integral did not. Based on results from (Gluck & Meyers, 1993) I have proposed that a predictive auto-encoder can account for the results found for separable dimensions. During categorization learning, the stimuli along with the assigned category is processed by a predictive auto-encoder. The result is that predictive dimensions acquire distinctiveness while redundant ones acquire similarity. Whether or not the irrelevant dimension acquire similarity will thus depend on whether it has previously been predictive. Language is another factor influencing perceptual discrimination. When language was introduced into the system it had a profound influence on the conceptual representations (Cangelosi & Parisi, 2001). The representations acquired within category similarity and between category distinctiveness. The effect was largest for verbs, but was also present during non-linguistic processing. Language thus helped the network perfect its conceptual skills with respect to non-linguistic behavior. In Cangelosi & Riga (2006) language was used to implement grounding transfer. This is a process where new behavior is acquired by grounding it in previously learned behavior. This was achieved in the guidance of language. This could also be seen as an implementation of Barsalou's (1999) productivity mechanism. The involvement of language in simulating abstract thought has also been discussed. With reference to cangelosi & Parisi (2001) and Cangelosi & Riga (2006) it seems that language has a profound effect on conceptual processing.

Dimensionality of the representation is another important factor in conceptual representations. As the dimensionality increases, the number of examples necessary to reach a given level of performance increases exponentially (Edelman & Intrator, 1997). Auto-encoders is a common method for unsupervised dimensionality reductions which also preserves the topology of the original domain. The dimensionality is reduced by compressing redundant information thus allowing conception to focus on the relevant aspect of the representation.

We have also reviewed a theory if prefrontal cortex function suggesting its implication in guiding computation along processing specific pathways and also in acquiring categories and rules (Miller & Freedman, et. al., 2002; Miller & Cohen, 2001; Braver & Cohen, 2000). The PFC thus seems essential in conception. However, as the rules learned in the PFC is executed frequently, they get "pushed" down to more autonomous areas of the brain and thus become more autonomous. The PFC will thus be most involved in behavior requiring attention, among which acquiring concepts certainly belongs.

A framework for higher level cognitive behavior from Veflingstad & Yildirim (2007) was introduced. This framework was introduced within three levels of cognition: the stimulus-resonse level, the conceptual level

and the language level. Within this framework it is proposed that algorithms exist in the brain and that they are represented non-symbolically at the conceptual level. They operate on non-symbolic concepts and makes decisions using feed forward networks modeling an if-then rule. By employing distributed representations these algorithms exhibit the properties we have this far discussed and will thus exhibit semantic task performance. These algorithms help experiencing more complex thought and are engaged in higher level cognitive tasks such as planning. A simulation of a non-symbolic summation algorithm was presented showing the feasibility of the approach. It was proposed that the PFC is in charge of learning these algorithms, but as they are frequently executed they get "pushed" down to more autonomous areas of the brain and thus no longer require as much attention to be executed.

Novelty was proposed as a means of autonomous exploration and a continuous "type checking" parameter. Novelty is an informative and important "signal" as it allows one to assess knowledge of a perceived instance without any explicit reference of memory. This was implemented in a simulation as the sum of differences between the input pattern and the output pattern of an auto-encoder. The simulation showed that novelty could be reliably assessed within and between modalities as long as the environment was noise free. When noise was introduced, the performance dropped. The simulation was, however, very constrained as the link between the modalities only supported "one to one" relationships. It was therefor suggested that novelty of associations was better assessed as the amount of selective attention the PFC must exert in order for a pattern of activity in a massively recurrent system to settle into a new attractor. It should be mentioned that novelty is here interpreted very broadly. It might be possible that a specific association has been observed many times but that some other association overrides it in the system. This association would thus not be novel in that it has not been experienced but in that it has not been learned to a sufficient degree. Novelty is here also used as an assessment of which of two associations are least familiar. Novelty in this respect would thus be a measure of the amount of stress a current line of processing introduces in the system.

From the material presented in this thesis I will in line with Barsalou (2003b) conclude that the concept arises from a skill for producing context-specific representations. This skill arises from interacting with the world and observing meaningful relationships and properties within it. As this skill improves, perception is affected in a way further facilitating this skill. Once this skill has reached a certain level, language can be acquired, improving this skill even more. This in turn, probably facilitates further acquisition of language. Within reference to the three levels proposed there seems to be a circular dependency between the layers with the concept arising from this interaction. However, since conception can arise simply by similarity based generalization, language would not seem necessary for conception. It

does seem important in the complex conceptual abilities to humans though. Even though it is here concluded that the concept emerges from the skill of the system, this does not mean that it can not be investigated as patterns of activation. As wee have seen, much can be learned from these patterns. They can also be employed in algorithms achieving more complex thought.

References

Barsalou, L. W. (1999) Perceptual symbol systems. *Nehavioral and Brain Sciences*, 22, 577-660.

Barsalou, L. W. (2003) Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 18(5/6), 513-562

Barsalou, L. W., Simmons, K. W., Barbey, A. K. & Wilson, C. D. (2003) Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7(2), pp. 84-91.

Braver, T. S. & Cohen, J. D. (2000) On the Control of Control: The Role of Dopamine in Regulating Prefrontal Function and Working Memory. *Control of Cognitive Processes*, ch. 31. ISBN: 0262133679

Brooks, R.A. (1991) Intelligence without Representation. *Artificial Intelligence*, 47, pp. 139-159.

Cangelosi, A. & Parisi, D. (2001) How nouns and verbs differentially affect the behavior of artificial organisms. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society.*

Cangelosi, A. & Riga, T. (2006) An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments with Epigenetic Robots *Cognitive Science,* 30(4), pp. 673-689

DeMers, D. & Cottrell, G. (1993) Non-Linear Dimensionality Reduction. *Advances in Neural Information Processing Systems*, 5, pp. 580–587.

Edelman, S. & Intrator, N. (1997) Learning as extraction of low-dimensional representations. Academic Press, 1997. in press.

Elman, J. L. & Zipser, D. (1988) Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, 83(4), pp. 1615-1626.

Elman, J. L. (1990) Finding Structure in Time. *Cognitive Science*, 14, pp. 179-211.

Farah, M., J. & McClelland, J., L. (1991) A Computational Model of Semantic Memory Impairment: Modality Speciificity and Emergent Category Specificity. *Journal of Experimental Psychology: General*, 120(4), pp. 339-357

Gluck, M., A. & Meyers, C., E. (1993) Hippocampal Mediation of Stimulus Representation: A Computational Theory. *Hippocampus*, 3(4), pp. 491-516.

Goldstone, R. (1994) Influences of Categorization on Perceptual Discrimination. *Journal of Experimental Psychology: General*, 123(2), pp. 178-200.

Harnard, S. (1993) Symbol Gounding is an Empirical Problem: Neural Nets are just a Candidate Component. *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society.*

Hinton, G. E. (1989) Connectionist learning procedures. *Artificial Intelligence*. 40(1-3), pp. 185 - 234 .

Hook, J., C. (2005) Agents of the Mind. *Biological Cybernetics*, 92(6), pp 427-437.

Japkowicz, N. et al. (1995) A Novelty Detection Approach to Classification. *in the proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pp. 518-523.

Jordan, K. E. & Brannon, E. M. (2006) The multisensory representation of number in infancy. *Proceedings of the National Academy of Sciences,* 103, pp. 3486-3489

Love, B., C. (1999) Utilizing time: Asynchronous Binding. *Proceedings of the 1998 conference on Advances in neural information processing systems*, pp. 38 - 44. ISBN:0-262-11245-0

Miller, E. K. & Cohen, J. D. (2001) An Integrative Theory of Prefrontal Cortex Function. *Annu. Rev. Neursci.*, 24, pp. 167-202.

Miller, E. K., Freedman, D. J. & Wallis, J. D (2002) The prefrontal cortex: categories, concepts and cognition. *Philos Trans R Soc Lond B Biol Sci.*, 357(1424), pp. 11231136.

Pollack, J. B. (1990) Recursive Distributed Representations. *Artificial Intelligence*, 46(1-2), pp. 77-105.

Rogers, T. T & McClelland, J. L. (2006) Semantic Cognition: A Parallel Distributed Processing Approach.
ISBN-10: 0-262-18239-4.

Shepard, R. N. (1987) Toward a Universal Law of Generalization for Psychological Science *Science, New Series*, 237(4820), pp. 1317-1323.

Touretzky, D. S. (1990) BoltzCONS: dynamic symbol structures in a connectionist network. *Artif. Intell.*, 46(1-2), pp. 5-16.

Veflingstad, H. & Yildirim, S. (2007) Non-Symbolic Algorithms within the Context of Different Levels of Cognition. *Proceedings of the Eighteenth Midwest Artificial Intelligence and Cognitive Science Conference*, pp. 29-36

Yildirim, S., Beachell, R. L. & Veflingstad, H. Novel Rock Detection Intelligence for Space Exploration Based on Non-Symbolic Algorithms and Concepts. *AIP Conference Proceedings*, 880, pp. 760-768

Non-Symbolic Algorithms within the Context of Different Levels of Cognition

Henning Veflingstad

Computer Science Department
Norwegian University of Science and Technology
7491 Trondheim NORWAY
veflings@stud.ntnu.no

Sule Yildirim

Computer Science Department
Hedmark University College
2451 Rena NORWAY
suley@osir.hihm.no

Abstract

In Artificial Intelligence community, algorithms and symbols are accepted as the two sides of the same coin. On the other hand, there is a long lasting debate between the two significant approaches to AI, namely Symbolic AI and connectionism, on whether the human brain functions symbolically or not. In this work, we are proposing that algorithms and symbols are not necessarily the different sides of the same coin and that they appear separately. Thus, non-symbolic algorithms can exist. We further proceed to extend the idea of existence of non-symbolic algorithms to their existence in the human brain. We also present the representation of steps in an algorithm and the concepts on which those steps operate. The non-symbolic algorithms are high level and they can be part of either conscious or non-conscious thinking. We also elaborate on different levels of cognition and especially on what we call the conceptual level where high level human thinking happens and where the proposed non-symbolic algorithms reside.

Introduction

Reasoning and behaviour are the two aspects that are assigned to cognitive systems. Proving theorems, thinking, planning, language production are tasks which belong to the reasoning aspect of cognition. On the other hand, movement generation and coordination are tasks that belong to behaviour aspect of cognition. Parallel distributed processing or connectionism has been more successful with tasks relevant to the behavior aspect of cognition. In this paper, we investigate into their role relevant to the reasoning aspect of cognition within the scope of non-symbolic algorithms and non-symbolic concepts.

More explicitly we can propose three levels of cognition:

Stimulus-Response Level: This is the level where there is a direct functional mapping from the sensations of a situation to behavioural outcomes. For example, a robot might be wired up to avoid obstacles conforming to a particular pattern of activations across proximity sensors.

Conceptual Level: This is the level where there is formation of concepts in parallel to a functional mapping from the sensations of a situation to behavioral outcomes. For example, a robot might be wired up to avoid obstacles conforming to a particular pattern of activations across proximity sensors and in the meantime it forms the concept of *obstacle*. The obtained concepts are employed in high level cognitive tasks e.g. thinking, planning, decision-making.

The Language Level: This is the symbolic level and there is a mapping from the conceptual level to the symbolic language level. For example, a robot maps a concept of *obstacle* to the word “obstacle”.

We believe that the current research in the field of cognitive science points to the above three levels in cognition. However, although we believe these levels are the major ones, the relation between levels is an issue under research in cognitive science including researchers from the fields of Artificial Intelligence, robotics, and complex systems. It is still vague how these levels are separated from each other. For example, if sensory-motor systems and conceptual representations are proven to be tightly related, the formation of concepts will be part of level 1, and level 2 will be left with the function of using formed concepts in higher level cognitive functioning.

The research relevant to level 1 is grouped under sensory-motor actions or reactive behaviors. If the foot of a new born infant touches cold lake water, the infant will take its foot away in a reactive way. Since it is newly born, it has not yet formed concepts such as “lake water”, “cold”, “move away” etc. As it grows, it will form these concepts and will be able to utilize these concepts in forming thoughts such as “lakewater is cold” and this capacity is the subject matter of level 2. Relevant work that points to the possibility of level 2 is found in [Ziemke et al., 2005; Tani and Nolfi, 1999]. Rogers and McClelland (2006) suggest hidden layer representations in artificial neural networks as conceptual representations. Cangelosi (2004) uses similar representations to conceptualize world and also addresses the mapping from the conceptual layer to the language level.

The relations between levels are formulated in the following questions:

What is the relation between sensory-motor systems to conceptual representations? Are conceptual representations different in kind from those computed within the perceptual input systems and motor output systems that feed into and out of them?

What is the relation of language to conceptual distinctions in thought? What is the dependency on language for the

determination of the content of a person's internal representation of a concept?

These questions are raised in [Hampton and Moss, 2003] without claiming the 3 levels above and giving references to relevant work. Barsolau (2003) proposes that perceptual simulations represent concepts. One other important point relevant to conceptual representations is whether they are modality specific or amodal.

In this paper, we address level 2 within the scope of non-symbolic algorithms and non-symbolic concepts. Our aim is to show the possible role of conceptual level in tasks other than conceptualizations during navigation. How can an agent utilize the concepts it forms during navigation in tasks that require cognition at a higher level than navigation? Such tasks do not require navigation necessarily, but are more directed towards pure thinking such as planning, algorithm learning, decision making, etc. As to our knowledge, there is not yet much work done that attempts to achieve high level forms of thinking utilizing the conceptual representations that are based on connectionism. On the other hand, Elman (1990) proposes work that moves from symbolic language to conceptual representations using connectionism. Our example to explain non-symbolic algorithms and non-symbolic concepts will involve an agent that senses the lakewater as cold, and conceptualizes the world it experiences in the meantime. However, our implementation is about employing concepts in a summation task.

Before embarking on this explanation an elaboration on what we mean by Non-Symbolic algorithms and Non-Symbolic concepts is in order. For this purpose we have devoted the next section. After this, we explain the basic elements of the Non-Symbolic algorithms in terms of their representations. The penultimate section presents our case study on summation which was implemented in the form of Non-Symbolic algorithms and Non-Symbolic concepts. Last, we give our conclusions.

Non-Symbolic Algorithms and Non-Symbolic Concepts

In this work, we claim that there can be non-symbolic algorithms; that is, algorithms need not necessarily be represented symbolically and they do not necessarily need to do symbol manipulation. This claim requires that algorithms and symbols need not be the two sides of the same coin. The claim is valid both for artificial and natural intelligence. Algorithms compose the higher levels of cognition.

If algorithms need not be represented symbolically and need not necessarily manipulate symbols, how are they represented and what do they manipulate? These questions are posed both for computational and natural intelligence.

We will first elaborate on the answer to the second question. We propose that non-symbolic algorithms can operate on non-symbolic concepts which are also represented non-symbolically. A non-symbolic concept representation is achieved by employing distributed representations which have been studied and applied for numerous kinds of problems in the connectionism literature (Hinton, McClelland & Rumelhart, 1986). In distributed representations, the particular pattern used to represent an item is determined by the nature of that item, and so similarities and differences among the items to be represented will be directly reflected in similarities and differences among the representations themselves [van Gelder, 1991, p41]. There is supportive evidence in [Chao, et al., 1999] for distributed representations in human temporal cortex.

Concepts can be considered as discrete categorizations (Gårdenfors, 2000), as opposed to sensation and perception that can be considered as continuous categorizations. Concepts can be seen as a discretization of the perceptual space [Gershenson, 2004].

In distributed representations, a cluster of neurons is involved in the representation of a concept (Dorffner, 1989). It is different from local representations where only a single neuron is responsible for representing a concept. Representing non-symbolic concepts by using distributed representations is valid both for natural and artificial intelligence.

Before we elaborate on the answer to the first question also, we will give a definition of an algorithm and indicate some of its important aspects. An algorithm is defined as follows:

"An algorithm is a step by step procedure for performing a task, solving a problem or accomplishing some end"

An important aspect of an algorithm in this definition is that it is composed of steps. We define a step as a process which starts and ends its execution at a certain point in time; that is, a step is carried out for a period in time and another step can be invoked for execution when its execution is finished. This aspect of a step does not necessarily require that steps always need to be carried out in sequence. Steps can be carried out in parallel and this aspect will neither conflict with the fact that a step starts and ends at a certain point in time nor with the fact that it lasts for a period of time.

In addition to the above stated aspects, an algorithm can be either learned or innate but we are not going to elaborate on this aspect any further in this paper. Also, since algorithms are supposedly already existent in the human brain according to our non-symbolic algorithm approach, then what is left is to carry out the steps in an algorithm when required.

In the next section, we will explain how non-symbolic algorithms might be represented in the agent's brain (answer to the first question) and how they might manipulate or operate on non-symbolic concepts.

Representation of Non-Symbolic Algorithms: Basic Elements

In this section we will choose to present examples to introduce our general model or framework which

represents non-symbolic algorithms and the concepts that the non-symbolic algorithms operate on.

The use of hidden layers in artificial neural networks has been proposed for implementing distributed representations in semantic cognition for representing concepts. However, as to our knowledge, utilizing represented concepts and investigating their role in high level cognition such as thinking, planning, algorithms etc. has rarely been worked on. A recent work which represents concepts using distributed representations and uses them for high level cognition is given in [Rogers and McClelland, 2006].

At this point, it is almost clear that a step in an algorithm can be an action. Also, we propose If-Then rules as steps in an algorithm.

In an If-Then rule, certain inputs are expected to cause certain outputs and for that reason, it embodies a decision-making process. As a result, in computational intelligence, an If-Then rule can be represented by using a feed forward neural network. Then, the concepts in the inputs of the neural network would represent the "If" part of the If-Then rule whereas the concepts in the outputs of the neural network would represent the "Then" part of the If-Then rule. Finally, correct decision-making would require training the neural network in such a way that, certain inputs of the neural network could be mapped to the corresponding outputs of the network.

On the other hand, such a decision-making mechanism is not necessary in representing an action. A representation that helps differentiate an action from another action would be enough to represent actions. For that reason, distributed representations are also useful in representing actions and concepts involve actions in our work. This approach is in line with some previous Artificial Intelligence work that presents ways of representing concepts and where actions are also accepted as concepts. Having said that, we can now propose that representations of objects in the real world, our sensations from the real world and actions are all concepts that we human beings form through our life times. In this paper, we do not discuss whether concepts are part of animal cognition although, in some animals such as mammals (dogs, cats, primates), we believe it to be so.

Returning back to our goal of presenting means of how algorithms can be represented in an agent's brain, we will examine a bathing algorithm. The algorithm consists of two steps:

1) Sense the lakewater.
2) If the lakewater is cold then get out of lakewater.

The first step of the algorithm consists of the concepts of "sense" and "lakewater". The second step is an If-Then rule and it has concepts of "lakewater" and "cold" in the precondition part of the rule and the concepts of "getoutof" and "lakewater" in the action part of the rule.

If the lakewater is really cold then sensing the lake water physically while being in the lakewater will activate the concepts of "lakewater" and "cold".

We assume that concepts of "lakewater" and "cold" are already in the agent's brain because we assume that the agent who senses the lakewater now has probably already had previous sensations of a lakewater and its being cold. On the other hand, we propose that forming the thought of "lakewater is cold" simply requires the activation of both of these concepts at the same time.

Since both of "lakewater" and "cold" are concepts and that they do not involve a decision making process, an activation pattern in a neuron cluster for each of these concepts can be dedicated to representing them. The concept of "cold" corresponds to a sensation drawn from the lakewater whereas the "lakewater" itself is an observation or a sensation from the real world.

As an example the activation pattern in Figure 1 might be representing the concept of "cold" whereas the activation pattern in Figure 2 might be representing the concept of "warm" in the same neuron cluster. Note that the representation of the concept of "warm" requires the activation of a different set of neurons if the same neuron cluster is used for the representation of the concept of "cold". It is also the case that some of the neurons in both representations overlap in their activations for representing the two concepts.

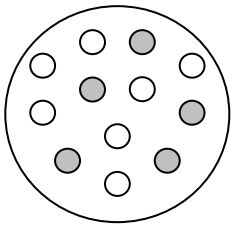

Figure 1. A cluster of neurons that represent the concept of "cold".

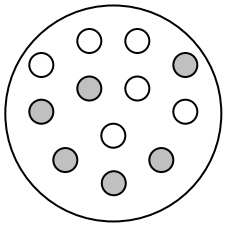

Figure 2. The same cluster of neurons that represent the concept of "warm".

First, "sense the lakewater" step in the algorithm is executed. The execution of this step will cause the two levels of cognition. At high level cognition (conceptual level), the concepts of "sense" and "lakewater" will be activated (Figure 3). This will correspond to the execution of the first step of the algorithm as well as the thought of "sensing lakewater". At the low level, sensing of lakewater will be achieved.

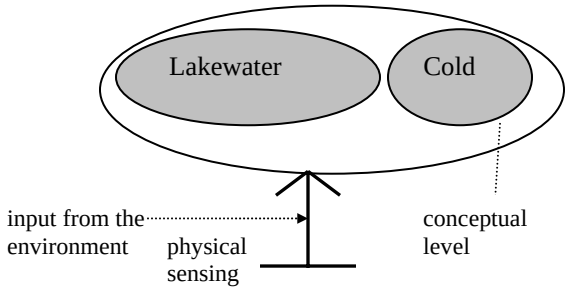

Figure 3. Sensing the lakewater activates the concepts of ”lakewater” and “cold” at the same time.

In Figure 4, it is shown that sensing the environment physically invokes the patterns of activations of the concepts of “lakewater” and “cold” at the same time (activation of the concepts are indicated by the gray color) meaning that these two concepts are related to each other, in this case “lakewater is cold”. Since the agent in the lakewater has possibly experienced a cold lakewater earlier, it is assumed that the representation of each of these concepts in a corresponding cluster already exists. Sensing the environment causes the activation of the corresponding representation for each of the concept. In addition to that, sensing each of these concepts simultaneously cause the simultaneous activation of these concepts. This way, not only the agent physically senses the lakewater as cold but also forms a high level thought which is “lakewater is cold” as part of its high level cognition and as part of its thinking while executing the if-then rule.

The thinking of each concept refers to the activation of gray neurons as one experiences the coldness through his/her early life.

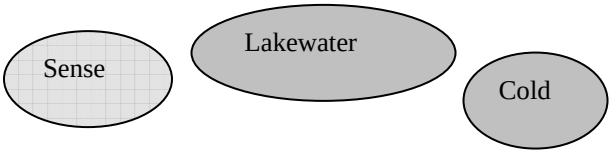

Figure 4. Sensing the lakewater activates the concepts of ”lakewater” and “cold” at the same time.

The next step to execute is the If-Then rule (Figure 5). The action part of the If-Then rule can be activated once its precondition is true. Its precondition became true after the activation of concepts of “lakewater” and “cold”. In the current situation, action “sense” is out of context and for that reason, even if it might be still active, it will not be part of the execution of the current step. Thus, it is highlighted with lighter gray color during the execution of the second step.

In the If-Then rule of Figure 5, each of the concepts of “lakewater” and “cold” will be activated as a result of sensing “lakewater” and sensing it to be “cold”. In addition to that, the physical incomfort of the agent’s body caused by the coldness of the lakewater will cause the action of “getting out of” the lakewater. The action of “getting out of” at the low level of cognition corresponds to the concept of “getoutof” at the high level of cognition. There is not a causal link from the concepts of “lakewater” and “cold” to the concept of “get out of” when for the first time, an agent feels a lakewater as cold and gets out of it.

In Figure 5, by the end of the execution of the first step, the “sense” concept will not be activated any longer.

However, all three of these concepts are activated as one senses or experiences them. Also, a causal link will begin to form between the concepts of “lakewater” and “cold” and the concept of “get out of” during the first simulatenous experience of these concepts as in Figure 6 (As the two concepts and the “get out of” concept are experienced simultaneously in time).

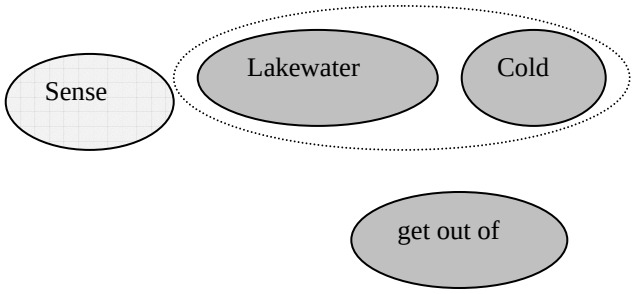

Figure 5. The execution of the second step. Note that the physical discomfort of an agent’s body causes the execution of the “get out of” action concept.

The forming of a causal link might help to experience a thought such as “if the lakewater is cold, I should get out of it”. In (Jordan & Brannon, 2006), a research on multisensory representation of numbers in 7 month old infants is reported. The multisensory representation reported seems to be the equivalent of what we refer to as distributedly represented concepts in this paper. The behaviour of the infants seem to correspond to an execution of an innate rule when they hear the voices of 2 or 3 women. The rule can be expressed as follows:

If 2 women’s voices are heard, look at 2 women’s faces.

We believe that an infant might also have the conceptual representations of “women”, “voice”, “face” in addition to the number they are reported to be representing. However, if one of the infants could reply in response to a question such as ”Why did you choose to look at 2 women’s faces instead of 3?” by saying “I heard 2 women’s voices and that is why I looked at 2 women’s faces”, then they would be forming a high level thought.

In the bathing algorithm example, a thought can appear for example just before we really get into the water and will help us plan ahead about what to do if the water is cold. This situation can also be interpreted as learning what to do under different circumstances.

We also need to remember that the action part of the If-Then rule in the bathing algorithm will be executed automatically when the concepts in its preconditions are activated after the formation of the causal link as in Figure 6. As a result, the model which we are presenting here both refers to how high level thought of an agent might be generated in addition to the proposal that it is a model of representing algorithms.

Also, in Figure 6, someone "gets out of" lake water and for that reason, a link (conceptual relevance link but not physical) must be present between "lakewater", "cold" and "get out of" concepts in addition to the earlier mentioned causality link so that the agent knows what to "get out of" (Figure 7).

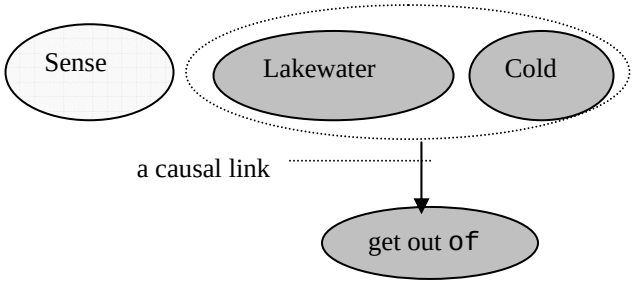

Figure 6. The execution of the second step. Note that the concepts of "lakewater" and "cold" together cause the execution of the "get out of" action concept.

In Figure 7, the conceptual relevance link is not composed of physically present links. The concepts of "lakewater" and "cold" will still be active at the time the "get out of" becomes activated. As we proposed earlier, simultaneously active concepts contribute to the generation of a particular thought and for that reason the three active concepts would correspond to the "get out of cold lake water" thought when they are all active simultaneously when once "get out of" concept becomes active as a result of the activation of the two other concepts, namely "lakewater" and "cold".

This approach presents a model of thinking and how thought might be produced without the use of any symbols.

On the other hand, there can be a physically present conceptual relevance link. A physically present conceptual relevance link is important when one concept needs to be activated as a result of the activation of another concept. We have already explained a situation as such while we gave the example of the automatic activation of "get out of" concept as a result of the activation of concepts of "lakewater" and "cold". A physically present conceptual relevance link is named as a causal link in this paper.

In addition to that example, a non-action concept might cause the activation of another concept automatically. As an example, the concept of "my bag" will activate the concept of "black" since my bag is black and it will stay black until I replace it with another bag with a different color. Sometimes the simultaneous activation of concepts will be enough to generate a thought but sometimes physical conceptual relevance links will be necessary to activate some other concepts as a result of activation of a particular concept.

Also, when it comes to the point that one thing is permanently a feature of the other and hence thinkable from the thought of the other, then it will be necessary to have physical causal links between two concepts. My red t-shirt is an example where red and my t-shirt are relevant concepts and red is a feature of my t-shirt.

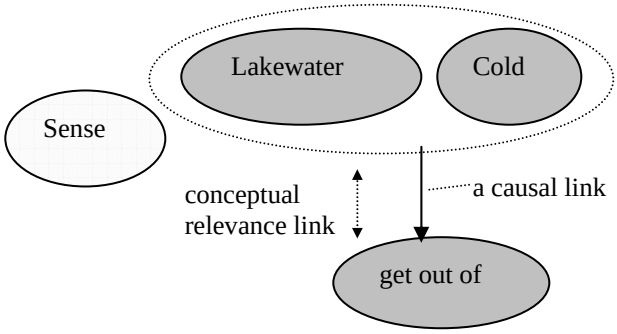

Figure 7. The execution of the second step. The conceptual relevance link between "lakewater", "cold" and "get out of" concepts.

A more complicated causally relevant concept activations is given in Figures 1 and 2 of (Yildirim, 2005). The arrows in the figures correspond to physical causally relevant concept links.

Non-symbolic Summation

In Figure 5, we have seen that the concepts that represent sensations can be associated with concepts that represent actions. In the figure, the action concept "get out of" is associated with the concepts of "lakewater" and "cold". The association provides the activation of the action concept "get out of" as a result of activations of concepts of "lakewater" and "cold" at the same time. The association is not bidirectional though.

We need to be able to simulate this association while representing the basic elements of algorithms using artificial neural networks. Towards this end we have implemented a Non-Symbolic summation algorithm simulating an agent doing long addition. The simulation employed two neural networks. One, an auto-encoder learning the identity-mapping on 256 pixel grayscale images of the digits from zero to nine deriving a conceptual representation of the numbers seen (Figure 8). These images are not intended to represent the retinal activation patterns produced when perceiving a digit. The simulation assumes a nonverbal numerical representational system extending across different modalities. Evidence supporting this contention is presented in (Jordan & Brannon, 2006). Since the auto-encoder is too simple a system to perform this task, the images are intended to

represent all instances having a numerical interpretation known to the agent. A simulation on how to achieve a common numerical representation for all instances perceived would deserve an article of its own.

The other network, henceforth called the summation network, was a four-layered feed-forward network taught to map the conceptual representations of two numbers and a carry to a conceptual representation of a number and a carry out (answer), thus carrying out the actual summation (Figure 9).

To add two numbers the following steps was performed:

Starting with the lowest order digits, one digit from each number was presented to the auto-encoder in turn. After each presentation, the conceptual representation derived on the conceptualization layer was copied to its corresponding position on the input neurons of the summation network. If these were the first digits to be added, carry in was set to the conceptualization of zero. If not, the conceptualization present at carry out was copied to carry in. Then activation was fed forward and the activation on the output neurons corresponding to the answer was copied to the conceptualization layer of the auto-encoder and decoder. This was the first digit in the answer. This process was carried out until there were no more digits to add and carry out was zero.

The corresponding Symbolic expression of the summation algorithm can be given as follows:

1. Sense two digits from two numbers visually.
2. If there are no more digits to add and carry out is zero, display summation in a visual form.
3. Else, sum the two digits and the carry into a number and a carry.
4. Go to Step 1.

The auto-encoder was a five layered neural network, the middle layer being the one representing the concepts. The input and output layers had 256 neurons each. The two layers on each side of the middle layer had 60 neurons each, and the middle layer had 15 neurons. The network was trained using back-propagation for 10000 epochs with a learning rate of 0.01. At each epoch, the digits from zero to ten were presented. Weights were updated after each presentation of a digit. After presentation of a digit and before feeding activation forward, Gaussian noise with a mean of zero and a standard deviation of 0.2 was added to the input neurons.

The summation network had 45 input neurons (3 * 15), 37 neurons in each of the two hidden layers, and 30 output neurons (15 * 2). It was trained using back-propagation for 10000 epochs with a learning rate of 0.01. At each epoch, every combination of two digits and a carry out of 200 possibilities were taught. Each digit having 10 possible values ranging from 0 to 9 leads to 100 possibilities for the digit pairs to be added. Having a carry value 0 or 1 increases the number of possibilities for digit pairs to 200. No Gaussian noise was added. Weights were updated after each combination. In both networks, every hidden and output neuron also had a constant bias of one with a modifiable weight.

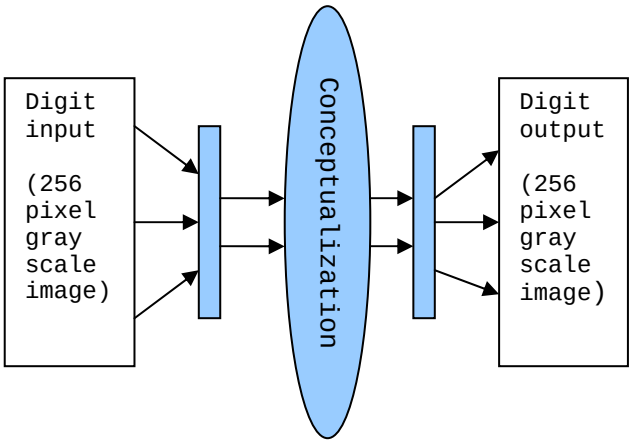

Figure 8. Auto-encoder.

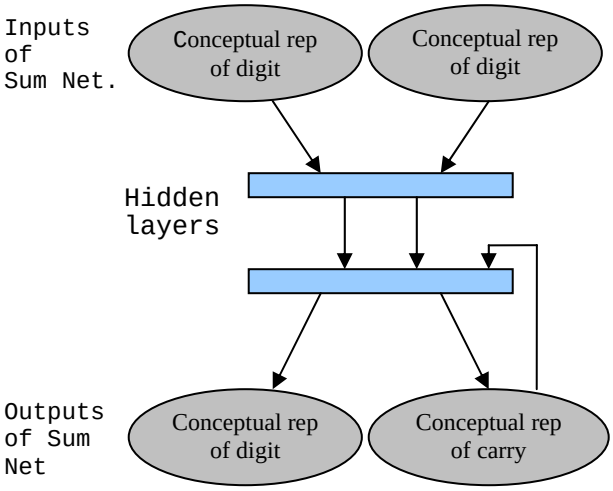

Figure 9. Summation Network.

We ran the simulation ten times and averaged the results. After 10000 epochs the sum of squared errors was 1.83 and 3.16 in the auto-encoder and the summation network respectively. To test the performance of the summation network, training was turned off each 500th epoch and 1000 summations trials were performed. At each summation trial two random numbers in the range 0-999 were selected and summed according to the procedure described above. The performance was recorded in the following manner:

Failure to produce the correct carry or answer was an error. Each answer and carry produced was decoded in the auto-encoder. Then, for each image in the dataset, the sum of absolute differences with respect to the decoded image was computed. The image in the data set with the smallest sum of absolute differences was interpreted as the summation network's intended answer/carry. If this was incorrect an error was recorded.

A reconstruction error was also computed as the average of the absolute differences computed above.

After 4500 epochs the number of errors stayed below 3 and at epoch 10000 it dropped to zero. The reconstruction error after 10000 epochs was 3.11. Figure 10 shows the activation pattern for the conceptualizations.

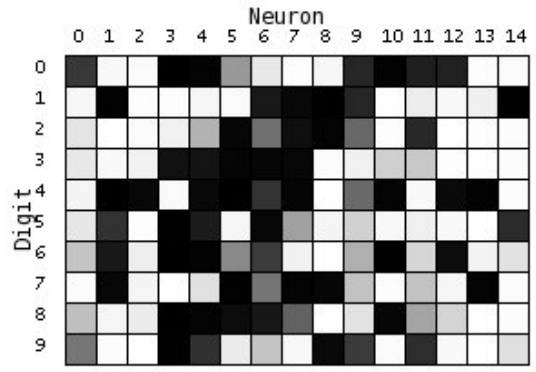

Figure 10. The activation pattern for the conceptualizations

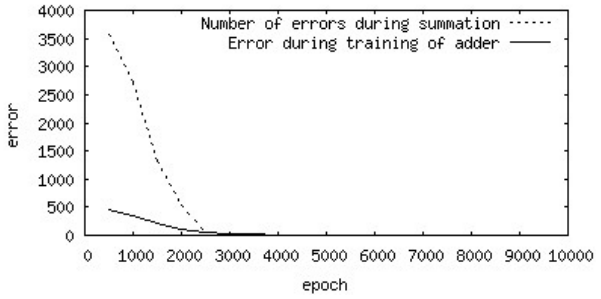

Figure 11. Errors during summation and training.

This shows that the network had almost perfectly learned non-symbolic summation, the only "error" being an inability to produce an answer identical in appearance to the digits learned (an absolute difference of 3.11 is very close though). Figure 11 shows the change in squared sum and summation error over 10000 epochs, maximum error obtained being around 3500 when averaged for 10 runs.

Conclusion

Our aim is to show the possible role of conceptual level in tasks that require conceptualizations. Navigation is a task where an agent can form conceptualizations. However, how can an agent utilize the concepts it forms during navigation in tasks that require cognition higher level than navigation? Such tasks do not especially require navigation but are more directed towards thinking such as planning, algorithm learning. As to our knowledge, there is not yet much of such work that attempts to achieve high level forms of thinking utilizing the conceptual representations that are based on connectionism. For that reason, we presented a way of how distributed representations can be employed in a higher level cognitive function which is representation of algorithms in an agent's brain and execution of them. The algorithms are non-symbolic in nature and they employ non-symbolic concepts. One such algorithm is a bathing algorithm and we have shown how that algorithm can be expressed non-symbolically instead of employing text-like symbolic representations. The execution of such an algorithm refers to the activation of concepts in it considering a current step and flow of activations through causal links from concepts to concepts. In general, concepts might be represented and associated with each other in order to represent steps in an algorithm and create thoughts. One another such algorithm is summation algorithm. We have implemented that algorithm to show that implementation of non-symbolic algorithms is feasible using non-symbolic concepts in a high level cognitive task such as summation. We employed associative neural networks for that purpose.

References

Barsalou, L.W. (2003) Situated simulation in the human conceptual system. Language and Cognitive Processes, 18, pp. 513–562.

Chao, L. L., Haxby, J. V.and Martin, A. (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects Nature Neuroscience 2, pp. 913 – 919.

Cangelosi, A. (2004) The sensorimotor bases of linguistic structure: Experiments with grounded adaptive agents. In S. Schaal et al., editor, SAB04, pp. 487--496. Los Angeles: Cambridge MA, MIT Press.

Dorffner, G. (1989) Replacing Symbolic Rule Systems with PDP Networks: NETZSPRECH: A German Example. *Applied Artificial Intelligence, 3 (1),* 45-67.

Elman, J. L. (1990) Finding Structure in Time. *Cognitive Science*, Volume 14.

Gershenson, C. (2004) Cognitive Paradigms: Which One is the Best? *Cognitive Systems Research*, 5(2), pp.135-156, June 2004.

Gårdenfors, P. (2000) Conceptual spaces. Cambridge, MA: MIT Press.

Hampton, J. A., & Moss, H. E. (2003) Concepts and meaning: Introduction to the special issue on conceptual

representation. Language and Cognitive Processes, 18(5/6), pp. 505-512.

Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. (1986) Distributed Representations. *Parallel Distributed Processing: Exploration In The Microstructure Of Cognition.* Vol. 1. MIT Press.

Jordan, K. E., & Brannon, E. M. (2006) The multisensory representation of number in infancy. *Proceedings of the National Academy of Sciences*, 103, pp. 3486-3489.

Rogers, T. T., McClelland, J. L. (2006) Semantic Cognition: A Parallel Distributed Processing Approach. ISBN-10: 0-262-18239-4.

Tani J. and Nolfi, S. (1999) "Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems", Proc. 5th Int. Conf. on Simulation of Adaptive Behavior, (Eds) R. Pfeifer, B. Blumberg, J.A. Meyer and S.W. Wilson, MA: The MIT Press, pp.270-279. The revised version is in Neural Networks, Vol.12, pp. 1131-1141.

Van Gelder, T. (1991) What is the "D" in "PDP"? A survey of the concept of distribution. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory*, pp. 33-59, Lawrence Erlbaum.

Yildirim, S. (2005) *"Innate Planning Mechanisms"*, MNAS'05, Modeling Natural Action Selection Workshop, Extended Abstract, 30-31 July, Edinburgh.

Ziemke, T., Jirenhed, D.-A., and Hesslow, G. (2005) Internal simulation of perception: a minimal neuro-robotic model. *Neurocoputing*. 68, pp. 85-104.

Novel Rock Detection Intelligence for Space Exploration Based on Non-Symbolic Algorithms and Concepts

Sule Yildirim[1a], Ronald L. Beachell[1b], Henning Veflingstad[2]

[1a, b] *Computer Science Department, Hedmark University College, 2451Rena, Norway*
[2] *Computer Science Department, Norwegian University of Science and Technology, Trondheim, Norway*
[1a] *+47 62430478, suley@osir.hihm.no*
[1b] *+47 99313855, ronald.beachell@osir.hihm.no*

Abstract. Future space exploration can utilize artificial intelligence as an integral part of next generation space rover technology to make the rovers more autonomous in performing mission objectives. The main advantage of the increased autonomy through a higher degree of intelligence is that it allows for greater utilization of rover resources by reducing the frequency of time consuming communications between rover and earth. In this paper, we propose a space exploration application of our research on a non-symbolic algorithm and concepts model. This model is based on one of the most recent approaches of cognitive science and artificial intelligence research, a parallel distributed processing approach. We use the Mars rovers, Sprit and Opportunity, as a starting point for proposing what rovers in the future could do if the presented model of non-symbolic algorithms and concepts is embedded in a future space rover. The chosen space exploration application for this paper, novel rock detection, is only one of many potential space exploration applications which can be optimized (through reduction of the frequency of rover-earth communications, collection and transmission of only data that is distinctive/novel) through the use of artificial intelligence technology compared to existing approaches.

Keywords: Space Rover, Novel Rock Detection, Artificial Intelligence, Non-Symbolic Algorithms and Concepts, Connectionism.
PACS: 87.17.Aa; 87.18.Sn**;** 87.19.La; 89.20.Ff; 89.75.-k; 91.60.-x**.**

INTRODUCTION

In this paper, we are proposing the use of an artificial intelligence model based on non-symbolic algorithm and concepts for performing appropriate tasks on space exploration missions. The specific space exploration application is novel rock detection. The model that we propose uses the Parallel and Distributed Processing (PDP) approach of Artificial Intelligence (AI) and Cognitive Science Research (Rumelhart et al., 1986). A detailed description of non-symbolic algorithms and concepts which form the basis for the model can be found in previous work (Yildirim and Beachell, 2006).

Parallel Distributed Processing or Connectionism is an alternative or complementary approach in Artificial Intelligence to a Symbolic or Classical approach (Garson, 2002). The reason for choosing Connectionism is that Symbolic approaches are unable to perform common sense reasoning or to exhibit knowledge of rudimentary physical reality, such as how things change over time (Luger, 2005).

In a space exploration application such as novel rock detection, a rover uses intelligence provided by our model to obtain knowledge of rudimentary physical reality and carry out common-sense reasoning. Thus, a rover will be able to make more decisions on it own (autonomy) and therefore use more efficiently the time it is awake to carry out more space exploration at a lower cost. This is due to the potential for rovers with a higher level of artificial intelligence to autonomously perform more exploration tasks (decide to use a rock abrasion tool, use instruments to scan geological samples, perform an imaging scenario, etc.) in a given time period. The model also enables a rover to optimize the use of on-board resources (Estlin et al., 2005) in order to complete or extend the mission and to adapt

to their environments to ensure their survival in space. Specifically, some of the ways that rover autonomy can be leveraged to increase efficiency and reduce costs are:

(1) By reducing the dependency of the rover on time consuming communications with earth and thus allowing a rover to carry out more tasks for longer periods of time when awake.

(2) By reducing the frequency of command and control communications between rover and earth, there may be a reduced requirement for resources for monitoring and controlling the rovers at the command centers on earth.

(3) By finding which geological samples are distinct or unique when there may be multiple discoveries of the same rocks, the rover reduce that duplicate scientific data is stored and downloaded to earth.

As stated above, Connectionism rather than Symbolism seems to be a better approach to equip rovers with intelligence and required autonomy. Utilization of the proposed model to the novel rock detection application emphasizes a rover's capacity to conceptualize its environment and, form associations between its conceptualizations while it is executing innate (embedded) algorithms. The algorithms are represented in the form of concepts and associations between concepts.

Connectionist or neural models of intelligence emphasizes the brain's ability to adapt to the environment in which it is situated by modifying the relationships between individual neurons. Rather than representing knowledge in explicit logical sentences, the computational model embedded in the rover mimics the human brain by capturing knowledge implicitly, as a property of patterns of relationships (Luger, 2005). The model represents reality in a distributed manner. For those reasons, it is much easier for a connectionist approach to represent reality although it is often perceived to have graded values or be incomplete.

Some of the work on building autonomous robots for space applications include large scale assembly in space (Simmons et al., 2000), exploration of Mars with a federation of Intelligent robots (Goldberg et al., 2003), and Mars autonomy project (Singh, 2000). All of these are based on symbolic approaches of AI. However, as to our knowledge, none of these work or any others present a way of enabling a rover to infer its own rules and algorithms and utilize them for adapting to its environment.

In the rest of the paper, we first present the application of the basic principles of the proposed model for a novel rock detection application. Secondly, we present the architecture of the proposed model. Lastly, we give conclusions and propose future work.

BASIC PRINCIPLES OF THE MODEL APPLIED TO NOVEL ROCK DETECTION

The novel rock detection application uses our non-symbolic algorithm and concepts model to analyze rock samples and determine if the rock has been encountered previously. The rover uses its sensory equipment to examine or sense the rock sample and provide inputs to the model to form conceptualizations of the rock samples.

If the rover determines that a rock sample is novel with one sensory analysis, then it can decide to do more extensive analysis and scanning of the rock sample. This requires that rovers are equipped with advanced visionary and sensory equipment. The assumption is that the data collected about novel rocks can contribute to the research on origins of solar systems and the universe. In the coming sections, we will present what in a rover's environment can be conceptualized, how it can be conceptualized, and what kind of algorithms such explorers can be equipped with for space exploration tasks (Smith, 2005).

Non-Symbolic Concepts

A non-symbolic concept representation is achieved by employing distributed representations which have been studied and applied for numerous kinds of problems in the Connectionism literature (Hinton, McClelland, and Rumelhart, 1986). In distributed representations, a cluster of neurons is involved in the representation of a concept

(Dorffner, 1989). It is different from local representations where only a single neuron is responsible for representing a concept.

As an example of distributed representations, the activation pattern in Figure 1 might be representing the concept of "Mineral Composition A" whereas the activation pattern in Figure 2 might be representing the concept of "Mineral Composition B" in the same neuron cluster. Note that the representation of the concept of "Mineral Composition A" requires the activation of a different set of neurons in the same neuron cluster where the representation of the concept of "Mineral Composition B" also exists. It is also the case that some of the neurons in both representations overlap in their activations for representing the two concepts. Then, the thinking of that concept refers to the activation of shaded/colored neurons which have been pre-wired together as a rover detects Mineral Composition A in a rock using a spectrometer for example.

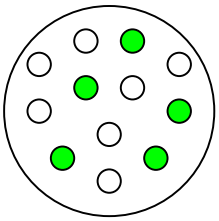

FIGURE 1. A Cluster of Neurons That Represent the Concept of "Mineral Composition A."

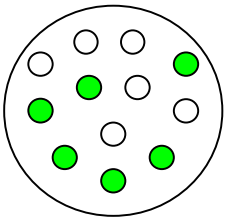

FIGURE 2. The Same Cluster of Neurons That Represent the Concept of "Mineral Composition B."

In the next section, we will explain how non-symbolic algorithms might be represented in a rover and how they might manipulate or operate on non-symbolic concepts

Non-Symbolic Algorithms

This section will give insight into how non-symbolic algorithms are represented in a space exploration rover such that the rover is able to detect novel rocks. The representation of steps in an algorithm is closely tied to representation of non-symbolic concepts on which a non-symbolic algorithm operates. A rover's non-symbolic algorithm can be composed of steps each of which is either an action step or an If-Then rule.

We will first explain how an If-Then rule can be represented non-symbolically. In an If-Then rule, certain inputs are expected to cause certain outputs and for that reason, it embodies a decision-making process. An autonomy which will be embodied in a rover in the form of computational intelligence can represent an If-Then rule by using a feed forward neural network. The inputs of the neural network then represent the values of the concepts in the preconditions of the "If" part of the If-Then rule. The outputs of the neural network represent the values of the concepts in the "Then" part of the If-Then rule. Correct decision-making requires training the neural network in such a way that certain inputs of the neural network are mapped to the corresponding outputs of the network.

On the other hand, such a decision-making mechanism is not necessary in representing an action. A representation that helps differentiate an action from another action would be enough to represent actions. For that reason, distributed representations which are used for representing concepts are also useful in representing actions.

In summary, we can take representations of objects in the real world, sensations from the real world and actions all as concepts and employ distributed representation to represent them. It is assumed that the following two mechanisms are in place for a rover to be able to represent non-symbolic algorithms and concepts:

a) A mechanism that help robot filter what exactly it is to categorize from what is out in the world. As sensory inputs are categorized and conceptualized, the resultant categories and concepts are perceptually grounded and refer to things in the real world. Relevant work can be found in (Steels, 1996; 1997).

b) A mechanism that holds concepts and categories. This mechanism can be hard-coded in advance in the rover's memory.

Regarding point a), human beings are good at selecting relevant and necessary information while categorizing the world as a necessary aspect of their cognition (Sloman, 2005). A rover can categorize rocks, mineral composition, and grain among others. Regarding point b), although it is not known how human beings achieve this aspect of cognition, it is possible to propose means of achieving it computationally and in a way that supports distributed representations. Relevant work is found in (Steels, 1997; Steels, 1996; Cangelosi, 2003; Ziemke, Jirenhed, and Hesslow, 2005).

Returning back to our goal of presenting means of how non-symbolic algorithms can be represented in a space exploration rover, we will examine a novel rock detection algorithm. This algorithm can be embedded in the rover as an innate non-symbolic algorithm. The algorithm consists of two steps which can be expressed in natural language representation:

- Sense the rock (with spectrometers, microscope, color imaging system, etc.).
- If the rock is a New Rock Type, has a New Mineral Composition or New Grain Size/Shape, then it is novel.

A space rover equipped with different sensory instruments like a camera (imaging), microscope and spectrometer would develop non-symbolic concepts in the respective modalities of its "brain" responsible for processing sensory input from these devices (Singularity, 2001). For each of the concepts in the above algorithm, an activation pattern in a neuron cluster is dedicated to represent each concept in the form of a distributed representation. The concepts are "Rock", "New Rock Type", "New Mineral Composition", "New Grain Size/Shape" and "Novel". The first step of the non-symbolic algorithm is an action concept "Sense" which operates on the concept of "Rock".

The rover will activate the concepts of Sense and Rock during the execution of the first step of the algorithm. At the end of execution of the first step, the rover will have sensed the rock by means of a camera, spectrometer and microscope. For this example, we will, for simplification purposes, only consider the concepts of "Rock" and "New Mineral Composition" which are produced and/or activated as a result of spectrometer analysis.

If the rock sample is found to have a unique combination of minerals (as a result of spectrometer analysis) at the point in time that it is "sensed", then a new non-symbolic concept is formed. For purposes of explanation, we will call this concept "New Mineral Composition" although that concept name has no significance in a non-symbolic concept representation. If in a rock sample, an existing (familiar) combination of minerals is detected, a previously formed concept is activated again. The "Sense" concept will become inactive by the end of the execution of the first step while the concepts of "Rock" and "New Mineral Composition" will remain active. An activated concept is indicated by an oval formed by dashed lines in Figure 3, 4 and 5.

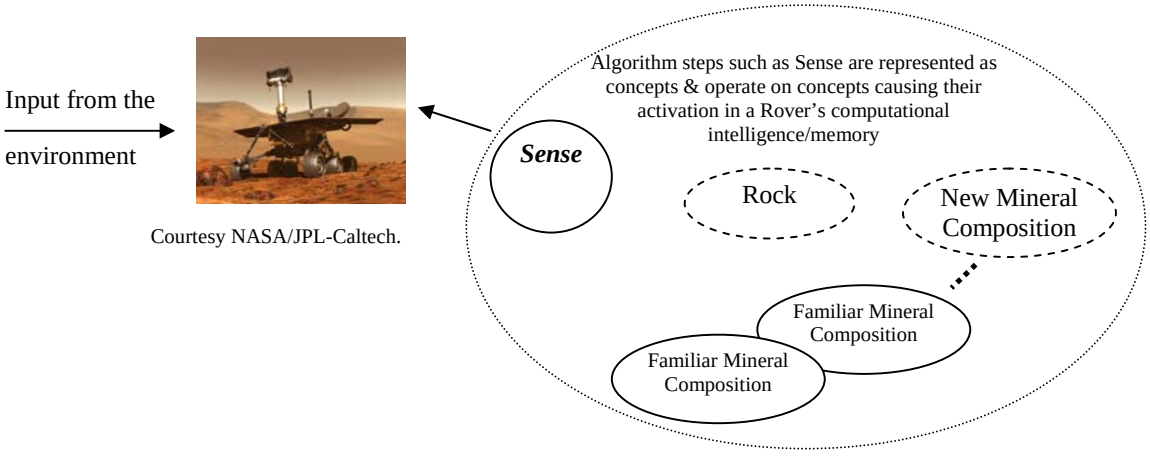

FIGURE 3. Sensing the Rock Activates the Concepts of "Rock" and "New Mineral Composition" at the Same Time.

The next step in the novel rock detection algorithm is to execute the If-Then rule as represented in Figure 4. The If-Then rule operates on concepts of "Rock", "New Rock Type", "New Mineral Composition", and "New Grain Size/Shape". All of these concepts except "Novel" are the concepts that occur in the precondition part of the rule.

The concept of “Novel” is in the action part of the rule. Once any of the concepts in the precondition part of the If-Then rule are activated, the action part of the If-Then rule can be activated, that is the concept of “Novel” becomes active. The level of activation is “graded” according to how many of the concepts of the precondition part are activated, i.e. the highest activation level occurs if all three of the concepts are activated and the lowest activation level occurs if only one of the three concepts is activated. The amplitude of activation can be digitalized to the desired degree of grading of the novel concept by using a threshold.

In the If-Then rule representation of Figure 4, only the spectrometer relevant concept “New Mineral Composition” is shown. Thus, the concept of “Novel” is activated as a result of activation of concepts of “Rock” and “New Mineral Composition”. This is due to a causal link from the precondition concepts to the action concept. In this example, the causal relation is innately embedded in the rover. However, there can be situations where a rover will have to associate some concepts as a result of other concepts and hence form its If-Then rules itself as it explores its environment autonomously.

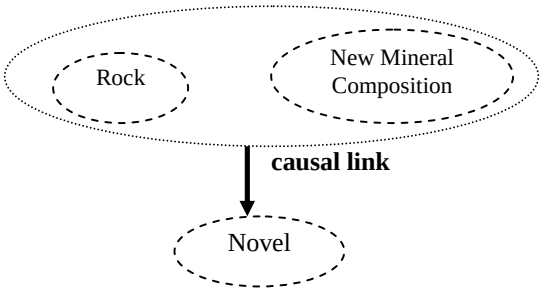

FIGURE 4. The Execution of the Second Step of a Non-Symbolic Algorithm - The Activation of Two Precondition Concepts Causes the Activation of the Novel Concept.

Figure 5 shows the modalities and related concepts involved in novel rock detection. The conditions under which the concept of “Novel” is activated or it is not activated have been explained previously for the Spectrometer modality. The other modalities function the same way. Familiar in the context of a concept means that the neural network has determined that the information from the sensors is identical or similar to a rock previously encountered in the mission.

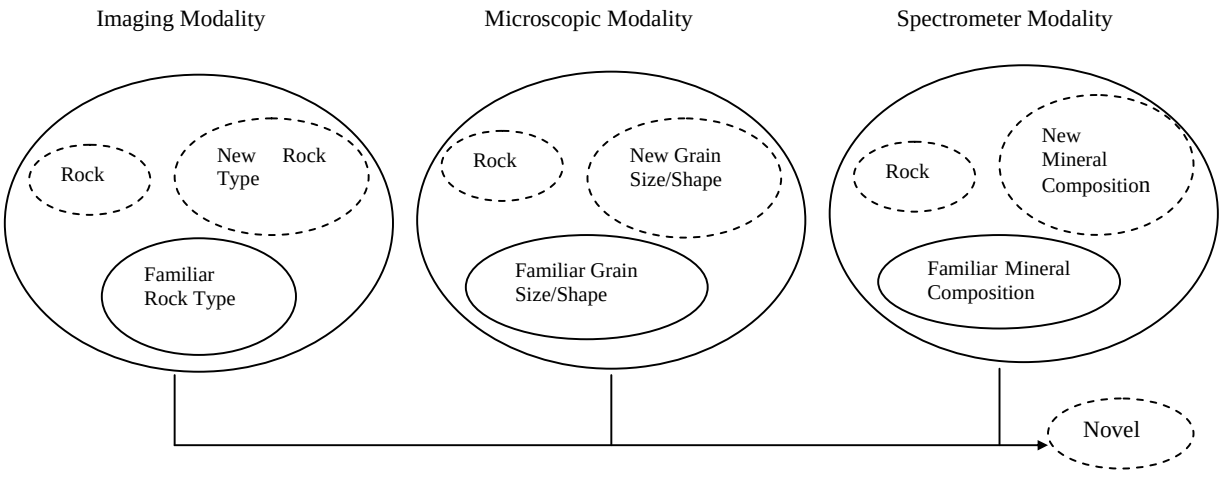

FIGURE 5. Different Modalities are Causally Linked to Concept “Novel.”

When a rover comes across a rock sample, it first investigates a rock sample by Color imaging modality. If this causes the activation of the novel concept, the rover investigates the rock further with other modalities.

Contrary to our model based on connectionism which can conceptualize features, a rover based on a symbolic approach like NASA's OASIS (Castafio, 2003) is more formalized since it makes predictions/decisions based on a feature vector extracted through some image processing routine (transduction). The designers therefore limit the cognitive world of the rover to the expressional power these feature vectors endow. Thus, if some unforeseen feature is important for successful cognitive processing, the rover would operate sub-optimally in its environment. A connectionist approach would only be limited by its sensory equipment and the associations it is capable to make.

ARCHITECTURE OF MODEL FOR NOVEL ROCK DETECTION

The ability of the model to detect novel rocks will require training of the neural networks. This can be done on earth before launch of the rover on a spacecraft by using data from earth or data collected by probes or rovers that have previously visited the destination planet or moon or similar celestial bodies. However, since a high incidence of rocks encountered in space exploration may be different than those rocks on which the data for the neural networks are trained on, the model may at the beginning of the mission falsely conclude that a rock is novel when in reality sensing information for a similar (now familiar) rock has been stored and downloaded to earth previously by the rover. However, as the mission progresses, the rover will become more and more accurate in determining what rocks have not been encountered before and avoid collecting, storing, and downloading duplicate data on rocks. Thus a rover can autonomously decide not to do more scanning of a particular rock and discard the duplicate data thus avoiding valuable resources (communication link, power, etc.) and time necessary for downloading the less useful information to earth. This time can be used to find more novel rocks or to perform other space exploration tasks. This is extremely important when taking into consideration that space rovers are resource limited compared to similar rovers or robots on earth.

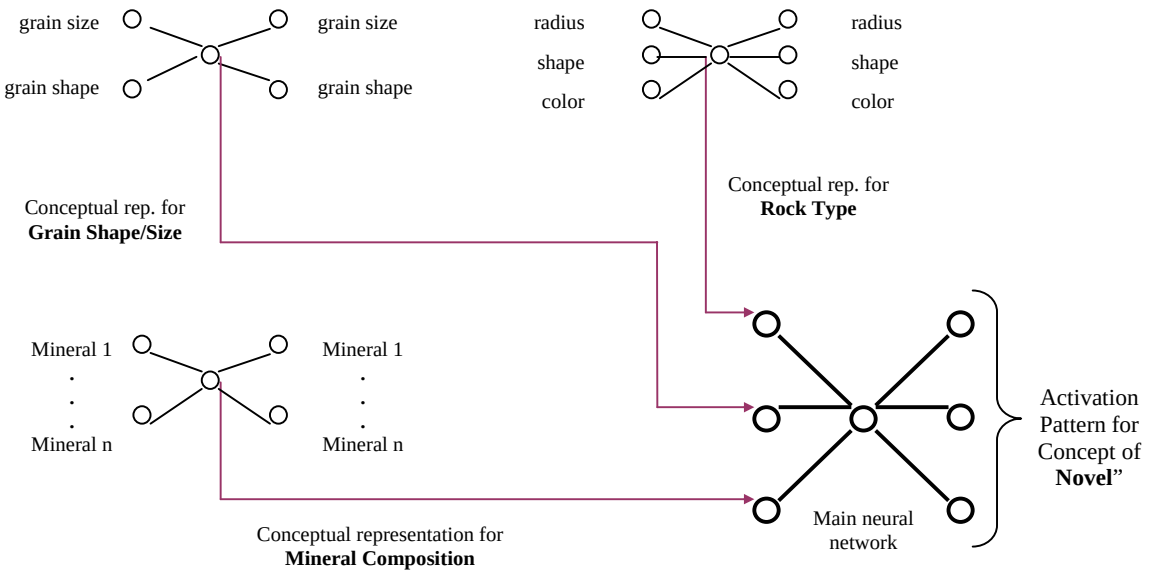

FIGURE 6. Three Modalities (Each Represented by a Neural Network) Representing Inputs from Sensing Systems for Activation of a Concept Called "Novel" in a Novel Rock Application.

The rover is exposed to a database of rock images and rock samples first on earth and then in space. The rock images are analyzed by Color imaging and rock samples are analyzed by Microscope and Spectrometer modalities. The analysis provides the features which are used by modality neural networks to generate conceptual representations. For example, the Color imaging modality will extract the features of color, shape and radius of a rock and generate conceptual representation for the features (Figure 6). Color imaging feature extraction is based on work described in Castafio. The Spectrometer modality will extract minerals that can exist in a rock sample as features and generate conceptual representation from these features obtainable from that modality. The mineral types

are indicated by Mineral 1, …, Mineral n labels in the Figure 6. The Microscopic modality will extract grain size and grain shape as features of a rock sample. Please note that in the three modality neural networks of Figure 6; the circles represent input, inner, and output layer nodes and each circle can stand for multiple nodes. More specifically for the input and output nodes, each circle can stand for multiple nodes of a feature and hold digital values. All of the inner layer nodes in the modality neural networks hold analog values.

Autoassociative neural networks are good at making associations and matching patterns (Callan, 1999) and therefore the modality neural networks are implemented as autoassociative networks to be able to detect a novel rock. That is, there is a corresponding autoassociative neural network for each modality where the representations at the inner/hidden layer of the neural network hold the conceptual representations obtained from the features of the corresponding modality. For example, the autoassociative network for Color imaging modality autoassociates (maps) the patterns (values) from the input nodes for the features of color, shape and radius to the same features at the output nodes. The neural network for this modality obtains a conceptual representation for "Rock Type" concept from these features at the hidden layer of the neural network shown as in Figure 6. The corresponding autoassociative networks of the other two modalities produce conceptual representations for "Mineral Composition" and "Grain Shape/Size" concepts relevant to the Spectrometer and Microscopic modalities respectively.

There is also a separate neural network where the conceptual representation for the "Novel" concept is held. The second step of the novel rock detection algorithm requires that there is a main neural network which maps conceptual representations from all modalities into the concept of "Novel". This network is the one shown by bold lines in Figure 6. Please note that in the main neural network, the circles represent input, inner, and output layer nodes and each circle can stand for multiple nodes. All of the nodes in the main neural network hold analog values.

All of the neural networks in Figure 6 are trained with existing knowledge of rocks on earth. For that reason, for example, the neural network which is used to generate conceptual representations of a Rock Type will generate the same hidden layer pattern for a previously trained (radius, shape, color) feature set. The feature set has already been autoassociated to the same feature set by training while the rover is on earth. However, a new feature set that was not used for training the neural network before can not be autoassociated with itself at the outputs of the network when it occurs the first time in operation on a mission. For that reason, training will be required to autoassociate the feature set with itself and generate a conceptual representation of it at the hidden layer of the network.

The conceptual representations that are either distinct or not and that are obtained from the modality neural networks are the inputs to the main network. The main network will be trained on earth to associate a familiar conceptual representation set from three of the modalities with a pattern of all zeros at the output layer. The training will be done for all familiar rock samples. A pattern of zeros at the output layer means that the conceptual representation set obtained from three modalities for a rock sample is not novel. If the representation set is new, then it will generate a pattern of activation at the output nodes of the network which is not all zeros. Then a training algorithm such as backpropogation algorithm will be used to train the network to obtain a pattern of activation at the output nodes that represents the concept of Novel. The pattern will be kept at the output nodes for a while and then the network will be trained to have a pattern of all zeros at the output layer. This is necessary in case a similar conceptual representation set from three modalities occur at the inputs again, the network does not recognize the set as novel.

After training the main network for novelty, the "Novel" concept will have the strongest level of activations in its active nodes in its pattern at the outputs of the neural network in case of all the conceptual representations are generated for the first time. Each of the conceptual representations will activate the Novel concept to a certain degree whereas activation of all of these concepts simultaneously will increase the levels of the active nodes in the concept of Novel to the highest degree meaning that the rock sample is novel to the highest degree.

The links between the input and the hidden layer nodes, the links from the hidden layer nodes to the output nodes and the hidden nodes themselves represent the causal link shown in Figure 5 by black color.

CONCLUSION

In this paper, we proposed a novel rock detection application for space exploration which could be implemented using our model based on non-symbolic algorithms and concepts. The model gives an insight into how a system onboard a space rover, in some ways similar to the human brain, can represent and utilize non-symbolic concepts

and algorithms to perform tasks. Being based on connectionism, the model makes use of the desirable properties of connectionist approach.

In general, neural networks which implement connectionism-based solutions to AI exhibit robust flexibility and decision making intelligence in the face of challenges posed by the real world or even space exploration. In this paper, we presented the ability of rovers to better use its resources for space exploration tasks by reducing the frequency of communications, most importantly exchanges related to control from earth. This makes a rover more productive when in an awakened mode. In the next paragraph, the other main advantage which is related to determining which geological samples are distinct or unique such that only distinct scientific data is stored and downloaded to earth.

While our model when applied to the novel rock applications is not better than other computational approaches at storing and downloading data on a particular rock, its strength is that it autonomously can determine what rocks are novel or distinct. The mission objective with our novel rock application is to collect scientific data on rocks that are encountered in space and at the same time avoid collection of data on rocks that are similar to or identical to rocks encountered earlier. This enables the rover to maximize its resources toward the objective of collecting as much unique data on rocks encountered on a planet or moon as possible.

Not all tasks of the novel rock detection application are suited to be realized by our model. Therefore an optimal rover should include different systems equipped with models of artificial intelligence based on Connectionist and Symbolic approaches. Additionally, the rover should include software based on multi-level abstraction models, object-oriented methodologies and design patterns that are found in today's robotic platforms (Nesnas et al., 2006). An example of a task that is better suited to a Symbolic approach is an image processing functionality such as the color imaging system on the rovers currently exploring Mars. This system or a new generation of it can supply the inputs to our model to generate conceptualizations of the rover's environment.

Using concepts represented in a distributed manner and using connectionism as a basis for making agents intelligent and adaptive to their environment is a state of the art research area in the Artificial Intelligence and cognitive science fields (Rogers and McClelland, 2004). We believe that our research contributes to enhancing the state of the art in this area. Our research is most likely among the first to present a model and an overall framework where distributed concepts, connectionism and our proposed non-symbolic algorithms are used for high level cognitive functioning/processing for a large scale application.

Possible future work includes the use of our model in other space exploration applications that can benefit from our approach, a threat detection and avoidance application, and further development of our model to provide rovers with better adaptation and autonomy. Within future work, the model can be extended to encompass prediction and planning cognitive faculties which may provide supplementary intelligence or backup systems to current solutions for autonomous planetary mobility.

ACKNOWLEDGMENTS

The authors would particularly like to thank the anonymous reviewers for their detailed comments that considerably improved the quality of the paper. Without the comments, we probably would not have realized the need for a rewrite with total focus on space exploration. We also would like to thank STAIF Administrative Staff, Claudia O'Keefe for her excellent assistance with mediating reviewer comments, and her positive attitude towards handling our questions. The last but not the least, we would like to thank Raymond W. Jensen for his encouragement in submission of a paper to STAIF07.

REFERENCES

Callan, R., *The Essence of Neural Networks*, Prentice-Hall, NY, NY, 1999, pp. 84-99.

Cangelosi A., "Neural network models of category learning and language," *Brain and Cognition*, **53**, 106-107, (2003).

Castafio, R., Anderson, R.C., Estlin, T., DeCoste, D., Fisher, F., Gaines, D., Mazzoni, D., Judd, M., "Rover Traverse Science for Increased Mission Science Return," in proceedings of *2003 IEEE Aerospace Conference*, Institute of Electrical and Electronics Engineers, Inc., Manhattan Beach, CA, 2003, pp. 8_3629-8_3636.

Dorffner, G., “Replacing Symbolic Rule Systems with PDP Networks: NETZSPRECH: A German Example,” *Applied Artificial Intelligence*, **3**, 45-67, (1989).

Estlin, T., Gaines, D., Chounard, C., Fisher, F., Castano R., Judd, M., Anderson, R., and Nesnas, I., “Enabling Autonomous Rover Science Through Dynamic Planning and Scheduling,” in proceedings of *2005 IEEE Aerospace Conference*, Institute of Electrical and Electronics Engineers, Inc., Manhattan Beach, CA, 2005, pp.385-396.

Garson, J., "Connectionism," (2002), http://plato.stanford.edu/archives/win2002/entries/connectionism/#6, accessed October 22, 2006.

Goldberg, D., Cicirello, V., Dias, M. B., Simmons, R., Smith, S. and Stentz, A., “Market-based multi-robot planning in a distributed layered architecture,” in proceedings of *the 2003 International Workshop on Multi-Robot Systems*, Kluwer Academic Publishers, Dordrecht, London, 2003, pp. 27-38.

Hinton, G.E., McClelland, J.L., and Rumelhart, D.E., *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, MIT Press, Cambridge, MA, 1986, pp. 77-109.

Luger, F., *Artificial Intelligence*, Harlow, England, Addison Wesley, 2005, pp. 16-19.

Nesnas, I.A., Simmons, R., Gaines, D., Kunz, C., Diaz-Calderon, A., Estlin, T., Madison, R., Guineau, J., McHenry, M., Shu, I. and Apfelbaum, D., “CLARAty: Challenges and Steps Toward Reusable Robotic Software,” International Journal of Advanced Robotic Systems,” **3**, 61-66, (2006).

Rogers T.T., and McClelland, J.L., Semantic Cognition: A Parallel Distributed Processing Approach, MIT Press, Cambridge, MA, 2004, pp. 1-114.

Rumelhart, D.E., McClelland J.L., and the PDP Research Group, Parallel Distributed Processing: Exploration in the Microstructure of Cognition, MIT Press, Cambridge, MA, 1986, pp. 10-76.

Simmons, R., Singh, S., Hershberger, D., Ramos, J., and Smith, T., “First Results in the Coordination of Heterogeneous Robots for Large-Scale Assembly,” in proceedings of *International Symposium on Experimental Robotics*, Springer-Verlag Lecture Notes in Control and Information Sciences, Berlin, Germany, 2001, pp. 323-332.

Singh, S., Simmons, R., Smith, T., Stentz, A., Verma, V., Yahja A. and Schwehr, K., "Recent Progress in Local and Global Traversability for Planetary Rovers," in proceedings of *the IEEE International Conference on Robotics and Automation*, Institute of Electrical and Electronics Engineers, Inc., Los Alamitos, CA, 2000, pp. 1194-1200.

Singularity Institute for Artificial Intelligence, Inc, “General Intelligence and Seed AI,” (2001), http://www.singinst.org/GISAI/mind/modality.html, accessed October 22, 2006.

Sloman S., *Causal Models: How People Think About the World and Its Alternative*s, Oxford University Press, NY, NY, 2005, pp. 13-17.

Smith, A., Woods, M., and Townend, M., “Advanced On-Board Software for Planetary Exploration,” (2006), http://www.ercim.org/publication/Ercim_News/enw65/smith.html, accessed October 22, 2006.

Steels, L., “Perceptually Grounded Meaning Creation,” in proceedings of *Second International Conference on Multi-Agent Systems* , the AAAI Press, Menlo Park, California, 1996, pp. 338-344.

Steels, L., “Constructing and Sharing Perceptual Distinctions: Coupling of Meaning Creation and Naming Game in Software Experiments,” in proceedings of *the European Conference on Machine Learning*, edited by van Someren, M. and Widmer G., Springer-Verlag, Berlin, Germany, 1997, pp. 4-13.

Yildirim, S. and Beachell, R.L., “Does the Human Brain Have Algorithms,” in the proceedings of *International Conference on Artificial Intelligence*, CSREA Press, Las Vegas, Nevada, 2006, pp. 486-492.

Ziemke T., Jirenhed D., Hesslow G., “Internal Simulation of Perception: A Minimal Neuro-Robotic Model,” *Neurocomputing*, **68,** 85-104, (2005).