Learning Representations for Multimodal Data with Deep Belief Nets

Nitish Srivastava, Ruslan Salakhutdinov
Department of Computer Science, University of Toronto

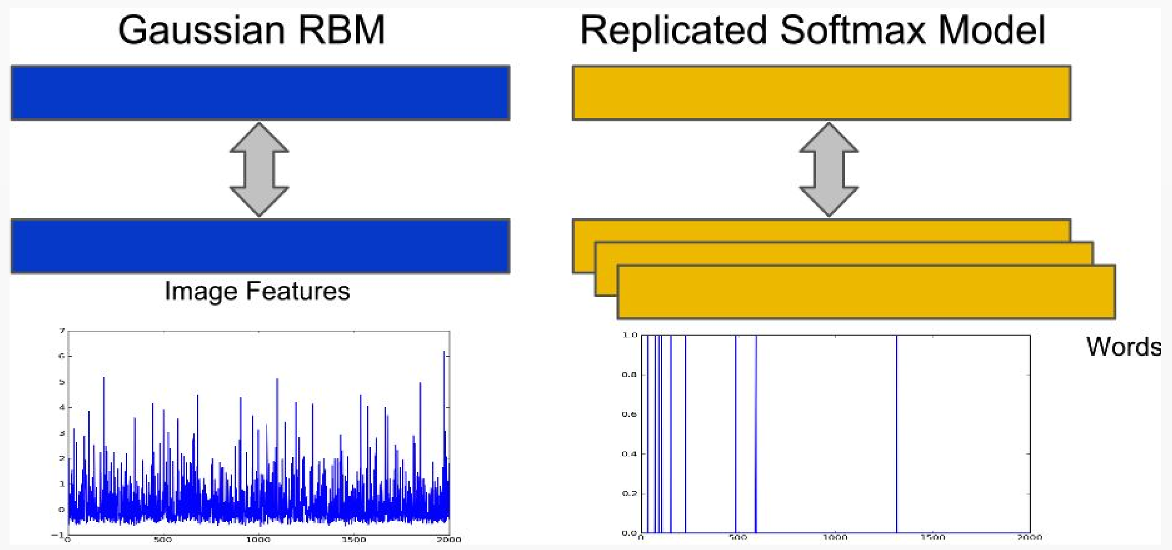
Introduction

- Real world data is often multimodal - Captioned images, video, sensory perception
- Strong associations exist across modalities but hard to discover in terms of low-level features
- Goal : Use unlabeled multimodal data to
  - Learn joint "modality-free" representation
  - Infer missing modalities given some observed ones
- Method : Build a joint density model using a DBN $P(\mathbf{v}_m, \mathbf{v}_t; \theta)$
  - Use states of top level hidden units as joint representation
  - Sample from conditional density model to fill in missing data
- Data : Multimedia Information Retrieval Flickr dataset
  - 1M images with noisy (sometimes missing) user-assigned tags
  - 25K annotated with 38 topics e.g. sky, tree, animals, baby, water (Used only for classification experiments).

Restricted Boltzmann Machines and their extensions

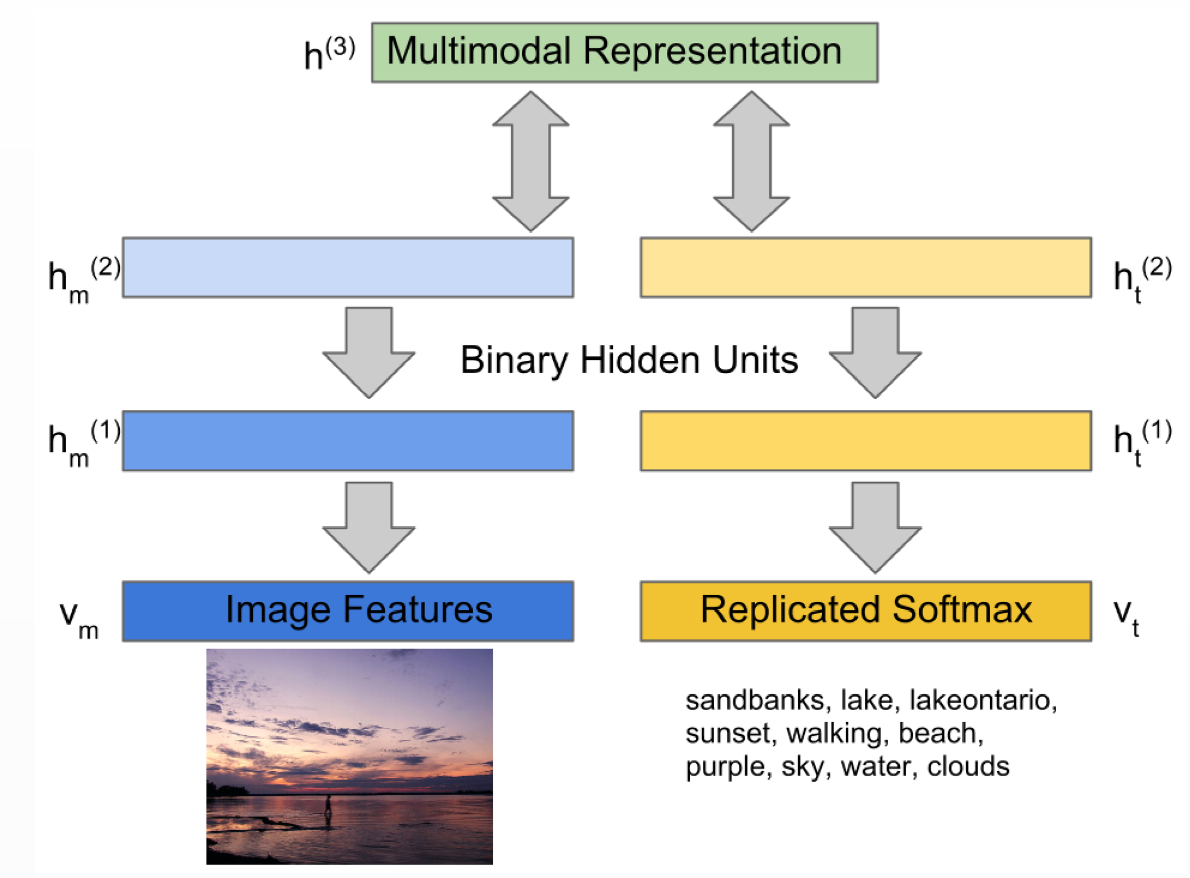
$P(v,h) = \exp(-E(v,h))/Z$

- Binary RBM
  $E(v,h) = -v^\top W h - b^\top v - a^\top h$
- Gaussian RBM
  $E(v,h) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} h_j - a^\top h$
- Replicated Softmax Model
  $E(v,h) = -\sum_{k,j} v_k W_{kj} h_j - \sum_k v_k b_k - M \sum_j h_j a_j$



Dense and real valued vs. sparse and discrete

Multimodal Deep Belief Net



Model Description

Multimodal DBN $\equiv$ Unimodal "pathways" combined with a top-level RBM

- First layer RBMs are modality-specific - Gaussian for images, Replicated Softmax for text
- Each successive layer learns higher-level features, abstracts away modality-specific correlations
- Top-level RBM jointly models high-level image and text features
- Easier to discover cross-modal relations since both sets of features are now binary and sparse.
- In contrast, input representations were widely different, which makes it difficult for shallow models to find cross-modal relations

Learning Greedy layer-wise training with Persistent Contrastive Divergence

Generative Tasks Sample conditional models using MCMC methods

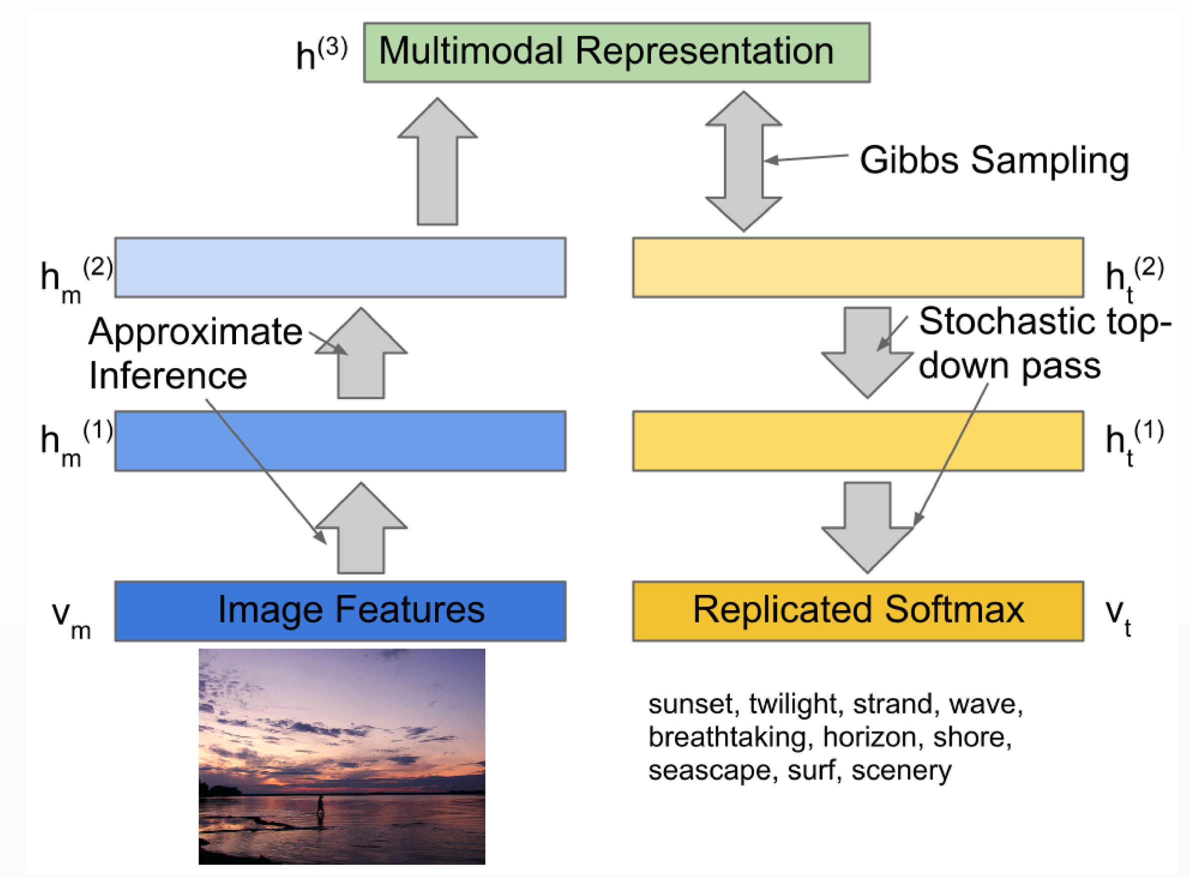
- Retrieve images using $P(\mathbf{v}_m|\mathbf{v}_t)$,
- Annotate images using $P(\mathbf{v}_t|\mathbf{v}_m)$

Discriminative Tasks Use DBN to initialize feed-forward network

- Multimodal Inputs - use both pathways
- Unimodal Inputs - infer unknown pathway with Gibbs Sampling (DBN-GenText)

Data pre-processing Images - extract SIFT, Gist, MPEG-7 descriptors, 2000 most frequent tags

Sampling from conditional model $P(\mathbf{v}_t|\mathbf{v}_m)$



Joint Distribution

The Multimodal DBN implies the following joint distribution

$$P(\mathbf{v}_m, \mathbf{v}_t) = \sum_{\mathbf{h}_m^{(2)}, \mathbf{h}_t^{(2)}, \mathbf{h}^{(3)}} P(\mathbf{h}_m^{(2)}, \mathbf{h}_t^{(2)}, \mathbf{h}^{(3)}) \times \sum_{\mathbf{h}_m^{(1)}} P(\mathbf{v}_m|\mathbf{h}_m^{(1)}) P(\mathbf{h}_m^{(1)}|\mathbf{h}_m^{(2)}) \times \sum_{\mathbf{h}_t^{(1)}} P(\mathbf{v}_t|\mathbf{h}_t^{(1)}) P(\mathbf{h}_t^{(1)}|\mathbf{h}_t^{(2)}).$$

Classification Results on MIR-Flickr dataset

Task : Predict whether input belongs to a user-annotated topic. Results are averaged over all topics.

Multimodal Inputs

Model	MAP	Prec@50
Random	0.124	0.124
Linear Discriminant Analysis	0.492	0.754
Support Vector Machines	0.475	0.758
DBN-Labeled-Data	0.503	0.741
Deep Autoencoder	0.547	**0.794**
DBN	**0.563**	0.785

Unimodal Inputs

Model	MAP	Prec@50
Image-SVM	0.375	-
Image-DBN	0.413	0.718
Text-DBN	0.471	0.723
DBN-ZeroText	0.484	0.730
DBN-GenText	**0.492**	**0.762**

Text Generation from Image Features

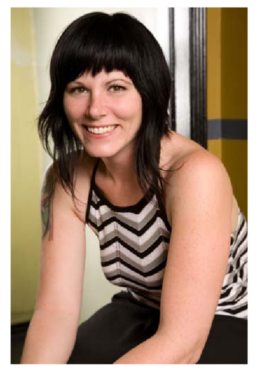
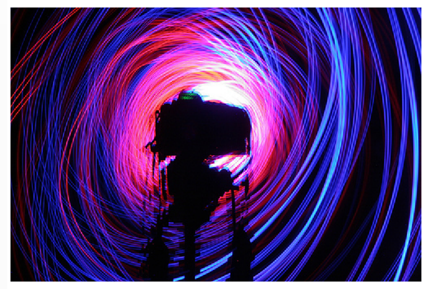
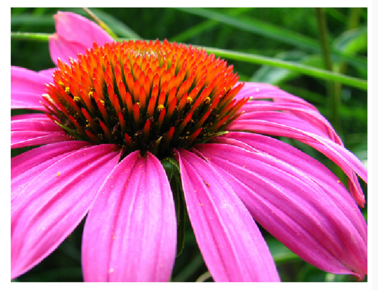
Image	Given Tags	Generated Tags
	pentax, k10d, kangarooisland, southaustralia, sa, australia, australiansealion, 300mm	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves
	<no text>	night, notte, traffic, light, lights, parking, darkness, lowlight, nacht, glow
	mickikrimmel, mickipedia, headshot	portrait, girl, woman, lady, blonde, pretty, gorgeous, expression, model
	camera, jahdakine, lightpainting, relection, doublepaneglass, wowiekazowie	blue, art, artwork, artistic, surreal, expression, original, artist, gallery, patterns

Image Retrieval from Text Queries

Input Text	2 nearest neighbours to generated image features	
nature, hill scenery, green clouds		
flower, nature, green, flowers, petal, petals, bud		
blue, red, art, artwork, painted, paint, artistic surreal, gallery bleu		
bw, blackandwhite, noiretblanc, biancoenero blancoynegro