Multimodal Symbolic Association using Parallel Multilayer Perceptron

Federico Raue[1,2], Sebastian Palacio[2], Thomas M. Breuel[1], Wonmin Byeon[1,2]
Andreas Dengel[1,2], Marcus Liwicki[1]
[1]University of Kaiserslautern, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Germany.
{federico.raue,sebastian.palacio}@dfki.de
{wonmin.byeon,andreas.dengel}@dfki.de,
{tmb,liwicki}@cs.uni-kl.de

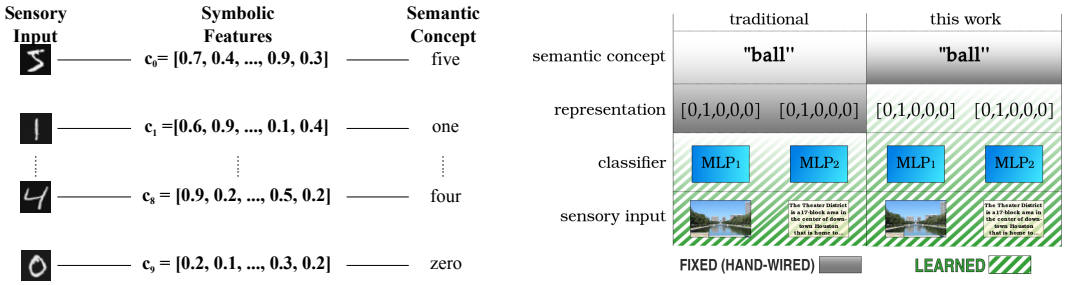
Abstract

The relation between abstract concepts and sensory input, has been extensively studied by multiple scientific communities, as it plays a fundamental role in language development. For instance, we learn to associate the abstract concept *zero* to the Arabic digit '0'. In this paper, we introduce a new model that not only learns to classify the sensory input but also learns the semantic relation and its internal representation. Our framework uses two parallel Multilayer Perceptron with an EM-training rule for learning the semantic association between two input signals that can be of the same modality or not. We assume two constraints: 1) there is a symbolic representation, 2) two input signals have the same symbolic structure. As a result, our model has less information than in the traditional classification problem where the association is fixed during training. We have tested the model using two multimodal datasets (TVGraz and Wikipedia). We observed that our model achieves similar classification accuracy compared to a single MLP with a standard training rule, despite the more challenging setup.

1 Introduction

Associating semantic concepts to sensory responses from the real world, is one of the key components in order to acquire language, and has been investigated in fields like Neuroscience, Cognitive Science and Artificial Intelligence. On infants, it has been shown that such a process takes place through an unconscious mapping going from sensory inputs to high-level concepts [1]. Cognitive researchers have found that the first words infants learn, consist mainly on nouns that correspond to visible objects [2]. Furthermore, infants exhibit a slower development of speech if one of their sensory mechanisms fails (e.g., deafness or blindness) [3, 4]. In Neuroscience, a correspondence was found in brain patterns when visual and audio stimuli were semantically correlated and when they were not [5].

In Psychology, such a mechanism has been described as the *symbol grounding problem* [2]. In general, the symbol grounding problem describes the process by which sensory signals can be coupled to high level concepts that, in turn, relate to each other to create more concepts. Such process can be seen as an interaction between three main components: input signals which is nothing but the objects that we want to ground; sensory signals which are *categorical representations* of input signals e.g., the response of cones and rods in the back of the retina and third, the concepts associated with the sensory information that is received. Note that sensory signals are not intrinsically coupled with high-level concepts. Instead, such signals get gradually associated to the different symbols through some learning process.



(a) (b)

Figure 1: (a) Relation between sensory inputs, symbolic features and semantic concepts. (b) Comparison between fixed and learned components in traditional classification problems and our learning task.

Artificial Intelligence has been involved in this context, developing models that replicate the proposed ideas from the fields of Cognitive Science and Neuroscience. Such models fall into a field known as multimodal machine learning, where there are mainly two tasks at hand: multimodal feature fusion [6, 7] and generation of textual descriptions [8, 9]. In contrast, this work introduces another scenario related to object-word associations based on a cognitive model, namely the symbol grounding problem. In our case, the semantic concept or relation does not have a pre-defined representation in the output space. Figure 1a shows an example of how sensory inputs, symbolic features and semantic concepts are related to each other. The symbolic features by themselves have no intrinsic meaning. Instead, they encode the information from the sensory inputs into a compressed numeric form. However, symbolic features are grounded once a consistent link exists between them and a semantic concept.

So far, these models have used a fixed relation between the semantic concepts and the sensory input. With this in mind, we propose an alternative learning rule for semantic association between different sensory input, where the output layer of a Multilayer Perceptron (MLP) is respectively used as a symbolic feature. As a result, the internal representation of each semantic concept is also learned. Figure 1b shows the components in a classification problem and the difference between the traditional case and the proposed learning scenario. It can be seen that our model has less information in relation to the traditional scenario as the latter relies on the availability of both the semantic concept and the corresponding representation in the output space. In contrast, our proposed model relies only of the semantic concept and learns its representation during training. In other words, our model learns to classify the sensory input signals *and* to compute the semantic association simultaneously. The proposed model is an adaptation from Raue *et al.* [10] where they used Long Short-Term Memory (LSTM) networks. Besides, we use two sensory input signals, whereas their model was applied to parallel sequences in only one modality.

In this work, we introduce our model, which uses two *parallel MLPs* with a modified training scheme for handling multiple sensory input signals. Both *parallel MLPs* learn to associate a pair of visual and textual inputs. The association rule is defined using *Expectation Maximization* (EM). An overview of the whole architecture can be found in Figure 2. Moreover, a comparison of performance between parallel MLPs and standard MLPs is also being conducted.

2 Methodology

As mentioned in Section 1, the main goal of the proposed model is to learn the relationship between two different sensory input signals which are associated with the same semantic concept. More formally, we define a set $\mathcal{S} = \{(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{c}) \mid \boldsymbol{x_1} \in \boldsymbol{X_1},\ \boldsymbol{x_2} \in \boldsymbol{X_2},\ \boldsymbol{c} \in \boldsymbol{C}\}$ where $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$

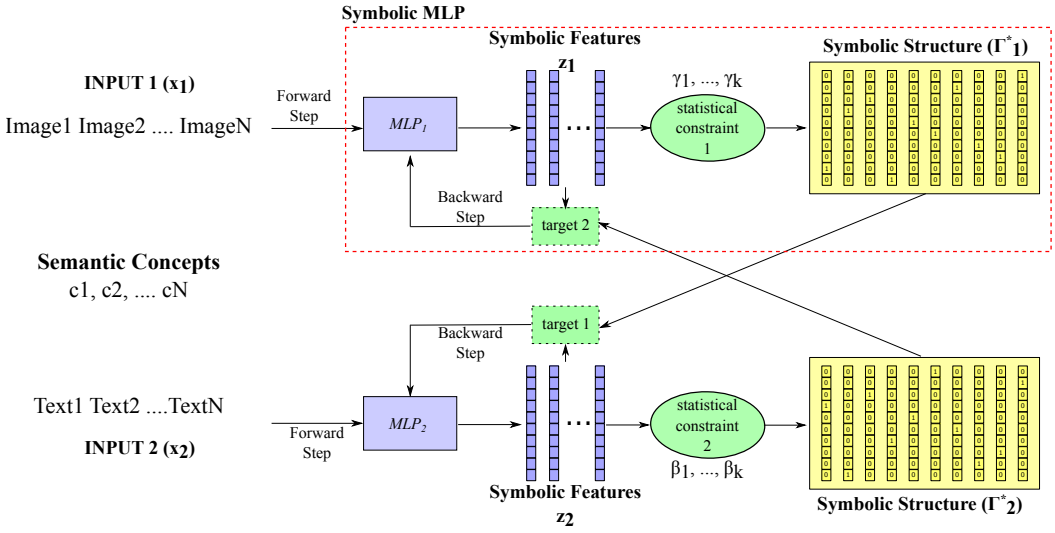


Figure 2: Overview for the *parallel MLP*. Parallel training sets are forwarded to each MLP. Then, the statistical constraint in each network is applied in order to find the structure between the *semantic concepts* and the *symbolic features*, given the output of the network. Later, the deltas are calculated with the following procedure: One of the networks is the target of the other network, and vice versa. Finally, the weights of both MLPs are updated using the deltas.

represent sets of sensory input signals, and $\boldsymbol{C}$ is the set of semantic concepts that link members of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. From now on, we also refer to this learned relationship as the *symbolic structure*.

First, our model processes sensory input signals to generate categorical representations. At this point, such a process consists of forwarding input signals through an MLP. In turn, such categorical representations are used to learn the symbolic structure.

Second, learning is achieved via *Expectation Maximization* (EM) [11]. On one hand, a statistical constraint first estimates the best mapping between categorical representations of input samples. Then, the estimated mapping adjusts according to a known cost based on the distribution of the input samples.

2.1 Symbolic MLP (sMLP)

In this paper, a modified version of MLP is introduced for learning the association between a set of sensory inputs with its semantic concepts. The general idea is to use the outputs of the network as symbolic features and simulate the same cognitive behaviour of associating semantic meaning to meaning-less objects (grounding) using a symbolic structure. The combination of the MLP and the symbolic structure is referred to as symbolic MLP (sMLP). Without loss of generality, we explain the training rule in terms of only one of the sensory input sets[1] in $\mathcal{S}$ and the set of semantic concepts $\boldsymbol{C}$. In addition to them, we also define the set of output vectors $\boldsymbol{Z}$ as the symbolic features calculated from $\boldsymbol{X}$. Note that the size of the output layer is equal to $|\boldsymbol{C}|$.

The learning rule for an *sMLP* is updated in mini-batches of size m. We represent the elements in the mini-batch using $\boldsymbol{X}_m$, $\boldsymbol{C}_m$ and $\boldsymbol{Z}_m$. First, a mini-batch of elements from $\boldsymbol{X}_m$ and $\boldsymbol{C}_m$ is forwarded to the sMLP. Then, output signals from $\boldsymbol{Z}_m$ are used as symbolic features constrained to a prior known distribution (Section 2.1.1). In more detail, we are using the raw values of $\boldsymbol{Z}_m$ for matching the statistical distribution of the mini-batch with a target distribution such as a uniform distribution. Consequently, this information can be used to find the most likely association in the

[1] we drop the index of the chosen set e.g. $\boldsymbol{X} \equiv \boldsymbol{X}_1$.

symbolic structure for the MLP. Finally, the symbolic structure provides the target vector for each semantic concept. After the sMLP is trained, the semantic concept can be retrieved from the sMLP outputs (Section 2.1.2).

2.1.1 Statistical Constraint

As we stated before, the output vectors $\boldsymbol{z}$ from the mini-batch $\boldsymbol{Z_m}$ are used as symbolic features. Moreover, all raw values ($\boldsymbol{z} \in \boldsymbol{Z}_m$) need to match a prior statistical distribution. With this in mind, we introduce a set of weighted concepts ($\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_c$), where each $\boldsymbol{\gamma}_c$ is a vector of size k. Each weighted concept represents the relation between a semantic concept (c) and its symbolic representation. We follow an EM-approach for training the weighted concepts.

The *E-Step* finds the best symbolic structure ($\boldsymbol{\Gamma}^*$) given the current symbolic features and the weighted concepts. First, we define the *average weighted symbolic feature* $\hat{\boldsymbol{z}}_c$ for each semantic concept as follows:

$$\hat{\boldsymbol{z}}_c = \frac{1}{m} \sum_{\boldsymbol{z} \in \boldsymbol{Z}_m} \boldsymbol{z}^{\boldsymbol{\gamma}_c},\ c \in C \tag{1}$$

where $\boldsymbol{z}^{\boldsymbol{\gamma}_c}$ is the element-wise power operation between the vector $\boldsymbol{z}$ and the vector $\boldsymbol{\gamma_c}$. More formally $\boldsymbol{z}^{\boldsymbol{\gamma}_c} = [z_0^{\gamma_{c,0}}, \ldots, z_k^{\gamma_{c,k}}]$. The measure of weighted responses of all inputs in a mini-batch accumulates the tendency for symbolic features to converge to their intrinsic semantic concepts.

We can now organize all $\hat{\boldsymbol{z}}_c$ into a square matrix $\boldsymbol{\Gamma} = \left[\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_c\right],\ c \in C$, which in turn, is used for constructing a matrix $\boldsymbol{\Gamma}^*$ by setting the maximum element $\Gamma_{i,j}$ in $\boldsymbol{\Gamma}$ to 1.0 followed by a row-column elimination. In other words, we set all elements in the i-th row and j-th column of $\boldsymbol{\Gamma}$ to 0.0 (except at $\Gamma_{i,j}$ which has been set to 1.0). This process is iteratively performed $|C|$ times. Thus, ensuring there is a one-to-one relation between semantic concepts and symbolic features. More formally, $\boldsymbol{\Gamma}^*$ can be seen as a permutation of the identity matrix. Also, $\boldsymbol{\Gamma}^*$ defines the symbolic structure, where semantic concepts are associated to the output space. The columns encode the information about semantic concepts, while the rows represent the different symbolic features. To map any given symbolic feature to a semantic concept, it now suffices to look up $\boldsymbol{\Gamma}^*$.

The *M-Step* updates the weighted concepts given the current symbolic structure. To that effect, we define the following loss function:

$$cost(\boldsymbol{\gamma}_c) = \left(\hat{\boldsymbol{z}}_c - \frac{1}{|C|}\boldsymbol{\Gamma}^*_{\boldsymbol{c}}\right)^2,\ c \in C \tag{2}$$

where $\boldsymbol{\Gamma}^*_{\boldsymbol{c}}$ denotes the c-th column vector of $\boldsymbol{\Gamma}^*$. Furthermore, we assume an uniform distribution among all elements in C. Thus, we normalize $\boldsymbol{\Gamma}^*_{\boldsymbol{c}}$ by $|C|$. Next, each weighted concept is updated using gradient descent.

In addition, this step not only learns the symbolic structure but also provides information for updating the weights in the symbolic MLP. The current symbolic structure provides the target vector for the back-propagation step of the MLP. In this case, the target for the semantic concept c is the column vector $\boldsymbol{\Gamma}^*_{\boldsymbol{c}}$.

2.1.2 Retrieving Semantic Concepts from Symbolic Features

After an sMLP is trained, the semantic concept can be extracted from the symbolic feature ($\boldsymbol{z}$). The decision rule of the MLP is the maximum value of the output vector $\boldsymbol{z}$. We denote this decision as k^*.

Furthermore, we can combine the learned symbolic structure with k^*. Hence, the semantic concept is retrieved by the maximum value of all weighted concepts at position k^* in $\boldsymbol{z}$.

2.2 Parallel Symbolic MLP

Parallel Symbolic MLP follows a similar EM-training mechanism as the symbolic MLP. In this case, two sMLP's ($sMLP_1$, $sMLP_2$) process each sensory input set on $\mathcal{S}$ in parallel, for learning a unified symbolic structure. First, the pair of samples ($\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{S}_m \subseteq \mathcal{S}$) are fed to each *sMLP*, where $\mathcal{S}_m$ refers to a mini-batch in $\mathcal{S}$. We want to point out that both networks do not share any

Table 1: Average accuracy (% on a 10-fold) of our model and the standard MLP. It is observed that our model reaches similar performance compared to the standard MLP. Note that our focus is not to outperform the standard MLP, but to solve the semantic association without hindering performance.

DATASET	FORMAT	METHOD	
		Our Model	MLP
Wikipedia	visual	27.44 ± 2.69	28.38 ± 1.60
	text	34.07 ± 2.96	37.25 ± 1.43
TVGraz	visual	53.19 ± 2.74	55.97 ± 1.86
	text	52.74 ± 2.45	53.65 ± 1.38

weights between them. Second, statistical constraints (γ, β) are applied to the outputs $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ for selecting the most likely symbolic structure (Γ_1^*, Γ_2^*). As a reminder, the symbolic structure is a square matrix that represents the relation between the *semantic concepts* and *symbolic features*. We want to indicate that the previous two steps are the same as with one symbolic MLP. Third, the symbolic structure (Γ_1^*) from $sMLP_1$ is used as a target for $sMLP_2$ in order to update the weights, and vice versa. This step ensures that, over time, both networks agree on the same symbolic structure. Moreover, the statistical constraints on both networks are updated in cases where the symbolic relation is the same i.e., $\Gamma_{1c}^* = \Gamma_{2c}^*$.

3 Experiments

As mentioned previously, the aim of our experiments is to evaluate the semantic association of two parallel MLPs based only on semantic concepts without affecting the classification performance. To that effect, we tested our model using two multimodal datasets: **Wikipedia Articles** [12] and **TVGraz** [13]. Each disjoint set corresponds to just one modality e.g., $\boldsymbol{X}_1$ corresponds to text features and $\boldsymbol{X}_2$ to image features. Figure 3 shows examples of the datasets.

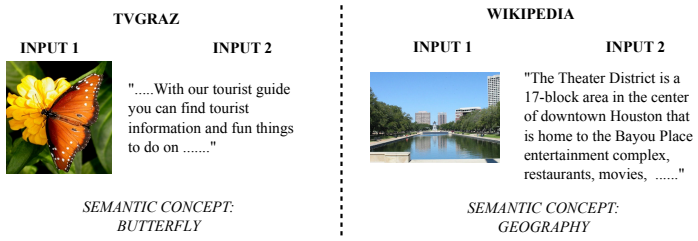


Figure 3: Example of the datasets. It is observed that each sample (input1-input2) has two instances of the same semantic concept.

We compared the accuracy of our model against standard MLPs, which have been trained independently on each modality. It can be seen in Table 1 that the performance of our model was consistent with respect to the standard MLP. This suggests that the sMLP's in our model are able to converge to a unified symbolic structure despite the dynamic scenario. By 'dynamic scenario' we refer to the oscillating structure of the semantic concepts and symbolic features during training. Figure 4 shows an example of several epochs and the components during training.

In the future, we plan to extend our model to scenarios where even semantic concepts are unknown. The intuition behind, is to translate from one modality to another modality thought the symbolic structure. Moreover, we are interested in exploiting deeper architectures for increasing performance.

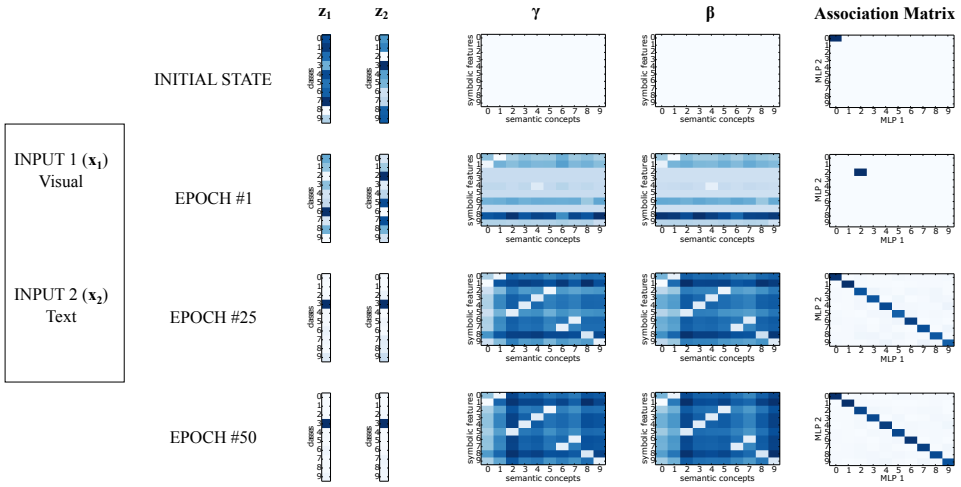


Figure 4: Example of the learning behavior for the symbolic association model at different stages. Initially, the presented model sets the weighted concepts (γ and β) to 1.0 for both networks. Also, the association matrix shows only one relation (0, 0) for all outputs. During training, the model starts learning the underlying symbolic structure represented by both weighted concepts. The last row (epoch 50) shows the semantic prediction step. Here, the maximum value (dark blue) of the output vector is '3', which is associated with the semantic concept *butterfly*. This behaviour is consistent between both weighted concepts. Hence, the association matrix results in a diagonal matrix which indicates that both networks have agreed on the same symbolic structure.

References

[1] P. C. Quinn, P. D. Eimas, and S. L. Rosenkrantz, "Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants," *PERCEPTION-LONDON-*, vol. 22, pp. 463–463, 1993.

[2] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, pp. 335–346, 1990.

[3] E. S. Andersen, A. Dunlea, and L. Kekelis, "The impact of input: language acquisition in the visually impaired," *First Language*, vol. 13, no. 37, pp. 23–49, Jan. 1993.

[4] P. E. Spencer, "Looking without listening: is audition a prerequisite for normal development of visual attention during infancy?" *Journal of deaf studies and deaf education*, vol. 5, no. 4, pp. 291–302, Jan. 2000.

[5] M. Asano, M. Imai, S. Kita, K. Kitajo, H. Okada, and G. Thierry, "Sound symbolism scaffolds language development in preverbal infants," *cortex*, vol. 63, pp. 196–205, 2015.

[6] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[7] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Advances in Neural Information Processing Systems*, 2014, pp. 2141–2149.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *arXiv preprint arXiv:1411.4555*, 2014.

[9] A. Karpathy, A. Joulin, and F. F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems*, 2014, pp. 1889–1897.

[10] F. Raue, W. Byeon, T. Breuel, and M. Liwicki, "Parallel Sequence Classification using Recurrent Neural Networks and Alignment," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*.

[11] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society.*, vol. 39, no. 1, pp. 1–38, 1977.

[12] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A New Approach to Cross-Modal Multimedia Retrieval," in *ACM International Conference on Multimedia*, 2010, pp. 251–260.

[13] I. Khan, A. Saffari, and H. Bischof, "Tvgraz: Multi-modal learning of object categories by combining textual and visual features," in *AAPR Workshop*, 2009, pp. 213–224.