Exploring Compositional High Order Pattern Potentials for Structured Output Learning

Supplementary Material

Yujia Li, Daniel Tarlow, Richard Zemel
University of Toronto
Toronto, ON, Canada, M5S 3G4
{yujiali, dtarlow, zemel}@cs.toronto.edu

1. Equating Pattern Potentials and RBMs

This section provides the detailed proof of the equivalence between pattern potentials and RBMs. The high level idea of the proof is to treat each hidden variable in an RBM as encoding a pattern.

We first introduce the definition of pattern potentials by Rother et al. in [1], a few necessary change of variable tricks, and two different ways to compose more general high order potentials, "sum" and "min".

Then we relates the composite pattern potentials to RBMs. We show in Section 1.2 that *minimizing out* hidden variables in RBMs are equivalent to pattern potentials. When there are no constraints on hidden variables, we recover the "sum" composite pattern potentials; when there is a 1-of-J constraint on hidden variables, we recover the "min" composite pattern potentials. In Section 1.3, we show that *summing out* hidden variables in RBMs approximates pattern potentials, and similarly with and without constraints on hidden variables would lead us to "min" and "sum" cases respectively.

The RBM formulation offers considerable generality via choices about how to constrain hidden unit activations. This allows a smooth interpolation between the "sum" and "min" composition strategies. Also, this formulation allows the application of learning procedures that are appropriate for cases other than just the "min" composition strategy.

In Section 2, we provide a way to unify minimizing out hidden variables and summing out hidden variables by introducing a temperature parameter in the model.

Notation. In this section, we use g for pattern potentials and $\hat{g}$ for the high order potentials induced by RBMs. Superscripts 's' and 'm' on g corresponds to two composition schemes, sum and min. Superscripts on $\hat{g}$ correspond to two types of constraints on RBM hidden variables, and subscripts on $\hat{g}$ correspond to minimizing out or summing out hidden variables.

1.1. Pattern potentials

In [1], a basis pattern potential for a clique of binary variables $\mathbf{y}_a$ is defined as

$$g(\mathbf{y}_a) = \min\{d(\mathbf{y}_a) + \theta_0, \theta_{\max}\} \tag{1}$$

where $d : \{0,1\}^{|a|} \to \mathbb{R}$ is a deviation function specifying the penalty for deviating from a specific pattern. The pattern potential penalizes configurations of $\mathbf{y}_a$ that deviates from the pattern, and the penalty is upper bounded by $\theta_{\max}$ while θ_0 is a base penalty.

For a specific pattern $\mathbf{Y}$, the deviation function $d(\mathbf{y}_a)$ is defined as[1]

$$d(\mathbf{y}_a) = \sum_{i \in a} abs(w_i)(y_i \neq \mathbf{Y}_i) \tag{2}$$

where $abs()$ is the absolute value function. This is essentially a weighted hamming distance of $\mathbf{y}_a$ from $\mathbf{Y}$. Since $\mathbf{y}_a$ and $\mathbf{Y}$ are both binary vectors, we have the following alternative formulation

$$\begin{aligned} d(\mathbf{y}_a) &= \sum_{i \in a: \mathbf{Y}_i = 1} (-w_i)(1 - y_i) + \sum_{i \in a: \mathbf{Y}_i = 0} w_i y_i \\ &= \sum_{i \in a} w_i y_i + \sum_{i \in a: \mathbf{Y}_i = 1} (-w_i) \end{aligned} \tag{3}$$

w_i specifies the cost of assigning y_i to be 1. $w_i > 0$ when $\mathbf{Y}_i = 0$ and $w_i < 0$ when $\mathbf{Y}_i = 1$.

We can subtract constant $\theta_{\max}$ from Eq. 1 to get

$$g(\mathbf{y}_a) = \min\left\{\sum_{i \in a} w_i y_i + \sum_{i \in a: \mathbf{Y}_i = 1} (-w_i) - \theta, 0\right\} \tag{4}$$

[1] Note that in [1], there is also a factor θ in this definition ($d(\mathbf{y}_a)$ is given by the product of factor θ and the sum), but actually the θ factor can always be absorbed in w_i's to get this equivalent formulation.

Making the change of variables $w_i' = -w_i$, $c = \theta + \sum_{i \in a: \mathbf{Y}_i = 1} w_i$, we can rewrite the above equation as

$$g(\mathbf{y}_a) = \min \left\{ -c - \sum_{i \in a} w_i' y_i, 0 \right\} \quad (5)$$

This formulation is useful for establishing connections with RBMs as shown later in this section.

[1] proposed two ways to compose more general high order potentials from basis pattern potentials defined above. One is to take the sum of different pattern potentials

$$\begin{aligned} g^s(\mathbf{y}_a) &= \sum_{j=1}^{J} \min\{d_j(\mathbf{y}_a) + \theta_j, \theta_{\max}\} \\ &= \sum_{j=1}^{J} \min\{d_j(\mathbf{y}_a) + \theta_j', 0\} + \text{const} \end{aligned} \quad (6)$$

and the other is to take the minimum of them, to get

$$g^m(\mathbf{y}_a) = \min_{1 \le j \le J} \{d_j(\mathbf{y}_a) + \theta_j\} \quad (7)$$

In both cases, $d_j(.)$'s are J different deviation functions, and θ_j's are base penalties for different patterns. In the "min" case, we can also fix one deviation function to be 0 (i.e. by setting all weights $w_i = 0$), to get a constant threshold.

Using the change of variable tricks introduced above, we can rewrite the "sum" composite pattern potential as

$$g^s(\mathbf{y}_a) = \sum_{j=1}^{J} \min \left\{ -c_j - \sum_{i \in a} w_{ij} y_i, 0 \right\} \quad (8)$$

where we ignored the constant term, and rewrite the "min" composite pattern potential as

$$g^m(\mathbf{y}_a) = \min_{1 \le j \le J} \left\{ -c_j - \sum_{i \in a} w_{ij} y_i \right\} \quad (9)$$

Since we always work on a clique of variables in this section, we drop the subscript a on $\mathbf{y}$ for the rest of this section.

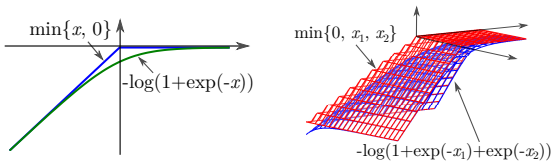
1.2. Minimizing out hidden variables in RBMs

We start from minimizing hidden variables out. The probability distribution defined by a binary RBM is given by

$$p(\mathbf{y}, \mathbf{h}) = \frac{1}{Z} \exp\left(-E(\mathbf{y}, \mathbf{h})\right) \quad (10)$$

where the energy

$$E(\mathbf{y}, \mathbf{h}) = -\sum_{i=1}^{I} \sum_{j=1}^{J} w_{ij} y_i h_j - \sum_{i=1}^{I} b_i y_i - \sum_{j=1}^{J} c_j h_j \quad (11)$$



(a) (b)

Figure 1. (a) $-\log(1+\exp(-x))$ is a smoothed approximation to $\min\{x, 0\}$; (b) $-\log(1+\exp(-x_1)+\exp(-x_2))$ is a smoothed approximation to $\min\{x_1, x_2, 0\}$.

Minimizing out the hidden variables, the equivalent high order potential is

$$\hat{g}_{\min}(\mathbf{y}) = \min_{\mathbf{h}} \left\{ -\sum_{j=1}^{J} \left(c_j + \sum_{i=1}^{I} w_{ij} y_i \right) h_j \right\} \quad (12)$$

When there is no constraint on hidden variables, i.e. they are independent binary variables, the minimization can be factorized and moved inside the sum

$$\hat{g}^{\text{uc}}_{\min}(\mathbf{y}) = \sum_{j=1}^{J} \min \left\{ -c_j - \sum_{i=1}^{I} w_{ij} y_i, 0 \right\} \quad (13)$$

The superscript "uc" is short for "unconstrained". This is exactly the same as the "sum" composite pattern potentials in Eq. 8.

When we put a 1-of-J constraint on hidden variables, i.e. forcing $\sum_{j=1}^{J} h_j = 1$, the minimization becomes

$$\hat{g}^{\text{1of}J}_{\min}(\mathbf{y}) = \min_{1 \le j \le J} \left\{ -c_j - \sum_{i=1}^{I} w_{ij} y_i \right\} \quad (14)$$

This is exactly the same as the "min" composite pattern potentials in Eq. 9.

1.3. Summing out hidden variables in RBMs

The key observation that relates the pattern potentials and RBMs with hidden variables summed out is the following approximation,

$$\min\{x, 0\} \approx -\log(1 + \exp(-x)) \quad (15)$$

It is easy to see that when x is a large positive value, the right hand side will be close to 0 and when x is a large negative value, the right hand side will be linear in x. This is illustrated in Fig 1 (a).

With this approximation, we can rewrite the basis pattern potential in Eq. 5 as

$$g(\mathbf{y}) \approx -\log\left(1 + \exp\left(c + \sum_{i=1}^{I} w_i' y_i\right)\right) \quad (16)$$

On the other hand, summing out hidden variables in an RBM with no constraints on hidden variables, the marginal distribution becomes

$$p(\mathbf{y}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{I} b_i y_i\right) \prod_{j=1}^{J} \left(1 + \exp\left(c_j + \sum_{i=1}^{I} w_{ij} y_i\right)\right) \tag{17}$$

Eq. 5 in the main paper is another equivalent form of this. Therefore the equivalent high order potential induced by summing out the hidden variables is

$$\hat{g}^{\text{uc}}_{\text{sum}}(\mathbf{y}) = -\sum_{j=1}^{J} \log\left(1 + \exp\left(c_j + \sum_{i=1}^{I} w_{ij} y_i\right)\right) \tag{18}$$

which is exactly a sum of potentials in the form of Eq. 16.

Now we turn to the "min" case. We show that the composite pattern potentials are equivalent to RBMs with a 1-of-J constraint on hidden variables and hidden variables summed out, up to the following approximation

$$\min\{x_1, x_2, ..., x_J, 0\} \approx -\log\left(1 + \sum_{j=1}^{J} \exp(-x_j)\right) \tag{19}$$

This is a high dimensional extension to Eq. 15. The 2-D case is illustrated in Fig 1 (b).

We use the definition of "min" composite pattern potentials in Eq. 7, but fix $d_J(\mathbf{y})$ to be 0, to make a constant threshold on the cost.

Then we can subtract constant θ_J from the potential and absorb θ_J into all other θ_j's (with the same change of variable tricks) to get

$$g^m(\mathbf{y}) = \min\left\{-c_1 - \sum_{i=1}^{I} w_{i1} y_i, ..., -c_{J-1} - \sum_{i=1}^{I} w_{i,J-1} y_i, 0\right\} \tag{20}$$

Using the approximation, this high order potential becomes

$$g^m(\mathbf{y}) \approx -\log\left(1 + \sum_{j=1}^{J-1} \exp\left(c_j + \sum_{i=1}^{I} w_{ij} y_i\right)\right) \tag{21}$$

In an RBM with J hidden variables, the 1-of-J constraint is equivalent to $\sum_{j=1}^{J} h_j = 1$. With this constraint, the energy (Eq. 11) can be transformed into

$$\begin{aligned} E(\mathbf{y}, \mathbf{h}) &= -\sum_{i=1}^{I} b_i y_i - \sum_{j=1}^{J-1} \left(c_j - c_J + \sum_{i=1}^{I} (w_{ij} - w_{iJ}) y_i\right) h_j \\ &\quad - \left(c_J + \sum_{i=1}^{I} w_{iJ} y_i\right) \\ &= -\sum_{i=1}^{I} (b_i - w_{iJ}) y_i \\ &\quad - \sum_{j=1}^{J-1} \left(c_j - c_J + \sum_{i=1}^{I} (w_{ij} - w_{iJ}) y_i\right) - c_J \end{aligned} \tag{22}$$

We can therefore use a new set of parameters $b'_i = b_i - w_{iJ}$, $c'_j = c_j - c_J$ and $w'_{ij} = w_{ij} - w_{iJ}$, and get

$$E(\mathbf{y}, \mathbf{h}) = -\sum_{i=1}^{I} b'_i y_i - \sum_{j=1}^{J-1} \left(c'_j + \sum_{i=1}^{I} w'_{ij} y_i\right) h_j \tag{23}$$

We ignored the constant c_J because it would cancel out when we normalize the distribution. Note that now the set of $J-1$ hidden variables can have at most one on, and they can also be all off, corresponding to the case that the Jth hidden variable is on.

Summing out $\mathbf{h}$, we get

$$p(\mathbf{y}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{I} b'_i y_i\right) \left(1 + \sum_{j=1}^{J-1} \exp\left(c'_j + \sum_{i=1}^{I} w'_{ij} y_i\right)\right) \tag{24}$$

The constant 1 comes from the Jth hidden variable. The equivalent high-order potential for this model is then

$$\hat{g}^{\text{1of}J}_{\text{sum}}(\mathbf{y}) = -\log\left(1 + \sum_{j=1}^{J-1} \exp\left(c'_j + \sum_{i=1}^{I} w'_{ij} y_i\right)\right) \tag{25}$$

which has exactly the same form as Eq. 21.

Our results in this section are summarized in Table 1.

2. The CHOPP

We define the CHOPP as

$$f(\mathbf{y}; T) = -T \log\left(\sum_{\mathbf{h}} \exp\left(\frac{1}{T} \sum_{j=1}^{J} \left(c_j + \sum_{i=1}^{I} w_{ij} y_i\right) h_j\right)\right) \tag{26}$$

where T is the temperature parameter. Summation over $\mathbf{h}$ is a sum over all possible configurations of hidden variables.

Setting $T = 1$, this CHOPP becomes

$$f(\mathbf{y}; 1) = -\log\left(\sum_{\mathbf{h}} \exp\left(\sum_{j=1}^{J} \left(c_j + \sum_{i=1}^{I} w_{ij} y_i\right) h_j\right)\right) \tag{27}$$

this is the equivalent RBM high order potential with hidden variables summed out. When there is no constraint on $\mathbf{h}$, the above potential becomes

$$f(\mathbf{y}; 1) = -\sum_{j=1}^{J} \log\left(1 + \exp\left(c_j + \sum_{i=1}^{I} w_{ij} y_i\right)\right) \tag{28}$$

When there is a 1-of-J constraint on $\mathbf{h}$, the above potential is

$$f(\mathbf{y}; 1) = -\log\left(\sum_{j=1}^{J} \exp\left(c_i + \sum_{i=1}^{I} w_{ij} y_i\right)\right) \tag{29}$$

Composition Scheme for Pattern Potentials	Operation on RBM Hidden Variables (T axis)		Constraint on $\mathbf{h}$ (sparsity axis)
	Minimizing out $\mathbf{h}$ $T \to 0$	Summing out $\mathbf{h}$ $T = 1$	
Min	$\min_{1 \le j \le J} \left\{ -c_j - \sum_{i=1}^{I} w_{ij} y_i \right\}$	$-\log \left(1 + \sum_{j=1}^{J-1} \exp \left(c_j + \sum_{i \in a} w_{ij} y_i \right) \right)$	1-of-J
Sum	$\sum_{j=1}^{J} \min \left\{ -c_j - \sum_{i=1}^{I} w_{ij} y_i, 0 \right\}$	$-\sum_{j=1}^{J} \log \left(1 + \exp \left(c_j + \sum_{i=1}^{I} w_{ij} y_i \right) \right)$	None

Table 1. Equivalent compositional high order potentials by applying different operations and constraints on RBMs. Minimizing out hidden variables results in high order potentials that are exactly equivalent to pattern potentials. Summing out hidden variables results in approximations to pattern potentials. 1-of-J constraint on hidden variables corresponds to the "min" compositional scheme. No constraints on hidden variables corresponds to "sum" compositional scheme. Corresponding temperature T in the CHOPP is also shown in the table.

Setting $T \to 0$, the CHOPP becomes

$$f(\mathbf{y}; 0) = \min_{\mathbf{h}} \left\{ -\sum_{j=1}^{J} \left(c_j + \sum_{i=1}^{I} w_{ij} y_i \right) h_j \right\} \qquad (30)$$

this is exactly the same as the high order potential induced by a RBM with hidden variables minimized out, and therefore equivalent to composite pattern potentials as shown in Section 1.2. When there are no constraints on hidden variables we will get the "sum" composite pattern potentials, while adding a 1-of-J constraint will give us the "min" composite pattern potentials.

Therefore, by using a temperature parameter T, CHOPPs can smoothly interpolate summing out hidden variables (usually used in RBMs) and minimizing out hidden variables (used in Rother et al.[1]). On the other hand, by using sparsity (the 1-of-J constraint), it interpolates the "sum" and "min" composition schemes.

Note that all experiments in the paper are done with $T = 1$. It would be interesting to try other temperature settings, which corresponds to operation on $\mathbf{h}$ in between summing out and minimize out.

3. Remark on LP Relaxation Inference

After summing out hidden variables when there are no sparsity constraints, the remaining energy function has a sum over J terms, one per hidden unit. It is possible to view each of these terms as a high order potential, then to use modern methods for MAP inference based on linear program (LP) relaxations [2]. In fact, we tried this approach, formulating the "marginal MAP" problem as simply a MAP problem with high order potentials, then using Dual Decomposition to solve the LP relaxation. The key computational requirement is a method for finding the minimum free energy configuration of visibles for an RBM with a single hidden unit, which we were able to do efficiently. However, we found that the energies achieved by this approach were worse than those achieved by the EM procedure described above. We attribute this to looseness in the resulting LP relaxation. This hypothesis is also supported by the results reported by Rother et al. [1], where ordinary belief propagation outperformed LP-based inference, which tends to occur when LP relaxations are loose. Going forward, it would be worthwhile to explore methods for tightening LP relaxations [3].

4. Convolutional Structures

We explored the convolutional analog to RBMs in our experiments. We tried two variants: (a) a vanilla pre-trained convolutional RBM, and (b) a pre-trained convolutional RBM with conditional hidden biases as described in Section 4.1 in the paper. We tried two different patch sizes (8×8, 12×12) and tiled the images densely. Though the conditional variant outperformed the unconditional variant, overall results were discouraging—performance was not even as good as the simple Unary+Pairwise model. This is surprising because a convolutional RBM should in theory be able to easily represent pairwise potentials, and convolutional RBMs have fewer parameters than their global counterparts, so overfitting should not be an issue. We believe the explanation for the poor performance is that learning methods for convolutional RBMs are not nearly as evolved as methods for learning ordinary RBMs, and thus the learning methods that we have at our disposal do not perform as well.

On the bright side, this can be seen as a challenge to overcome in future work.

5. Real Data Sets

Images in the three real data sets are shown in Fig. 2 and Fig. 3.

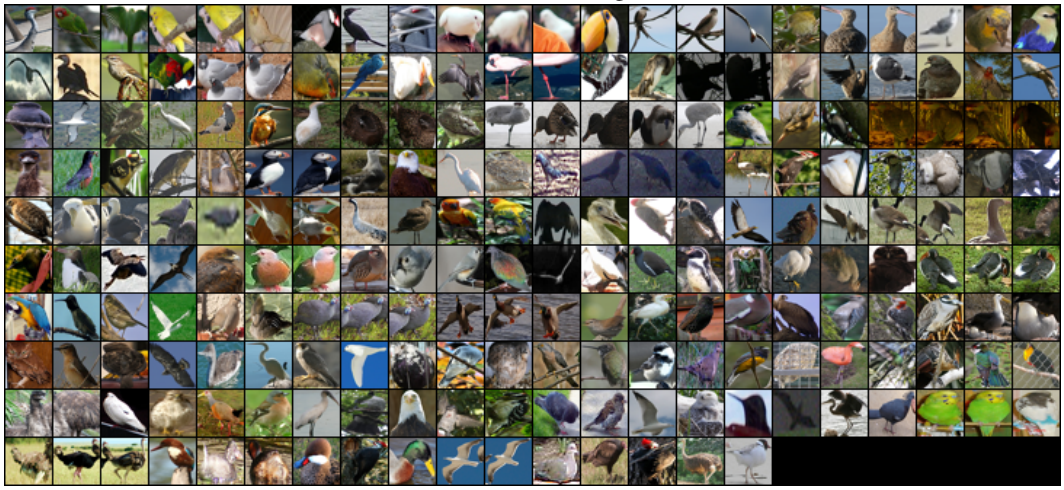
You can find the original Weizmann horses data set from http://www.msri.org/people/members/eranb/ and PASCAL VOC data set from http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/.

Our version of the three data sets as well as the 6 synthetic data sets will be available online.

6. Learned Filters

The learned filters, i.e. weights w_{ij}, with a pretrained RBM for each of the three data sets, are shown in Fig. 4. Filters for 6 synthetic data sets are shown in Fig. 5 and Fig. 6.

(a) Horse data set, 328 images in total.



(b) Bird data set, 224 images in total.

Figure 2. Horse and bird data sets.

For each filter, the weights are positive for bright regions and negative for dark regions. In other words, filters favor bright regions to be on and dark regions to be off.

We can see the compositional nature of RBMs from these filters. For example, each single horse filter is actually expressing soft rules like "if the head of a horse is here, then the legs are likely to be there". Any single filter would not make too much sense, but only when a few different filters are combined can we recover a horse.

7. Prediction Results

Some example segmentations for horse, bird and person data sets are given in Fig. 7, Fig. 8 and Fig. 9.

References

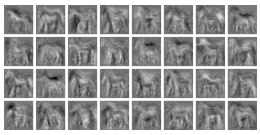
[1] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, 2009. 1, 2, 4

[2] D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. In S. Sra, S. Nowozin,

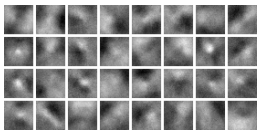
Figure 3. Person data set.

and S. J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011. 4

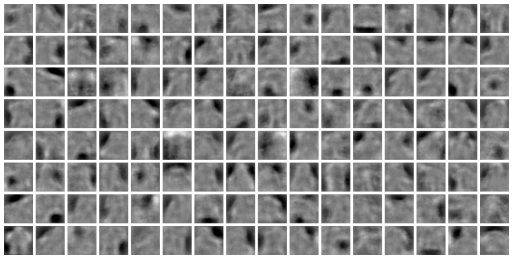
[3] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening lp relaxations for MAP using message passing. In *UAI*, 2008. 4



(a) Horse filters

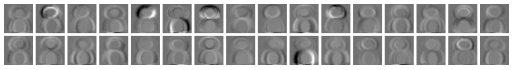


(b) Bird filters

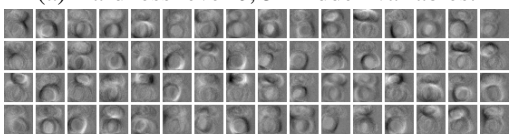


(c) Person filters.

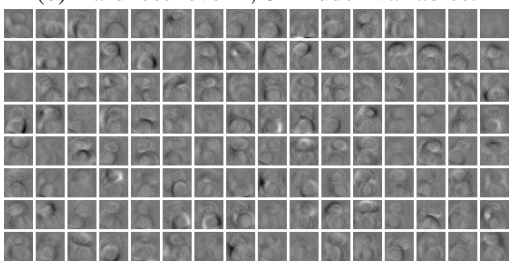
Figure 4. Filters learned on three real data sets.



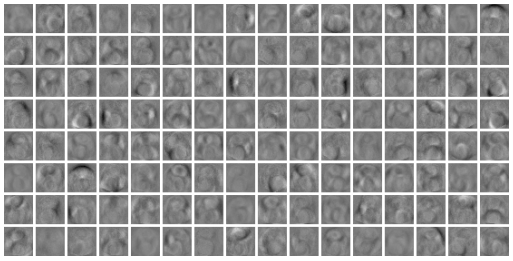
(a) Hardness level 0, 32 hidden variables.



(b) Hardness level 1, 64 hidden variables.

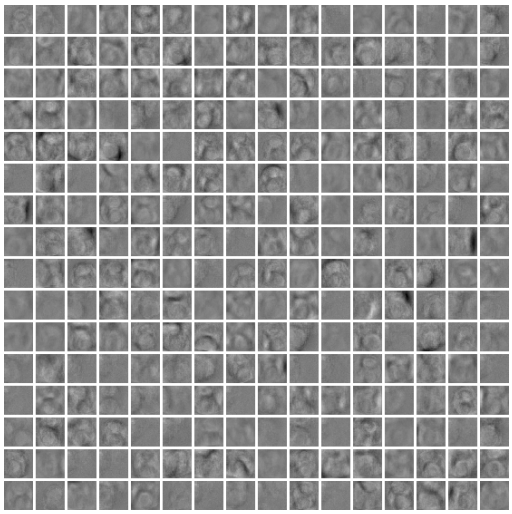


(c) Hardness level 2, 128 hidden variables.

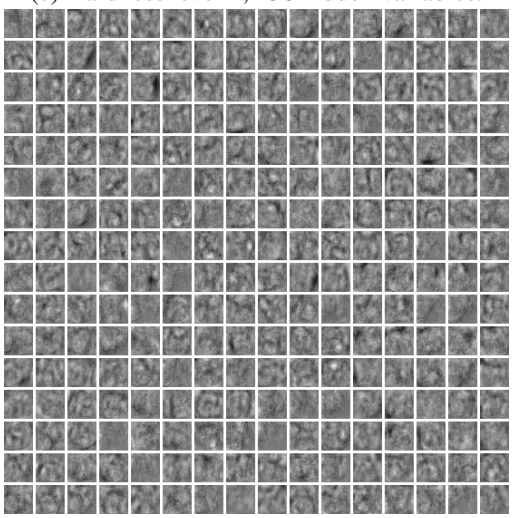


(d) Hardness level 3, 128 hidden variables.

Figure 5. Filters learned on synthetic data sets.



(e) Hardness level 4, 256 hidden variables.

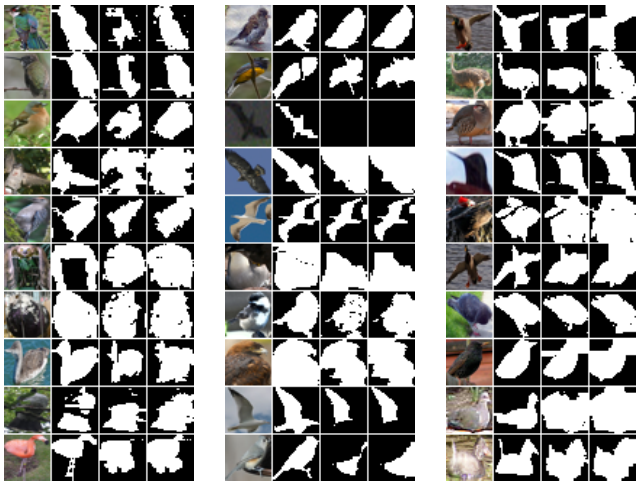


(f) Hardness level 5, 256 hidden variables.

Figure 6. Filters learned on synthetic data sets, continued.

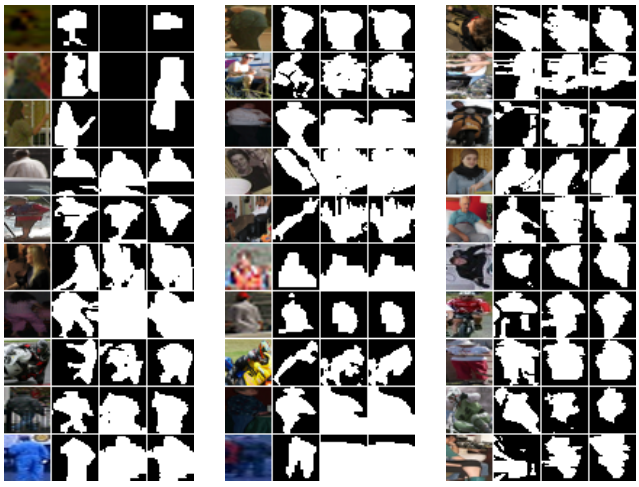
(a) Best (b) Average (c) Worst

Figure 7. Prediction results on horse data set. The three categories best, average and worst are measured by the improvement of Unary+Pairwise+RBM over Unary+Pairwise. Each row left to right: original image, ground truth, Unary+Pairwise prediction, Unary+Pairwise+RBM prediction.



(a) Best (b) Average (c) Worst

Figure 8. Prediction results on bird data set.



(a) Best (b) Average (c) Worst

Figure 9. Prediction results on person data set.