Deep and Sparse Learning in Speech and Language Processing: An Overview

Dong Wang[1,2*], Qiang Zhou[1,2], and Amir Hussain[3]

[1]CSLT, RIIT, Tsinghua University, Beijing 100084, China
[2]Tsinghua National Lab for Information Science and Technology, Beijing, China
[3]University of Stirling, FK9 4LA, Scotland, UK
wangdong99@mails.tsinghua.eud.cn
zq-lxd@mail.tsinghua.edu.cn
hussain.doctor@gmail.com

Abstract. Large-scale deep neural models, e.g., deep neural networks (DNN) and recurrent neural networks (RNN), have demonstrated significant success in solving various challenging tasks of speech and language processing (SLP), including speech recognition, speech synthesis, document classification and question answering. This growing impact corroborates the neurobiological evidence concerning the presence of layer-wise deep processing in the human brain. On the other hand, sparse coding representation has also gained similar success in SLP, particularly in signal processing, demonstrating sparsity as another important neurobiological characteristic. Recently, research in these two directions is leading to increasing cross-fertlisation of ideas, thus a unified Sparse Deep or Deep Sparse learning framework warrants much attention. This paper aims to provide an overview of growing interest in this unified framework, and also outlines future research possibilities in this multi-disciplinary area.

Keywords: Deep learning, sparse coding, speech processing, language processing

1 Introduction

How the human brain processes information so effectively and efficiently is a long-standing mystery. Although still far from a full understanding, physiological studies seem to support two hypotheses: a sparse coding scheme that can represent information succinctly, redundantly and robustly, and a layer-wise hierarchical processing pipeline, gradually forming high-level abstraction with clear and rich semantic meaning [**?**,**?**]. This section summarizes findings from research on mammal neural systems, relating to sparse and hierarchical characteristics, and will then review the machine learning research inspired by each. Finally we will discuss the concept of combining these two characteristics.

1.1 Sparsity and hierarchy in the brain

The sparse information representation in the human brain has been recognized by researchers for a long time [**?**,**?**,**?**]. By scrutinizing cellular recordings physiologists found

that neurons represent external stimuli in a rather sparse way. Specifically, many stimuli may activate the same neuron, and each stimulus is represented by only a few neurons [**?**]. This suggests a sparse coding scheme in our brain, where each stimulus is represented by a distribution of the activity it triggers over the neurons, and the number of activated neurons is small. In other words, stimuli are represented by *sparse codes* in our brain.

The layer-wised hierarchical structure, or the deep architecture, is another basic assumption for the neural system since early research on connectionist models [**?**]. The key advantage of this architecture is that high-level abstraction can be learned layer-by-layer. The high-level abstraction is assumed to be a key ingredient in perception and cognition [**?**]. Interestingly, the mammal brain was found to be organized in a deep architecture [**?**], where a given input is processed at multiple levels, and each level corresponds to a different area of cortex.

1.2 Sparsity and hierarchy in machine learning

Both sparse representations and hierarchical architectures have attracted much attention in machine learning and artificial intelligence (AI) research. On the sparse coding part, researchers usually cast the sparse coding problem to a constrained optimization problem for over-complete linear equations, and developed numerous optimization approaches to solve the problem.

On the hierarchical architecture part, researchers in machine learning have demonstrated the brilliant success of deep learning methods over the past decade. A key point of deep learning is the layer-wised information processing within the hierarchical architecture [**?**], for example the deep neural network (DNN) and its convolutional and recurrent variants, i.e., convolutional neural networks (CNN) and recurrent neural networks (RNN). Deep learning has delivered remarkable performance on numerous machine learning tasks, including speech processing [**?**,**?**,**?**,**?**,**?**] and language processing [**?**]. A high-level summary was recently published in Nature [**?**], and more details can be found in the review papers written by Bengio [**?**,**?**].

1.3 Deep sparse or sparse deep models?

Both sparsity and hierarchy are important properties of the brain and play fundamental roles in perception, cognition and other functions that comprise human intelligence. An interesting question is: How these two properties are integrated together to support these fundamental functions? A simple picture is that sparse coding provides efficient and robust codes, while the hierarchical architecture offers a constrained structure where the sparse codes are processed. Although not fully confirmed by physiological studies, machine learning researchers are investigating in this direction and have achieved some promising results, e.g., [**?**,**?**,**?**,**?**,**?**]. In this paper, we give a quick review for the recent development of deep *and* sparse models, and describe some applications of this new technique in speech and language processing. Note that the sparse models we are concerned with are not limited to sparse coding, but any models with sparsity regularization. Additionally, we distinguish deep sparse models and sparse deep models: the former refers to sparse models that are stacked to form a deep structure, while the later

refers to deep models (usually neural networks) that involve sparsity regularization. Our review covers both categories, but sparse deep models are clearly dominant at present.

The paper is organized as follows: Section 2 describes sparse deep models, and Section 3 describes deep sparse models. Their application in speech and language processing is presented in Section 4, and some ideas for future research are presented in Section 5.

2 Sparse deep models

Sparse deep models refer to deep models (e.g., DNNs) that involve certain sparsity regularization. This regularization can be applied to various components of the deep model, e.g., units, weights, or gradients. It is often in the form of sparsity-oriented norms (e.g., ℓ-0 or ℓ-1), as well as special activation functions (e.g., rectifier) and some pre-training procedure.

2.1 Unit sparsity by norms

The deep model with unit sparsity is closely related to sparse coding, and can be regarded as a deep and non-linear extension to the conversional sparse coding model. The most straightforward way to produce sparse units is to impose certain sparse-oriented norms on the hidden units and add the norms into the objective function. Although ℓ-0 is the most ideal, l-1 is used more often due to its smoothness. For example, Olshausen et al. [**?**] used ℓ-1 to generate sparse units to model activities of the retina in mammals [**?**]. Another commonly used sparse-oriented regularization is the ℓ-1/ℓ-2 norm, which often leads to group sparsity [**?**,**?**].

An early work from Ranzato and colleagues [**?**] employed a sparse regularization in the form $\sum_i^M log(1 + z_i^2)$ on the hidden units when training deep encoder-decoding models, where z_i is the nonlinear-transformed activation of the i-th hidden unit, and M is the number of hidden units. The main purpose of this sparsity regularization was to limit the input space where the energy surface has a low value, which is a cheap approximation to the partition function and leads to an efficient training algorithm for deep models.

Another work presented by Lee et al. [**?**] imposed sparsity regularization when training RBMs, where the sparsity regularization was defined as follows:

$$\sum_{j=1}^{n} |p - \frac{1}{m} \sum_{l=1}^{m} \mathbb{E}[h_j^{(l)} | \mathbf{v}^{(l)}]|^2$$

where $\mathbf{v}^{(l)}$ represents the observed variables of the l-th example, $h_j^{(l)}$ denotes the j-th hidden variable of the l-th example. The authors found that the sparse RBM components can be stacked to form a sparse deep belief net (DBN). Particularly, they found a two-layer sparse DBN can model the cell receptive fields of the visual area V1 and V2. Note that similar sparsity regularization was also employed in the discriminative or semi-supervised RBM framework, as proposed by Larochelle et al. [**?**].

Luo et al. [**?**] proposed to use the $\ell 1/\ell 2$ mix norm to yield sparse units in RBMs. This norm is formulated as follows:

$$\sum_k \sqrt{\sum_{m \in G_k} P(h_m = 1|x^{(l)})^2}$$

where G_k is the k-th group of the hidden units. It can be seen that the regularization on the groups is ℓ-1, while the regularization on the hidden units (within a particular group) is ℓ-2. This leads to group-level sparsity, while keeping the firing probability of the units within one group equally small. The authors stacked the group-level sparse RMBs to construct sparse deep Boltzmann machines (DBMs). Experiments on letter recognition tasks with the MNIST and OCR databases confirmed that the sparse DBM can learn reasonable hierarchical features and provide good pre-training for DNNs. The ℓ-1/ℓ-2 norm was also used by Li et al. [**?**] to construct deep stacking networks (DSN). Experiments conducted on image classification confirmed the efficiency of their model.

The popular ℓ-1 regularization is equal to a Laplace prior distribution over the hidden units. Another sparsity-oriented prior is the spike-and-slab prior recently proposed by Goodfellow et al. [**?**]. Experiments on image pattern learning and classification tasks demonstrated that this prior can lead to better sparse representations.

2.2 Unit sparsity by activation functions

The second approach to deriving sparse units is through special activation functions. Perhaps the most popular sparsity-deriving activation function is the rectifier function $g(x) = max(0, x)$, which suppresses the negative part of the activation to zero. Interestingly, this function was found to resemble the true activation function of human neurons, according to the leaky integrate-and-fire (LIF) model [**?**]. Xaiver et al. [**?**] presented an empirical study for deep neural networks with the rectifier activation function and found several merits associated with this function, particularly that the training is much easier, as the segment linearity of this activation avoids the notorious gradient vanishing and explosion problem. They found that with the rectifier activation, competitive performance can be obtained even without pre-training.

Poultney et al. [**?**] studied another sparse-oriented activation function called 'sparsifying logistic'. They found that with this type of activation function, simple and complex cell receptive fields can be derived, leading to a topographic layout of the filters, which is reminiscent of the topographic maps found in area V1 of the visual cortex.

Another sparse-oriented activation function called 'win-take-all' was proposed by Makhzan et al. [**?**]. This function applies to the convolutional layers of a CNN model and retains the largest activation while setting others to zero. This approach was tested on the MNIST image classification task and obtained competitive performance.

2.3 Unit sparsity by pre-training

Recently, Li et al. [**?**] presented an interesting analysis and showed that various pre-training methods lead to sparse units, where the sparsity is measured by the ℓ-1 norm. They identified a sufficient condition and demonstrated theoretically and empirically

that if the condition is satisfied, some popular pre-training approaches lead to sparse hidden units, including the pre-training methods based on denoising auto-encoders (DAEs) and RMBs. They also argued that pre-training improves DNN training because of the sparsenes. This argument seems to explain why pre-training is less important for DNNs with the rectifier activation function: the units have been sufficiently sparse already with this type of activation function so the sparsity contributed by the pre-training is less useful.

2.4 Weight sparsity and pruning

The sparsity regularization can also be applied to weights. This is not directly related to sparse coding, but it does help to learn prominent patterns, and can significantly reduce redundant parameters. This reduction of parameters leads to improved efficiency in terms of both statistics and computation.

Weight decay [**?**] and the variants (e.g., [**?**]) employ ℓ-2 norm to encourage weights of small values, while a true sparsity-oriented regularization is based on the ℓ-1 norm [**?**]. Connection pruning is a more efficient process for obtaining weight sparsity. The simplest pruning approach is to remove connections with small weights [**?**]; a more sophisticated approach considers the impact of the removal on the cost function, e.g., the optimal brain damage (OBD) method [**?**,**?**]. The primary advantage of the pruning approach is that it can yield very compact models. For example, it was reported by Liu et al. [**?**] that 90% of connections can be removed with a 1.5% frame accuracy reduction in a speech recognition task. A particular problem with the compact model is that on modern CPUs the operations on sparse matrices are generally much slower than the operations on dense matrices. Recently, Liu et al. [**?**] presented an interesting approach that can speed up sparse matrix multiplication.

2.5 Gradient sparsity

Another interesting approach related to sparse deep neural models is the contractive auto-encoder (CAE) proposed by Rifai et al. [**?**]. In this, the Frobenius norm of the Jacobian of the hidden units is used as a regularization. This 'contraction' penalty is imposed on the gradients and ensures robustness against minor change in the input. This is related to sparse coding, since if the activation function is sigmoid, the contraction penalty tends to drive the hidden unit activations to either zero or one where the gradient is zero.

Alain et al. [**?**] showed that the DAE model [**?**] is closely related to the CAE model, except that the contractive penalty for DAE is imposed on the Jacobian of the output units. Arpit et al. [**?**] presented a further study and showed that under certain conditions, deep neural models encourage sparse representations. Some popular models including DAE and CAE satisfy these conditions and therefore tend to produce sparse representations.

3 Deep sparse models

The sparse deep models described in the previous section focus on deep neural models with sparse regularization. In contrast, deep sparse models focus on sparse models, whilst borrowing the idea of hierarchical processing from deep learning and constructing multi-level sparse models.

For example, Yu et al. [**?**] presented a two-level sparse coding approach. In the first layer, sparse codes are derived from raw bits of an input image. Covariance matrcies of the sparse codes are then computed for neighbouring patches. The second-level sparse codes are then derived from the diagonal vectors of the covariance matrices. These set-level codes are used as features for image classification. By using two-level sparse coding, features are extracted in a hierarchical way, leading to additional abstraction that can represent features of a larger scope.

He et al. [**?**] presented a similar multi-layer sparse coding framework. An innovation is that each sparse coding component is followed by a dense code conversion layer. This framework ensures that higher-level features cover larger scopes of the input than lower-level features, and the neighboring patches are placed close to each other in the dense code space. The entire framework is purely supervised. After the feature learning, a classifier is trained to conduct object recognition.

An interesting work presented by Kavukcuoglu [**?**,**?**] employs neural models to learn sparse codes. This predicted sparse decomposition (PSD) method is much more efficient at run-time and inherits most of the advantages of the conventional sparse coding method.

4 Application in speech and language processing

Both sparse coding and deep learning have been widely employed in speech and language processing, however thorough investigation of deep sparse or sparse deep models is still limited. We start by summarizing the work in sparse coding and deep learning, and then present some recent studies which combine these two models.

Sparse coding Sparse coding has been widely employed in a wide range of speech processing tasks, including but not limited to speech coding [**?**,**?**], speech enhancement [**?**,**?**,**?**], source separation [**?**,**?**], music coding, classification and retrieval [**?**,**?**,**?**], speech recognition [**?**,**?**,**?**,**?**,**?**], overlap detection [**?**], voice activity detection [**?**], and sound localization [**?**].

Sparse coding is also applied to language processing. For example, Zhu et al. [**?**] presented a sparse topic model by introducing ℓ-1 regularization on both the document and word representations. This was later extended to an online version [**?**]. Note that sparse topic models can be extended to a hierarchical structure [**?**]. The probabilistic formulation was later changed to a non-probabilistic formulation that can effectively control the sparsity [**?**]. The idea of sparse topic models was also presented in [**?**], where ℓ-1 regularization was introduced to the conventional latent semantic analysis (LSA). Recently, Liu et al. [**?**] presented a document summary approach using two-level sparse representation.

The application of sparse coding in language processing is far from extensive, when compared to speech processing. A particular reason is that the vanilla sparse coding approach assumes a Gaussian residual, whereas language processing often uses discrete representations that are generally not Gaussian distributed. To solve this problem, Lee [**?**] extended the conventional Gaussian sparse coding to an exponential family sparse coding. This approach was successfully applied to text classification.

Deep learning Deep learning has obtained exceptional results in both speech and language processing, including speech recognition [**?**,**?**], speech synthesis [**?**], speech enhancement [**?**,**?**], speech separation [**?**], language modeling [**?**], semantic parsing [**?**], paraphrase detection [**?**], machine translation [**?**], and sentiment prediction [**?**]. There are numerous papers on deep learning methods for speech and language processing. Readers are referred to the recent book from Goodfellow et al. [**?**] and the deep learning website http://deeplearning.net.

Deep and parse models Just very recently, deep sparse or sparse deep models were employed in speech and language processing. For example, Sivaram et al. [**?**] proposed a regularization in the form $log(1 + v^2)$ in DNN model training, where v is the activation of the unit to regularize. This regularization, is equivalent to a student-t prior on v, was found to deliver better performance in phone recognition [**?**]. Yogatam [**?**] introduced a hierarchical group sparse regularization to derive sparse word vectors. They reported better performance with the group-sparsity in a text classification task. Sun et al. [**?**] proposed an ℓ-1 regularized word embedding algorithm and found that sparsity leads to better performance on a bunch of analogy tasks, and the resulting embedding is more expressive and interpretable. Vyas et al. [**?**] recently presented a sparse bilingual word representation approach for cross-lingual lexical entailment, i.e., detect whether the meaning of a word in one language can be inferred from the meaning of a word in another language.

It should be emphasized that the studies mentioned are just a part of the contemporary work towards deep sparse and sparse deep learning. There may be some interesting research missed in our review, but generally the work in this direction is rather limited.

5 Conclusions and future directions

We gave a quick review of sparse and deep learning methods in machine learning. Both approaches can find strong physiological support and have demonstrated success in a wide range of machine learning tasks. It has been commonly adopted that sparse models can learn prominent patterns and are highly robust due to the redundancy in the code. Deep learning, on the other hand, has an advantage as it can learn high-level abstraction, which is more invariant and transferable across conditions and domains.

A particulary interesting question is how sparsity and hierarchy interact and complement each other to support the complex functions of human brains. It seems natural to believe that sparse coding plays the role of robust information representation, while the hierarchical structure plays the role of knowledge deduction and induction. However,

how the two ingredients are integrated in a biological neural system and the mechanism driving hierarchical sparse information processing is far from known and requires further investigation.

Although the physiological mechanism remains unclear for machine learning researchers, pioneering work has been conducted in several groups, with the aim to leverage the respective advantages of deep and sparse models, as discussed in our paper. Interestingly, these studies demonstrate that both sparse deep models and deep sparse models are highly promising: in the former case, sparse regularizations encourage more plausible local patterns, while in the later case, deep structures yield large-scope representations.

It seems that most of the successful deep and sparse research resides in the unsupervised learning paradigm, e.g., with (stacked) RBMs or DAEs. Investigations into how the patterns can be learned with explicit supervision seem interesting and can be demonstrated in a semi-supervised framework, as shown in Larochelle's work [**?**]. Additionally, it is reasonable to hypothesize that different types of sparse regularization may contribute to the neural system in different but collaborative ways. For example, it is possible that the code sparsity and weight sparsity collaborate together to determine the characteristics of the human neural system. More investigation should be conducted on the impact of multiple sparse regularizations. Another interesting question is how sparsity functions differently at different layers in a deep structure. It is known that in deep neural models, the responses at high-level layers are more sparse than at low-level layers. This sparsity is mainly attributed to the layer-wised information disentanglement [**?**]. It is then interesting to know if sparse coding still contributes to form representations at high levels. If the answer is 'yes', the challenge is to understand how the sparsity is derived from the two forces, i.e., sparse-oriented regularization and the nature of high-level abstraction.

We also reviewed some papers on speech and language processing that employ deep sparse or sparse deep models. For speech processing, there are numerous studies on both sparse and deep models, however we didn't find much literature combining the two techniques. It is common practice for speech researchers to try ℓ-1 or ℓ-2 regularization when training deep neural models, but little work treats sparse representations seriously in the deep architecture. For language processing, neither sparse nor deep models were widely studied until very recently. One reason is that traditional language processing methods focus on symbolic representations, e.g., words, phrases, tags. These representations are discrete and are not amiable to both sparse and deep models. Thanks to the embedding technique, continuous representations (e.g., word vectors) have become popular, which in turn have motivated the investigation and application of deep and sparse learning in language processing, as already discussed in the paper.

We expect that more interesting findings will be obtained for speech and language processing, through novel combinations of sparse and deep models. In the past, researchers hoped to learn acoustic or semantic patterns by either sparse coding or deep learning, and both directions seem fruitful, e.g., [**?**,**?**]. However, knowledge is still limited on how the patterns learned by the two very different methods differ from each other and which approach delivers more 'plausible' patterns. Importantly, can we combine the two approaches together to better learn patterns? If we accept the argument

that the two mechanisms function in a collaborative way in our brain, then we posit exploiting them in a unified framework to decipher the complex information in human speech and language, as our brain seamlessly does every day.

Acknowledgments.

This research was supported by the RSE-NSFC joint project (No. 61411130162), the National Science Foundation of China (NSFC) under the project No. 61371136, the UK Engineering and Physical Sciences Research Counci Grant (EPSRC) Grant No. EP/M026981/1, and the MESTDC PhD Foundation Project No. 20130002120011. It is also supported by Huilan Ltd, Tongfang Corp., and FreeNeb.

References

1. Alain, G., Bengio, Y.: What regularized auto-encoders learn from the data-generating distribution. Journal of Machine Learning Research 15(1), 3563–3593 (2014)
2. Arpit, D., Zhou, Y., Ngo, H., Govindaraju, V.: Why regularized auto-encoders learn sparse representation? arXiv preprint arXiv:1505.05561 (2015)
3. Asaei, A., Taghizadeh, M.J., Haghighatshoar, S., Raj, B., Bourlard, H., Cevher, V.: Binary sparse coding of convolutive mixtures for sound localization and separation via spatialization. IEEE Transactions on Signal Processing 64(3), 567–579 (2016)
4. Barlow, H.: Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1, 371–394 (1972)
5. Benesty, J.: Springer handbook of speech processing. Springer Science & Business Media (2008)
6. Bengio, Y.: Learning deep architectures for ai. Foundations and trends® in Machine Learning 2(1), 1–127 (2009)
7. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: ICML Unsupervised and Transfer Learning (2012)
8. Blumensath, T., Davies, M.: Sparse and shift-invariant representations of music. IEEE Transactions on Audio, Speech, and Language Processing 14(1), 50–57 (2006)
9. Bordes, A., Glorot, X., Weston, J.: Joint learning of words and meaning representations for open-text semantic parsing. In: International Conference on Artificial Intelligence and Statistics (2012)
10. Cho, K., Merrienboer, B.V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. Computer Science (2014)
11. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. pp. 160–167. ACM (2008)
12. Cun, Y.L., Denker, J.S., Solla, S.A.: Optimal brain damage. In: Proc. NIPS'90 (1990)
13. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4688–4691 (2011)
14. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20(1), 30–42 (2012)

15. Dayan, P., Abbott, L.F.: Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT press (2001)
16. Fahlman, S.E., Hinton, G.E.: Connectionist architectures for artificial intelligence. Computer;(United States) 20(1) (1987)
17. Földiák, P., Young, M.P.: Sparse coding in the primate cortex. The handbook of brain theory and neural networks 1, 1064–1068 (1995)
18. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the 14th international conference on Artificial Intelligence and Statistics (AISTATS). pp. 315–323 (2011)
19. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning (2016), http://www.deeplearningbook.org, book in preparation for MIT Press
20. Goodfellow, I., Courville, A., Bengio, Y.: Large-scale feature learning with spike-and-slab sparse coding. In: International Conference on Machine Learning. pp. 1439–1446 (2012)
21. He, Y., Kavukcuoglu, K., Wang, Y., Szlam, A., Qi, Y.: Unsupervised feature learning by deep sparse coding. arXiv preprint arXiv:1312.5783 (2013)
22. Huang, P., Kim, M., Hasegawajohnson, M., Smaragdis, P.: Deep learning for monaural speech separation. In: ICASSP 2014 (2014)
23. Jaitly, N.: Exploring Deep Learning Methods for Discovering Features in Speech Signals. Ph.D. thesis, University of Toronto (2014)
24. Kavukcuoglu, K., Fergus, R., LeCun, Y., et al.: Learning invariant features through topographic filter maps. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1605–1612. IEEE (2009)
25. Kavukcuoglu, K., Ranzato, M., LeCun, Y.: Fast inference in sparse coding algorithms with applications to object recognition. arXiv preprint arXiv:1010.3467 (2010)
26. Klein, D.J., König, P., Körding, K.P.: Sparse spectrotemporal coding of sounds. EURASIP Journal on Advances in Signal Processing 2003(7), 1–9 (2003)
27. Krogh, A., Hertz, J.A.: A simple weight decay can improve generalization. In: Advances in Neural Information Processing Systems (NIPS). vol. 4, pp. 950–957 (1992)
28. Larochelle, H., Bengio, Y.: Classification using discriminative restricted boltzmann machines. In: Proceedings of the 25th international conference on Machine learning. pp. 536–543. ACM (2008)
29. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)
30. Lee, H.: Unsupervised feature learning via sparse hierarchical representations. Ph.D. thesis, STANFORD UNIVERSITY (2010)
31. Lee, H., Ekanadham, C., Ng, A.Y.: Sparse deep belief net model for visual area v2. In: Advances in neural information processing systems. pp. 873–880 (2008)
32. Li, J., Zhang, T., Luo, W., Yang, J., Yuan, X.T., Zhang, J.: Sparseness analysis in the pretraining of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems PP(99), 1–14 (2016)
33. Li, J., Chang, H., Yang, J.: Sparse deep stacking network for image classification. arXiv preprint arXiv:1501.00777 (2015)
34. Liu, B., Wang, M., Foroosh, H., Tappen, M., Pensky, M.: Sparse convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 806–814 (2015)
35. Liu, C., Zhang, Z., Wang, D.: Pruning deep neural networks by optimal brain damage. In: Interspeech'14 (2014)
36. Liu, H., Yu, H., Deng, Z.: Multi-document summarization based on two-level sparse representation model. In: National Conference on Artificial Intelligence (2015)
37. Luo, H., Shen, R., Niu, C.: Sparse group restricted boltzmann machines. arXiv preprint arXiv:1008.4988 (2010)

38. Luo, Y., Bao, G., Xu, Y., Ye, Z.: Supervised monaural speech enhancement using complementary joint sparse representations. IEEE Signal Processing Letters 23(2), 237–241 (2016)
39. Makhzani, A., Frey, B.: A winner-take-all method for training sparse convolutional autoencoders. In: NIPS Deep Learning Workshop (2014)
40. Martin, J.H., Jurafsky, D.: Speech and language processing. International Edition (2000)
41. Mikolov, T.: Statistical Language Models Based on Neural Networks. Ph.D. thesis, Brno University of Technology (2012)
42. Nam, J., Herrera, J., Slaney, M., Smith, J.O.: Learning sparse feature representations for music annotation and retrieval. In: ISMIR. pp. 565–570 (2012)
43. Northoff, G.: Unlocking the Brain, Volume 1: Coding. Oxford (2014)
44. Ogrady, P.D., Pearlmutter, B.A.: Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. Neurocomputing 72(1), 88–101 (2008)
45. O'Grady, P.D., Pearlmutter, B.A., Rickard, S.T.: Survey of sparse and non-sparse methods in source separation. International Journal of Imaging Systems and Technology 15(1), 18–33 (2005)
46. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision Research 37(23), 3311–3325 (1997)
47. Plumbley, M.D., Blumensath, T., Daudet, L., Gribonval, R., Davies, M.E.: Sparse representations in audio and music: from coding to source separation. Proceedings of the IEEE 98(6), 995–1005 (2010)
48. Poultney, C., Chopra, S., Cun, Y.L., et al.: Efficient learning of sparse representations with an energy-based model. In: Advances in neural information processing systems. pp. 1137–1144 (2006)
49. aurelio Ranzato, M., lan Boureau, Y., Cun, Y.L.: Sparse feature learning for deep belief networks. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (eds.) Advances in Neural Information Processing Systems 20, pp. 1185–1192. Curran Associates, Inc. (2008), http://papers.nips.cc/paper/3363-sparse-feature-learning-for-deep-belief-networks.pdf
50. Rifai, S., Vincent, P., Muller, X., Glorot, X., Bengio, Y.: Contractive auto-encoders: Explicit invariance during feature extraction. In: International Conference on Machine Learning, 2011 (2011)
51. Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., Poggio, T.: A quantitative theory of immediate visual recognition. Progress in brain research 165, 33–56 (2007)
52. Setiono, R.: A penalty function approach for prunning feedforward neural networks. Neural computation 9(1), 185–204 (1994)
53. Sigg, C.D., Dikk, T., Buhmann, J.M.: Speech enhancement with sparse coding in learned dictionaries. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4758–4761. IEEE (2010)
54. Sivaram, G.S.V.S., Nemala, S.K., Elhilali, M., Tran, T.D., Hermansky, H.: Sparse coding for speech recognition. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4346–4349 (March 2010)
55. Sivaram, G.S., Hermansky, H.: Multilayer perceptron with sparse hidden outputs for phoneme recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 5336–5339. IEEE (2011)
56. Socher, R., Huang, E.H., Pennington, J., Ng, A.Y., Manning, C.D.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Advances in Neural Information Processing Systems 24, 801–809 (2011)
57. Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John Mcintyre Con-

ference Centre, Edinburgh, Uk, A Meeting of Sigdat, A Special Interest Group of the ACL. pp. 151–161 (2011)
58. Sun, F., Guo, J., Lan, Y., Xu, J., Cheng, X.: Sparse word embeddings using l-1 regularized online learning. In: IJCAI 2016. pp. 2915–2921 (2016)
59. Teng, P., Jia, Y.: Voice activity detection using convolutive non-negative sparse coding. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7373–7377. IEEE (2013)
60. Utgoff, P.E., Stracuzzi, D.J.: Many-layered learning. Neural Computation 14(10), 2497–2529 (2002)
61. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11(Dec), 3371–3408 (2010)
62. Vinyals, O., Deng, L.: Are sparse representations rich enough for acoustic modeling? In: INTERSPEECH. pp. 2570–2573 (2012)
63. Vipperla, R., Bozonnet, S., Wang, D., Evans, N.: Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization. Proc. CHiME pp. 74–79 (2011)
64. Vipperla, R., Geiger, J.T., Bozonnet, S., Wang, D., Evans, N., Schuller, B., Rigoll, G.: Speech overlap detection and attribution using convolutive non-negative sparse coding. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4181–4184. IEEE (2012)
65. Vyas, Y., Carpuat, M.: Sparse bilingual word representations for cross-lingual lexical entailment. In: NAACL 2016. pp. 1187–1197 (2016)
66. Wang, D., Tejedor, J.: Heterogeneous convolutive non-negative sparse coding. In: INTERSPEECH. pp. 2150–2153 (2012)
67. Wang, D., Vipperla, R., Evans, N., Zheng, T.F.: Online non-negative convolutive pattern learning for speech signals. IEEE Transactions on Signal Processing 61(1), 44–56 (2013)
68. Wang, D., Vipperla, R., Evans, N.W.: Online pattern learning for non-negative convolutive sparse coding. In: INTERSPEECH. pp. 65–68 (2011)
69. Wu, C., Yang, H., Zhu, J., Zhang, J., King, I., Lyu, M.R.: Sparse poisson coding for high dimensional document clustering. In: IEEE International Conference on Big Data (2013)
70. Xu, T., Wang, W., Dai, W.: Sparse coding with adaptive dictionary learning for underdetermined blind speech separation. Speech Communication 55(3), 432–450 (2013)
71. Xu, Y., Du, J., Dai, L., Lee, C.: A regression approach to speech enhancement based on deep neural networks. IEEE Trans. on audio, speech and langauge processing (2015)
72. Yogatama, D.: Sparse Models of Natural Language Text. Ph.D. thesis, Carnegie Mellon University (2015)
73. Yu, D., Deng, L.: Automatic Speech Recognition: A Deep Learning Approach. Springer Publishing Company, Incorporated (2014)
74. Yu, D., Seide, F., Li, G., Deng, L.: Exploiting sparseness in deep neural networks for large vocabulary speech recognition. In: Proc. ICASSP2012 (2012)
75. Yu, K., Lin, Y., Lafferty, J.: Learning image representations from the pixel level via hierarchical sparse coding. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 1713–1720. IEEE (2011)
76. Zen, H., Senior, A.W., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: ICASSP2013 (2013)
77. Zhang, A., Zhu, J., Zhang, B.: Sparse online topic models. In: WWW 2013 (2013)
78. Zhao, M., Wang, D., Zhang, Z., Zhang, X.: Music removal by denoising autoencoder in speech recognition. In: APSIPA 2015 (2015)
79. Zhu, J., Xing, E.P.: Sparse topical coding. In: UAI 2012 (2012)