Representational Power of Restricted Boltzmann Machines and Deep Belief Networks

Nicolas Le Roux and Yoshua Bengio
Dept. IRO, Université de Montréal
{*lerouxni,bengioy*}*@iro.umontreal.ca*

Deep Belief Networks (DBN) are generative models with many layers of hidden causal variables, recently introduced by Hinton et al, along with a greedy layer-wise unsupervised learning algorithm. The building block of a DBN is a probabilistic model called a Restricted Boltzmann Machine (RBM), used to represent one layer of the model. We show that RBMs are universal approximators of discrete distributions. We then study the question of whether DBNs with more layers are strictly more powerful in terms of representational power. This suggests another criterion for DBNs, obtained by considering that the top layer can perfectly fit its input.

Introduction

Learning algorithms that learn to represent functions with many levels of composition are said to have a *deep architecture*. Bengio and Le Cun (2007) point to results in computational theory of circuits to strongly suggest that deep architectures are much more efficient in terms of representation (number of computational elements, number of parameters) than shallow ones. Hinton, Osindero, and Teh (2006) introduced a greedy layer-wise *unsupervised* learning algorithm for Deep Belief Networks (DBN). The training strategy for such networks may hold great promise as a principle to help address the problem of training deep networks. Upper layers of a DBN are supposed to represent more "abstract" concepts that explain the input observation x, whereas lower layers extract "low-level features" from x.

Background on RBMs and DBNs

A RBM with n hidden units is a parametric model of the joint distribution between hidden variables h_i and observed variables x_j, of the form

$$P(x, \mathbf{h}) \propto e^{\mathbf{h}' W x + b' x + c' \mathbf{h}}$$

with parameters $\theta = (W, b, c)$. We consider here the simpler case of binary units. It is straightforward to show that $P(x|\mathbf{h}) = \prod_i P(x_i|\mathbf{h})$ and $P(x_i = 1|\mathbf{h}) = \mathrm{sigm}(b_i + \sum_j W_{ji} h_j)$, and $P(\mathbf{h}|x)$ has a similar form. Although $P(x)$ is not tractable, it can be computed easily up to a normalizing constant, and a good stochastic approximation of $\frac{\partial \log P(x)}{\partial \theta}$ can be computed, called the Contrastive Divergence gradient.
A DBN with i layers models the joint distribution between observed variables x_j and i hidden layers $\mathbf{h}^k$ made of binary units $\mathbf{h}_l^k$ (here all binary variables), as follows:

$$P(x, \mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^i) = P(x|\mathbf{h}^1) P(\mathbf{h}^1|\mathbf{h}^2) \ldots P(\mathbf{h}^{i-2}|\mathbf{h}^{i-1}) P(\mathbf{h}^{i-1}, \mathbf{h}^i)$$

Denoting $x = \mathbf{h}^0$, $P(\mathbf{h}^k|\mathbf{h}^{k+1})$ has the form $P(\mathbf{h}^k|\mathbf{h}^{k+1}) = \prod_i P(\mathbf{h}_i^k|\mathbf{h}^{k+1})$ and $P(\mathbf{h}_i^k = 1|\mathbf{h}^{k+1}) = \mathrm{sigm}(b_i + \sum_j W_{jk}^k \mathbf{h}_j^{k+1})$, and $P(\mathbf{h}^{i-1}, \mathbf{h}^i)$ is a RBM.

RBMs are Universal Approximators

RBMs with data-selected number of hidden units become non-parametric and possess universal approximation properties relating them closely to neural networks:
Theorem 0.1. *Any distribution over $\{0,1\}^n$ can be approximated arbitrary well with a RBM with $k+1$ hidden units where k is the number of input vectors whose probability is not 0.*
Theorem 0.2. *Let u be an arbitrary distribution over $\{0,1\}^n$ and let P be a RBM with marginal distribution p over the visible units such that $KL(u||p) > 0$. Then there exists a RBM Q composed of P and an additional hidden unit with marginal distribution q over the visible units such that $KL(u||q) < KL(u||p)$.*

Open Questions

Let R_i^n be a Deep Belief Network with $i+1$ layers, each of them composed of n units. Can we say something about the representational power of R_i^n as i increases? Let us denote D_i^n the set of distributions one can obtain with R_i^n. It is shown in Hinton et al. (2006) that $D_i^n \subseteq D_{i+1}^n$. Two questions remain:

- do we have $D_i^n \subset D_{i+1}^n$, at least for $i = 1$?
- what is D_∞^n?

Trying to Anticipate and Memory-Based Top Layer

The proposed greedy training of Deep Belief Networks means that only one layer is trained at a time. In that greedy phase, one does not take into account the fact that other layers will be added next.
Instead of directly maximizing the likelihood, this greedy strategy maximizes a lower bound on it, called the **variational bound** (Hinton et al., 2006):

$$\log P(\mathbf{h}^0) \geq \sum_{\mathbf{h}^1} Q(\mathbf{h}^1|\mathbf{h}^0) \left[\log P(\mathbf{h}^1) + \log P(\mathbf{h}^0|\mathbf{h}^1)\right] - \sum_{\mathbf{h}^1} Q(\mathbf{h}^1|\mathbf{h}^0) \log Q(\mathbf{h}^1|\mathbf{h}^0)$$

Once the weights of the first layer are frozen, the only element that is optimized is $P(\mathbf{h}^1)$.
We can show that there is an analytic formulation for the distribution $P^*(\mathbf{h}^1)$ that maximizes this variational bound (but not necessarily the likelihood $P(\mathbf{h}^0)$):

$$P^*(\mathbf{h}^1) = \sum_{\mathbf{h}_0} p^0(\mathbf{h}^0) Q(\mathbf{h}^1|\mathbf{h}^0)$$

where p^0 is the empirical distribution of input examples. One can sample from it by first randomly sampling an $\mathbf{h}^0$ from the empirical distribution and then propagating it stochastically through $Q(\mathbf{h}^1|\mathbf{h}^0)$. Using the first theorem stated before, there exists an RBM that can achieve this optimal distribution $P^*(\mathbf{h}^1)$.
At that point, we can make a very important comment:
Using a RBM that achieves this "optimal" $P^*(\mathbf{h}^1)$ (in terms of the variational bound), the Kullback-Leibler divergence between the empirical distribution and the distribution of our model is equal to $KL(p^0||p^1)$ where p^0 is the empirical distribution and p^1 is the distribution one obtains when starting from p^0 clamped in the visible units of the lower layer ($\mathbf{h}^0$), sampling the hidden units $\mathbf{h}^1$ given $\mathbf{h}^0$ and then sampling a $\mathbf{h}^0$ given $\mathbf{h}^1$. This is equivalent to making one "forward-backward pass" in the first RBM trained.

One might wonder why this is important. Even with the best possible model for $P(\mathbf{h}^1, \mathbf{h}^2)$ (according to the variational bound), i.e., the model that can achieve $P^*(\mathbf{h}^1)$, we obtain a KL divergence equal to $KL(p^0||p^1)$. For this KL divergence to be 0, one should have $p^0 = p^1$. But $p^0 = p^1$ could have been obtained with a one-level DBN (i.e. a single RBM) that perfectly fits the data, so that the second layer $\mathbf{h}^2$ seems useless.
Does that mean that adding layers is useless? We believe the answer is no; even if adding layers does not allow to perfectly fit the data (which is not sure because we optimize the variational bound rather than the likelihood), the distribution of our model is closer to the empirical distribution than a single RBM (we do only one "forward-backward pass" instead of doing an infinity of them). Furthermore, the extra layers allow to regularize and hopefully obtain a representation in which even a memory-based top layer could generalize well. This approach suggests using alternative criteria to train DBNs, that approximate $KL(p^0||p^1)$, and which can be computed before $\mathbf{h}^2$ is added, but that unlike Contrastive Divergence, take into account the fact that more layers will be added later.

References

Bengio, Y., & Le Cun, Y. (2007). Scaling learning algorithms towards AI. In Bottou, L., Chapelle, O., DeCoste, D., & Weston, J. (Eds.), *Large Scale Kernel Machines*. MIT Press.

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