Learning Structured, Robust, and Multimodal Deep Models

Russ Salakhutdinov

Department of Statistics and Computer Science
University of Toronto

VELUT ARBOR fiVO
UNIVERSITY OF TORONTO

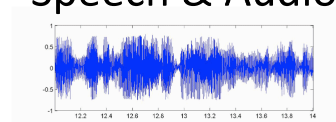
Mining for Structure

Massive increase in both computational power and the amount of data available from web, video cameras, laboratory measurements.

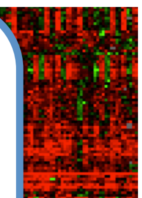
Images & Video

Text & Language

Speech & Audio



Gene Expression

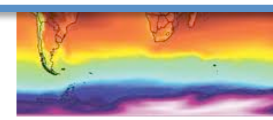


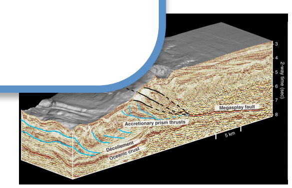
Prod

Reco

Hierarchical Generative Models that support inferences and discover structure at multiple levels.

al Data

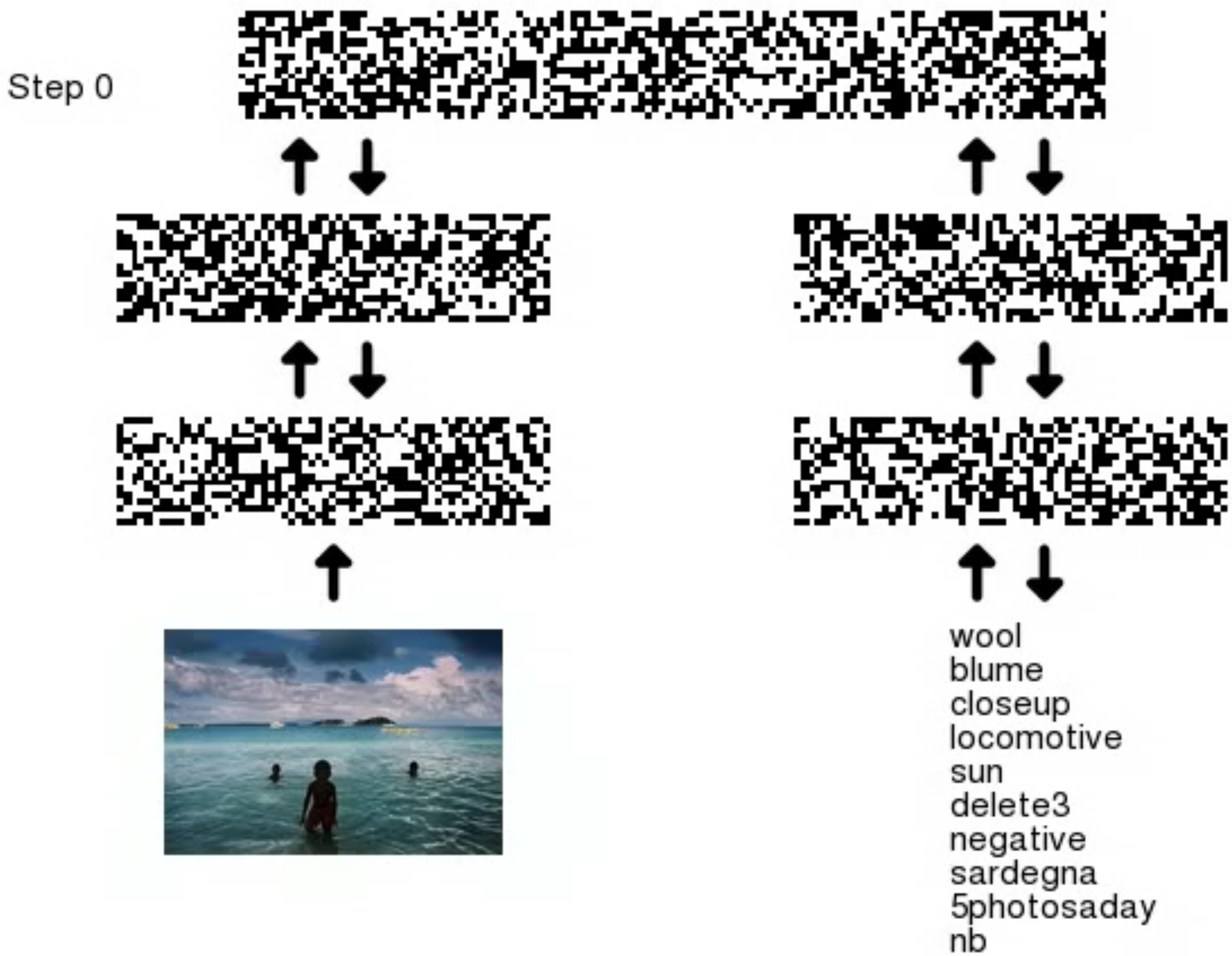




Mostly Unlabeled

- Develop statistical models that can discover underlying structure, cause, or statistical correlation from data in **unsupervised** or **semi-supervised** way.
- Multiple application domains.

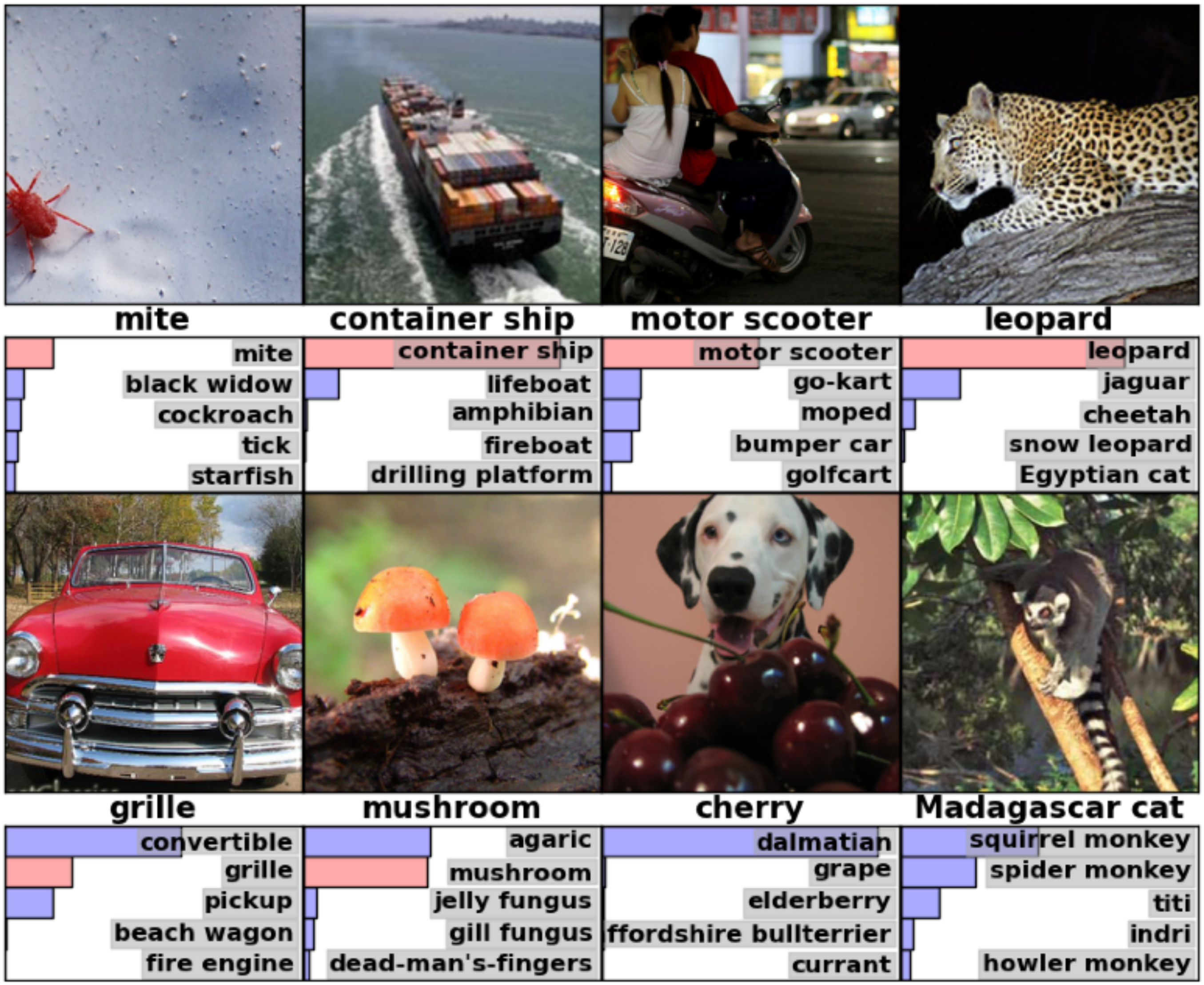
Generating Text from Images



Samples drawn after every 50 steps of Gibbs updates

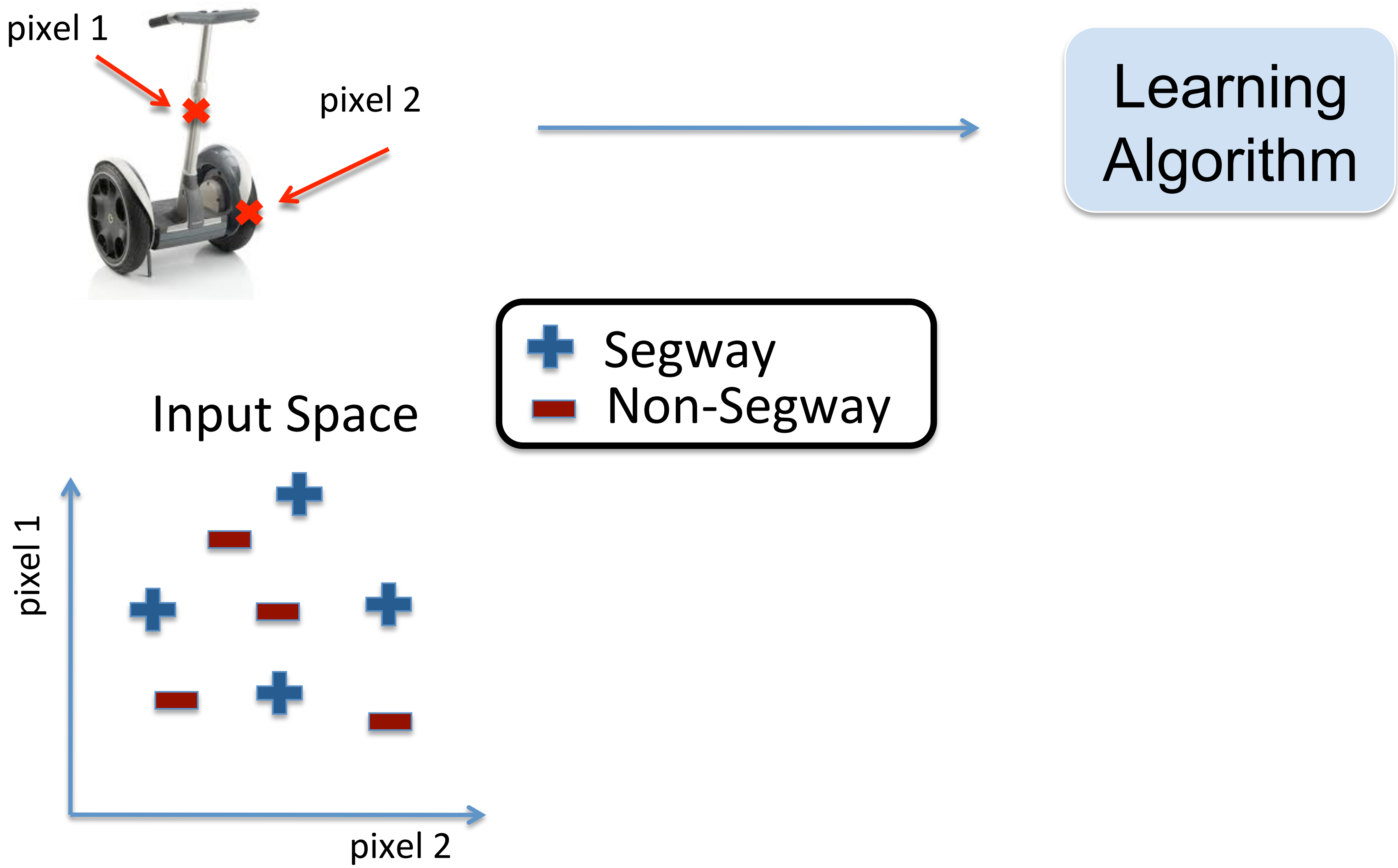
Sample at step 0
wool
blume
closeup
locomotive
sun
delete3
negative
sardegna
5photosaday
nb

Convolutinal Deep Models for Image Recognition

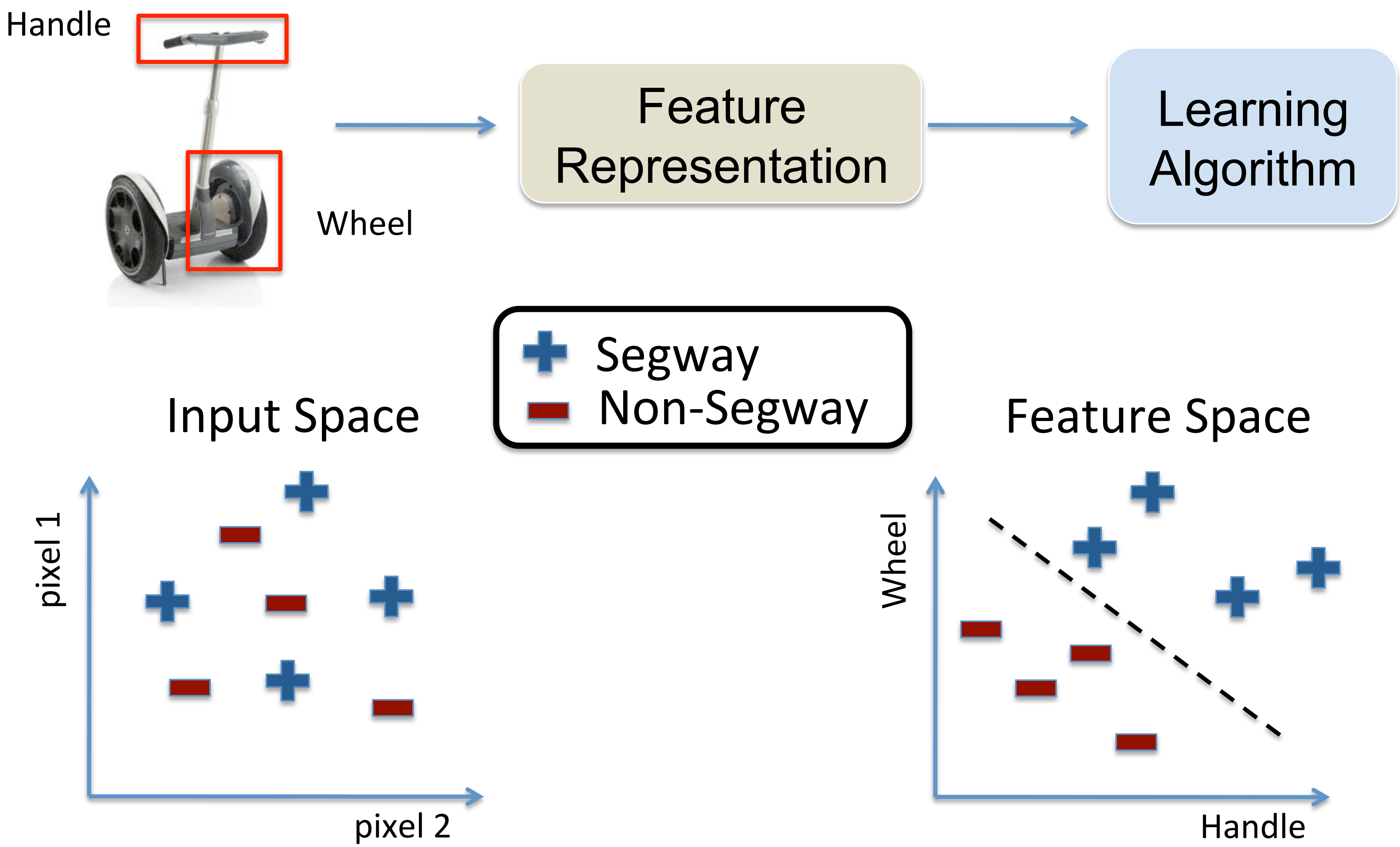


(Krizhevsky et. al., NIPS 2012)

Learning Feature Representations



Learning Feature Representations

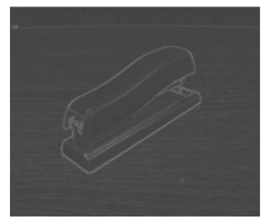


Computer Perception

Input Data → Low-level features → Learning Algorithm

Object detection

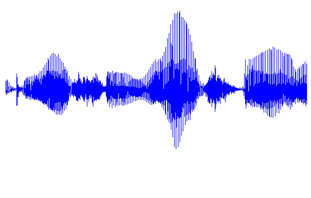
Image



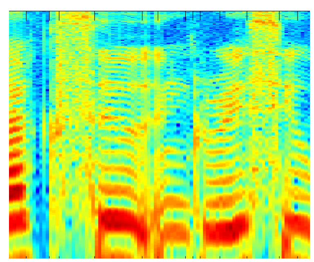
Low-level vision features

Recognition

Audio classification



Audio

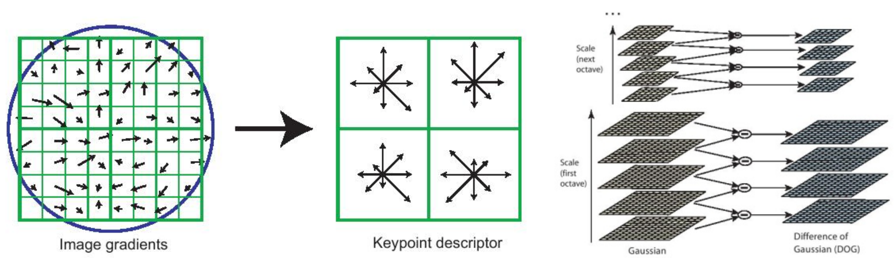


Low-level audio features

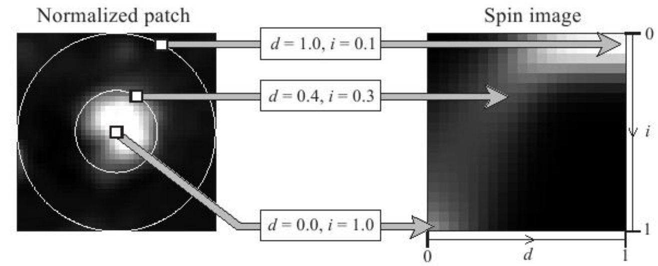
Speaker identification

Slide Credit: Honglak Lee

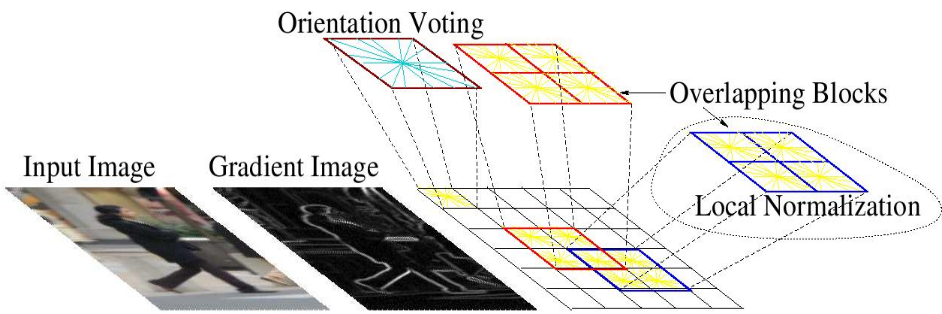
Computer Vision Features



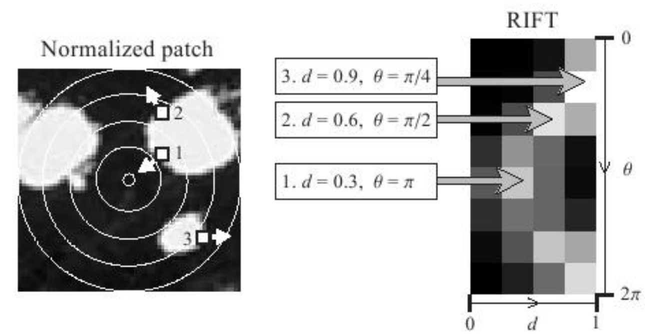
SIFT



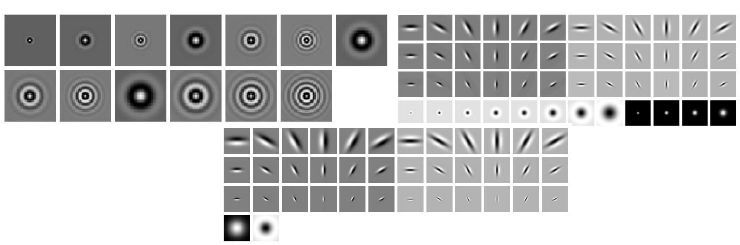
Spin image



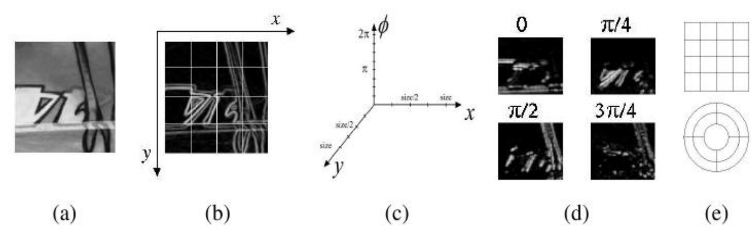
HoG



RIFT



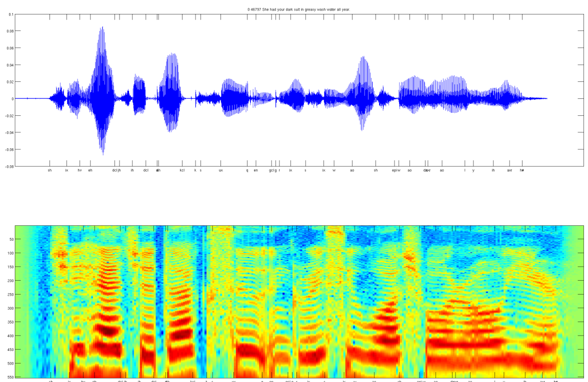
Textons



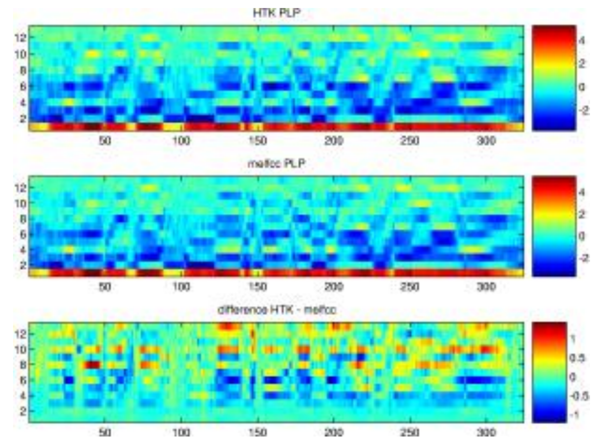
GLOH

Slide Credit: Honglak Lee

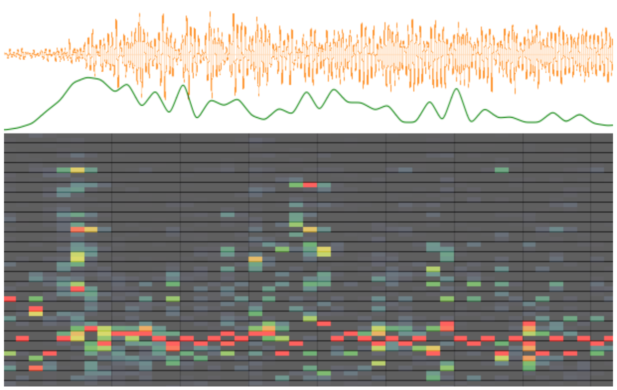
Audio Features



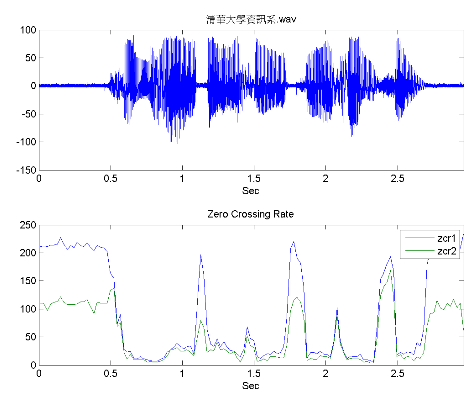
Spectrogram



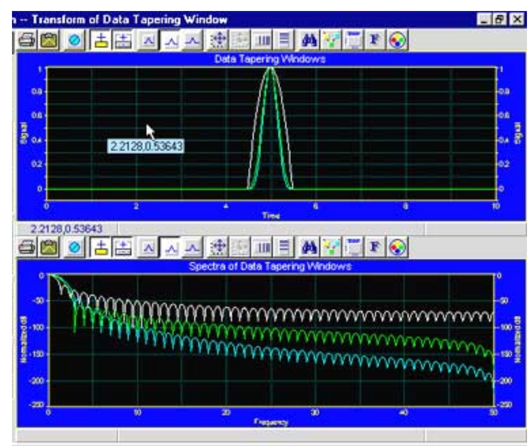
MFCC



Flux

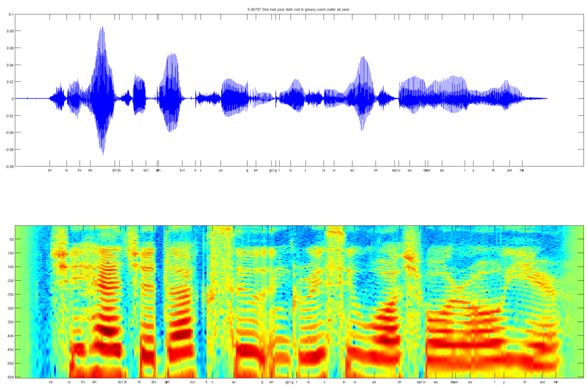


ZCR

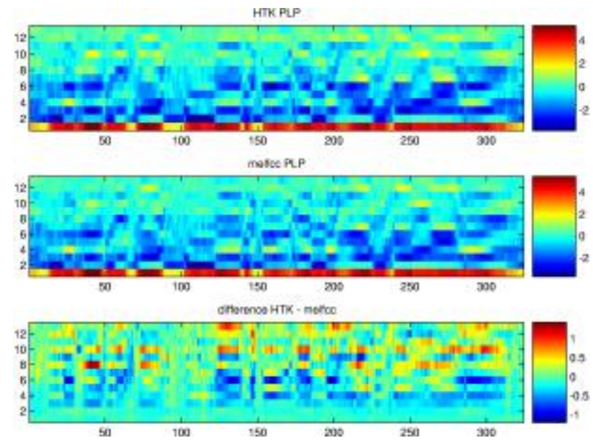


Rolloff

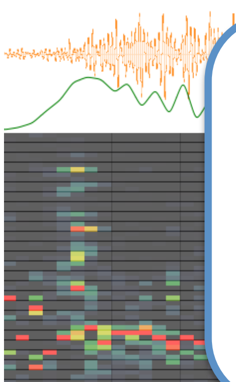
Audio Features

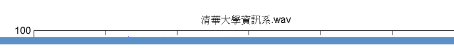


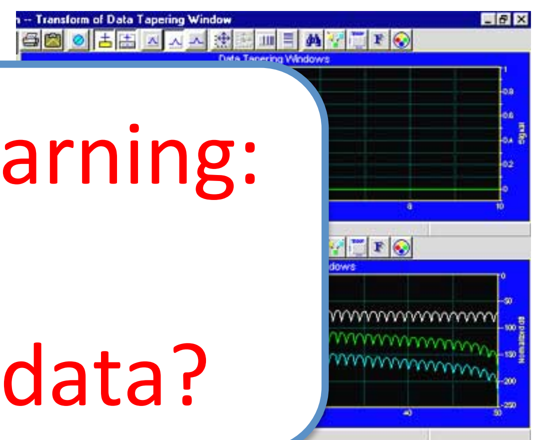
Spectrogram



MFCC







Unsupervised Feature Learning:
Can we learn meaningful features from unlabeled data?

Flux

ZCR

Rolloff

Talk Roadmap

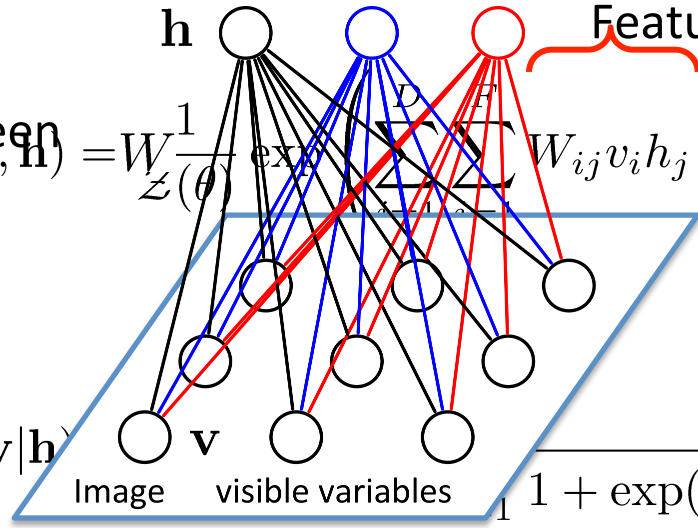
- Learning Deep Models
  - Restricted Boltzmann Machines
  - Deep Boltzmann Machines

- Learning Structured and Robust Models

- Multi-Modal Learning

Restricted Boltzmann Machines

Graphical Models: Powerful framework for representing dependency structure between random variables.



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

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

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

RBM is a Markov Random Field with:

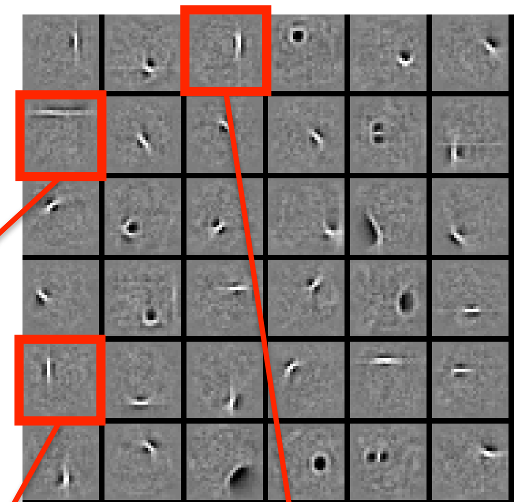
- Stochastic binary visible variables $\mathbf{v} \in \{0,1\}^D$.
- Stochastic binary hidden variables $\mathbf{h} \in \{0,1\}^F$.
- Bipartite connections.

Markov random fields, Boltzmann machines, log-linear models.

Learning Features

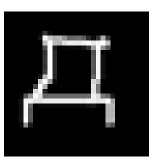
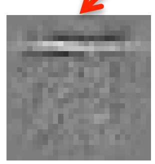
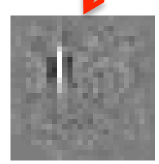
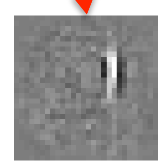
Observed Data
Subset of 25,000 characters

Learned W: "edges"
Subset of 1000 features



Sparse representations

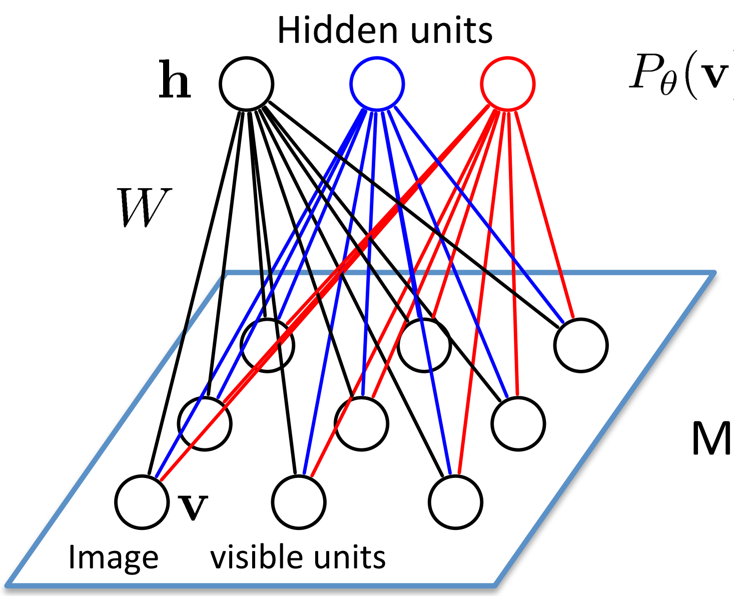
New Image: $p(h_7 = 1|v)$ $p(h_{29} = 1|v)$

 = $\sigma\Big($ $0.99 \times$  $+ \; 0.97 \times$  $+ \; 0.82 \times$  $\ldots \Big)$

$\sigma(x) = \frac{1}{1+\exp(-x)}$

Logistic Function: Suitable for modeling binary images

Model Learning



$$P_\theta(\mathbf{v}) = \frac{P^*(\mathbf{v})}{\mathcal{Z}(\theta)} = \frac{1}{\mathcal{Z}(\theta)} \sum_{\mathbf{h}} \exp\left[\mathbf{v}^\top W \mathbf{h} + \mathbf{a}^\top \mathbf{h} + \mathbf{b}^\top \mathbf{v}\right]$$

Given a set of *i.i.d.* training examples $\mathcal{D} = \{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, ..., \mathbf{v}^{(N)}\}$, we want to learn model parameters $\theta = \{W, a, b\}$.

Maximize log-likelihood objective:

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log P_\theta(\mathbf{v}^{(n)})$$

Derivative of the log-likelihood:

$$\frac{\partial L(\theta)}{\partial W_{ij}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial W_{ij}} \log\left(\sum_{\mathbf{h}} \exp\left[\mathbf{v}^{(n)\top} W \mathbf{h} + \mathbf{a}^\top \mathbf{h} + \mathbf{b}^\top \mathbf{v}^{(n)}\right]\right) - \frac{\partial}{\partial W_{ij}} \log \mathcal{Z}(\theta)$$

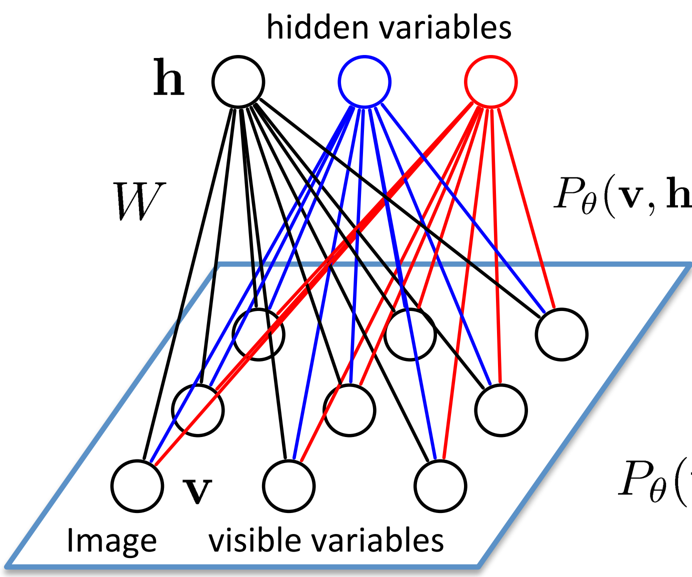
$$= \mathrm{E}_{P_{data}}[v_i h_j] - \underbrace{\mathrm{E}_{P_\theta}[v_i h_j]}$$

Difficult to compute: exponentially many configurations

$$P_{data}(\mathbf{v}, \mathbf{h}; \theta) = P(\mathbf{h}|\mathbf{v}; \theta) P_{data}(\mathbf{v})$$

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

RBMs for Real-valued Data



Pair-wise **Unary**

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

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

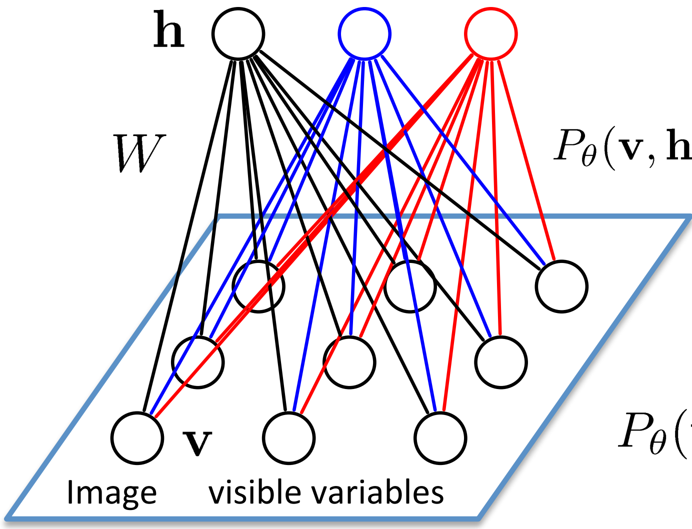
$$P_\theta(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{D} P_\theta(v_i|\mathbf{h}) = \prod_{i=1}^{D} \mathcal{N}\left(b_i + \sum_{j=1}^{F} W_{ij} h_j, \sigma_i^2 \right)$$

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.

(Salakhutdinov & Hinton, NIPS 2007; Salakhutdinov & Murray, ICML 2008)

RBMs for Real-valued Data



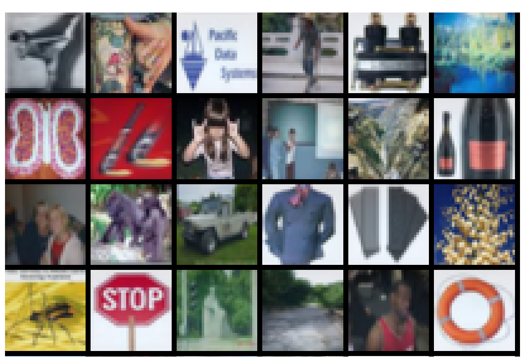
Pair-wise **Unary**

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

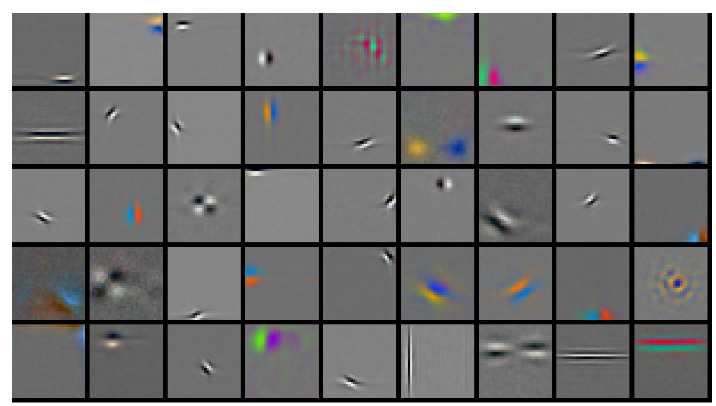
$$\theta = \{W, a, b\}$$

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

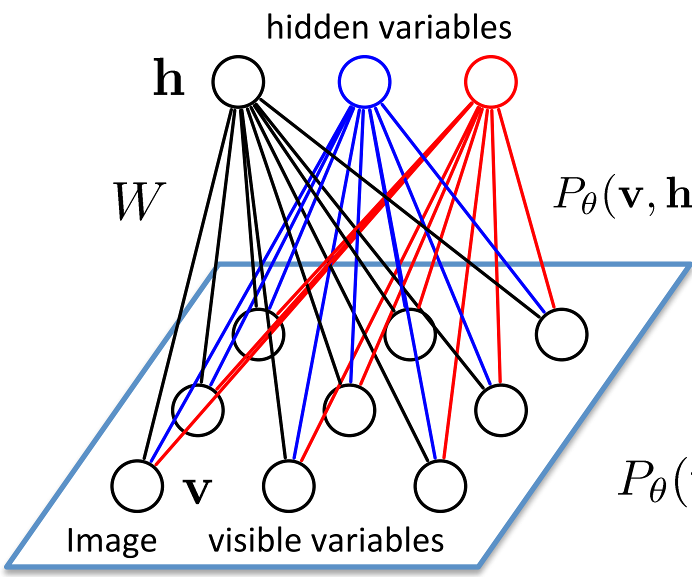
4 million **unlabelled** images



Learned features (out of 10,000)



RBMs for Real-valued Data



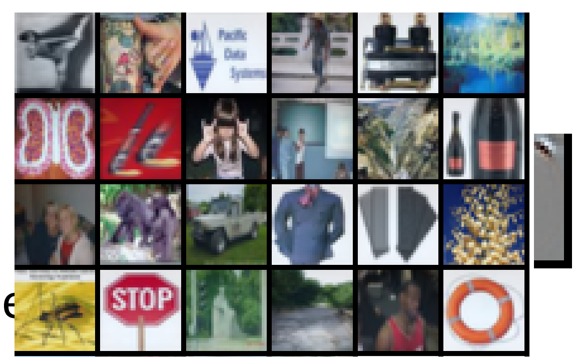
Pair-wise **Unary**

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

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

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

4 million **unlabelled** images

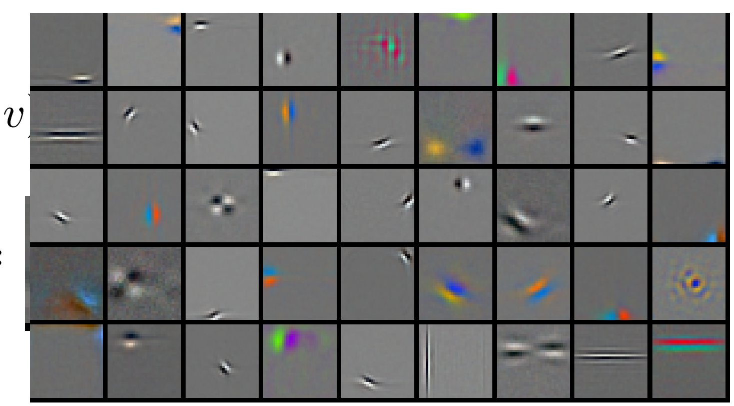


Ne

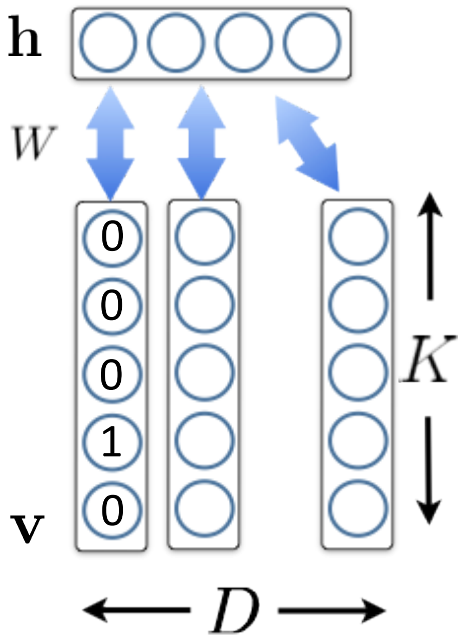
$p(h_{29} = 1|v$

$+ \ 0.8 \ *$

Learned features (out of 10,000)



RBMs for Word Counts



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

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

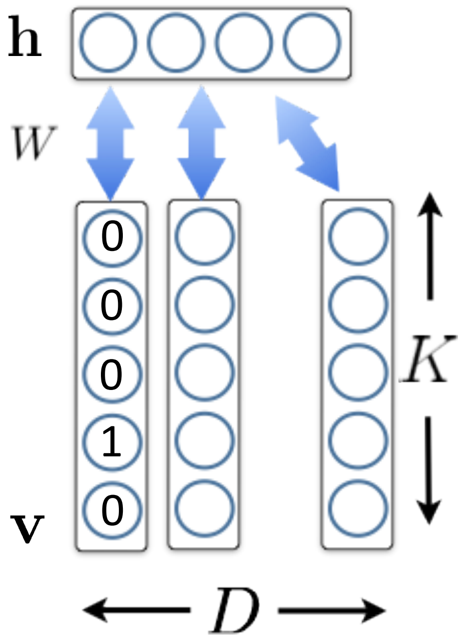
$$P_\theta(v_i^k = 1|\mathbf{h}) = \frac{\exp\left(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k\right)}{\sum_{q=1}^{K} \exp\left(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q\right)}$$

Replicated Softmax Model: undirected topic model:

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

(Salakhutdinov & Hinton, NIPS 2010, Srivastava & Salakhutdinov, NIPS 2012)

RBMs for Word Counts



Pair-wise **Unary**

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

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

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

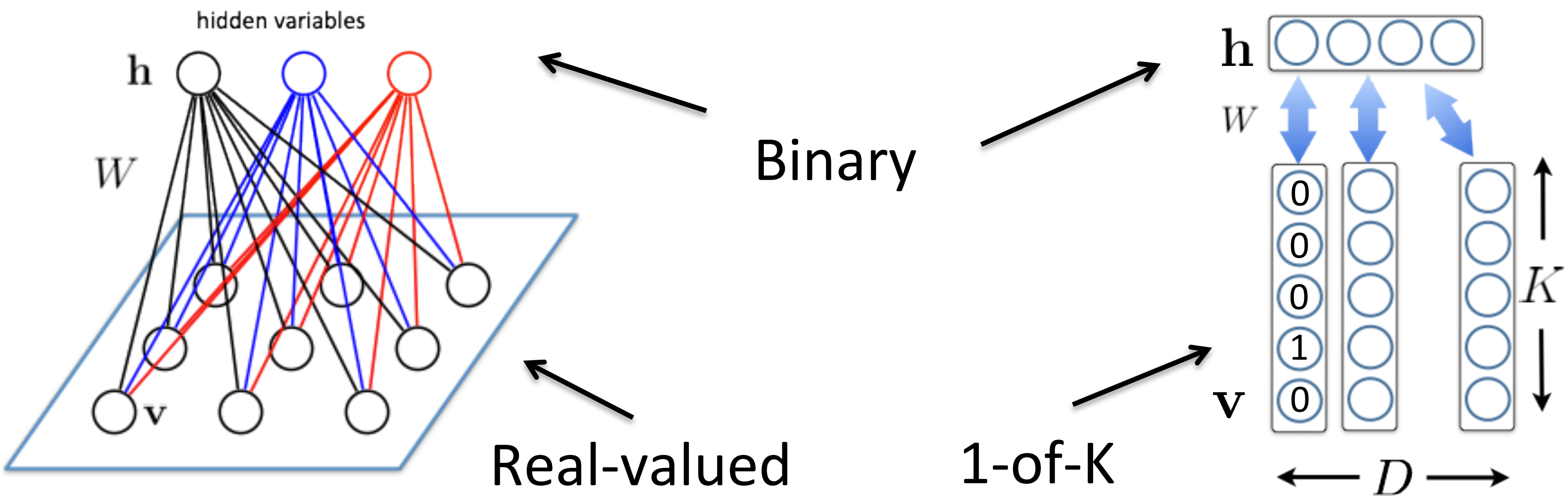
Reuters dataset:
804,414 **unlabeled**
newswire stories
Bag-of-Words

Learned features: ``topics''

russian	clinton	computer	trade	stock
russia	house	system	country	wall
moscow	president	product	import	street
yeltsin	bill	software	world	point
soviet	congress	develop	economy	dow

Different Data Modalities

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



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

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

Product of Experts

The joint distribution is given by:

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big(\sum_{ij} W_{ij} v_i h_j + \sum_i b_i v_i + \sum_j a_j h_j\Big)$$

Marginalizing over hidden variables:

Product of Experts

$$P_\theta(\mathbf{v}) = \sum_{\mathbf{h}} P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \prod_i \exp(b_i v_i) \prod_j \Big(1 + \exp(a_j + \sum_i W_{ij} v_i)\Big)$$

government	clinton	bribery	oil	stock	...
auhority	house	corruption	barrel	wall	
power	president	dishonesty	exxon	street	
empire	bill	putin	putin	point	
putin	congress	fraud	drill	dow	

Putin

Topics "government", "corruption" and "oil" can combine to give very high probability to a word "Putin".

(Salakhutdinov & Hinton, NIPS 2010, Srivastava & Salakhutdinov, NIPS 2012)

Product of Experts

The joint distribution is given by:

$$P_\theta(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big(\sum_{ij} W_{ij} v_i h_j + \sum_i b_i v_i + \sum_j a_j h_j\Big)$$

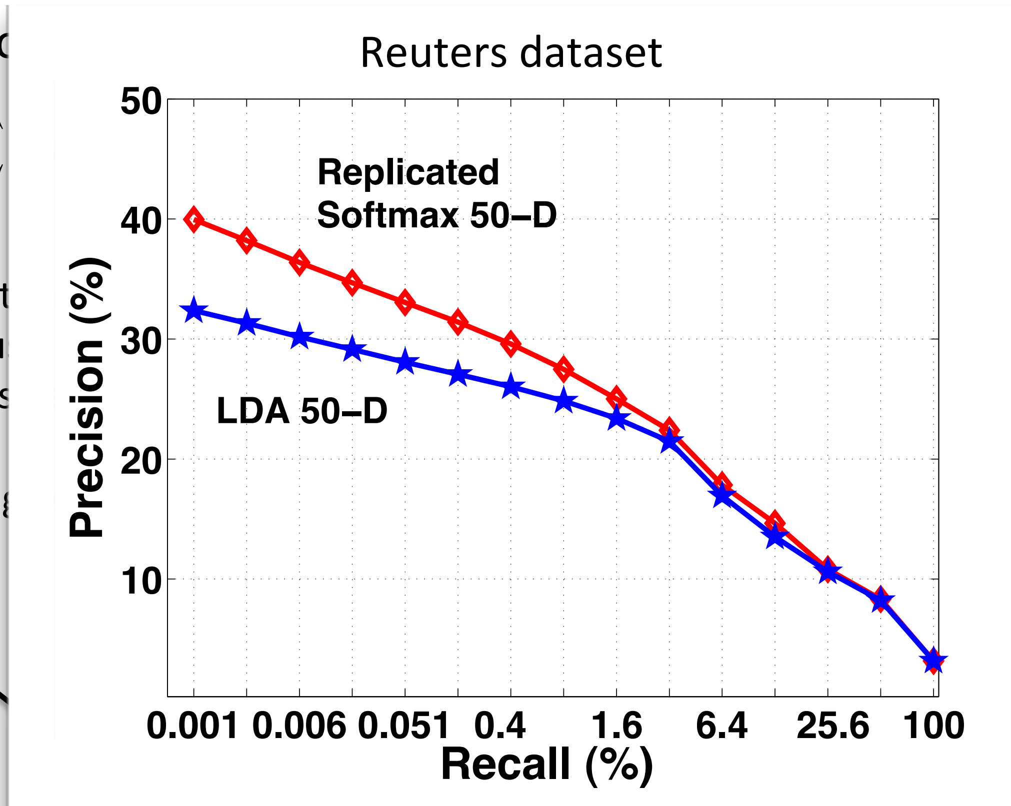
Marginalizing c

$$P_\theta(\mathbf{v}) = \sum_{\mathbf{h}}$$

duct of Experts

$W_{ij} v_i)\Big)$

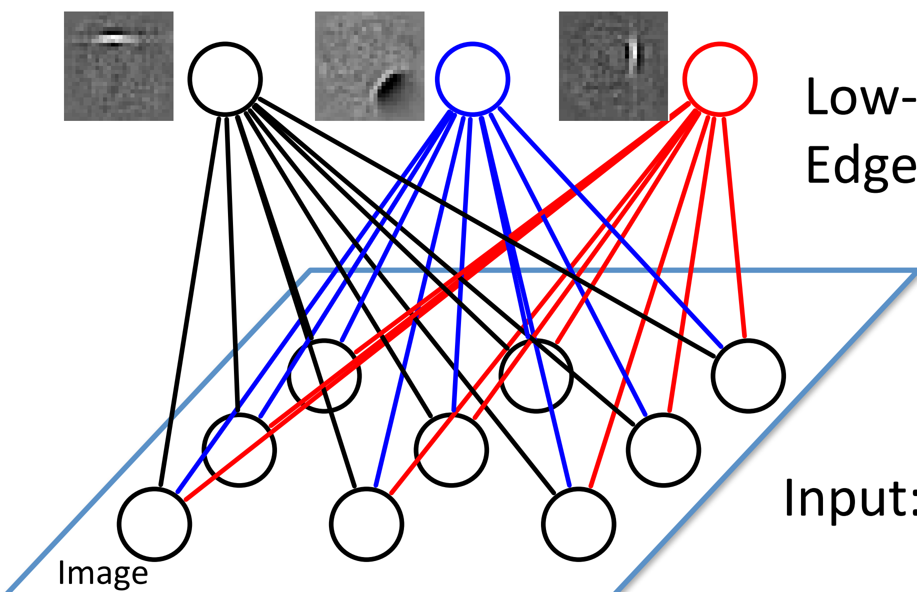
government	clint
auhority	hou
power	pres
empire	bill
putin	con



tations allow the
, "corruption" and
ive very high
probability to a word "Putin".

(Salakhutdinov & Hinton, NIPS 2010, Srivastava & Salakhutdinov, NIPS 2012)

Deep Boltzmann Machines



Low-level features: Edges

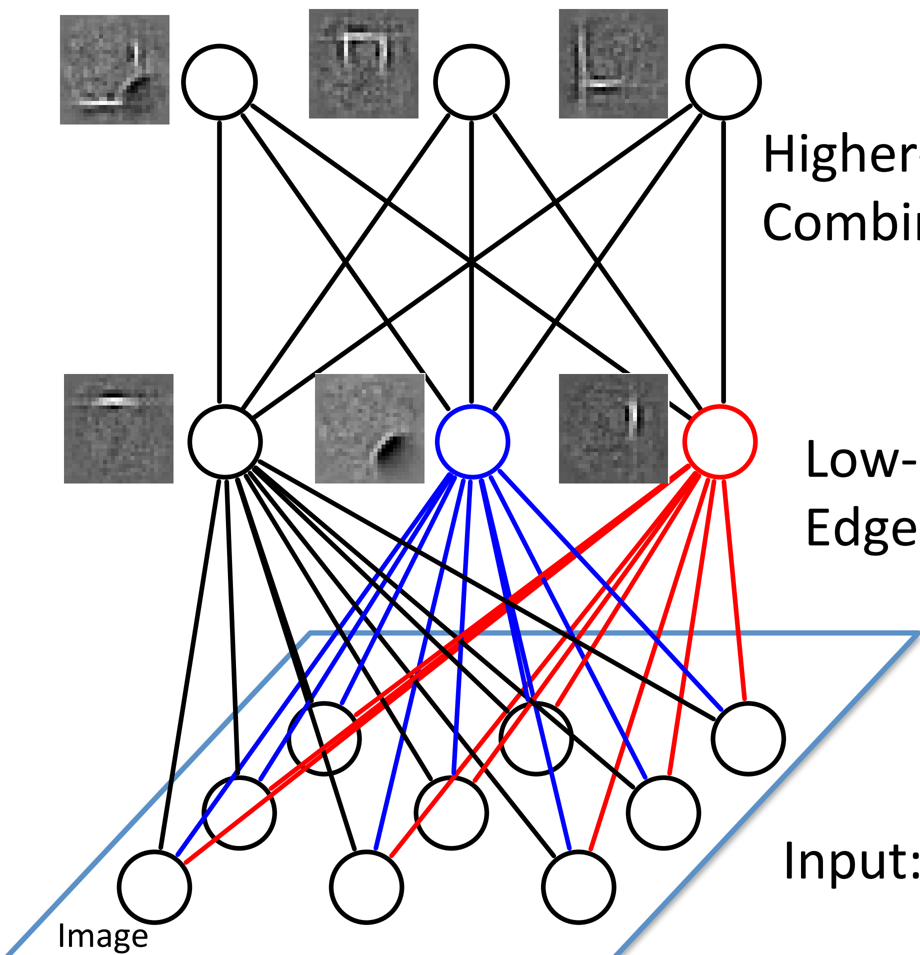
Built from **unlabeled** inputs.

Input: Pixels

(Salakhutdinov & Hinton, Neural Computation 2012)

Deep Boltzmann Machines

Learn simpler representations, then compose more complex ones

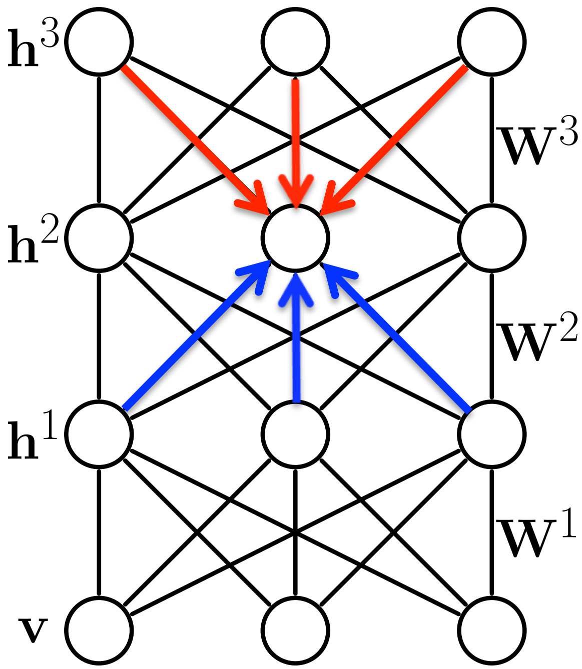


Built from **unlabeled** inputs.

(Salakhutdinov & Hinton, Neural Computation 2012)

Model Formulation

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



Same as RBMs

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

- Dependencies between hidden variables.
- All connections are undirected.
- Bottom-up and Top-down:

$$P(h_j^2 = 1 | \mathbf{h}^1, \mathbf{h}^3) = \sigma\left(\sum_k W_{kj}^3 h_k^3 + \sum_m W_{mj}^2 h_m^1\right)$$

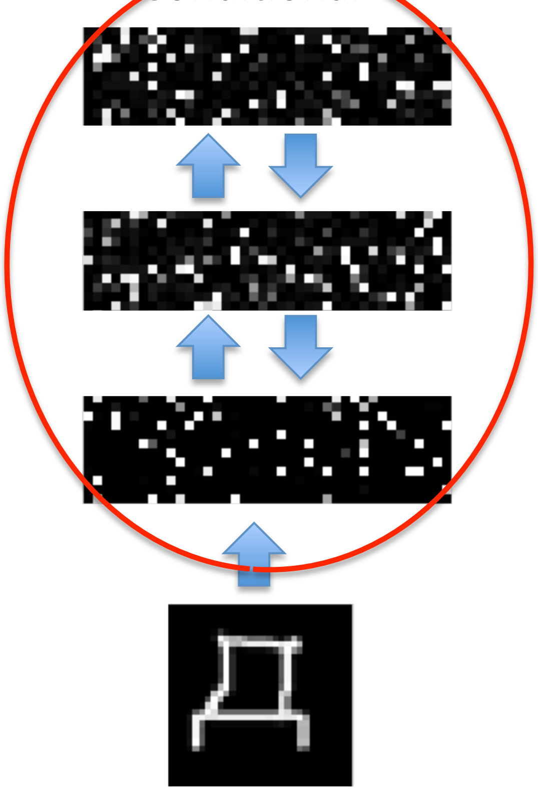
Top-down — Bottom-up

- Hidden variables are dependent even when **conditioned on the input**.

New Learning Algorithm

Posterior Inference

Conditional



Approximate conditional

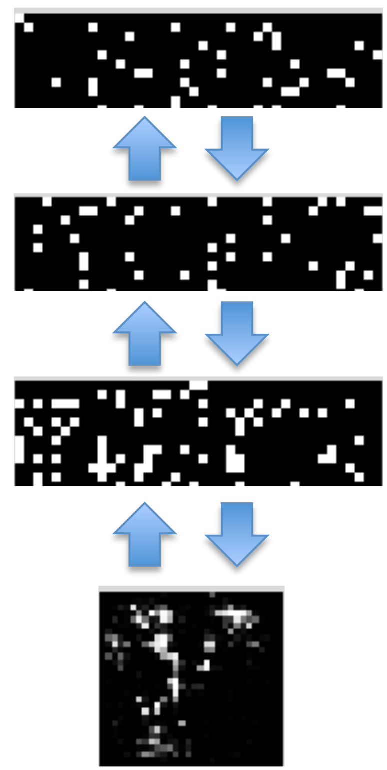
$$P_{data}(\mathbf{h}|\mathbf{v})$$

Simulate from the Model

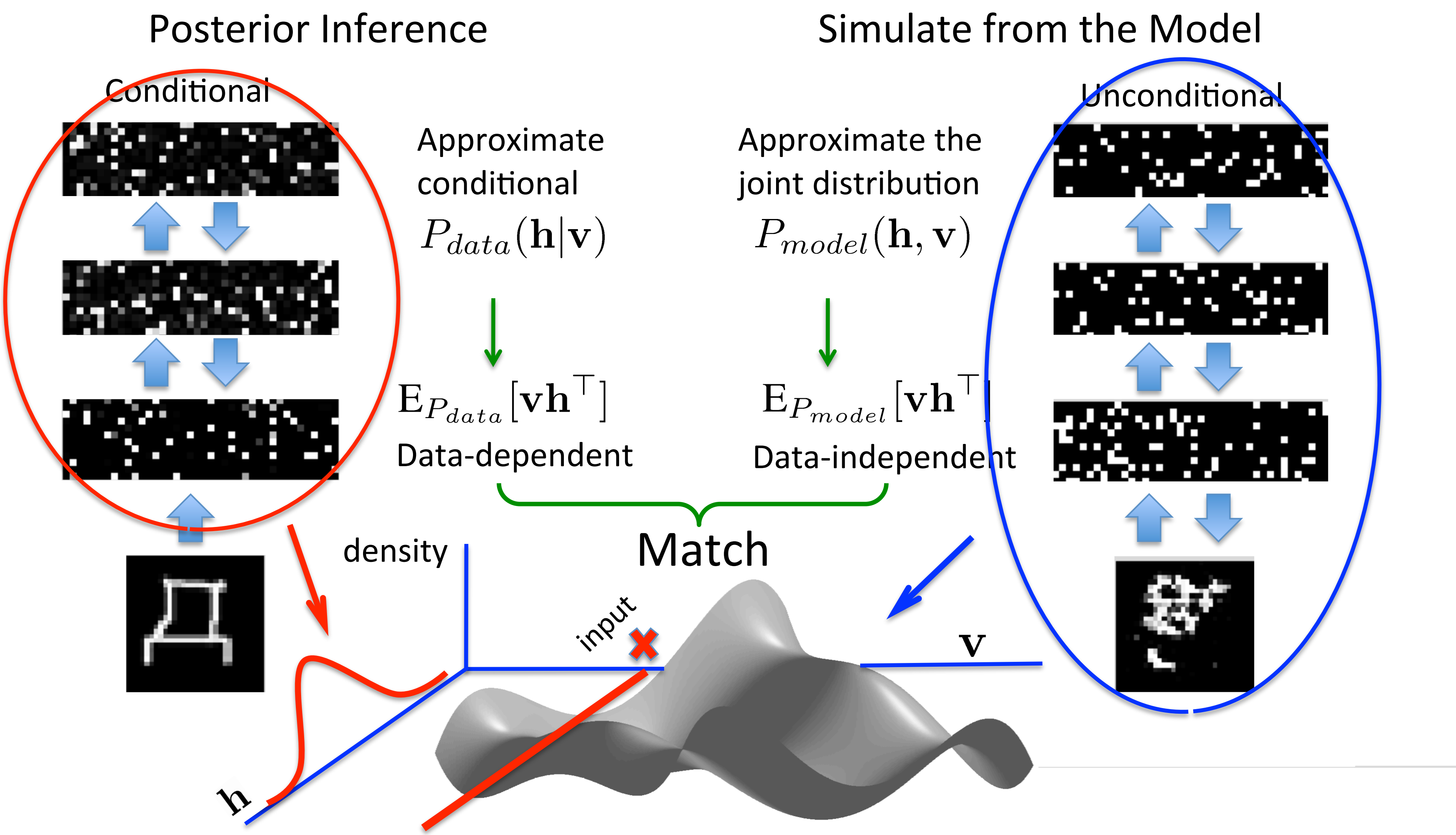
Approximate the joint distribution

$$P_{model}(\mathbf{h}, \mathbf{v})$$

Unconditional



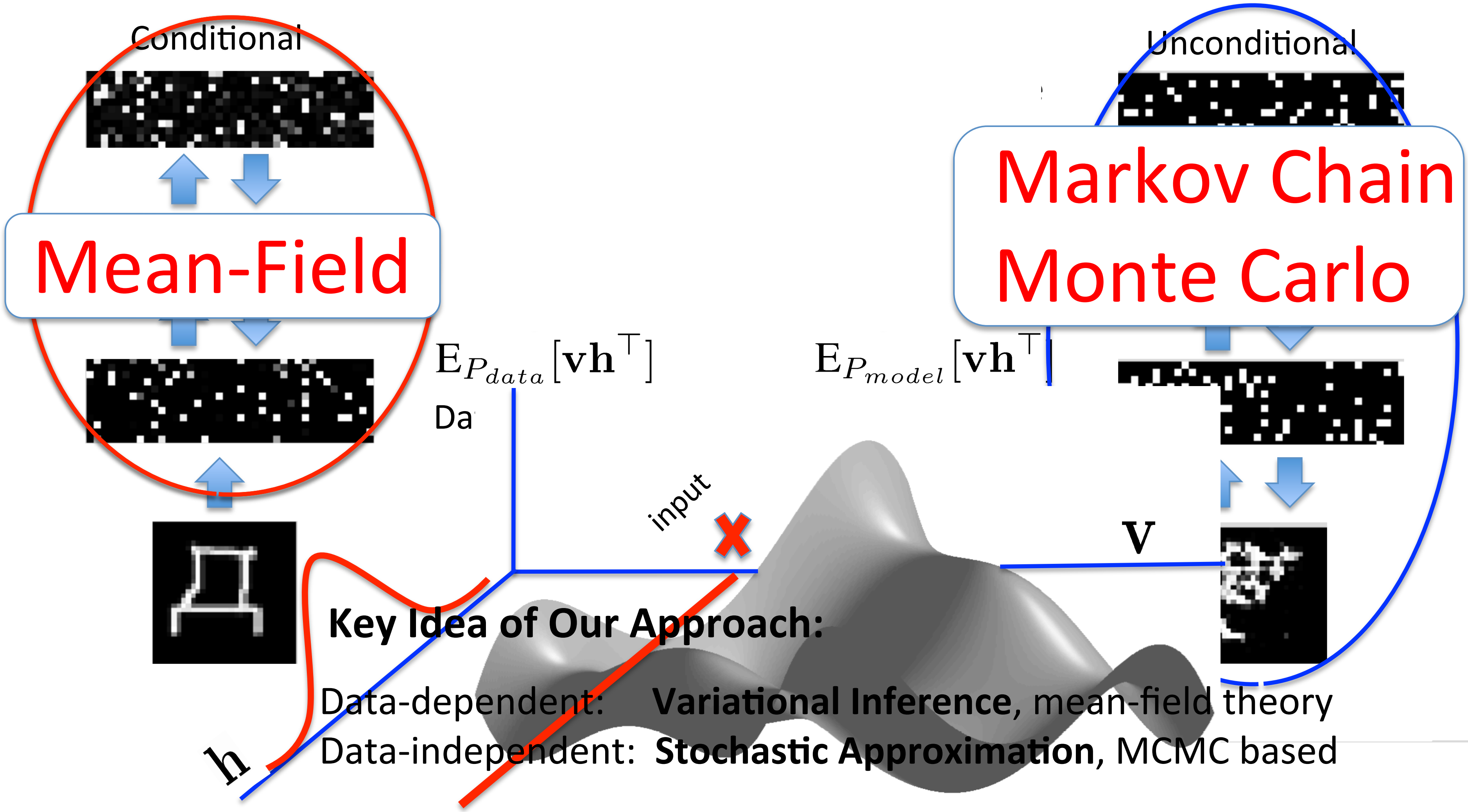
New Learning Algorithm



New Learning Algorithm

Posterior Inference

Simulate from the Model



Key Idea of Our Approach:

Data-dependent: **Variational Inference**, mean-field theory

Data-independent: **Stochastic Approximation**, MCMC based

Variational Inference

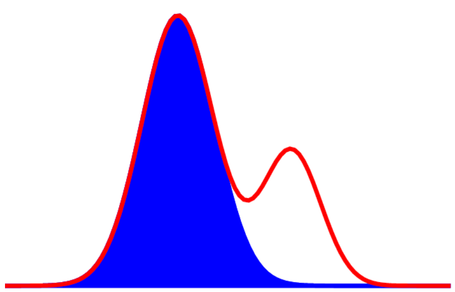
(Salakhutdinov, 2008; Salakhutdinov & Larochelle, AI & Statistics 2010)

Approximate intractable distribution $P_\theta(\mathbf{h}|\mathbf{v})$ with simpler, tractable distribution $Q_\mu(\mathbf{h}|\mathbf{v})$:

$$\mathrm{KL}(Q||P) = \int Q(x) \log \frac{Q(x)}{P(x)} dx$$

$$\log P_\theta(\mathbf{v}) \geq \log P_\theta(\mathbf{v})$$

$$= \sum_{\mathbf{h}} Q_\mu$$



Mean-F

Variational In
lower bound w.r.t. variational parameters μ .

Nonlinear fixed-point equations:

$$\mu_k^{(2)} = \sigma\left(\sum_j W_{jk}^2 \mu_j^{(1)} + \sum_m W_{km}^3 \mu_m^{(3)}\right)$$

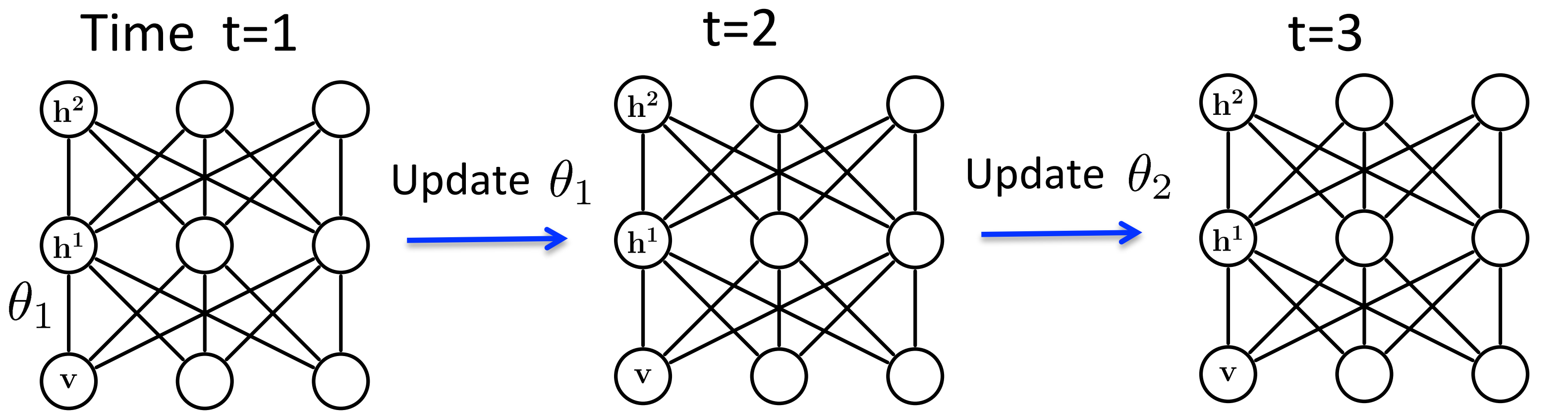
$$\mu_m^{(3)} = \sigma\left(\sum_k W_{km}^3 \mu_k^{(2)}\right)$$

(Approximate) Maximum Likelihood:

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

Variational Inference

Stochastic Approximation



$\mathbf{x}_1 \sim \mathrm{T}_{\theta_1}(\mathbf{x}_1 \leftarrow \mathbf{x}_0)$ $\mathbf{x}_2 \sim \mathrm{T}_{\theta_2}(\mathbf{x}_2 \leftarrow \mathbf{x}_1)$ $\mathbf{x}_3 \sim \mathrm{T}_{\theta_3}(\mathbf{x}_3 \leftarrow \mathbf{x}_2)$

Update θ_t and $\mathbf{x}_t$ sequentially, where $\mathbf{x} = \{\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2\}$

- Generate $\mathbf{x}_t \sim \mathrm{T}_{\theta_t}(\mathbf{x}_t \leftarrow \mathbf{x}_{t-1})$ by simulating from a Markov chain that leaves P_{θ_t} invariant (e.g. Gibbs or M-H sampler)
- Update θ_t by replacing intractable $\mathrm{E}_{P_{\theta_t}}[\mathbf{v}\mathbf{h}^\top]$ with a point estimate $[\mathbf{v}_t \mathbf{h}_t^\top]$

In practice we simulate several Markov chains in parallel.

Robbins and Monro, Ann. Math. Stats, 1957
L. Younes, Probability Theory 1989

Learning Algorithm

Update rule decomposes:

$$\theta_{t+1} = \theta_t + \alpha_t \Big(\mathbb{E}_{P_{data}}[\mathbf{v}\mathbf{h}^\top] - \frac{1}{M}\sum_{m=1}^{M} \mathbf{v}_t^{(m)} \mathbf{h}_t^{(m)\top} \Big)_{P_{\theta_t}}[\mathbf{v}\mathbf{h}^\top]$$

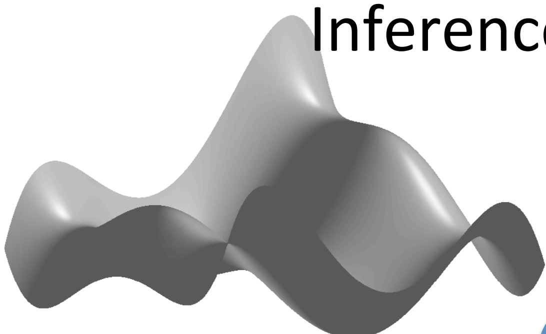
True gradient

Perturbation term ϵ_t

Variational Inference

MCMC

Almost sure convergence guarantees as learning rate $\alpha_t \to 0$



Problem: High-dimensional data:

...y multimodal.

Fast Inference

Key insight: The transition operator can be

Learning can scale to millions of examples

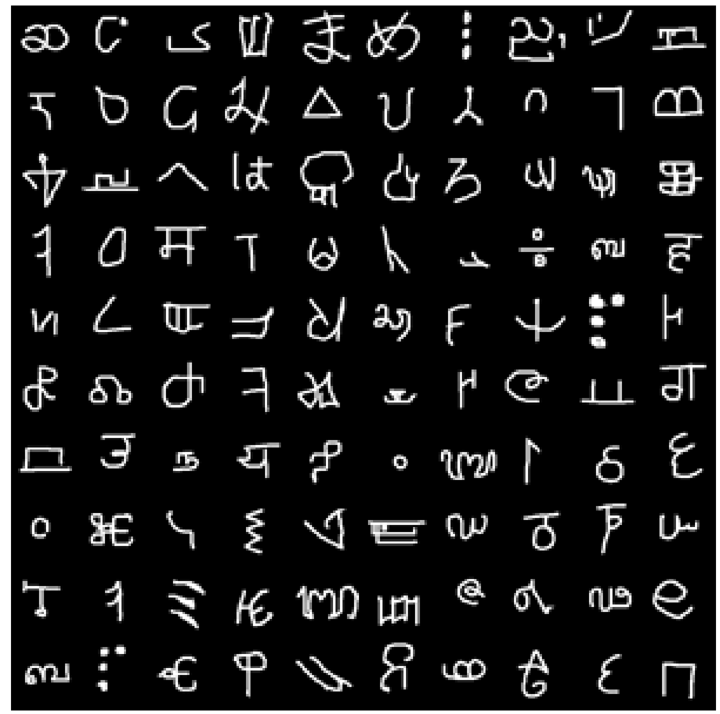
Connections to the theory of stochastic approximation and adaptive MCMC.

Good Generative Model?

Handwritten Characters

Good Generative Model?

Handwritten Characters



Good Generative Model?

Handwritten Characters

Simulated

Real Data

Good Generative Model?

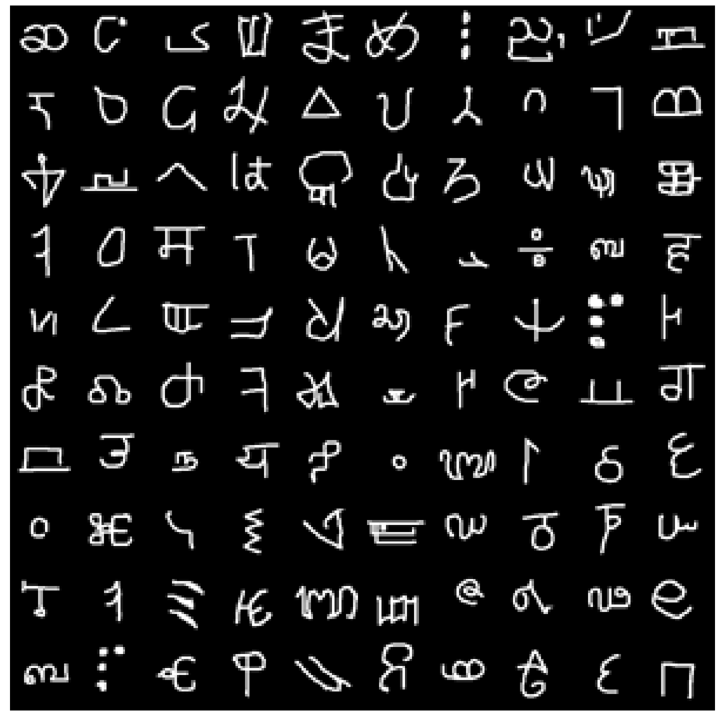
Handwritten Characters

Real Data

Simulated

Good Generative Model?

Handwritten Characters



Handwriting Recognition

MNIST Dataset
60,000 examples of 10 digits

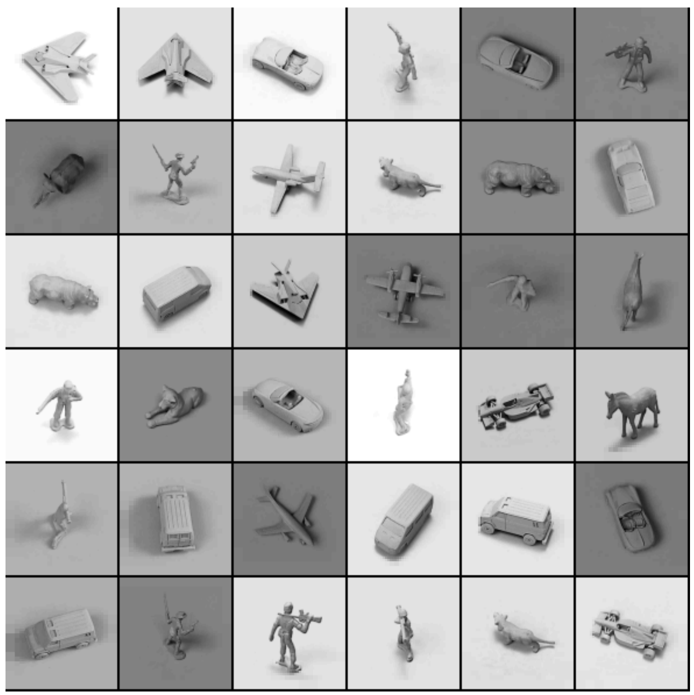
Learning Algorithm	Error
Logistic regression	12.0%
K-NN	3.09%
Neural Net (Platt 2005)	1.53%
SVM (Decoste et.al. 2002)	1.40%
Deep Autoencoder (Bengio et. al. 2007)	1.40%
Deep Belief Net (Hinton et. al. 2006)	1.20%
DBM	**0.95%**

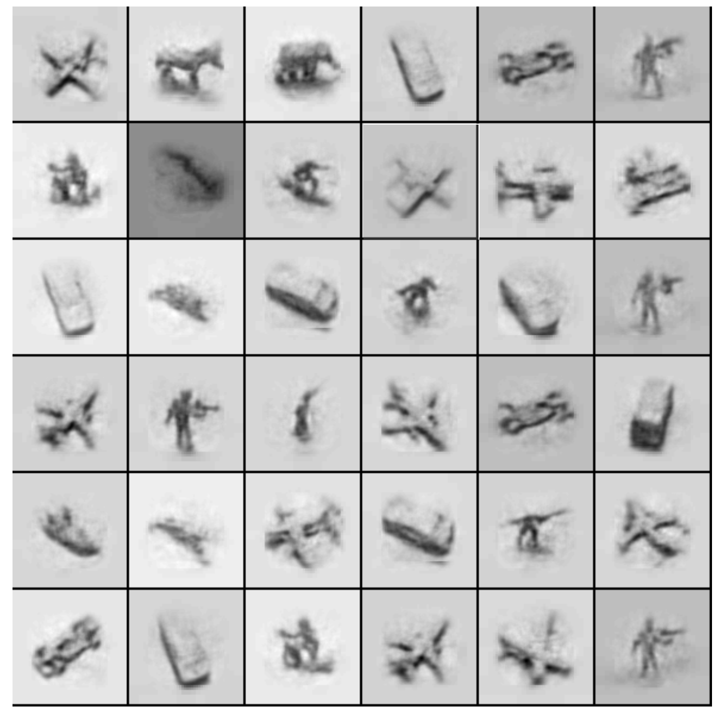
Optical Character Recognition
42,152 examples of 26 English letters

Learning Algorithm	Error
Logistic regression	22.14%
K-NN	18.92%
Neural Net	14.62%
SVM (Larochelle et.al. 2009)	9.70%
Deep Autoencoder (Bengio et. al. 2007)	10.05%
Deep Belief Net (Larochelle et. al. 2009)	9.68%
DBM	**8.40%**

Permutation-invariant version.

Generative Model of 3-D Objects

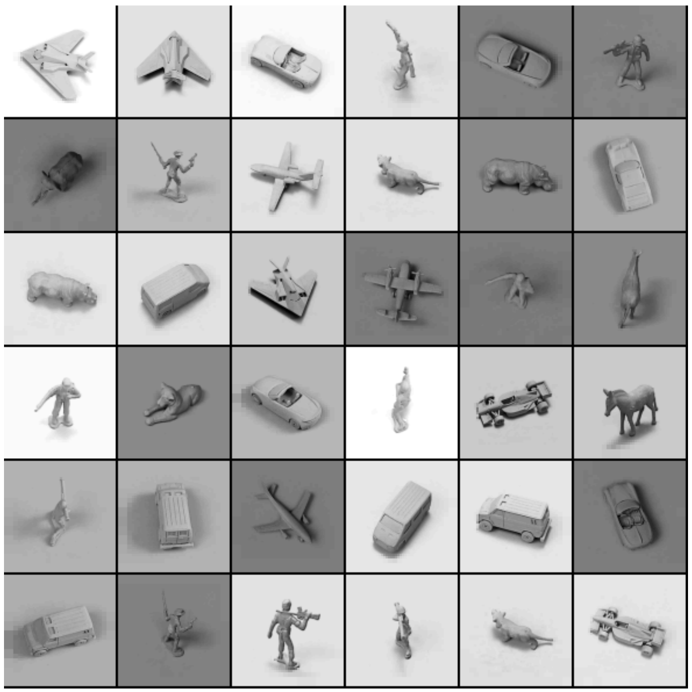




24,000 examples, 5 object categories, 5 different objects within each category, 6 lightning conditions, 9 elevations, 18 azimuths.

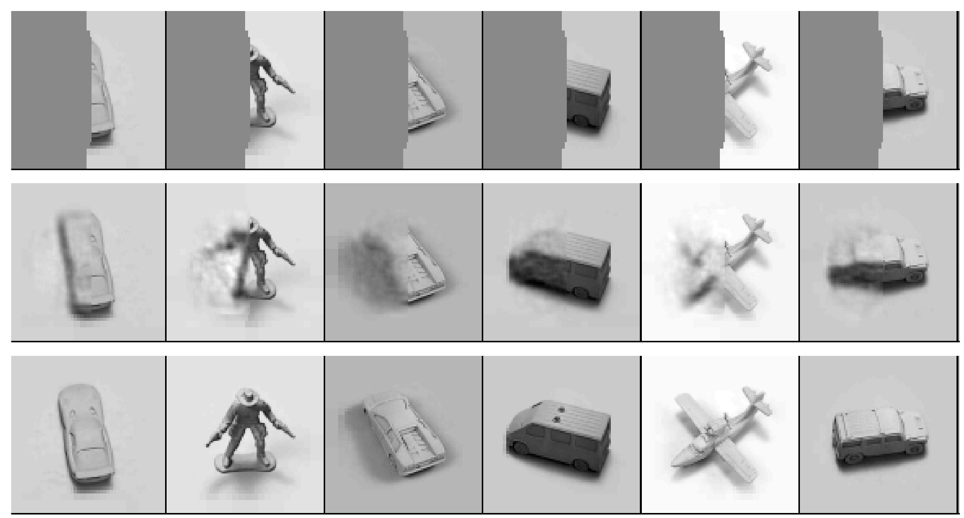
3-D object Recognition

NORB Dataset: 24,000 examples



Learning Algorithm	Error
Logistic regression	22.5%
K-NN (LeCun 2004)	18.92%
SVM (Bengio & LeCun 2007)	11.6%
Deep Belief Net (Nair & Hinton 2009)	9.0%
DBM	**7.2%**

Pattern Completion



Learning Hierarchical Representations

Deep Boltzmann Machines:

Learning Hi

in Features

of edges.

Need more structured and robust models

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

Talk Roadmap

- Learning Deep Models
  - Restricted Boltzmann Machines
  - Deep Boltzmann Machines
- Learning Structured and Robust Models
- Multi-Modal Learning

Face Recognition

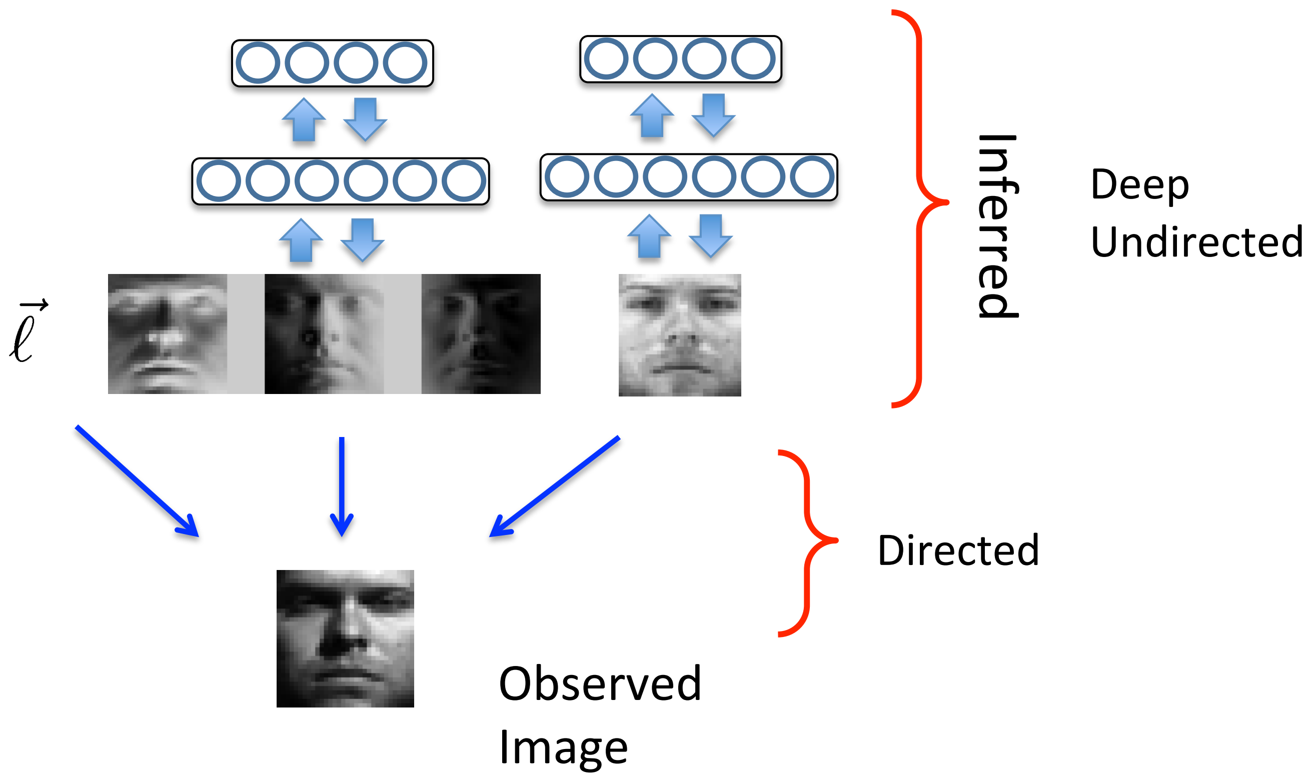
Yale B Extended Face Dataset
4 subsets of increasing illumination variations

Subset 1
Subset 2
Subset 3
Subset 4

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.

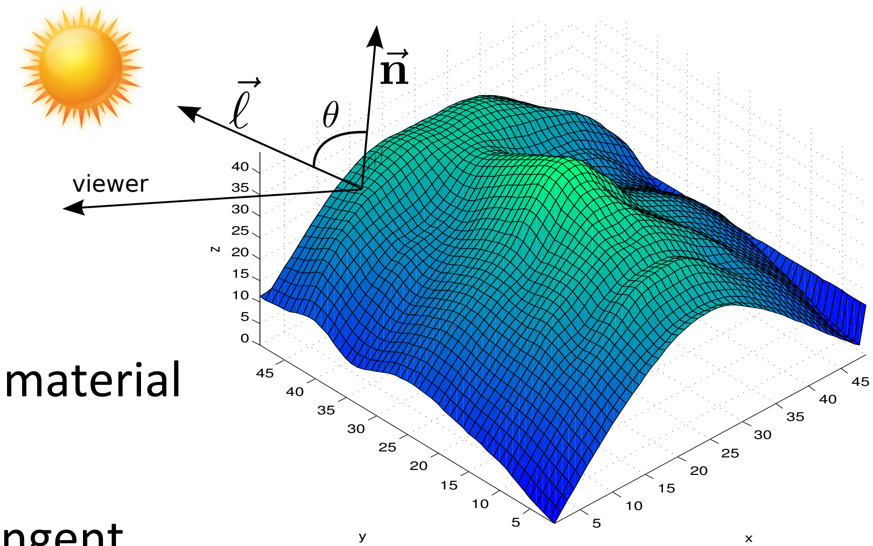
(Tang et. Al., ICML 2012, Tang et. al. CVPR 2012)

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

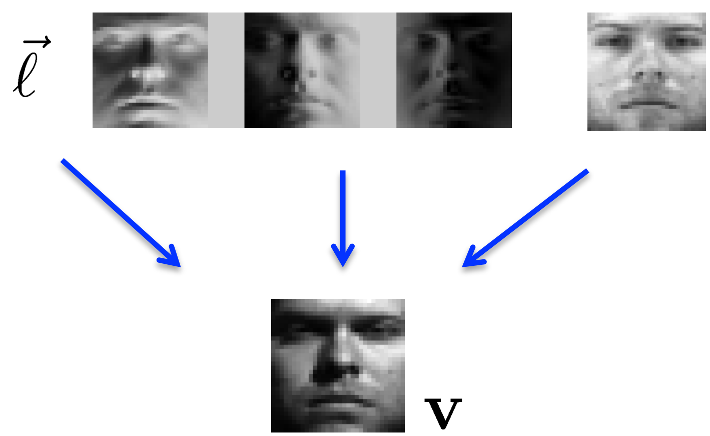


- Albedo -- diffuse reflectivity of a surface, material dependent, illumination independent.

- 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



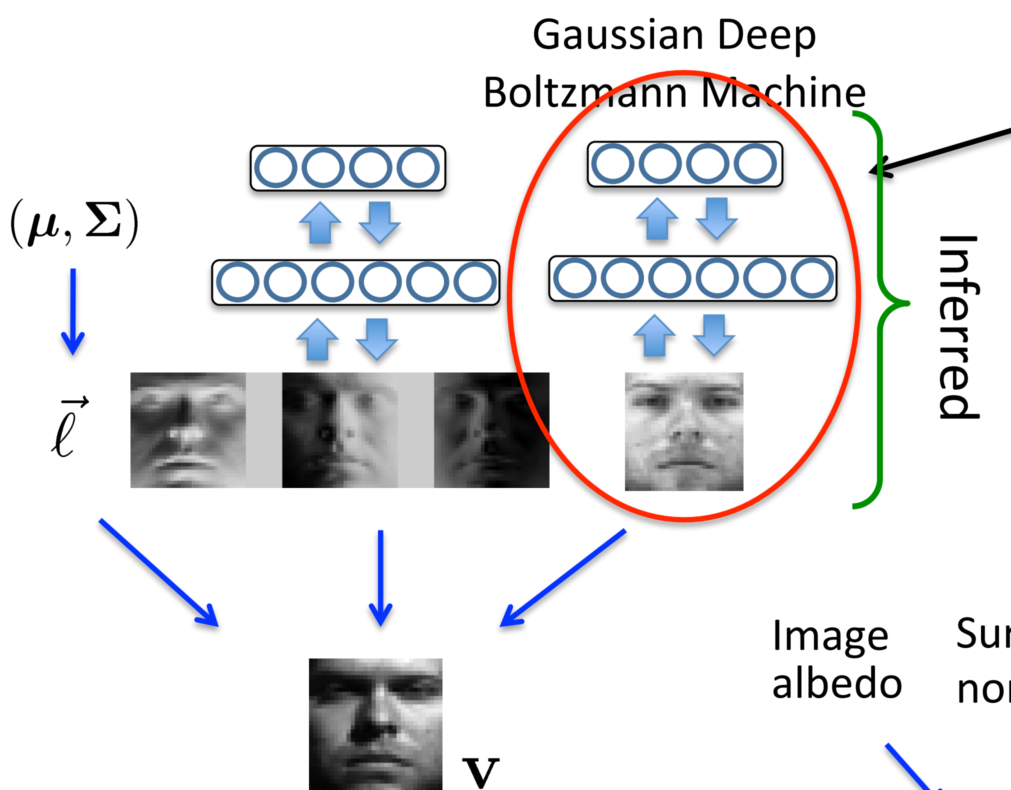
Observed Image

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



Observed
Image

Transfer Learning

Albedo DBM:
Pretrained using
Toronto Face Database

$$\mathbf{a} \quad \sim \quad \text{GDBM}(\mathbf{a})$$

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

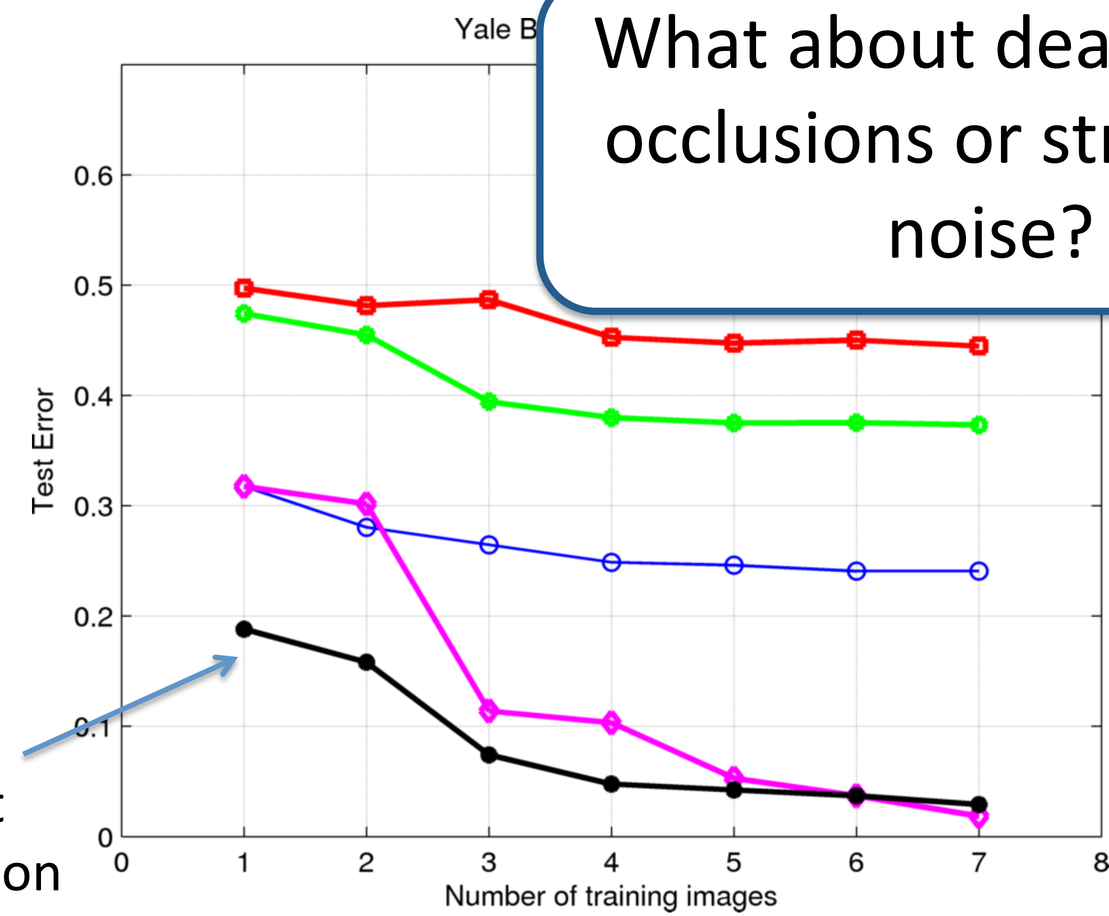
Observed

Inferred albedo

Face Relighting

Recognition Results

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



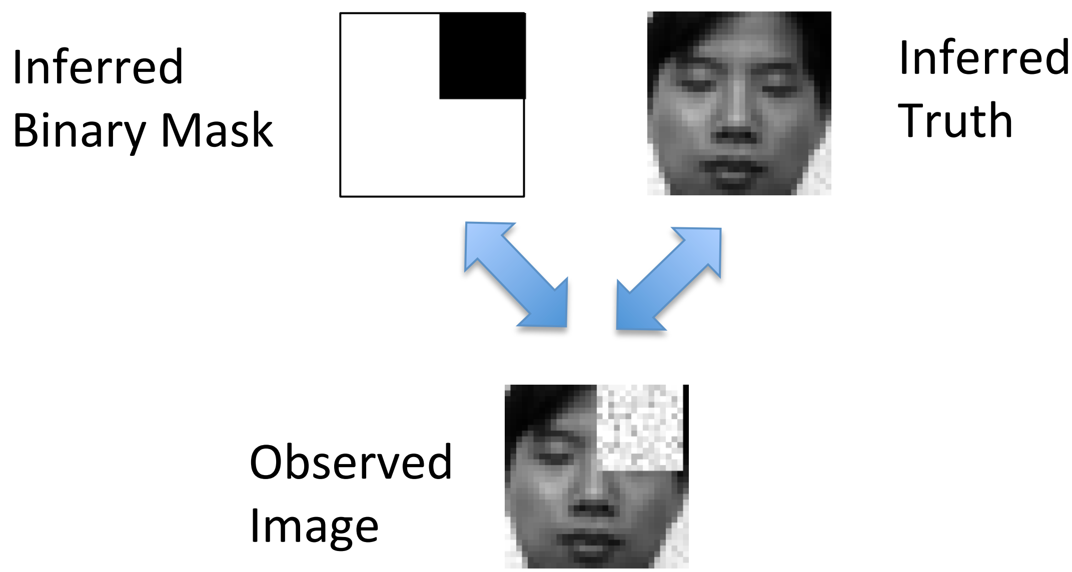
What about dealing with occlusions or structured noise?

One-Shot Recognition

Robust Boltzmann Machines

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

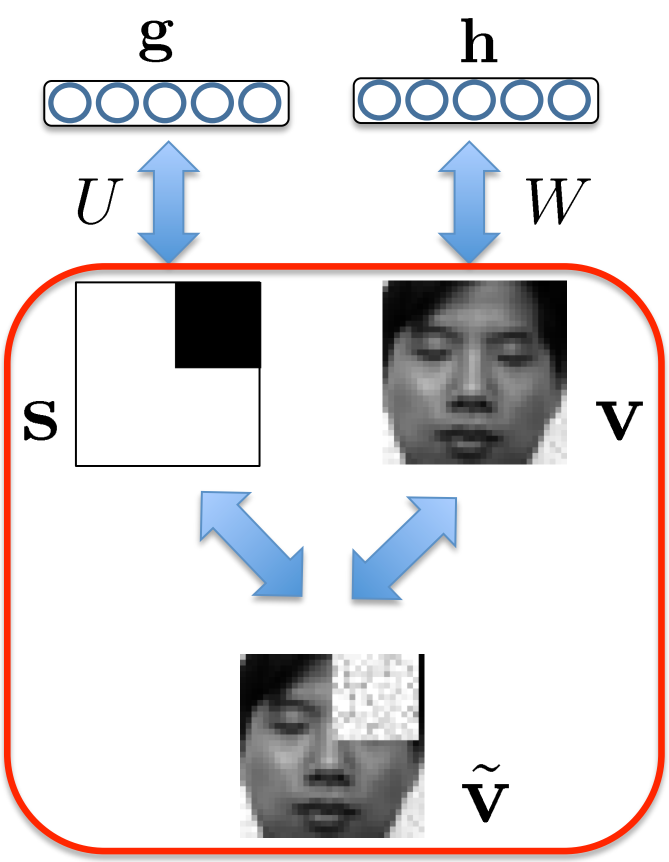
$$\log P(\tilde{\mathbf{v}}, \mathbf{v}, \mathbf{s}, \mathbf{h}, \mathbf{g}) \sim$$



(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

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

$$-\frac{1}{2} \sum_{i \in \text{pixels}} \frac{(v_i - b_i)^2}{\sigma_i^2} + \mathbf{v}^\top W \mathbf{h}$$

Gaussian RBM, modeling clean faces

Binary RBM modeling occlusions

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

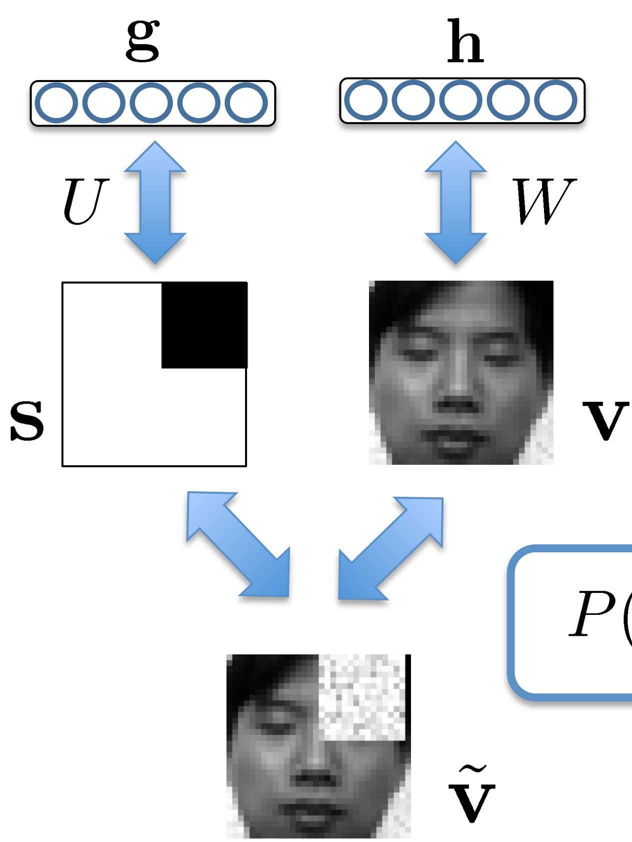
Binary pixel-wise Mask

Gaussian noise model

(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.



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

$$-\frac{1}{2} \sum_{i \in \text{pixels}} \frac{(v_i - b_i)^2}{\sigma_i^2} + \mathbf{v}^\top W \mathbf{h} + \mathbf{s}^\top U \mathbf{g}$$

Gaussian RBM, modeling clean faces

Binary RBM modeling occlusions

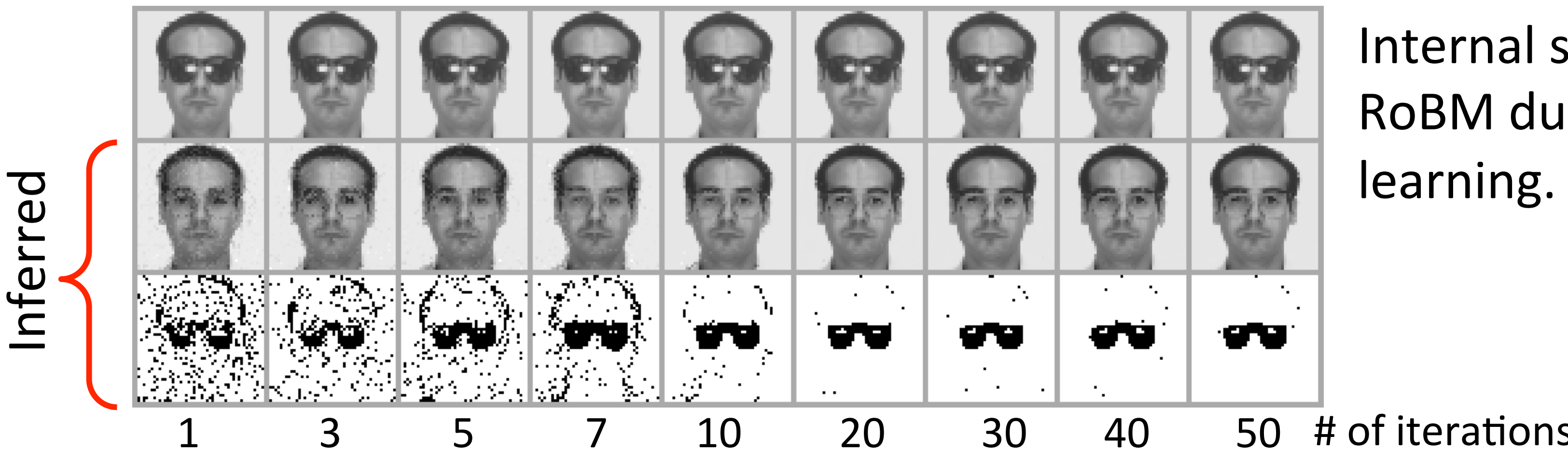
$P(\tilde{\mathbf{v}}|\mathbf{h}, \mathbf{g})$ is a heavy-tailed distribution

Binary pixel-wise Mask

Gaussian noise model

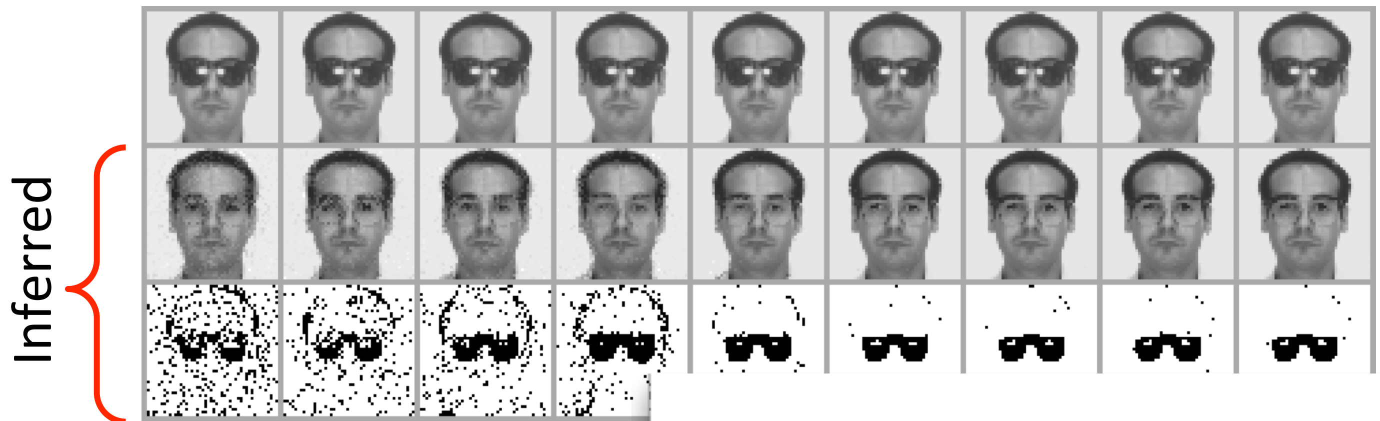
Inference: Variational Inference.
Learning: Stochastic Approximation

Recognition Results on AR Face Database

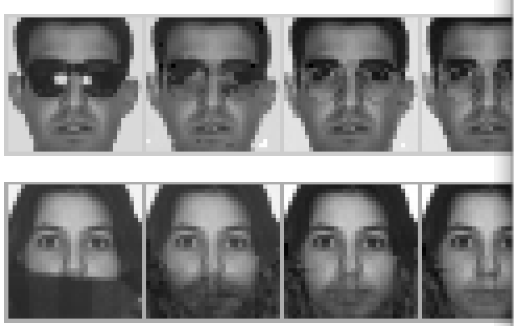


Internal states of RoBM during learning.

Recognition Results on AR Face Database



Internal states of RoBM during learning.



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%

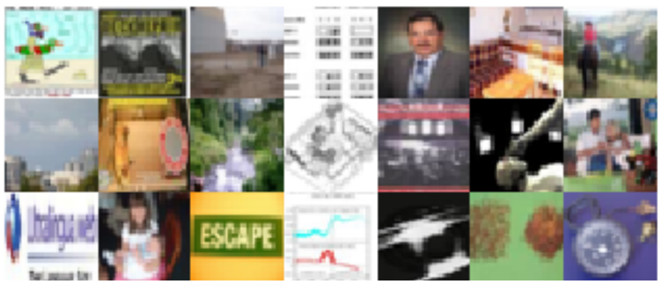
Transfer Learning

How can we learn a novel concept – a high dimensional statistical object – from few examples.

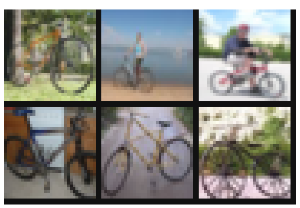
Transfer Learning

Background Knowledge

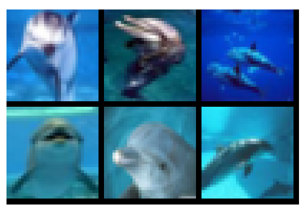
Millions of unlabeled images



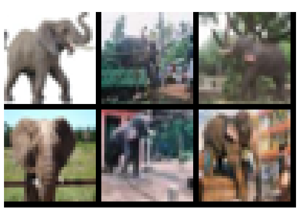
Some labeled images



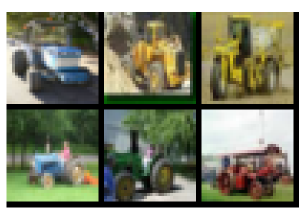
Bicycle



Dolphin



Elephant



Tractor

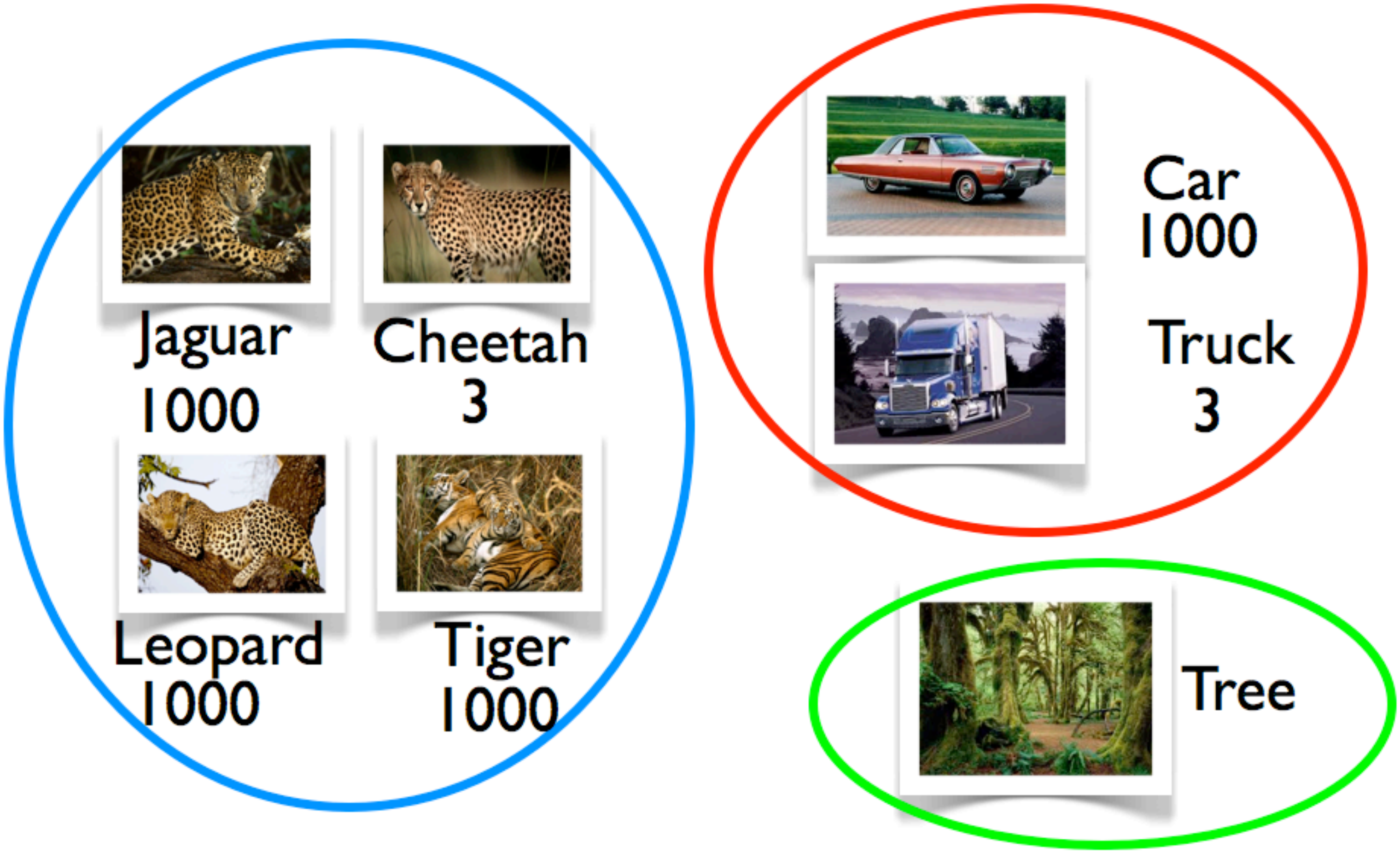
Learn to Transfer Knowledge

Learn novel concept from one example

Test:
What is this?

An Example

Structure in classes!



Slide Credit: Nitish Srivastava

Hierarchical-Deep Models

(Salakhutdinov, Tenenbaum, Torralba, NIPS 2011, PAMI 2013)

HD Models: Integrate hierarchical Bayesian models with deep models.

Hierarchical Bayes:

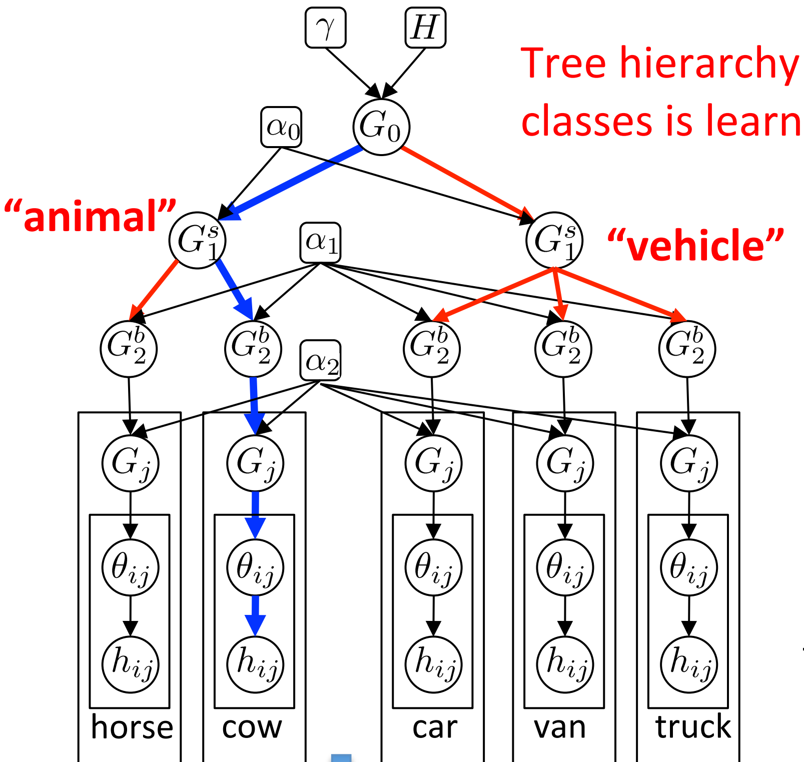
- Learn **hierarchies of categories** for sharing abstract knowledge.

Deep Models:

- Learn **hierarchies of features.**
- **Unsupervised feature learning** – no need to rely on human-crafted input features.

Hierarchical-Deep Models

(Salakhutdinov, Tenenbaum, Torralba, NIPS 2011, PAMI 2013)

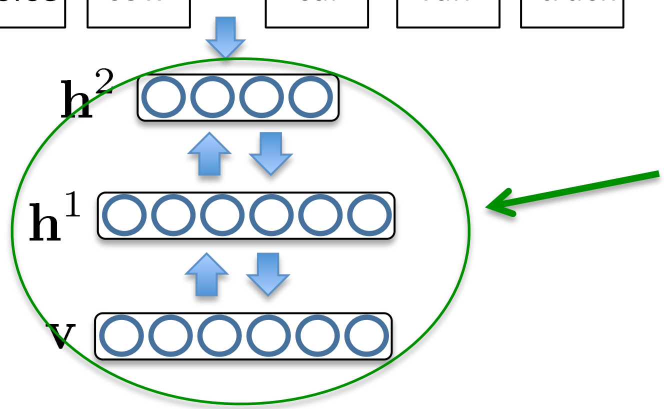


Tree hierarchy of classes is learned

$\mathbf{z} \sim \mathrm{nCRP}$ (**Nested Chinese Restaurant Process**) prior: a nonparametric prior over tree structures

$\mathbf{h}^3|\mathbf{z} \sim \mathrm{HDP}$ (**Hierarchical Dirichlet Process**) prior: a nonparametric prior allowing categories to share higher-level features, or parts.

$\mathbf{v}|\mathbf{h}^3 \sim \mathrm{DBM}$ **Deep Boltzmann Machine**



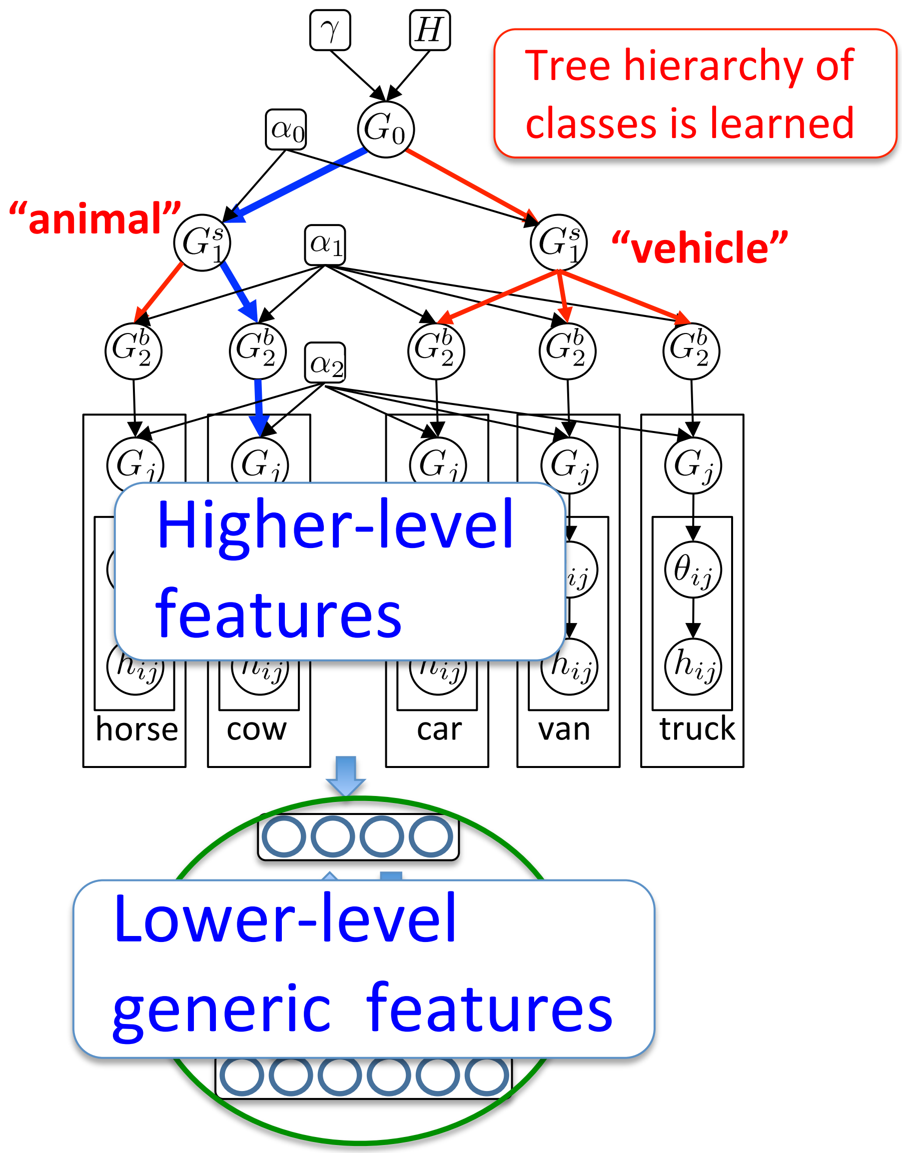
Enforce approximate global consistency through many local constraints.

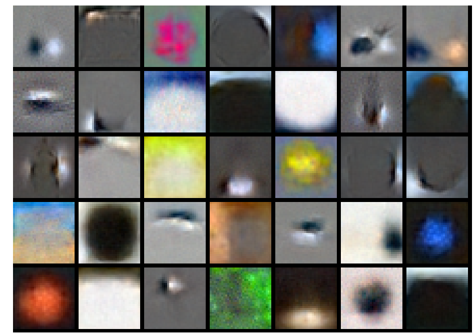
Incorporate prior knowledge to deal with occlusions, corrupted or missing data.

Images, Handwritten characters, Motion capture datasets.

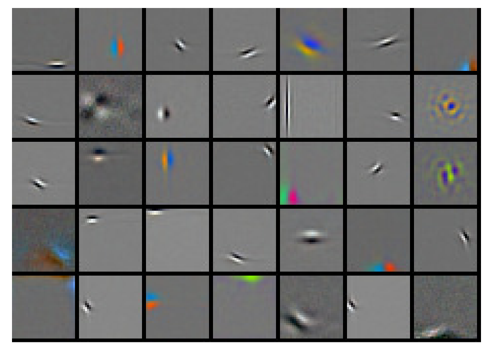
CIFAR Object Recognition

(Salakhutdinov, Tenenbaum, Torralba, NIPS 2011, PAMI 2013)



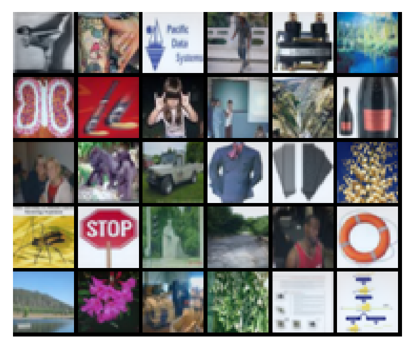


Learned high-level features



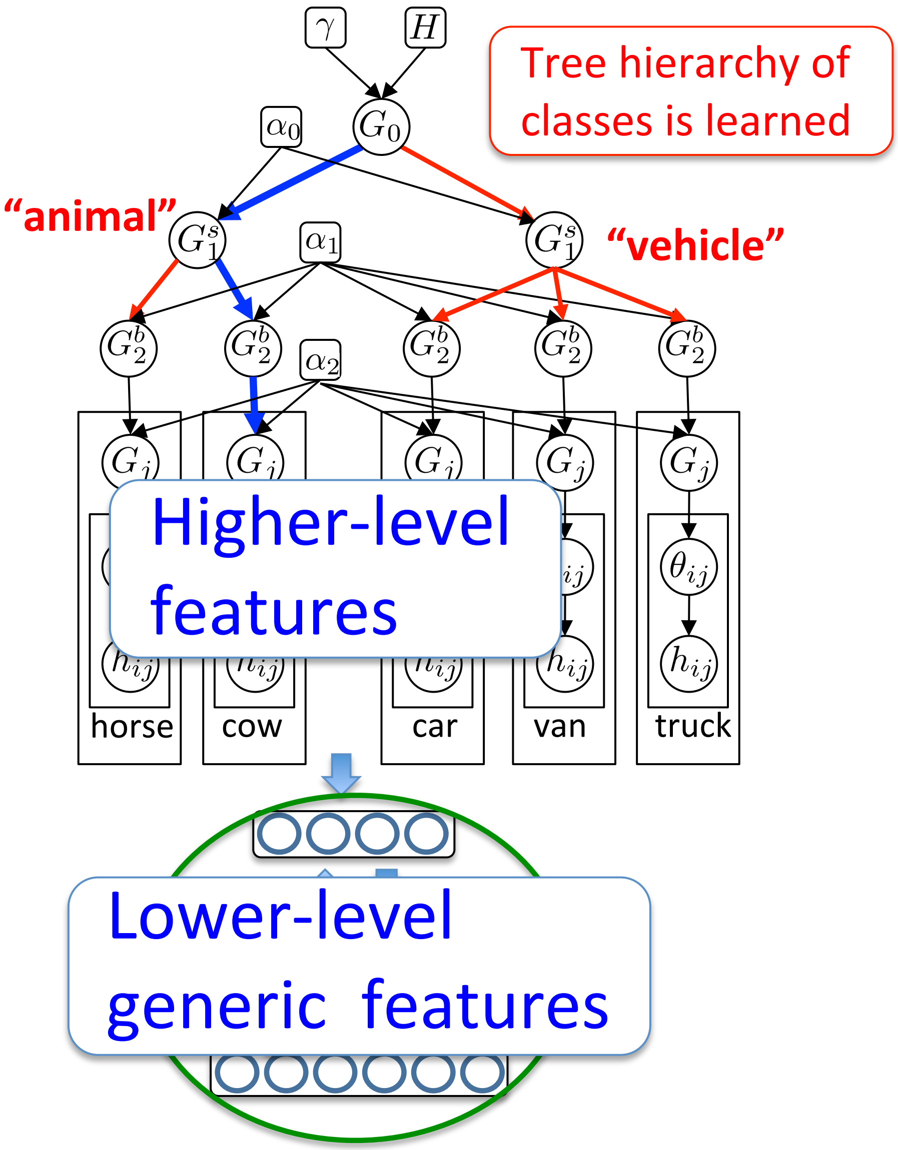
DBM generic features

4 million Images

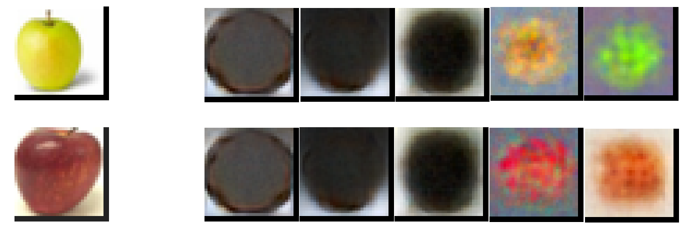


CIFAR Object Recognition

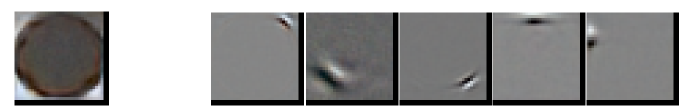
(Salakhutdinov, Tenenbaum, Torralba, NIPS 2011, PAMI 2013)



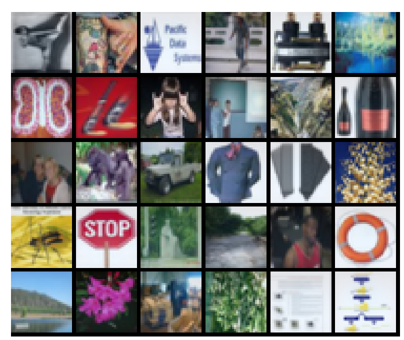
Each image is made up of learned high-level features features.



Each higher-level feature is made up of lower-level features.

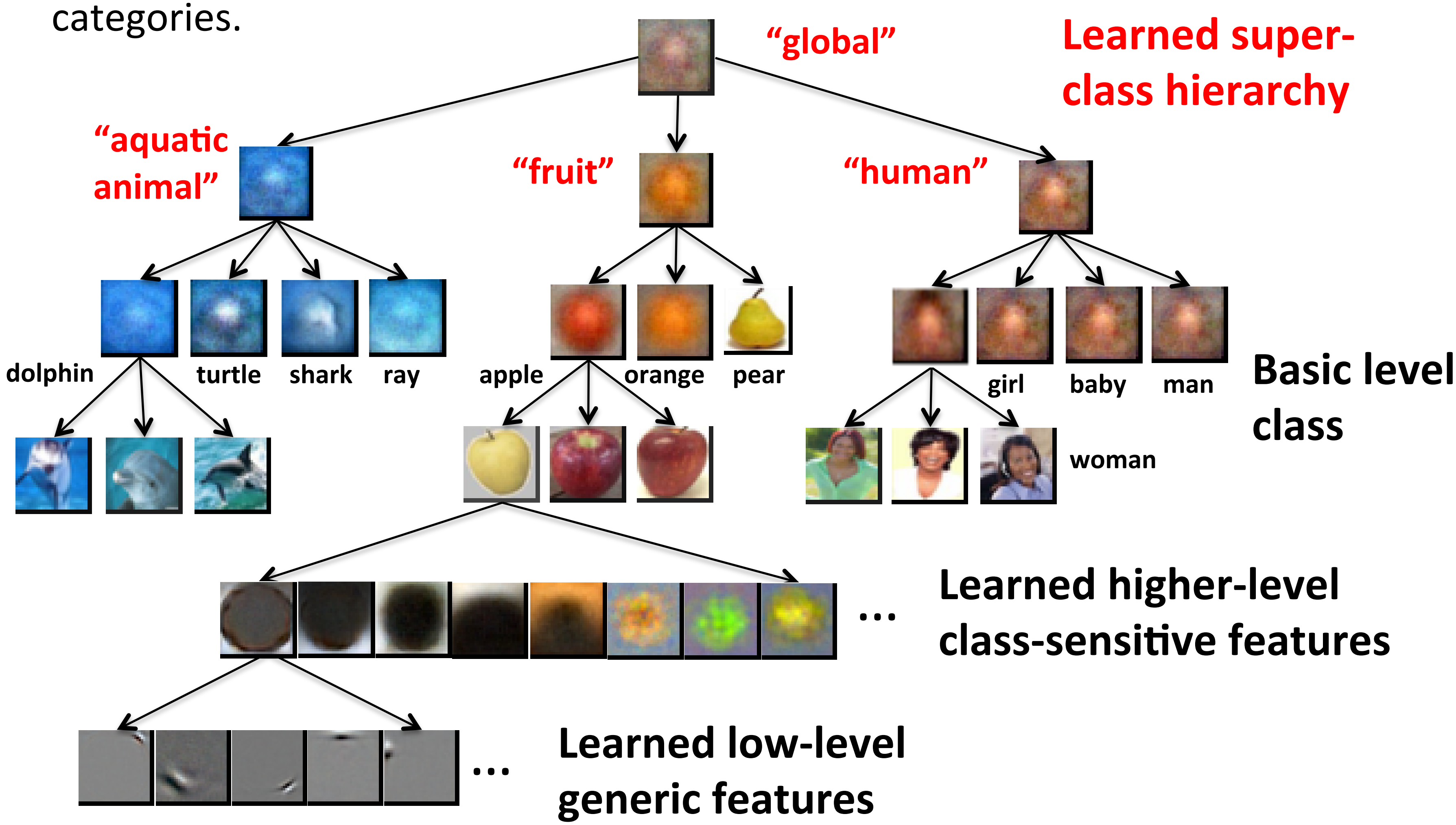


4 million Images



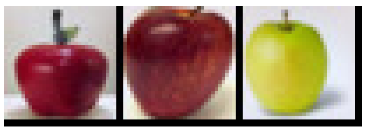
Learning Category Hierarchy

The model learns how to share the knowledge across many visual categories.



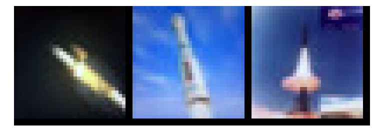
Learning from 3 Examples

Given only 3 Examples

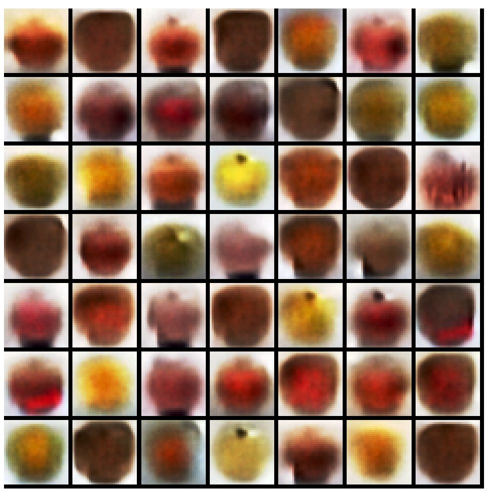


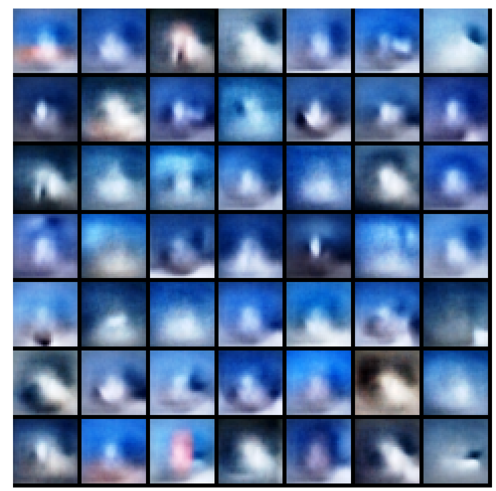
Willow Tree

Rocket

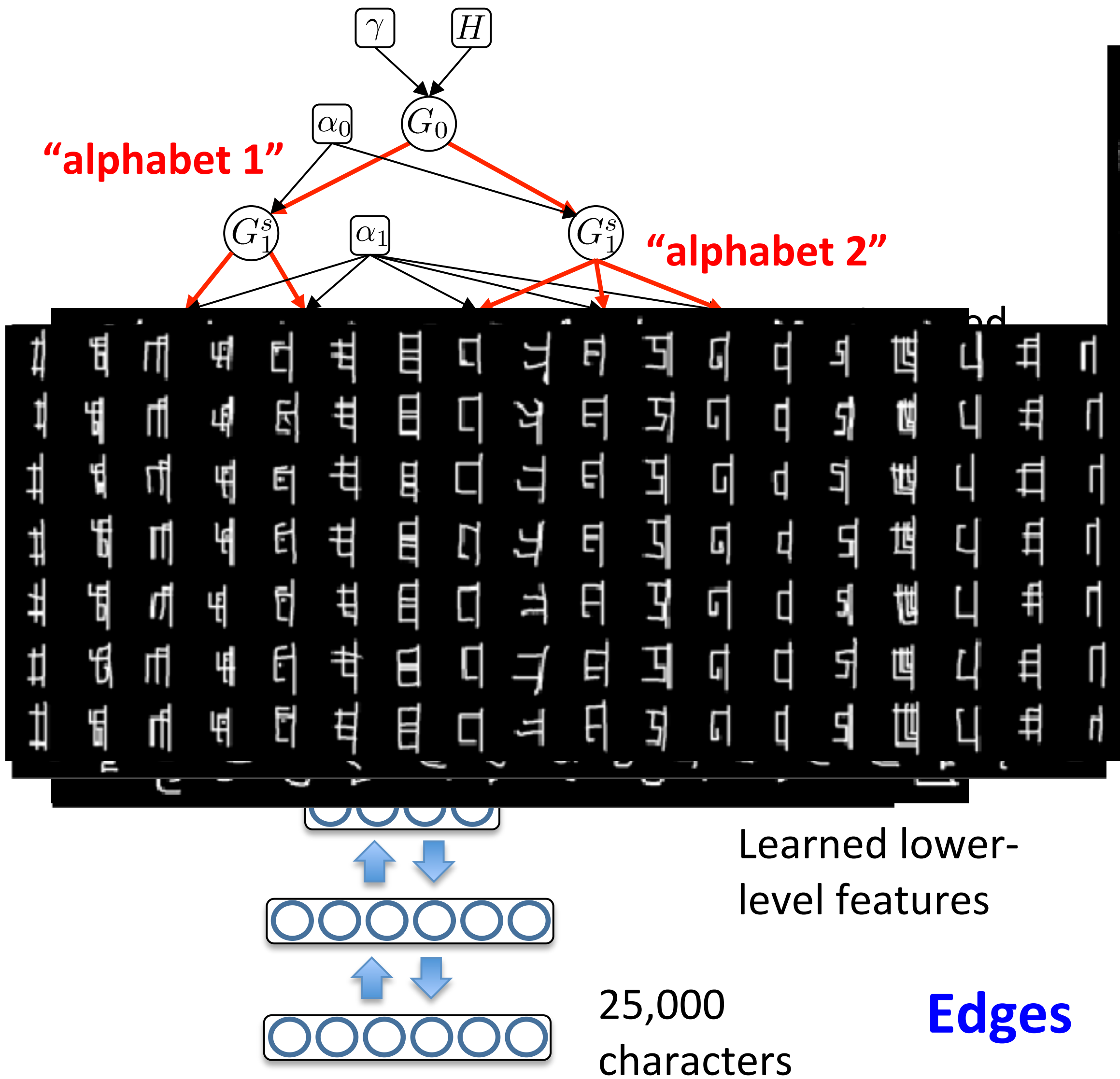


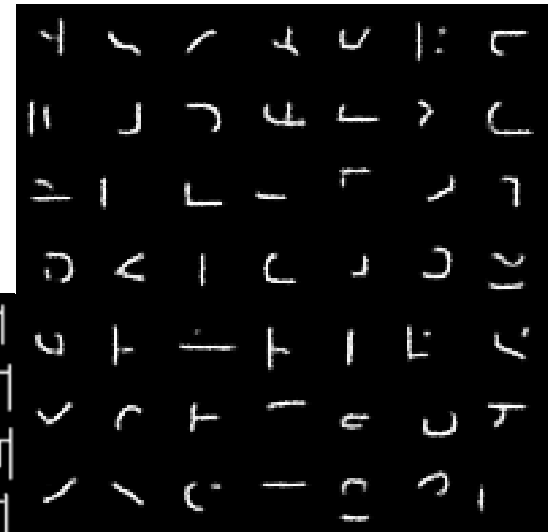
Generated Samples

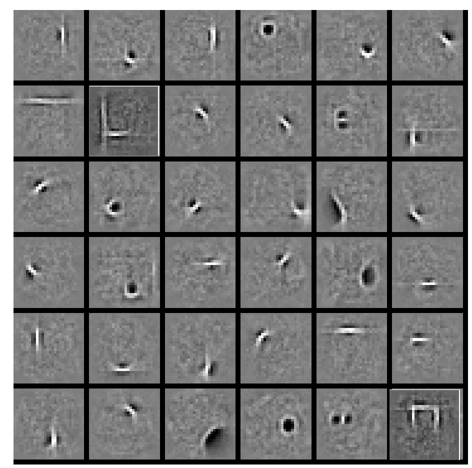




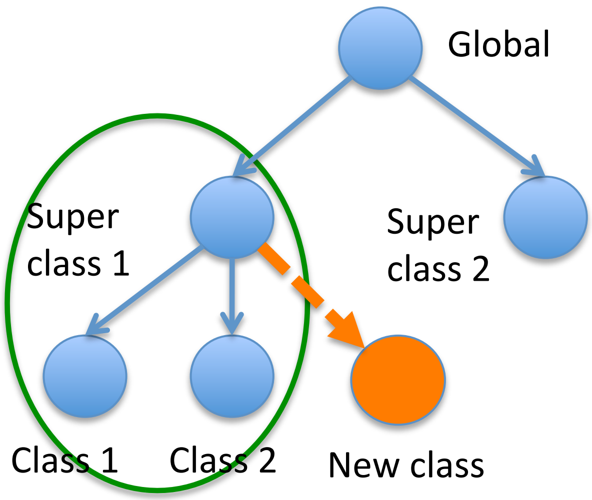
Handwritten Character Recognition



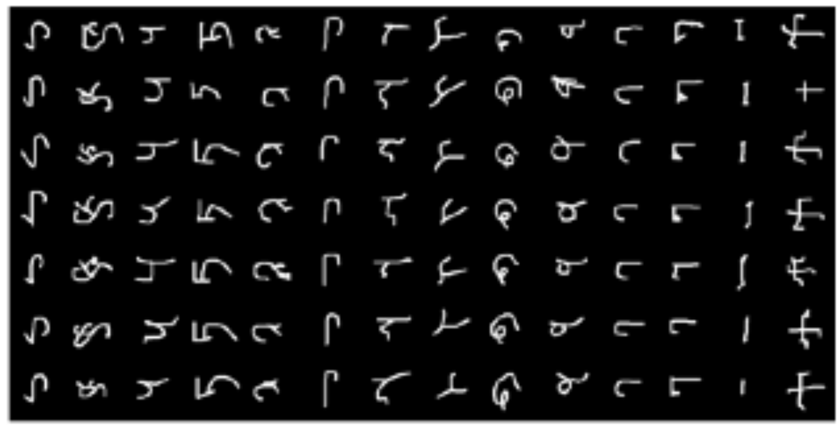




Simulating New Characters

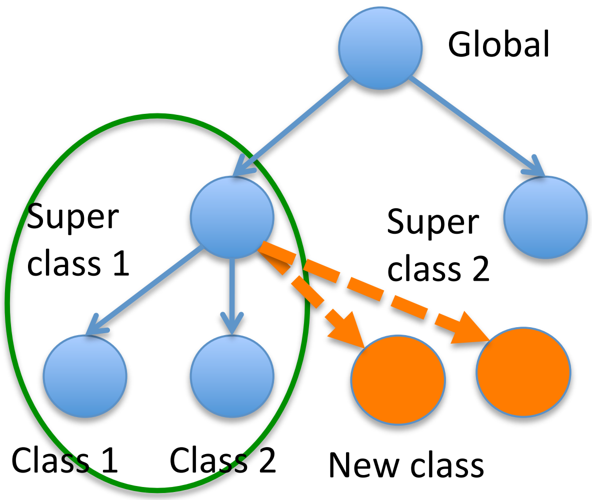


Real data within super class

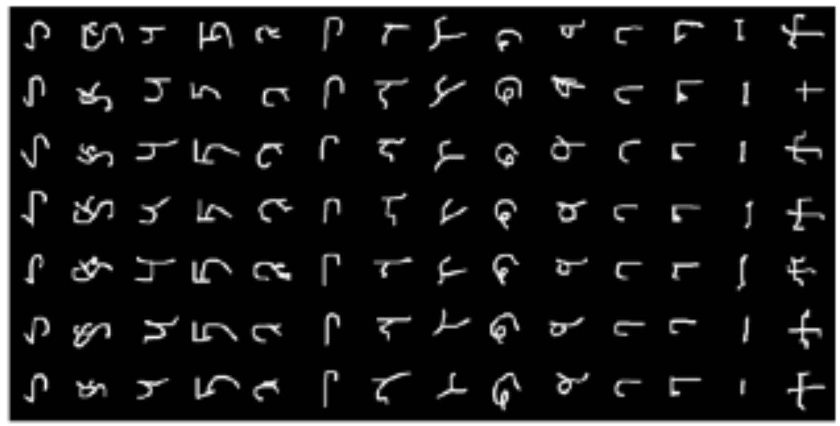


Simulated new characters

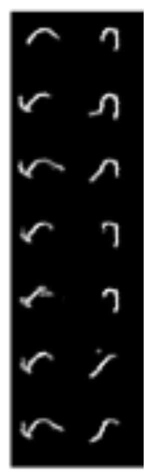
Simulating New Characters



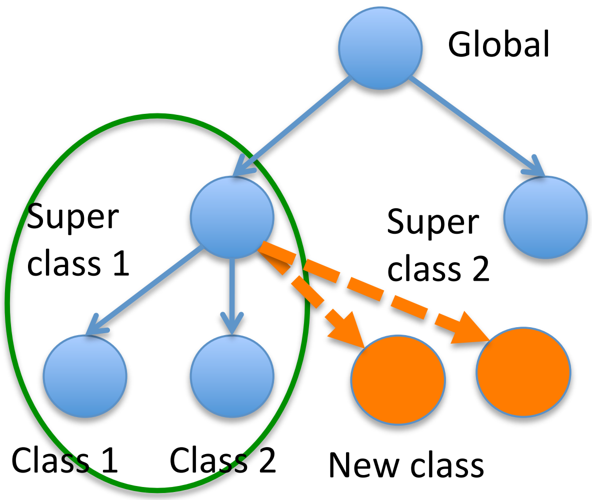
Real data within super class



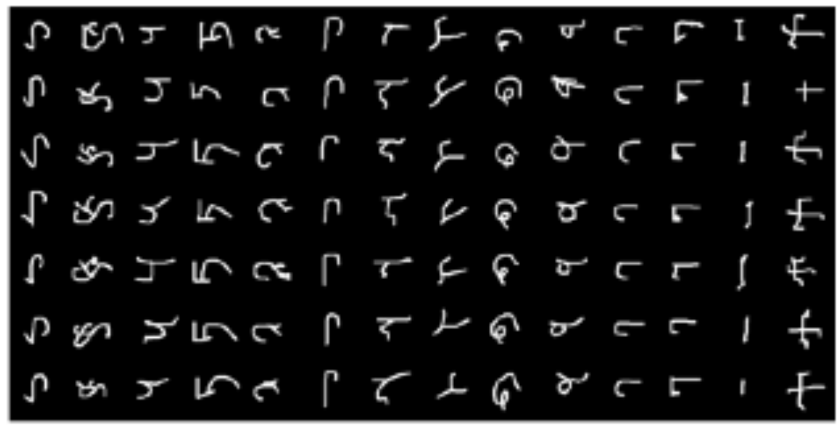
Simulated new characters



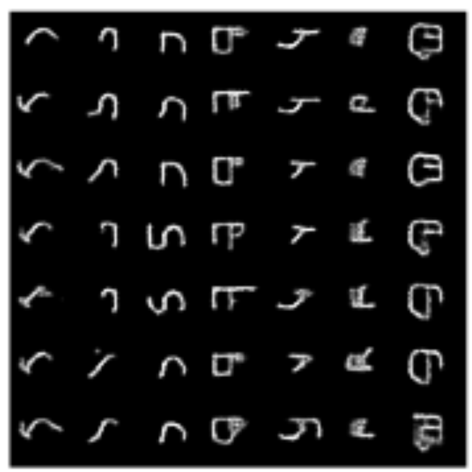
Simulating New Characters



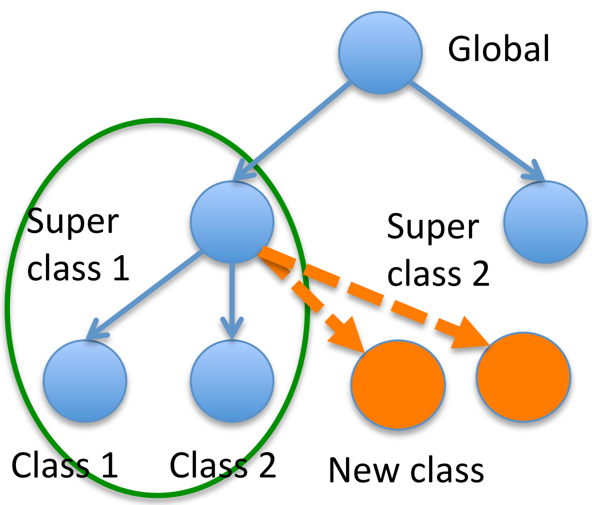
Real data within super class



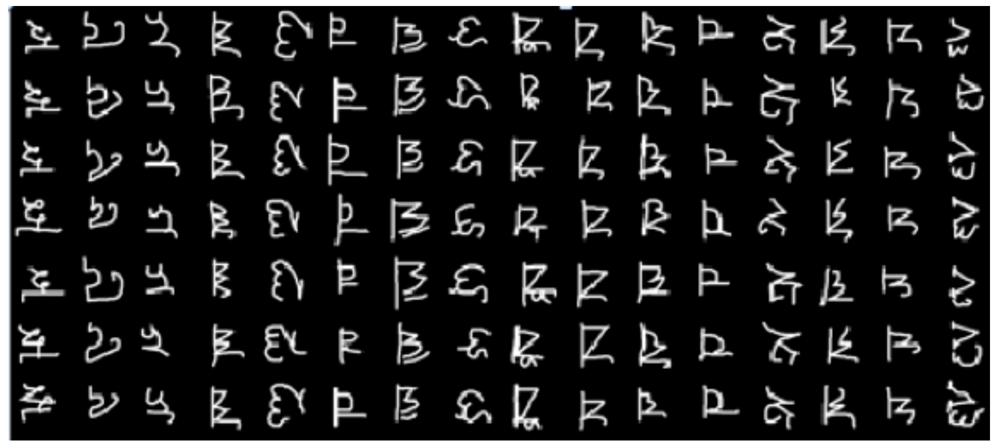
Simulated new characters



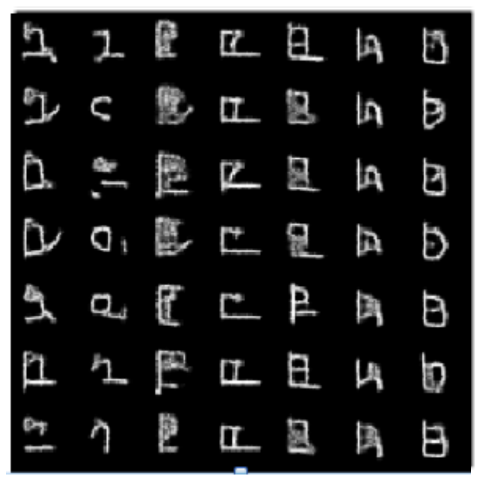
Simulating New Characters



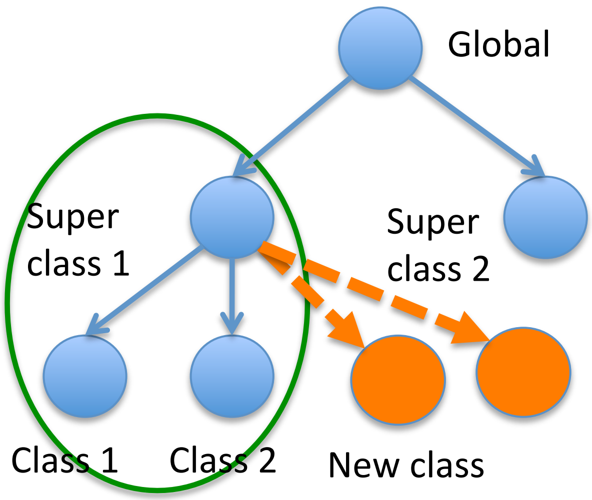
Real data within super class



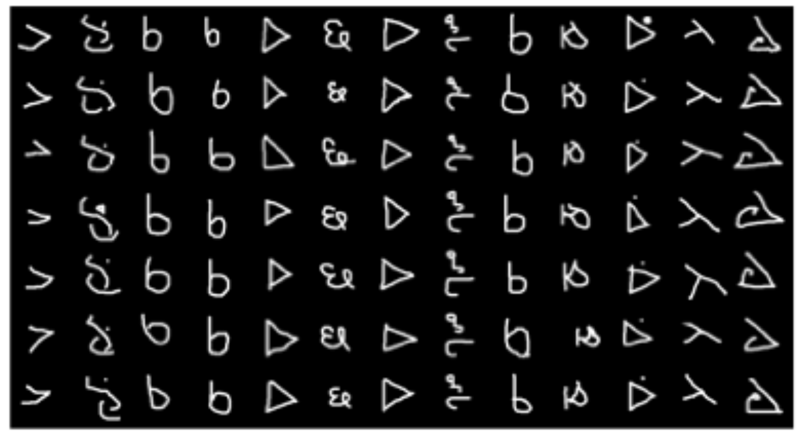
Simulated new characters



Simulating New Characters



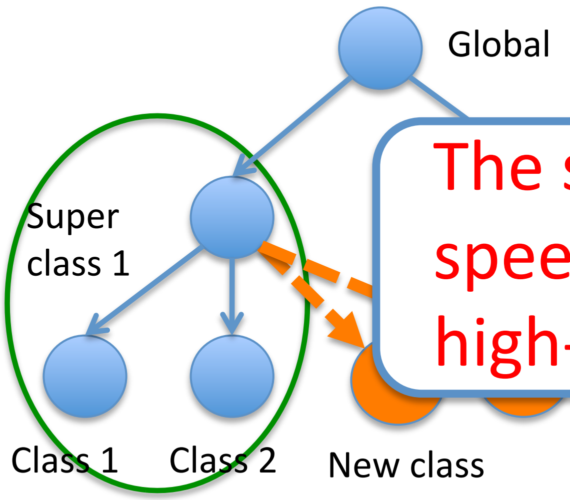
Real data within super class

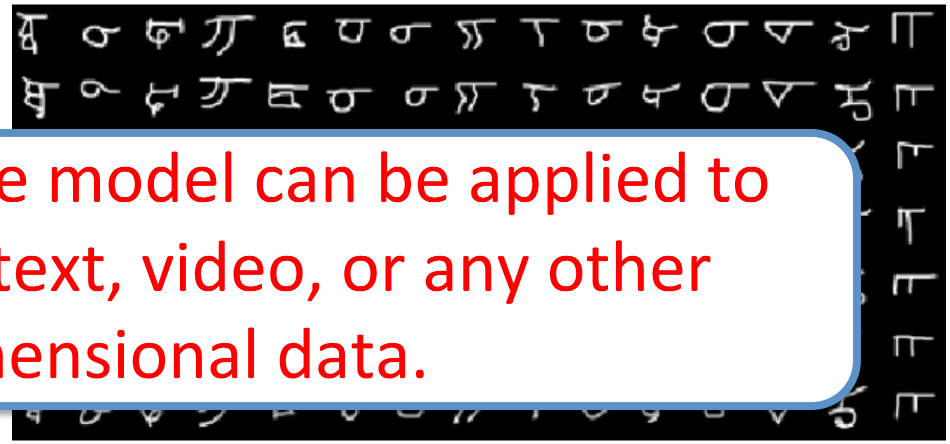


Simulated new characters

Simulating New Characters

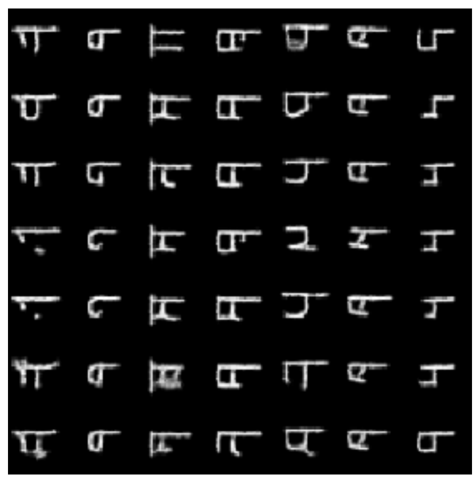
Real data within super class





The same model can be applied to speech, text, video, or any other high-dimensional data.

Simulated new characters



Talk Roadmap

- Learning Deep Models
  - Restricted Boltzmann Machines
  - Deep Boltzmann Machines
- Learning Structured and Robust Models
- Multi-Modal Learning

Data – Collection of Modalities

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

- Product recommendation systems.

- Robotics application

M

Shared Concept

"Modality-free" representation

"Concept"

sunset, pacific ocean, baker beach, seashore, ocean

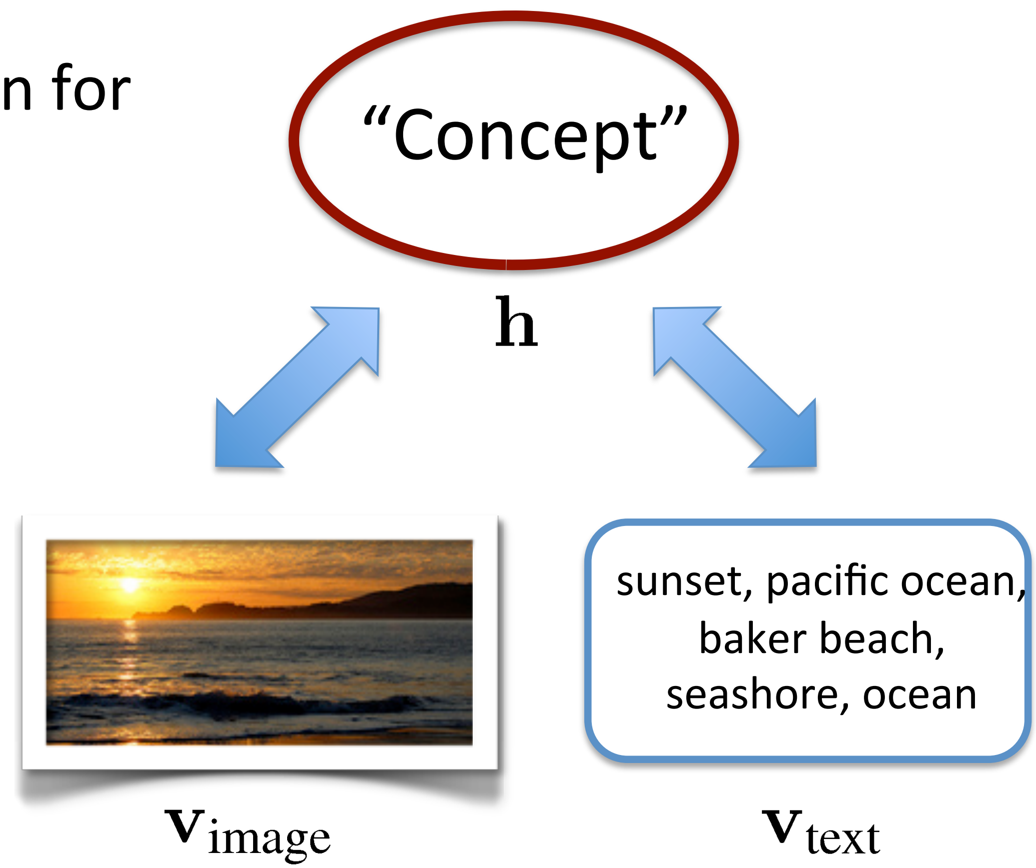
"Modality-full" representation

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

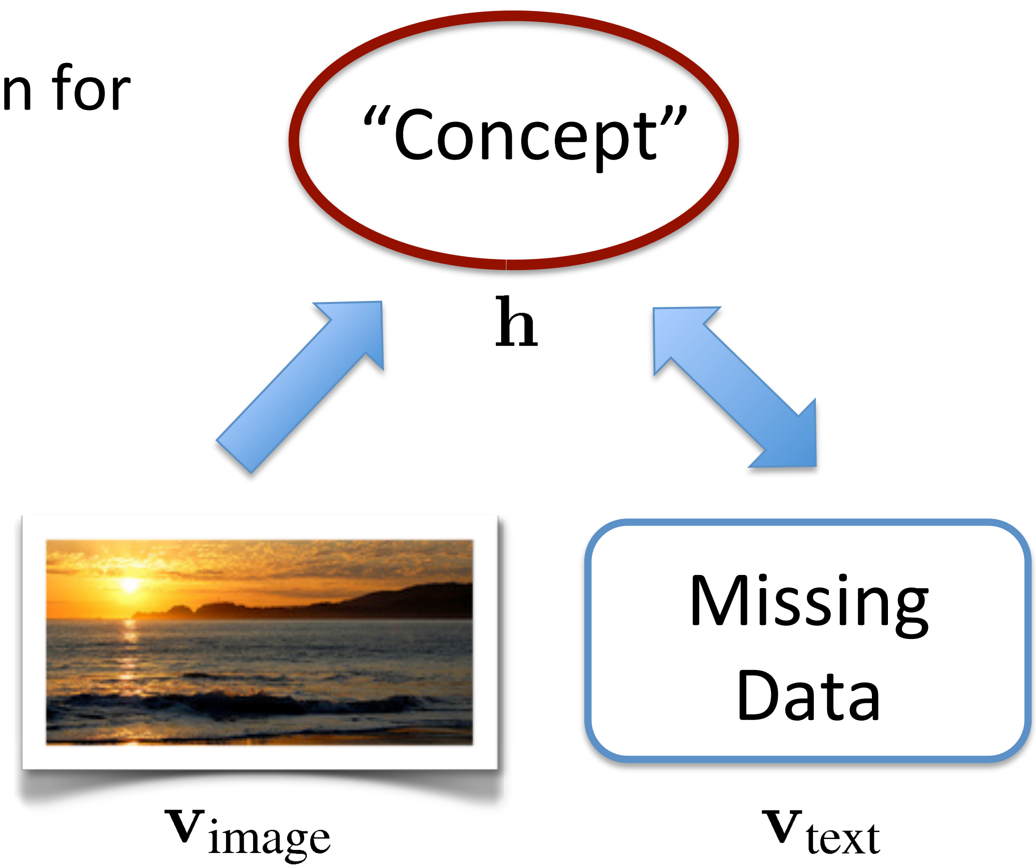
- Learn a joint density model: $P(\mathbf{h}, \mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}})$.

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

- Generate data from conditional distributions for

  - Image Annotation

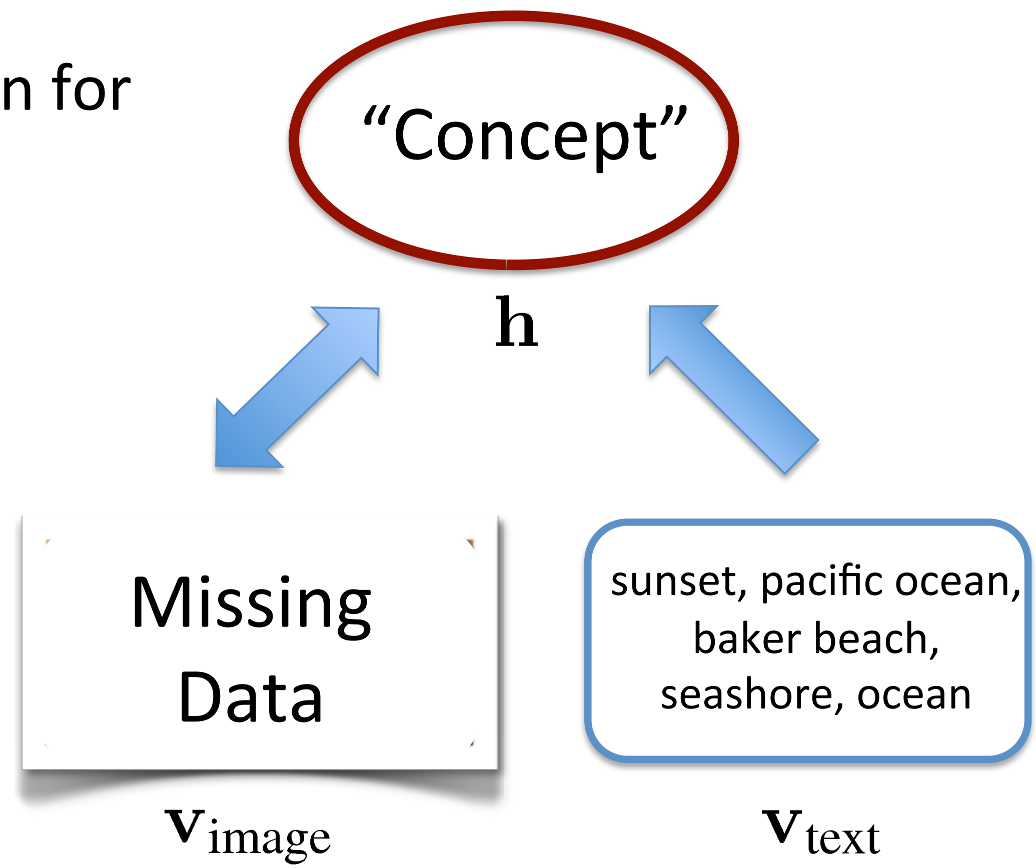
$$P(\mathbf{h}, \mathbf{v}_{\text{text}} | \mathbf{v}_{\text{image}})$$



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

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



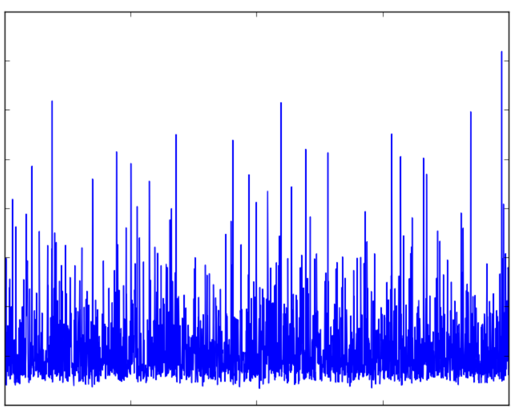
Challenges - I

Image

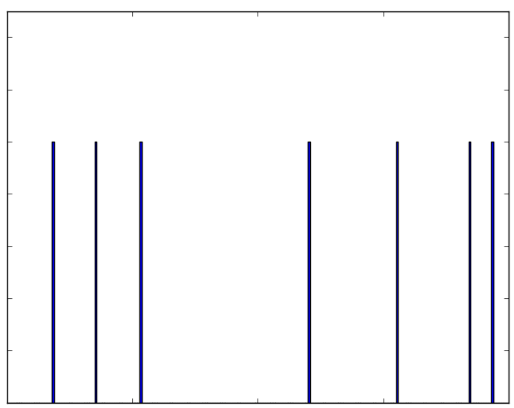
Text

sunset, pacific ocean, baker beach, 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, crap

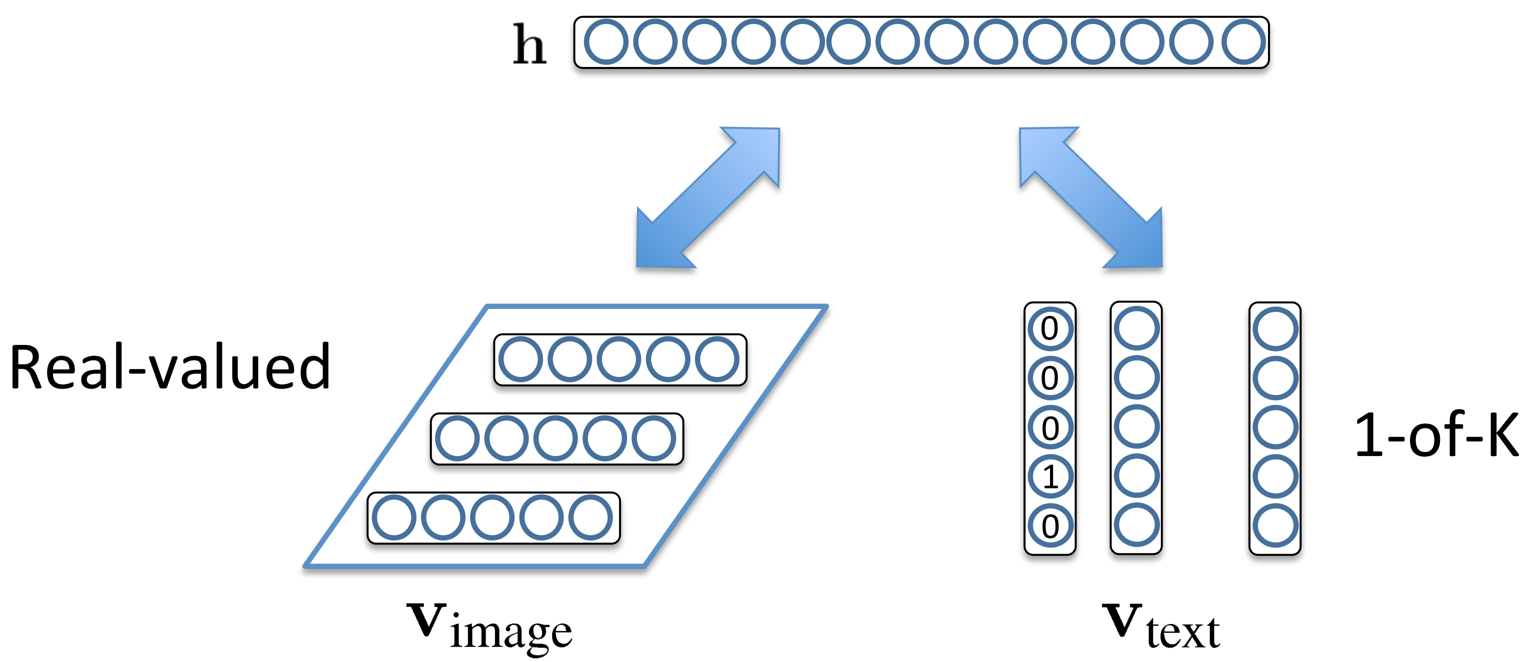
Noisy and missing data

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, crap	fall, autumn, trees, leaves, foliage, forest, woods, branches, path

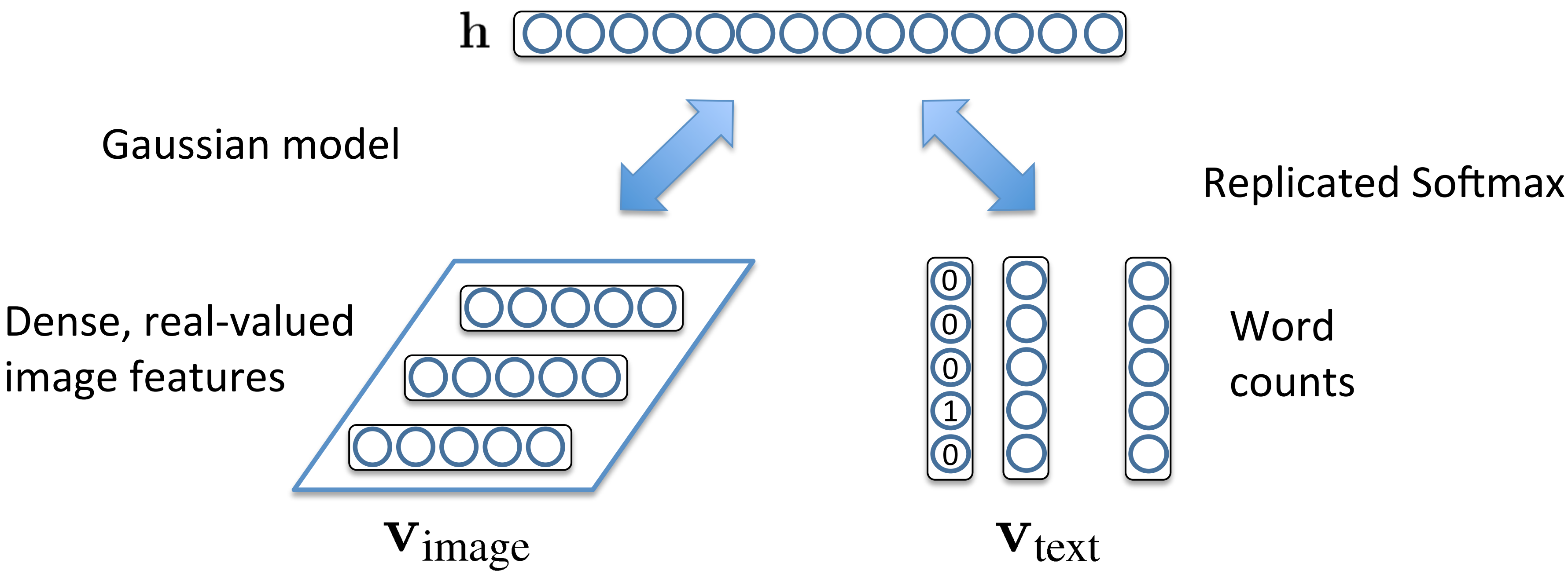
A Simple Multimodal Model

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



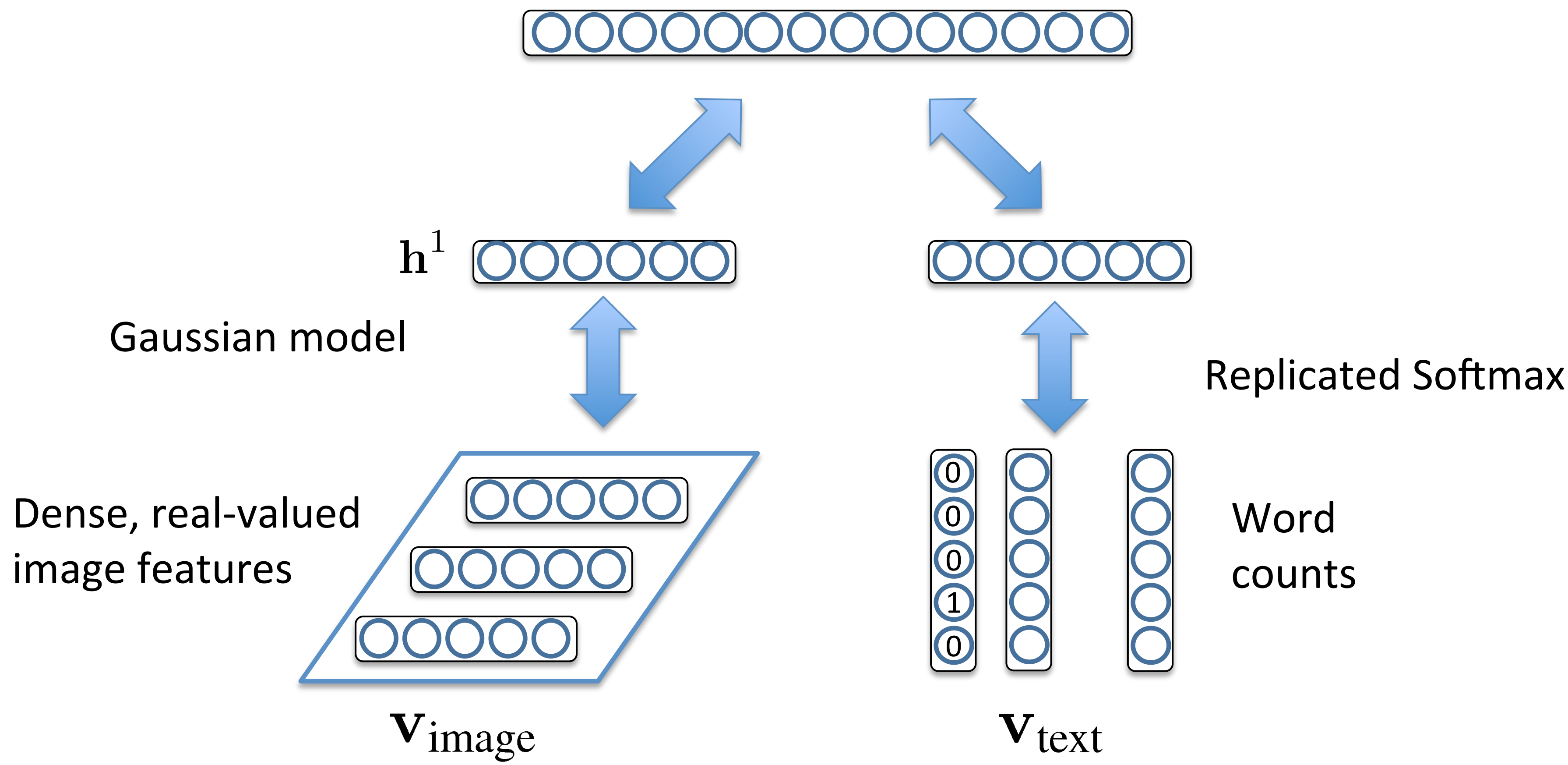
Multimodal DBM

(Srivastava & Salakhutdinov, NIPS 2012)



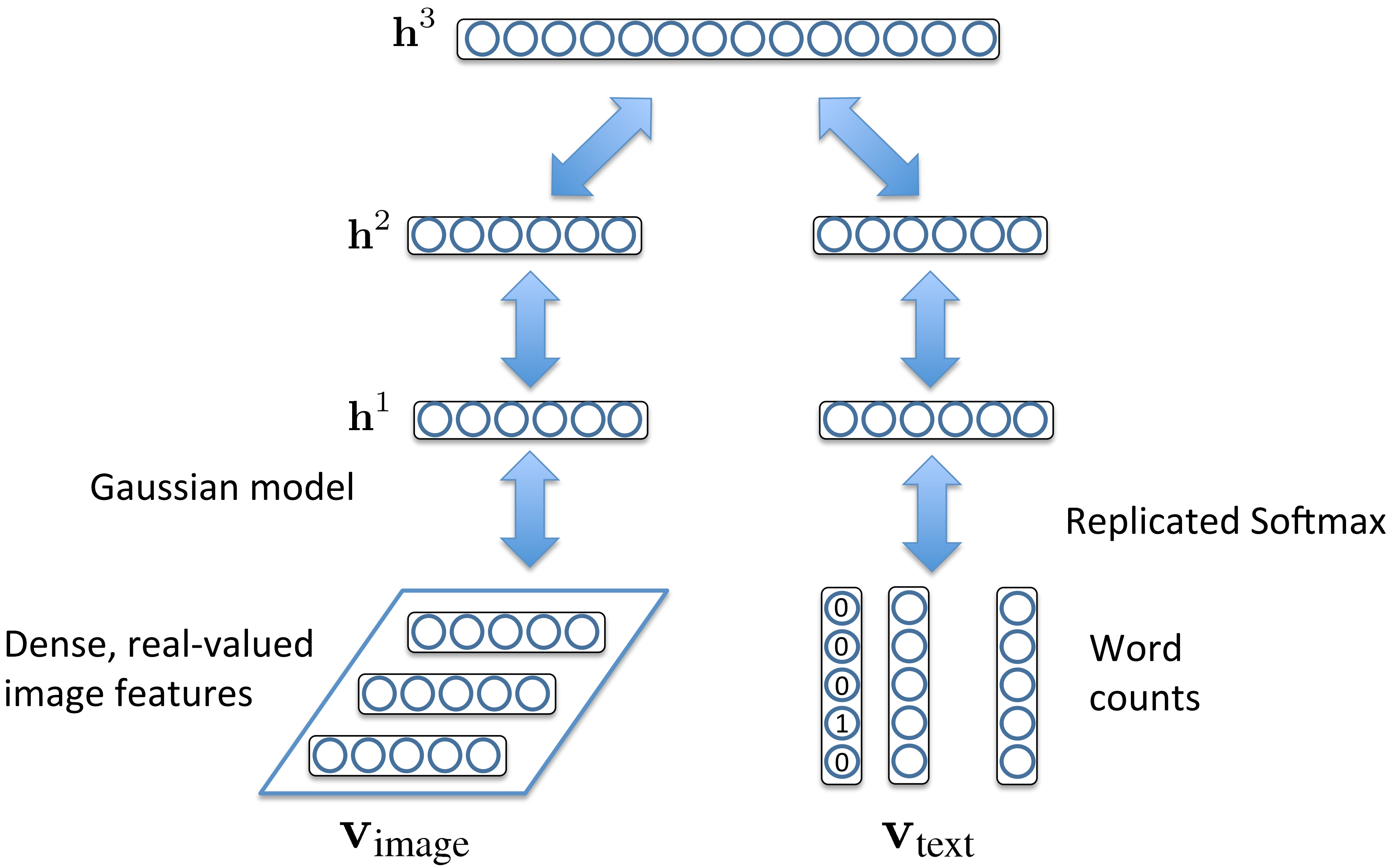
Multimodal DBM

(Srivastava & Salakhutdinov, NIPS 2012)



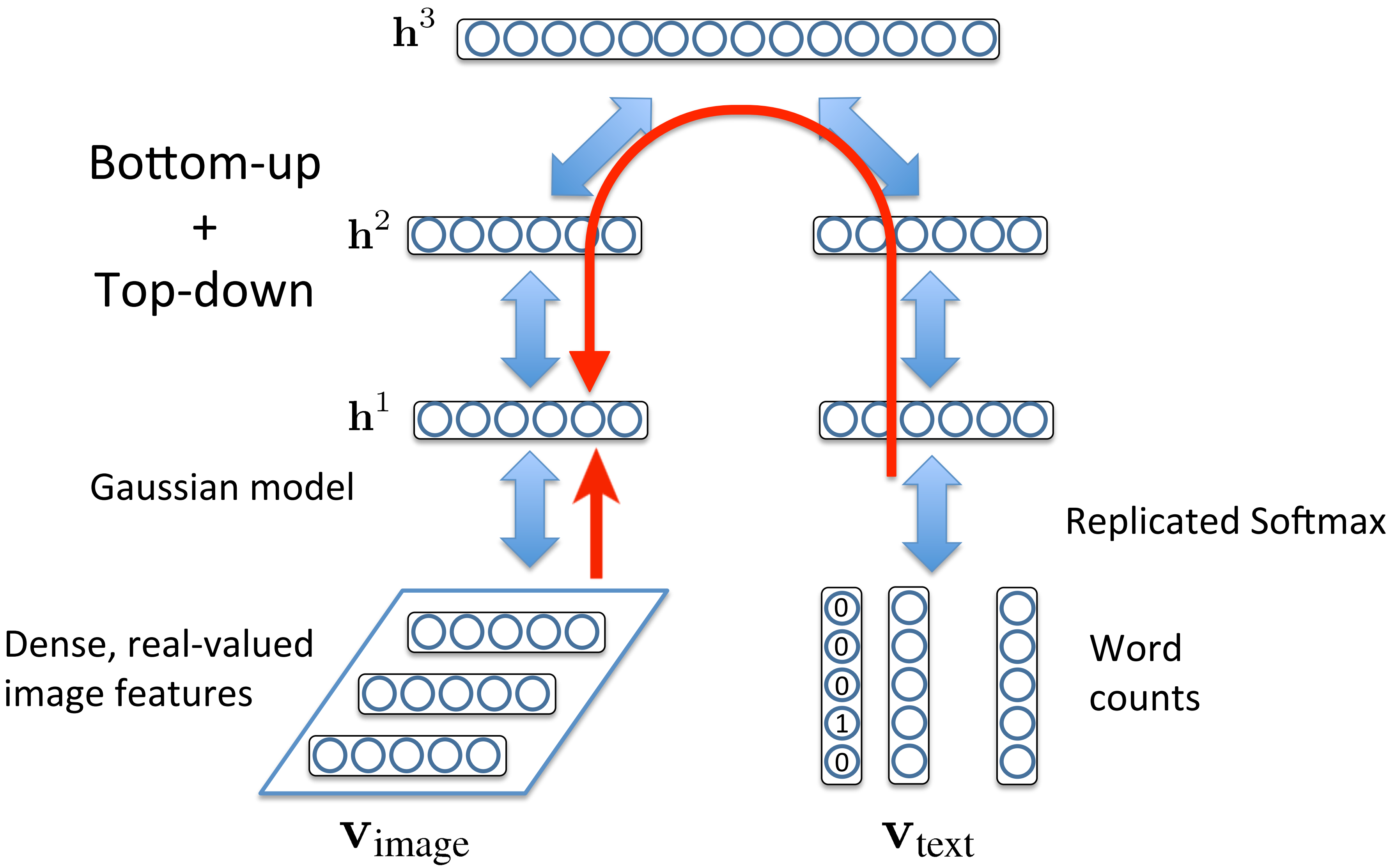
Multimodal DBM

(Srivastava & Salakhutdinov, NIPS 2012)



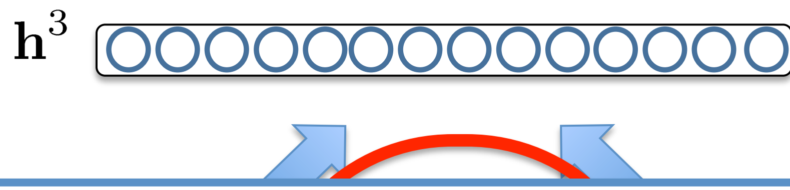
Multimodal DBM

(Srivastava & Salakhutdinov, NIPS 2012)



Multimodal DBM

(Srivastava & Salakhutdinov, NIPS 2012)



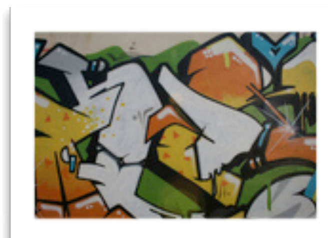
$$P(\mathbf{v}^m, \mathbf{v}^t; \theta) = \sum_{\mathbf{h}^{(2m)}, \mathbf{h}^{(2t)}, \mathbf{h}^{(3)}} P(\mathbf{h}^{(2m)}, \mathbf{h}^{(2t)}, \mathbf{h}^{(3)}) \left(\sum_{\mathbf{h}^{(1m)}} P(\mathbf{v}_m, \mathbf{h}^{(1m)} | \mathbf{h}^{(2m)}) \right) \left(\sum_{\mathbf{h}^{(1t)}} P(\mathbf{v}^t,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

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

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

Text Generated from Images

Given	Generated
	dog, cat, pet, kitten, puppy, ginger, tongue, kitty, dogs, furry
	sea, france, boat, mer, beach, river, bretagne, plage, brittany
	portrait, child, kid, ritratto, kids, children, boy, cute, boys, italy

Given	Generated
	insect, butterfly, insects, bug, butterflies, lepidoptera
	graffiti, streetart, stencil, sticker, urbanart, graff, sanfrancisco
	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

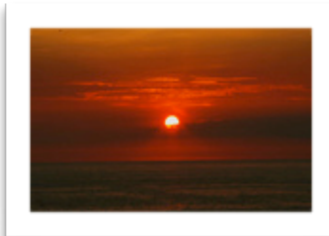
water, glass, beer, bottle, drink, wine, bubbles, splash, drops, drop

Images from Text

Given

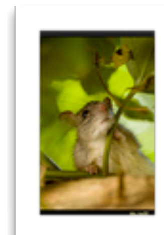
Retrieved

water, red, sunset



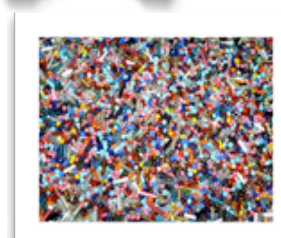
nature, flower, red, green



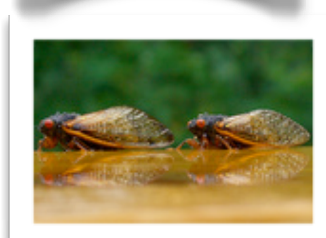



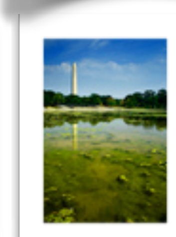
blue, green, yellow, colors



chocolate, cake



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

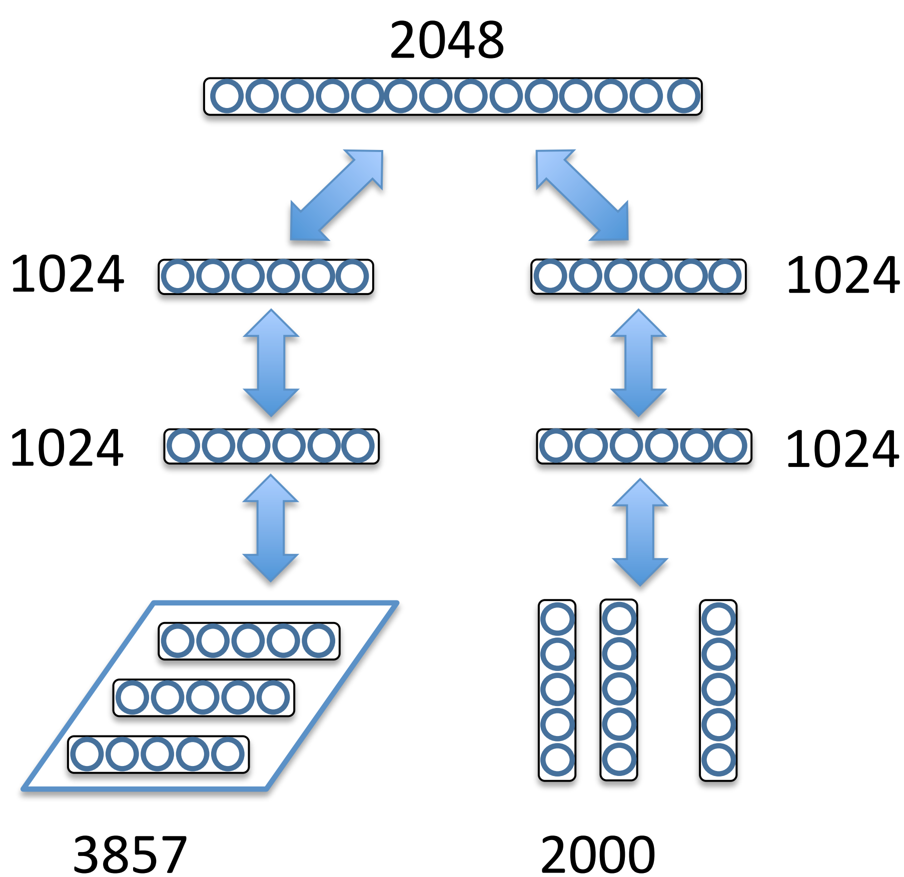
white, yellow, abstract, lines, bus, graphic

sky, geotagged, reflection, cielo, bilbao, reflejo

Huiskes et. al.

Data and Architecture

$\approx$ 12 Million parameters



- 200 most frequent tags.
- 25K labeled subset (15K training, 10K testing)
- Additional 1 million unlabeled data
- 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

Similar Features, 25K

Results

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

Mean Average Precision → MAP

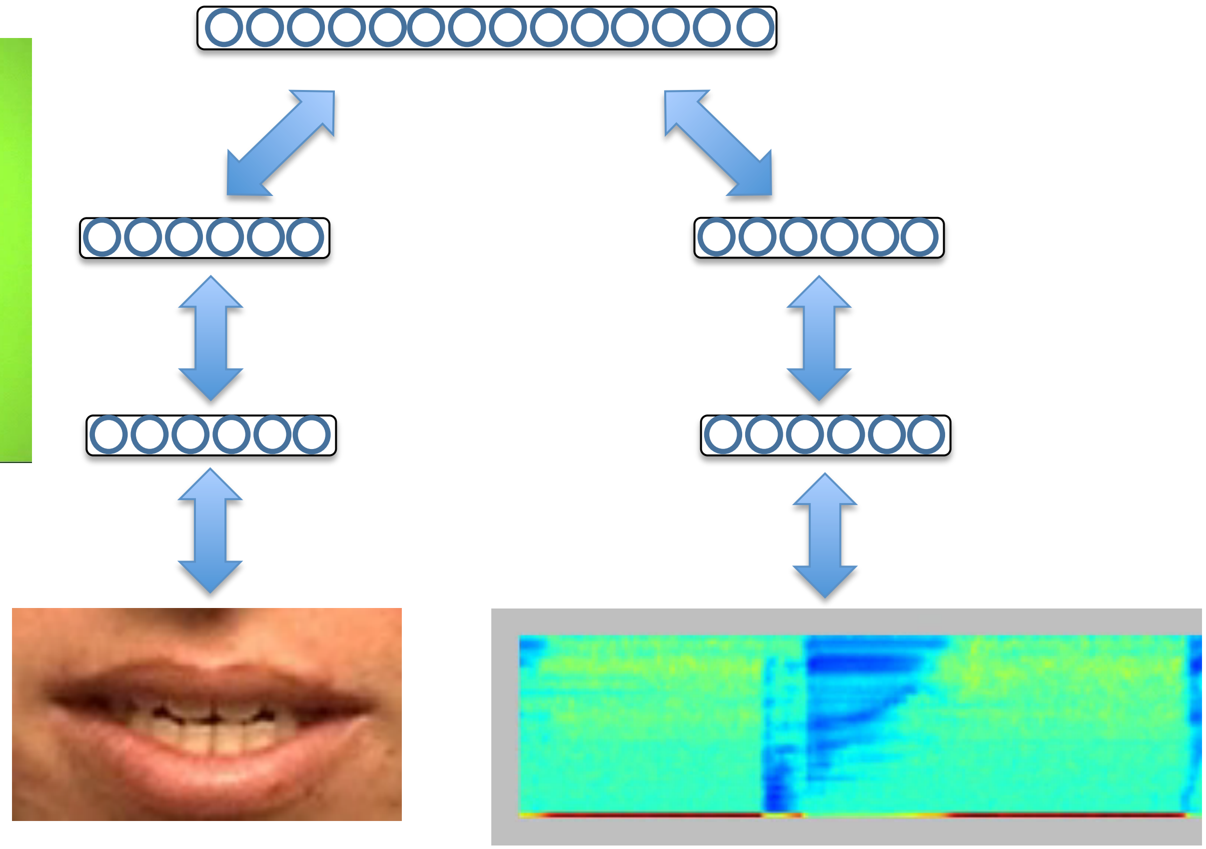
Learning Algorithm	MAP	Precision@50
Random	0.124	0.124
LDA [Huiskes et. al.]	0.492	0.754
SVM [Huiskes et. al.]	0.475	0.758
DBM-Labelled	0.526	0.791
DBM	0.609	0.863
Deep Belief Net	0.599	0.867
Autoencoder	0.600	0.875

Similar Features, 25K (LDA, SVM, DBM-Labelled)

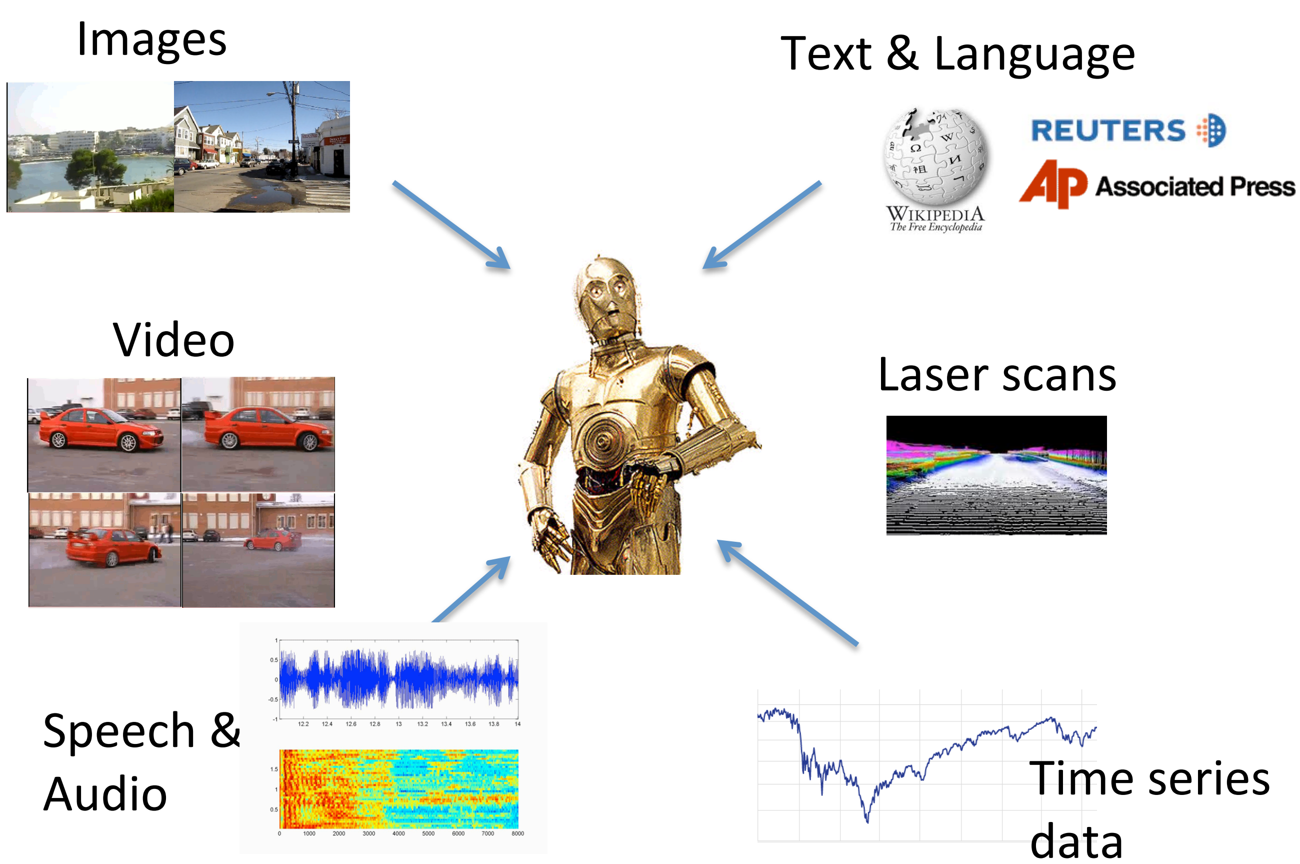
+ 1 Million Unlabelled (DBM, Deep Belief Net, Autoencoder)

Video and Audio

Cuave Dataset



Multi-Modal Models

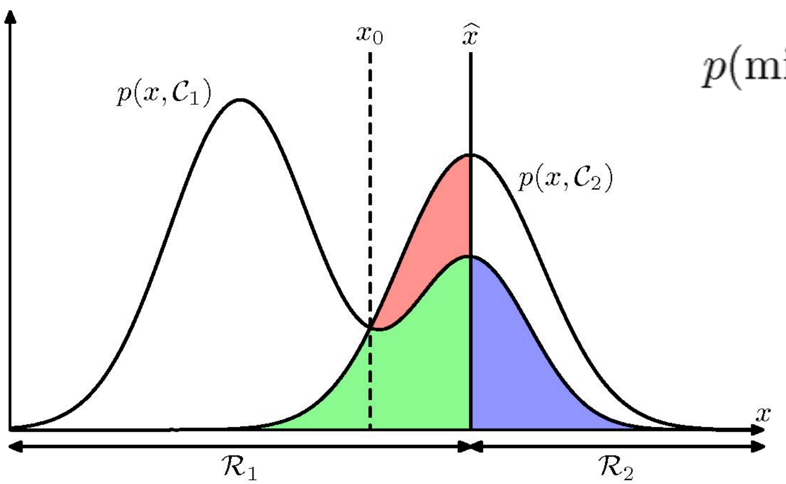


Develop learning systems that come
closer to displaying human like intelligence

One of Key Challenges:
Inference

Midterm Review

- Polynomial curve fitting – generalization, overfitting
- Decision theory:
    - Minimizing misclassification rate / Minimizing the expected loss



$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}. \end{aligned}$$

- Loss functions for regression

$$\mathbb{E}[L] = \int\int \left(t - y(\mathbf{x})\right)^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x} \mathrm{d}t.$$

Midterm Review

- Bernoulli, Multinomial random variables (mean, variances)
- Multivariate Gaussian distribution (form, mean, covariance)
- Maximum likelihood estimation for these distributions.

- Exponential family / Maximum likelihood estimation / sufficient statistics for exponential family.

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

- Linear basis function models / maximum likelihood and least squares:

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta)$$

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi).$$

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

Midterm Review

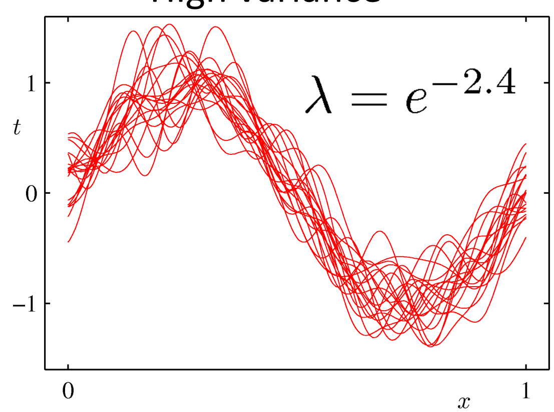
- Regularized least squares:

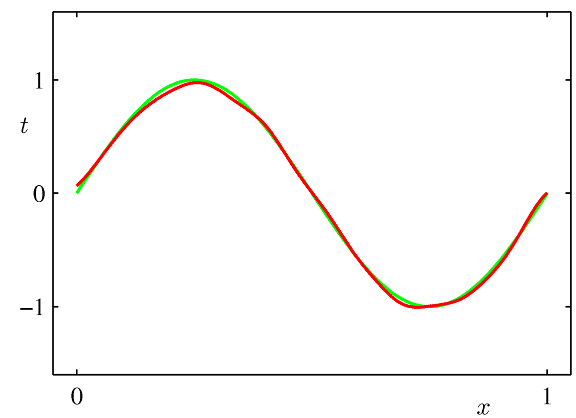
Ridge regression

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \qquad \mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$$

- Bias-variance decomposition.

High variance





Low bias

Midterm Review

- Bayesian Inference: likelihood, prior, posterior:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})P(\mathbf{w})}{P(\mathcal{D})}$$

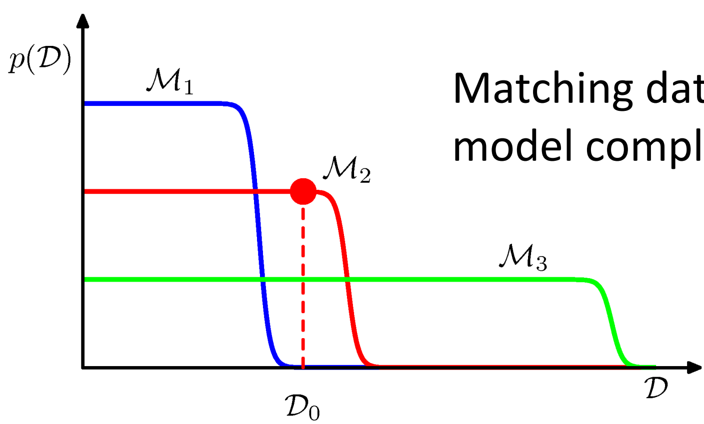
Marginal likelihood (normalizing constant):

$$P(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})P(\mathbf{w})\mathrm{d}\mathbf{w}$$

- Marginal likelihood / predictive distribution.

- Bayesian linear regression / parameter estimation / posterior distribution / predictive distribution

- Bayesian model comparison / Evidence approximation



Matching data and model complexity

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}}) + M \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right).$$

Midterm Review

- Classification models:
  - Discriminant functions
  - Fisher's linear discriminant
  - Perceptron algorithm

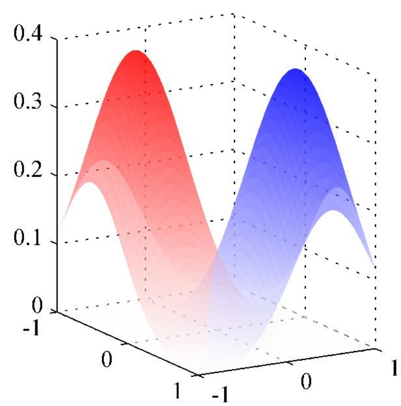
- Probabilistic Generative Models / Gaussian class conditionals / Maximum likelihood estimation:

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right).$$

$$p(\mathcal{C}_k|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0),$$

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

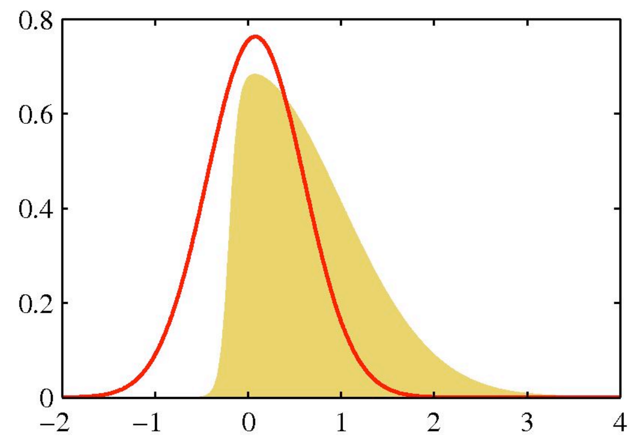
$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}.$$



Midterm Review

- Discriminative Models / Logistic regression / maximum likelihood estimation

- Laplace approximation



- BIC

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\theta_{\text{MAP}}) + \ln P(\theta_{\text{MAP}}) + \frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|A|,$$

- Bayesian logistic regression / predictive distribution

Midterm Review

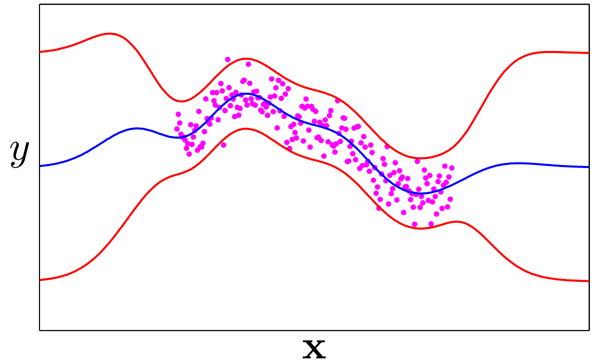
- Gaussian processes, definition:

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \right)$$

- GPs for regression.

- Marginal/predictive distributions. Making predictions using GPs.

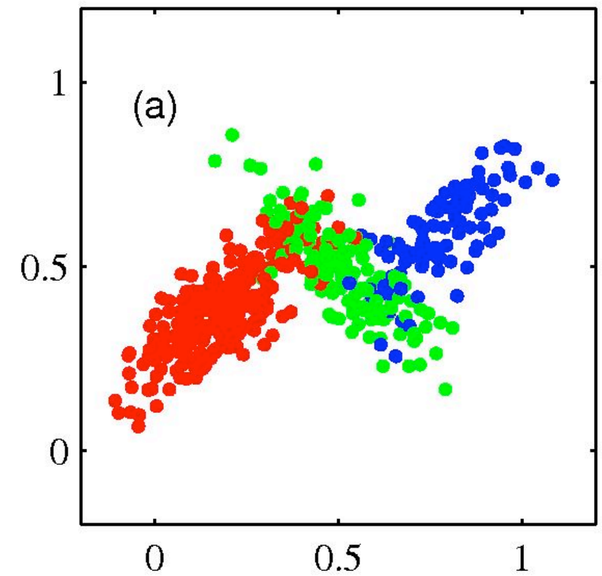
- Covariance functions, automatic relevance determination, role of hyperparameters



$$p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})}$$

Midterm Review

- Mixture Models, k-means, Mixture of Gaussians
- Mixture of Gaussians: Maximum likelihood estimation.
- EM algorithm: definition of E-step, definition of M-step, relationship to k-means.
- Alternative view of EM: expected complete data log-likelihood:



- E-step: Compute posterior over latent variables: p(Z|X,θ^{old}).
- M-step: Find the new estimate of parameters θ^{new}:

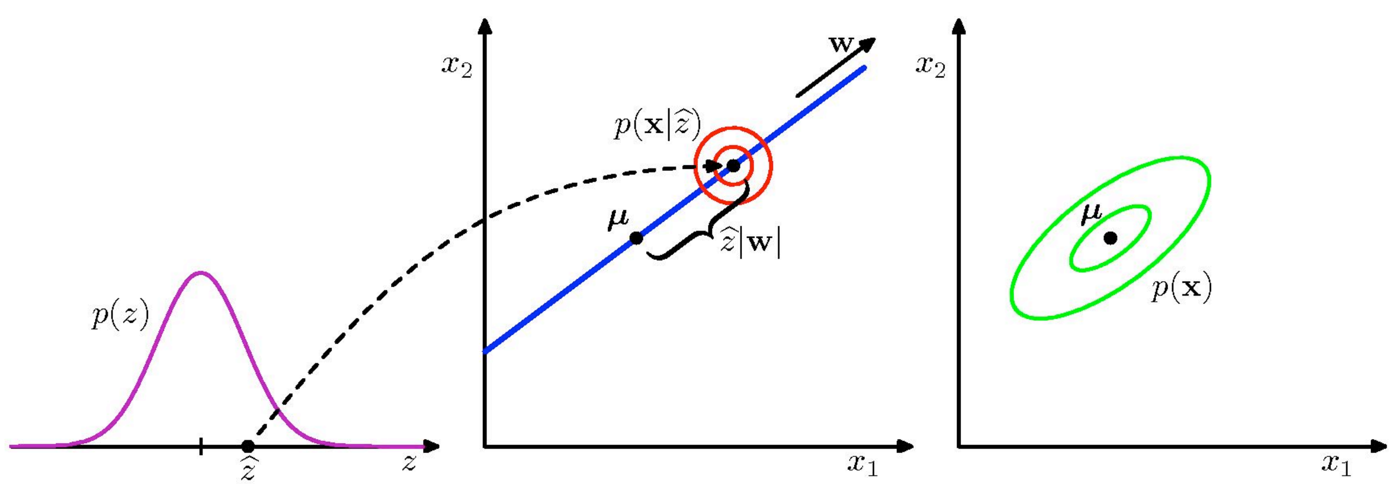
$$\theta^{new} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}).$$

where

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

Midterm Review

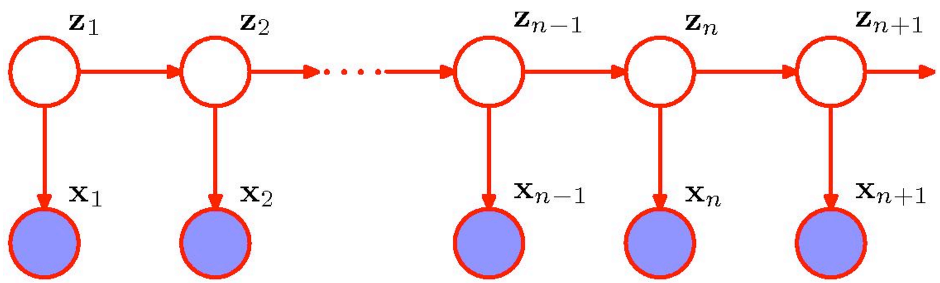
- Continuous latent variable models: Probabilistic PCA, Factor Analysis
- PCA, PCA for high-dimensional data
- Probabilistic PCA: definition of probabilistic model, Joint/Marginal density, posterior over latent variables, relationship to standard PCA
- Probabilistic PCA: Maximum likelihood estimation, zero noise limit.
- Factor analysis, definition, marginal/joint/posterior. Relationship to PPCA.
- Autoencoders: definition



Midterm Review

• Sequential data: Markov models, maximum likelihood estimation

• State Space models: definition, transition model, observation model.



• Hidden Markov models: definition, transition model, observation model.

• Maximum likelihood estimation for HMMs, basics of EM algorithm.

• Basics of EM algorithm for HMMs: interring posterior over latent paths and parameter estimation for the transition and observation model.

• Dynamic programming (understanding of alpha-beta recursions)

• Viterbi decoding.