Deep Narrow Boltzmann Machines are Universal Approximators

Guido Montúfar (author)
montufar@mis.mpg.de
Max Planck Institute for Mathematics in the Sciences

KyungHyun Cho (presenter)
kyunghyun.cho@umontreal.ca
Université de Montréal

INTRODUCTION

- It is an interesting question how the representational power of deep artificial neural networks compares with that of shallow neural networks.
- Furthermore, it is interesting how the representational power of layered networks compares in the cases of undirected and directed connections.
- A basic question in this respect is whether a given network type can reach any degree of representation accuracy, when endowed with sufficiently many units.
- Universal approximation has been verified for many types of neural networks, but has remained an open problem for deep narrow Boltzmann machines.

DEFINITION

A deep Boltzmann machine with \( n_1, \ldots, n_3 \) units is a model of probability distributions of the form

\[
p_W(s) = \frac{1}{Z(W, b)} \exp(\sum_{l=0}^{L-1} \sum_{i=0}^{n_l} W_{i,l} s_i + \sum_{j=0}^{n_l} b_j). 
\]

The model is narrow, when all layers have about the same number of units.

OVERVIEW

- At an intuitive level, undirected networks are expected to be more powerful than directed networks, since “they allow information to flow both ways.”
- This intuition is not straightforward to verify. Feedforward networks can be naturally studied in a sequential way, but undirected networks are more subtle.
- We develop a method to study undirected architectures in a sequential way.

Figure. Restricted Boltzmann machine (RBM), deep belief network (DBN), and deep Boltzmann machine (DBM).

SEQUENTIAL ANALYSIS

- Express the visible probability distribution of a DBM in terms of the distributions of two smaller DBMs.

\[
q_{W, b}(x) = \frac{1}{Z(W, b)} \exp(W x + b). 
\]

- Problem: shared parameters of intermediate marginal and feedforward map.

- Solution: restrict attention to special marginals \( s \) from the top DBM to obtain independent parameters for the feedforward maps.

- In this way, with each additional layer we can transform the visible distribution by an independent feedforward map.

UNIVERSAL APPROXIMATION

Theorem. A deep and narrow Boltzmann machine with a visible layer of \( n \) units and \( L \) hidden layers of \( n_i \) units each is a universal approximator of probability distributions on the states of the visible layer, provided \( L \) is large enough.

- Sufficient condition:

\[
L \geq \frac{2^n}{2^n - \log(n' - 1)},
\]

for any \( n' = 2^k + k + 1 \geq n, k \in \mathbb{N} \).

- Necessary condition:

\[
L \geq \frac{2^n - (n + 1)}{n(n + 1)}.
\]

- For universal approximation, the first hidden layer must have at least as many units as the visible layer (minus one).

- Similar results for discriminative and multinomial models.

REFERENCES


ACKNOWLEDGMENTS

I am grateful to the Santa Fe Institute, Santa Fe, NM, USA, where I was hosted while working on this article. Many thanks to KyungHyun Cho for presenting this ICLR conference paper.

CONCLUSIONS

- We investigated the compositional structure of DBMs and presented a trick to separate the activities on the upper part of the network from those on the lower part of the network.
- Within certain parameter regions, deep Boltzmann machines can be studied as feedforward networks.
- We showed that deep narrow Boltzmann machines are universal approximators, and provided upper and lower bounds on the sufficient depth and width.
- In a specific sense, deep narrow Boltzmann machines are at least as powerful as narrow sigmoid belief networks and restricted Boltzmann machines.
- The methods appear valuable for studying the effects of training undirected networks sequentially, from layer to layer.