Auxiliary unsupervied loss for supervised learning

Sabína Samporová a Kristína Malinovská

Katedra aplikovanej informatiky, FMFI,
Univerzita Komenského v Bratislave
Mlynská dolina, 84248 Bratislava
Email: samporova1, rebrova1@uniba.sk

Abstrakt

We present the design of a novel learning method for improved neural network learning that takes into account the data similarities determined by a self-organizing network. We incorporate self-organization as a part of the loss function and show the use of our concept in a new neural network for supervised learning. We compare how the success rate of a classical perceptron increases with this added component of the loss in a classification task. Our results show that with a minimalist model architecture in terms of the hidden layer size, our model convergence is more reliably, compared to the classical multi-layer perceptron. Even the classification accuracy of our model is higher. Our model achieves mean accuracy 89%, while the MLP only 82%.

1 Introduction and motivation

Our overall goal is to improve the semi-supervised learning of deep neural networks. In this paradigm, we use a limited amount of labeled data, but we are also looking for ways to incorporate unlabeled data into the learning. Our improvement lies in the application of topological self-organization. Self-organizing models can learn from unlabeled data, allowing us to make use of all available data. We believe that the additional information from the unlabeled data will support more efficient model learning.

2 SOM

For our task of augmenting the supervised loss with unsupervised mechanism we selected the self-organizing map (SOM) (Kohonen, 1990) as a suitable model for implementing self-organization over the input data. Using SOM, the data is nonlinearly transformed into a 2D map and divided into clusters according to the classes to which they belong. In a well-trained map, similar inputs are represented by data models, prototypes (winning neurons), that are close to each other, within the same cluster, therefore preserving the similarities in the topology of the network. Since this is an unsupervised learning model, there is no need to know the labels of the input data.

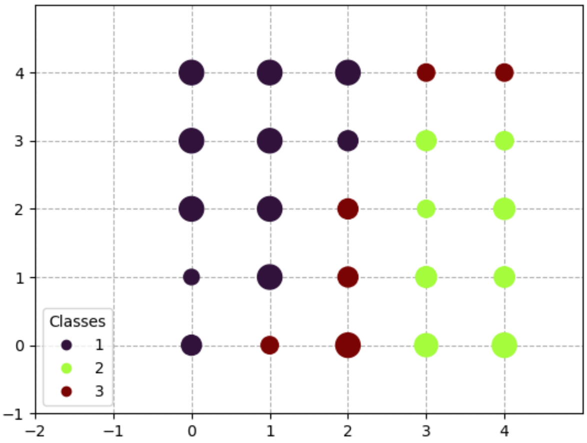

Fig. 1: Self-organizing map trained on Wine dataset

3 Loss function design

We present auxiliary SOM-based loss function J_S. The input of the function are triplets of input data points. Such an approach using triples of data inputs is well known in the field of deep neural networks and has been used primarily for the visual data classification task (Chechik et al., 2010). The first item of i-th triplet is a particular input from a class, which we denote by x_i. The second part is a different input also from the same class, denoted by z_i. The third part is an input that is from a different class than the other two, let us denote it as ξ_i.

We train the SOM using all the items from triplets. During model training, we compute the Euclidean distances D_E between the SOM prototypes (denoted as p) of pairs x_i, z_i and x_i, ξ_i, and combine these distances in one batch as in Eq. 1 and Eq. 2, denoted as D_C and D_I (congruent and incongruent). In a well-trained network, D_C should be small and D_I should be large. Therefore we compute their relative relationship, as given in Eq. 3.

$$D_C = \frac{1}{n}\sum_{i}^{n} D_E(p(x_i), p(z_i)) \tag{1}$$

$$D_I = \frac{1}{n}\sum_{i}^{n} D_E(p(x_i), p(\xi_i)) \tag{2}$$

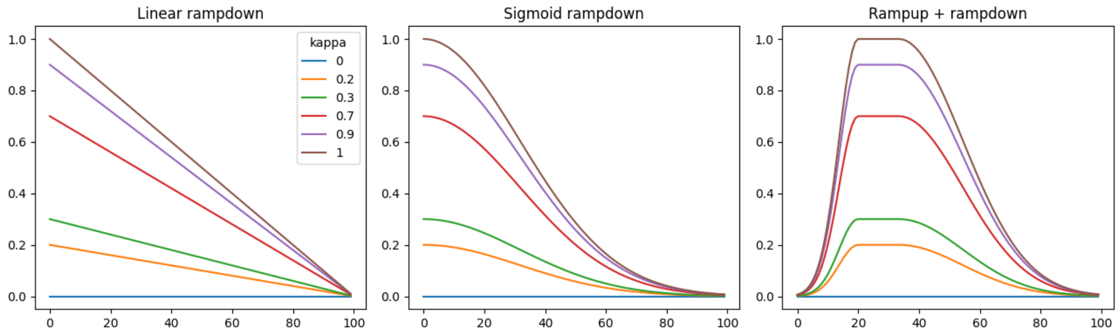

Fig. 2: Investigated ways of ramping the hyperparameter κ in time

Subsequently we rescale this relation to the interval $[0, 1]$ according to the equation 4 and obtain the formula of the loss function whose value is close to zero if D_C is small and D_I is large. If D_C and D_I are different, the value of the loss function increases and approaches to the value 1.

$$D = \frac{D_I - D_C}{D_I + D_C} \tag{3}$$

$$J_S(\tau) = \frac{1}{2} - \frac{1}{2} \cdot D = \frac{1}{2}\left(1 - \frac{D_I - D_C}{D_I + D_C}\right) =$$

$$\frac{1}{2}\left(\frac{D_I + D_C}{D_I + D_C} - \frac{D_I - D_C}{D_I + D_C}\right) = \frac{D_C}{D_I + D_C} \tag{4}$$

4 MLP-SOM model

We propose the new MLP-SOM model that links the multi-layer perceptron (MLP) and the distance-based SOM loss function J_S from pre-trained SOM. The combined loss function (Eq. 6) is a combination of the supervised loss function S and proposed unsupervised loss J_S. The supervised loss is the mean squared error of the model prediction of the first part of the triplet, denoted $g(x_i, \theta)$ and the expected label $\hat{y}$. The loss function J_S is scaled by the parameter $\kappa(t)$, which is linearly ramped down. Its value decreases linearly from a maximum value equal to κ to a value of zero in time. We also take into consideration different ramping strategies such as the Sigmoid function. The $\kappa(t)$ denotes the value of the parameter used in the relation at time t. The maximum value of κ is a hyperparameter that we experimentally explore different values.

$$S(\theta) = \frac{1}{n} \sum_{i}^{n} ||g(x_i, \theta), \hat{y}||^2 \tag{5}$$

$$Loss(\theta) = S(\theta) + \kappa(t) \cdot J_S(\tau) \tag{6}$$

5 Experiments and results

We decided to validate the proposed MLP-SOM model experimentally on a multi-class classification task. The famous tabular dataset we used, Wine dataset, was created by Aeberhard and Forina (1991). We used the pretrained SOM illustrated in Fig.1. Looking at the projection of this SOM we see the map is quite well organized with the individual clusters corresponding to the 3 classes of the learned dataset.

5.1 Wine dataset

The data in the Wine dataset were the result of chemical analysis of wines produced in the same region in Italy, from three different varieties. Each data sample is vector of 13 values. We divided the data into training (75%) and test (25%). Within the training set, we created triples from the data. Each sample was paired 25 times with sample from the same class and sample from a different class. In this way, we obtained a training set containing 6650 triples.

5.2 Compared models

In the experiment, we compared the accuracy of a multilayer perceptron (MLP) baseline with our MLP-SOM model that uses the combined loss function. In both cases, the perceptrons had the same architecture, namely 13 neurons in the input layer, 15 neurons in the hidden layer, and 3 neurons in the output layer. The activation function for the neurons in the input and hidden layers was the Sigmoid and for the output layer we used the Softmax function. The models were implemented using the PyTorch library (Paszke et al., 2019) using the default weight initialization. Other hyperparameters were the learning rate set to 0.0001 and the type of optimizer, which was Adam.

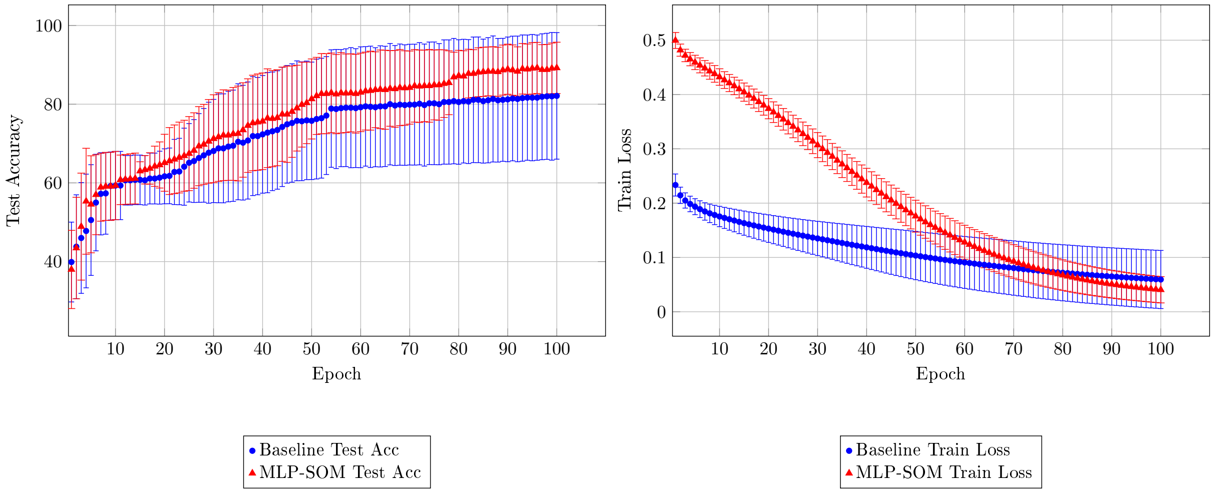

Fig. 3: Evolution of baseline and MLP-SOM model metrics during training (testing accuracy, training loss function)

6 Results of experiments

The results of experiments are the average performance of 20 models after a fixed amount of 100 training epochs. We experimented with different kinds of weight initialization and even smaller architectures. In all experiments, our best performing model was at least 2% better in accuracy than the baseline.

In the main experiment, we investigated different kinds of ramping of the κ and maximum value of κ. Ramps are shown in Fig. 2. The resulting test accuracies of the baseline and models with 5 different κ values are shown in table **??**. In this experiment, we obtained the best result so far with the dataset, namely with the MLP-SOM model ($\kappa = 0.7$) and Sigmoid ramping. The ramping method combining rampup and rampdown was the least successful. For the best model, the test accuracy is $89.22\% \pm 6.53\%$, which is more than 7% better than the baseline model without using the SOM loss function. Fig.3 shows a comparison of the evolution of the two models during training.

7 Conclusion

In our research, we have been developing and testing a method to improve the classification accuracy of a neural network trained with supervised learning. Our method was based on the use of an unsupervised neural model - a self-organizing map that is trained from data that does not need to have labels. We proposed an MLP-SOM model, using predictions from the pre-trained self-organizing map as part of the loss function. We experimentally tested this model on the tabular Wine classification dataset. We investigated various hyperparameters and the results have shown that our model had better testing accuracy, compared to a model that did not use an additional SOM-based loss function. Given this result, we believe that even in the case of semi-supervised learning, where most of the data does not have an associated label, the SOM loss function can be a useful tool to improve the training and increase the classification success of the model.

Acknowledgement

This research was supported by grants VEGA 1/0373/23 and KEGA 022UK-4/2023 and Slovak Society for Cognitive Science[1].

Literatúra

Aeberhard, S. and Forina, M. (1991). Wine. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC7J.

Chechik, G., Sharma, V., Shalit, U., and Bengio, S. (2010). Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3).

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., An-

[1] https://cogsci.fmph.uniba.sk/sskv/

tiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.