On the Training Algorithms for Restricted Boltzmann Machines

Leandro Aparecido Passos*, João Paulo Papa
Department of Computing
São Paulo State University
Bauru, Brazil
E-mail:{leandro.passos, joao.papa}@unesp.br

***Abstract*—Deep learning techniques have been studied extensively in the last years due to their good results related to essential tasks on a large range of applications, such as speech and face recognition, as well as object classification. Restrict Boltzmann Machines (RBMs) are among the most employed techniques, which are energy-based stochastic neural networks composed of two layers of neurons whose objective is to estimate the connection weights between them. Recently, the scientific community spent much effort on sampling methods since the effectiveness of RBMs is directly related to the success of such a process. Thereby, this work contributes to studies concerning different training algorithms for RBMs, as well as its variants Deep Belief Networks and Deep Boltzmann Machines. Further, the work covers the application of meta-heuristic methods concerning a proper fine-tune of these techniques. Moreover, the validation of the model is presented in the context of image reconstruction and unsupervised feature learning. In general, we present different approaches to training these techniques, as well as the evaluation of meta-heuristic methods for fine-tuning parameters, and its main contributions are: (i) temperature parameter introduction in DBM formulation, (ii) DBM using adaptive temperature, (iii) DBM meta-parameter optimization through meta-heuristic techniques, and (iv) infinity Restricted Boltzmann Machine (iRBM) meta-parameters optimization through meta-heuristic techniques.**

***Index Terms*—Machine Learning; Restricted Boltzmann Machine; Optimization**

I. Introduction

In the last decades, machine learning techniques have grown exponentially in a wide range of applications, mainly the ones regarding decision-making procedures. Such tasks are of extreme interest in environments that involve large amounts of data, such as automated diagnosis, image and video processing, and data mining, to cite a few.

Usually, the traditional data flow employed to "solve" machine learning-related problems tend to follow four main steps: (i) data processing, (ii) feature extraction, (iii) feature selection/transformation, and (iv) pattern recognition. Although each of the aforementioned steps had evolved in the last decades, a new set of techniques based on deep learning (DL) strategies provide an approach that mimics the brain-behavior while processing visual information, where the data extraction is performed on distinct layers, when each one is responsible for extracting different types of information.

*Ph.D. Thesis

Restricted Boltzmann Machines (RBMs) [1] are classified as stochastic neural networks composed of a set of "hidden" or latent units employed to encode a representation of input data. Roughly speaking, RBMs are not considered a DL method, though their "stacking" process is. In a nutshell, RBMs are used as building blocks for deep learning models, such as the well-known Deep Belief Network (DBNs) [2] and the Deep Boltzmann Machines(DBMs) [3].

One of the major constraints regarding RBMs stands on the training step, which can be interpreted as an optimization problem where the minimization of the system's energy implies directly in an increase of the posterior probability of activating a hidden neuron. Such assumption led many studies towards a more efficient manner of solving this optimization problem and to approximate the output to the log-likelihood, which is considered the "perfect result"; however intractable when the number of variables is relatively large. Since the number of visible units generally stand for the number of pixels when dealing with image problems, the number of visible units tends to be large enough to convert such log-likelihood approximation into a prohibitive task.

Recently, many works addressed the task of modeling such log-likelihood approximation as a sampling over a Markov chain [4]–[8], where the initial solution, i.e., the input model, stands for some data sample, as well as the expected output stands for the corresponding sample approximation. Such a process is then repeated over the training dataset until some stopping criterion is met.

The hypothesis and main contributions of the present thesis concern answering the following question: which strategies could one adopt towards enhancing the training process of RBM-based models? Two approaches are proposed to accomplish such task: (i) the application of meta-heuristic optimization algorithms to fine-tune hyper-parameters, and (ii) the introduction of the temperature parameter into the DBM-based formulation.

The works presented in the next sections aim towards the optimization of Restricted Boltzmann Machines-based machine learning algorithms. The proposed approaches employ meta-heuristic techniques for such tasks, as well as an approximation of the computational formulation to the original Boltzmann formulation by introducing the temperature parameter in the DBM domain.

Section II presents a brief referential background regarding RBMs, DBNs, DBMs, and infinity Restricted Boltzmann Machines (iRBMs). The temperature meta-parameter is introduced for the very first time into the DBM formulation in the paper presented in Section III. A continuation of this work is provided in Section IV. The paper presented in Section V introduces the problem of DBMs meta-parameter fine-tuning aided by meta-heuristic optimization techniques. Following the same idea, the work presented in Section VI introduces a similar approach for meta-parameter optimization regarding the ordered Restricted Boltzmann Machines (oRBM) and iRBM domains. Finally, Section VII presents a continuation of the work presented in Section VI, applying iRBM for Barret's Esophagus lesions detection. Finally, Section VIII provides the conclusions and the main contributions of this work.

II. Theoretical Background

This chapter presents the theoretical background regarding RBM-based models.

A. Restricted Boltzmann Machines

Invented under the name "Harmonium" by Paul Smolensky in 1986, [9] and renamed in the mid-2000s by Geoffrey Hinton, Restricted Boltzmann Machines are energy-based stochastic neural networks composed of two layers of neurons (visible and hidden), in which the learning phase is conducted by means of an unsupervised fashion. A naïve architecture of a Restricted Boltzmann Machine comprises a visible layer $\mathbf{v}$ with m units and a hidden layer $\mathbf{h}$ with n units. Additionally, a real-valued matrix $\mathbf{W}_{m \times n}$ models the weights between the visible and hidden neurons, where w_{ij} stands for the weight between the visible unit v_i and the hidden unit h_j. Figure 1 depicts the RBM architecture.

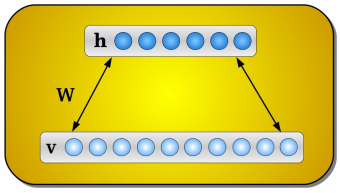


Fig. 1. The RBM architecture.

Let us assume both $\mathbf{v}$ and $\mathbf{h}$ as being binary-valued units. In other words, $\mathbf{v} \in \{0,1\}^m$ e $\mathbf{h} \in \{0,1\}^n$. The learning process is conducted using the minimization of the systems energy, analogous to the Maxwell-Boltzmann distribution law of thermodynamics. The energy function of a Restricted Boltzmann Machine is given by:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m} \sum_{j=1}^{n} v_i h_j w_{ij}, \quad (1)$$

where $\mathbf{a}$ e $\mathbf{b}$ stand for the biases of the visible and hidden units, respectively.

Since the RBM is a bipartite graph, the probabilities of activating both visible and hidden units are mutually independent, thus leading to the following conditional probabilities:

$$P(v_i = 1|\mathbf{h}) = \phi\left(\sum_{j=1}^{n} w_{ij} h_j + a_i\right), \quad (2)$$

and

$$P(h_j = 1|\mathbf{v}) = \phi\left(\sum_{i=1}^{m} w_{ij} v_i + b_j\right). \quad (3)$$

Note that $\phi(\cdot)$ stands for the logistic-sigmoid function. One can solve the aforementioned equation using Contrastive Divergence [4], for instance.

B. Deep Belief Networks

Deep Belief Network [2] is a generative graphical model composed of multiple layers of latent variables ("hidden units"), with connections between the layers but not between units within each layer. In a nutshell, DBNs are composed of a set of stacked RBMs, being each of them trained using the same learning algorithm of RBMs, but in a greedy fashion, which means an RBM at a certain layer does not consider others during its learning procedure. In this case, we have a DBN composed of L layers, being $\mathbf{W}^i$ the weight matrix of the RBM at layer i. Additionally, we can observe the hidden units at layer i become the input units to the layer $i+1$. Figure 2 depicts the model.

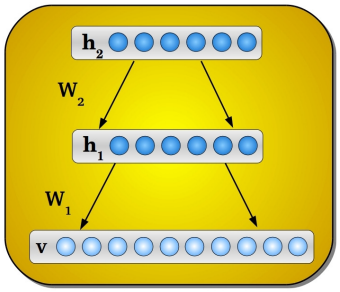


Fig. 2. The DBN architecture.

C. Deep Boltzmann Machines

The DBM formulation is rather similar to the DBN one, but with some slightly differences. Suppose we have a DBM with two layers, where $\mathbf{v}$ stand for the visible units, as well as $\mathbf{h}^1$ and $\mathbf{h}^2$ stand for the hidden units at the first and second layer, respectively. Figure 3 depicts the architecture of a standard DBM, which formulation is slightly different from a DBN one.

The energy of a DBM can be computed as follows:

$$E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2) = -\sum_{i=1}^{m^1} \sum_{j=1}^{n^1} v_i h_j^1 w_{ij}^1 - \sum_{i=1}^{m^2} \sum_{j=1}^{n^2} h_i^1 h_j^2 w_{ij}^2, \quad (4)$$

where m^1 and m^2 stand for the number of visible units in the first and second layers, respectively, and n^1 and n^2 stand for the number of hidden units in the first and second layers,

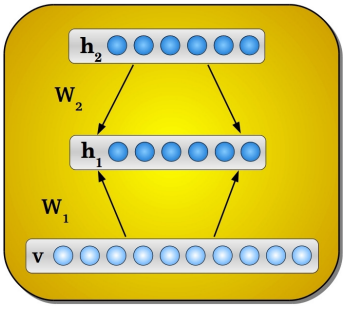


Fig. 3. The DBM architecture with two hidden layers.

respectively. In addition, we have the weight matrices $\mathbf{W}^1_{m^1 \times n^1}$ and $\mathbf{W}^2_{m^2 \times n^2}$, which encode the weights of the connections between vectors $\mathbf{v}$ and $\mathbf{h}^1$, and vectors $\mathbf{h}^1$ and $\mathbf{h}^2$, respectively. For the sake of simplification, we dropped the bias terms out.

D. Infinity Restricted Boltzmann Machines

The Infinity Restricted Boltzmann Machine is a variant of the RBM such that the hidden units are trained sequentially, from the left to the right, where the maximum number of hidden units is not specified. This number increases automatically until its capacity is sufficiently high, which is possible by taking the limit of $n \rightarrow \infty$. The model is presented in Figure 4.

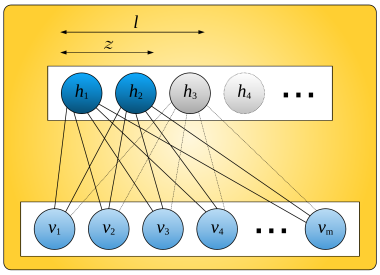


Fig. 4. An iRBM with $z = 2$ trained units. There are some non-zero (dashed lines) values connecting the third unit ($l = 3$) that is going to be used for training. All remaining hidden units (i.e., $l > 3$) have zero-valued weights.

E. Sampling Methods

Initially, the strategy adopted to estimate $E[\mathbf{hv}]^{model}$, which is the representation of the data learned by the system, is basically to start the visible units with random values and run alternating Gibbs chain until equilibrium, (i.e., convergence). However, this approach is computationally expensive, since a good model is obtained when the number of Gibbs steps $k \rightarrow \infty$. Figure 5 depicts the model.

To tackle the aforementioned problem, some alternatives to Gibbs sampling were presented in the following years. The next sections discuss some of the most used techniques for such purpose.

1) Contrastive Divergence: Basically, the idea is to initialize the visible units with a training sample, to compute the states of the hidden units using Equation 3, and then to compute the states of the visible unit (reconstruction step)

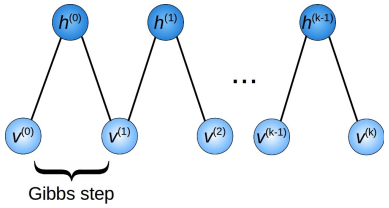


Fig. 5. Gibbs sampling.

using Equation 2. In short, this is equivalent to perform Gibbs sampling using $k = 1$ and initializing the chain with the the training samples.

Based on the above assumption, we can now compute $E[\mathbf{hv}]^{model}$ as follows:

$$E[\mathbf{hv}]^{model} = P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T, \tag{5}$$

where $\tilde{\mathbf{v}}$ stands for the reconstruction of the visible layer given $\mathbf{h}$, and $\tilde{\mathbf{h}}$ denotes a estimation of the hidden vector $\mathbf{h}$ given $\tilde{\mathbf{v}}$.

Therefore, the equation below leads to a simple learning rule for updating the weight matrix $\mathbf{W}$, as follows:

$$\begin{aligned} \mathbf{W}^{t+1} &= \mathbf{W}^t + \eta(E[\mathbf{hv}]^{data} - E[\mathbf{hv}]^{model}) \\ &= \mathbf{W}^t + \eta(P(\mathbf{h}|\mathbf{v})\mathbf{v}^T - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T), \end{aligned} \tag{6}$$

where $\mathbf{W}^t$ stands for the weight matrix at time step t, and η corresponds to the learning rate. Additionally, we have the following formulae to update the biases of the visible and hidden units:

$$\begin{aligned} \mathbf{a}^{t+1} &= \mathbf{a}^t + \eta(\mathbf{v} - E[\mathbf{v}]^{model}) \\ &= \mathbf{a}^t + \eta(\mathbf{v} - \tilde{\mathbf{v}}), \end{aligned} \tag{7}$$

and

$$\begin{aligned} \mathbf{b}^{t+1} &= \mathbf{b}^t + \eta(E[\mathbf{h}]^{data} - E[\mathbf{h}]^{model}) \\ &= \mathbf{b}^t + \eta(P(\mathbf{h}|\mathbf{v}) - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})), \end{aligned} \tag{8}$$

where $\mathbf{a}^t$ and $\mathbf{b}^t$ stand for the visible and hidden units biases at time step t, respectively. In short, Equations 6, 7 and 8 are the standard formulation for updating the RBM parameters.

Later on, Hinton [10] introduced a weight decay parameter λ, which penalizes weights with large magnitude, as well as a momentum parameter α to control possible oscillations during the learning process. Therefore, we can rewrite Equations 6, 7 and 8 as follows:

$$\mathbf{W}^{t+1} = \mathbf{W}^t + \underbrace{\eta(P(\mathbf{h}|\mathbf{v})\mathbf{v}^T - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T) - \lambda\mathbf{W}^t + \alpha\Delta\mathbf{W}^{t-1}}_{=\Delta\mathbf{W}^t}, \tag{9}$$

$$\mathbf{a}^{t+1} = \mathbf{a}^t + \underbrace{\eta(\mathbf{v} - \tilde{\mathbf{v}}) + \alpha\Delta\mathbf{a}^{t-1}}_{=\Delta\mathbf{a}^t} \tag{10}$$

and

$$\mathbf{b}^{t+1} = \mathbf{b}^t + \underbrace{\eta(P(\mathbf{h}|\mathbf{v}) - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})) + \alpha\Delta\mathbf{b}^{t-1}}_{=\Delta\mathbf{b}^t}. \quad (11)$$

2) Persistent Contrastive Divergence: Most of the issues related to Contrastive Divergence approach are related to the number of iterations employed to approximate the model to the real data. Although the approach proposed by Hinton [4] takes $k = 1$ and works well for real world problems, one can settle different values for k [11][1].

Notwithstanding contrastive divergence provides a good approximation to the likelihood gradient, i.e., it provides a good approximation of the model to the data when $k \rightarrow \infty$. However, its convergence might become poor when the Markov chain has a "low mixing". Furthermore, contrastive divergence has a good convergence only in the early iterations, getting slower as iterations go by, thus demanding the use of a parameter decay term (as shown in equations 9, 10 and 11, for instance).

Therefore, an interesting alternative for contrastive divergence would be using higher values for k, usually named CD-k. However, a major problem related to this approach is due to its computational burden, since a greater number of iterations are required to approximate the model to the data. Given such premise, Tieleman [5] proposed the Persistent Contrastive Divergence - PCD for short - which aims to approximate the model to CD-k, but with a lower computational burden. The idea is quite simple: on CD-1, each training sample is employed to start an RBM and rebuilds a model after a single Gibbs sampling iteration. Once every training sample is presented to the RBM, we have a so-called "epoch". The process is repeated for each next epoch, i.e., the same training samples are used to feed the RBM and the Markov chain is restarted at each epoch. PCD aims to achieve an "ideal" approximation of the model to the data given CD-k (when $k \rightarrow \infty$) by means of not restarting the Markov chain, but using the model built in the former epoch to feed the RBM in the current epoch. Therefore, as the number of epochs increases, the model tends to be similar to the one obtained through CD-k. The only problem related to this technique concerns the number of epochs demanded for convergence, but yet the reconstruction error rate is generally still lower than CD.

III. Temperature-Based Deep Boltzmann Machines

This section presents the content published in the journal Neural Processing Letters [12], and it introduces the concept of temperature in DBMs, which play a key role in Boltzmann-related distributions, but it has never been considered in this context up to date. Therefore, the main contribution of this work is to take into account this information, as well as the impact of replacing a standard Sigmoid function by another one and to evaluate their influence in DBMs considering the task of binary image reconstruction. Its impact is evaluated through the learning steps, and the results are compared even with distinct activation functions, once such parameter added to the energy function can be interpreted as a scalar multiplication of the Sigmoid function input. Provided results confirm the hypothesis suggested by Li et al. [13] that lower temperatures tend to reach more accurate results, as presented in Table I. Furthermore, one can observe that lower temperatures also support sparseness representations of the hidden layer, which leads to a dropout like regularization.

	0.1	0.2	0.5	0.8	1.0	1.5	2.0	Gompertz
DBM-CD	**0.18518**	**0.18503**	**0.18504**	0.19087	0.19718	0.21495	0.21591	0.26833
DBM-PCD	**0.18527**	**0.18606**	**0.18655**	0.19154	0.19735	0.21423	0.21532	0.27248
DBN-CD	0.21613	0.21977	0.21814	0.21465	0.21352	0.21725	0.22455	0.22142
DBN-PCD	0.21051	0.21155	0.21660	0.21104	0.21012	0.21080	0.21431	0.21617

TABLE I
AVERAGE MSE OVER THE TEST SET CONSIDERING SEMEION HANDWRITTEN DIGIT DATASET.

IV. Deep Boltzmann Machines Using Adaptive Temperatures

Section IV is continuity of the work started in Section III. Here, one can observe the behavior of DBMs under adaptive temperatures. The work was presented in the 17th International Conference on Computer Analysis of Images and Patterns [14] and proposes an adaptive temperature that increases smoothly while the training progresses. Such approach can be compared to the behavior observed in meta-heuristic algorithms, where each agent initially explores the search space in the quest for better solutions, and later converges to the points whose results are more promising as training advances. In a nutshell, the training procedure of such models concerns the minimization of the energy of each training sample in order to increase its probability. Therefore, such an optimization process needs to be regularized in order to reach the best trade-off between exploitation and exploration. The idea is to provide an adaptive regularization approach based on temperatures which implies advantages considering Deep Belief Networks and Deep Boltzmann Machines. The main contribution of the work is the exemption of the task of fine-tuning the temperature parameter, providing a friendly interface for less experienced users. Additionally, it presents results at least competitive with the ones where the temperature is fine-tuned in the context of binary image reconstruction, thus outperforming temperature-fixed DBNs and DBMs, as presented in Table II.

	0.1	0.5	0.8	1.0	1.5	2.0	Linear	Curve
DBM-CD	0.16048	0.16048	0.16049	0.16048	0.16049	0.15983	**0.15822**	0.16053
DBM-PCD	0.16049	0.16050	0.16048	0.16049	0.16049	0.15983	**0.15929**	0.16039
DBN-CD	0.16049	0.16049	0.16050	0.16049	0.16249	0.17040	**0.15822**	0.16523
DBN-PCD	0.16048	0.16049	0.16049	0.16048	0.16081	0.16120	**0.15929**	0.16321

TABLE II
AVERAGE DBM/DBN MSE OVER THE TEST SET CONSIDERING CALTECH 101 SILHOUETTES DATASET WITH 200 ITERATIONS.

The impact of adaptive temperatures during the convergence process is depicted in Figure 6, where the MSE of the first layer during the learning process converged faster during the

[1]Usually, contrastive divergence with a single iteration is called CD-1.

first 50 iterations, and they did not get stuck in local optima, as one can observe in the experiment with the fixed temperature.

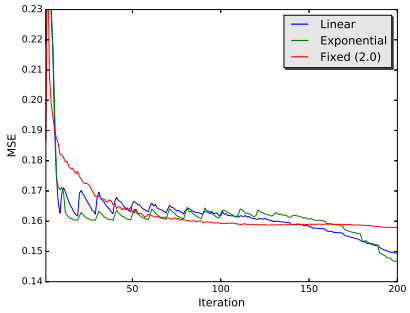


Fig. 6. MSE during the learning step of the first layer considering Caltech 101 Silhouettes dataset for DBM.

V. A Metaheuristic-Driven Approach to Fine-Tune Deep Boltzmann Machines

This section presents a paper accepted in the Applied Soft Computing journal. It introduces the problem of DBMs meta-parameter fine-tuning aided by meta-heuristic optimization techniques, since one of the main shortcomings of these techniques involves the choice of their hyperparameters, which have a significant impact on the final results. The work addresses the issue using metaheuristic optimization techniques with different backgrounds, such as swarm intelligence and memory- and evolutionary-based approaches, i.e., IHS, AIWPSO, CS, FA, BSA, JADE, and CoBiDE, as well as a random search. Experiments conducted in three public datasets for binary image reconstruction showed that metaheuristic techniques can obtain reasonable results. DBM's performance is compared against the DBN, outperforming the results of the latter in two out of three datasets.

Table III presents the average values of the minimum squared error over the MNIST dataset considering DBM, being the values in bold the best results considering the Wilcoxon signed-rank test. One can observe the metaheuristic techniques obtained the best results, with special attention to IHS, JADE, and CoBiDE. Also, one can not figure a considerable difference between shallow and deep models.

	1L		2L		3L	
	CD	PCD	CD	PCD	CD	PCD
IHS	**0.08744**	0.08766	0.08761	0.08761	0.08760	0.08761
AIWPSO	0.08765	0.08771	0.08762	0.08761	**0.08759**	0.08760
CS	0.08767	0.08770	0.08760	0.08760	0.08762	0.08761
FA	0.08766	0.08762	0.08761	0.08763	0.08761	0.08761
BSA	0.08774	0.08766	0.08761	0.08762	0.08762	0.08762
JADE	**0.08754**	**0.08749**	0.08761	0.08761	0.08761	0.08761
CoBiDE	**0.08757**	0.08765	0.08762	0.08760	0.08761	0.08760
RS	0.08780	0.08782	0.08761	0.08760	0.08761	0.08761

TABLE III
AVERAGE MSE VALUES CONSIDERING MNIST DATASET.

VI. Fine-Tuning Infinite Restricted Boltzmann Machines

One of the main concerns about RBMs is related to the number of hidden units, which is application-dependent. Infinite RBM was proposed as an alternative to the regular RBM, where the number of units in the hidden layer grows as long as it is necessary, dropping out the need for selecting a proper number of hidden units [15]. However, a less sensitive regularization parameter is introduced as well.

The paper proposed in this section was published in the 30th Conference on Graphics, Patterns, and Images [16], and it follows the idea developed in Section V, now applied in the infinite Restricted Boltzmann Machine domain. It proposes to fine-tune iRBM hyper-parameters using meta-heuristic techniques such as Particle Swarm Optimization, Bat Algorithm, Cuckoo Search, and the Firefly Algorithm. The main objective of iRBM is to ease the proper selection of its meta-parameters, setting automatically the number of hidden units that best fit the model. The proposed approach is validated in the context of binary image reconstruction over two well-known datasets, i.e., MNIST and Silhouettes Datasets. Furthermore, the experimental results compare the robustness of the iRBM against the RBM and Ordered RBM (oRBM) using two different learning algorithms, showing the suitability in using meta-heuristics for hyper-parameter fine-tuning in RBM-based models. Table IV presents the average NLL results concerning Caltech 101 Silhouettes dataset. The iRBM achieved the best results with all meta-heuristic techniques using CD for learning, except for CS. Additionally, oRBM obtained the best results with the FA algorithm. Actually, iRBM trained with CD and optimized by FA achieved the best result so far.

	RBM		oRBM		iRBM	
	CD	PCD	CD	PCD	CD	PCD
RS	384.30±29.94	432.38±140.15	267.42±28.39	386.03±94.26	274.36±33.99	424.30±187.62
BA	292.08±77.24	609.27±170.72	243.72±24.93	458.95±216.99	**229.32±32.14**	593.33±229.98
CS	349.60±47.13	455.83±104.28	267.82±29.60	448.20±126.03	255.15±18.67	579.29±254.97
FA	279.88±57.13	629.06±170.37	**237.85±23.63**	420.77±163.16	**218.36±28.54**	486.86±110.73
PSO	315.42±85.29	599.11±140.47	240.40±26.29	411.74±66.69	**237.83±37.83**	554.60±254.15

TABLE IV
AVERAGE NLL VALUES CONSIDERING CALTECH 101 SILHOUETTES DATASET.

VII. Barrett's Esophagus Analysis Using Infinity Restricted Boltzmann Machines

This chapter presents the paper entitled "Barretts Esophagus Analysis Using Infinity Restricted Boltzmann Machines", accepted in the Journal of Visual Communication and Image Representation as an extension from the idea presented in [16] applied to medical issues.

Since the number of patients with Barret's esophagus (BE) has increased in the last decades, and considering the dangerousness of the disease and its evolution to adenocarcinoma, an early diagnosis of BE may provide a high probability of cancer remission. However, limitations regarding traditional methods of detection and management of BE demand alternative solutions. As such, computer-aided tools have been recently used to assist in this problem, but the challenge still persists.

To manage the problem, we introduce the infinity Restricted Boltzmann Machines to the task of automatic identification of Barrett's esophagus from endoscopic images of the lower esophagus. Moreover, since iRBM requires a proper selection of its meta-parameters, we also present a discriminative iRBM fine-tuning using six meta-heuristic optimization techniques. We showed that iRBMs are suitable for the context since it provides competitive results, as well as the meta-heuristic techniques showed to be appropriate for such a task. Considering the very best results obtained for all the techniques, Table VI presents the sensitivity (SE) and the specificity (SP) results. Notice the best values are in bold. values are in bold.

	Accuracy	Sensitivity	Specificity
iRBM-FA	**66.35%**	**0.644**	0.687
SVM-RBF	65.60%	0.612	**0.706**
SVM-Linear	58.60%	0.582	0.593
Bayes	59.98%	0.593	0.605

TABLE VI
MEAN SE AND SP VALUES FOR THE SELECTED BEST RESULTS OBTAINED USING DICTIONARIES OF 500 WORDS.

VIII. CONCLUSIONS

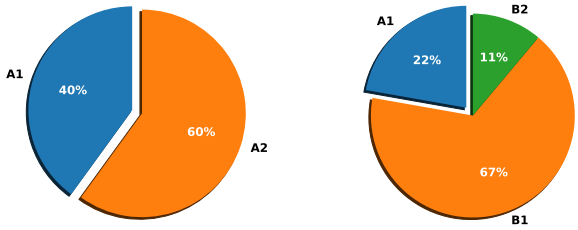
The present thesis was organized into eight sections, described as follows: the introduction exposed the context of the research, as well as the motivation and main contribution to the proposed subject, while Section II briefly presented the theoretical background regarding the objective of the research. Section III and IV presented works published in the journal Neural Processing Letters (NPL) [12] entitled "Temperature-Based Deep Boltzmann Machines", as well as the paper "Deep Boltzmann Machines Using Adaptive Temperatures", presented at the 17th International Conference on Computer Analysis of Images and Patterns (CAIP) [14], respectively. The former introduced the temperature parameter into the DBM formulation, while the latter proposed to use the previously mentioned parameter in an adaptive fashion.

Section V presented the work accepted in the journal Applied Soft Computing (ASoC), which introduced the concepts of meta-heuristic parameters optimization into the DBM domain. Similarly, Section VI employed the idea to the Infinity Restricted Boltzmann Machine (iRBM) context on a paper presented at the 30th Conference on Graphics, Patterns and Images (SIBGRAPI) [16]. Moreover, Section VII applied iRBM for Barret's Esophagus lesions detection. The latter was published in the Journal of Visual Communication and Image Representation (JVCIR) as an invited extension from [16].

The results obtained in the aforementioned sections confirm the hypothesis of this works, evidencing that both the application of meta-heuristic optimization algorithms to fine-tune the hyper-parameters, as well as the introduction of the temperature parameter into the RBM-based formulation are suitable strategies concerning the enhancement of RBM-based models training process.

A. Publications

Table V presents a complete list of the works produced during the study period, which is composed of 5 journals, 9 conferences, and one book chapter, denoting a total of 15 papers. Further, Figure 7 depicts the distribution of journals and conferences published in the period distributed by their 'Qualis' status.



(a) (b)

Fig. 7. Distribution of the publications by Qualis: (a) journals and (b) conferences.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and FAPESP/Microsoft grant #2017/25908-6.

Name	Type	Qualis	Year	Status
Learning Parameters in Deep Belief Networks Through Firefly Algorithm [17]	Conference	B2	2016	Published
Deep Boltzmann Machines Using Adaptive Temperatures [14]	Conference	B1	2017	Published
Parkinsons Disease Identification Using Restricted Boltzmann Machines [18]	Conference	B1	2017	Published
Fine-Tuning Infinity Restricted Boltzmann Machines [16]	Conference	B1	2017	Published
A Metaheuristic-Driven Approach to Fine-Tune Deep Boltzmann Machines [19]	Journal	A1	2017	Published
Temperature-based Deep Boltzmann Machines [12]	Journal	A2	2018	Published
Parkinson Disease Identification Using Residual Networks and Optimum-Path Forest [20]	Conference	B1	2018	Published
Enhancing Brain Storm Optimization Through Optimum-Path Forest [21]	Conference	B1	2018	Published
Fine Tuning Deep Boltzmann Machines Through Meta-Heuristic Approaches [22]	Conference	B1	2018	Published
Intelligent Network Security Monitoring based on Optimum-Path Forest Clustering [23]	Journal	A1	2018	Published
Adaptive Improved Flower Pollination Algorithm for Global Optimization [24]	Book Chapter	-	2018	Published
Barrett's Esophagus Analysis Using Infinity Restricted Boltzmann Machines [25]	Journal	A2	2018	Published
Exudate Detection in Fundus Images Using Deeply-learnable Features [26]	Journal	A2	2018	Published
Quaternion-Based Backtracking Search Optimization Algorithm [27]	Conference	A1	2019	Published
κ-Entropy Based Restricted Boltzmann Machines [28]	Conference	A1	2019	Accepted

TABLE V
WORKS DEVELOPED DURING THE STUDY PERIOD

REFERENCES

[1] P. Smolensky, "Parallel distributed processing: Explorations in the microstructure of cognition," D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281.

[2] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[3] R. Salakhutdinov and G. E. Hinton, "An efficient learning procedure for deep boltzmann machines," *Neural Computation*, vol. 24, no. 8, pp. 1967–2006, 2012.

[4] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[5] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 1064–1071.

[6] T. Tieleman and G. E. Hinton, "Using fast weights to improve persistent contrastive divergence," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 1033–1040.

[7] P. Brakel, S. Dieleman, and B. Schrauwen, "Training restricted boltzmann machines with multi-tempering: Harnessing parallelization," in *Artificial Neural Networks and Machine Learning*, ser. Lecture Notes in Computer Science, A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm, Eds. Springer Berlin Heidelberg, 2012, vol. 7553, pp. 92–99.

[8] J. Xu, H. Li, and S. Zhou, "Improving mixing rate with tempered transition for learning restricted boltzmann machines," *Neurocomputing*, vol. 139, pp. 328–335, 2014.

[9] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," DTIC Document, Tech. Rep., 1986.

[10] G. E. Hinton, "Neural networks: Tricks of the trade: Second edition," G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. A Practical Guide to Training Restricted Boltzmann Machines, pp. 599–619.

[11] M. A. Carreira-Perpiñán and G. E. Hinton, "On Contrastive Divergence Learning," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, R. G. Cowell and Z. Ghahramani, Eds. Society for Artificial Intelligence and Statistics, 2005, pp. 33–40.

[12] L. A. Passos and J. P. Papa, "Temperature-based deep boltzmann machines," *Neural Processing Letters*, pp. 1–13, 2017.

[13] G. Li, L. Deng, Y. Xu, C. Wen, W. Wang, J. Pei, and L. Shi, "Temperature based restricted boltzmann machines," *Scientific reports*, vol. 6, 2016.

[14] L. A. Passos, K. A. Costa, and J. P. Papa, "Deep boltzmann machines using adaptive temperatures," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2017, pp. 172–183.

[15] M.-A. Côté and H. Larochelle, "An infinite restricted boltzmann machine," *Neural computation*, 2016.

[16] L. A. Passos and J. P. Papa, "Fine-tuning infinity restricted boltzmann machines," in *Electronic Proceedings of the 30th Conference on Graphics, Patterns and Images (SIBGRAPI'17)*, M. Lage, L. A. F. Fernandes, R. Marroquim, and H. Lopes, Eds., Niteri, RJ, Brazil, october 2017. [Online]. Available: http://sibgrapi2017.ic.uff.br/

[17] G. H. Rosa, J. P. Papa, K. A. P. Costa, L. A. Passos, C. R. Pereira, and X.-S. Yang, *Learning Parameters in Deep Belief Networks Through Firefly Algorithm*. Cham: Springer International Publishing, 2016, pp. 138–149.

[18] C. R. Pereira, L. A. Passos, R. R. Lopes, S. A. Weber, C. Hook, and J. P. Papa, "Parkinsons disease identification using restricted boltzmann machines," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2017, pp. 70–80.

[19] L. A. Passos and J. P. Papa, "A metaheuristic-driven approach to fine-tune deep boltzmann machines," *Applied Soft Computing*, p. 105717, 2019.

[20] L. A. Passos, C. R. Pereira, E. R. Rezende, T. J. Carvalho, S. A. Weber, C. Hook, and J. P. Papa, "Parkinson disease identification using residual networks and optimum-path forest," in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2018, pp. 000 325–000 330.

[21] L. C. Afonso, L. A. Passos, and J. a. P. Papa, "Enhancing brain storm optimization through optimum-path forest," in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2018, pp. 000 183–000 188.

[22] L. A. Passos, D. R. Rodrigues, and J. P. Papa, "Fine tuning deep boltzmann machines through meta-heuristic approaches," in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2018, pp. 000 419–000 424.

[23] R. R. Guimaraes, L. A. Passos, R. Holanda Filho, V. H. C. de Albuquerque, J. J. Rodrigues, M. M. Komarov, and J. P. Papa, "Intelligent network security monitoring based on optimum-path forest clustering," *IEEE Network*, 2018.

[24] D. Rodrigues, G. H. de Rosa, L. A. Passos, and J. P. Papa, "Adaptive improved flower pollination algorithm for global optimization," in *Nature-Inspired Computation in Data Mining and Machine Learning*. Springer, 2020, pp. 1–21.

[25] L. A. Passos, L. A. de Souza Jr, R. Mendel, A. Ebigbo, A. Probst, H. Messmann, C. Palm, and J. P. Papa, "Barretts esophagus analysis using infinity restricted boltzmann machines," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 475–485, 2019.

[26] P. Khojasteh, L. A. P. Júnior, T. Carvalho, E. Rezende, B. Aliahmad, J. P. Papa, and D. K. Kumar, "Exudate detection in fundus images using deeply-learnable features," *Computers in biology and medicine*, vol. 104, pp. 62–69, 2019.

[27] L. A. Passos, D. Rodrigues, and J. P. Papa, "Quaternion-based backtracking search optimization algorithm," in *2019 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2019, pp. 3014–3021.

[28] L. A. Passos, M. C. Santana, T. Moreira, and J. P. Papa, "κ-entropy based restricted boltzmann machines," in *The 2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.