Supplemental Material

July 5, 2016

1 Multimodal Restricted Boltzmann Machine

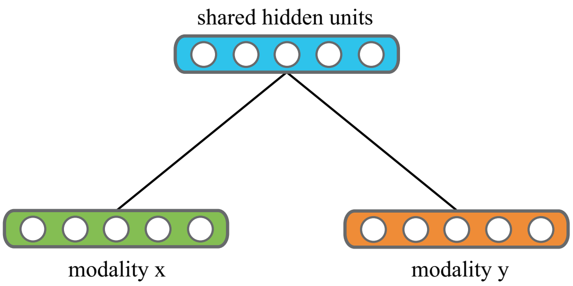

Figure 1: An illustration of the MRBM.

The RBM is an undirected graphical model that defines a probability distribution of visible units using hidden units. Under the case of multimodal input, MRBM (Figure 1) defines the joint distribution over modality $\mathbf{x}$, visual modality $\mathbf{y}$, and shared hidden units $\mathbf{h}$,

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

where Z is the partition function and E is an energy function given by

$$\begin{aligned} E\left(\mathbf{x}, \mathbf{y}, \mathbf{h}\right) = &-\mathbf{x}^T \mathbf{W}^x \mathbf{h} - \mathbf{y}^T \mathbf{W}^y \mathbf{h} \\ &- \mathbf{x}^T \mathbf{b}^x - \mathbf{y}^T \mathbf{b}^y - \mathbf{h}^T \mathbf{b}^h, \end{aligned} \tag{2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the binary visible units of audio and visual input, and $\mathbf{h}$ is the binary shared hidden units. $\mathbf{W}^x$ is a matrix of pairwise weights between elements of $\mathbf{x}$ and $\mathbf{h}$, and similar for $\mathbf{W}^y$. $\mathbf{b}^x$, $\mathbf{b}^y$, $\mathbf{b}^h$ are bias vectors for $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{h}$, respectively. To obtain the joint likelihood $p(\mathbf{x}, \mathbf{y})$, $\mathbf{h}$ is marginalized out from the distribution,

$$p(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{h}} \exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{h}))/Z. \tag{3}$$

For the MRBM model, similar to the standard RBM, *Contrastive Divergence* (CD) or *Persistent CD* (PCD) is used to approximate the gradient to maximize the joint likelihood, i.e., $p(\mathbf{x}, \mathbf{y})$. This is the typical maximum joint likelihood learning for MRBM. Finally, the learned shared hidden units $\mathbf{h}$ is treated as the joint representation across modalities.

2 Problem

The *Semantic Similarity Learning* (SSL) aims at training the MRBM model via reinforcing the similarity between the high-level features of $\mathbf{H}^x$, $\mathbf{H}^y$ for modality $\mathbf{x}$ and $\mathbf{y}$, which is combined with the joint probability distribution over the modalities as follows,

$$\text{minimize}_{\left\{\mathbf{W}^x, \mathbf{W}^y, \mathbf{b}^x, \mathbf{b}^y, \mathbf{b}^h\right\}} \quad - \log p\left(\mathbf{x}, \mathbf{y}\right) + \lambda \ \text{sim}(\mathbf{H}^x, \mathbf{H}^y), \tag{4}$$

where the parameters $\left\{\mathbf{W}^x, \mathbf{W}^y, \mathbf{b}^x, \mathbf{b}^y, \mathbf{b}^h\right\}$ are the same as the ones in the last section and λ in Eq. 4 is a regularization constant. And the sigmoid activation function is adopted in this paper. The similarity function $\text{sim}(\mathbf{H}^x, \mathbf{H}^y)$ is described in different methods, i.e., cross-entroy and *Canonical Correlation Analysis* (CCA). In the following section, we will introduce the specific optimization of Eq. 4 wrt each kind of the similarity function.

3 Optimization

3.1 Optimization wrt cross-entropy

In this section, we use the cross-entropy function to describe the similarity between modalities, which can be written as

$$\text{minimize}_{\left\{\mathbf{W}^x, \mathbf{W}^y, \mathbf{b}^x, \mathbf{b}^y, \mathbf{b}^h\right\}} \quad - \log p(\mathbf{x}, \mathbf{y}) + \lambda \ \mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y), \tag{5}$$

where

$$\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y) = - \sum_{j,k} \left[\mathbf{H}^x_{jk} \log \mathbf{H}^y_{jk} + (1 - \mathbf{H}^x_{jk}) \log(1 - \mathbf{H}^y_{jk})\right]. \tag{6}$$

For the optimization, CD is employed to approximated the gradient of the term $p\left(\mathbf{x}, \mathbf{y}\right)$, followed by the gradient of the regularization term. Specifically, the gradient of the second term in Eq. 5 wrt parameter $\{\mathbf{W}^x, \mathbf{W}^y\}$ is

$$\mathbf{W}^x := \mathbf{W}^x + \alpha \cdot \mathbf{x}^T(\log\mathbf{H}^y - \log(1 - \mathbf{H}^y)) \odot \mathbf{H}^x \odot (1 - \mathbf{H}^x), \tag{7}$$

$$\mathbf{W}^y := \mathbf{W}^y + \alpha \cdot \mathbf{y}^T(\mathbf{H}^x - \mathbf{H}^y), \tag{8}$$

where α is the step size and $\odot$ is the dot product. Besides, as $\mathbb{H}(\mathbf{H}^x \parallel \mathbf{H}^y) \neq \mathbb{H}(\mathbf{H}^y \parallel \mathbf{H}^x)$, similar optimization is also performed to the latter one but in symmetric fashion.

3.2 Optimization wrt CCA

For the similarity function in terms of CCA, the objective function can be written as follows,

$$\text{minimize}_{\{\mathbf{W}^x,\mathbf{W}^y,\mathbf{b}^x,\mathbf{b}^y,\mathbf{b}^h\}} \quad -\log p\left(\mathbf{x},\mathbf{y}\right) - \lambda\left\|\mathbf{T}\right\|_{tr}, \tag{9}$$

where the term $\|\mathbf{T}\|_{tr}$ is the matrix trace norm of $\sum_{xx}^{-1/2}\sum_{xy}\sum_{yy}^{-1/2}$. The three terms denote the covariance of $\mathbf{H}^x$, $\mathbf{H}^y$, and both of them,

$$\begin{aligned}
\sum\nolimits_{xx} &= \frac{1}{n-1}\mathbf{H}^x(\mathbf{H}^x)^T, \\
\sum\nolimits_{yy} &= \frac{1}{n-1}\mathbf{H}^y(\mathbf{H}^y)^T, \text{ and} \\
\sum\nolimits_{xy} &= \frac{1}{n-1}\mathbf{H}^x(\mathbf{H}^y)^T,
\end{aligned} \tag{10}$$

where n is the size of the training samples. The above added term of $\|\mathbf{T}\|_{tr}$ is one of several methods to explain the solutions of CCA, where the correlation of the components of $\mathbf{H}^x$ and $\mathbf{H}^y$ is the singular values of $\mathbf{T}$. And using the chain rule, we can derive the gradient of the regularization term,

$$\begin{aligned}
\frac{\partial\|\mathbf{T}\|_{tr}}{\partial\mathbf{W}^x} &= \frac{\partial\|\mathbf{T}\|_{tr}}{\partial\mathbf{H}^x}\cdot\frac{\partial\mathbf{H}^x}{\partial\mathbf{W}^x} \\
&= \left(\frac{\partial\|\mathbf{T}\|_{tr}}{\partial\sum_{xx}}\cdot\frac{\partial\sum_{xx}}{\partial\mathbf{H}^x} + \frac{\partial\|\mathbf{T}\|_{tr}}{\partial\sum_{xy}}\cdot\frac{\partial\sum_{xy}}{\partial\mathbf{H}^x}\right)\cdot\frac{\partial\mathbf{H}^x}{\partial\mathbf{W}^x} \\
&= \left(\nabla_{xx}\|\mathbf{T}\|_{tr}\cdot\frac{\partial\sum_{xx}}{\partial\mathbf{H}^x} + \nabla_{xy}\|\mathbf{T}\|_{tr}\cdot\frac{\partial\sum_{xy}}{\partial\mathbf{H}^x}\right)\cdot\frac{\partial\mathbf{H}^x}{\partial\mathbf{W}^x}.
\end{aligned} \tag{11}$$

In Eq. 11, the term $\nabla_{xx}\|\mathbf{T}\|_{tr}$ and $\nabla_{xy}\|\mathbf{T}\|_{tr}$ have been derived in [1] (in the Reference of the paper) as follows,

$$\nabla_{xx}\|\mathbf{T}\|_{tr} = -\frac{1}{2}\sum\nolimits_{xx}^{-1/2}\mathbf{U}\mathbf{D}\mathbf{U}^T\sum\nolimits_{xx}^{-1/2}, \tag{12}$$

$$\nabla_{xy}\|\mathbf{T}\|_{tr} = \sum\nolimits_{xx}^{-1/2}\mathbf{U}\mathbf{V}^T\sum\nolimits_{yy}^{-1/2}, \tag{13}$$

where the matrix $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{D}$ are obtained via the singular value decomposition of $\mathbf{T}$,

$$\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{14}$$

Thus, Eq. 11 becomes,

$$\frac{\partial\|\mathbf{T}\|_{tr}}{\partial\mathbf{W}^x} = \frac{1}{n-1}(\mathbf{x}((2\nabla_{xx}\|\mathbf{T}\|_{tr}\mathbf{H}^x + \nabla_{xy}\|\mathbf{T}\|_{tr}\mathbf{H}^y)\odot\mathbf{H}^x\odot(1-\mathbf{H}^x))^T). \tag{15}$$

The gradient wrt $\mathbf{W}^y$ takes a symmetric form as Eq. 15. Thus, we can employ the gradient of the regularization term for optimizing the parameters in Eq. 9.