

SELF-ORGANIZATION ADDS APPLICATION ROBUSTNESS TO DEEP LEARNERS

Anonymous authors

Paper under double-blind review

ABSTRACT

While self-organizing principles have motivated much of early learning models, such principles have rarely been included in deep learning architectures. Indeed, from a supervised learning perspective it seems that topographic constraints are rather detrimental to optimal performance. Here we study a network model that incorporates self-organizing maps into a supervised network and show how gradient learning results in a form of a self-organizing learning rule. Moreover, we show that such a model is robust in the sense of its application to a variety of areas, which is believed to be a hallmark of biological learning systems.

1 INTRODUCTION

Machine learning has made significant improvements, specifically with deep neural network models (LeCun et al., 2015), (Bengio, 2009), (Goodfellow et al., 2016). Deep learning was made possible by much faster computer technology such as GPUs, and with algorithmic advancement such as (Srivastava et al., 2014) (Goodfellow et al., 2013) (Hochreiter, 1998), (Glorot et al., 2011). Learning tasks that, due to their complexity or data volume, were impossible to execute a decade ago, now can be run in reasonable time scale. The improvements allow the applications of Deep Learning to many real world problems.

Learning good internal representations is a key aspect of deep learning. Indeed, it is interesting to recall that the first breakthrough in deep learning came from an application of unsupervised pre-training with gradient-based fine tuning (Hinton & Salakhutdinov, 2006). Restricted Boltzmann Machines (RBMs) (Hinton, 2002) and Autoencoders (Bourland & Kamp, 1988) (Hinton & Zemel, 1994) are utilized for constructing the hidden layers of early models such as Deep Belief Networks (DBN) and Deep Boltzmann Machine (DBM), (Hinton et al., 2006), (Salakhutdinov, 2015). While much research of deep learning research focuses on learning efficiency and running performances, there is much less research into the understanding of the formation of internal representation in hierarchical neural networks. Moreover, while self-organizing maps have been an integral part of biologically motivated learning theories since the 1970s (Willshaw & Von Der Malsburg, 1976), (Kohonen, 1982) the role of such self-organizing mechanisms are less understood in modern deep learning theories. Topographical self-organization is often observed in biological neural networks (Hubel & Wiesel, 1962), (Romani et al., 1982) and thus may give new insights in understanding learning and self-organization in artificial neural networks.

Here, we propose a network that combines aspects of self-organization into a supervised network model for classification. More specifically, in this study we modify the previously proposed Restricted Radial Basis Function Networks (rRBF) (Hartono et al., 2015) with a softmax output layer that is trained on a crossentropy cost function. This is more consistent with a probabilistic interpretation of the class membership output function than the previous implementation which allows a more clear derivation of the emergence of the self-organizing learning aspects of this network. We call this modified network the Softmax Restricted Radial Basis Function Networks (S-rRBF). Through this network we argue that it is possible to build a learning model in which unsupervised self-organization and supervised learning are just different aspects of a single learning mechanism. We show that the network achieves compatible performance with other deep network architectures while having the added feature of robustness in the sense that it compares favorably with the best performers in the studied examples, while the best performer changes for different applications. While the results are consistent with Wolpert & Macready (1997) ‘No free lunch theorem’, they also

highlight that robustness against variation of applications and not the best performance is an important part in flexible learner, which is thought to be of importance when understanding biological learning systems.

We highlight our ideas here with well understood applications examples of moderate complexity. However, the proposed architecture can also be scaled to deeper layers and hence applied to deeper learning problems. The main contribution here is showing algebraically the emergence of the self-organizing structures from supervised gradient learning. We believe that this research opens new insights into the relation between unsupervised and supervised learning. Also, we illustrate on some examples the internal representation in the competitive layer and compare it to a standard self-organizing map (SOM) (Kohonen, 1982) and to t-Stochastic Neighborhood Embedding (t-SNE) (van der Maaten, 2008) that represents a deeper transformation of the feature space.

2 SOFTMAX RESTRICTED RADIAL BASIS FUNCTION NETWORKS

Softmax Restricted Radial Basis Function Networks (S-rRBF) is a hierarchical neural network that has one or more hidden layers where the neurons are aligned in a two dimensional grid. For simplicity we will restrict our discussion to networks with one hidden layer as shown in Fig. 1. The S-rRBF is developed based on Restricted Radial Basis Function Networks (rRBF) introduced in (Hartono et al., 2015). Here, unlike the original rRBF that has a sigmoidal output layer and quadratic cost function, S-rRBF adopts the softmax output layer with a cross entropy cost function. These modifications yields clearer understanding on the relation between the internal self-organization with supervised learning process. So far, most studies treat self-organizing and supervised learning as two unrelated learning mechanisms. Here, we argue that with the proposed S-rRBF it is possible to build a learning model in which self-organization is an integrated process of supervised learning, and thus giving a new perspective on the learning process of artificial neural networks.

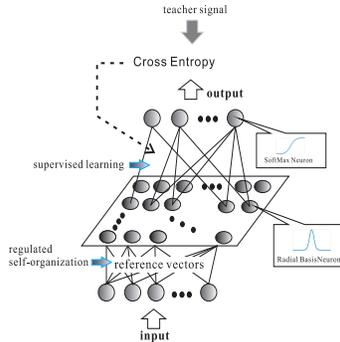


Figure 1: Outline

The dynamics of the S-rRBF is as follows. Suppose the S-rRBF is trained on a data set $\{(\mathbf{X}^i, Y^i)\} (i = 1, 2, \dots, m)$, in which $\mathbf{X}^i \in \mathbb{R}^d$ and $Y^i \in \{1, 2, \dots, C\}$, and d is the dimension of the input while C is the number of classes and thus the number of output neurons.

Given input, \mathbf{X}^i , the j -th hidden neuron generates output, h_j^i , as

$$\begin{aligned} h_j^i &= \sigma(\text{win}^i, j) e^{-\|\mathbf{X}^i - \mathbf{W}_j\|^2} \\ \text{win}^i &= \arg \min_j \|\mathbf{X}^i - \mathbf{W}_j\|^2. \end{aligned} \quad (1)$$

In Eq. 1, $\sigma()$ is a neighborhood function defined as

$$\sigma(\text{win}^i, j, t) = e^{-\frac{\text{dist}(\text{win}^i, j)}{S(t)}} \quad (2)$$

$$\begin{aligned} S(t) &= S_{start} \left(\frac{S_{end}}{S_{start}} \right)^{\frac{t}{t_{end}}} \\ (0 \leq t \leq t_{end}, S_{start} > S_{end}), \end{aligned} \quad (3)$$

where $dist(win, j, t)$ is the Euclidean distance between the winning neuron and the j -th neuron on the two-dimensional grid of the hidden layer, while t , and t_{end} , is the current epoch, and the target epoch when the learning process is terminated.

The activation function of a hidden neuron in S-rRBF is similar to that of Radial Basis Function Networks (RBF) (Poggio & Girosi, 1990) except that in S-rRBF it is topologically restricted by the neighborhood function $\sigma(win, j)$.

The output of the hidden neurons are then propagated to the output layers, where the k -th output, O_k , in the output layer is defined as

$$O_k = e^{\mathbf{V}_k^T \mathbf{h}^i}. \quad (4)$$

The conditional probability that the S-rRBF classifies the input into the class k is given by

$$P(Y^i = k | \mathbf{W}, \mathbf{V}, \mathbf{X}^i) = \frac{e^{\mathbf{V}_k^T \mathbf{h}^i}}{\sum_l e^{\mathbf{V}_l^T \mathbf{h}^i}}. \quad (5)$$

The S-rRBF is then trained to minimize the cross entropy,

$$J(\mathbf{W}, \mathbf{V}) = - \sum_i P(Y^i) \log P(Y^i | \mathbf{W}, \mathbf{V}, \mathbf{X}^i) \quad (6)$$

Considering that $Y^i \in \{1, \dots, C\}$, Eq. 6 can be rewritten as

$$J(\mathbf{W}, \mathbf{V}) = - \sum_i \sum_k \Pi(Y^i = k) \log \frac{e^{\mathbf{V}_k^T \mathbf{h}^i}}{\sum_l e^{\mathbf{V}_l^T \mathbf{h}^i}}. \quad (7)$$

In Eq. 7, $\Pi(Y^i = k) = 1$ when $Y^i = k$ is true, and otherwise $\Pi(Y^i = k) = 0$ otherwise.

To minimize the cross entropy, its gradients are calculated as

$$\frac{\partial J}{\partial \mathbf{V}_j} = - \sum_i (\Pi(Y^i = j) \mathbf{h}^i - (\sum_k \Pi(Y^i = k)) \frac{e^{\mathbf{V}_j^T \mathbf{h}^i}}{\sum_l e^{\mathbf{V}_l^T \mathbf{h}^i}} \mathbf{h}^i).$$

Because $(\sum_k \Pi(Y^i = k)) = 1$,

$$\frac{\partial J}{\partial \mathbf{V}_j} = - \sum_i (\Pi(Y^i = j) - P(Y^i = j | \mathbf{W}, \mathbf{V}, \mathbf{X}^i)) \mathbf{h}^i. \quad (8)$$

Hence, the modification of the weight vector leading to the j -th output neuron is as follows.

$$\mathbf{V}_j(t+1) = \mathbf{V}_j(t) + \eta \sum_i (\Pi(Y^i = j) - P(Y^i = j | \mathbf{W}, \mathbf{V}, \mathbf{X}^i)) \mathbf{h}^i. \quad (9)$$

Eq. 9 shows that the values of connection weights leading to an output neuron are increased if that neuron is associated with the true label of the input and are decreased otherwise. Consequently these modifications increase the probability that the S-rRBF predicts the correct class.

Also,

$$\frac{\partial J}{\partial \mathbf{W}_n} = \frac{\partial J}{\partial \mathbf{h}^i} \frac{\partial \mathbf{h}^i}{\partial \mathbf{W}_n}. \quad (10)$$

In calculating Eq. 10, considering the weight vector \mathbf{W}_n is only relevant to the output of the n -th hidden neuron, h_n , the equation can be rewritten as

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{W}_n} &= \frac{\partial J}{\partial h_n^i} \frac{\partial h_n^i}{\partial \mathbf{W}_n} \\ &= - \sum_i \sum_k \frac{\partial}{\partial h_n^i} (\Pi(Y^i = k) (\log e^{\mathbf{V}_k^T \mathbf{h}^i} - \log \sum_l e^{\mathbf{V}_l^T \mathbf{h}^i})) \frac{\partial h_n^i}{\partial \mathbf{W}_n} \\ &= - \sum_i \sum_k \Pi(Y^i = k) (v_{kn} - \sum_l v_{ln} P(Y^i = l | \mathbf{W}, \mathbf{V}, \mathbf{X}^i)) \frac{\partial h_n^i}{\partial \mathbf{W}_n}. \end{aligned} \quad (11)$$

In Eq. 11, v_{kn} is the weight connecting the n -th hidden neuron with the k -th output neuron. Hence, $\sum_l v_{ln} P(Y^i = l | \mathbf{W}, \mathbf{V}, \mathbf{X}^i)$ is the weighted average of the connection weights from the n -th hidden neuron to the output layer, with the conditional probabilities of the all possible classes being selected as the predicted class by the S-rRBF as the weighting coefficients.

Defining, $\tilde{v}_n = \sum_l v_{ln} P(Y^i = l | \mathbf{W}, \mathbf{V}, \mathbf{X}^i)$, Eq. 11 can be expressed as

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{W}_n} &= - \sum_i \sum_k \Pi(Y^i = k) (v_{kn} - \tilde{v}_n) \frac{\partial h_n^i}{\partial \mathbf{W}_n} \\ &= -2 \sum_i \sum_k \Pi(Y^i = k) (v_{kn} - \tilde{v}_n) h_n(\mathbf{X}^i - \mathbf{W}_n). \end{aligned} \quad (12)$$

When the true class of the given input \mathbf{X}^i is K , hence $\Pi(Y^i = K) = 1$ and 0 for all other classes, Eq. 12 becomes as follows,

$$\frac{\partial J}{\partial \mathbf{W}_n} = -2 \sum_i (v_{Kn} - \tilde{v}_n) h_n^i(\mathbf{X}^i - \mathbf{W}_n) \quad (13)$$

Hence the modification of the n -th reference vector is given by

$$\begin{aligned} \mathbf{W}_n(t+1) &= \mathbf{W}_n(t) + \eta \sum_i (v_{Kn} - \tilde{v}_n) h_n^i(\mathbf{X}^i - \mathbf{W}_n) \\ &= \mathbf{W}_n(t) + \eta \sum_i (v_{Kn} - \tilde{v}_n) \sigma(win^i, n) e^{-\|\mathbf{X}^i - \mathbf{W}_n\|^2} (\mathbf{X}^i - \mathbf{W}_n). \end{aligned} \quad (14)$$

Eq. 14 shows that a self-organizing process, similar to that of Kohonen's SOM (Kohonen, 1982) shown in Eq. 15, occurs in the internal layer during the supervised training process of S-rRBF. The self-organization occurs as a mathematical implication of the cost function minimization. It shows that it is possible to link topological self-organization and supervised learning, which are often treated as different learning mechanism, in a single supervised learning model.

$$\mathbf{W}_n(t+1) = \mathbf{W}_n(t) + \eta \sum_i \sigma(win^i, n) (\mathbf{X}^i(t) - \mathbf{W}_n(t)) \quad (15)$$

The recent surge of deep learning triggers a natural interest in learning representations. Representations extracted by Autoencoders, RBMs and other recent deep models have been extensively studied. However, although often observed in biological neural networks, so far topographical representations in hierarchical learning models of artificial neural networks have not been well studied. Here, we show that a topographical structure is feasible for internal representation in hierarchical supervised learning of neural networks. It is important to mention that the internal self-organization here is different from that of a SOM, in that in a SOM, as shown in Eq. 15, the reference vector is always modified towards the input vector while the direction of self-organization in S-rRBF is regulated by the sign of $(v_{Kn}(t) - \tilde{v}_n(t))$. The sign of this regularization term is decided by the relative value of the weight connecting the neuron associated with the n -th reference with the output associated with the true class of the input. If the weight leading to the output neuron associated with the true class is larger than the expected value of weight connection leading from the neuron associated with the n -th reference vector, a "positive" self-organization as in SOM occurs, while if the value of the weight is below average a "negative" self-organization that moves the reference vector away from the input occurs. In this context, the self-organization process in SOM is label-free, while in S-rRBF it is label-oriented.

While the internal self-organization in this study is not fully unsupervised, as it depends on the labels of the input, it does not require the exact information of the output error but only relative value of connection weight from a particular hidden neuron leading to the output neuron. Hence, it is not supervised in the strict sense either. We consider that the semi-supervised self-organization here is a good starting point in further study to connect unsupervised learning with the supervised learning scheme.

Furthermore, the term $e^{-\|\mathbf{X}^i - \mathbf{W}_n(t)\|^2}$ in Eq. 14 also triggers a dropout effect (Glorot et al., 2011), (Srivastava et al., 2014), resulting in a sparse network in that hidden neurons associated with reference vectors that differ greatly from the input \mathbf{X} are inhibited.

Table 1: Performance in terms of Error Rate (%) (Standard Deviation)

of the different methods on a series of standard machine learning benchmark programs.

Dataset	S-rRBF	DBN	SAEs	ReLU MLP
Abalone	27.8 (1.9)	27.8 (2.2)	26.4 (2.4)	26.6 (3.6)
Activity Log	6.4 (1.4)	11.4 (1.8)	1.3 (0.3)	21.9 (8.4)
Balance	9.6 (2.6)	11.4 (5.3)	1.1 (2.5)	2.9 (2.7)
Bank Marketing	10.5 (1.2)	11.5 (1.4)	13.3 (1.6)	10.8 (1.5)
Breast Cancer	2.5 (1.8)	2.6 (2.3)	3.7 (2.6)	3.2 (2.3)
Cardiotocography	9.6 (2.4)	13.3 (3.2)	9.6 (2.1)	11.6 (4.2)
Hearth	15.2 (5.4)	16.3 (7.0)	22.2 (8.4)	15.6 (10.1)
Iris	1.3 (2.8)	2.7 (4.7)	3.3 (3.5)	4.7 (6.4)
Spambase	7.6 (1.5)	8.9 (1.4)	6.0 (1.3)	11.8 (8.4)
Waveform	13.4 (1.6)	14.2 (1.0)	15.0 (1.9)	13.8 (2.2)
Wine Quality	21.7 (2.3)	25.4 (1.7)	22.5 (2.2)	22.1 (1.7)
MNIST	7.0 (0.3)	18.4 (0.9)	8.3 (0.6)	16.2 (5.5)
MNIST Fashion	11.3 (0.68)	13.9 (0.6)	10.9 (0.7)	41.6 (10.6)

3 EXPERIMENTS

We tested the S-rRBF on a variety of standard machine learning benchmark problems and compared it against three deep learning models, namely a Deep Belief Network (DBN), a Stacked Autoencoder (SAE), and a ReLU MLP with softmax output layers and crossentropy cost function. The average classification error rates over 15-fold cross validation test are shown in Table 1. In those experiments, the number of hidden neurons, as well as the structures for the deep neural networks were empirically tried, and the results of the best settings were registered for comparison. In Table 1 the performance of the best algorithm is highlighted in bold. The results indicate that although S-rRBF does not always outperform the three deep networks, it generally compares favorably with the best performing deep model.

To show the properties of the internal representation of the S-rRBF, the resulting internal representations of the S-rRBF for some of the benchmark problems are shown. The first one is the internal representation for the Iris problem, a well known 3-class problem where one of the classes are linearly separable from the other non-linearly separable two. The self-organized internal topographical representation of the S-rRBF for this problem is shown in Fig. 2a. For comparison, 2-D visualizations of SOM and t-SNE are respectively shown in Fig. 2b and Fig. 2c. In these 2-D maps, each class is represented with different marker and color, while \times on the maps show the overlapping representation of some data belonging to contrasting classes. The size of the marker reflects the number of data that it represents. It should be noted that for SOM and t-SNE the 2-D representations are constructed based only on the similarities of the data while their labels are irrelevant. The internal representation of the S-rRBF is different in that during the learning process, the directions of the topological self-organization are regulated by the labels of the data. Hence, the internal representation of the S-rRBF is context-dependent. The three representations reflect the problem’s separability well. This is an easy problem in that the data distribution is consistent with the labels distribution. The simplicity of the problem is also obvious from the high classification performances of the compared algorithms in Table 1.

The second example is Bank Marketing Data, a 48-D, 3-class problem. The low dimensional representations of SOM in Fig. 3b and t-SNE in Fig. 3c indicate that there are many overlapping data belonging to contrasting classes that make this problem a relatively difficult one. Figure 3a indicates that the S-rRBF generates a nice topographical internal representation that illustrates how the classifier separates the two classes. The \times in S-rRBF’s representation indicates the area where data are likely to be misclassified.

The third example is the recently proposed ”Fashion MNIST”, an apparel-related image classification problem (Xiao et al., 2017). Some of the data of this data set are shown in Fig. 5. This data set has the same dimensionality, class number and data size than the traditional MNIST data set. The SOM and t-SNE representations of this problem are respectively shown in Fig. 5b and Fig.

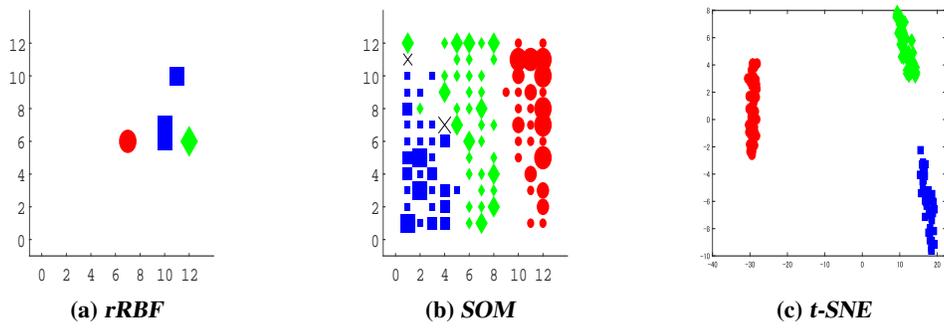


Figure 2: Iris (dim:4, class:3)

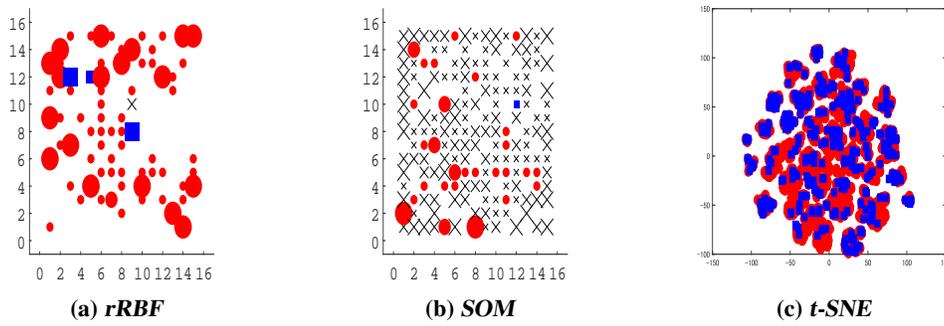


Figure 3: Bank Credit

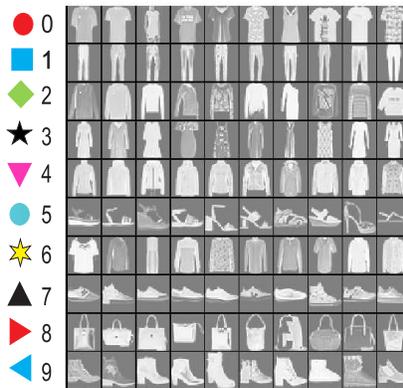


Figure 4: Fashion MNIST.

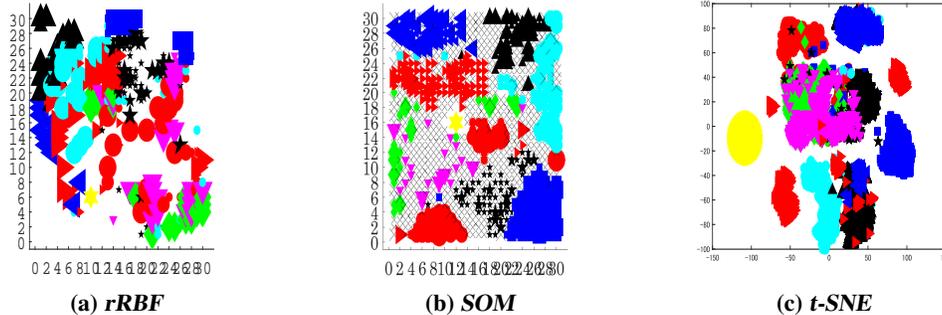


Figure 5: Fashion MNIST

5c. These figures illustrate that some of the classes are distinctively represented by well separated clusters but there are also many overlapping classes. The internal representation of the S-rRBF in Fig. 5a shows that it did not form a clearly distinctive class representations as in the previous two examples.

The internal representations offer understanding on how the S-rRBF self-organizes the data to be further classified in the output layer.

4 CONCLUSIONS

In this research we showed that it is possible to build a hierarchical neural network that self-organizes with context-relevant topographical internal representation. More specifically, we showed that topographical self-organization can emerge as an implication of the supervised learning. Thus, the two learning processes of self-organization and supervised learning, which are often considered to be unrelated, are can be viewed as two different aspects of a single learning mechanism. The two learning processes are only distinguished by the layers where they occurs. The internal self-organization in this network is not fully unsupervised. However, the direction of the self-organization process in a hidden neuron is only decided by the relative value of the connection weight leading from the neuron to the output neuron relevant to the true label of the input and thus not dependent on the supervised error.

The experiments show that the classification performance of the proposed model is comparable to that of standard supervised networks. While the proposed model does not always outperform existing conventional models, we found that the performance was comparable to the best performer for most of the diverse benchmark applications. Specific machine learning methods often perform well on datasets for which they have been designed, but it is well acknowledged that sufficient performance in a variety of tasks is useful in many applications such as robotics and probably to understand better human abilities. Another advantage of our system is its 2-dimensional internal layer offers auxiliary visual information on its learning representations. The S-rRBF can readily expanded into deep networks. As layered networks transfer transform inputs (physical stimuli) into labels (concepts) in a layer by layer manner, the visualization of internal layers in multi-layered S-rRBF can be considered as concept-forming visualization. The visualization can potentially offer new insights for machine learning.

REFERENCES

- Y. Bengio. *Learning Deep Architectures for AI*. now Publishers, Hanover, MA, 2009.
- H. Boursland and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59:291–294, 1988.
- X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudik (eds.), *Proceeding of the Fourteenth International Conference*

- on *Artificial Intelligence and Statistics, AISTAT 2011, Fort Lauderdale, FLorida, USA, April 11-13*, volume 15 of *JMLR Workshop and Conference Proceedings*, pp. 315–323. JMLR.org, 2011. URL <http://www.jmlr.org/proceedings/papers/v15/>.
- I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of The 30th International Conference on Machine Learning*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 1319–1327. JMLR.org, 2013.
- I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, Cambridge, MA, 2016.
- P. Hartono, P. Hollensen, and T. Trappenberg. Learning-regulated context relevant topographical map. *IEEE Trans. on Neural Networks and Learning Systems*, 26(10):2323–2335, 2015. doi: 10.1109/TNNLS.2014.2379275.
- G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1711–1800, 2002.
- G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- G. Hinton and R.S. Zemel. Autoencoders, minimum description length, and helmholtz free energy. In J.D. Cowan, G. Tesauro, and J. Alspector (eds.), *Advances in Neural Information Processing Systems 6 (NIPS 1993)*, 1994. URL <https://papers.nips.cc/book/advances-in-neural-information-processing-systems-6-1993>.
- G. Hinton, S. Osindero, and Teh. Y-W. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.
- S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(2): 107 – 116, 1998.
- D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.
- T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
- Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.
- T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of IEEE*, 87:1484–1487, 1990.
- G.L. Romani, S.J. Williamson, and L. Kaufman. Tonotopic organization of the human auditory cortex. *Science*, 216(4552):1339–1340, 1982.
- R. Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2:361–385, 2015.
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15: 1929–1958, 2014.
- L.P.J. van der Maaten. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- D.J. Willshaw and C. Von Der Malsburg. How patterned neural connections can be set up by self-organization. *Proc. Royal Society of London*, 194(1117):431–445, 1976.
- D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Trans. on Evolutionary Computation*, 1(1):67–81, 1997.
- H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017. URL <https://arxiv.org/abs/1708.07747>.