Multimodal Learning with Deep Boltzmann Machines

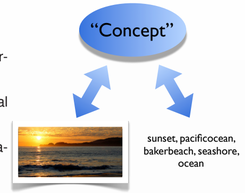
Nitish Srivastava Ruslan Salakhutdinov

UNIVERSITY OF TORONTO

Introduction

Data is often a collection of modalities.

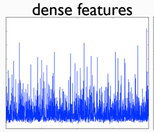
- Images and captions, audio and video, sensory perception.
- Each input channel often has very different statistical properties.
- Need a way to fuse modalities into a joint representation that captures the "concept".



Challenges

Very different input representations make it hard to learn cross-modal features.

sunset, pacificocean, bakerbeach, seashore, ocean



Real-valued, dense features

Sparse, discrete word counts

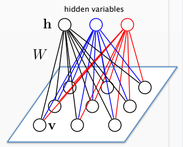
Noisy and missing data.

Image	User assigned Tags
	⟨ no text ⟩
	pentax, k10d, pantaxda50200, kangarooisland, sa, australiansealion
	unseulpixel, naturey, crap
	mickikrimmel, mickipedia, headshot

Restricted Boltzmann Machines

A Restricted Boltzmann Machine is a Markov Random Field with

- Stochastic visible units $\mathbf{v} \in \{0,1\}^D$.
- Stochastic hidden units $\mathbf{h} \in \{0,1\}^F$.
- Bipartite connections.



Binary RBMs $\mathbf{v} \in \{0,1\}^D$

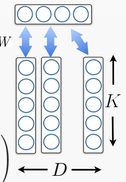
$$P_\theta(\mathbf{v},\mathbf{h};\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F} v_i W_{ij} h_j + \sum_{i=1}^{D} b_i v_i + \sum_{j=1}^{F} a_j h_j\right)$$

Gaussian RBMs $\mathbf{v} \in \mathbb{R}^D$

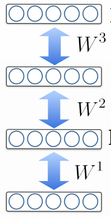
$$P_\theta(\mathbf{v},\mathbf{h};\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{i=1}^{D}\sum_{j=1}^{F} \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma_i^2} + \sum_{j=1}^{F} a_j h_j\right)$$

Replicated Softmax Model $\mathbf{v} \in \{1,\ldots,K\}^D$ (1-of-K representation).

$$P_\theta(\mathbf{v},\mathbf{h};\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{i=1}^{D}\sum_{k=1}^{K}\sum_{j=1}^{F} v_i^k W_{ij}^k h_j + \sum_{i=1}^{D}\sum_{k=1}^{K} b_i^k v_i^k + \sum_{j=1}^{F} a_j h_j\right)$$



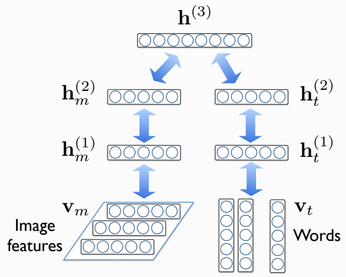
Deep Boltzmann Machines



- Markov Random Field with layers of binary hidden variables.
- Pair-wise potentials between variables across adjacent layers.
- Introduces dependencies between hidden variables.
- Learns high-order features.

$$P_\theta(\mathbf{v},\mathbf{h}^1,\mathbf{h}^2,\mathbf{h}^3;\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\mathbf{v}^\top W^1 \mathbf{h}^1 + {\mathbf{h}^1}^\top W^2 \mathbf{h}^2 + {\mathbf{h}^2}^\top W^3 \mathbf{h}^3\right)$$

Multimodal Deep Boltzmann Machines



- Joint density model of images and text.
- At the first layer use modality-specific models - Gaussian RBMs for image features, Replicated Softmax for word counts.
- Higher levels combine information from both modalities, create fused representations.
- Bottom-up and Top-down influence.

Learning DBMs

(Approximate) Maximum likelihood learning

$$\frac{\partial \log P_\theta(\mathbf{v})}{\partial W^1} = \mathbb{E}_{P_{data}}[\mathbf{v}{\mathbf{h}^1}^\top] - \mathbb{E}_{P_\theta}[\mathbf{v}{\mathbf{h}^1}^\top]$$

- Both expectations intractable.
- Mean-field inference for estimating $\mathbb{E}_{P_{data}}[\,]$.
- MCMC-based stochastic approximation for estimating $\mathbb{E}_{P_\theta}[\,]$.

Approximate the true posterior $P(\mathbf{h}|\mathbf{v};\theta)$, where $\mathbf{v} = \{\mathbf{v}_m, \mathbf{v}_t\}$, with a fully factorized approximating distribution over the five sets of hidden units $\{\mathbf{h}_m^{(1)}, \mathbf{h}_m^{(2)}, \mathbf{h}_t^{(1)}, \mathbf{h}_t^{(2)}, \mathbf{h}^{(3)}\}$:

$$Q(\mathbf{h}|\mathbf{v};\boldsymbol{\mu}) = \left(\prod_{j=1}^{F_1} q(h_{mj}^{(1)}|\mathbf{v}) \prod_{l=1}^{F_2} q(h_{ml}^{(2)}|\mathbf{v})\right)\left(\prod_{j=1}^{F_1} q(h_{tj}^{(1)}|\mathbf{v}) \prod_{l=1}^{F_2} q(h_{tl}^{(2)}|\mathbf{v})\right) \prod_{k=1}^{F_3} q(h_k^{(3)}|\mathbf{v}),$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_m^{(1)}, \boldsymbol{\mu}_m^{(2)}, \boldsymbol{\mu}_t^{(1)}, \boldsymbol{\mu}_t^{(2)}, \boldsymbol{\mu}^{(3)}\}$ are the mean-field parameters with $q(h_i^{(l)} = 1) = \mu_i^{(l)}$ for $l = 1, 2, 3$.

- Find the value of $\boldsymbol{\mu}$ that maximizes the variational lower bound for the current value of model parameters θ by iterating a set of the mean-field fixed-point equations.
- Given the variational parameters $\boldsymbol{\mu}$, update the model parameters θ to maximize the variational bound using an MCMC-based stochastic approximation.

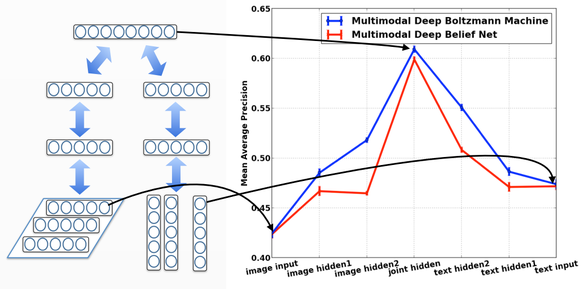
The models are initialized with stacks of RBMs trained with Persistent Contrastive Divergence.

Classification Results

The MIR-Flickr dataset consists of 1 million images and user-assigned tags. 25K images have been annotated for 38 topics.

Multimodal Inputs

Method	MAP	Precision@50	
Random	0.124	0.124	
LDA [Huiskes et. al.]	0.492	0.754	Similar features, 25K + 1 Million unlabelled
SVM [Huiskes et. al.]	0.475	0.758	
DBM-Labelled	0.526	0.791	
DBM-Unlablled	0.585	0.836	
Deep Belief Net	0.599	0.867	+ SIFT features
Autoencoder	0.600	0.875	
DBM	0.609	0.873	

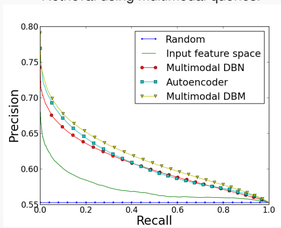


Unimodal Inputs

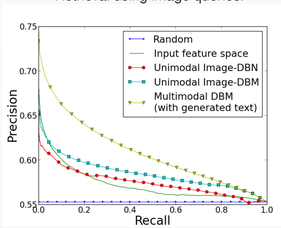
Method	MAP	Precision@50
Image-LDA [Huiskes et. al.]	0.315	-
Image-SVM [Huiskes et. al.]	0.375	-
Image-DBM	0.469	0.803
Multimodal-DBM (missing text)	0.531	0.832

Image Retrieval Results

Retrieval using multimodal queries.



Retrieval using image queries.



Generating Text Conditioned on Images.

Image	Generated Text	Image	Generated Text
	sea, france, boat, mer, beach, river, bretagne, plage, brittany		food, art, dessert, cooking, delicious, cake, lunch, sugar
	insect, butterfly, insects, bug, butterflies, lepidoptera		architecture, reflection, window, building, facade, architektur
	graffiti, streetart, stencil, sticker, urbanart, street, mural, nyc, graff, sanfrancisco		portrait, women, army, soldier, mother, postcard, soldiers
	portrait, child, kid, ritratto, kids, children, boy, cute, boys, italy		obama, barackobama, election, politics, president, hope, change, convention, rally

Image Retrieval from Text Queries

Query	Retrieved Images		
water, red, sunset			
nature, flower, red, green			
car, auto, automobile			
chocolate, cake			

Conclusions

Deep Boltzmann Machines are an effective way of fusing modalities.
Samples from conditional distributions can be used for annotation and retrieval.
Learning multimodal models helps even when only unimodal data is present at test time.

References

Ngiam, Khosla, Kim, Nam, Lee, and Ng, Multimodal deep learning. ICML 2011.
Xing, Yan, and Hauptmann, Mining associated text and images with dual-wing harmoniums. UAI 2005.
Huiskes, Thomee, and Lew, New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative. In Multimedia Information Retrieval, 2010.
Guillaumin, Verbeek, and Schmid, Multimodal semi-supervised learning for image classification. CVPR 2010.