Document downloaded from:

http://hdl.handle.net/10251/54425

This paper must be cited as:

Mansanet Sandin, J.; Albiol Colomer, A.; Paredes Palacios, R.; Villegas, M.; Albiol Colomer, AJ. (2014). Restricted Boltzmann Machines for Gender Classification. Lecture Notes in Computer Science. 8814:274-281. doi:10.1007/978-3-319-11758-4_30.

The final publication is available at

http://dx.doi.org/10.1007/978-3-319-11758-4_30

Copyright Springer Verlag (Germany)

Additional Information

Restricted Boltzmann Machines for Gender Classification

Jordi Mansanet[1], Alberto Albiol[1], Roberto Paredes[2], Mauricio Villegas[2], and Antonio Albiol[1]

[1] iTEAM - Instituto de Telecomunicaciones y Aplicaciones Multimedia
[2] PRHLT Research Centre
Universitat Politècnica de València
Valencia Spain

Abstract. This paper deals with automatic feature learning using Gaussian Restricted Boltzmann Machines (GRBM) for the problem of gender recognition in face images. The GRBM is presented together with some practical learning tricks to improve the learning capabilities and speedup the training process. The performance of the features obtained is compared against several linear methods using the same dataset and the same evaluation protocol. The results show a classification accuracy improvement compared with classical linear projection methods. Moreover, in order to increase even more the classification accuracy, we have run some experiments where an SVM is fed with the non-linear mapping obtained by the GRBM in a tandem configuration.

Keywords: Representation learning, RBM, gender classification

1 Introduction

Gender recognition of face images is an important task in computer vision as many applications depend on the correct gender assessment. Examples of applications of gender recognition include visual surveillance, marketing, intelligent user interfaces, demographic studies, etc.

There exist many approaches in the literature that deal with the problem of gender recognition [18]. In most cases, the first stage of gender recognition is to extract a set of handcrafted features, such as Haar [13], LBP [16], IDP [17], that are fed into a suitable classifier. The problem of this paradigm is that it is based on the expertise of the researcher to find the best feature set for a given problem. For this reason, representation learning emerged as a promising research field. The main goal of representation learning is to automatically convert data into a form that makes it easier to extract useful information when building classifiers [1]. The success of representation learning will be the key to board complex problems in the future.

Classical methods for representation learning were usually focused on dimensionality reduction techniques that preserve the representation capability (principal component analysis, independent component analysis, etc). When class

information was available, classical techniques focused on obtaining discriminative features (discriminant analysis) as well as a reduction of dimensionality. All these techniques have been widely used because of their simplicity and effectiveness [3].

In this paper, we propose the use of a powerful generative graphical model called Restricted Boltzman Machine (RBM) for feature learning. Recently, RBMs has become very popular for its success in an impressive variety of applications [7] [15] [5]. RBMs model non-linear statistical dependencies of observed variables by introducing binary latent variables. These latent variables are assumed to be independent given the observed variables. Although, the idea of extracting independent features is also common to other algorithms (such as PCA), the main contribution of RBMs is that its non linear nature is able to find more complex relations between input variables. Also, another important difference is that the number of learned binary features will be much higher than the number of learned features extracted using PCA or LDA.

To our knowledge, this is the first paper that analyzes the performance of RBMs applied to gender recognition. Moreover, we will discuss some practical issues that illustrate how to train the RBMs in a practical application.

The remainder of the paper is organized as follows. Section 2 describes the RBM's and the main notation used throughout the paper. In section 3 and 4 we describe the dataset used and the set of experiments carried out. The final section draws the conclusions and directions for future research.

2 Restricted Boltzmann machine

2.1 Generative models

A Restricted Boltzman Machine (RBM) is an stochastic generative model that can learn probability distributions over its inputs. This generative model can be implemented as a neural network where the set of inputs is called "visible" layer and this visible layer is connected to a set of "hidden" units. Every RBM is characterized by an energy model function that assigns low energy values to high probability samples. The standard type of RBM uses binary visible and hidden units. The problem of using binary visible units in RBM is that they are not appropriate for real-valued data, such as pixel intensities in images. To deal with this situation, a new model called Gaussian RBM (GRBM) [12] is defined. In this case, the energy function is defined as:

$$E_{GRBM}(\mathbf{v},\mathbf{h}) = \sum_{i \in vis} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in hid} b_j h_j - \sum_{i,j} v_i h_j w_{ij}/\sigma_i \qquad (1)$$

where v_i denotes the *real-valued* activity of visible unit i, σ_i is its corresponding standard deviation and h_j is the binary state of hidden unit j. The parameters of the model are the biases a_i,b_j and the weights w_{ij} that connect visible and hidden units.

A nice property of GRBMs is that the hidden units are mutually independent given the visible units and vice-versa. This is a consequence of the lack of intra-layer connections. Therefore, the conditional distribution over the hidden units can be factorized given the visible units:

$$p(h_j = 1|\mathbf{v}) = \frac{1}{1 + exp(-\sum_i w_{ij} v_i/\sigma_i - b_j)} \tag{2}$$

Likewise, the conditional distribution over the visible units given the hidden units also factorizes:

$$p(v_i|\mathbf{h}) = \mathcal{N}(v_i|\mu_i, \sigma_i^2) \tag{3}$$

where $\mu_i = a_i + \sigma_i^2 \sum_j w_{ij} h_j$. The previous equation is important because it shows explicitly the Gaussian nature of the visible units.

During the training process, the parameters of the model are adjusted so that the log-likelihood of the training data is maximized using stochastic gradient descent. It is important to note that the log-likelihood definition does not depend on the labels of samples, so the training process of the GRBM model is completely unsupervised.

The derivative of the log probability with respect to the weights leads to a very simple weight update rule:

$$\Delta w_{ij} = \epsilon \left(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \right) \tag{4}$$

where ϵ is a learning rate and the angle brackets are used to denote expectations under the distribution specified by the subscript that follows. A simplified version of the same learning rule is used for the biases. It is important to mention that to accelerate the learning process it is essential to approximate the unbiased samples of $\langle v_i h_j \rangle_{model}$ using Contrastive Divergence (CD) [10].

Although the RBM is a powerful model, it is possible to improve its performance and speed up the learning procedure [11]. One common trick to increase de speed of learning is to use the momentum method that takes into account the update rule from the previous state $\Delta w_{ij}^{(t-1)}$. Weight-decay is another trick that usually improves the performance of the RBMs. The main reason is that improves generalization to new data by reducing overfitting to the training data, but also makes the receptive fields of the hidden units smoother and more interpretable by shrinking useless weights. The simplest form of weight-decay, called L_2, adds an extra term to the normal gradient that penalizes large weights. Finally, encouraging sparse hidden activities it is important to easily interpret the function of each hidden unit. This trick can be achieved by adding a penalty term that fixes a "sparsity target" which is the desirable probability of being active. Also, discriminative performance is sometimes improved by using features that are only rarely active [14]. In the results section, we will show the effect of these tricks on the classification performance.

3 Dataset

Although there are several works on gender recognition of human face images [3] [8], there is no standard database or protocol for experimentation in this task.

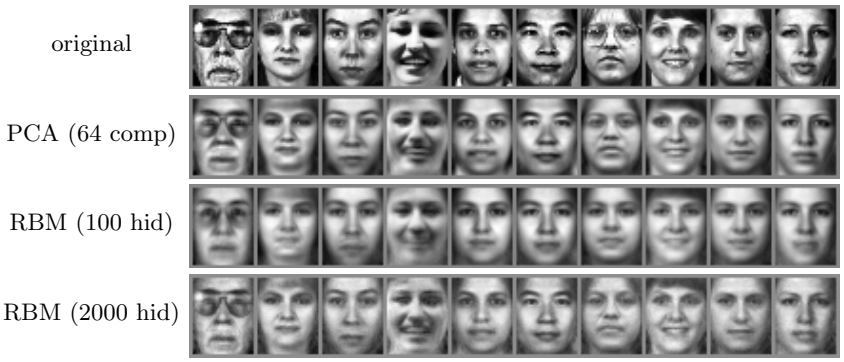

Fig. 1: Original images and reconstructions for different models

LFW (Labeled Faces in the Wild) was compiled to aid the study of unconstrained face recognition. The dataset contains faces that show a large range of variation typically encountered in everyday life, exhibiting natural variability in factors such as pose, lighting, race, accessories, occlusions, and background [6]. The problem of LFW is that number of males is much higher than the number of females, with some individual having appearing more than once.

In many datasets, the images are not annotated with gender information. Therefore researchers had to manually label the ground truth using visual inspection, either by themselves or with the help of others. Also, it is very important to take special care that any person appears in both training and test sets to prevent that the classifier learns the identity instead of the gender.

As a conclusion, no large, publicly available dataset specifically designed for the problem of face gender recognition has been established. For our experiments we have taken a set of 1892 images (946 males and 946 females) from many public face databases (FERET, BANCA, FRGC, AR ...) using the first frontal view from each subject only. For the details about the composition of the dataset see [18].

4 Experiments

We have carried out three different sets of experiments. First we have run experiments in order to assess the gender classification performance of GRBM w.r.t. the number of hidden units and the application of sparsity and regularization terms. Second, we have compared these results with those from [18] where different linear methods are applied to the same dataset and the same evaluation protocol. Finally, in order to increase even more the classification accuracy we run some experiments where an SVM is fed with the non-linear mapping obtained by the GRBM's in a *tandem* configuration.

In general, in all the experiments the GRBM's were trained using the $CD-1$ algorithm for 100 epochs using the training set, without the class-labels infor-

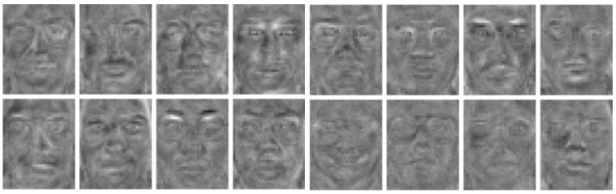

Fig. 2: Examples of features learned by the RBM model using 2000 hidden units

mation, i.e. unsupervised learning. The weights of the GRBM's were initialized with small random values sampled from a normal distribution with zero mean and standard deviation of 0.05. The learning rate value was set to 0.001 for both weights and biases. Optionally, in some experiments, we have applied a "sparsity target" in the binary hidden units and a weight decay term, as it is explained in 2. The sparsity target was fixed to 0.01 and L_2 regularization was used as a weight decay term.

4.1 Gaussian RBM

In order to evaluate the behaviour of the GRBM we have run experiments varying the number of hidden units from 100 to 2000. Note that the GRBM is an unsupervised technique that leads to a non-linear mapping of the original representation space. Figure 2 shows a few examples of the type of features learned by the RBM model using 2000 hidden units. This features correspond to some weight vectors $\mathbf{w_j}$ associated to the hidden units. Note that these features might not be orthogonal as in the case of PCA. Moreover a sigmoid function is applied to the result of the projection $\mathbf{v} \cdot \mathbf{w_j}$ obtaining a non-linear mapping of $\mathbf{v}$. Another important difference w.r.t. PCA projection is that the GRBM features are more spatially localized so that each feature explains a part of the input sample.

To visually asses the quality of the non-linear mapping obtained by the GRBM, Figure 1 shows a few examples and the corresponding reconstructions using PCA and GRBM with different number of hidden units. It can be seen that the quality of the reconstruction using PCA is very good using only 64 principal components. In the case of GRBMs each hidden unit carries exactly one bit of information due to the saturation produced by the sigmoid function, for this reason the number of hidden units required to capture the input information must be much higher than in the case of PCA. Another interesting result is that the information about the gender (and identity) is lost in some cases for low number of hidden units. This fact explains the poor results obtained for the GRBM when the number of hidden units is too low.

A quantitative assessment of the performance of the GRBM for gender classification is carried out by means of adding a discriminative layer (a linear classifier) after the GRBM layer. This is the standard procedure using GRBM's for

Regularization	Number of hidden units			
	100	500	1000	2000
None	14.3 ± 2.3	11.3 ± 1.2	10.6 ±1.4	**10.1** ± 1.1
sparsity $+L_2$ reg	14.2 ± 2.7	11.1 ± 2.1	10.2 ± 1.4	10.3 ± 1.0

Table 1: Face gender recognition results depending on the number of hidden units in the GRBM

classification. This discriminative layer is trained using supervised data. However it is important to note that the non-linear projection is learned from unsupervised data, normally easier to obtain and leading to very large training sets, while for the discriminative layer we can use smaller datasets, even different. Table 1 shows the gender classification results of the GRBM w.r.t the number of hidden units and the application of sparsity and regularization. Normally this sparsity and regularization are used to improve the results but for the 2000 hidden units the best results is obtained without sparsity and regularization.

4.2 GRBM as a non-linear projection technique

In this section, we aim at comparing the classification performance of GRBM's versus other projection methods. For the other projection methods a k-NN classifier was used in order to provide a classification. In each case, the corresponding algorithm parameters were properly adjusted, and only the best result obtained is shown for each algorithm.

We propose to compare the GRBM's performance with the following well-known linear mappings: Locality Preserving Projections (SLPP) [9], Locality Sensitive Discriminant Analysis (LSDA) [4] and Non-parametric Discriminant Analysis (NDA) [2].

Essentially, we want to test whether the non-linear mapping of the GRBM's together with a plain (linear) discriminative layer provides any benefit in front of a linear projection mapping and a non-linear classifier (k-NN). Table 2 shows the results of GRBM using 2000 hidden units. In general the linear techniques tend to work bad handling the original high dimensional space (except PCA), making these techniques inadequate for high-dimensional problems. Note that in [18] these linear techniques worked better using a previous PCA. However the GRBM despite of being a non-linear mapping is able to manage adequately the original high dimensionality representation and to obtain an adequate mapping, from the discriminative point of view.

Technique	PCA	LSDA	SLPP	NDA	GRBM
Error rate(%)	17.7 ± 2.0	35.7 ± 2.6	34.0 ± 2.9	29.6 ± 2.3	**10.1** ± 1.1

Table 2: Face gender recognition results for different projection techniques.

4.3 Tandem classification: GRBM + SVM

In this section we aim at increasing the classification performance of the GRBM-based representation and compare with the state-of-the-art results in the same dataset with the same evaluation protocol. After the unsupervised pre-train, we get a new representation of each sample in the data set, given by its hidden unit outputs after the sigmoid function. This new feature vector (and its label) is used as an input to feed an SVM with a RBF kernel. To set the best parameters of the SVM a grid search over the parameters was performed using a five-fold cross validation set in each subset.

Table 3 shows a comparison where different number of hidden units has been tested and the GRBM results are compared with the LDPP algorithm [18] and with the tandem PCA+SVM as well.

Technique	Error rate(%)
LDPP	8.5 ± 1.3
PCA+SVM	10.4 ± 1.6
GRBM+SVM	
100 units	11.6 ± 1.9
100 units + L_2 + sparse	11.5 ± 1.9
500 units	8.6 ± 1.5
500 units + L_2 + sparse	8.9 ± 1.4
1000 units	8.4 ± 1.6
1000 units + L_2 + sparse	7.9 ± 1.6
2000 units	8.2 ± 1.9
2000 units + L_2 + sparse	**7.8** ± 1.7

Table 3: Face gender recognition results using SVM

The best results are obtained using a GRBM with 2000 hidden units, better that the LDPP algorithm. Note that in this case, the SVM classification accuracy is higher when the GRBM are trained using sparsity and regularization. Moreover it is important to note the good performance of PCA+SVM but still worse that the LDPP algorithm.

5 Conclusions

This paper presents a new scheme to perform gender classification using a Gaussian Restricted Boltzman Machine as a non-linear feature extractor method. First of all we have carried out a comparison of the GRBM classification performance varying some parameters of the model: number of hidden units, using a sparsity criterion on hidden units and weight decay with L_2 regularization. We have evaluated the performance of the GRBM as a non-linear projection method jointly with a linear classifier. The results show an important improvement compared with a classical linear projection mapping methods (PCA, LSDA, SLPP, NDA) followed by a non-linear classifier (k-NN). Finally, in order to increase even

more the classification accuracy, we have run some experiments where a SVM is fed with the non-linear mapping obtained by the GRBM in a tandem configuration. This model outperforms the best gender classification performance published with this database.

Future research will be focused on the use of deep architectures based on stacking GRBM as a pre-training for the entire network. Usually this deep models are able to yield more abstract (and useful) representations.

References

1. Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *PAMI, IEEE Trans. on*, 35(8):1798–1828, 2013.
2. M. Bressan and J. Vitrià. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 24(15):2743–2749, 2003.
3. S. et al. Buchala. Dimensionality reduction of face images for gender classification. *Intelligent Systems, Proceedings.*, 1:88–93 Vol.1, 2004.
4. Deng Cai, Xiaofei He, Yuxiao Hu, Jiawei Han, and T. Huang. Learning a spatially smooth subspace for face recognition. *CVPR*, pages 1–7, 2007.
5. Aarron Courville, James Bergstra, and Yoshua Bengio. Unsupervised models of images by spike-and-slab rbms. In *ICML*, pages 1145–1152, 2011.
6. Gary B. Huang et. al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, Univ. of Massachusetts, October 2007.
7. Tanya Schmah et al. Generative versus discriminative training of rbms for classification of fmri images. In *NIPS*, pages 1409–1416. 2008.
8. Arnulf B. A. Graf and Felix A. Wichmann. Gender classification of human faces. In *BMCV*, pages 491–500, London, UK, 2002.
9. Xiaofei He and Partha Niyogi. Locality preserving projections. In *NIPS*. 2004.
10. G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002.
11. G. E. Hinton. A practical guide to training restricted boltzmann machines. Technical report, University of Toronto, 2010.
12. G. E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
13. B. Moghaddam and Ming-Hsuan Yang. Learning gender with support faces. *PAMI, IEEE Trans. on*, 24(5):707–711, 2002.
14. V. Nair and G. E. Hinton. 3d object recognition with deep belief nets. In *NIPS*, pages 1339–1347, 2009.
15. Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *ICML*, pages 791–798, 2007.
16. Caifeng Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4):431 – 437, 2012.
17. A. Shobeirinejad and Yongsheng Gao. Gender classification using interlaced derivative patterns. In *ICPR*, pages 1509–1512, 2010.
18. Mauricio Villegas and Roberto Paredes. Dimensionality reduction by minimizing nearest-neighbor classification error. *Pattern Recognition Letters*, 32(4):633 – 639, 2011.