Pattern Recognition 47 (2014) 3179–3191

Contents lists available at ScienceDirect

Pattern Recognition

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

A sparse-response deep belief network based on rate distortion theory

Nan-Nan Ji, Jiang-She Zhang*, Chun-Xia Zhang

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an Shaanxi 710049, China

ARTICLE INFO

Article history:
Received 26 April 2013
Accepted 26 March 2014
Available online 12 April 2014

Keywords:
Deep belief network
Kullback–Leibler divergence
Information entropy
Rate distortion theory
Unsupervised feature learning

ABSTRACT

Deep belief networks (DBNs) are currently the dominant technique for modeling the architectural depth of brain, and can be trained efficiently in a greedy layer-wise unsupervised learning manner. However, DBNs without a narrow hidden bottleneck typically produce redundant, continuous-valued codes and unstructured weight patterns. Taking inspiration from rate distortion (RD) theory, which encodes original data using as few bits as possible, we introduce in this paper a variant of DBN, referred to as sparse-response DBN (SR-DBN). In this approach, Kullback–Leibler divergence between the distribution of data and the equilibrium distribution defined by the building block of DBN is considered as a distortion function, and the sparse response regularization induced by L_1-norm of codes is used to achieve a small code rate. Several experiments by extracting features from different scale image datasets show that our approach SR-DBN learns codes with small rate, extracts features at multiple levels of abstraction mimicking computations in the cortical hierarchy, and obtains more discriminative representation than PCA and several basic algorithms of DBNs.

1. Introduction

Recent neuroscience findings [1–6] have provided insight into the principles governing information representation in the mammal brain, leading to new ideas for designing systems to effectively represent information. One of the key findings is that the mammal brain is organized in a deep architecture. Given an input percept, it is represented with multiple levels of abstraction, each level corresponding to a different area of cortex. By using these abstract features learned through the deep architecture, human achieves perfect performance on many real-world tasks such as object recognition, detection, prediction, and visualization. In current machine learning community, how to imitate this hierarchical architecture of mammal brain to obtain good representation of data in order to improve the performance of a learning algorithm has become an essential issue.

Recently, deep learning has become the dominant technique to learn good information representation that exhibits similar characteristics to that of the mammal brain. It has gained significant interest for building hierarchical representations from unlabeled data. A deep architecture consists of feature detector units arranged in multiple layers: lower layers detect simple features and feed into higher layers, which in turn detect more complex features. In particular, deep belief network (DBN), the most popular approach of deep learning, is a multilayer generative model in which each layer encodes statistical dependencies among the units in the layer below, and it can be trained to maximize (approximately) the likelihood of its training data. So far, there have been a great deal of DBN models being proposed. For example, Hinton et al. [7] proposed an algorithm based on learning individual layers of a hierarchical probabilistic graphical model from the bottom up. Bengio et al. [8] proposed a similar greedy algorithm on the basis of auto-encoders. Ranzato et al. [9] developed an energy-based hierarchical algorithm, using a sequence of sparsified auto-encoders/decoders. Particularly, the model proposed by Hinton et al. is a breakthrough for training deep networks. It can be viewed as a composition of simple learning modules, each of which is a restricted Boltzmann machine (RBM) that contains a layer of visible units representing observable data and a layer of hidden units learned to represent features that capture higher-order correlations in the data [10]. Nowadays, DBNs have been successfully applied to a variety of real-world applications, including hand-written character recognition [7,11], text representation [12], audio event classification, object recognition [13], human motion capture data [14,15], information retrieval [16], machine transliteration [17], speech recognition [18–20] and various visual data analysis tasks [21–23].

Although DBNs have demonstrated promising results in learning good codes or representations, DBNs without constraints on the hidden layers may produce redundant, continuous-valued codes and unstructured weight patterns. Some scholars attempted to further improve DBNs' performance according to add constraints

* Corresponding author.
E-mail address: jszhang@mail.xjtu.edu.cn (J.-S. Zhang).

http://dx.doi.org/10.1016/j.patcog.2014.03.025

on the representations [24–30]. Among them, introducing the notion of sparsity makes DBN present state-of-the-art results because sparse representation is able to obtain succinct codes and structured weight patterns. Given images, sparse DBN is able to discover low-level structures such as edges, as well as high-level structures such as corners, local curvatures, and shapes.

Sparsity was first proposed as a model of simple cells in the visual cortex [31]. Up to now, it has been a key element of DBNs learning through exploiting a variant of auto-encoders, RBM, or other unsupervised learning algorithms. In practice, there are many ways to enforce some form of sparsity on the hidden layer representation of a deep architecture. The first successful deep architecture exploiting sparsity of representation involved auto-encoders [9]. Sparsity was achieved with a so-called sparsifying logistic, by which the codes are obtained with a nearly saturating logistic whose offset is adapted to maintain a low average number of times that the code is significantly non-zero. One year later the same group introduced a somewhat simple variant [28] through assigning a Student-*t* prior on the coders. In the past, the Student-*t* prior was used to obtain sparse MAP estimates for the codes generating an input [29] in computational neuroscience models of the V1 visual cortex area. Another approach also related to computational neuroscience involves two levels of sparse RBMs [24]. Sparsity is achieved with a regularization term that penalizes a deviation of the expected activation of hidden units from a fixed low-level. One level of sparse coding of images results in filters very similar to those seen in V1. When training a sparse deep belief network, the second level appears to detect visual features similar to those observed in area V2 of visual cortex.

From the point of view of information theory, one of the major principles for finding concise representations is rate distortion (RD) theory [32], which focuses on the problem of determining minimal amount of information that should be communicated over a channel, so that a compressed representation of the original data can be approximately reconstructed at the output data without exceeding a given distortion. Sparse coding methods can be interpreted as special cases of RD theory [33]. For deep multi-layer neural networks, hidden layers without narrow bottleneck may result in redundant and continuous-valued codes [34]. We hold that incorporating the constraint of a minimum rate of information flow into the training process of multi-layer neural networks is able to make networks obtain succinct representations. From this point of view, we propose in this paper a novel version of sparse DBNs for unsupervised feature extraction by taking inspiration from the idea of RD theory. In DBNs, activation probability of the hidden units over a data vector is always regarded as its representation or code. Therefore, in our approach, a small code rate is achieved by adding a constraint on the activation probability of hidden units. The used constraint is L_1-norm of this activation probability. Meanwhile, Kullback–Leibler divergence between the distribution of data and the equilibrium distribution defined by the building block of DBN is considered as a measurement of distortion. More specifically, the novel approach is implemented by a trade-off between the L_1 regularizer and the Kullback–Leibler divergence. The novel approach has the advantages that representations with small information rate can be automatically learnt and the hierarchical representations (which mimics computations in the cortical hierarchy) can be obtained. Furthermore, compared to PCA and several basic algorithms of DBN, the new approach learns more discriminative representations.

The remainder of this paper is organized as follows. We first describe the DBN's structure and building block, RBM, with their learning rules in Section 2. Section 3 introduces the novel sparse DBN based on RD theory, and provides its learning rule. In Section 5, the novel sparse DBN is compared with several methods qualitatively (the hierarchical bases learned by algorithms) and quantitatively (the classification performance of subsequently built classifier, starting from the representations obtained by unsupervised learning). Finally, this paper is concluded with a summary and some directions for further research in Section 6.

2. Deep belief network (DBN) and its building block

DBNs are probabilistic generative models that contain many layers of hidden variables, in which each layer captures high-order-correlations between the activities of hidden features in the layer below. A key feature of this algorithm is its greedy layer-by-layer training that can be repeated several times to learn a deep, hierarchical model. The main building block of a DBN is a bipartite undirected graphical model called the Restricted Boltzmann Machine (RBM). In this section, we provide a brief technical overview of RBM and the greedy learning algorithm for DBNs.

2.1. Restricted Boltzmann machine (RBM)

RBM [10,11,35,36] is a two-layer, bipartite, undirected graphical model with a set of (binary or real-valued) visible units (random variables) $\mathbf{v}$ of dimension D representing observable data, and a set of binary hidden units (random variables) $\mathbf{h}$ of dimension K learned to represent features that capture higher-order correlations in the observable data. These two layers are connected by a symmetrical weight matrix $W \in R^{D\times K}$, whereas there are no connections within a layer. Fig. 1 illustrates the undirected graphical model of an RBM.

RBM can be viewed as a Markov random field that tries to represent input data with hidden units. Here, the weights encode a statistical relationship between the hidden units and the visible units. The joint distribution over the visible and hidden units is defined by

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

$$Z = \sum_{v}\sum_{h}\exp(-E(\mathbf{v},\mathbf{h})), \tag{2}$$

where Z is a normalization constant. $E(\mathbf{v},\mathbf{h})$ denotes the energy of the state $(\mathbf{v},\mathbf{h})$. If the visible units are binary-valued, the energy function can be defined as

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{D}\sum_{j=1}^{K} v_i W_{ij} h_j - \sum_{j=1}^{K} b_j h_j - \sum_{i=1}^{D} c_i v_i, \tag{3}$$

where b_j and c_i are respectively hidden and visible unit biases. If the visible units are real-valued, we can define the energy function by adding a quadratic term to make the distribution well defined, that is,

$$E(\mathbf{v},\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{D} v_i^2 - \sum_{i=1}^{D}\sum_{j=1}^{K} v_i W_{ij} h_j - \sum_{j=1}^{K} b_j h_j - \sum_{i=1}^{D} c_i v_i. \tag{4}$$

From the energy function, we can see that the hidden units h_j are independent of each other when conditioning on $\mathbf{v}$ since there are no direct connections between hidden units. Similarly, the visible units v_i are also independent of each other when conditioning on $\mathbf{h}$.

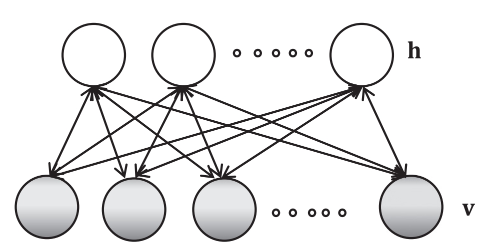

Fig. 1. Undirected graphical model of an RBM.