UNSUPERVISED LEARNING WITH MULTI-LAYER BOLTZMANN MACHINES:
A STATISTICAL MECHANICS PERSPECTIVE

Scientific relevance of the project

In the past the generation of data, for example as a result of some experiment, was to test our knowledge and comprehension about the world. Nowadays, we have experience of data generators practically all around us and any kind of data is a priori collected even without a specific goal. For this reason we are witnessing to a change of perspective in which not only much of the information about the world we are living is hidden in the form of data, but also the data itself are becoming part of the world we feel the need to discover.

Data mining is exactly the process of detecting patterns of information in a large dataset, extracting features and coding them into an understandable structure for further use. It involves methods at the intersection of Machine Learning, Statistics and Artificial Intelligence. In recent years we are witnessing a formidably fast progress and Neural networks are playing an important and renewed role in this trend mainly because of the success of Deep Networks [11, 6] in several applications, ranging from engineering and computer science to neuro and computational biology, as for example in computer vision, speech recognition, and natural language processing

Despite their remarkable success, relatively little is understood theoretically about why these techniques are so successful at feature learning and compression. A clear and exhaustive theoretical scaffold is still needed and most of the advances in technology are often achieved by specific home made recipes but without a global and systematic increasing understanding of the methodology.

The Statistical Mechanics perspective is proving to be a fruitful point of view both in formulating general and paradigmatic questions then for providing answers: several problems in statistical inference involve the study of the Statistical Mechanics and the equilibrium properties of disordered Multi-Species systems of spins.

On the opposite point of view, many open problems in Statistical Mechanics of spin glasses and neural networks arise exactly from the study of the so called Non-Convex Multi-Species systems. The statistical inference and data mining perspective stimulates new ideas and suggests different approaches to this mathematical challenge.

Originality and Innovative nature of the project

Deep Neural Networks (DNNs), a set of deep learning algorithms, are biologically-inspired graphical statistical models that consist of multiple layers of neurons, with units in one layer receiving inputs from units in the layer before them. Despite their enormous success, it is still unclear what advantages these deep, multi-layer architectures possess over shallower architectures with a similar number of parameters. In particular, it is still not well understood theoretically why DNNs are so successful at extracting features from structured data.

Training a DNN for a specific task usually needs two different steps of learning: a preliminary preprocessing phase of unsupervised learning, in which parameters are adjusted to get an internal representation of input data, is followed by a supervised process of parameter's

refinement, in which the error between the expected output and the one given by the network is minimized through a gradient descent like procedure. The last step makes use of the feed-forward nature of the network, from the input to the output, thus being task dependent. The former is the problem's core and depends on how much the training input set is representative of the whole dataset as well as on the representational power of the architecture.

Typically the performance of a DNN gets measured after the supervised learning process using cross validation on a given database of inputs/outputs. For this reason most of the results on DNN capabilities and performances are not universal, being task and dataset dependent and oriented to application: here are some architectures performing better on pictures without corners or others that struggle in classifying greyscale objects and so on.

The point of view of this project is on the contrary focused just on the unsupervised learning process, i.e. the ability of these multi-layers architectures to approximate a probability distribution of input data, using an internal representation able to capture different levels of features and patterns correlations. A DNN undressed of the feed-forward interaction among layers becomes simply a probabilistic model with many hidden units called Deep Boltzmann Machine. As a Gaussian Mixture model uses the centers and the variances of a suitable amount of Gaussians to detect and characterize clusters of points, as a Deep Boltzmann Machine uses its parameters to extract hierarchies of features from a set of data.

Aim of this project is twofold. First of all we want to describe what kind of structures and hierarchies of features a Boltzmann Machine is able to represent and thus to extract, focusing on the role of the number of layers and their relative size. Second of all we'll ask which properties a training set must satisfy in order to be representative of the whole dataset, so that feature's detection become statistically possible for the network.

To takle the first problem we will study the equilibrium states of Random Boltzmann Machines with quenched couplings, focusing on the relationship between the number of layers and the shape of the quenched disorder. To give answers to the second question we will set our analysis in some controlled teacher-student scenarios in which one knows in advance the structure of the input data the network is trying to infer.

Both these aims involve the study of the equilibrium Statistical Mechanics of systems with many species (or layers) of units and frustrated interactions and most of the answers can be formulated in terms of phase transitions of their order parameters.

Research methodology and state of the art

Deep Boltzmann Machines are multilayered structures, where each couple of units belonging to consecutive layers are connected by a coupling. No couplings between units of the same layer are present: this property is called non-convexity in the Statistical Mechanics of multi-species systems. Each couple of consecutive layers of which they are composed, define an architecture called *Restricted Boltzmann Machine* (RBM). A RBM mirrors exactly the structure of what is called bipartite spin glass in the context of statistical mechanics of disordered systems. It is a Gibbs probability density over N visible units $\{\sigma_i\}_{i=1}^{N}$ and N_h hidden units $\{h_\mu\}_{\mu=1}^{N_h}$, parameterized by a (real) $N \times N_h$ matrix $\boldsymbol{\xi}$ as

$$P(\boldsymbol{\sigma}, \boldsymbol{h}|\boldsymbol{\xi}) = \frac{P_\sigma(\boldsymbol{\sigma}) P_h(\boldsymbol{h}) e^{\sum_{i=1}^{N}\sum_{\mu=1}^{N_h}\xi_i^\mu \sigma_i h_\mu}}{Z(\boldsymbol{\xi})}, \tag{1}$$

where P_σ and P_h are generic priors and the partition function $Z(\boldsymbol{\xi})$ is a normalisation factor. As proposed by Hinton as a preprocessing learning step for DNN, a RBM and its multi-layers generalizations can be thought as probabilistic models with many hidden units whose parameters $\boldsymbol{\xi}$ can be optimized to mimic a generic probability distribution of input data $P_0(\boldsymbol{\sigma})$, living on the first (visible) layer. To do that one typically tries to minimize the Kullback-Leibler distance $D(P_0(\boldsymbol{\sigma})||P(\boldsymbol{\sigma}|\boldsymbol{\xi}))$. If we assume priors factorize over units we can compute the marginal distribution of a RBM over the visible layer as

$$P(\boldsymbol{\sigma}|\boldsymbol{\xi}) = Z_H^{-1}(\boldsymbol{\xi}) P_\sigma(\boldsymbol{\sigma}) \exp\left(\sum_{\mu=1}^{N_h} u\left(\sum_{i=1}^{N} \xi_i^\mu \sigma_i\right)\right), \tag{2}$$

with $u(x) = \log \mathbb{E}_h e^{xh}$ the cumulant generating function of the hidden unit prior. It is a family of models called *Generalized Hopfield models* (GHM) because they include the Standard Hopfield model [9] when the hidden units are standard gaussians. Hopfield models [1, 2, 3] are recurrent neural networks that have been proposed a long time ago as models of content addressable memories, i.e. systems that are able to retrieve patterns of memory from partial information. Their introduction came from the observation that in large physical systems, interactions between the elementary degrees of freedom are able to generate collective phenomena, such low temperature magnetization in Ising models. Any physical system whose dynamics is dominated by a number of locally stable states can act as a content addressable memory as long as these states can be controlled. In the case of the Hopfield models there exist a region of the phase diagram, the so called retrieval phase, where the stable states are exactly described in terms of the vectors $\vec{\xi}^\mu$ of the interaction: each stable states contain a cluster of configurations correlated with one of these vectors.

The underlying phylosophy of this project is that the performance of the RBMs are connected with the phase diagram of the GHMs: the retrieval phase of the GHM is intimately related to the significance and representational power of the RBM in extracting features and classifying data; the transition from paramagnetic phase to ordered phase in the GHM helps to determine the size of the training set necessary in the RBM for a good estimate of the data distribution.

The phase diagram of the Standard Hopfield model is well known by physicists even if a complete rigorous mathematical theory is still missing, as for the whole class of Non-Convex Multi layers systems. The Parisi theory [14] for the free energy and the structure of the equilibrium states, proved rigorously [7, 17, 16] for convex systems as the celebrated Sherrington-Kirkpatrick model, need to be extended.

Work plan

The project is divided into three main research streams, each of them taking approximatively four months. They are ideally consecutive in the philosophy, from 2 to 3 to many layers analysis, but they are actually independent: the starting of one is not subjected to the success of the previous. In the following they are briefly described, stressing which are the pilot studies and the goals.

Patterns learning through Restricted Boltzmann Machines

The relation between Restricted Boltzmann Machines and Generalized Hopfield networks reveals the kind of correlations a two-layers Machine is able to extract from data and where they are encoded. In the retrieval regime, they are essentially classifiers: they detect directions, i.e. patterns, in the input configurations space to each of which a non-negligible cluster of configurations is correlated, in the spirit of a Principal Component Analysis. The patterns can be read in the interaction matrix between the two layers, each hidden unit in principle can be associated to a different pattern, stored in its couplings.

We want to study the reliability of patterns learning depending on the training set properties. To do that we introduced a teacher-student framework in which the input data (of N units) are generated exactly from a Generalized Hopfield probabiility distribution with P patterns $\boldsymbol{\xi}$ and at a given temperature. The student will train a RBM with N_h hidden units trying to infer the original patterns from the only observation of M input configurations $\{\boldsymbol{\sigma}_b\}_{b=1}^M$. Using a Bayes framework, the posterior of the problem reads as

$$P(\boldsymbol{\xi}|\boldsymbol{\sigma}) = \frac{P_\xi(\boldsymbol{\xi}) \prod_{b=1}^M P(\boldsymbol{\sigma}^b|\boldsymbol{\xi})}{P(\boldsymbol{\sigma})} \propto Z^{-M} \prod_{\mu=1}^P P_\xi(\boldsymbol{\xi}^\mu) \exp\left(\sum_{b=1}^M u\left(\boldsymbol{\sigma}^b \cdot \boldsymbol{\xi}^\mu\right)\right).$$

and we expect a detectability phase transition in terms of $\alpha = M/N$ defining a Bayes optimal threshold α_c. The case of $P = 1$ boolean pattern is relatively simple: the threshold is related to the disorder-order phase transition of a dual Generalized Hopfield model in which the samples $\{\boldsymbol{\sigma}_b\}_{b=1}^M$ plays the role of the patterns of information.

We want to generalize this result to the case $P > 1$ and generic priors. Moreover we want to understand the relation between the threshold α_c on the training set size and the Machine's architecture, i.e. the hidden units priors and the ratio N_h/P between the number of hidden units and the number of planted (by the teacher) patterns.

In this facilitated setting we have also the possibility to compare the different algorithms used to train RBMs, i.e. Contrastive Divergence, Montecarlo based and Belief Propagation based methods, studying their accuracy, efficiency and Bayes optimality.

Pilot studies: I've already studied RBM and GHM with generic priors in [1, 2]. The T-S scenario with $P = 1$ was mentioned by R.Monasson [4, 20] and H.Huang [10]. Works on different learning algorithms are for example [19, 8]

Structured patterns learning through 3-layers Boltzmann Machines

The following step is to understand the meaning of becoming deep. Thus we will study what differs in the learning process when a further layer is added to the machine. The idea is that a deep architectures is able to disentangle different levels of correlations, i.e. patterns of informations composed in turn by the combination of some features, being the superposition of other sub-features and so on.

A model with such a structure of correlation is for example an Hopfield model with combinatorial disorder, i.e. structured patterns, meaning that the patterns ξ_i^μ are not completely independent random vectors but random superposition of a given set of N_f features, i.e.

$\vec{\xi}^{\mu} = \sum_f^{N_f} v_f^{\mu} \vec{u}^f$. It was recently shown by M.Mezard that an Hopfield model with combinatorial disorder is equivalent to a 3-layers Boltzmann Machine, where a further intermediate layer of suitable units $\{t_f\}_{f=1}^{N_f}$ is introduced between the visible and hidden layers, and where u_i^f and v_f^{μ} are respectively the visible/intermediate and intermediate/hidden couplings. From our perspective it means that a 3-layers Boltzmann Machines can work as an higher level patterns classifier, from whose couplings we can read both features and weights from which getting the patterns.

As for the standard RBM it is worth to study the phase diagram of the model, checking if a retrieval phase does really exist and if it's robust as a function of the model's parameters: relative size of the layers, unit's priors and temperature. Then we want to study again the learning reliability from the observation of a training set in a teacher student framework. We will use a Bayes framework to detect again the detectability threshold as a transition from disorder (paramagnetic undetectability phase) to order ("ferromagnetic" detectability phase). Finally we can compare the performance of learning by using a 3-layers respect to a 2-layers Boltzmann Machine in the case of input data sampled from an Hopfield model with planted structured patterns, for which both the probabilistic models are optimal.
Pilot studies: M.Mezard [13] and partial results for non convex multilayers spin-glass are for example the one by Panchenko [15] and myself with collaborators [3]

Deep Boltzmann Machines and Renormalization Group

The relation between Deep Boltzmann Machines and Renormalization Group has been recently discussed. The basic idea of this analogy is that the coarse-grain operated by Kadanoff's RG is very similar to the marginalization over Restricted Boltzmann Machine's hidden units. In order to follow this direction, it is possible to study particular statistical mechanical models where RG can be made exactly and compare the action of RG and RBM-learning. It is the case of spin systems living on a hierarchical lattice, like the diamond lattice introduced by Derrida [5]. We want to study a learning process in which the input data are configurations sampled from the equilibrium distribution of such hierarchical systems. It is again a teacher-student scenario in which we know the structure of the data correlations and we are interested in studying how they are detected and encoded in the network's parameters. Controlling the input data correlations, in terms of an exact RG procedure, allows to follow analytically the unsupervised learning process of parameters optimization.
Pilot studies: The original idea is by P.Mehta and D.J. Schwab [12]. Works on learning thermodynamics systems with RBM are for example by R.Melko [18].

References

[1] A.Barra, G.Genovese, P.Sollich, D.Tantari, *Phase transition in Restricted Boltzmann Machines with generic priors*, arXiv preprint arXiv:1612.03132 (2016)

[2] A.Barra, G.Genovese, P.Sollich, D.Tantari, *Phase diagram of Restricted Boltzmann Machines and Generalised Hopfield networks with generic priors*, arXiv preprint arXiv:1702.05882 (2017)

[3] A. Barra, P.Contucci, E.Mingione, D.Tantari, *Multi-species mean-field spin-glasses. Rigorous results*, Annales Henri Poincaré 16 (2015) 691-708.

[4] S. Cocco, R. Monasson, V. Sessak, *Phys. Rev. E* **83**:051123, (2011).

[5] B.Derrida, L. De Seze, C. Itzykson, *Fractal structure of zeros in hierarchical models*, Journal of Statistical Physics, 1983 - Springer

[6] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, Google book (2016).

[7] F Guerra, *Broken replica symmetry bounds in the mean field spin glass model*, Communications in mathematical physics 233 (1), 1-12

[8] G.E. Hinton, S. Osindero, Y.W. Teh, *A fast algorithm for deep belief nets*, Neural Comp. **18**, 1527-1554, (2006).

[9] J.J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Nat. Acad. Sci. USA **79**, 2554-2558 (1982).

[10] H. Huang, T. Toyoizumi, *Unsupervised feature learning from finite data by message passing: discontinuous versus continuous phase transition*, Physical Review E 94 (6), 062310

[11] Y. LeCun, Y. Bengio, G. Hinton, *Deep learning*, Nature **521**(7553): 436-444, (2015).

[12] P. Mehta, D.J. Schwab , *An exact mapping between the variational renormalization group and deep learning* arXiv preprint arXiv:1410.3831, 2014

[13] M Mezard, *Mean-field message-passing equations in the Hopfield model and its generalizations*, Physical Review E 95 (2), 022117

[14] M. Mézard, G. Parisi, M. A. Virasoro, *Spin glass theory and beyond*, World Scientific, Singapore, (1987).

[15] D Panchenko, *The free energy in a multi-species Sherrington-Kirkpatrick model*, Annals of Probability 43 (6), 3494–3513

[16] D Panchenko, *The Sherrington-Kirkpatrick Model*, Springer-Verlag, New York

[17] M. Talagrand, *Mean Field Models for Spin Glasses*, Vol. 1,2, Springer-Verlag Berlin Heidelberg (2011).

[18] G. Torlai, R.G. Melko, *Learning thermodynamics with Boltzmann machines*, Physical Review B 94 (16), 165134

[19] E. W. Tramel, A. Manoel, F. Caltagirone, M. Gabrie, and F. Krzakala, *Inferring sparsity: Compressed sensing using generalized restricted Boltzmann machines*, arXiv preprint arXiv:1606.03956, (2016).

[20] J Tubiana, R Monasson, *Emergence of Compositional Representations in Restricted Boltzmann Machines*, Physical Review Letters, 2017 - APS