Evolving Culture vs Local Minima

Yoshua Bengio, Dept. IRO, U. Montreal

Introduction. We propose a theory that relates difficulty of learning in deep architectures to culture and language. The theory posits that language and the combination of old ideas into new ideas provide an efficient cross-over evolutionary operator, and this allows rapid search in the space of communicable ideas that help humans build up better high-level internal representations of their world. The theory implies that human culture and the evolution of ideas have been crucial to counter an optimization difficulty: this optimization difficulty would otherwise make it very difficult for human brains to capture high-level knowledge of the world. The theory is grounded in experimental observations of the difficulties of training deep artificial neural networks. Plausible consequences of this theory for the efficiency of cultural evolutions are sketched.

Minimizing a Criterion. The theory is articulated around the following observations and hypotheses. First of all, we assume that biological learners approximately optimize a criterion (e.g. related to survival and ability to predict the sensory stream, and which can only be estimated stochastically) by a local descent process using a stochastic gradient estimator.

Hypothesis H1. When the brain of a single biological agent learns, it relies on approximate local descent in order to gradually improve itself.

High-Level Abstractions and Depth. We call *high-level abstraction* the kind of concept or feature that could be computed efficiently only through a deep structure in the brain (i.e., by the sequential application of several different transformations, each associated with an area of the brain or large group of neurons). Deeper architectures can be much more efficient in terms of representation of functions (or distributions) than shallow ones, as shown with theoretical results where for specific families of functions a too shallow architecture can require exponentially more resources than necessary (Bengio, 2009; Bengio and Delalleau, 2011). The basic intuition why this can be true is that in a deep architecture there is *re-use* of parameters and *sharing* of sub-functions to build functions, similarly to the way we decompose programs into subroutines calling each other rather than as a flat main program.

Hypothesis H2. Higher-level abstractions in brains are represented by deeper computations (going through more areas or more computational steps).

Observation O1: training deep architectures is easier if hints are provided about the function that intermediate levels should compute (Hinton, Osindero and Teh, 2006; Weston, Ratle and Collobert, 2008; Salakhutdinov and Hinton, 2009; Bengio, 2009).

Observation O2: from the work on artificial neural networks, it is clearly much easier to teach a network with supervised learning (where we provide it examples of when a concept is present and when it is not present in a variety of examples) than to expect unsupervised learning to discover the concept (which may also happen but usually leads to poorer renditions of the concept).

Observation O3: directly training all the layers of a deep network together not only makes it difficult to exploit all the extra modeling power of a deeper architecture but it actually get worse results *as the number of layers is increased* (Larochelle et al., 2009; Erhan et al., 2010).

Observation O4: in (Erhan et al., 2010), we observed that no two training trajectories end up in the same local minimum, out of hundreds of runs. This suggests that the number of functional local minima (i.e. corresponding to different functions, each of which possibly corresponding to many instantiations in parameter space) must be huge.

Observation O5: a training trick (unsupervised pre-training) which changes the initial conditions of the descent procedure allows to reach much better local minima (in terms of generalization error!), and these better local minima do not appear to be reachable by chance alone (as visually clear in the 2-D figures of trajectories in (Erhan et al., 2010)).

Hypothesis H3. Learning of a single human learner is limited by effective local minima.

Hypothesis H4. The effect of local minima tends to be more pronounced when training deeper architectures (by an optimization method based on iteratively descending the training criterion).

Hypothesis H5. A single human learner is unlikely to discover high-level abstractions by chance because these are represented by a deep subnetwork in the brain.

Hypothesis H6. A human brain can learn high-level abstractions if guided by the signals produced by other humans, which act as hints or indirect supervision for these high-level abstractions.

Culture and Evolution of Ideas for Efficient Optimization. The idea of *memes* is old (Dawkins, 1976): it is anything that can be copied from one mind to another. Like for genes, the copy can be imperfect, and some form of cross-over can occur when new memes are created by the combination of old memes. Culture is a snapshot of a population of memes spread across the brains of a population of individuals. Memes are the units of selection for cultural evolution. Whereas pure parallel search (based only on mutation, copy error) would be very slow if the number of local minima is huge (as we hypothesized), the recombination of old memes to form new memes would allow, like the cross-over operator, a much more efficient search, because it allows a form of divide-and-conquer, the combination of independendly optimized sub-solutions to form solutions to larger problems. Whereas the practical application of genetic algorithms is sensitive to the choice of representation of the units of selection (genes), we hypothesize that memes are excellent units of selection, by construction.

Hypothesis H7. Language and the combination of old ideas into new ideas provide an efficient cross-over evolutionary operator, and this allows rapid search in the space of communicable ideas that help humans build up better high-level internal representations of their world.

Conclusion. How do we test these hypotheses? A first step is to firm up the observations, continuing the analysis of learning trajectories in deep networks, to validate and extend the observations made in the cited articles. Further along the way, one should test the validity of mechanisms for "escaping" local minima thanks to "hints" from another agent. This work also suggests the exploration of massively parallelized learning algorithms (which could take advantage of loosely-coupled clusters) that exploit parallel search and recombination of learned abstractions in the style discussed here, inspired by the evolution of ideas and culture. We want to test if group learning can be more efficient than isolated learning. Finally, this work would suggest conclusions regarding the efficiency of cultural evolution, which would be influenced by efficiency of exploration of new memes and the rate of spread of good memes. The former would be enhanced by investments in scientific research (especially in high-risk high-potential areas), and encouraging all forms of diversity (in education, scientific schools of thought, and beliefs in general). The latter would be enhanced by open and free access to information and scientific results, investing more in education, and an internet where everyone can publish easily, but where multiple decentralized peer evaluation systems help the most interesting new ideas to bubble up and spread faster.

References

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127. Also published as a book. Now Publishers, 2009.

Bengio, Y. and Delalleau, O. (2011). On the expressive power of deep architectures. In *ALT'2011*.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660.

Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554.

Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2009). Exploring strategies for training deep neural networks. *JMLR*, 10:1–40.

Salakhutdinov, R. and Hinton, G. (2009). Deep Boltzmann machines. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, volume 8.

Weston, J., Ratle, F., and Collobert, R. (2008). Deep learning via semi-supervised embedding. In *Proc. ICML 2008*, pages 1168–1175, New York, NY, USA.