Restricted Boltzmann Machines

Chaimaa Kadaoui chaimaa.kadaoui*
Othman SBAI othman.sbai*
Xi Shen xi.shen*

1 Abstract

For our project in Probabilistic Graphical Models course, we studied the use and the training of Restricted Boltzmann Machines as an instance of undirected graphical models. We mainly base our approach on the practical guide for training Restricted Boltzmann Machines (Hinton 2010 [4]). We start by an introduction to RBM and the interest that they have raised in the last years, then we define the theoretical aspects of main training methods of RBM. Finally we present our results by comparing our implementation of training RBM to an implementation from another deep learning library, on two popular image datasets MNIST and CIFAR-10. We show our convergence results and investigate the influence of different parameters of the training algorithm.

2 Introduction

2.1 Definition of an RBM

A Restricted Boltzmann Machine (RBM) is an undirected energy-based probabilistic graphical model which can be seen as a two layer neural network (visible and hidden units). RBMs can be used to model and learn important aspects of a probability distribution of a training data. They are called restricted, because we impose restrictions on the network topology, by allowing only connections between units of different layers. RBM are energy-based, since they define probability distribution through an energy function. Learning corresponds to shaping the energy so that desirable configurations have a low energy and thus maximize probability of training data under the model. Maximum likelihood learning is challenging for undirected graphical models because MLE parameters cannot be found analytically and the log likelihood gradient based optimization is not tractable. This optimization requires obtaining samples through Markov Chain Monte Carlo, which is computationally demanding.

2.2 Interest in RBM

A major interest have been dedicated to directed graphical models with one layer of observed variables and one or more layers of hidden variables such Mixture of Gaussians, probabilistic PCA, factor analysis, latent dirichlet analysis LDA... However, the undirected two layered analogue which was first introduced and called Harmonium by Smolensky P. [9] and renamed as "Restricted Boltzmann Machines" by Hinton G. [5] have also seen advances through new training methods, which unlike directed graphical models, makes inference and learning very fast.

2.3 Applications of RBM

Restricted Boltzmann machines have successfully been applied to problems involving high dimensional data such as images, text [6], and as unsupervised feature extractors [11].

Recent developments have demonstrated the capacity of RBM to be powerful generative models, able to extract useful features from input data or construct deep artificial neural networks. In this setting

*mail extension @ens-paris-saclay.fr for all authors

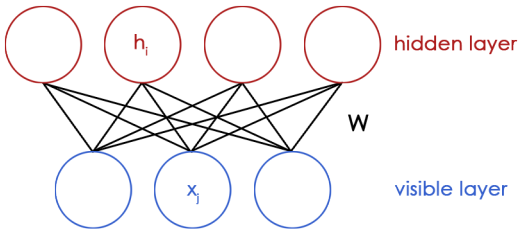

Figure 1: The network graph of an RBM

they can be seen as good preprocessing or initialization for some other models. They can also be used as self-contained classifiers as shown in [8].

- **RBM as a generative model**: learns and generates samples from data distribution
- **RBM as feature extractor**: unsupervised feature extraction from data, instead of handcrafting features
- **RBM for computer vision**: such as object recognition, image denoising and inpainting.
- **RBM for collaborative filtering**: given a set of N users and M movies, recommend movies to the users.

3 Theory of RBM

We can represent the graphical model of RBM with the schema in the figure 1.

Each node of one layer is connected to all the other nodes of the other layer ($h_j \leftrightarrow x_k$). The matrix W characterizes the connections between the two layers: $W_{j,k}$ is the value for the connection between h_j and x_k. For simplification, we suppose that the variables $\mathbf{x}$ and $\mathbf{h}$ are binary.

3.1 Energy and probability

We introduce the two bias vectors c and b to define the energy of this graph and the joint probability of $\mathbf{x}$ and $\mathbf{h}$:

$$E(x,h) = -h^\top W x - c^\top x - b^\top h \quad , \quad p(x,h) = \frac{1}{Z}\exp(-E(x,h))$$

Z is the partition parameter which can be computed by calculating all the possibles values of $\mathbf{x}$ and $\mathbf{h}$. In practice, this parameter is intractable.

We see that to increase the probability, we want to decrease the energy. We have:

- if $c_k < 0$ and $x_k = 1$, the energy is higher; therefore we prefer having $x_k = 0$ (the probability of $x_k = 1$ decreases).
- if $c_k > 0$, similarly, we prefer $x_k = 1$ over $x_k = 0$.

We can dedude similar intepretations for b and W.

3.2 Inference: Conditional inference

Because $p(x)$ is intractable, the only possible type of inference is computing conditional inference. We will prove that computing the conditional probabilities $p(x|h)$ and $p(h|x)$ is easy.

Because of the construction of the graph, the variables $h_1, \ldots, h_H$ are independant conditionally on $\mathbf{x}$:

$$p(h|x) = \prod_j p(h_j|x)$$

We can also prove that given $\mathbf{x}$, h_j follows a Bernoulli:

$$p(h_j = 1|x) = sigm(b_j + W_{j.}x) \qquad \text{with} \qquad sigm(x) = \frac{1}{1+e^{-x}}$$

We can prove that either by derivation of $p(h|x)$ or by applying the *Local Markov Property*: $p(z_i|z_j,\ j \neq i) = p(z_i|\mathcal{N}(z_i))$ where $\mathcal{N}(z_i)$ are the neighbors of z_i.

3.3 Free energy

To have a better understanding of the parameters $W,\ c,\ b$, let's introduce the concept of free energy. Using a marginalization and some derivations, we can write the probability $p(x)$ as :

$$p(x) = \sum_h p(x,h) = \exp\left(c^\top x + \sum_{j=1}^{H} \log\left(1 + \exp(b_j + W_{j.}x)\right)\right)/Z$$

The function $u \mapsto \log(1+\exp(u))$ is called `softplus` and is a soft approximation of $u \mapsto \max(u, 0)$.

We can see that $p(x) = \exp(-F(x))/Z$ where F is called the free energy. We want to increase the probability of our data x and thus decreases its energy. Let's look at each term of this energy:

- As seen in the previous subsection (3.1), the sign of c_k impacts wether x_k would be a 0 or 1. Thus, having $c^\top x$ in the probability means that we want $\mathbf{x}$ and the bias c to be aligned. This bias characterizes the probability.
- Similarly, $\mathbf{x}$ and $W_{j.}$ should align for each j. And we can see b_j as a control parameter: $W_{j.}x$ has to be big enough to have $b_j + W_{j.}x > 0$. Therefore, we can consider the hidden variables as features and b_j the bias for each feature (it defines the importance of each feature). The goal of the RBM is to represent meaningful features.

4 Contrastive divergence method for training RBM efficiently

To train the RBM with our data $\mathbf{x}^{(t)}$, we would like to minimize the average loss:

$$\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log p(\mathbf{x}^{(t)})$$

We want to apply a stochastic gradient descent. We derive each term of the sum with respect to our model parameter θ as follow (where $\mathbb{E}_h$ and $\mathbb{E}_{x,h}$ are expectations of h and (x,h) respectively):

$$\frac{\partial - \log p(\mathbf{x}^{(t)})}{\partial\theta} = \mathbb{E}_h\left[\frac{\partial E(\mathbf{x}^{(t)}, h)}{\partial\theta}|\mathbf{x}^{(t)}\right] - \mathbb{E}_{x,h}\left[\frac{\partial E(\mathbf{x}, h)}{\partial\theta}\right]$$

The first term is called the *positive phase* and the second term the *negative phase*. Because of the difficulty to compute the seconde term, we will use an algorithm called *Contrastive Divergence*. The idea is:

1. We estimate the expectation $\mathbb{E}_{x,h}$ by sampling a single point $\tilde{\mathbf{x}}$.
2. To do so, we use Gibbs sampling in chain (we apply it k times).
3. We initialize our Gibbs sampling with $\mathbf{x^{(t)}}$.

Therefore we can rewrite each term as:

$$\mathbb{E}_h\left[\frac{\partial E(\mathbf{x}^{(t)}, h)}{\partial\theta}|\mathbf{x}^{(t)}\right] \simeq \frac{\partial E(\mathbf{x}^{(t)}, h(\mathbf{x}^{(t)}))}{\partial\theta} \quad , \quad \mathbb{E}_{x,h}\left[\frac{\partial E(\mathbf{x}, h)}{\partial\theta}\right] \simeq \frac{\partial E(\tilde{\mathbf{x}}, h(\tilde{\mathbf{x}}))}{\partial\theta}$$

The goal is to minimize the energy of our training data and to increase the energy of our samplings at each iteration until we have samplings similar to our training data.

We define for a given x, the vector $h(x)$ as:

$$h(x) = \begin{bmatrix} p(h_1 = 1|x) \\ \vdots \\ p(h_H = 1|x) \end{bmatrix} \quad , \quad h(x) = sigm(b + Wx)$$

Looking at the derivative of the energy w.r.t the parameters $W,\ b,\ c$ we can find their updates:

$$W = W + \alpha \left(h(\mathbf{x^{(t)}}) \mathbf{x^{(t)}}^\top - h(\tilde{\mathbf{x}}) \tilde{\mathbf{x}}^\top \right)$$

$$b = b + \alpha \left(h(\mathbf{x^{(t)}}) - h(\tilde{\mathbf{x}}) \right) \qquad , \qquad c = c + \alpha \left(\mathbf{x^{(t)}} - \tilde{\mathbf{x}} \right)$$

5 Implementation and results

We implemented the RBM training algorithm in python (Github repository available at [2]) and then we compared our implementation performance with an existing deep learning library using tensorflow (YADLT [1]), we applied it to the MNIST dataset (images of handwritten digits) and CIFAR-10 (color images of different classes). In the case of image data, each pixel is considered as a node; the intensity $p \in [0, 1]$ of a pixel is considered as a probability.

5.1 Implementation of RBM in python

The implementation of RBM training consists of implementing the constrastive divergence learning strategy through Gibbs sampling. We only needed few steps of Contrastrive divergence for a good training. We compare the difference between using different values of k steps for CD-k algorithm 5.4.3. We also used mini batches during training in order to compute the approximate gradient in every step. The influence of the batch size is studied in subsection 5.4.1.

We show the images reconstructed by the learned model starting from samples from each dataset in the figure 5.

5.2 Implementation of RBM in tensorflow using a GPU

We use the implementation in tensorflow and python provided by Gabriele Angeletti as a library gathering many deep learning models and algorithms such as convolutional networks, RNN, RBM, Deep Belief Networks... We observe the difference in training time between the two implementations. The difference is not very important because of small batch size, and small matrices size in multiplication [2]: 25 seconds for training 1 epoch on GPU with batch size of 10 versus 35 to 40 seconds with our CPU implementation using numpy.

We show the filters obtained after 20 epochs in the figure 6

5.3 Convergence plot

In general, it is difficult to directly evaluate the performance of RBM. This is due to the fact that the parameter Z is intractable and we are unable to compute the log-likelihood. One way to monitor the behavior of the algorithm is to calculate at each step the average stochastic reconstruction which corresponds to the euclidean distance between the samples generated with Gibbs sampling and our data: $|\mathbf{x} - \mathbf{x}'|$. When performing the training we make sure that we do not overfit by checking that the validation error is continuously decreasing for different values of the parameters studied.

[2] https://github.com/sbaio/Restricted-Boltzmann-Machine

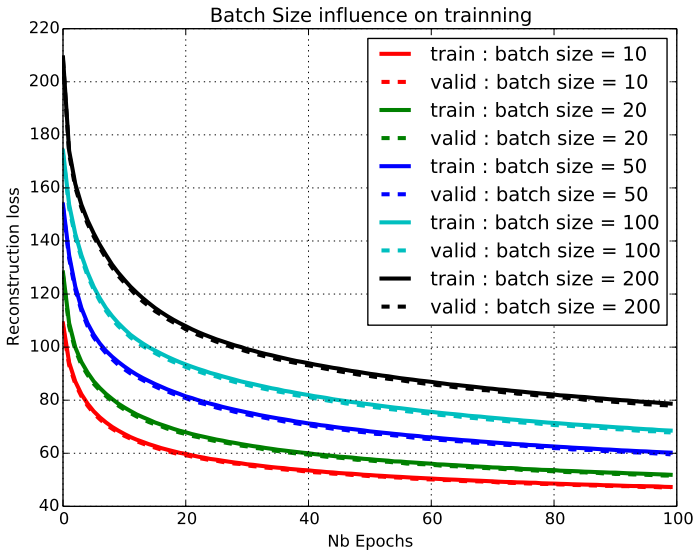

Figure 2: $|\mathbf{x} - \mathbf{x}'|$ for different batch sizes

5.4 Influence of different algorithm parameters

5.4.1 Influence of batch size

The figure 2 shows the performance of RBM with different batch sizes. We can see that we have better results for smaller batch sizes. In fact, even though increasing the batch size leads to better gradient estimate, it tends to decrease the weight updates per gradient evaluation. For example, for 500 images, at each epoch we would have 25 updates for a batch size of 20 while we would have only 5 updates for a batch size of 100. On the other side, the computation time is more important with small batch sizes.

5.4.2 Influence of number of hidden units

As we can see in the figure 3, the number of hidden units has a direct impact on the performance of the algorithm. By increasing the number of hidden units, we can decrease the average stochastic reconstruction. In the figure, we can see that clearly 100 hidden units isn't enough; however for bigger values the curves become closer.

Since we deal with high dimensional data (images of abouts 1000 pixels each), we don't risk overfitting with a number of hidden units between 100 and 500 (the valiation error keep decreasing in each case). Hence, our main constraint for the choice of this parameter was the time complexity. We found that 200 or 300 was a good compromise between computing time and performances.

5.4.3 Influence of number of steps in contrastive divergence algorithm

5.4.4 Influence of the steps of Gibbs sampling

In general, using more steps for the Gibbs sampling doesn't influence much the behavior of the algorithm. The figure 4 summarizes our results for different choices of this parameter. We can see in the zoom that we have better performances for greater number of steps but the improvement is negligible.

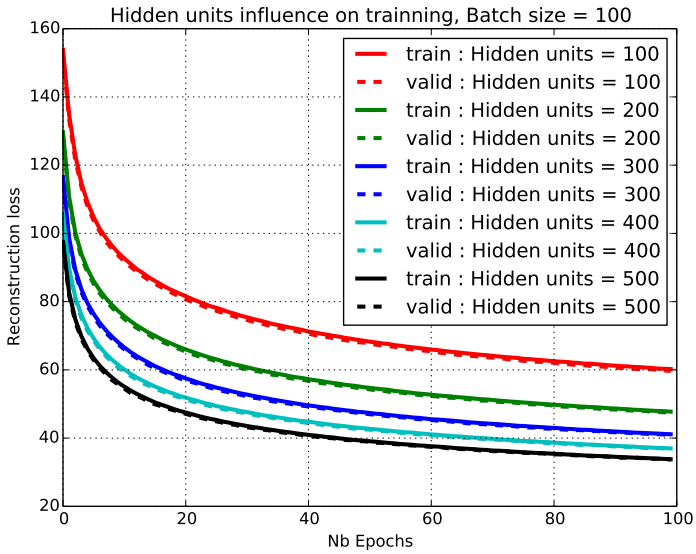

Figure 3: $|\mathbf{x} - \mathbf{x}'|$ for different number of hidden units

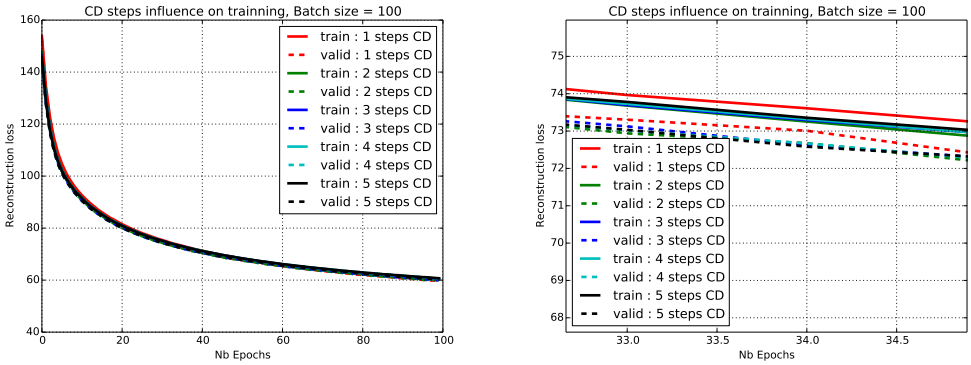

Figure 4: $|\mathbf{x} - \mathbf{x}'|$ for different number of Gibbs samling's steps. The right figure is a zoom of the left one.

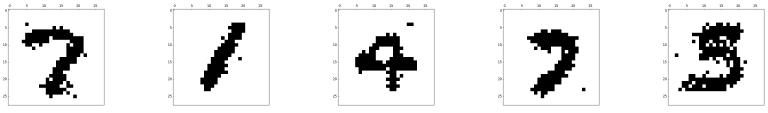

Figure 5: Generated samples after 20 epochs

Figure 6: Filters obtained after 20 epochs: Left for MNIST dataset, right for CIFAR-10 dataset.

6 Conclusion and future work

As for other deep learning techniques, RBM implementations are costly. Hence, the use of GPU achitectures is necessary for fast computation time. For high dimensional data such as images, we can avoid overfitting by choosing optimal hyperparameters. Moreover, even for 20 epochs, we could extract meaningful features from the algorithm which can serve for classification for example. By stacking RBMs into Deep Belief Nets, as specified in [6], we can benefit from the feature extraction capability at different levels and achieve even better classification results.

There are many alternatives to the RBM training techniques studied in this project. For example, we can use Persistent Contrastive Divergence [4] where a set of samples drawn from the model distribution is maintained at each update of the model, it works generally better than CD-1.

Furthermore, it is possible to adapt the RBM model for different types of data. In the case of multi-class data, instead of a Bernoulli, the data can follow a Multinomial probability. For unbounded reals, we can add a quadratic term ($\frac{1}{2}\mathbf{x}^\top\mathbf{x}$) to the energy. Then, $p(\mathbf{x}|h)$ becomes a Gaussian distribution.

References

[1] Gabriele Angeletti. Deep-learning-tensorflow. `https://github.com/blackecho/Deep-Learning-TensorFlow`, 2015.

[2] Kayvon Fatahalian, Jeremy Sugerman, and Pat Hanrahan. Understanding the efficiency of gpu algorithms for matrix-matrix multiplication. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*, pages 133–137. ACM, 2004.

[3] Asja Fischer and Christian Igel. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

[4] Geoffrey Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010.

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

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

[7] Hugo Larochelle. Neural networks: Restricted boltzmann machine.

[8] Hugo Larochelle, Michael Mandel, Razvan Pascanu, and Yoshua Bengio. Learning algorithms for the classification restricted boltzmann machine. *Journal of Machine Learning Research*, 13(Mar):643–669, 2012.

[9] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, DTIC Document, 1986.

[10] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008.

[11] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.