Denoising Deep Boltzmann Machines: Compression for Deep Learning

Qing Li [1] Chen Yang [2]

[1]Western Digital Research (Qing.Li7@wdc.edu)

[2]University of Michigan

DCC'20

Motivation and Main Results

- In theory, Deep Boltzmann Machines (DBM) are universal approximators;
- In practice, they are not ...
- What compression can do for DBM?

Challenges for DBM: Gap between theory and practice

Figure 1: Sample from LFW.

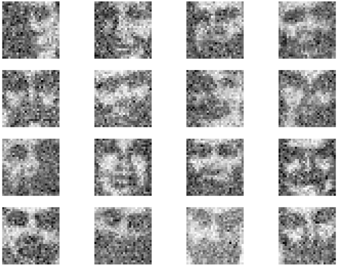


Figure 2: Sample from DBM.

Motivated by Denoising

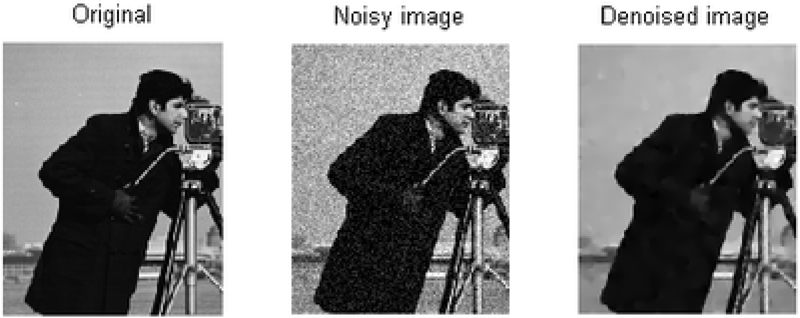


Figure 3: An example of denoising.

Denoising DBM results

Figure 4: Denoising 784-500-784-500 GBDBM on LFW images. From top to bottom, and left to right, the images are samples from noisy GBDBM, denoised GBDBM with $\beta = 5, 50, 100, 200, 250$, respectively.

Outline

1. Background

2. Denoising DBM

3. Experimental results

4. Conclusion and Future Work

Background

- DBM refers to the following four Boltzmann Machines:
    1. Restricted Boltzmann Machines (RBM) [1];
    2. Bernoulli Deep Boltzmann Machines (BDBM) [2];
    3. Gaussian-Bernoulli RBM (GBRBM); [3]
    4. Gaussian-Bernoulli Deep Boltzmann Machines (GBDBM) [4].
- Lossy compression & rate distortion.

Background: RBM & BDBM

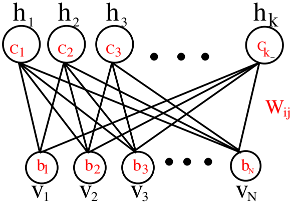


Figure 5: A Restricted Boltzmann Machines.

- Parameterized by $(\mathbf{W}, \mathbf{b}, \mathbf{c})$;
- $P(\mathbf{v}, \mathbf{h}) = Z^{-1}\exp(-E(\mathbf{v}, \mathbf{h}))$, where

$$
\begin{aligned}
E(\mathbf{v}, \mathbf{h}) &= -(\mathbf{v}^T\mathbf{b} + \mathbf{h}^T\mathbf{c} + \mathbf{v}^T\mathbf{W}\mathbf{h}), \\
Z &= \sum_{\mathbf{h},\mathbf{v}} \exp(-E(\mathbf{v}, \mathbf{h})).
\end{aligned}
$$

- BDBM: represented by $\{(\mathbf{W}^l, \mathbf{b}, \mathbf{b}^l)\}$ for $l = 1, 2, \cdots, L$.

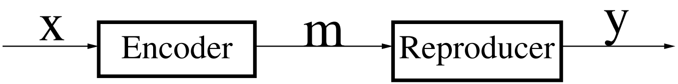
Background: GBRBM & GBDBM

- Parameterized by $(\mathbf{W}, \mathbf{b}, \mathbf{c}, \boldsymbol{\sigma})$;
- $P(\mathbf{v}, \mathbf{h}) = Z^{-1} \exp\left(-E(\mathbf{v}, \mathbf{h})\right)$, where

$$\begin{aligned} E(\mathbf{v}, \mathbf{h}) &= \sum_{i=1}^{N} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{v_i}{\sigma_i^2} c_j w_{ij} - \sum_{j=1}^{K} c_j h_j, \qquad (1) \\ \text{and} \quad \mathrm{Z} &= \sum_{\mathbf{v}, \mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right). \end{aligned}$$

- GBDBM: Represented by $\{(\mathbf{W}^l, \mathbf{b}, \mathbf{b}^l, \boldsymbol{\sigma}), l = 1, 2, \cdots, L\}$.

Background: Lossy compression



- $\mathbf{x} \in \mathbb{R}^N$, $m \in \{0, 1, \cdots, M-1\}$, $\mathbf{y} \in \mathbb{R}^N$;
- Rate: $R = \frac{\log_2 M}{N}$;
- Distortion: $D = E(\varphi(\mathbf{x}, \mathbf{y}))$; e.g.,$\varphi(\cdot)$ can be Hamming distance.
- N^{th}-order rate distortion $R_N(D)$:
  $\mathcal{L}(P(\mathbf{y}|\mathbf{x})) = I(\mathbf{x}, \mathbf{y}) + \beta\mathbb{E}(\varphi(\mathbf{x}, \mathbf{y}))$, i.e.,
  $R_N(D) = \min_{P(\mathbf{y}|\mathbf{x})} \mathcal{L}(P(\mathbf{y}|\mathbf{x}))$;
- $P^*(\mathbf{y}|\mathbf{x}) = \arg\min_{P(\mathbf{y}|\mathbf{x})} \mathcal{L}(P(\mathbf{y}|\mathbf{x}))$, and let $P^*(\mathbf{y})$ be the resulting marginal distribution.

Denoising DBM: problem statement

Given $\mathcal{T} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n_t}\} \in \mathbb{R}^{n_t \times n_v}$, where $\mathbf{x}_i \in \mathbb{R}^{n_v}$ is from some unknown distribution P_{data} and some noisy $\{(\mathbf{W}^l, \mathbf{b}, \mathbf{b}^l, \boldsymbol{\sigma}), l = 1, 2, \cdots, L\}$, our goal is to fine-tune or denoise $\{(\mathbf{W}^l, \mathbf{b}, \mathbf{b}^l, \boldsymbol{\sigma}), l = 1, 2, \cdots, L\}$ representing $\mathbf{v}$ to some less noisy $\{(\hat{\mathbf{W}}^l, \hat{\mathbf{b}}, \hat{\mathbf{b}}^l, \boldsymbol{\sigma}), l = 1, \hat{2}, \cdots, L\}$ representing $\mathbf{y}$ such that $I(\mathbf{x}, \mathbf{y}) > I(\mathbf{x}, \mathbf{v})$.

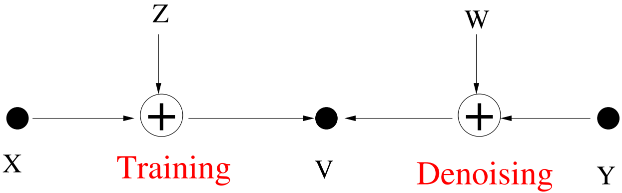


Figure 6: Denoising DBM illustration.

Denoising DBM: Motivated by compress-based denoising

Lossy compression of a noisy signal, under the **right distortion measure** and at the **right distortion level**, leads to an effective denoising. [5, 6, 7, 8].

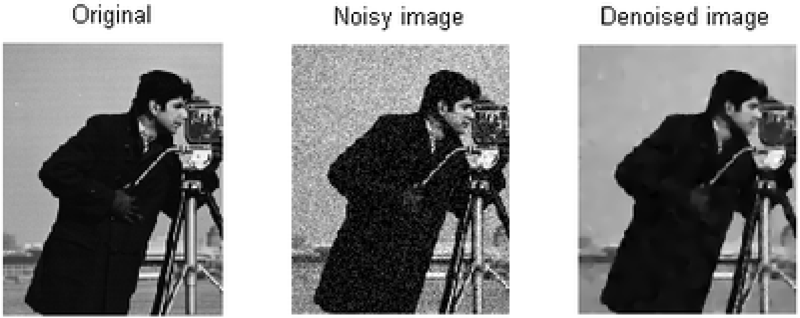


Figure 7: An example of denosing.

Denoising DBM: distortion measure and distortion level

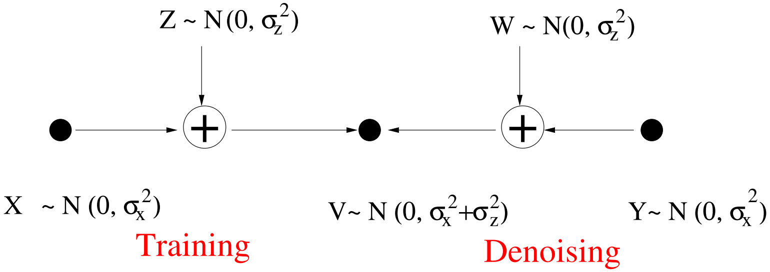


Figure 8: Denoising DBM illustration.

- Assume $\mathbf{v}_i = \mathbf{x}_i + \mathbf{z}_i$, $\mathbf{z}_i \sim P_{\mathbf{z}}(\mathbf{z})$.
- $\varphi(\mathbf{y}, \mathbf{v}) = -\log_2 P_{\mathbf{z}}(\mathbf{y} - \mathbf{v})$, and $D = \sum_{i=1}^{n} \varphi(\mathbf{x}_i, \mathbf{v}_i)$.
- $\mathbf{z}_n \sim \mathrm{Ber}(\mathbf{p}_\mathrm{n})$, $\varphi(\mathbf{x}_n, \mathbf{y}_n)$ is *weighted Hamming distortion*;
- $p(\mathbf{v}|\mathbf{x}) \sim \mathcal{N}(\mathbf{v}|\mathbf{x}, \varsigma)$,$\varphi(\mathbf{x}_n, \mathbf{y}_n)$ is approximately $(\mathbf{x} - \mathbf{v})^2$.

Denoising DBM: compression with DBM

- Given $\varphi(\mathbf{x}, \mathbf{y})$ and D, need to design a lossy compression;
- Given $\varphi(\mathbf{x}, \mathbf{y})$ and D, it associates $R_N(D)$, $P^*(\mathbf{y}|\mathbf{x})$ and $P^*(\mathbf{y})$;
- DBM is universal, thus train it to learn $P^*(\mathbf{y}|\mathbf{x})$ and $P^*(\mathbf{y})$.

Denoising DBM: DBM interpretation of rate-distortion

Lemma [9, Chapter 13.7, pp. 362] 1.

$$
\begin{aligned}
P^*(\mathbf{y}|\mathbf{x}) &= \frac{1}{Z'_\beta(\mathbf{x})} P^*(\mathbf{y}) \exp(-\beta\varphi(\mathbf{y},\mathbf{x})), && (2) \\
R_N(D) &= \frac{E(-\log_2 Z'_\beta(\mathbf{x}))}{N} - \frac{\beta D}{\ln 2}, && (3)
\end{aligned}
$$

where the expectation is with respect to the probability distribution on $\mathbf{x}$,

$$
Z'_\beta(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{\mathbf{y}} P^*(\mathbf{y}) \exp(-\beta\varphi(\mathbf{x},\mathbf{y})), \qquad (4)
$$

where β is the Lagrange multiplier that minimizes $I(\mathbf{x},\mathbf{y}) + \beta E(\varphi(\mathbf{x},\mathbf{y}))$.

Denoising DBM: DBM rate-distortion for binary case

Theorem 1.
For the distortion $\varphi(0,0) = \varphi(1,1) = 0$, $\varphi(0,1) = a$, $\varphi(1,0) = b$, $a \geq b > 0$, *assume that* $P^*(\mathbf{y})$ *can be represented by a BDBM,* $\{(\mathbf{W}^l_{\mathbf{Y}}, \mathbf{b}_{\mathbf{Y}}, \mathbf{b}^l_{\mathbf{Y}})\}$. *Then* $P^*(\mathbf{y}|\mathbf{x})$ *can be represented by the BDBM* $\{(\mathbf{W}^l_{\mathbf{Y}|\mathbf{X}}, \mathbf{b}_{\mathbf{Y}|\mathbf{X}}, \mathbf{b}^l_{\mathbf{Y}|\mathbf{X}})\}$, *where*

$$\begin{cases} \mathbf{W}^1_{\mathbf{Y}|\mathbf{X}} &= \mathbf{W}^1_{\mathbf{Y}}, \\ \mathbf{b}_{\mathbf{Y}|\mathbf{X}} &= \mathbf{b}_{\mathbf{Y}}, \\ b^1_{\mathbf{Y}|\mathbf{X},1} &= b^1_{\mathbf{Y},1} - \beta a 1_{x_i=0} + \beta b 1_{x_i=1}, \end{cases}$$
$$\begin{cases} \mathbf{W}^l_{\mathbf{Y}|\mathbf{X}} &= \mathbf{W}^l_{\mathbf{Y}}, \\ b^l_{\mathbf{Y}|\mathbf{X}} &= b^l_{\mathbf{Y}}, \end{cases} \quad l \geq 2. \tag{5}$$

Denoising DBM: DBM rate-distortion for Gaussian case

Theorem 2.
For a squared error distortion, $\varphi(x, y) = (x - y)^2$ where $x, y \in R$, assume that $P^*(\mathbf{y})$ can be represented by a GBDBM, $\{(\mathbf{W}^l_{\mathbf{Y}}, \mathbf{b}_{\mathbf{Y}}, \mathbf{b}^l_{\mathbf{Y}}, \boldsymbol{\sigma}_{\mathbf{Y}})\}$. Then, $P^*(\mathbf{y}|\mathbf{x})$ can be represented by the GBDBM $\{(\mathbf{W}^l_{\mathbf{Y}|\mathbf{X}}, \mathbf{b}_{\mathbf{Y}|\mathbf{X}}, \mathbf{b}^l_{\mathbf{Y}|\mathbf{X}}, \boldsymbol{\sigma}_{\mathbf{Y}|\mathbf{X}})\}$, where

$$\left\{\begin{array}{ccc} \mathbf{W}^1_{\mathbf{Y}|\mathbf{X},i,j} & = & \mathbf{W}^1_{\mathbf{Y},i,j}\frac{(\sigma_i^{'})^2}{\sigma_i^2} \\ b_{\mathbf{Y}|\mathbf{X},i} & = & \frac{\mathbf{b}_{\mathbf{Y},i}\sigma_T^2 + x_i\sigma^2_{\mathbf{Y},i}}{\gamma_i^2}, \\ \sigma_{\mathbf{Y}|\mathbf{X},i} & = & \frac{\sigma_{\mathbf{Y},\mathbf{i}}\sigma_T}{\gamma_i}. \end{array}\right.$$

$$\left\{\begin{array}{ccc} \mathbf{W}^l_{\mathbf{Y}|\mathbf{X}} & = & \mathbf{W}^l_{\mathbf{Y}}, \\ b^l_{\mathbf{Y}|\mathbf{X}} & = & b^l_{\mathbf{Y}}, \end{array}\right. \quad l \geq 2, \tag{6}$$

where $\sigma_i', \sigma_T, \gamma_i$ are defined in the Appendix.

Denoising DBM: DBM-Blahut-Arimoto

Algorithm 1 DBM Blahut-Arimoto

procedure $\mathrm{DBM\text{-}BA}(\mathcal{T}, \beta, \varphi(\cdot))$
    initialize $\{(\mathbf{W}^{l,0}_{\mathbf{Y}}, \mathbf{b}^{0}_{\mathbf{Y}}, \mathbf{b}^{l,0}_{\mathbf{Y}}, \sigma^{0}_{\mathbf{Y}})\}$ arbitrarily.
    for $t = 1, \ldots, t_{\mathsf{max}}$ **do**
        sample $\mathbf{y}^t_n$ for $\mathbf{x}_n \in \mathcal{T}$ from $P^*(\mathbf{y}|\mathbf{x})$.
        train $\{\mathbf{W}^{l,t}_{\mathbf{Y}}, \mathbf{b}^{t}_{\mathbf{Y}}, \mathbf{b}^{l,t}_{\mathbf{Y}}, \sigma^{t}_{\mathbf{Y}}\}$ with $\mathcal{T}^t_{ba} \stackrel{def}{=} \{\mathbf{y}^t_1, \mathbf{y}^t_2, \cdots, \mathbf{y}^t_{n_t}\}$.
    end for
    return $\{\mathbf{W}^{l,t_{max}}_{\mathbf{Y}}, \mathbf{b}^{t_{max}}_{\mathbf{Y}}, \mathbf{b}^{l,t_{max}}_{\mathbf{Y}}, \sigma^{t_{max}}_{\mathbf{Y}}\}$.
end procedure

Denoising DBM: algorithm and theoretical results

The denoising DBM scheme is to transform $\{(\mathbf{W}^l, \mathbf{b}, \mathbf{b}^l, \boldsymbol{\sigma}), l = 1, 2, \cdots, L\}$ to $\{(\hat{\mathbf{W}}^l, \hat{\mathbf{b}}, \hat{\mathbf{b}^l}, \hat{\boldsymbol{\sigma}}), l = 1, 2, \cdots, L\}$ via the DBM-BA algorithm with $\varphi(\mathbf{y}, \mathbf{v}) = -\log_2 P_{\mathbf{z}}(\mathbf{y} - \mathbf{v})$, the D defined above, and some β.

Theorem 3.
For strictly convex $R_N(D)$, if $\{(\mathbf{W}^l, \mathbf{b}, \mathbf{b}^l, \boldsymbol{\sigma}), l = 1, 2, \cdots, L\}$ converges to $R_N(D)$ with DBM-BA, the denoised $\{(\hat{\mathbf{W}}^l, \hat{\mathbf{b}}, \hat{\mathbf{b}^l}, \hat{\boldsymbol{\sigma}}), l = 1, 2, \cdots, L\}$ fully recovers all information about training data, i.e., $D_{\mathrm{KL}}\left(P(\mathbf{y}) \| P(\mathbf{x})\right) \to 0$.

Olivetti

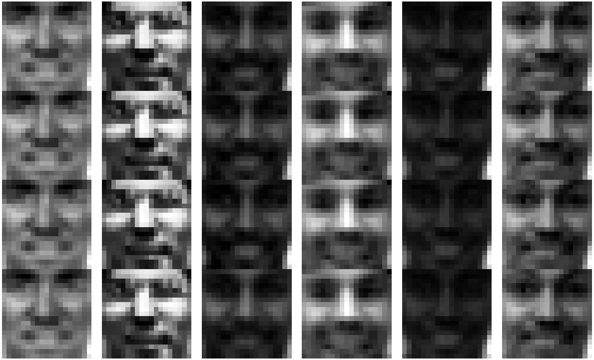


Figure 9: Denoising GBRBM on Olivetti face dataset. The first, the third and the fifth columns are images sampled from noisy GBRBM and the second, the fourth and the sixth columns are images sampled from denoised GBRBM with β's 5, 2.5, and 2, respectively.

LFW

Figure 10: Denoising 784-500-784-500 GBDBM on LFW images. From top to bottom, and left to right, the images are samples from noisy GBDBM, denoised GBDBM with $\beta = 5, 50, 100, 200, 250$, respectively.

Conclusion and Future Work

- Conclusion: propose denoising DBM to better train DBM;
- Future work: Is it possible to generalize the idea to other generative models?

Thank you!

Asja Fischer and Christian Igel.
Training restricted Boltzmann machines: An introduction.
Pattern Recognition, 47(1):25–39, 2014.

Ruslan Salakhutdinov and Geoffrey Hinton.
Deep Boltzmann machines.
In *Artificial Intelligence and Statistics*, pages 448–455, 2009.

Geoffrey E Hinton and Ruslan R Salakhutdinov.
Reducing the dimensionality of data with neural networks.
science, 313(5786):504–507, 2006.

KyungHyun Cho, Alexander Ilin, and Tapani Raiko.
Improved learning of Gaussian-Bernoulli restricted Boltzmann machines.
In *International conference on artificial neural networks*, pages 10–17. Springer, 2011.

Balas K Natarajan.
Filtering random noise from deterministic signals via data compression.

IEEE transactions on signal processing, 43(11):2595–2605, 1995.

Juan Liu and Pierre Moulin.
Complexity-regularized image denoising.
IEEE Transactions on Image Processing, 10(6):841–851, 2001.

S Grace Chang, Bin Yu, and Martin Vetterli.
Adaptive wavelet thresholding for image denoising and compression.
IEEE transactions on image processing, 9(9):1532–1546, 2000.

Tsachy Weissman and Erik Ordentlich.
The empirical distribution of rate-constrained source codes.
IEEE Transactions on Information Theory, 51(11):3718–3733, 2005.

T. M. Cover and J. A. Thomas.
Elements of Information Theory.
Wiley, New York, 1991.