Motivation

The overall goal of our project is design of modular architecture that demonstrates sensory-motor learning, social learning and language learning. We focus on neural network approach, with a dominant role of self-organizing maps that are adopted in various parts of the model. Our emphasis on unsupervised learning is cognitively appealing and will be coupled with supervised learning only in cases when the teaching signal originates from the environment itself (e.g. linguistic labels entered via auditory modality), rather than being provided by an external teacher (designer), to avoid the grounding problem.

Main objective is to test the limitations of machine learning in the process of building representations solely from the sensory inputs. We propose a hierarchical architecture that is able to represent information from different modalities and to find the mapping between unimodal representations. This approach imitates the nature of human learning capabilities within the development.

We test the radical version of embodied cognition (Varela et al., 1991), arguing that the co-occurrence of inputs from the environment is a sufficient source of information to create an intrinsic representational system. These representations preserve constant attributes of the environment. We propose an alternative solution to the classical architecture (Hinton, 1990). The difference is the way of processing symbolic input by a separate auditory subsystem and further integration of auditory and visual information in a multimodal layer. Multimodal layer incorporates the process of identification. Our approach is similar to “grounding transfer” (Ripa, Camurri, & Ghezzi, 2006) based on SOM maps and supervised multi-layer perception, but our system works in fully unsupervised manner that implies different way of symbolic level creation.

Our model is based on hierarchical processing. Top (multimodal) level is justified and modified from both modalities (the ‘symbolic’ assumption of the multimodal level) and the ‘representation’ level. The latter level of the multimodal layer provides platform for the development of subsequent stages of the system (inference mechanisms etc.).

The models

We focus on learning spatial locations of two objects in 2D space and their linguistic description. This conceptual level is represented by the visual subsystem, the symbolic level is represented by the auditory subsystem. We use unimodal visual layers (28x28 neurons) to study the difference in the discrimination of ‘what’ and ‘where’ (vocabulary). Our model learns to separate object and their position from the environment.

Visual layer

The visual layer in the environment is formed by artificial retina (28x28 neurons) that processes the input data and learns to separate object and their position from the environment. The SOM was trained for 100 epochs and tested it using a novel set of inputs. Then we measured the effectiveness of this system, based on the percentage of correctly classified test inputs. The error rate was lower for the auditory layer compared to the visual subsystem in the Model I. This should be attributed to the smaller variability of the inputs because there is the same sentence, describing the spatial location of the trajector that varies in the different position in the specific area. There is also a difference between the what and where system effectiveness in the Model II. The where system is more accurate in the representation of spatial relations and the what system represents color and shape of trajector with smaller error.

Multimodal layer

In agreement with the theory of perceptual symbol systems (Barsalou, 1999), a multimodal layer forms the core of the system. Its role is to identify unique combinations of internal representations of information. We tested two implementations of the multimodal layer: a SOM and Neural Gas. The layer has primary role in the unimodal recognition problem, implemented by the unsupervised SOM, to build representations of visual scenes in topographic manner. The SOM was expected to differentiate various positions of two objects, as well as object types and their color in Model I. Model II consists of separate SOM for spatial locations (resembling ‘whole’ system) and separate SOM for color and shape of objects (resembling ‘what’ system). The SOM was trained for 100 epochs (training data size varied from 6400 to 45000 depending on the complexity of environment) with decreasing parameter values (neighborhood radius, learning rate).

Auditory layer

Auditory input (English sentence), in the form phonetic feature vectors feeds into the primary RecSOM (Voegtlin, 2002), a recurrent SOM-based architecture, that learns to represent inputs (words) in temporal contexts (hence capturing sequential information). The sentences are presented one word at a time, RecSOM output, in terms of map activation, feeds to the multimodal layer, to be integrated with the visual pathway. Like SOM, RecSOM is trained by competitive, Hebbian-type of learning.

Multimodal layer

In agreement with the theory of perceptual symbol systems (Barsalou, 1999), a multimodal layer forms the core of the system. Its role is to identify unique combinations of internal representations of information. We tested two implementations of the multimodal layer: a SOM and Neural Gas. The layer has primary role in the unimodal recognition problem, implemented by the unsupervised SOM, to build representations of visual scenes in topographic manner. The SOM was expected to differentiate various positions of two objects, as well as object types and their color in Model I. Model II consists of separate SOM for spatial locations (resembling ‘whole’ system) and separate SOM for color and shape of objects (resembling ‘what’ system). The SOM was trained for 100 epochs (training data size varied from 6400 to 45000 depending on the complexity of environment) with decreasing parameter values (neighborhood radius, learning rate).

Auditory layer

The auditory input (English sentence), in the form phonetic feature vectors feeds into the primary RecSOM (Voegtlin, 2002), a recurrent SOM-based architecture, that learns to represent inputs (words) in temporal contexts (hence capturing sequential information). The sentences are presented one word at a time, RecSOM output, in terms of map activation, feeds to the multimodal layer, to be integrated with the visual pathway. Like SOM, RecSOM is trained by competitive, Hebbian-type of learning.

Future steps

- Representation of causal relations both in visual (spatiotemporal) and auditory (linguistic) domains
- Inference processes based on multimodal layer
- Backward transfer from the symbolic to the conceptual level
- Representation of reference frames and frame-dependent linguistic descriptions of the scene
- Representation of homonymic words (see Fig. 3)
- Hierarchical representation of abstract terms
- Implementation into a robotic system and real robot for testing in the real environment (see Fig. 4)

We are open for cooperation and looking for partners.