Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Sparse Boltzmann Machines with Structure Learning as Applied to Text Analysis

Zhourong Chen,* Nevin L. Zhang,* Dit-Yan Yeung, Peixian Chen
Hong Kong University of Science and Technology
{zchenbb,lzhang,dyyeung,pchenac}@cse.ust.hk

Abstract

We are interested in exploring the possibility and benefits of structure learning for deep models. As the first step, this paper investigates the matter for *Restricted Boltzmann Machines (RBMs)*. We conduct the study with Replicated Softmax, a variant of RBMs for unsupervised text analysis. We present a method for learning what we call *Sparse Boltzmann Machines*, where each hidden unit is connected to a subset of the visible units instead of all of them. Empirical results show that the method yields models with significantly improved model fit and interpretability as compared with RBMs where each hidden unit is connected to all visible units.

Introduction

Deep learning has achieved great successes in recent years. It has produced superior results in a range of applications, including image classification (Krizhevsky, Sutskever, and Hinton 2012), speech recognition (Hinton et al. 2012; Mikolov et al. 2011), language translation (Sutskever, Vinyals, and Le 2014) and so on. It is now time to ask whether it is possible and beneficial to learn structures for deep models.

To learn the structure of a deep model, we need to determine the number of hidden layers and the number of hidden units at each layer. More importantly, we need to determine the connections between neighboring layers. This implies that we need to talk about sparse models where neighboring layers are not fully connected.

Sparseness is desirable and full connectivity is unnecessary. In fact, (Han et al. 2015) have shown that many weak connections in the fully connected layers of *Convolutional Neural Networks (CNNs)* (Lecun et al. 1998) can be pruned without incurring any accuracy loss. The convolutional layers of CNNs are sparse, and the fact is considered one of the key factors that have led to the success of CNNs. Moreover, it is well known that overfitting is a serious problem in deep models. One method to address the problem is dropout (Srivastava et al. 2014), which randomly drops out units (while keeping full connectivity) during training. The possibility of randomly dropping connections has also been explored in (Wan et al. 2013). Sparseness offers an interesting alternative. It amounts to deterministically dropping out connections.

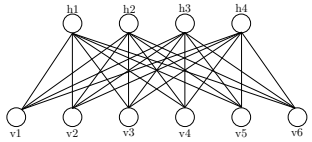

Figure 1: An example RBM with $K = 6$ and $F = 4$.

How can one learn sparse deep models? One method is to first learn a fully connected model and then prune weak connections (Han et al. 2015). The drawbacks of this method are that it is computationally wasteful and does not provide a way to determine the number of hidden units. We would like to develop a method that determines the number of hidden units and the connections between units automatically. The key intuition is that a hidden unit should be connected to a group of strongly correlated units at the level below. This idea is used in the convolutional layers of CNNs, where a unit is connected to pixels in a small patch of an image. In image analysis, spatial proximity implies strong correlation.

To apply the intuition to applications other than image analysis, we need to identify groups of strongly correlated variables for which latent variables should be introduced. *Hierarchical Latent Tree Analysis (HLTA)* (Liu et al 2014, Chen et al 2016) offers a plausible solution. HLTA first partitions all the variables into groups such that the variables in each group are strongly correlated and the correlations can be properly modelled using a single latent variable. It then introduces a latent variable for each group. After that it converts the latent variables into observed variables via data completion and repeats the process to produce a hierarchy. The output of HLTA is a hierarchical latent tree model where the observed variables are at the bottom and there are multiple layers of latent variables on top. To obtain a non-tree sparse deep model, we propose to use the tree model as a skeleton and introduce additional connections to model the residual correlations not captured by the tree.

In this paper, we fully develop and test the idea in the context of RBMs, which have a single layer of hidden units and are the building blocks of Deep Belief Networks (Hin-

*Corresponding authors.

ton, Osindero, and Teh 2006). The target domain is unsupervised text analysis. We present an algorithm for learning what we call *Sparse Boltzmann Machines (SBMs)*. Empirically, we show that the full-connectivity restriction of RBMs can easily lead to overfitting, and that SBMs are effective in avoiding overfitting. We also demonstrate that Sparse Boltzmann Machines are more interpretable than RBMs.

Related Works

The term sparse RBMs first appeared in (Lee, Ekanadham, and Ng 2008), where it was used to refer to sparse hidden unit activations rather than sparse connections. (Adams, Wallach, and Ghahramani 2010) proposed to learn sparse structure for deep directed belief networks by introducing the cascading Indian buffet process, which was very time-consuming.

Network pruning is also a potential way to optimize the structure of a neural network. Biased weight decay was the early approach to pruning. Later, Optimal Brain Damage (Cun, Denker, and Solla 1990) and Optimal Brain Surgeon (Hassibi, Stork, and Com 1993) suggested that magnitude-based pruning may not be the best strategy and they proposed pruning methods based on the Hessian of the loss function. With respect to deep neural networks, (Han et al. 2015) proposed to compress a network through a three-step process: train, prune connections, and retrain. We call it redundancy pruning. In contrast, (Srinivas and Babu 2015) proposed to prune redundant neurons directly. They all reduced the number of parameters vastly with slight or even no performance loss. The drawback of network pruning is that the original networks should be large enough and hence some computation would be wasted on those unnecessary parameters during pre-training.

Restricted Boltzmann Machines

An *Restricted Boltzmann Machine (RBM)* (Smolensky 1986) is a two-layer undirected graphical model with a layer of K visible units $\{v^1, \ldots, v^K\}$ and a layer of F hidden units $\{h_1, \ldots, h_F\}$. The two layers are fully connected to each other, while there are no connections between units at the same layer. An example is shown in Figure 1. In the simplest case, all the units are assumed to be binary. An energy function is defined over all the units as follows:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{j=1}^{F}\sum_{k=1}^{K} W_j^k h_j v^k - \sum_{k=1}^{K} v^k b^k - \sum_{j=1}^{F} h_j a_j, \quad (1)$$

where a_j and b^k are bias parameters for the hidden and visible units respectively, while W_j^k is the connection weight between hidden unit h_j and visible unit v^k. The energy function defines a joint probability over $\mathbf{v}$ and $\mathbf{h}$ as follows:

$$P(\mathbf{v}, \mathbf{h}) = \exp(-E(\mathbf{v}, \mathbf{h}))/Z, \quad (2)$$

where $Z = \sum_{\mathbf{v}', \mathbf{h}'} \exp(-E(\mathbf{v}', \mathbf{h}'))$ is a normalization term called the partition function. An important property of RBM is that the conditional distributions $P(\mathbf{h}|\mathbf{v})$ and $P(\mathbf{v}|\mathbf{h})$ factorize as below:

$$P(\mathbf{h}|\mathbf{v}) = \prod_j P(h_j|\mathbf{v}) \qquad P(\mathbf{v}|\mathbf{h}) = \prod_k P(v^k|\mathbf{h}) \quad (3)$$

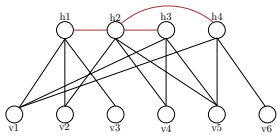

Figure 2: An example SBM with $K = 6$ and $F = 4$.

$$P(h_j = 1|\mathbf{v}) = \sigma(a_j + \sum_{k=1}^{K} W_j^k v^k) \quad (4)$$

$$P(v^k = 1|\mathbf{h}) = \sigma(b^k + \sum_{j=1}^{F} W_j^k h_j) \quad (5)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. The model parameters of an RBM are learned using the *Contrastive Divergence (CD)* algorithm (Hinton 2002), which maximizes the data likelihood via stochastic gradient descent.

In (Hinton and Salakhutdinov 2009), RBM was used for topic modeling and the proposed model was called Replicated Softmax. Suppose the vocabulary size is K. Let us represent a document with D tokens as a binary matrix $\mathcal{U}$ of size $K * D$ with $u_i^k = 1$ if the i^{th} token is the k^{th} word in the vocabulary. The energy function of a document $\mathcal{U}$ and hidden units $\mathbf{h}$ is defined as follows:

$$E(\mathcal{U}, \mathbf{h}) = -\sum_{j=1}^{F}\sum_{k=1}^{K} W_j^k h_j \hat{u}^k - \sum_{k=1}^{K} \hat{u}^k b^k - D\sum_{j=1}^{F} h_j a_j, \quad (6)$$

where $\hat{u}^k = \sum_{i=1}^{D} u_i^k$ denotes the count for the k^{th} word. The conditional probability $P(h_j = 1|\mathcal{U})$ is:

$$P(h_j = 1|\mathcal{U}) = \sigma(Da_j + \sum_{k=1}^{K} W_j^k \hat{u}^k). \quad (7)$$

The motivation behind Replicated Softmax is to properly model word counts in documents of varying lengths through weight sharing. It was shown to generalize better than *Latent Dirichlet Allocation (LDA)* (Blei, Ng, and Jordan 2003) in terms of log-probability on held-out documents and accuracy on retrieval tasks. In this paper, we use Replicated Softmax for text analysis.

Sparse Boltzmann Machines

In this section, we propose our new models, Sparse Boltzmann Machines (SBMs). An SBM is a two-layer undirected graphical model with a layer of K visible units $\{v^1, \ldots, v^K\}$ and a layer of F hidden units $\{h_1, \ldots, h_F\}$. The hidden units in SBMs are directly linked up to form a tree structure, while each hidden unit is also individually connected to a subset of the visible units. See Figure 2 for an example SBM. In SBMs, the number of hidden units and the connectivities are both learned from data.

One technical difference between SBMs and RBMs is that there are direct connections among the hidden units in SBMs. We call them hidden connections. The reason we introduce the hidden connections into our models is that, the hidden connections provide a way to relate a hidden unit to

a visible unit without a direct connection. For example, in Figure 2, hidden unit h_1 is not directly connected to visible unit v_4. However, the existence of the hidden connection between h_1 and h_2 introduces a path connecting h_1 and v_4, which can help us to better model the correlation between the two units. This is crucial in reducing the number of connections between hidden units and visible units. To avoid the connections among the hidden units becoming too dense, we restrict them to form a tree structure. Tree structures among hidden units were used before in *Boltzmann Trees* (Saul and Jordan 1994). In *Restricted Boltzmann Forest* (Larochelle, Bengio, and Turian 2010), the activations of hidden units were also constrained to follow a tree-based rule. However those trees were determined manually rather than learned from data. Moreover, the hidden and visible layers were fully connected.

Parameter Learning

SBMs also can be extended for text analysis as RBMs are extended to Replicated Softmax. Here we introduce SBMs in the context of Replicated Softmax and use the same notations as in the previous section. Let $\mathcal{G}$ be a graph representing the model structure. Edge (j,k) belongs to $\mathcal{G}$ if and only if there is a link between visible unit v^k and hidden unit h_j. Edge (j,l) belongs to $\mathcal{G}$ if and only if there is a link between hidden unit h_j and hidden unit h_l ($j<l$). Also let W_{jl} be the weight on the connection between h_j and h_l. Then the energy function of an SBM for a document $\mathcal{U}$ and hidden units $\mathbf{h}$ is as below:

$$\begin{aligned} E(\mathcal{U},\mathbf{h}) = &- \sum_{(j,k)\in\mathcal{G}} W_j^k h_j \hat{u}^k - \sum_{k=1}^{K} \hat{u}^k b^k \\ &- D\sum_{j=1}^{F} h_j a_j - D\sum_{(j,l)\in\mathcal{G}} W_{jl} h_j h_l. \end{aligned} \tag{8}$$

Similar to Replicated Softmax, our model defines the joint distribution as:

$$P(\mathcal{U},\mathbf{h}) = \frac{1}{Z}\exp(-E(\mathcal{U},\mathbf{h})), \tag{9}$$

where $Z = \sum_{\mathcal{U}'}\sum_{\mathbf{h}}\exp(-E(\mathcal{U}',\mathbf{h}))$. Note that the summation over $\mathcal{U}'$ is done over all the possible documents with the same length as $\mathcal{U}$.

Let $\bar{\mathcal{U}} = \{\mathcal{U}_n\}_{n=1}^N$ be a collection of N documents with potentially different lengths $D_1, \ldots, D_N$. We assume that $P(\bar{\mathcal{U}}) = \prod_{n=1}^N P(\mathcal{U}_n)$, where $P(\mathcal{U}_n) = \sum_{\mathbf{h}} P(\mathcal{U}_n,\mathbf{h})$. The objective of training an SBM for $\bar{\mathcal{U}}$ is to maximize the log-likelihood of the documents $\log P(\bar{\mathcal{U}})$. We maximize the objective function via stochastic gradient descent. The partial derivatives of $\log P(\bar{\mathcal{U}})$ w.r.t the parameters W_j^k, b^k and a_j remain the same as in Replicated Softmax:

$$\frac{\partial \log P(\bar{\mathcal{U}})}{\partial W_j^k} = \sum_{n=1}^{N}(E_{P(h_j|\mathcal{U}_n)}[h_j\hat{u}_n^k] - E_{P(\mathcal{U},\mathbf{h})}[h_j\hat{u}^k]) \tag{10}$$

$$\frac{\partial \log P(\bar{\mathcal{U}})}{\partial b^k} = \sum_{n=1}^{N}(\hat{u}_n^k - E_{P(\mathcal{U})}[\hat{u}^k])$$

$$\frac{\partial \log P(\bar{\mathcal{U}})}{\partial a_j} = \sum_{n=1}^{N} D_n(E_{P(h_j|\mathcal{U}_n)}[h_j] - E_{P(h_j)}[h_j])$$

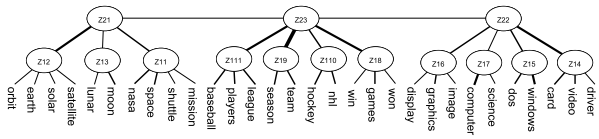

Figure 3: An example HLTM from (Chen et al. 2016).

while the partial derivative of $\log P(\bar{\mathcal{U}})$ w.r.t the new parameter W_{jl} for fixed j and l is:

$$\frac{\partial \log P(\bar{\mathcal{U}})}{\partial W_{jl}} = \sum_{n=1}^{N} D_n(E_{P(\mathbf{h}|\mathcal{U}_n)}[h_j h_l] - E_{P(\mathbf{h})}[h_j h_l]) \tag{11}$$

The first terms in these partial derivatives require the computation of the conditional probabilities $P(h_j|\mathcal{U}_n)$ and $P(\mathbf{h}|\mathcal{U}_n)$. In Replicated Softmax, $P(h_j|\mathcal{U}_n)$ can be calculated using Equation (7). While in SBMs, due to the connections between hidden units, $P(\mathbf{h}|\mathcal{U}_n)$ no longer factorizes and hence Equation (7) cannot be applied. Nevertheless, since the hidden units in SBMs are linked as a tree, we can easily compute the value of $P(h_j|\mathcal{U}_n)$ and $P(\mathbf{h}|\mathcal{U}_n)$ by conducting message propagation (Murphy 2012) in the model.

The second terms in these derivatives require taking an expectation with respect to the distribution defined by the model, which is intractable. Thus as in Replicated Softmax, we adopt the CD algorithm to approximate the second terms by running Gibbs sampling chains in the model. Specifically, the Gibbs chains are initialized at the training data and run for T full steps to draw samples from the model. In SBMs, given a document $\mathcal{U}$ and the value of all the other hidden units $\mathbf{h}_{-j}$, the conditional probability to sample h_j is:

$$\begin{aligned} P(h_j = 1|\mathcal{U},\mathbf{h}_{-j}) = &\sigma(\sum_{(j,k)\in\mathcal{G}} W_j^k\hat{u}^k + Da_j + \\ &D\sum_{(j,l)\in\mathcal{G}} W_{jl}h_l + D\sum_{(l,j)\in\mathcal{G}} W_{lj}h_l), \end{aligned}$$

while the conditional probability to sample a visible unit remains the same as in Replicated Softmax.

Structure Learning

We regard SBMs as a method to model correlations among the visible units. Learning an SBM hence amounts to building a latent structure to explain the correlations. Recently, (Liu, Zhang, and Chen 2014) and (Chen et al. 2016) proposed a method, called HLTA, for learning a *Hierarchical Latent Tree Model (HLTM)* from data. Our structure learning algorithm for SBMs is built upon their work. We expand the tree model from HLTA to obtain the structure of an SBM.

HLTA learns a tree model $\mathcal{T}$ with a layer of observed variables at the bottom and multiple layers of latent variables. Note that the visible units and hidden units in SBMs are called observed variables and latent variables in HLTM respectively. Figure 3 and the left panel in Figure 4 illustrate example models that HLTA produces. Each latent variable in the model is connected to a set of highly-correlated variables in the layer below. The number of latent variables at each layer is determined automatically by the algorithm. The

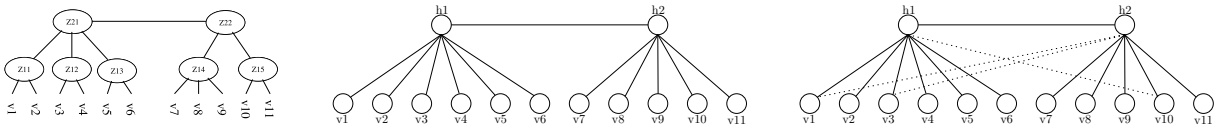

Figure 4: Structure learning for SBMs: A three layer HLTM is first learned (left). The hidden variables at the top level are used to build a skeleton for an SBM (middle). An SBM is finally obtained by adding connections to the skeleton (right).

number L of latent layers in $\mathcal{T}$ is controllable. In this paper, we set $L = 2$. Let H_l be the l^{th} latent layer in $\mathcal{T}$. Also let $\mathbf{V}_Z$ be the set of observed variables which are located in the subtree rooted at latent variable Z in $\mathcal{T}$.

To build the structure of an SBM from $\mathcal{T}$, we first remove all the latent layers except the top layer H_L. Then we connect each latent variable Z in H_L to the set of observed variables $\mathbf{V}_Z$. We use the resulting structure as a skeleton $\mathcal{T}'$ of the corresponding SBM. This is illustrated in Figure 4, where the hidden units h_1, h_2 in SBM correspond to Z_{21}, Z_{22} in $\mathcal{T}$ respectively. Note that the skeleton is still a tree structure, where each node has only one parent.

As to remove the tree-structure constraint, we conduct an expansion step to increase the number of "fan-out" connections for each hidden unit in $\mathcal{T}'$. The key question is how to determine the new set of visible units that a hidden unit should be connected to. We introduce our method using Z_{21} (correspondingly h_1 in $\mathcal{T}'$) and v_7 in Figure 4 as an example. To determine whether Z_{21} should also be connected to v_7, we consider the empirical conditional mutual information $I(Z_{21}, v_7|Z_{22}, \bar{\mathcal{U}})$, where Z_{22} is the root of the subtree that v_7 is in. To estimate the value, we first estimate the empirical joint distribution $\hat{p}(Z_{21}, Z_{22}, v_7)$. We go through all the documents and compute $p(Z_{21}, Z_{22}|\mathcal{U}_n)$ for each document $\mathcal{U}_n$ in $\bar{\mathcal{U}}$ by conducting inference in $\mathcal{T}$. Then we collect the statistics of Z_{21}, Z_{22} and v_7 to get $\hat{p}(Z_{21}, Z_{22}, v_7)$. After that, $I(Z_{21}, v_7|Z_{22}, \bar{\mathcal{U}})$ can be estimated as:

$$I(Z_{21}, v_7|Z_{22}, \bar{\mathcal{U}}) = \sum_{Z_{22}} \hat{p}(Z_{22}) \sum_{v_7} \sum_{Z_{21}} \hat{p}(Z_{21}, v_7|Z_{22}) log \frac{\hat{p}(Z_{21}, v_7|Z_{22})}{\hat{p}(Z_{21}|Z_{22})\hat{p}(v_7|Z_{22})}.$$

All the distributions in the above formula can be derived from the joint distribution $\hat{p}(Z_{21}, Z_{22}, v_7)$.

If the correlation between Z_{21} and v_7 is properly modeled in $\mathcal{T}$, the two variables should be conditionally independent given Z_{22}, and hence $I(Z_{21}, v_7|Z_{22}, \bar{\mathcal{U}})$ should be 0. Therefore, if $I(Z_{21}, v_7|Z_{22}, \bar{\mathcal{U}})$ is not 0, then we can conclude that the correlation between Z_{21} and v_7 is not properly modeled in the model, and the model needs to be expanded by adding new connections between the two variables.

Our algorithm, called *SBM-SFC (SBM-Structure from Correlation)*, is given in Algorithm 1. It considers the latent variables one at a time. For a given latent variable Z (suppose the corresponding hidden unit in $\mathcal{T}'$ is h), it computes the conditional mutual information between Z and each unconnected observed variable, and sorts the observed variables in descending order with respect to the conditional mutual information. Then in $\mathcal{T}'$, it connects hidden unit h to the visible units corresponding to the top M observed variables with the highest conditional mutual information. M is a predefined parameter, which normally is set to the value such that each hidden unit is connected to $0.2 * K$ visible units. After the above expansion step is done for each hidden unit in $\mathcal{T}'$, the whole structure of an SBM is determined.

Algorithm 1 SBM-SFC($\mathcal{T}$)

Inputs: $\mathcal{T}$—Graph of an HLTM, $\bar{\mathcal{U}}$—Collection of training documents, M—Number of new connections for each hidden unit.

Outputs: Graph $\mathcal{T}'$ of a corresponding SBM.

$\mathcal{T}' \leftarrow \emptyset$, $H_L \leftarrow$ graph of the top latent layer in $\mathcal{T}$
$V \leftarrow$ observed variables in $\mathcal{T}$
$\mathcal{T}'.add\_graph(H_L), \mathcal{T}'.add\_units(V)$
for variable Z in H_L **do**
  $V_Z \leftarrow$ observed variables in subtree rooted at variable Z
  $\mathcal{T}'.add\_edges(Z, V_Z), I \leftarrow \emptyset$
  for V' in $(V - V_Z)$ **do**
    $Z' \leftarrow$ root of the subtree containing V'
    $I_{Z,V'} \leftarrow I(Z, V'|Z', \bar{\mathcal{U}})$
    $I.add(I_{Z,V'})$
  end for
  $I \leftarrow$ sort(I, 'descend')
  for V' in the top M_Z pairs in I **do**
    $\mathcal{T}'.add\_edge(V', Z)$
  end for
end for
return $\mathcal{T}'$

Experiments

In this section we test the performance of SBMs on three text datasets of different scales: NIPS proceeding papers[1], CiteULike articles[2], and New York Times dataset[3]. Experimental results show that SBMs perform consistently well over the three datasets in terms of model generalizability, and SBMs always give much better interpretability.

Datasets

NIPS proceeding papers consist of 1,740 NIPS papers published from 1987 to 1999. We randomly sample 1,640 papers as training data, 50 as validation data and the remaining 50 as test data. We choose the 1,000 most frequent words and each document is represented as a vector of 1,000 dimensions, with each element being the number of times a word appears in the current document.

[1] http://www.cs.nyu.edu/ roweis/data.html
[2] http://www.wanghao.in/data/ctrsr_datasets.rar
[3] http://archive.ics.uci.edu/ml/datasets/Bag+of+Words

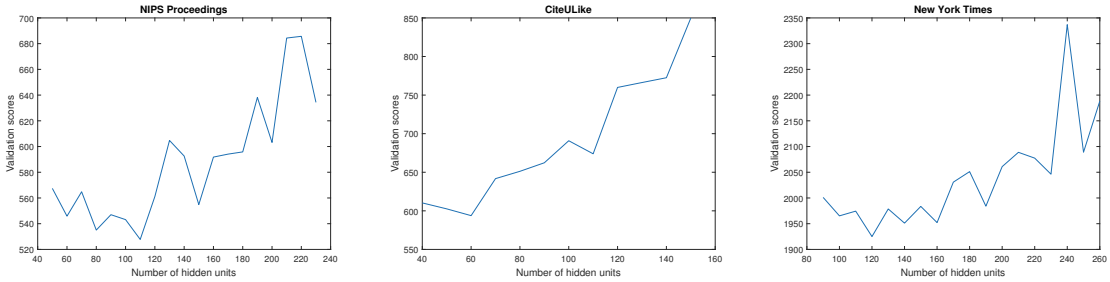

Figure 5: The generalization performance of Replicated Softmax with different number of hidden units.

CiteULike article collection contains 16,980 articles. Similarly, we randomly divide it into training data with 12,000 articles, validation data with 1,000 articles and test data with 3,980 articles. Then, 2,000 words with the highest average TF-IDF values are chosen to represent the articles.

The New York Times dataset includes 300,000 documents, among which we randomly pick 290,000 documents for training, 1,000 for validation and 9,000 for testing. Then, 10,000 words with the highest average TF-IDF values are chosen to represent the documents.

Training

We divide the training data into mini-batches for training. The batch sizes of dataset NIPS, CiteULike and New York Times are 10, 100 and 1,000 respectively. Model parameters are updated after each mini-batch. We set the maximum number of training epochs to 50. And we train all the models using the CD algorithm with $T = 10$ full Gibbs steps.

As for RBM-based Replicated Softmax, we determine the optimal number of hidden units over the validation data with 10 units as the step size. While for SBMs, we firstly train a two-layer HLTM and then increase the number of connections such that every hidden unit is connected to 20% of the visible units that are most correlated. A mask matrix is applied to the connection matrix after each parameter update so as to force the sparse connectivity. The numbers of hidden units automatically determined by our algorithm are 112, 194 and 326 for dataset NIPS, CiteULike and New York Times respectively.

Evaluations

The log-probability on held-out data is used to gauge the generalization performance of different models. As computing these values exactly is intractable, *Annealed Importance Sampling (AIS)* (Neal 2001; Salakhutdinov and Murray 2008) was used in (Hinton and Salakhutdinov 2009) to estimate the partition function of Replicated Softmax. We extend AIS to SBMs in our experiments. In AIS, we use 500 "inverse temperatures" β_k spaced uniformly from 0 to 0.5, 3,000 β_k spaced uniformly from 0.5 to 0.9, and 6,500 β_k spaced uniformly from 0.9 to 1.0, with a total of 10,000 intermediate distributions. The estimates are averaged over 100 AIS runs for each held-out document. Then we calculate the average per-word perplexity as $\exp(-\frac{1}{N}\sum_{n=1}^{N}\frac{1}{D_n}logP(\mathcal{U}_n))$. A smaller score indicates better generalization performance. Due to the high computation cost, we follow the experiments in (Hinton and Salakhutdinov 2009) and randomly sample 50 documents from the validation data to calculate the score. While for testing, we use all of the 50 test documents in the NIPS dataset, and randomly sample 500 documents from test data in CiteULike and New York Times datasets.

Results

Overfitting of Fully-Connected RBMs We first empirically show that, the fully-connected structure in Replicated Softmax can easily lead to overfitting once the number of hidden units (and hence the number of parameters) gets too large. Figure 5 depicts the average perplexity scores over validation data for Replicated Softmax with different number of hidden units after 30 epochs. We can see that the optimal numbers of hidden units for the three datasets are 110, 60 and 120 respectively. After that, the performances of the models worsen when the numbers of hidden units gradually increase. Therefore, selecting a proper number of hidden units is crucial to Replicated Softmax since the model is very likely to overfit the training data.

Generalizability of SBMs and Replicated Softmax In this part, we compare the generalization performance of SBMs with Replicated Softmax. We denote our method as *SBM-SFC*. Two variants of Replicated Softmax included in comparison are *RS** and *RS*$^+$. *RS** trains Replicated Softmax with the optimal number of hidden units. *RS*$^+$ produces Replicated Softmax with the same number of hidden units as *SBM-SFC*. Since this number is normally larger than the optimal number, we denote the method as *RS*$^+$. As we can see in Table 1, *SBM-SFC* consistently outperforms *RS** and *RS*$^+$ over the three datasets. This confirms that Replicated Softmax with full connectivity is prone to overfitting. It also shows that SBMs can lead to better model fit than fully connected RBMs. This is true even when the number of hidden units in RBMs is optimized through held-out validation. Moreover, the poor performance of *RS*$^+$ shows that the performance gain of *SBM-SFC* cannot be attributed to the larger number of hidden units.

Comparisons with Redundancy Pruning We also compare our method with the redundancy pruning method which

Table 1: Average per-word perplexity achieved by different methods on different datasets.

	NIPS		CiteULike		New York Times	
	Validation	Test	Validation	Test	Validation	Test
RS*	518	547	591	636	1,865	1,809
RS$^+$	505	538	795	913	2,129	1,985
RS$^+$ SFC	532	551	632	668	2,021	1,910
RS$^+$ Pruned	542	565	**534**	**584**	1,697	1,608
SBM-SFC	**476**	**488**	545	597	**1,624**	**1,583**

produces Replicated Softmax with sparse connections (Han et al. 2015). We denote the method as *RS$^+$ Pruned*. It starts from a fully trained model, produced by *RS$^+$*, and prunes the connections gradually until the number of connections is reduced to be the same as the model by *SBM-SFC*. For each hidden unit, it prunes the set of connections with the smallest absolute weight value. Then it retrains the pruned model for 1 epoch, and conducts pruning again. The pruning and retraining process is repeated until the desired sparsity is reached. In our experiments, the pruning process took 80, 40 and 40 epochs on the three datasets respectively. As shown in Table 1, *SBM-SFC* achieves comparable model fit as *RS$^+$ Pruned*. It shows that our structure learning algorithm is effective and can ease the overfitting problem of fully connected structure as well as the pruning method does. Our method has three advantages over *RS$^+$ Pruned*. First, the iterative pruning process of *RS$^+$ Pruned* is computationally expensive. Second, it does not offer a way to determine the number of hidden units. One can do this using held-out validation, but that would be computationally prohibitive. Third, as will be seen later, the models produced by *RS$^+$ Pruned* are not as interpretable as those obtained by our method.

Necessity of Hidden Connections In SBMs, we impose a tree structure among the hidden units. Is this necessary? To answer the question, we compare *SBM-SFC* with a method for Replicated Softmax denoted as *RS$^+$ SFC*. The model produced by *RS$^+$ SFC* is the same as that by *SBM-SFC*, except that there are no connections among the hidden units. As we can see in Table 1, *SBM-SFC* always performs better than *RS$^+$ SFC*. This supports our conjecture that the hidden connections are necessary in our models. The result is not surprising. In a multiple layer model, units at a layer are connected via units at higher layers. In a two layer model, there are no higher layers. Hence it is natural to connect the second-layer units directly. To generalize our work to multiple layers, we will need to add connections only among the hidden units at the top layer.

Interpretability of SBMs and Replicated Softmax Next we compare the interpretability of SBMs and Replicated Softmax. Here is how we interpret hidden units. For each hidden unit, we sort the words in descending order of the absolute value of the connection weights between the words and the hidden unit. The top 10 words with the highest absolute weights are chosen to characterize the hidden unit. We propose to measure the "interpretability" of a hidden unit by considering how similar pairs of words in the top-10 list are. The similarity between two words is determined using a word2vec model (Mikolov et al. 2013a; 2013b) trained on part of the Google News datasets [4], where each word is mapped to a high dimensional vector. The similarity between two words is defined as the cosine similarity of the two corresponding vectors. High similarity suggests that the two words appear in similar contexts. Let $\mathcal{L}$ be the list of words representing a hidden unit. We define the *compactness* of $\mathcal{L}$ to be the average similarity between pairs of words in $\mathcal{L}$. We also call it the *interpretability score* of the hidden unit. Note that some of the words in $\mathcal{L}$ might not be in the vocabulary of the word2vec model we use. This happens infrequently, and when it does, the words are simply skipped. Suppose there are F hidden units in a model. Let $C_1, ...C_F$ be the interpretability scores of hidden units. We define the *interpretability score* of the model as: $Q = \frac{1}{F}\sum_{j=1}^{F} C_j$. Obviously the score depends heavily on the number of hidden units.

Table 2: Interpretability scores of models: The three models included have the same number of hidden units.

	NIPS	CiteULike	NYTimes
RS$^+$	0.1102	0.1499	0.1407
RS$^+$ Pruned	0.1006	0.1449	0.1420
SBM-SFC	**0.1235**	**0.1725**	**0.1433**

Table 3: Characterizations of selected hidden units in models produced by *SBM-SFC*. Only top 5 words are listed.

NIPS	spike neuron pruning weight rules
	pixel pca image pixels images
	markov likelihood conditional posterior probabilities
CiteULike	models model modeling causal modelling
	ancestral species selection duplication evolution
	network networks connected topology connectivity
NYTimes	china beijing south_africa mexican chinese
	george_bush laura_bush bill_clinton tournament jew
	gene patient doctor medical physician

Table 2 reports the interpretability scores of the models produced by *RS$^+$*, *RS$^+$ Pruned* and *SBM-SFC*. The models all have the same number of hidden units and hence their in-

[4] https://code.google.com/archive/p/word2vec/

terpertability scores are comparable. *SBM-SFC* consistently performs the best over the three datasets, showing superior coherency and compactness in the characterizations of the hidden units and thus better model interpretability. Table 3 shows the characterizations of selected hidden units in models produced by *SBM-SFC*. They are clearly meaningful.

Conclusions

Overfitting in deep models is caused not only by excessive amount of hidden units, but also excessive amount of connections. In this paper we have developed, for models with a single hidden layer, a method to determine the number of hidden units and the connections among the units. The models obtained by the method are significantly better, in terms of held-out likelihood, than RBMs where the hidden and observed units are fully connected. This is true even when the number of hidden units in RBMs is optimized by held-out validation. In comparison with redundancy pruning, our method is more efficient and is able to determine the number of hidden units. Moreover, it produces more interpretable models. In the future, we will generalize the structure learning method to models with multiple hidden layers.

Acknowledgments

Research on this article was supported by Hong Kong Research Grants Council under grants 16202515 and 16212516.

References

Adams, R. P.; Wallach, H. M.; and Ghahramani, Z. 2010. Learning the structure of deep sparse graphical models. In *AISTATS*, 1–8.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chen, P.; Zhang, N. L.; Poon, L. K. M.; and Chen, Z. 2016. Progressive EM for latent tree models and hierarchical topic detection. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 1498–1504.

Cun, Y. L.; Denker, J. S.; and Solla, S. A. 1990. Optimal brain damage. In *Advances in Neural Information Processing Systems*, 598–605.

Han, S.; Pool, J.; Tran, J.; and Dally, W. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc. 1135–1143.

Hassibi, B.; Stork, D. G.; and Com, S. C. R. 1993. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5*, 164–171.

Hinton, G. E., and Salakhutdinov, R. R. 2009. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems 22*. 1607–1614.

Hinton, G. E.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29(6):82–97.

Hinton, G. E.; Osindero, S.; and Teh, Y.-W. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18(7):1527–1554.

Hinton, G. E. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation* 14(8):1771–1800.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105.

Larochelle, H.; Bengio, Y.; and Turian, J. 2010. Tractable multivariate binary density estimation and the restricted boltzmann forest. *Neural computation* 22(9):2285–2307.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 2278–2324.

Lee, H.; Ekanadham, C.; and Ng, A. Y. 2008. Sparse deep belief net model for visual area v2. In *Advances in Neural Information Processing Systems*, 873–880.

Liu, T.; Zhang, N. L.; and Chen, P. 2014. Hierarchical latent tree analysis for topic detection. In *Machine Learning and Knowledge Discovery in Databases 2014*, 256–272.

Mikolov, T.; Deoras, A.; Povey, D.; Burget, L.; and Černockỳ, J. 2011. Strategies for training large scale neural network language models. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 196–201.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations Workshops*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, 3111–3119.

Murphy, K. P. 2012. *Machine learning: a probabilistic perspective*.

Neal, R. M. 2001. Annealed importance sampling. *Statistics and Computing* 11(2):125–139.

Salakhutdinov, R., and Murray, I. 2008. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th International Conference on Machine Learning*, 872–879.

Saul, L., and Jordan, M. I. 1994. Learning in boltzmann trees. *Neural Computation* 6(6):1174–1184.

Smolensky, P. 1986. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory, 194–281.

Srinivas, S., and Babu, R. V. 2015. Data-free parameter pruning for deep neural networks. In *Proceedings of the British Machine Vision Conference*, 31.1–31.12.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 3104–3112.

Wan, L.; Zeiler, M.; Zhang, S.; Cun, Y. L.; and Fergus, R. 2013. Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning*, 1058–1066.