Deep Belief Networks are Compact Universal Approximators

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

Since Deep Belief Networks (DBN) have been introduced by Hinton, Osindero, and Teh (2006), we have started to wonder what their true modeling power was. Le Roux and Bengio (2008) attempted to give some insight on the relative representational powers of DBNs and Restricted Boltzmann Machines, their building blocks, while leaving some questions open. Sutskever and Hinton (2008) partially answered one of these questions by proving that DBNs with 2^{n+1} layers composed of $n+1$ units each were universal generative models on the vectors of n bits. This paper improves this result by proving the same property for DBNs with $\frac{2^n}{n}$ layers of n units each. We also prove that Gaussian DBNs are universal generative models of distributions on $\mathbb{R}^n$ and that deep sigmoid networks with layers of size n can model any function from $\{0,1\}^n$ to $\{0,1\}^k$.

Introduction

Deep models are receiving increased attention in the community (Lee, Ekanadham, & Ng, 2008; Salakhutdinov & Hinton, 2008; Ranzato, Boureau, & LeCun, 2008), but the motivation behind them remains obscure; Bengio and Le Cun (2007) point to results in computational theory suggesting the use of deep architectures and Bengio (2007) proposes an explanation of what happens in the upper layers of a deep network.

Background on RBMs and DBNs

An 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

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

with parameters $\theta = (W, b, c)$. We consider here the 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. 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 $\mathbf{h}^k_l$ (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 $P(\mathbf{h}^k|\mathbf{h}^{k+1}) = \prod_i P(\mathbf{h}^k_i|\mathbf{h}^{k+1})$ and $P(\mathbf{h}^k_i = 1|\mathbf{h}^{k+1}) = \mathrm{sigm}(b_i + \sum_j W^k_{jk}\mathbf{h}^{k+1}_j)$, and $P(\mathbf{h}^{i-1}, \mathbf{h}^i)$ is a RBM.

DBNs are Compact Universal Approximators

Sutskever and Hinton (2008) provided a proof of the universal approximation property of Deep Belief Networks using 2^{n+1} layers of $n+1$ bits each. They thus gave partial answers to the open questions asked by Le Roux and Bengio (2008):

> Let R^n_i be a Deep Belief Network with $i+1$ layers, each of them composed of n units and D^n_i be the set of distributions one can obtain with R^n_i.
>
> - do we have $D^n_i \subset D^n_{i+1}$?
> - what is D^n_∞?

Their proof used a switch, placed at every layer, which would change an arbitrary vector x_0 into another arbitrary vector x_1 with the appropriate probability to cover the whole set of vectors of n bits. Using a similar idea, we were able to prove the following theorem:

Theorem 1. *Let n be an arbitrary positive integer and let $k = \lfloor \log_2 n \rfloor$. Then it is possible to model any distribution over n bits with a DBN composed of $2^{n-k} + 1$ layers of n units.*

Proof sketch. Let x_0 be an arbitrary vector over n bits. Then, for all k in $\{1, \ldots, n\}$ and all p in $[0, 1]$, there exists a sigmoid belief network composed of two layers h and v of size n such that:

- if x_0 is clamped in h, v is equal to x_0, except for the k-th bit which is flipped with probability p
- if $x \neq x_0$ is clamped in h, v is equal to x with probability 1.

We will therefore build a DBN with $2^n + 1$ layers of n bits, such that, at every layer, a particular vector is potentially transformed into another vector which differs by one bit. Using a Gray code, we can construct a sequence of vectors differing by one bit to cover the set of sequences of n bits, thus achieving the universal approximation property of DBNs. Since a Gray code only changes one bit at a time, we can split the set of sequences into n subsequences of almost equal length, each of them being a Gray code, such that no two subsequences change the same bit at the same time. This divides the number of layers needed by n (approximately).

□

The network presented there has a total number of parameters equal to $n2^n$, similar to the universal approximator RBM first presented by Freund and Haussler (1994). Even though there are only 2^n different vectors, thus raising the possibility of using only as many parameters, storing the address of every vector takes n bits. It is therefore reasonable to think that the network achieved is of the same order of magnitude as the most compact network.
Using a similar construction, we can prove another theorem:

Theorem 2. *Let n and k be two arbitrary positive integers. A DBN with $2^n + 2^k$ layers of n units can model any function from $\{0, 1\}^n$ to $\{0, 1\}^k$.*

Proof sketch. At every layer, one can arbitrarily change an arbitrary vector into another arbitrary vector while leaving the other unchanged. The idea is thus to move all the elements of one class to the same position and to do so for every class. Once we are reduced to as many points as the number of classes, we move them so that k hyperplanes can separate them as needed. □

Finally, we can also prove the following general approximation property:

Theorem 3. *Let n be an arbitrary positive integer. Then it is possible to approximate arbitrarily well any distribution over $\mathbb{R}^n$ with a DBN composed of 2 layers whose number of units is not restricted.*

Proof sketch. Since any distribution can be approximated arbitrarily well with a mixture of Gaussians, we create a DBN with as many hidden units as the number of Gaussians, their associated weights being the centers of the Gaussians. The second layer ensures that the hidden units are mutually exclusive, thus avoiding to create unwanted extra components in the mixture. □

References

Bengio, Y. (2007). Learning deep architectures for AI. Tech. rep. 1312, Université de Montréal, dept. IRO.

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.

Freund, Y., & Haussler, D. (1994). Unsupervised learning of distributions of binary vectors using two layer networks. Tech. rep. CRL-94-25, UCSC.

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

Le Roux, N., & Bengio, Y. (2008). Representational power of restricted boltzmann machines and deep belief networks. *Neural Computation*, *to appear*.

Lee, H., Ekanadham, C., & Ng, A. (2008). Sparse deep belief net model for visual area V2. In Platt, J., Koller, D., Singer, Y., & Roweis, S. (Eds.), *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA.

Ranzato, M., Boureau, Y.-L., & LeCun, Y. (2008). Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems (NIPS 2007)*. MIT Press.

Salakhutdinov, R., & Hinton, G. (2008). Using deep belief nets to learn covariance kernels for gaussian processes. In Platt, J., Koller, D., Singer, Y., & Roweis, S. (Eds.), *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA.

Sutskever, I., & Hinton, G. (2008). Deep narrow sigmoid belief networks are universal approximators.. Vol. to appear.