A Spike and Slab RBM Approach to Modeling Natural Images

Aaron C. Courville AARON.COURVILLE@UMONTREAL.CA
James Bergstra JAMES.BERGSTRA@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
Université de Montréal, Montréal QC H3T 1J4 Canada

The *spike and slab* Restricted Boltzmann Machine (ssRBM) is defined by having both a real valued "slab" variable and a binary "spike" variable associated with each unit in the hidden layer. Earlier work exploring a simple incarnation of the ssRBM model family demonstrated its utility as a model of natural images. In this work, we explore an extension of the ssRBM model – the μ-ssRBM – that includes additional terms to the energy function which we use, in part, to address one of the potential drawback of the original ssRBM introduced in (Courville et al., 2010), specifically, the lack of a guarantee that the model defines a valid density over the whole data domain. We find that while it is possible to parametrize the model to guarantee that all conditionals of the model are well defined, loosening this constraint empirically yields better classification performance with the CIFAR-10 image dataset.

The μ-Spike and Slab RBM

The μ-ssRBM describes the interaction between three random vectors: the visible vector v, the binary "spike" variables h and the real-valued "slab" variables s. In what follows, we assume there are N hidden units: $h \in [0,1]^N$, $s \in \mathbb{R}^N$ and a visible vector of dimension D: $v \in \mathbb{R}^D$. The μ-ssRBM model is defined via the energy function:

$$E(v,s,h) = -\sum_{i=1}^{N} v^T W_i s_i h_i \; + \frac{1}{2} v^T \left(\Lambda + \sum_{i=1}^{N} \Phi_i h_i\right) v + \frac{1}{2}\sum_{i=1}^{N} \alpha_i s_i^2 \; - \sum_{i=1}^{N} \alpha_i \mu_i s_i h_i \; - \sum_{i=1}^{N} b_i h_i + \sum_{i=1}^{N} \alpha_i \mu_i^2 h_i, \tag{1}$$

with model parameters: $(\forall i)$ $W_i \in \mathbb{R}^D$, $b_i \in \mathbb{R}$, $\alpha_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}$ and diagonal matrices $\Lambda \in \mathbb{R}^{D\times D}$ and $\Phi_i \in \mathbb{R}^{D\times D}$. From the energy function it is straightforward to derive a set of conditional distributions $p(s \mid v,h)$ and $p(v \mid s,h)$:

$$p(s \mid v,h) = \prod_{i=1}^{N} \mathcal{N}\left(\left(\alpha_i^{-1} v^T W_i + \mu_i\right) h_i \;,\; \alpha_i^{-1}\right) \tag{2}$$

$$p(v \mid s,h) = \mathcal{N}\left(C_{v|s,h} \sum_{i=1}^{N} W_i s_i h_i, C_{v|s,h}\right) \tag{3}$$

where $\mathcal{N}(\xi, C)$ denotes a Gaussian with mean ξ and covariance C. The covariance of $p(v \mid s,h)$: $C_{v|s,h} = \left(\Lambda + \sum_{i=1}^{N} \Phi_i h_i\right)^{-1}$ is diagonal.

We now consider the effect of marginalizing out the slab variables s to retrieve the traditional RBM conditionals $p(v \mid h)$ and $p(h \mid v)$:

$$p(v \mid h) = \mathcal{N}(C_{v|h} \sum_{i=1}^{N} W_i \mu_i h_i, C_{v|h}) \tag{4}$$

where $C_{v|h} = \left(\Lambda + \sum_{i=1}^{N} \Phi_i h_i - \sum_{i=1}^{N} \alpha_i^{-1} h_i W_i W_i^T\right)^{-1}$, the last equality holds only if the covariance matrix $C_{v|h}$ is positive definite. Note that in marginalizing over s, the visible vector v remains Gaussian distributed but the covariance matrix is no longer diagonal (due to the $\sum_{i=1}^{N} \alpha_i^{-1} h_i W_i W_i^T$ term).

The final conditional that we will consider is $P(h \mid v) = \prod_i^N P(h_i \mid v)$ and

$$P(h_i = 1|v) = \sigma\left(\frac{1}{2}\alpha_i^{-1}(v^T W_i)^2 + v^T W_i \mu_i - \frac{1}{2} v^T \Phi_i v + b_i\right),$$

where σ represents a logistic sigmoid. As with the conditionals $p(v \mid s,h)$ and $p(s \mid v,h)$, the distribution of h given v factorizes over the elements of h. Learning and inference in the μ-ssRBM is rooted in the ability to efficiently draw samples by iteratively sampling from the factorial conditionals $P(h \mid v)$, $p(s \mid v,h)$ and $p(v \mid s,h)$ with a Gibbs sampling procedure. In training the μ-ssRBM, we use persistent contrastive divergence (Tieleman, 2008) to update the model parameters.

Positive Definite Constraints

The conditional $p(v \mid h)$ is only a well defined Gaussian if the covariance matrix $C_{v|h}$ is positive definite (PD). In order to ensure that, we need to con-

Model	Accuracy (%)
no PD, μ free, Φ free	73.10 ± 0.9
no PD, μ free, $\Phi = \mathbf{0}$	71.43 ± 0.9
no PD, $\mu = \mathbf{0}$, Φ free	71.19 ± 0.9
no PD, $\mu = \mathbf{0}$, $\Phi = \mathbf{0}$	68.92 ± 0.9
PD by Diag. W (Eqn. 6)	69.10 ± 0.9
PD by scal. mat. (Eqn. 5)	67.10 ± 0.9

Table 1. The performance of μ-ssRBM variants with 256 hidden. Lines labeled PD correspond to models that were constrained to have "PD" precision of $p(v \mid h)$ while the lines labeled "no DP" are not. The notation $\mu = \mathbf{0}$ corresponds to models trained with μ fixed to zero. 95% confidence intervals are given for each score assuming the official test set size of ten thousand.

strain $\Lambda + \sum_{i=1}^{N} \Phi_i h_i$ to be large enough to offset $\sum_{i=1}^{N} \alpha_i^{-1} W_i W_i^T h_i$, accomplished by constraining Φ_i. We consider two options that are both shown to satisfy the PD constraint. The most straightforward is to restrict Φ_i to a scalar matrix, in which case:

$$\Phi_{ij} = \zeta_{ij} + \alpha^{-1} \sum_j W_{ij}^2 I \tag{5}$$

where W_{ij} is the jth element of filter W_i and Φ_{ij} is the jjth element of diagonal matrix Φ_i. Alternatively, we could choose the following parametrization of Φ_i:

$$\Phi_{ij} = \zeta_{ij} + \alpha^{-1} D W_{ij}^2. \tag{6}$$

Experiments

We evaluate the μ-ssRBM as a feature-extraction algorithm by plugging it into the classification pipeline developed by (Coates et al., 2010). In broad strokes, the μ-ssRBM is fit to whitened ZCA (192-dimensional) representations of 8x8 RGB image patches, and applied convolutionally. Classification was done with an L2-regularized SVM, using $p(h \mid v)$ extracted from every 8x8 image patch.Prior to classification, our conditional h values were spatially pooled into 9 regions.

It is possible to constrain Φ to enforce that $C_{x|h}$ is PD and achieve classification results that matches that of the original (Courville et al., 2010) ssRBM (Table 1). However if we take the same μ-ssRBM form and loosen the PD constraint, the model performs better. Also of note, is that both the μ and the Φ terms seem to contribute approximately equally to improving the classification accuracy.

The best performing μ-ssRBM (1024 units, not 256 like in the table) has **an accuracy of 76.2** ± 0.9, performing better than the most closely-related models in literature (copied from (Ranzato & Hinton, 2010)): the GRBM (59.7 ± 1.0) and mcRBM (225 factors; 68.2 ± 0.9). Recent work by (Coates et al., 2010) has shown that a K-means based approach out-performs these energy-based approaches to feature extraction on CIFAR-10, in the limit of a large hidden layer (4000 units; 79.6 ± 0.9). With a smaller number of elements in the representation, the μ-ssRBM model outperforms the K-means-based approach: μ-ssRBM with 256 units; 73.1 ± 0.9 while K-means with 400 units; 72.7 ± 0.9, and with 200 units 70.1 ± 0.9. We believe the relatively poor performance of the RBM-based models at the large hidden layer limit is due to our inability to effectively train them.

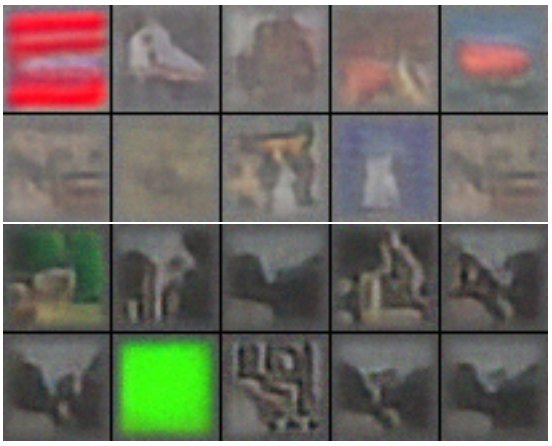

Figure 1. Samples as in (Krizhevsky, 2010) from a convolutionally trained μ-ssRBM exhibit global coherence, and sharp region boundaries, a range of colours, and more natural-looking shading than (Ranzato et al., 2010).

References

Coates, Adam, Lee, Honglak, and Ng, Andrew Y. An analysis of single-layer networks in unsupervised feature learning. NIPS*2010 Workshop on Deep Learning, 2010.

Courville, Aaron, Bergstra, James, and Bengio, Yoshua. Modeling natural image covariance with a spike and slab restricted boltzmann machine. NIPS*2010 Workshop on Deep Learning, 2010.

Krizhevsky, Alex. Convolutional deep belief networks on cifar-10, Aug 2010.

Ranzato, Marc'Aurelio and Hinton, Geoffrey E. Modeling pixel means and covariance using factorized third-order boltzmann machines. In *CVPR'2010*. IEEE Press, 2010.

Ranzato, Marc'Aurelio, Mnih, Volodymyr, and Hinton, Geoffrey. Generating more realistic images using gated MRF's. In *NIPS 23*. 2010.

Tieleman, Tijmen. Training restricted boltzmann machines using approximations to the likelihood gradient. In *ICML 2008*, pp. 1064–1071, 2008.