Grounding Language in Descriptions of Scenes

Paul Williams (pwilly@cs.utexas.edu)
Department of Computer Sciences, The University of Texas at Austin
1 University Station C0500, Austin, Texas 78712 USA

Risto Miikkulainen (risto@cs.utexas.edu)
Department of Computer Sciences, The University of Texas at Austin
1 University Station C0500, Austin, Texas 78712 USA

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 paper presents the GLIDES model, a neural network architecture that shows how this symbol-grounding problem can be solved through learned relationships between simple visual scenes and linguistic descriptions. Unlike previous models of symbol grounding, the model's learning is completely unsupervised, utilizing the principles of self organization and Hebbian learning and allowing direct visualization of how concepts are formed and grounding occurs. Two sets of experiments were conducted to evaluate the model. In the first set, linguistic test stimuli were presented and the scenes that were generated by the model were evaluated as the grounding of the language. In the second set, the model was presented with visual test samples and its language generation capabilities based on the grounded representations were assessed. The results demonstrate that symbols can be grounded based on associations of perceptual and linguistic representations, and the grounding can be made transparent. This transparency leads to unique insights into symbol grounding, including how many-to-many mappings between symbols and referents can be maintained and how concepts can be formed from cooccurrence relationships.

Introduction

In order to create an intelligent symbol system, symbols must be grounded in perceptual information (Harnad, 1990; Barsalou, 1999). 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 to have achieved understanding. For understanding to occur, the symbols must have inherent meaning in terms of the system's experiences of the external world. In order to develop a symbol system, it is therefore necessary to understand how symbols become grounded in their perceptual correlates (Cottrell, Bartell, & Haupt, 1990; Chalmers, 1992).

Technically, symbol grounding means establishing perceptual categories and associating these categories with abstract tokens. In order to do that, it is first necessary to determine the commonalities of all the external objects to which a symbol refers that are distinct from attributes of objects in other categories. This process involves emphasizing the differences between categories and minimizing the differences within categories, a process called "categorical perception" (Harnad, 1987). Once the boundaries of a category have been established, it can be associated with an abstract token, at which point symbol grounding has occurred. 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.

Artificial neural network (ANN) architectures provide strong candidates for computational models of symbol grounding. Several such architectures have been previously proposed, as will be reviewed in the following section. While these models have provided many insights, the problem is by no means solved. First, most of these models are based on supervised learning, utilizing corrective feedback. Assuming that symbol grounding is a developmental cognitive process, it is unclear what the source of the error signals might be. Second, the previous models are often opaque, i.e. difficult to interpret. A model of symbol grounding should ideally do more than simply show that grounding can be achieved; it should demonstrate how the grounding occurs and what grounding looks like on a conceptual level.

This paper presents the GLIDES (Grounding Language in DEscriptions of Scenes) model, a neural network architecture that learns to ground linguistic descriptions into visual scenes. The model uses an unsupervised learning procedure based on self-organizing maps and Hebbian adaptations, learning associations between descriptions and scenes. It allows directly examining the representations and associations that are formed from the linguistic and visual inputs. The model therefore provides a unique framework for studying the grounding task.

GLIDES was evaluated in two sets of experiments. The first set assesses the model's symbol grounding by evaluating the scenes it generates for linguistic test inputs of three types: (1) single words/concepts, (2) complex descriptions present in the training set, and (3) complex novel descriptions. The second set validates the model's grounding by evaluating its language generation ability when describing visual samples from two test sets: (1) scenes from the training set and (2) novel scenes. The results demonstrate unique insights into symbol grounding, including how many-to-many mappings between symbols and referents can be maintained and how concepts can be formed from cooccurrence relationships.

Prior Grounding Research

An early solution to the symbol-grounding problem was proposed by Harnad (1993), a combined connectionist/symbolic model trained by supervised learning. Similar models have been used in several studies since, successfully demonstrating the strength of connectionist learning in the grounding task (Cangelosi, Greco, & Harnad, 2000; Riga, Cangelosi, &

Greco, 2004). Symbols can be grounded with connectionist networks, allowing for transfer of meaning from grounded symbols to higher level symbols. Others have proposed that connectionist models may be capable of learning grounding for complex concepts and even syntactic structure (Gasser, 1993).

A valuable touchstone task for the symbol-grounding problem was proposed by Feldman, Lakoff, Stolcke, & Weber (1990). The task is 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 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, several of which are discussed by Feldman, Lakoff, Bailey, Narayanan, Regier & Stolcke (1996). One of the most compelling models in this line of research is the DETE architecture, an ANN model that learns relationships between sequences of scenes and descriptions (Nenov & Dyer, 1993; 1994). While this model learned impressive performance, it was difficult to understand how it developed conceptual representations, i.e. how its symbols became grounded.

In sum, although previous models have shown that grounding is possible, it has been difficult to show how exactly it is achieved and what the grounded representations look like. This paper presents a new model with the goal of providing such an account.

The GLIDES Model

GLIDES is a neural network architecture that accomplishes symbol grounding by learning correlations between visual scenes and linguistic descriptions. The model consists of two memory modules, one each for the linguistic and visual modalities, as well as associative connections between the two that store learned relationships (Figure 1).

This design was inspired by similar architectures that have been used for tasks analogous to symbol grounding. In one study, such an architecture was shown to successfully learn associations between linguistic modalities (phonological and orthographic) and the semantic meanings of words (Miikkulainen, 1997). Another study used a like architecture in modeling the acquisition of verb semantics (Li, 1999). These models were able to form robust prototypes and learn meaningful associations between different modalities such that they could later be used for generalization. The architectures also allow direct inspection of what prototypes are formed and how these prototypes are associated with each other. These properties are precisely what are needed to accomplish grounding as well, as will be discussed in this section.

Network Architecture

The GLIDES architecture consists of two memory modules, one each for the linguistic and visual modalities, with associative connections between them. The memory modules are implemented as self-organizing maps. A self-organizing map (SOM) is a system for unsupervised learning that maps high-dimensional input vectors onto a two-dimensional feature map (Kohonen, 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 is used to determine the similarity between the vectors. 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:

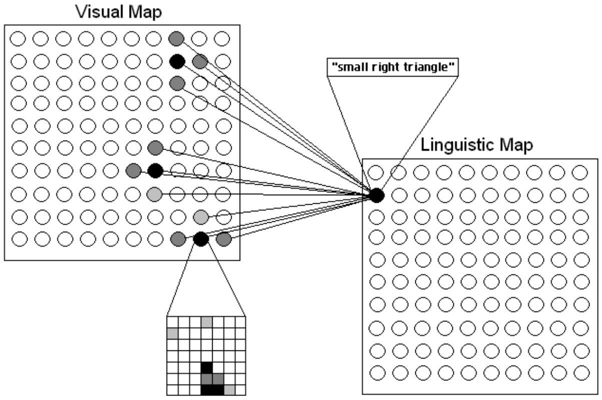


Figure 1: The network consists of two SOMs, a linguistic map and a visual map. The units of the maps are connected with many-to-many associative connections. Each unit of the maps contains a prototype formed from the inputs to the network. The associative connections between the units represent the learned relationships between linguistic and visual information. The network is able to form prototypes and learn associations from inputs in an unsupervised fashion.

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

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 is a square area around the BMU; nodes within a square neighborhood around the BMU are adjusted towards the input vector with a uniform learning rate.

When the map is trained, nodes of the maps become prototypes of the input vectors. Additionally, similar inputs are mapped onto topologically proximal nodes. The result is that nodes on the maps that are close together contain prototypes for similar inputs. SOM models often lead to cognitively valid behavior, as has been demonstrated in several areas of cognition, including vision, audition, memory, decision making, and language processing (Oja, Kaski, & Kohonen, 2001).

In GLIDES, the two SOMs are connected to each other with many-to-many associative connections. Each node on one map has unidirectional connections to each 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 Euclidean 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:

$$\Delta w_{ij,uv} = \alpha(t) n_{S,ij} n_{D,uv}, \tag{2}$$

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. Thus, the connections between simultaneously active nodes are strengthened. 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.

The model is trained by presenting complementary [scene, description] pairs to both SOMs simultaneously. The data stored in the SOMs is modified based on the inputs 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.

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 (Figure 2). 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. The objects varied in their sizes and positions. The size of an object was described as either "small", "medium", or "large". There were 13 possible descriptions for the position of an object: "in the top left", "in the top middle", "in the top right", "in the middle left", "in the middle", "in the middle right", "in the bottom left", "in the bottom middle", "in the bottom right", "on the left", "on the right", "on the top", and "on the bottom". 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.

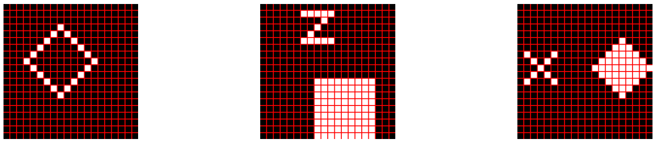


Figure 2: The visual inputs were 400-dimensional vectors representing 20×20 grayscale bitmaps. The vector components were either 0 (black) or 1 (white). The scenes had either one or two objects. These inputs provided the model with a significantly complex visual domain for grounding.

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 are shown in Table 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 complex mapping 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 learning task amounts to categorical perception and, if successfully accomplished, allows the network to generalize the meanings of the symbols to novel situations.

Table 1: Generative Rules for Scene Descriptions

Description = **NP** | **REL**
REL = **NP relterm NP**
NP = [**size**] **object** [**position**]
object = **specobj** | "object"
relterm = "above" | "to the left of" | "inside of" | ...
specobj = "open square" | "filled diamond" | ...
size = "small" | "medium" | "large"
position = "in the top left" | "on the right" | ...

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 technique for representing sequential information through activation decay was inspired by the SARDNET model (James & Miikkulainen, 1995).

Experiments

The GLIDES model was first trained with a corpus of [scene, description] pairs and its symbol grounding and language generation abilities were then evaluated. This section describes the procedure for training the model, the experiments conducted with the trained model, and the results.

Training Procedure

The network was trained for 2000 epochs with 2500 training pairs. The 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. 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.

Testing Procedure

During testing, a test stimulus (either linguistic or visual) was presented to its appropriate map and the best matching unit (BMU) was determined. The associative connections for the BMU were then displayed. The strongest associative connection for the BMU was propagated through and the corresponding unit on the other map was considered the network’s response. The network’s responses were then scored subjectively based on their relevance for the given test stimulus, as will be discussed for each experiment below.

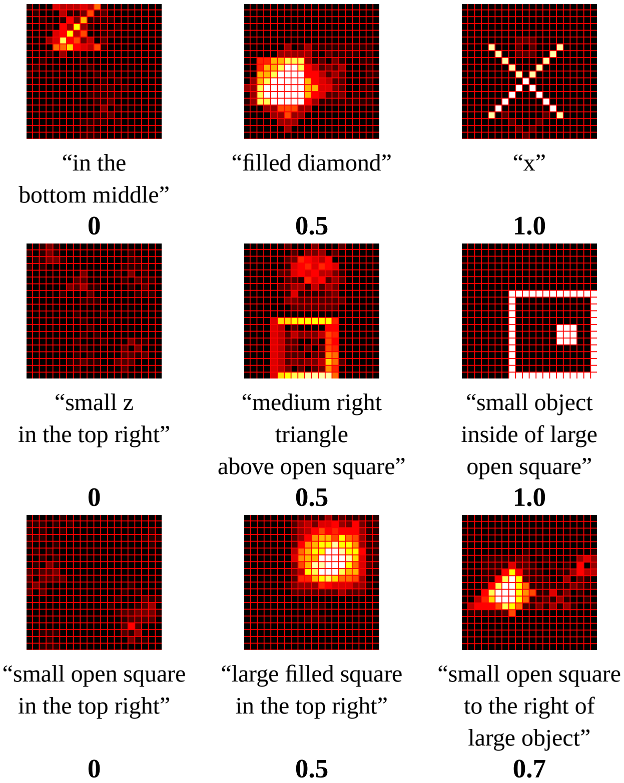
Experiment 1: Symbol Grounding

In order to evaluate the symbol grounding in the model, linguistic test samples were presented and the scenes that the network generated in response were evaluated. Because visual scenes are highly variable, there is no mechanical procedure that could be used to automatically judge how appropriate a visual scene is for a given description. Thus, the responses can only be scored subjectively. If done systematically, however, such scoring can provide useful information about the performance of the system. Therefore, each scene was assigned a score from 0 to 1 in increments of 0.1 based on how appropriate it was for the description. Sample responses and their corresponding scores are presented in Figure 3 to provide a sense for this system.

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. 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 third testing set, the network was presented with complex descriptions which it had not seen in training. This testing set examined the network’s ability to generalize to novel stimuli.

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

Simple Symbols: μ = 0.48, σ = 0.31
Complex Descriptions: μ = 0.62, σ = 0.23
Novel Descriptions: μ = 0.25, σ = 0.14.



“in the bottom middle”
0

“filled diamond”
0.5

“x”
1.0

“small z in the top right”
0

“medium right triangle above open square”
0.5

“small object inside of large open square”
1.0

“small open square in the top right”
0

“large filled square in the top right”
0.5

“small open square to the right of large object”
0.7

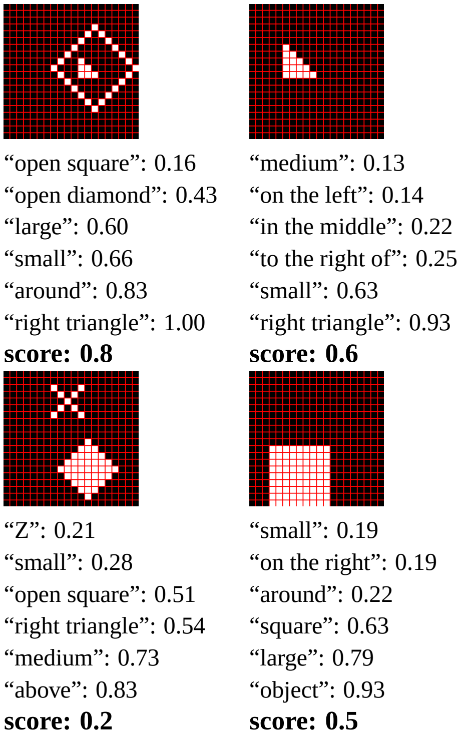
Figure 3: The scenes generated by the network were assigned scores between 0 and 1 in increments of 0.1 based on their relevance to the description. Sample scenes, descriptions, and scores are shown for each of the three test sets: simple, complex, and novel descriptions. The top three images are from the simple test set, the middle three are from the complex test set, and the bottom three are from the novel test set. These samples provide a sense of the scoring system that was used, as well as the level of performance achieved.

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 well. Its performance on novel descriptions indicate that the network was somewhat capable of generalizing the meanings it had learned, as indicated by the examples in Figure 3. The evaluation of the network’s performance was by necessity subjective, but indicates that the network is capable of accomplishing symbol grounding.

Experiment 2: Language Generation

The second set of experiments analyzed the ability of the GLIDES model to generate linguistic descriptions when presented with scenes as input. The idea was to test whether the grounding was functionally adequate. If grounding was sufficiently established the model should be able to apply its knowledge about word meanings to describe novel scenes. The model was presented with samples from two test sets: (1) scenes from the training set and (2) novel scenes.

Again there is no straightforward mechanistic technique for evaluating how appropriate the linguistic responses are.



"open square": 0.16
"open diamond": 0.43
"large": 0.60
"small": 0.66
"around": 0.83
"right triangle": 1.00
score: 0.8

"medium": 0.13
"on the left": 0.14
"in the middle": 0.22
"to the right of": 0.25
"small": 0.63
"right triangle": 0.93
score: 0.6

"Z": 0.21
"small": 0.28
"open square": 0.51
"right triangle": 0.54
"medium": 0.73
"above": 0.83
score: 0.2

"small": 0.19
"on the right": 0.19
"around": 0.22
"square": 0.63
"large": 0.79
"object": 0.93
score: 0.5

Figure 4: The descriptions generated by the model were assigned scores between 0 and 1 in increments of 0.1 based on their relevance to the visual input. The top two scenes are images from the training set and the bottom two are novel test samples. The activation strengths are listed for all those that were over 0.10. These samples provide a feel for the scoring system and performance of the network.

For each scene, the network should ideally retain numerous descriptions because many such descriptions can be deemed appropriate. It is therefore unclear what "correct" description a given response should be compared against. Even if it were possible to discern what the desired answer should be, the encoding of sequential information in the descriptions additionally complicates the scoring process. It does not suffice to directly compare activation strengths of corresponding units in the response description and the "correct" description because decreasing activation strength is used to encode sequential information. Therefore, the linguistic response were scored similarly to the visual responses, using a systematic subjective valuation based on appropriateness for a given test scene. The scoring system assigned values between 0 and 1 in increments of 0.1, taking into consideration the activation strengths for appropriate words in describing a given scene and weighing them against strong activations for inappropriate words. Additionally, if the sequence indicated by the activation values was meaningful, the score was improved 20%.

Due to the large number of possible scenes and computational constraints with training the system, the training set only covered a small portion of possible images. In order to account for the poverty of training stimuli, the novel test samples were generated so that they overlapped with at least ten samples in the training set in at least two of the size, shape and position dimensions. The model was presented with 30 test samples from each testing group: training samples and novel samples. To provide a sense of the scoring system, sample test scenes, descriptions, and their corresponding scores are presented in Figure 4. The means and standard deviations for the scores were:

Training Samples: $\mu = 0.64$, $\sigma = 0.24$
Novel Samples: $\mu = 0.32$, $\sigma = 0.13$.

The performance, together with samples in Figure 4, shows that the model learned the training data well and was somewhat capable of generalizing to scenes resembling those in the training corpus. These results provide evidence that the model can ground meanings and, given sufficient training data, generalize them to novel situations.

Insights on Grounded Representations

Direct examination of representations and associations learned by the GLIDES model leads to interesting insights into how grounding is attained. One such insight is the way in which the many-to-many mappings between symbols and referents is retained by the model. The associative connections for a certain concept, such as "large square", have strong links to scenes containing large squares in different positions (Figure 5). Another insight concerns how concepts can be learned from their cooccurrence. The visual responses for various linguistic inputs show clearly the information that was extracted from the training data and grounded in the concepts. As the description of a concept becomes more specific, the associations similarly narrow in scope and the grounded image becomes more precise. In this way, the model retains concepts of both a coarse and fine granularity. Insights such as these have been difficult to obtain with previous models of grounding, which did not have the transparent representation of categories and associations that GLIDES has. Such observations may prove valuable in future work in building grounded systems.

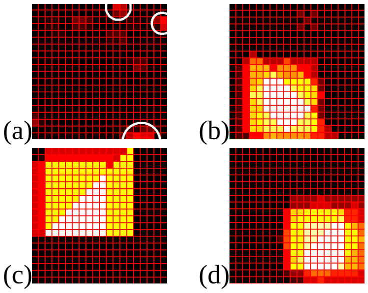


Figure 5: Image (a) displays the associative connections for the linguistic input "large square" with the three strongest areas of activation identified. Images (b), (c), and (d) display the strongest scenes stored for each of these activation areas. This figure demonstrates how the model is able to retain many-to-many mappings between symbols and referents. Such visualizations have been difficult to do with previous models but the two-map architecture of GLIDES makes them explicit.

Discussion and Future Work

The GLIDES model is a neural network architecture that demonstrates how symbol grounding can be accomplished by learning relationships between visual scenes and linguistic descriptions. The model learns in an unsupervised fashion and allows inspecting concepts and mappings that it has learned. This ability provides a unique perspective on the grounding problem that has been difficult to achieve with previous models. The architecture provides a potential platform for further investigations of symbol grounding and early language acquisition.

A possible direction for future work is to compare the learning in the model to child language acquisition. For example, the network can be analyzed at different times during its training to determine how well it has learned different concepts. These results can then be analyzed to see if the network exhibits observed phenomena from child language studies, such as the over- and undergeneralization of the meanings of words. Such a study could serve to verify or falsify the network as a cognitive model.

Another possible extension is to implement a more robust representation for sequential information. Such an extension would allow more complex scenes and descriptions to be represented and more complex phenomena to be studied, such as complex grammatical constructs and moving objects. 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 & Miikkulainen, 1995). Such encodings of complex scenes and descriptions could be used to identify the limits of what can be effectively grounded in perceptual input and what can be more effectively represented as higher-level symbolic constructs.

Conclusion

The GLIDES model provides a way for accomplishing the grounding task in a straightforward and explicit manner. The model learns to associate linguistic descriptions and visual scenes in an unsupervised process, which results in transparent representations for grounded symbols. Such transparency provides unique insights into the grounding process, and can serve as a foundation for future psychological studies of grounding as well as implementations of grounded artificial systems.

References

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.

Chalmers, D. J. (1992). Subsymbolic computation and the Chinese room. In J. Dinsmore (Ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap*. Hillsdale, NJ: Lawrence Erlbaum

Cottrell, G.W., Bartell, B., & Haupt, C. (1990). Grounding Meaning in Perception. *Proceedings of the German Workshop on Artificial Intelligence (GWAI)* (pp. 307–321).

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*, *10*(1-2), 103–129.

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* (pp. 686–693). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gasser, M. (1993). The Structure Grounding Problem. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 149–152). Hillsdale, NJ: Lawrence Erlbaum Associates.

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. *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 308–313). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Oja, M., Kaski, S., & Kohonen, T. (2001). Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum. *Neural Computer Surveys*, *3*, 1–156.

Riga, T., Cangelosi, A., & Greco, A. (2004). Symbol Grounding Transfer with Hybrid Self-Organizing/Supervised Neural Networks. *IJCNN04 International Joint Conference on Neural Networks* (pp. 686–693). Budapest, July 2004.