CLASSLESS ASSOCIATION USING NEURAL NETWORKS

Federico Raue[1,2], Sebastian Palacio[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,andreas.dengel}@dfki.de,
liwicki@cs.uni-kl.de

ABSTRACT

In this paper, we propose a model for the *classless* association between two instances of the same *unknown* class. This scenario is inspired by the *Symbol Grounding Problem* and the *association learning* in infants. Our model has two parallel Multilayer Perceptrons (MLPs) and relies on two components. The first component is a *EM-training rule* that matches the output vectors of a MLP to a statistical distribution. The second component exploits the output classification of one MLP as target of the another MLP in order to learn the agreement of the *unknown* class. We generate four *classless* datasets (based on MNIST) with uniform distribution between the classes. Our model is evaluated against totally supervised and totally unsupervised scenarios. In the first scenario, our model reaches good performance in terms of accuracy and the *classless constraint*. In the second scenario, our model reaches better results against two clustering algorithms.

1 INTRODUCTION

Infants are able to learn the binding between *abstract concepts* to the real world via their sensory input. For example, the abstract concept *ball* is binding to the visual representation of a rounded object and the auditory representation of the phonemes */b/ /a/ /l/*. This scenario can be seen as the *Symbol Grounding Problem* (Harnad, 1990). Moreover, infants are also able to learn the *association* between different sensory input signals while they are still learning the binding of the abstract concepts. Several results have shown a correlation between object recognition (visual) and vocabulary acquisition (auditory) in infants (Balaban & Waxman, 1997; Asano et al., 2015). One example of this correlation is the first hundred words that infants have learned. In that case, the words are mainly nouns, which are *visible concepts*, such as, dad, mom, ball, dog, cat (Gershkoff-Stowe & Smith, 2004). As a result, we can define the previous scenario in terms of a machine learning tasks: learning the association between two parallel streams of data that represent the same *unknown* class (or semantic concept).

2 CLASSLESS ASSOCIATION MODEL

In this work, we are interested in the *classless* association where sample pairs represent different instances of the same *unknown* class. With this in mind, our model has two parallel Multilayer Perceptrons (MLPs) with an EM-training rule (Dempster et al., 1977). We present a novel training rule that matches mini-batches of raw output vectors of each MLP and a target statistical distribution as alternative loss function because of the lack of labels. Moreover, each MLP classifies the raw output vectors based on the statistical distribution. Note that *pseudo-classes* obtained by the classification step change during training. As a result, similar input samples are classified by the same *pseudo-classes*. With this in mind, we have introduced a *weighting vector* that modifies the raw output vector in order to match with the statistical constraint. For learning the agreement between both MLPs, the *pseudo-classes* of one MLP are used as target of the other MLP, and vice versa.

More formally, our task is defined by two disjoint input streams $\boldsymbol{x}^{(1)} \in R^{n1}$ and $\boldsymbol{x}^{(2)} \in R^{n2}$ that represent the same *unlabeled* class. The goal is to learn the association by classifying both with the same *pseudo-class* $c^{(1)} = c^{(2)}$ where $c^{(1)}$ and $c^{(2)} \in R^{n3}$.

Initially, all input samples $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ have random *pseudo-classes* $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$. The histogram of *pseudo-classes* is similar to the desired statistical distribution $\boldsymbol{\phi} \in R^{n3}$ (i.e. uniform). Also, the *weighting* vectors $\boldsymbol{\gamma}^{(1)} \in R^{n3}$ and $\boldsymbol{\gamma}^{(2)} \in R^{n3}$ are initialized to one. For explanation purposes, we have defined two MLPs with one hidden layer

$$\boldsymbol{z}^{(1)} = MLP^{(1)}(\boldsymbol{x}^{(1)}; \theta^{(1)}) \tag{1}$$

$$\boldsymbol{z}^{(2)} = MLP^{(2)}(\boldsymbol{x}^{(2)}; \theta^{(2)}) \tag{2}$$

where $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)} \in R^{n3}$ are the output vectors of each MLP, $\theta^{(1)}$ and $\theta^{(2)}$ are the parameters of each network. We want to point out that the ***E-step*** and ***M-step*** are applied to each MLP indepently.

The ***E-step*** obtains the *pseudo-classes* for each MLP and estimates the current statistical distribution based on a mini-batch of output vectors and weighting vectors[1]. In this case, an approximation of the distribution is obtained by the following equation

$$\hat{\boldsymbol{z}} = \frac{1}{M} \sum_{i=1}^{M} power(\boldsymbol{z}_i, \boldsymbol{\gamma}) \tag{3}$$

where $\boldsymbol{\gamma}$ is the weighting vector, $\boldsymbol{z}_i$ is the output vector of the network, M is the number of elements, and the function $power$[2] is the element-wise power operation between the output vector and the weighting vector. We have used the *power function* because the output vectors are quite similar between them at the initial state of the network, and the *power function* provides an initial boost for learning to separate the input samples in different *pseudo-classes* in the first iterations. Furthermore, we can classify each output vector by retrieving the maximum value of the following equation

$$c^* = arg\ max_c\ power(\boldsymbol{z}_i, \boldsymbol{\gamma}) \tag{4}$$

where c^* is the *pseudo-class*, which are used in the ***M-step*** for updating the MLP parameters. Also, note that the *pseudo-classes* are not updated in an online manner. Instead, the *pseudo-classes* are updated after a certain number of iterations. The reason is the network requires a number of iterations to learn the common features.

The ***M-step*** updates the weighting vector and the MLP parameters. The cost function is the variance between the distribution and the desired statistical distribution, which is defined by

$$cost = (\hat{\boldsymbol{z}} - \boldsymbol{\phi})^2 \tag{5}$$

where $\hat{\boldsymbol{z}}$ is the current statistical distribution of the output vectors, and $\boldsymbol{\phi}$ is a vector that represent the desired statistical distribution, e.g. uniform distribution. Then, the weighting vector is updated via gradient descent, the network parameters are updating using the *pseudo-classes* generated by the other network, and vice versa.

3 Experiments and Results

Our model has been evaluated in four *classless* datasets that were generated from MNIST (Lecun & Cortes, 2010). Each dataset has two disjoint sets *input 1* and *input 2*. The first dataset has two different instances of the same *classless* digit. The other three datasets have a transformation that is applied only to *input 2*, such as, *fix rotation to 90 degrees*, *inverted*, and *random rotation between* 0 *and* 2π. All datasets have a uniform distribution between the digits. The dataset size is 21,000 pair samples for training and 4,000 pair samples for validation and testing. Ten different folds for each dataset has been random generated, and we report the average results of two metrics: *Association Accuracy*[3] and *Purity*. Table 1 shows the *Association Accuracy* between our model and the supervised association task and the *Purity* between our model and two clustering algorithms. First, the supervised association task performances better that the presented model. This was expected because our task is more complex in relation to the supervised scenario. However, we can infer from our results

[1]For explanation purposes, we drop the super-indexes (1) and (2) that represent each stream

[2]We decide to use $power$ function instead of $\boldsymbol{z}_i^{\boldsymbol{\gamma}}$ in order to simplify the index notation

[3]$Association\ Accuracy = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(c_i^{(1)} = c_i^{(2)})$ where N is the number of elements, and $c^{(2)}$, $c^{(1)}$ are the output classification of each network, respectively

Table 1: Association Accuracy (%) and Purity (%) results. Our model is compared against the supervised and unsupervised scenarios.

Dataset	Model	Association Accuracy (%)	Purity (%)	
			input 1	input 2
MNIST	supervised association	96.7 ± 0.3	96.7 ± 0.2	96.6 ± 0.3
	classless association	86.1 ± 3.2	89.6 ± 4.5	89.0 ± 4.2
	K-means	-	63.9 ± 2.2	62.5 ± 3.7
	Hierarchical Agglomerative	-	64.9 ± 4.7	64.3 ± 5.5
Rotated-90 MNIST	supervised association	93.2 ± 0.3	96.4 ± 0.2	96.6 ± 0.2
	classless association	86.5 ± 2.5	82.9 ± 4.5	82.9 ± 4.3
	K-means	-	65.0 ± 2.8	64.0 ± 3.6
	Hierarchical Agglomerative	-	65.4 ± 3.5	64.1 ± 4.1
Inverted MNIST	supervised association	93.2 ± 0.3	96.5 ± 0.2	96.5 ± 0.2
	classless association	89.2 ± 2.4	89.0 ± 6.8	89.1 ± 6.8
	K-means	-	64.8 ± 2.0	65.0 ± 2.5
	Hierarchical Agglomerative	-	64.8 ± 4.4	64.4 ± 3.8
Random Rotated MNIST	supervised association	88.0 ± 0.5	96.5 ± 0.3	90.9 ± 0.5
	classless association	69.3 ± 2.2	75.8 ± 7.3	65.3 ± 5.0
	K-means	-	64.8 ± 2.6	14.8 ± 0.4
	Hierarchical Agglomerative	-	65.9 ± 2.8	15.2 ± 0.5

that the presented model has a good performance in terms of the classless scenario and supervised method. Second, our model not only learns the association between input samples but also finds similar elements covered under the same *pseudo-class*. Also, we evaluate the purity of our model and found that the performance of our model reaches better results than both clustering methods for each set ($input$ 1 and $input$ 2).

4 Conclusion

In this paper, we have shown the feasibility to train a *classless* model that has two parallel MLPs under the following scenario: pairs of input samples that represent the same unknown classes. This scenario was motivated by the *Symbol Grounding Problem* and the association learning between sensory input signal in infants development. Our model relies on the EM-training rule that matches the network's output against a statistical distribution and uses one network as a target of the other network. Our model reaches better performance than two clustering algorithms and good results with respect to the supervised method in terms of unlabeled data. We want to point out that our model was evaluated in an optimal case where the input samples are uniform distributed and the number of classes is known. However, we will explore the performance of our model if the number of classes and the statistical distribution are unknown. One way is to change the number of *pseudo-classes*. This can be seen as changing the number of clusters k in k-means. Furthermore, we are interested in replicating our findings in multimodal datasets like TVGraz (Khan et al., 2009) or Wikipedia featured articles (Rasiwasia et al., 2010).

Acknowledgments

We would like to thank Damian Borth, Christian Schulze, Jörn Hees, Tushar Karayil, and Philipp Blandfort for helpful discussions.

References

Michiko Asano, Mutsumi Imai, Sotaro Kita, Keiichi Kitajo, Hiroyuki Okada, and Guillaume Thierry. Sound symbolism scaffolds language development in preverbal infants. *cortex*, 63:196–205, 2015.

M T Balaban and S R Waxman. Do words facilitate object categorization in 9-month-old infants? *Journal of experimental child psychology*, 64(1):3–26, January 1997. ISSN 0022-0965.

AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.*, 39(1):1–38, 1977.

Lisa Gershkoff-Stowe and Linda B Smith. Shape and the first hundred nouns. *Child development*, 75 (4):1098–114, 2004. ISSN 0009-3920.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

Inayatullah Khan, Amir Saffari, and Horst Bischof. Tvgraz: Multi-modal learning of object categories by combining textual and visual features. In *AAPR Workshop*, pp. 213–224, 2009.

Yann Lecun and Corinna Cortes. The MNIST database of handwritten digits. 2010.

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

SUPPLEMENTAL MATERIAL

We have included several examples of the *classless* training. In addition, we have generated some demos that show the training algorithm (`https://goo.gl/xsmkFD`)

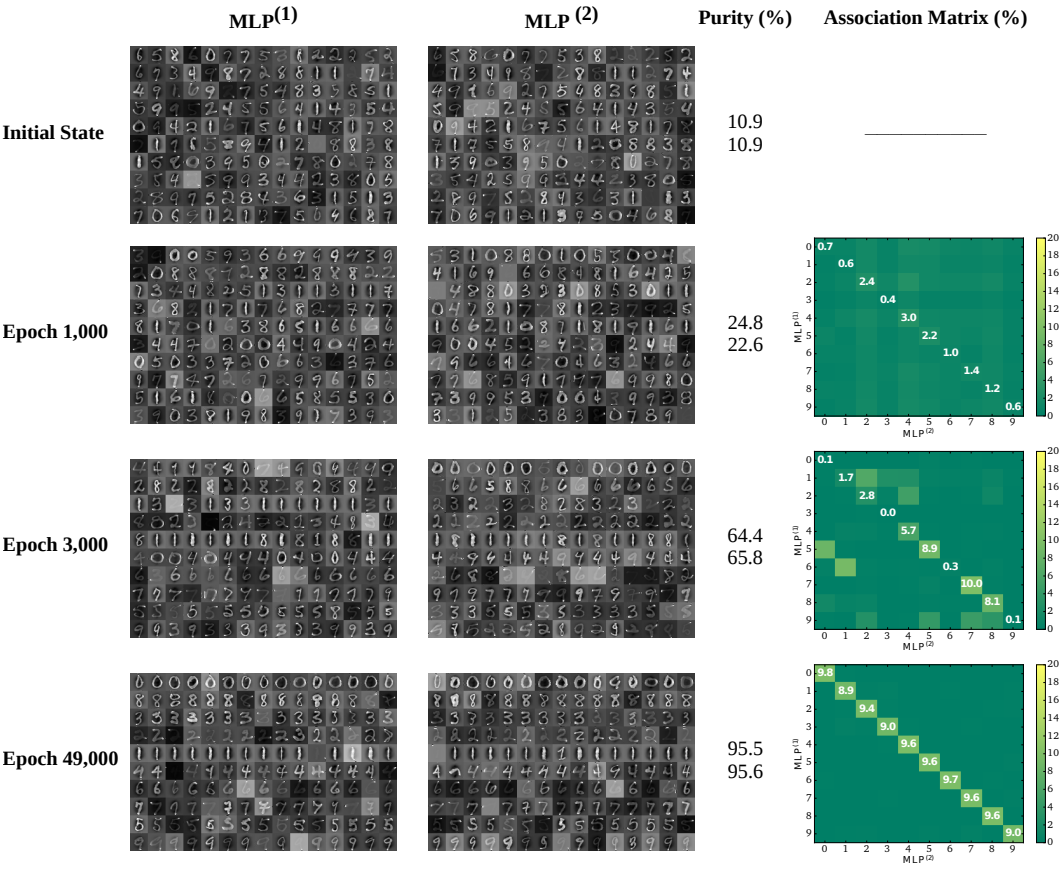

Figure 1: Example of the presented model during *classless* training. In this example, there are ten *pseudo-classes* represented by each row of $MLP^{(1)}$ and $MLP^{(2)}$. Note that the output classification are randomly selected (not cherry picking). Initially, the *pseudo-classes* are assigned randomly to all input pair samples, which holds a uniform distribution (first row). Then, the *classless association* model slowly start learning the features and grouping similar input samples. Afterwards, the output classification of both MLPs slowly agrees during training, and the association matrix shows the relation between the occurrences of the *pseudo-classes*.

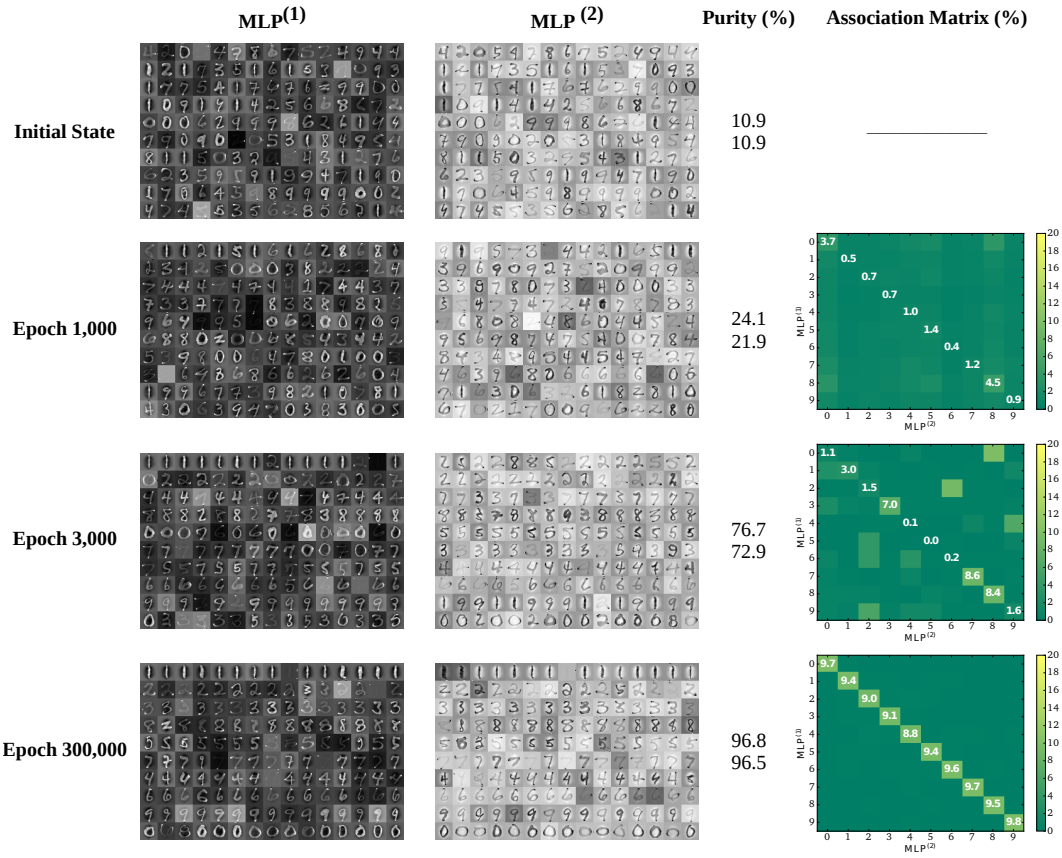

Figure 2: Example of the *classless* training using Inverted MNIST dataset.

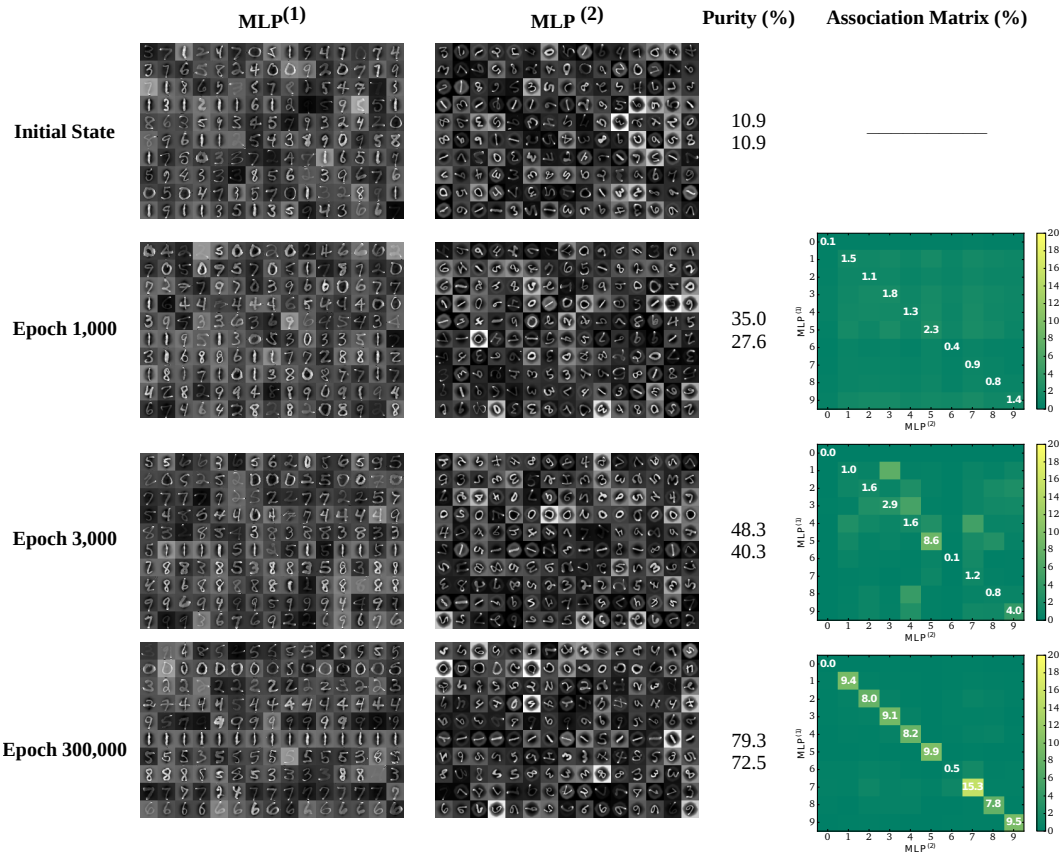

Figure 3: Example of the *classless* training using Random Rotated MNIST dataset.