Online Representation Learning with Multi-layer Hebbian Networks for Image Classification Tasks

Yanis Bahroun and Andrea Soltoggio

Loughborough University - Department of Computer Science
Loughborough - United Kingdom

Abstract. Unsupervised learning allows algorithms to adapt to different data thanks to the autonomous discovery of discriminating features during the training. When these algorithms are reducible to cost-function minimisation, better interpretations of their learning dynamics are possible. Recently, new Hebbian-like plasticity, bio-inspired, local and unsupervised learning rules for neural networks, have been shown to minimise a cost-function while performing online sparse representation learning. However, it is unclear to what degree such rules are effective to learn features from images. To investigate this point, this study introduces a novel multi-layer Hebbian network trained by a rule derived from a non-negative classical multidimensional scaling cost-function. The performance is compared to that of other fully unsupervised learning algorithms.

1 Introduction

Synaptic plasticity stands as one of the main phenomena responsible for learning and memory. One mechanism of synaptic weight update is inspired by the Hebbian learning principle which strengthens the connections between two units when they are simultaneously active. This principle makes the network learn recurring patterns. Extensions of the Hebb’s rule such as the Oja’s rule [1] or the Generalized Hebbian rule, also called Sanger’s rule [2] proved particularly efficient at tasks such as online dimensionality reduction. Two important properties of brain-inspired models, namely competitive learning [3] and sparse coding [4, 5] can be performed using Hebbian and anti-Hebbian learning rules. Such properties can be achieved with inhibitory connections, which extend the capabilities of such learning rules beyond simple pattern recognition and dimensionality reduction tasks [6]. Formulating the learning problem under a minimisation principle offers a rigorous framework [5] to study the learning dynamics of the network. Moreover, the way the cost-function proposed in [7] is derived leads to corresponding network architectures.

This study employs a Hebbian learning rule derived from a cost-function and applies it to perform online unsupervised learning of features from multiple image datasets. One of the main focus of this work is to find an online learning principle that is suitable for dealing with a continuous stream of data. Such an architecture can take one image at a time with memory requirements that are independent of the number of samples. A recent model of Hebbian/anti-Hebbian neural network [7] was reproduced here and applied for the first time to online feature learning for image classification. The quality of the features is assessed visually and by performing classification with a linear classifier working on the learned features. The simulations show that a simple single-layer Hebbian network can outperform more complex models such as Sparse Autoencoders (SAE) and Restricted Boltzmann machines (RBM) for image classifications.
tasks. When tested with with the novel multi-layer architecture, the features learned in the different layers improve the classification accuracy, indicating that further layers help to learn additional relevant features.

This study is the first of its kind to perform deep sparse dictionary learning based on the similarity matching principle developed in [7]. Unlike SAE or RBM, the different layers of the network evaluate the pairwise similarity of the input and not a reconstruction error. Moreover, the learning principle based on the multidimensional scaling cost-function considers input similarities also strongly relates to the representational similarity analysis developed in [8] which appears critical for understanding the IT cortex. One further advantage of the algorithm is that it is fully unsupervised and does not require any semi labelling nor data-augmentation.

2 Hebbian/anti-Hebbian network derived from a similarity matching cost-function

The classical multidimensional scaling (CMDS) is a fundamental information analysis tool with applications widely spread from ecology, education to neuroscience [9]. It takes as input a matrix of distances or dissimilarity, and generates a set of embedding coordinates in a lower dimensional, Euclidean space [10]. In its simplest form, CMDS and PCA are equivalent [10], producing dense features maps which are often less suitable for image classification than sparse features maps. With sparse encoding, the representations are less correlated than with dense features, which in turn leads to higher classification accuracy when linear classifier are used. Learning sparse encoding is often implemented with non-negativity constraints, which ensure efficiency and biological plausibility.

Recently, [7] introduced a non-negative classical multidimensional scaling model which allowed the derivation of a new biologically plausible Hebbian model. The Hebbian/anti-Hebbian rule introduced in [7, 11] is explained in the following. For a set of input $x^t \in \mathbb{R}^n$ for $t \in \{1, \ldots, T\}$, the concatenation of the inputs defines an input matrix $X \in \mathbb{R}^{n \times T}$. Similarly, the output matrix is $Y \in \mathbb{R}^{m \times T}$. The objective function proposed by [7] is:

$$Y^* = \arg \min_{Y \geq 0} \|X'X - Y'Y\|_F^2.$$  \hfill (1)

Solving Eq.1 requires to store $Y \in \mathbb{R}^{m \times T}$ which increases with time $T$ making online learning difficult. The online learning version of Eq.1 is expressed as:

$$(y^T)^* = \arg \min_{y^{T \geq 0}} \|X'X - Y'Y\|_F^2.$$  \hfill (2)

The solution of Eq.2, as shown in [7], can be solved by coordinate descent

$$(y^T_i)^* = \max \left( W^T_i x^T - M^T_i y^T \right) \forall i \in \{1, \ldots, m\}$$  \hfill (3)
The update rules of $W^T$ and $M^T$ can be expressed using recursive formulations:

$$W_{ij}^T = W_{ij}^{T-1} + \frac{\left( y_{i}^{T-1} (x_{j}^{T-1} - W_{ij}^{T-1} y_{i}^{T-1}) \right)}{\hat{Y}_{i}^{T}}$$ (5)

$$M_{ij \neq i}^T = M_{ij}^{T-1} + \frac{\left( y_{i}^{T-1} (y_{j}^{T-1} - M_{ij}^{T-1} y_{i}^{T-1}) \right)}{\hat{Y}_{i}^{T}}$$ (6)

$$\hat{Y}_{i}^{T} = \hat{Y}_{i}^{T-1} + (y_{i}^{T-1})^2$$ (7)

$W^T$ and $M^T$ can be interpreted respectively as feed-forward synaptic connections and lateral synaptic inhibitory connections. The weight matrices are of fixed size and updated sequentially, which makes the model suitable for online learning. The architecture of the Hebbian/anti-Hebbian network is represented in Figure 1.

![Hebbian/anti-Hebbian network](image)

Fig. 1: Hebbian/anti-Hebbian network with lateral connections derived from Eq.2

### 3 Learning features from images

The learning rule expressed by Eqs. 3-5-6-7 is applied here with single and multi-layer architectures and tested for the first time on image classification tasks. The learning principle used in the following resembles a variant of the Hebbian principle presented in [5] performing alternating minimization. When a new input $x^T$ is presented, the model first computes a sparse post-synaptic activity $y^T$, second the synaptic weights are modified based on local Hebbian/anti-Hebbian learning rules requiring only the current activity of its pre- ($x^T$) and postsynaptic ($y^T$) neuronal activities. The model presented can be seen as a sparse encoding followed by a recursive updating scheme, which are both well suited to solve large-scale online problems.

When used with multiple resolutions, the model proposed here can be seen as a biologically plausible, fully unsupervised and online inception network proposed in the GoogLeNet ([12]). An important aspect of this new learning network is that it does not require batch learning, which in turn reduces the memory requirement of the learning system.
3.1 Multi-layer Hebbian/anti-Hebbian neural network combined with a linear classifier

In the proposed approach, layers of Hebbian/anti-Hebbian network are stacked similarly to the Convolutional DBN [13], and Hierarchical K-means [14]. However, in the multi-layer Hebbian network, both the weights of the first layer and second layer are continuously updated. In between layers, a simple average pooling is used to down-sample the feature maps. Unlike other convolutional neural networks, the non-linearity used in each layer is not only due to the positivity constraint. In this case, it is due to the combination of a rectified linear unit (ReLU) activation function and of interneurons competition. Unlike with the K-means [15] or with CNNs, the algorithm does not use a cross-validated fixed threshold function which reduces the number of meta-parameters of the model. This model combines the powerful architecture of convolutional neural networks using ReLU activation with interneurons competition, while all synaptic weights are updated using online local learning rules.

A (L2) multi-class SVM classifies the pictures using the features learned by the neural network. The regularisation parameter is determined by cross-validation. A simple form of pooling is used to feed the classifier. The output vector is pooled over the four equal-sized quadrant of the image [15].

3.2 Overcompleteness of the representation and multi-resolution

If the number of neurons exceeds the size of the input, the representation is overcomplete. Overcompleteness may be beneficial but requires increased computation, particularly for deep networks in which the number of neurons in layer \( N > 2 \) has to be bigger than the product of the size of a feature patch and the number of neurons in layer \( N = 1 \). One motivation for overcompleteness is that it may allow more flexibility in matching the output structure to the input. However, not all learning algorithms can learn and take advantage of overcomplete representations. Overcompleteness is often a characteristic shared by models performing sparse coding. The behaviour of the algorithm is analysed in the transition between undercomplete and overcomplete representations.

The advantages of the algorithm are still mitigated by the number of operations required by the coordinate descent when the number of neurons increases. The multi-resolution model proposed trains simultaneously three single-layer neural networks, each of them having different receptive field sizes (\( 4 \times 4, 6 \times 6, \) and \( 8 \times 8 \) pixels). The model would produce three different dictionaries requiring less computational time and memory since the synaptic weights only connect neurons within each neural network.

4 Results

The effectiveness and validity of the algorithm is assessed by measuring the performance on an image classification task. We acknowledge that classification accuracy is at best an implicit measure evaluating the performance of representation learning algorithms, but provides a standardised way of comparing them.

A single-layer Hebbian/anti-Hebbian neural network combined with the standard multi-class SVM is trained on the CIFAR-10 dataset [16]. Brightness and contrast
normalisation were applied to the input images. Although there exist Hebbian networks that can perform online input decorrelation [6, 11], an offline whitening technique based on singular value decomposition [15] is applied in these experiments. Figure 2a and Figure 2b show the features learned by the network from raw input and whitened input respectively. The features learned from raw data (Fig 2a) are neither sharp nor localised filters and just slightly capture edges. With whitened data (Fig 2b), the features are sharp, localised, and resemble Gabor filters. These results match those reported with clustering algorithms [15].

Fig. 2: Sample of features learned from raw (2a) and whitened input (2b). Classification accuracy with (2c) raw and (2d) whitened input.

The performance of the network was tested for varying receptive field sizes (Fig. 2c-2d) and varying network sizes (400, 500, 600, and 800 neurons). The results show that the performance peaks at a receptive field size of 7 pixels and then begins to decline. Similar results were reported with other unsupervised learning algorithms [15]. One could justify this result by the difficulty of learning spatially extended features. Figures 2c and 2d also show that for every configuration, the performance of the algorithm is largely and uniformly improved when whitening is applied to the input. Finally, the step-size between extracted features (stride) is a meta-parameter influencing the classification accuracy. Preliminary experiments (not shown) indicated that a stride of 1 was optimal, which is used for all the experiments presented in this work.

4.1 Comparison to state-of-the-art performances and online training

Various unsupervised learning algorithms were tested on the CIFAR-10 dataset. Spherical K-means, in particular, proved in [15] to outperform autoencoders and restricted Boltzmann machines, providing a very simple and efficient solution for dictionary learning for image classification. Thus, spherical K-means is used here as a benchmark to evaluate the performance of a single-layer Hebbian/anti-Hebbian neural network. Similarly to other unsupervised learning algorithms, increasing the number of output
neurons to reach overcompleteness also improved classification performance (Fig 3a). Although the Hebbian network has theoretically a higher degree of sparsity than the K-means proposed in [15] (results not shown here), they appear to have the same performance in their optimal configurations (Fig 3a). The model can be set to have varying sparsity by influencing the matrix $M^T$ in the feature extraction phase, therefore decreasing the competition between neurons. In the case of $M^T$ being set to zero, the model Eq 3 becomes a simple linear neuron model with ReLU activation function for which performance is well known when trained using back-propagation.

The classification accuracy of the network during training is shown in Fig 3b. The graph (Fig 3b) suggests that the features learned by the network over time help the system improve the classification accuracy. This is significant because it demonstrates for the first time the effectiveness of features learned with a Hebb-like cost-function minimisation. It is not obvious a priori that optimising online a cost-function for sparse similarity matching (Eq 2) produces features suitable for image classification.

Fig. 3: Classification accuracy, Hebbian vs K-means.

As shown in Table 1, the multi-resolution network outperforms the single resolution Hebbian and K-means algorithm [15], reaching 80.42% accuracy on the CIFAR-10. It also outperforms the single layer NOMP [17], sparse TIRBM [18], the CKN-GM, and the CKN-PM [19], which are far more complex models. The multi-resolution model proves to show better performances while requiring less computation and memory than the single resolution model.

4.2 Multi-layer Hebbian Network

Learning overcomplete representations at every stage of a multi-layer network is a challenging task due to the increasing number of neurons required. A double-layer neural network with different numbers of neurons in each layer was trained similarly to the single-layer network in the previous section. In Table 2, $\phi_1$ and $\phi_2$ represent respectively the features learned by the first and second layer. The results show that $\phi_2$ alone are less discriminative than $\phi_1$ for the same number of neurons. However, when the combined ($\phi_1 + \phi_2$) the model achieves better performance than each layer considered separately. A future test may analyse whether the second layer learns larger-scale features than the first layer, a finding that could explain the improved classification accuracy with the multi-layer neural network.
Table 1: Comparison with unsupervised learning algorithms on CIFAR-10.

| Algorithm                                      | Accuracy  |
|------------------------------------------------|-----------|
| Single Layer Hebbian, Single Resolution        | 79.58 %   |
| Single-Layer Hebbian, Multi-Resolution         | 80.42 %   |
| Single-layer K-means [15]                      | 79.60 %   |
| Multi-layer K-means [15]                       | 82.60 %   |
| Sparse RBM                                     | 72.40 %   |
| Convolutional DBN [13]                         | 78.90 %   |
| Sparse TIRBM [18]                              | 80.10 %   |
| TIOMP-1/T [18]                                 | 82.20 %   |
| Single Layer NOMP [17]                         | 78.00 %   |
| Multi-Layer NOMP [17]                          | 82.90 %   |
| Multi-Layer CKN-GM [19]                        | 74.84 %   |
| Multi-Layer CKN-PM [19]                        | 78.30 %   |
| Multi-Layer CKN-CO [19]                        | 82.18 %   |

Table 2: Classification accuracy for a two-layer Hebbian/anti-Hebbian network

| #Neurons Layer 2 | 50 | 100 | 200 | 400 | 800 |
|------------------|----|-----|-----|-----|-----|
| 100 Neurons Layer 1 | $\phi_2$ | 54.9% | 59.7% | 64.7% | 68.7% | 71.45% |
|                   | $\phi_1 + \phi_2$ | 67.2% | 68.1% | 69.9% | 72.4% | 73.81% |
| 200 Neurons Layer 1 | $\phi_2$ | 55.8% | 60.6% | 65.3% | 70.3% | 72.7% |
|                   | $\phi_1 + \phi_2$ | 69.9% | 70.8% | 71.9% | 73.7% | 75.1% |

5 Conclusion

This work proposes a multi-layer neural network exploiting a set of Hebbian/anti-Hebbian rules to learn features for image classification. The network is trained on the CIFAR-10 image dataset prior to feeding a linear classifier. The model successfully learns online more discriminative sparse representations of the data when the number of neurons and the number of layers increase. We observed in this work that the over-completeness of the representation is critical for learning relevant features. We have also shown that a minimum unsupervised learning time is needed to optimise the Hebbian network for sparse feature extraction. Finally, one key factor in improving image classification is the appropriate choice of the receptive field size used for training the network.

Such findings prove that neural networks can be trained to solve problems as complex as sparse dictionary learning with Hebbian learning rules, delivering competitive accuracy compared to other encoder, including deep neural networks. This makes deep Hebbian networks attractive for building large-scale image classification systems. While showing competitive performances on the CIFAR-10, the network can offer an alternative to batch trained neural networks. Ultimately, thanks to its bio-inspired architecture and learning rules, it also stands as a good candidate for memristive devices.
Moreover, if a “decaying” factor is added to the proposed model that might result in an algorithm that can deal with complex datasets with temporal variations of the distributions.

References

[1] E. Oja. Neural networks, principal components, and subspaces. *International journal of neural systems*, 1(01):61–68, 1989.

[2] T. D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989.

[3] D. E. Rumelhart and D. Zipser. Feature discovery by competitive learning. *Cognitive science*, 9(1):75–112, 1985.

[4] B. A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[5] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

[6] M. D. Plumley. A Hebbian/anti-Hebbian network which optimizes information capacity by orthonormalizing the principal subspace. In *Artificial Neural Networks, 1993.*, *Third International Conference on*, pages 86–90. IET, 1993.

[7] C. Pehlevan and D. B. Chklovskii. A Hebbian/anti-Hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 769–775. IEEE, 2014.

[8] N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[9] I. Borg and P. J. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[10] T. F. Cox and M. A. Cox. *Multidimensional scaling*. CRC press, 2000.

[11] C. Pehlevan and D. Chklovskii. A normative theory of adaptive dimensionality reduction in neural networks. In *Advances in Neural Information Processing Systems*, pages 2269–2277, 2015.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[13] A. Krizhevsky and G. Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40, 2010.

[14] X. Hu, P. Qi, and B. Zhang. Hierarchical k-means algorithm for modeling visual area v2 neurons. In *International Conference on Neural Information Processing*, pages 373–381. Springer, 2012.

[15] A. Coates, H. Lee, and A. Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS 2011*, volume 1001, 2011.

[16] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[17] T.-h. Lin and H. Kung. Stable and efficient representation learning with nonnegativity constraints. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1323–1331, 2014.

[18] K. Sohn and H. Lee. Learning invariant representations with local transformations. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1311–1318, 2012.

[19] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid. Convolutional kernel networks. In *Advances in Neural Information Processing Systems*, pages 2627–2635, 2014.

[20] J. H. Poikonen and M. Laiho. Online linear subspace learning in an analog array computing architecture. *CNNA 2016*, 2016.