Neural Networks 105 (2018) 405–418

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

On the importance of hidden bias and hidden entropy in representational efficiency of the Gaussian-Bipolar Restricted Boltzmann Machines

Altynbek Isabekov, Engin Erzin *

College of Engineering, Koç University, Rumelifeneri yolu, Istanbul 34450, Turkey

ARTICLE INFO

Article history:
Received 18 July 2017
Received in revised form 12 March 2018
Accepted 3 June 2018

Keywords:
RBM
Hidden entropy
Hidden bias
Representational efficiency
Autoencoder
Deep learning

ABSTRACT

In this paper, we analyze the role of hidden bias in representational efficiency of the Gaussian-Bipolar Restricted Boltzmann Machines (GBPRBMs), which are similar to the widely used Gaussian-Bernoulli RBMs. Our experiments show that hidden bias plays an important role in shaping of the probability density function of the visible units. We define hidden entropy and propose it as a measure of representational efficiency of the model. By using this measure, we investigate the effect of hidden bias on the hidden entropy and provide a full analysis of the hidden entropy as function of the hidden bias for small models with up to three hidden units. We also provide an insight into understanding of the representational efficiency of the larger scale models. Furthermore, we introduce Normalized Empirical Hidden Entropy (NEHE) as an alternative to hidden entropy that can be computed for large models. Experiments on the MNIST, CIFAR-10 and Faces data sets show that NEHE can serve as measure of representational efficiency and gives an insight on minimum number of hidden units required to represent the data.

1. Introduction

Recently, the subject of Restricted Boltzmann Machines (RBMs) and deep learning became the focus of attention in machine learning research. Application of deep learning in different areas such as image processing, computer vision, and natural language processing has proved its efficiency (Collobert & Weston, 2008; Krizhevsky, 2009; Ranzato, Poultney, Chopra, & Cun, 2007). RBMs are probabilistic generative models which are used to obtain new (usually compressed) representation of the data. Different types of RBMs are used as building blocks for deep neural architecture by means of unsupervised layer-wise pre-training (Bengio, Lamblin, Popovici, & Larochelle, 2006). However, RBMs with real-valued inputs are of primarily importance as most of the analyzed data is real-valued. Conventional Bernoulli–Bernoulli RBMs have been studied in Cuartas (2012), Le Roux and Bengio (2008) and Martens, Chattopadhya, Pitassi, and Zemel (2013) where they are referred as universal approximators of any binary distribution. One of the first Gaussian-Bernoulli RBM (GBLRBM) models with real-valued inputs was proposed in Bengio et al. (2006) and Hinton and Salakhutdinov (2006), and was explicitly analyzed in Krizhevsky (2009). Another version of a GBLRBM with a more intuitive energy function was proposed in Cho, Ilin, and Raiko (2011). Moreover, a more simplified sub-type of the latter model was analyzed in Melchior (2012) and Wang, Melchior, and Wiskott (2012).

Despite the ongoing research in this field, still not much is known about the principle of operation of GBLRBMs. Combinatorial nature of the model makes the analysis even harder. Nevertheless, conceptual understanding of GBLRBMs is given in Melchior (2012). The thesis has a well-described comparison to a Gaussian mixture model and a good visualization of the modeled distribution that gives an insight into the principle of operation of GBPRBMs. However, the thesis lacks analysis of hidden bias and its effect on the modeled probability density function. The effect of the biases and the mean of the data on the learning process was investigated in Melchior, Fischer, and Wiskott (2016) and Montavon and Müller (2012). Visible and hidden offsets are used to center the RBM model and make learning more stable.

Another interesting visualization of RBMs is given in Yosinski and Lipson (2012). Debugging of the RBMs is done by visualizing weight parameters as a tensor in a cube filled with small cells. Evaluation of histograms of the parameters on the mini-batch helps finding optimal stopping point for the training process. Disappearance of Gaussian-like shapes of the histograms indicates that training has converged to a stationary phase. This phenomenon was analyzed in Dieleman and Schrauwen (2012). A measure of non-Gaussianity based on negentropy and excess kurtosis was proposed as a stopping criterion for the training.

* Corresponding author.
E-mail addresses: aisabekov@ku.edu.tr (A. Isabekov), eerzin@ku.edu.tr (E. Erzin).

https://doi.org/10.1016/j.neunet.2018.06.002
0893-6080/

The problem of measuring usefulness of the hidden neurons was investigated in Berglund, Raiko, and Cho (2014), Le Roux and Bengio (2008) and Martens et al. (2013). The first two papers describe the effect of augmenting hidden layer on the representational efficiency of Bernoulli–Bernoulli RBMs. In the latter paper, mutual information between visible and hidden units is suggested as a measure of relevant activity of the hidden units. Usefulness of the hidden neurons is also tested by pruning neurons after training and by adding neurons during training. The results show that models initialized with a large number of hidden units can be simplified by pruning neurons without decreasing classification performance.

Nowadays, most of the research in deep learning is concentrated on application of RBMs and speeding up the training process. Fundamental questions remain still unanswered. What is a good measure for usefulness of the hidden neurons? How does the hidden bias affect the representational efficiency of the RBM model? What is the number of hidden neurons needed to represent the data? We try to answer these questions by introducing a new Gaussian-Bipolar RBM (GBPRBM) model, in which we investigate representational efficiency of hidden units in defining distribution of the visible units. This model is very similar to Gaussian-Bernoulli RBM except that it has a more symmetrical geometry which facilitates hidden entropy analysis described in Section 3.

Our contributions are summarized as follows:

- In Section 3, we define hidden entropy function and propose it as a measure of representational efficiency of GBPRBM models. We demonstrate how hidden bias shapes probability distribution of visible units. Moreover, we present a list of conditions needed to attain maximum hidden entropy. Also we provide a full analysis of the hidden entropy function for models with up to three hidden units. In this analysis, regions of high hidden entropy are given analytically in terms of other model parameters. This analysis provides an intuition to visualize hidden entropy space in higher dimensions.
- In Section 4, we propose a technique to measure activations of hidden units by defining Normalized Empirical Hidden Entropy (NEHE) function as an upper bound to the hidden entropy. This function allows to analyze models with higher number of hidden units. By measuring NEHE on the GBPRBM models trained using MNIST, CIFAR-10 and Faces data sets, we illustrate how number of hidden units affects representational efficiency of the GBPRBM models. This experiment gives an insight on the minimum number of hidden units needed to represent the data.

Findings and derivations in the paper are presented using examples. The reference GBPRBM model given in Section 2.1 and its derivative models with smaller number of hidden and visible units are used in visualization of the probability of visible units and the hidden entropy function.

2. Gaussian-Bipolar restricted Boltzmann machines

Gaussian-Bipolar Restricted Boltzmann Machine (GBPRBM) is an undirected graphical model which is used to model relation between *visible* and *hidden* units in a probabilistic way. GBPRBMs have real-valued inputs in the visible layer and binary units in the hidden layer.

Let the input vector with real-valued visible units be of size V such that $\mathbf{v} = [v_1\ v_2\ \dots\ v_V]^T$. Binary hidden units are constrained to have antipode values $\{-1, 1\}$ and grouped into a column vector $\mathbf{h} = [h_1\ h_2\ \dots\ h_H]^T$ with H being the number of hidden units. For notational consistency, visible units are represented by vectors $\mathbf{v}$, $\mathbf{u}$, hidden units — by vectors $\mathbf{f}$, $\mathbf{h}$, $\mathbf{g}$ throughout the paper. Subscripts i, j are reserved for visible and hidden units, respectively.

Two more parameters are associated with visible units. The first one is visible bias term b_i^v and the second one is visible variance term σ_i where $i \in \{1, \dots, V\}$. Bias terms are also present in the hidden units as: $b_j^h,\ j \in \{1, \dots, H\}$. Visible and hidden units are connected using weights $w_{ij},\ i \in \{1, \dots, V\},\ j \in \{1, \dots, H\}$. The relationship between visible and hidden units is described by the energy function

$$E(\mathbf{v}, \mathbf{h}) = \frac{1}{2}(\mathbf{v} - \mathbf{b}_v)^T \boldsymbol{\Sigma}^{-1}(\mathbf{v} - \mathbf{b}_v) - \mathbf{v}^T \boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{h} - \mathbf{b}_h^T\mathbf{h}, \tag{1}$$

which is defined similarly for the Gaussian-Bernoulli RBM model in Cho et al. (2011) with parameters

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,H} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,H} \\ \vdots & \vdots & \ddots & \vdots \\ w_{V,1} & w_{V,2} & \cdots & w_{V,H} \end{bmatrix},\ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_V^2 \end{bmatrix},$$

$$\mathbf{b}_v = \left[b_1^v, b_2^v, \dots, b_V^v\right]^T, \qquad \mathbf{b}_h = \left[b_1^h, b_2^h, \dots, b_H^h\right]^T. \tag{2}$$

The energy function is used to define joint probability density function (pdf) of $\mathbf{v}$ and $\mathbf{h}$

$$p(\mathbf{v}, \mathbf{h}) = \frac{\exp(-E(\mathbf{v}, \mathbf{h}))}{\int_{\mathbf{u}} \sum_{\mathbf{g}} \exp(-E(\mathbf{u}, \mathbf{g}))d\mathbf{u}}, \tag{3}$$

where $\int_{\mathbf{u}}(\dots)d\mathbf{u}$ is integration over all space of visible units and $\sum_{\mathbf{g}}$ is summation over all 2^H configurations of hidden vector $\mathbf{g}$. Likewise, conditional probability of the visible vector $\mathbf{v}$ given hidden vector $\mathbf{h}$ is defined as

$$\begin{aligned} p(\mathbf{v}|\mathbf{h}) &= \frac{p(\mathbf{v}, \mathbf{h})}{p(\mathbf{h})} = \frac{\exp(-E(\mathbf{v}, \mathbf{h}))}{\int_{\mathbf{u}} \exp(-E(\mathbf{u}, \mathbf{h}))d\mathbf{u}} \\ &= \mathcal{N}\left(\mathbf{v};\ [\mathbf{b}_v + \mathbf{W}\mathbf{h}],\ \boldsymbol{\Sigma}\right), \end{aligned} \tag{4}$$

where $\mathcal{N}(\mathbf{v};\ \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Since $\boldsymbol{\Sigma}$ is diagonal, the conditional pdf can be represented as product of marginal conditional pdfs of each visible unit:

$$p(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{V} \mathcal{N}\left(v_i;\ [b_i^v + \mathbf{W}_{(i,:)}\mathbf{h}],\ \sigma_i^2\right) = \prod_{i=1}^{V} p(v_i|\mathbf{h}). \tag{5}$$

Detailed derivations of $p(\mathbf{v}|\mathbf{h})$ can be found in Appendix A.1.

2.1. Data modeling using probability of visible units

Restricted Boltzmann machines have been used as unsupervised learning algorithms to extract latent features and to model the data distribution. This corresponds to clustering in the space of visible units and encoding each cluster using hidden units. Nevertheless, RBMs are probabilistic models, and a more straightforward interpretation of the data modeling is representing the data distribution as probability of visible units $p(\mathbf{v})$. The proposed GBPRBM has a probability of visible units given as

$$\begin{aligned} p(\mathbf{v}) &= \sum_{\mathbf{h}} p(\mathbf{h})p(\mathbf{v}|\mathbf{h}) \\ &= \sum_{\mathbf{h}} p(\mathbf{h}) \times \mathcal{N}\left(\mathbf{v};\ [\mathbf{b}_v + \mathbf{W}\mathbf{h}],\ \boldsymbol{\Sigma}\right). \end{aligned} \tag{6}$$

This implies that a GBPRBM models the probability of observing $\mathbf{v}$ as a Gaussian Mixture Model (GMM). Every Gaussian component with covariance matrix $\boldsymbol{\Sigma}$ is scaled by mixture weight $p(\mathbf{h})$ and located at $[\mathbf{b}_v + \mathbf{W}\mathbf{h}]$.

Visualizations of $p(\mathbf{v})$ of submodels with dimension V equal to 1, 2 and 3 are shown in Fig. 1. In order to exemplify and visualize