

Multimodal Learning with Deep Boltzmann Machines

Article authors: N. Srivastava & R. Salakhutdinov

Presenter: Thomas Flynn

Overview

- Problem: Multimodal learning
- Model: Boltzmann machines
- Extensions for multiple modalities
- Experiments

Multimodal Learning

- Stuff can be recorded/expressed in different modalities:
 - Text
 - Digital Images
 - Speech
- Each signal has very unique characteristics, impossible to directly compare sensor outputs (e.g. comparing pixel intensities to word-count vectors)
- Idea: Use multi-layer network to map input to a “modality independent” representation

Multimodal Learning

- Multimodal learning problems:
 - How to take advantage of multiple modalities to build better classifiers?
 - How to translate between modalities? i.e. given an image, create a text description.

Multimodal Learning

- The idea is to “learn” the joint distribution $P(v,t)$ of visual and textual representations of stuff

– Find a P such that $P(v =$



$| t = \text{Cat})$ is high

- We can use P to perform some of previous tasks:
 - Getting a description of an image corresponds to sampling from the conditional distribution $P(t | v)$

Multimodal learning













Image	Given Tags	Generated Tags	Input Tags	Nearest neighbors to generated image features
	pentax, k10d, kangarooisland, southaustralia, sa, 300mm, australia, australiansealion	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves	nature, hill, scenery, green, clouds	 
	< no text >	night, lights, christmas, nightshot, nacht, nuit, notte, longexposure, noche, nocturna	flower, nature, green, flowers, petal, petals, bud	 
	aheram, 0505, sarahc, moo	portrait, bw, balckandwhite, people, faces, girl, blackwhite, person, man	blue, red, art, artwork, painted, paint, artistic, surreal, gallery, bleu	 
	unseulpixel, naturey crap	fall, autumn, trees, leaves, foliage, forest, woods, branches, path	bw, blackandwhite, noiretblanc, bianconero, blancoynegro	 

Figure 1: **Left:** Examples of text generated from a Deep Boltzmann Machine by sampling from $P(\mathbf{v}_{txt}|\mathbf{v}_{img};\theta)$. **Right:** Examples of images retrieved using features generated from a Deep Boltzmann Machine by sampling from $P(\mathbf{v}_{img}|\mathbf{v}_{txt};\theta)$.

Boltzmann Machines

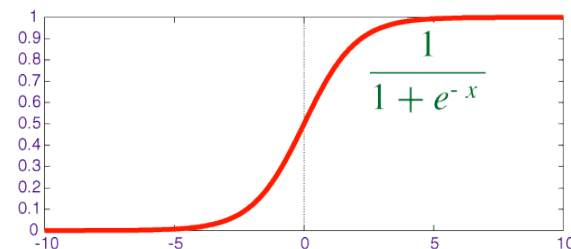
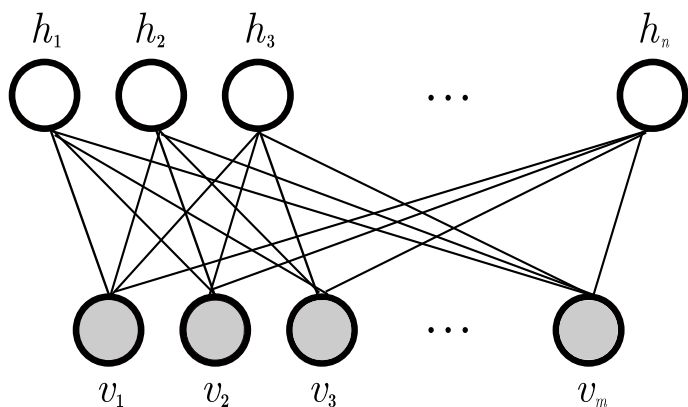
- The joint distribution $P(v,t)$ will be represented by a “Deep Multimodal Boltzmann Machine”, a type of stochastic neural network.
- Given many examples $\{v_i, t_i\}$, the training procedure (approximately) optimizes average likelihood:
$$\sum_i \log P(v_i, t_i; \theta)$$
- At runtime, use efficient approximations to perform tasks of interest: sampling $P(v | t)$ or $P(t | v)$

Boltzmann Machines

- Restricted Boltzmann Machines
- Gaussian-Bernoulli RBM
 - Modification for real valued inputs (e.g. images)
- Replicated Softmax RBM
 - Modification for categorical inputs (e.g. word count vectors)
- Deep Boltzmann Machine
 - A boltzmann machine with multiple hidden layers
- Gaussian + Replicated Softmax + Deep = Multimodal Deep Boltzmann Machine

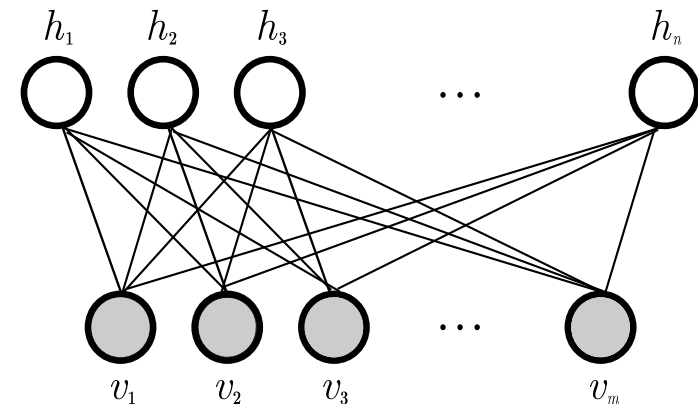
Restricted Boltzmann Machine

- A binary-valued stochastic neural network
- Units partitioned into hidden & visible groups
- “Computation” proceeds by selecting a unit at random and updating its value stochastically:
 - $P(h_i(t+1)=1 \mid h(t),v(t)) = \sigma(\sum_j w_{i,j}v_j(t)+ a_i)$
 - $P(v_i(t+1)=0 \mid h(t),v(t)) = \sigma(\sum_j w_{j,i}h_j(t)+ b_i)$



Restricted Boltzmann Machine

- RBM can also be thought of as an undirected graphical model.
- The stationary distribution of the update rule is $P(v,h) = \exp(\sum_{i,j} w_{ij} v_j h_i + \sum_i h_i a_i + \sum_i v_i b_i) / Z$
- Larger value of $P(v)$ \leftrightarrow RBM visits v frequently
- Sampling from $P(v | U = u)$ is done by clamping the group of units to the state u .



Replicated Softmax RBM

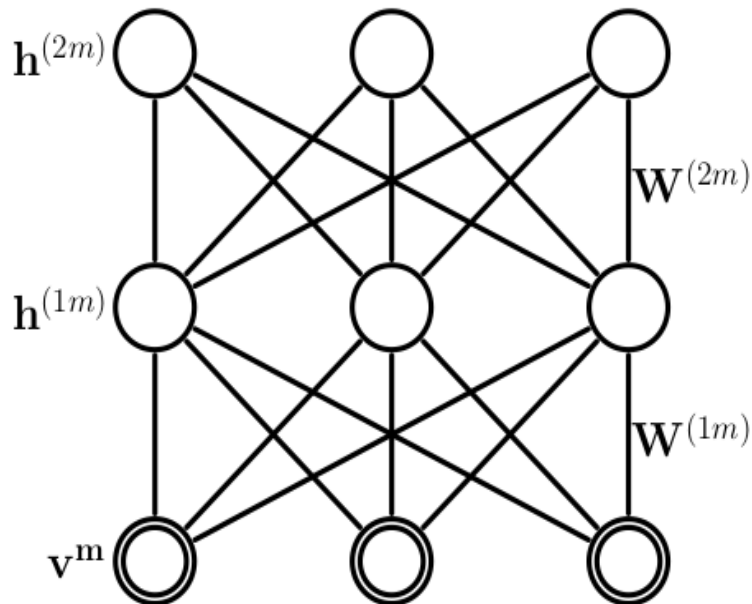
- A modification that is useful for modeling count data.
- For example, one can represent a document with words from a dictionary of size K using a K dimensional vector v , where $v_i = \#$ of times dictionary word i occurs.

Input	Reconstruction
chocolate, cake	cake, chocolate, sweets, dessert, cupcake, food, sugar, cream, birthday
nyc	nyc, newyork, brooklyn, queens, gothamist, manhattan, subway, streetart
dog	dog, puppy, perro, dogs, pet, filmshots, tongue, pets, nose, animal
flower, high, 花	flower, 花, high, japan, sakura, 日本, blossom, tokyo, lily, cherry
girl, rain, station, norway	norway, station, rain, girl, oslo, train, umbrella, wet, railway, weather
fun, life, children	children, fun, life, kids, child, playing, boys, kid, play, love
forest, blur	forest, blur, woods, motion, trees, movement, path, trail, green, focus
españa, agua, granada	españa, agua, spain, granada, water, andalucía, naturaleza, galicia, nieve

Table 1: Some examples of one-step reconstruction from the Replicated Softmax Model.

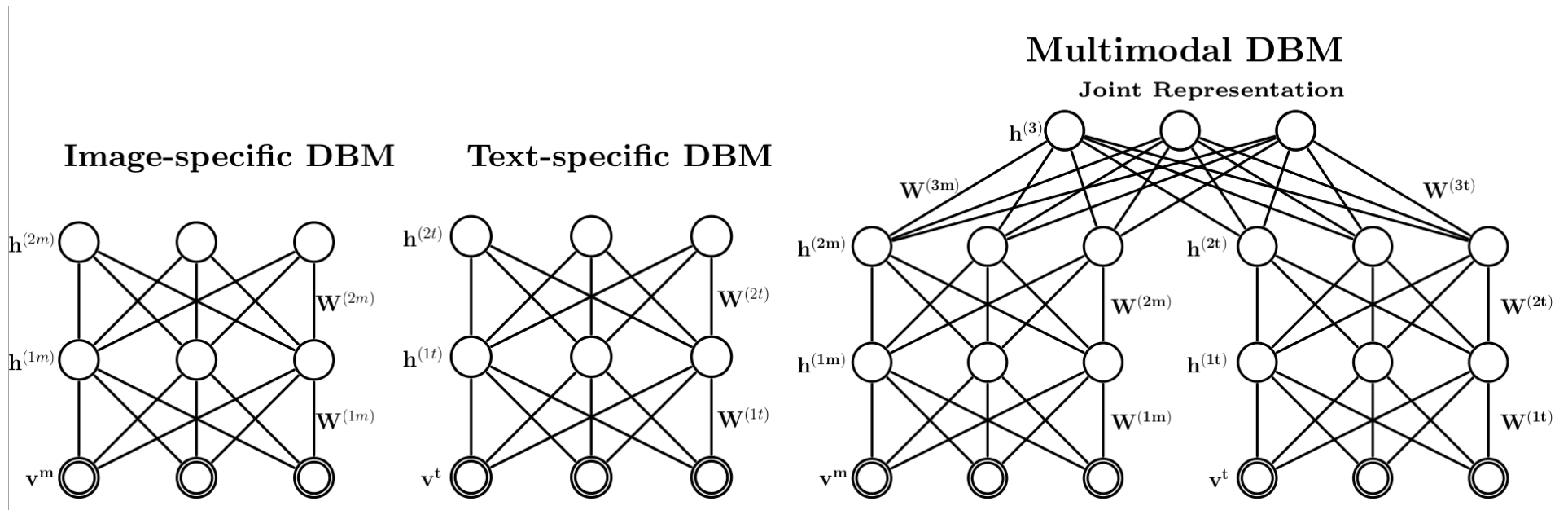
Deep Boltzmann Machine

- Multiple layers of hidden units
- $P(v, h^1, h^2)$ is similar to RBM:
- $P(v, h^1, h^2) = \exp(-E(v, h^1, h^2))/Z$
- $E(v, h^1, h^2) = \sum_{i,j} w^1_{ij} v_j h^1_i + \sum_{i,j} w^2_{ij} h^2_i h^1_j + \text{bias terms}$

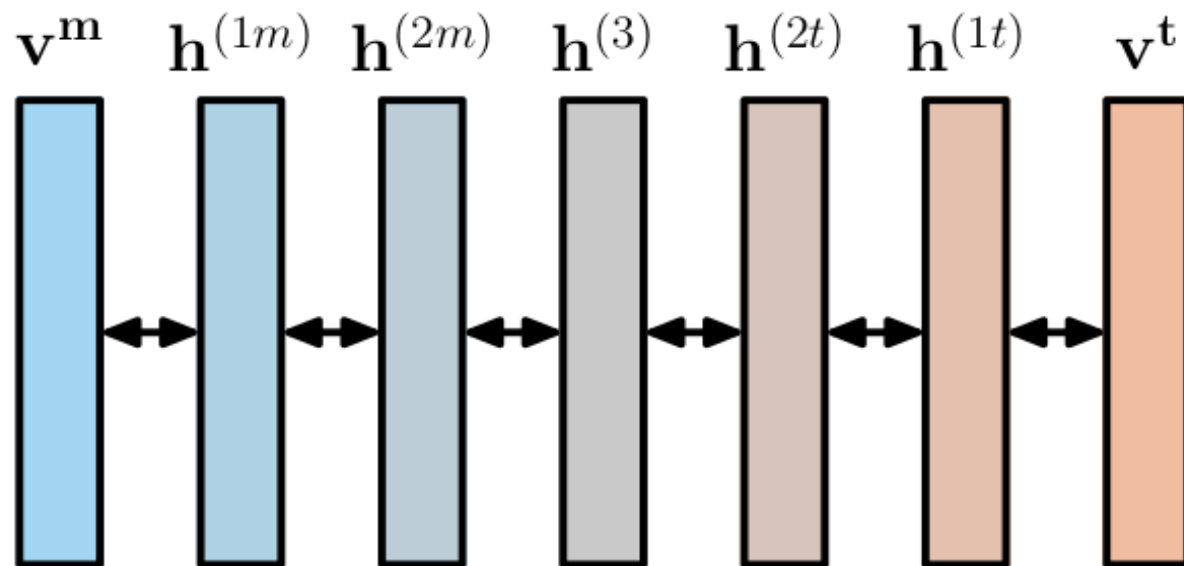


Multimodal DBM

- Image layer v^m units are Gaussian units.
- Text units v^t are Replicated-Softmax units
- All others are binary



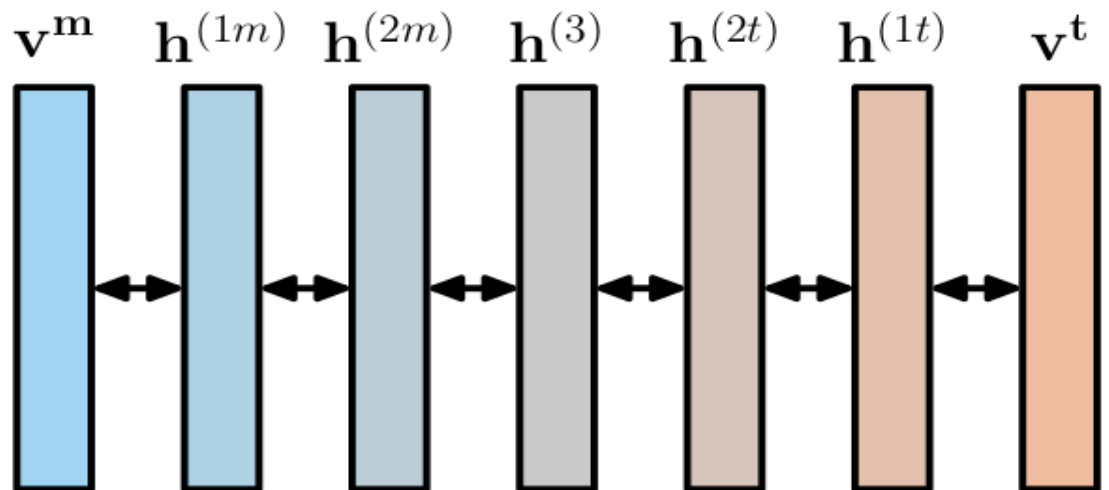
Multimodal DBM



Multimodal DBM

- An energy function $E(v^t, v^m, h^{1m}, h^{2m}, h^{1t}, h^{2t})$ describes the stationary distribution:
$$p(v^t, v^m, h^{1m}, \dots) = \exp(-E(v^t, v^m, h^{1m}, \dots))/Z$$
- Optimization problem: $\text{Max } p(v^t, v^m; \theta)$
- Requires MCMC techniques to approximate

$(d/d\theta) \log p(v^t, v^m; \theta)$



Experiments

- Task: Generate tags for an image, find image by tags
- Dataset: Flickr. Pairs of (image,tags). Each tag is an (unordered) collection of words describing the image.
- DBM trained to represent $P(\text{features}, \text{tags})$
- Features are standard computer vision features, represented by ~ 4000 dimensions
- Tags are represented by word counts of the 2000 most common words in the Flickr Data

Experiments

- Filling in missing modalities:













Image	Given Tags	Generated Tags	Input Tags	Nearest neighbors to generated image features	
	pentax, k10d, kangarooisland, southaustralia, sa, 300mm, australia, australiansealion	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves	nature, hill, scenery, green, clouds		
	< no text >	night, lights, christmas, nightshot, nacht, nuit, notte, longexposure, noche, nocturna	flower, nature, green, flowers, petal, petals, bud		
	aheram, 0505, sarahc, moo	portrait, bw, balckandwhite, people, faces, girl, blackwhite, person, man	blue, red, art, artwork, painted, paint, artistic, surreal, gallery, bleu		
	unseulpixel, naturey crap	fall, autumn, trees, leaves, foliage, forest, woods, branches, path	bw, blackandwhite, noiretblanc, bianconero, blancoynegro		

Figure 1: **Left:** Examples of text generated from a Deep Boltzmann Machine by sampling from $P(\mathbf{v}_{txt}|\mathbf{v}_{img};\theta)$. **Right:** Examples of images retrieved using features generated from a Deep Boltzmann Machine by sampling from $P(\mathbf{v}_{img}|\mathbf{v}_{txt};\theta)$.

Experiments


	Step 50	Step 100	Step 150	Step 200	Step 250
	travel	beach	sea	water	italy
	trip	ocean	beach	canada	water
	vacation	waves	island	bc	sea
	africa	sea	vacation	britishcolumbia	boat
	earthasia	sand	travel	reflection	italia
	asia	nikon	ocean	alberta	mare
	men	surf	caribbean	lake	venezia
	2007	rocks	tropical	quebec	acqua
	india	coast	resort	ontario	ocean
	tourism	shore	trip	ice	venice

Figure 5: Text generated by the DBM conditioned on an image by running a Gibbs sampler. Ten words with the highest probability are shown at the end of every 50 sampling steps.


Input tags	Step 50	Step 100	Step 150	Step 200	Step 250
purple, flowers					
car, automobile					

Figure 6: Images retrieved by running a Gibbs sampler conditioned on the input tags. The images shown are those which are closest to the sampled image features. Samples were taken after every 50 steps.

Other stuff

- Classification: Use the weights obtained from optimizing probabilities $p(v,t)$ to initialize a neural network
- “Pretraining”
- Approximations used in optimization procedure
- Experiments on audio/video data e.g. learning $p(\text{audio}, \text{video})$

End

- Thanks!