Learning Representations for Multimodal Data with Deep Belief Nets

Nitish Srivastava,Ruslan Salakhutdinov-ICML2012

Yunfei Wang

Department of Computer Science & Technology
Huazhong University of Science & Technology

April 9, 2013

Table of contents

Multimodal data

- Information in the world comes from multiple input channels.

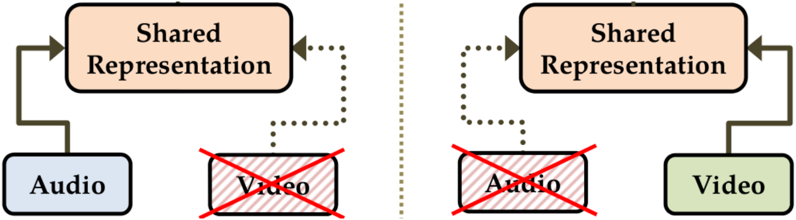
- Information content of any modality is unlikely to be independent of the others.
- How to dig out the joint representation of multimodal data?
- How to handle missing data modalities?

Challenges I

- Different modalities have different representation.

Challenges II

- Observations are noisy and may have missing modalities.



Restricted Boltzman Machines

Visible units $v \in \{0,1\}^D$,hidden units $h \in \{0,1\}^F$
The energy of joint distribution:

$$E(v,h;\theta) = -v^T W h - a^T v - b^T h \tag{1}$$

where $\theta = W, a, b$ are the model parameters.
The joint probability over visible and a hidden units:

$$P(v,h;\theta) = \frac{1}{Z(\theta)} \exp(-E(v,h;\theta)) \tag{2}$$

where $Z(\theta)$ is the normalizing constant:

$$Z(\theta) = \sum_{v,h} \exp(-E(v,h;\theta)) \tag{3}$$

Multinomial RBM

Visible units $v \in \mathbb{N}^K$,hidden units $h \in \{0,1\}^F$.
The energy function is defined as follows:

$$E(v,h;\theta) = -\sum_{k=1}^{K}\sum_{j=1}^{F} v_i W_{kj} h_j - \sum_{k=1}^{K} a_k v_k - M\sum_{j=1}^{F} b_j h_j \tag{4}$$

where v_k is frequency of word k in a document,K is vocabulary size,$M = \sum_{k=1}^{K} v_k$ is total number of words in the document.
This leads to the following conditional distribution:

$$P(v_k = 1|h;\theta) = \frac{\exp(-a_k + \sum_{j=1}^{F} W_{kj} h_j)}{\sum_{k=1}^{K} \exp(-a_k + \sum_{j=1}^{F} W_{kj} h_j)} \tag{5}$$

Modelling sparse count data,such as word count vectors in a document.

Gaussian RBM

Visible units $v \in \mathbb{R}^D$,hidden units $h \in \{0,1\}^F$.
The energy function is defined as follows:

$$E(v,h;\theta) = \sum_{i=1}^{D} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{F} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j W_{ij} \tag{6}$$

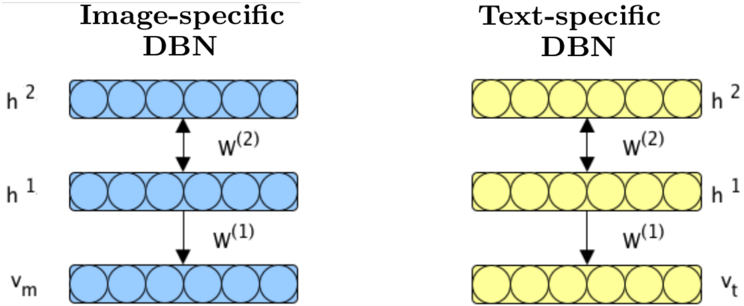
where $\theta = \{a, b, W, \sigma\}$ are the model parameters. This leads to the following conditional distribution:

$$P(v_i, h;\theta) = \mathcal{N}(a_i + \sigma_i \sum_{j=1}^{F} W_{ij} h_j, \sigma_i^2) \tag{7}$$

Modelling real-valued data,such as density value in a image.

Modality-free Features I

Model each data modality using a separate two-layer DBN.



Modality-free Features II

Image visible units $v_m \in \mathbb{R}^D$,test visible units $v_t \in \mathbb{N}^K$.
Image-specific DBN uses Gaussian RBM to model the distribution over real-valued image features:

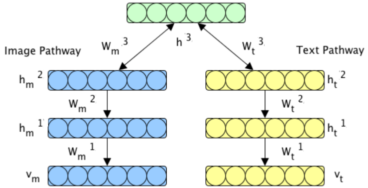
$$P(v_m) = \sum_{h^1,h^2} P(h^1, h^2)P(v_m|h^1) \tag{8}$$

Text-specific DBN uses multinomial RBM to model the distribution over word count vectors:

$$P(v_t) = \sum_{h^1,h^2} P(h^1, h^2)P(v_t|h^1) \tag{9}$$

Multimodal DBN

Multimodal DBN:learning a joint RBM on top of two models.



The joint distribution can be written as:

$$P(v_m, v_t) = \sum_{h_m^2, h_t^2, h^3} P(h_m^2, h_t^2, h^3) \times \sum_{h_m^1} P(v_m|h_m^1) P(h_m^1|h_m^2) \times \sum_{h_t^1} P(v_t|h_t^1) P(h_t^1|h_t^2) \quad (10)$$

Handle missing modalities I

Infer missing values by drawing samples from conditional model.

Generate text conditioned on a given image v_m

1. Infer the values of hidden variables h_m^2.
2. Perform Gibbs sampling using following conditional distributions:

$$P(h^3|h_m^2, h_t^2) = \sigma(W_m^3 h_m^2 + W_t^3 h_t^2 + b) \tag{11}$$

$$P(h_t^2|h^3) = \sigma((W_t^3)^T h^3 + a) \tag{12}$$

where $\sigma(x) = 1/(1 + \exp(-x))$.

3. h_t^2 can be propagated back to generate text.

Handle missing modalities II

Figure: Procedure of infering missing values