Deep Learning II

Russ Salakhutdinov

KDD Tutorial

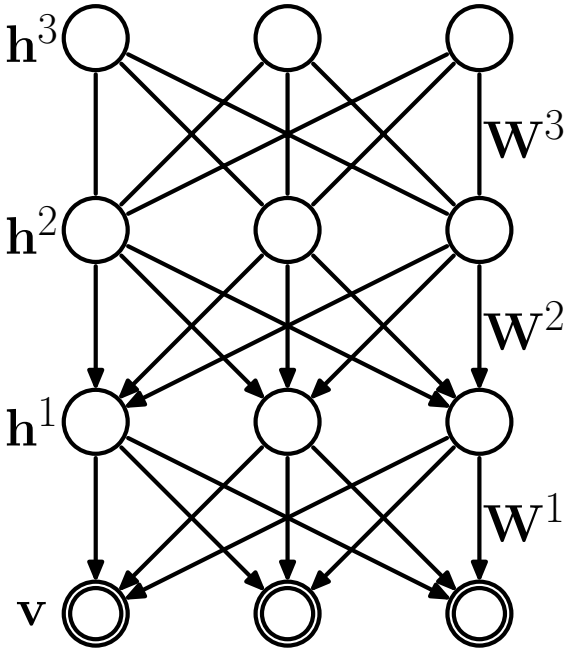
Department of Computer Science
Department of Statistics
University of Toronto
Canadian Institute for Advanced Research

Talk Roadmap

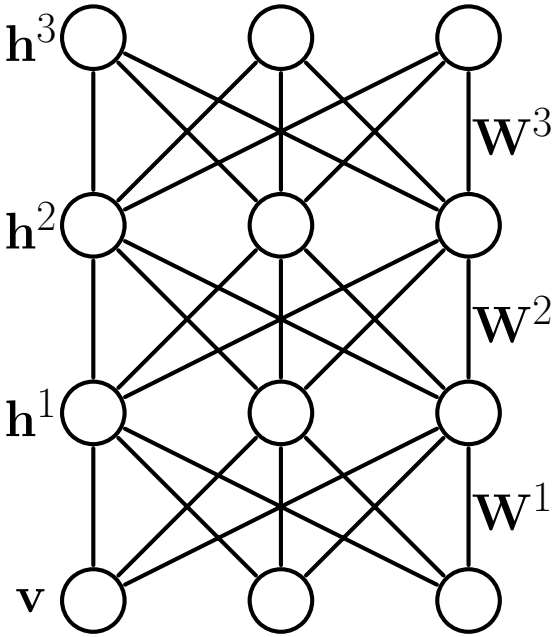
- Advanced Deep Models
  - Deep Boltzmann Machines
  - Learning Structured and Robust Deep Models
  - One-Shot and Transfer Learning
- Multimodal Learning
- Conclusions

DBNs vs. DBMs

Deep Belief Network



Deep Boltzmann Machine



DBNs are hybrid models:

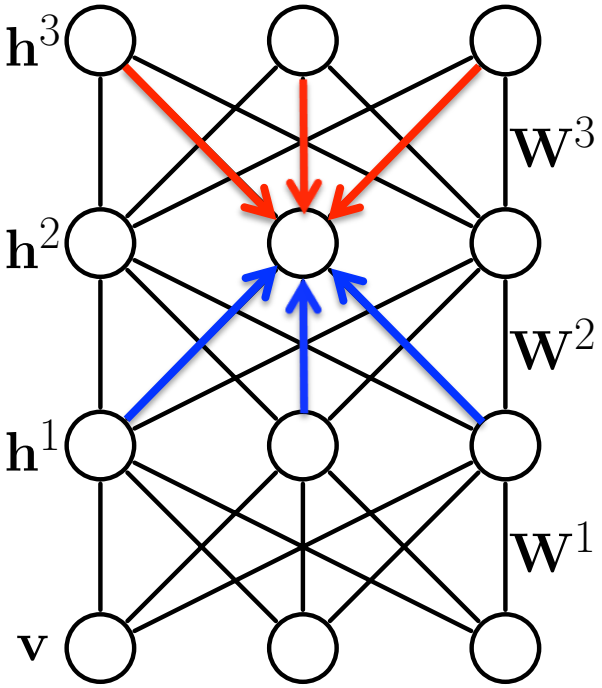
- Inference in DBNs is problematic due to **explaining away**.
- Only greedy pretrainig, **no joint optimization over all layers**.
- Approximate inference is feed-forward: **no bottom-up and top-down**.

Introduce a new class of models called Deep Boltzmann Machines.

Mathematical Formulation

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

Deep Boltzmann Machine



Input

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

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

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

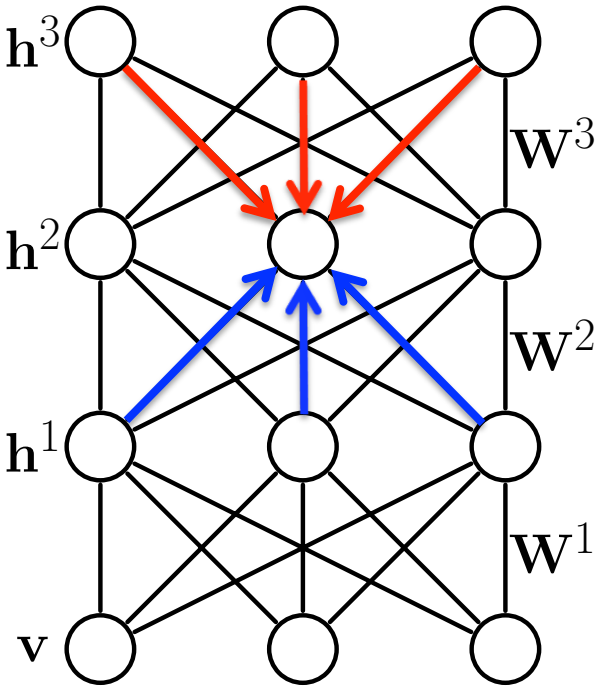
Bottom-up Top-Down

Unlike many existing feed-forward models: ConvNet (LeCun), HMAX (Poggio et.al.), Deep Belief Nets (Hinton et.al.)

Mathematical Formulation

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

Deep Boltzmann Machine



- Conditional Distributions:

$$P(h_j^1 = 1|\mathbf{v}, \mathbf{h}^2) = \sigma\left(\sum_i W_{ij}^1 v_i + \sum_k W_{jk}^2 h_k^2\right)$$

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

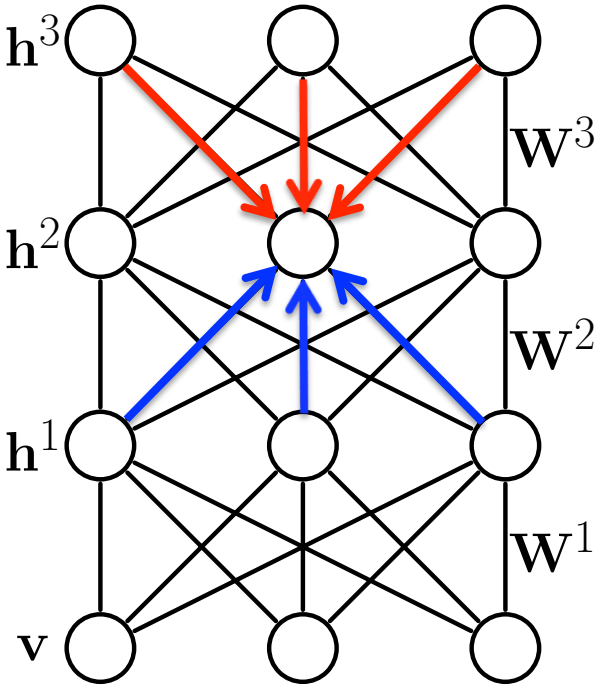
$$P(h_m^3 = 1|\mathbf{h}^2) = \sigma\left(\sum_k W_{km}^3 h_k^2\right)$$

- Note that exact computation of $P(\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3|\mathbf{v})$ is intractable.

Mathematical Formulation

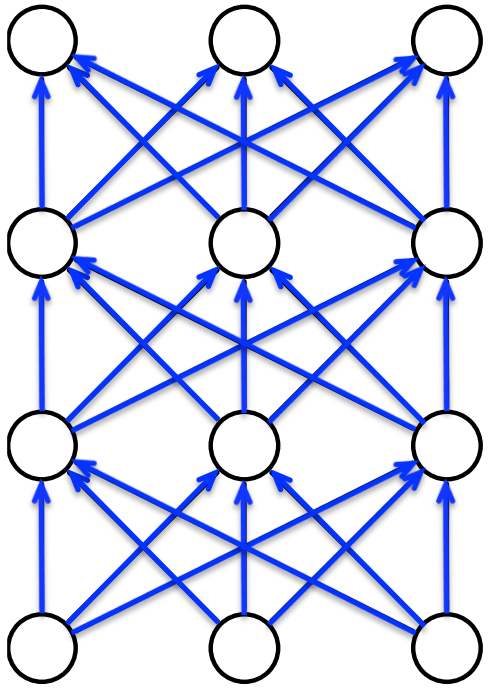
$$P_\theta(\mathbf{v}) = \frac{P^*(\mathbf{v})}{\mathcal{Z}(\theta)} = \frac{1}{\mathcal{Z}(\theta)} \sum_{\mathbf{h}^1,\mathbf{h}^2,\mathbf{h}^3} \exp\left[\mathbf{v}^\top W^1 \mathbf{h}^1 + {\mathbf{h}^1}^\top W^2 \mathbf{h}^2 + {\mathbf{h}^2}^\top W^3 \mathbf{h}^3\right]$$

Deep Boltzmann Machine

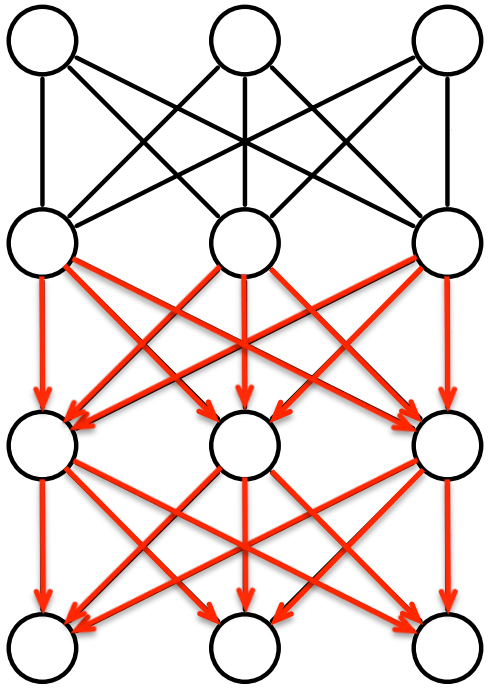


Input

Neural Network Output



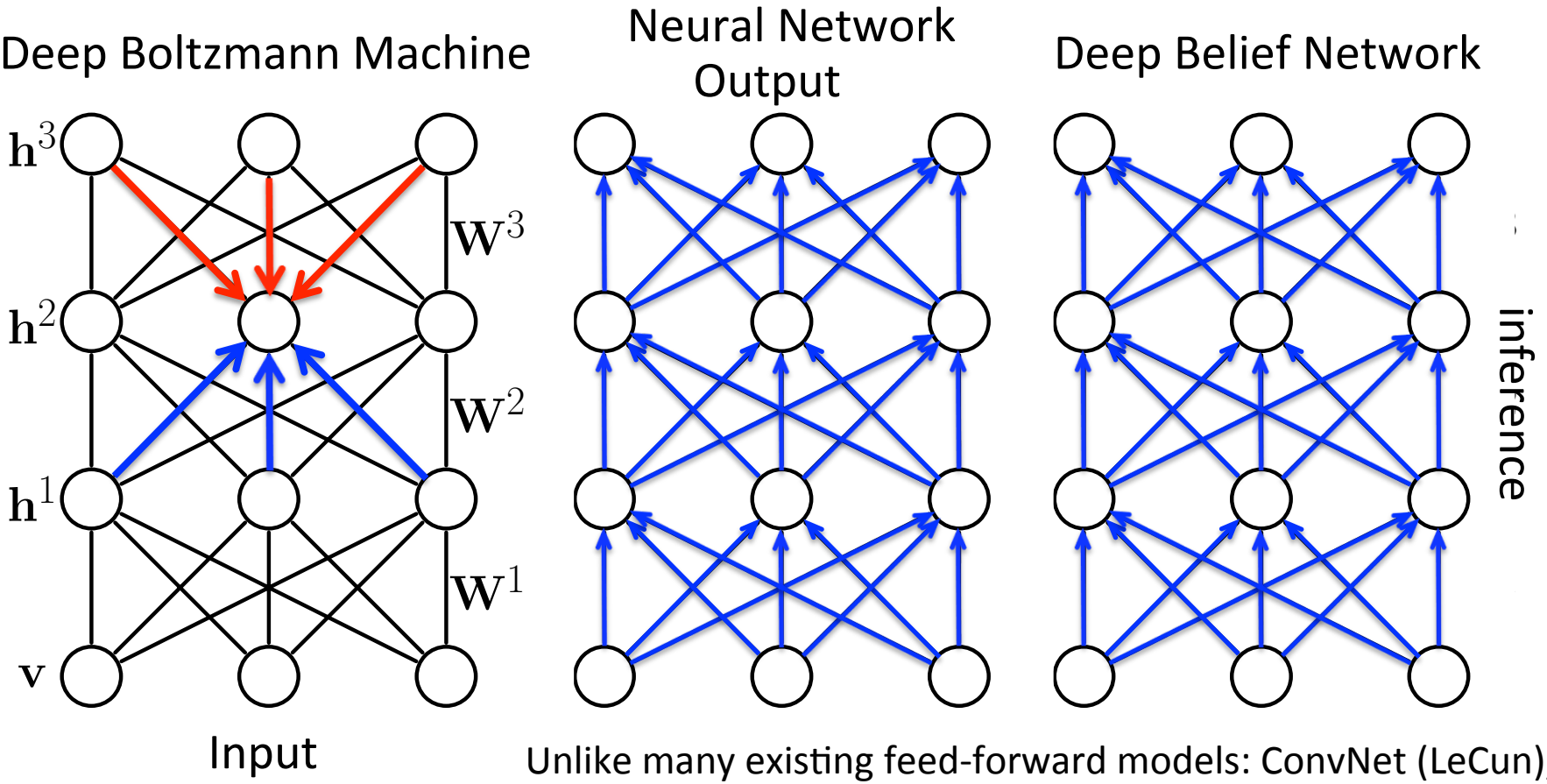
Deep Belief Network



Unlike many existing feed-forward models: ConvNet (LeCun), HMAX (Poggio), Deep Belief Nets (Hinton)

Mathematical Formulation

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

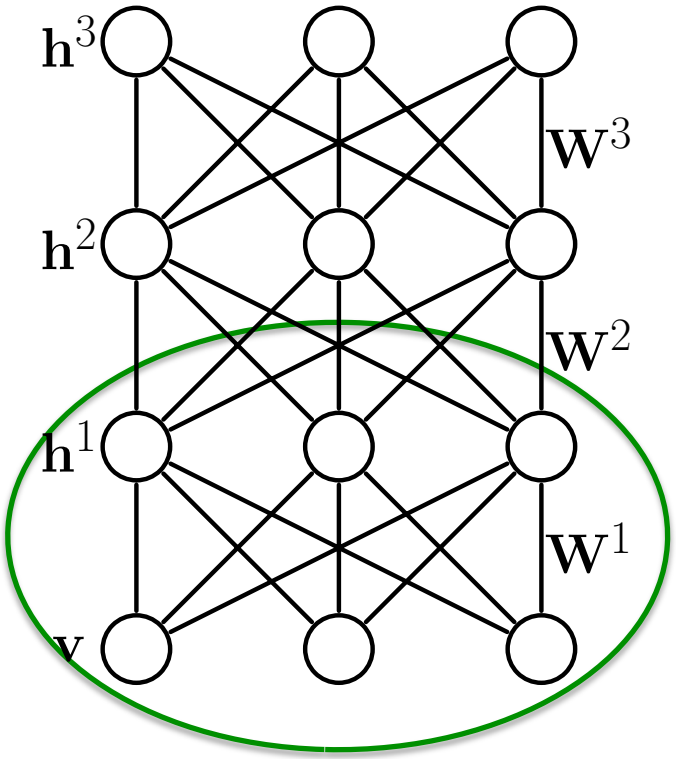


Unlike many existing feed-forward models: ConvNet (LeCun), HMAX (Poggio), Deep Belief Nets (Hinton)

Mathematical Formulation

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

Deep Boltzmann Machine



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

- Dependencies between hidden variables.

Maximum likelihood learning:

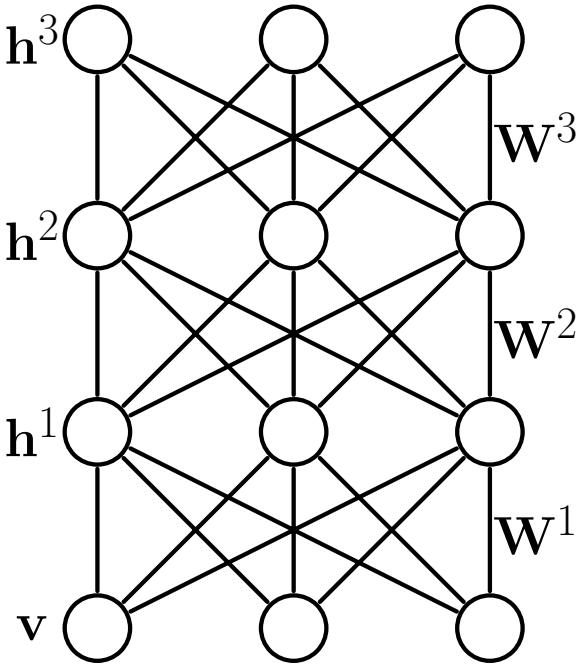
$$\frac{\partial \log P_\theta(\mathbf{v})}{\partial W^1} = \mathrm{E}_{P_{data}}[\mathbf{v}\mathbf{h}^{1\top}] - \mathrm{E}_{P_\theta}[\mathbf{v}\mathbf{h}^{1\top}]$$

Problem: Both expectations are intractable!

Learning rule for undirected graphical models: MRFs, CRFs, Factor graphs.

Approximate Learning

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



(Approximate) Maximum Likelihood:

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

- Both expectations are intractable!

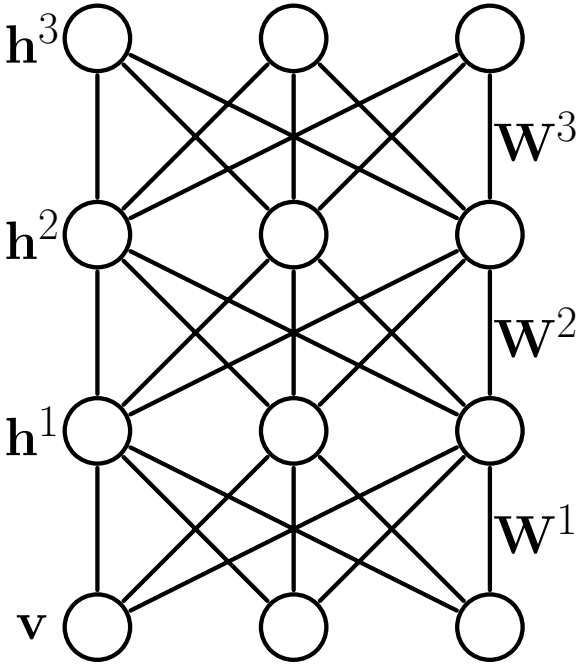
$$P_{data}(\mathbf{v}, \mathbf{h^1}) = P_\theta(\mathbf{h^1}|\mathbf{v}) P_{data}(\mathbf{v})$$

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

Not factorial any more!

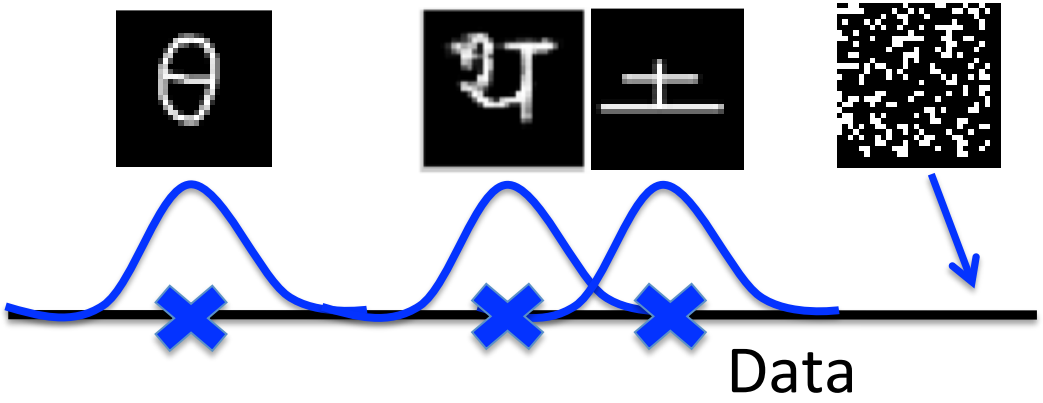
Approximate Learning

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



(Approximate) Maximum Likelihood:

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



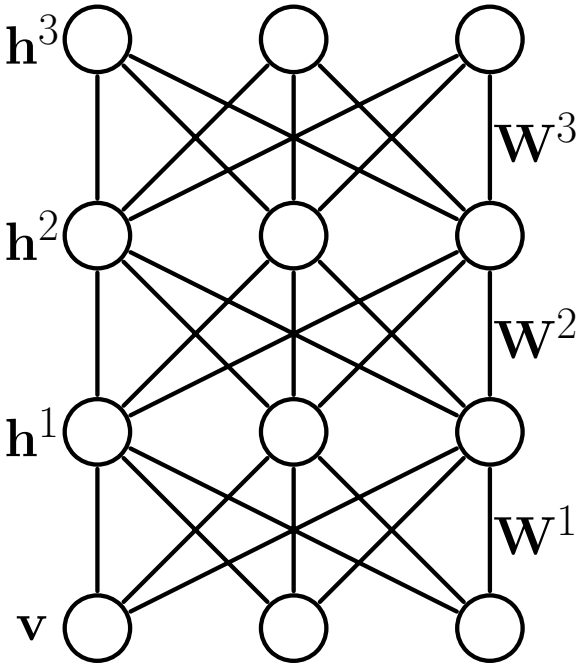
$$P_{data}(\mathbf{v}, \mathbf{h^1}) = P_\theta(\mathbf{h^1}|\mathbf{v}) P_{data}(\mathbf{v})$$

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

Not factorial any more!

Approximate Learning

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



(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 (MCMC-based)

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

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

Not factorial any more!

Previous Work

Many approaches for learning Boltzmann machines have been proposed over the last 20 years:

- Hinton and Sejnowski (1983),
- Peterson and Anderson (1987)
- Galland (1991)
- Kappen and Rodriguez (1998)
- Lawrence, Bishop, and Jordan (1998)
- Tanaka (1998)
- Welling and Hinton (2002)
- Zhu and Liu (2002)
- Welling and Teh (2003)
- Yasuda and Tanaka (2009)

Real-world applications – thousands of hidden and observed variables with millions of parameters.

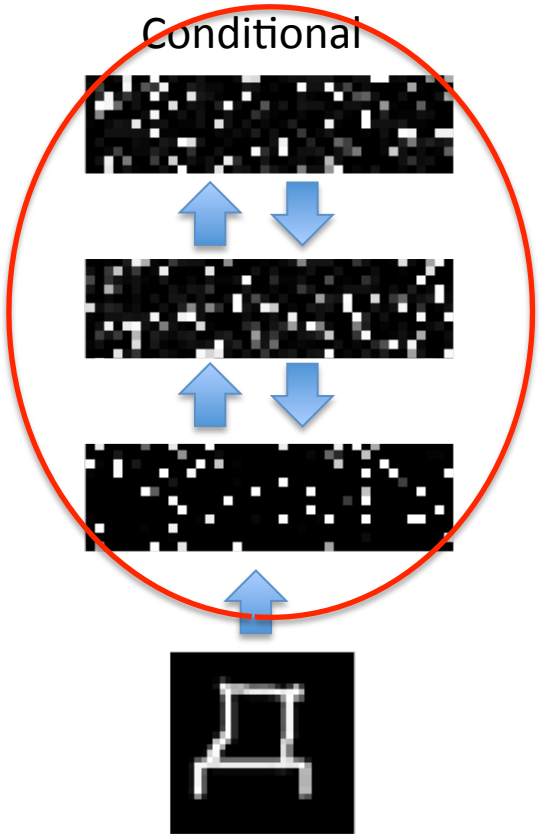
Many of the previous approaches were not successful for learning general Boltzmann machines with **hidden variables.**

Algorithms based on Contrastive Divergence, Score Matching, Pseudo-Likelihood, Composite Likelihood, MCMC-MLE, Piecewise Learning, cannot handle multiple layers of hidden variables.

New Learning Algorithm

Posterior Inference

Conditional



Approximate conditional

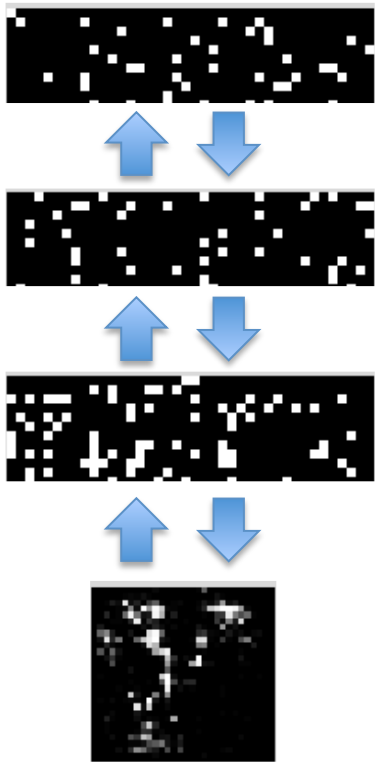
$P_{data}(\mathbf{h}|\mathbf{v})$

Simulate from the Model

Approximate the joint distribution

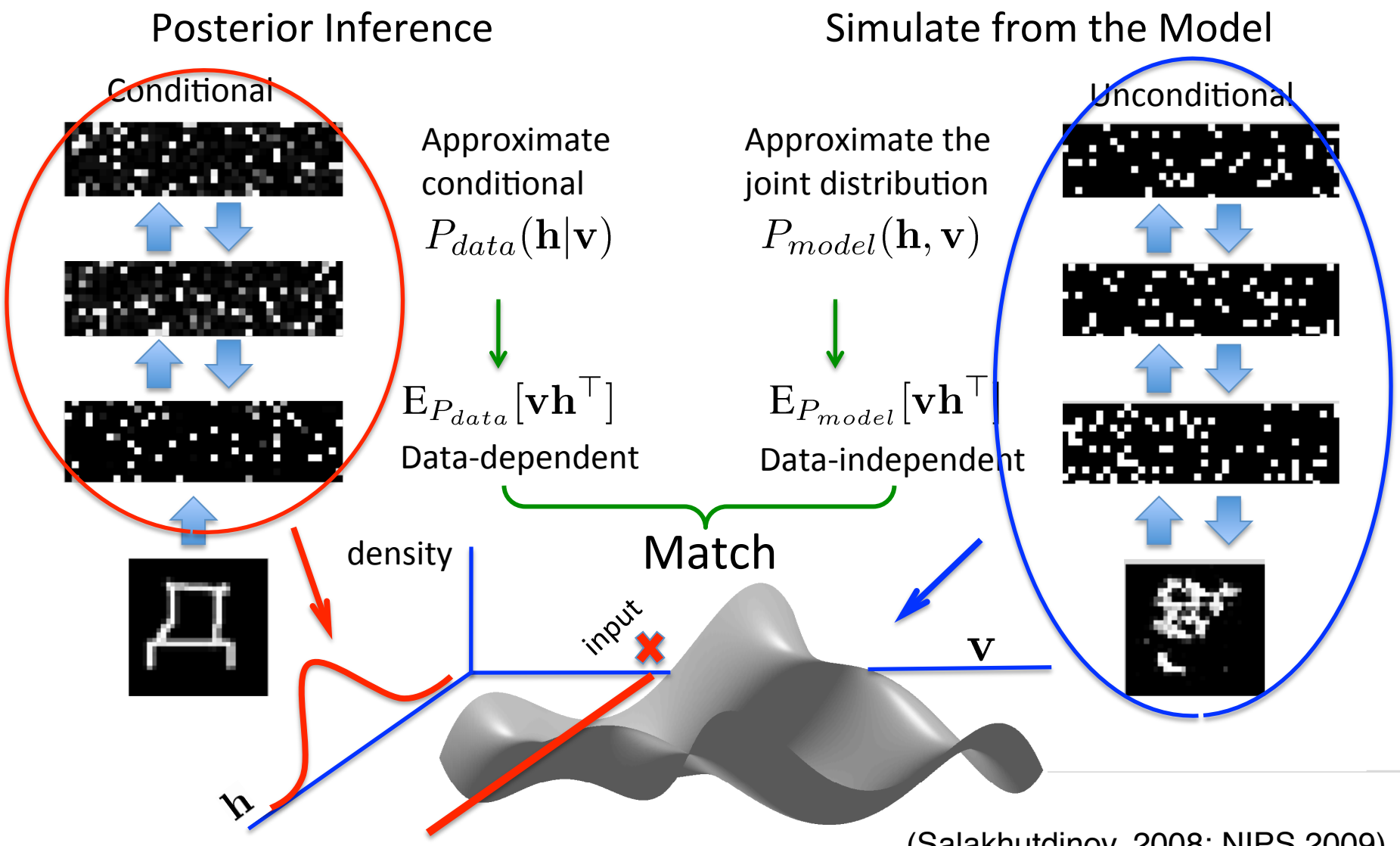
$P_{model}(\mathbf{h}, \mathbf{v})$

Unconditional



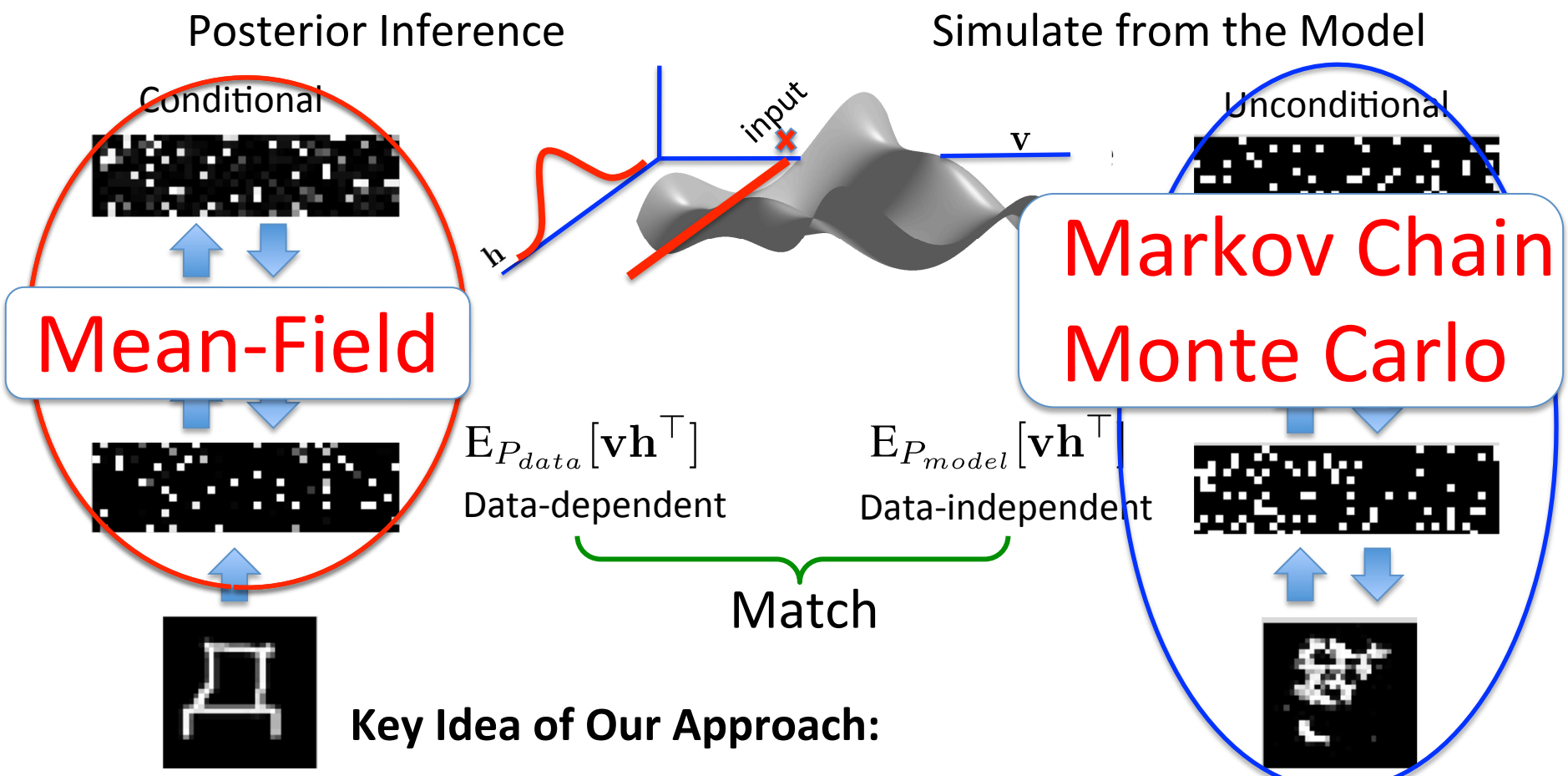
(Salakhutdinov, 2008; NIPS 2009)

New Learning Algorithm



(Salakhutdinov, 2008; NIPS 2009)

New Learning Algorithm



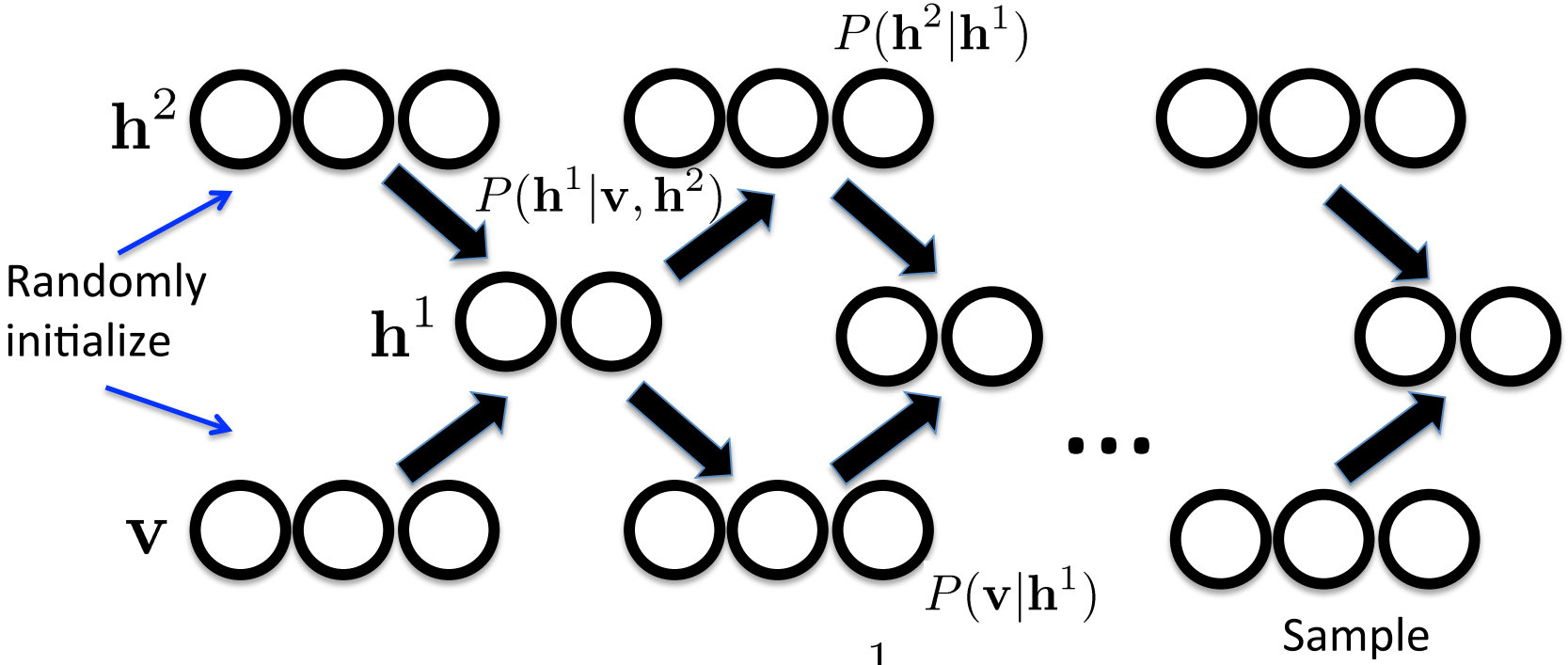
Key Idea of Our Approach:

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

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

Sampling from DBMs

Sampling from two-hidden layer DBM by running a Markov chain:

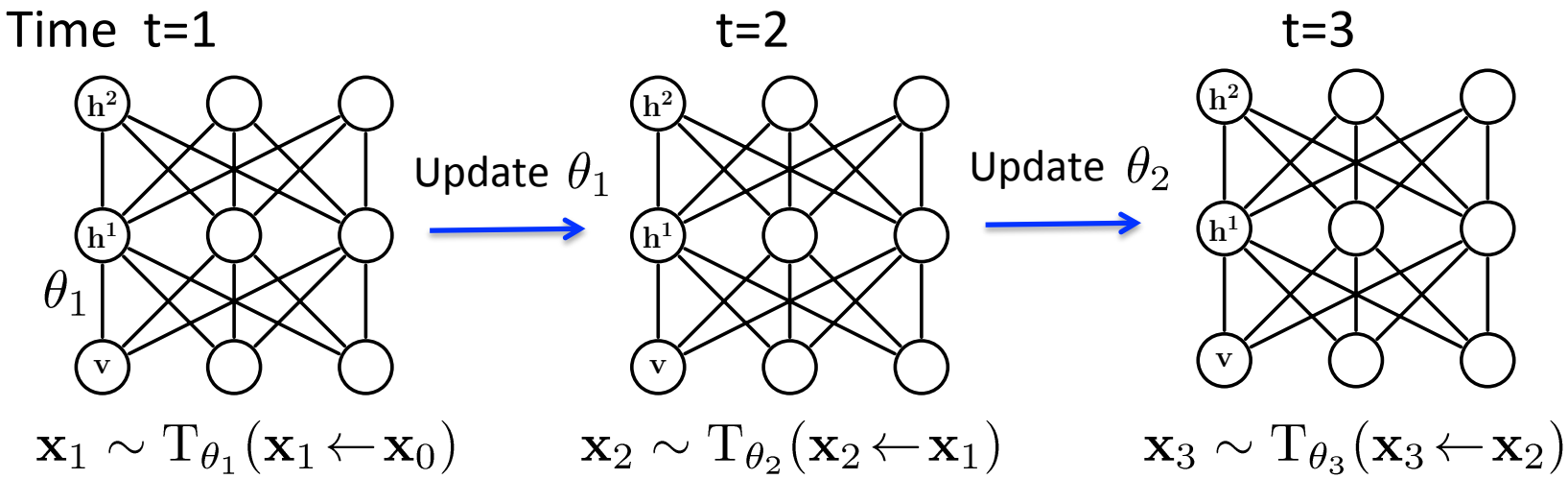


$$P(h_m^1 = 1|\mathbf{v},\mathbf{h}^2) = \frac{1}{1+\exp(-\sum_i W_{im}^1 v_i - \sum_j W_{mj}^2 h_j^2)}$$

$$P(h_j^2 = 1|\mathbf{h}^1) = \frac{1}{1+\exp(-\sum_m W_{mj}^2 h_m^1)}$$

$$P(v_i = 1|\mathbf{h}^1) = \frac{1}{1+\exp(-\sum_m W_{im}^1 h_m^1)}$$

Stochastic Approximation



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:

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

Variational Inference

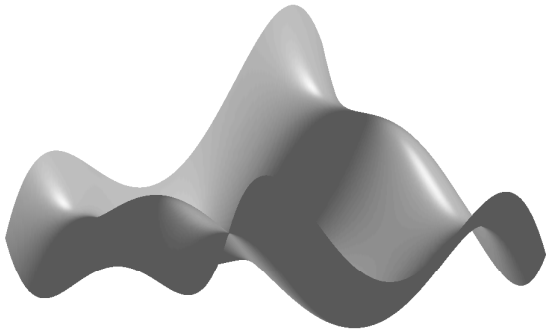
MCMC

Learning Algorithm

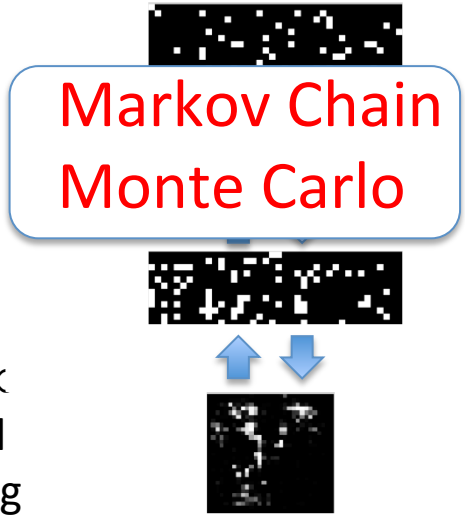
Update rule decomposes:

$$\theta_{t+1} = \theta_t + \alpha_t \underbrace{\left(\mathbb{E}_{P_{data}}[\mathbf{v}\mathbf{h}^\top] - \mathbb{E}_{P_{\theta_t}}[\mathbf{v}\mathbf{h}^\top] \right)}_{\text{True gradient}} + \alpha_t \underbrace{\left(\mathbb{E}_{P_{\theta_t}}[\mathbf{v}\mathbf{h}^\top] - \frac{1}{M} \sum_{m=1}^{M} \mathbf{v}_t^{(m)} \mathbf{h}_t^{(m)\top} \right)}_{\text{Perturbation term } \epsilon_t}$$

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



Problem: High-dimensional data: the probability landscape is highly multimodal.



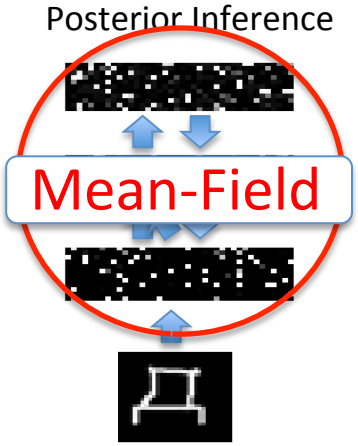
Key insight: The transition operator can b any valid transition operator – Tempered Transitions, Parallel/Simulated Tempering

Connections to the theory of stochastic approximation and adaptive MCMC.

Variational Inference

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

$$\log P_\theta(\mathbf{v}) = \log \sum_{\mathbf{h}} P_\theta(\mathbf{h}, \mathbf{v}) = \log \sum_{\mathbf{h}} Q_\mu(\mathbf{h}|\mathbf{v}) \frac{P_\theta(\mathbf{h}, \mathbf{v})}{Q_\mu(\mathbf{h}|\mathbf{v})}$$



$$\geq \sum_{\mathbf{h}} Q_\mu(\mathbf{h}|\mathbf{v}) \log \frac{P_\theta(\mathbf{h}, \mathbf{v})}{Q_\mu(\mathbf{h}|\mathbf{v})}$$

$$= \sum_{\mathbf{h}} Q_\mu(\mathbf{h}|\mathbf{v}) \underbrace{\log P^*_\theta(\mathbf{h}, \mathbf{v})}_{\mathbf{v}^\top W^1 \mathbf{h}^1 + {\mathbf{h}^1}^\top W^2 \mathbf{h}^2 + {\mathbf{h}^2}^\top W^3 \mathbf{h}^3} - \log \mathcal{Z}(\theta) + \sum_{\mathbf{h}} Q_\mu(\mathbf{h}|\mathbf{v}) \log \frac{1}{Q_\mu(\mathbf{h}|\mathbf{v})}$$

Variational Lower Bound

$$= \log P_\theta(\mathbf{v}) - \mathrm{KL}\big(Q_\mu(\mathbf{h}|\mathbf{v}) || P_\theta(\mathbf{h}|\mathbf{v})\big)$$

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

Minimize KL between approximating and true distributions with respect to variational parameters μ .

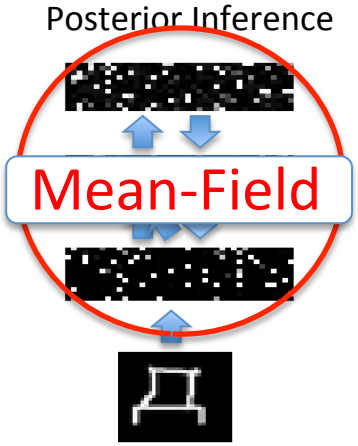
(Salakhutdinov, 2008; Salakhutdinov & Larochelle, AI & Statistics 2010)

Variational Inference

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

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

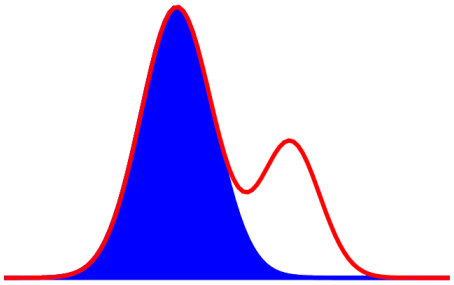
$$\log P_\theta(\mathbf{v}) \geq \underbrace{\log P_\theta(\mathbf{v}) - \mathrm{KL}\big(Q_\mu(\mathbf{h}|\mathbf{v})||P_\theta(\mathbf{h}|\mathbf{v})\big)}_{\text{Variational Lower Bound}}$$



Mean-Field: Choose a fully factorized distribution:

$$Q_\mu(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{F} q(h_j|\mathbf{v}) \text{ with } q(h_j = 1|\mathbf{v}) = \mu_j$$

Variational Inference: Maximize the lower bound w.r.t. Variational parameters μ.



Nonlinear fixed-point equations:

$$\mu_j^{(1)} = \sigma\Big(\sum_i W_{ij}^1 v_i + \sum_k W_{jk}^2 \mu_k^{(2)}\Big)$$
$$\mu_k^{(2)} = \sigma\Big(\sum_j W_{jk}^2 \mu_j^{(1)} + \sum_m W_{km}^3 \mu_m^{(3)}\Big)$$
$$\mu_m^{(3)} = \sigma\Big(\sum_k W_{km}^3 \mu_k^{(2)}\Big)$$

Variational Inference

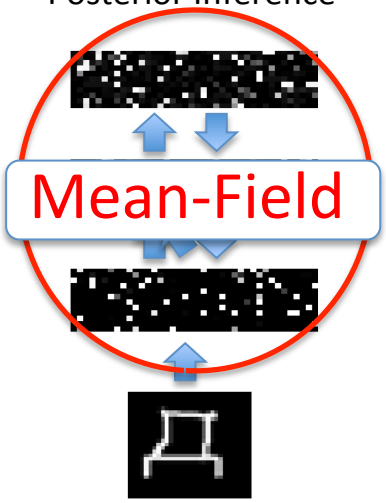
Approximate intractable distribution $P_\theta(\mathbf{h}|\mathbf{v})$ with simpler, tractable distribution $Q_\mu(\mathbf{h}|\mathbf{v})$:

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

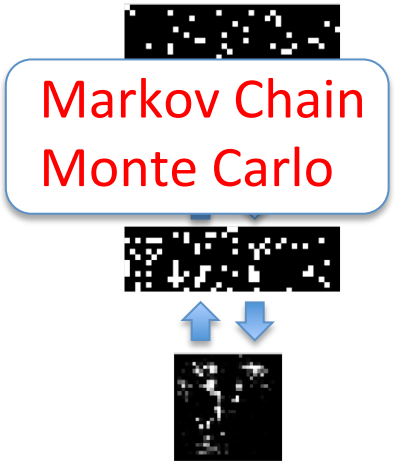
$$\log P_\theta(\mathbf{v}) \geq \log P_\theta(\mathbf{v}) - \mathrm{KL}\big(Q_\mu(\mathbf{h}|\mathbf{v})||P_\theta(\mathbf{h}|\mathbf{v})\big)$$

Variational Lower Bound

Posterior Inference



Unconditional Simulation



1. Variational Inference: Maximize the lower bound w.r.t. variational parameters

2. MCMC: Apply stochastic approximation to update model parameters

Almost sure convergence guarantees to an asymptotically stable point.

Variational Inference

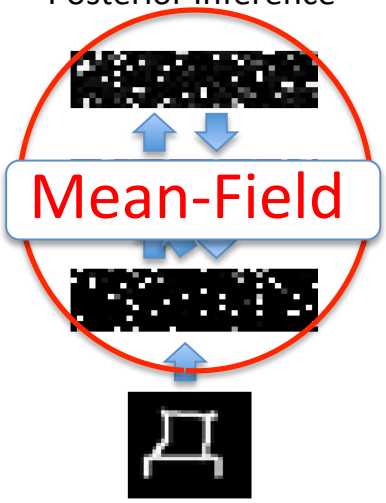
Approximate intractable distribution $P_\theta(\mathbf{h}|\mathbf{v})$ with simpler, tractable distribution $Q_\mu(\mathbf{h}|\mathbf{v})$:

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

$$\log P_\theta(\mathbf{v}) \geq \log P_\theta(\mathbf{v}) - \mathrm{KL}\big(Q_\mu(\mathbf{h}|\mathbf{v})||P_\theta(\mathbf{h}|\mathbf{v})\big)$$

Variational Lower Bound

Posterior Inference



Unconditional Simulation

1. V wer bou

Fast Inference

2. M to u

Learning can scale to millions of examples

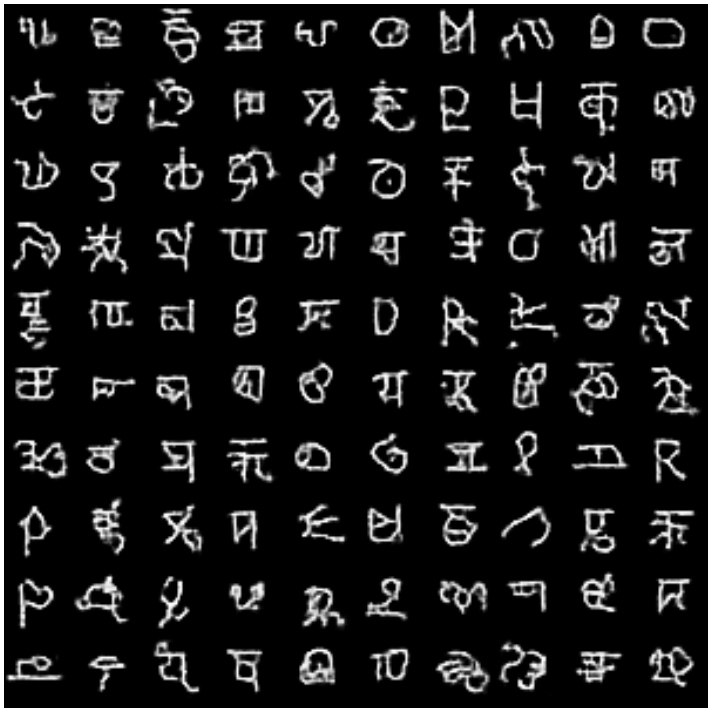
Almost sure convergence guarantees to an asymptotically stable point.

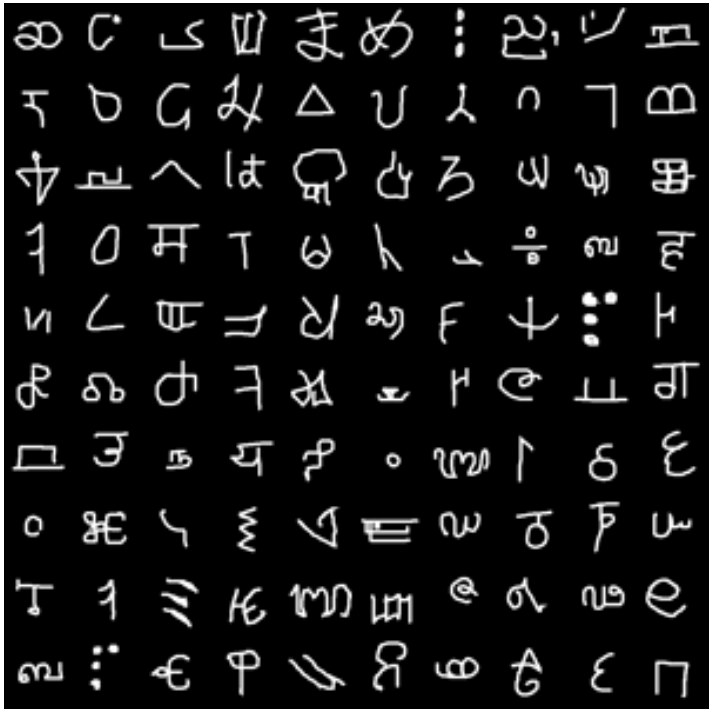
Good Generative Model?

Handwritten Characters

Good Generative Model?

Handwritten Characters





Good Generative Model?

Handwritten Characters

Simulated

Real Data

Good Generative Model?

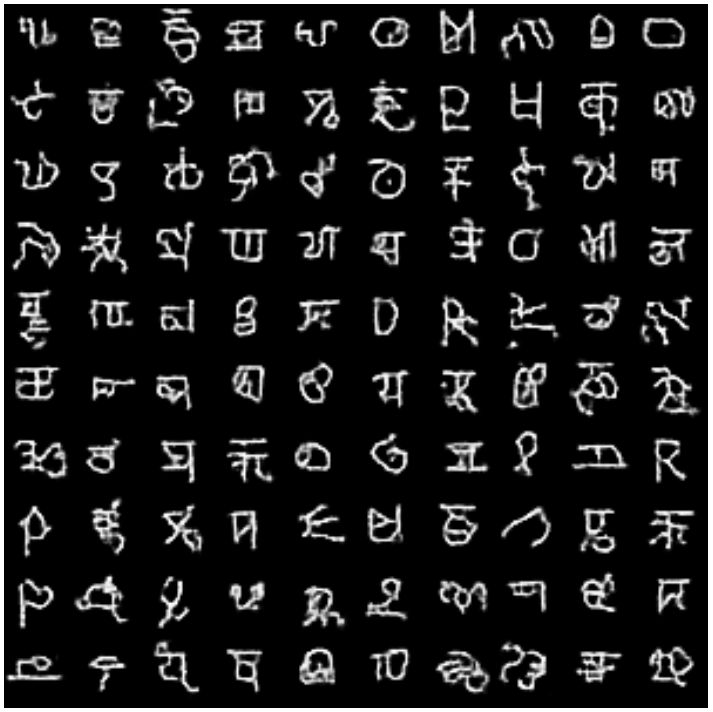
Handwritten Characters

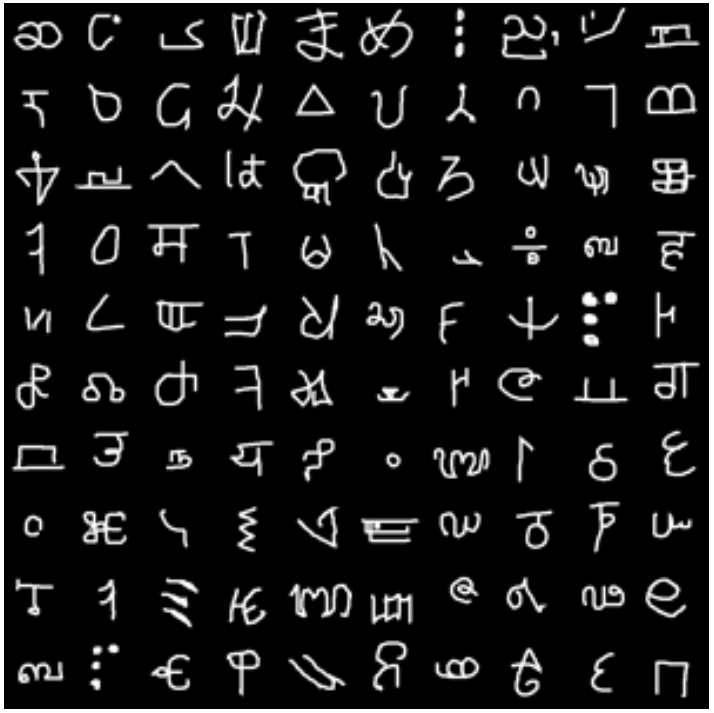
Real Data

Simulated

Good Generative Model?

Handwritten Characters





Good Generative Model?

MNIST Handwritten Digit Dataset

Handwriting Recognition

MNIST Dataset
60,000 examples of 10 digits

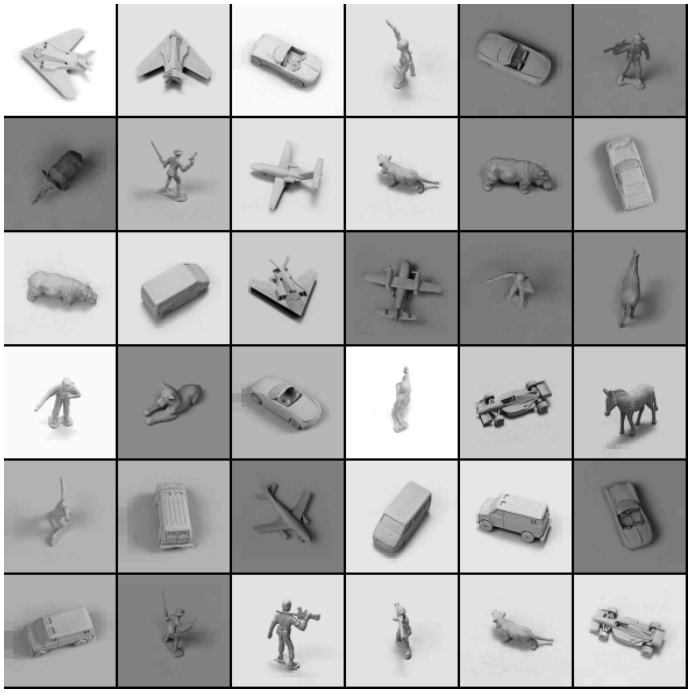
Learning Algorithm	Error
Logistic regression	12.0%
K-NN	3.09%
Neural Net (Platt 2005)	1.53%
SVM (Decoste et.al. 2002)	1.40%
Deep Autoencoder (Bengio et. al. 2007)	1.40%
Deep Belief Net (Hinton et. al. 2006)	1.20%
DBM	**0.95%**

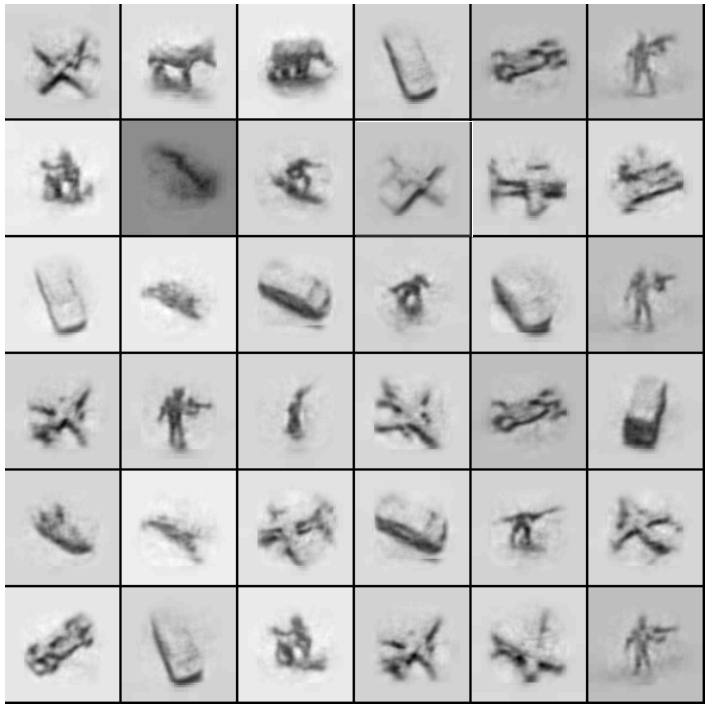
Optical Character Recognition
42,152 examples of 26 English letters

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



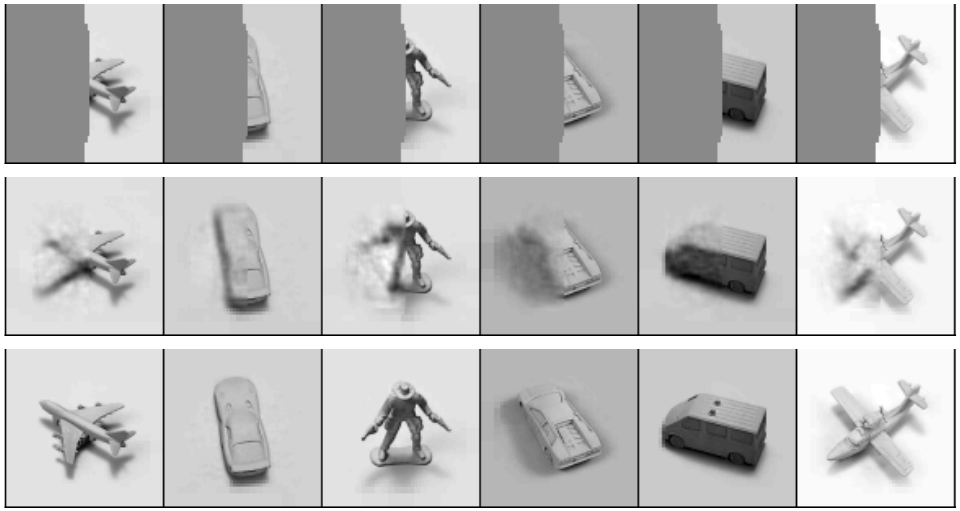


24,000 examples, 5 object categories, 5 different objects within each category, 6 lightning conditions, 9 elevations, 18 azimuths.

3-D Object Recognition

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

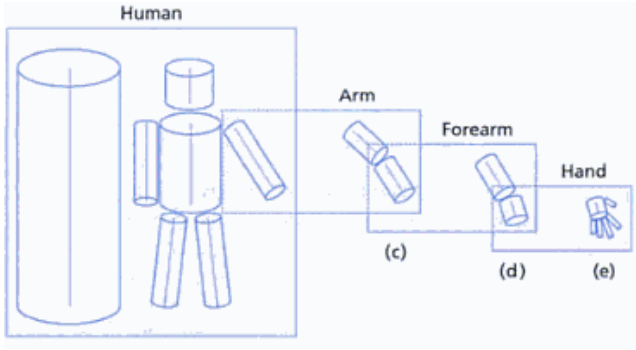


Permutation-invariant version.

Learning Hierarchical Representations

Deep Boltzmann Machines:

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

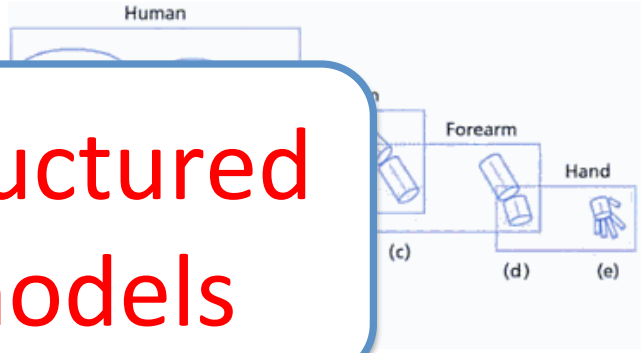


- Performs well in many application domains
- Fast Inference: fraction of a second
- Learning scales to millions of examples

Learning Hierarchical Representations

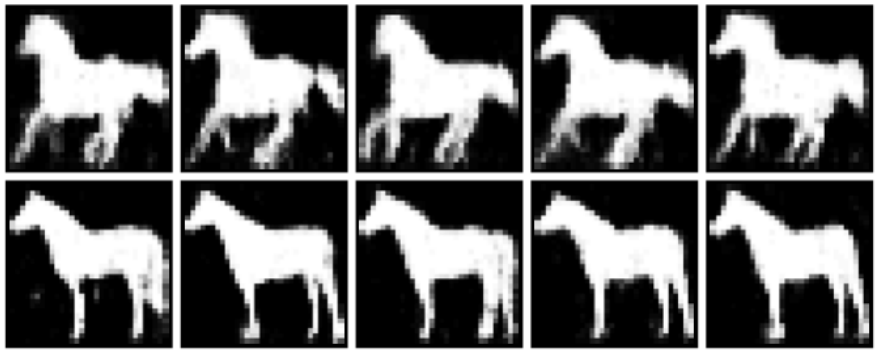
Deep Boltzmann Machines:

Learning Hi
in Features
of edges.



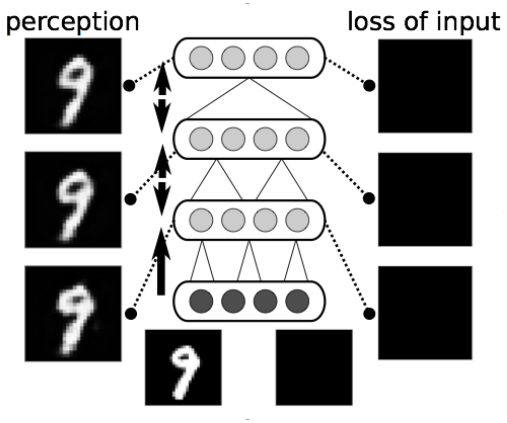
Need more structured and robust models

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)



Talk Roadmap

- Advanced Deep Models
  - Deep Boltzmann Machines
  - Learning Structured and Robust Deep Models
  - One-Shot and Transfer Learning
- Multimodal Learning
- Conclusions

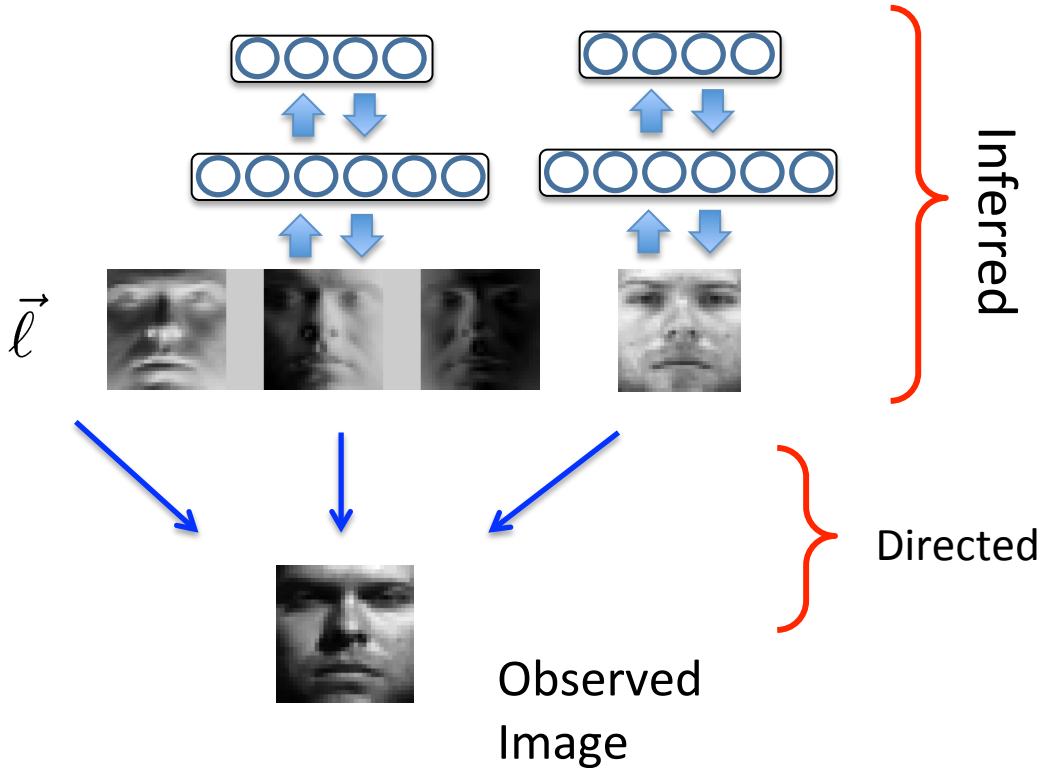
Face Recognition

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

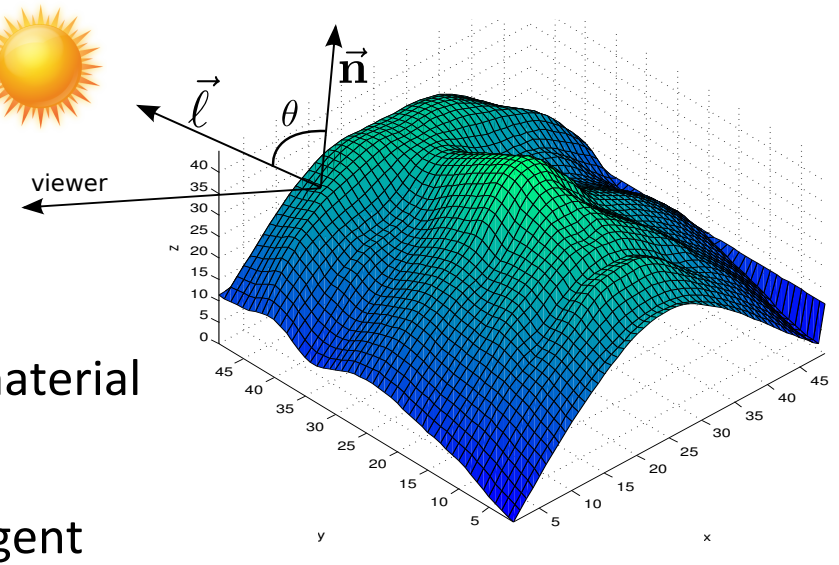
- A simple model of the image formation process.

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

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

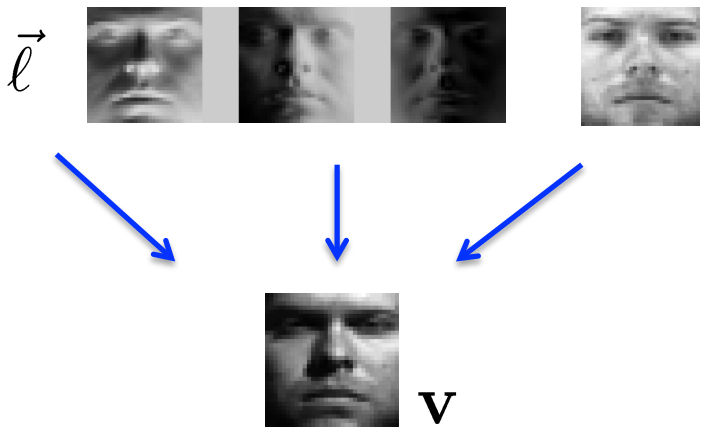


Image albedo

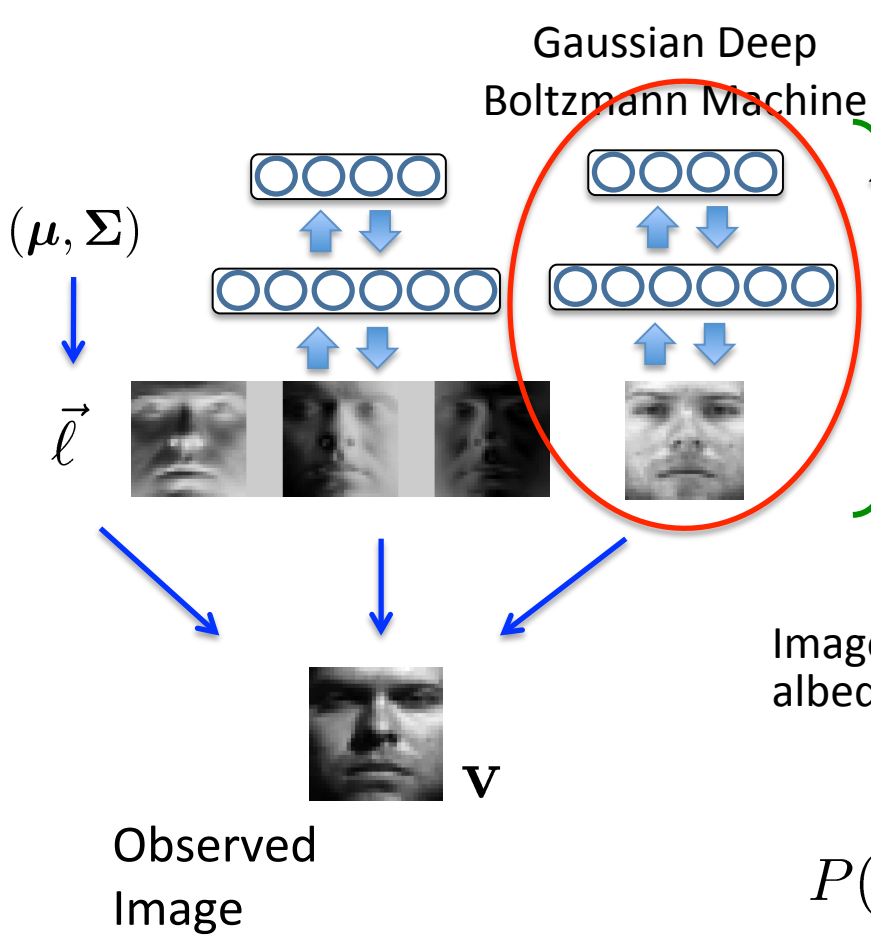
Surface normals

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

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

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

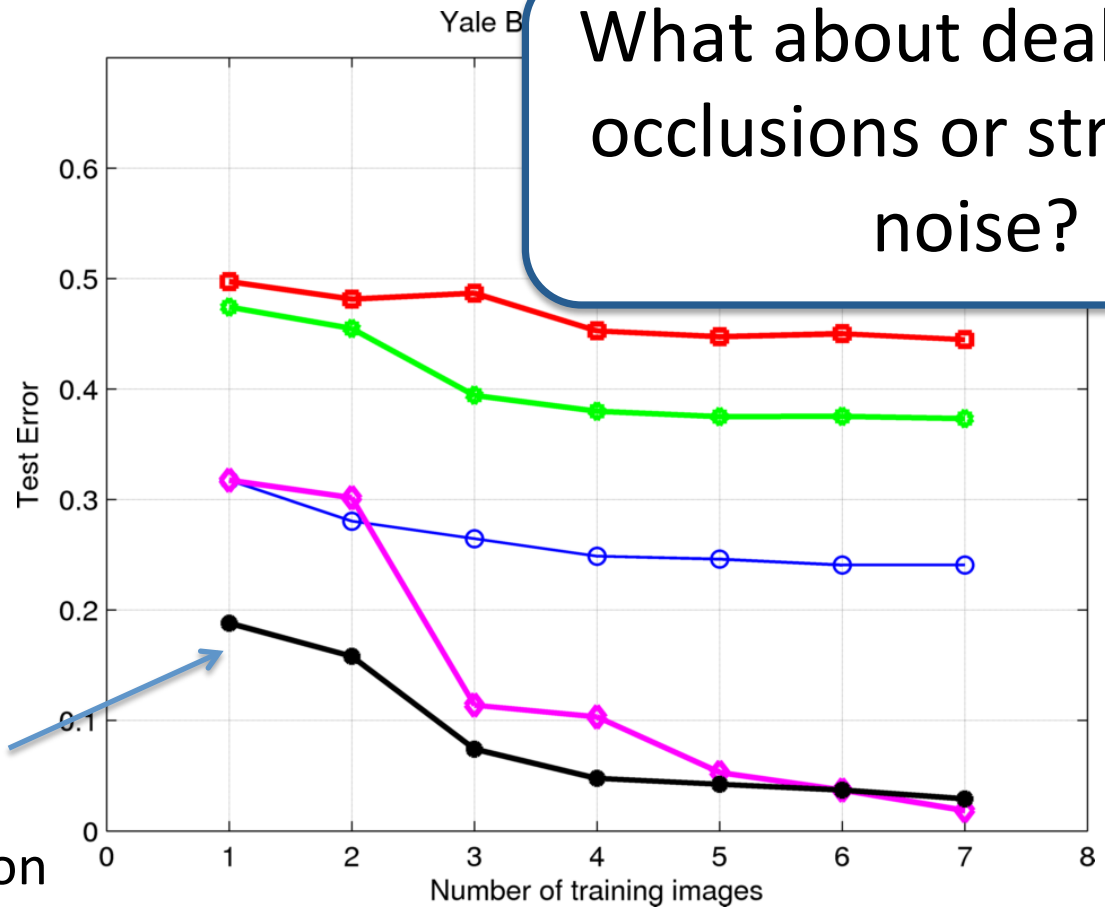
Observed

Inferred albedo

Face Relighting

Recognition Results

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



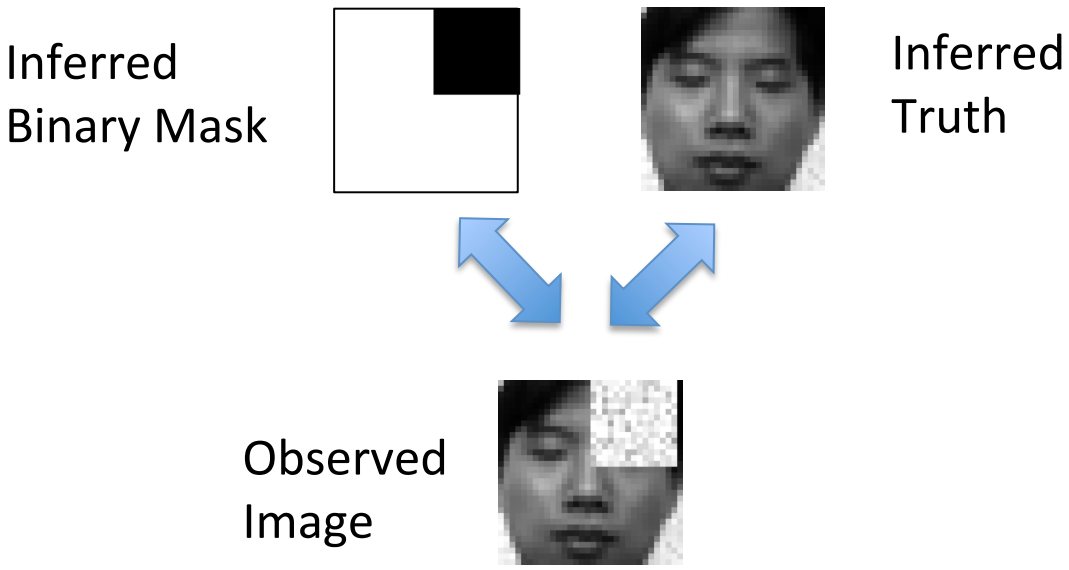
What about dealing with occlusions or structured noise?

One-Shot Recognition

Robust Boltzmann Machines

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

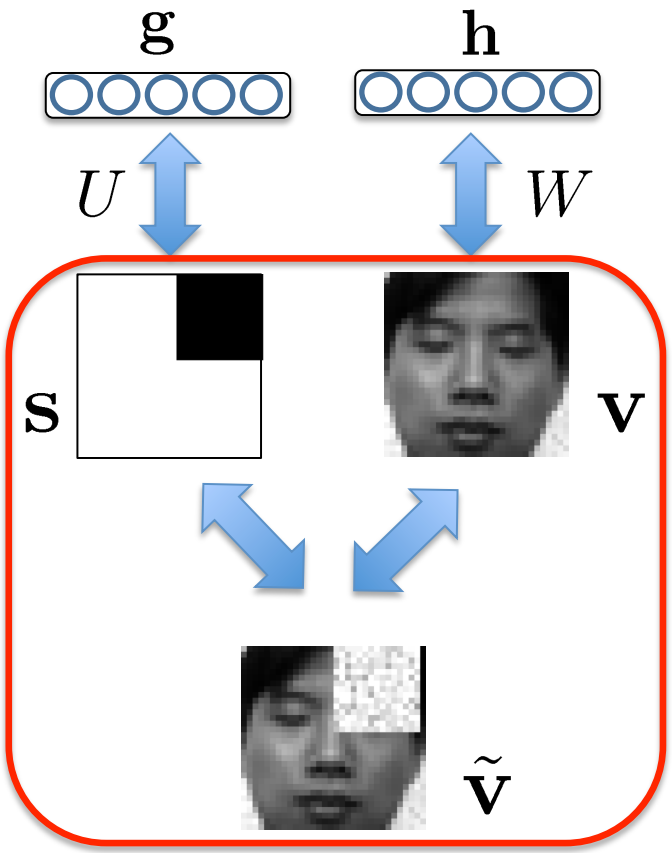
$$\log P(\tilde{\mathbf{v}}, \mathbf{v}, \mathbf{s}, \mathbf{h}, \mathbf{g}) \sim$$



(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

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

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

Gaussian RBM, modeling clean faces

Binary RBM modeling occlusions

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

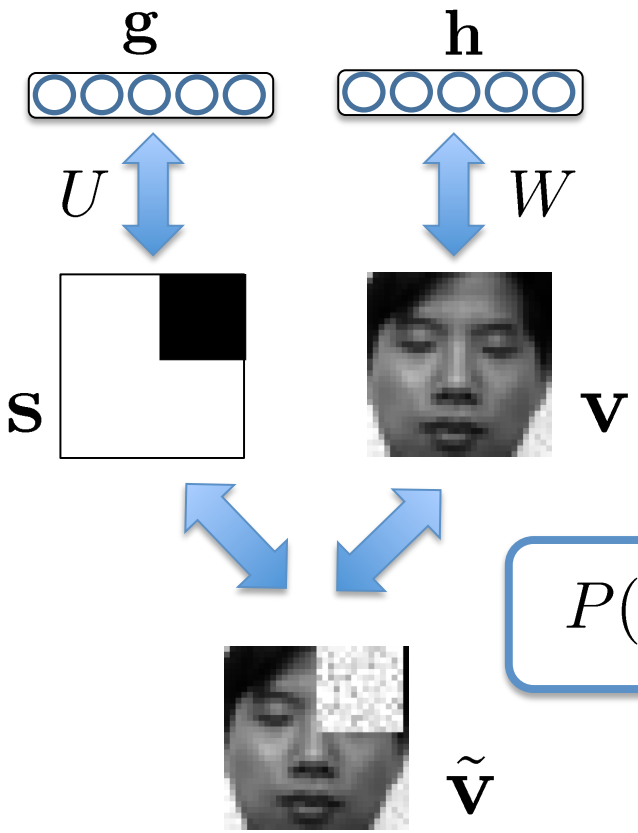
Binary pixel-wise Mask

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.



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

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

Gaussian RBM, modeling clean faces

Binary RBM modeling occlusions

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

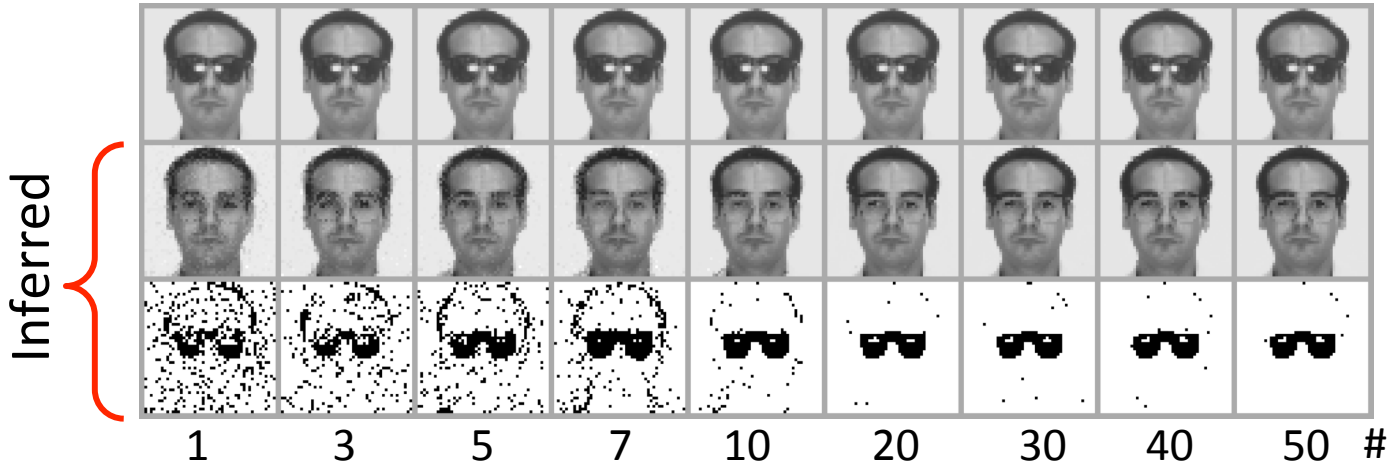
Binary pixel-wise Mask

Gaussian noise model

Inference: Variational Inference.

Learning: Stochastic Approximation

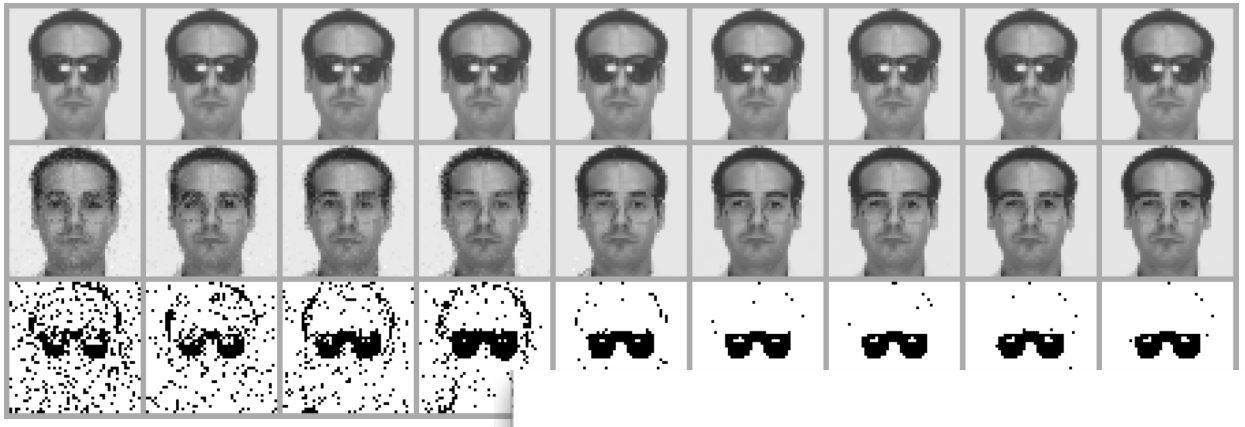
Recognition Results on AR Face Database



Internal states of RoBM during learning.

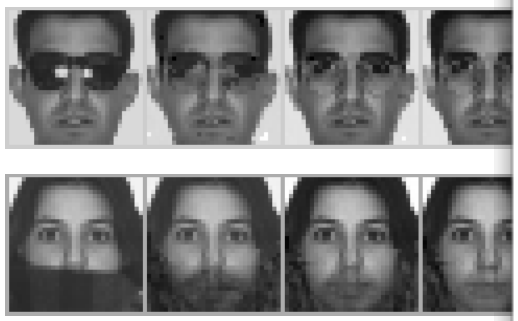
Inference on the test subjects

Recognition Results on AR Face Database



Internal states of RoBM during learning.

Inference on the

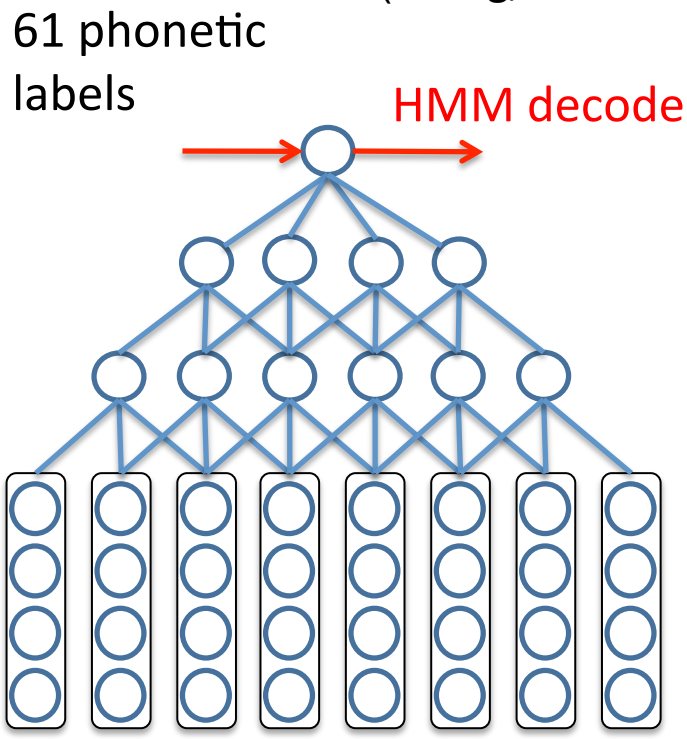


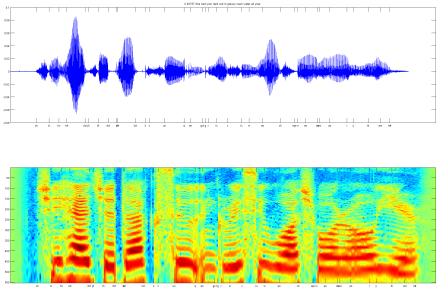
# of iteratio

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)





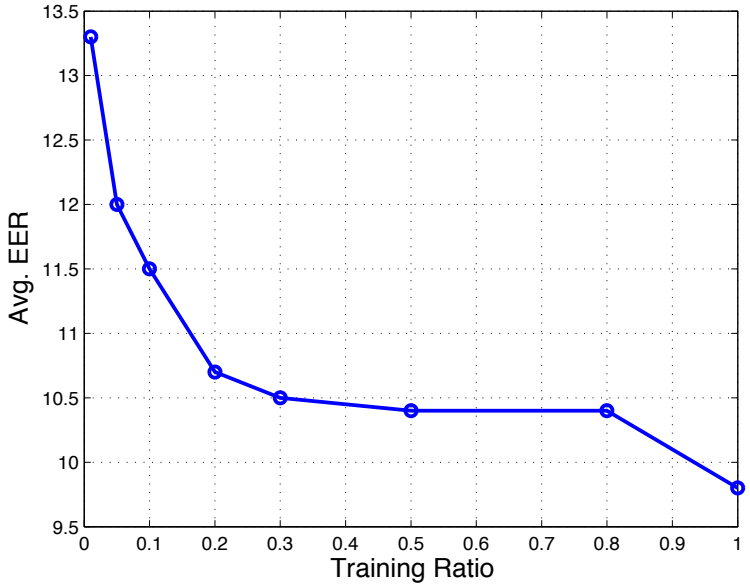
- 630 speaker TIMIT corpus: 3,696 training and 944 test utterances.
- **Spoken Query Detection**:
  For each keyword, estimate utterance's probability of containing that keyword.
- Performance: Average equal error rate (EER).

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)

Talk Roadmap

- Advanced Deep Models
  - Deep Boltzmann Machines
  - Learning Structured and Robust Deep Models
  - One-Shot and Transfer Learning
- Multimodal Learning
- Conclusions

Transfer Learning

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

Supervised Learning

Segway

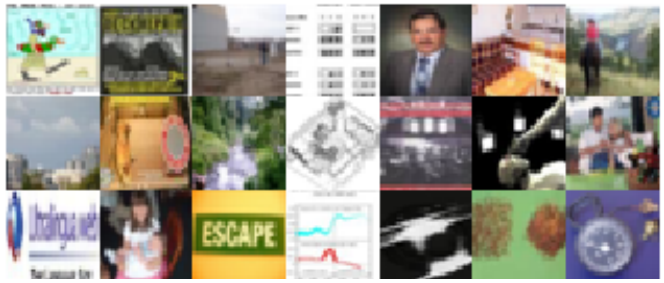
Motorcycle

Test:

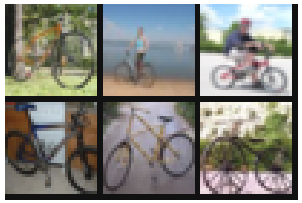
Learning to Learn

Background Knowledge

Millions of unlabeled images



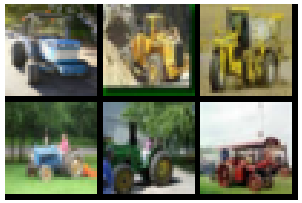
Some labeled images



Bicycle

Dolphin

Elephant



Tractor

Learn to Transfer Knowledge

Learn novel concept from one example

Test:

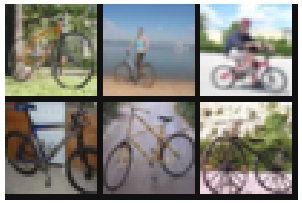
Learning to Learn

Background Knowledge

Millions of unlabeled images

Key problem in computer vision, speech perception, natural language processing, and many other domains.

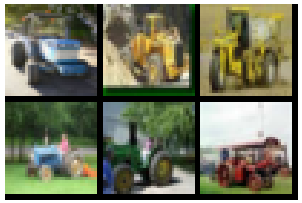
Some labeled images



Bicycle

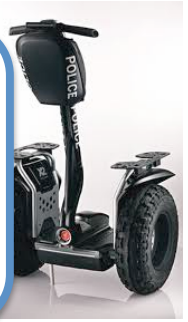
Dolphin

Elephant



Tractor

Learn to Transfer Knowledge



Learn novel concept from one example

Test:

Hierarchical-Deep Model

HD Models: Integrate hierarchical Bayesian models with deep models.

One-Shot Learning

Super-category

Hierarchical Bayes:

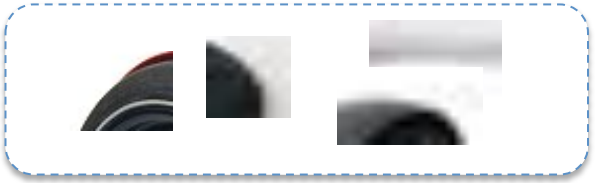
- Learn **hierarchies of categories** for sharing abstract knowledge.

Shared higher-level features

Deep Models:

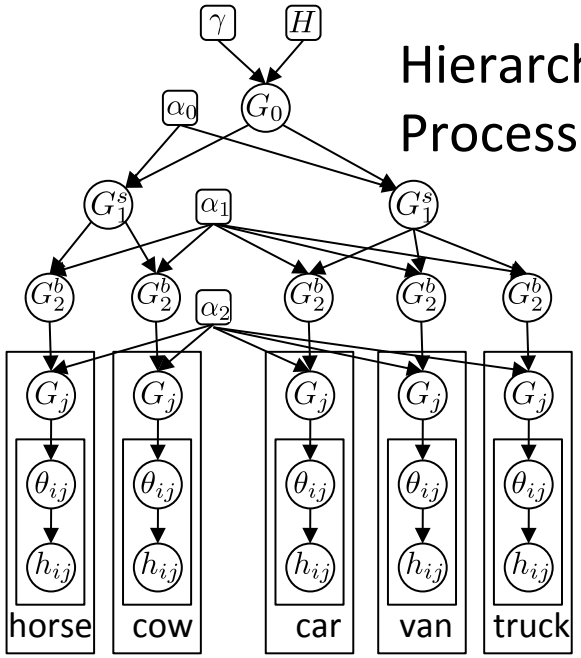
- Learn **hierarchies of features.**
- **Unsupervised feature learning** – no need to rely on human-crafted input features.

Shared low-level features

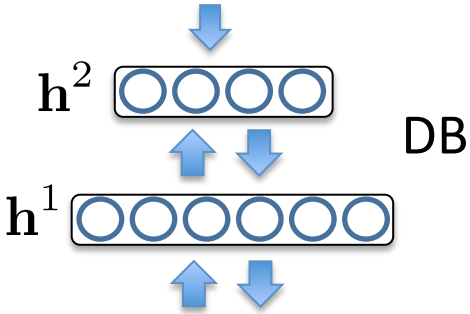


(Salakhutdinov, Tenenbaum, Torralba, NIPS 2011, PAMI 2013)

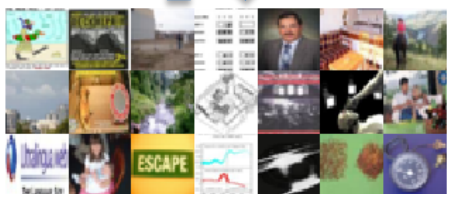
Hierarchical-Deep Model



Hierarchical Dirichlet Process Model



DBM Model



Images

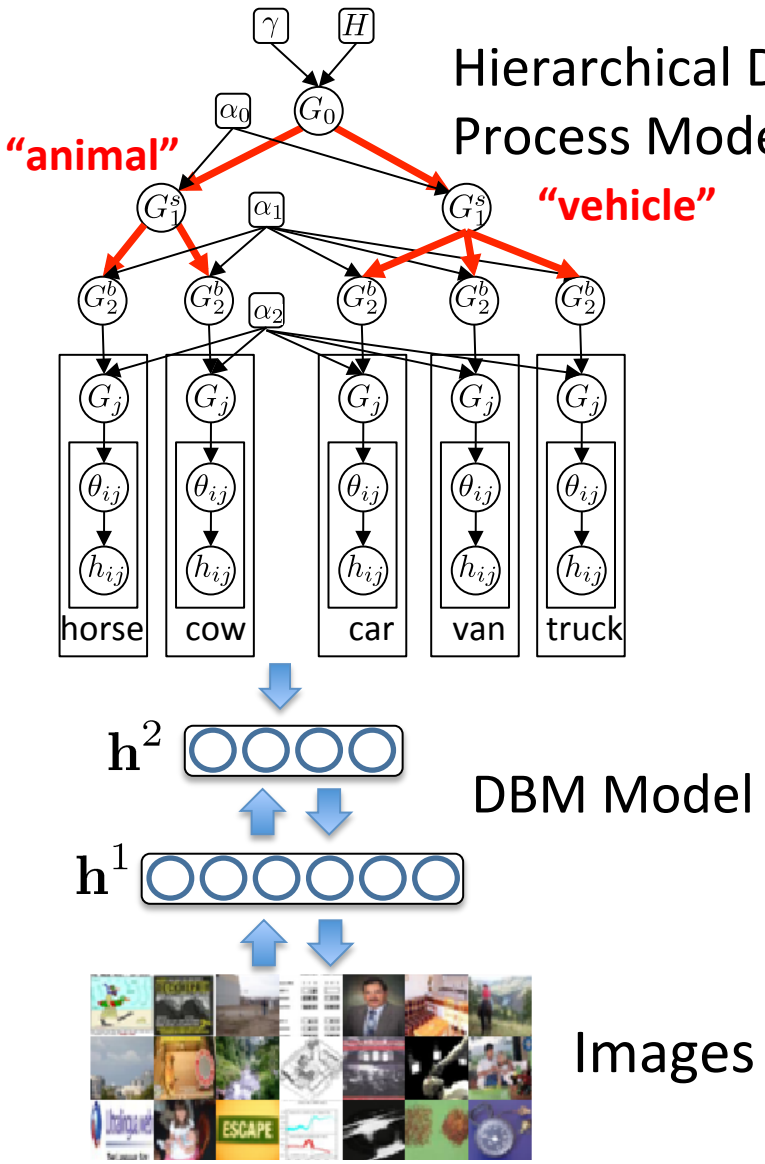
Higher-level class-sensitive features:

- capture distinctive perceptual structure of a specific concept

Lower-level generic features:

- edges, combination of edges

Hierarchical-Deep Model



Hierarchical Dirichlet Process Model

DBM Model

Images

Hierarchical Organization of Categories:

- express priors on the features that are typical of different kinds of concepts
- modular data-parameter relations

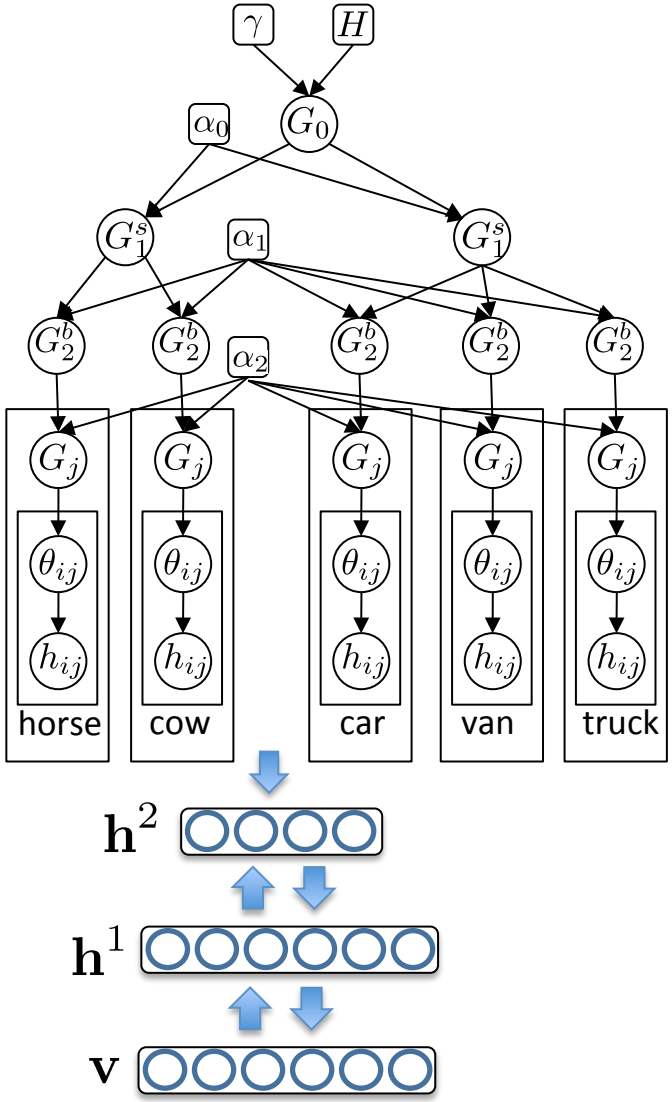
Higher-level class-sensitive features:

- capture distinctive perceptual structure of a specific concept

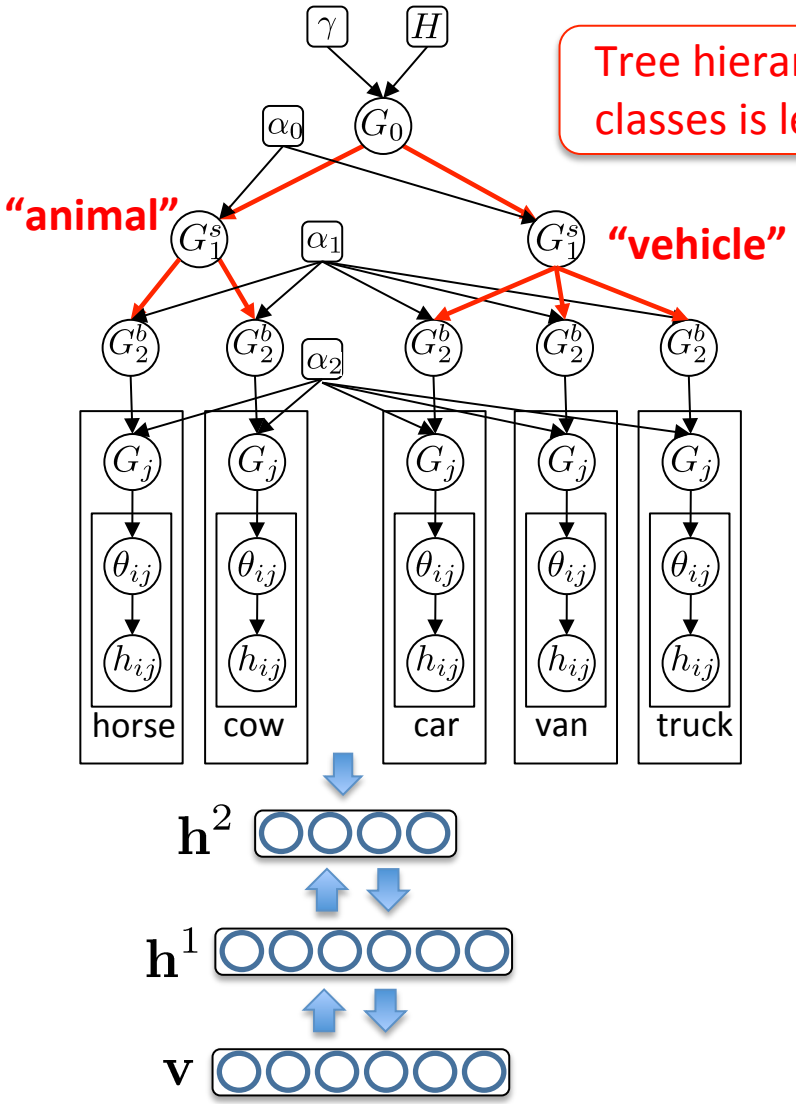
Lower-level generic features:

- edges, combination of edges

Hierarchical-Deep Model



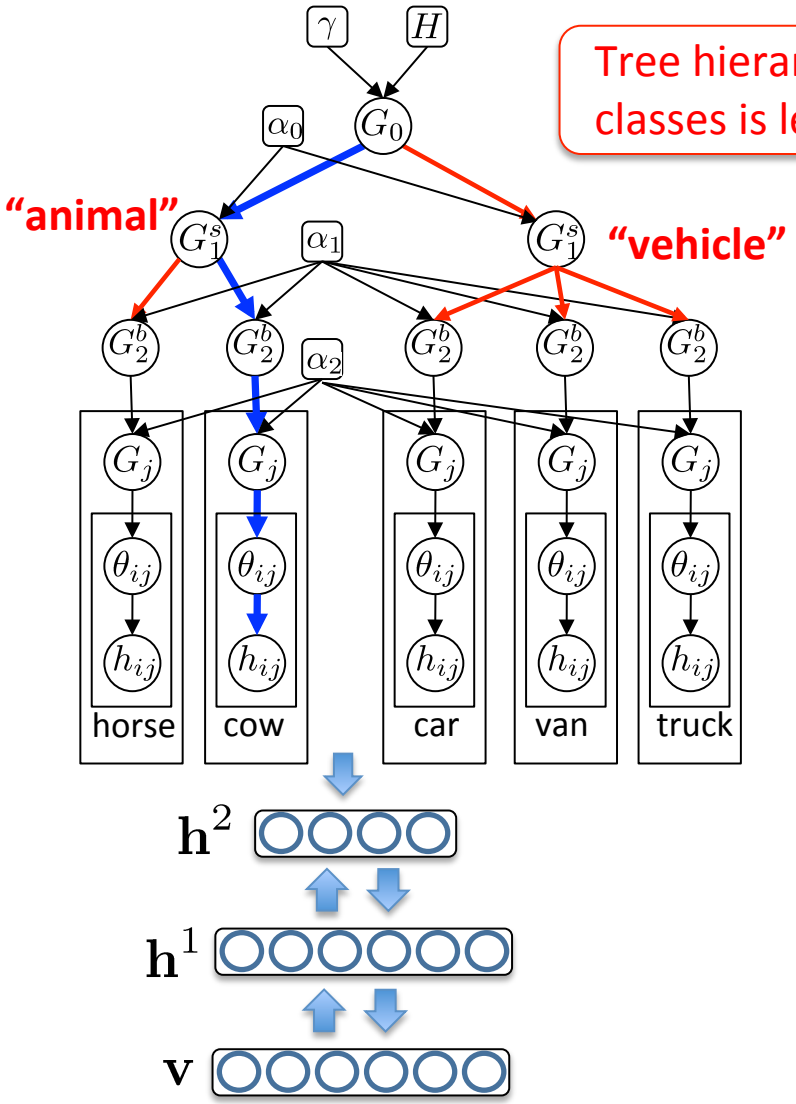
Hierarchical-Deep Model



$\mathbf{z} \sim \text{nCRP}$ (**Nested Chinese Restaurant Process**)

prior: a nonparametric prior over tree structures

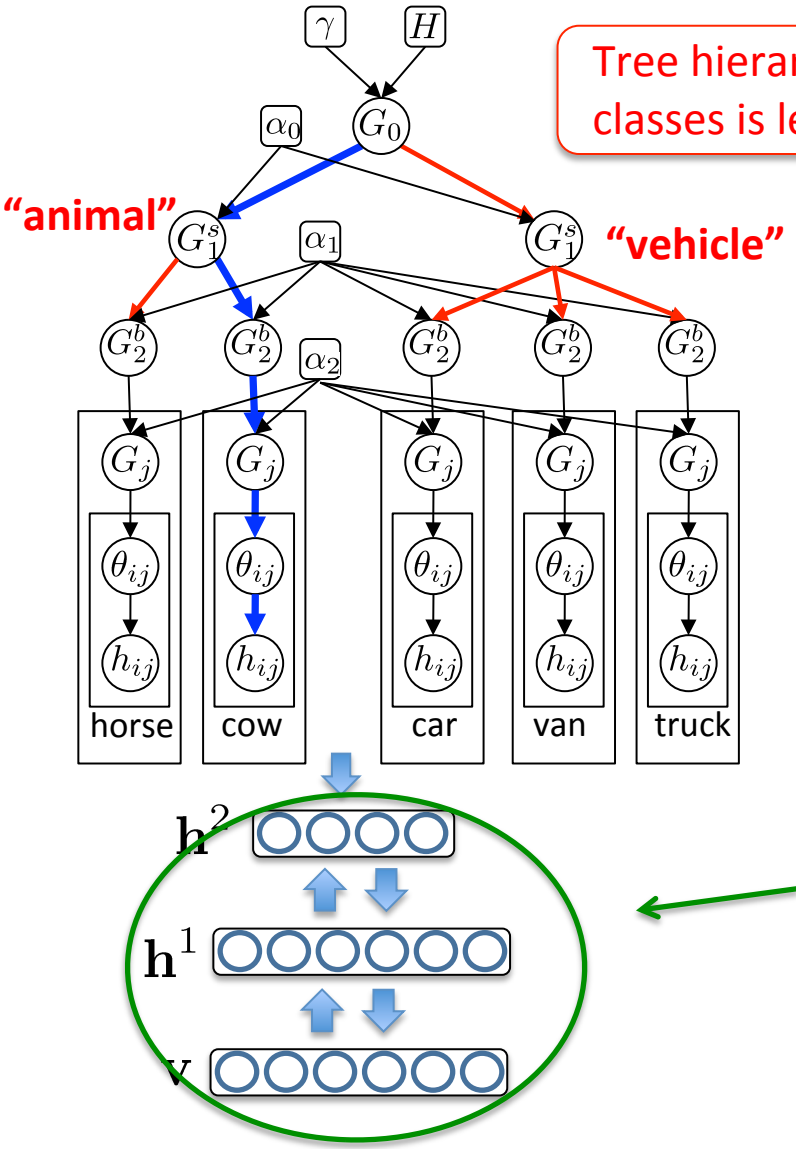
Hierarchical-Deep Model



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

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

Hierarchical-Deep Model



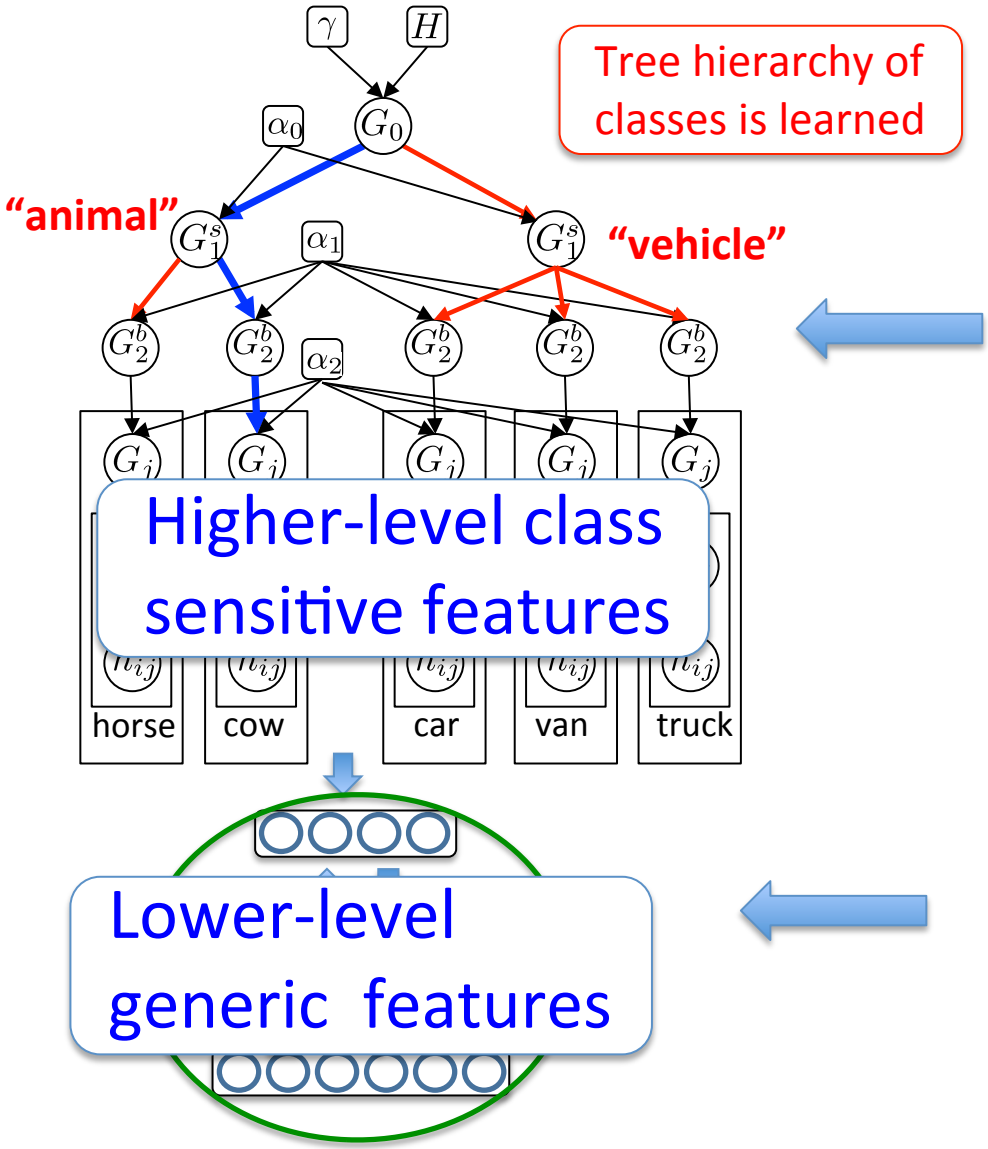
$\mathbf{z} \sim \text{nCRP}$ (**Nested Chinese Restaurant Process**)
prior: a nonparametric prior over tree structures

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

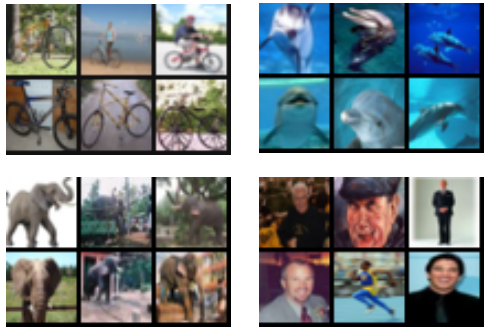
$\mathbf{v}|\mathbf{h}^3 \sim \text{DBM}$ **Deep Boltzmann Machine**

Enforce (approximate) global consistency through many local constraints.

CIFAR Object Recognition

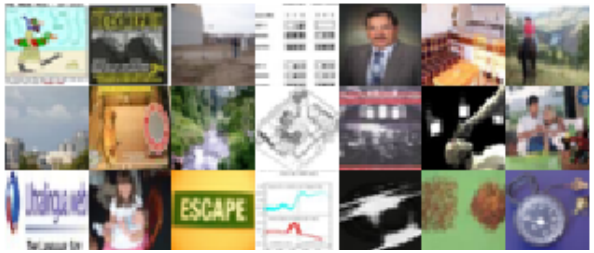


50,000 images of 100 classes



Inference: Markov chain Monte Carlo

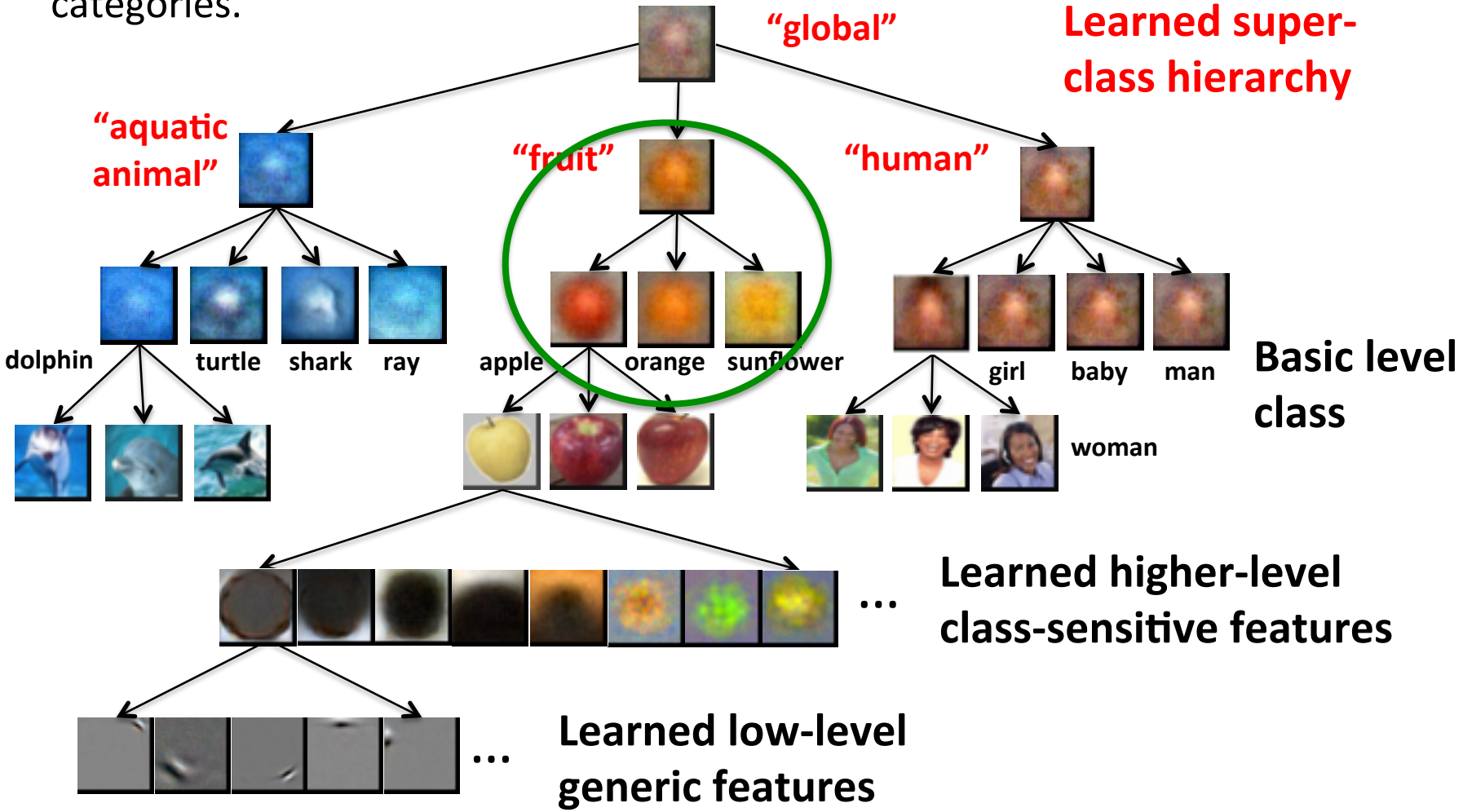
4 million unlabeled images



32 x 32 pixels x 3 RGB

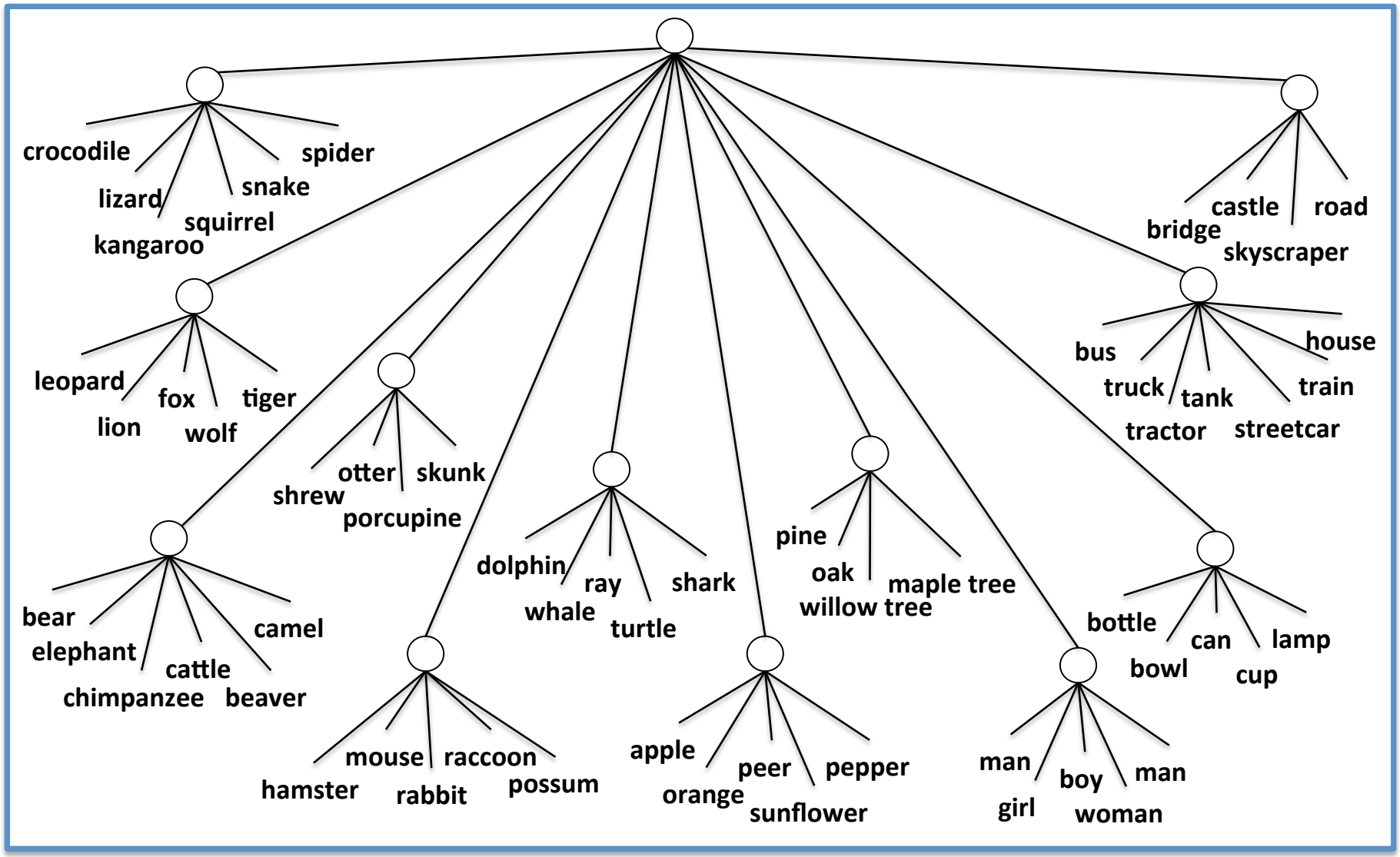
Learning the Hierarchy

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

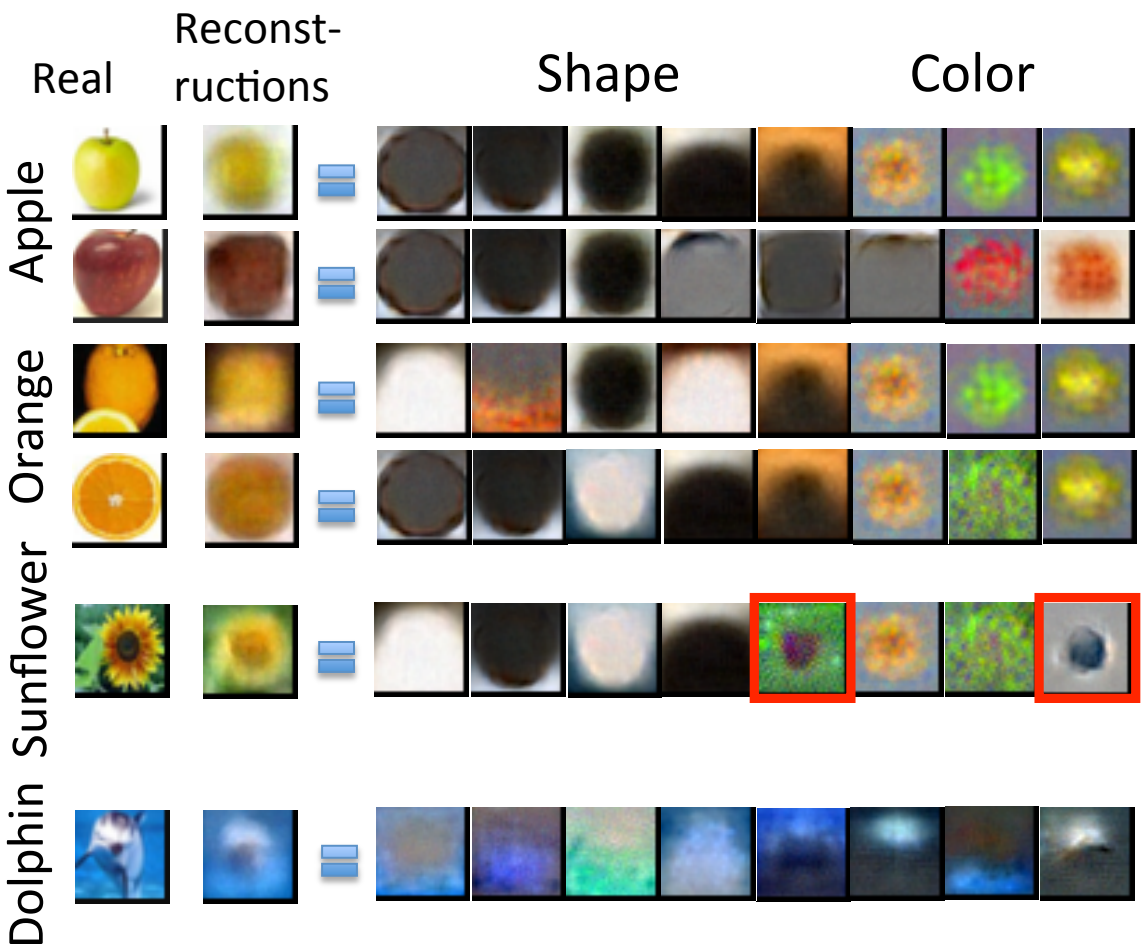


Learning the Hierarchy

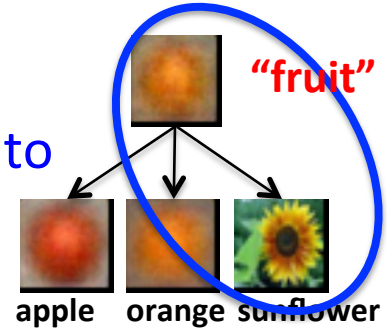
The model learns how to share the knowledge across many visual



Sharing Features

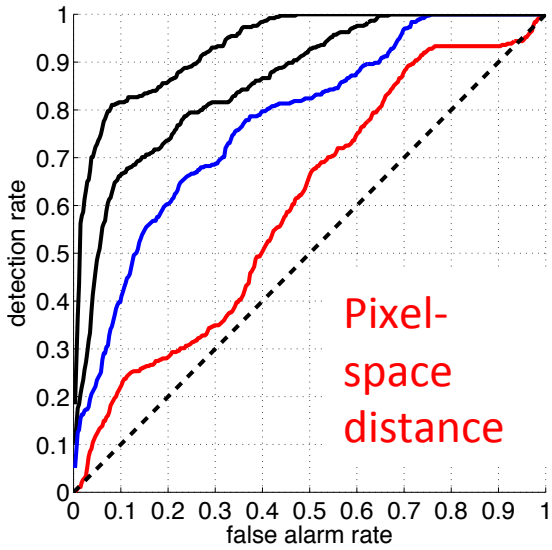


Learning to Learn



Sunflower ROC curve

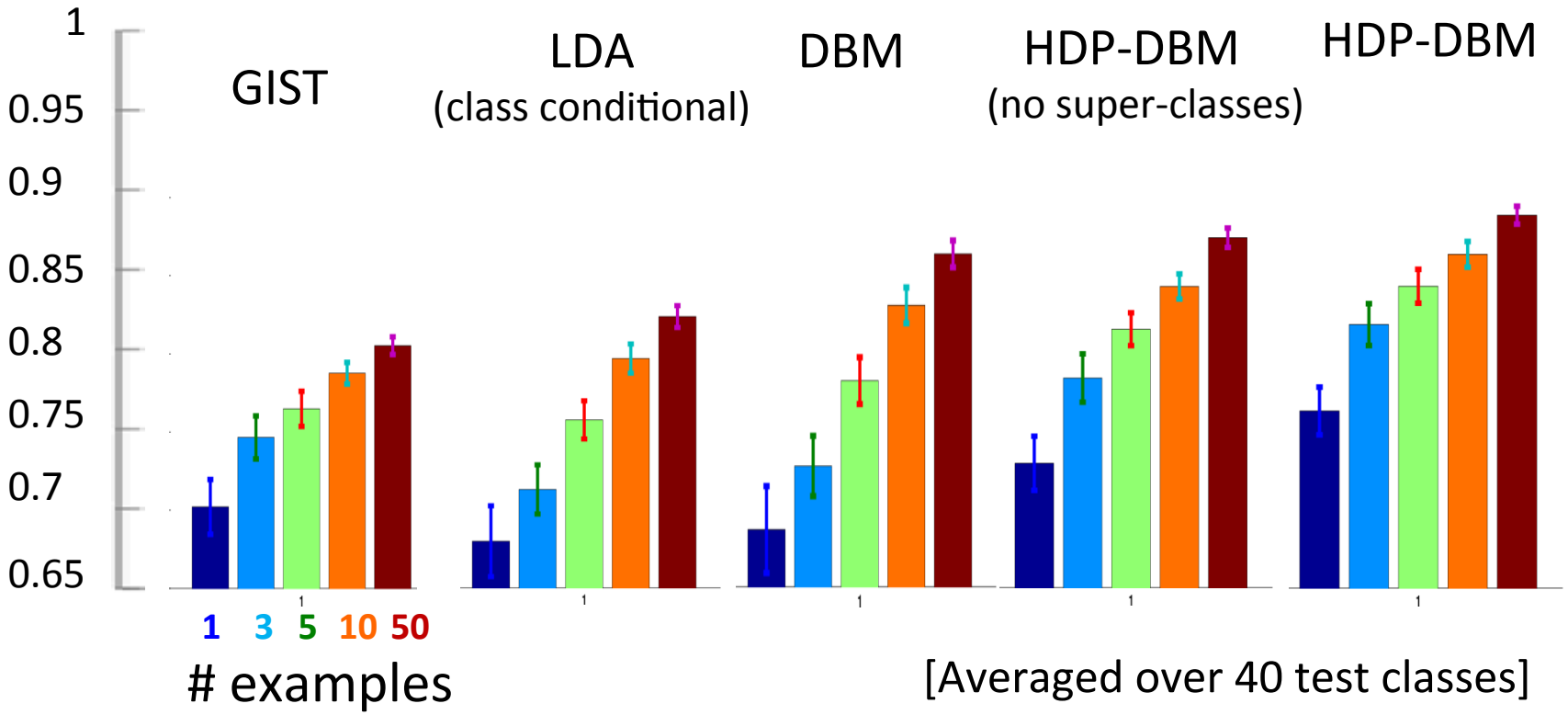
5 3 1 ex's



Learning to Learn: Learning a hierarchy for sharing parameters – rapid learning of a novel concept.

Object Recognition

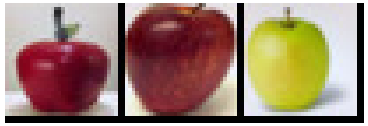
Area under ROC curve for same/different
(1 new class vs. 99 distractor classes)



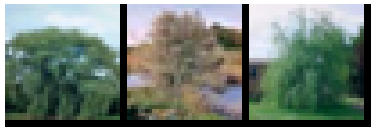
Our model outperforms standard computer vision features (e.g. GIST).

Learning from 3 Examples

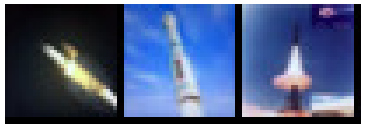
Given only 3 Examples



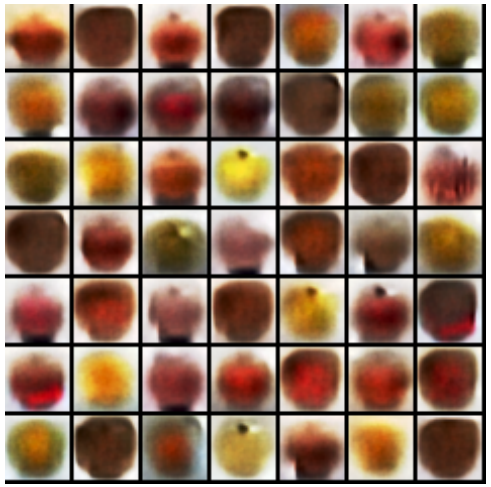
Willow Tree

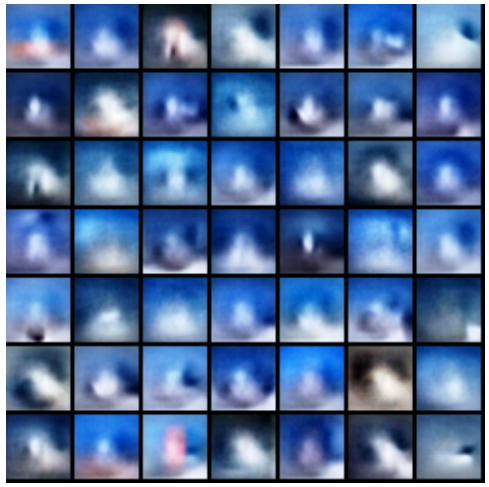


Rocket

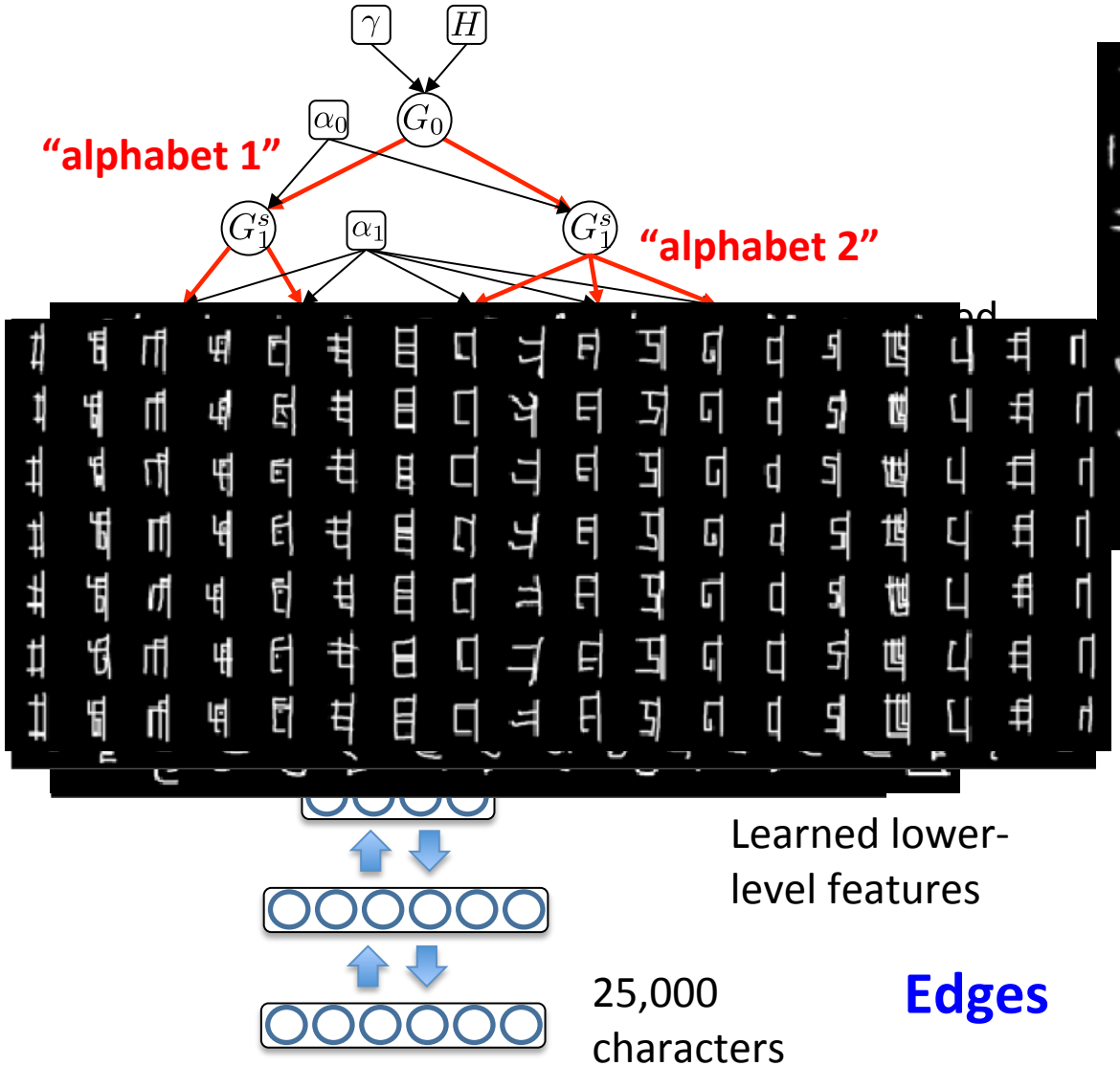


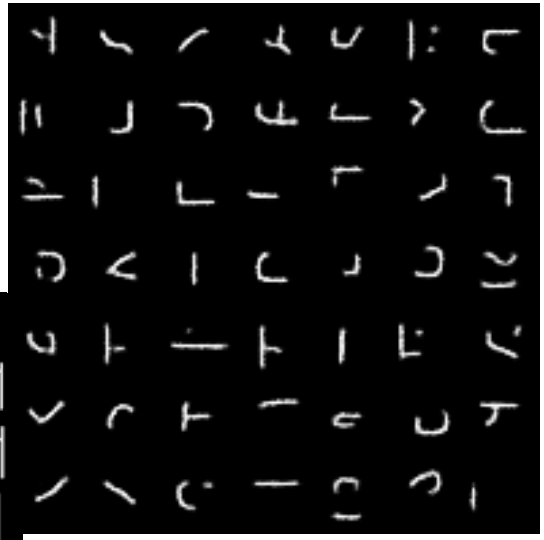
Generated Samples



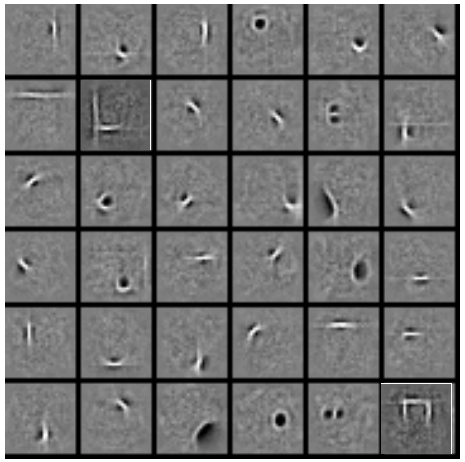


Handwritten Character Recognition





Learned lower-level features

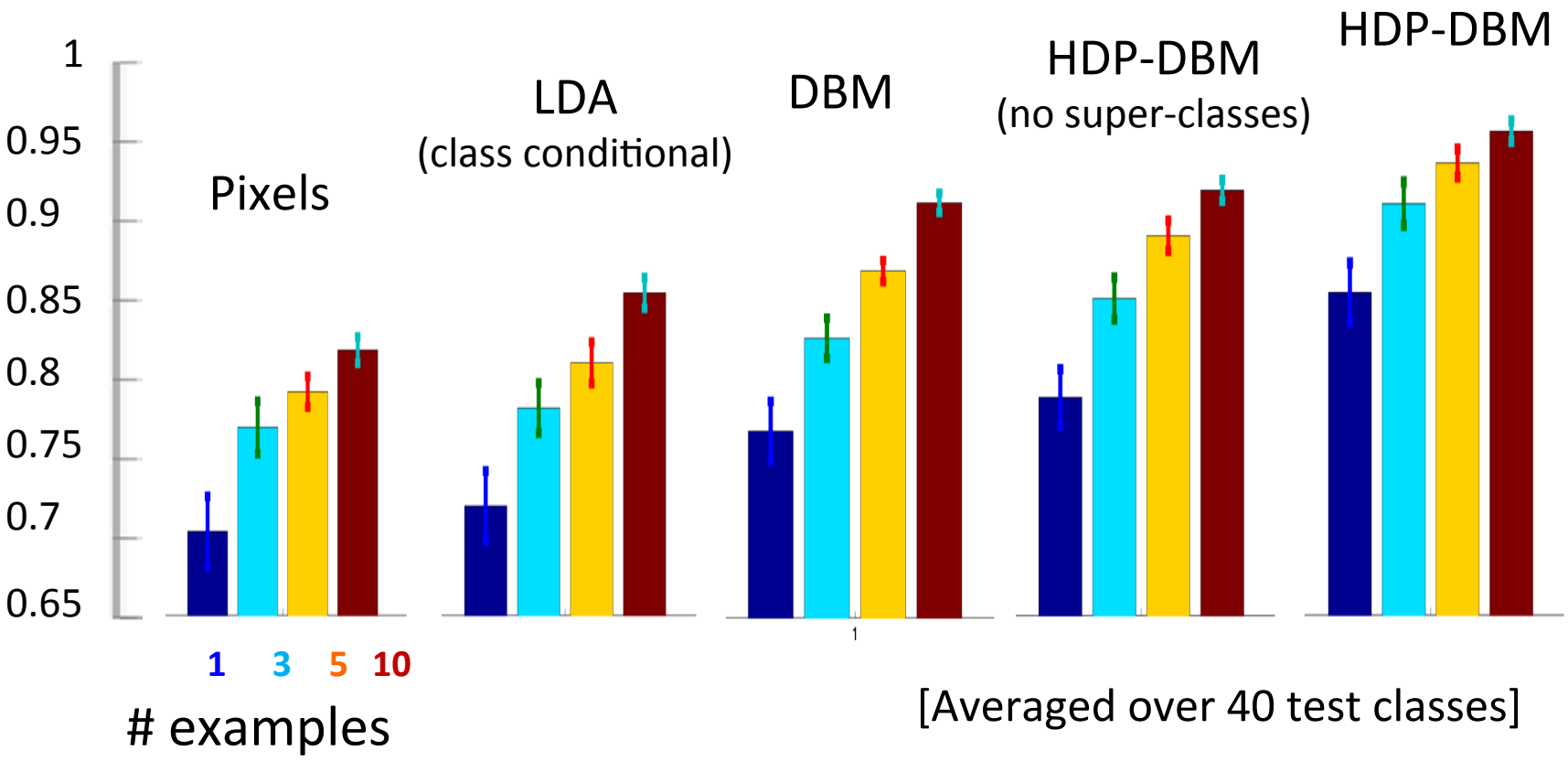


25,000 characters

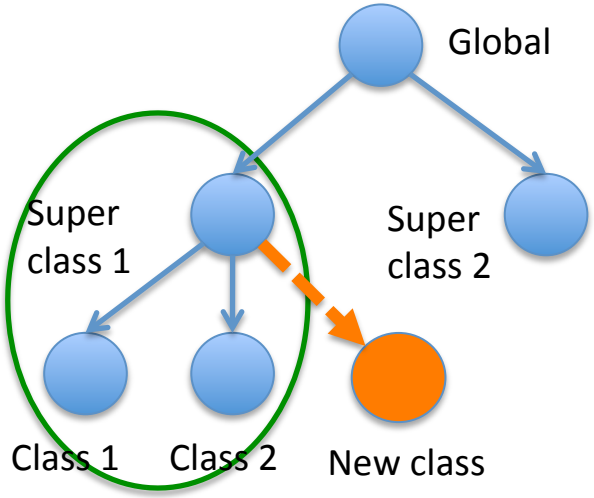
Edges

Handwritten Character Recognition

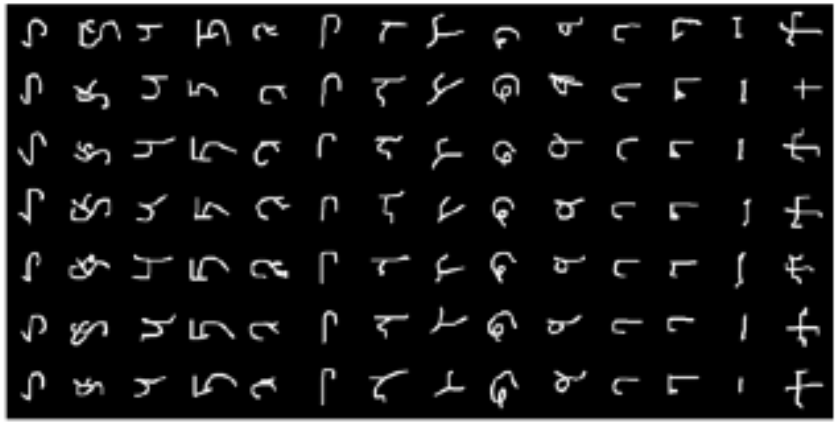
Area under ROC curve for same/different
(1 new class vs. 1000 distractor classes)



Simulating New Characters

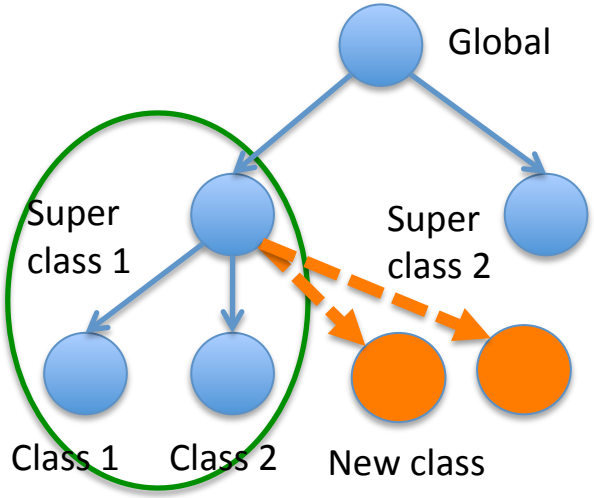


Real data within super class

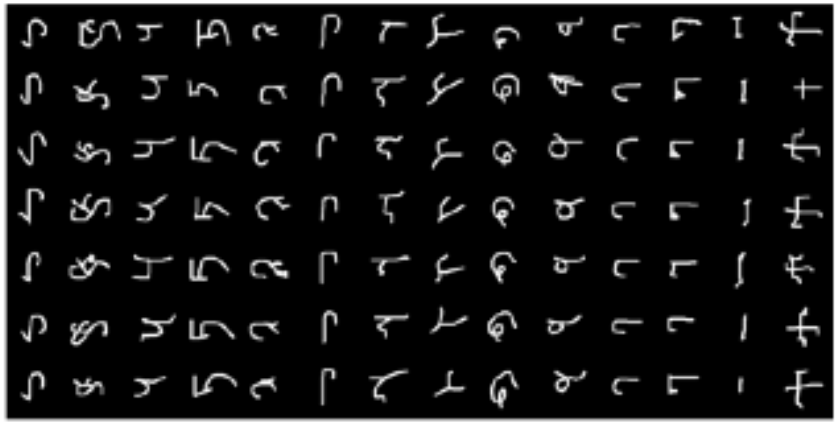


Simulated new characters

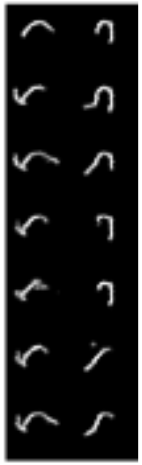
Simulating New Characters



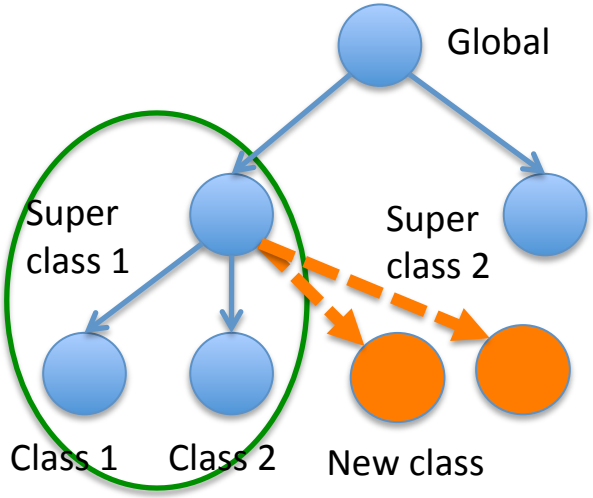
Real data within super class



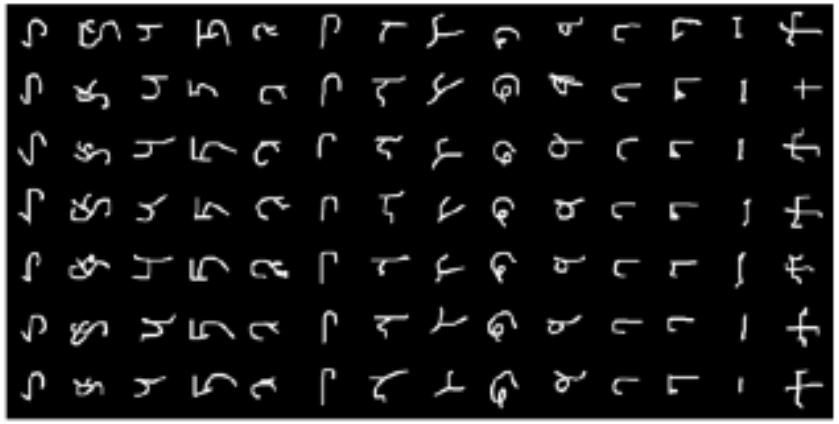
Simulated new characters



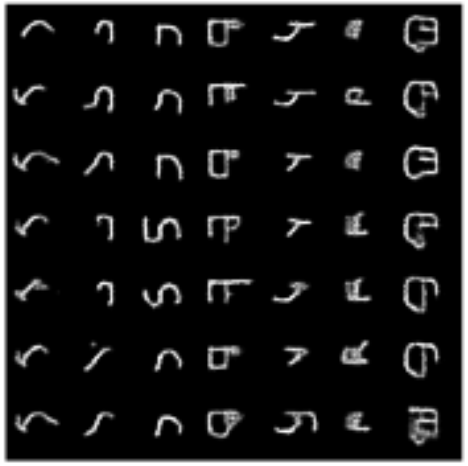
Simulating New Characters



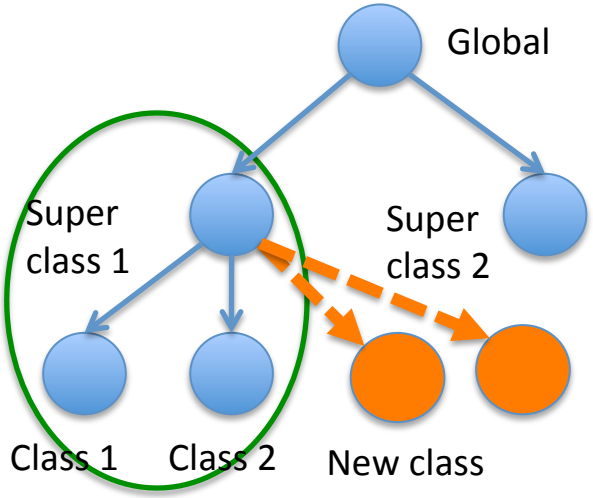
Real data within super class



Simulated new characters

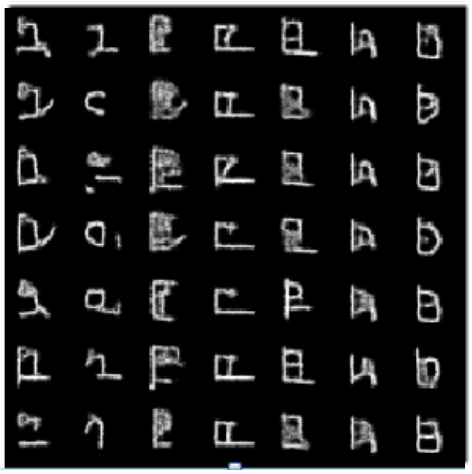


Simulating New Characters

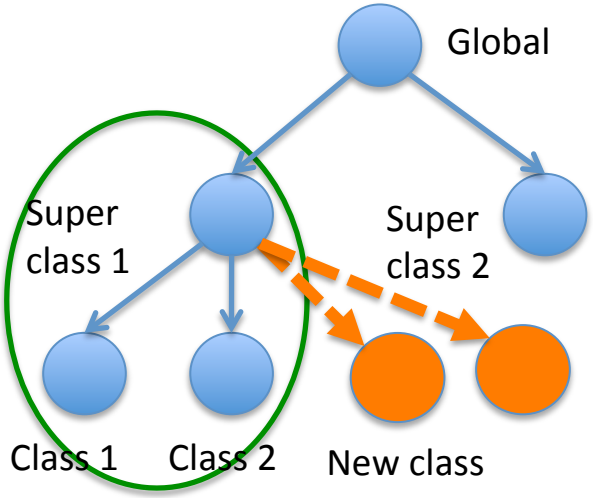


Real data within super class

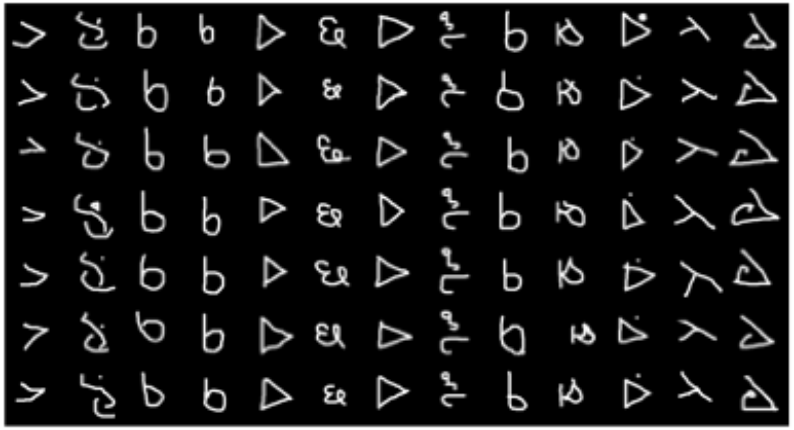
Simulated new characters



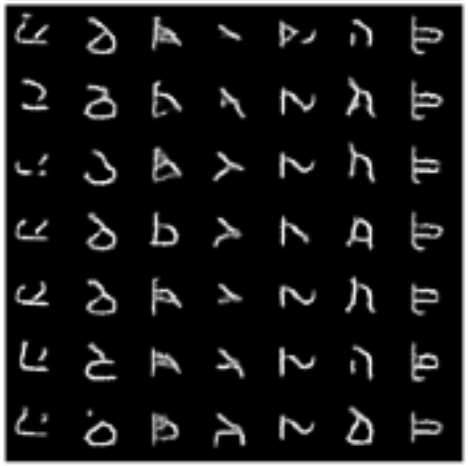
Simulating New Characters



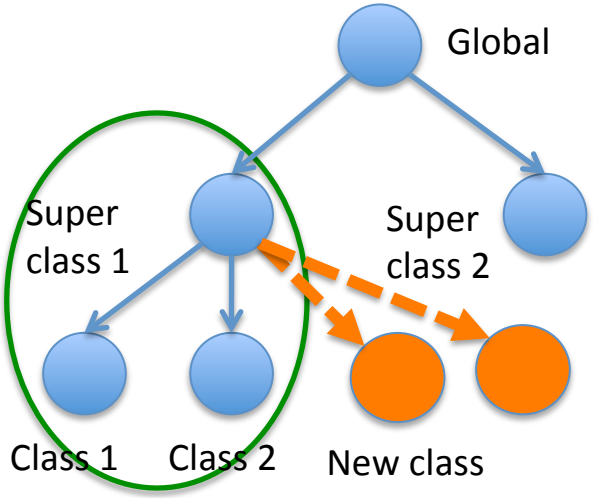
Real data within super class



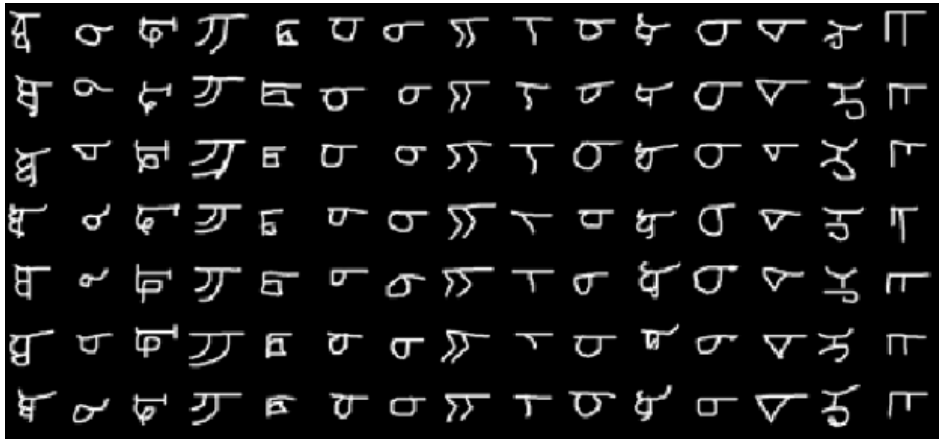
Simulated new characters



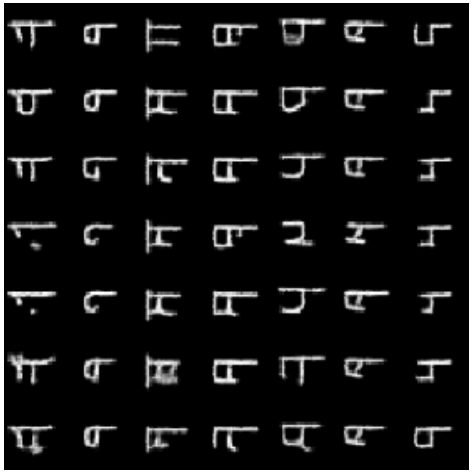
Simulating New Characters



Real data within super class

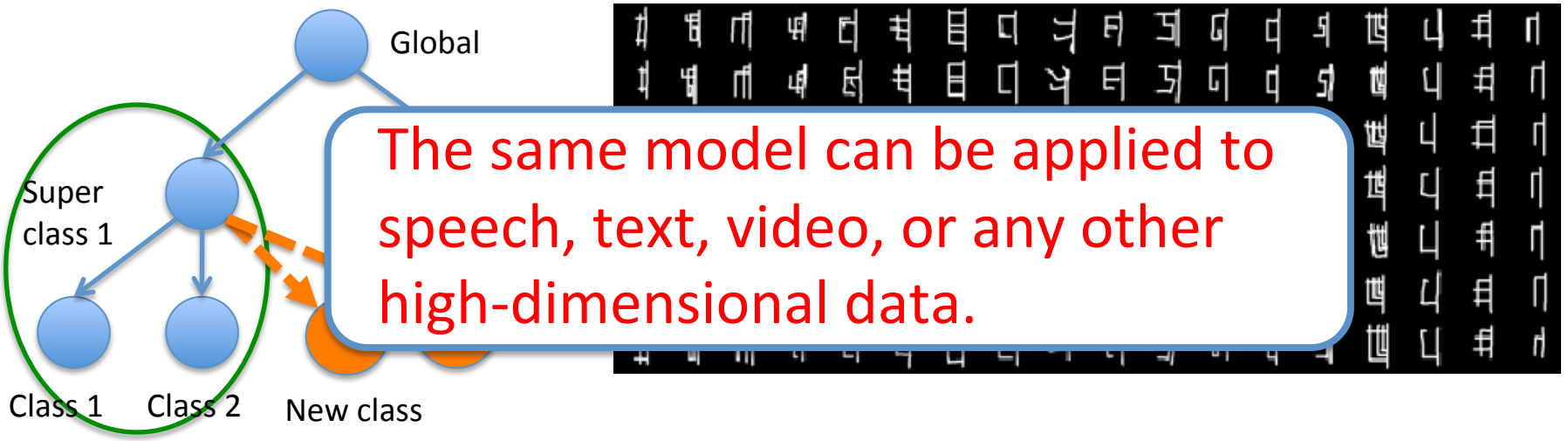


Simulated new characters

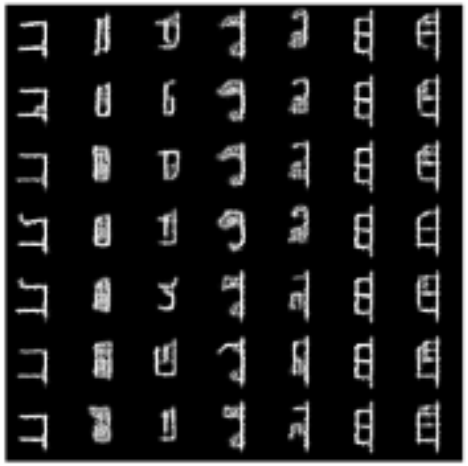


Simulating New Characters

Real data within super class

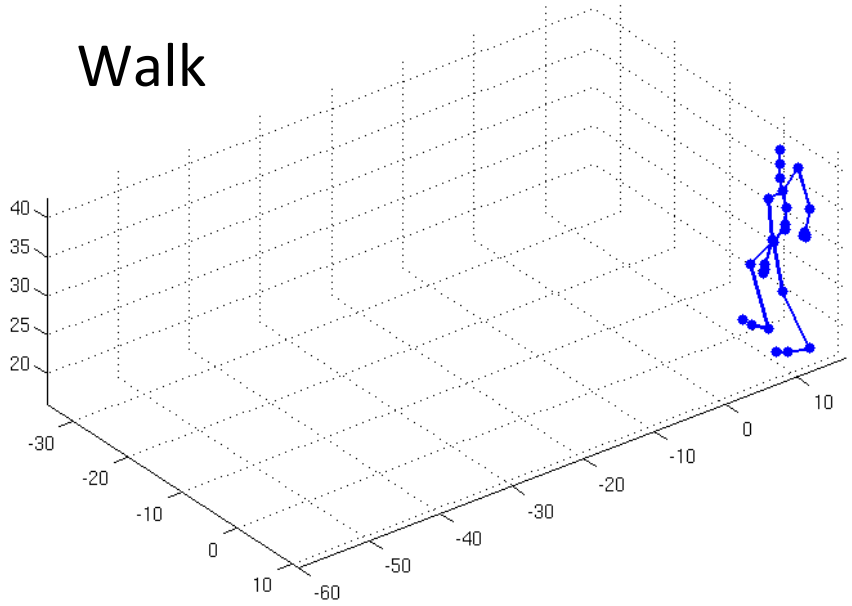


Simulated new characters

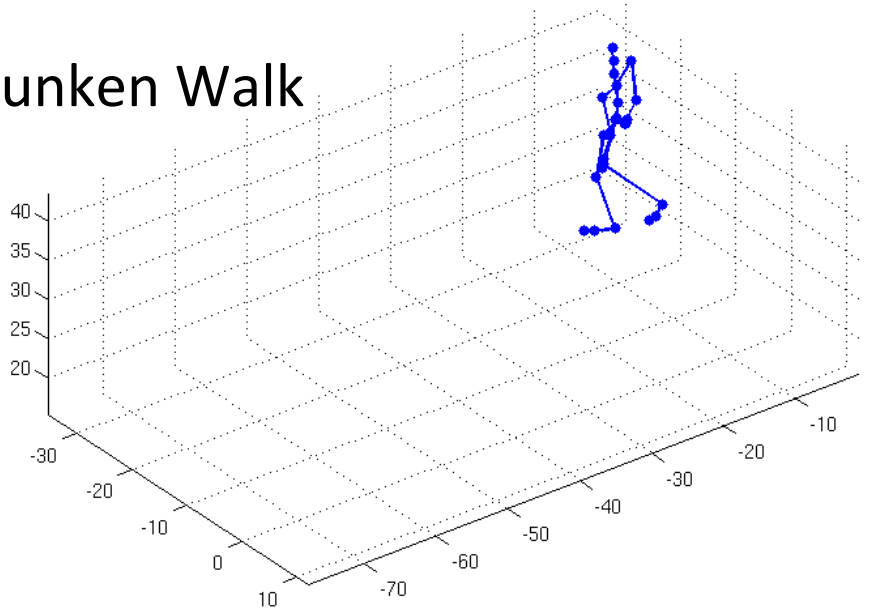


Motion Capture

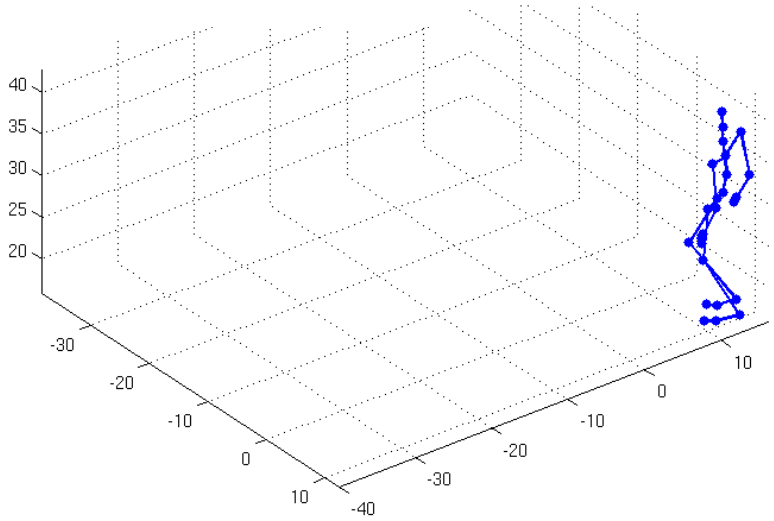
Walk



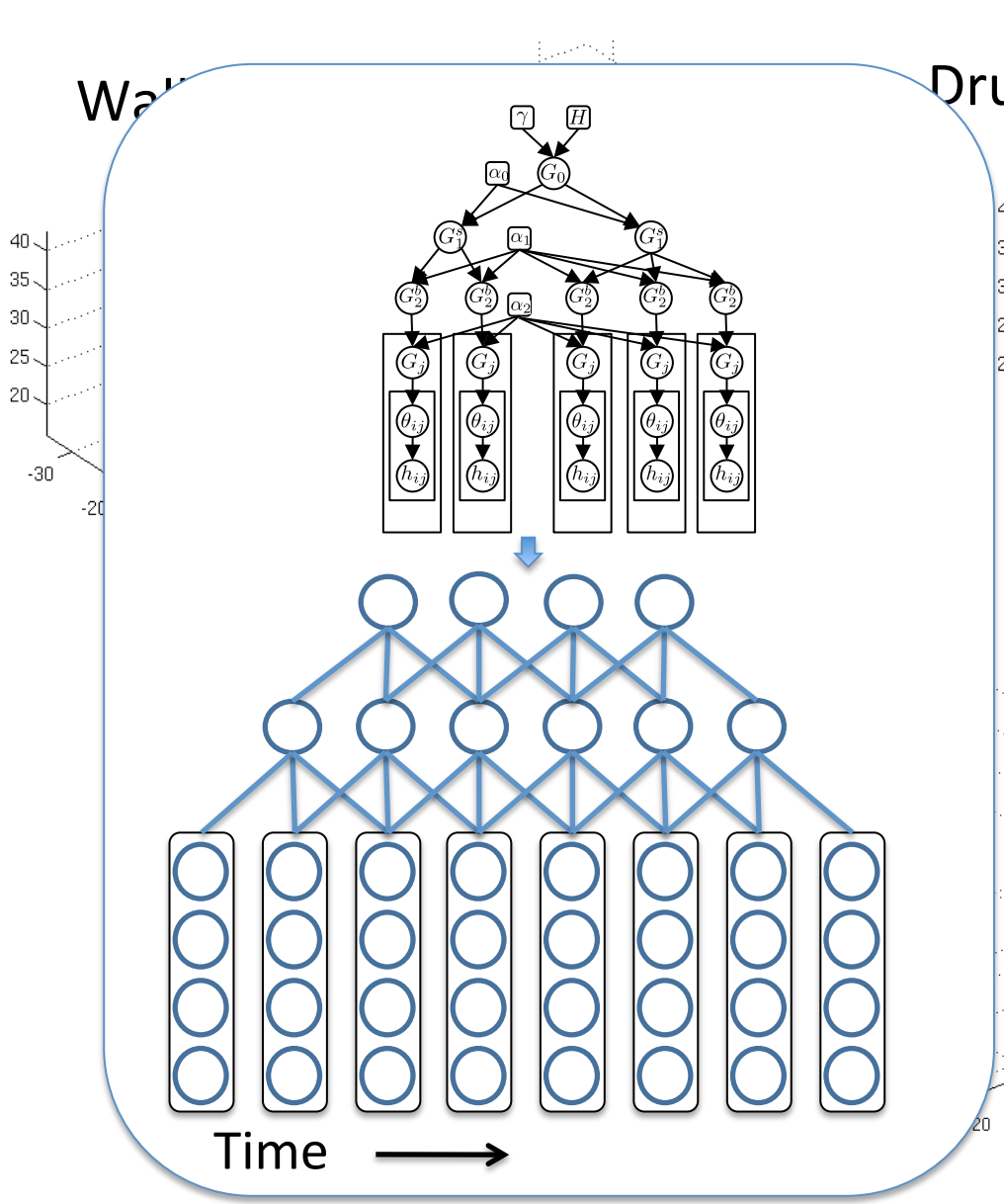
Drunken Walk



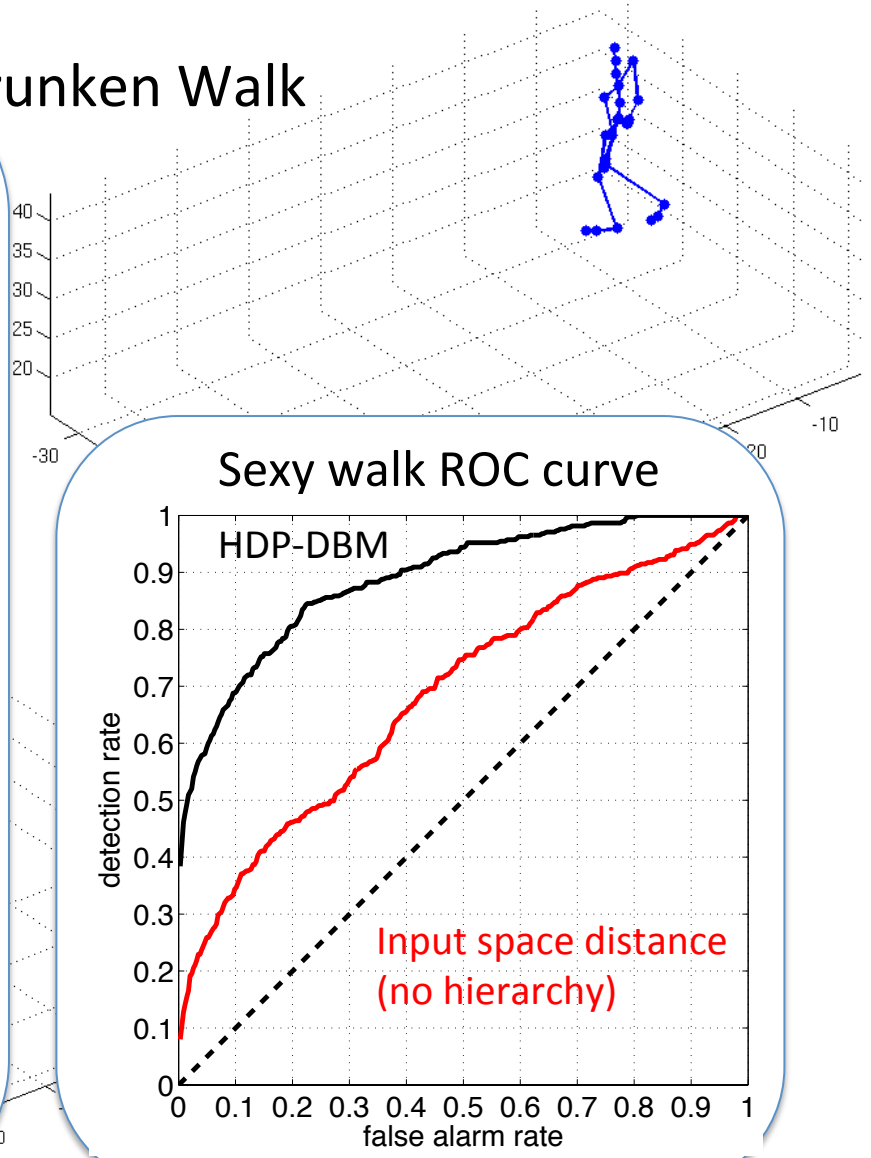
Sexy Walk



Motion Capture



Drunken Walk



Sexy walk ROC curve

Talk Roadmap

- Advanced Deep Models
  - Deep Boltzmann Machines
  - One-Shot and Transfer Learning
  - Learning Structured and Robust Deep Models

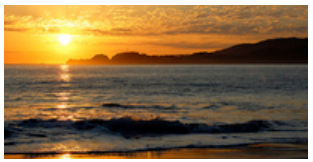
- Multimodal Learning

- Conclusions

Data – Collection of Modalities

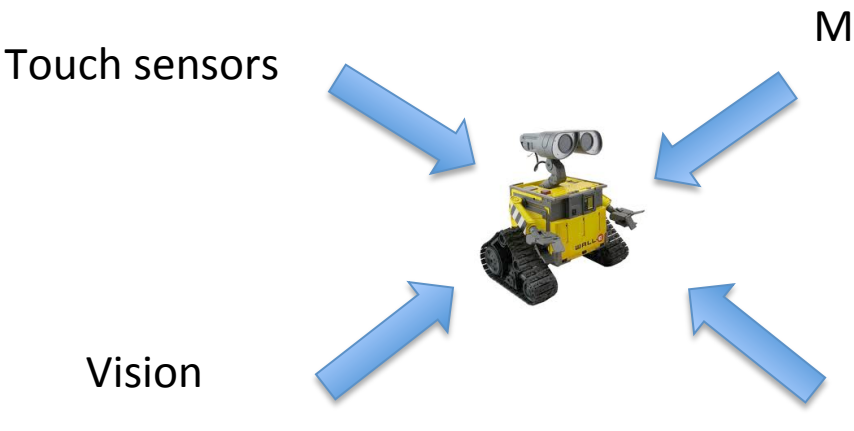
- Multimedia content on the web - image + text + audio.
- Product recommendation systems.
- Robotics applications.

car, automobile



sunset, pacificocean, bakerbeach, seashore, ocean

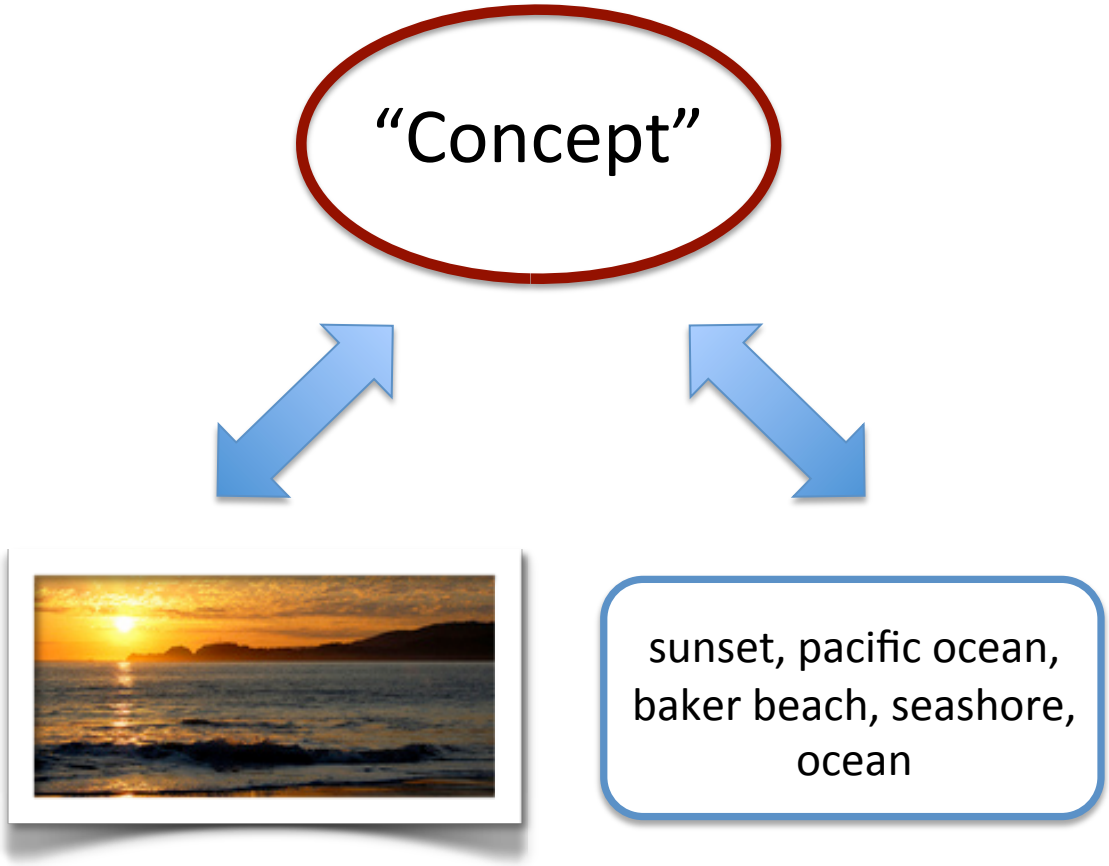
Google ebay YouTube flickr



NETFLIX amazon

Shared Concept

"Modality-free" representation



"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

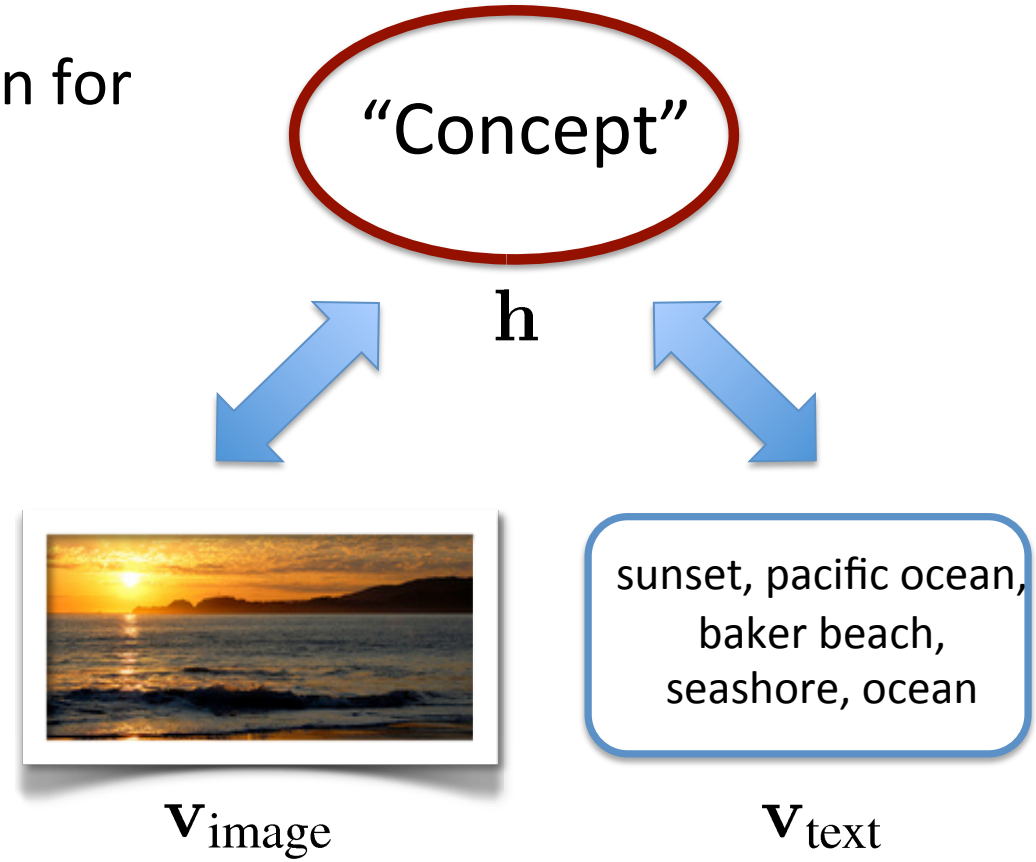
→ 

Building a Probabilistic Model

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

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

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



Building a Probabilistic Model

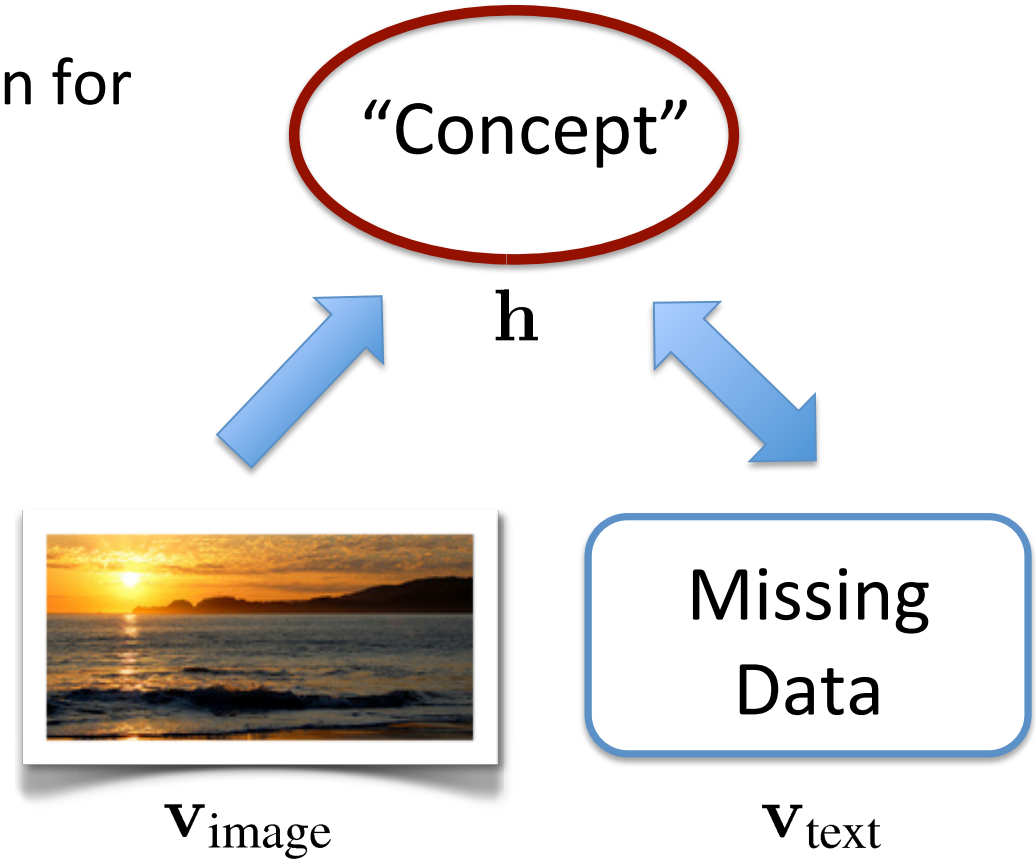
- Learn a joint density model: $P(\mathbf{h}, \mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}})$.

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

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

- Generate data from conditional distributions for

  - Image Annotation



Building a Probabilistic Model

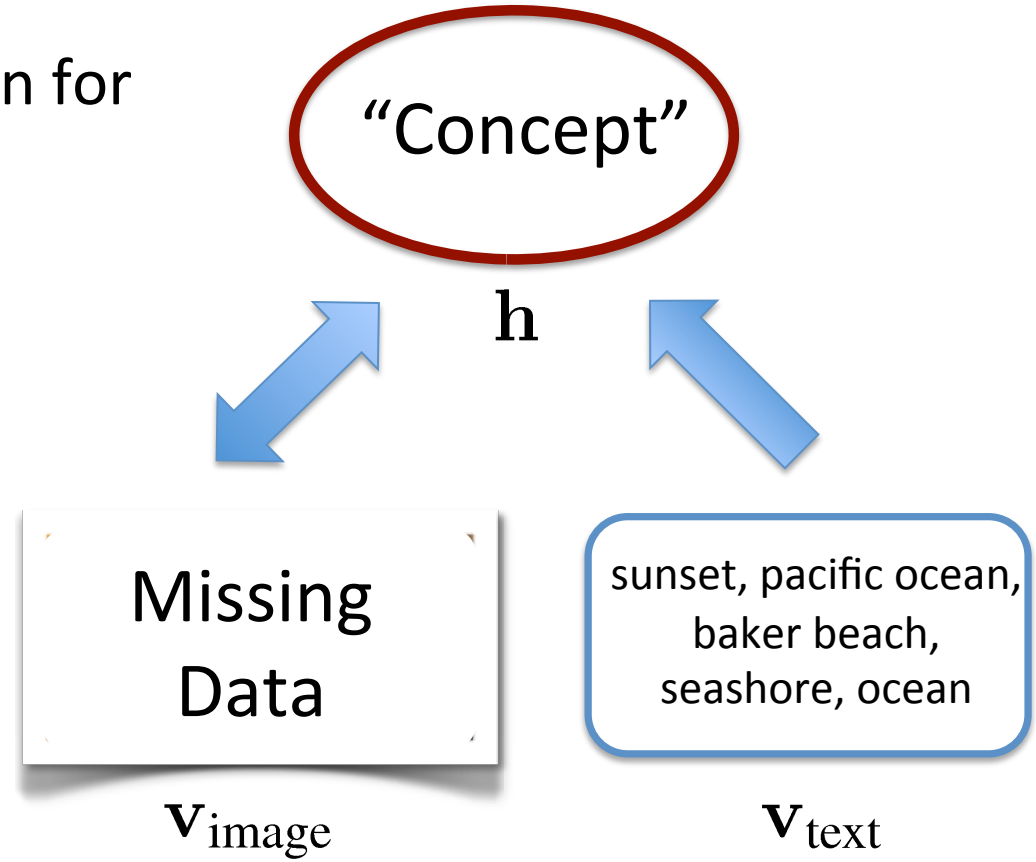
- Learn a joint density model: $P(\mathbf{h}, \mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}})$.

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

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

- Generate data from conditional distributions for

  - Image Annotation
  - Image Retrieval



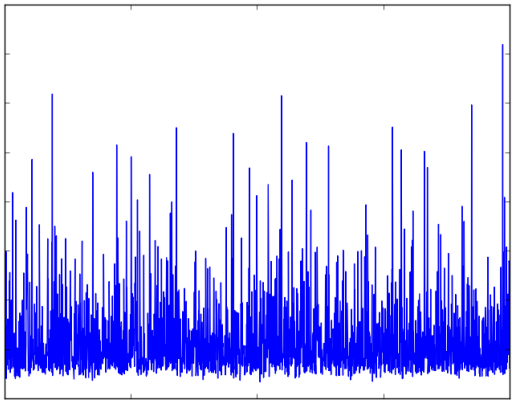
Challenges - I

Image

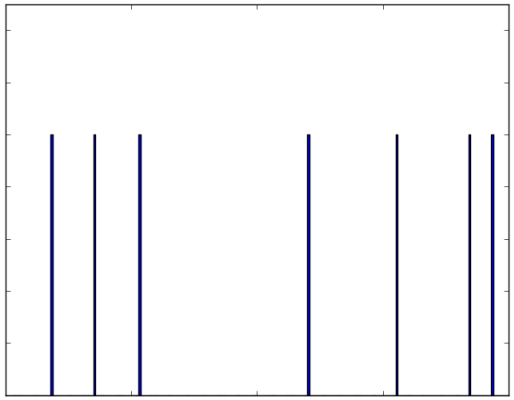
Text

sunset, 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

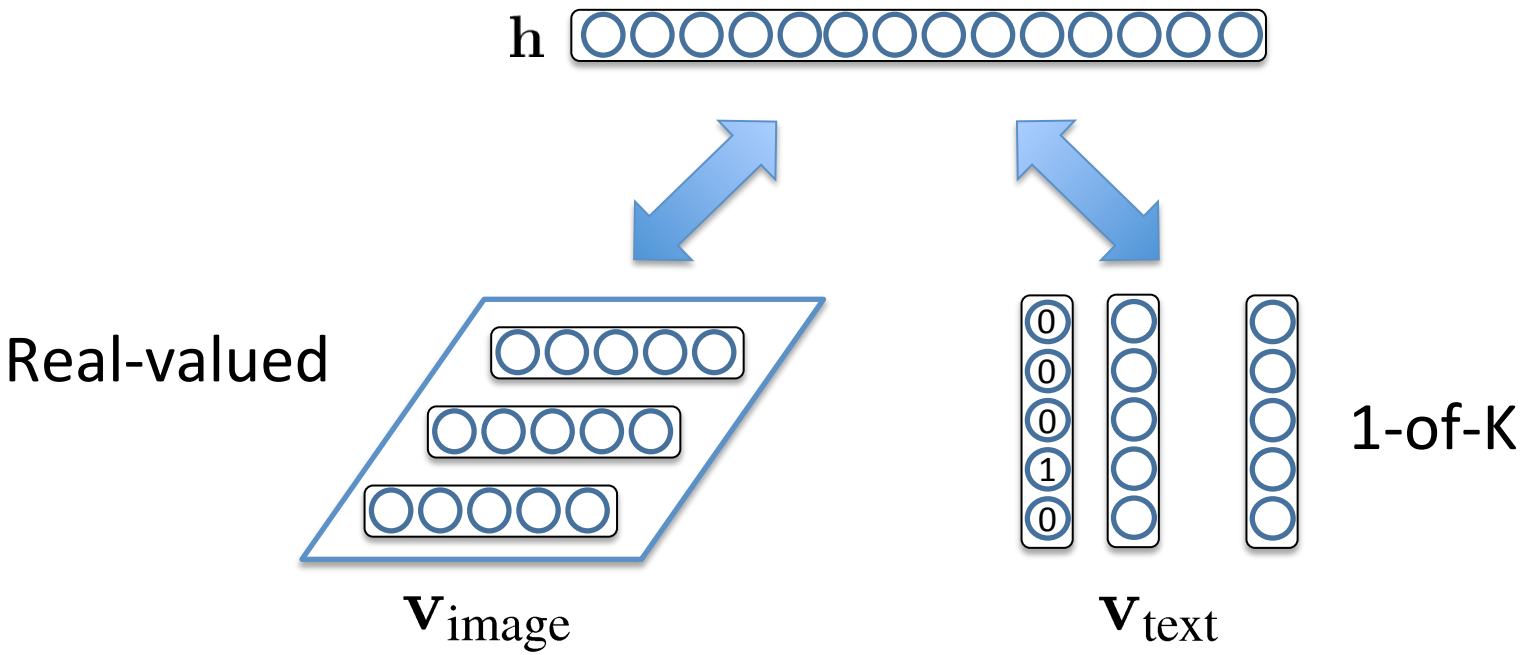
Noisy and missing data

Challenges - II

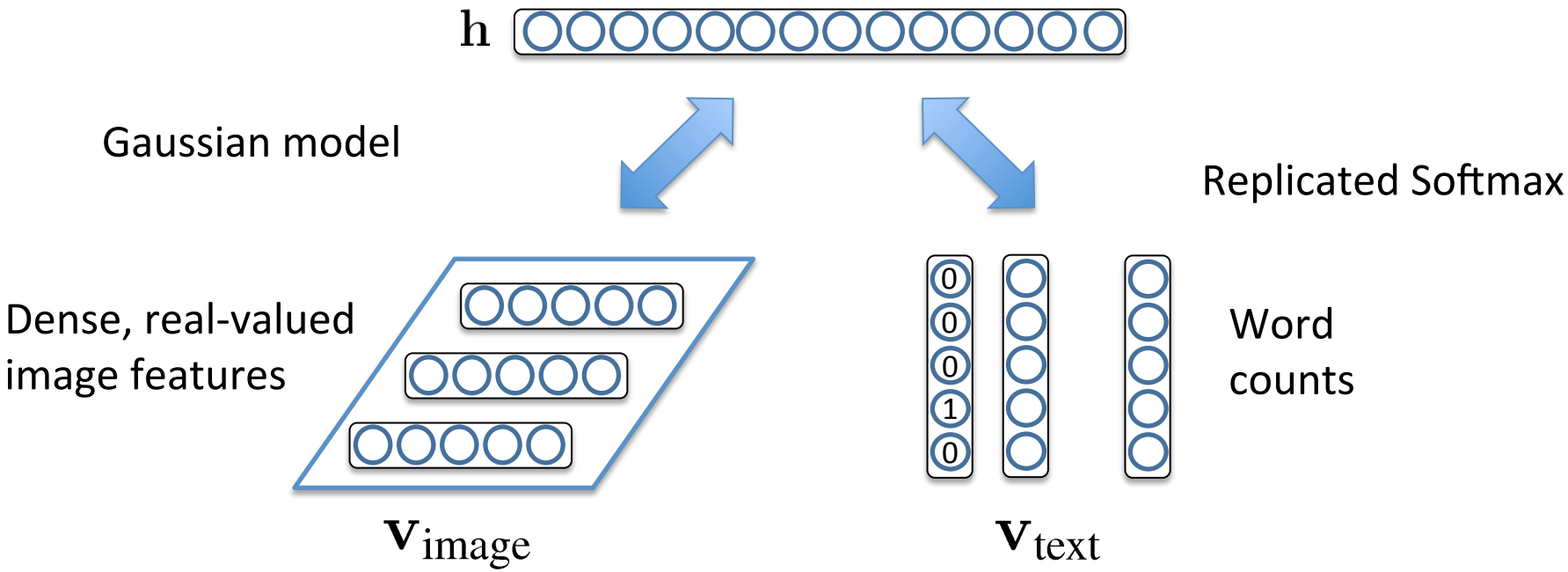
Image	Text	Text generated by the model
	pentax, k10d, pentaxda50200, kangarooisland, sa, australiansealion	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves
	mickikrimmel, mickipedia, headshot	portrait, girl, woman, lady, blonde, pretty, gorgeous, expression, model
	< no text>	night, 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.

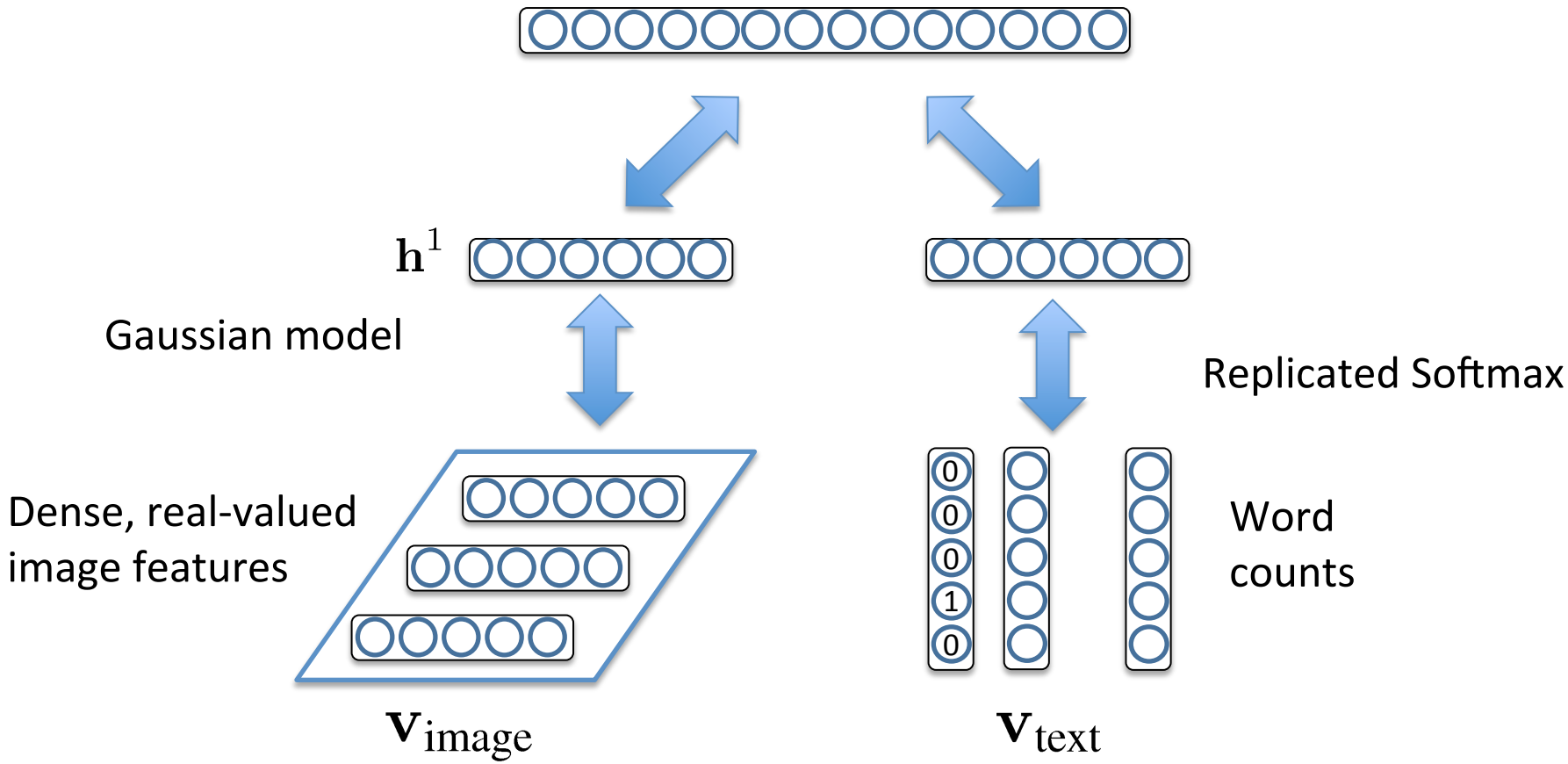


Multimodal DBM



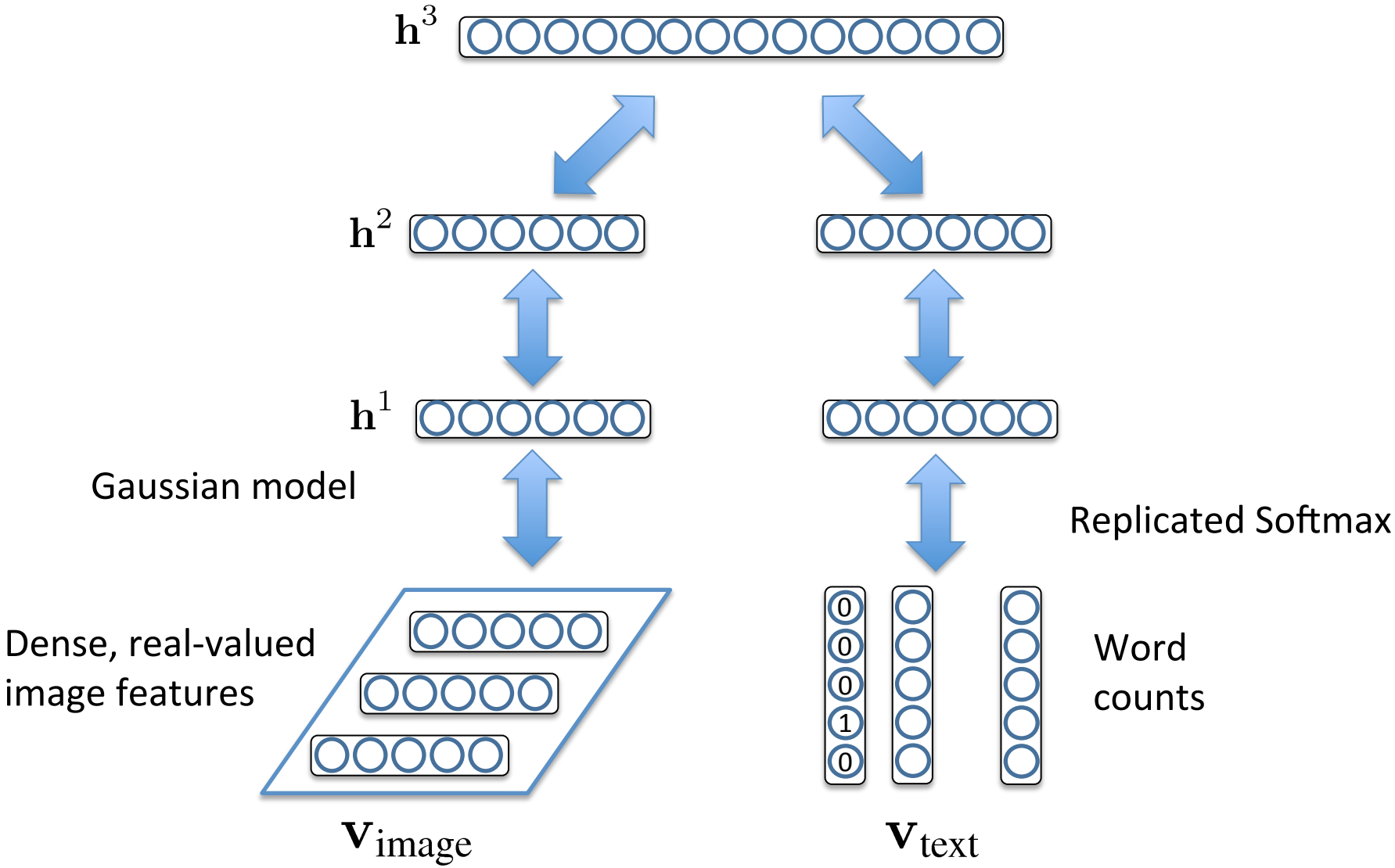
(Srivastava & Salakhutdinov, NIPS 2012)

Multimodal DBM



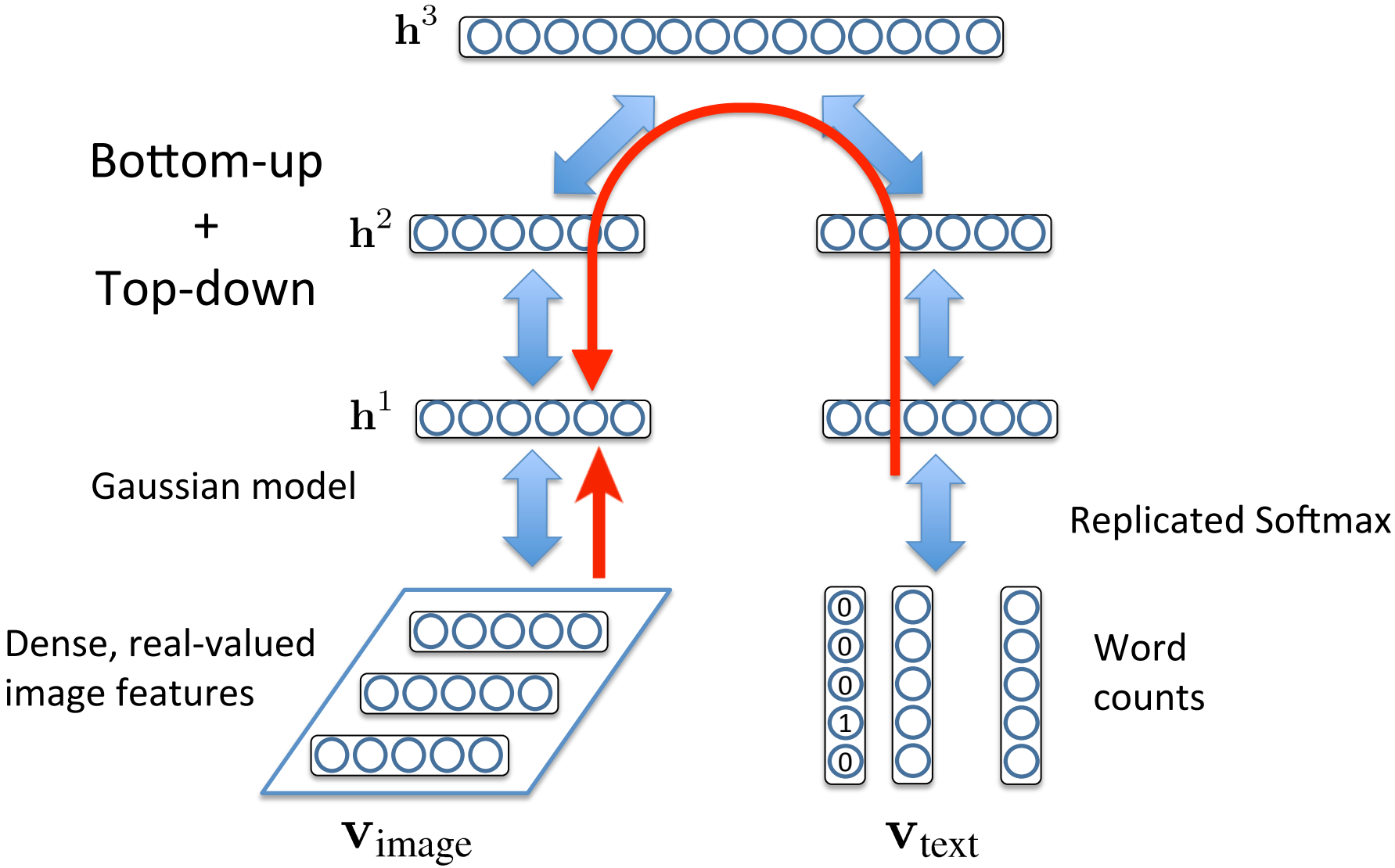
(Srivastava & Salakhutdinov, NIPS 2012)

Multimodal DBM

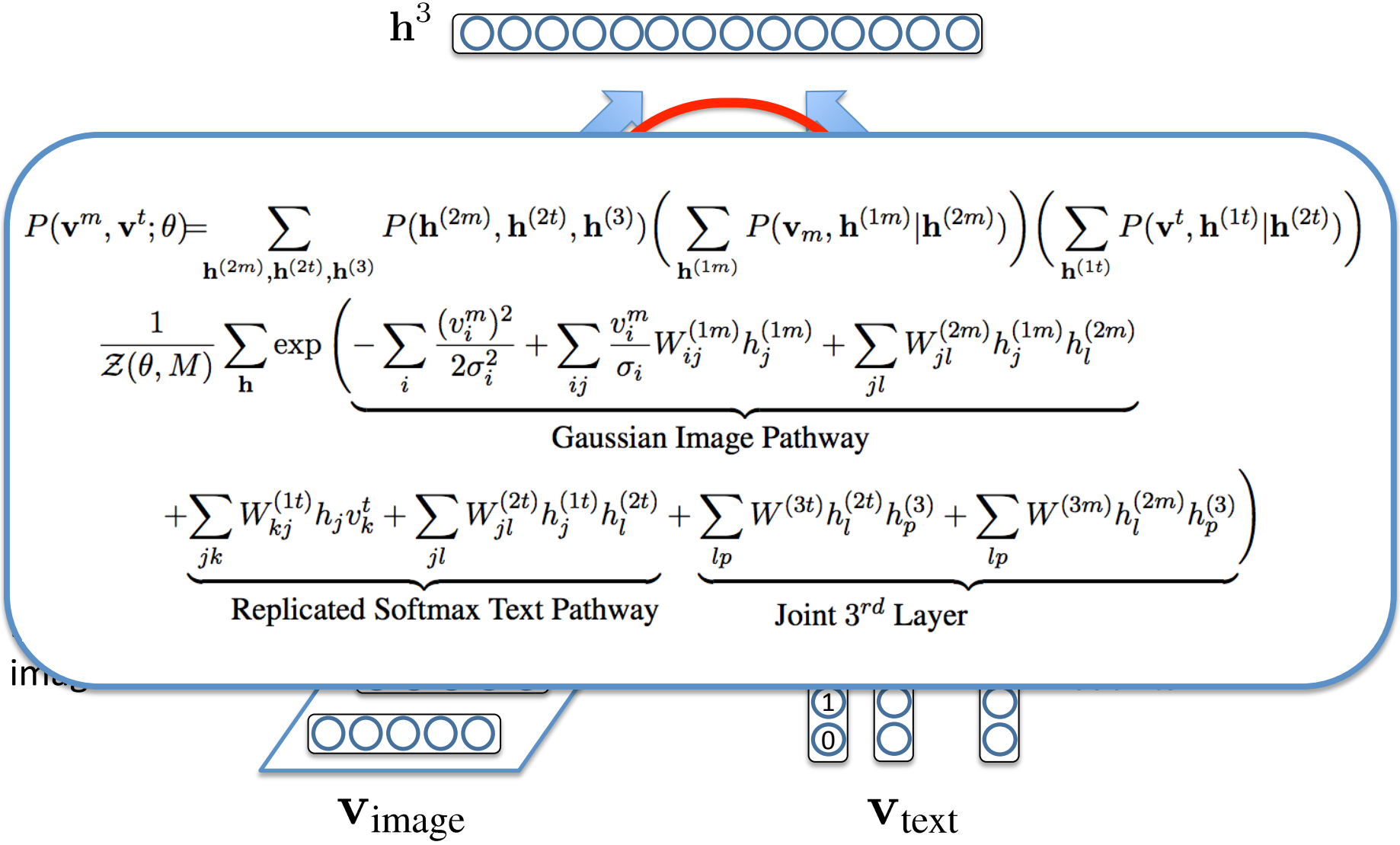


(Srivastava & Salakhutdinov, NIPS 2012)

Multimodal DBM



Multimodal DBM



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

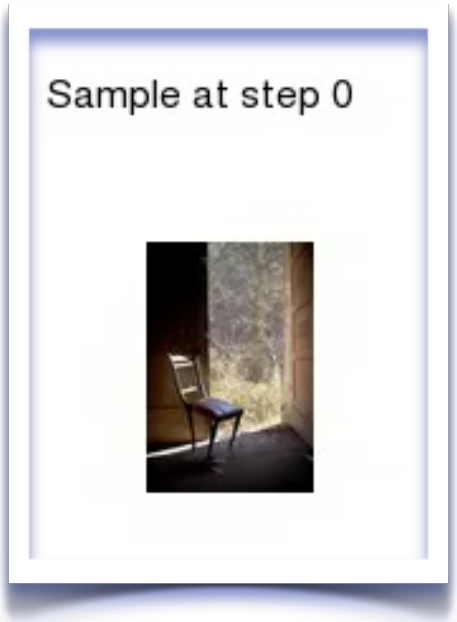
Text Generated from Images

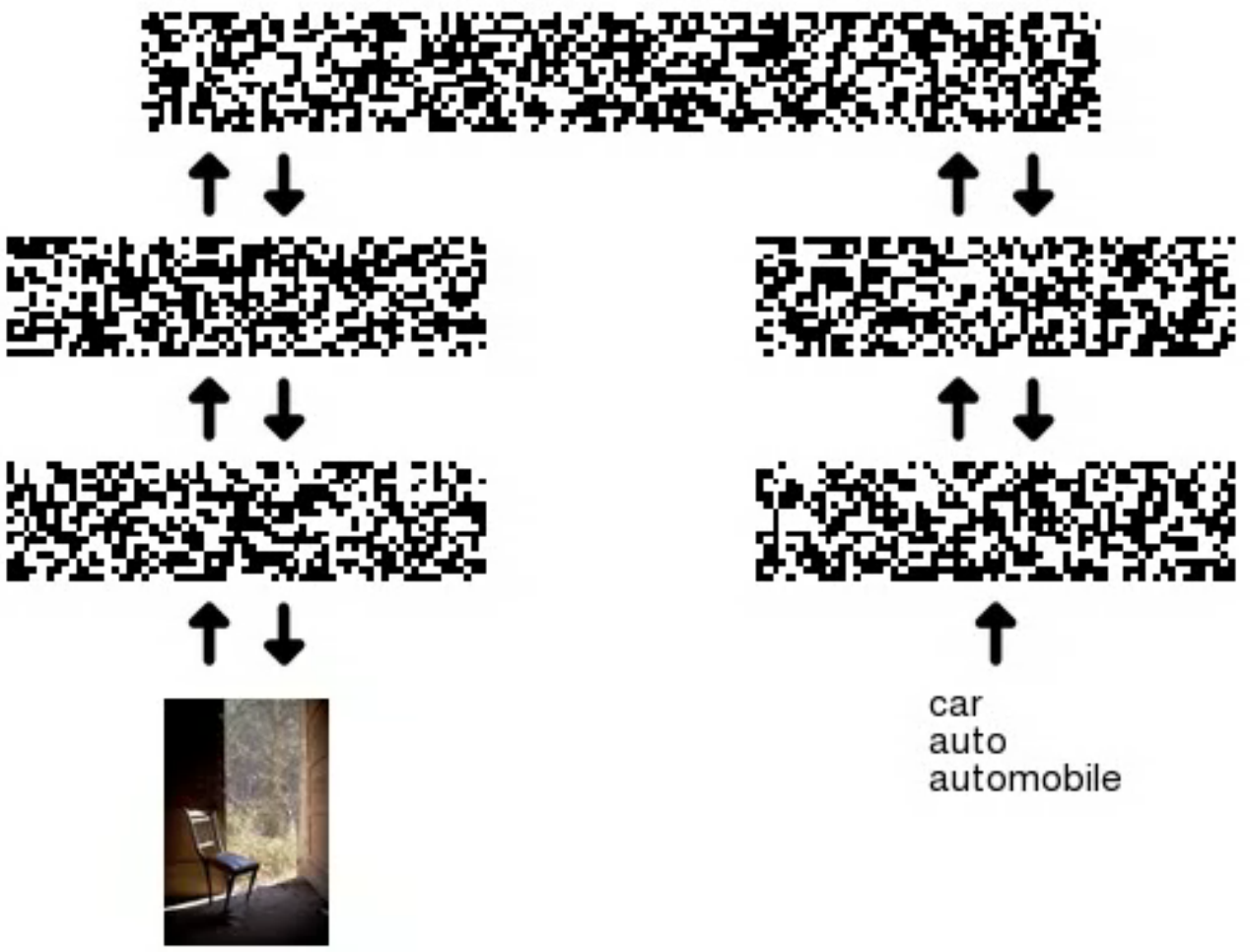
Given	Generated
	portrait, women, army, soldier, mother, postcard, soldiers
	obama, barackobama, election, politics, president, hope, change, sanfrancisco, convention, rally
	water, glass, beer, bottle, drink, wine, bubbles, splash, drops, drop

Images from Text

Step 0

Sample drawn after every 50 steps of Gibbs sampling





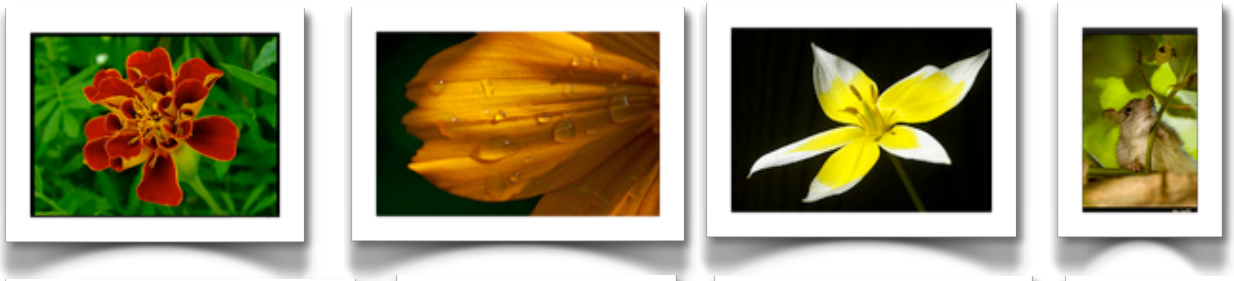
Images from Text

Given

Retrieved

water, red, sunset

nature, flower, red, green



blue, green, yellow, colors

chocolate, cake

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

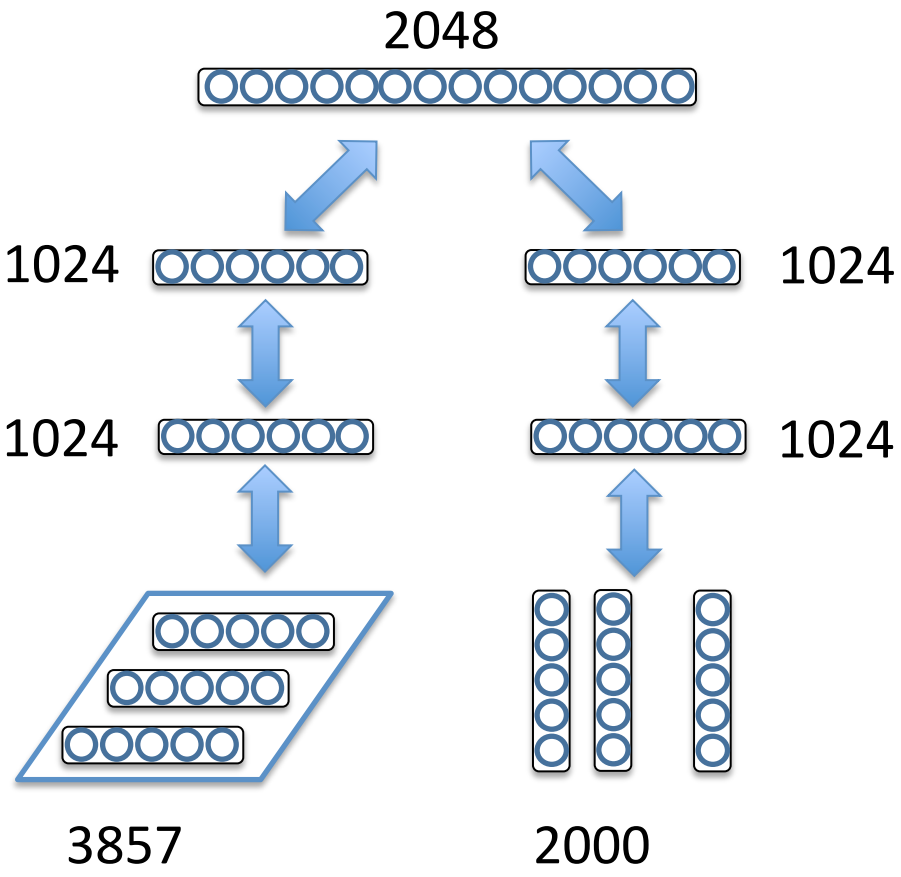
white, yellow, abstract, lines, bus, graphic

sky, geotagged, reflection, cielo, bilbao, reflejo

Huiskes et. al.

Data and Architecture

≈ 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

Labeled 25K examples

Results

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

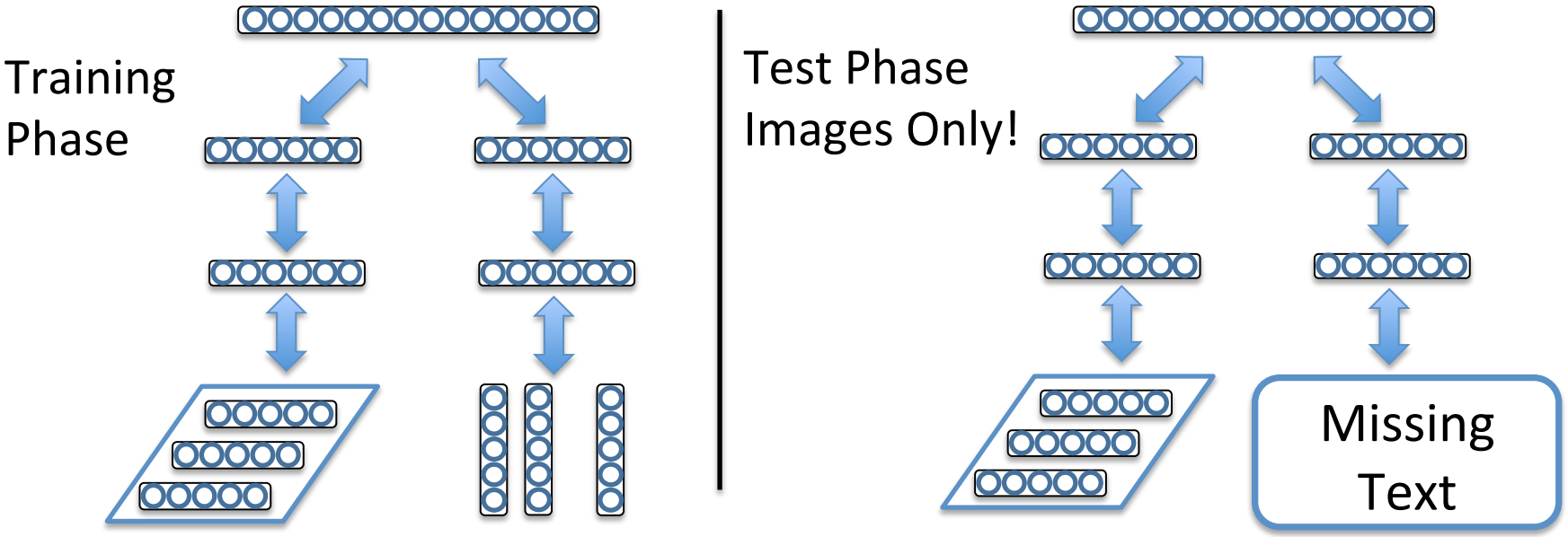
Mean Average Precision

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
Deep Belief Net	0.638	0.867
Autoencoder	0.638	0.875
DBM	0.641	0.873

Labeled 25K examples

+ 1 Million unlabelled

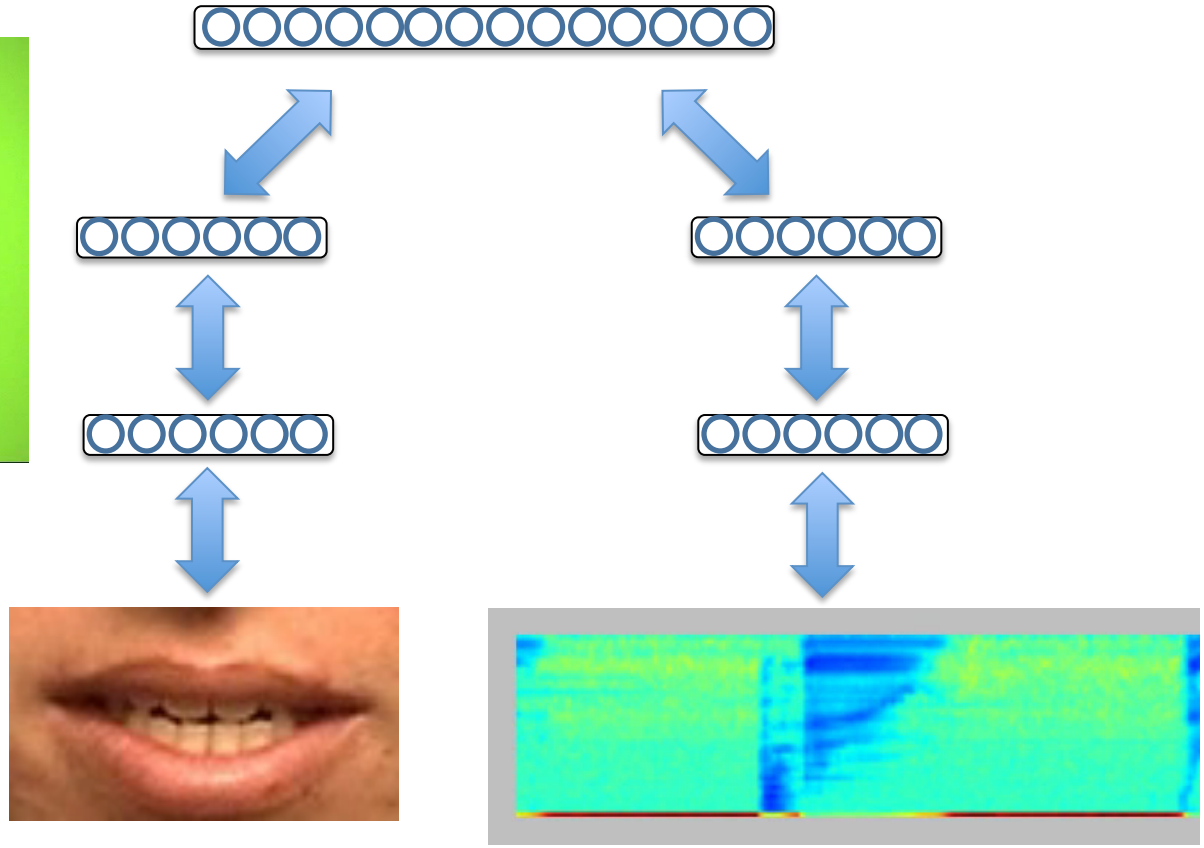
Benefits of using Multimodal Data



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



Generating Sentences

- More challenging problem.
- How can we generate complete descriptions of images?

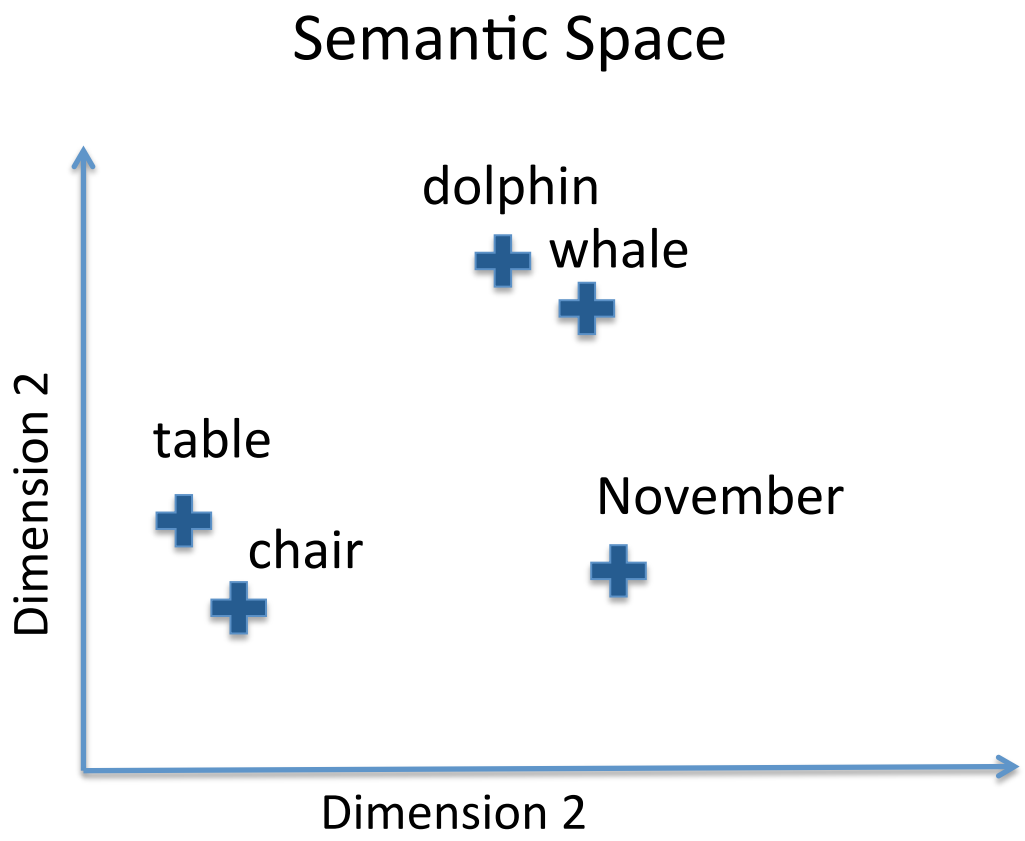
Input

Output

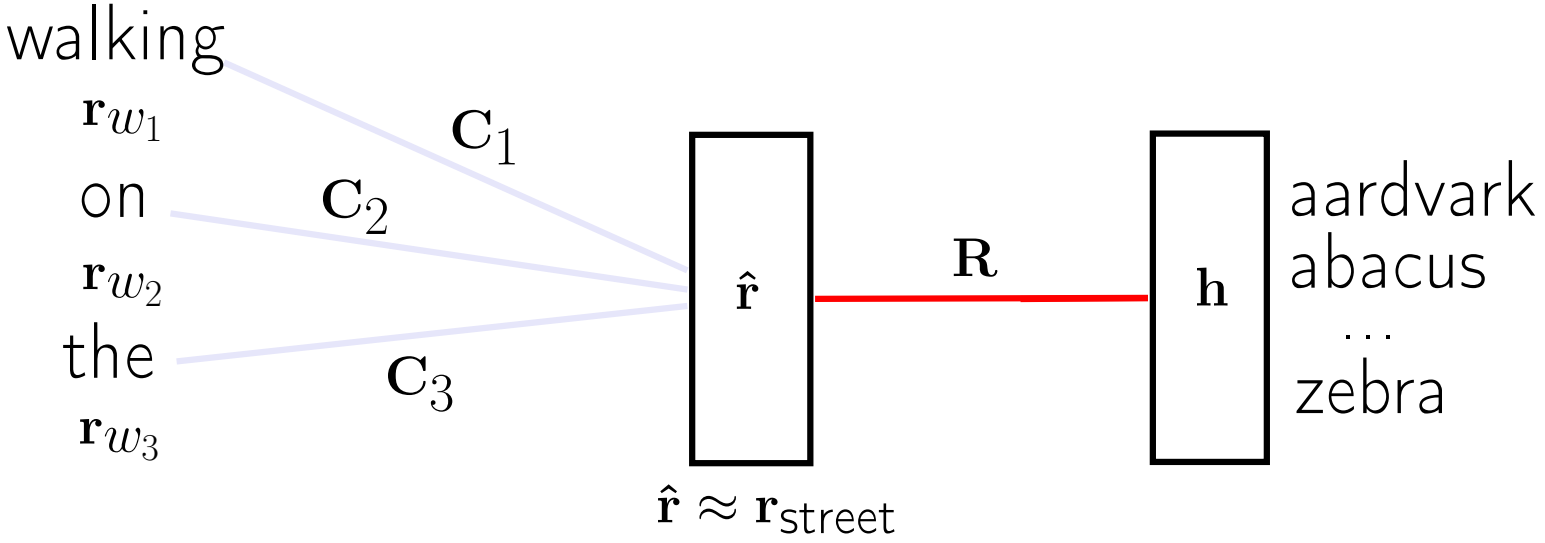
A man skiing down the snow covered mountain with a dark sky in the background.

Log-bilinear Neural Language Model

- Key Idea: Each word w is represented as a D-dimensional real-valued vector $r_w \in R^D$.



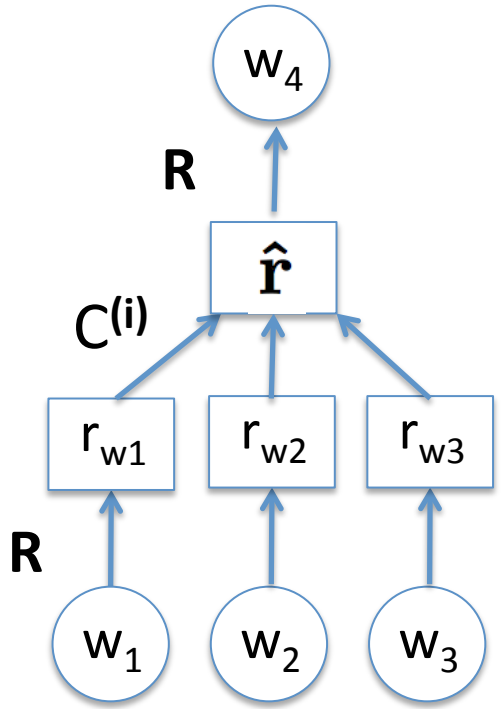
Log-bilinear Neural Language Model



- Let (w_1, …, w_{n-1}) be a tuple of n-1 words, where n-1 is the context size.
- Then, a linear prediction of the next word representation is:

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \underbrace{\mathbf{C}^{(i)}}_{\text{K} \times \text{K context parameter matrices}} \mathbf{r}_{w_i},$$

Log-bilinear Neural Language Model



1-of-V encoding of words

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{r}_{w_i},$$

Predicted representation of r_{wn}.

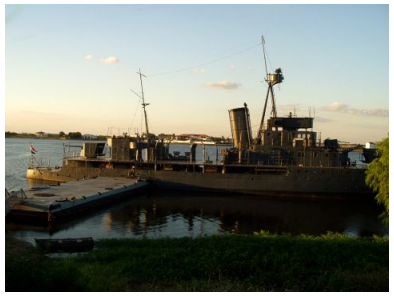
- The conditional probability of the next word given by:

$$P(w_n = i | w_{1:n-1}) = \frac{\exp(\hat{\mathbf{r}}^T \mathbf{r}_i + b_i)}{\sum_{j=1}^{V} \exp(\hat{\mathbf{r}}^T \mathbf{r}_j + b_j)}$$

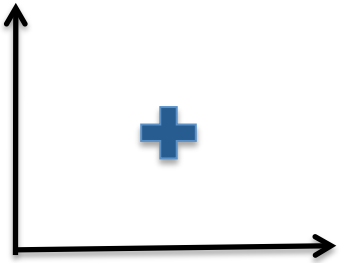
V: vocabulary size

Bengio, 2003, Mnih 2008, Mikolov, 2009, Kiros 2014

An Image-Text Encoder-Decoder



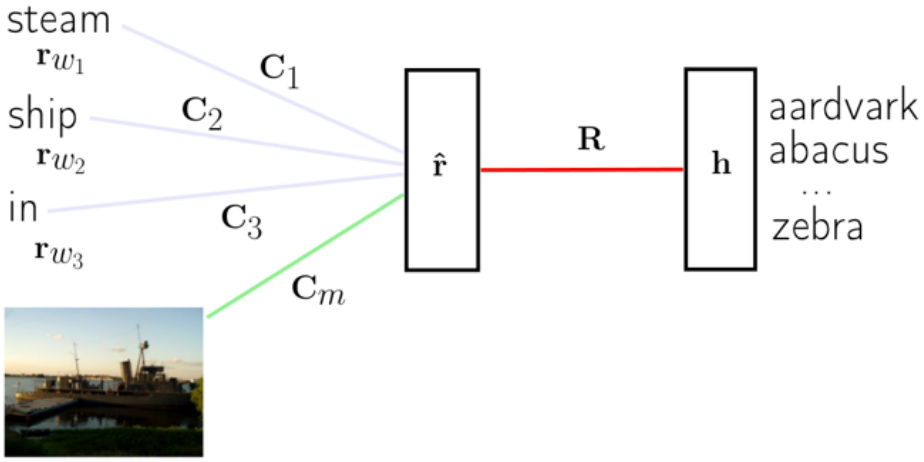
Feature space



steam ship in the water

Encoder: convolutional neural net

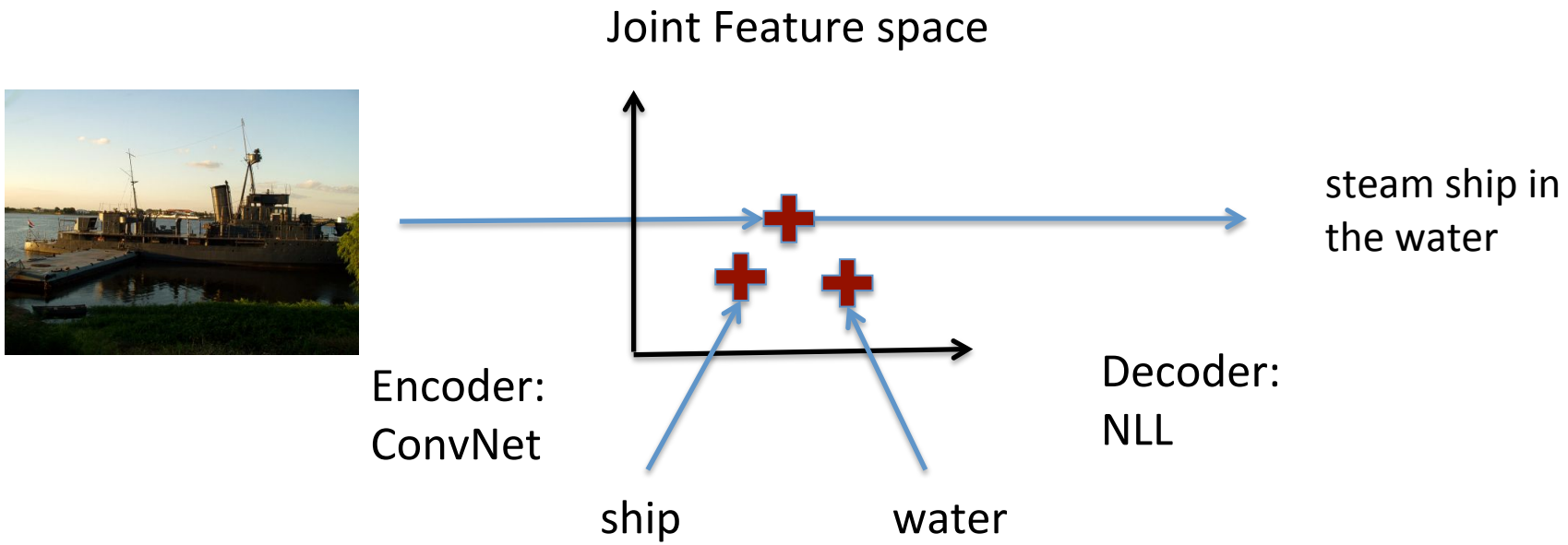
Decoder: neural language model



Use the image features to additively bias the prediction of the next word representation

Multimodal Neural Language Models (Kiros, et.al., ICML 2014)

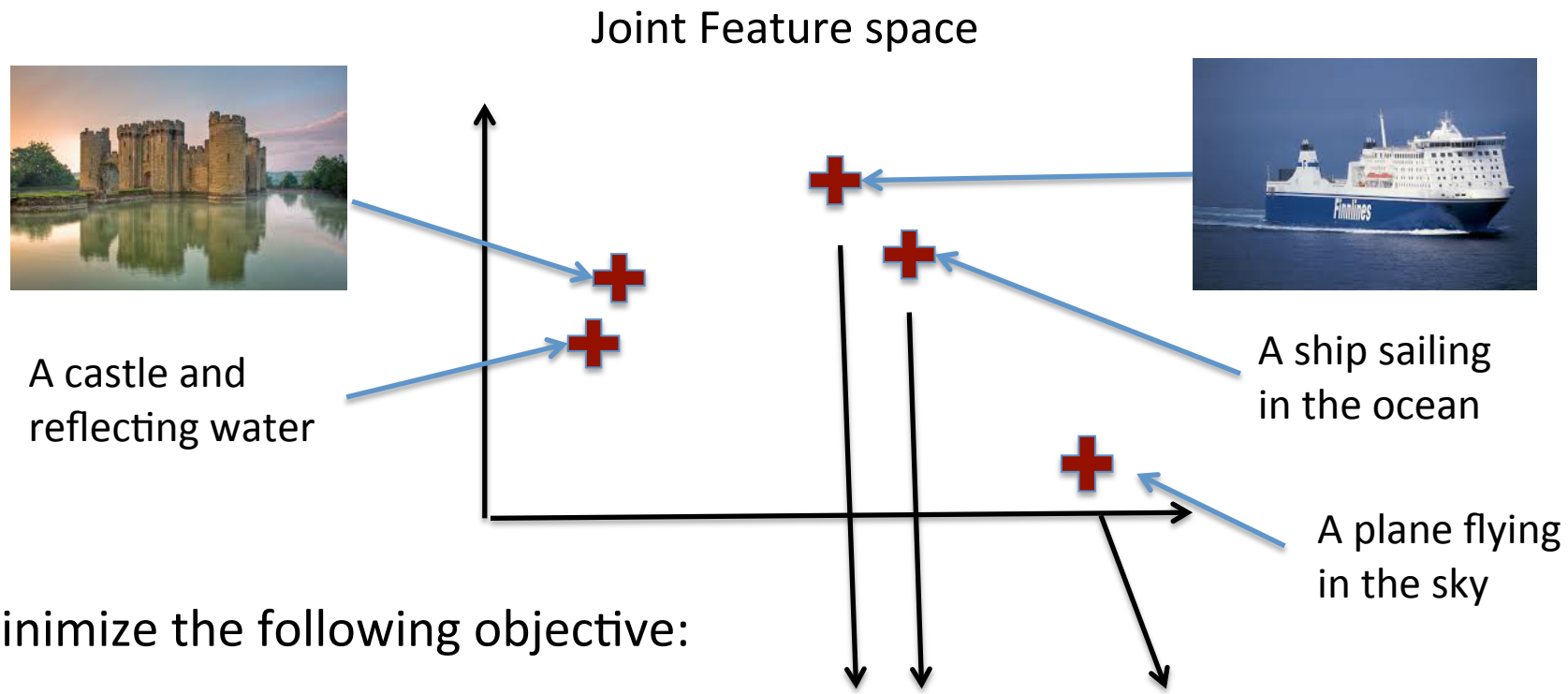
An Image-Text Encoder-Decoder



- Learn a joint embedding space of images and text:
  - Can condition on anything (images, words, phrases, etc)
  - Natural definition of a scoring function (inner products in the joint space)
  - Use a new language model that incorporates additional structure

Socher 2013, Frome 2013, Kiros 2014

An Image-Text Encoder-Decoder



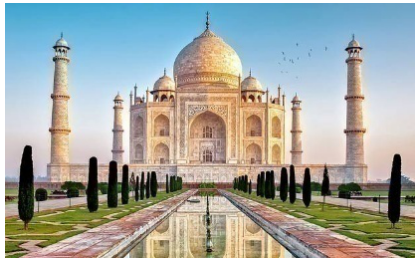
Minimize the following objective:

Images: $$\sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} +$$

Text: $$\sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\}$$

Slide Credit: Ryan Kiros

Tagging and Retrieval



mosque, tower, building, cathedral, dome, castle

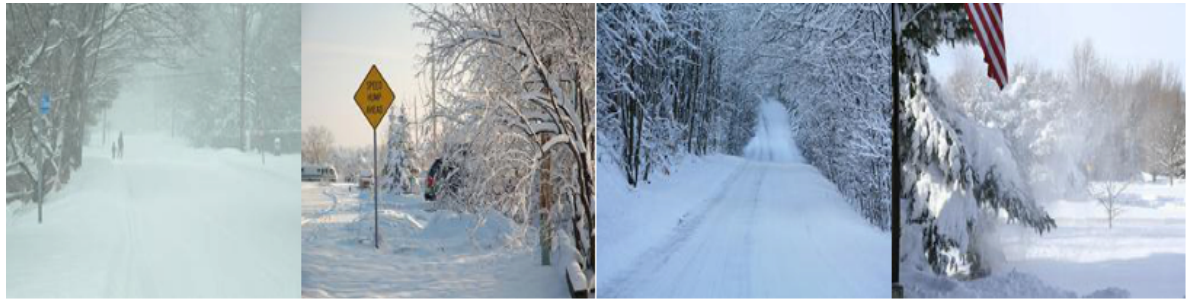
ski, skiing, skiers, skiiers, snowmobile

kitchen, stove, oven, refrigerator, microwave

bowl, cup, soup, cups, coffee

beach

snow



Ryan Kiros, 2014

Retrieval with Adjectives

fluffy

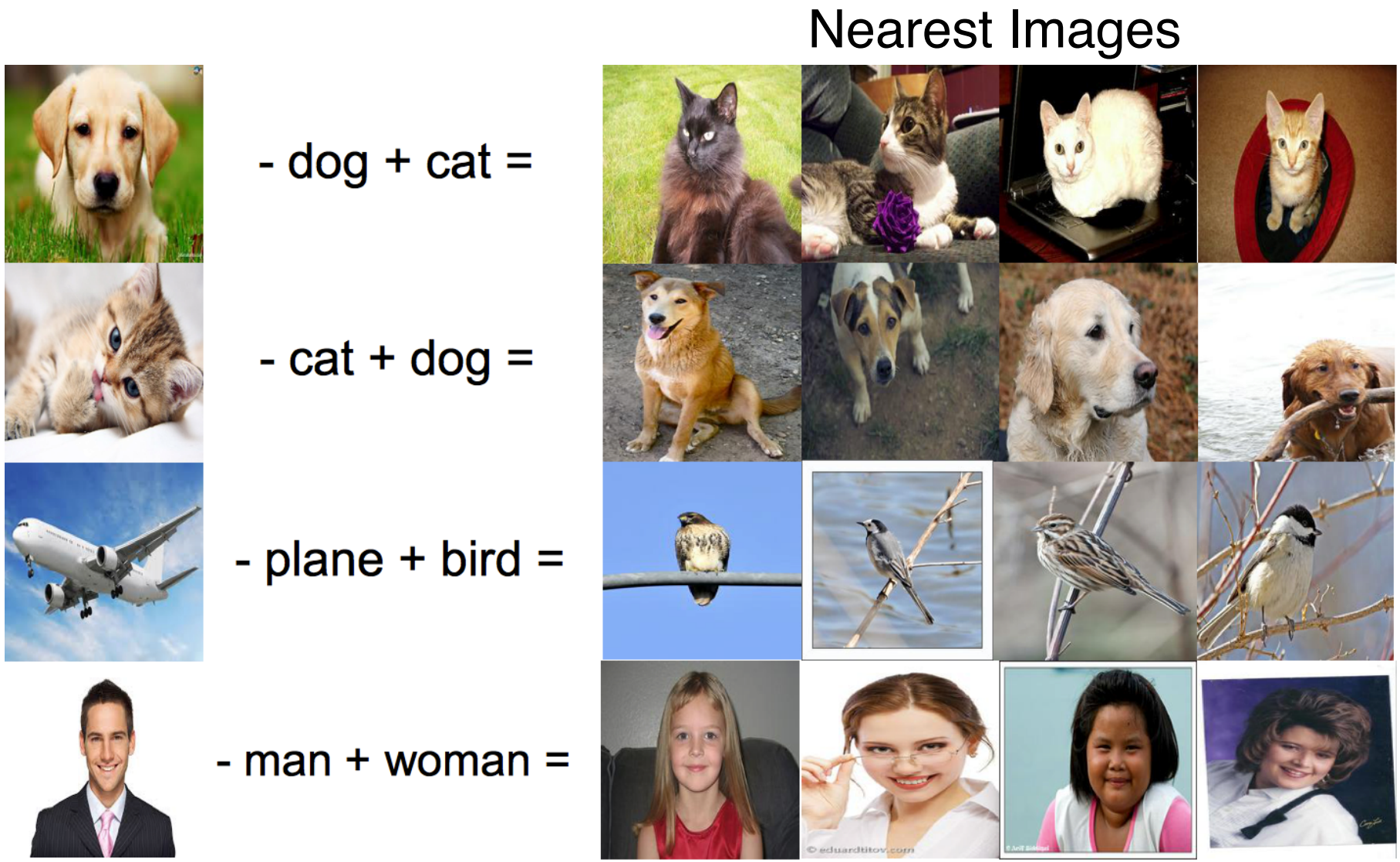
delicious

adorable

sexy

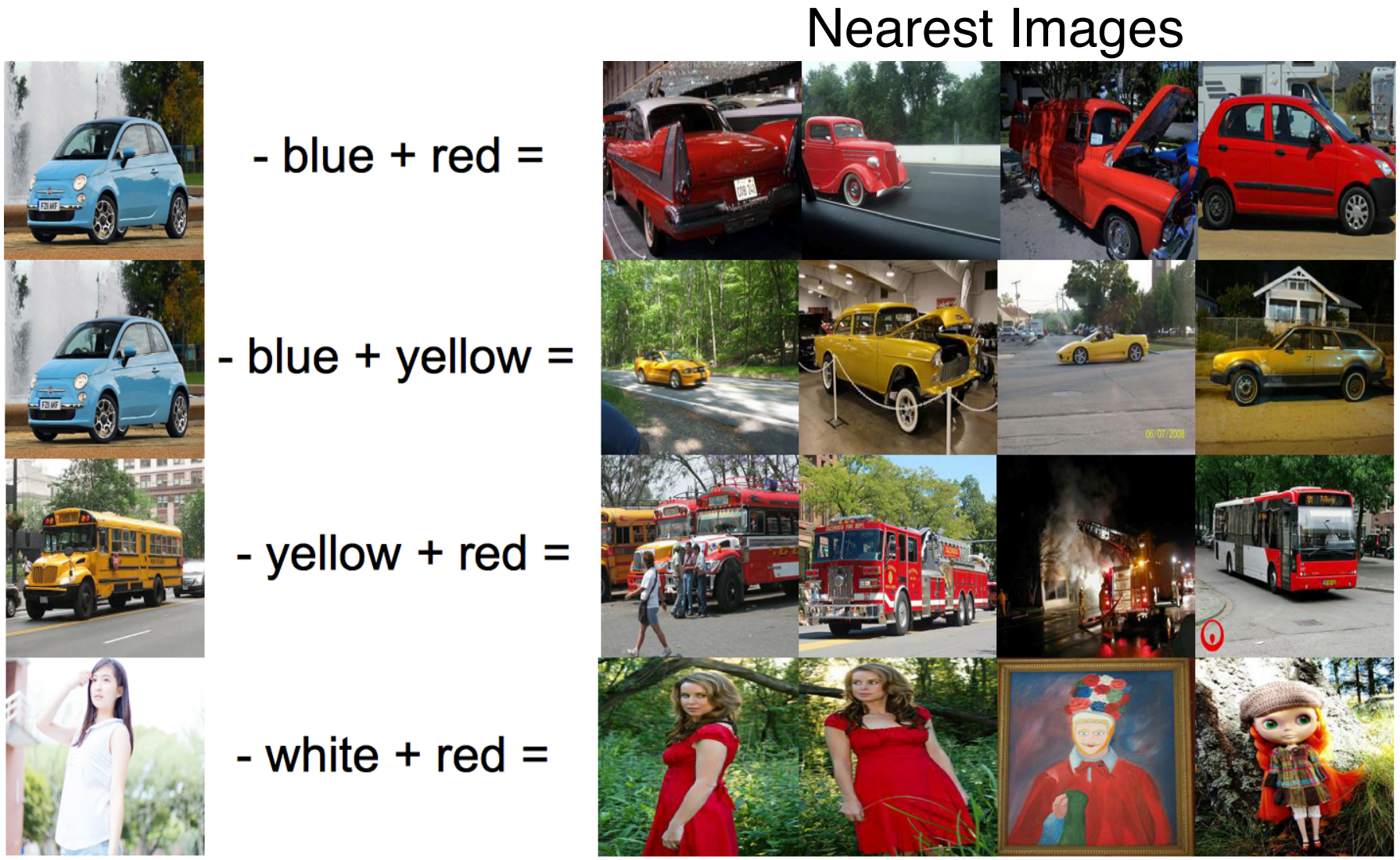
Ryan Kiros, 2014

Multimodal Linguistic Regularities



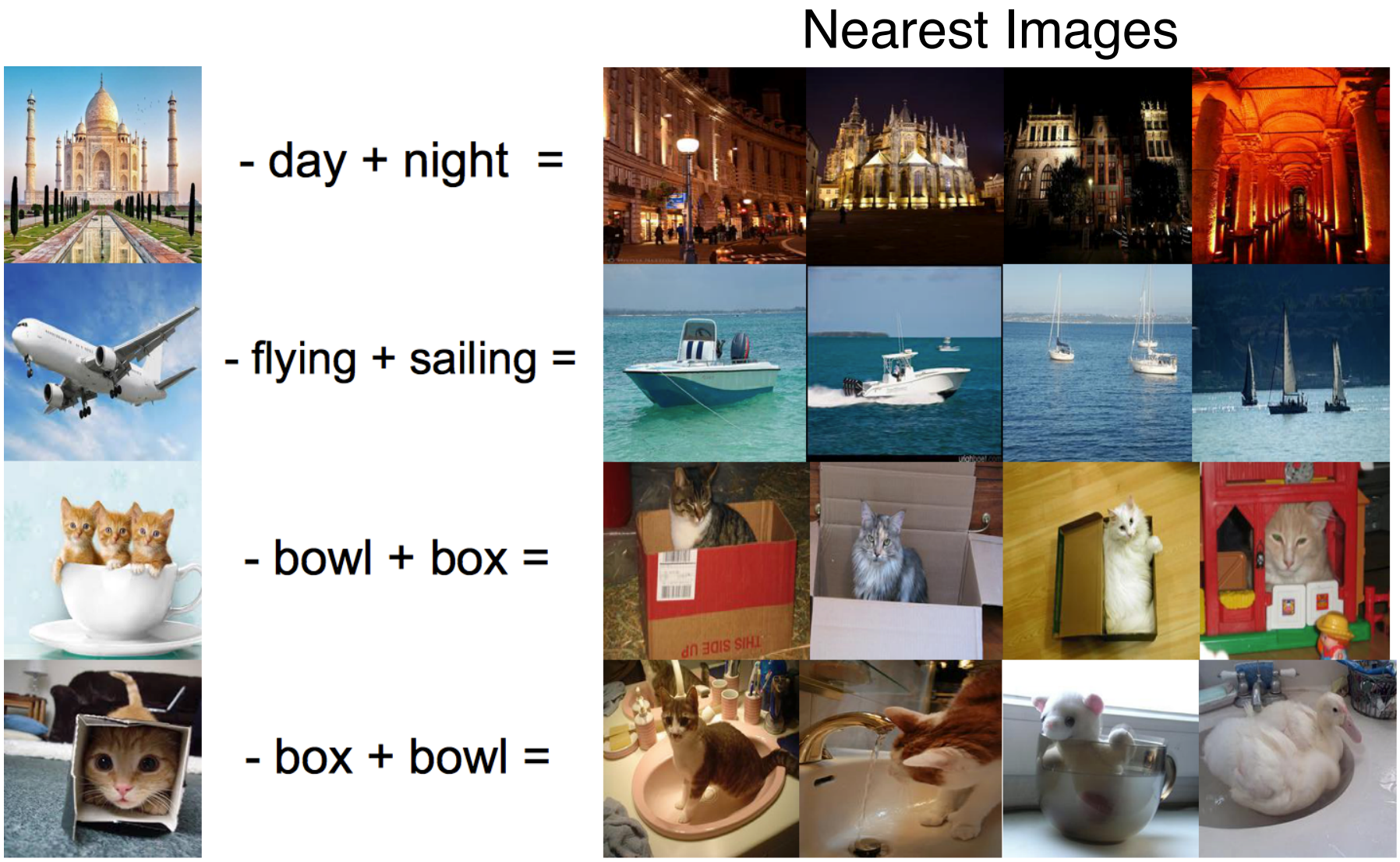
Ryan Kiros, 2014

Multimodal Linguistic Regularities



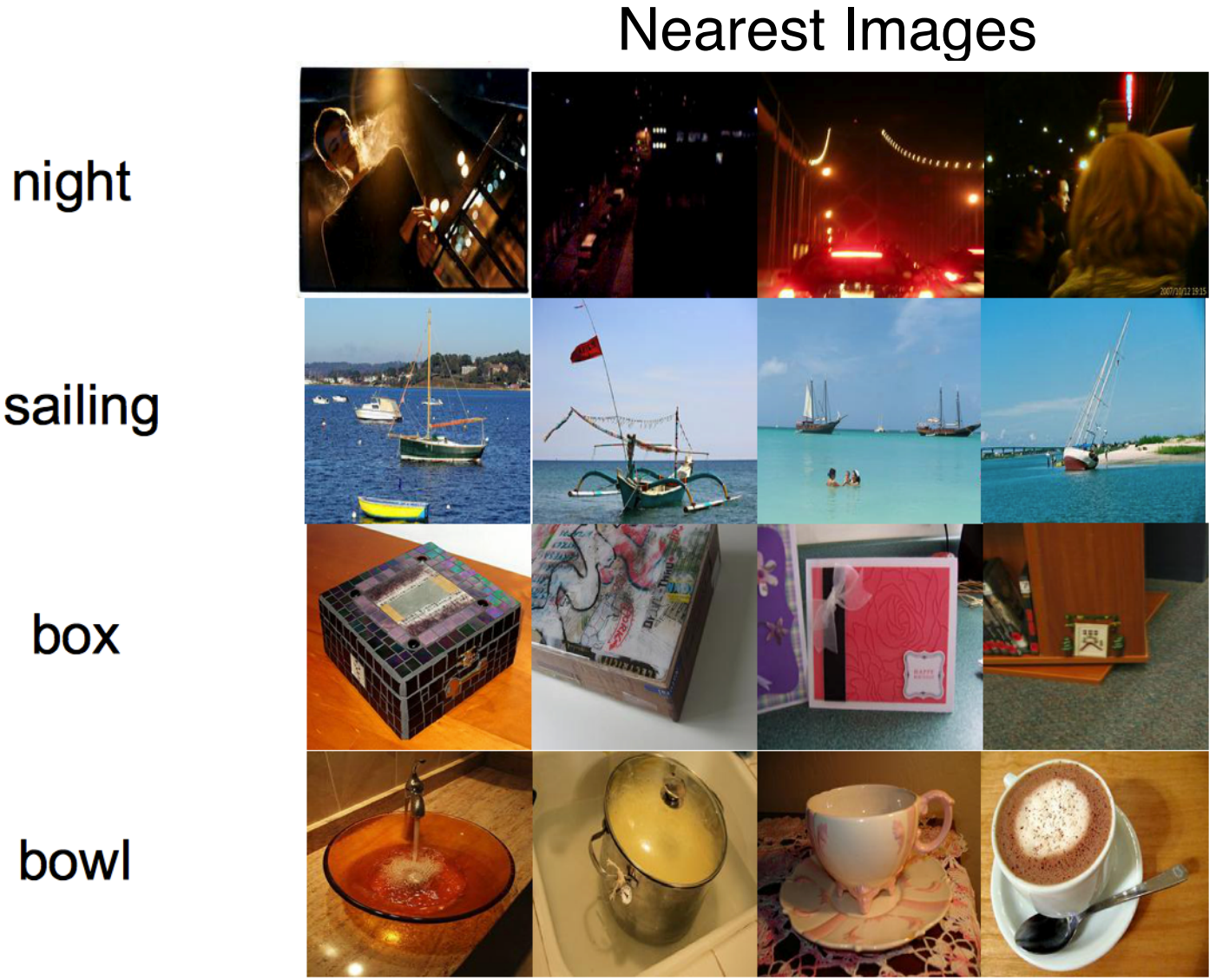
Ryan Kiros, 2014

Multimodal Linguistic Regularities



Ryan Kiros, 2014

Sanity Check



More Examples

cup, bowl, coffee, soup, cups

yummy, delicious, plastic, foamy, savory

Denotations:

coffee cup
cup of coffee
espresso cup
styrofoam coffee cup
cup of espresso

Top-5 Model Samples

from a plastic cup of tea.
sweet cup of tea in my kitchen.
we had a cup of delicious.
cups of tea in red wine.
cup of red wine in our kitchen.

More Examples

cat, ferret, hamster, weasel, puppy

cute, furry, cuddly, adorable, naughty

Denotations:

cat hide
toy cat
cat swat
baby pull cat
cat doll

Top-5 Model Samples

my cat who lives in a box.
I put his cat in the world.
cute little cat in the box.
hanging around the cat in Santa Monica.
kitty cat toys in the box.

More Examples

spider, spiders, arachnid,
insects, insect

male, female, creepy, spooky,
elfin

Denotations:

spider web
giant spider
have spider web
toy spider
hold spider

Top-5 Model Samples

giant spider found in the Netherlands.
look at the new spider web.
this was near the black spider web.
I like the spider.
the pattern of one spider web.

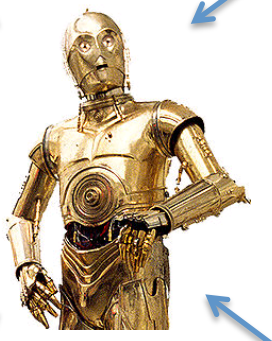
Multi-Modal Models

Images

Text & Language

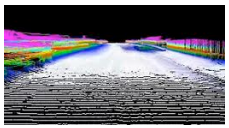
REUTERS

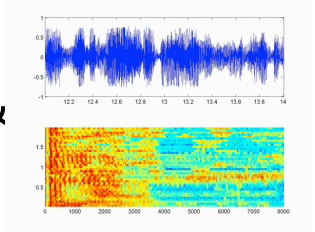
AP Associated Press



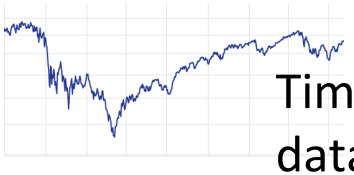
Video

Laser scans





Speech & Audio



Time series data

Develop learning systems that come closer to displaying human like intelligence

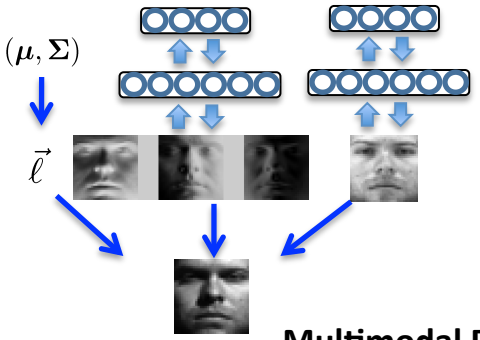
One of Key Challenges:
Inference

Summary

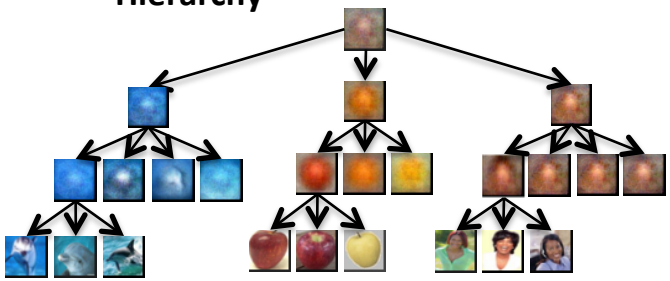
- Efficient learning algorithms for Hierarchical Generative Models. Learning more adaptive, robust, and structured representations.

Text & image retrieval / Object recognition

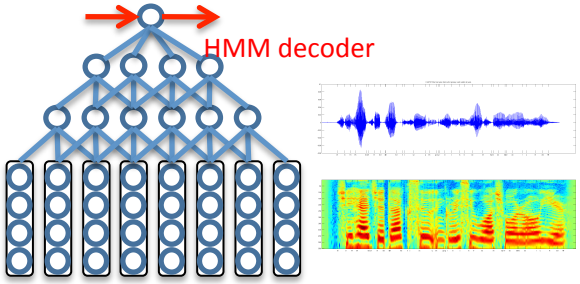
Dealing with missing/ occluded data



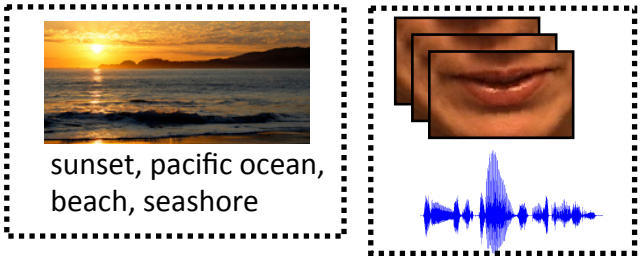
Learning a Category Hierarchy



Speech Recognition



Multimodal Data



Object Detection

- Deep models can improve current state-of-the art in many application domains:
  - Object recognition and detection, text and image retrieval, handwritten character and speech recognition, and others.

Thank you

Code is available at:
http://deeplearning.cs.toronto.edu/