Learning Visual Scene Descriptions: An Approach to Symbol Grounding

by

Paul Leighton Williams

Honors Thesis

Turing Scholars Program

Department of Computer Sciences

The University of Texas at Austin

December 2005

Learning Visual Scene Descriptions: An Approach to Symbol Grounding

Paul Leighton Williams
The University of Texas at Austin, 2005

Supervisor: Risto Miikkulainen
Additional Readers: Raymond Mooney and Greg Plaxton

Abstract

The problem of how abstract symbols, such as those in systems of natural language, may be grounded in perceptual information presents a significant challenge to several areas of research. This thesis presents an unsupervised learning model that allows analysis of the symbol-grounding problem. The model learns associations between visual scenes and linguistic descriptions and provides means for direct examination of what it has learned. By analyzing the system, it is possible to assess how well symbols can be grounded in perceptual information with an unsupervised neural network architecture. The model demonstrates potential for accomplishing grounding in artificial systems and provides valuable insight into the grounding task.

Contents

Abstract **ii**

Contents **iii**

Chapter 1 Introduction **1**

1.1 Motivation 1

1.2 Approach 2

1.3 Overview of the Thesis 3

Chapter 2 Background **5**

2.1 Philosophical Foundations 5

2.2 Symbol Grounding Research 7

2.3 Conclusion 8

Chapter 3 Model **10**

3.1 Network Architecture 10

3.2 Visual Inputs 13

3.3 Linguistic Descriptions 14

3.4 The Self-Organizing Map 15

3.5 Associative Connections between Maps 17

3.6 Conclusion 18

Chapter 4 Experiment and Results **19**

4.1 Training Procedure 19

4.2 Testing Procedure 20

4.3 Results and Analysis 22

Chapter 5 Future Work **25**

5.1 Analysis of Network Performance 25

5.2 Comparison With Child Language Acquisition 25

5.3 Comparison with Other Learning Models 26

5.4 Representing Sequential Information 27

Chapter 6 Conclusion **28**

Bibliography **30**

1. Introduction

In order for symbols to possess intrinsic meaning, they must have relationships to what they represent (Harnad 1990). If a symbol system is not grounded in this manner, then the meanings of its symbols can only be defined in terms of other symbols, never establishing a relationship to the outside world. To illustrate this point, consider trying to learn a language solely by reading a dictionary written in that language. This would be an impossible task because none of the symbols in the language would have any bindings to the external world, and so the symbols would be meaningless. Thus, to some extent symbols must be directly linked to their referents.

1.1 Motivation

The symbol-grounding problem presents a significant challenge for several fields, including philosophy, robotics, and cognitive science. It is as of yet unknown to what extent abstract symbols may be grounded directly in perceptual experience and a cognitively plausible model of how this is accomplished is still lacking. This motivates the work presented in this thesis, which describes and examines an unsupervised neural network model that addresses the symbol-grounding problem.

The necessity for grounded symbols has been put forth as a criticism of attempts to model human intelligence with purely symbolic systems (Harnad 1990). According to the criticism, regardless of how intelligent the behavior of a system seems, if its symbols depend on external interpretation to attain meaning then it cannot be said that the system has achieved understanding. For understanding to occur, the

symbols must have inherent meaning for the system in terms of its experiences of the external world. In other words, the symbols must relate to the system's perceptual experience. In order to develop a symbol system, it is necessary then to understand how symbols may become grounded in their perceptual correlates.

The grounding of symbols requires establishing perceptual categories and associating these categories with abstract tokens. Consider as an example how one may learn the meaning of the symbol "square". First it is neccessary to determine the commonalities of all the external objects to which the symbol refers that are distinct from attributes of objects in other categories. It must be determined what is meant by "squareness". This process involves emphasizing the differences between categories and minimizing the differences within categories, a process referred to as "categorical perception" (Harnad 1987). A square refers to an object with four sides of equal length and not just to a closed figure. Once the boundaries of a category have been established, it can be bound to an abstract token, at which point symbol-grounding has occurred. In our example, once acquired, the concept of "squareness" can be associated with the token "square" and the meaning of the token becomes established. Successfully modeling this process computationally could allow symbols used by machines to have directly grounded meanings as well as provide insight into how grounding may be accomplished by the human brain.

1.2 Approach

Neural network architectures provide strong candidates for such computational models. Several neural network models of symbol-grounding have been presented and studied, some of which will be discussed in Chapter 2. This thesis continues

this strain of research, proposing a new model. Unlike previous architectures, the model attempts to accomplish symbol-grounding through a completely unsupervised learning procedure. The network learns correlations between visual scenes and linguistic descriptions. After the network is trained, it will be analyzed to see how well it has grounded the meanings of linguistic tokens in their corresponding visual inputs. The model presented by this thesis was chosen because like architectures have been shown to perform well on learning tasks similar to symbol grounding (Miikkulainen 1997; Li 1999), as will be discussed in Section 3.2. The architecture uses self-organizing maps (Section 3.4) and associative connections between them (Section 3.5) to accomplish the learning task in a completely unsupervised process.

The lack of corrective error feedback makes the model more neurologically plausible (Section 3.1). Additionally, the vastly simplified problem domain presented by this thesis allows for direct examination of what the model has learned. This ability to examine representations directly could prove valuable for future attempts at understanding symbol grounding and for building grounded systems.

1.3 Overview of the Thesis

This thesis consists of three parts: Introduction and Background (Chapters 1 and 2), Model and Experiment (Chapters 3 and 4), and Evaluation (Chapters 5 and 6).

Chapter 2 discusses the symbol-grounding problem and the philosophical ideas behind it. Several studies examining the issue of grounding are also discussed.

Chapter 3 describes the neural network model which is examined in this thesis. The implementation of the learning task is also explained.

In **Chapter 4**, an experiment which examines the grounding capabilities of the model is described and the results are discussed.

Chapter 5 reviews the accomplishments of this thesis and outlines several possible extensions to be undertaken in the future.

Finally, **Chapter 6** offers a summary of the research efforts presented in this thesis and concludes.

2. Background

The purpose of this thesis is to present and analyze an unsupervised learning model of how linguistic symbols may be grounded directly in visual information. This chapter provides background information on the symbol-grounding problem and several research efforts that have addressed the challenge that it presents. The first section of this chapter discusses some of the philosophical ideas behind the symbol-grounding problem. The second section discusses several other research investigations that have examined the issue of grounding. The third and final section summarizes the significant results from the previous two sections and explains the significance of the work described in this thesis in light of these findings.

2.1 Philosophical Foundations

The idea behind the symbol-grounding problem is well expressed in the thought experiment of the Chinese Room Argument (Searle 1980). Consider a man with no knowledge of the Chinese language placed in a room with two openings. Through one of the openings, questions written in Chinese are passed. The man's task is to write appropriate responses to the questions and pass them through the other opening. The man has with him a book containing an exhaustive list of possible questions written in Chinese and corresponding appropriate responses, also written in Chinese. With this resource, the man is able to look up the questions which are passed in and write down appropriate responses from his book. The point of consideration then is whether the man in the room can be said to understand Chinese. Clearly this is not the case because none of the Chinese symbols have any meaning for the man. The man has no idea what any of the symbols refer to because

they are not grounded in his perceptual experiences. This situation is analogous to the computation done by purely symbolic systems. Such a system has no way of binding the meanings of its symbols to the outside world. Thus the symbols have no intrinsic meanings and the system is incapable of gaining understanding. This is the symbol-grounding problem, as presented by Harnad (1990).

Harnad argues that abstract symbols must be grounded bottom-up in two forms of representations, both deriving directly from perceptual experience. The first type, iconic representations, are analogs of the direct sensory experience of an external object or concept. These correspond to the actual sensory projections made by a scene on our perceptual systems. The second type of representation, a categorical representation, is formed when the unique features of a concept category are extracted by feature detectors. These representations are still based directly on sensory experience but represent prototypes of an object or concept category. With these forms of representations in hand, abstract symbolic representations can be grounded in representations formed directly from perceptual experience. Additionally, Harnad argues that once a set of "entry-level" symbols are grounded in this fashion, "higher-level" symbols can be grounded in terms of those entry-level symbols without being directly grounded in experience. For example, once the category "horse" and the concept "striped" have been directly grounded, the category "zebra" can be grounded by learning that it corresponds to a striped horse. This ability to transfer learned meanings facilitates purely symbolic learning once a set of simple symbols have been grounded.

Harnad's formulation of the symbol-grounding problem has motivated a large number of studies examining how symbols may become grounded. Several such studies

will now be examined.

2.2 Symbol Grounding Research

Harnad proposes his own solution to the symbol-grounding problem: a combined connectionist/symbolic model (Harnad 1990, 1993). He suggests that a connectionist network is a strong candidate for the grounding of symbols and that once grounded, a symbolic system can be used to learn the meanings of complex symbols from symbolic descriptions. This type of model has been examined by several studies that have successfully demonstrated the strength of connectionist learning in the grounding task (Cangelosi, et al. 2000; Riga, et al. 2004). These models demonstrated that symbols could be grounded with connectionist networks, allowing for transfer of meaning from grounded symbols to higher level symbols. In these studies, supervised learning procedures were used to train the networks. While these models transfer from connectionist to symbolic learning after simple symbols have been grounded, others have proposed that connectionist models may be capable of learning grounding for complex concepts and even syntactic structure (Gasser 1993). It remains unclear to what extent connectionist models are capable of achieving grounding.

In the study of grounding, a task for learning the meanings of symbols from corresponding visual information has been put forth as an important challenge for cognitive science (Feldman, et al. 1990). The task is for a system to learn the meanings of symbols from pairings of visual scenes and linguistic descriptions. With minimal complexity of scenes and descriptions, this task addresses many important facets of the grounding problem. The task allows for a vast simplification of the symbol-

grounding task accomplished by human infants. However, if completed successfully, the results could provide valuable insight into how more complex grounding is accomplished.

Numerous studies have adopted this methodology for examining grounding (see Feldman, Lakoff, et al. for a review of several studies). One such study presented a model, called the DETE architecture, that learned relationships between sequences of scenes and descriptions (Nenov and Dyer 1993, 1994). The visual scenes consisted of simple shapes of varying sizes and positions. Sequences of scenes demonstrated actions such as "bounce" and "push" and the linguistic sequences were descriptions of the scenes. The model demonstrated strong grounding capabilities; it was able to learn descriptions for considerably complex visual scenes. However, the complexity of the model makes it difficult to examine exactly what meanings have been learned for different symbols. The visual scenes were also preprocessed by feature extractors, so the model did not learn directly from the scene bitmaps. The learning task used by Nenov and Dyer provided the primary motivation for the study presented in this thesis.

2.3 Conclusion

This chapter reviewed the philosophical underpinnings of the symbol-grounding problem as well as a discussion of some research efforts that have addressed it. The task of learning correspondences between visual scenes and linguistic descriptions and its significance for symbol-grounding were also discussed. The extent to which symbols can be directly grounded in perceptual information and a neurally-inspired model of how this may be accomplished is still an open question. The work

presented in this thesis proposes such a model and demonstrates how its symbol-grounding capabilities can be analyzed. Unlike previous models, it learns in a completely unsupervised fashion. Additionally, the model allows for direct examination of what has been learned, which is greatly beneficial in a study of symbol grounding.

3. Model

In this thesis, the symbol-grounding problem is approached by analyzing the grounding capabilities of an unsupervised neural network learning architecture. The task is to learn correlations between simple visual scenes and corresponding linguistic descriptions. This chapter describes in detail the design and implementation of the network architecture and learning task. The first section describes the network architecture, the second and third sections describe the visual scenes and linguistic descriptions used in the learning task, the fourth section describes self-organizing maps (SOMs), the fifth section describes the associative connections between the SOMs, and the last section summarizes the network model and learning task.

3.1 Network Architecture

The neural network model consists of two self-organizing maps (SOMs, Section 3.4), one each for the linguistic and visual inputs. The SOMs are connected to each other with many-to-many associative connections between the nodes of the maps (Figure 3.1). The model is trained to learn correspondences between scenes and descriptions by presenting complementary pairs of scenes and descriptions to both SOMs simultaneously. The inputs are used to modify the data stored in the SOMs and the associative connections between the maps are updated. After training, it is possible to present a description to the linguistic SOM, propagate through the associative connections, and generate the visual scene which the network associates with that description. Similarly, it is possible to present a visual scene and view the corresponding linguistic description or descriptions.

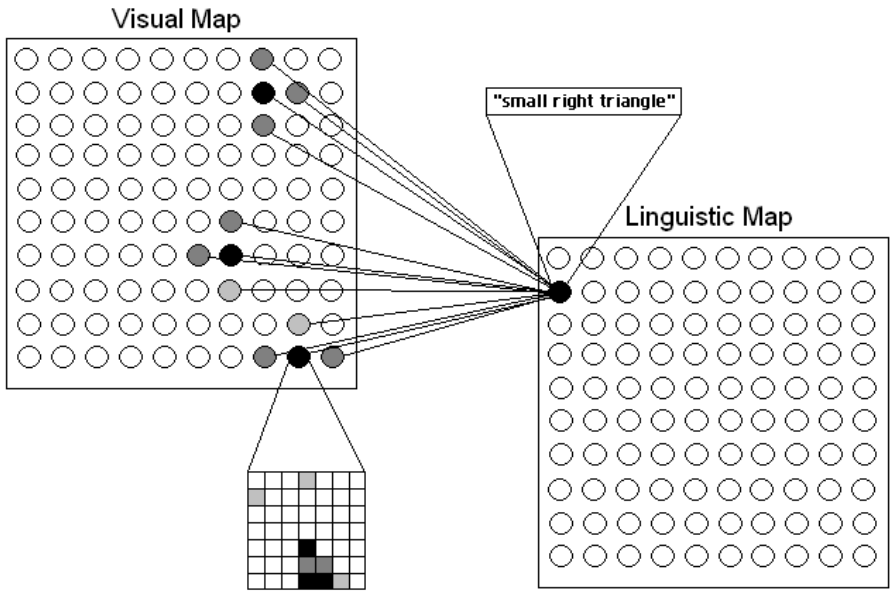

Figure 3.1: **Network architecture.** The network consists of two SOMs, a linguistic map and a visual map. The units of the maps are connected with many-to-many associative connections. Each unit of the maps contain a prototype formed from the inputs to the network. The associative connections between the units represent the learned relationships between linguistic and visual information.

Each SOM functions as a memory module for its respective input type, either visual or linguistic. Each node on a SOM has an associated representation vector of the same dimensionality as that map's input vectors. When trained, these representation vectors are modified to more closely resemble similar inputs from the training examples. In this way, nodes of the maps become prototypes of the input vectors. An additional characteristic of the SOM is that similar inputs are mapped onto topologically proximal nodes of the map. The result is that nodes on the maps that are close together contain prototypes for similar inputs. The SOM generates prototypes from the input vectors without any corrective error feedback, or in other words, it is an unsupervised learning procedure. This is a desirable characteristic

when attempting to model human developmental learning because there is little evidence that children receive clear corrective error signals from their environments. SOMs have also demonstrated wide applicability in modeling other areas of cognition, including vision and audition (Kohonen 1997).

The associative connections between nodes of the SOMs represent the learned connections between the scenes and descriptions. The connections from a visual node represent the strengths of associations between that node and possible linguistic descriptions, and the converse is true for connections from a linguistic node. In other words, the strongest connections from a visual node represent the model's best descriptions for that scene and the connections from a linguistic node represent images for that description. An important aspect of these connections for the purposes of symbol-grounding are their many-to-many connectivities. These many-to-many mappings allow for several scenes to be strongly associated with a given description, and vice versa. Such connectivity is important for grounding because it is necessary to maintain many possible associations for a given scene or description. For example, for the description "small square", the model should have strong associations for scenes containing squares in many positions, not just one. Similarly, it is possible for a given scene to be described in numerous ways, and the model should retain these various descriptions. It may be possible to describe a certain scene as "a square", "a small object", "a small object in the middle", etc., and the model should retain these descriptions. The associative connections in the model are effective for this purpose.

Network architectures similar to the one presented in this thesis have been shown to be successful at tasks that are analagous to symbol grounding. In one study, such

a model was shown to successfully learn associations between linguistic modalities (phonological and orthographic) and the semantic meanings of words (Miikkulainen 1997). The model was used successfully as a lexicon module in a larger story-processing system and also exhibited behavior similar to dyslexia when lesioned. Another study used a like architecture in modeling the acquisition of verb semantics (Li 1999). In this study, the model was shown to successfully exhibit various attributes that make it desireable for studies of language acquisition, including the abilities to generalize and form representations. Both of these studies supported the hypothesis that the model presented in this thesis would be a strong candidate for the symbol-grounding task.

3.2 Visual Inputs

The visual inputs for the model were 20×20 grayscale bitmaps, represented by 400-dimensional vectors with each component between 0 and 1. The visual inputs consisted of one-object and two-object scenes. There were 8 types of objects used in the scenes: open squares, filled squares, open diamonds, filled diamonds, left triangles, right triangles, X's and Z's (Figure 3.2. The objects varied in their sizes and positions. Scenes with two objects presented various relationships between objects: "inside of", "around", "to the left of", "to the right of", "above" and "below". These dimensions for variation provided a significantly large set of possible scenes, making the learning task considerably difficult.

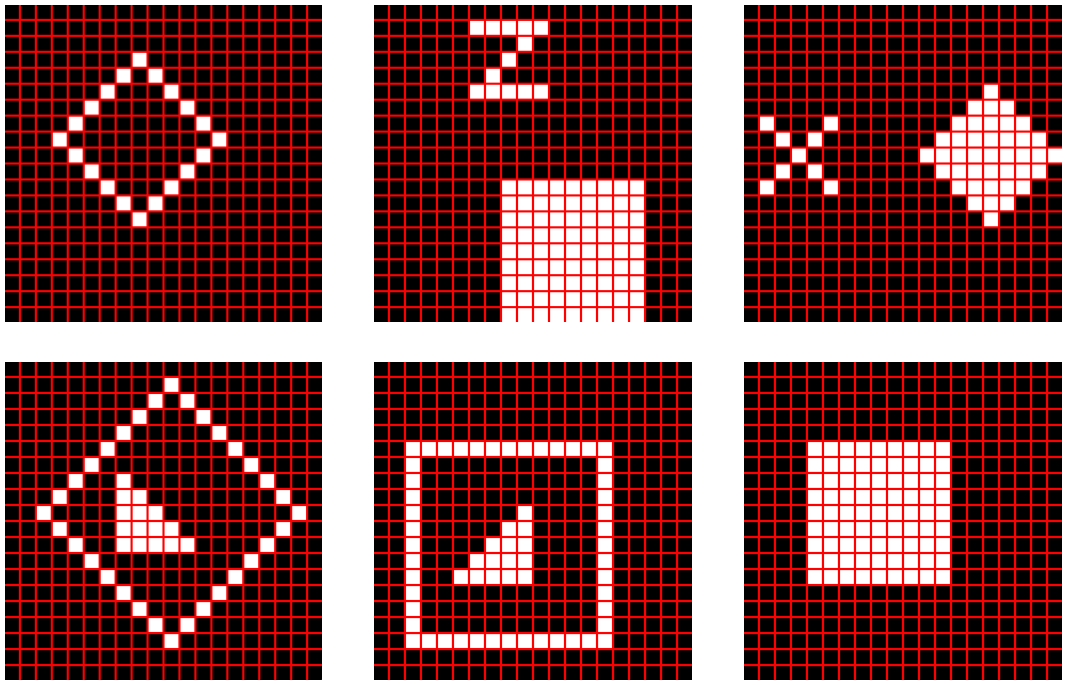

Figure 3.2: **Sample Visual Inputs.** The visual inputs were 400-dimensional vectors representing 20×20 grayscale bitmaps. The vector components were either 0 (black) or 1 (white). The scenes had either one or two objects.

3.3 Linguistic Descriptions

The linguistic descriptions for the visual scenes were generated from a 31 word vocabulary. The vocabulary contains words for describing attributes of individual objects (size, shape, and position) as well as relationships between objects. The generative rules for scene descriptions can be found in Table 3.1. A given scene may be described by a number of possible descriptions, and similarly there may be a number of different scenes given the same description. This is a key aspect of the symbol-grounding problem. There are many-to-many mappings between scenes and descriptions, as is the case in the real world. In order to successfully complete the task, it is necessary to learn the meanings of individual symbols. This

Generative Rules for Scene Descriptions
Description = **NP** | **REL**
REL = **NP relterm NP**
NP = [**size**] **object** [**position**]
object = **specobj** | "object"
relterm = "above" | "below" | "to the left of" | "inside of" | etc.
specobj = "open square" | "filled diamond" | "left triangle" | etc.
size = "small" | "medium" | "large"
position = "in the top left" | "in the middle" | "on the right" | etc.

Table 3.1: **Generating a Linguistic Description.** These rules describe the potential descriptions that may be generated for a given scene. These various descriptions for a given scene create many-to-many mappings between scenes and descriptions. Therefore, the learning task requires extracting the meanings of individual symbols rather than simply learning (scene, description) pairings.

learning task amounts to categorical perception (Section 2) and, if successfully accomplished, would allow the network to generalize the meanings of the symbols to novel situations. The linguistic descriptions were represented by 31-dimensional vectors, with each unit of the vector corresponding to a distinct word in the vocabulary. The vector components were between 0 and 1. The sequential information of the descriptions was represented by decaying the activations of the vector components linearly with respect to their positions in the sequences. This process for generating the linguistic descriptions is described in Table 3.2. This technique for representing sequential information through activation decay was inspired by the SARDNET model (James and Miikkulainen 1995).

3.4 The Self-Organizing Map

The Self-Organizing Map (SOM) is a system for unsupervised learning that maps high-dimensional input vectors onto a two-dimensional feature map (Kohonen

Procedure for Generating a Linguistic Representation
INITIALIZE all components of description vector to 0 WHILE linguistic description is not complete multiply all vector components by 0.9 set vector component for next word to 1 ENDWHILE

Table 3.2: **Generating a Linguistic Representation.** This pseudocode describes the process for creating a linguistic description vector that embodies the temporal information of the linguistic sequence. The activation for a word in the sequence is decayed linearly with respect to its position in the sequence.

1989, 1997). The input vectors contain descriptions of observations about the environment. The map consists of an array of interconnected nodes, where each node i has an associated representation vector $m_i = [\mu_i 1, \mu_i 2, ..., \mu_i n]^T \in R^n$. During training, an input vector $x = [\xi_1, \xi_2, ..., \xi_n]^T \in R^n$ is compared with the representation vectors for all map nodes in parallel and the node whose representation vector is most similar to the input vector is chosen to represent the input on the map. This node is referred to as the Best Matching Unit (BMU). A standard Euclidean distance measure was used to determine the similarity between the vectors. Thus, the BMU is determined by finding c such that:

$$c = \arg \min_i |x - m_i|. \tag{3.1}$$

Once the BMU has been determined, the map is modified by updating the BMU and the nodes in its neighborhood to reflect the new input. The reference vectors for the nodes are adjusted so that they more closely resemble the input vector. The adjustment is determined by topological distance from the BMU. The nodes are modified such that:

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t)[x(t) - m_i(t)] \tag{3.2}$$

where t is an integer discrete-time coordinate, α is the learning rate, and h_{ci} is a neighborhood function. The function h_{ci} determines the amount of modification for map nodes with respect to their distance from the BMU. The neighborhood function used for this thesis was a square area around the BMU. In other words, nodes within a square neighborhood around the BMU were adjusted towards the input vector with a uniform learning rate and nodes outside the neighborhood were left unchanged.

3.5 Associative Connections between Maps

Each node on a map, either linguistic or visual, has unidirectional connections to every node on the other map. The strengths of these connections represent the strength of associations between a given description and possible scenes, or conversely between a given scene and possible descriptions. When training samples are presented to the maps, each map node produces an activity strength. This strength is proportional to the similarity between the input vector and the node's representation vector, as determined by Eudlidean distance. The associative connections between map nodes are then adjusted with respect to their activity strengths, using Hebbian learning (Hebb 1949). The connection between two nodes is strengthened proportional to their activity levels:

$$\Delta w_{ij,uv} = \alpha(t) n_{S,ij} n_{D,uv}, \tag{3.3}$$

where $w_{ij,uv}$ is the unidirectional weight between the source map node at location (i, j) and the destination node at location (u, v), and $n_{S,ij}$ and $n_{D,uv}$ represent the activations of these units, respectively. This serves to strengthen the connections between nodes that are simultaneously active. The associative weight vectors are

then normalized, which serves to decrease the strengths of connections to inactive units. In this way, the model is able to learn cooccurrence relationships between nodes on the different maps.

3.6 Conclusion

The network architecture described in this chapter uses an unsupervised learning procedure to learn relationships between visual scenes and linguistic descriptions. The architecture uses self-organizing maps as memory and prototype-formation modules and associative connections between the maps are trained with Hebbian learning. The associative connections represent the learned relationships between scenes and descriptions. The technique for encoding visual scenes and linguistic descriptions as feature vectors was also described.

4. Experiment and Results

In this chapter, an analysis of the network's symbol-grounding capabilities is presented. The first section describes the protocol that was used to train the network. The second section describes how the model was tested and its performance was measured. The final section presents and discusses the results from experimentation with the model.

4.1 Training Procedure

The network was trained for 2000 epochs with 2500 (scene, description) training pairs. The training pairs consisted of 50% one-object scenes and 50% two-object scenes. The descriptions were generated so that size, shape, and position information was included in 80% of the samples. This was done in hopes that it would expedite the network's learning. The idea was that the network can learn more from a description such as "small open square on the left" than it can from the description "object", which lacks any meaningful information. The training pairs were presented in random order. The same learning rate $\alpha(t)$ was used for both maps and the associative connections. The learning rate was decreased linearly from 0.1 to 0.05 over the first 500 epochs and then decreased linearly to 0 during the remaining epochs. At the same time, the neighborhood size for both maps decreased linearly from 4 to 1 and then from 1 to 0.

4.2 Testing Procedure

One strong attribute of the model presented in this thesis is that it facilitates several forms of testing to assess what the network has learned. It is possible to test the network by examining its ability to produce an appropriate scene when given a description or vice versa. The former corresponds more with language generation capabilities and the later is a more effective examination of symbol-grounding, and so for the purposes of this thesis the later measure was used. The network was presented with descriptions and its ability to produce appropriate scenes was analyzed. However, both forms of examination are useful in assessing the network's performance, as grounding is necessary for both directions, and so analysis of the former type is planned for the future.

During testing, a description was presented to the linguistic map and the BMU was determined. The associative connections for the BMU were then displayed. Here again there were different options for analyzing the network's performance. Either the network's strongest association for the BMU could be chosen as the response or an average of all responses weighted by their associative connections could be formed as a composite response. Both ways of interpreting the results offered valuable insight into the learning accomplished by the network. For the purposes of this thesis, responses are determined by the strongest associative connection from the BMU.

Although the model presents multiple ways of analysis, interpreting its results and performance poses a difficult challenge. When judging how appropriate a visual scene is for a given description there is no quantitative measure of performance that

is readily apparent. Judging performance when generating descriptions of scenes, while potentially easier to analyze, still presents a nontrivial task. For each scene, the network will, hopefully, generate numerous descriptions. This generation of multiple descriptions is desired because for any given scene there are numerous potential descriptions that are accurate. In analyzing the network's responses, it is unclear which valid descriptions the responses should be compared with and how this similarity should be measured. Therefore, both ways of analyzing the network's performance present significant challenges to deriving quantitative measures, and formulating such measures is a future goal for this project. At present, the analysis used is to grade the network's visual responses subjectively. Each response was assigned a score from 0 to 1 in increments of 0.1 based on how appropriate it was for the description. These scores were determined using visual inspection of the scenes. Sample responses and their corresponding grades are presented in Figure 4.1.

In testing, the network was presented with three groups of stimuli: simple symbols, complex descriptions, and novel descriptions. The first group, simple symbols, consisted of individual words from the network's vocabulary. In other words, the network was tested with inputs such as "small", "square", and "to the right of". This form of testing examined the network's grounding of individual concepts. The second testing set, complex descriptions, was composed of examples from the network's training corpus. This form of testing examined how well the network had learned the information it was given. For the final testing set, the network was presented with descriptions which it had not seen in training. This testing set examined the network's ability to generalize to novel stimuli.

4.3 Results and Analysis

For each of the three testing sets, 30 samples were presented. The means and standard deviations for the scores were calculated and were as follows:

Simple Symbols: μ = 0.48, σ = 0.31

Complex Descriptions: μ = 0.62, σ = 0.23

Novel Descriptions: μ = 0.25, σ = 0.14.

These results indicate that the network performed best on examples which were in its training set, as is to be expected. The network also learned the meanings of simple symbols fairly well and its performance on novel descriptions indicate that the network was somewhat capable of generalizing the meanings it had learned. The evaluation of the network's performance was subjective, but as a preliminary examination it indicates that the network may be capable of successfully modeling symbol-grounding. Additionally, the model can also give insights into the mechanisms of grounding.

Visual inspection of the network after training also provides valuable insight into what it has learned. For example, for simple symbols the network has multiple groups of strong associative connections. For the description "small", several strong groups of activations are present, each containing images of different small objects in various positions. Similar mappings from a symbol to several groups of scenes exist for other atomic symbols and is a desirable result because it indicates that the network retains the meanings of simple symbols in different scene contexts. Observations like this are hard to encapsulate in a quantifiable metric of the network's performance. However, they can offer valuable insight into how well the network

is performing symbol-grounding.

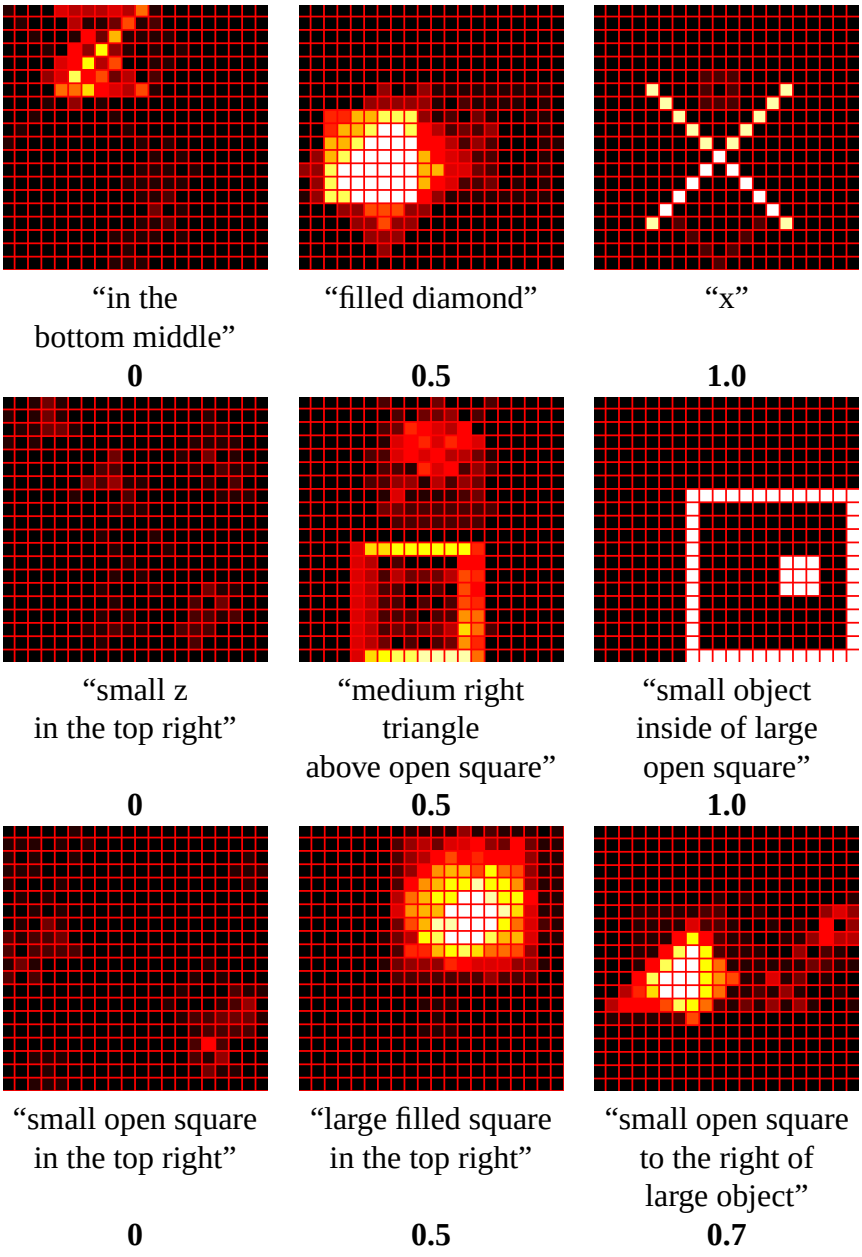

Figure 4.1: **Sample Scorings.** The scenes generated by the network were assigned scores between 0 and 1 in increments of 0.1 based on their relevance to the description. Sample scenes, descriptions, and scores are shown for each of the three test sets: simple, complex, and novel descriptions. The top three images are from the simple test set, the middle three are from the complex test set, and the bottom three are from the novel test set.

5. Future Work

In this thesis, a neural network model for learning correspondences between visual scenes and linguistic descriptions was presented and its symbol-grounding capabilities were examined. This chapter discusses several directions for future work that would further analyze the model and facilitate extension of the architecture to allow for examination of more complex problem domains.

5.1 Analysis of Network Performance

As was described in section 4.2, there are numerous ways in which the network's analysis could be extended. The most direct extension would be to analyze the network's performance when presented with scenes and generating descriptions. Combining this with the current form of analysis would allow the network's performance at symbol-grounding to be assessed bidirectionally. It may also be possible to formulate meaningful quantitative measures of the network's performance, either in one direction or both. Ideally, this measure could be automated and the network's performance could be examined over a large testing corpus. This would be superior to the current technique of visual inspection which limits the number of testing samples that can be analyzed.

5.2 Comparison With Child Language Acquisition

If the network's performance when generating descriptions of scenes could be meaningfully quantified, an interesting future study would be to examine the network's performance at language learning relative to results from developmental studies of

child language acquisition. For example, the network could be analyzed at different discete time points during its training to determine how well it had learned different concepts at specific times. These results could then be analyzed to see if the network exhibits observed phenomenon from child language studies, such as the over- and undergeneralization of the meanings of words. Such a study could serve to validate or reject the network as a cognitively valid model.

5.3 Comparison with Other Learning Models

The network's efficacy at acquiring an understanding of language could be compared to the performances of other models applied to similar learning tasks. For example, it would be interesting to compare the current model with models that use different learning procedures, such as backpropagation, to examine their relative performances. A model based on Recursive Auto-Associative Memory (RAAM; Pollack 1988) was created and tested as part of a preliminary examination of the grounding task. Several shortcomings of the RAAM model were identified, such as an inability to retain many-to-many mappings and reliance on corrective error-feedback. These shortcomings motivated the design of the model presented in Chapter 3. However, a quantitative analysis of the relative performances of the RAAM-based network and the model presented in Chapter 3 is left for future work. It would be interesting to see which network exhibited the best performance overall and whether the different architectures were better suited for learning certain types of concepts.

5.4 Representing Sequential Information

If a strong representation for sequential information was used it would be possible to vastly increase the complexity of the scenes and descriptions and push the limits of the model. If the current representation of sequential information in the descriptions was modified, the network could be examined with complex grammatical constructs. If sequential information could also be well represented for visual scenes it would be possible for the model to learn from sequences of images rather than stills, which would increase the complexity of the grounding task. The model could attempt to learn verbs and changes in object states over time. One possibility for efficiently representing sequential information in both the visual and linguistic input domains would be to create SARDNET encodings of the input sequences and then present those encodings to the SOMs(James and Miikkulainen 1995). These encodings of complex scenes and descriptions could be used to examine the limits of what can be effectively grounded.

6. Conclusion

This thesis presented a model with which to explore the issue of how abstract symbols may become grounded directly in perceptual information. More specifically, an unsupervised learning architecture that learned relationships between visual scenes and linguistic descriptions was presented and then used to explore symbol grounding.

Chapter 2 reviewed the philosophical ideas behind the symbol-grounding problem. Several studies which sought to address this problem were discussed, including the research which inspired the work in this thesis.

Chapter 3 presented and described the network model which was the focus of this thesis. The network consisted of two self-organizing maps with associated connections between them. The implementation of the learning task was also described. The task consisted of learning correspondences between simple visual scenes and corresponding linguistic descriptions.

Chapter 4 discussed the experimental procedure that was used to examine the network and analyze the network's performance. The network was trained with (scene, description) pairs and then examined to see how well it had learned the meanings of simple symbols, complex descriptions, and novel descriptions. The network was analyzed by visually inspecting the generated scenes and assessing how relevant they were for the test descriptions. The results from this analysis indicate that the network learned its training set well and had accomplished some level of grounding. The difficulties with examining the network's performance were also described.

Chapter 5 proposed several possible extensions to the current analysis of the model. The derivation of strong quantitative measures of the network's performance would facilitate more detailed analysis of how well it accomplished its learning task. Additionally, comparison of the model to results from other studies could examine its strength as a model of language learning and child language acquisition. Representations for visual and linguistic sequences could allow the model to be examined with increasingly complex learning tasks.

Together, these chapters presented a model which provides a strong platform for examinations of the symbol-grounding problem. The model learns associations between linguistic descriptions and visual scenes and allows examining what it has learned. Preliminary results suggest that the model effectively learns associations and may be capable of performing generalizations with the grounded meanings of symbols.

Bibliography

Barsalou, L. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences, 22*, 577-609.

Cangelosi, A., Greco, A., & Harnad, S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science, 12(2),* 143-162.

Feldman, J.A., Lakoff, G., Bailey, D.R., Narayanan, S., Regier, T. & Stolcke, A. (1996). L_0–The first five years of an automated language acquisition project. *Artificial Intelligence Review, 8*.

Feldman, J.A., Lakoff, G., Stolcke, A., & Weber, S.H. (1990). Miniature Language Acquisition: A Touchstone for Cognitive Science. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 686-693.

Gasser, M. (1993). The Structure Grounding Problem. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 149-152.

Harnad, S. (ed.) (1987). *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D, 42*, 335-346.

Harnad, S. (1993). Grounding Symbols in the Analog World with Neural Nets. *Think, 2*, 12-78.

Hebb, D.O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

James, D.L. & Miikkulainen, R. (1995). SARDNET: A self-organizing feature map for sequences. *Advances in Neural Information Processing Systems, 7*, 577-584.

Kohonen, T. (1989). *Self-Organization and Associative Memory*. New York:Springer. Third Edition.

Kohonen, T. (1997). *Self-organizing maps*. Berlin: Springer-Verlag.

Li, P. (1999). Generalization, Representation, and Recovery in a Self-Organizing Feature-Map Model of Language Acquisition. In M. Hahn & S.C. Stoness (Eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 308-313) Mahwah, NJ: Lawrence Erlbaum.

Miikkulainen, R. (1997). Dyslexic and Category-Specific Aphasic Impairments in a Self-Organizing Feature Map Model of the Lexicon. *Brain and Language, 59*, 334-366.

Nenov, V.I. & Dyer, M.G. (1993). Perceptually Grounded Language Learning: Part 1 – A Neural Network Architecture for Robust Sequential Association. *Connection Science, 5 (2)*, 115-138.

Nenov, V.I. & Dyer, M.G. (1994). Perceptually Grounded Language Learning: Part 2 – DETE: A Neural/Procedural Model. *Connection Science, 6 (1)*, 3-41.

Pollack, J.B. (1988). Recursive auto-associative memory: Devising compositional distributed representations. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 33-39) Hillsdale, NJ: Lawrence Erlbaum.

Riga, T., Cangelosi, A., & Greco, A. (2004). Symbol Grounding Transfer with Hybrid Self-Organizing/Supervised Neural Networks. *IJCNN04 International Joint Conference on Neural Networks*. Budapest, July 2004

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3 (3)*, 417-457.