Supplementary Material: Learning to combine foveal glimpses with a third-order Boltzmann machine

Hugo Larochelle and Geoffrey Hinton
Department of Computer Science, University of Toronto
6 King's College Rd, Toronto, ON, Canada, M5S 3G4
{larocheh,hinton}@cs.toronto.edu

Abstract

We provide here additional details relatively to our paper.

1 Reminder of properties of multi-fixation RBM

Here we write down explicitly the main properties of the multi-fixation RBM, specifically the form of its conditional distributions:

$$\begin{aligned}
p(\mathbf{h}|\mathbf{y},\mathbf{x}) &= \prod_j p(h_j|\mathbf{y},\mathbf{x}) \\
p(h_j = 1|\mathbf{y},\mathbf{x}_{1:K}) &= \mathrm{sigm}(c_j + \mathbf{U}_{j\cdot}\mathbf{y} + \sum_{k=1}^{K} \mathbf{P}_{j\cdot}\ \mathrm{diag}(\mathbf{z}(i_k,j_k))\ \mathbf{F}\ \mathbf{x}_k) \\
p(\mathbf{x}_k|\mathbf{h}) &= \prod_i p(x_{ki}|\mathbf{h}) \quad \forall\, k \in \{1,\ldots,K\} \\
p(x_{ki} = 1|\mathbf{h}) &= \mathrm{sigm}(b_i + \mathbf{h}^\top \mathbf{P}\ \mathrm{diag}(\mathbf{z}(i_k,j_k))\ \mathbf{F}_{\cdot i}) \quad \forall\, k \in \{1,\ldots,K\} \\
p(\mathbf{y} = \mathbf{e}_l|\mathbf{h}) &= \frac{\exp(d_l + \mathbf{h}^\top \mathbf{U}_{\cdot l})}{\sum_{l^*=1}^{C} \exp(d_{l^*} + \mathbf{h}^\top \mathbf{U}_{\cdot l^*})} \\
p(\mathbf{y} = \mathbf{e}_l|\mathbf{x}_{1:K}) &= \frac{\exp(d_l + \sum_j \mathrm{softplus}(c_j + U_{jl} + \sum_{k=1}^{K} \mathbf{P}_{j\cdot}\ \mathrm{diag}(\mathbf{z}(i_k,j_k))\ \mathbf{F}\ \mathbf{x}_k))}{\sum_{l^*=1}^{C} \exp(d_{l^*} + \sum_j \mathrm{softplus}(c_j + U_{jl^*} + \sum_{k=1}^{K} \mathbf{P}_{j\cdot}\ \mathrm{diag}(\mathbf{z}(i_k,j_k))\ \mathbf{F}\ \mathbf{x}_k))}
\end{aligned}$$

where each glimpse $\mathbf{x}_k$ is a binary vector.

2 Detailed description of the hybrid cost gradient

We start with the hybrid cost:

$$\textbf{Hybrid cost:} \qquad \mathcal{C}_{\mathrm{hybrid}} = -\log p(\mathbf{y}^t|\mathbf{x}^t_{1:K}) - \alpha \log p(\mathbf{y}^t, \mathbf{x}^t_{1:K})\ . \tag{1}$$

The gradient with respect to any parameter θ has the following simple form:

$$\begin{aligned}
\frac{\partial \mathcal{C}_{\mathrm{hybrid}}}{\partial \theta} &= \mathrm{E}_{\mathbf{h}|\mathbf{y}^t,\mathbf{x}^t_{1:K}}\left[\frac{\partial}{\partial \theta} E(\mathbf{y}^t, \mathbf{x}^t_{1:K}, \mathbf{h})\right] - \mathrm{E}_{\mathbf{y},\mathbf{h}|\mathbf{x}^t_{1:K}}\left[\frac{\partial}{\partial \theta} E(\mathbf{y}, \mathbf{x}^t_{1:K}, \mathbf{h})\right] \\
&\quad + \alpha\left(\mathrm{E}_{\mathbf{h}|\mathbf{y}^t,\mathbf{x}^t_{1:K}}\left[\frac{\partial}{\partial \theta} E(\mathbf{y}^t, \mathbf{x}^t_{1:K}, \mathbf{h})\right] - \mathrm{E}_{\mathbf{y},\mathbf{x}_{1:K},\mathbf{h}}\left[\frac{\partial}{\partial \theta} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h})\right]\right)
\end{aligned}$$

Algorithm 1 Gibbs sampling in Contrastive Divergence, to obtain samples $\mathbf{x}_{1:K}^{\text{neg}}$ and $\mathbf{y}^{\text{neg}}$ for the multi-fixation RBM, for the hybrid cost

Input: training pair $(\mathbf{y}^t, \mathbf{x}_{1:K}^t)$
% Notation: $a \sim p$ means a is sampled from p
$\mathbf{h}^{\text{neg}} \sim p(\mathbf{h}|\mathbf{y}^t, \mathbf{x}_{1:K}^t)$
$\mathbf{y}^{\text{neg}} \sim p(\mathbf{y}|\mathbf{h}^{\text{neg}})$
for k from 1 to K **do**
  $\mathbf{x}_k^{\text{neg}} \sim p(\mathbf{x}_k|\mathbf{h}^{\text{neg}})$
end for

The expectations with respect to $\mathbf{h}$ only are tractable. Because $\mathbf{h}$ is binary and the energy function is linear in $\mathbf{h}$, we have that

$$\mathrm{E}_{\mathbf{h}|\mathbf{y}^t,\mathbf{x}_{1:K}^t}\left[\frac{\partial}{\partial\theta}E(\mathbf{y}^t,\mathbf{x}_{1:K}^t,\mathbf{h})\right] = \frac{\partial}{\partial\theta}E(\mathbf{y}^t,\mathbf{x}_{1:K}^t,\mathbf{h}(\mathbf{y}^t,\mathbf{x}_{1:K}^t))$$

where we defined

$$\mathbf{h}(\mathbf{y}^t,\mathbf{x}_{1:K}^t) = \mathrm{sigm}\left(\mathbf{c} + \mathbf{U}\mathbf{y}^t + \sum_{k=1}^{K}\mathbf{P}\,\mathrm{diag}(\mathbf{z}(i_k,j_k))\,\mathbf{F}\,\mathbf{x}_k^t\right) .$$

In other words, the stochastic value of $\mathbf{h}$ is simply replaced by its expectation given $\mathbf{y}^t$ and $\mathbf{x}_{1:K}^t$.

The expectation with respect to $\mathbf{y}$ and $\mathbf{h}$ can also be done exactly:

$$\begin{aligned}
\mathrm{E}_{\mathbf{y},\mathbf{h}|\mathbf{x}_{1:K}^t}\left[\frac{\partial}{\partial\theta}E(\mathbf{y},\mathbf{x}_{1:K}^t,\mathbf{h})\right] &= \mathrm{E}_{\mathbf{y}|\mathbf{x}_{1:K}^t}\left[\mathrm{E}_{\mathbf{h}|\mathbf{y},\mathbf{x}_{1:K}^t}\left[\frac{\partial}{\partial\theta}E(\mathbf{y},\mathbf{x}_{1:K}^t,\mathbf{h})\right]\right] \\
&= \mathrm{E}_{\mathbf{y}|\mathbf{x}_{1:K}^t}\left[\frac{\partial}{\partial\theta}E(\mathbf{y},\mathbf{x}_{1:K}^t,\mathbf{h}(\mathbf{y},\mathbf{x}_{1:K}^t))\right] \\
&= \sum_{\mathbf{y}\in\{\mathbf{e}_l|l\in\{1,\ldots,C\}\}} p(\mathbf{y}|\mathbf{x}_{1:K}^t)\frac{\partial}{\partial\theta}E(\mathbf{y},\mathbf{x}_{1:K}^t,\mathbf{h}(\mathbf{y},\mathbf{x}_{1:K}^t))
\end{aligned}$$

where C is the number of classes, and $p(\mathbf{y}|\mathbf{x}_{1:K}^t)$ can be computed tractably.

However, the expectation with respect to $\mathbf{y}$, $\mathbf{x}_{1:K}$ and $\mathbf{h}$ is intractable. Contrastive Divergence provides a good approximation however, by replacing the expectation over the input units $\mathbf{x}_{1:K}$ with a point estimate at a sample $\mathbf{x}_{1:K}^{\text{neg}}$. We also replace the expectation over $\mathbf{y}$ by a point estimate at a sample $\mathbf{y}^{\text{neg}}$ (while not necessary, it is more efficient to do so):

$$\begin{aligned}
\mathrm{E}_{\mathbf{y},\mathbf{x}_{1:K},\mathbf{h}}\left[\frac{\partial}{\partial\theta}E(\mathbf{y},\mathbf{x}_{1:K},\mathbf{h})\right] &= \mathrm{E}_{\mathbf{y},\mathbf{x}_{1:K}}\left[\mathrm{E}_{\mathbf{h}|\mathbf{y},\mathbf{x}_{1:K}}\left[\frac{\partial}{\partial\theta}E(\mathbf{h}|\mathbf{y},\mathbf{x}_{1:K})\right]\right] \\
&= \mathrm{E}_{\mathbf{h}|\mathbf{y}^{\text{neg}},\mathbf{x}_{1:K}^{\text{neg}}}\left[\frac{\partial}{\partial\theta}E(\mathbf{y}^{\text{neg}},\mathbf{x}_{1:K}^{\text{neg}},\mathbf{h})\right] \\
&= \frac{\partial}{\partial\theta}E(\mathbf{y}^{\text{neg}},\mathbf{x}_{1:K}^{\text{neg}},\mathbf{h}(\mathbf{y}^{\text{neg}},\mathbf{x}_{1:K}^{\text{neg}}))
\end{aligned}$$

In Contrastive Divergence, the samples $\mathbf{x}_{1:K}^{\text{neg}}$ and $\mathbf{y}^{\text{neg}}$ are obtained by running a brief MCMC chain, initialized at the training data observation $\mathbf{x}_{1:K}^t$ and $\mathbf{y}^t$. In particular, we use one step of Gibbs sampling, first sampling a value of $\mathbf{h}^{\text{neg}}$ for $\mathbf{h}$ given $\mathbf{x}_{1:K}^t$, and then sampling a new value for all glimpses $\mathbf{x}_{1:K}^{\text{neg}}$ and for the target $\mathbf{y}^{\text{neg}}$ given $\mathbf{h}^{\text{neg}}$. Algorithm 1 gives a pseudocode of this sampling procedure.

Algorithm 2 Gibbs sampling in Contrastive Divergence, to obtain samples $\mathbf{x}_{1:k}^{\mathrm{neg}}$ and $\mathbf{y}^{\mathrm{neg}}$ for the multi-fixation RBM, for the k^{th} term of the hybrid-sequential cost

Input: training pair $(\mathbf{y}^t, \mathbf{x}_{1:k}^t)$
% Notation: $a \sim p$ means a is sampled from p
$\mathbf{h}^{\mathrm{neg}} \sim p(\mathbf{h}|y^t, \mathbf{x}_{1:k}^t)$
$\mathbf{y}^{\mathrm{neg}} \sim p(\mathbf{y}|\mathbf{h}^{\mathrm{neg}})$
$\mathbf{x}_k^{\mathrm{neg}} \sim p(\mathbf{x}_k|\mathbf{h}^{\mathrm{neg}})$

All that is left to derive are the gradients of the energy function with respect to all parameters, which are simply:

$$\begin{aligned}
\frac{\partial}{\partial d_{l^*}} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h}) &= -y_{l^*} \\
\frac{\partial}{\partial c_j} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h}) &= -h_j \\
\frac{\partial}{\partial b_i} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h}) &= -\sum_{k=1}^{K} x_{ki} \\
\frac{\partial}{\partial U_{jl^*}} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h}) &= -h_j y_{l^*} \\
\frac{\partial}{\partial P_{ji}} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h}) &= -h_j \sum_{k=1}^{K} z(i_k, j_k)_i \, \mathbf{F}_{i\cdot} \, \mathbf{x}_k \\
\frac{\partial}{\partial F_{ji}} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h}) &= -\mathbf{h}^\top \sum_{k=1}^{K} \mathbf{P}_{\cdot j} z(i_k, j_k)_j x_{ki} \\
\frac{\partial}{\partial \bar{z}(i_k, j_k)_a} E(\mathbf{y}, \mathbf{x}_{1:K}, \mathbf{h}) &= -(\mathbf{h}^\top \mathbf{P}_{\cdot a}) \, z(i_k, j_k)_a \, (1 - z(i_k, j_k)_a) \, (\mathbf{F}_{a\cdot} \, \mathbf{x}_k)
\end{aligned}$$

3 Detailed description of the hybrid-sequential cost gradient

We now move to the hybrid-sequential cost:

$$\textbf{Hybrid-sequential cost:} \;\; \mathcal{C}_{\mathrm{hybrid-seq}} = \sum_{k=1}^{K} -\log p(\mathbf{y}^t|\mathbf{x}_{1:k}^t) - \alpha \log p(\mathbf{y}^t, \mathbf{x}_k^t|\mathbf{x}_{1:k-1}^t)$$

It has the following gradient:

$$\begin{aligned}
\frac{\partial \mathcal{C}_{\mathrm{hybrid-seq}}}{\partial \theta} &= \sum_{k=1}^{K} \Bigg\{ \mathrm{E}_{\mathbf{h}|\mathbf{y}^t, \mathbf{x}_{1:k}^t} \left[\frac{\partial}{\partial \theta} E(\mathbf{y}^t, \mathbf{x}_{1:k}^t, \mathbf{h}) \right] - \mathrm{E}_{\mathbf{y}, \mathbf{h}|\mathbf{x}_{1:k}^t} \left[\frac{\partial}{\partial \theta} E(\mathbf{y}, \mathbf{x}_{1:k}^t, \mathbf{h}) \right] \\
&\quad + \alpha \left(\mathrm{E}_{\mathbf{h}|\mathbf{y}^t, \mathbf{x}_{1:k}^t} \left[\frac{\partial}{\partial \theta} E(\mathbf{y}^t, \mathbf{x}_{1:k}^t, \mathbf{h}) \right] - \mathrm{E}_{\mathbf{y}, \mathbf{x}_k, \mathbf{h}|\mathbf{x}_{1:k-1}} \left[\frac{\partial}{\partial \theta} E(\mathbf{y}, \mathbf{x}_{1:k}, \mathbf{h}) \right] \right) \Bigg\}
\end{aligned} \tag{2}$$

The expectations with respect to $\mathbf{h}$ only and with respect to $\mathbf{h}$ and $\mathbf{y}$ are still tractable. The only difference is that we have K such expectations, one for every subsequence $\mathbf{x}_{1:k}$ where $k \in 1, \ldots, K$. Hence, the formulas of the previous section still apply, the only difference being that the number of glimpses k changes in the visible layer.

As for the expectations with respect to $\mathbf{y}$, $\mathbf{x}_k$ and $\mathbf{h}$, it is intractable but Contrastive Divergence can also be used, much like for the expectations with respect to $\mathbf{h}$, $\mathbf{y}$ and $\mathbf{x}_{1:K}$ in the previous section. The only difference is that a sample $\mathbf{x}_k^{\mathrm{neg}}$ for the k^{th} glimpse only is needed, instead of for the whole sequence of glimpses, since we are conditioning on the previous glimpses $\mathbf{x}_{1:k-1}$. Algorithm 2 gives a pseudo-code for sampling $\mathbf{x}_k^{\mathrm{neg}}$.

The training update for the hybrid-sequential cost can just proceed sequentially. For $k = 1$ to K, the k^{th} glimpse $\mathbf{x}_k$ is obtained and then gradients for the corresponding k^{th} group of terms in the summation of Equation 2 are estimated and accumulated. Once all gradients have been accumulated, the multi-fixation RBM is updated by a gradient step.