10707
Deep Learning

Russ Salakhutdinov

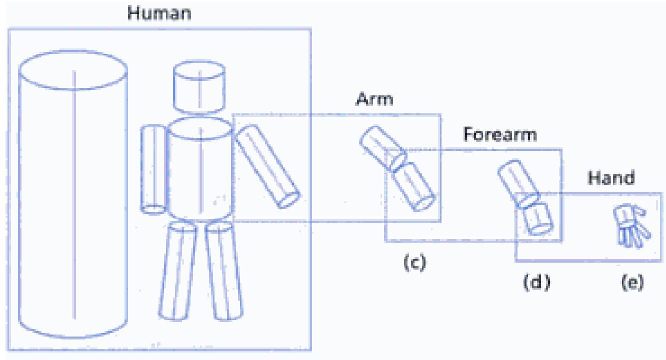
Machine Learning Department
rsalakhu@cs.cmu.edu
http://www.cs.cmu.edu/~rsalakhu/10707/

Deep Boltzmann Machines II

Learning Hierarchical Representations

Deep Boltzmann Machines:

Learning Hierarchical Structure in Features: edges, combination of edges.



- Performs well in many application domains
- Fast Inference: fraction of a second
- Learning scales to millions of examples

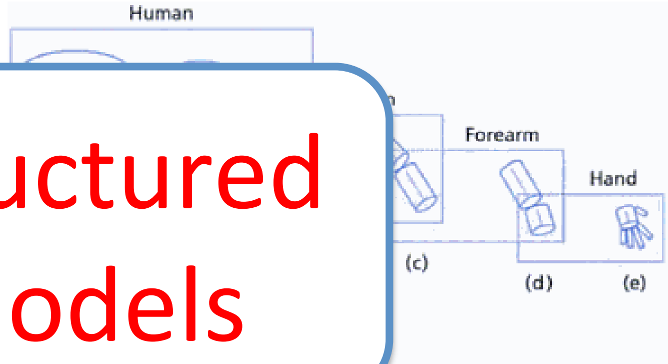
Learning Hierarchical Representations

Deep Boltzmann Machines:

Learning Hi

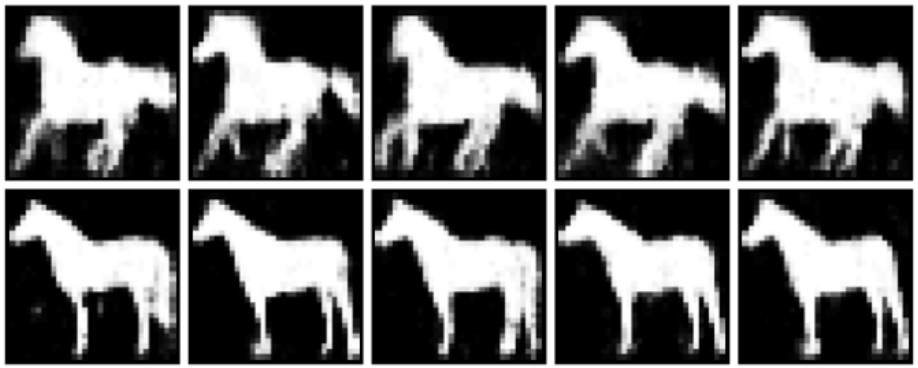
in Features

of edges.



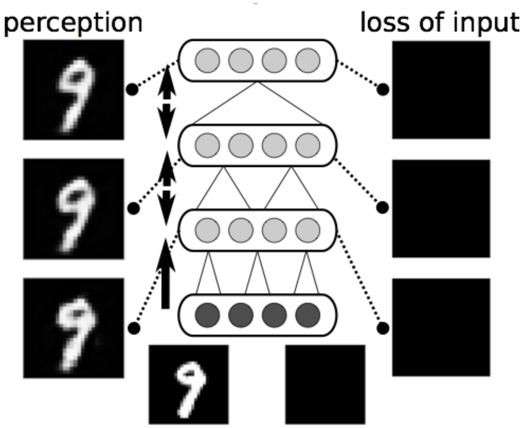
Need more structured and robust models

The Shape Boltzmann Machine: a Strong Model of Object Shape (Eslami, Heess, Winn, CVPR 2012).



[Demo DBM]

Hallucinations in Charles Bonnet Syndrome Induced by Homeostasis: a Deep Boltzmann Machine Model (Reichert, Series, Storkey, NIPS 2012)



Face Recognition

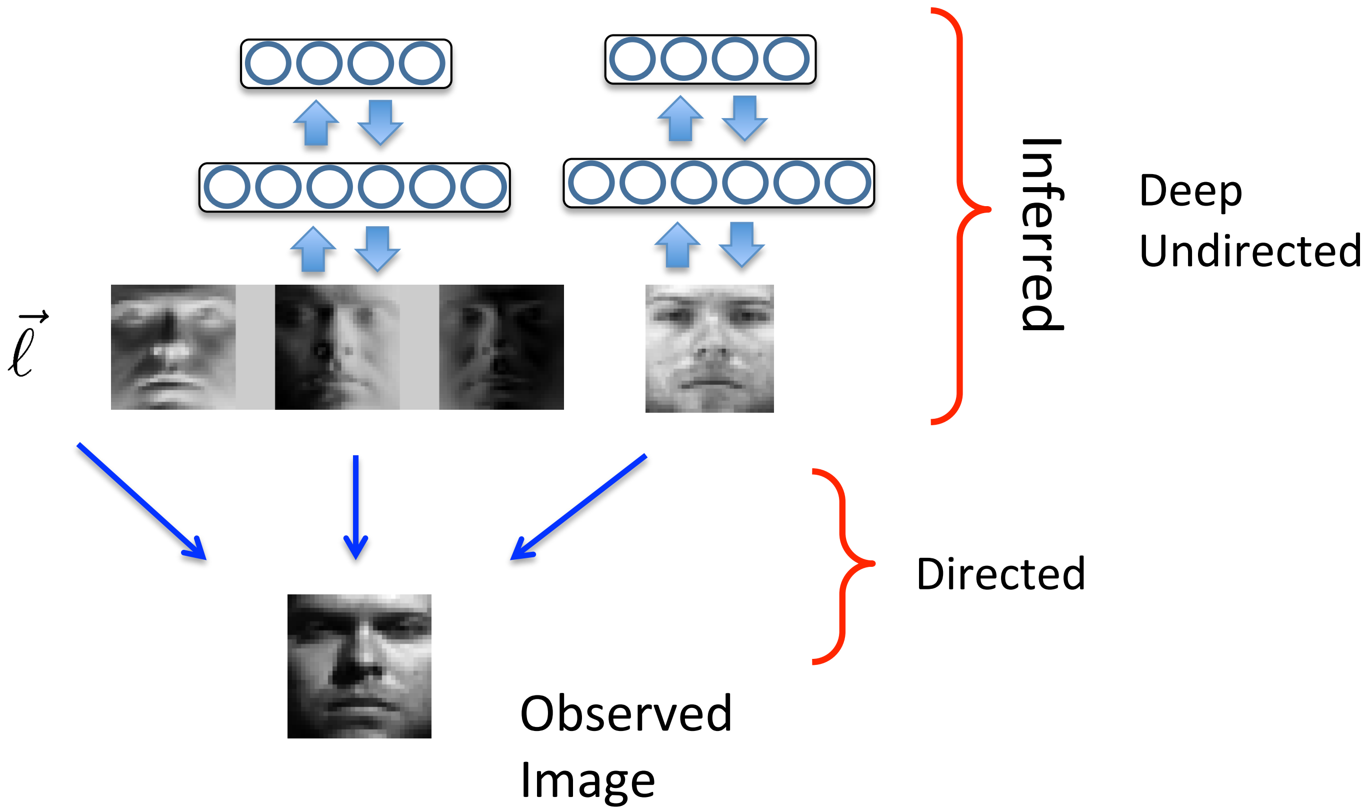
Yale B Extended Face Dataset

4 subsets of increasing illumination variations

Due to extreme illumination variations, deep models perform quite poorly on this dataset.

Deep Lambertian Model

Consider More Structured Models: undirected + directed models.



Combines the elegant properties of the Lambertian model with the Gaussian DBM model.

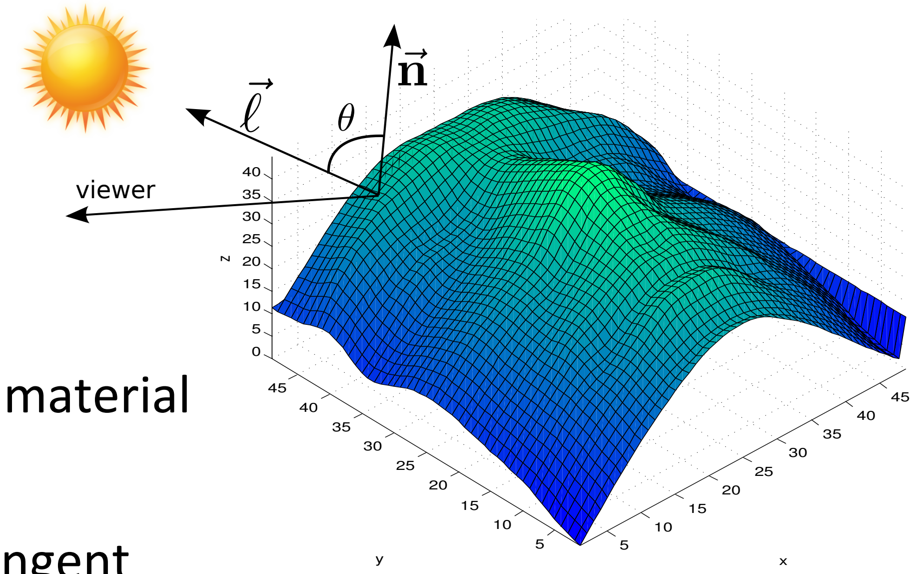
(Tang et. Al., ICML 2012, Tang et. al. CVPR 2012)

Lambertian Reflectance Model

- A simple model of the image formation process.

$$I = a \times |\vec{\ell}||\vec{\mathbf{n}}| \cos(\theta)$$

Image albedo   Light source   Surface normal

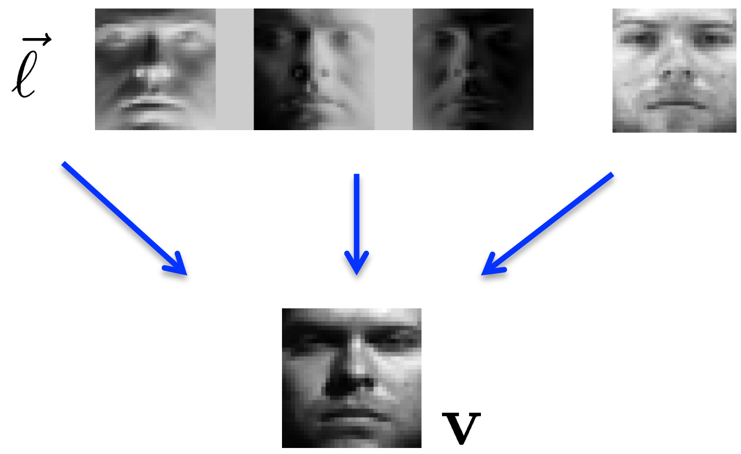


- Albedo -- diffuse reflectivity of a surface, material dependent, illumination independent.

- Surface normal -- perpendicular to the tangent plane at a point on the surface.

- Images with different illumination can be generated by varying light directions

Deep Lambertian Model



Observed Image

Image albedo

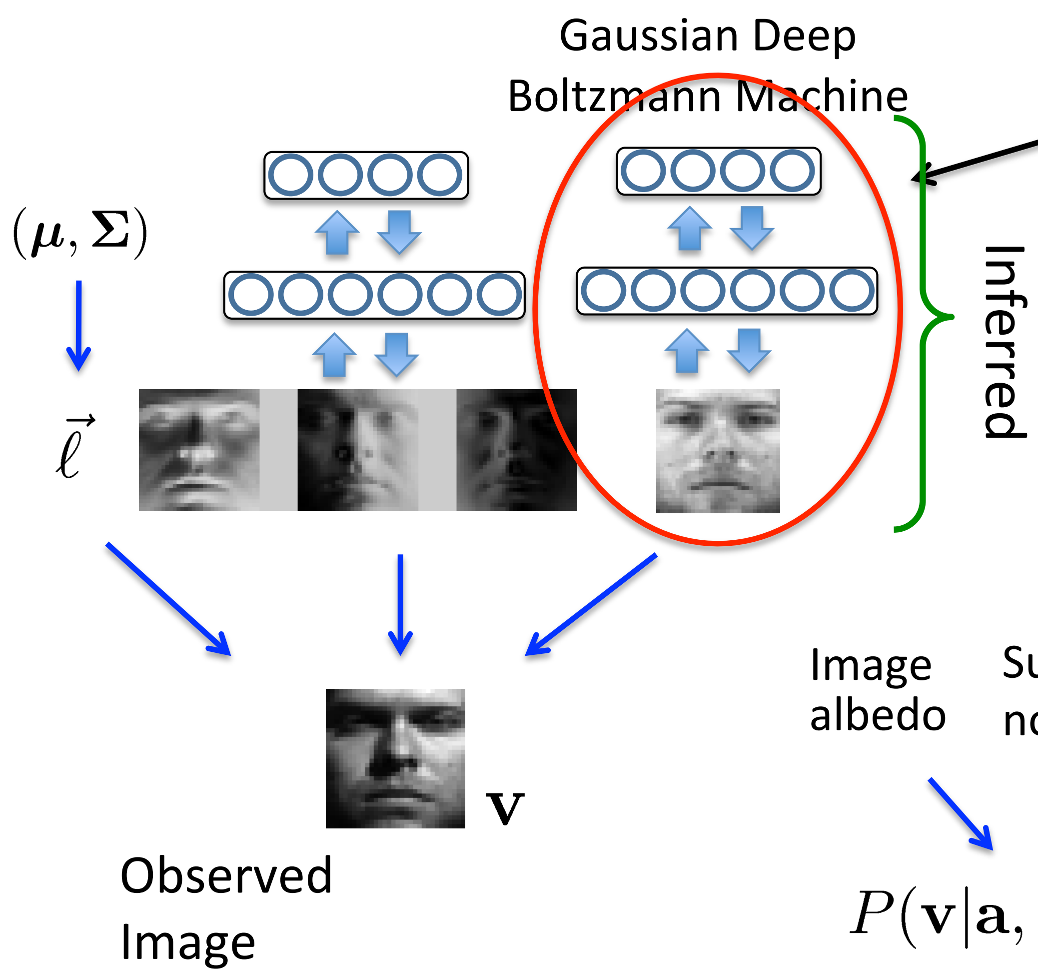
Surface normals

Light source

$$P(\mathbf{v}|\mathbf{a}, \mathbf{N}, \vec{\ell}) = \prod_{i \in \text{pixels}} \mathcal{N}\left(v_i | a_i(\vec{\mathbf{n}}_i^\top \vec{\ell}), \sigma^2\right)$$

$$\mathbf{a} \in \mathbb{R}^D, \quad \mathbf{N} \in \mathbb{R}^{D \times 3}, \quad \ell \in \mathbb{R}^3$$

Deep Lambertian Model



Transfer Learning

Albedo DBM:
Pretrained using
Toronto Face Database

$$\begin{aligned} \mathbf{a} &\sim \mathrm{GDBM}(\mathbf{a}) \\ \mathbf{N} &\sim \mathrm{GDBM}(\mathbf{N}) \\ \vec{\ell} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

Image albedo

Surface normals

Light source

$$P(\mathbf{v}|\mathbf{a}, \mathbf{N}, \vec{\ell}) = \prod_{i \in \text{pixels}} \mathcal{N}\left(v_i | a_i(\vec{\mathbf{n}}_i^\top \vec{\ell}), \sigma^2\right)$$

$$\mathbf{a} \in \mathbb{R}^D, \quad \mathbf{N} \in \mathbb{R}^{D \times 3}, \quad \ell \in \mathbb{R}^3$$

Inference: Variational Inference.

Learning: Stochastic Approximation

Yale B Extended Face Dataset

• 38 subjects, ~ 45 images of varying illuminations per subject, divided into 4 subsets of increasing illumination variations.

• 28 subjects for training, and 10 for testing.

Face Relighting

One Test Image

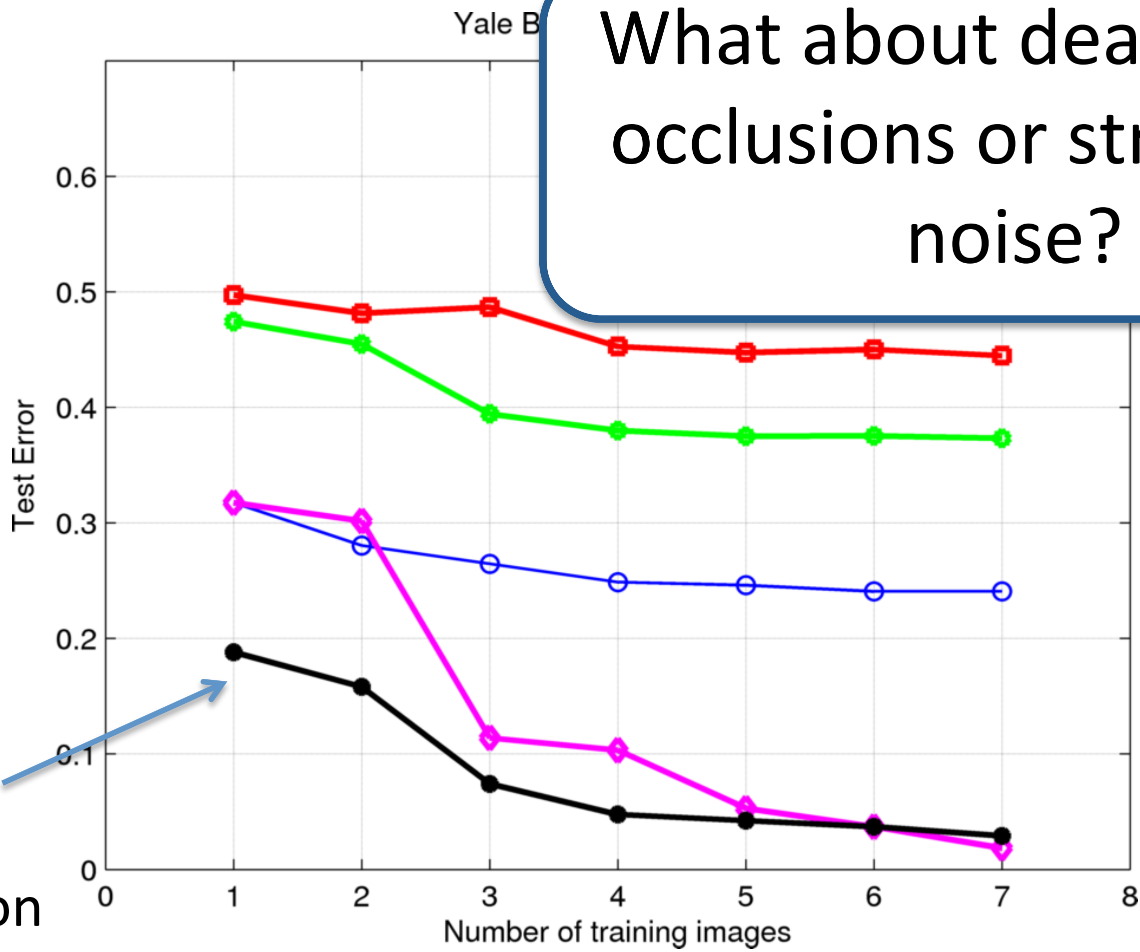
Observed

Inferred albedo

Face Relighting

Recognition Results

Recognition as function of the number of training images for 10 test subjects.



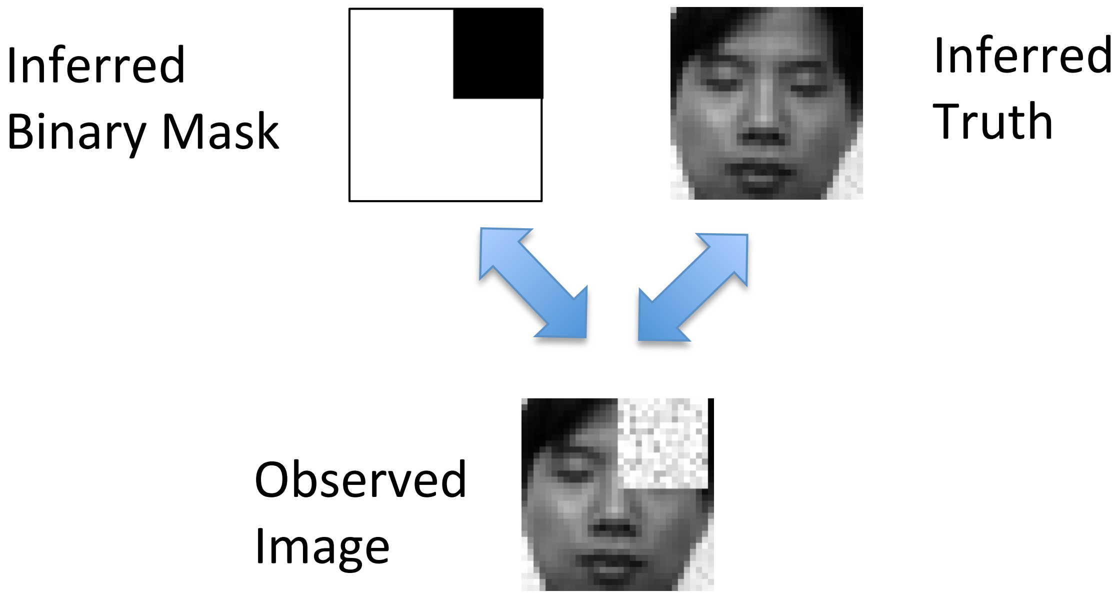
What about dealing with occlusions or structured noise?

One-Shot Recognition

Robust Boltzmann Machines

- Build more structured models that can deal with occlusions or structured noise.

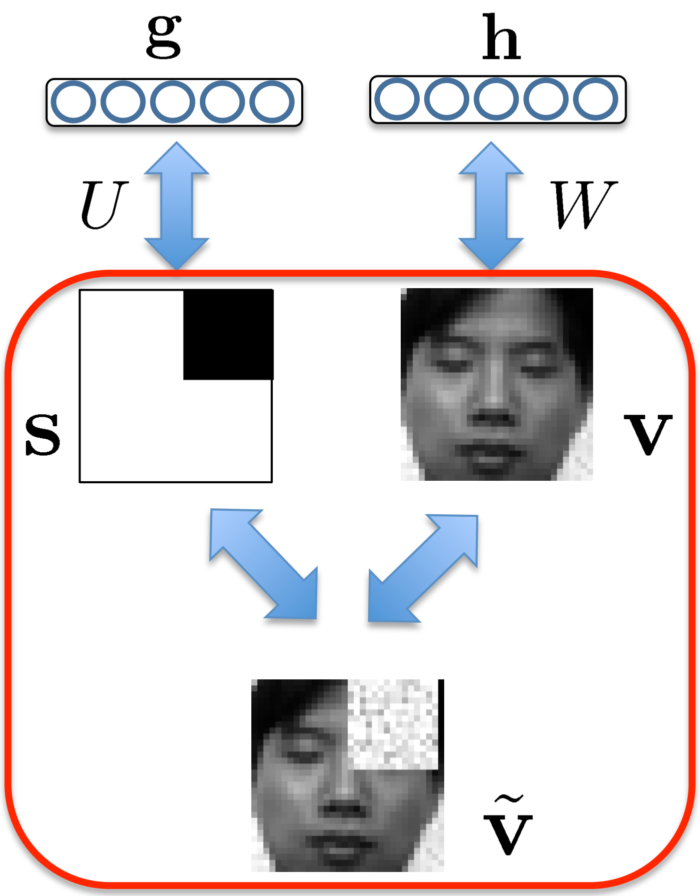
$$\log P(\tilde{\mathbf{v}}, \mathbf{v}, \mathbf{s}, \mathbf{h}, \mathbf{g}) \sim$$



(Tang et. Al., ICML 2012, Tang et. al. CVPR 2012)

Robust Boltzmann Machines

- Build more structured models that can deal with occlusions or structured noise.



Observed Image

$$\log P(\tilde{\mathbf{v}}, \mathbf{v}, \mathbf{s}, \mathbf{h}, \mathbf{g}) \sim$$

$$\underbrace{-\frac{1}{2}\sum_{i \in \text{pixels}} \frac{(v_i - b_i)^2}{\sigma_i^2} + \mathbf{v}^\top W \mathbf{h}}_{\text{Gaussian RBM, modeling clean faces}} + \underbrace{\mathbf{s}^\top U \mathbf{g}}_{\text{Binary RBM modeling occlusions}}$$

$$-\frac{1}{2}\sum_{i \in \text{pixels}} \gamma_i s_i (v_i - \tilde{v}_i)^2 - \frac{1}{2}\sum_{i \in \text{pixels}} \frac{(\tilde{v}_i - \tilde{b}_i)^2}{\tilde{\sigma}_i^2}$$

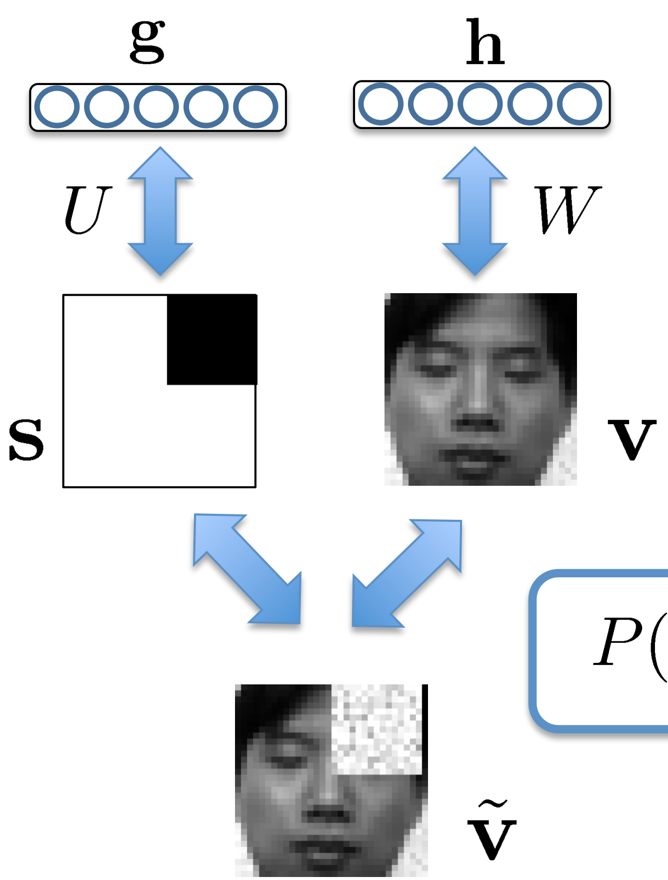
Binary pixel-wise Mask

Gaussian noise model

Robust Boltzmann Machines

(Tang et. Al., ICML 2012, Tang et. al. CVPR 2012)

- Build more structured models that can deal with occlusions or structured noise.



$$\log P(\tilde{\mathbf{v}}, \mathbf{v}, \mathbf{s}, \mathbf{h}, \mathbf{g}) \sim$$

$$\underbrace{-\frac{1}{2}\sum_{i \in \text{pixels}} \frac{(v_i - b_i)^2}{\sigma_i^2} + \mathbf{v}^\top W \mathbf{h}}_{\text{Gaussian RBM, modeling clean faces}} + \underbrace{\mathbf{s}^\top U \mathbf{g}}_{\text{Binary RBM modeling occlusions}}$$

$P(\tilde{\mathbf{v}}|\mathbf{h}, \mathbf{g})$ is a heavy-tailed distribution

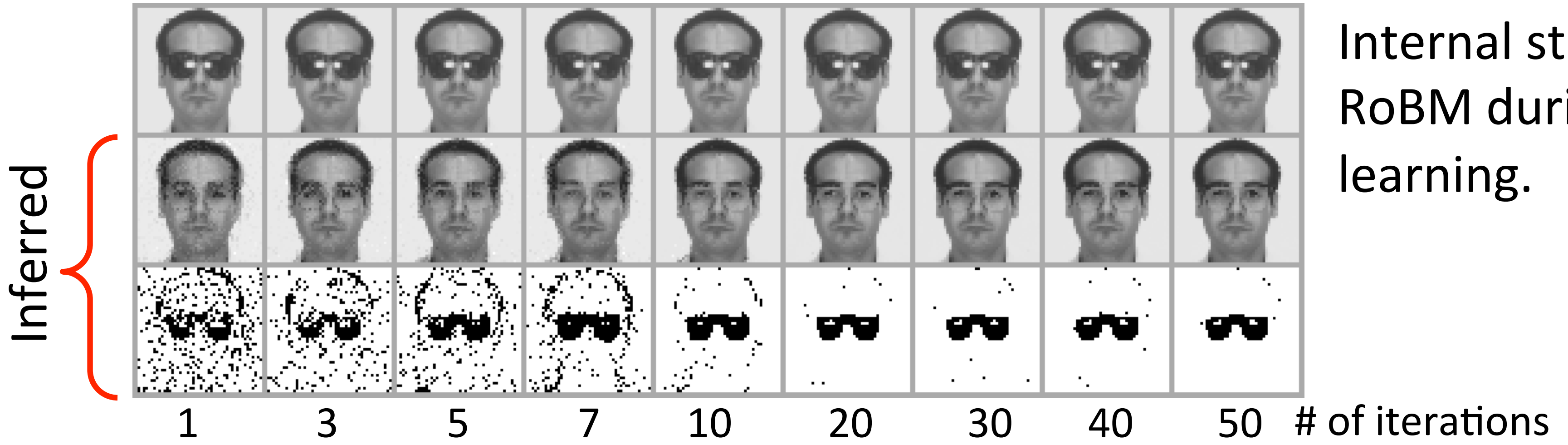
Binary pixel-wise Mask

Gaussian noise model

Inference: Variational Inference.

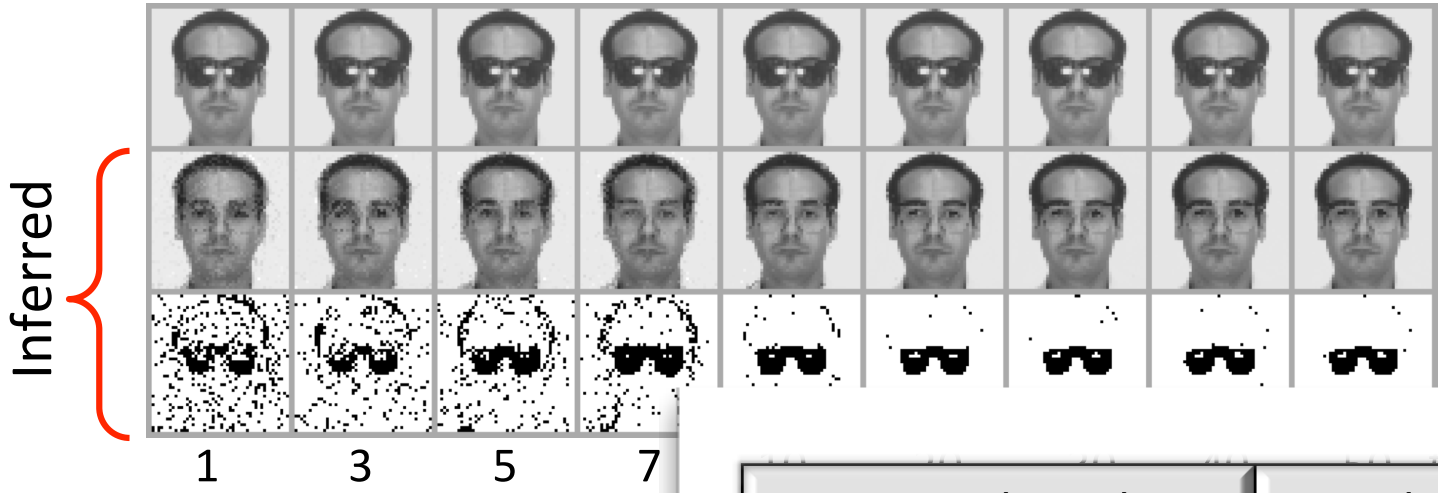
Learning: Stochastic Approximation

Recognition Results on AR Face Database



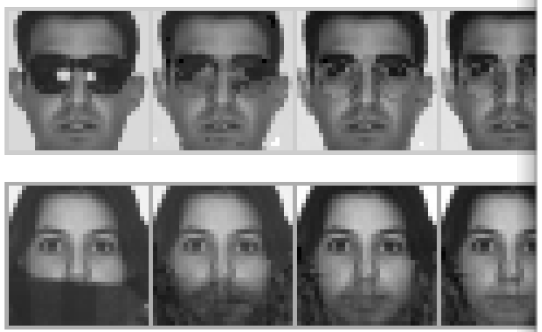
Internal states of RoBM during learning.

Recognition Results on AR Face Database



Internal states of RoBM during learning.

Inference on the

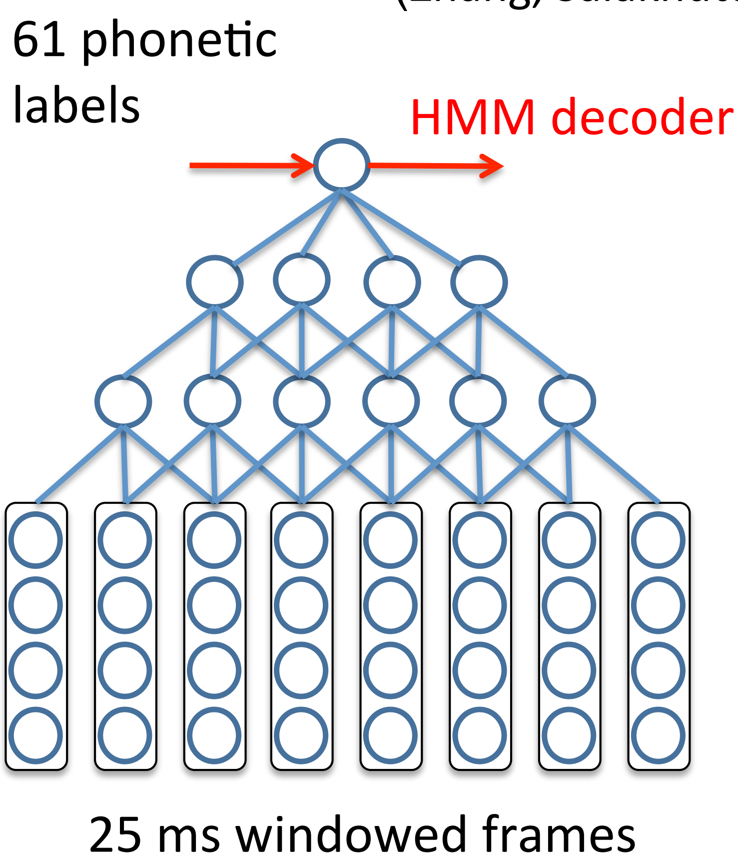


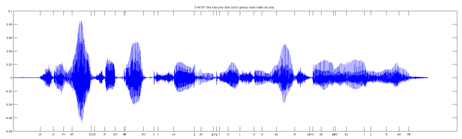
# of iteratio

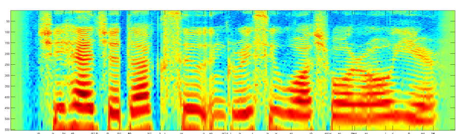
Learning Algorithm	Sunglasses	Scarf
Robust BM	84.5%	80.7%
RBM	61.7%	32.9%
Eigenfaces	66.9%	38.6%
LDA	56.1%	27.0%
Pixel	51.3%	17.5%

Speech Recognition

(Zhang, Salakhutdinov, Chang, Glass, ICASSP 2012)







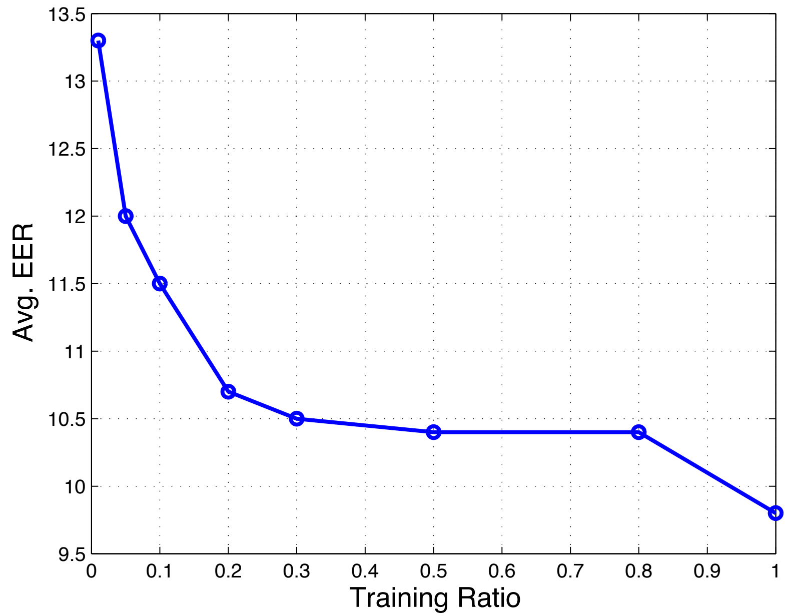
- 630 speaker TIMIT corpus: 3,696 training and 944 test utterances.
- **Spoken Query Detection**:
  For each keyword, estimate utterance's probability of containing that keyword.
- Performance: Average equal error rate (EER).

Learning Algorithm	AVG EER
GMM Unsupervised	16.4%
DBM Unsupervised	14.7%
DBM (1% labels)	13.3%
DBM (30% labels)	10.5%
DBM (100% labels)	9.7%

Spoken Query Detection

- 630 speaker TIMIT corpus: 3,696 training and 944 test utterances.
- 10 query keywords were randomly selected and 10 examples of each keyword were extracted from the training set.
- **Goal**: For each keyword, rank all 944 utterances based on the utterance's probability of containing that keyword.
- Performance measure: The average equal error rate (EER).

Learning Algorithm	AVG EER
GMM Unsupervised	16.4%
DBM Unsupervised	14.7%
DBM (1% labels)	13.3%
DBM (30% labels)	10.5%
DBM (100% labels)	9.7%

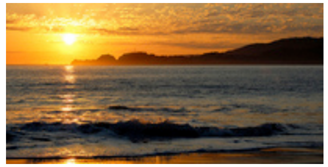


(Yaodong Zhang et.al. ICASSP 2012)

Data – Collection of Modalities

- Multimedia content on the web - image + text + audio.

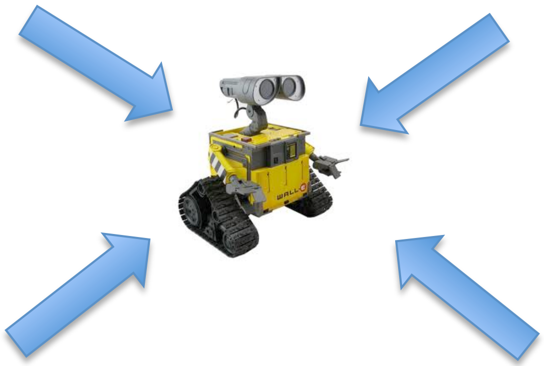
car, automobile



sunset, pacificocean, bakerbeach, seashore, ocean

- Product recommendation systems.

- Robotics applications.



Ngiam et. al. 2011
Huiskes, Thomee, Lew 2010
Guillaumin, Verbeek, Schmid 2010
Xing, Yan, and Hauptmann. 2005

Shared Concept

“Modality-free” representation

“Concept”

sunset, pacificocean, bakerbeach, seashore, ocean

“Modality-full” representation

Multi-Modal Input

- Improve Classification

pentax, k10d, kangarooisland
southaustralia, sa australia
australiansealion 300mm

SEA / NOT SEA

- Fill in Missing Modalities

beach, sea, surf,
strand, shore,
wave, seascape,
sand, ocean, waves

- Retrieve data from one modality when queried using data from another modality

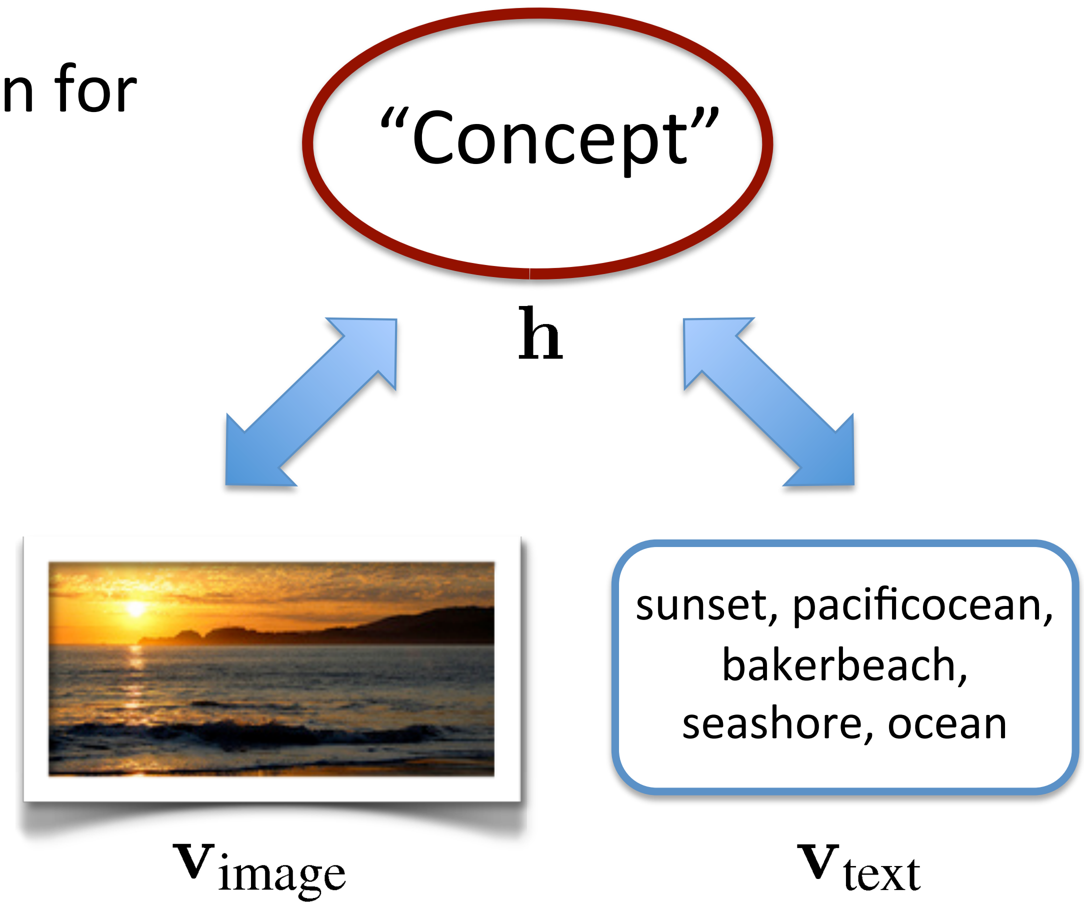
beach, sea, surf,
strand, shore,
wave, seascape,
sand, ocean, waves

Building a Probabilistic Model

- Learn a joint density model: $P(\mathbf{h}, \mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}})$.

- **h**: “fused” representation for classification, retrieval.

$$P(\mathbf{h} | \mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}})$$



Building a Probabilistic Model

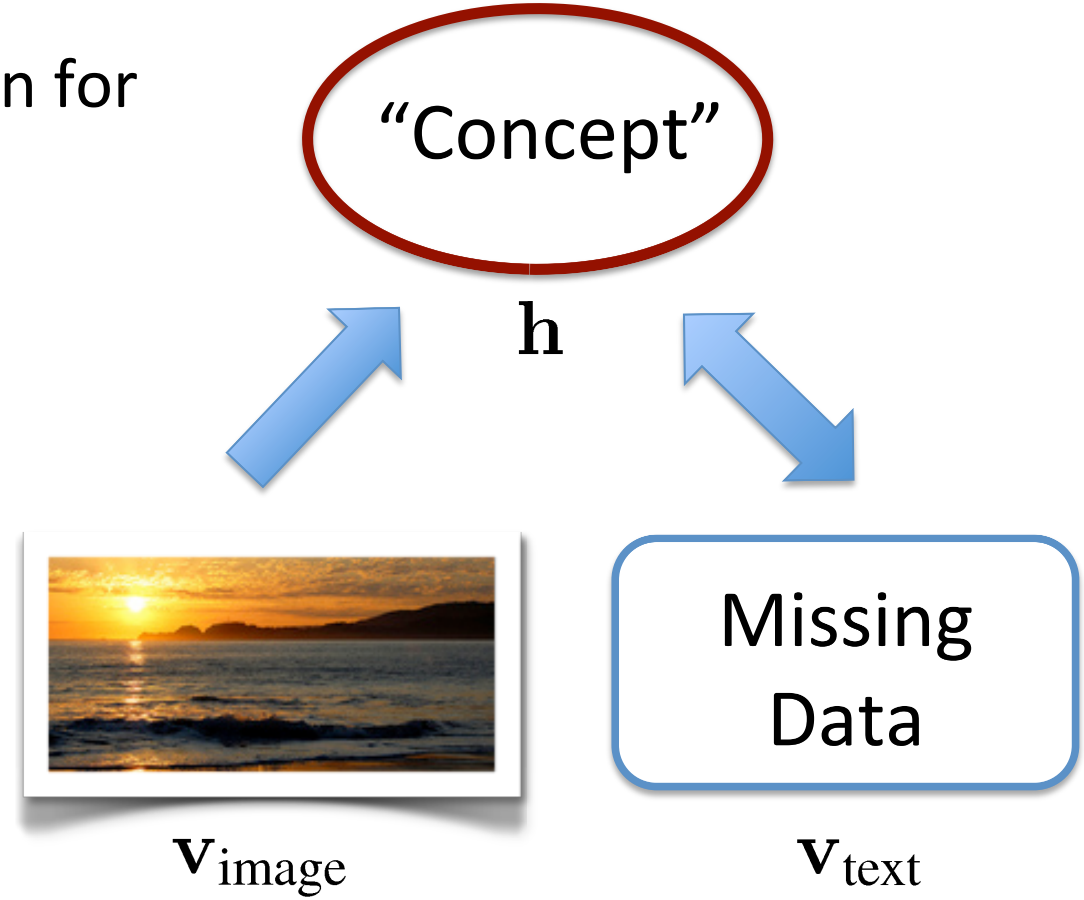
- Learn a joint density model: $P(\mathbf{h}, \mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}})$.

- **h**: "fused" representation for classification, retrieval.

- Generate data from conditional distributions for

  - Image Annotation

$$P(\mathbf{h}, \mathbf{v}_{\text{text}} | \mathbf{v}_{\text{image}})$$

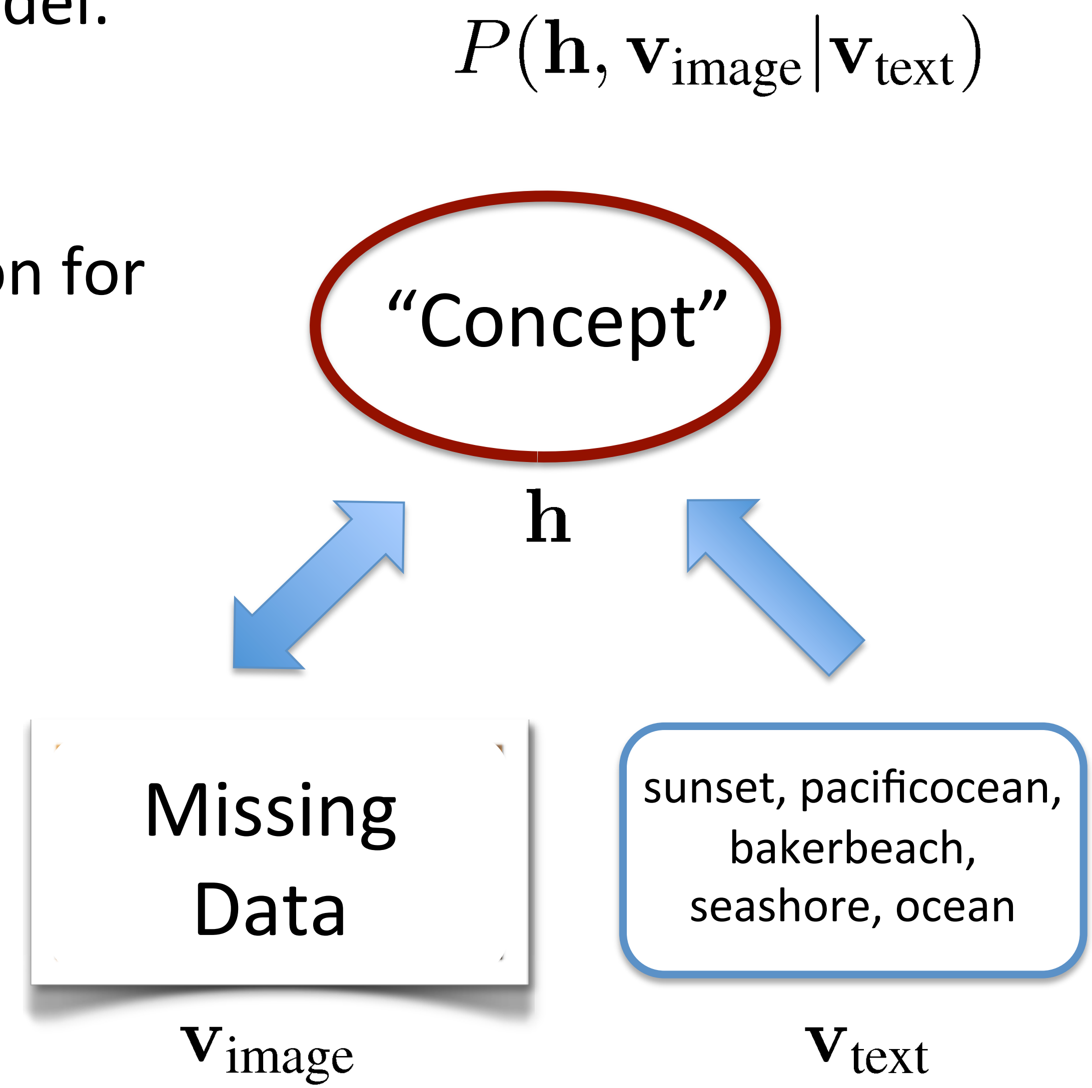


Building a Probabilistic Model

- Learn a joint density model: $P(\mathbf{h}, \mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}})$.

- **h**: "fused" representation for classification, retrieval.

- Generate data from conditional distributions for
  - Image Annotation
  - Image Retrieval



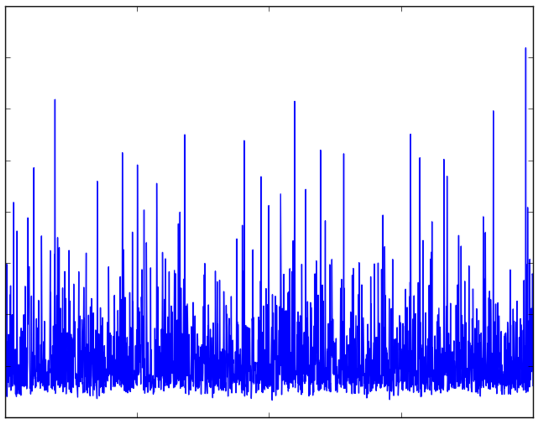
Challenges - I

Image

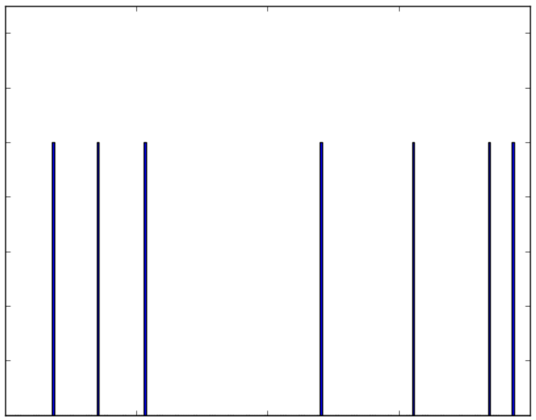
Text

sunset, pacificocean, bakerbeach, seashore, ocean

Dense



Sparse



Very different input representations

- Images – real-valued, dense
- Text – discrete, sparse

Difficult to learn cross-modal features from low-level representations.

Challenges - II

Image | Text

pentax, k10d,
pentaxda50200,
kangarooisland, sa,
australiansealion

mickikrimmel,
mickipedia,
headshot

< no text>

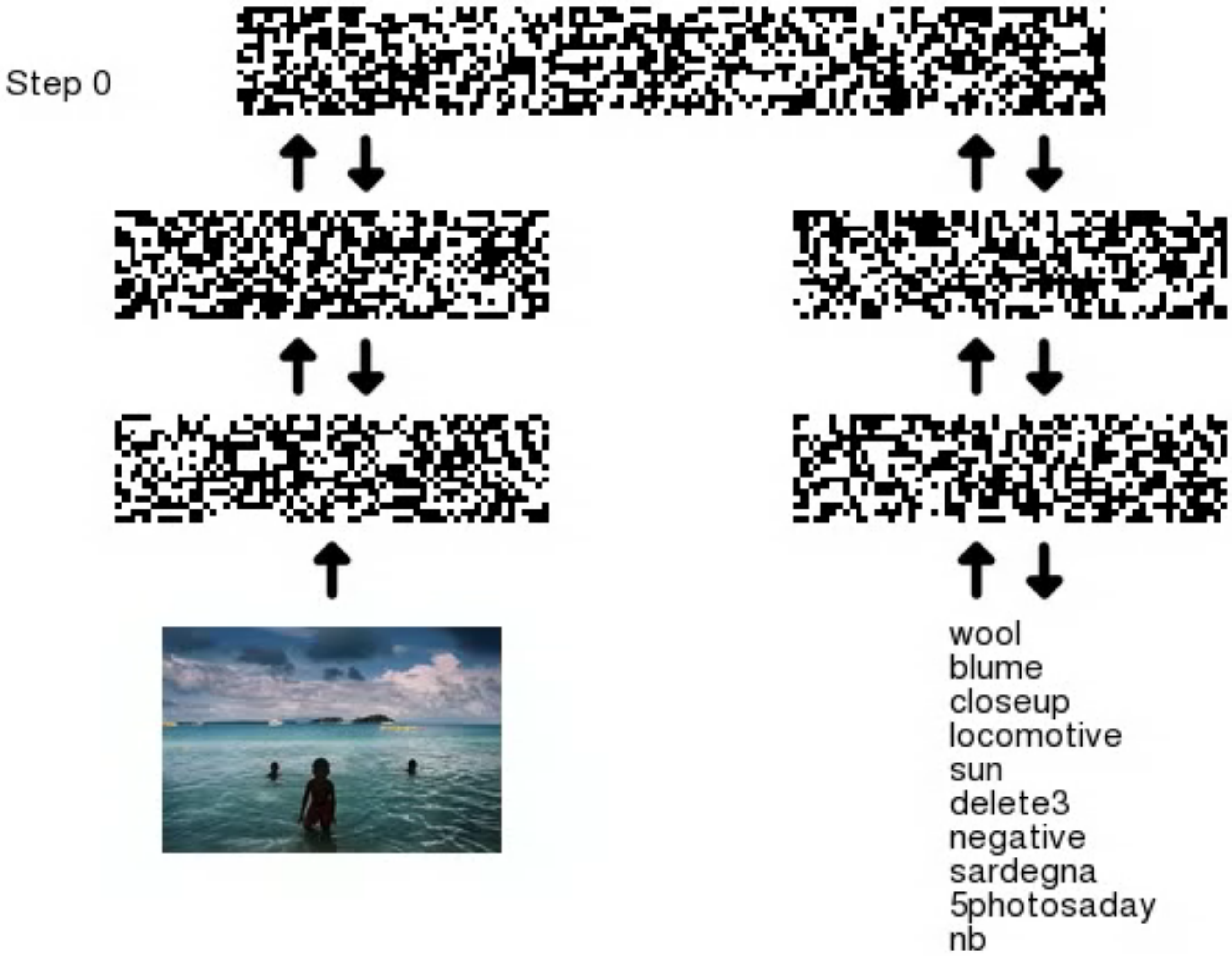
unseulpixel,
naturey,

Noisy and missing data

Challenges - II

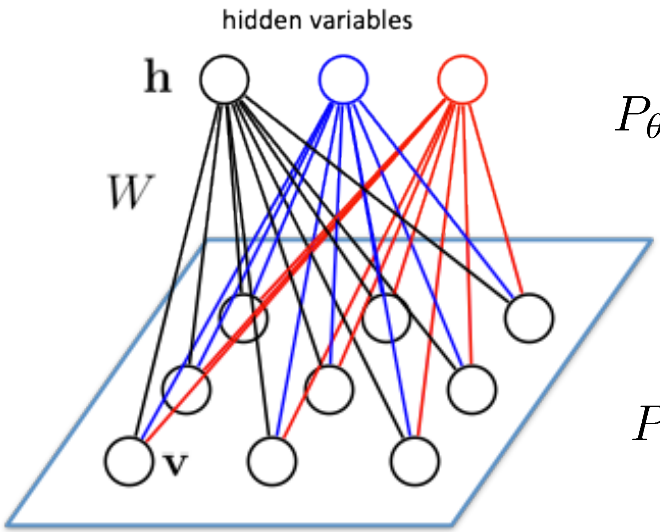
Image	Text	Text generated by the model
	pentax, k10d, pentaxda50200, kangarooisland, sa, australiansealion	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves
	mickikrimmel, mickipedia, headshot	portrait, girl, woman, lady, blonde, pretty, gorgeous, expression, model
	< no text>	night, notte, traffic, light, lights, parking, darkness, lowlight, nacht, glow
	unseulpixel, naturey,	fall, autumn, trees, leaves, foliage, forest, woods, branches, path

Generating Text from Images



Samples drawn after every 50 steps of Gibbs updates

Restricted Boltzmann Machines



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

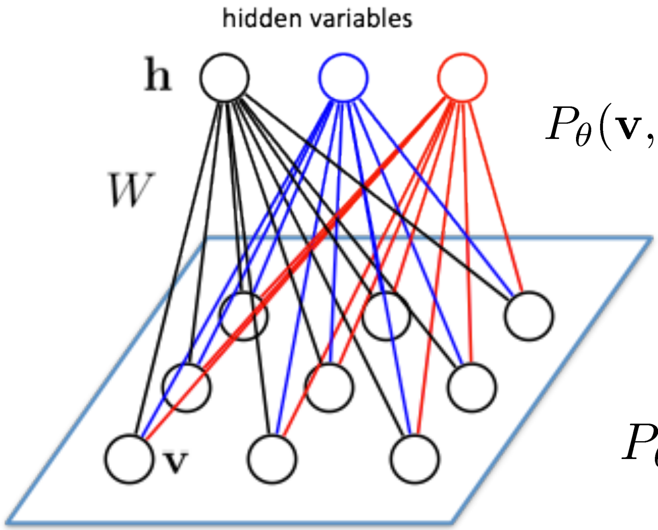
$$\theta = \{W, a, b\}$$

$$P_\theta(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{D} P_\theta(v_i|\mathbf{h}) = \prod_{i=1}^{D} \frac{1}{1+\exp(-\sum_{j=1}^{F} W_{ij} v_i h_j - b_i)}$$

RBM is a Markov Random Field with:

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

RBMs for Real-valued Data



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

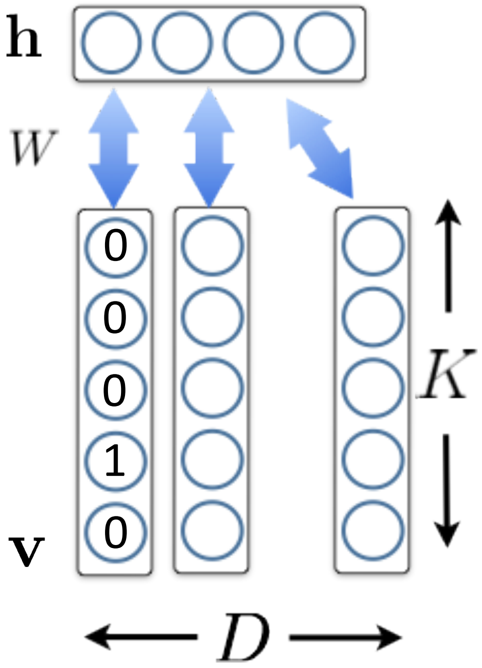
$$\theta = \{W, a, b\}$$

$$P_\theta(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{D} P_\theta(v_i|\mathbf{h}) = \prod_{i=1}^{D} \mathcal{N}\left(b_i + \sum_{j=1}^{N} W_{ij} h_j, \sigma_i^2 \right)$$

Gaussian-Bernoulli RBM:

- Stochastic real-valued visible variables $\mathbf{v} \in \mathbb{R}^D$.
- Stochastic binary hidden variables $\mathbf{h} \in \{0,1\}^F$.
- Bipartite connections.

RBMs for Word Counts



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

$$\theta = \{W, a, b\}$$

$$P_\theta(v_i^k = 1|\mathbf{h}) = \frac{\exp\left(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k\right)}{\sum_{q=1}^{K} \exp\left(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q\right)}$$

RBM Replicated Softmax Model: undirected topic model:

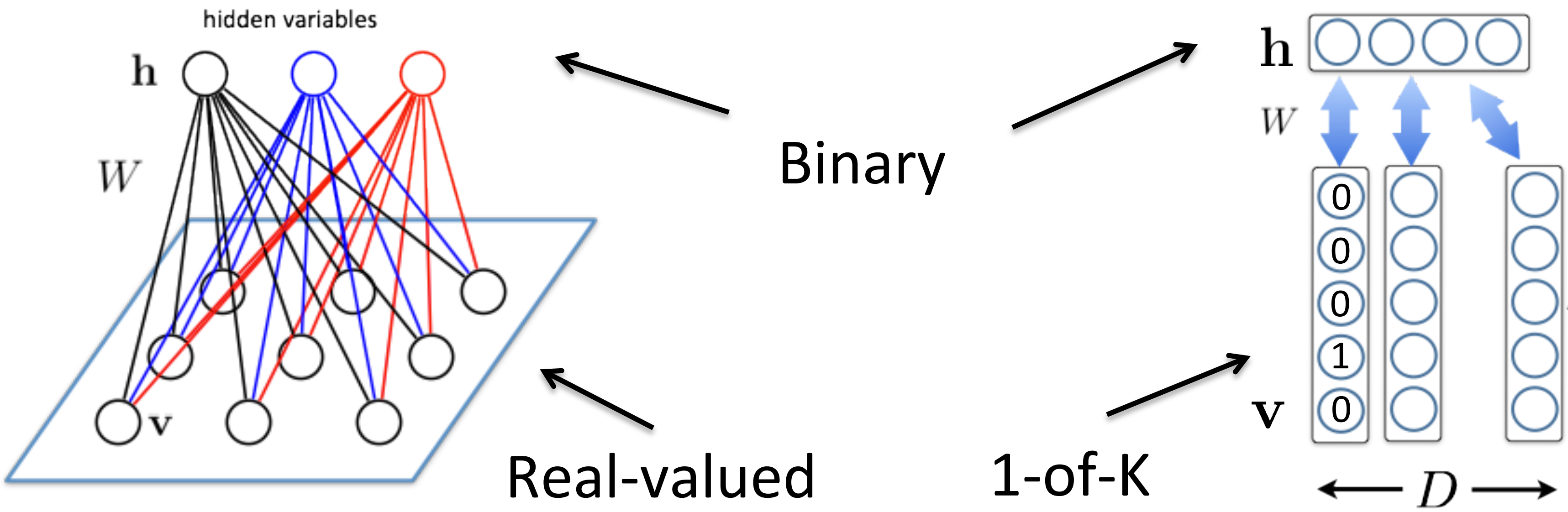
- Stochastic 1-of-K visible variables.
- Stochastic binary hidden variables $\mathbf{h} \in \{0,1\}^F$.
- Bipartite connections.

A Nice Thing about RBMs

- It is easy to infer the states of the hidden variables:

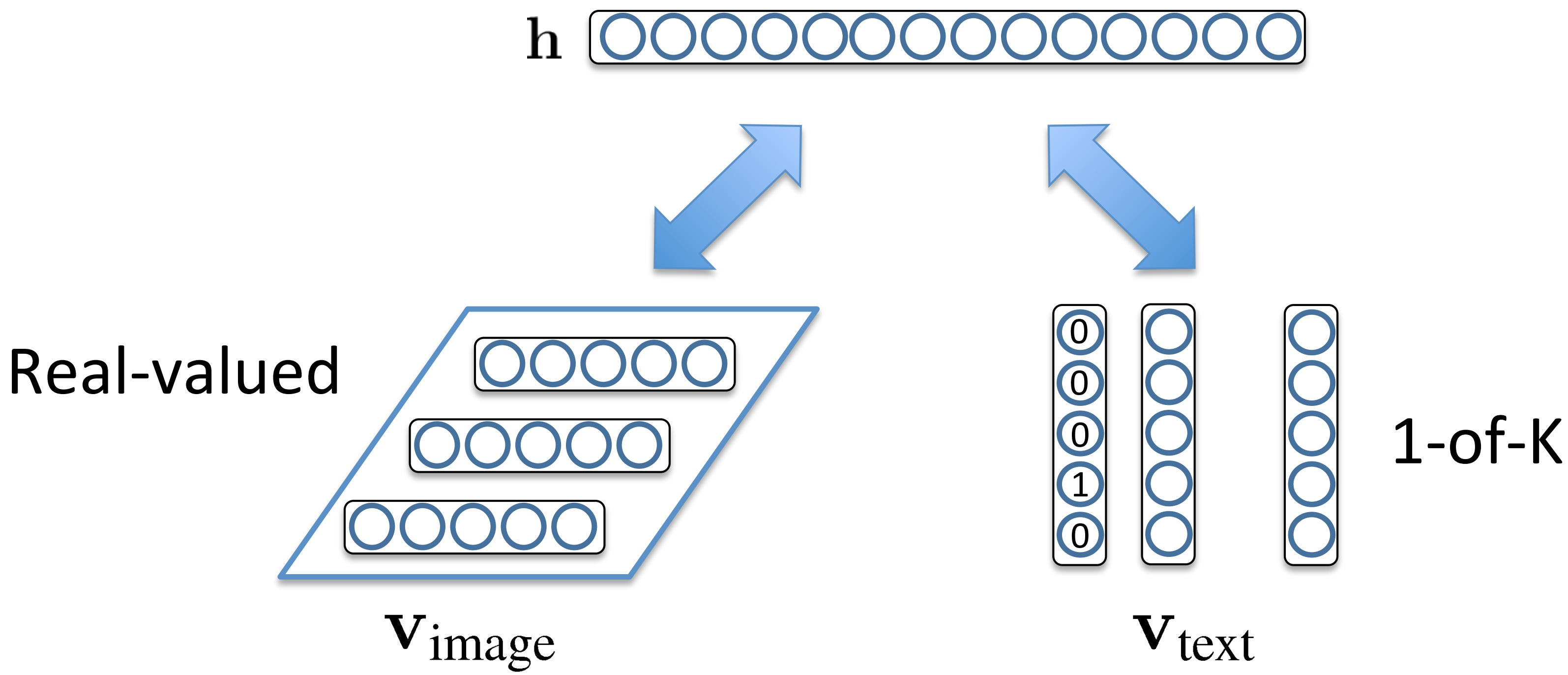
$$P_\theta(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{F} P_\theta(h_j|\mathbf{v}) = \prod_{j=1}^{F} \frac{1}{1+\exp(-a_j - \sum_{i=1}^{D} W_{ij} v_i)}$$

- Binary/Gaussian/Softmax RBMs: All have binary hidden variables but use them to model different kinds of data.



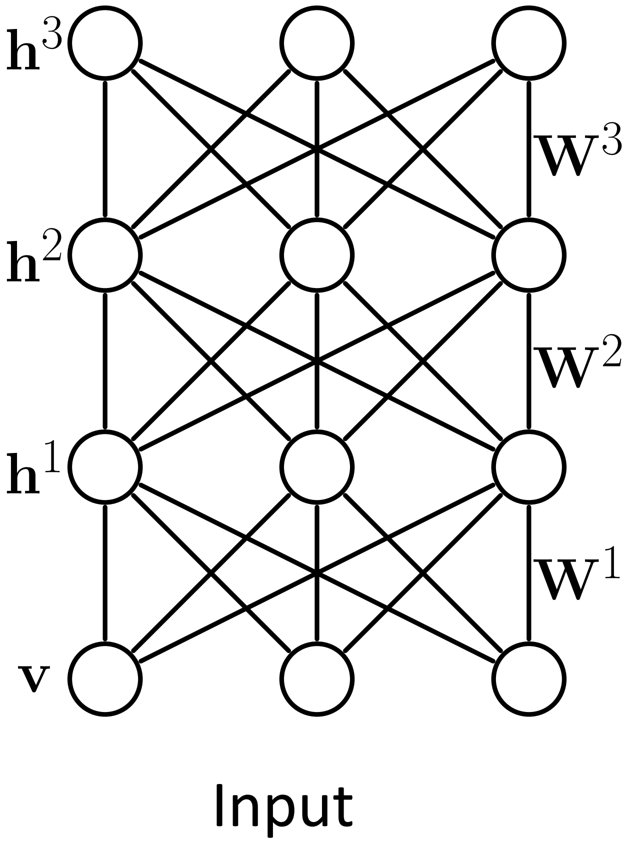
A Simple Multimodal Model

- Use a joint binary hidden layer.
- **Problem**: Inputs have very different statistical properties.
- Difficult to learn cross-modal features.



Deep Boltzmann Machines

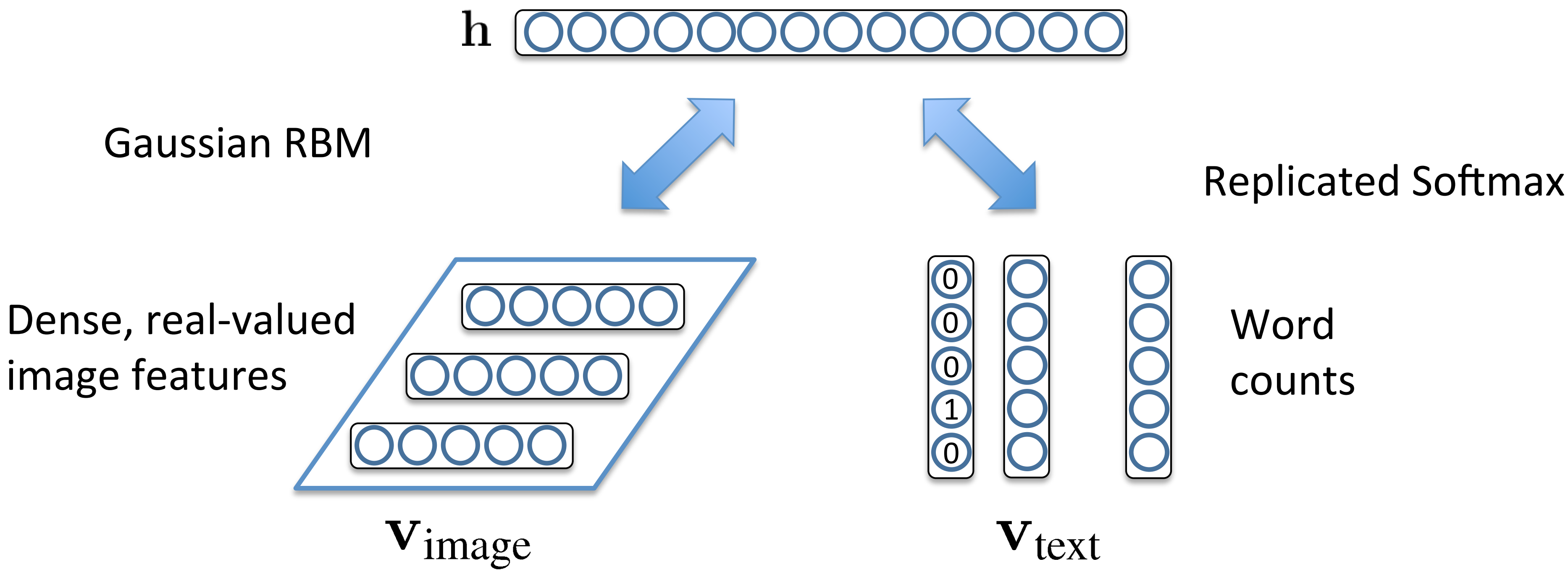
$$P_\theta(\mathbf{v}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{h}^{(3)}) = \frac{1}{\mathcal{Z}(\theta)} \exp \left[\underbrace{\mathbf{v}^\top W^{(1)} \mathbf{h}^{(1)}}_{\text{Same as RBMs}} + \underline{\mathbf{h}^{(1)\top} W^{(2)} \mathbf{h}^{(2)}} + \underline{\mathbf{h}^{(2)\top} W^{(3)} \mathbf{h}^{(3)}} \right]$$



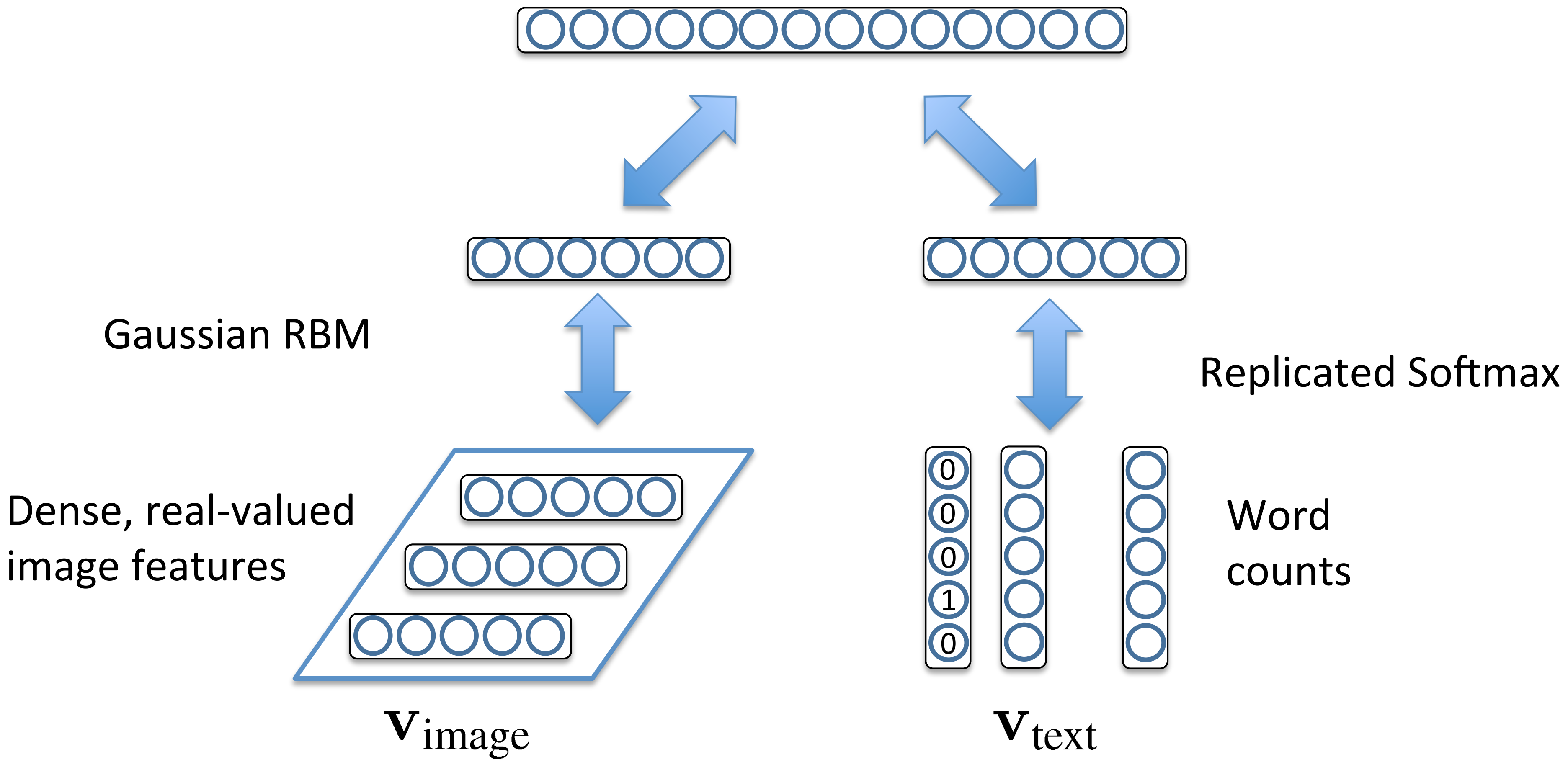
$\theta = \{W^1, W^2, W^3\}$ model parameters.

- Dependencies between hidden variables.
- All connections are undirected.
- Hidden variables are dependent even when **conditioned on the input**.

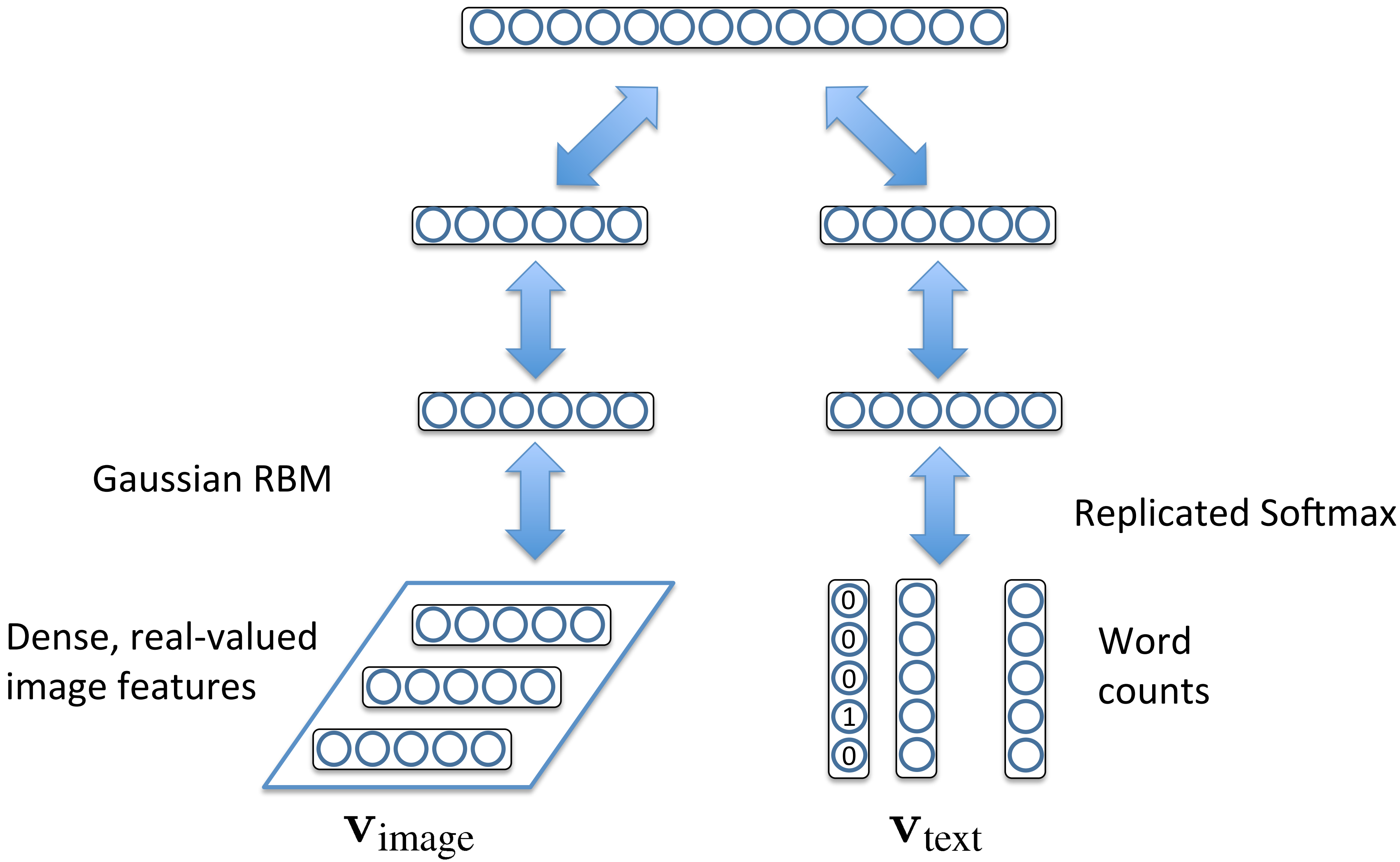
Multimodal DBM



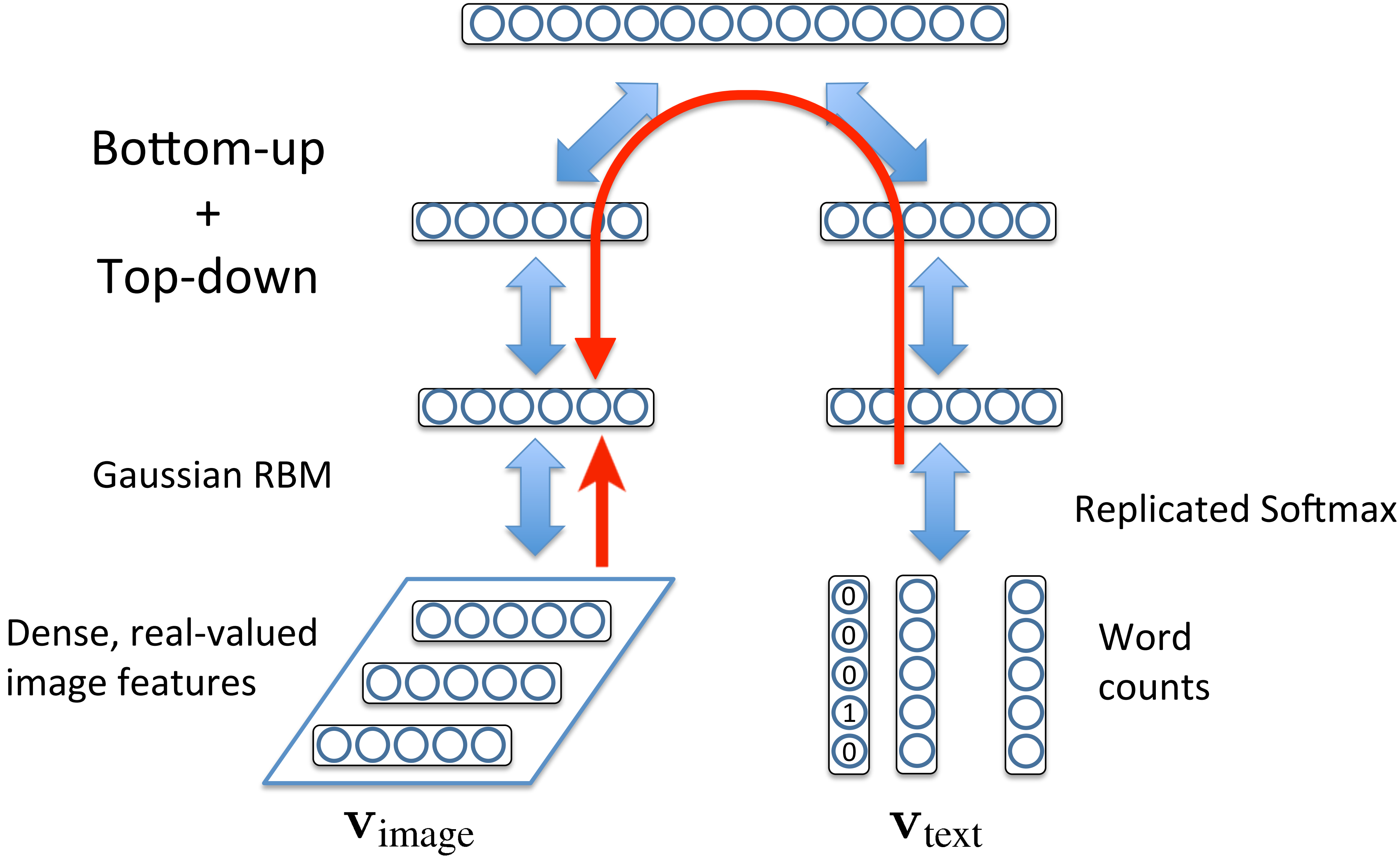
Multimodal DBM



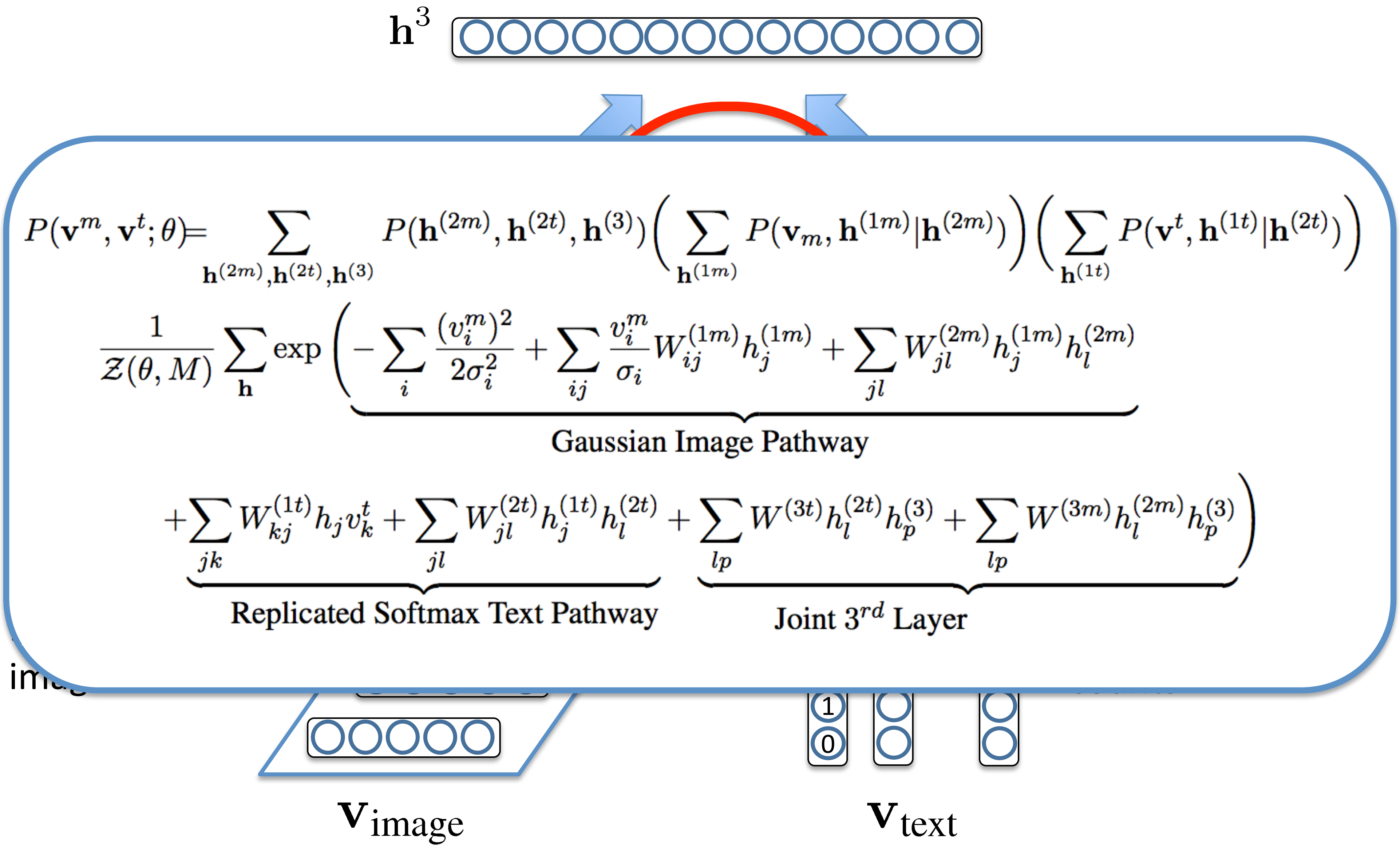
Multimodal DBM



Multimodal DBM



Multimodal DBM

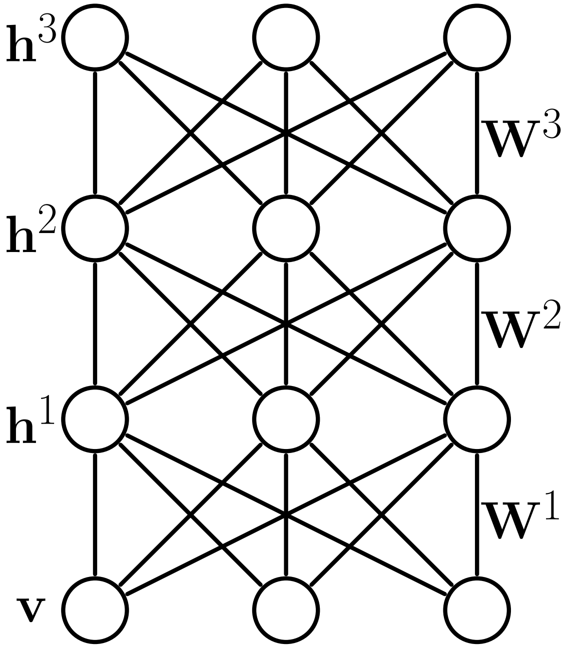


$$P(\mathbf{v}^m, \mathbf{v}^t; \theta) = \sum_{\mathbf{h}^{(2m)}, \mathbf{h}^{(2t)}, \mathbf{h}^{(3)}} P(\mathbf{h}^{(2m)}, \mathbf{h}^{(2t)}, \mathbf{h}^{(3)}) \Big(\sum_{\mathbf{h}^{(1m)}} P(\mathbf{v}_m, \mathbf{h}^{(1m)} | \mathbf{h}^{(2m)}) \Big) \Big(\sum_{\mathbf{h}^{(1t)}} P(\mathbf{v}^t, \mathbf{h}^{(1t)} | \mathbf{h}^{(2t)}) \Big)$$

$$\frac{1}{\mathcal{Z}(\theta, M)} \sum_{\mathbf{h}} \exp \Bigg(\underbrace{- \sum_i \frac{(v_i^m)^2}{2\sigma_i^2} + \sum_{ij} \frac{v_i^m}{\sigma_i} W_{ij}^{(1m)} h_j^{(1m)} + \sum_{jl} W_{jl}^{(2m)} h_j^{(1m)} h_l^{(2m)}}_{\text{Gaussian Image Pathway}}$$

$$+ \underbrace{\sum_{jk} W_{kj}^{(1t)} h_j v_k^t + \sum_{jl} W_{jl}^{(2t)} h_j^{(1t)} h_l^{(2t)}}_{\text{Replicated Softmax Text Pathway}} + \underbrace{\sum_{lp} W^{(3t)} h_l^{(2t)} h_p^{(3)} + \sum_{lp} W^{(3m)} h_l^{(2m)} h_p^{(3)}}_{\text{Joint } 3^{rd} \text{ Layer}} \Bigg)$$

Learning DBMs



(Approximate) Maximum Likelihood:

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

Mean-field

MCMC
(Gibbs sampling)

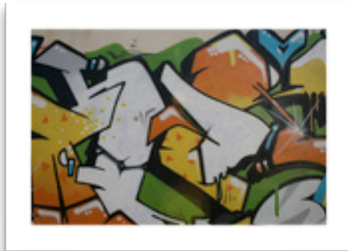
$$P_{data}(\mathbf{v}, \mathbf{h^1}) = P_\theta(\mathbf{h^1}|\mathbf{v}) P_{data}(\mathbf{v})$$

$$P_{data}(\mathbf{v}) = \frac{1}{N}\sum_{n=1}^{N} \delta(\mathbf{v} - \mathbf{v_n})$$

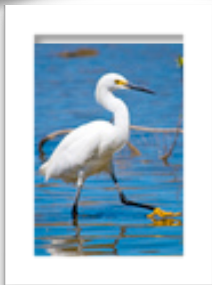
Not factorial any more!

Pretraining using a stack of PCD trained RBMs.

Text Generated from Images

Given	Generated	Given	Generated
	dog, cat, pet, kitten, puppy, ginger, tongue, kitty, dogs, furry		insect, butterfly, insects, bug, butterflies, lepidoptera
	sea, france, boat, mer, beach, river, bretagne, plage, brittany		graffiti, streetart, stencil, sticker, urbanart, graff, sanfrancisco
	portrait, child, kid, ritratto, kids, children, boy, cute, boys, italy		canada, nature, sunrise, ontario, fog, mist, bc, morning

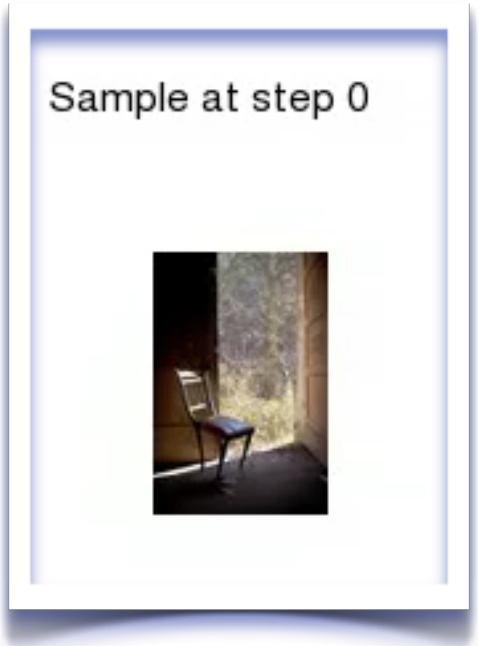
Text Generated from Images

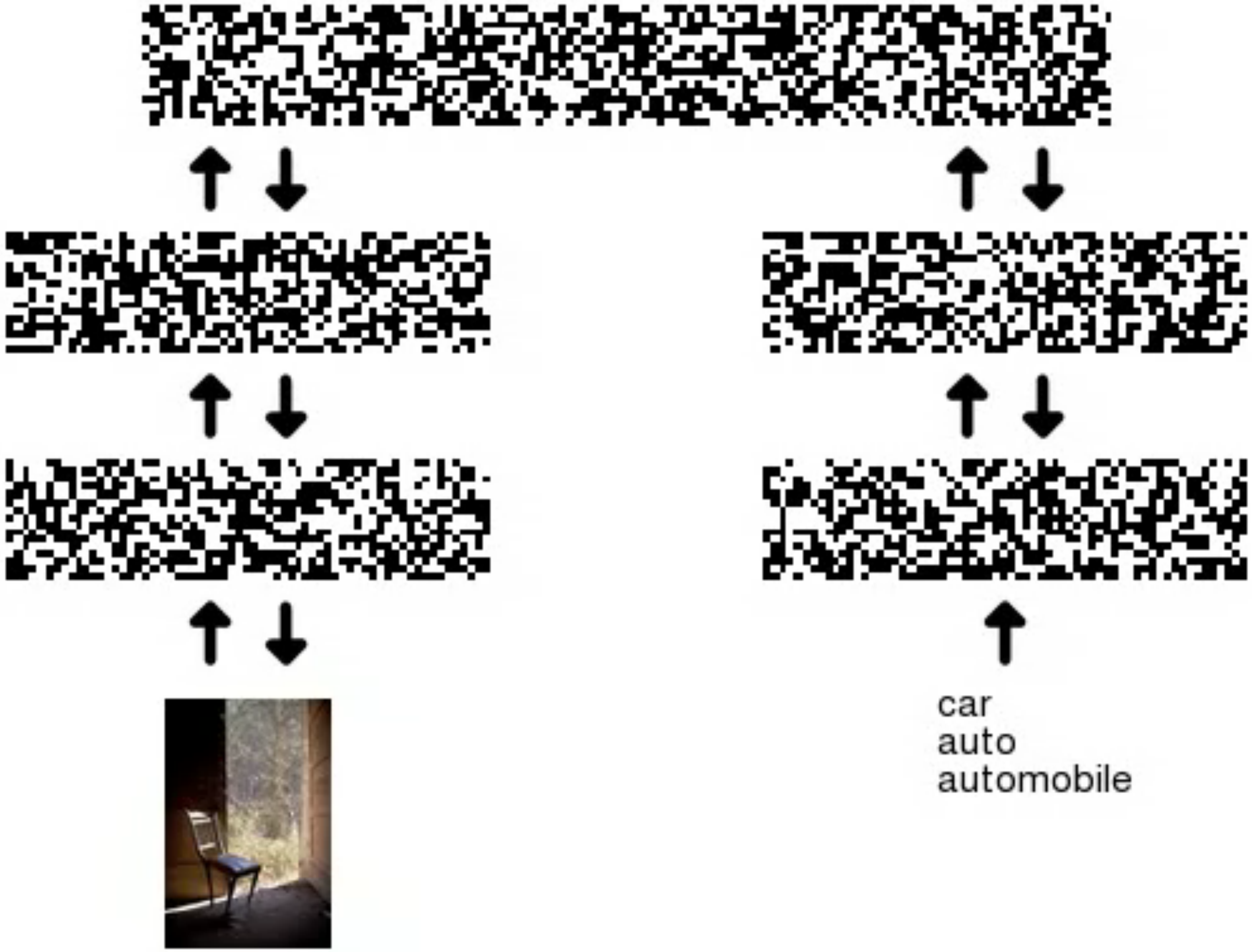
Given	Generated
	portrait, women, army, soldier, mother, postcard, soldiers
	obama, barackobama, election, politics, president, hope, change, sanfrancisco, convention, rally
	water, glass, beer, bottle, drink, wine, bubbles, splash, drops, drop

Images from Text

Step 0

Sample drawn after every 50 steps of Gibbs sampling





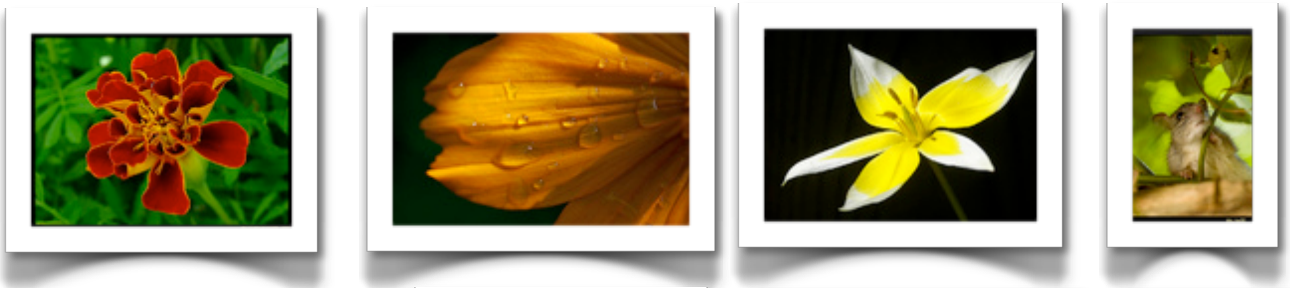
Images from Text

Given

Retrieved

water, red, sunset

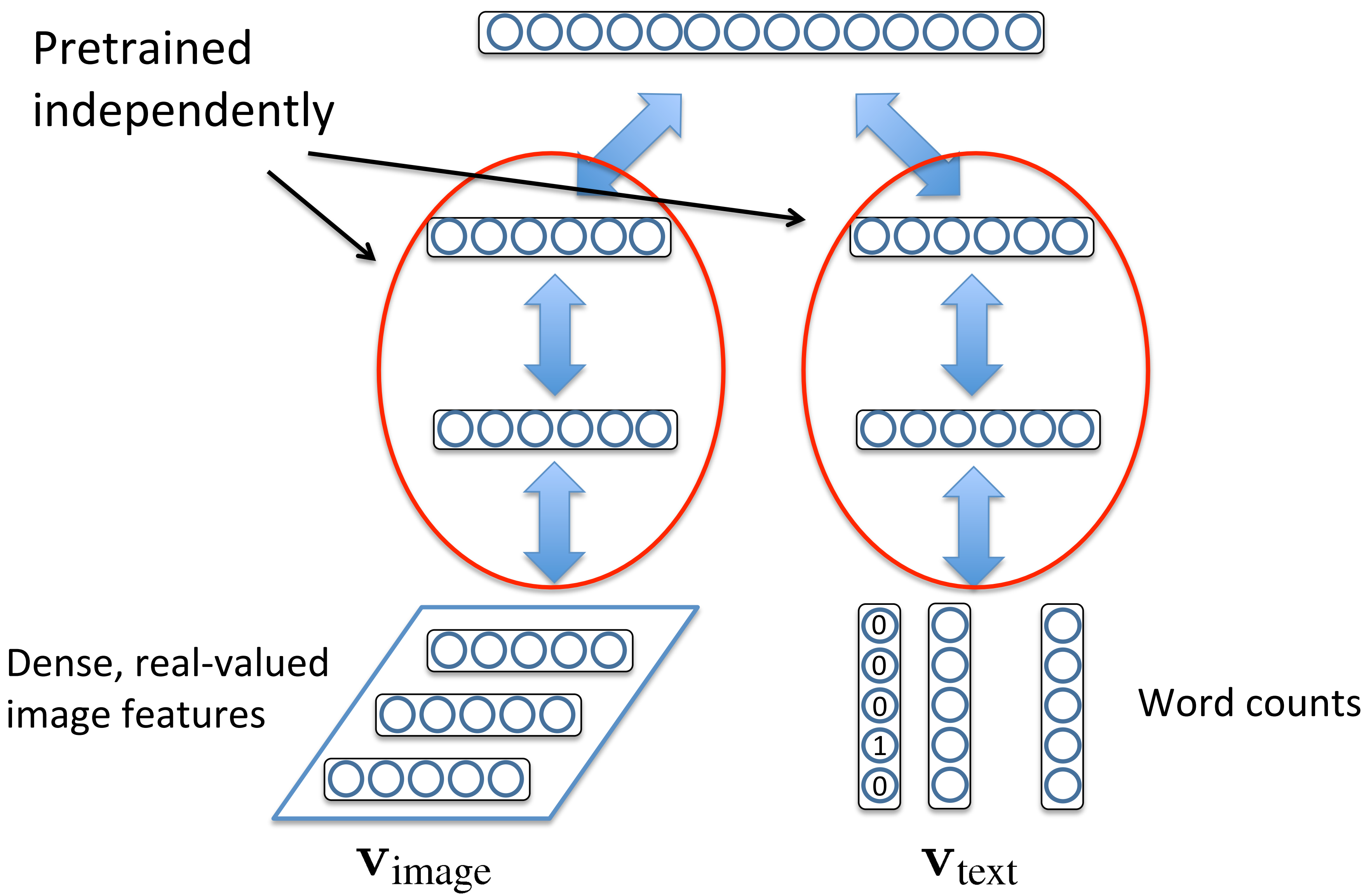
nature, flower, red, green



blue, green, yellow, colors

chocolate, cake

Pretraining



MIR-Flickr Dataset

- 1 million images along with user-assigned tags.

sculpture, beauty, stone

d80

nikon, abigfave, goldstaraward, d80, nikond80

food, cupcake, vegan

anawesomeshot, theperfectphotographer, flash, damniwishidtakenthat, spiritofphotography

nikon, green, light, photoshop, apple, d70

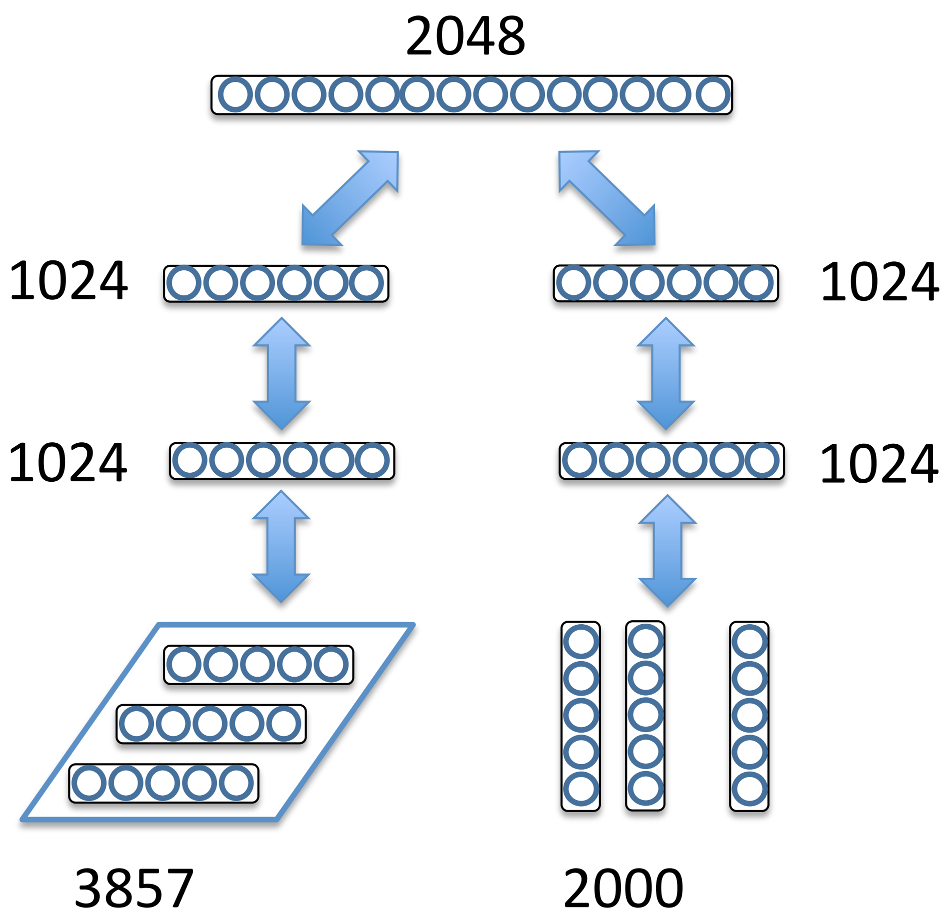
white, yellow, abstract, lines, bus, graphic

sky, geotagged, reflection, cielo, bilbao, reflejo

Huiskes et. al.

Data and Architecture

$\approx$ 12 Million parameters



- Image features: Gist, SIFT, MPEG-7 descriptors - 3857-dims.
- 200 most frequent tags.
- 25K labeled subset (15K training, 10K testing)
- 38 classes - *sky, tree, baby, car, cloud …*

Results

- Logistic regression on top-level representation.
- Multimodal Inputs

Mean Average Precision

Learning Algorithm	MAP	Precision@50
Random	0.124	0.124
LDA [Huiskes et. al.]	0.492	0.754
SVM [Huiskes et. al.]	0.475	0.758
DBM-Labelled	0.526	0.791

Same Features, 25K

Results

- Logistic regression on top-level representation.
- Multimodal Inputs

Mean Average Precision

Learning Algorithm	MAP	Precision@50
Random	0.124	0.124
LDA [Huiskes et. al.]	0.492	0.754
SVM [Huiskes et. al.]	0.475	0.758
DBM-Labelled	0.526	0.791
DBM-Unlablled	0.585	0.836

Similar Features, 25K

+ 1 Million unlabelled

Results

- Logistic regression on top-level representation.
- Multimodal Inputs

Mean Average Precision

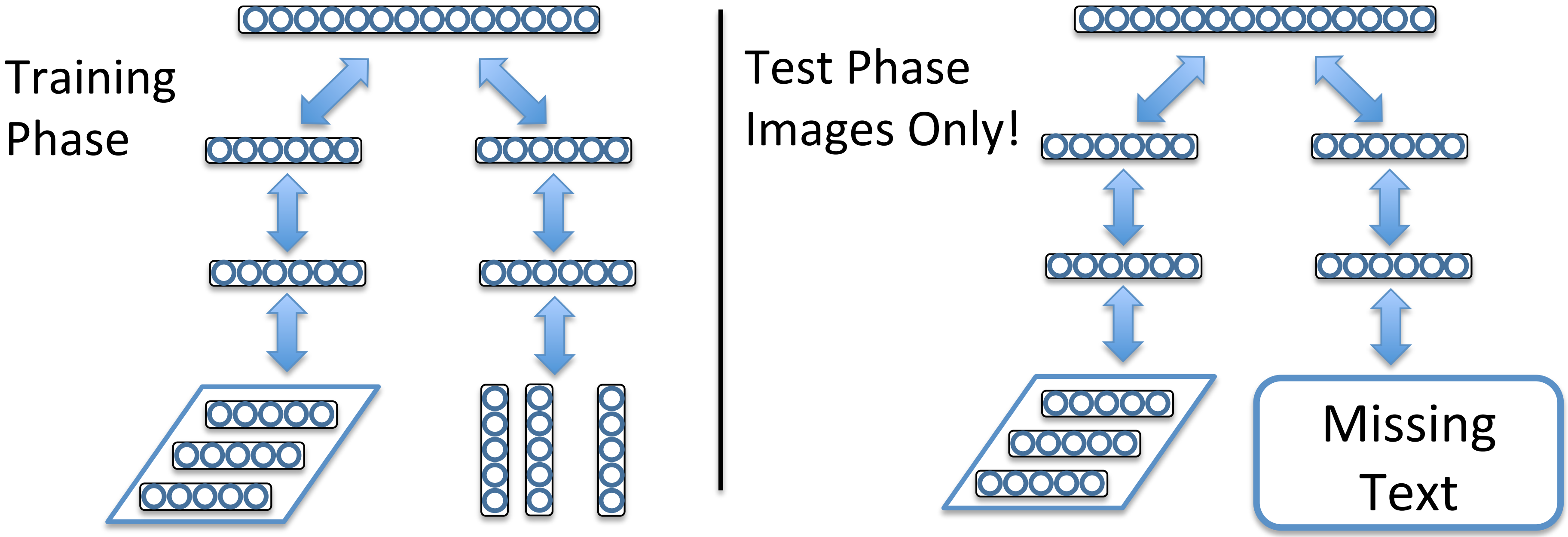
Learning Algorithm	MAP	Precision@50
Random	0.124	0.124
LDA [Huiskes et. al.]	0.492	0.754
SVM [Huiskes et. al.]	0.475	0.758
DBM-Labelled	0.526	0.791
DBM-Unlablled	0.585	0.836
Deep Belief Net	0.599	0.867
Autoencoder	0.600	0.875
DBM	0.609	0.873

Similar Features, 25K

+ 1 Million unlabelled

+ SIFT features

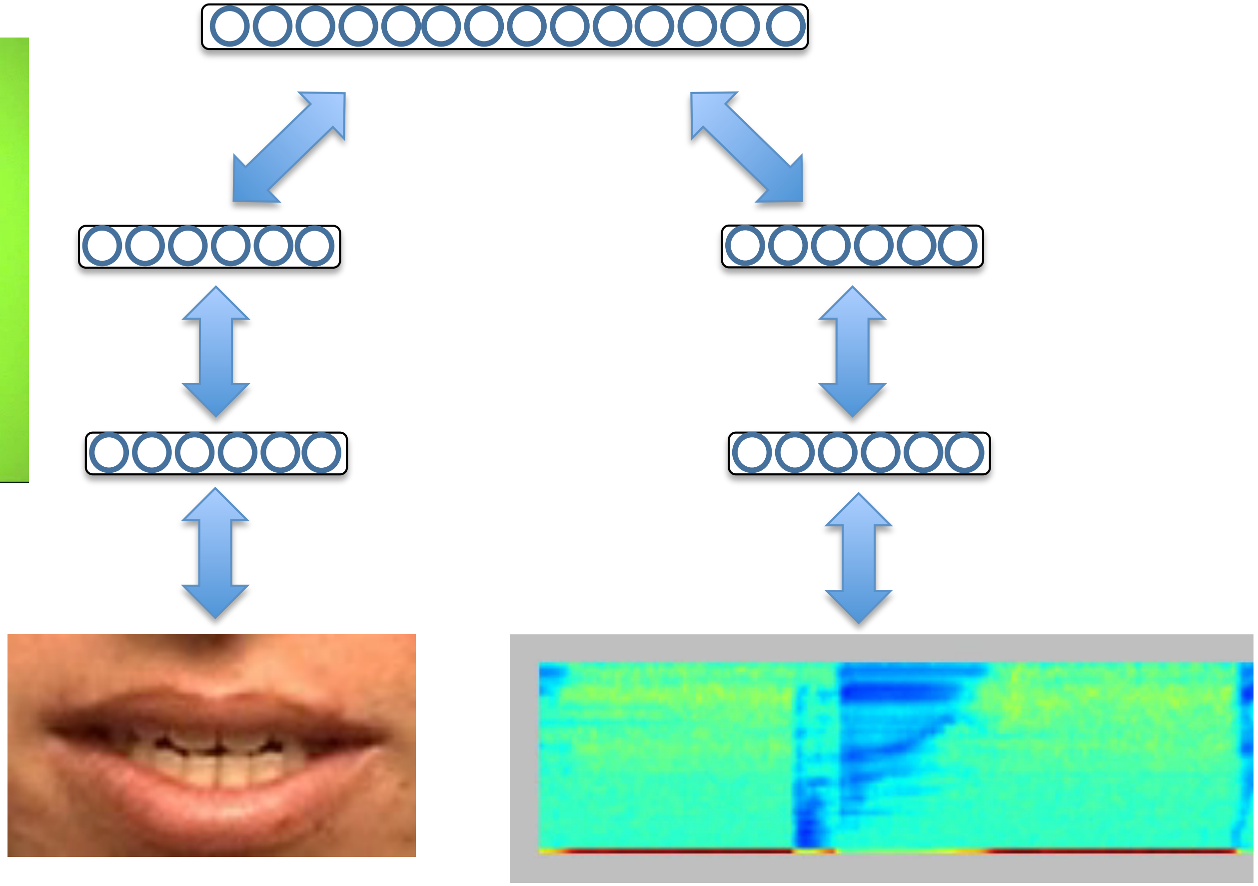
Results



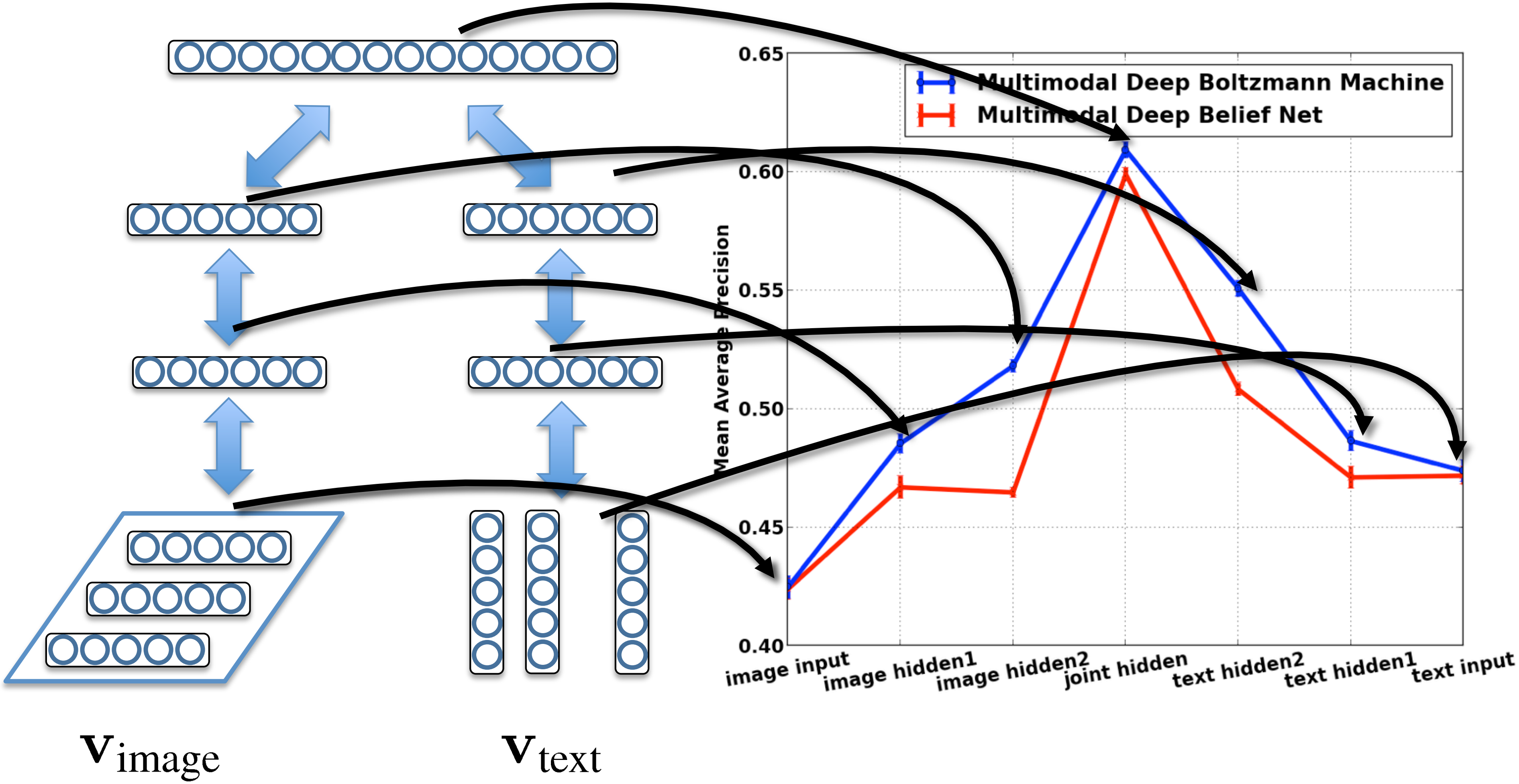
Learning Algorithm	MAP	Precision@50
Image-LDA [Huiskes et. al.]	0.315	-
Image-SVM [Huiskes et. al.]	0.375	-
Image-DBM	0.469	0.803
Multimodal-DBM (missing text)	0.531	0.832

Video and Audio

Cuave Dataset

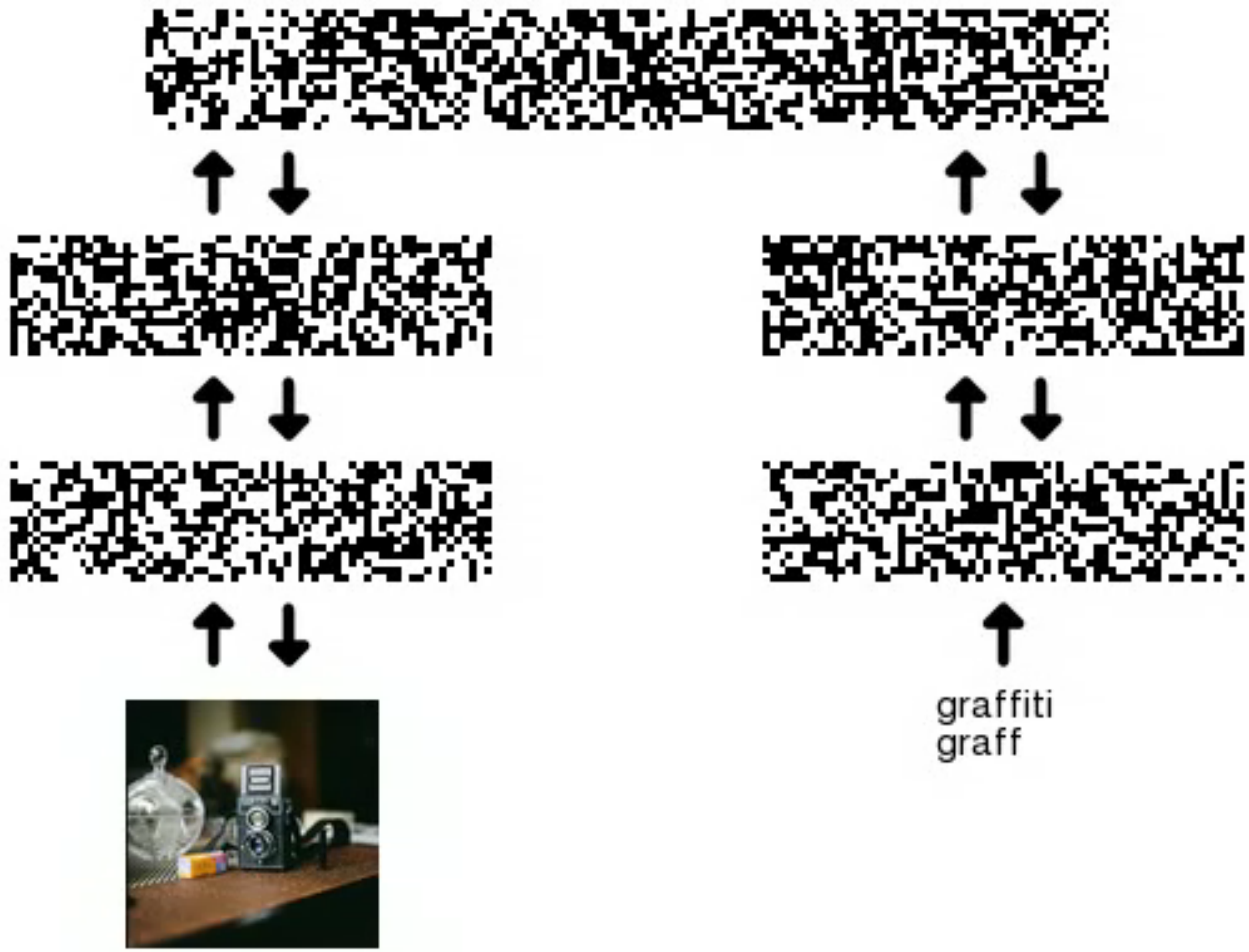


Classification Layer-wise



Images from Text

Step 0

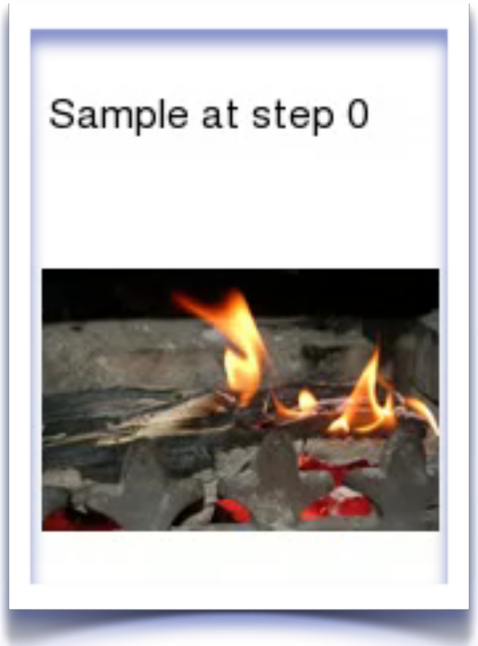


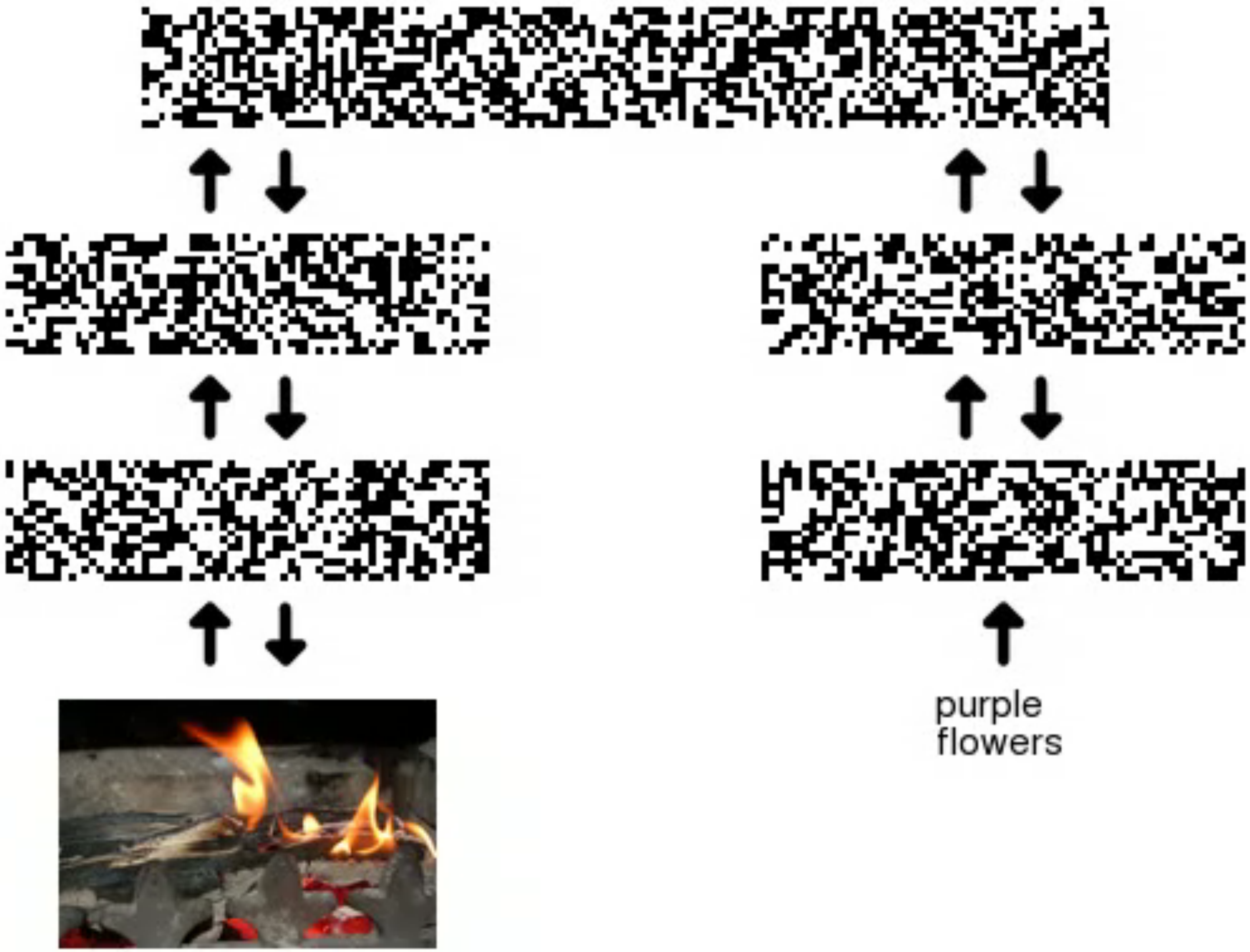
Sample drawn after every 50 steps of Gibbs sampling

More Videos

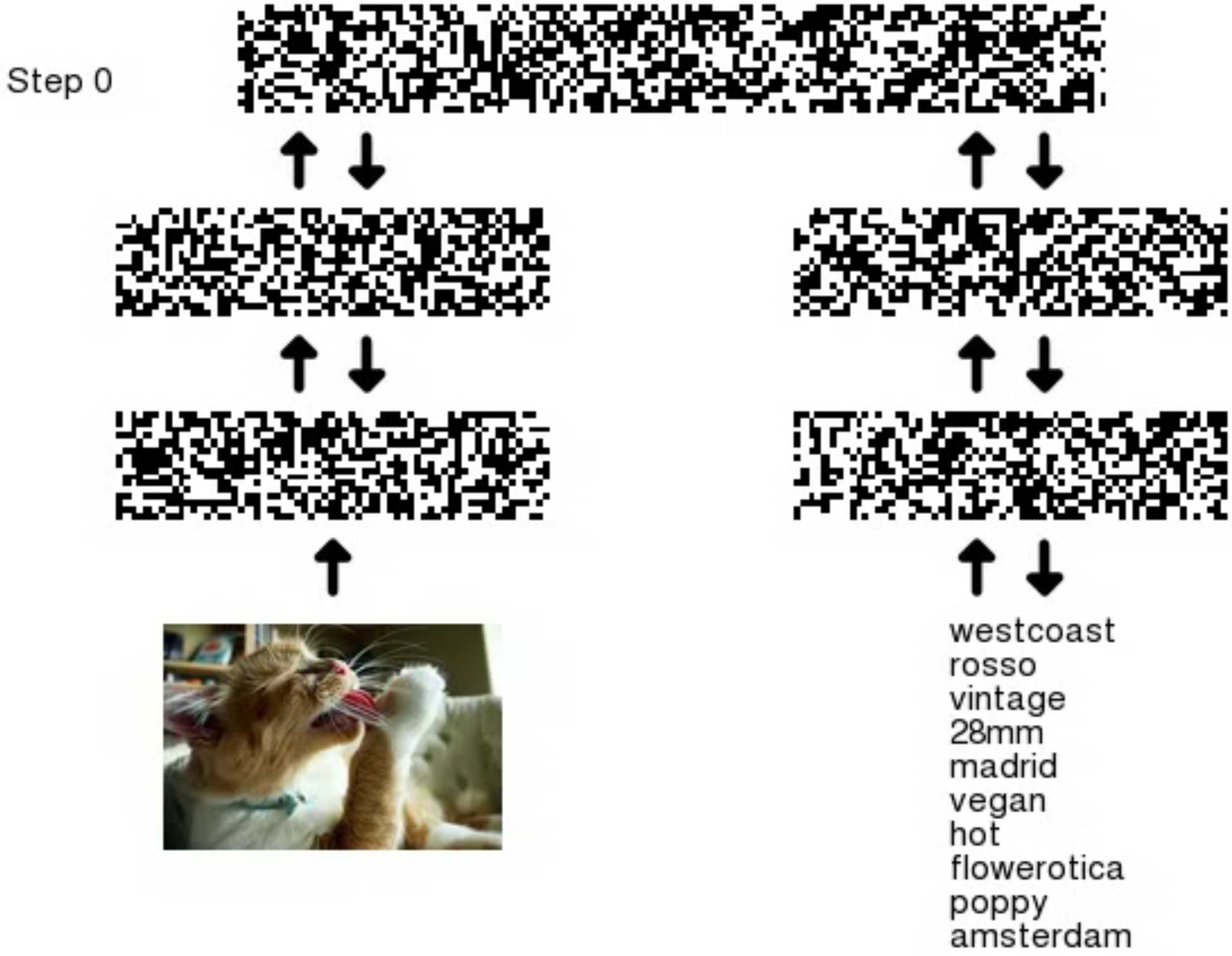
Step 0

Sample drawn after every 50 steps of Gibbs sampling





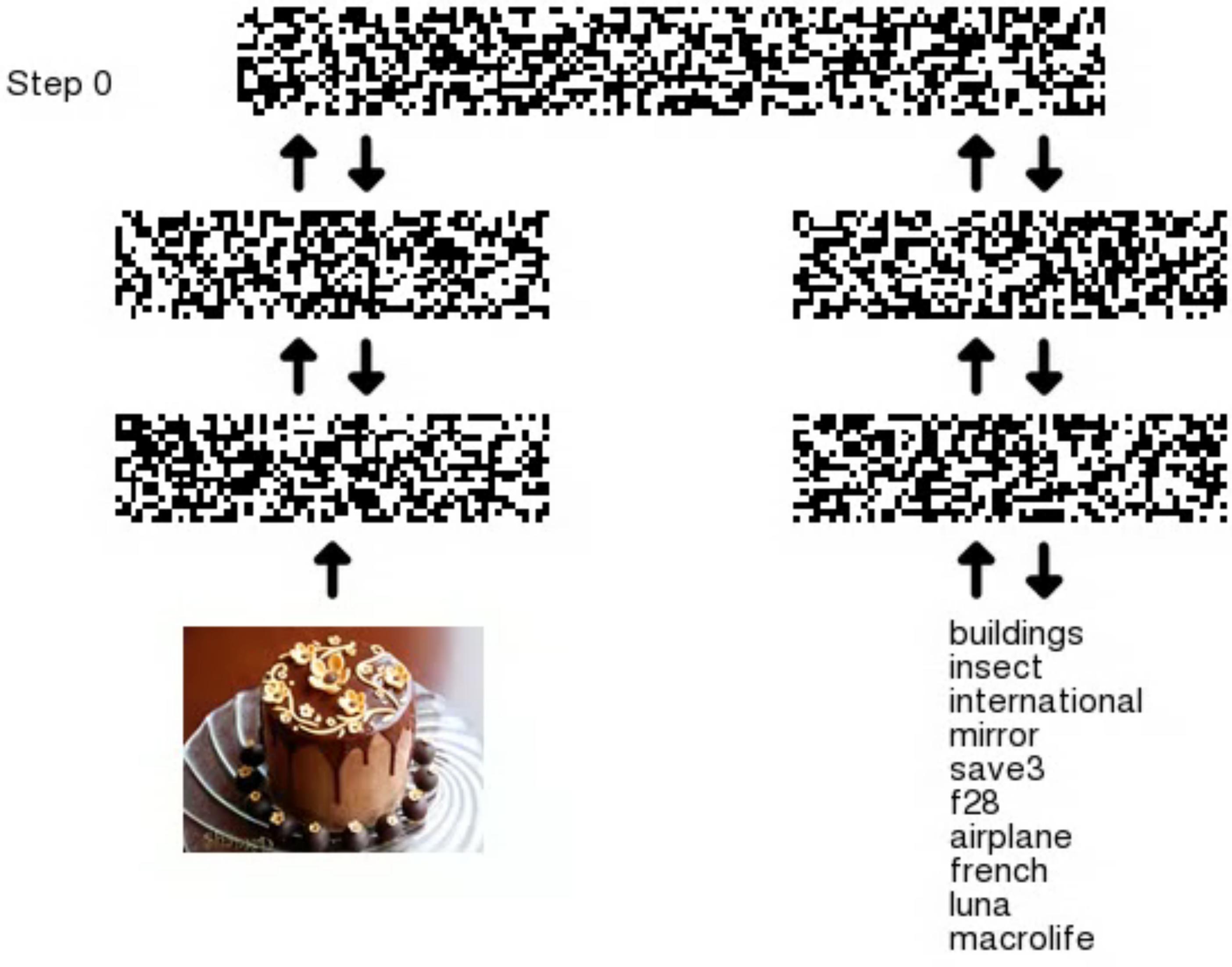
More Videos



Samples drawn after every 50 steps of Gibbs updates

Sample at step 0
westcoast
rosso
vintage
28mm
madrid
vegan
hot
flowerotica
poppy
amsterdam

More Videos



Samples drawn after every 50 steps of Gibbs updates

Sample at step 0
buildings
insect
international
mirror
save3
f28
airplane
french
luna
macrolife