Neural Simpletrons – Learning in the Limit of Few Labels with Directed Generative Networks

Dennis Forster
Machine Learning Group
University of Oldenburg
26129 Oldenburg, Germany
FIAS, Goethe-University
60438 Frankfurt, Germany
dennis.forster@uol.de

Abdul-Saboor Sheikh
Zalando Research
11501 Berlin, Germany
Cluster of Excellence H4a
University of Oldenburg
26129 Oldenburg, Germany
sheikh.abdulsaboor
@gmail.com

Jörg Lücke
Machine Learning Group
Cluster of Excellence H4a &
RC Neurosensory Sciences
University of Oldenburg
26129 Oldenburg, Germany
joerg.luecke@uol.de

Abstract

We investigate the task of classifier training for data sets with limited labels using a neural network derived from a hierarchical normalized Poisson mixture model. With the single objective of likelihood optimization, both labeled and unlabeled data are naturally incorporated into learning. Using standard benchmarks for non-negative data, such as text document representations (20 Newsgroups), MNIST and NIST SD19, we study the classification performance when only very few labels are used for training and parameter tuning. In different settings, the network's performance is compared to standard and recently suggested semi-supervised classifiers. While other recent approaches are more competitive for many labels or fully labeled data sets, we find that the here studied network can be applied to limits of few labels where no other system has been reported to operate so far.

1 Introduction

Large data sets, e.g., in the form of digital texts, images or sounds, become increasingly ubiquitous. However, acquisition of fully labeled data becomes increasingly costly with larger amounts of data points, as correct labeling requires ground-truth or a human who can hand-label the data. Consequently, classifiers leveraging information from both labeled and unlabeled data have in recent years shifted into the focus of many research groups (Liu et al., 2010; Weston et al., 2012; Pitelis et al., 2014; Kingma et al., 2014; Forster et al., 2015; Rasmus et al., 2015; Miyato et al., 2015). Typically, a supervised deep neural network (DNN) is used in combination with additional mechanisms that allow usability when labels are limited. However, such systems generally come with large numbers of free parameters, which in the semi-supervised setting increases the risk of overfitting to very small validation sets. Alternatively, standard probabilistic networks, e.g., in the form of deep directed graphical models (DDMs) can in principle be trained using unlabeled and labeled data. However, while being potentially very powerful information processors, typical directed models are limited in size (compare, e.g. Larochelle and Murray, 2011; Bornschein and Bengio, 2015; Gan et al., 2015).

In contrast to DNNs and SVMs, DDMs are primarily used for unsupervised learning. For the targeted limit of few labels, DDMs thus appear as a more natural starting point if we are able to address scalability for classification applications. In order to do so, we base our study on a directed graphical model which is sufficiently richly structured to give rise to a good classifier, while it allows for efficient training on large data sets and with large network sizes. Scalability will be realized by the derivation of a neural network equivalent for maximum likelihood learning of the considered graphical model. The emerging compact and local inference and learning equations of the network can then be parallelized and scaled using the same tools as were originally developed for conventional deep neural networks. By additionally considering a minimalistic network architecture, the number of free parameters will, at the same time, be kept low and easily tunable on few labels.

LLD Workshop, 31st Conf. on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 A Hierarchical Neural Network for Learning with Limited Labels

A classification problem can be modeled as an inference task based on a probabilistic hierarchical mixture model (e.g., Duda et al., 2001). For our goal of semi-supervised learning with limited labels, we restrain the model complexity to the minimum of no more than three layers:

$$p(k) = {}^{1}\!/\!{}_{K}, \qquad p(l|k) = \delta_{lk} \tag{1}$$

$$p(c|k,\mathcal{R}) = \mathcal{R}_{kc}, \qquad \textstyle\sum_c \mathcal{R}_{kc} = 1 \tag{2}$$

$$p(\vec{y}\,|c,\mathcal{W}) = \textstyle\prod_d \text{Poisson}(y_d;\mathcal{W}_{cd}), \qquad \textstyle\sum_d \mathcal{W}_{cd} = A \tag{3}$$

The parameters of the model, $\mathcal{W}\in\mathbb{R}_{\geq 0}^{C\times D}$ and $\mathcal{R}\in\mathbb{R}_{\geq 0}^{K\times C}$, will be referred to as generative weights, which are normalized to constants A and 1, respectively. The top node represents K abstract concepts or super classes k (for example, ten classes of digits). The middle node represents any of the occurring C subclasses c (like different writing styles of the digits). And the bottom nodes represent an observed data sample $\vec{y}$ with an according data label l (e.g., ranging from '0' to '9'). Our model assumes non-negative observed data, and we use the Poisson distribution as an elementary distribution for non-negative observations.

For the purposes of this study, we specify a neural network formulation that corresponds to learning and inference in this hierarchical generative model. Consider the neural network in Fig. 1 with neural activities $\vec{y}$, $\vec{s}$ and $\vec{t}$ and learning rules as in Tab. 1. We assume the values of $\vec{y}$ to be obtained from a set of unnormalized data points $\vec{\tilde{y}}$ by Eq. (T1.3), and the label information to be presented as top-down input vector $\vec{u}$ as given in Eq. (T1.2). For the neural weights (W,R) of the network, we consider Hebbian learning with a subtractive synaptic scaling term (see for example Abbott and Nelson, 2000) as in Eqs. (T1.7) and (T1.8), where $\epsilon_W > 0$ and $\epsilon_R > 0$ are learning rates. These learning rules are local, can integrate both supervised and unsupervised learning, are highly parallelizable and they result in normalized neural weights (W,R). Those we can relate to the generative weights $(\mathcal{W},\mathcal{R})$ by identifying s_c with the posterior probability $p(c|\vec{y}^{\,(n)},l^{(n)},\Theta)$ by Eq. (T1.4) and t_k with $p(k|\vec{y}^{\,(n)},l^{(n)},\Theta)$ by Eq. (T1.6), using the neural weights as model parameters Θ. Such neural learning converges to the same fixed points as EM for the hierarchical Poisson mixture model (compare Lücke and Sahani, 2008; Keck et al., 2012; Forster et al., 2015; Forster and Lücke, 2017). In other words, executing the online neural network of Tab. 1 optimizes the likelihood of the generative model Eqs. (1) to (3). The network's neural activities therein provide the posterior probabilities, which we use for classification with the MAP estimate of t_k (Eq. 1.6) giving the inferred classes. The computation of posteriors is in general a difficult and computationally intensive endeavor, and their interpretation as neural activation rules is usually difficult. In our case, because of a specific interplay between introduced constraints, categorical distribution and Poisson noise, the posteriors and their neural interpretability however greatly simplify. The equations defining the neural network are elementary, very compact, and contain a total number of only four free parameters: the number of hidden units C, an input normalization constant A, and learning rates ϵ_W and ϵ_R. Because of its compactness, we call the network *Neural Simpletron* (NeSi).

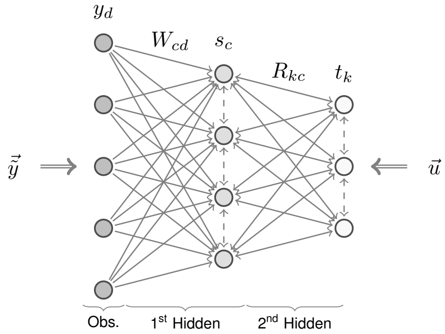

Figure 1: Graphical illustration of the hierarchical recurrent neural network.

Table 1: Neural network formulation of probabilistic inference and maximum likelihood learning.

Neural Simpletron			
Input			
Bottom-Up:	$\tilde{y}_d$	unnormalized data	(T1.1)
Top-Down:	$u_k = \begin{cases} \delta_{kl} & \text{for labeled data} \\ \frac{1}{K} & \text{for unlabeled data} \end{cases}$		(T1.2)
Activation Across Layers			
Obs. Layer:	$y_d = (A-D)\frac{\tilde{y}_d}{\sum_{d'}\tilde{y}_{d'}} + 1$		(T1.3)
1st Hidden:	$s_c = \frac{\exp(I_c)}{\sum_{c'}\exp(I_{c'})}$, with		(T1.4)
	$I_c = \sum_d \log(W_{cd})y_d + \log(\sum_k u_k R_{kc})$		(T1.5)
2nd Hidden:	$t_k = \begin{cases} u_k & \text{labeled data} \\ \sum_c \frac{R_{kc}}{\sum_{k'} R_{k'c}} s_c & \text{unlabeled data} \end{cases}$		(T1.6)
Learning of Neural Weights			
1st Hidden:	$\Delta W_{cd} = \epsilon_W(s_c y_d - s_c W_{cd})$		(T1.7)
2nd Hidden:	$\Delta R_{kc} = \epsilon_R(t_k s_c - t_k R_{kc})$	labeled data	(T1.8)

2.1 Simpletron Variants

We investigate different NeSi versions for learning with limited labels. All variants accord with the likelihood objective and were chosen such that the number of tunable parameters remains small: (I) The complete formulas for the first hidden layer, given in Eqs. (T1.4) and (T1.5), define a recurrent network (‘r-NeSi’), i.e., combine both bottom-up and top-down information. By removing the second term in Eq. (T1.5), we gain a feedforward network (‘ff-NeSi’) that is equivalent to treating the $p(c|k, R)$ in the first hidden layer as uniform $1/C$. (II) Using the posterior t_k, the network can itself provide missing training labels (often termed ‘self-labeling’; see, e.g., Lee, 2013; Triguero et al., 2015) and the ‘Best versus Second Best’ (BvSB) measure on t_k gives an index for classification certainty. We then train the top layer also on those inferred labels with high certainty, i.e., where the BvSB lies above a given threshold ϑ. We mark NeSi networks using self-labeling by a superscript ‘$+$’ (‘r$^+$-NeSi’ and ‘ff$^+$-NeSi’). (III) We apply TV-EM (Lücke, 2016) to ff$^+$-NeSi by keeping only the C' highest values in s_c (Eq. T1.4) and setting lower values to hard zero with subsequent renormalization. Such truncation can be shown to maximize the variational free energy of the mixture model with a significantly lower computational cost (Forster and Lücke, 2017). We refer to simpletrons with truncated middle layer activations as ‘t-NeSi’. This represents a further development of the network in Forster et al. (2015) with enhanced training and significantly improved results.

3 Parameter Tuning with Limited Labels

It is customary to regard limited numbers of labels as restriction only on training itself and not on the preceding optimization of free model parameters by using validations sets with (much) more labels than available during training. For model comparison, this however introduces a bias towards more complex models, that would be otherwise more prone to overfitting to small validation sets. We therefore train our models given a strictly limited total number of labels for the complete tuning and training procedure. For parameter tuning, we use 10 labeled training data points per class (the setting with the lowest number of labels on which models are generally compared on MNIST) with a half/half split into training and validation set. Once optimized, we keep the free parameters fixed for all following experiments. In doing so, we ensure that all our results of 10 labels per class and more are achievable by using no more labels in total than provided within each training setting. Furthermore, using only training data for parameter tuning guarantees a fully blind test set, such that the test error gives a reliable index for generalization.

4 Numerical Experiments

We apply the NeSi networks to three standard benchmarks for classification on non-negative data: the 20 Newsgroups text data set (Lang, 1995), the MNIST data set of handwritten digits (LeCun et al., 1998) and the NIST Special Database 19 of handwritten characters (Grother, 1995). We perform experiments for different proportions of randomly chosen, class-balanced labeled data and measure the mean classification error on the blind test set.

4.1 Document Classification (20 Newsgroups)

We preprocess the newsgroups data using only tf-idf weighting (Sparck Jones, 1972). No stemming, removals of stop words or frequency cutoffs were applied. We investigate semi-supervised settings of 20, 40, 200, 800 and 2000 labels in total – that is 1, 2, 10, 40 and 100 labels per class – as well as the fully labeled setting. For each setting, we present the mean test error averaged over 100 independent runs and the standard error of the mean (SEM). On each new run, a new set of class balanced labels is chosen randomly from the training set. We train our model on the full 20-class problem without any feature selection.

Table 2: Test error on 20 Newsgroups.

#labels	ff-NeSi	r-NeSi	HDRBM
20	70.64 $\pm$ 0.68 $^{(*)}$	**68.68** $\pm$ 0.77 $^{(*)}$	
40	55.67 $\pm$ 0.54 $^{(*)}$	**54.24** $\pm$ 0.66 $^{(*)}$	
200	30.59 $\pm$ 0.22	**29.28** $\pm$ 0.21	
800	28.26 $\pm$ 0.10	**27.20** $\pm$ 0.07	31.8 $^{(*)}$
2000	27.87 $\pm$ 0.07	**27.15** $\pm$ 0.07	
11269	28.08 $\pm$ 0.08	**17.85** $\pm$ 0.01	23.8

As reported results on the full 20-class task in the semi-supervised setting are rare, we here only compare to a hybrid of generative and discriminative RBMs (HDRBM) trained by Larochelle and Bengio (2008) using stochastic gradient descent to perform semi-supervised learning. The test errors of the methods are compared in Tab. 2. For results marked with ‘$(*)$’, free parameters were optimized using additional labels: NeSi used the same parameter setting in all semi-supervised experiments on 20 Newsgroups, which was tuned with 200 labels in total; HDRBM used 1000 labels in total for tuning in the semi-supervised setting (200 additional labels for the validation set).

4.2 Handwritten Digit Recognition (MNIST)

We perform experiments on MNIST in the semi-supervised setting using 10, 100, 600, 1000 and 3000 labels in total – that is 1, 10, 60, 100 and 300 labels per class –, which are randomly and class balanced chosen from the 10 classes. Results are given as the mean and standard error (SEM) over 100 independent repetitions, with randomly drawn, class-balanced labels.

Table 3: Test error of NeSi algorithms on permutation invariant MNIST for different semi-supervised settings.

#labels	ff-NeSi	r-NeSi	ff$^+$-NeSi	r$^+$-NeSi	t-NeSi
10	55.5 ± 0.6 $^{(*)}$	29.6 ± 0.6 $^{(*)}$	10.9 ± 0.9 $^{(*)}$	17.9 ± 0.9 $^{(*)}$	$\mathbf{7.2} \pm 0.5$ $^{(*)}$
100	19.1 ± 0.3	12.4 ± 0.2	4.96 ± 0.08	4.93 ± 0.05	$\mathbf{4.23} \pm 0.07$
600	7.27 ± 0.05	6.94 ± 0.05	4.08 ± 0.02	4.34 ± 0.01	$\mathbf{3.65} \pm 0.01$
1000	5.88 ± 0.03	6.07 ± 0.03	4.00 ± 0.01	4.26 ± 0.01	$\mathbf{3.63} \pm 0.01$
3000	4.39 ± 0.02	4.68 ± 0.02	3.85 ± 0.01	4.05 ± 0.01	$\mathbf{3.52} \pm 0.01$
60000	3.27 ± 0.01	$\mathbf{2.94} \pm 0.01$	3.27 ± 0.01	$\mathbf{2.94} \pm 0.01$	$\mathbf{2.94} \pm 0.01$

Tab. 3 shows the results of the NeSi algorithms. Again, for results marked with '$(*)$', the free parameters were optimized using more labels than available in the given setting. We used the same parameter setting for all experiments shown here, which was tuned using 100 labels in total. As the NeSi model has no prior knowledge about spatial relations in the data, the given results are invariant to pixel permutation.

Fig. 2 shows a comparison to standard and recent state-of-the-art approaches for 100 labels and more. The performance of the models is given with respect to the number of labels used during training (left-hand side) and with respect to the total number of labels used for the complete tuning and training procedure (right-hand side). For the NeSi algorithms, these plots are identical, as we only use maximally as many labels in the tuning phase as in the training phase for the shown results of 100 labels and more. All other algorithms (for lack of more comparable findings) either use a validation set with a substantial amount of additional labels than available during training or (explicitly) use the test set for parameter tuning. Also, some of the shown results (namely the TSVM, AGR, AtlasRBF and the Em-networks) were achieved in the transductive setting, where the (unlabeled) test data is included into the training process. The NeSi approaches are to our knowledge so far the closest to our goal of a competitive algorithm in the limit of as few labels as possible. Regarding classification performance, the NeSi networks achieve competitive results, surpassing even deep belief networks ('DBN-rNCA') and other recent approaches (like the 'Embed'-networks,

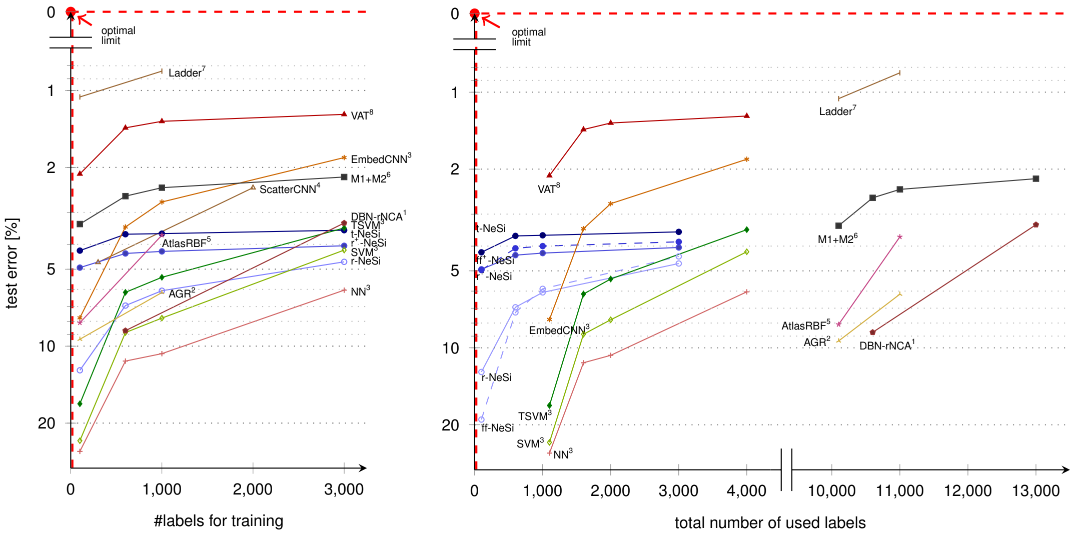

Figure 2: Classification performance of different algorithms compared against varying proportion of labeled training data. The algorithms are described in detail in the corresponding papers: [1]Salakhutdinov and Hinton (2007), [2]Liu et al. (2010), [3]Weston et al. (2012), [4]Bruna and Mallat (2013). [5]Pitelis et al. (2014), [6]Kingma et al. (2014), [7]Rasmus et al. (2015), [8]Miyato et al. (2015). All algorithms except ours use 1000 or 10 000 additional data labels (from the training or test set) for parameter tuning. For ScatterCNN (Bruna and Mallat, 2013) the validation set size is not reported.

‘AGR’ and ‘AtlasRBF’). In the light of reduced model complexity and effectively used labels, we can furthermore compare to the few very recent algorithms with a lower error rate (‘M1+M2’, ‘VAT’ and the ‘Ladder’-networks).

One competing model that so far comes closest to our limit setting of as few labels as possible is an approach which combines 10 generative adversarial networks (GANs) (Salimans et al., 2016) with 5 layers each. With down to 20 labels for training (and an however unknown additional number of labels for the validation set) the classification error of the full ensemble of 10 GANs the error was reported to be $(11.34 \pm 4.45)\%$. This compares to an error of $(7.22 \pm 0.53)\%$ for t-NeSi which was trained with only 10 labels (one per class) and used 100 labels during parameter tuning.

4.3 Large Scale Handwriting Recognition (NIST SD19)

Table 4: Test error on NIST SD19 data set on the task of digit and letter recognition for different total amounts of labeled data.

#lbls/class	1	10	60	100	300	fully labeled
digits (10 classes)						
#lbls total	10	100	600	1000	3000	344 307
ff$^+$-NeSi	7.6 ± 1.8	6.2 ± 0.2	6.0 ± 0.1	6.0 ± 0.1	5.70 ± 0.03	5.11 ± 0.01
r$^+$-NeSi	9.8 ± 2.4	6.1 ± 0.2	5.8 ± 0.1	5.9 ± 0.1	5.7 ± 0.1	4.52 ± 0.01
t-NeSi	**5.7** ± 0.4	**5.3** ± 0.2	**4.84** ± 0.02	**4.86** ± 0.03	**4.84** ± 0.02	4.50 ± 0.01
35c-MCDNN						**0.77**
letters (52 classes)						
#lbls total	52	520	3120	5200	15600	387361
ff$^+$-NeSi	55.7 ± 0.6	46.2 ± 0.4	44.2 ± 0.2	43.7 ± 0.2	43.0 ± 0.3	34.66 ± 0.05
r$^+$-NeSi	65.0 ± 0.9	54.1 ± 0.4	43.7 ± 0.2	**41.6** ± 0.1	**38.0** ± 0.1	31.93 ± 0.06
t-NeSi	**52.1** ± 1.1	**45.6** ± 0.4	**41.9** ± 0.3	41.8 ± 0.4	41.1 ± 0.3	33.34 ± 0.04
35c-MCDNN						**21.01**

To show large scale applicability, we show experiments on the NIST Special Database 19, containing over $800\,000$ binary 128×128 images. We preprocess the data similar to MNIST, which allows us to use the same setting for our free model parameters without retuning, and perform experiments on digit recognition (10 classes) and case-sensitive letter recognition (52 classes). The experiments are done using 1, 10, 60, 100, 300, or all labels per class. In Tab. 4, we report the mean and standard error over 10 experiments and compare to the state-of-the-art 35c-MCDNN (Cireşan et al., 2012). For the NeSi networks, the results are given for the permutation invariant task. To the best of our knowledge, this is the first system to report results for NIST SD19 in the semi-supervised setting.

5 Discussion

In this study, we explored classifier training on data sets with limited labels. We put a special emphasize on adhering to this restriction not only for the training phase, but the complete tuning, training and testing procedure. Our tool was a novel neural network with learning rules based on a maximum likelihood objective. Starting from hierarchical Poisson mixtures, the derived three layer directed data model can be observed to take on a form similar to learning in standard DNNs. The parameters of the network can be optimized with a very limited amount of labels and training in the same setting showed to achieve competitive results, giving the first network shown to operate using no more than 10 labels per class in total and down to a single training label per class on the investigated data sets.

Our main empirical results for the NeSi systems were obtained using the 20 Newsgroups, the MNIST and the NIST SD19 datasets. Tabs. 2 to 4 and Fig. 2 summarize the results and provide comparison to other approaches. The r-NeSi system is the best performing system for the 20 Newsgroups data set (Tab. 2), but comparative results were only available for HDRBM in the semi-supervised setting. Both on MNIST and NIST the performance of our 3-layer network in the fully labeled setting is not competitive to state-of-the-art fully supervised algorithms. Our results however do apply for the permutation invariant setting and do not take prior knowledge about two-dimensional image data into account (like convolutional networks do). More importantly, we only see a relatively mild decrease in test error when we strongly decrease the total number of used labels, with the t-NeSi consistently performing best. It has so far not been shown that other classifiers can be trained with similarly low total numbers of labels, as all comparable approaches use at least 1000 additional labels to optimize the free parameters of their respective systems (Fig. 2, right-hand-side). Furthermore, as all better performing approaches on MNIST combine different objectives to perform well with limited labels, the NeSi networks can be considered as the best performing non-hybrid approaches even if we only consider exclusively the labels for training (Fig. 2, left-hand-side).

References

Abbott, L. F. and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature Neuroscience*, 3:1178–1183.

Bornschein, J. and Bengio, Y. (2015). Reweighted wake-sleep. In *International Conference on Learning Representations (ICLR)*.

Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 35(8):1872–1886.

Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3642–3649. IEEE.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley-Interscience (2nd Edition).

Forster, D. and Lücke, J. (2017). Truncated variational EM for semi-supervised Neural Simpletrons. In *IJCNN 2017, in press.*

Forster, D., Sheikh, A.-S., and Lücke, J. (2015). Neural Simpletrons – Minimalistic directed generative networks for learning with few labels. *arXiv preprint arXiv:1506.08448*.

Gan, Z., Henao, R., Carlson, D., and Carin, L. (2015). Learning deep sigmoid belief networks with data augmentation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 268–276.

Grother, P. J. (1995). NIST special database 19 handprinted forms and characters database. *National Institute of Standards and Technology*.

Keck, C., Savin, C., and Lücke, J. (2012). Feedforward inhibition and synaptic scaling – two sides of the same coin? *PLoS Computational Biology*, 8:e1002432.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3581–3589.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning (ICML)*, pages 331–339.

Larochelle, H. and Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. In *International Conference on Machine Learning (ICML)*, pages 536–543.

Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 29–37.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *IEEE*, 86(11):2278–2324.

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2.

Liu, W., He, J., and Chang, S.-F. (2010). Large graph construction for scalable semi-supervised learning. In *International Conference on Machine Learning (ICML)*, pages 679–686.

Lücke, J. (2016). Truncated variational expectation maximization. *arXiv preprint, arXiv:1610.03113*.

Lücke, J. and Sahani, M. (2008). Maximal causes for non-linear component extraction. *Journal of Machine Learning Research (JMLR)*, 9:1227–1267.

Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., and Ishii, S. (2015). Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677v6*.

Pitelis, N., Russell, C., and Agapito, L. (2014). Semi-supervised learning using an unsupervised atlas. In *Machine Learning and Knowledge Discovery in Databases*, pages 565–580. Springer.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. (2015). Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3532–3540.

Salakhutdinov, R. and Hinton, G. E. (2007). Learning a nonlinear embedding by preserving class neighbourhood structure. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 412–419.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2226–2234.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Triguero, I., García, S., and Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42(2):245–284.

Weston, J., Ratle, F., Mobahi, H., and Collobert, R. (2012). Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer.