Deep Learning : *a new layer*

Jason Weston and Ronan Collobert
NEC Laboratories America, Princeton, NJ
jasewestonagmail.com and ronanacollobert.com

A slew of semi-supervised learning algorithms have been developed in the past few years, most of them based on the premise that one can improve performance on a supervised task by somehow combining it with an unsupervised learning technique.

However, many of these architectures are *disjoint* and *shallow* by which we mean the unsupervised learning algorithm is trained on unlabeled data separately as a first step, and then its results are fed to a supervised classifier which has a shallow architecture such as a (kernelized) linear model. For example, several methods learn a clustering or a distance measure based on a nonlinear manifold embedding as a first step [1]. Transductive SVMs [6] (which employs a kind of clustering directly as a regularizer) and LapSVM [5] (which employs a kind of embedding directly as a regularizer) are examples of methods that are *joint* in their use of labeled and unlabeled data, but they are still *shallow*.

As argued by several researchers, one is not likely to ever have enough labeled data to perform well at hard AI tasks such as scene or language understanding, making a compelling argument for using unlabeled data. Equally, the idea of sharing information learnt across complex sub-tasks (*multi-task learning*) seems an economical use of data, but necessitates a *deep architecture*, where all tasks are learnt *jointly*.

In a seemingly unrelated body of work to the previously mentioned semi-supervised learning algorithms, several authors have recently proposed methods for using unlabeled data in deep neural network-based architectures [4, 2, 3]. These methods either perform a greedy layer-wise learning of weights or learn an unsupervised auxiliary task at multiple levels of the architecture *jointly*. However, the methods proposed so far in our opinion are somewhat complicated and restricted. They include a particular kind of generative model (built from Restricted Boltzman Machines) [2], autoassociators [3], and a method of sparse encoding [4]. Moreover, in all cases these methods are not compared with, and appear on the surface to be completely different to, algorithms developed by researchers in the field of semi-supervised learning.

In this work we advocate that there are simpler ways of performing deep learning by leveraging many of the *existing* ideas in unsupervised and semi-supervised algorithms so far developed in *shallow* architectures. We show these methods can actually be (i) trained by stochastic gradient descent and used in multi-layered networks; and (ii) are also valid in the *deep* learning framework given above. Specificially, we propose three new regularizers (auxiliary tasks) for deep architectures: (i) a clustering-based regularizer; (ii) an embedding-based regularizer; and (iii) by learning the support of the density. These tasks can be trained either at the output layer, or on each layer of the architecture.

What do we get? State-of-the-art, fast, simple, deep-learning for dummies.

References

[1] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., USA, 09 2006.

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

[3] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. *ICML2007*, pages 473–480, 2007.

[4] Marc'Aurelio Ranzato, Y-Lan Boureau, and Yann LeCun. Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems (NIPS 2007)*, 2007.

[5] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *International Conference on Machine Learning, ICML*, 2005.

[6] Vladimir N. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, New York, 1998.

Category: learning algorithms
Preference: oral