Neurocomputing 139 (2014) 328–335

Contents lists available at ScienceDirect

Neurocomputing

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

Improving mixing rate with tempered transition for learning restricted Boltzmann machines

Jungang Xu *, Hui Li, Shilong Zhou

School of Computer and Control Engineering, University of Chinese Academy of Sciences, 101408 Beijing, China

ARTICLE INFO

Article history:
Received 4 December 2013
Received in revised form
19 February 2014
Accepted 27 February 2014
Communicated by zhi yong Liu
Available online 8 April 2014

Keywords:
Restricted Boltzmann machines
Tempered transition
Mixing rate
Deep learning

ABSTRACT

Recently, as the building block of deep generative models such as Deep Belief Networks (DBNs), Restricted Boltzmann Machines (RBMs) have attracted much attention. RBM is a Markov Random Field (MRF) associated with a bipartite undirected graph which is famous for powerful expression and tractable inference. While training an RBM, we need to sample from the model. The larger the mixing rate is, the smaller the bias of the samples is. However, neither Gibbs sampling based training methods such as Contrastive Divergence (CD) nor Parallel Tempering based training methods can achieve satisfying mixing rate, which causes poor rendering of the diversity of the modes captured by these trained models. This property may hinder the existing methods to approximate the likelihood gradient. In order to alleviate this problem, we attempt to introduce Tempered Transition, an advanced tempered Markov Chain Monte Carlo method, into training RBMs to replace Gibbs sampling or Parallel Tempering for sampling from RBMs. Experimental results show that our proposed method outperforms the existing methods to achieve better mixing rate and to help approximate the likelihood gradient.

1. Introduction

Learning algorithms for deep architectures such as Deep Belief Networks [1,2] and Deep Boltzmann Machines [3] have recently been proposed and successfully applied to various machine learning tasks such as image processing [4,5], speech recognition [6,7], natural language processing [8] and so on. Deep neural networks are distinguished from shallow architectures such as Support Vector Machine (SVM) by a large number of layers of neurons and characterized by using layer-wise unsupervised pre-training to learn a more accurate model.

Restricted Boltzmann Machines (RBMs) [9–12] can be interpreted as neural network models which consist of two types of units called visible neurons and hidden neurons. And there are only connections between different types of neurons, therefore RBMs can be divided into two layers. The first layer is constituted by the visible neurons and corresponds to an observation, the second layer is constituted by the hidden neurons and models the dependencies between the components of the observation. So, RBMs can be viewed as non-linear feature detectors [9,10]. By stacking RBMs, such multistage learning methods have been empirically confirmed as good as, or in many cases better than, conventional learning methods, such as back propagation with random initialization [1,3,13–15]. Therefore, RBMs play a very important role in deep learning and it is necessary to explore more efficient method for learning RBMs.

Although RBMs are famous for their powerful expression and tractable inference [16], training an RBM can be difficult in practice. The difficulties come from the intractability of the log-likelihood gradient which is composed of a positive phase term and a negative phase term. Calculating the exact value of the negative phase term requires unbiased sampling from the model distribution for a long time to ensure convergence to stationarity, which is of exponential complexity. Therefore, additional approximations are usually introduced into the learning methods to yield more efficient algorithms.

Gibbs sampling based approximations of the negative phase term in the log-likelihood gradient often lead to divergence of the training procedure and result in spurious probability modes far from the training data [17]. Therefore, RBM learning algorithms based on Gibbs sampling, such as Contrastive Divergence show very poor mixing, as we can see in Section 5. Parallel Tempering based approximations can suppress the diverging problem, but the bias of the approximations still exists. Neither the learning algorithms based on Gibbs sampling nor the ones based on Parallel Tempering are ideal enough to train full-fledged generative models of data because of the poor mixing rate.

To improve the mixing rate of negative phase, we propose an RBM training algorithm based on Tempered Transition [18],

* Corresponding author. Tel./fax: +86 10 82681221.
E-mail addresses: xujg@ucas.ac.cn, xujungang@hotmail.com (J. Xu), lihui211@mails.ucas.ac.cn (H. Li), zhoushilong12@mails.ucas.ac.cn (S. Zhou).

http://dx.doi.org/10.1016/j.neucom.2014.02.024
0925-2312/

another extended ensemble Monte Carlo method besides Parallel Tempering [19], which is characterized by strong ability of handling multimodal distributions. Different from existing methods, we sample from RBM distribution from multiple chains in serial, but always keep current state in the objective distribution. The experimental results on MNIST data set of handwritten digits show that the training algorithm based on Tempered Transition obtains better mixing rate than existing methods and improves the learning procedure to some extent.

The rest part of this paper is organized as follows. The related works is described in Section 2. The structure of the RBMs and log-likelihood gradient approximation for training RBMs are introduced in Section 3. The details of our proposed approach based on Tempered Transition are described in Section 4. In Section 5, we show the experimental results of our approach on the MNIST data set. Finally, we conclude the paper and discuss the directions of the future work.

2. Related works

Recently, quite a few efforts are made on training RBMs. Most of the existing methods are based on Gibbs sampling and some are based on Parallel Tempering, as described below.

The Contrastive Divergence (CD) algorithm [11,20,21] is the most popular learning algorithm for RBMs, which approximates the negative phase term of log-likelihood gradient by sampling from a Gibbs chain that only runs for k steps (and usually $k=1$) starting from the observed training example. The strategy of CD is effective in learning representations or features of data in practice, especially when k is larger. However, the data-centric focus of CD training can yield biased estimates of the gradient and the theoretical results from Ref. [22] can explain this phenomenon.

The Persistent Contrastive Divergence (PCD) algorithm [23] was proposed to improve upon CD's limitation (which is a kind of biased estimate method). Similar to CD, PCD approximates the negative phase of the gradient with samples drawn from a Gibbs chain runs for k steps. However, in PCD, the Gibbs chain is initialized by the state in which it ends for the previous model instead of the training sample. The fundamental idea underlying PCD is that one could assume that the initialization is close to the model distribution, even though the model has changed a bit in the parameter update. However, the reliance on a single persisting Markov chain often leads to degenerative training [24]. When faced with multimodal target distributions, Gibbs sampling used in the PCD negative phase estimation is easily to be stuck in the local optimum, leading to a chain that mixes slowly, over-representing certain modes of the distribution while under-representing others. This produces a biased estimate of the gradient, and the mini-batch strategy is only helpful for small-scale problems with simple distributions [25].

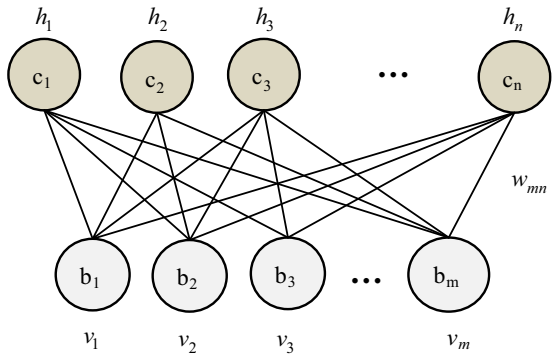

Fig. 1. The undirected graph of an RBM with n hidden and m visible variables.

The Fast Persistent Contrastive Divergence (FPCD) algorithm [24] attempts to improve upon PCD's mixing properties by introducing a group of additional parameters called fast parameters that are only used for sampling. FPCD tries to get out of any single mode of the distribution by these fast learning parameters and achieves better results in approximating the RBM gradient which are reported in Ref. [24]. However, neither PCD nor FPCD seem to enlarge the mixing rate (or decrease the bias of the approximation) sufficiently to avoid the divergence problem as can be seen in the empirical analysis in Refs. [17,25].

The Parallel Tempering [26] based training algorithm [27,28] replaces the single Gibbs chain used in PCD with a series of chains implementing a Parallel Tempering scheme. Parallel Tempering is one of a collection of extended ensemble Monte Carlo methods [19] introducing multiple versions of the same distribution under different temperatures into the sampling procedure. By sampling from the multiple chains and switching the states among the chains under different temperatures, the sampling procedure can get rid of the local maxima by means of smoothed distribution under high temperature. When approximating the negative phase term of RBM gradient, Parallel Tempering samples from a series of chains in parallel and switches the states between distributions under consecutive temperatures. The results from Refs. [27,28] show that Parallel Tempering improves mixing between multiple modes of the distribution and helps approximate the RBM gradient. However, only switching the states between distributions under consecutive temperatures is not enough to confirm shaking off the control of the local maxima.

In order to alleviate the problem listed above, we propose an RBM training method based on Tempered Transition [18,29], another extended ensemble Monte Carlo method besides Parallel Tempering [19]. Similar to Parallel Tempering, Tempered Transition approximates the negative phase term of the RBM gradient by sampling from the multiple chains under different temperatures. However, Tempered Transition moves from the desired distribution to easily-sampled distribution which is under high temperatures, and back to the desired distribution. This strategy is better for getting rid of local maxima to improve mixing and to help training RBMs.

3. Training of restricted Boltzmann machines

3.1. Restricted Boltzmann machines

A Restricted Boltzmann Machine is a restricted type of Boltzmann Machines (BM) which have been introduced as bidirectionally connected networks of stochastic processing units [9,24]. A BM can be used to learn important aspects of an unknown probability distribution based on samples from this distribution. However, there are practical limitations in using BM due to difficult and time-consuming learning process. RBM is proposed to alleviate this problem by imposing restrictions on the network topology [28].

Specifically, an RBM is a Markov Random Field (MRF) associated with a bipartite undirected graph as shown in Fig. 1. There are m visible units $v=(v_1,\ldots,v_m)$ to represent observable data and n hidden units $h=(h_1,\ldots,h_n)$ to capture dependencies between observed variables. We focus on binary RBMs where the random variables (v,h) take values from $\{0,1\}$. The probability distribution of (v,h) configuration is given by (1).

$$p(v,h)=\frac{1}{Z}e^{-E(v,h)} \tag{1}$$

And the probability of observable variables v is denoted as follows:

$$p(v)=\sum_h p(v,h)=\frac{1}{Z}\sum_h e^{-E(v,h)} \tag{2}$$