RESTRICTED BOLTZMANN MACHINES FOR IMAGE TAG SUGGESTION

Jiří Král, Michal Hradiš
Doctoral Degree Programme (1,5), FIT BUT
E-mail: ikral@fit.vutbr.cz, ihradis@fit.vutbr.cz

Supervised by: Pavel Zemčík
E-mail: zemcik@fit.vutbr.cz

Abstract: In this paper, we propose to model dependencies among binary variables in semantic tagging and similar tasks by Restricted Boltzmann Machines (RBM). In the proposed approach, Gibbs sampling allows learning RBMs even on data with large portion of missing values. Similarly, Gibbs sampling is used to estimate marginal probabilities of tags. The results show that the tag predictions become more certain with higher portion of known tags, and that the approach could be used for tag suggestion or semi-supervised learning.

Keywords: Tag suggestion, Restricted Boltzmann Machine, semantic indexing, Conditional Restricted Boltzmann Machine

1 INTRODUCTION

Automatic tagging of image data has wide applications ranging from supporting search in multi-media databases [7] to various classification tasks (e.g. video genre recognition [4]) and semi-supervised annotation [1]. Existing approaches to estimating probability of presence of semantic categories in images [7] rely mostly on content-based features extracted from the image to provide information about present classes. Such approaches are suitable for fully automatic scenarios where no user input is required. However, in certain scenarios the user input is available and, in fact, needed to provide reliable enough results. An example of such scenario is tagging of images uploaded by users to online multi-media repositories - current automatic tagging systems do not provide results reliable enough for this task. Another example is semi-supervised learning of the semantic classification system itself.

In this paper, we propose a novel approach which is able to model dependencies between semantic tags, as well as the dependency of tags on content-based features. By using Conditional Restricted Boltzmann Machines, we model the dependencies in an unified probabilistic framework which allows unknown variables (presence of semantic tags) to be estimated by Gibbs sampling. The proposed approach uses content-based features for a first rough estimate of tag probability. These estimates can be subsequently refined by making some of the tag variables visible (a user selects a presence or absence of some tags by hand).

The proposed approach is well suited for example for tagging of image data when uploading images to a database such as Flickr. There it can be used to provide suggestions of tags appropriate for the images from which the user picks the correct ones. The suggestions become more accurate as the user hand-selects some of the tags for the image.

A short introduction to Restricted Boltzmann Machines (RBM) and their conditional variant (CRBM) is given in Section 2 respective Section 3. Section 4 then introduces a method for learning RBM in cases when large portion of variables in training data is not known. Experiments and their results are discussed in Section 5. Finally, the paper is concluded in Section 6.

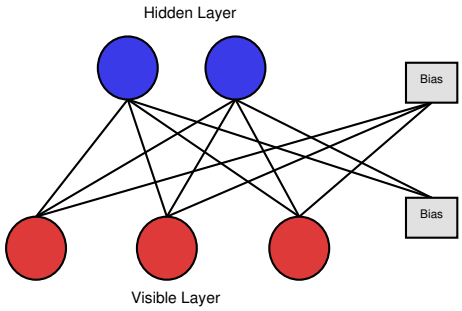

Figure 1: Restricted Boltzmann Machine - bipartite graph of visible and hidden variables linked by bidirectional connections.

2 RESTRICTED BOLTZMANN MACHINE

Restricted Boltzmann Machine [2] is an undirected bipartite graphical model. It defines a probability distribution over a vector of visible variables $\mathbf{v}$ and a vector of hidden variables $\mathbf{h}$ as shown in Figure 1. In this paper we consider the simplest version of RBM where $\mathbf{v}$ and $\mathbf{h}$ contain both binary variables. The visible variables $\mathbf{v}$ are fully defined by states of the hidden variables $\mathbf{h}$ and vice versa.

The joint probability over $\mathbf{v}$ and $\mathbf{h}$ is defined as

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

where Z is a normalization constant and E is energy function given by

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{v}^\top \mathbf{W}\mathbf{h} - \mathbf{v}^\top \mathbf{b}^v - \mathbf{h}^\top \mathbf{b}^h, \tag{2}$$

where (W) is a matrix of weights between elements of $\mathbf{v}$ and $\mathbf{h}$, and $\mathbf{b}^v$ and $\mathbf{b}^h$ are biases of visible respective hidden variables. Dependencies between the variables are expressed as

$$p(\mathbf{h}|\mathbf{v}) = \sigma\left(\mathbf{W}\mathbf{v} - \mathbf{b}^v\right) \text{ and } p(\mathbf{v}|\mathbf{h}) = \sigma\left(\mathbf{W}^\top \mathbf{h} - \mathbf{b}^h\right), \tag{3}$$

where $\sigma()$ is a sigmoid function.

As a generative model, RBM could be trained using maximum likelihood. However, derivatives of the likelihood are intractable. Hinton [2] introduced a practical approximation called *Contrastive Divergence* (CD). The CD algorithm computes gradients for optimization as

$$\nabla W = \langle \mathbf{v}\mathbf{h} \rangle_{data} - \langle \mathbf{v}\mathbf{h} \rangle_{recon} \tag{4}$$

$$\nabla \mathbf{b}^v = \langle \mathbf{v} \rangle_{data} - \langle \mathbf{v} \rangle_{recon} \tag{5}$$

$$\nabla \mathbf{b}^h = \langle \mathbf{h} \rangle_{data} - \langle \mathbf{h} \rangle_{recon}, \tag{6}$$

where $\langle . \rangle_{data}$ are expectations with respect to the distribution of data and $\langle . \rangle_{recon}$ are expectations with respect to the distribution of reconstructed data. The reconstructed data is obtained by starting with a data vector on visible variables, and alternatively sampling from distribution $p(\mathbf{h}|\mathbf{v})$ and then $p(\mathbf{v}|\mathbf{h})$ (Equation 3).

In the context of Image tag suggestion we use RBM as a generative model which captures dependencies between the semantic tags - each visible variable from $\mathbf{v}$ indicates presence of a semantic class in an image.

3 CONDITIONAL RESTRICTED BOLTZMANN MACHINE

Conditional Restricted Boltzmann Machine [8] (CRBM) is an extension of RBM which models joint distribution of $\mathbf{v}$ and $\mathbf{h}$ conditioned on data $\mathbf{c}$ - $p(\mathbf{v},\mathbf{h}|\mathbf{c})$. In CRBM, the Equations 3 are extended by computing the biases as

$$\mathbf{b}^v = \mathbf{A}\mathbf{c} + \mathbf{a} \text{ and } \mathbf{b}^h = \mathbf{B}\mathbf{c} + \mathbf{b}. \tag{7}$$

In CRBM, the CD gradient of W is still computed according to Equation 4. The other gradients are:

$$\begin{aligned}
\nabla A &= \langle \mathbf{v}\mathbf{c} \rangle_{data} - \langle \mathbf{v}\mathbf{c} \rangle_{recon} && (8)\\
\nabla B &= \langle \mathbf{h}\mathbf{c} \rangle_{data} - \langle \mathbf{h}\mathbf{c} \rangle_{recon} && (9)\\
\nabla \mathbf{a} &= \langle \mathbf{v} \rangle_{data} - \langle \mathbf{v} \rangle_{recon} && (10)\\
\nabla \mathbf{b} &= \langle \mathbf{h} \rangle_{data} - \langle \mathbf{h} \rangle_{recon}. && (11)
\end{aligned}$$

For tag suggestion, the conditioning data $\mathbf{c}$ are content-based features extracted from an image or other media.

4 HANDLING UNOBSERVED VISIBLE DATA

In the context of image tag suggestion, the task of RBM and CRBM is to provide marginal probabilities of unobserved tags which constitute the visible variables $\mathbf{v}$ as more and more tags become observed (by actions of a user). Due to a large number of possible tags (hundreds or thousands), it is not possible to obtain a large training dataset where presence or absence of all tags for all images would be known. Such dataset has to have sparse annotations and the learning algorithm has to handle the unobserved tags.

Estimation of probabilities of unobserved visible variables can be achieved by using *Gibbs sampling* to draw several samples from the RBM distribution and computing means of marginal distributions $E(p(v_i))$ using the samples. Gibbs sampling starts by assigning random values to unobserved variables and a sample is obtained by iterating between computing $p(\mathbf{h}|\mathbf{v})$ (Equation 3) and sampling from it, followed by computing $p(\mathbf{v}|\mathbf{h})$.

Several methods for handling missing training data in the context of RBM were proposed. Single missing value can be easily filled by sampling from its exact conditional distribution (it is known for single unobserved variable). More missing values can be treated in the same way as other parameter [3] if they are updated often during learning. This approach is efficient only on training sets of limited size. Salakhundinov et al. [6] introduce a radical way of dealing with missing values by using RBM's with different numbers of visible units for different training cases. This approach is able to handle very sparse data, but it no longer produces a single RBM model.

In our work, we decided to use Gibbs sampling to fill the unobserved values in the training data. For the CD gradients (Equation 4), the data means $\langle . \rangle_{data}$ have to be computed. This can be done by drawing samples from the distribution of the unobserved visible variables conditioned on the observed visible variables. This distribution is not known during learning of the RBM model. However, current imperfect RBM model can be used instead as an approximation. When a sample from the distribution of the visible data is obtained, the CD algorithm proceeds exactly as described in Section 2 and Section 3.

5 EXPERIMENTS AND RESULTS

We tested the proposed approach on a training dataset for semantic indexing task from TRECVID 2011 evaluations. The dataset consists of 400 hours of video from which over 260 thousand images

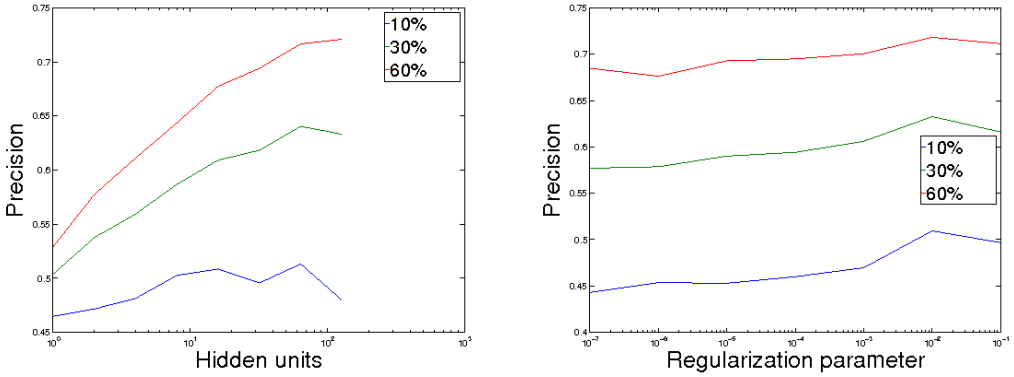

Figure 2: Dependency of average Precision of tag suggestion on number of hidden units (left) and on regularization weight decay parameter (higher values equal to stronger regularization).

(key-frames) were extracted. 345 semantic classes were annotated by active learning [1]. Total 14M shots-level annotations were collected (approximately 16 %), from which only 400 thousand are positive. On average, there is over 1100 positive and 42 thousand negative annotations for each class. Examples of the classes are Actor, Airplane Flying, Bicycling, Canoe, Doorway, Ground Vehicles, Stadium, Tennis, Armed Person, Door Opening, George Bush, Military Buildings, Researcher, Synthetic Images, Underwater and Violent Action.

The conditional content-based features are Bag-Of-Visual-Words representations (BOW [5]). The particular feature extraction method for BOW consists of dense sampling, RGB-SIFT descriptor, and soft assignment to BOW [4]. The dimensionality of the content-based features is 4098.

For the experiments, the TRECVID dataset was divided into two parts. First 200 thousand key-frames were used for training and from the remaining 60 thousand key-frames 20 thousand were randomly selected for testing. All tests were performed for probabilities 10 %, 30 % and 60 % that the annotated tags are known - known tags were sampled randomly for each key-frame. Note that even for 60 %, only small number of tags per key-frame are known due to sparse initial annotation. Average precision (across all tags) is used as an evaluation measure.

The first experiment explored the effect of dimension of the hidden layer (results shown in Figure 2). The optimal strength of L2 regularization (*weight decay*) was selected by grid search using cross-validation. The training process iterated twenty times over the training set, and the marginal probabilities of tags were estimated using fifty samples. The results show that there is a positive correlation between the dimension of the hidden layer and average precision in all percentages of known annotations (more notably for higher percentages). A turning point in this correlation can be seen at 64 hidden units. This can be interpreted as a saturation point and adding more hidden units would result in lower precisions due to overfitting. Most importantly, the results show that the proposed approach can utilize the information provided by the known tags, and that the average precision significantly improves with the number of known tags.

The second test measured dependency of precision of CRBM tag suggestion on a regularization parameter for the conditional data (regularization of A and B from Equation 7). The *weight decay* parameter was set to 0.00055 and the dimension of hidden layer was set to 64 according to results of the first experiment. Figure 2 shows that the highest precisions were achieved when the weight decay parameter for conditional data was set to the highest value 0.1 (when the contribution of the conditional data was reduced the most). In fact, the best achieved results are worse compared to simple RBM. The reason could be that the large number of CRBM parameters (namely matrix $\mathbf{A}$ and $\mathbf{B}$) can

not be reliably fitted using the CD algorithm when large number of training data is missing.

6 CONCLUSIONS

We have proposed an approach which uses RBM to model dependencies among binary variables in semantic tagging and possibly similar tasks. We have shown that RBMs can be efficiently learned by filling unobserved data using Gibbs sampling even when large portion of the training data is missing. Further, we have proposed to use Gibbs sampling to infer marginal probabilities of unobserved data variables in order to predict presence of tags based on other known tags. The results show that the predictions become more certain with higher portion of known tags. The RBM with the proposed inference could be used to suggest semantic tags for image annotation, or in active learning frameworks.

In addition, we have shown how CRBM can integrate content-based features and tag dependencies in a single probabilistic model. However, the experiments suggest that CRBM is not able to utilize the BOW content-based features in the semantic image tagging task. One possible reason is the high number of parameters in the model. Reducing dimensionality of the content-based features could improve results as it would reduce the number of model parameters. Alternatively, logistic regression could be used to pre-train the conditional part of CRBM. Deep belief network [3] could be constructed to improve performance over the one layer RBM.

ACKNOWLEDGEMENT

This work has been supported by EU-7FP-IST - Decipher and BUT FIT grant No. FIT-11-S-2 and EU-7FP-IST - GLOCAL - EEU - 248984 and the Research and Development Council of the Czech Republic - CEZ MŠMT, MSM0021630528

REFERENCES

[1] Stéphane Ayache and Georges Quénot. Evaluation of active learning strategies for video indexing. *Image Commun.*, 22(7-8):692–704, August 2007.

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

[3] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[4] Michal Hradis, Ivo Reznicek, and Kamil Behu. Semantic Class Detectors in Video Genre Recognition. In *Proceedings of VISAPP 2012*, page 7, 2012.

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR'06*, pages 2169–2178. IEEE, 2006.

[6] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted Boltzmann machines for collaborative filtering. *Proceedings of the ICML'07*, pp:791–798, 2007.

[7] Cees G M Snoek et al. The MediaMill TRECVID 2010 Semantic Video Search Engine. In *TRECVID 2010: Participant Notebook Papers and Slides*, 2010.

[8] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Two Distributed-State Models For Generating High-Dimensional Time Series. *The Journal of Machine Learning Research*, 12:1025–1025–1068–1068, February 2011.