Cardinality Restricted Boltzmann Machines

Kevin Swersky[1], Daniel Tarlow[1], Ilya Sutskever[1], Ruslan Salakhutdinov[1], Richard S. Zemel[1], and Ryan P. Adams[2]

1. University of Toronto 2. Harvard University

Goal and Motivation

- Develop a principled approach for incorporating sparsity in the hidden units of a Restricted Boltzmann Machine (RBM).
- Motivation: sparse feature representations often lead to better classification performance, robustness, interpretability.
- Problem: inference is seemingly intractable.

The Cardinality RBM (CaRBM)

- Visible units $v \in \{0,1\}^{N_v}$ (can be Gaussian)
- Hidden units $h \in \{0,1\}^{N_h}$
- Weights $W \in \mathbb{R}^{N_v \times N_h}$
- Distribution $P(v,h) = \frac{\exp(-E(v,h))}{\mathcal{Z}}$

$$E(v,h) = \underbrace{-v^\top W h}_{\text{RBM}} - \underbrace{f(\sum_j h_j)}_{\text{cardinality potential}}$$

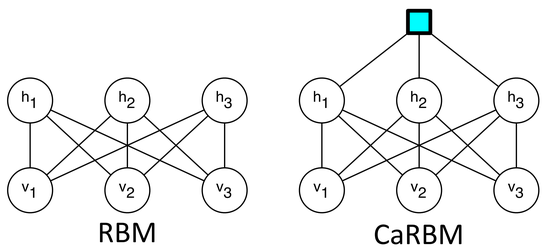
- Trained with contrastive divergence, uses conditional distributions.

	RBM	CaRBM
$P(v\|h)$	$\sigma(Wh)$	$\sigma(Wh)$
$P(h\|v)$	$\sigma(W^\top v)$	$\mu(W^\top v)$

$$\mu_j = P(h_j = 1|v) = \frac{1}{\mathcal{Z}_v} \sum_{\substack{h, \\ h_j=1}} \exp(v^\top W h + f(\sum_{j'} h_{j'}))$$

This work: $f(\sum_j h_j) = \begin{cases} 0 \text{ if } \sum_j h_j \le k \\ -\infty \text{ otherwise} \end{cases}$

- At most k hidden units are allowed to be simultaneously active.
- This is a form of *competition* or *lateral inhibition.*



Inference Over Hidden Variables

Let $x = W^\top v, \quad \phi_j(h_j) = \exp(h_j x_j), \quad \psi(\cdot) = \exp(f(\cdot))$

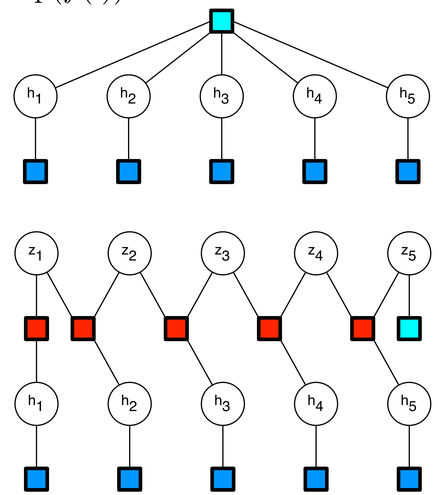
Original Distribution: unary potentials and a global cardinality potential.

$$P(h|v) \propto \psi(\sum_j h_j) \prod_j \phi_j(h_j)$$

Introduce auxiliary z variables to store cumulative sums. Global potential becomes a local potential.

$$P(h,z|v) \propto \psi(z_{N_h}) \prod_{j=1}^{N_h} \phi_j(h_j) \prod_{j=2}^{N_h} \gamma(h_j, z_j, z_{j-1})$$

$$\gamma(h_j, z_j, z_{j-1}) = \begin{cases} 1 \text{ if } z_j = z_{j-1} + h_j \\ 0 \text{ otherwise} \end{cases}$$



- We can compute marginals and sample from $P(h|v)$ using dynamic programming.
- Computational complexity is $O(N_h k)$.

(Gail et al., 1981, Tarlow et al., 2012)

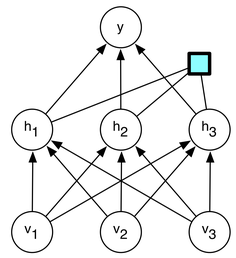
The Cardinality Nonlinearity

- We construct a neural network using μ as the outputs of the hidden layer.
- Corresponds to the *expected value* of a distribution over *sparse vectors*.

Logistic	Softmax	Cardinality
$\sigma(x_j)$	$\frac{\exp(x_j)}{\sum_{j'} \exp(x_{j'})}$	$\mu_j(x)$

- Backpropagate error of cost L using finite differences (Domke, 2010).

$$\frac{\partial L}{\partial x_j} \approx \frac{\mu_j(x + \epsilon \frac{\partial L}{\partial \mu}) - \mu_j(x - \epsilon \frac{\partial L}{\partial \mu})}{2\epsilon}$$

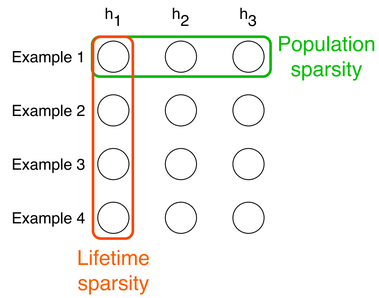


Sparsity Penalty for RBMs (SpRBM)

- Let q be the average activity of a hidden unit across all training examples, and λ be a penalty strength.
- Penalize the KL between q and a desired activation ρ.

$$L = -\log P(v) + \lambda \mathrm{KL}(\rho || q)$$

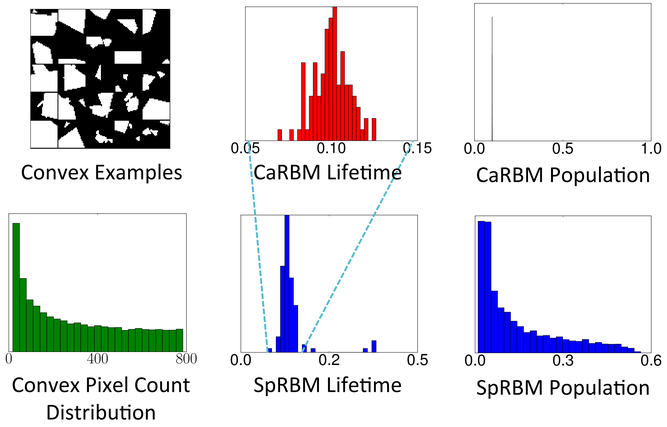
- Encourages *lifetime sparsity*, sparse activations across examples.
- *Population sparsity*: every example is encoded by a sparse vector.
- SpRBM + CaRBM = lifetime sparsity + population sparsity.



(Lee et al., 2007, Nair & Hinton, 2009)

Experiments

Hypothesis: when the data has a highly variable number of active inputs, SpRBM alone is not enough to guarantee population sparsity.



Classification on binary datasets

- CaRBM and SpRBM perform similarly.
- Sparsity generally improves classification performance.

Dataset	RBM	SpRBM	CaRBM	Dataset	RBM	SpRBM	CaRBM
rectangles	4.05%	2.66%	5.60%	convex	20.66%	18.52%	21.13%
background im	23.78%	23.49%	22.16%	mnist basic	4.42%	3.84%	3.65%
background im rot	58.21%	56.48%	56.39%	mnist rot	14.83%	13.11%	12.40%
recangles im	24.24%	22.50%	22.56%	background rand	12.96%	12.97%	12.67%

Topic modeling with the NIPS dataset

- All methods learn meaningful topics.
- Ordinary RBM topics tend to include the most common words.
- SpRBM is sensitive to the KL strength λ, and can end up with *dead examples*, training examples that are ignored by the model.

Model	Computer Vision	Neuroscience	Bayesian Inference
RBM	images, pixel, computer, quickly, stanford	inhibitory, organization, neurons, synaptic, explain	probability, bayesian, priors, likelihood, covariance
SpRBM	visual, object, objects, images, vision	neurons, biology, spike, synaptic, realistic	conditional, probability, bayesian, hidden, mackay
CaRBM	image, images, pixels, objects, recognition	membrane, resting, inhibitory, physiol, excitatory	likelihood, hyperparameters, monte, variational, neal