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], and Marcus Liwicki[1,2]

[1]University of Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Germany
{federico.raue,sebastian.palacio, wonmin.byeon, andreas.dengel}@dfki.de,
{tmb,liwicki}@cs.uni-kl.de

Abstract. The goal of our paper is to learn the association and the semantic grounding of two sensory input signals that represent the same semantic concept. The input signals can be or cannot be the same modality. This task is inspired by infants learning. We propose a novel framework that has two *symbolic* Multilayer Perceptron (MLP) in parallel. Furthermore, both networks learn to ground semantic concepts and the same coding scheme for all semantic concepts in both networks. In addition, the training rule follows EM-approach. In contrast, the traditional setup of association task pre-defined the coding scheme before training. We have tested our model in two cases: mono- and multi-modal. Our model achieves similar *accuracy association* to MLPs with pre-defined coding schemes.

Keywords: symbol grounding, neural network, cognitive model

1 Introduction

The relation between the real world via sensory input and abstract concepts helps humans to develop language. More formally, Harnad [5] investigated the process of coupling high level concepts and multimodal sensory signals. He called this process the *Symbol Grounding Problem*.

All modalities (visual, audio, and haptic) are important for language acquisition by infants. Cognitive researchers found that nouns are the first acquired words by infants [1]. In more detail, nouns correspond to visible elements, such as dog, cat, etc. In contrast, infants acquire vocabulary slower if one of their sensory input fails i.e. deafness, blindness [1, 17]. Also, Neuroscience researchers discovered different patterns in infants' brain related to multimodal signals and abstract concepts [2]. The patterns showed different behavior depending on the existence or absence of a semantic relation between visual and audio signals. This finding shows a relation between both modalities.

Previous work has been inspired by the *Symbol Grounding Problem*. One of the first model was proposed by Plunket *et al.* [13]. The authors suggested a feed-forward network for associating a visual stimuli and a label. Since then,

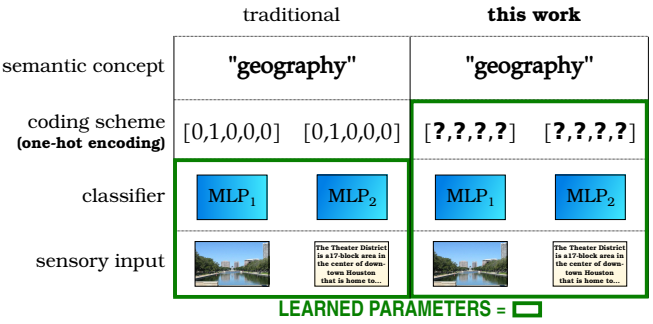

Fig. 1. Components of our learning problem. The coding scheme is unknown in this work and is learned during training.

more complex scenarios have been proposed. Yu and Hallard [18] presented a multimodal model for grounding spoken languages using Hidden Markov Models. Nakamura *et al.* [9] developed a model that ground the word meanings in a multimodal scenario based on Latent Dirichlet Allocation.

In this paper, we are interested in a different setup of the *Symbol Grounding Problem* for two sensory input signals. Moreover, abstract concepts are represented by the sensory input, which can or cannot be of the same modality. Usually, each abstract concept is represented by a pre-defined coding scheme, which is used for training classifiers. Figure 1 shows an example to explain the difference between the traditional setup and this work for the association problem of two sensory input. This problem setup was introduced by Raue *et al.* [15], who only evaluated visual sequences, which was represented by text lines in an OCR case. Our contributions in this paper are

- We define a *symbolic* Multilayer Perceptron (MLP), which is trained without specifying a coding scheme. In this case, an EM-training algorithm is used for learning simultaneously the classification and the coding scheme during training. Hence, the abstract concepts are grounded to the input signals during training (Section 2).
- We propose (mono- and multi-modal) associations via symbol grounding, where two parallel symbolic MLPs learn to agree on the same coding scheme. As a result, the unknown agreements is learned using the information of one network as target of the other network. Moreover, the association is gradient based and can be extended to deeper architectures (Section 3).
- The *Association Accuracy* of the presented model reaches similar results to MLP training with a pre-defined coding scheme in two scenarios: mono-modal and multi-modal (Sections 4 and 5).

2 Symbolic Multilayer Perceptron

In this paper, a new training rule for Multilayer Perceptron (MLP) is introduced. For explanation purposes, we define a MLP with one hidden layer, where $\boldsymbol{x}$, $\boldsymbol{y}$,

and $\boldsymbol{z}$ are vectors that represent the input, hidden, and output layers, respectively. In addition, we define a set of *weighted concepts* $\boldsymbol{\gamma}_c$ where $c \in \{1, \ldots, C\}$. Each *weighted concept* learns the relation between the semantic concept and the output layer. In this case, the output layer is used as a symbolic feature at which the size of vectors $\boldsymbol{z}$ and $\boldsymbol{\gamma}_c$ is the same. The cost function matches the output vectors $\boldsymbol{z_1}, \ldots, \boldsymbol{z_m}$ in a mini-batch of size m with a uniform distribution. The proposed learning rule follows an *Expectation Maximization* approach [4].

2.1 Training

The *E-step* finds suitable candidates for the *coding scheme* given the network outputs and the *weighted concepts*. Initially, the *weighted concepts* are set to 1.0. First, we define an approximation vector $\hat{\boldsymbol{z}}_c$ for each semantic concept c. It is defined as follows

$$\hat{\boldsymbol{z}}_c = \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{z_i}, \boldsymbol{\gamma}_c), \tag{1}$$

where $\boldsymbol{z_i}$ is the output vectors, $\boldsymbol{\gamma}_c$ is a weighted concept vector c, m is the size of the mini-batch, and the function f is the element-wise power operator between vectors $\boldsymbol{z_i}$ and $\boldsymbol{\gamma}_c$. Equation 1 provides an approximation of all semantic concepts. Second, all approximation vectors $\hat{\boldsymbol{z}}_c$ are concatenated in order to obtain the array $\boldsymbol{\Gamma}$

$$\boldsymbol{\Gamma} = g\Big(\big[\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_C\big] \Big), \tag{2}$$

where function g represents a row-column elimination procedure. In other words, all elements in the i-th row and j-th column of the input array are set to 0 (except at position (i,j), which are set to 1). This process is iteratively performed c times. As a result, $\boldsymbol{\Gamma}$ is a set of *one-hot vectors* and represents a one-to-one relation between semantic concepts and symbolic features. Consequently, $\boldsymbol{\Gamma}$ is an array where 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 coding scheme. To that effect, we define the following loss function:

$$cost(\boldsymbol{\gamma}_c) = \left(\hat{\boldsymbol{z}}_c - \frac{1}{|C|} \boldsymbol{\Gamma_c} \right)^2, \tag{3}$$

where $\boldsymbol{\Gamma_c}$ denotes the c-th column vector of $\boldsymbol{\Gamma}$. Furthermore, we assume a uniform distribution among all elements in c. Thus, we normalize $\boldsymbol{\Gamma_c}$ by the number of semantic concepts c. Next, each weighted concept is updated using gradient descent

$$\boldsymbol{\gamma}_c = \boldsymbol{\gamma}_c - \alpha * \nabla cost(\boldsymbol{\gamma}_c), \tag{4}$$

where $\nabla cost(\boldsymbol{\gamma}_c)$ is the derivative w.r.t. $\boldsymbol{\gamma}_c$ and α is the learning rate. In addition, this step not only learns the *coding scheme* but also provides information for

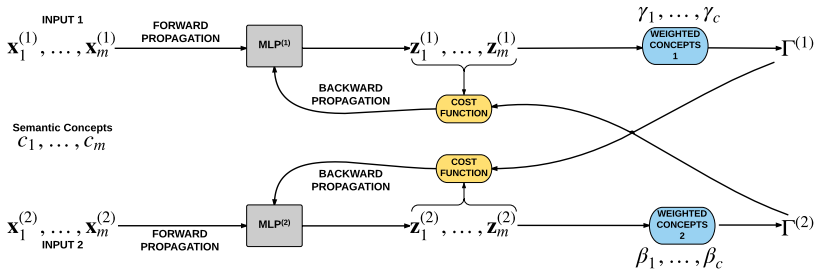

Fig. 2. Overview for the *parallel symbolic MLPs*. Parallel training sets are forwarded to each MLP. The EM-training rule learns to agree on the same coding scheme for both networks, where the coding schemes are unknown before training.

updating the weights in the symbolic MLP. The current *coding scheme* provides the target vectors for propagating backward. In this case, the target vectors for the semantic concept c is the column vector $\boldsymbol{\Gamma_c}$.

2.2 Semantic Concept Prediction

After the symbolic MLP is trained, the semantic concept can be retrieved by a similar decision rule of the standard MLP. With this in mind, the decision rule is defined by

$$c^* = arg\ max_c\ f(z_{k^*}, \gamma_{c,k^*}), \quad where \quad k^* = arg\ max_k\ \boldsymbol{z}, \tag{5}$$

z_{k^*} is the value from output vector $\boldsymbol{z}$ at index k^*, γ_{c,k^*} is the value from *weighed concept* vector $\boldsymbol{\gamma_c}$ at index k^*, and function f is the power operator.

3 Parallel Symbolic MLP

As we mentioned in Section 1, our problem is defined by the association of two different sensory input signals, which represent the same semantic concept with an *unknown* coding scheme. Note that, the sensory input signals may be or may not be the same modality. More formally, the input set is defined by $\mathcal{S} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, c) | \boldsymbol{x}^{(1)} \in \boldsymbol{X}^{(1)}, \boldsymbol{x}^{(2)} \in \boldsymbol{X}^{(2)}, c \in C\}$, where $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are the set of elements for each input, and C is the set of all semantic concepts. We want to point out that our model does not have a pre-defined target vector via coding-scheme.

The proposed architecture combines two symbolic MLPs in parallel, where the information of one network is used as a target of the other network, and vice versa. Figure 2 shows an overview of the proposed model. The training follows a similar approach to the symbolic MLP (*cf.* Section 2).

Initially, two symbolic MLPs propagates forward each sensory input ($\boldsymbol{x_i}^{(1)}$ and $\boldsymbol{x_i}^{(2)}$ where $i = 1, \ldots, m$) in the mini-batch of size m. Afterwards, the *weighted*

Table 1. Sampling of datasets for training and testing. Each sample represents a pair of input signals.

DATASET	CONCEPT	TRAIN	TEST
MNIST	10	25000	4000
COIL-20	20	360	360
TVGraz	10	1942	652
Wikipedia	10	2146	720

concepts of both networks ($MLP^{(1)}$: $\boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_c}$ and $MLP^{(2)}$ $\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_c}$) are applied to network outputs ($\boldsymbol{z_i}^{(1)}$ and $\boldsymbol{z_i}^{(2)}$) in order to obtain the candidates for the coding scheme for each network ($\boldsymbol{\Gamma}^{(1)}$ and $\boldsymbol{\Gamma}^{(2)}$). As a reminder, the coding scheme represents the relation between the semantic concepts and the symbolic features. Finally, the generated coding scheme from one network is used as a target for the other network in order to update the network weights, and vice versa. This step forces both networks to learn the same coding scheme. Figure 2 illustrates the presented architecture.

4 Experimental Design

4.1 Datasets

As we mentioned, our goal was to evaluate the symbolic association of two entities that represent the same semantic concept, where the coding scheme is not pre-defined before training. To that effect, we tested our model in two scenarios: mono-modal and multi-modal. Furthermore, we compared the presented model against the traditional classification problem, where the coding-scheme is already defined.

For the case of mono-modal input signals, two instances represented the same semantic concept, e.g., two images showing different instances of the same digit. With this in mind, we used MNIST [7] and COIL-20 [11] for generating the training and the testing set. We want to indicate that COIL-20 does not define a training and a testing set as MNIST does. However, we applied a common practice, which is to use the even view angles for training and the odd view angles for testing. For the multi-modal case, each input represents one modality of the same concept, e.g., image or text. We tested two multi-modal datasets: Wikipedia Articles [14] and TVGraz [6], where each multi-modal dataset represents the semantic concept using an image and a description of the image. All datasets were evaluated using training and testing sets of randomly sampled pairs with the constraint that all semantic concepts follow a uniform distribution. Table 1 gives an overview of such sampling.

4.2 Features and Network Setup

For each mono-modal dataset, we used the raw pixel values as input. For multi-modal datasets, we extracted *Latent Dirichlet Allocation* [3] features for text,

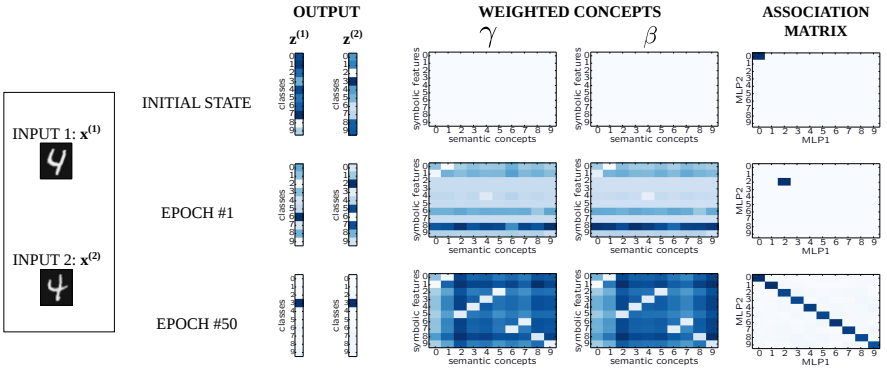

Fig. 3. Example of the learning behavior for the symbolic association model at different stages.

based on a model with 100 topics and, *Bag-of-Visual-Words* [16] based on SIFT [8] using a codebook of size 1024 for the corresponding visual input. Moreover, we used NLTK[1] for extracting LDA features and VLFeat[2] for computing SIFT features. These are the same features used by Pereira and Vasconcelos [12] for the multi-modal datasets. Note that we rescaled the feature values to mean zero and standard deviation one, in the multi-modal datasets. These steps were not required for the mono-modal datasets.

The following parameters were used in MNIST and COIL-20 datasets for each symbolic MLP: hidden layer was set to 40 neurons, learning rate to 0.0001, momentum to 0.9, and learning rate for *weighted concepts* to 0.01. Moreover, the size of the mini-batch was set 1000 and 360 for MNIST and COIL-20, respectively. For multi-modal datasets, the following parameters were used: the size of the hidden layer was 150 neurons, the learning rate was 0.00001, momentum was 0.9, and the learning rate for *weighted concepts* was 0.01. The size of the mini-batch was 300 samples. In both cases, the same parameters were used for the standard MLP with a pre-defined coding scheme as upper bound.

5 Results and Discussion

In this paper, we compared the association accuracy of our model against an MLP with a pre-defined coding scheme. The association accuracy is defined by

$$\textit{Association Accuracy} = \frac{1}{N}\sum_{i=1}^{N} h\big(\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_i^{(2)}, gt_i\big) \tag{6}$$

where $\boldsymbol{z}_i^{(1)}$ and $\boldsymbol{z}_i^{(2)}$ are the output classification from each network, gt_i is the ground-truth label, N is the total number of elements, and the function h is

[1] http://www.nltk.org/
[2] http://www.vlfeat.org/

Table 2. Association accuracy (%) of our model and the traditional approach using MLP.

Dataset	Our Model	Standard MLP
MNIST	94.61 ± 0.24	95.02 ± 0.32
COIL-20	92.86 ± 1.65	92.94 ± 0.62
TVGraz	28.30 ± 1.45	31.50 ± 1.16
Wikipedia	11.82 ± 2.25	12.97 ± 1.11

defined by 1 if $\boldsymbol{z}_i^{(1)} == \boldsymbol{z}_i^{(2)} == gt_i$, and 0 otherwise. We can see in Table 2 that the performance of our model was consistent with respect to the standard MLP. This suggests that the *symbolic* MLPs in our model were able to learn a unified coding scheme.

Figure 3 shows an example of several epochs and the components of our model during training for MNIST. Initially, the association matrix between $MLP^{(1)}$ and $MLP^{(2)}$ shows only one relation at position (0, 0). During training, the model starts learning the underlying *coding scheme* 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 the index '3', which is associated with the semantic concept *four*. This behavior 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.

6 Conclusions

The association between abstract concepts and parallel multimodal signals contributes to language development. In this work, we have shown a model that learns the association of two parallel sensory input signals, which both signals can or cannot be the same modality. Unlike the traditional approach where the coding scheme is pre-defined, we associate two parallel symbolic MLPs that learn a common coding scheme for each semantic concept. Hence, a new dimension is added to the association problem, which makes more sense because we are including the process that abstract concepts are grounded to their sensory representations. We have shown that our model achieved similar results to MLP with traditional training. This holds for both mono- and multi-modal association. *Symbol Grounding* is still an open problem, but reveals potential to understand more the development in this area [10]. One limitation of our work is to learn the association assuming a uniform distribution between the semantic concepts. We will extend our model with different statistical distributions. Another limitation is related to semantic concepts. The model requires more time to converge when the number of semantic concepts increases. Moreover, we are interested in exploiting robustness of deeper architectures and to learn the association when both networks have a different number of semantic concepts.

References

1. Andersen, E.S., Dunlea, A., Kekelis, L.: The impact of input: language acquisition in the visually impaired. First Language 13(37), 23–49 (Jan 1993)
2. Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., Thierry, G.: Sound symbolism scaffolds language development in preverbal infants. cortex 63, 196–205 (2015)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003)
4. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. 39(1), 1–38 (1977)
5. Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena 42(1), 335–346 (1990)
6. Khan, I., Saffari, A., Bischof, H.: Tvgraz: Multi-modal learning of object categories by combining textual and visual features. In: AAPR Workshop. pp. 213–224 (2009)
7. Lecun, Y., Cortes, C.: The MNIST database of handwritten digits
8. Lowe, D.: Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision pp. 1150–1157 vol.2 (1999)
9. Nakamura, T., Araki, T., Nagai, T., Iwahashi, N.: Grounding of word meanings in latent dirichlet allocation-based multimodal concepts. Advanced Robotics 25(17), 2189–2206 (2011)
10. Needham, C.J., Santos, P.E., Magee, D.R., Devin, V., Hogg, D.C., Cohn, A.G.: Protocols from perceptual observations. Artificial Intelligence 167(1), 103–136 (2005)
11. Nene, S.A., Nayar, S.K., Murase, H.: Columbia Object Image Library (COIL-20). Tech. rep. (Feb 1996)
12. Pereira, J.C., Vasconcelos, N.: Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems. Computer Vision and Image Understanding 124, 123 – 135 (2014), large Scale Multimedia Semantic Indexing
13. Plunkett, K., Sinha, C., Møller, M.F., Strandsby, O.: Symbol Grounding or the Emergence of Symbols? Vocabulary Growth in Children and a Connectionist Net. Connection Science 4(3-4), 293–312 (Jan 1992)
14. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G., Levy, R., Vasconcelos, N.: A New Approach to Cross-Modal Multimedia Retrieval. In: ACM International Conference on Multimedia. pp. 251–260 (2010)
15. Raue, F., Byeon, W., Breuel, T., Liwicki, M.: Parallel Sequence Classification using Recurrent Neural Networks and Alignment. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on
16. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2. pp. 1470–. ICCV '03, IEEE Computer Society, Washington, DC, USA (2003)
17. Spencer, P.E.: Looking without listening: is audition a prerequisite for normal development of visual attention during infancy? Journal of deaf studies and deaf education 5(4), 291–302 (Jan 2000)
18. Yu, C., Ballard, D.H.: A multimodal learning interface for grounding spoken language in sensory perceptions. ACM Transactions on Applied Perception (TAP) 1(1), 57–80 (2004)