Sujet de thèse – 2010-2013

Unsupervised Learning in Deep Neural Nets: From Curriculum to Multi-Task Learning

Responsable: Michèle Sebag
Laboratoire: LRI, CNRS & INRIA
Equipe: TAO, Machine Learning and Optimisation
Adresse: Université Paris-Sud, 91405 Orsay Cedex
Mail: Michele.Sebag@lri.fr

Context

Deep Neural Networks (DNN) have emerged since the mid 2000's due to Y. Bengio and G. Hinton [1, 5], based on the hybridization of unsupervised and supervised learning. The claim behind Deep NN can be schematized as: several levels of representations, stacked on top of each other, are required to represent complex concepts in a tractable way; a single-level representation, though in principle able to do the job, will not make it in practice. While the greater expressiveness and compacity obtained through the composition of representations had been known for decades, deep architectures were discarded as they could not be trained in an efficient way. The training bottleneck of deep architectures was overcome using layerwise unsupervised learning [1, 5]: each network layer is trained in turn after some unsupervised criterion, e.g. minimizing the reconstruction error for Auto-Associators or the log-likelihood for Stacked Boltzman Machines. Only in the last stage is the supervised goal taken into account, using e.g. back-propagation to optimize the whole set of weights in the DNN using a supervised criterion. This last stage is conducted using gradient approaches, thus prone to end up in a local optimum; the number and low quality of the local optima precisely is the bottleneck of standard deep architectures.

DNNs however overcome this bottleneck; they have demonstrated outstanding performances in difficult application domains, e.g. vision, and the reason for these good performances still is under study. A first conjecture was that unsupervised learning would draw the DNN into a basin of attraction conducive to the discovery of a good local optimum, akin a continuation method. It was later shown however by Erhan et al. (2009) that unsupervised learning rather acts as a regularization term, preventing the network from overfitting the discrimination task at hand [4].

In practice, the DNN unsupervised phase considers a sequence of examples for computational tractability. The order of examples in the sequence has been shown to have a significant impact on the final predictive accuracy of the network by Bengio et al. [2], leading to the so-called Curriculum Learning approach. The principle of Curriculum Learning can be viewed as another continuation method: by training the DNN on "simple examples" first, and considering more complex examples in later stages, one favors the discovery of a good basin of attraction. Notably, this principle reflects plain common sense: one should learn to walk before running, a good teacher should start with simple topics/problems before coming to more complex issues. The difficulty, naturally, is that "simplicity" has to be defined in the perspective of the learning algorithm.

Playing with the distribution of the training examples, through ordering or weighting, also is at the core of Boosting approaches [6], uniformly considering all examples in first learning stages, and gradually focussing on more complex ones (e.g., misclassified ones) in later stages; the main weakness of the approach comes from noisy examples, misleading the search.

Yet another domain where example distribution plays a major role is Multi-task learning [3], where the hypothesis trained on one sample distribution is meant to be exploited on another distribution.

Thesis

The goal of the PhD will be to investigate from a theoretical and algorithmic perspective the principles of Curriculum Learning, and bridge the gap between Curriculum and Multi-Task Learning. In Philippe Rolet's PhD (to be defended in December 2010), the selection of the most informative examples has been tackled in the Active Learning context, and principled selection criteria have been proposed.

The difference between Active Learning and the current PhD topic is that all considered examples are labelled. The point thus is to take into account both unsupervised criteria, such as the density w.r.t. the current distribution (or the target distribution in the Multi-task learning) and supervised criteria (e.g. uncertainty).

The starting point of the theoretical analysis will be to consider the law of the random variables defined by the gradient updates, depending on the sequence of examples considered, starting with the simple linear case where the problem involves a fraction of relevant features, and varying the fraction.

The application domain used to validate the proposed approach will be the control of a Brain Computer Device.

References

[1] Y. Bengio, P. Lamblin, V. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, Cambridge, MA, 2007.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proc. ICML 2009*, 2009.

[3] W.W. Cohen, A. McCallum, and S. T. Roweis, editors. *Multi-task learning for HIV therapy screening*, volume Proc. of Int. Conf. on Machine Learning. ACM, 2008.

[4] D. Erhan, P.A. Manzago, Y. Bengio, S. Bengio, and P. Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Proc. AISTATS*, pages 153–160, 2009.

[5] G.E. Hinton, S. Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Conputation*, 18:1527–1554, 2006.

[6] R.E. Schapire. Theoretical views of boosting. In *Proceedings of EuroCOLT-99, European Conference on Computational Learning Theory*, pages 1–10, 1999.