Compositional RBMs, a Bird's Eye view (based on the work of J. Tubiana and R. Monasson [3])

Simone Maria Giancola [1]

[1]Bocconi University

Unsupervised Restricted Bolzmann Machine on MNIST

RBMs are well suited for learning distributions of data. Take N visible binary units $\mathbf{v} = (v_1, \dots, v_N)$ and M hidden units $\mathbf{h} = (h_1, \dots, h_M)$ with `ReLu` activations.

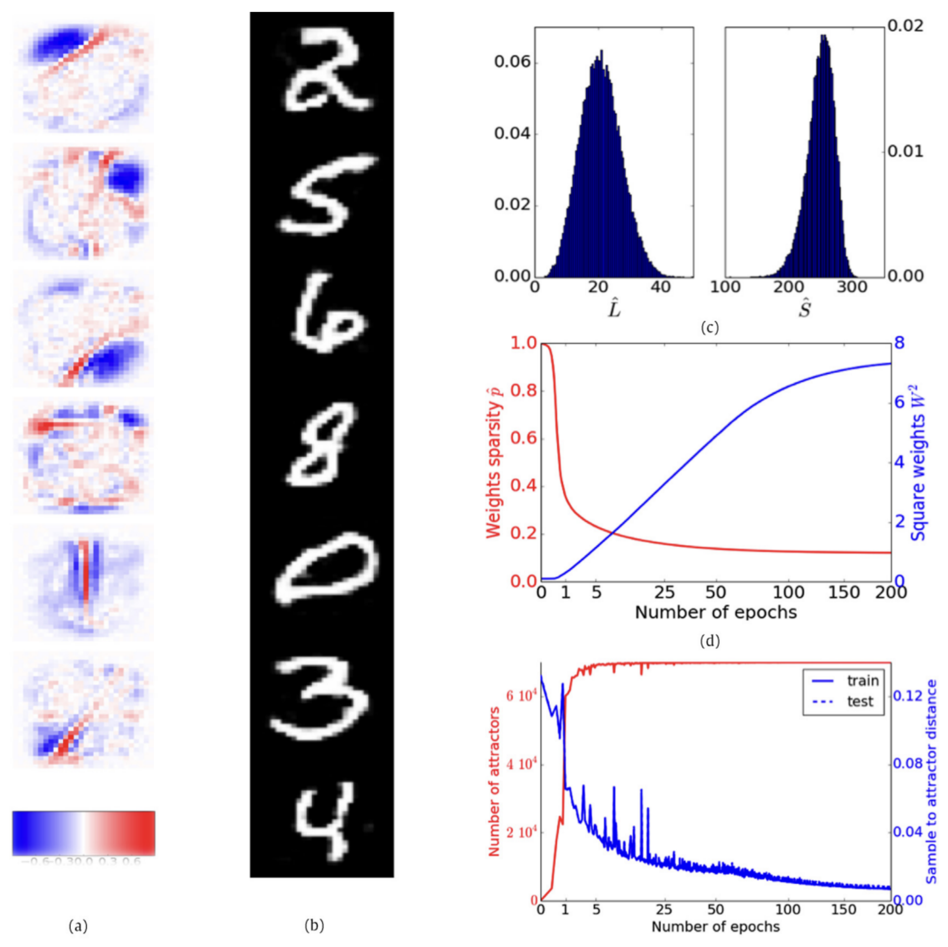

Figure 1. MNIST trained RBM. Source [1]

After training, weights of hidden units $\mathbf{w}_\mu$ are strokes on the grid (a), combined to generate representations of numbers (b). There is a *small* (resp. large) number of strongly activated (silent) hidden units (c). Sparsity $\widehat{p}$ and high inverse temperature W_2 are induced by likelihood maximization, and not forced a priori (d). Representative power is high: many local minima are at low distance from starting configuration (e).

Results

Empirical Simulations and Theoretical Analysis suggest that:

- **Structural Changes** on $\widehat{p}, W_2$ generalize to any RBM training dynamics to reach
- **a Compositional Phase** (in contrast with Glassy & Ferromagnetic),
- where **generated samples** are combinations of strongly activated units
- Additionally a **Random RBM ensemble** thermodynamically favours Compositionality

A Word about Phases

Dominating hidden configurations at end of training can be:

- **Ferromagnetic**, all weakly activated units but one, no variability
- **Glassy**, all weakly activated units, no representative power
- **Compositional**, enough strongly activated units, that access different low energy visible configurations

Training & Sampling

Assume reasonable approximate Likelihood training, well designed MCMC sampling techniques and high computational power.

R-RBM Ensemble

Weights $w_{i\mu}$ are **quenched** random variables wp $\left(\frac{p_i}{2}, 1-p_i, \frac{p_i}{2}\right)$ and values $\left\{-\frac{1}{\sqrt{N}}, 0, \frac{1}{\sqrt{N}}\right\}$, uniform visible field and hidden activations $g_i \equiv g, \mathcal{U}_\mu \equiv \mathcal{U}^{\texttt{Relu}}$ with threshold θ. Set $\alpha = \frac{M}{N} \in \mathcal{O}(1)$ and $M, N \to \infty$. Strongly activated hidden units will scale as $m\sqrt{N}$ to be larger than their inputs $I_\mu = \mathbf{w}_\mu \cdot \mathbf{v}$. In the analysis of a subsequent work [2] we find that the ground state energy is:

$$E_{GS} = f(\texttt{order params}) = f(m, L, r, q, B, C) \quad (1)$$

to optimize averaging over quenched weights.

R-RBM Compositional Phase

Letting $p \to 0$, the typical ground state energy presents a critical (α, θ) unbounded relation to have L magnetized hidden units without entering the Glassy Phase, upon adjusting θ.

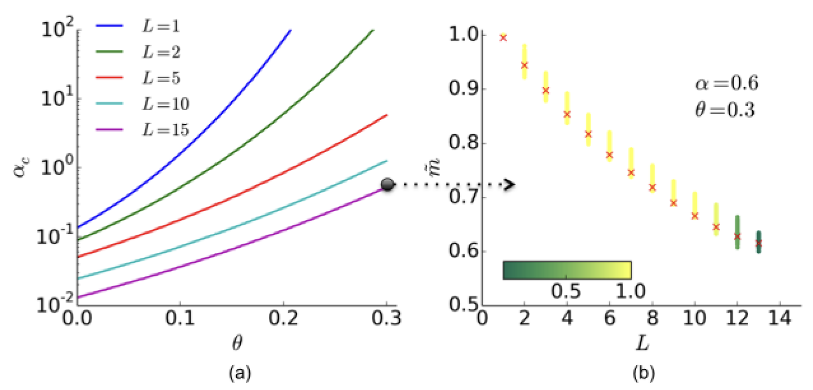

Figure 2. Compositional R-RBM & MCMC Simulation. Source [3]

Below the critical lines L hidden units can be strongly magnetized (a). MCMC simulations (yellow-green scale) of RBM are in great accordance when aligned at the start and the number of hidden units is as theorized (b). Normalized magnetization is non null and in accordance with R-RBM prediction in red (b).

R-RBMs Energy and Hidden at high sparsity

Strongly activated hidden units scale as $L \sim \frac{\ell}{p} > 0$, and normalized energy is $e_\ell = \frac{E_{GS}}{p}$ is non-monotonic function of ℓ (can be minimized). Non-magnetized hidden units shut down by choosing $\theta \sim \sqrt{p} \sim \sqrt{r}$ avoiding cross talk.

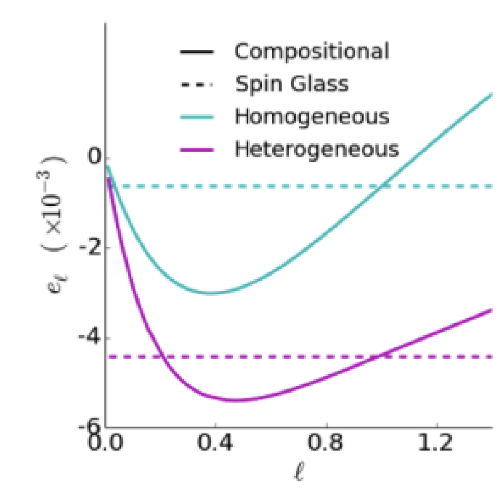

Figure 3. Compositional R-RBM & MCMC Simulation. Source [3]

RBMs Spectrum

Sparsity & Participation Ratios

RBMs at different sparsity obtained by adding a regularization on weights $\propto \sum_\mu (\sum_i |w_{i\mu}|)^x; x \geq 1$. To avoid using a threshold for estimating null activities, use Participation Ratios with clever exponents (see [3], Supp. Material).

Parameters

Control: α, p, g, θ, order: L, m, r, q, B, C.

RBM and R-RBM scaling

Simulate at equilibrium and average result to get $\widehat{L}, \widehat{p}$ scaling to confront with $L \sim \frac{\ell}{p}$ theoretical (b). Normalized magnetization vs ℓ, predicted to be such that decreasing linearly in $\widehat{p}$, i.e. decreasing in $\ell = \widehat{L} \times \widehat{p}$, (c).

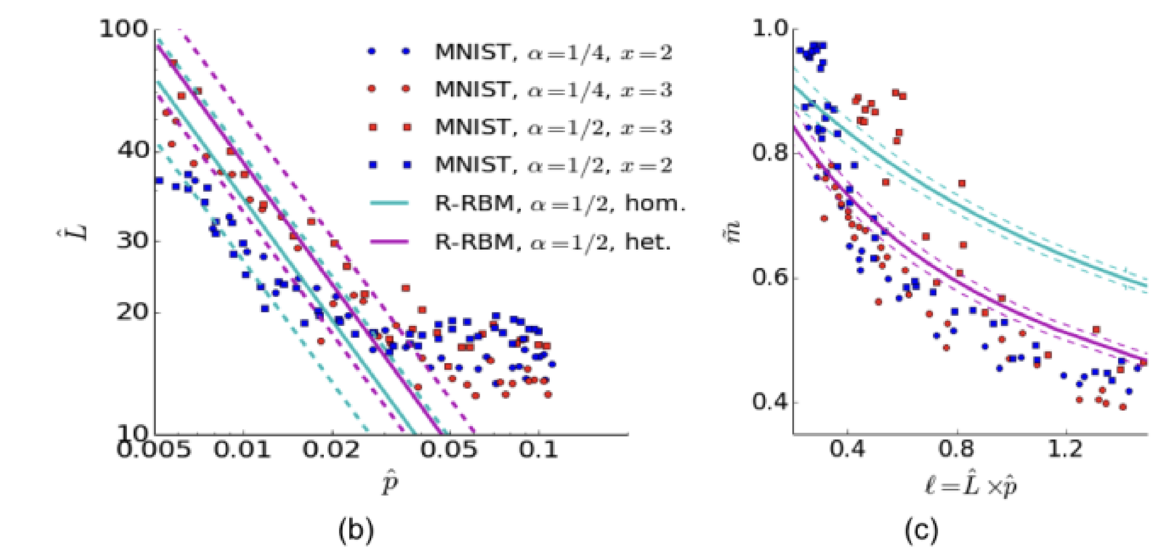

Figure 4. Behavior in accordance with expectations. Source [3]

Difference quantitatively by heterogeneity of number of neighbors of pixels, which justifies using a Heterogeneous model with p_i sparsities fitted from MNIST-trained RBMs. Heterogeneous model is even more in line with simulations.

General Recipe

A **Compositional Phase** for RBMs appears to exist if weights are sparse, escaping the other two phases, and justifying the generative power of the model.

References

[1] Simona Cocco, Rémi Monasson, Lorenzo Posani, Sophie Rosay, and Jérôme Tubiana. Statistical Physics and Representations in Real and Artificial Neural Networks. *Physica A: Statistical Mechanics and its Applications*, 504:45–76, August 2018.

[2] Jérôme Tubiana. *Restricted Boltzmann Machines : From Compositional Representations to Protein Sequence Analysis*. These de doctorat, Paris Sciences et Lettres (ComUE), November 2018.

[3] Jérôme Tubiana and Rémi Monasson. Emergence of Compositional Representations in Restricted Boltzmann Machines. *Physical Review Letters*, 118(13):138301, March 2017.