

# Learning Parts-based Representations with Nonnegative Restricted Boltzmann Machine

Tu Dinh Nguyen<sup>†</sup>

NGTU@DEAKIN.EDU.AU

Truyen Tran<sup>†‡</sup>

TRUYEN.TRAN@DEAKIN.EDU.AU

Dinh Phung<sup>†</sup>

DINH.PHUNG@DEAKIN.EDU.AU

Svetha Venkatesh<sup>†</sup>

SVETHA.VENKATESH@DEAKIN.EDU.AU

<sup>†</sup>*Center for Pattern Recognition and Data Analytics*

*School of Information Technology, Deakin University, Geelong, Australia.*

<sup>‡</sup>*Institute for Multi-Sensor Processing and Content Analysis*

*Curtin University, Australia*

**Editor:** Cheng Soon Ong and Tu Bao Ho

## Abstract

The success of any machine learning system depends critically on effective representations of data. In many cases, especially those in vision, it is desirable that a representation scheme uncovers the parts-based, additive nature of the data. Of current representation learning schemes, restricted Boltzmann machines (RBMs) have proved to be highly effective in unsupervised settings. However, when it comes to parts-based discovery, RBMs do not usually produce satisfactory results. We enhance such capacity of RBMs by introducing nonnegativity into the model weights, resulting in a variant called *nonnegative restricted Boltzmann machine* (NRBM). The NRBM produces not only controllable decomposition of data into interpretable parts but also offers a way to estimate the intrinsic nonlinear dimensionality of data. We demonstrate the capacity of our model on well-known datasets of handwritten digits, faces and documents. The decomposition quality on images is comparable with or better than what produced by the nonnegative matrix factorisation (NMF), and the thematic features uncovered from text are qualitatively interpretable in a similar manner to that of the latent Dirichlet allocation (LDA). However, the learnt features, when used for classification, are more discriminative than those discovered by both NMF and LDA and comparable with those by RBM.

**Keywords:** parts-based representation, nonnegative, restricted Boltzmann machines, learning representation, semantic features

## 1. Introduction

Learning meaningful representations from data is often critical<sup>1</sup> to achieve high performance in machine learning tasks (Bengio et al., 2012). An attractive approach is to estimate representations that explain the data best without the need of labels. One important class of such methods is the restricted Boltzmann machine (RBM) (Smolensky, 1986; Freund

---

1. The importance of the topic is best explained by the event of the first conference on learning representation (ICLR'13), following a series of workshops in NIPS/ICML.

and Haussler, 1994), an undirected probabilistic bipartite model in which a representational hidden layer is connected with a visible data layer. The weights associated with connections encode the strength of influence between hidden and visible units. Each unit in the hidden layer acts as a binary feature detector, and together, all the hidden units form a *distributed representation* of data (Hinton and Ghahramani, 1997; Bengio et al., 2012). This distributed representation is highly compact: for  $K$  units, there are  $2^K - 1$  non-empty configurations that can explain the data.

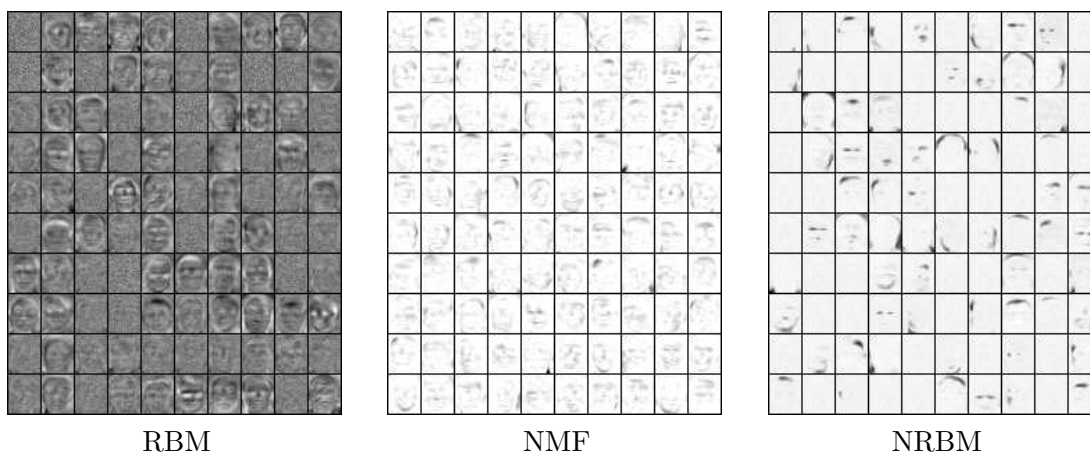


Figure 1: Representations learnt from the ORL face image database (AT&T at Cambridge) using the ordinary RBM, NMF and NRBM on the left, middle, right, respectively. Darker pixels show larger weights.

However, fully distributed representation learnt by RBMs may not interpretably disentangle the factors of variation since the learnt features are often global, that is, all the data units must play the role in one particular feature. As a result, learnt features do not generally represent parts and components (Teh and Hinton, 2001), a property often seen as desirable in real-life applications (Agarwal et al., 2004). For example the facial features learnt by RBM depicted in Fig. 1(left) are generally global and it is hard to explain how a face is constructed from these parts. One of the best known techniques to achieve parts-based representation is nonnegative matrix factorisation (NMF) (Lee and Seung, 1999). In NMF, the data matrix is approximately factorised into a basis matrix and a coding matrix, where all the matrices are assumed to be nonnegative. Each column of the basis matrix is a learnt feature, which could be sparse under appropriate regularisation (Hoyer, 2004). The NMF, however, has a fundamental drawback: it does not generalise to unseen data since there is no mechanism by which a new data point can be generated from the learnt model. Instead, new representations must be learnt from the expensive “fold-in” procedure. The RBM, on the other hand, is a proper generative model – once the model has been learnt, new samples can be drawn from the model distribution. Moreover, due to the special bipartite structure, estimating representation from data is efficient with a single matrix operation.

In this paper, we demonstrate derivation of useful parts-based representations whilst retaining the discriminative capacity of the RBM. As inspired by the NMF, we propose to

enforce nonnegativity in the connection weight matrix of the RBM. Our method integrates a barrier function into the objective function so that the learning is skewed toward nonnegative weights. As the contribution of the visible units towards a hidden unit is additive, there exists competition among visible units to activate the hidden unit leading to a small portion of connections survived. Using the same facial example, the method could achieve parts-based representation of faces (eyes, nose, mouth, forehead), which is, surprisingly, even better than what learnt by the standard NMF (see Fig. 1(right)). We term the resulting model the *Nonnegative Restricted Boltzmann Machine* (NRBM).

In addition to parts-based representation, there are several benefits with this nonnegativity constraint. First, in many cases, it is often easier to make sense of addition of new latent factors (due to nonnegative weights) than of subtraction (due to negative weights). For instance, clinicians may be more comfortable with the notion that a risk factor either contributes positively to a disease development or not at all (e.g., the connections have zeros weights). Second, as weights can be either positive or zero, the parameter space is highly constrained leading to potential robustness. This can be helpful when there are many more hidden units than those required to represent all factors of variation: extra hidden units will automatically be declared “dead” if all connections to them cannot compete against others in explaining the data.

We demonstrate the effectiveness of the proposed model through comprehensive evaluation on four real datasets of very different natures. Three image datasets are the MNIST for handwritten digits (Lecun et al.), CBCL (CBCL at MIT) and ORL (AT&T at Cambridge) for human faces. Our primary target is to decompose images into interpretable parts (and receptive fields), e.g., dots and strokes in handwritten digits, and facial components in faces. The last dataset is in text, which is extracted from the TDT2 corpus (Cai et al., 2005). The goal is to discover plausible latent thematic features, which are groups of semantically related words. For the MNIST and TDT2 datasets, the learnt features are then fed into standard classifiers. The experiments reveal that the classification performance is comparable with the standard RBM, and competitive against NMF (on both images and text) and latent Dirichlet allocation (on text) (Blei et al., 2003).

Our main contributions are: (i) the derivation of the Nonnegative Restricted Boltzmann Machine, a probabilistic machinery that has the capacity of learning parts-based representations; and (ii) a comprehensive evaluation the capability of our method as a representational learning tool on image and text data, both qualitatively and quantitatively.

The rest of paper is structured as follows. Section 2 presents the derivation and properties of our nonnegative RBM. We then report our experimental results on visual receptive fields learning and textual semantic features discovering in Section 3. Section 4 provides in-depth discussion of the work and related literature. Finally, Section 5 concludes the paper.

## 2. Nonnegative Restricted Boltzmann Machines

### 2.1. Representing data using RBM

Our model is based on Restricted Boltzmann Machines (RBM’s) (Smolensky, 1986; Freund and Haussler, 1994; Hinton, 2002). A RBM is a bipartite undirected graphical

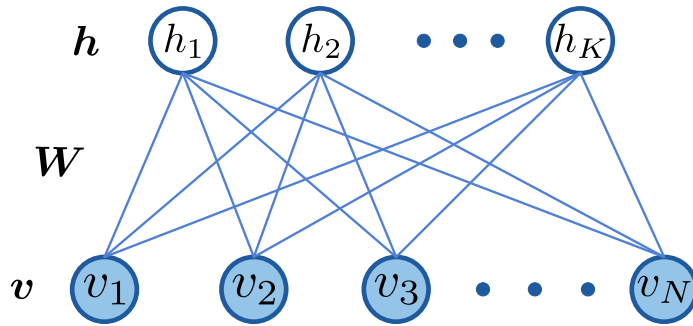


Figure 2: Graphical illustration of a RBM that models the joint distribution of  $N$  visible units and  $K$  hidden units. The connections are undirected and the shaded nodes are observed.

model in which the bottom layer contains observed variables called visible units and the top layer consists of latent *representational variables*, known as hidden units. A graphical illustration of RBM is presented in Fig. 2.

More formally, let  $\mathbf{v}$  denote the set of visible variables:  $\mathbf{v} = (v_1, v_2, \dots, v_N) \in \{0, 1\}^N$  and  $\mathbf{h}$  indicate the set of hidden ones:  $\mathbf{h} = (h_1, h_2, \dots, h_K) \in \{0, 1\}^K$ . The RBM defines an energy function of a joint configuration  $(\mathbf{v}, \mathbf{h})$  as:

$$E(\mathbf{v}, \mathbf{h}; \psi) = -\left(\mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h} + \mathbf{v}^\top \mathbf{W} \mathbf{h}\right) \quad (1)$$

where  $\psi = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$  is the set of parameters.  $\mathbf{a}, \mathbf{b}$  are the biases of hidden and visible units respectively, and  $\mathbf{W}$  represents the weights connecting the hidden and visible units. The model assigns a probability to the joint configuration as:

$$p(\mathbf{v}, \mathbf{h}; \psi) = \frac{1}{\mathcal{Z}(\psi)} e^{-E(\mathbf{v}, \mathbf{h}; \psi)} \quad (2)$$

where  $\mathcal{Z}(\psi)$  is the normalisation constant.

Since the network has no intra-layer connections, the conditional distributions over visible and hidden units are factorised as:

$$p(\mathbf{v} | \mathbf{h}; \psi) = \prod_{n=1}^N p(v_n | \mathbf{h}); \quad p(\mathbf{h} | \mathbf{v}; \psi) = \prod_{k=1}^K p(h_k | \mathbf{v}) \quad (3)$$

These factorisations allow fast layer-wise sampling, a property which proves crucial for MCMC-based model estimation. Once the model is fully specified, the new representation of an input data can be achieved by computing the posterior vector  $\hat{\mathbf{h}} = (\hat{h}_1, \hat{h}_2, \dots, \hat{h}_K)$ , where  $\hat{h}_k$  is shorthand for  $\hat{h}_k = p(h_k = 1 | \mathbf{v}) = \sigma(b_k + \sum_n W_{nk} v_n)$ , where  $\sigma(x)$  is the sigmoid function  $\sigma(x) = [1 + e^{-x}]^{-1}$ .

## 2.2. Deriving parts-based representation

Parts-based representations imply that column vectors of the connection weight matrix  $\mathbf{W}_{\bullet k}$  must be sparse, e.g., only a small portion of entries is non-zeros. Recall that the activation of a hidden unit is based on the probability  $\hat{h}_k = \sigma(b_k + \sum_n W_{nk} v_n)$ . The connection weight  $W_{nk}$  is the association strength between an visible unit  $n$  and an hidden unit  $k$ , that is the tendency of the two units being co-active when  $W_{nk} > 0$ . Due to asymmetric parameter initialisation, the learning process tends to increase some associations more than the others. Under nonnegative weights, i.e.,  $W_{nk} \geq 0$ , one can expect that for a given activation probability  $\hat{h}_k$  and bias  $b_k$ , such increase must cause other associations to degrade, since  $v_n \geq 0$ . As the lower bound of weights are now zeros, there is a natural tendency for many weights to be driven to zeros as learning progresses.

Recall that learning in the standard RBM is usually based on maximising the data likelihood with respect to the parameter  $\psi$  as:

$$\mathcal{L}(\mathbf{v}; \psi) = p(\mathbf{v}; \psi) = \sum_{\mathbf{h}} p(\mathbf{v}, \mathbf{h}; \psi) \quad (4)$$

To encourage nonnegativity in  $\mathbf{W}$ , we propose to use quadratic barrier function (Nocedal and Wright, 2000, pp. 497–506), that is, the objective function is now the following regularised likelihood

$$\mathcal{L}_{reg} = \log \mathcal{L}(\mathbf{v}; \psi) - \frac{\alpha}{2} \sum_{n=1}^N \sum_{k=1}^K f(W_{nk}) \quad (5)$$

where

$$f(x) = \begin{cases} x^2 & x < 0 \\ 0 & x \geq 0 \end{cases}$$

and  $\alpha \geq 0$ . Maximising this regularised likelihood is equivalent to simultaneously maximising the data likelihood  $\mathcal{L}(\mathbf{v}; \psi)$  and minimising the penalty of sum squared of negative weights. The level of nonnegativity is controlled by the choice of  $\alpha$ . When we set  $\alpha$  to zero, it means that there is no barrier function. The model turns back to the ordinary RBM without nonnegative constraint. The larger  $\alpha$  is, the tighter the barrier restriction becomes.

Finally the parameter update rule reads

$$W_{nk} \leftarrow W_{nk} + \eta (\langle v_n h_k \rangle_{\tilde{P}} - \langle v_n h_k \rangle_P - \alpha [W_{nk}]^-) \quad (6)$$

where  $\eta > 0$  is the learning rate,  $\tilde{P}(\mathbf{v}, \mathbf{h}; \psi) = \tilde{P}(\mathbf{v}) P(\mathbf{h} | \mathbf{v}; \psi)$  is the data distribution with  $\tilde{P}(\mathbf{v})$  representing the empirical distribution,  $P(\mathbf{v}, \mathbf{h}; \psi)$  is the model distribution defined in Eq. (2) and  $[W_{nk}]^-$  denotes the negative part of the weight. Following the learning of standard RBMs (Hinton, 2002),  $\tilde{P}(\mathbf{v})$  is computed from data observations and  $P$  can be efficiently approximated by running multiple Markov short chains starting from observed data at each update.

### 2.3. Estimating the intrinsic dimensionality

An issue in the RBM is that there are currently no easy methods to determine the appropriate number of hidden units needed for a particular problem. This is unlike PCA where the amount of variance explained by the principal components can be used to determine the number of components. The nonnegativity constraint in the NRBM, interestingly, leads to a similar capacity by examining the “dead” hidden units.

To see how, recall that the probability of a visible unit being active is given as:

$$p(v_n = 1 | \mathbf{h}) = \sigma \left( a_n + \sum_k W_{nk} h_k \right) \quad (7)$$

Since this probability, in general, is constrained by the data variations, the hidden units must “compete” with each other to explain the data. This is because the contribution towards the explanation is nonnegative, an increase on power of one unit must be at the cost of others. If  $K^* < K$  hidden units are intrinsically enough to account for all variations in the data, then one can expect that either the other  $K - K^*$  hidden units are always deactivated (e.g., with very large negative biases) or their connection weights are almost zeros since  $W_{nk} \geq 0$ . In either cases, the hidden units become permanently inoperative in data generation. Thus by examining the dead units, we may be able to uncover the intrinsic dimensionality of the data variations.

## 3. Experiments

In this section, we evaluate the capacity of the Nonnegative RBM on (i) unsupervised decomposing images into parts and discovering semantic features from texts, and (ii) discovering discriminative features that will be useful for supervised classification. Three image popular datasets were used, one for handwritten digits (MNIST, (Lecun et al.)), and two for faces (CBCL, (CBCL at MIT) and ORL, (AT&T at Cambridge)). The last dataset is in text where we take 30 categories subset of the TDT2 corpus<sup>2</sup>.

The MNIST dataset consists of 60,000 training and 10,000 testing  $28 \times 28$  images, each of which contains a handwritten digit. The CBCL database contains facial and non-facial well-aligned images, of which only 2,429 facial images in the training set would be used in our experiments. The ORL data consists of 40 subjects with ten  $92 \times 112$  faces each on different variations of illumination and facial expressions and details. These images are all in the grayscale, which is then normalised into the range  $[0, 1]$ . Since the image pixels are not exactly binary data, following the previous work (Hinton and Salakhutdinov, 2006), we treat the normalised intensity as empirical probabilities on which the NRBM is naturally applied. As the empirical expectation  $\langle v_n h_k \rangle_{\tilde{p}}$  in Eq. (6) requires the probability  $\langle v_n \rangle$ , the normalised intensity is a good approximation.

The TDT2 corpus was collected from news sources, (newswires: APW, NYT; radio programs: VOA, PRI; and television programs: CNN, ABC). It contains 11,201 on-topic documents arranged into 96 semantic categories. Following the preprocessing in (Cai et al., 2005), all multiple category documents are removed and the largest 30 categories are kept.

2. NIST Topic Detection and Tracking corpus is at <http://www.nist.gov/itl/>.

This retains 9,394 documents and 36,771 unique words in total. We further the preprocessing of data by removing common stopwords. Only 1,000 most frequent words are then kept and one blank document are removed. For NRBMs, word presence is used rather than their counts.

To speed up the training phase, we divide training samples into “mini-batches” of  $B = 100$  samples. Hidden, visible and visible-hidden learning rates are fixed to 0.1. Visible biases are initialised so that the marginal distribution, when there are no hidden units, matches the empirical distribution. Hidden biases are first set to some reasonable negative values to offset the positive activating contribution from the visible units. Mapping parameters are randomly drawn from positive values in  $[0, 0.01]$ . Parameters are then updated after every mini-batch. Learning is terminated after 100 epochs. The regularisation hyperparameter  $\alpha$  (Eq. 5) is empirically tuned so that the data decomposition is both meaningful (e.g., by examining visually, or by computing the parts similarity) and accurate (e.g., by examining the reconstruction quality).

### 3.1. Decomposing images into parts-based representations

We now show that the nonnegative constraints enable the NRBMs to produce meaningful parts-based receptive fields. Fig. 3 (left and right) depict the 100 filters learnt from the MNIST images. It can be seen that basic structures of handwritten digits such as strokes and dots are discovered by NRBMs, but it is much less clear in standard RBMs.

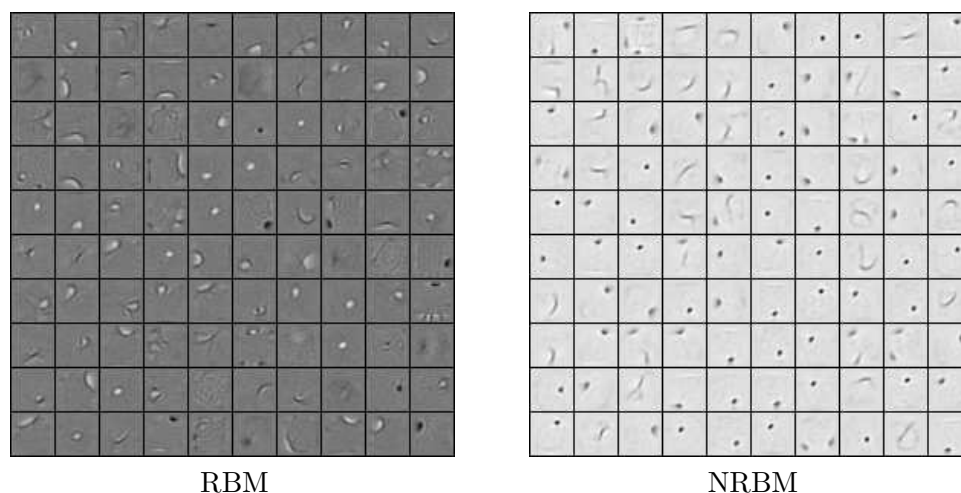


Figure 3: Receptive fields learnt from the MNIST handwritten digits database using the RBM on the left and NRBMs on the right. Darker pixels illustrate larger weights. Both RBMs and NRBMs produce stroke-based features. However, the features that NRBMs learn are simpler whilst the ones learnt by RBMs are more difficult to interpret.

For the CBCL dataset, the facial parts (eyes, mouth, nose, eyebrows etc.) uncovered by NRBMs (Fig. 4 (right)) are visually interpretable along the line with classical NMF (Lee

and Seung, 1999) (Fig. 4 (middle)). The RBM, on the other hand, produces global facial structures (Fig. 4 (left)). On the more challenging facial set with higher variation such as the ORL (see Section 1), NMF fails to produce parts-based representation (Fig. 1(middle)), and this is consistent with previous findings (Hoyer, 2004). In contrast, the NRBM is still able to learn facial components (Fig. 1(right)).

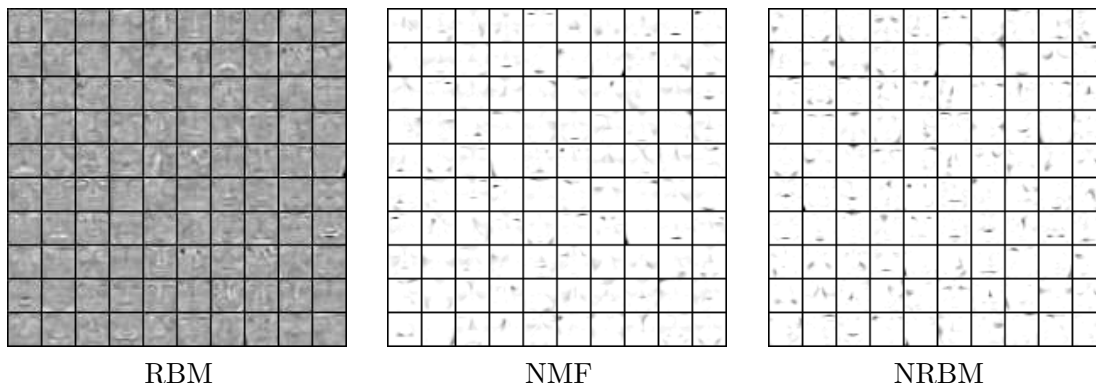


Figure 4: Receptive fields learnt from the CBCL face image database using RBM, NMF and NRBM on the left, middle and right. The NRBM and NMF yield parts-based representations of faces whereas the RBM produces holistic representations. Darker pixels show larger weights.

The capacity to decompose data in NRBM is controlled by a single hyperparameter  $\alpha$ . As shown in Fig. 5, there is a smooth transition from the holistic decomposition as in standard RBM (when  $\alpha$  is near zero) to truly parts-based representations (when  $\alpha$  is larger).

### 3.2. Dead factors and dimensionality estimation

We now examine the ability of NRBM to estimate the intrinsic dimensionality of the data, as discussed in Section 2.3. We note that by “dimensionality” we roughly mean the degree of variations, not strictly the dimension of the data manifold. This is because our latent factors are discrete binary variables, and thus they may be less flexible than real-valued coefficients.

For that purpose, we compute the number of dead or unused hidden units. The hidden unit  $k$  is declared “dead” if the normalised  $\ell_1$ -norm of its connection weight vector is lower than a threshold  $\tau$ :  $|\mathbf{W}_{\bullet k}|_1 N^{-1} \leq \tau$ , where  $N$  is the dimension of the original data. We also examine the hidden biases which, however, do not cause dead units in this case. In Fig. 6, the number of used hidden units is plotted against the total number of hidden units  $K$  by taking the average over a set of thresholds ( $\tau \in \{0.01; 0.02; \dots; 0.06\}$ ). With the NRBM, the number of hidden units which explicitly represents the data saturates at about 150 whilst all units are used by the RBM.



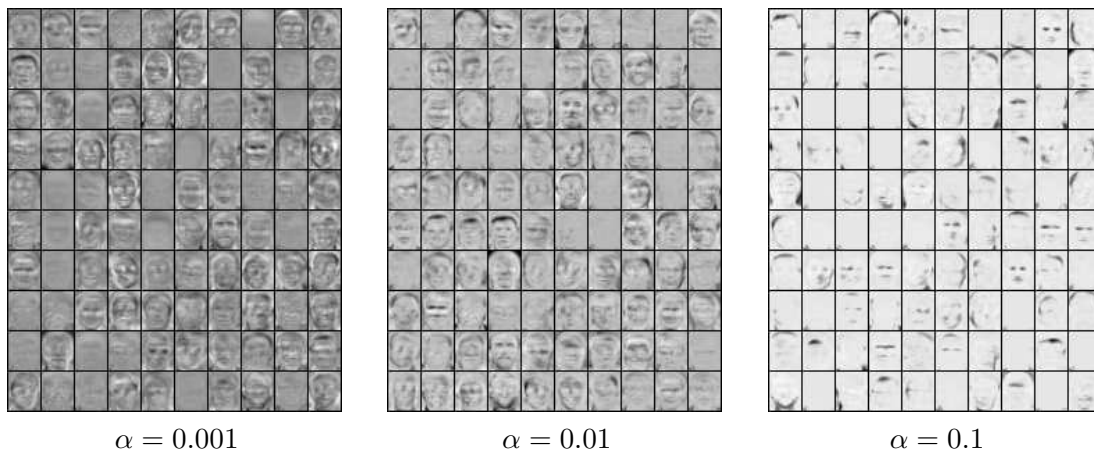


Figure 5: Receptive fields learnt from the ORL face image database using NRBM with varied barrier costs. The barrier cost  $\alpha$  is tightened up from left to right. Darker pixel indicate larger weights.

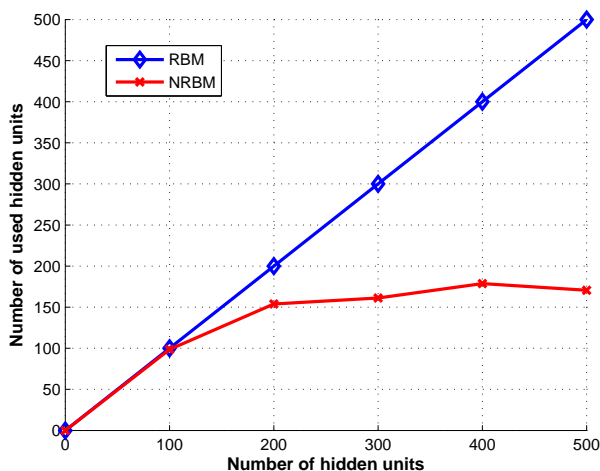


Figure 6: The number of used hidden units with different hidden layer sizes of RBM-based models.

### 3.3. Semantic features discovering on text data

The next experiment investigates the applicability of the NRBM on decomposing text into meaningful “parts”, although this notion does not convey the same meaning as those in vision. This is because the nature of text may be more complex and high-level, and it is hard to know whether the true nature of word usage is additive. Following literature in topic modelling (e.g., see (Blei et al., 2003)), we start from the assumption that there are latent themes that govern the choice of words in a particular document. Our goal is to

examine whether we can uncover such themes for each document, and whether the themes are corresponding to semantically coherent subset of words.

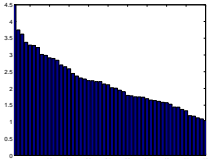
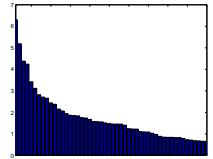
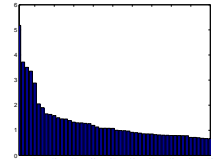
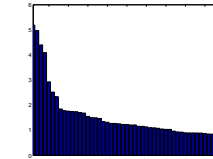
| <b>Asian Economic Crisis</b>  | <b>Current Conflict with Iraq</b>  | <b>1998 Winter Olympics</b>  | <b>India - A Nuclear Power?</b>  |
|---|--|--|--|
| FINANCIAL<br>FUND<br>MONETARY<br>INVESTMENT<br>FINANCE<br>WORKERS<br>INVESTORS<br>DEBT<br>TREASURY<br>CURRENCY<br>RATES<br>TOKYO<br>MARKETS<br>IMF<br>ASIAN | JURY<br>GRAND<br>IRAQ<br>SEVEN<br>IRAQI<br>GULF<br>BAGHDAD<br>SADDAM<br>PERSIAN<br>HUSSEIN<br>KUWAIT<br>IRAQS<br>INSPECTOR<br>STANDOFF<br>BIOLOGICAL | BUTLER<br>RICHARD<br>NAGANO<br>INSPECTOR<br>CHIEF<br>OLYMPICS<br>RISING<br>GAMES<br>COMMITTEE<br>WINTER<br>OLYMPIC<br>CHAIRMAN<br>JAPANESE<br>EXECUTIVE<br>JAKARTA | COURT<br>BAN<br>TESTS<br>INDIAS<br>TESTING<br>INDIA<br>SANCTIONS<br>ARKANSAS<br>RULED<br>INDIAN<br>PAKISTAN<br>NUCLEAR<br>JUDGE<br>LAW<br>ARMS |
|    |   |    |   |

Table 1: An example of 4 most distinguished categories, i.e., economics, politics, sport and armed conflict associated with top 15 words (ranked by their weights) discovered from the TDT2 subset. The charts at the bottom row illustrate the weight impact of words on the category. These weights are sorted in descending order.

Using the TDT2 corpus, we learn the NRBMs from the data and examine the mapping weight matrix  $\mathbf{W}$ . For each latent factor  $k$ , the entry to column  $\mathbf{W}_{\bullet k}$  reflects the association strength of a particular word with the factor, where zero entry means distant relation. Table 1 presents four noticeable semantic features discovered by our model. The top row lists the top 15 words per feature, and the bottom row plots the distribution of association strengths in decreasing order. It appears that the words under each feature are semantically related in a coherent way.

The next aspect is whether the NRBMs effectively disentangle the factors, that is, the factors should be least correlated. To assess the correlation between factors, we compute the pairwise cosine similarities between the weight columns  $S(k, k') = \text{cosine}(\mathbf{W}_{\bullet k}, \mathbf{W}_{\bullet k'})$  for  $k \neq k'$ . The correlation between a factor and the rest is

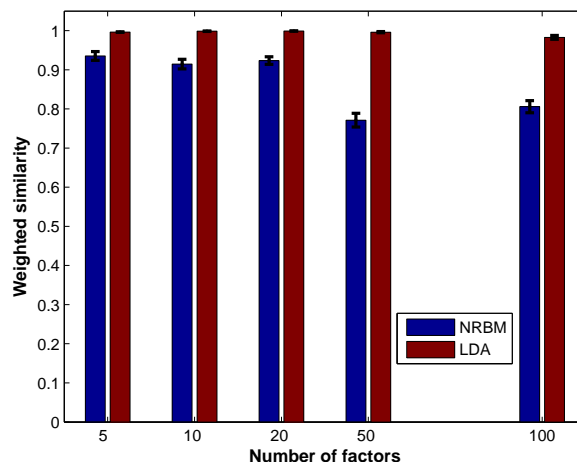


Figure 7: The means and standard deviations of cosine similarities among weight vectors of the NRBM and LDA with varying numbers of factors. This score evaluates the similarity of latent factors learnt using the model. The lower score indicates the better factor discovering.

$$S^*(k) = \frac{1}{K-1} \sum_{k' \neq k} S(k, k')$$

Fig. 7 shows the mean and standard deviation of the correlations for both NRBM and LDA of the same latent sizes<sup>3</sup>, where the “weights” for the LDA are the word generation probability matrix. The result is favourable toward NRBM (with smaller average correlation), that is, it has higher power of disentangling factors of variation.

### 3.4. Classification performance on learnt representations

Our next target is to evaluate whether the ability to decompose data into parts and to disentangle factors of variation could straightforwardly translate into better predictive performance. Although the NRBM can be easily turned into a nonnegative neural network and the weights are tuned further to best fit the supervised setting (e.g., see (Hinton and Salakhutdinov, 2006)), we choose not to do so because our goal is to see if the decomposition separates data well enough. Instead we apply standard classifiers on the learnt features, or more precisely the hidden posteriors.

The first experiment is with the MNIST, the 500 factors have been previously learnt in Section 3.1 and Fig. 3. Support Vector Machines (SVM, with Gaussian kernels, using the LIBSVM package (Chang and Lin, 2011)) and  $k$ -nearest neighbours ( $k$ NN, where  $k = 4$ , with cosine similarity measures) are used as classifiers. For comparison, we also apply the

3. Strictly speaking, the latent spaces of NRBM and LDA are very different: one is the discrete, while the other is the positive simplex.

same setting to the features discovered by the NMF. The error rate on test data is reported in Table 2. It can be seen that (i) compared to standard RBM, the nonnegativity constraint used in NRBM does not lead to degrade of predictive performance, suggesting that the parts are also indicative of classes; and (ii) nonlinear decomposition in NRBM can lead to better data separation than the linear counterpart in NMF.

|      | SVM         | 4-NN        |
|------|-------------|-------------|
| RBM  | <b>1.38</b> | 2.74        |
| NMF  | 3.25        | 2.64        |
| NRBM | 1.4         | <b>2.34</b> |

Table 2: The classification errors (%) on testing data of MNIST dataset.

The second experiment is on the text data TDT2. Unlike images, words are already conceptual and thus using standard bag-of-words representation is often sufficient for many classification tasks. The question is therefore whether the thematic features discovered by the NRBM could further improve the performance, since it has been a difficult task for topic models such as LDA (e.g., see experimental results reported in (Blei et al., 2003)). To get a sense of the capability to separate data into classes without the class labels, we project the 100 hidden posteriors onto 2D using t-SNE<sup>4</sup> (van der Maaten and Hinton, 2008). Fig. 8 (left) depicts the distribution of documents, where class information is only used for visual labelling. The separation is overall satisfactory.

For the quantitative evaluation, the next step is to run classifiers on learnt features. For comparison, we also use those discovered by NMF and LDA. For all models, 100 latent factors are used, and thus the dimensions are reduced 10-fold. We split TDT2 text corpus into 80% for training and 20% for testing. We train linear SVMs<sup>5</sup> on all word features and low-dimensional representations provided by LDA, NMF and NRBM with various proportions of training data. Fig. 8 (right) shows the classification errors on testing data for all methods. The learnt features of LDA and NMF improve classification performance when training label is limited. This is expected because the learnt representations are more compact and thus less prone to overfitting. However, as more labels are available, the word features catch up and eventually outperform those by LDA/NMF. Interestingly, this difficulty does not occur for the features learnt by the NRBM, although it does appear that the performance saturates after seeing 20% of training labels. Note that this is significant given the fact that computing learnt representations in NRBM is very fast, requiring only a single matrix-vector multiplication per document.

## 4. Discussion

Our work was partly motivated by the capacity of nonnegative matrix factorisation (NMF) (Lee and Seung, 1999) to uncover parts-based representations. Given a nonnegative data matrix  $\mathbf{V} \in \mathbb{R}^{N \times D}$ , NMF attempts to factorise into two low-rank real-valued nonnegative

4. Note that the t-SNE does not do clustering, it only reduces the dimensionality into 2D for visualisation while still try to preserve the local properties of the data.

5. SVM with Gaussian kernels did not perform well.

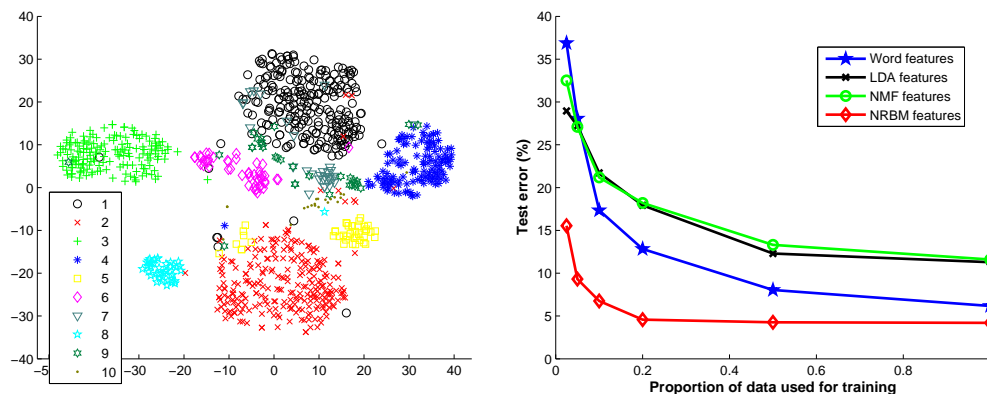


Figure 8: An example visualisation of 10 categories and classification performance on TDT2 text corpus. On the left figure, t-SNE (van der Maaten and Hinton, 2008) projection performs on 100 higher representations of documents mapped using the NRBM. Categories are labelled using the ground truth. (Best viewed in colours). The right figure represents the classification results for different proportions of training data.

matrices, the basis  $\mathbf{W} \in \mathbb{R}^{N \times K}$  and the coefficient  $\mathbf{H} \in \mathbb{R}^{K \times D}$ , i.e.,  $\mathbf{V} \approx \mathbf{W}\mathbf{H}$ . Thus  $\mathbf{W}$  plays the same role as NRBM’s connection weights, and each column of  $\mathbf{H}$  assumes the “latent factors”. However, it has been pointed out that unless there are appropriate sparseness constraints or certain conditions, NMF is not guaranteed to produce parts (Hoyer, 2004; Donoho and Stodden, 2003). Our experiment shows, on the contrary, NRBM can still produce parts-based representation when NMF fails (Fig. 1, also reported in (Hoyer, 2004)).

On the theoretical side, the main difference is that our cases the latent factors are stochastic binary that are inferred from the model, but not learnt as in the case of NMF. In fact this seemingly subtle difference is linked to a fundamental drawback suffered by the NMF: The learnt latent factors are limited to seen data only, and must be relearnt for every new data point. The NRBM, on the other hand, is a fully generative model in that it can generate new samples from its distribution, and at the same time, the representations can be efficiently computed for unseen data (see Eq. (3)) using one matrix operation.

Recently, there has been work closely related to NMF that does not require re-estimation on unseen data (Lemme et al., 2012). In particular, the coefficient matrix  $\mathbf{H}$  is replaced by the mapping from the data itself, that is  $\mathbf{H} = \sigma(\mathbf{W}^\top \mathbf{V})$ , resulting in the so-called autoencoder structure (AE), that is  $\mathbf{V} \approx \mathbf{W}\sigma(\mathbf{W}^\top \mathbf{V})$ , where  $\sigma(x)$  is a vector of element-wise nonlinear transforms and  $\mathbf{W}$  is nonnegative. A new representation estimated using  $\sigma(\mathbf{W}^\top \mathbf{V})$  now plays the role of the posteriors in NRBM, although it is non-probabilistic. The main difference from NRBM is that the nonnegative AE does not model data distribution, and thus cannot generate new samples. Also, it is still unclear how the new

representation could be useful for classification in general and on non-vision data in particular.

For the semantic analysis of text, our proposed model is able to discover plausible thematic features. Compared with those discovered by topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003), we found that they are qualitatively similar. We note that the two approaches are not directly comparable because the notion of association strength between a latent factor and a word, as captured in the nonnegative weight  $W_{nk}$ , cannot be translated into the properly normalised probability  $P(v_n = 1 \mid z_n = k)$  as in LDA, where  $z_n$  is the topic that generates the word  $v_n$ . Nevertheless, the NRBM offers many advantages over the LDA: (i) the notion that each document is generated from a subset of themes (or semantic features) in NRBM is an attractive alternative to the setting of topic distribution as assumed in LDA (see also (Ghahramani and Griffiths, 2005)); (ii) inference to compute the latent representation given an input is much faster in NRBM with only one matrix multiplication step, which it typically requires an expensive sampling run in LDA; (iii) learning in NRBM can be made naturally incremental, whilst estimating parameter posteriors in LDA generally requires the whole training data; and (iv) importantly, as shown in our experiments, classification using the learnt representations can be more accurate with NRBM (see Fig. 8).

This work can be considered along the line of imposing structures on standard RBM so that certain regularities are explicitly modelled. Our work has focused on nonnegativity as a way to ensure sparseness on the weight matrix, and consequently the latent factors. An alternative would be enforcing sparseness on the latent posteriors, e.g., (Hinton, 2012). Another important aspect is that the proposed NRBM offers a way to capture the so-called “explaining away” effect, that is the latent factors compete with each other as the most plausible explanatory causes for the data (see also Section 2.3). The competition is encouraged by the nonnegative constraints, as can be seen from the generative model of data  $p(v_n = 1 \mid \mathbf{h}) = \sigma(a_n + \sum_k W_{nk} h_k)$ , in that some large weights (strong explaining power) will force other to degrade or even vanish (weak explaining power). This is different from standard practice in neural networks, where complex inhibitory lateral connections must be introduced to model the competition (Hinton and Ghahramani, 1997).

One important question is that under such constraints, besides the obvious gains in structural regularisation, do we lose representational power of the standard RBM? On one hand, our experience has indicated that yes, there is certain loss in the ability to reconstruct the data, since the parameters are limited to be nonnegative. On the other hand, we have demonstrated that this does not away translate into loss of predictive power. In our setting, the degree of constraints can also be relaxed by lowering down the regularisation parameter  $\alpha$  in Eq. (5), and this would allow some parameters to be negative. Finally we note that our study has been confined to binary data, but it is not the inherent limitation and could be extended to continuous (Hinton and Salakhutdinov, 2006) and count data (Gehler et al., 2006).

## 5. Conclusion

To summarise, this paper has introduced a novel variant of the powerful restricted Boltzmann machine, termed nonnegative RBM (NRBM), where the mapping weights are con-

strained to be nonnegative. This gives the NRBM the new capacity to discover interpretable parts-based representations and semantically plausible high-level features for additive data such as images and texts. In addition, the NRBM can be used to uncover the intrinsic dimensionality of the data, the ability not seen in standard RBM. This is because under the nonnegativity constraint, the latent factors “compete” with each other to best represent data, leading to some unused factors. At the same time, the NRBM retains nearly full strength of the standard RBM, namely, compact and discriminative distributed representation, fast inference and incremental learning. Compared with the well-studied parts-based decomposition scheme, the nonnegative matrix factorisation (NMF), the NRBM could work in places where the NMF fails. When it comes to classification using the learnt representations, the features discovered by the NRBM are more discriminative than those by the NMF and the latent Dirichlet allocation (LDA). Thus, we believe that the NRBM is a seriously fast alternative to the NMF and LDA for a variety of data processing tasks.

## References

- Shivani Agarwal, Aatif Awan, and Dan Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(11):1475–1490, 2004.
- AT&T at Cambridge. The ORL Database of Faces, AT&T Laboratories of Cambridge. URL <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538*, 2012.
- David M Blei, Andrew Y Ng, and Michael Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.
- Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(12):1624–1637, December 2005.
- CBCL at MIT. CBCL Face Database #1, MIT Center For Biological and Computation Learning. URL <http://cbcl.mit.edu/software-datasets/FaceData2.html>.
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems (NIPS)*, 2003.
- Yoav Freund and David Haussler. Unsupervised learning of distributions on binary vectors using two layer networks. Technical report, Santa Cruz, CA, USA, 1994.
- P.V. Gehler, A.D. Holub, and M. Welling. The rate adapting Poisson model for information retrieval and object recognition. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 337–344. ACM, 2006.
- Zoubin Ghahramani and Thomas L Griffiths. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems (NIPS)*, pages 475–482, 2005.

- Geoffrey Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006.
- Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.
- Geoffrey E Hinton. A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade*, pages 599–619. Springer, 2012.
- Geoffrey E Hinton and Zoubin Ghahramani. Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1177–1190, 1997.
- Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research (JMLR)*, 5:1457–1469, 2004.
- Yann Lecun, Corinna Cortes, and Christopher J.C. Burges. The MNIST database of handwritten digits. URL <http://yann.lecun.com/exdb/mnist/>.
- Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- Andre Lemme, René Felix Reinhart, and Jochen Jakob Steil. Online learning and generalization of parts-based image representations by non-negative sparse autoencoders. *Neural Networks*, 33(0): 194 – 203, 2012. ISSN 0893-6080.
- Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, August 2000, pp. 497–506. ISBN 0387987932.
- P. Smolensky. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Information processing in dynamical systems: Foundations of harmony theory, pages 194–281. MIT Press, Cambridge, MA, USA, 1986. ISBN 0-262-68053-X.
- Yee Whye Teh and Geoffrey E Hinton. Rate-coded restricted Boltzmann machines for face recognition. *Advances in Neural Information Processing Systems (NIPS)*, pages 908–914, 2001.
- L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *The Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008.