Bio-inspired Model of Spatial Cognition

Michal Vavrečka[1], Igor Farkaš[2], and Lenka Lhotská[1]

[1] Faculty of Electrical Engineering
Czech Technical University, Prague
{vavrecka,lhotska}@fel.cvut.cz
[2] Faculty of Mathematics, Physics and Informatics
Comenius University, Bratislava
farkas@fmph.uniba.sk

Abstract. We present the results of an ongoing research in the area of symbol grounding. We develop a biologically inspired model for grounding the spatial terms that employs separate visual *what* and *where* subsystems that are integrated with the symbolic linguistic subsystem in the simplified neural model. The model grounds color, shape and spatial relations of two objects in 2D space. The images with two objects are presented to an artificial retina and five-word sentences describing them (e.g. "Red box above green circle") with phonological encoding serve as auditory inputs. The integrating multimodal module is implemented by Self-Organizing Map or Neural Gas algorithms in the second layer. We found out that using NG leads to better performance especially in case of the scenes with higher complexity, and current simulations also reveal that splitting the visual information and simplifying the objects to rectangular monochromatic boxes facilitates the performance of the *where* system and hence the overall functionality of the model.

Keywords: self-organization, categorization, symbol grounding, spatial relations, linguistic description.

1 Introduction

The core problem of embodied cognitive science is how to ground symbols to the external world. We are looking for a system interacting with the environment that is able to understand its internal representations which should preserve constant attributes of the environment, store them as concepts, and connect these to the symbolic level. This approach to the meaning representation is different from the classical symbolic theory based on formal semantics of truth values, which cannot guarantee correspondence of the symbolic level with the external world.

In this article we propose an extended version of the classical grounding architecture [1] that implements the multimodal representations in the framework of the perceptual symbol system proposed by Barsalou [2]. The main innovation is the processing of symbolic input by a separate auditory subsystem and the integration of auditory and visual information in a multimodal layer that

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part I, LNCS 7062, pp. 443–450, 2011.

incorporates the process of identification of symbols with concepts. Our theory is similar to grounding transfer approach [3] but unlike it, our model works in a fully unsupervised manner.

Our approach was tested in the area of spatial cognition. In our models, we consider the evidence that the information about the location and identification of an object in space are processed separately. Studies with humans [4] revealed two separate pathways involved in processing of visual and spatial information: The dorsal *where* pathway is assumed to be responsible for spatial representation of the object location, while the ventral *what* stream is involved in object recognition and form representation.

In our previous experiment [5] we compared two versions of the visual subsystem, analyzing the distinction between *what* and *where* pathways, by proposing different ways how to represent object features (shape and color) and object position (in a spatial quadrant). Model I contained a single self-organizing map (SOM; [6]) that learned to capture both *what* and *where* information. Model II consisted of two SOMs for processing *what* information (foveal input) and *where* information (retinal input). Comparison of both models confirmed the effectiveness of separate visual processing of shape and spatial properties that led to a significant decrease of errors in the multimodal layer.

Both models assume the existence of the higher layer that integrates the information from two primary modalities. This assumption makes the units in the higher layer bimodal (i.e. they can be stimulated by any of the primary layers) and their activation can be forwarded for further processing. Bimodal (and multimodal) neurons are known to be ubiquitous in the association areas of the brain [7]. See also discussion in [5] for the relation of our model to several other connectionist models.

2 Motivation

The first goal of experiments presented here is the more detailed analysis of the information processing in the *where* system. We tested two types of inputs for this subsystem, namely full retinal images projected to *where* system (the same as previously) and simplified version of retinal projections (the color information was omitted and the object shapes were simplified to rectangular monochromatic boxes. The results should help us decide, whether this simplification is important for enhancing the overall model performance.

In [5] we also identified the difference in the effectiveness of the SOM comparing to Neural Gas (NG) algorithm [8] in the multimodal layer in favour of lower NG error rates. The higher error rate in SOM should be attributed to its fixed neighborhood function (while NG uses flexible neighborhood) that imposes constraints to the learning process in multimodal layer. The second goal of the current experiment is hence the analysis of the neighborhood function in SOM. We presented stimuli with increasing fuzziness in the spatial location (see Fig. 2) and compared error rates of SOM and NG algorithms.

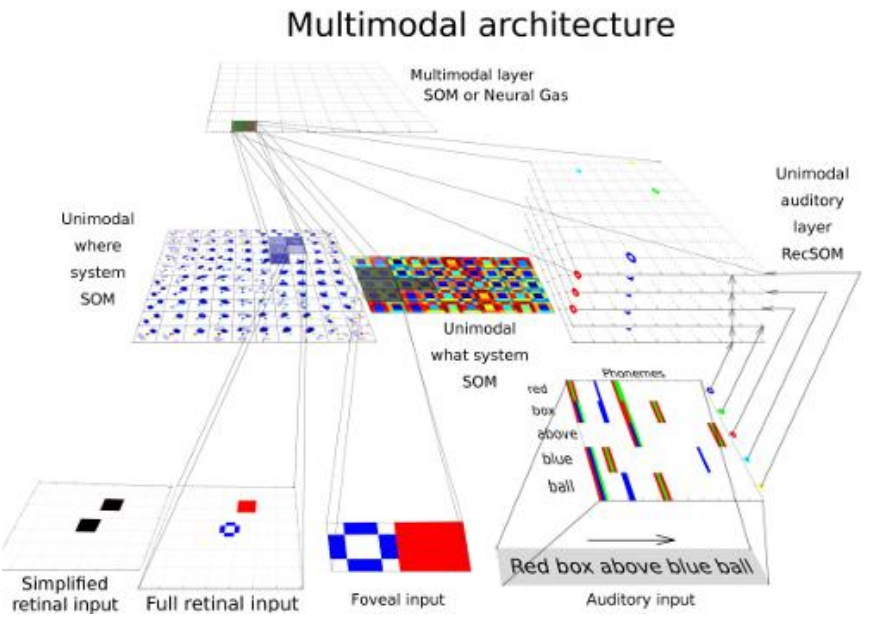

Fig. 1. The two-layer multimodal architecture used in our experiments

3 The Model

We adopted the model from [5] to test the architecture with a simplified type of inputs and variable level of fuzziness. The inputs (similar to previous models) consisted of two objects in 2D environment and their linguistic descriptions. The scenes contained the trajector and the base object in different spatial configurations. The position of the base is fixed in the center of the scene and the trajector position is fuzzy with variable level of fuzziness (Fig. 2). We trained the model using scenes with 3 colors (red, green, blue), 5 object types (box, ball, table, cup, bed) and 4 spatial relations (above, bellow, left, right) that means 840 combinations of two different objects in the scene. There were 42000 examples (50 instances per spatial configuration) in the training set.

3.1 Visual Subsystem

The visual subsystem is formed by an artificial retina (28×28 neurons) and an artificial fovea (two visual fields consisting of 4×4 neurons) that project visual and spatial information about the trajector and the base to the the primary unimodal visual layers. These layers are both made of SOMs that differentiate various positions of two objects (resembling *where* pathway) from retinal projection and color and shape of objects (resembling *what* pathway) from foveal projection. The color of each pixel was encoded by the activity level. Both maps were trained for 100 epochs with decreasing parameter values (unit neighborhood radius, learning rate).

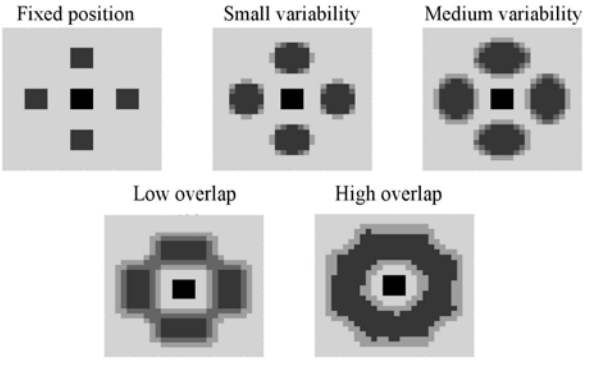

Fig. 2. Simplified visual inputs with varying levels of spatial fuzziness

3.2 Auditory Subsytem

Auditory inputs (English sentences) were encoded as phonological patterns representing word forms using PatPho, a generic phonological pattern generator that fits every word (up to trisyllables) onto a template according to its vowel-consonant structure [9]. It uses the concept of syllabic template: a word representation is formed by combinations of syllables in a metrical grid, and the slots in each grid are made up by bundles of features that correspond to consonants and vowels. In our case, each sentence consists of five 54-dimensional vectors with component values in the interval (0,1). These inputs are sequentially fed to RecSOM [10] that learns to represent inputs (words) in the temporal context (hence capturing sequential information). RecSOM output, in terms of map activation, feeds to the multimodal layer, to be integrated with the visual pathway. RecSOM units become sequence detectors after training, topographically organized according to the suffix (the last words).

Since RecSOM, unlike SOM, is not common, we provide its mathematical description here. Each neuron $i \in \{1, 2, ..., N\}$ in RecSOM has two associated weight vectors: $\mathbf{w}_i \in \mathbb{R}^n$ – linked with an n-dimensional input $\mathbf{s}(t)$ (in our case, the current word, $n = 54$) feeding the network at time t and $\mathbf{c}_i \in \mathbb{R}^N$ – linked with the context $\mathbf{y}(t-1) = [y_1(t-1), y_2(t-1), ..., y_N(t-1)]$ containing map activations $y_i(t-1)$ from the previous time step.

The output of a unit i at time t is $y_i(t) = \exp(-d_i(t))$, where

$$d_i(t) = \alpha \cdot \|\mathbf{s}(t) - \mathbf{w}_i\|^2 + \beta \cdot \|\mathbf{y}(t-1) - \mathbf{c}_i\|^2.$$

Here, $\|\cdot\|$ denotes the Euclidean norm, $\alpha > 0$ and $\beta > 0$ are model parameters that respectively influence the effect of the input and the context upon neuron's profile. Their suitable values are usually found heuristically (in our model, we used $\alpha = \beta = 0.1$). Both weight vectors are updated using the same form of SOM learning rule:

$$\Delta\mathbf{w}_i = \gamma \cdot h_{ik} \cdot (\mathbf{s}(t) - \mathbf{w}_i),$$

$$\Delta \mathbf{c}_i = \gamma \cdot h_{ik} \cdot (\mathbf{y}(t-1) - \mathbf{c}_i),$$

where b is an index of the best matching unit at time t, $b = \text{argmin}_i\{d_i(t)\}$, and $0 < \gamma < 1$ is the learning rate. Neighborhood function h_{ib} is a Gaussian (of width σ) on the distance $d(i, b)$ of units i and b in the map: $h_{ib} = \exp(-d(i, b)^2/\sigma^2)$. The 'neighborhood width', σ, linearly decreases in time to allow for forming topographic representation of input sequences.

3.3 Multimodal Integration

Outputs from both visual SOMs and auditory RecSOM are projected to the multimodal layer (SOM or NG). The main task for the multimodal layer is to find and learn the categories by merging different sources of information. We compared SOM and NG algorithms that are both unsupervised and based on the competition among units, but NG uses a flexible neighborhood function, as opposed to the fixed neighborhood in SOM.

For clarity, we explain NG algorithm briefly here. NG shares with SOM a number of fetaures. In each iteration t, an input vector $\mathbf{m}$(t) is randomly chosen from the training dataset. Subsequently, for all units in the multimodal layer we compute $d_i(t) = \|\mathbf{m}(t) - \mathbf{z}_i\|$ and sort the NG units according to their increasing distances d_i, using indices $l = 0, 1, ...$ (where $l(0)$ corresponds to unit b, the current winner). Then we update all weight vectors $\mathbf{z}_i$ according to

$$\Delta \mathbf{z}_i = \epsilon \cdot \exp(-l(i)/\lambda) \cdot (\mathbf{m}(t) - \mathbf{z}_i) \tag{1}$$

with ϵ as the adaptation step size and λ as the so-called neighborhood range. We used $\epsilon = 0.5$ and $\lambda = n/2$ where n is number of neurons. Both parameters are reduced with increasing t. It is known that after sufficiently many adaptation steps the feature vectors cover the data space with minimum representation error [8]. The adaptation step of the NG can be interpreted as gradient descent on a cost function.

Inputs for the multimodal layer are taken as unimodal activations (from both modalities) using the k-WTA (i.e. winner-takes-all) mechanism, where k most active units are proportionally turned on, and all other units are reset to zero (in the models, we used $k = 6$ for visual layers). The motivation for this type of output representation consists in introducing overlaps between similar patterns to facilitate generalization. On the other hand, the output representation in the multimodal layer is chosen to be localist for better interpretation of results and the calculation of error rate.

4 Results

We trained the system with the fixed sizes of unimodal layers (30×30 units) and the multimodal layer (29×29). After the training phase, the system was tested by a novel set of inputs. All inputs were indexed for the error calculation in the second layer. Then we measured the effectiveness of this system, based on the

percentage of correctly classified test inputs. To calculate the accuracy of neuron responses, we applied a voting algorithm after training to label each neuron in the layer based on its most frequent response. Then we measured the accuracy of this system, based on the percentage of correctly classified test inputs. At first, we compared the error rate in unimodal *where* layer trained with full retinal images or simplified monochromatic rectangles standing for objects in the scene. Results are shown in Fig. 3. It can be seen that simplified input significantly reduce the error for all levels of spatial fuzziness which could be explained by reduced variability of inputs that are topographically mapped in the SOM.

The analysis of the model behavior revealed that the trajector shape and the spatial term representations are the most difficult task components for visual unimodal systems which is caused by their variability and fuzziness. The model analysis also confirmed that simplified projection of retinal images to the *where* system resulted in lower error rates compared to full retinal images (Fig. 3). This leads us to the conclusion that it is possible to simplify the information projected to the *where* system to optimize the speed and effectiveness of our architecture.

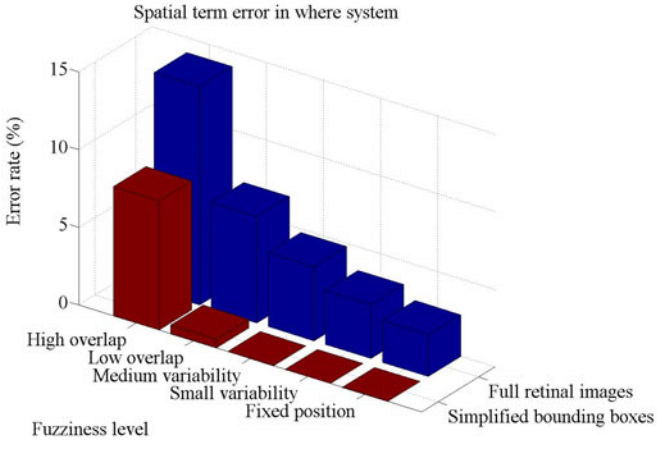

Fig. 3. The error rates in the *where* system as a function of input types and the levels of spatial fuzziness

Next we compared SOM and NG algorithms in the multimodal layer using the simplified *where* system. The calculation of the error rates was the same as for the unimodal layers. Fig. 4 shows a lower error rate for NG in all levels of fuzziness and the high error rates for SOM regardless of the fuzziness level.

The poorer result of multimodal SOM compared to NG could most probably be attributed to the fixed neighborhood function which imposes constraints the learned nonlinear mapping. There was a 70% error rate for all levels of fuzzy inputs, so the multimodal SOM is able to represent neither fuzzy inputs nor distinct inputs. We observed a different type of clustering in unimodal layers

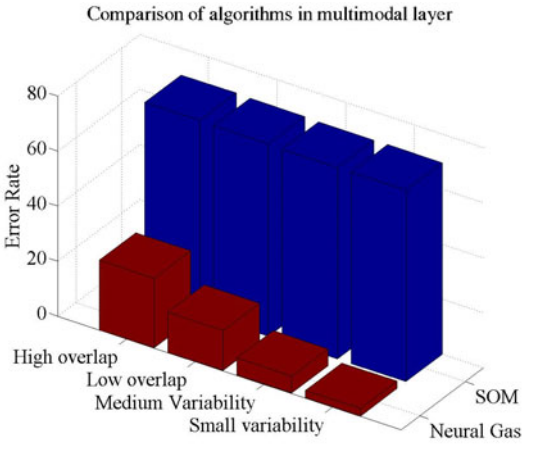

Fig. 4. Comparison of two models for different type of inputs. The error rates in the *where* system as a function of levels of spatial fuzziness.

that are transferred to the multimodal layer at which the SOM is not able to adapt to the joint outputs from unimodal layers. The results of NG algorithm for the same input data confirm this hypothesis. There was a 25% error rate only for highly overlaping inputs (compared to 70% error rate for all type of inputs for SOM. The effectiveness of NG for less fuzzy inputs was even better.

5 Conclusion

Previous models of symbol grounding (see Discussion in [5]) deal with the lexical level but our model goes beyond words because it is able to represent sentences with fixed grammar via RecSOM. It finds the mapping of the particular words to the concepts in the multimodal layer without any prior knowledge, so the system proposes the solution to the binding problem. The system design allows us in principle to append other modalities and still represent discrete multimodal categories. The hierarchical representation of the sign components is the important advantage of our model. It guarantees better processing and storing of representations because the sign (multimodal level) is modifiable from both modalities (the sequential "symbolic" auditory level and the parallel "conceptual" visual level). The separate multimodal level provides a platform for the development of subsequent stages of this system (e.g. inference mechanisms). Further tests of this approach should focus on scaling up this architecture to more complex mappings.

Acknowledgment. This research was supported by the research program MSM 6840770012 of the Czech Technical University in Prague, and by SAIA scholarship (M.V.) and by Slovak Grant Agency for Science, no. 1/0439/11 (I.F.).

References

1. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
2. Barsalou, L.: Perceptual symbol systems. Behavioral and Brain Sciences 22(4), 577–660 (1999)
3. Riga, T., Cangelosi, A., Greco, A.: Symbol grounding transfer with hybrid self-organizing/supervised neural networks. In: Int. Joint Conf. on Neural Networks (2004)
4. Millner, A., Goodale, M.: The Visual Brain in Action. Oxford University Press (1995)
5. Vavrečka, M., Farkaš, I.: Unsupervised grounding of spatial relations. In: Proceedings of European Conference on Cognitive Science, Sofia, Bulgaria (2011)
6. Kohonen, T.: Self-Organizing Map (3rd, extended edition). Springer, New York (2011)
7. Stein, B., Meredith, M.: Merging of the Senses. MIT Press, Cambridge
8. Martinetz, T., Schulten, K.: A neural-gas network learns topologies. In: Kohonen, T., et al. (eds.) Int. Conf. on Artificial Neural Networks, pp. 397–402. North-Holland, Amsterdam (1991)
9. Li, P., McWhinney, B.: PatPho: A phonological pattern generator for neural networks. Behavior Research Methods, Instruments and Computers 34(3), 408–415 (2002)
10. Voegtlin, T.: Recursive self-organizing maps. Neural Networks 15(8-9), 979–991 (2002)