Charlie Tang
Department of Computer Science, University of Toronto, Canada

Gated Boltzmann Machine for Recognition under Occlusion

Introduction

- Unconstrained real world environments are often full of clutter
- Deep Boltzmann Machines (DBMs) are good at generative modeling of objects
- We extend the DBM architecture to explicitly handle occlusion:
  1. Indicator variables are introduced to represent the occluder
  2. Inference tries to infer both the object and the occluder
  3. Learned occluder model can be easily combined with other object models, e.g. faces

Denoising Gated Boltzmann Machine

Formulation

Energy for a Deep Boltzmann Machine (biases omitted):

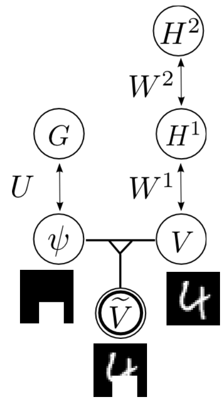
$$E_{DBM}(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2) = -\mathbf{v}^\mathsf{T}\mathbf{W}^1\mathbf{h}^1 - (\mathbf{h}^1)^\mathsf{T}\mathbf{W}^2\mathbf{h}^2$$

$$p(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2) = \frac{p^*(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2)}{Z(\theta)} = \frac{\exp^{-E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2)}}{Z(\theta)}$$

The DGBM is still an undirected graphical model defined by energy:

$$E_{DGBM} = E_{DBM} - \psi^\mathsf{T}\mathbf{U}\mathbf{g} + \textstyle\sum_i^D \gamma_i\psi_i \log(1 + (v_i - \tilde{v}_i)^2)$$

$$\mathbf{v}, \tilde{\mathbf{v}}, \psi \in \{0, 1\}^D$$



Inference

Conditional distribution of interest: $p(\mathbf{v}, \psi | \mathbf{h}^1, \mathbf{g}, \tilde{\mathbf{v}})$

We can compute the energy for all 4 states:

$$E(v_i = 0, \psi_i = 0) = 0$$
$$E(v_i = 0, \psi_i = 1) = \gamma_i \log(1 + \tilde{v}_i^2) - \textstyle\sum_k U_{ik} g_k$$
$$E(v_i = 1, \psi_i = 0) = -\textstyle\sum_j W_{ij}^1 h_j^1$$
$$E(v_i = 1, \psi_i = 1) = \gamma_i \log(1 + (1 - \tilde{v}_i)^2) - \textstyle\sum_k U_{ik} g_k - \textstyle\sum_j W_{ij}^1 h_j^1$$

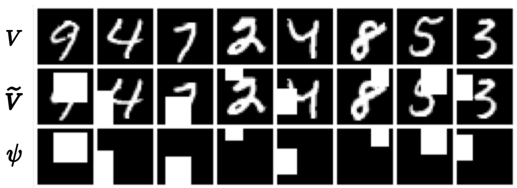
Learning

objective: $\max_\theta \frac{1}{N} \log p(\mathbf{v}, \tilde{\mathbf{v}}, \psi)$

data: $\{(\mathbf{v}_1, \tilde{\mathbf{v}}_1, \psi_1), \ldots, (\mathbf{v}_N, \tilde{\mathbf{v}}_N, \psi_N)\}$

gradient of log-likelihood:

$$\frac{\partial l(\theta)}{\partial \theta} = -\left\langle \frac{\partial E_{DGBM}}{\partial \theta} \right\rangle_{data} + \left\langle \frac{\partial E_{DGBM}}{\partial \theta} \right\rangle_{model}$$



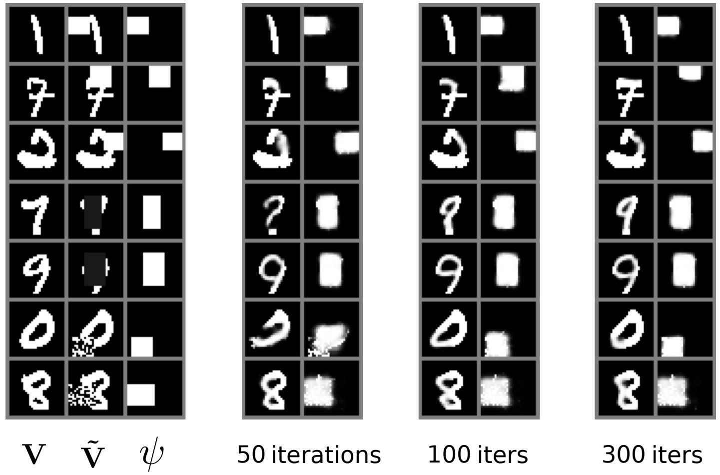
Variational approximation using mean-field iterations for expectation over the data

Persistent Contrastive Divergence used for expectation over the model

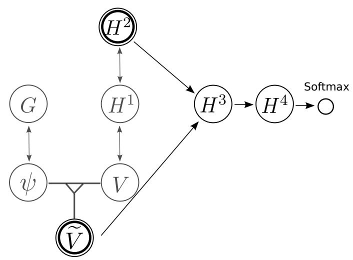
W and U can be pretrained with "clean" object images and occluder images

Experiments

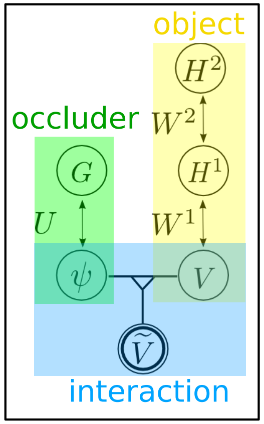
Denoising using Gibbs sampling   network size: [784-500-1000]



Recognition



- *For recognition, 3 additional layers are added with a softmax output on top*
- *150 Gibbs iterations was run to determine the activation of H^2 of the DGBM*
- *The activity of $\tilde{V}$ and H^2 are concatenated to form the input to H^3*
- *30 epochs of nonlinear Conjugate gradient method is used to fine-tune the weights and biases of the net while fixing $\tilde{V}$ and H^2*
- *Test error of* ***6.44%*** *on MNIST+occlusion; compared to 7.49% for the 2 layer DBM; and 8.39% for sparsely connected DBN*



Conclusions

- DGBM uses 3rd order interactions to specify the image pixels which are occluded
- Simple Gibbs sampling is able to infer the occluder and object during recognition
- Recognition using denoised hidden representation achieves lower error
- Occluder model learns the shape but not the appearance of the occluder
- Interaction weights can be easily combined with other object and occluder models, e.g. faces with sunglasses as occluders