Deep Learning

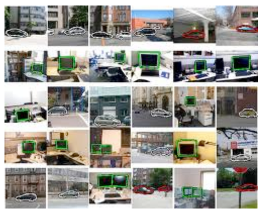
Russ Salakhutdinov

Department of Statistics and Computer Science
University of Toronto

Mining for Structure

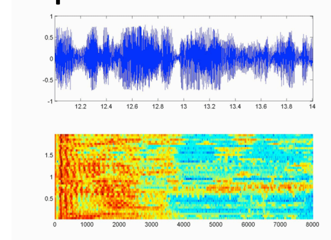
Massive increase in both computational power and the amount of data available from web, video cameras, laboratory measurements.

Images & Video

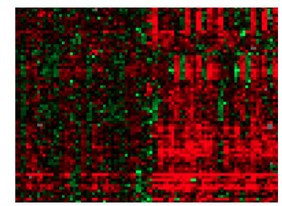


Text & Language

Speech & Audio

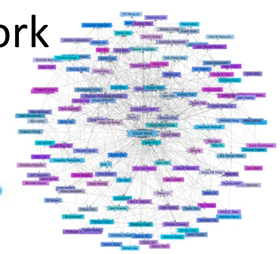


Gene Expression

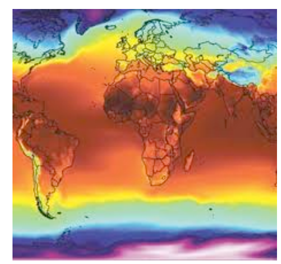


Product Recommendation

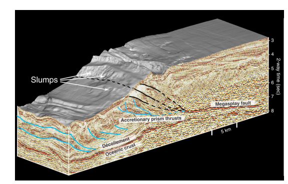
Relational Data/ Social Network



Climate Change



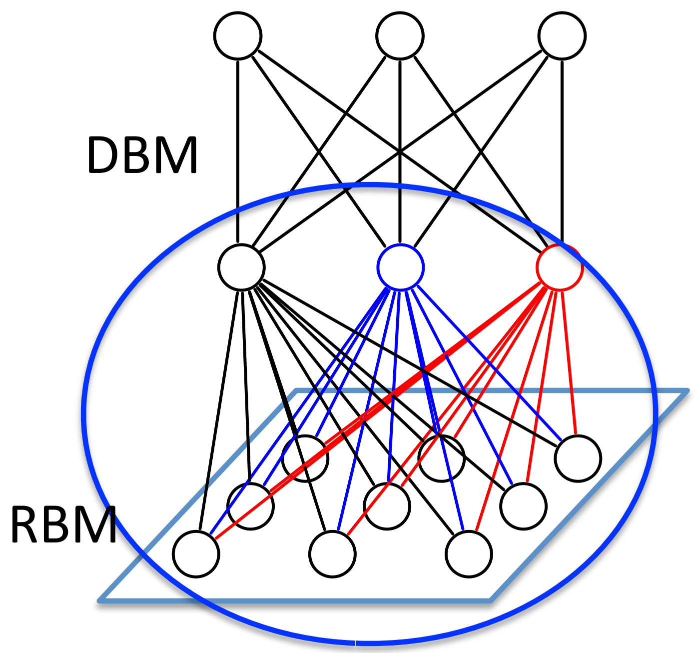
Geological Data



Mostly Unlabeled

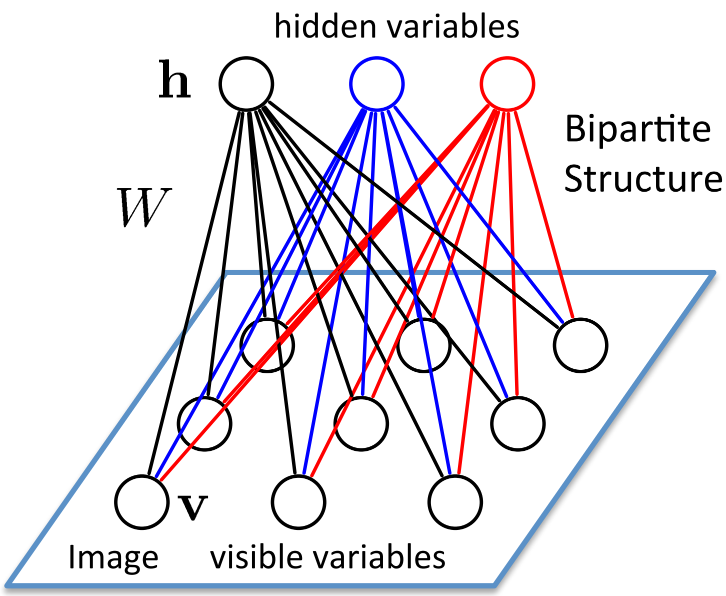
- Develop statistical models that can discover underlying structure, cause, or statistical correlation from data in **unsupervised** or **semi-supervised** way.
- Multiple application domains.

Talk Roadmap



- Unsupervised Feature Learning
  - Restricted Boltzmann Machines
  - Deep Belief Networks
  - Deep Boltzmann Machines
- Transfer Learning with Deep Models
- Multimodal Learning

Restricted Boltzmann Machines



Stochastic binary visible variables $\mathbf{v} \in \{0,1\}^D$ are connected to stochastic binary hidden variables $\mathbf{h} \in \{0,1\}^F$.

The energy of the joint configuration:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{ij} W_{ij} v_i h_j - \sum_i b_i v_i - \sum_j a_j h_j$$

$\theta = \{W, a, b\}$ model parameters.

Probability of the joint configuration is given by the Boltzmann distribution:

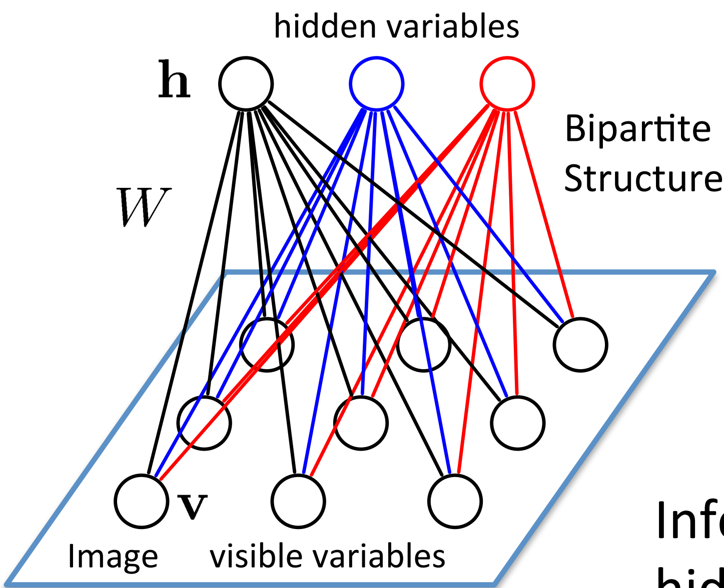
$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big) = \underbrace{\frac{1}{\mathcal{Z}(\theta)}}_{\text{partition function}} \prod_{ij} \underbrace{e^{W_{ij} v_i h_j}}_{\text{potential functions}} \prod_i e^{b_i v_i} \prod_j e^{a_j h_j}$$

$$\mathcal{Z}(\theta) = \sum_{\mathbf{h},\mathbf{v}} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big)$$

partition function   potential functions

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

Restricted Boltzmann Machines



Restricted: No interaction between hidden variables

Inferring the distribution over the hidden variables is easy:

$$P(\mathbf{h}|\mathbf{v}) = \prod_j P(h_j|\mathbf{v}) \quad P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp(-\sum_i W_{ij} v_i - a_j)}$$

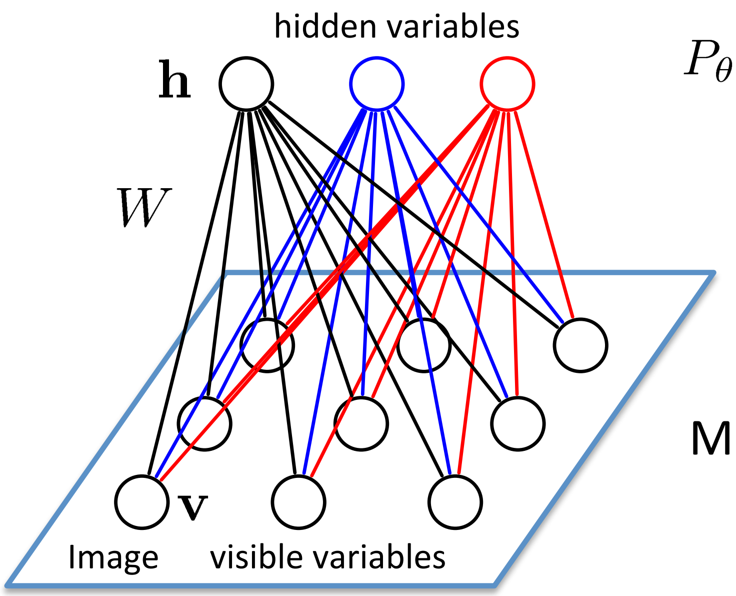
Factorizes: Easy to compute

Similarly:

$$P(\mathbf{v}|\mathbf{h}) = \prod_i P(v_i|\mathbf{h}) \quad P(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(-\sum_j W_{ij} h_j - b_i)}$$

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

Model Learning



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

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 (penalized) log-likelihood objective:

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N} \log P_\theta(\mathbf{v}^{(n)}) - \underbrace{\frac{\lambda}{N}||W||_F^2}_{\text{Regularization}}$$

Regularization

Derivative of the log-likelihood:

$$\frac{\partial L(\theta)}{\partial W_{ij}} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial W_{ij}} \log\left(\sum_{\mathbf{h}} \exp\left[\mathbf{v}^{(n)\top} W \mathbf{h} + \mathbf{a}^\top \mathbf{h} + \mathbf{b}^\top \mathbf{v}^{(n)}\right]\right) - \frac{\partial}{\partial W_{ij}} \log \mathcal{Z}(\theta) - \frac{2\lambda}{N} W_{ij}$$

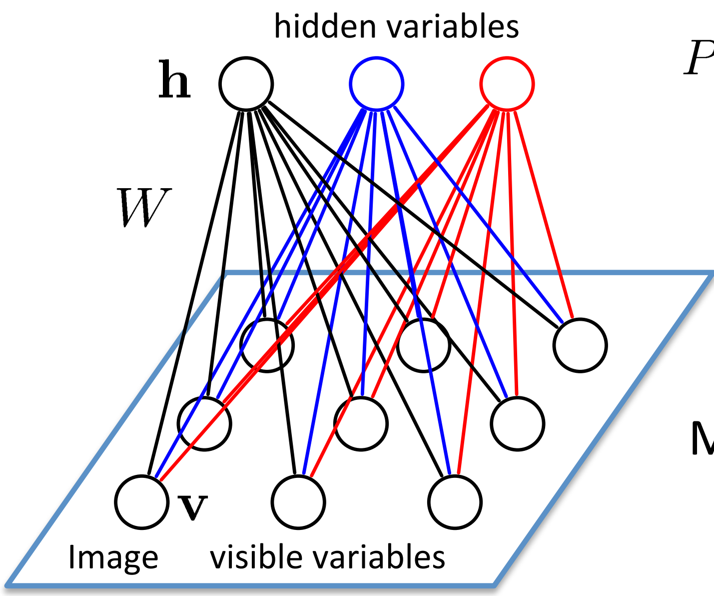
$$= \mathrm{E}_{P_{data}}[v_i h_j] - \underbrace{\mathrm{E}_{P_\theta}[v_i h_j]} - \frac{2\lambda}{N} W_{ij}$$

Difficult to compute: exponentially many configurations

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

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

Model Learning



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

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 (penalized) log-likelihood objective:

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log P_\theta(\mathbf{v}^{(n)}) - \frac{\lambda}{N} ||W||_F^2$$

Derivative of the log-likelihood:

$$\frac{\partial L(\theta)}{\partial W_{ij}} = \mathrm{E}_{P_{data}}[v_i h_j] - \mathrm{E}_{P_\theta}[v_i h_j] - \frac{2\lambda}{N} W_{ij}$$

Approximate maximum likelihood learning:

Contrastive Divergence (Hinton 2000)
MCMC-MLE estimator (Geyer 1991)

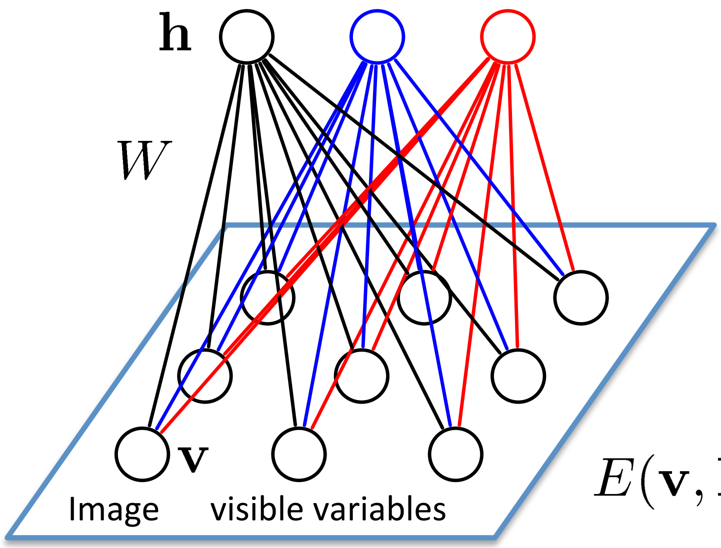
Tempered MCMC
(Salakhutdinov, NIPS 2009)

Pseudo Likelihood (Besag 1977)
Composite Likelihoods (Lindsay, 1988; Varin 2008)

Adaptive MCMC
(Salakhutdinov, ICML 2010)

RBMs for Images

Gaussian-Bernoulli RBM:



$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)}\exp(-E(\mathbf{v},\mathbf{h};\theta))$$

Define energy functions for various data modalities:

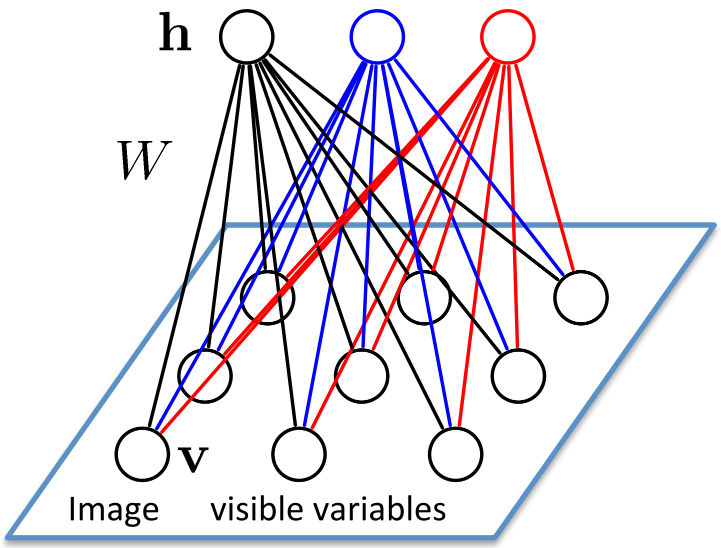
$$E(\mathbf{v},\mathbf{h};\theta) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{ij} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_j a_j h_j$$

$$P(v_i = x|\mathbf{h}) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x - b_i - \sigma_i \sum_j W_{ij}h_j)^2}{2\sigma_i^2}\right)$$ Gaussian

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1+\exp(-\sum_i W_{ij}\frac{v_i}{\sigma_i} - a_j)}$$ Bernoulli

RBMs for Images

Gaussian-Bernoulli RBM:



$$P_\theta(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}; \theta))$$

Interpretation: Mixture of exponential number of Gaussians

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

where

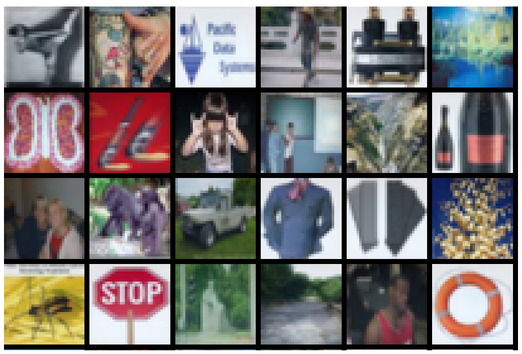
$$P_\theta(\mathbf{h}) = \int_{\mathbf{v}} P_\theta(\mathbf{v}, \mathbf{h}) d\mathbf{v}$$ is an implicit prior, and

$$P(v_i = x|\mathbf{h}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - b_i - \sigma_i \sum_j W_{ij} h_j)^2}{2\sigma_i^2}\right)$$ Gaussian

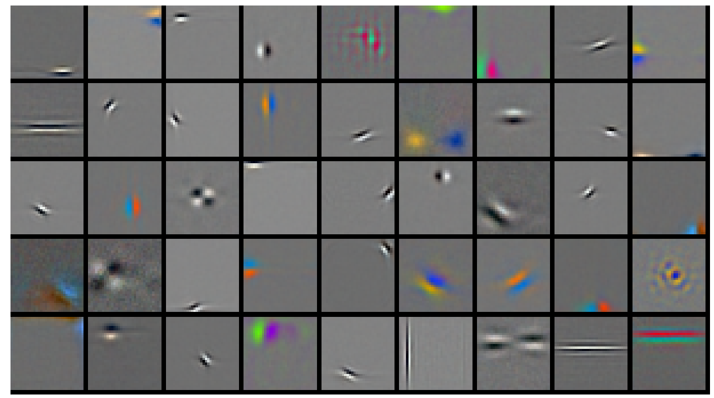
RBMs for Images and Text

Images: Gaussian-Bernoulli RBM

4 million **unlabelled** images



Learned features (out of 10,000)



Text: Multinomial-Bernoulli RBM

WIKIPEDIA
The Free Encyclopedia

REUTERS

AP Associated Press

Reuters dataset:
804,414 **unlabeled**
newswire stories

Bag-of-Words

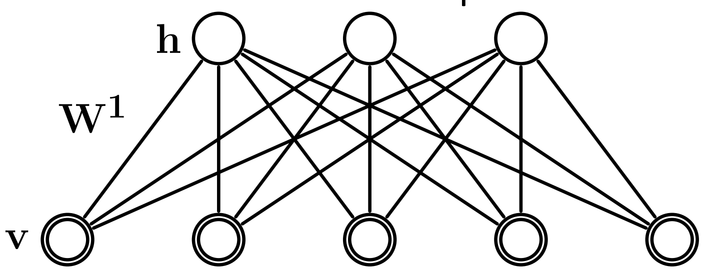
Learned features: ``topics''

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

Collaborative Filtering

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)}\exp\left(\sum_{ijk} W_{ij}^k v_i^k h_j + \sum_{ik} b_i^k v_i^k + \sum_j a_j h_j\right)$$

Bernoulli hidden: user preferences



Multinomial visible: user ratings

Netflix dataset:
480,189 users
17,770 movies
Over 100 million ratings

Learned features: ``genre''

Fahrenheit 9/11
Bowling for Columbine
The People vs. Larry Flynt
Canadian Bacon
La Dolce Vita

Independence Day
The Day After Tomorrow
Con Air
Men in Black II
Men in Black

Friday the 13th
The Texas Chainsaw Massacre
Children of the Corn
Child's Play
The Return of Michael Myers

Scary Movie
Naked Gun
Hot Shots!
American Pie
Police Academy

State-of-the-art performance on the Netflix dataset.

Relates to **Probabilistic Matrix Factorization**

(Salakhutdinov & Mnih ICML 2007)

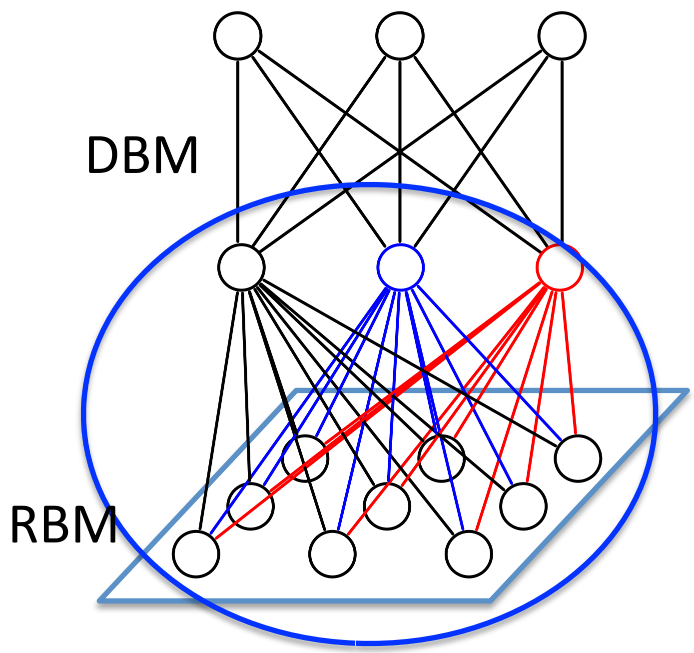
Multiple Application Domains

- Natural Images
- Text/Documents
- Collaborative Filtering / Matrix Factorization
- Video (Langford et al. ICML 2009 , Lee et al.)
- Motion Capture (Taylor et.al. NIPS 2007)
- Speech Perception (Dahl et. al. NIPS 2010, Lee et.al. NIPS 2010)

Same learning algorithm --
multiple input domains.

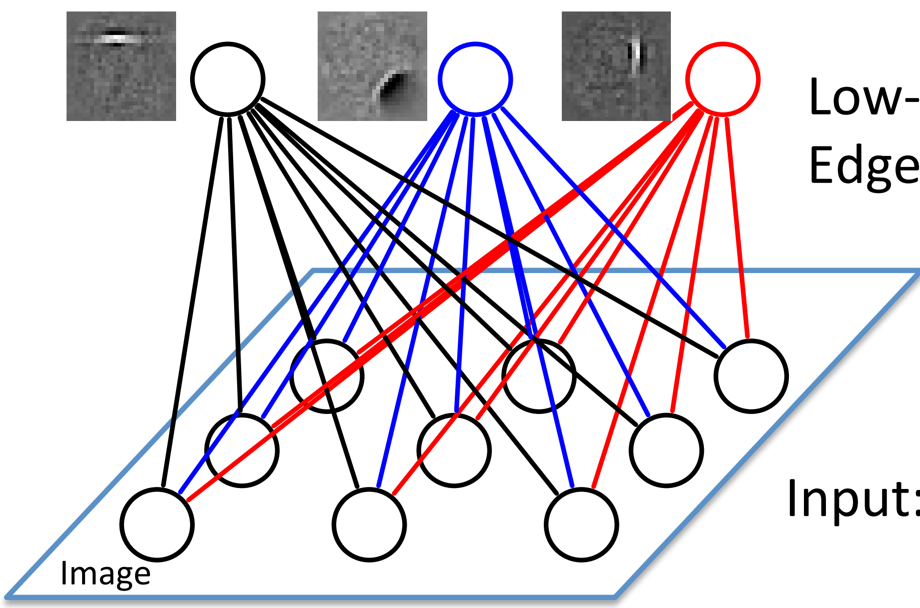
Limitations on the types of structure that can be represented by a single layer of low-level features!

Talk Roadmap



- Unsupervised Feature Learning
  - Restricted Boltzmann Machines
  - Deep Belief Networks
  - Deep Boltzmann Machines
- Transfer Learning with Deep Models
- Multimodal Learning

Deep Belief Network



Low-level features:
Edges

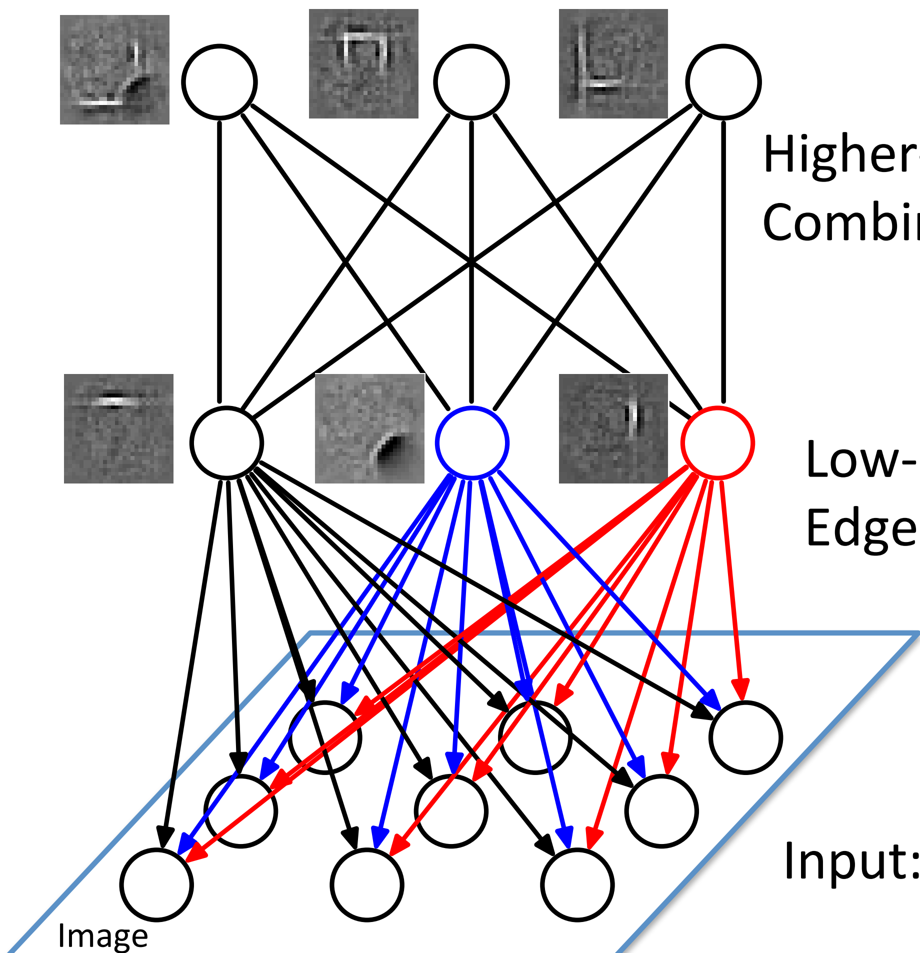
Built from **unlabeled** inputs.

Input: Pixels

Deep Belief Network

Unsupervised feature learning.

Internal representations capture higher-order statistical structure



Higher-level features: Combination of edges

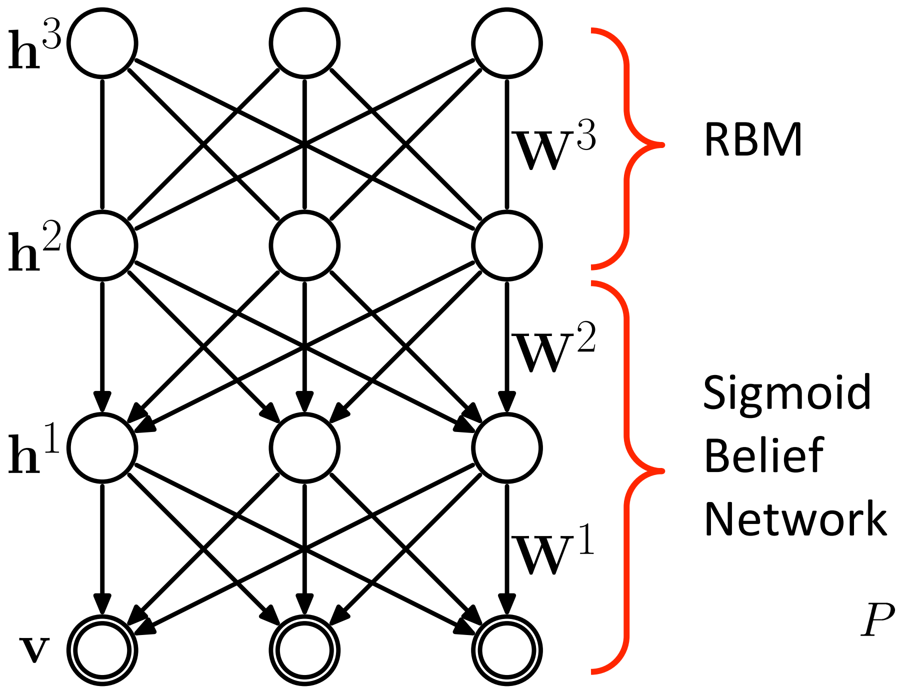
Low-level features: Edges

Built from **unlabeled** inputs.

Input: Pixels

(Hinton et.al. Neural Computation 2006)

Deep Belief Network



The joint probability distribution factorizes:

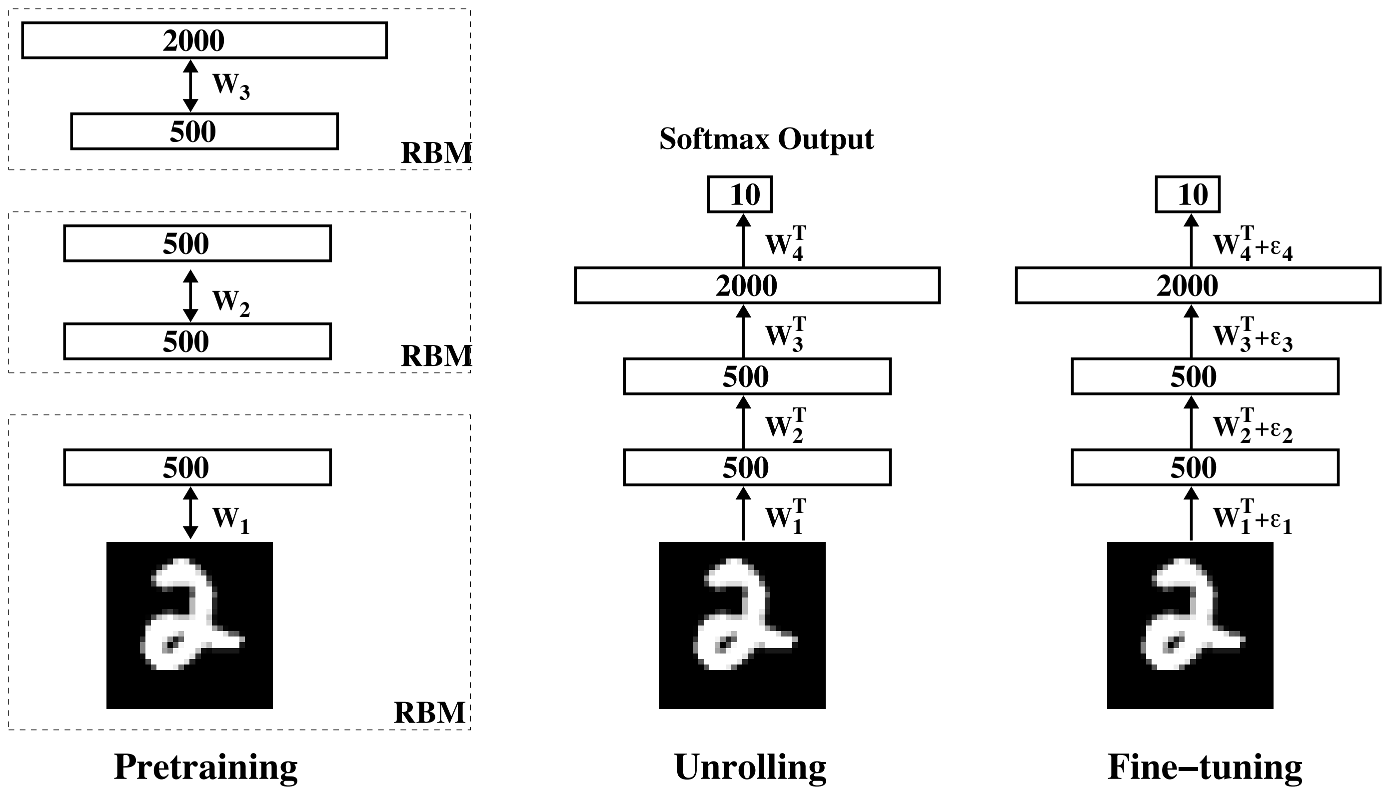
$$P(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3) = \underbrace{P(\mathbf{v}|\mathbf{h}^1) P(\mathbf{h}^1|\mathbf{h}^2)}_{\text{Sigmoid Belief Network}} \underbrace{P(\mathbf{h}^2, \mathbf{h}^3)}_{\text{RBM}}$$

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

$$P(\mathbf{h}^1|\mathbf{h}^2) = \prod_j P(h_j^1|\mathbf{h}^2) \qquad P(h_j^1 = 1|\mathbf{h}^2) = \frac{1}{1 + \exp\left(-\sum_k W_{jk}^2 h_k^2\right)}$$

$$P(\mathbf{v}|\mathbf{h}^1) = \prod_i P(v_i|\mathbf{h}^1) \qquad P(v_i = 1|\mathbf{h}^1) = \frac{1}{1 + \exp\left(-\sum_j W_{ij}^1 h_j^1\right)}$$

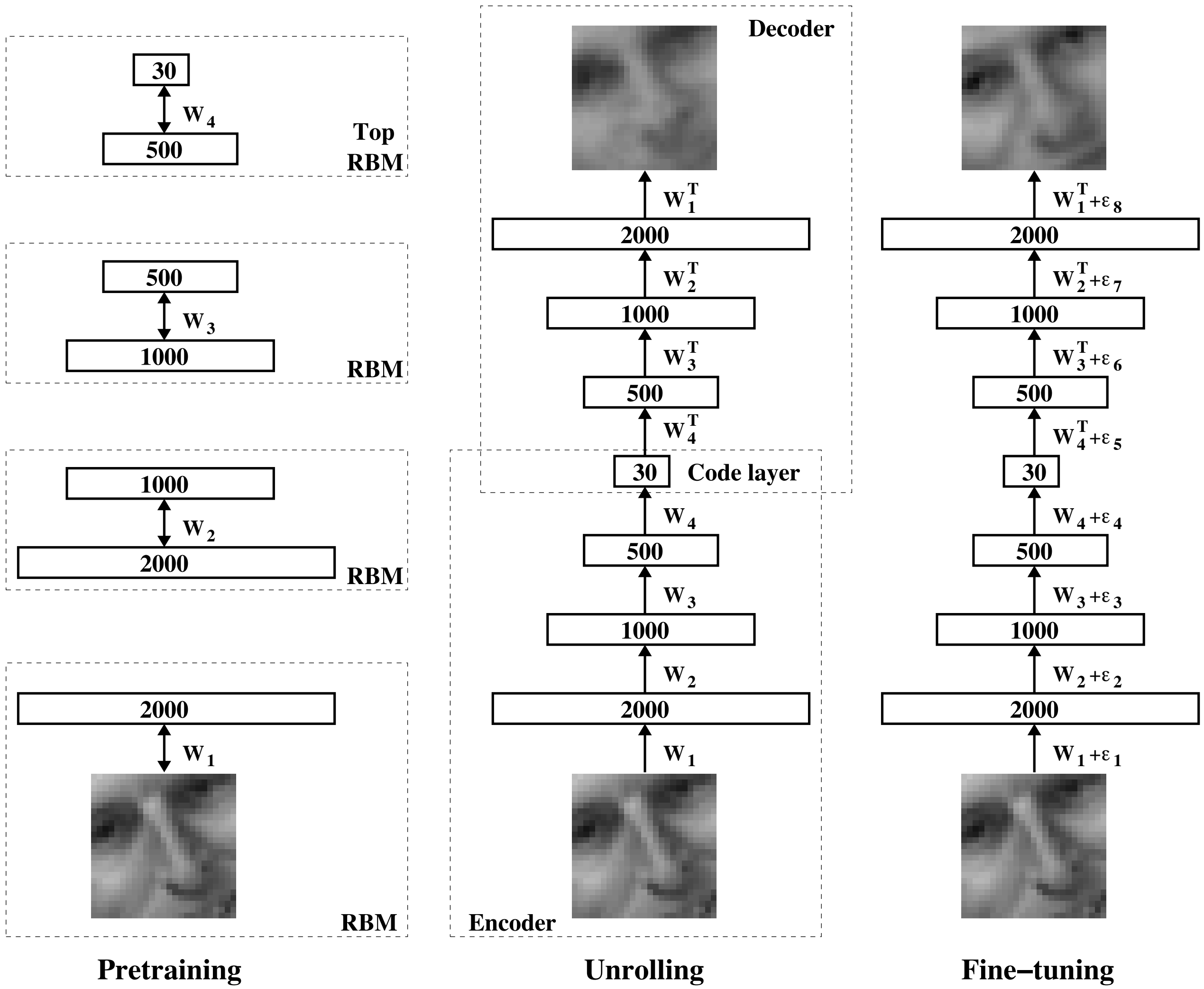
DBNs for Classification



• After layer-by-layer **unsupervised pretraining**, discriminative fine-tuning by backpropagation achieves an error rate of 1.2% on MNIST. SVM's get 1.4% and randomly initialized backprop gets 1.6%.

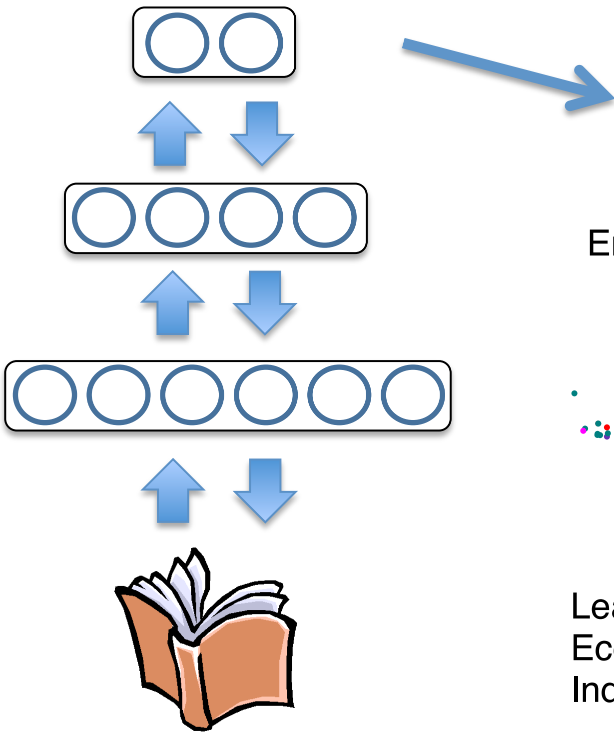
• Clearly unsupervised learning helps generalization. It ensures that most of the information in the weights comes from modeling the input data.

Deep Autoencoders

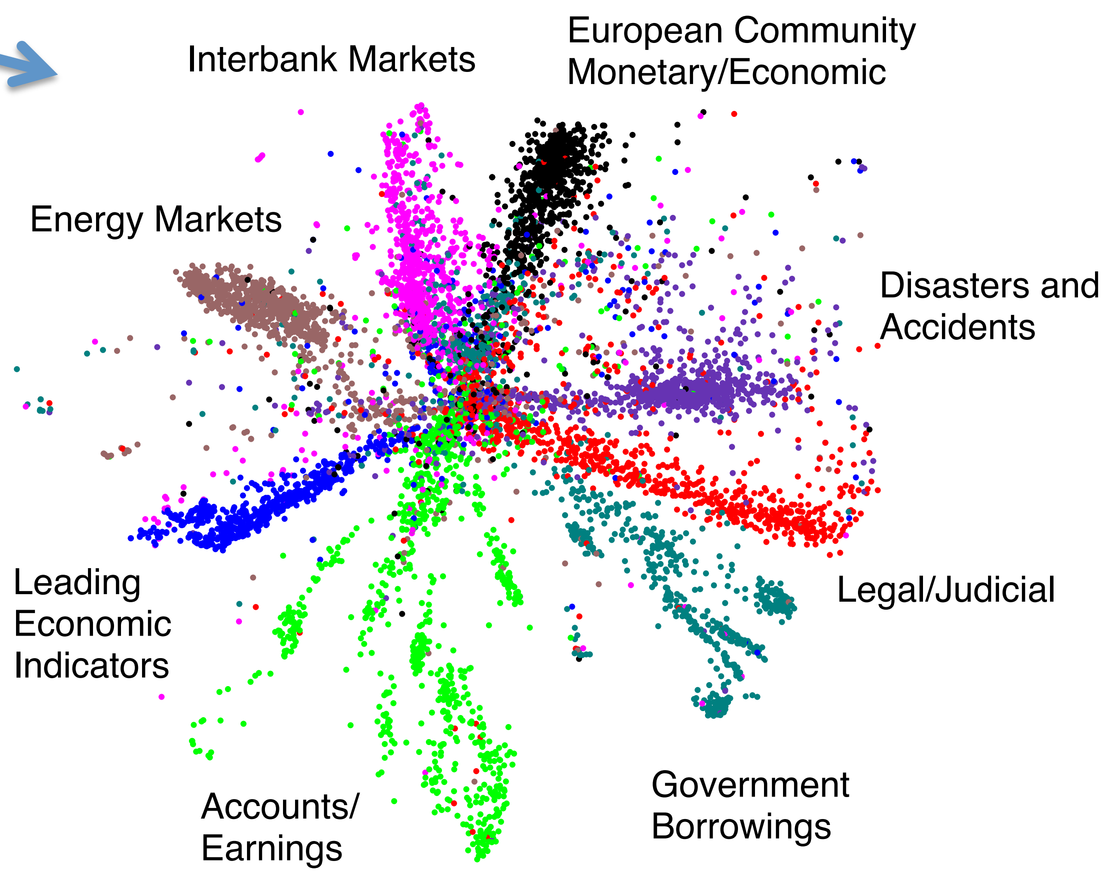


Deep Generative Model

Model P(document)

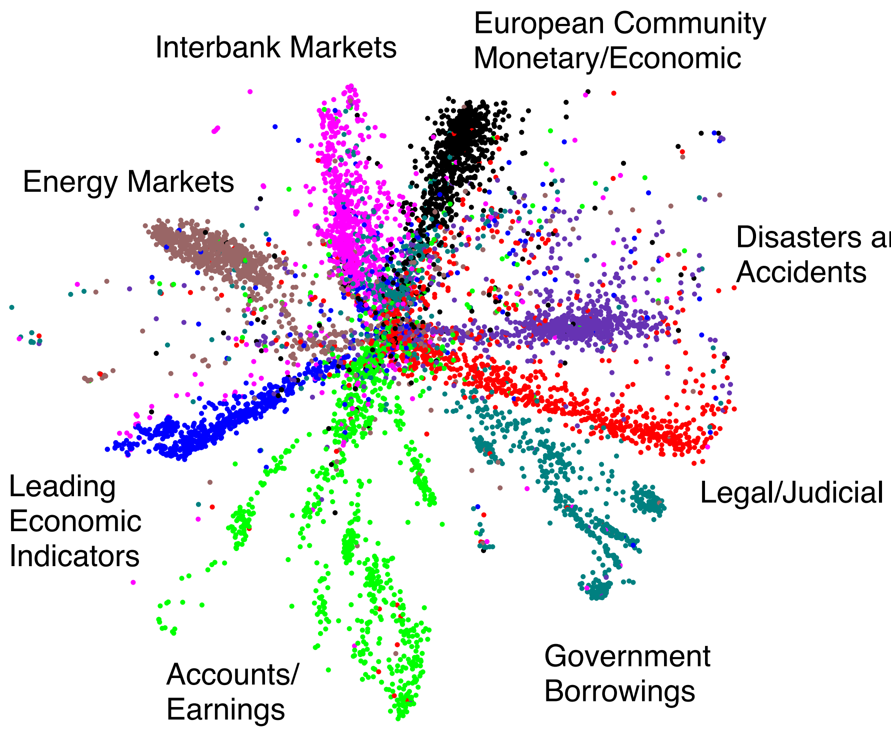


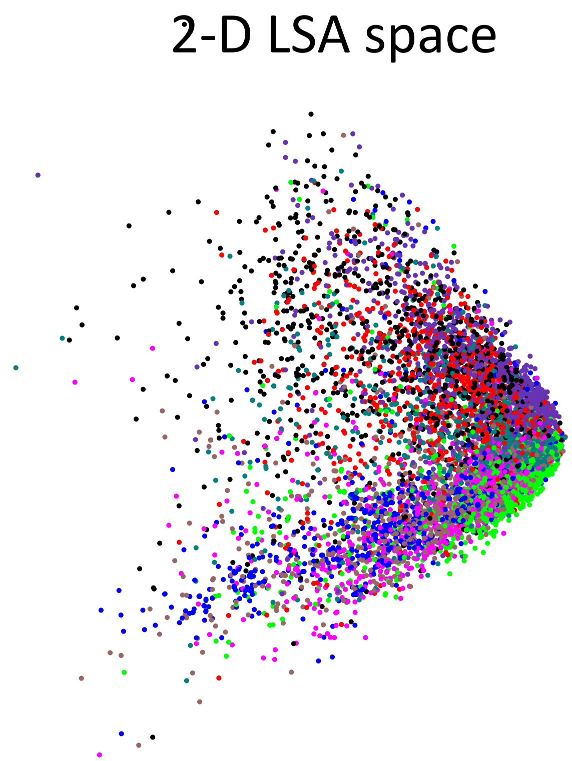
Reuters dataset: 804,414 newswire stories: **unsupervised**



(Hinton & Salakhutdinov, Science 2006)

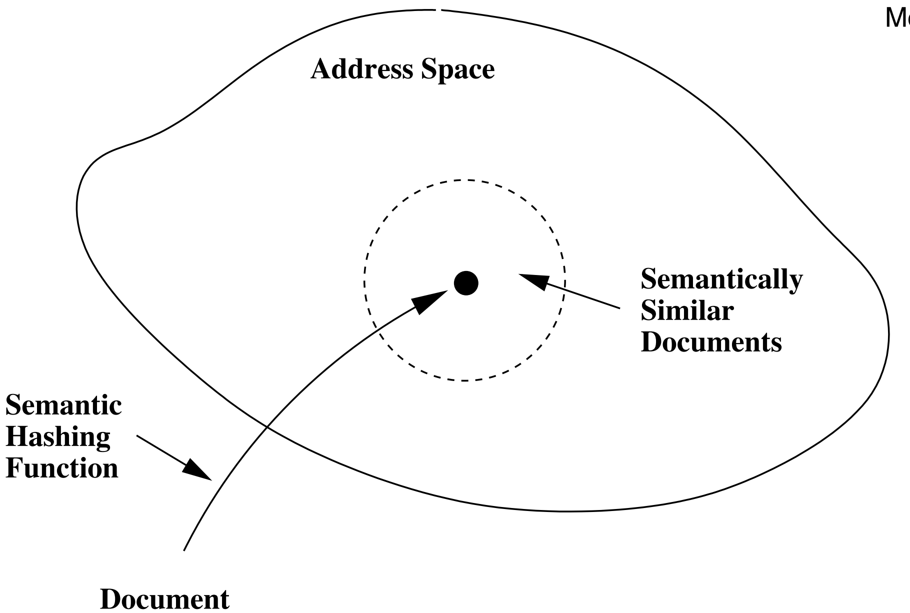
Information Retrieval

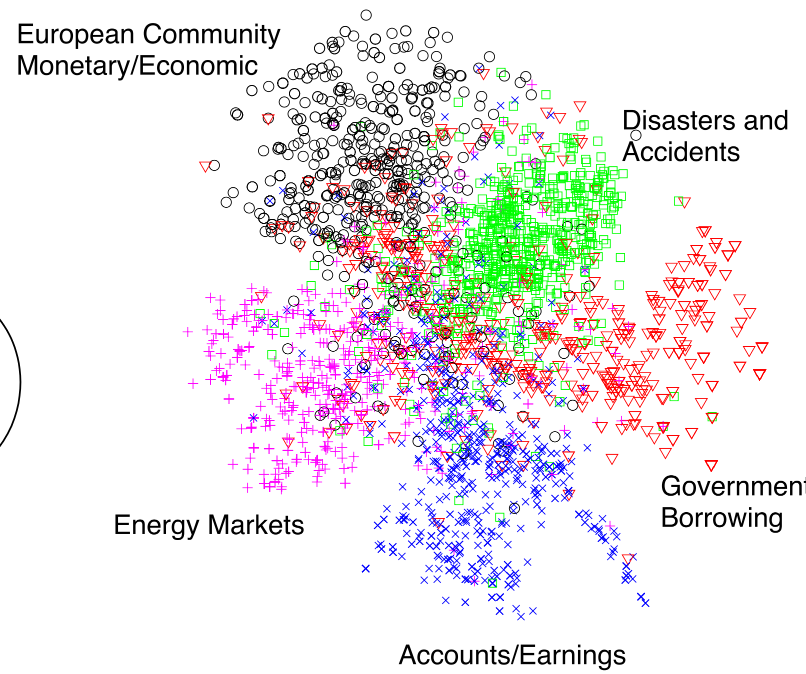




- The Reuters Corpus Volume II contains 804,414 newswire stories (randomly split into **402,207 training** and **402,207 test).**
- "Bag-of-words": each article is represented as a vector containing the counts of the most frequently used 2000 words in the training set.

Semantic Hashing

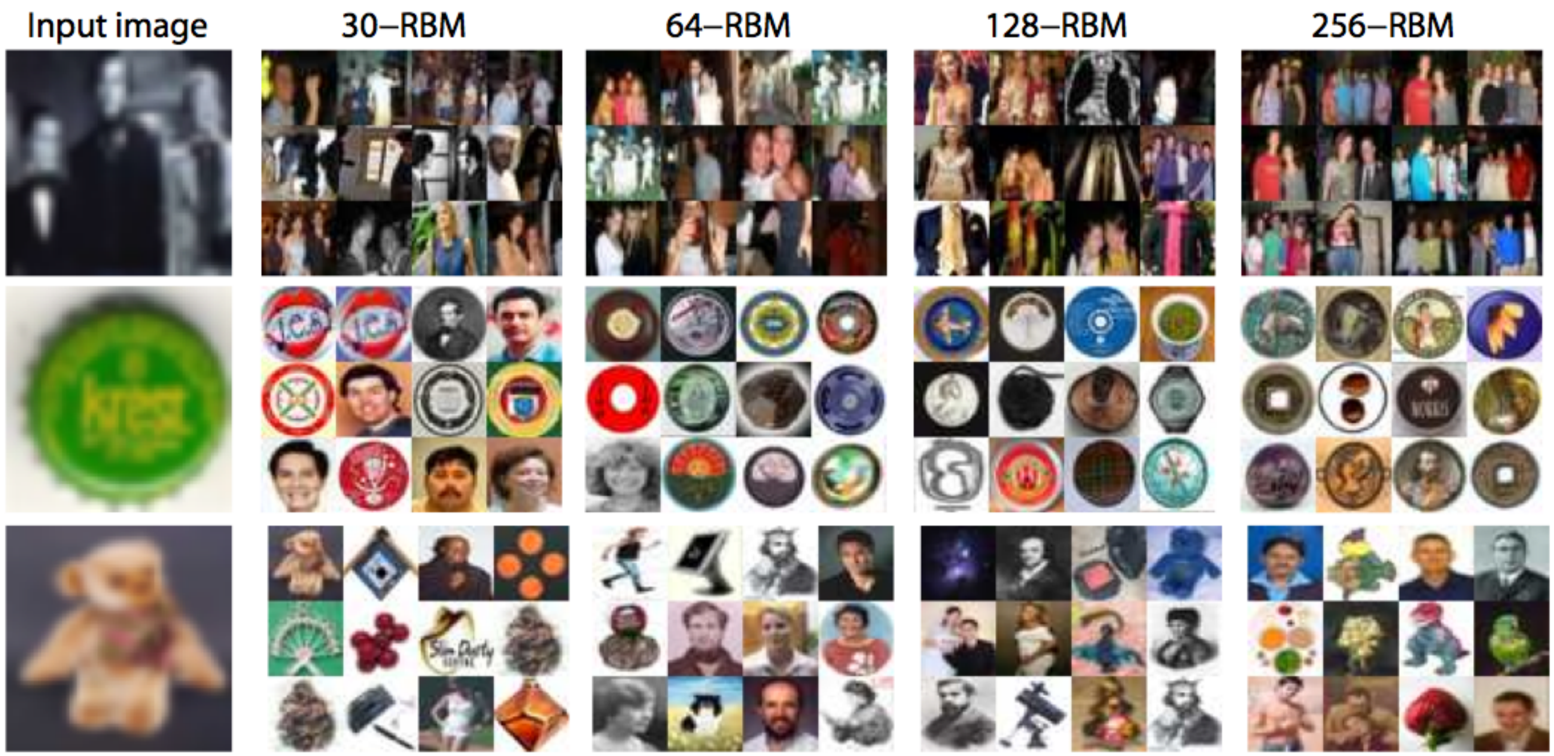




- Learn to map documents into **semantic 20-D binary codes.**
- Retrieve similar documents stored at the nearby addresses **with no search at all.**

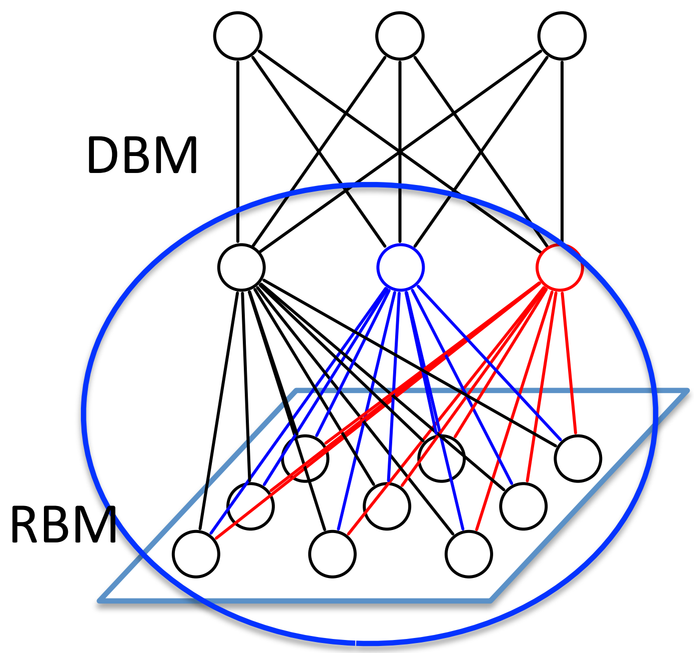
Searching Large Image Database using Binary Codes

- Map images into binary codes for fast retrieval.



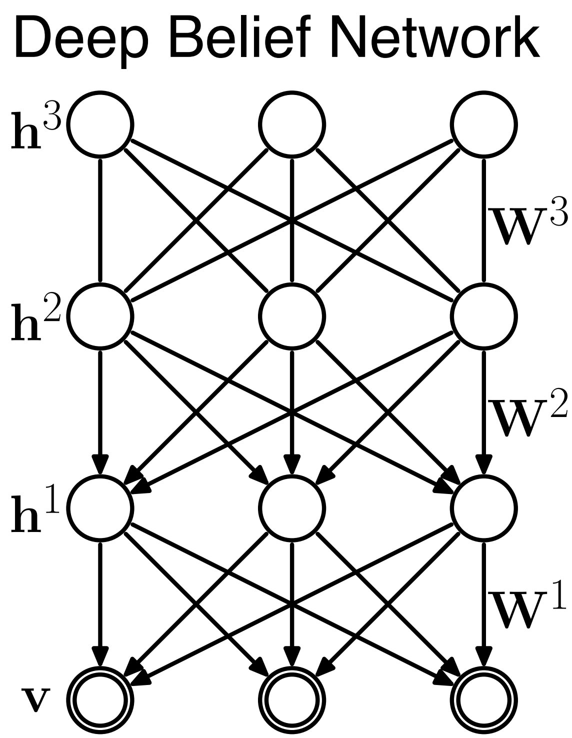
- Small Codes, Torralba, Fergus, Weiss, CVPR 2008
- Spectral Hashing, Y. Weiss, A. Torralba, R. Fergus, NIPS 2008
- Kulis and Darrell, NIPS 2009, Gong and Lazebnik, CVPR 20111
- Norouzi and Fleet, ICML 2011,

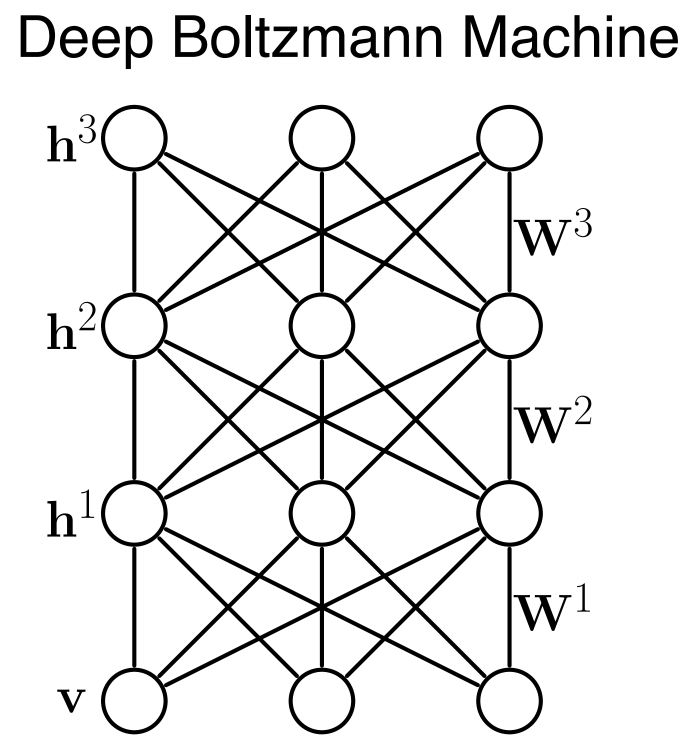
Talk Roadmap



- Unsupervised Feature Learning
  - Restricted Boltzmann Machines
  - Deep Belief Networks
  - Deep Boltzmann Machines
- Transfer Learning with Deep Models
- Multimodal Learning

DBNs vs. DBMs





DBNs are hybrid models:

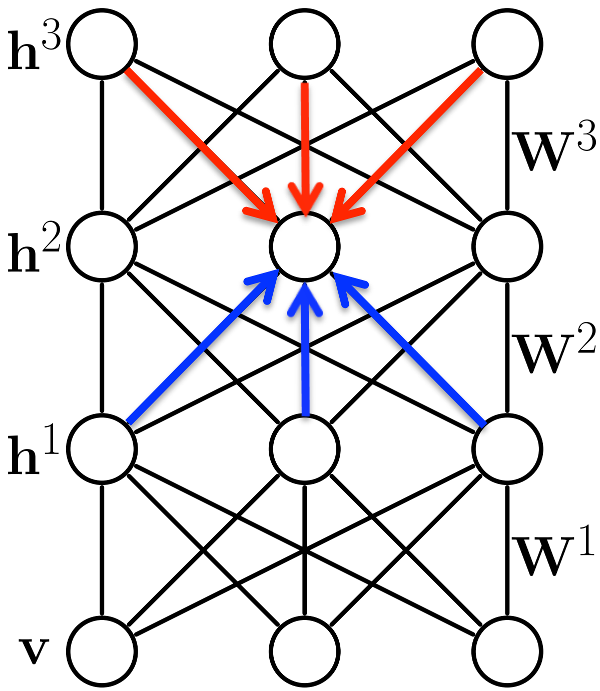
- Inference in DBNs is problematic due to **explaining away**.
- Only greedy pretrainig, **no joint optimization over all layers**.
- Approximate inference is feed-forward: **no bottom-up and top-down**.

Introduce a new class of models called Deep Boltzmann Machines.

Mathematical Formulation

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

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:

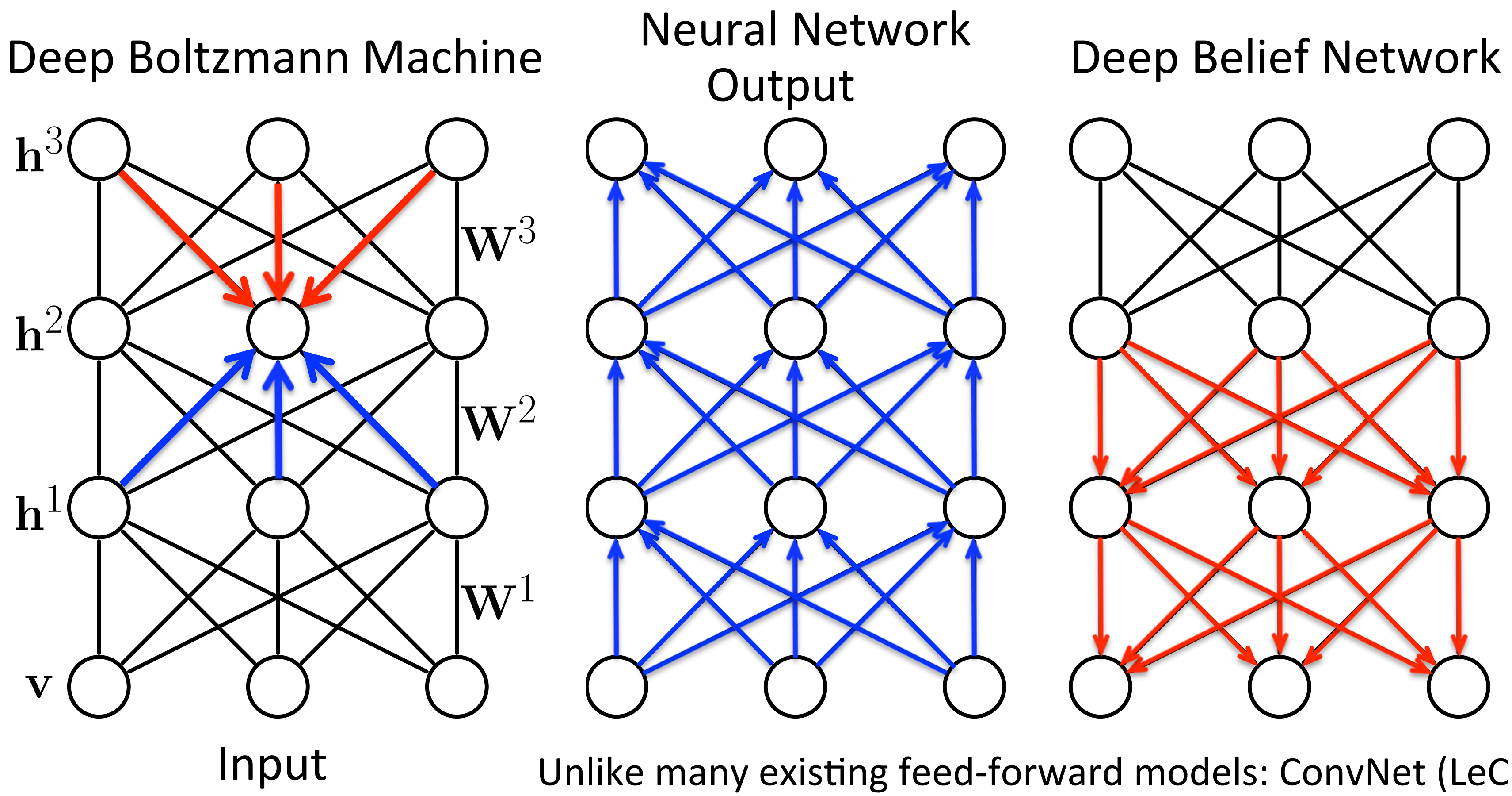
$$P(h_j^2 = 1 | \mathbf{h}^1, \mathbf{h}^3) = \sigma\left(\sum_k W_{kj}^3 h_k^3 + \sum_m W_{mj}^2 h_m^1\right)$$

Top-down   Bottom-up

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

Mathematical Formulation

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



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

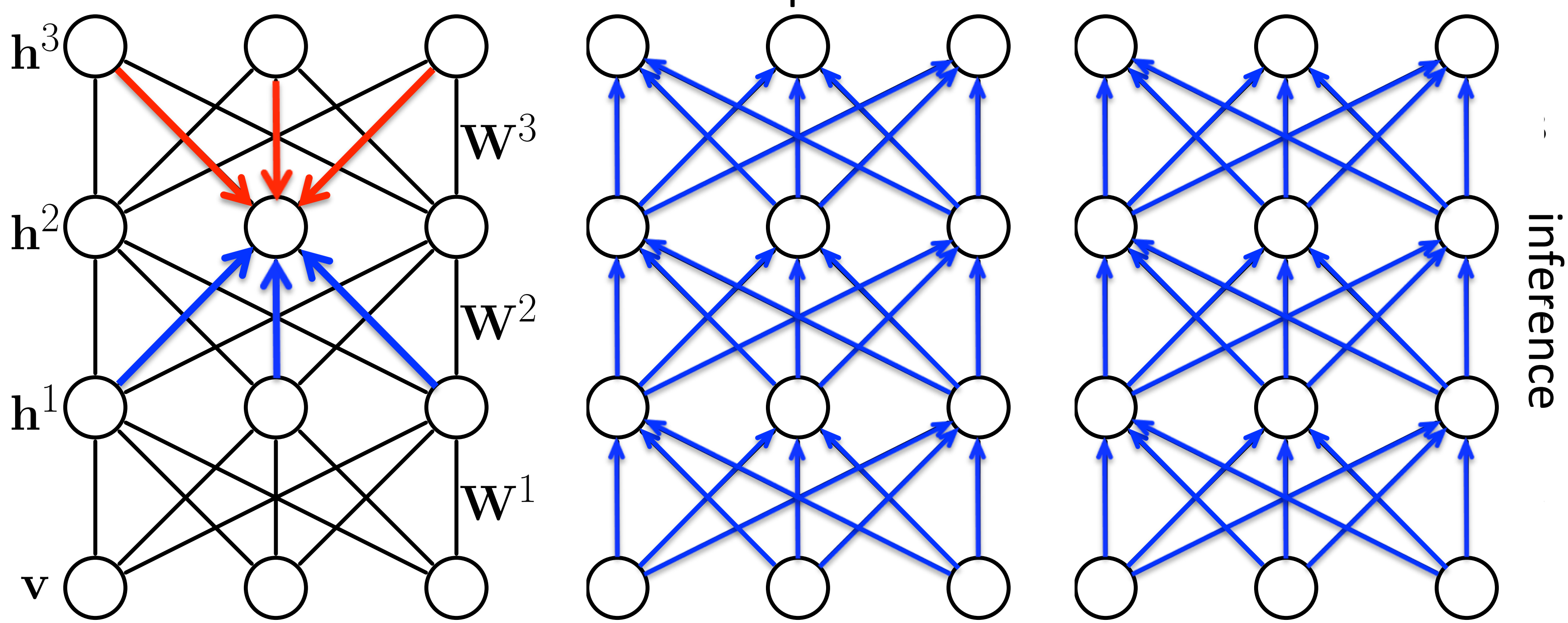
Mathematical Formulation

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

Deep Boltzmann Machine

Neural Network
Output

Deep Belief Network



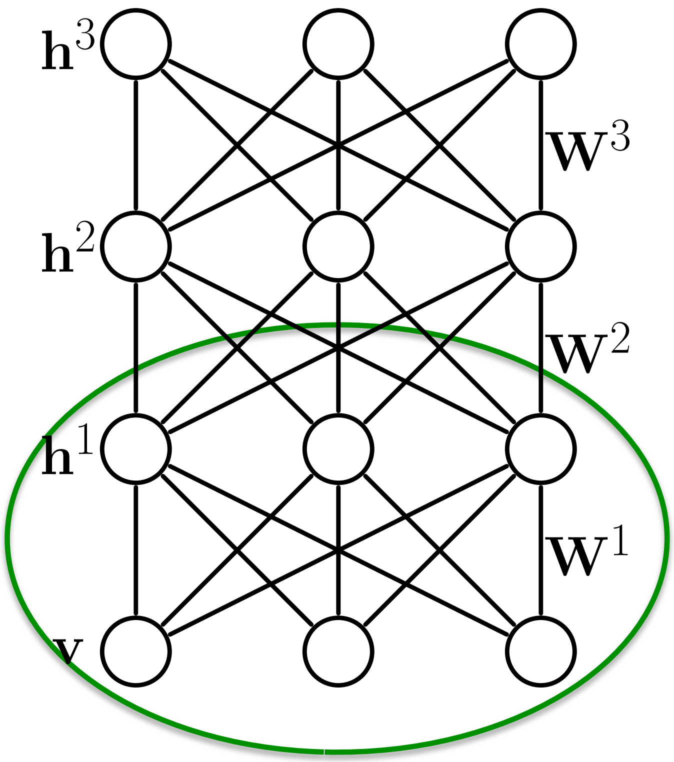
Input

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.

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.

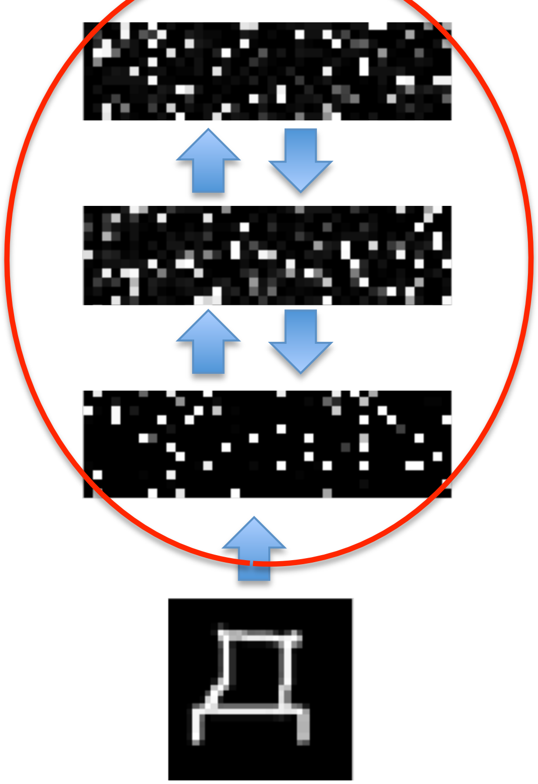
Many of the previous approaches were not successful for learning general Boltzmann machines with **hidden variables.**

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

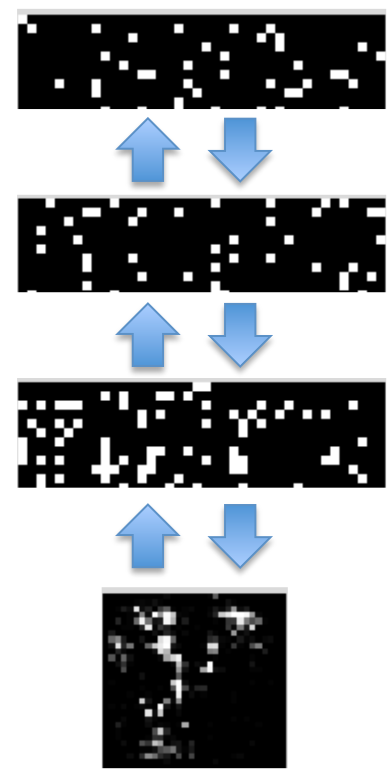
$P_{data}(\mathbf{h}|\mathbf{v})$

Simulate from the Model

Approximate the joint distribution

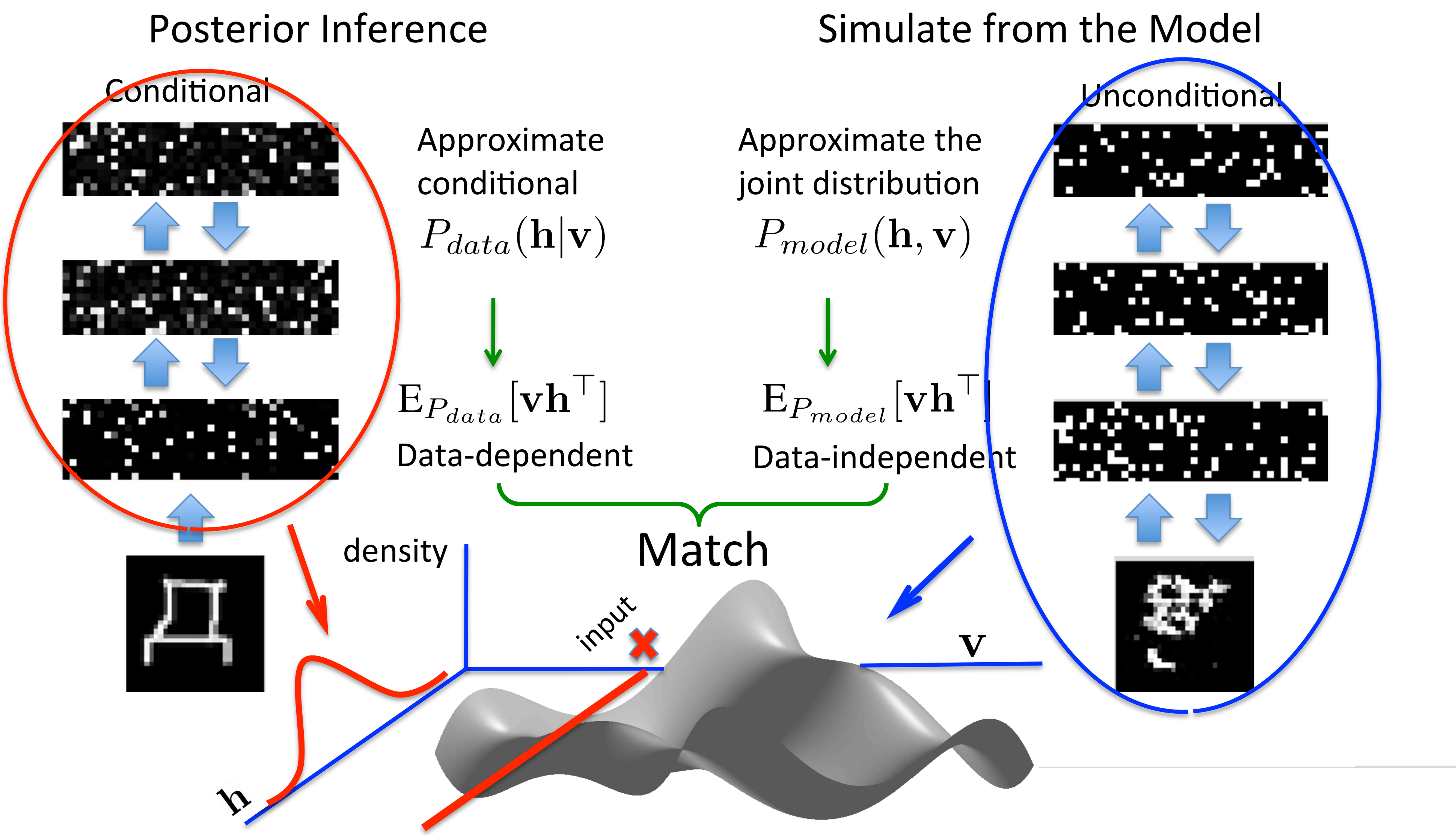
$P_{model}(\mathbf{h}, \mathbf{v})$

Unconditional

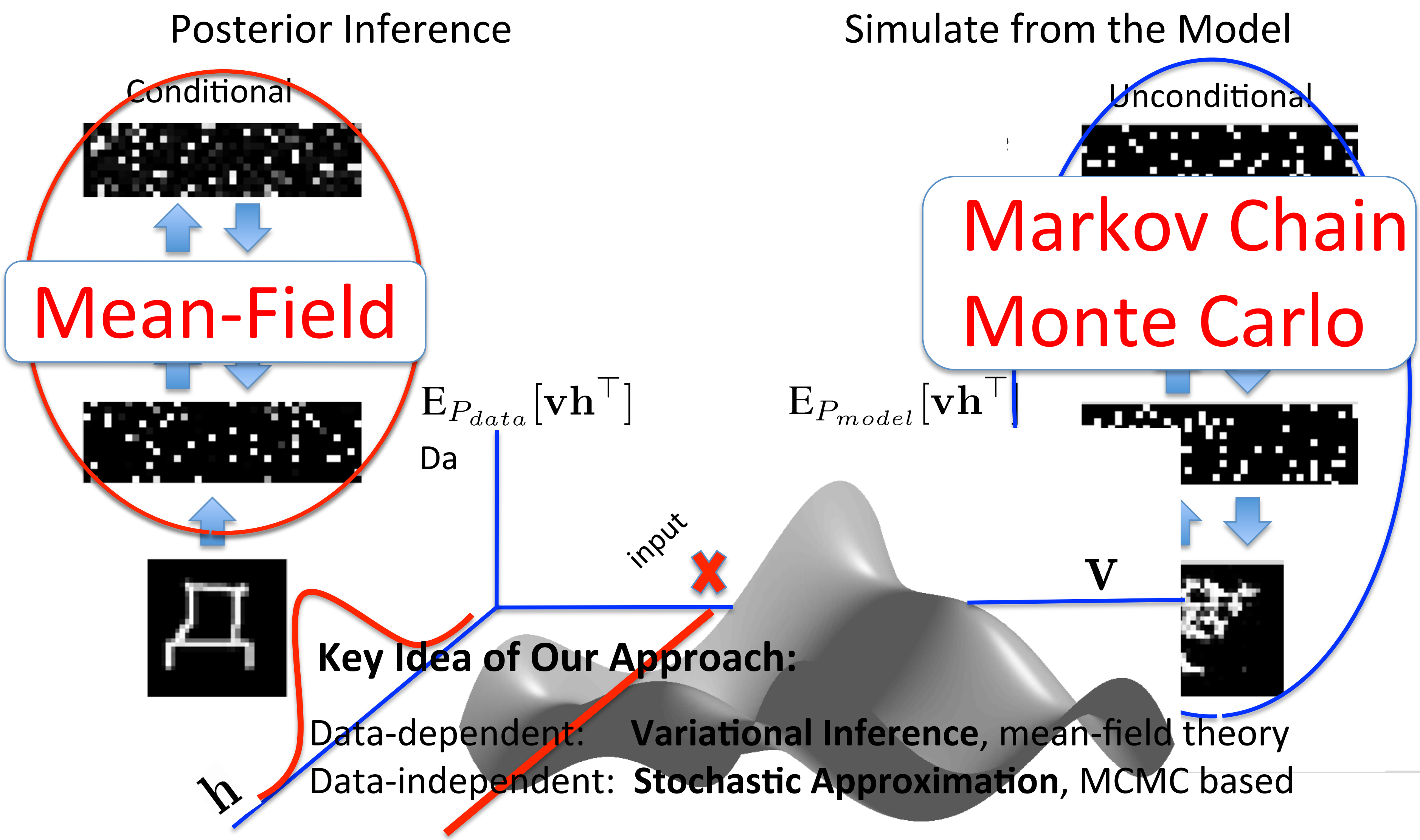


(Salakhutdinov, 2008; NIPS 2009)

New Learning Algorithm



New Learning Algorithm



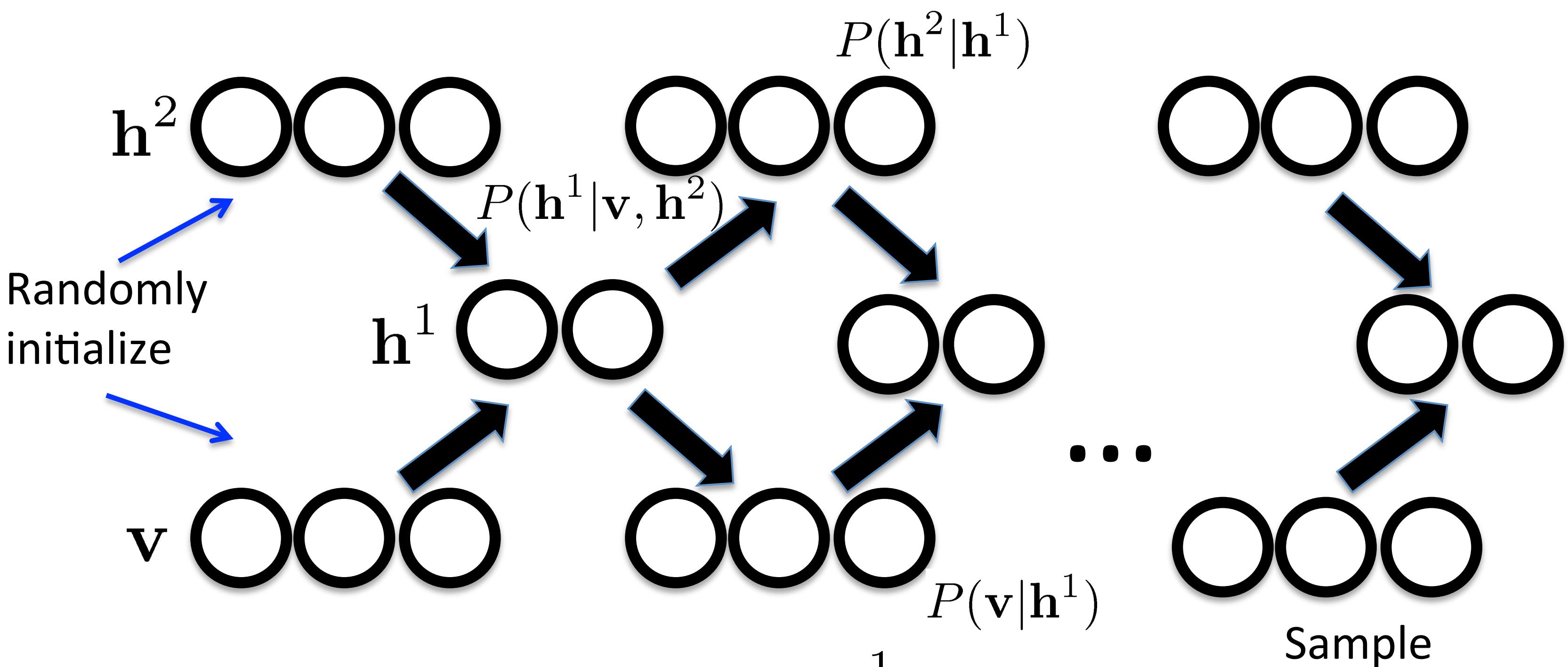
Key Idea of Our Approach:

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

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

Sampling from DBMs

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

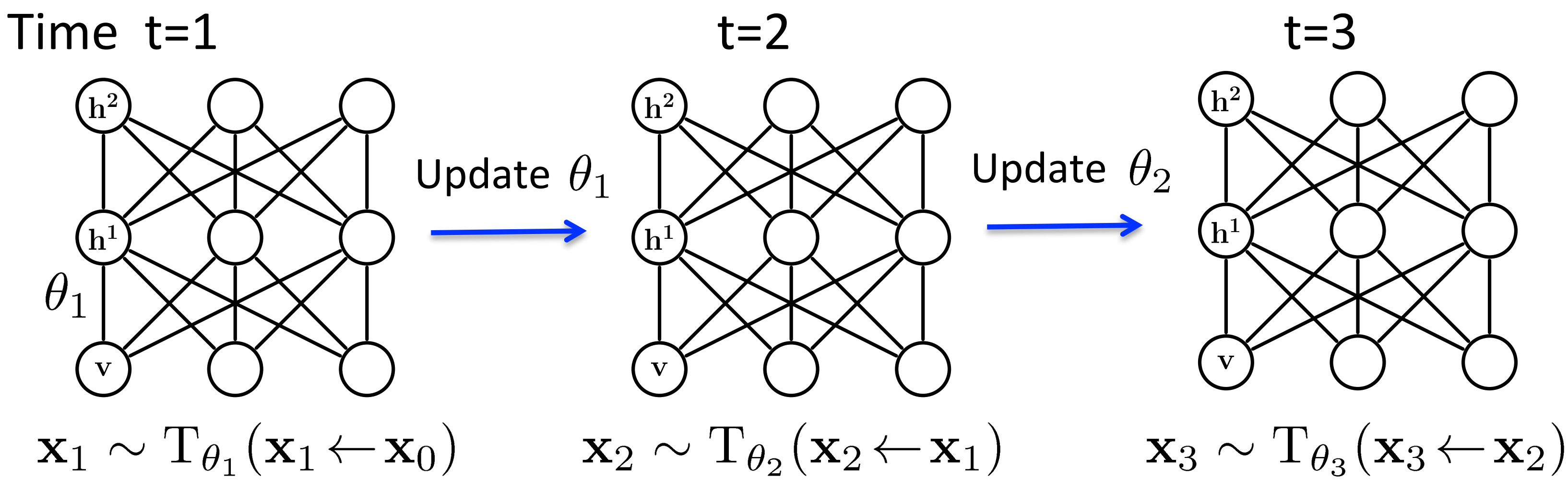


$$P(h^1_m = 1|\mathbf{v},\mathbf{h}^2) = \frac{1}{1+\exp(-\sum_i W^1_{im} v_i - \sum_j W^2_{mj} h^2_j)}$$

$$P(h^2_j = 1|\mathbf{h}^1) = \frac{1}{1+\exp(-\sum_m W^2_{mj} h^1_m)}$$

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

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.

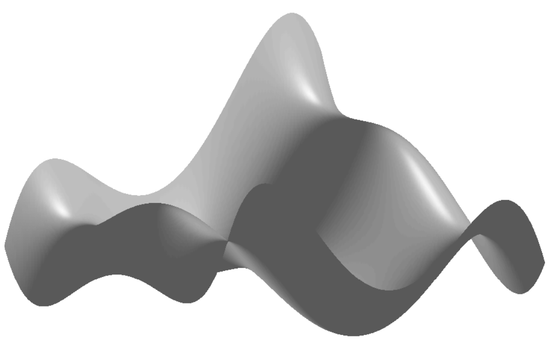
Robbins and Monro, Ann. Math. Stats, 1957
L. Younes, Probability Theory 1989, Tieleman, ICML 2008.

Stochastic Approximation

Update rule decomposes:

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

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



Problem: High-dimensional data: the energy landscape is highly multimodal

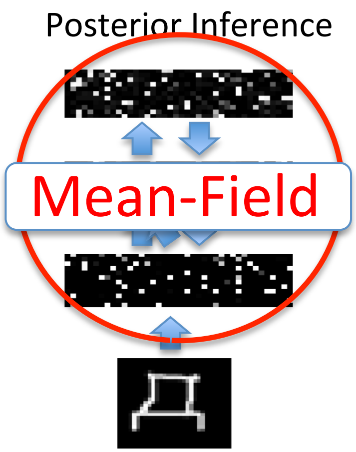
Key insight: The transition operator can be any valid transition operator – Tempered Transitions, Parallel/Simulated Tempering.

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 $Q_\mu(\mathbf{h}|\mathbf{v})$:

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



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

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

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

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

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

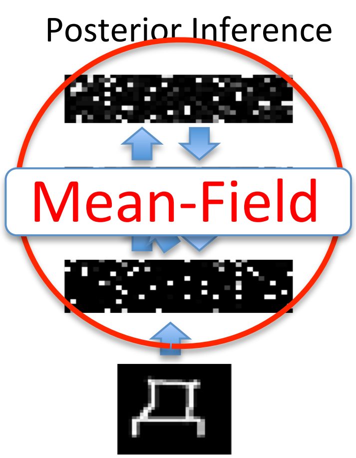
(Salakhutdinov & Larochelle, AI & Statistics 2010)

Variational Inference

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

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

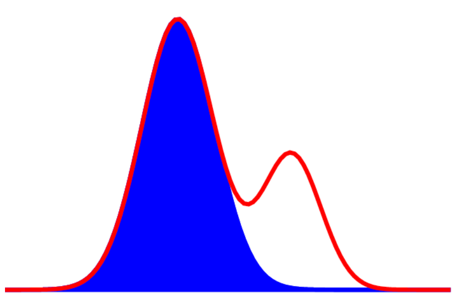
$$\log P_\theta(\mathbf{v}) \geq \underbrace{\log P_\theta(\mathbf{v}) - \mathrm{KL}\big(Q_\mu(\mathbf{h}|\mathbf{v}) || P_\theta(\mathbf{h}|\mathbf{v})\big)}_{\text{Variational Lower Bound}}$$



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:

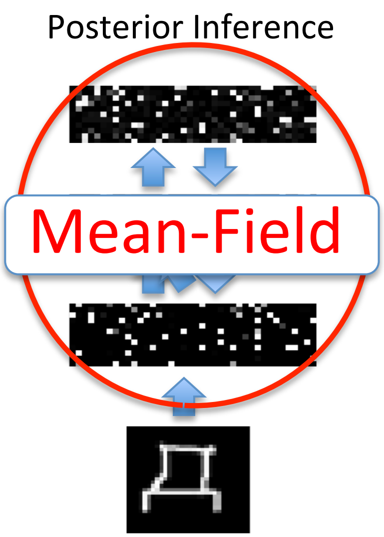
$$\mu_j^{(1)} = \sigma\Big(\sum_i W_{ij}^1 v_i + \sum_k W_{jk}^2 \mu_k^{(2)}\Big)$$
$$\mu_k^{(2)} = \sigma\Big(\sum_j W_{jk}^2 \mu_j^{(1)} + \sum_m W_{km}^3 \mu_m^{(3)}\Big)$$
$$\mu_m^{(3)} = \sigma\Big(\sum_k W_{km}^3 \mu_k^{(2)}\Big)$$

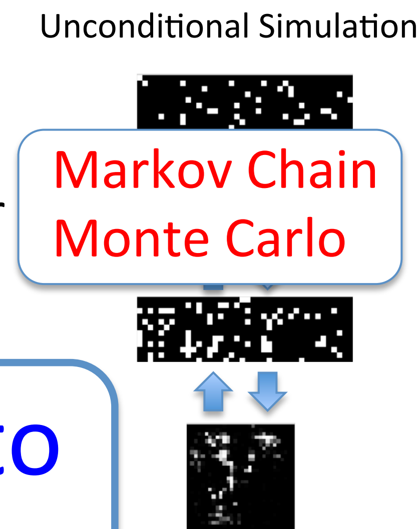
Variational Inference

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

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

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





1. V … wer
bou …

Fast Inference

2. M …
to u …

Learning can scale to millions of examples

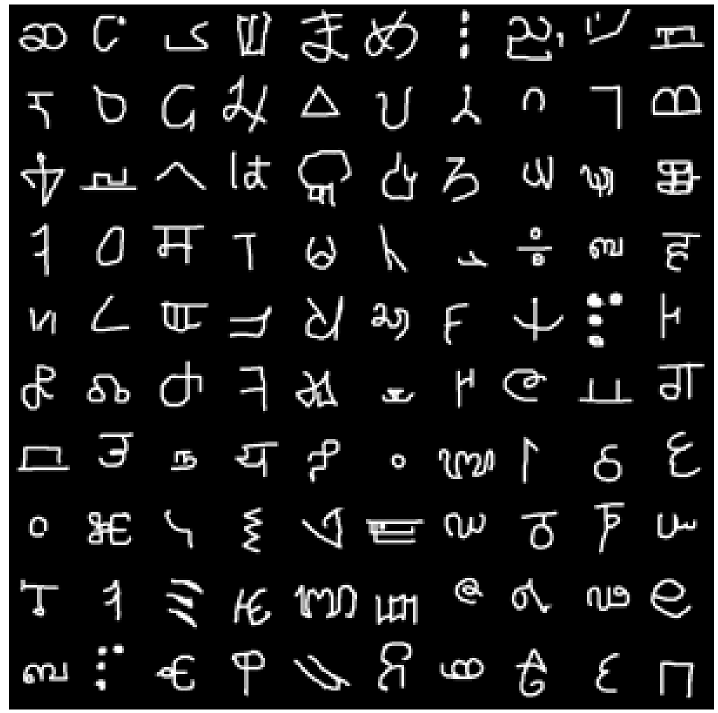
Almost sure convergence guarantees to an asymptotically stable point.

Good Generative Model?

Handwritten Characters

Good Generative Model?

Handwritten Characters



Good Generative Model?

Handwritten Characters

Simulated

Real Data

Good Generative Model?

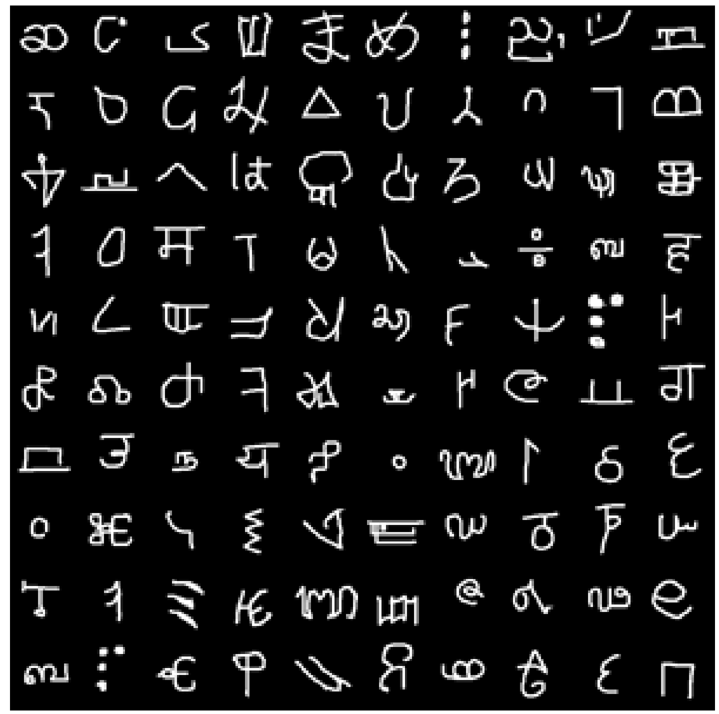
Handwritten Characters

Real Data

Simulated

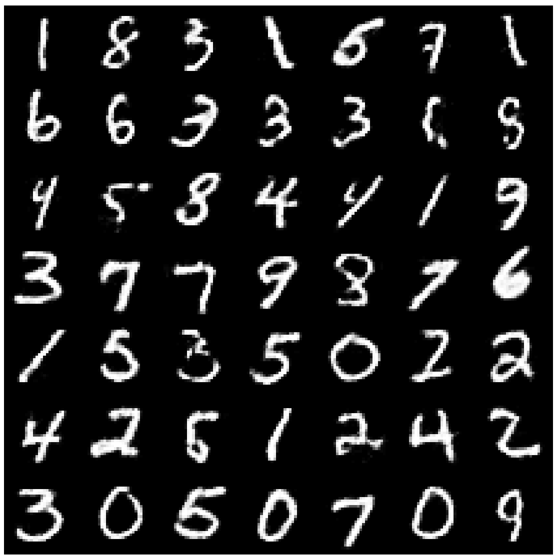
Good Generative Model?

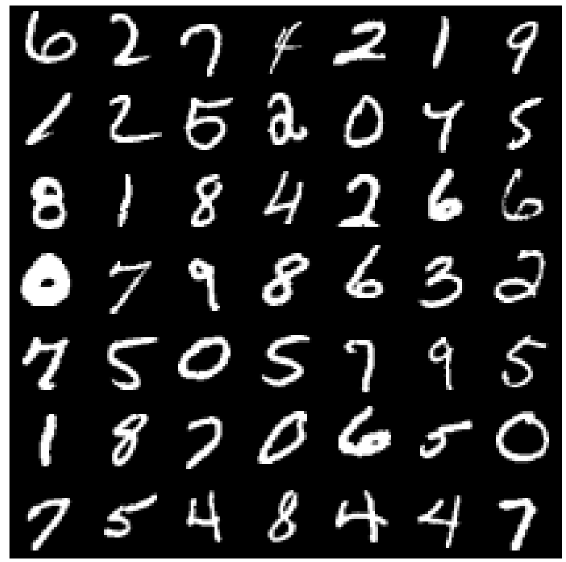
Handwritten Characters



Good Generative Model?

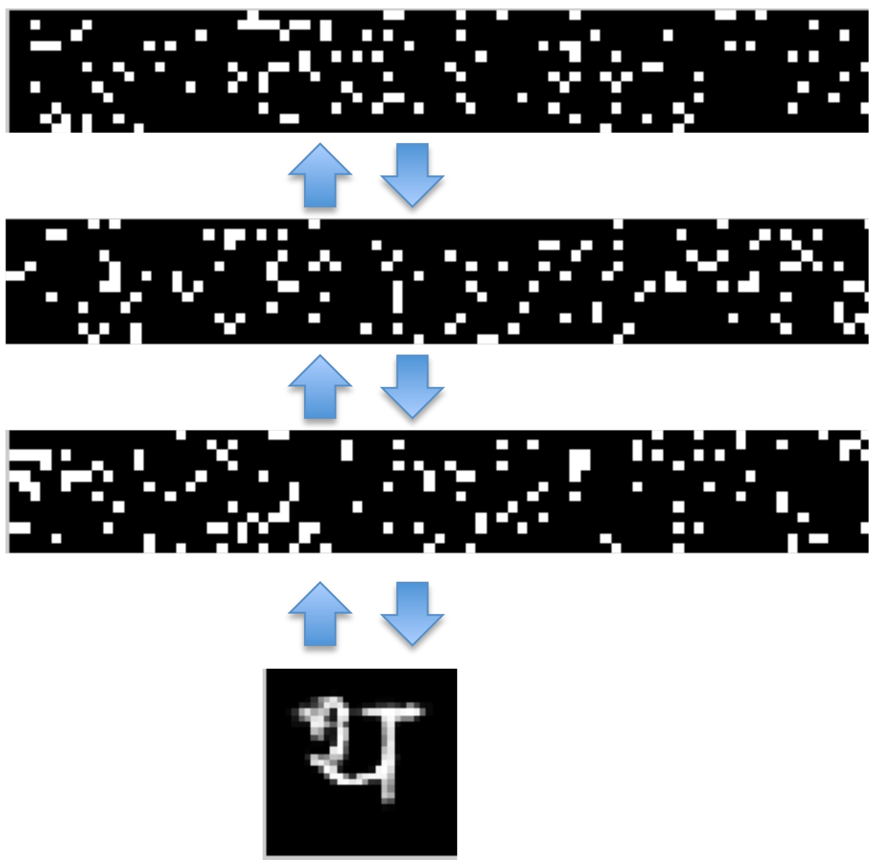
MNIST Handwritten Digit Dataset



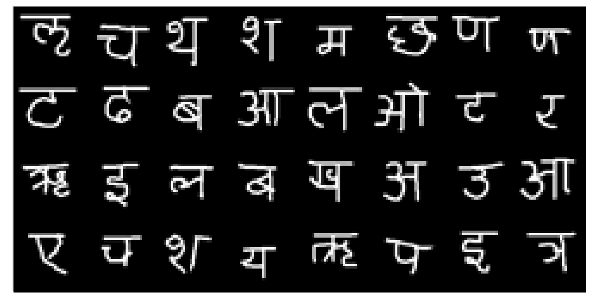


Deep Boltzmann Machine

Sanskrit



Model P(image)

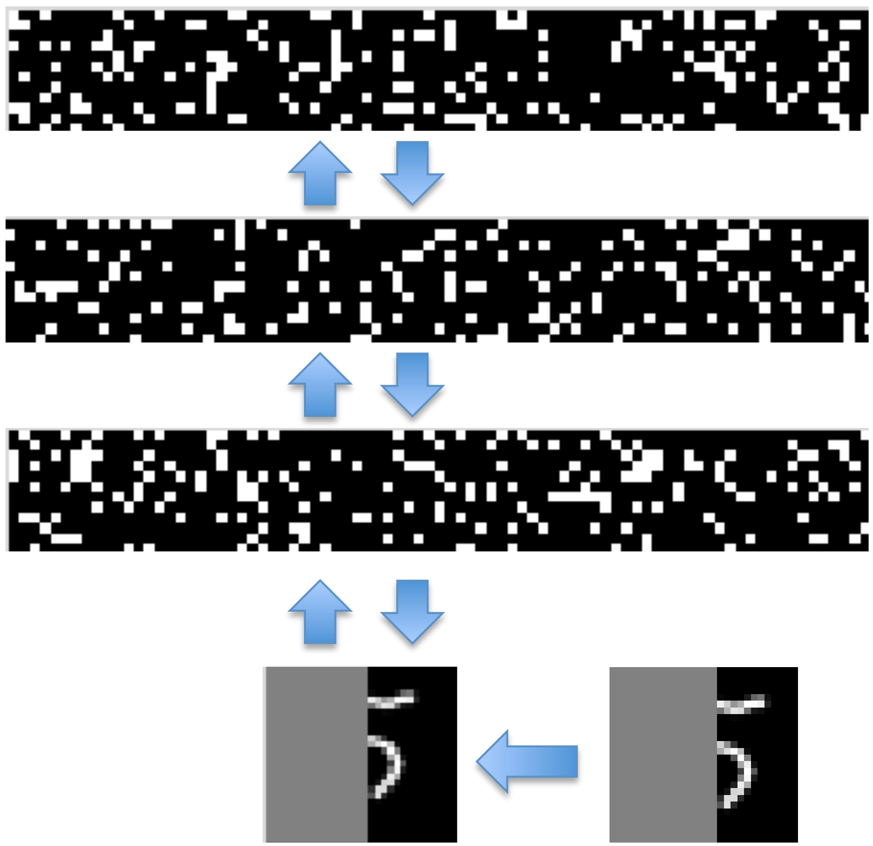


25,000 characters from 50 alphabets around the world.

- 3,000 hidden variables
- 784 observed variables (28 by 28 images)
- Over 2 million parameters

Bernoulli Markov Random Field

Deep Boltzmann Machine



Conditional
Simulation

P(image|partial image)

Bernoulli Markov Random Field

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

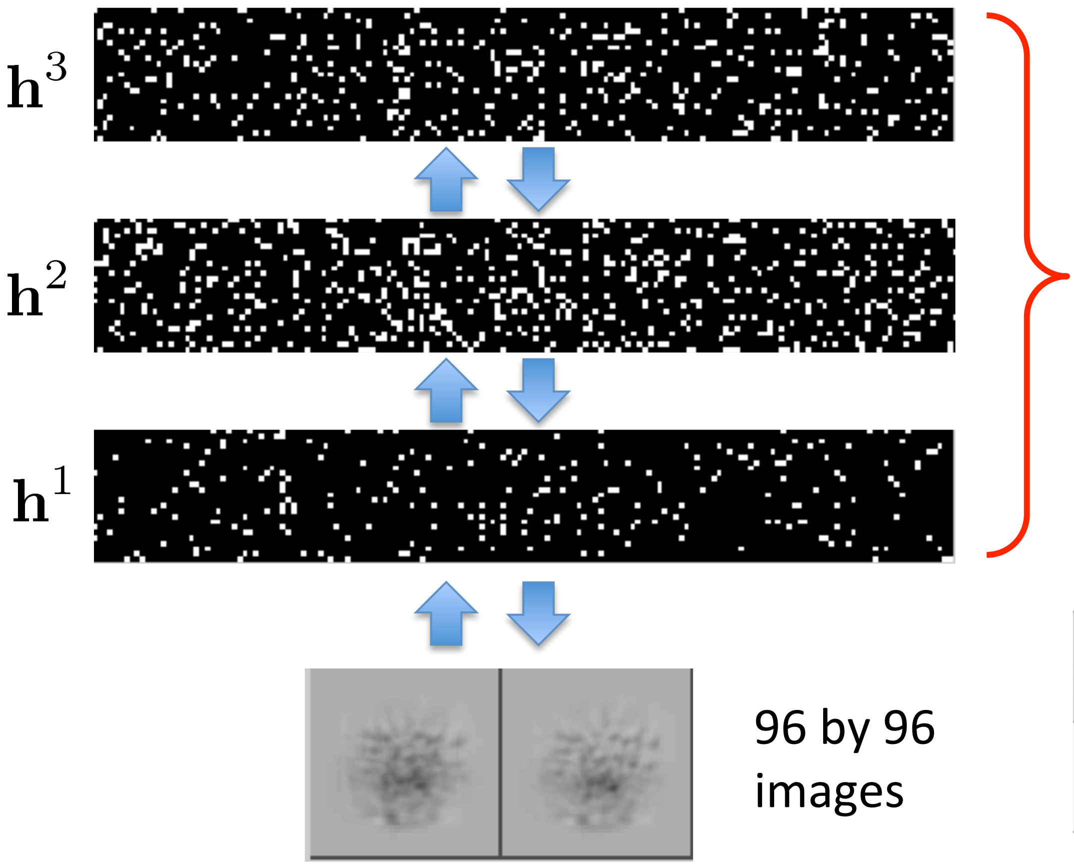
Learning Algorithm	Error
Logistic regression	22.14%
K-NN	18.92%
Neural Net	14.62%
SVM (Larochelle et.al. 2009)	9.70%
Deep Autoencoder (Bengio et. al. 2007)	10.05%
Deep Belief Net (Larochelle et. al. 2009)	9.68%
DBM	**8.40%**

Permutation-invariant version.

Deep Boltzmann Machine

Deep Boltzmann Machine

Gaussian-Bernoulli Markov Random Field

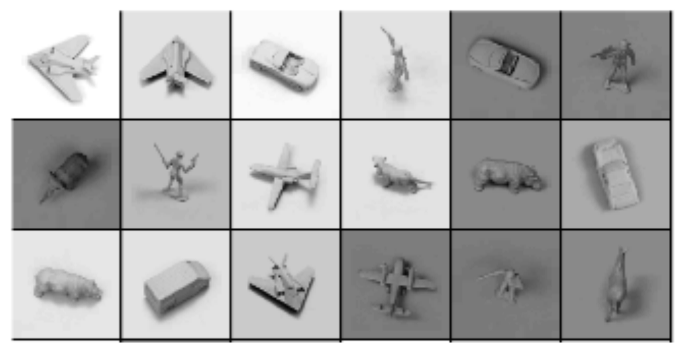


12,000 Latent Variables

96 by 96 images

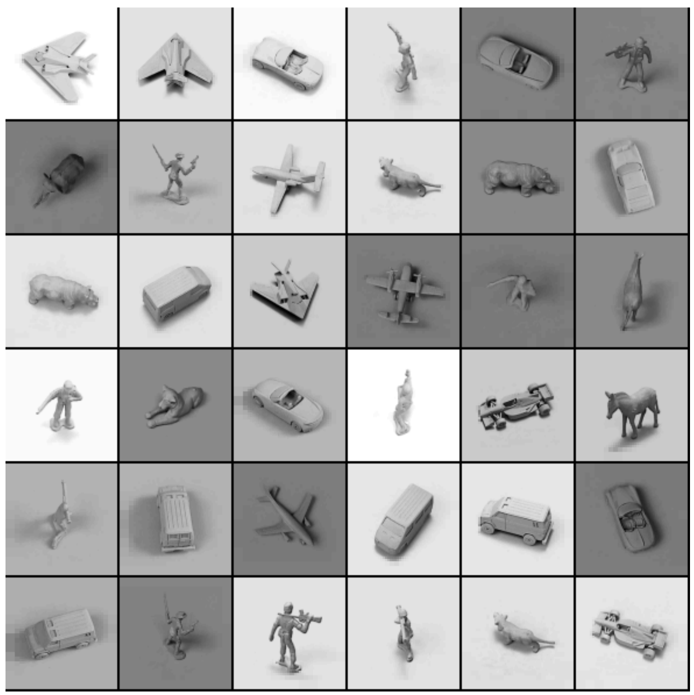
Stereo pair

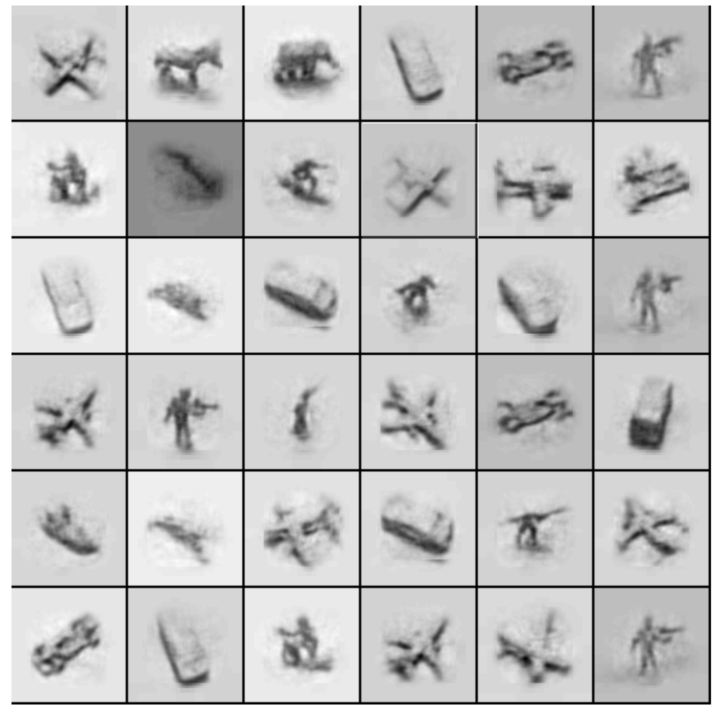
Model P(image)



24,000 Training Images

Generative Model of 3-D Objects



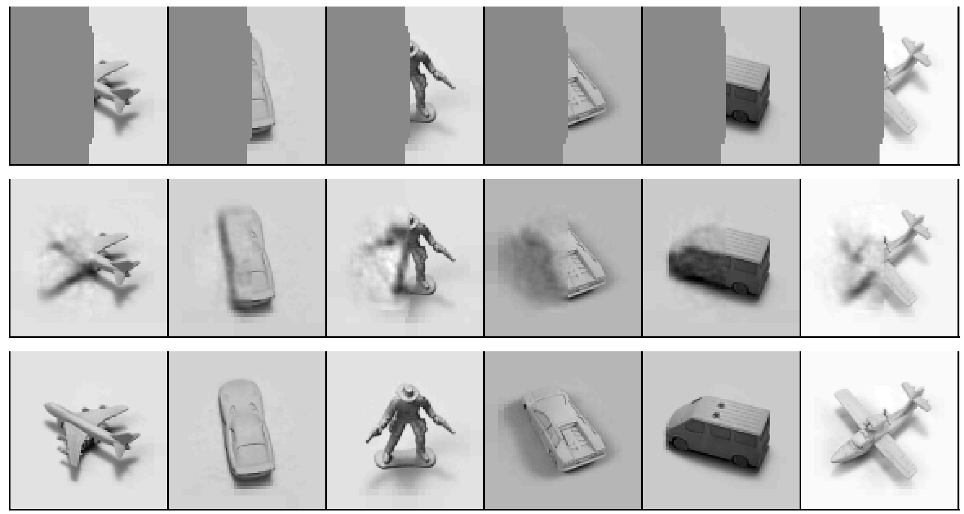


24,000 examples, 5 object categories, 5 different objects within each category, 6 lightning conditions, 9 elevations, 18 azimuths.

3-D Object Recognition

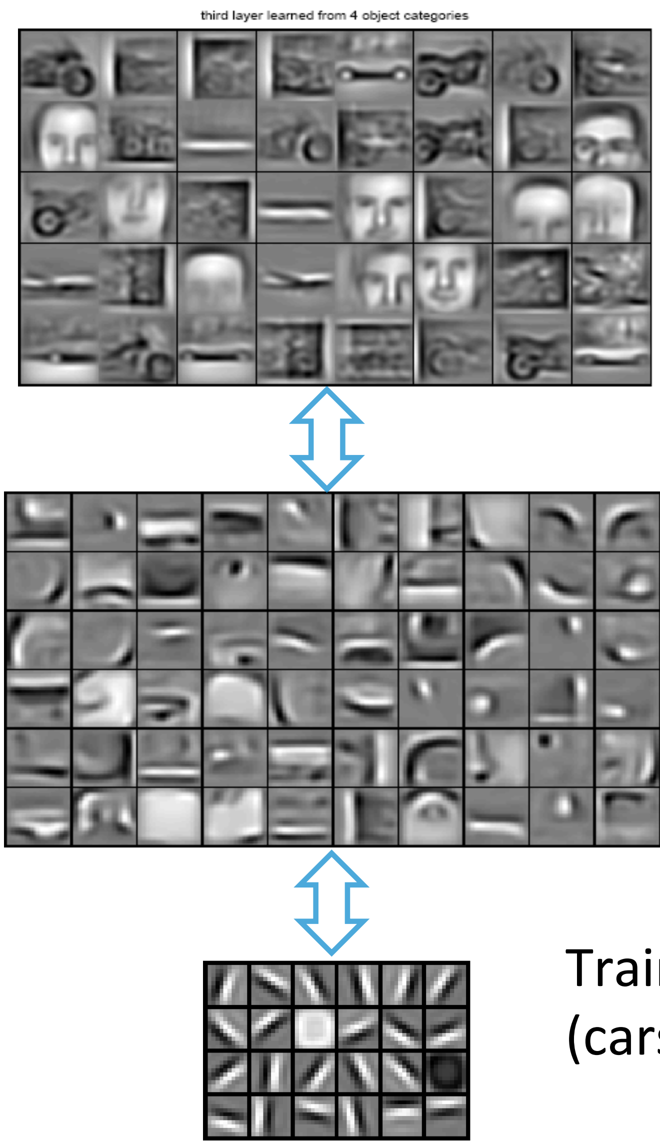
Learning Algorithm	Error
Logistic regression	22.5%
K-NN (LeCun 2004)	18.92%
SVM (Bengio & LeCun 2007)	11.6%
Deep Belief Net (Nair & Hinton 2009)	9.0%
DBM	**7.2%**

Pattern Completion



Permutation-invariant version.

Learning Part-based Hierarchy



Object parts.

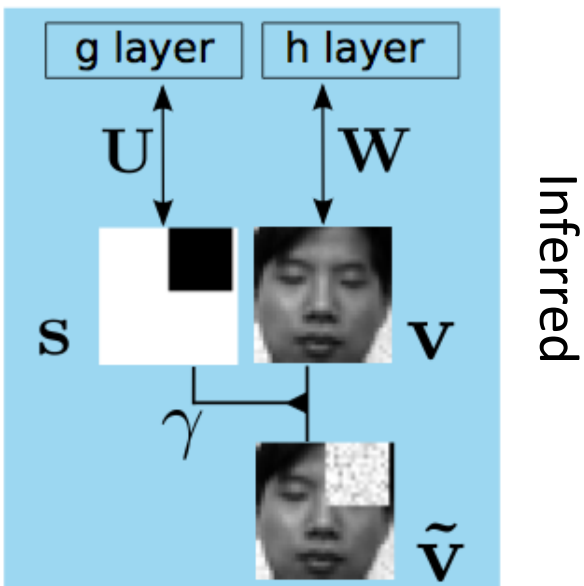
Combination of edges.

Trained from multiple classes
(cars, faces, motorbikes, airplanes).

Lee et.al., ICML 2009

Robust Boltzmann Machines

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



Observed

Gaussian RBM, modeling clean faces

Binary RBM modeling occlusions

$$E = \frac{1}{2}\sum \frac{(v_i - b_i)^2}{\sigma^2} - \mathbf{v}^\top W \mathbf{h} - \mathbf{s}^\top U \mathbf{g} + \frac{1}{2}\sum_i \gamma_i s_i (v_i - \tilde{v}_i)^2 + \frac{1}{2}\sum_i \frac{(\tilde{v}_i - \tilde{b}_i)^2}{\tilde{\sigma}_i^2}$$

Binary pixel-wise Mask

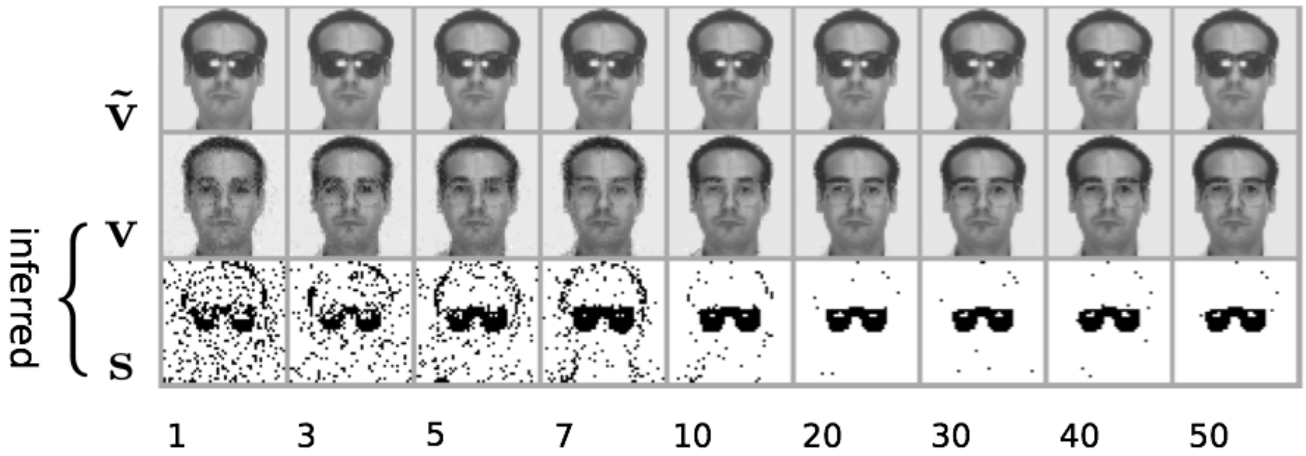
Gaussian noise

Relates to Le Roux, Heess, Shotton, and Winn, Neural Computation, 2011

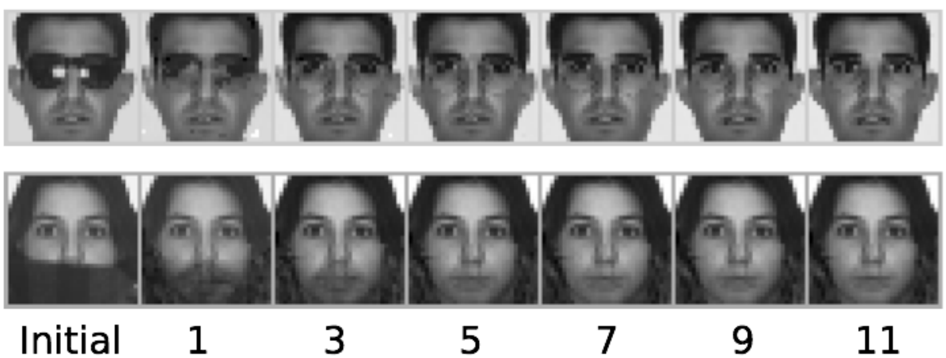
Eslami, Heess, Winn, CVPR 2012

Tang et. al., CVPR 2012

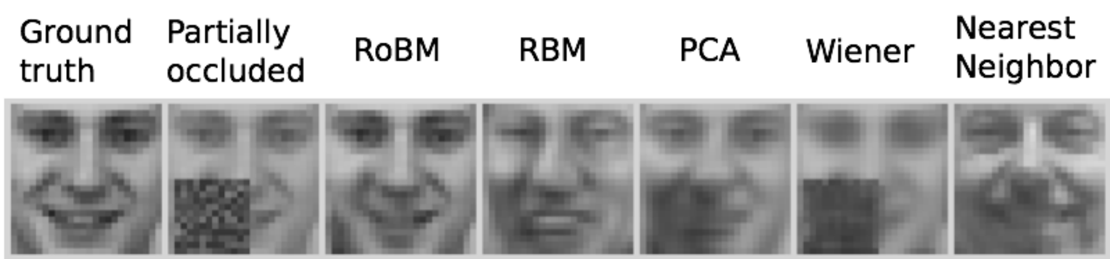
Robust Boltzmann Machines



Internal States of RoBM during learning.



Inference on the test subjects

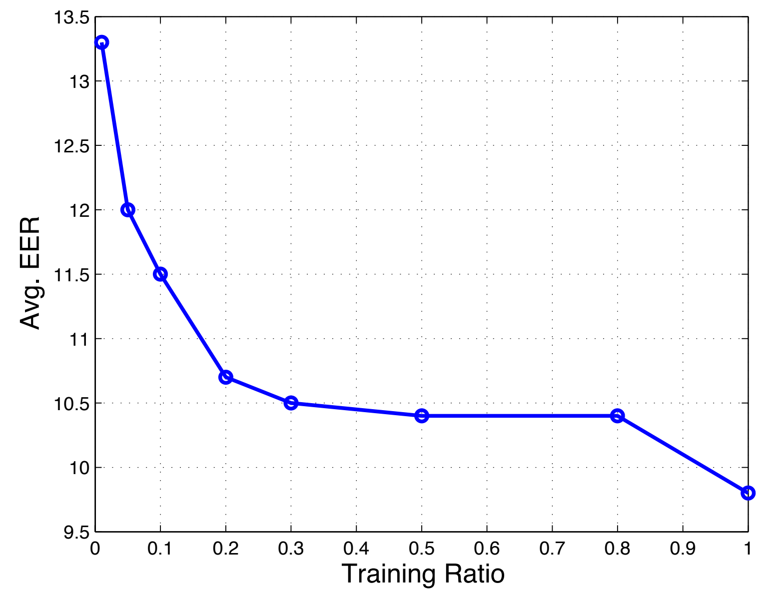


Comparing to Other Denoising Algorithms

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%

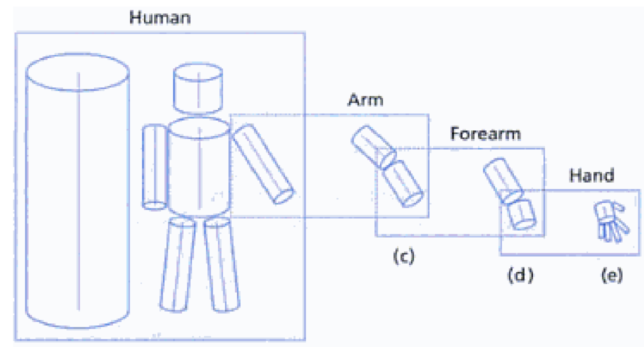


(Yaodong Zhang et.al. ICASSP 2012)

Learning Hierarchical Representations

Deep Boltzmann Machines:

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

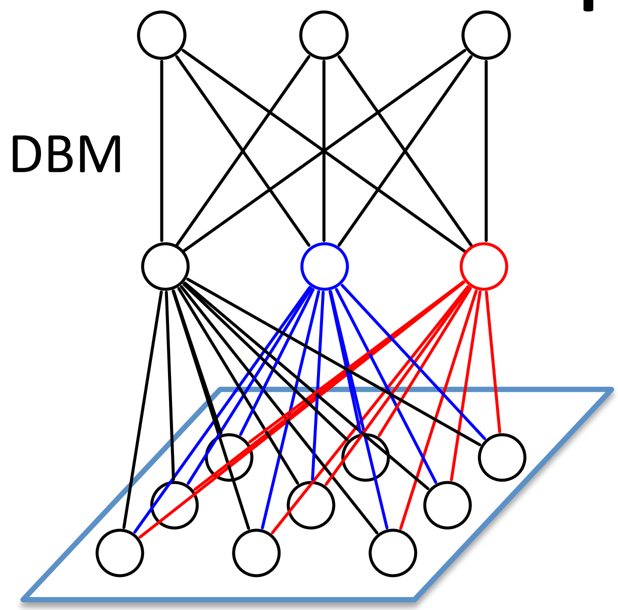


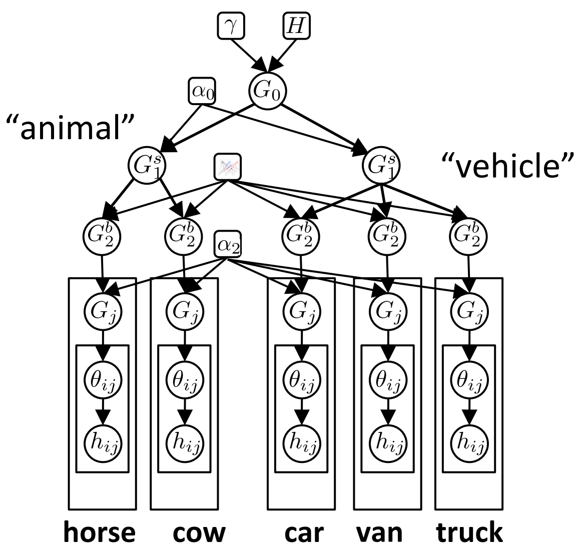
- Performs well in many application domains
- Combines bottom and top-down
- Fast Inference: fraction of a second
- Learning scales to millions of examples

Many examples, few categories

Next: Few examples, many categories – Transfer Learning

Talk Roadmap





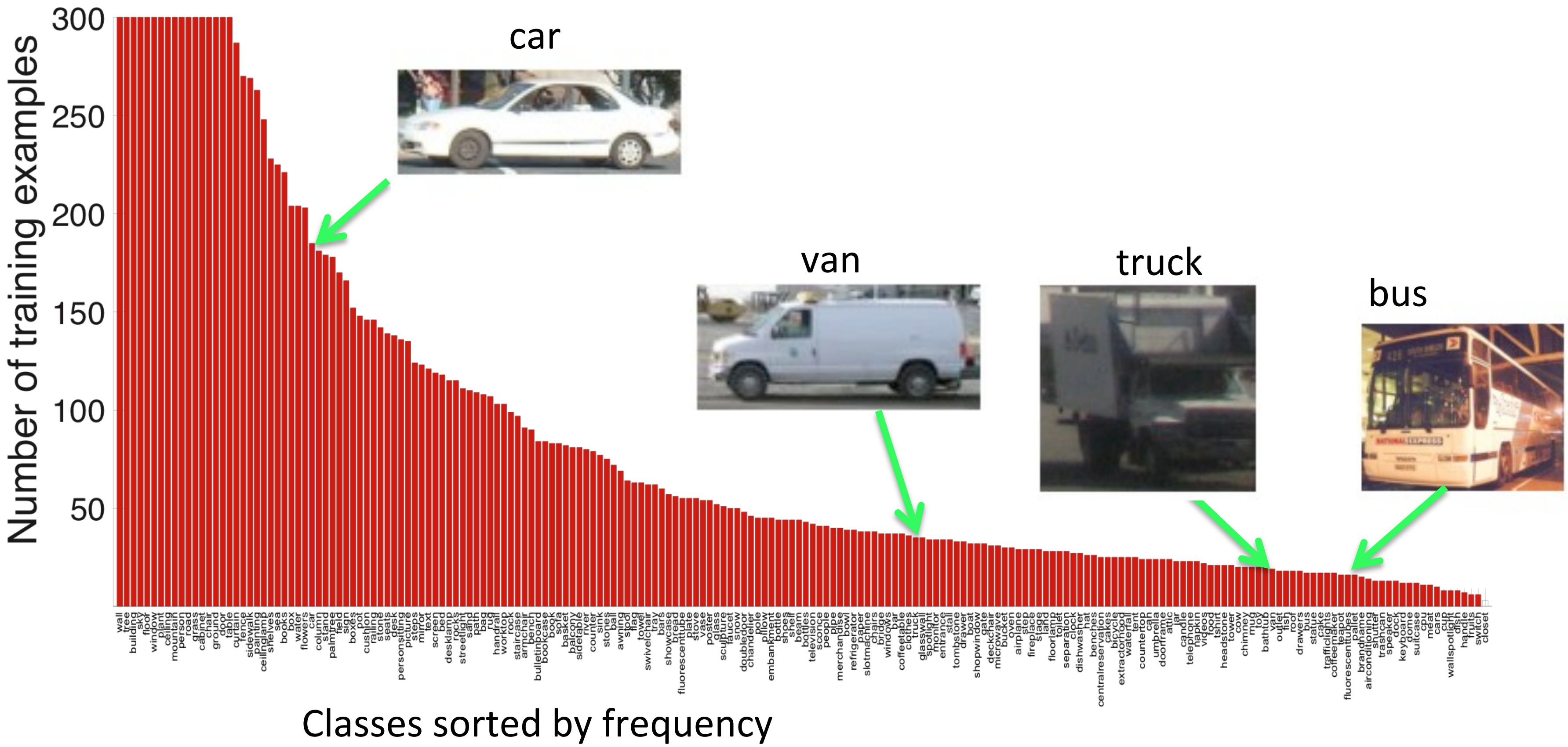
- Unsupervised Feature Learning
  - Restricted Boltzmann Machines
  - Deep Belief Networks
  - Deep Boltzmann Machines
- Transfer Learning with Deep Models
- Multimodal Learning

One-shot Learning

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

Learning from Few Examples

SUN database



Rare objects are similar to frequent objects

Traditional Supervised Learning

Segway

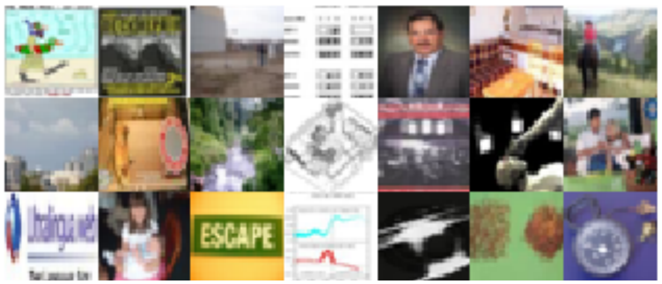
Motorcycle

Test:
What is this?

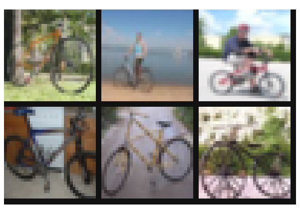
Learning to Transfer

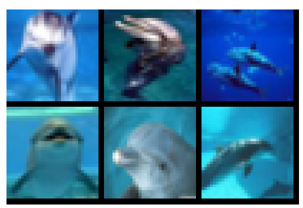
Background Knowledge

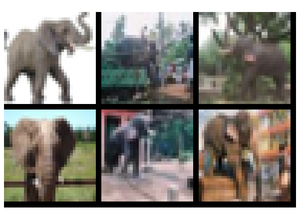
Millions of unlabeled images

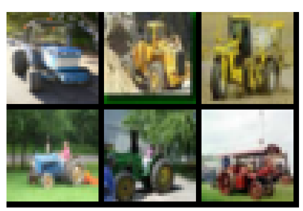


Some labeled images


Bicycle


Dolphin


Elephant


Tractor

Learn to Transfer Knowledge

Learn novel concept from one example

Test:
What is this?

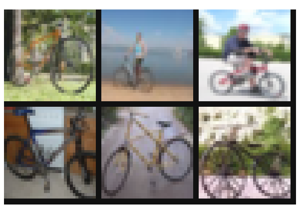
Learning to Transfer

Background Knowledge

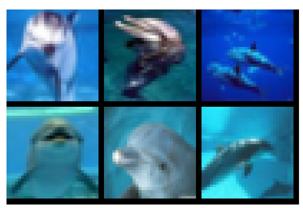
Millions of unlabeled images

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

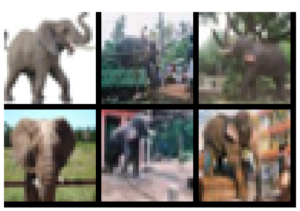
Some labeled images



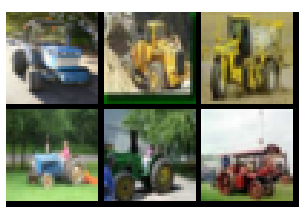
Bicycle



Dolphin



Elephant



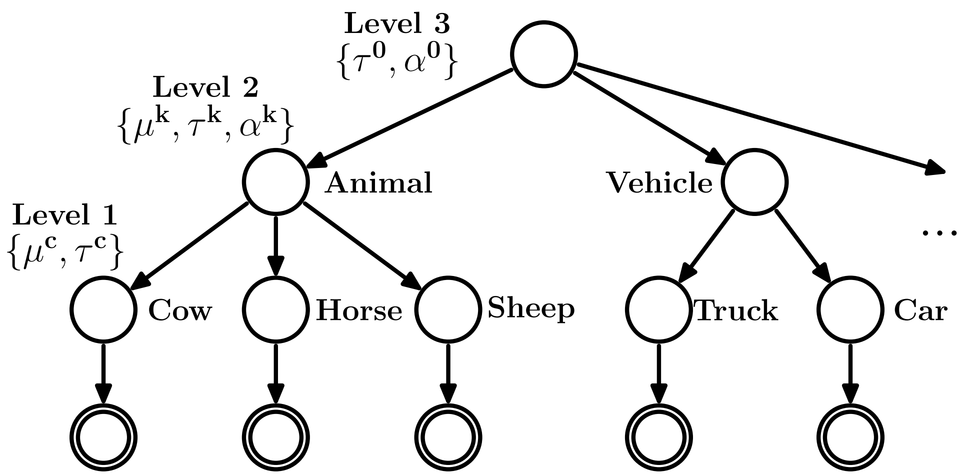
Tractor

Learn to Transfer Knowledge

Learn novel concept from one example

Test:
What is this?

One-Shot Learning



Hierarchical Bayesian Models

Hierarchical Prior.

Probability of observed data given parameters

Prior probability of weight vector W

Posterior probability of parameters given the training data D.

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

- Fei-Fei, Fergus, and Perona, TPAMI 2006
- E. Bart, I. Porteous, P. Perona, and M. Welling, CVPR 2007
- Miller, Matsakis, and Viola, CVPR 2000
- Sivic, Russell, Zisserman, Freeman, and Efros, CVPR 2008

Hierarchical-Deep Models

HD Models: Compose hierarchical Bayesian models with deep networks, two influential approaches from unsupervised learning

Deep Networks:

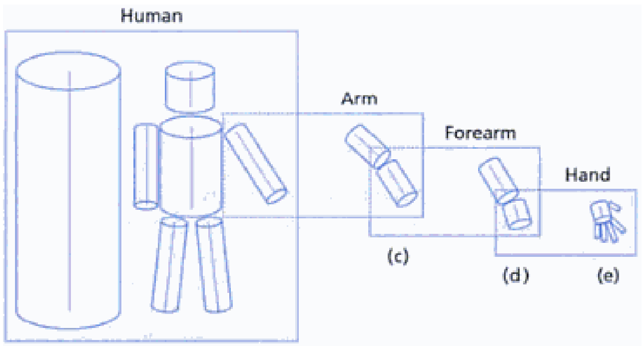
- learn multiple **layers of nonlinearities.**
- trained in unsupervised fashion -- **unsupervised feature learning** – no need to rely on human-crafted input representations.
- **labeled data** is used to slightly adjust the model for a specific task.

Hierarchical Bayes:

- **explicitly represent category hierarchies** for sharing abstract knowledge.
- explicitly identify only a **small number of parameters** that are relevant to the new concept being learned.

Deep Nets

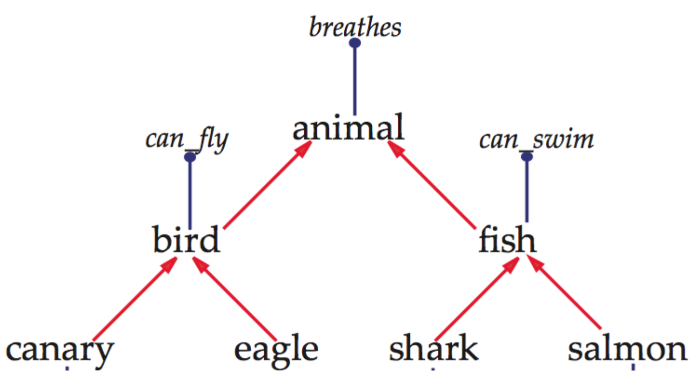
Part-based Hierarchy



Marr and Nishihara (1978)

Hierarchical Bayes

Category-based Hierarchy



Collins & Quillian (1969)

Motivation

Learning to transfer knowledge:

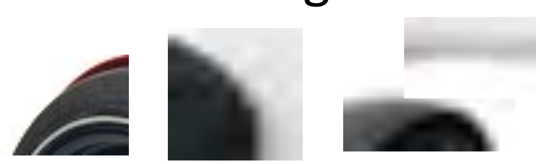
Hierarchical

- Super-category: “A segway looks like a funny kind of vehicle”.

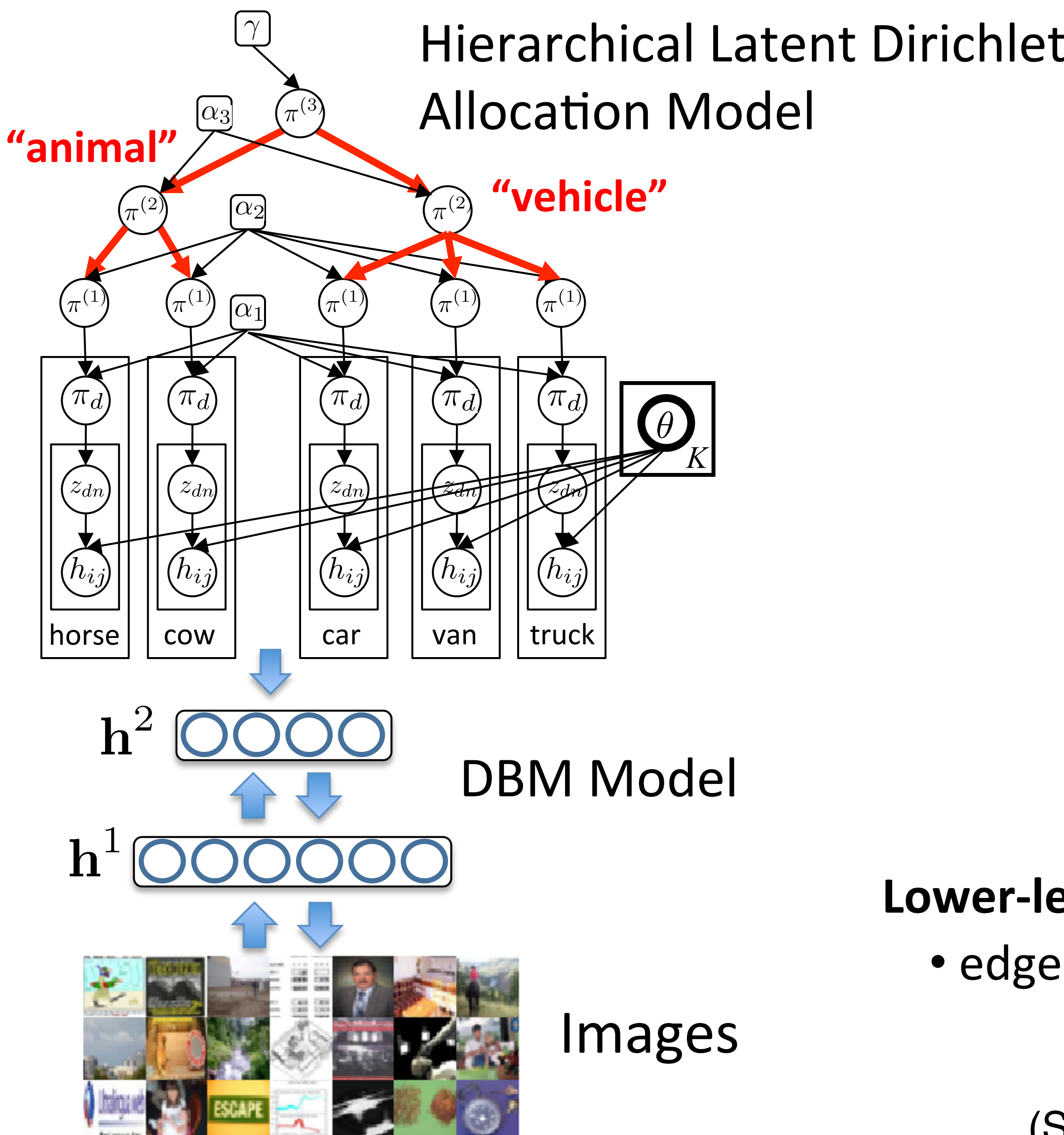
- Higher-level features, or parts, shared with other classes:
  - ➢ wheel, handle, post

- Lower-level features:
  - ➢ edges, composition of edges

Deep



Hierarchical Generative Model

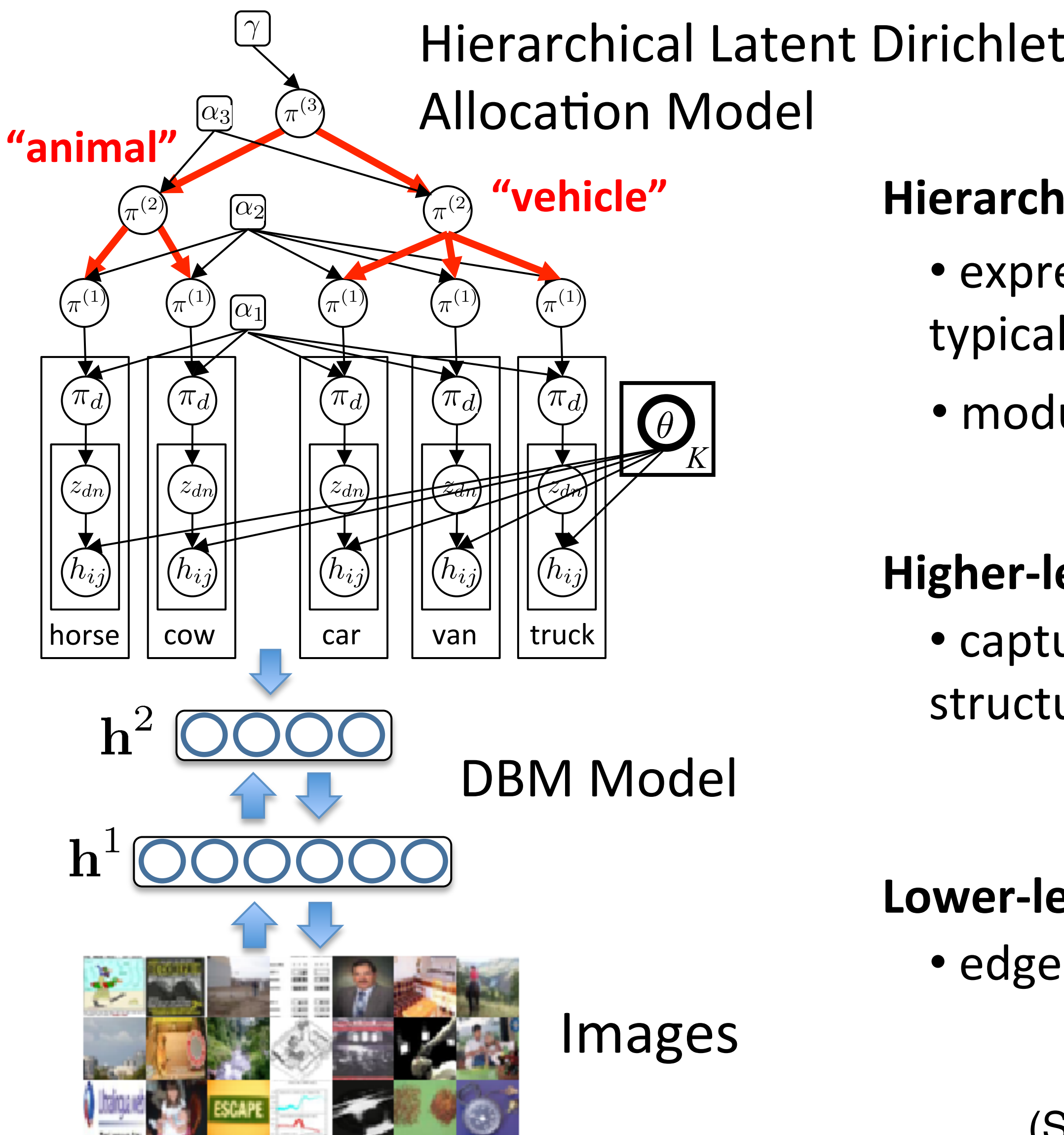


Lower-level generic features:

- edges, combination of edges

(Salakhutdinov, Tenenbaum, Torralba, 2011)

Hierarchical Generative Model



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:

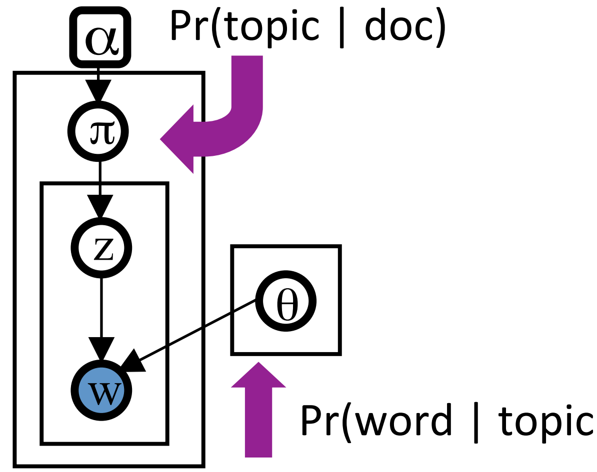
- edges, combination of edges

(Salakhutdinov, Tenenbaum, Torralba, 2011)

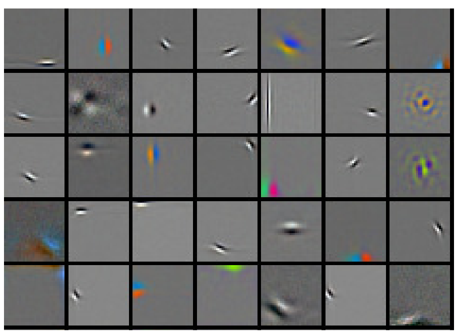
Intuition

$\mathbf{h}^3 \sim \text{LDA prior}$

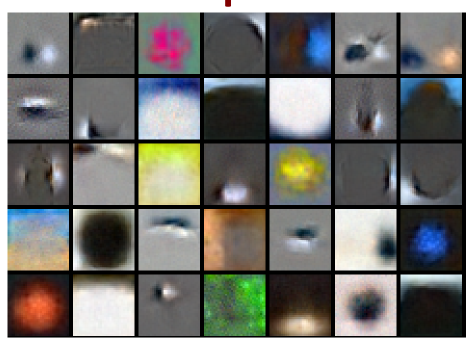
Words ⇔ activations of DBM's top-level units.
Topics ⇔ distributions over top-level units, or higher-level parts.



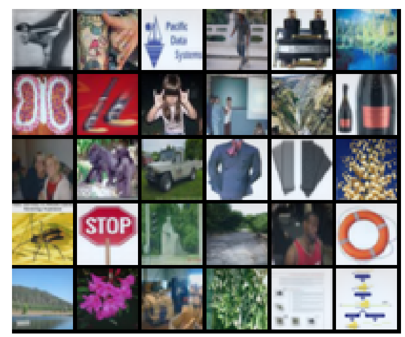
DBM generic features:
Words



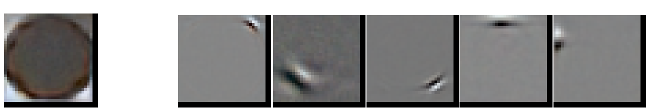
LDA high-level features:
Topics



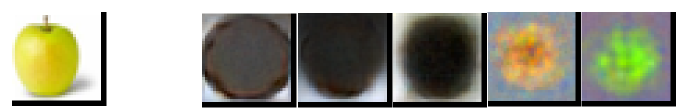
Images
Documents



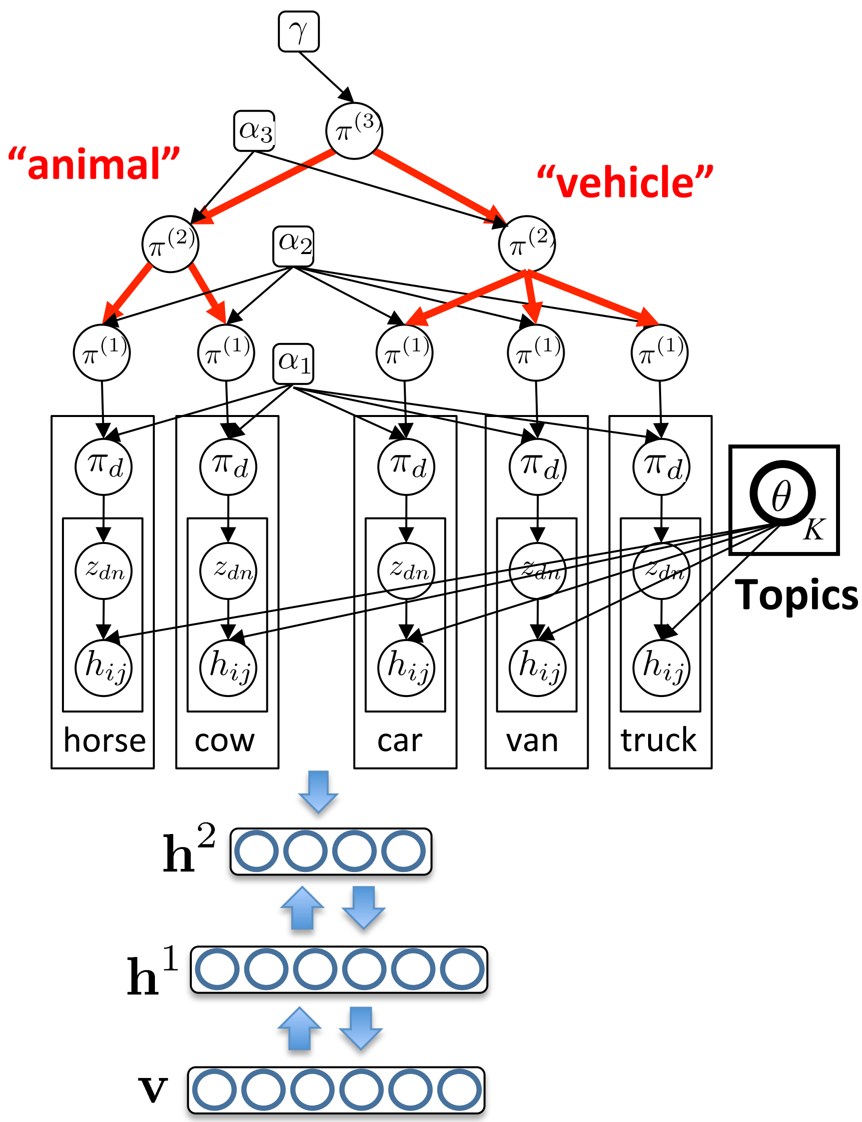
Each topic is made up of words.



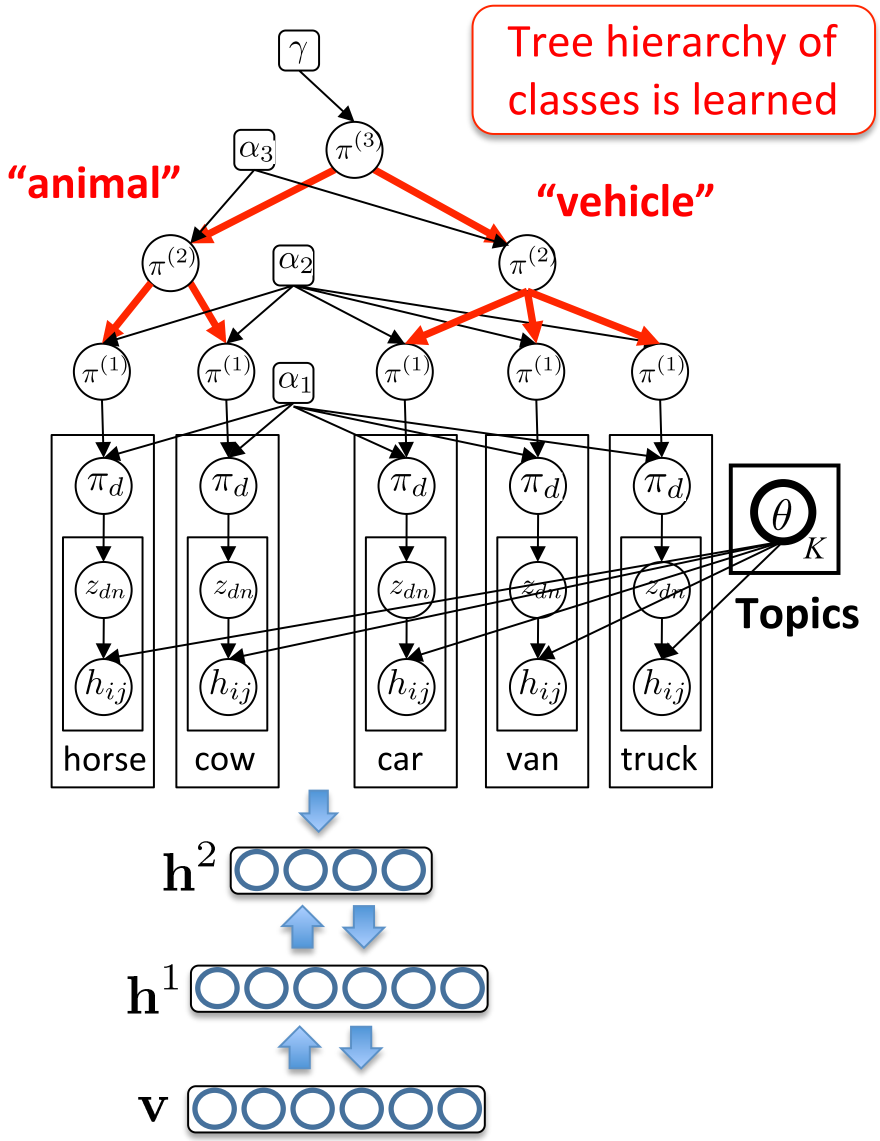
Each document is made up of topics.



Hierarchical Deep Model



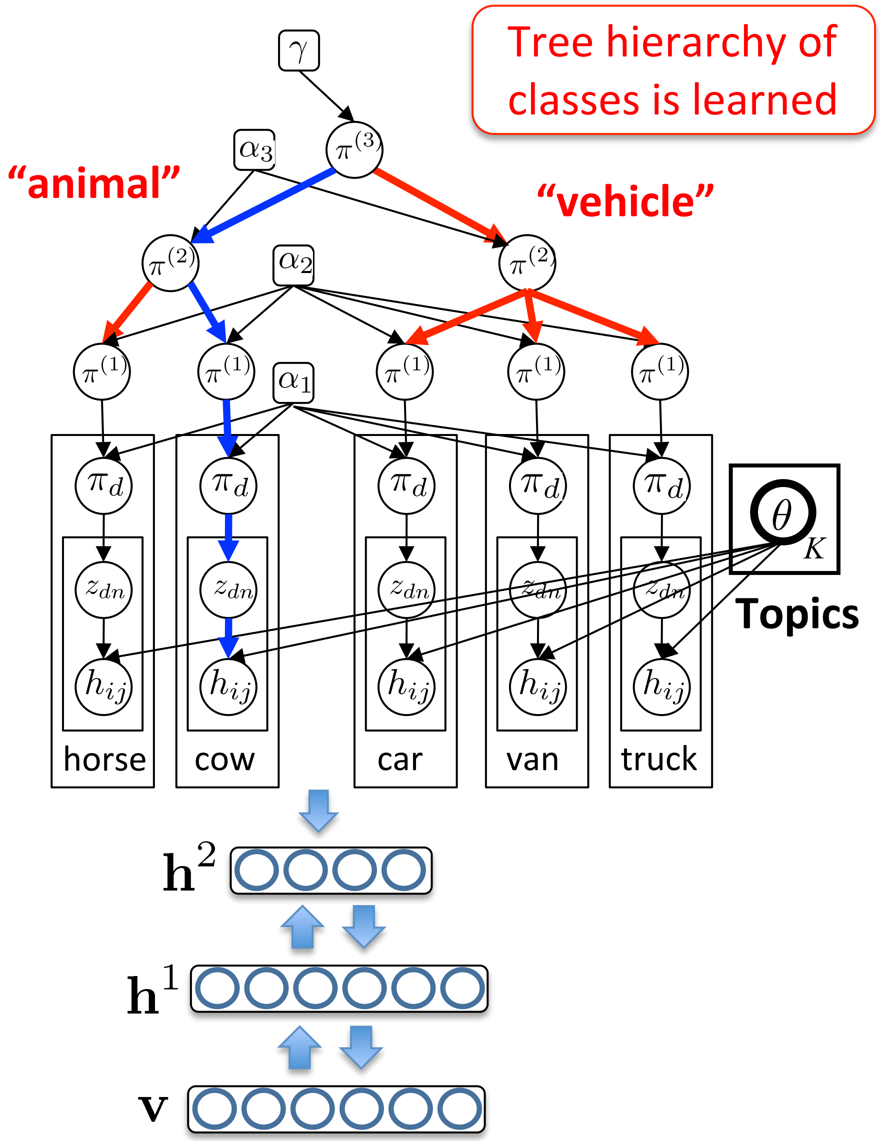
Hierarchical Deep Model



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

prior: a nonparametric prior over tree structures.

Hierarchical Deep Model

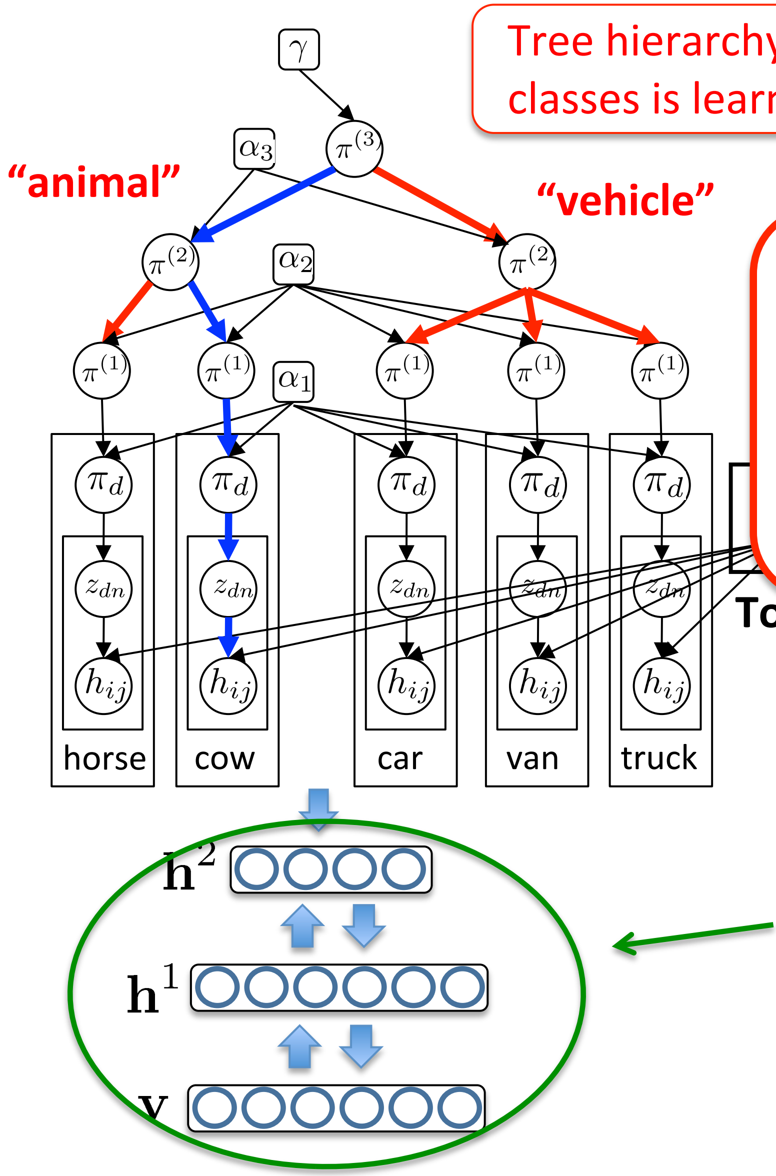


$\mathbf{z} \sim \mathrm{nCRP}$ (**Nested Chinese Restaurant Process**)
prior: a nonparametric prior over tree structures.

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

Hierarchical Deep Model

Tree hierarchy of classes is learned



Unlike standard statistical models, in addition to inferring parameters, we also infer the hierarchy for sharing those parameters.

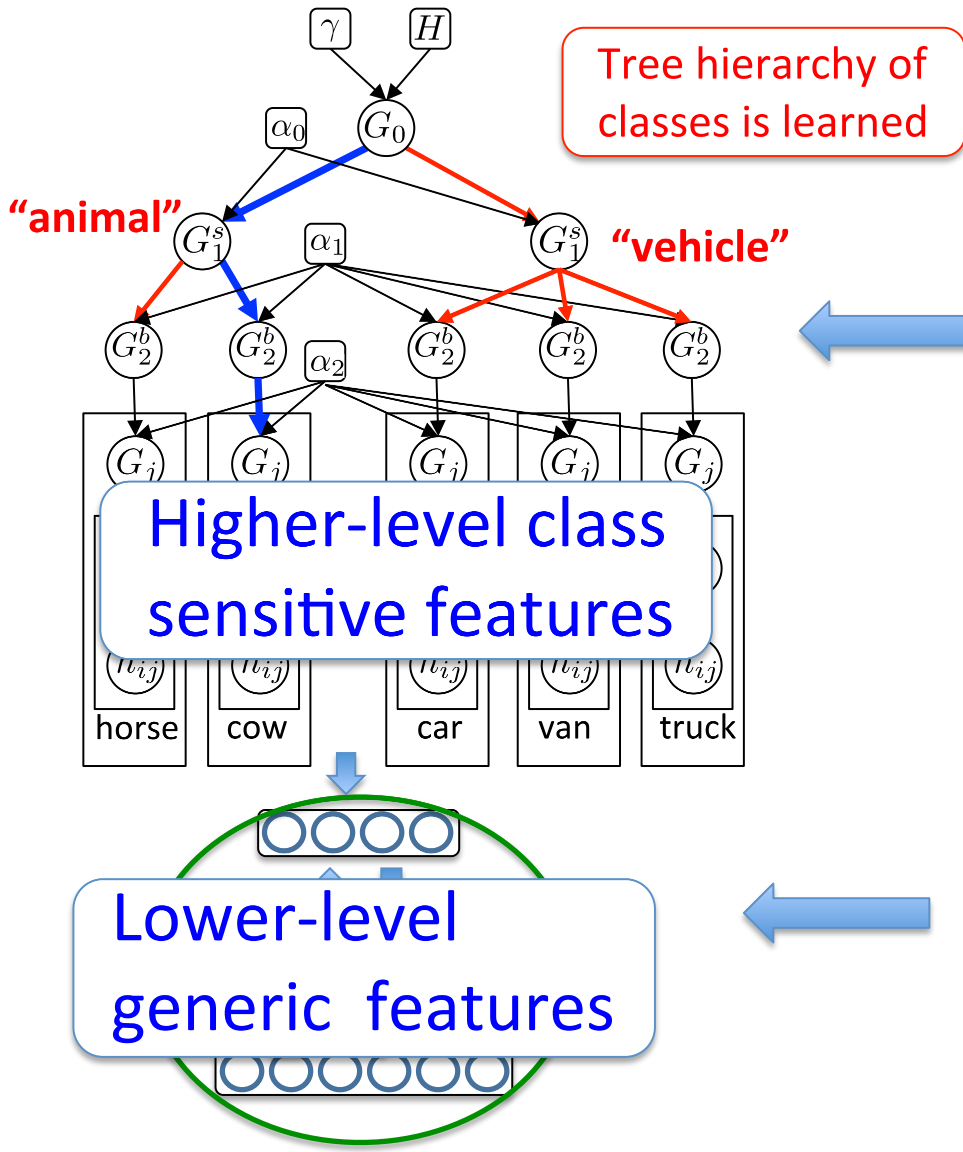
Topics

share higher-level features, or parts.

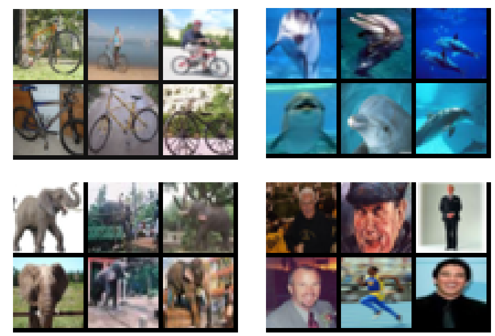
$\mathbf{v}|\mathbf{h}^3 \sim \mathrm{DBM}$ **Conditional Deep Boltzmann Machine.**

Enforce (approximate) global consistency through many local constraints.

CIFAR Object Recognition

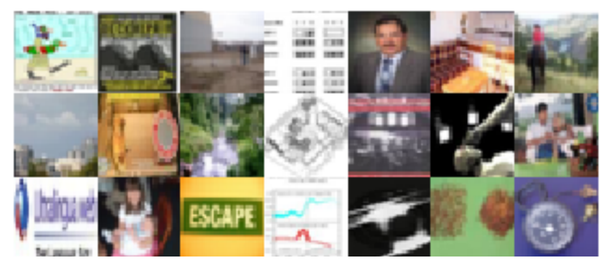


50,000 images of 100 classes



Inference: Markov chain Monte Carlo – Later!

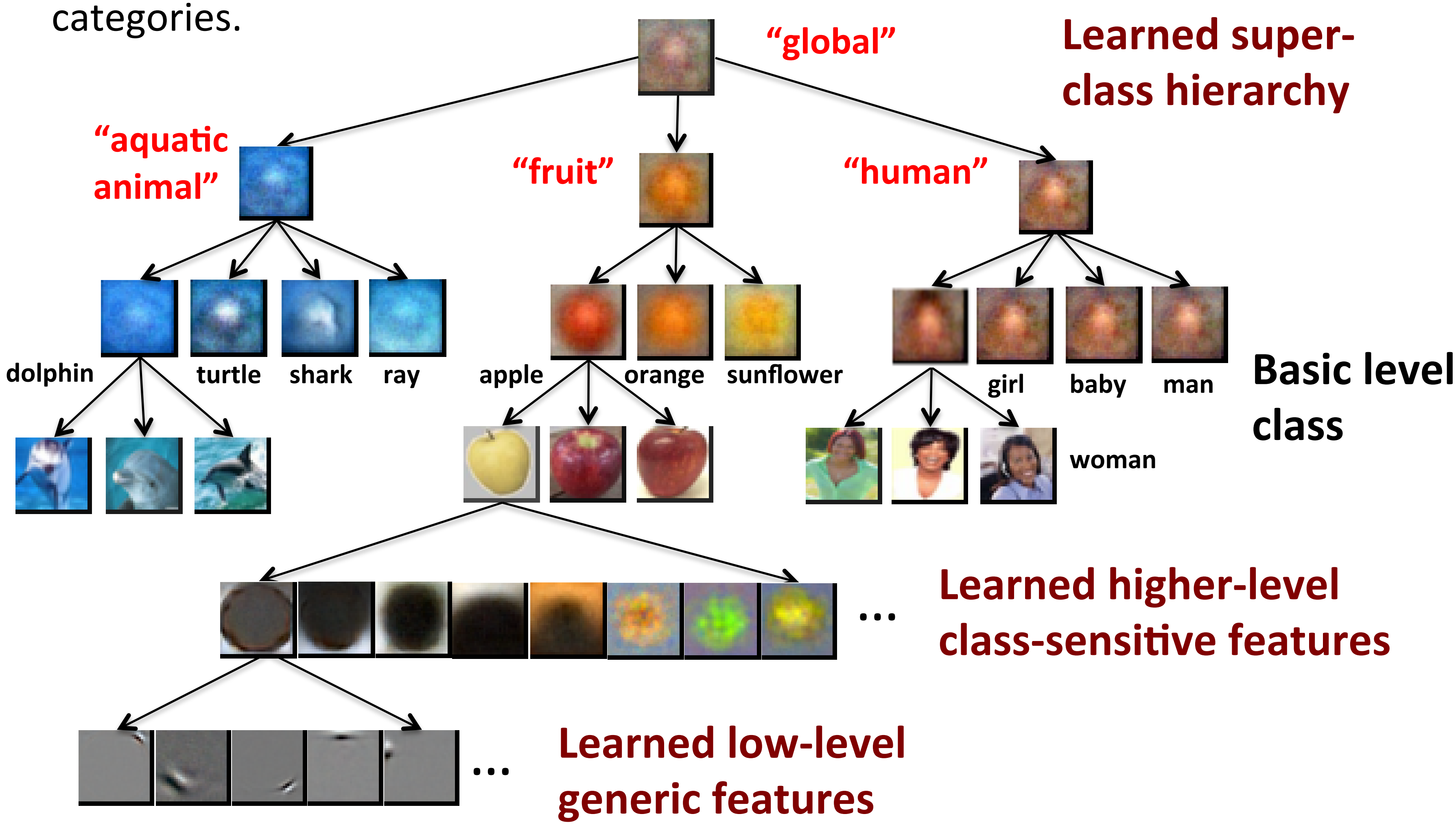
4 million unlabeled images



32 x 32 pixels x 3 RGB

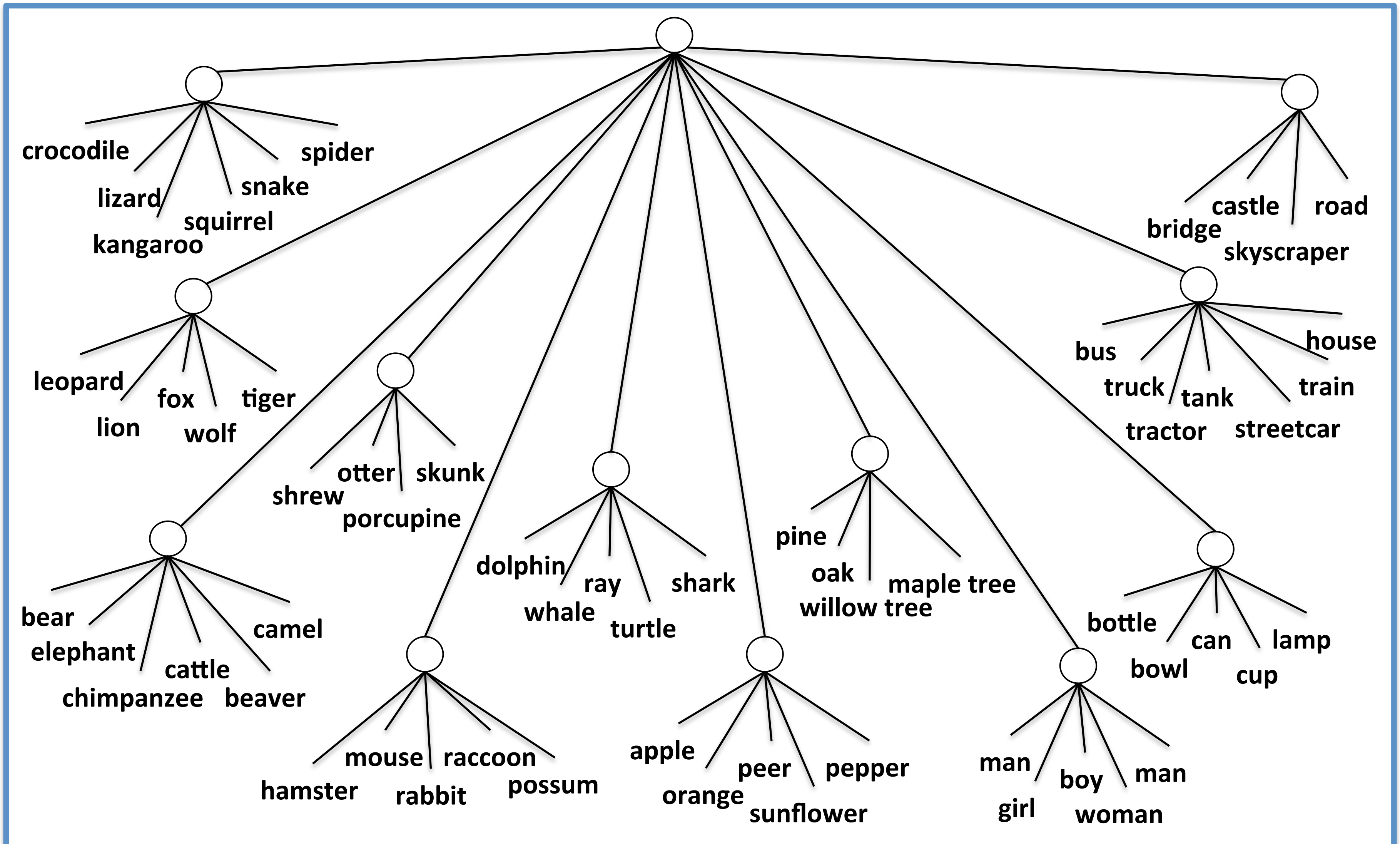
Learning to Learn

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

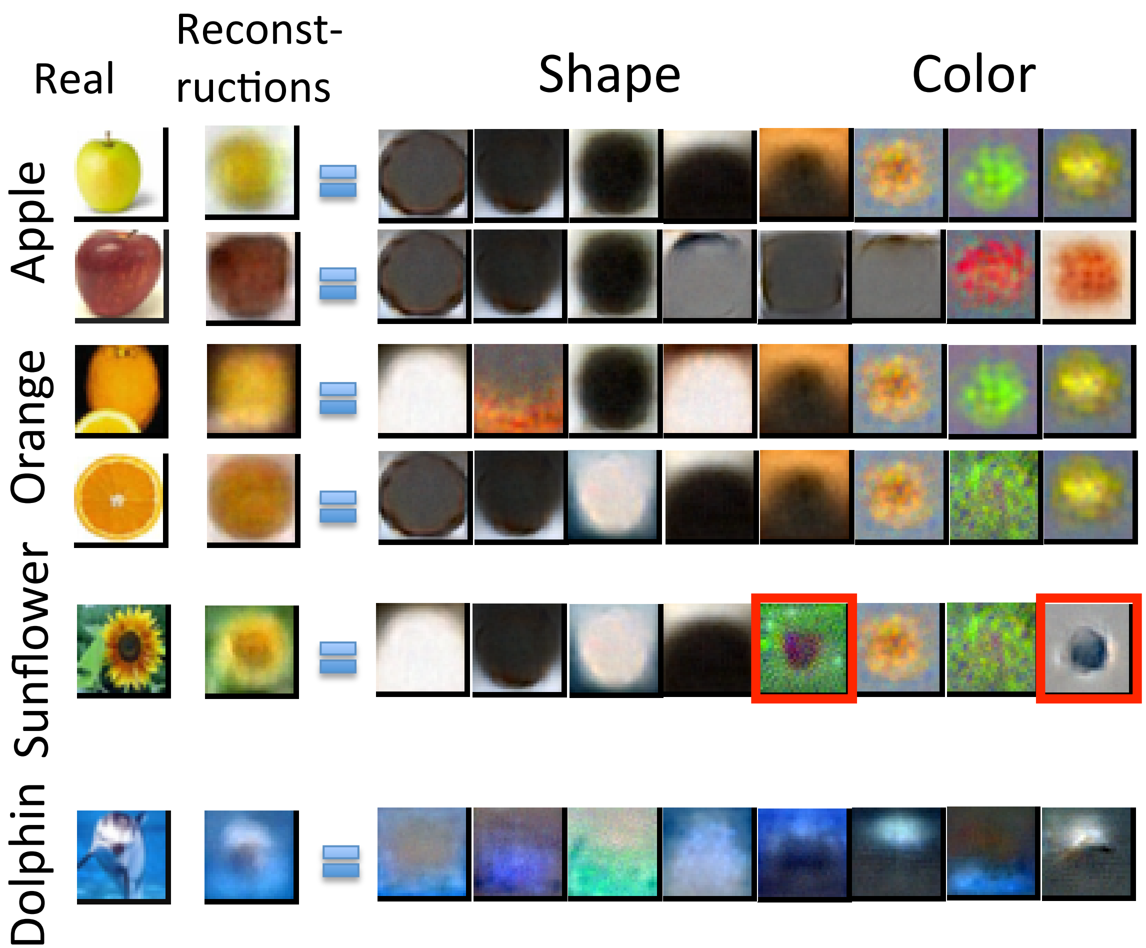


Learning to Learn

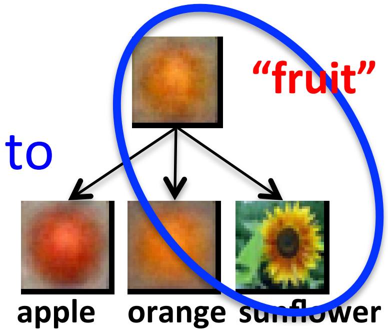
The model learns how to share the knowledge across many visual

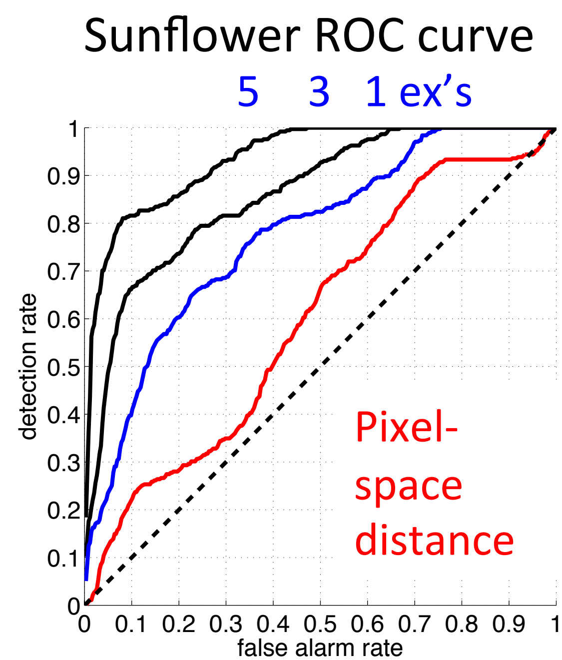


Sharing Features



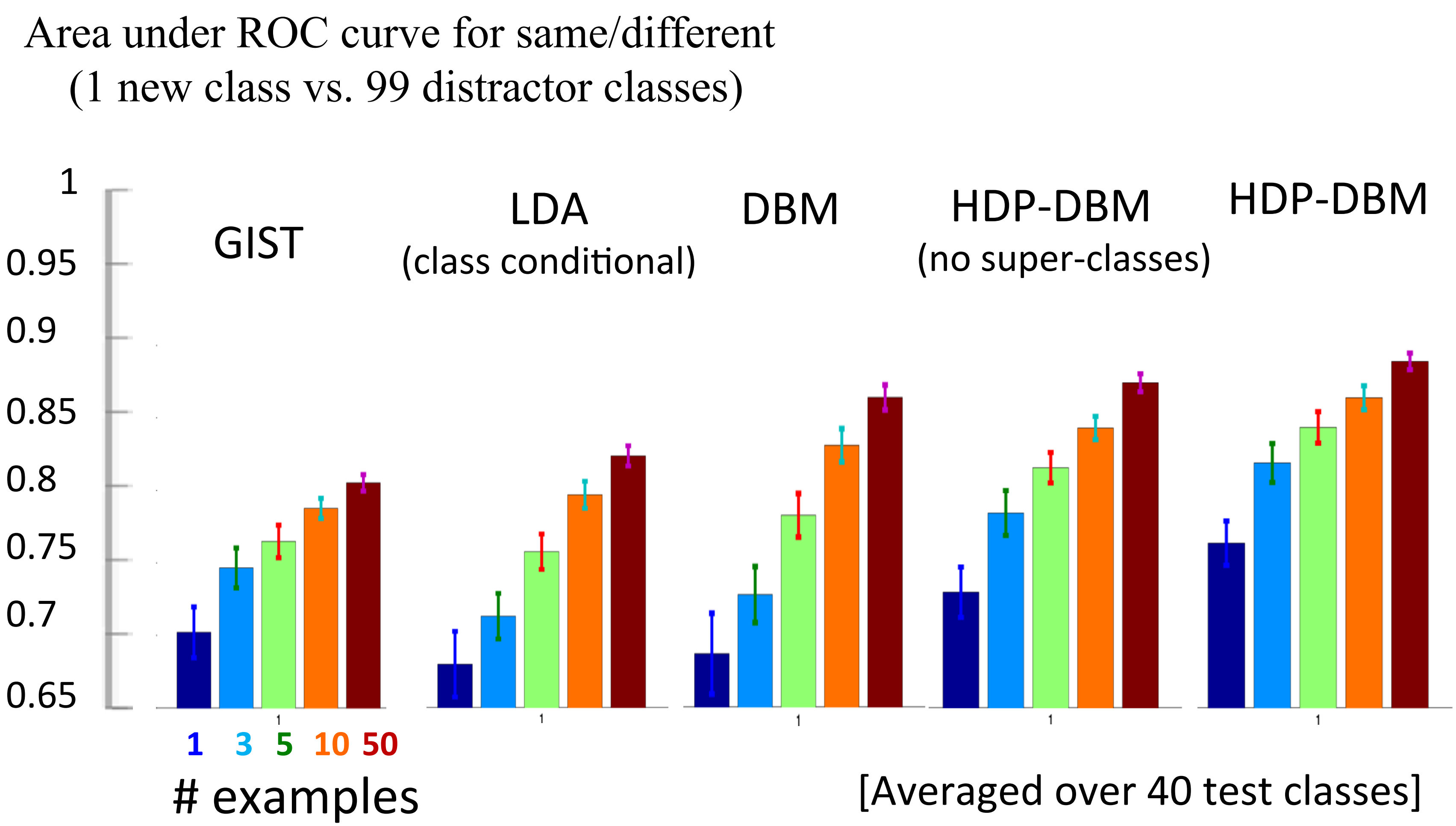
Learning to Learn





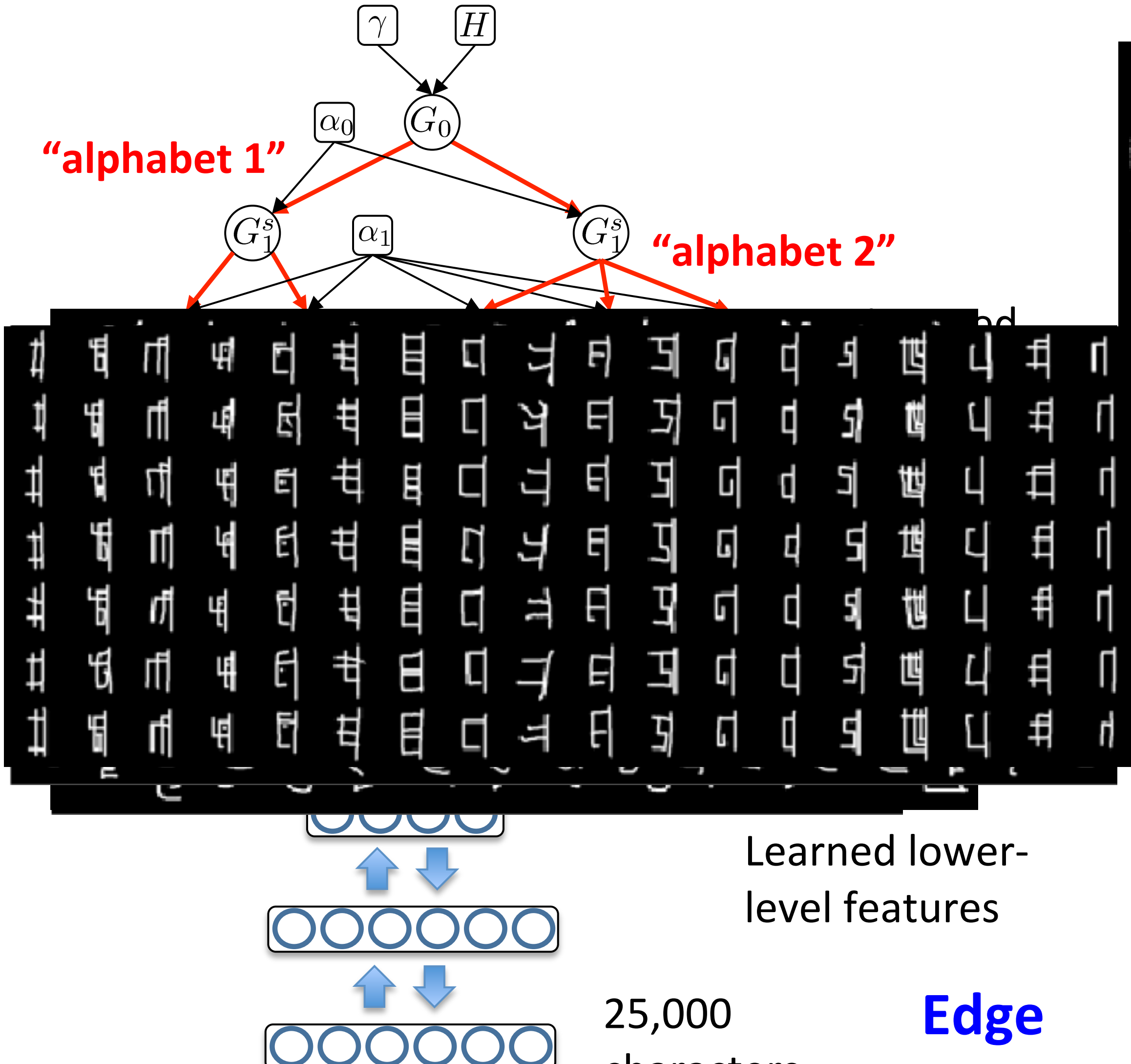
Learning to Learn: Learning a hierarchy for sharing parameters – rapid learning of a novel concept.

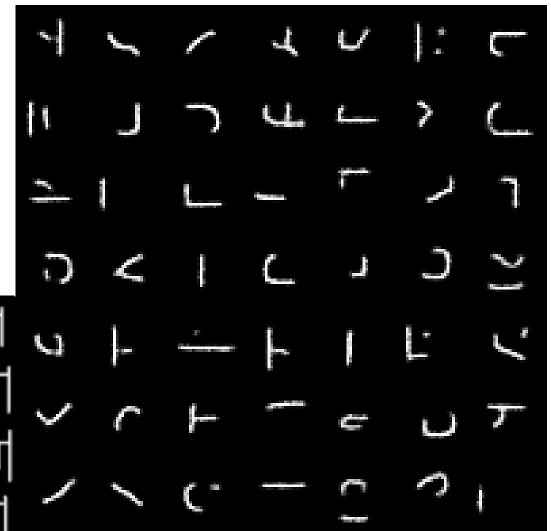
Object Recognition

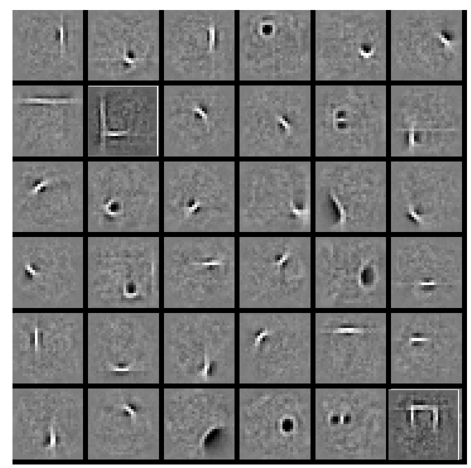


Our model outperforms standard computer vision features (e.g. GIST).

Handwritten Character Recognition







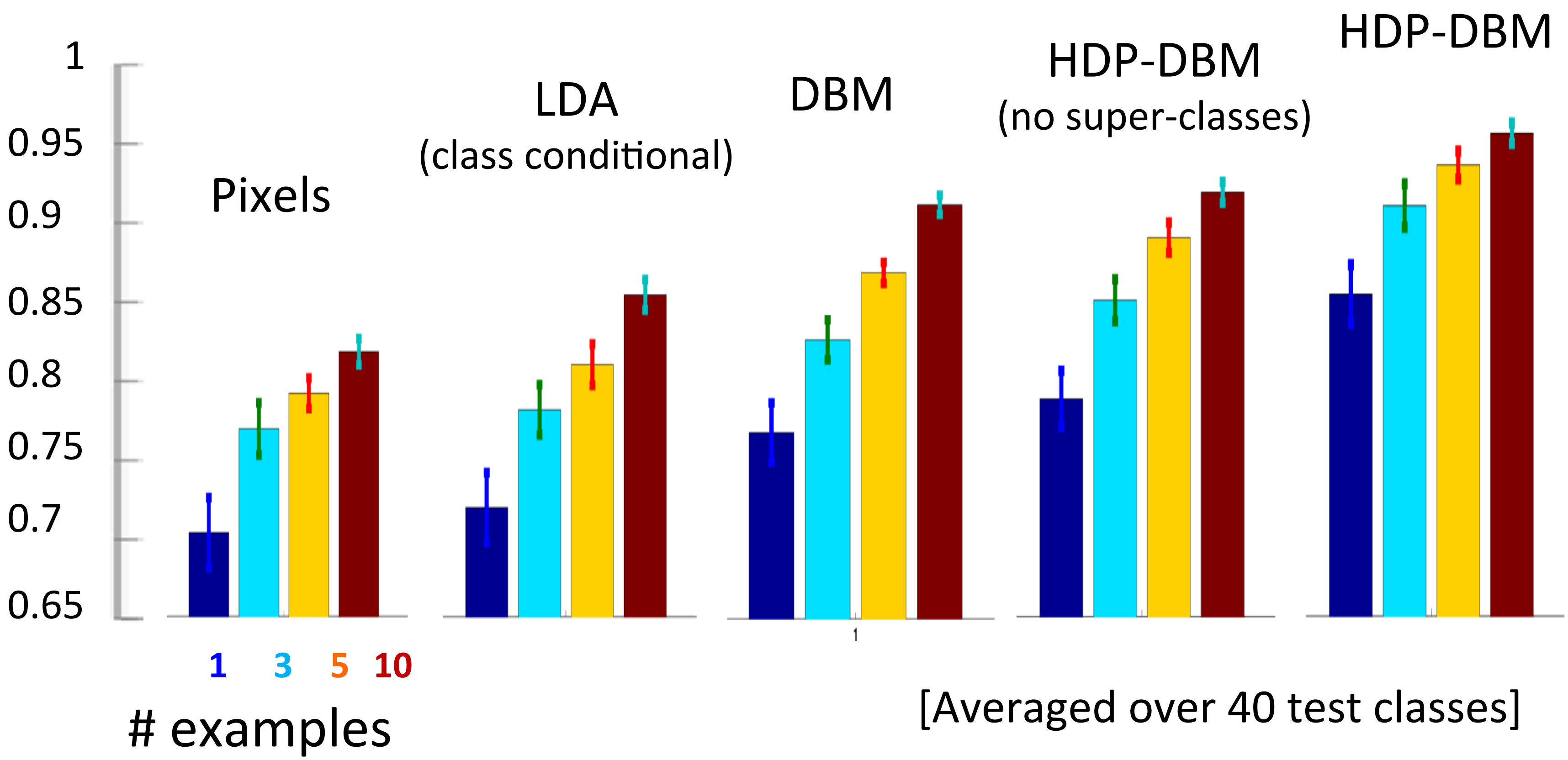
Learned lower-level features

25,000 characters

Edges

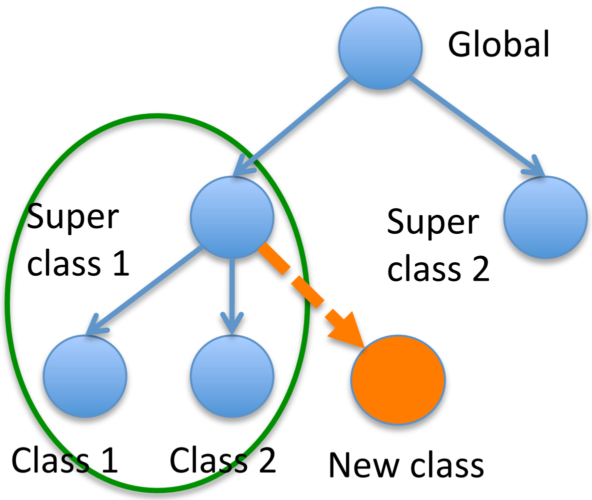
Handwritten Character Recognition

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

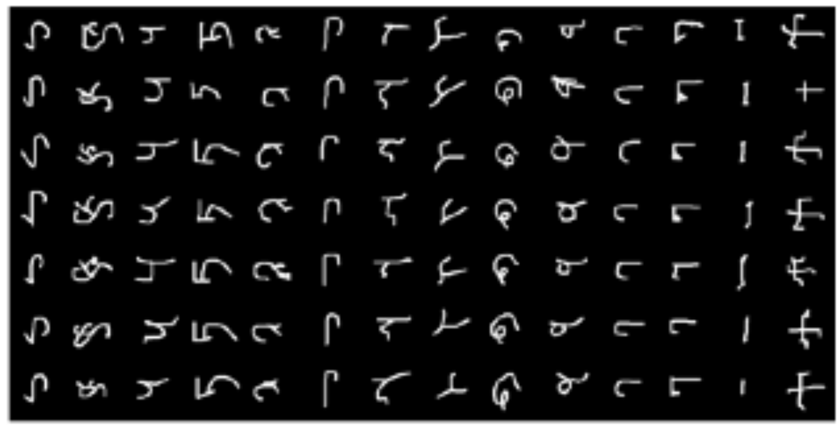


[Averaged over 40 test classes]

Simulating New Characters

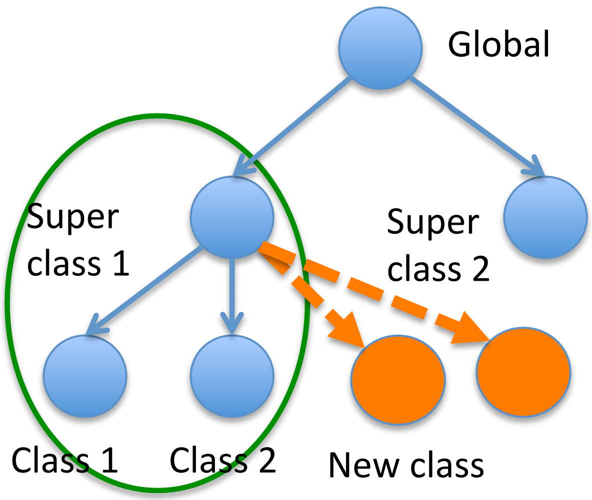


Real data within super class

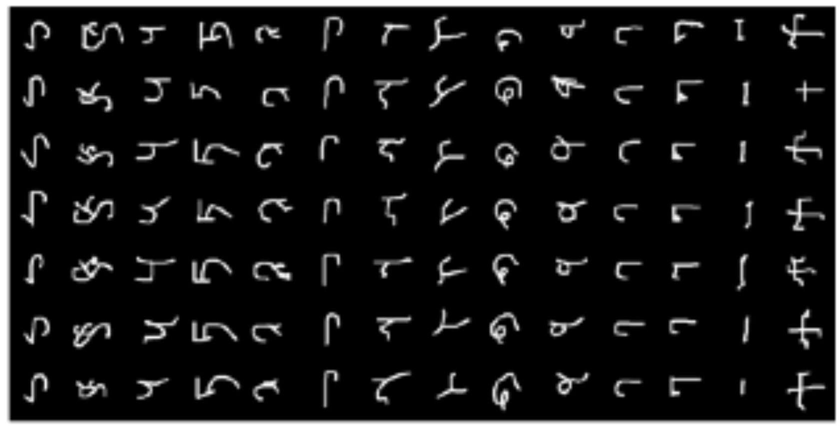


Simulated new characters

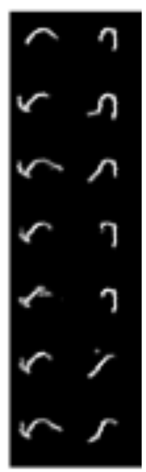
Simulating New Characters



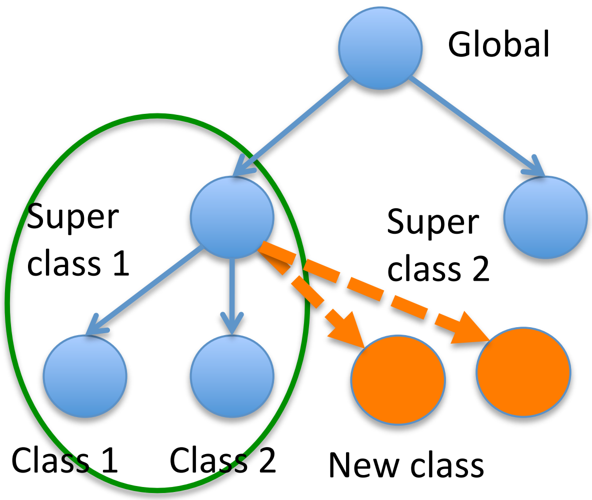
Real data within super class



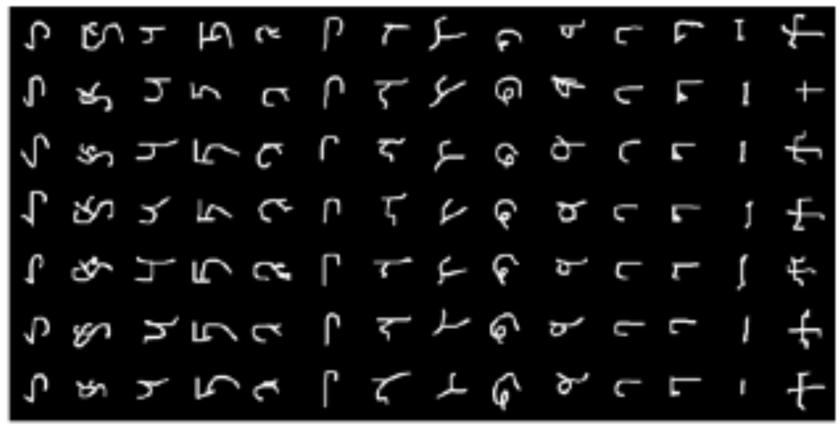
Simulated new characters



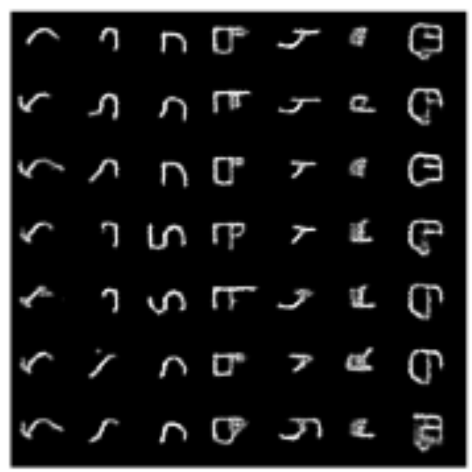
Simulating New Characters



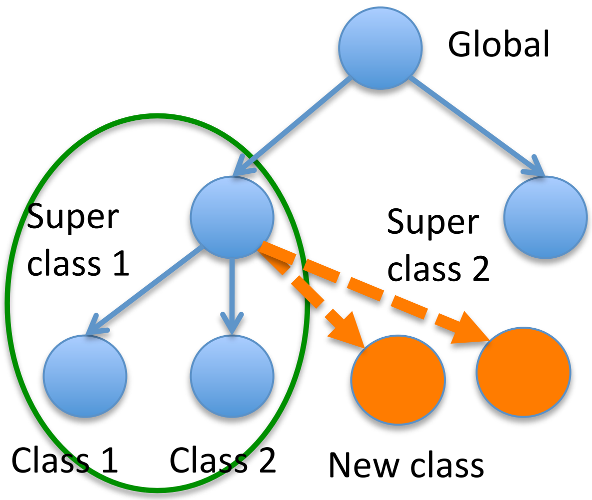
Real data within super class



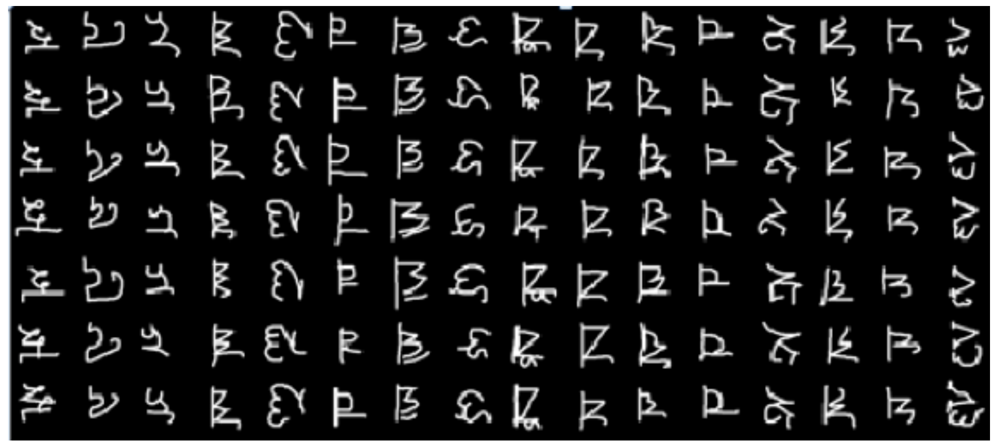
Simulated new characters



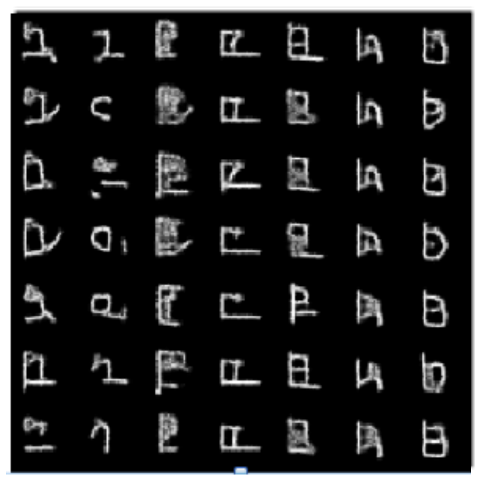
Simulating New Characters



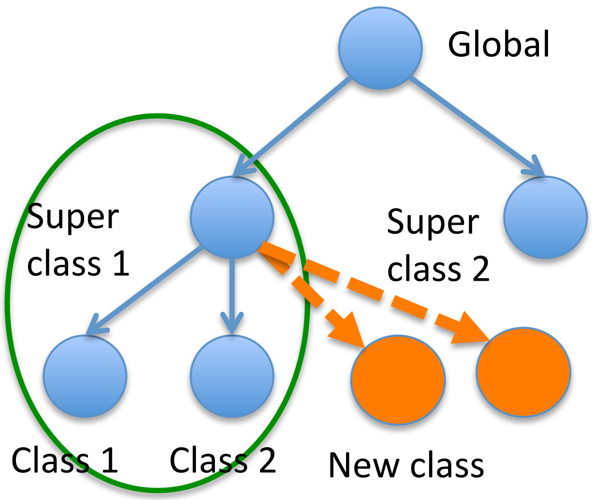
Real data within super class



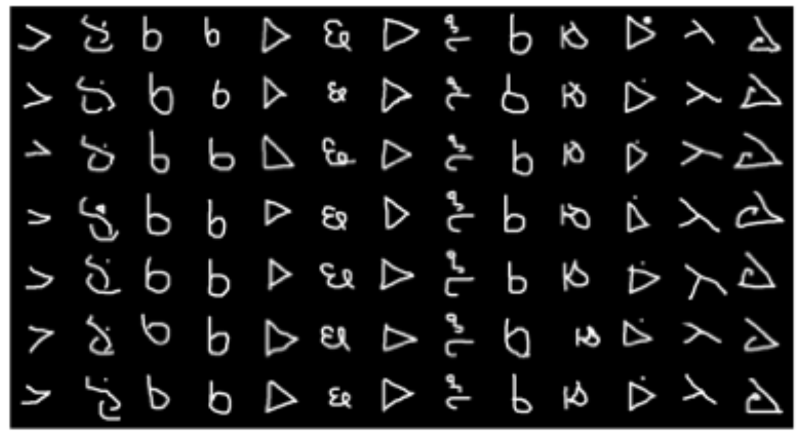
Simulated new characters



Simulating New Characters

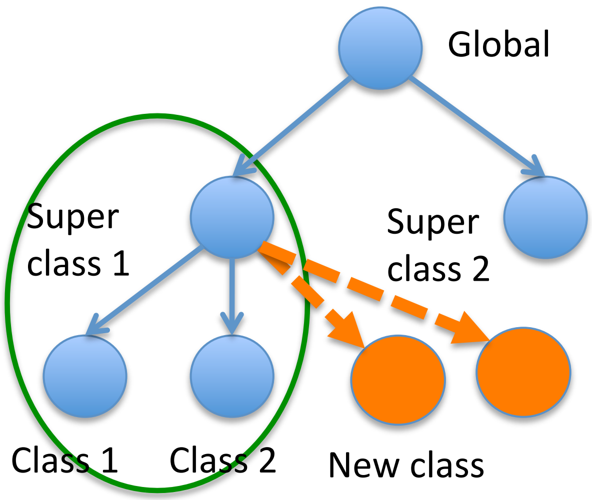


Real data within super class

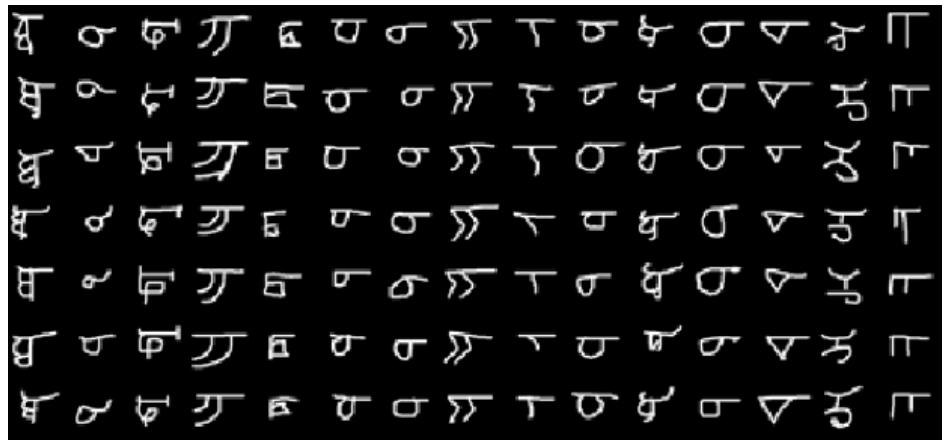


Simulated new characters

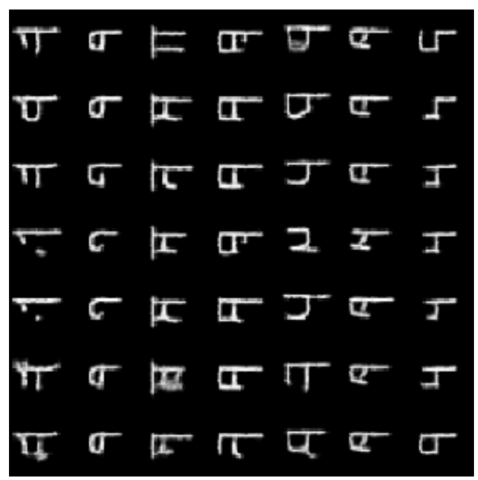
Simulating New Characters



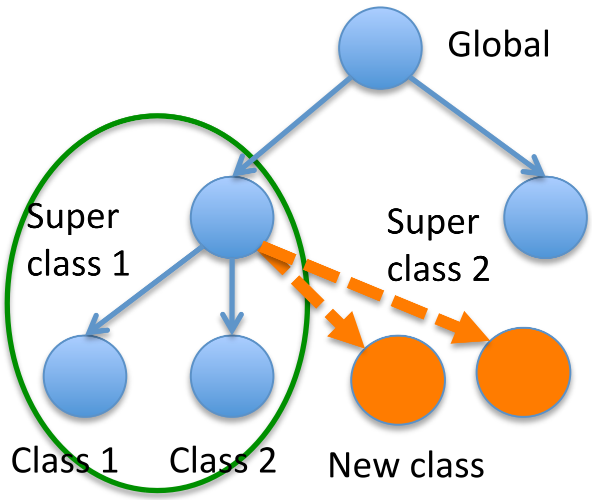
Real data within super class



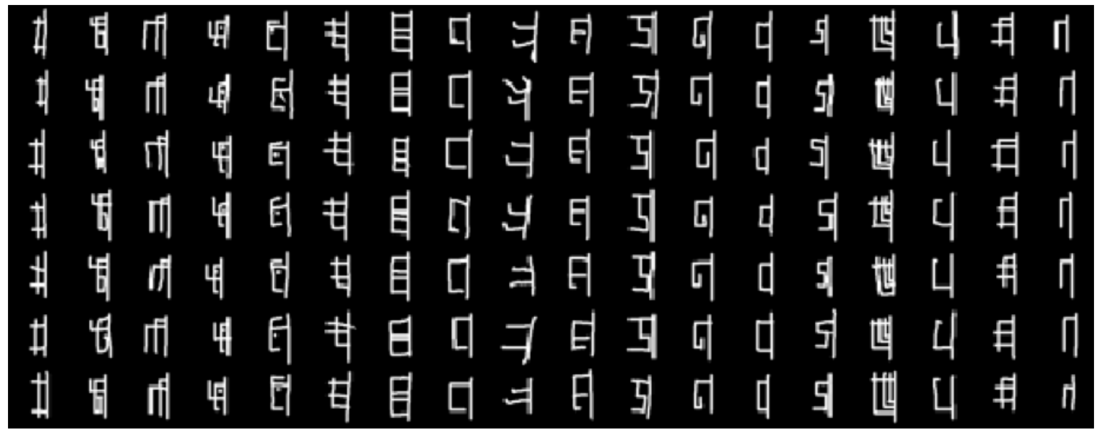
Simulated new characters



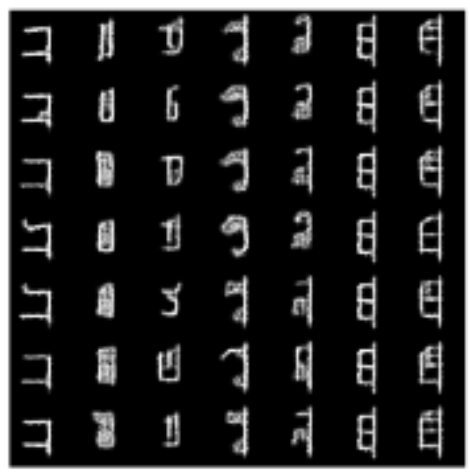
Simulating New Characters



Real data within super class



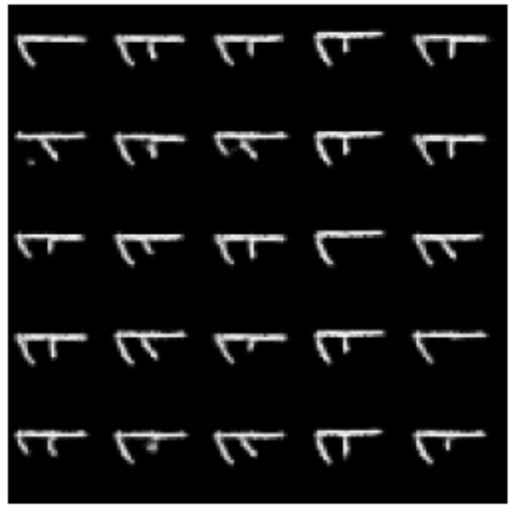
Simulated new characters



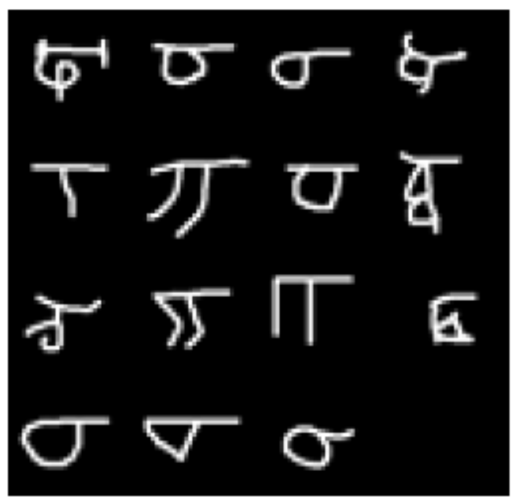
Learning from very few examples

3 examples of
a new class

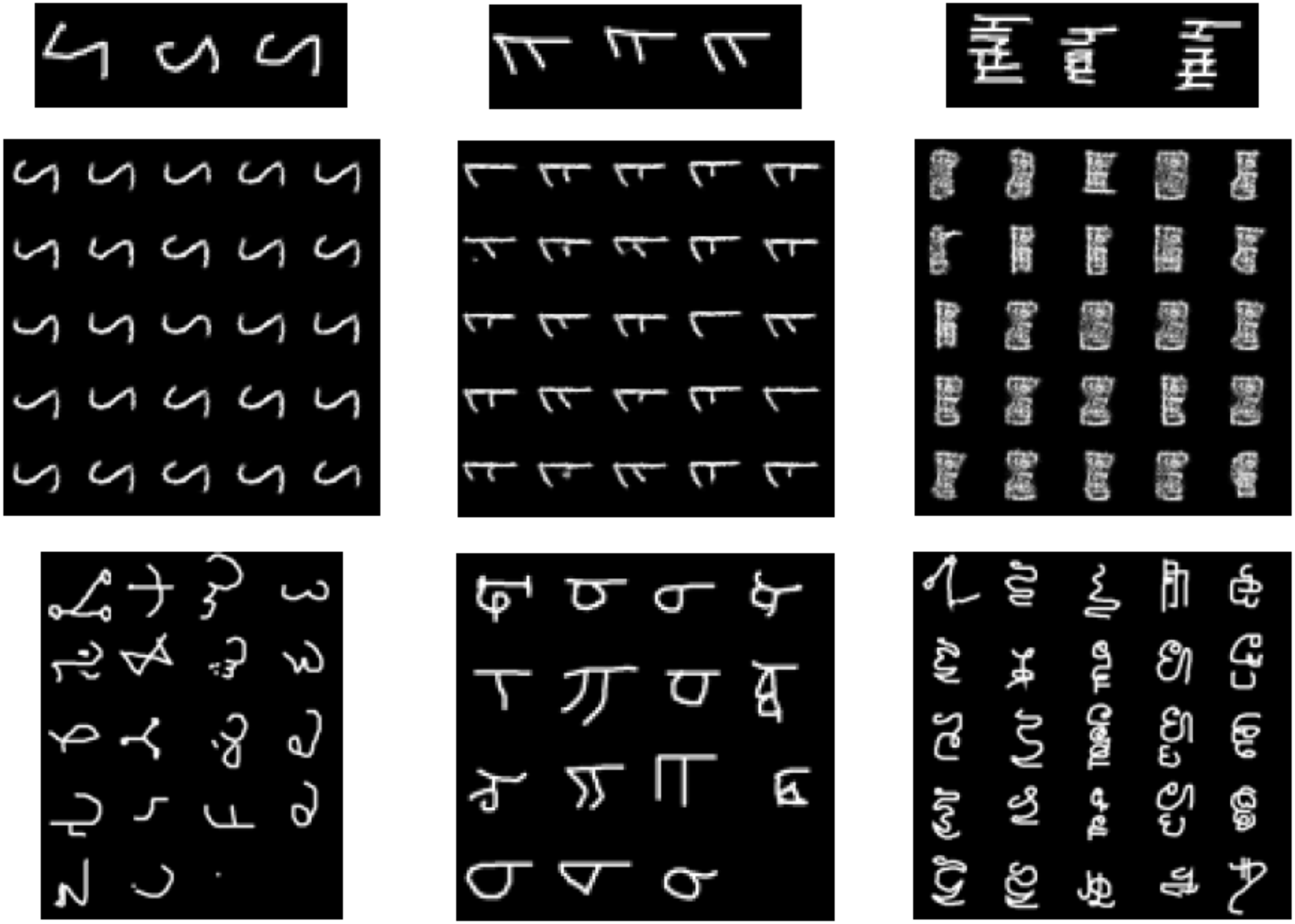
Conditional samples
in the same class



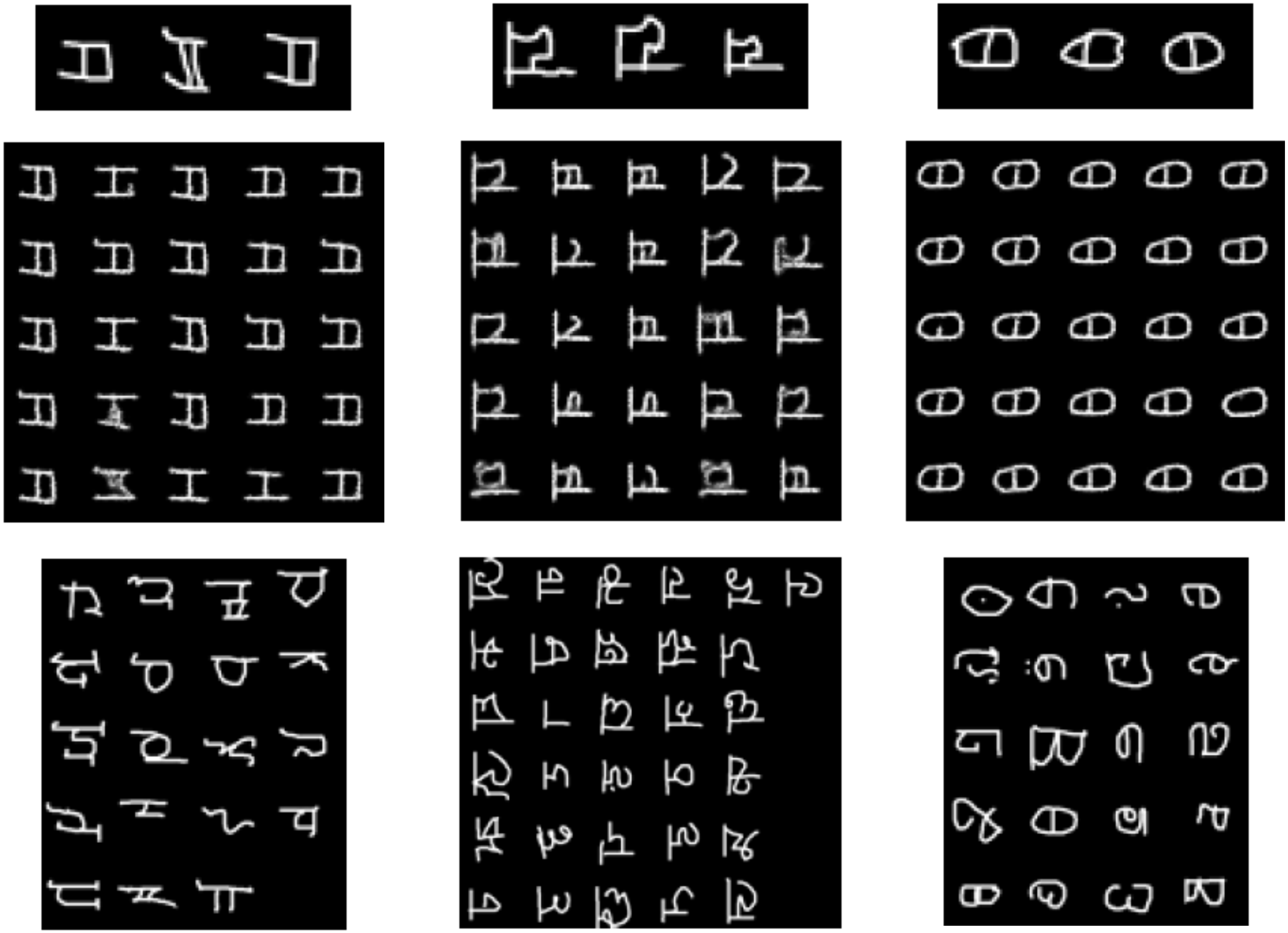
Inferred super-class



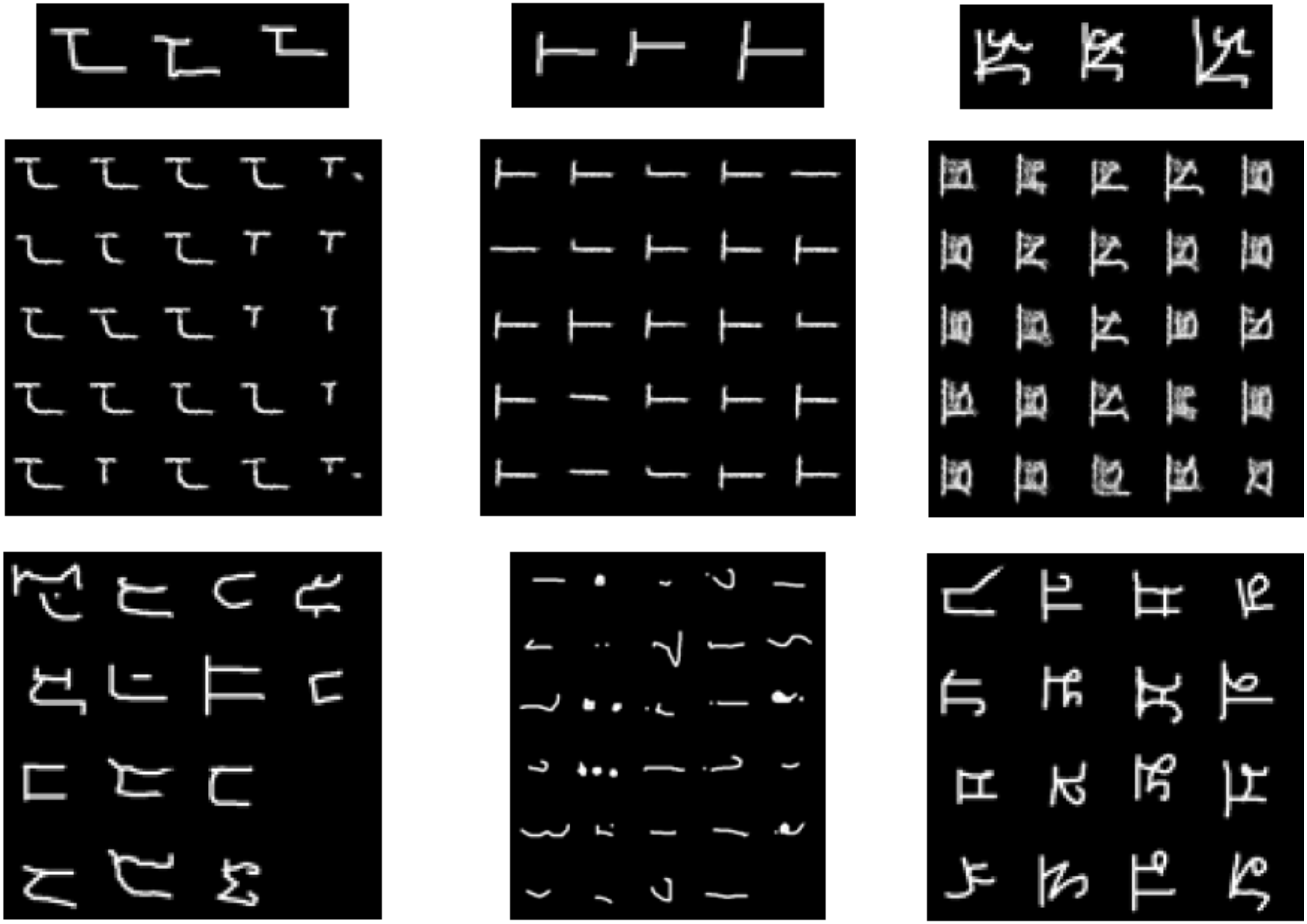
Learning from very few examples



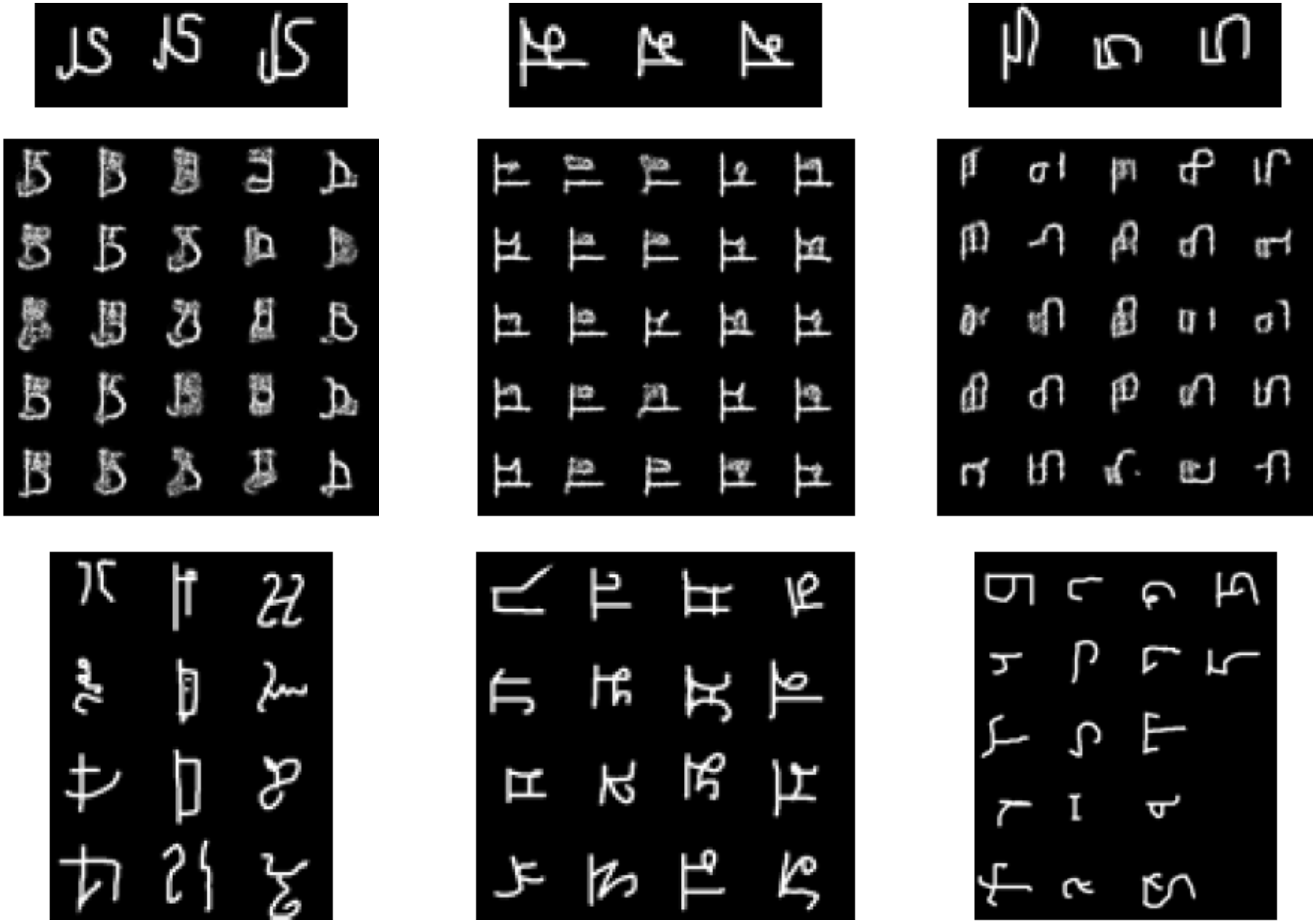
Learning from very few examples



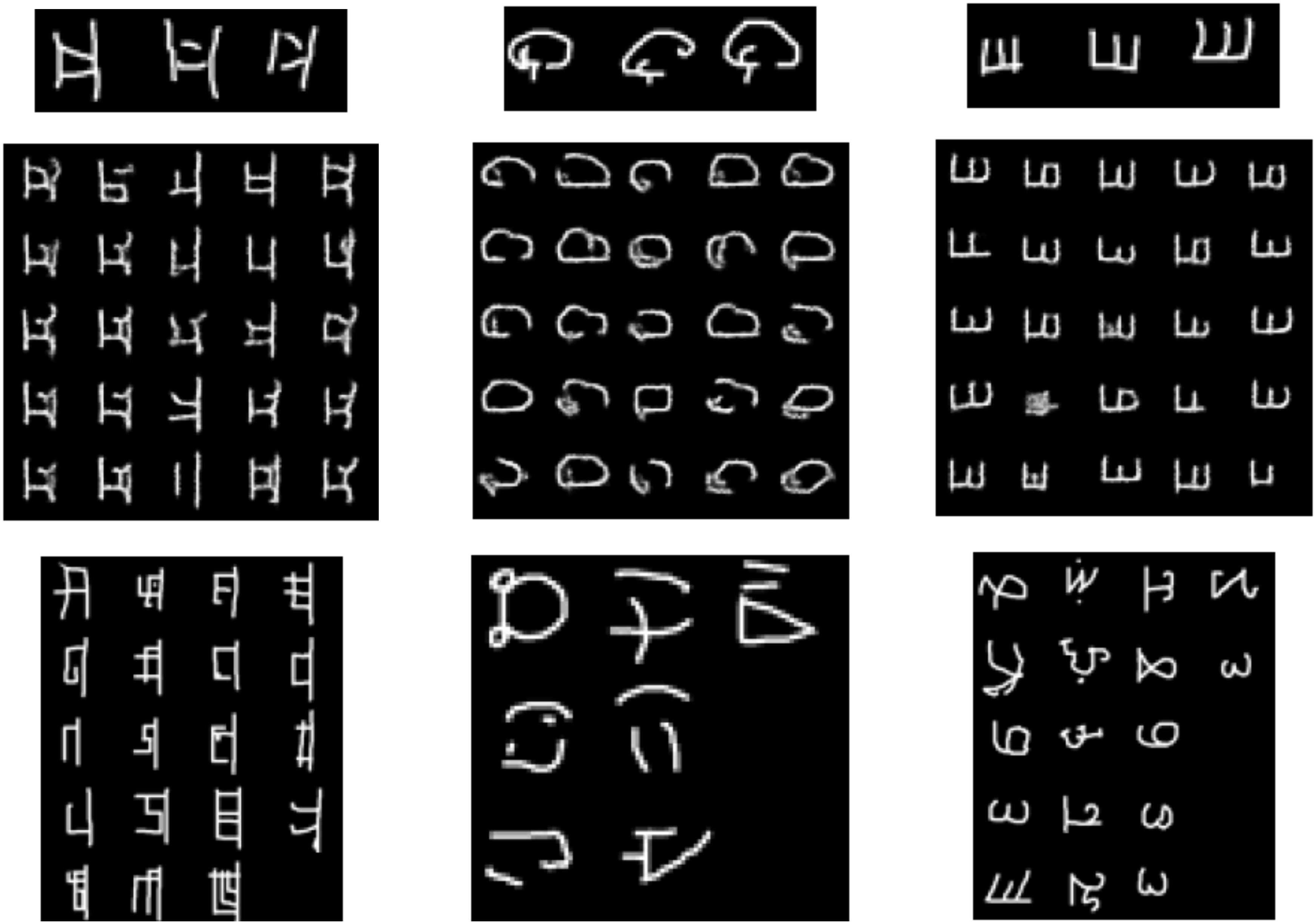
Learning from very few examples



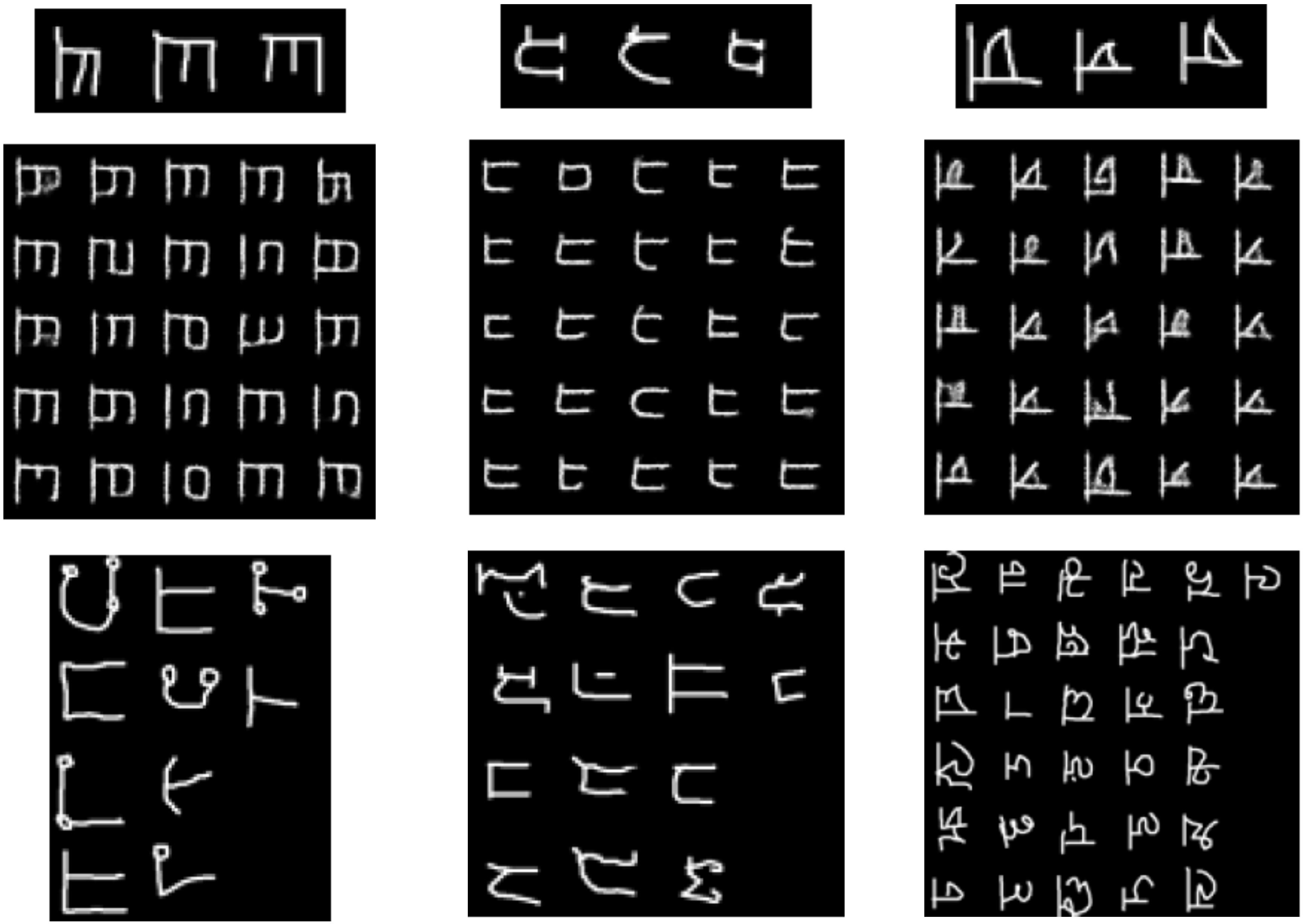
Learning from very few examples



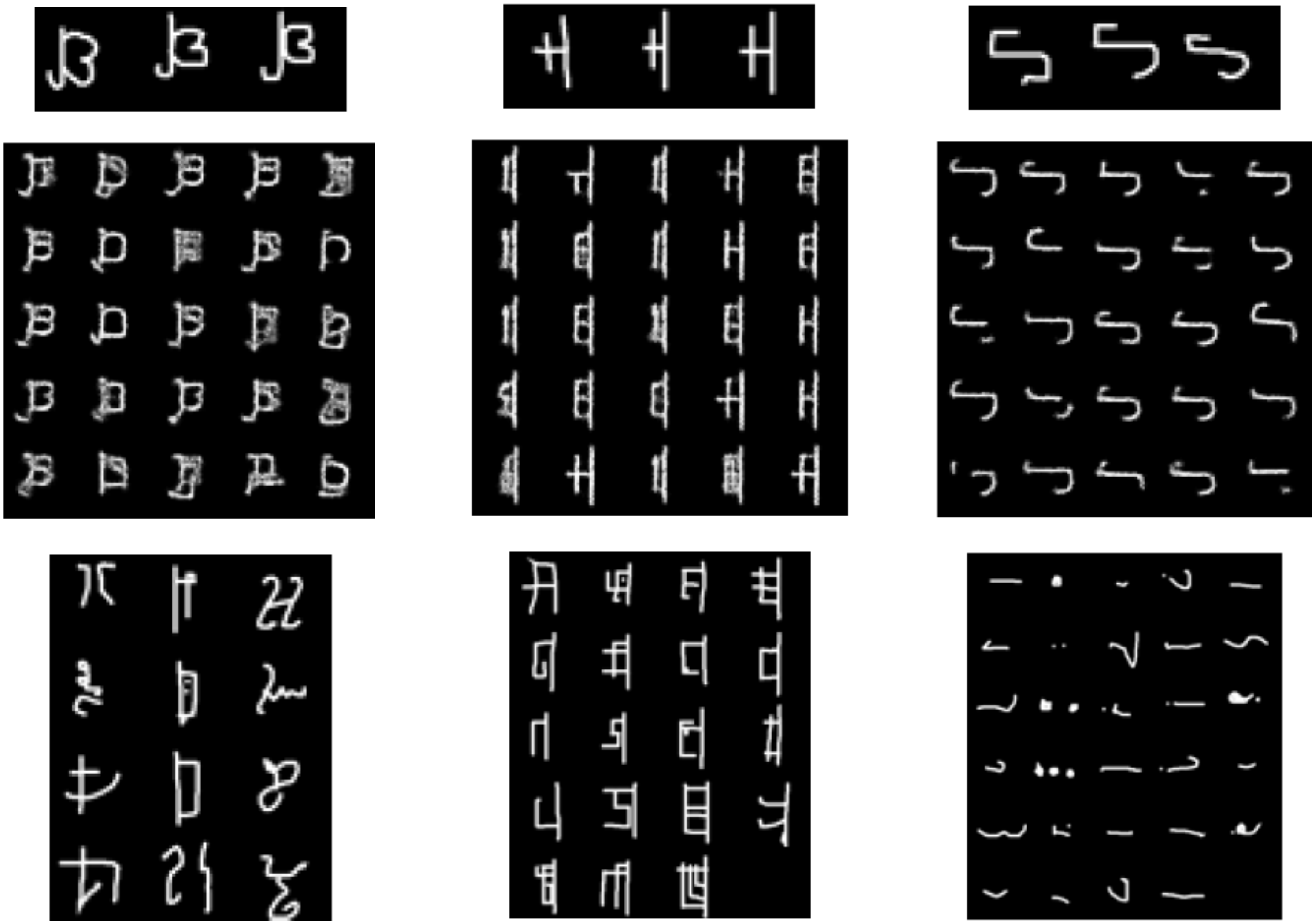
Learning from very few examples



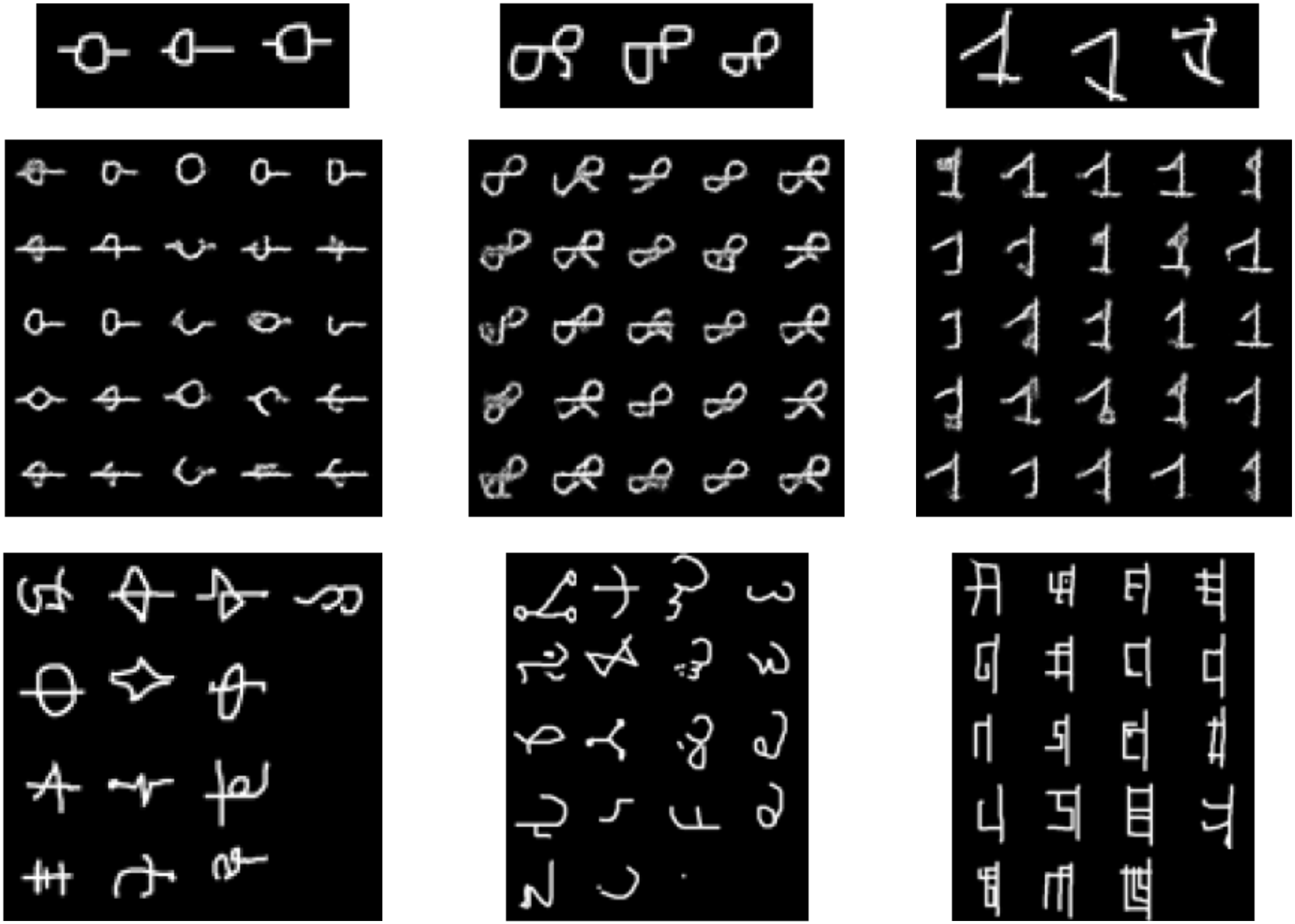
Learning from very few examples



Learning from very few examples

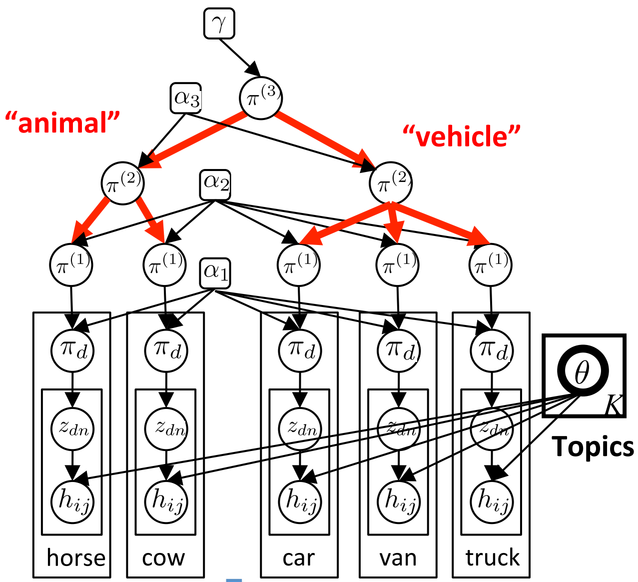


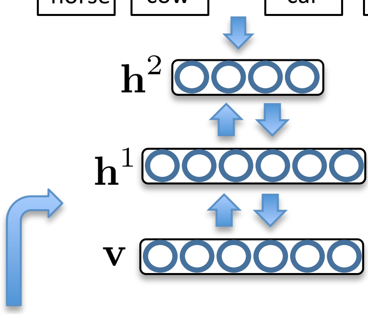
Learning from very few examples



Hierarchical-Deep

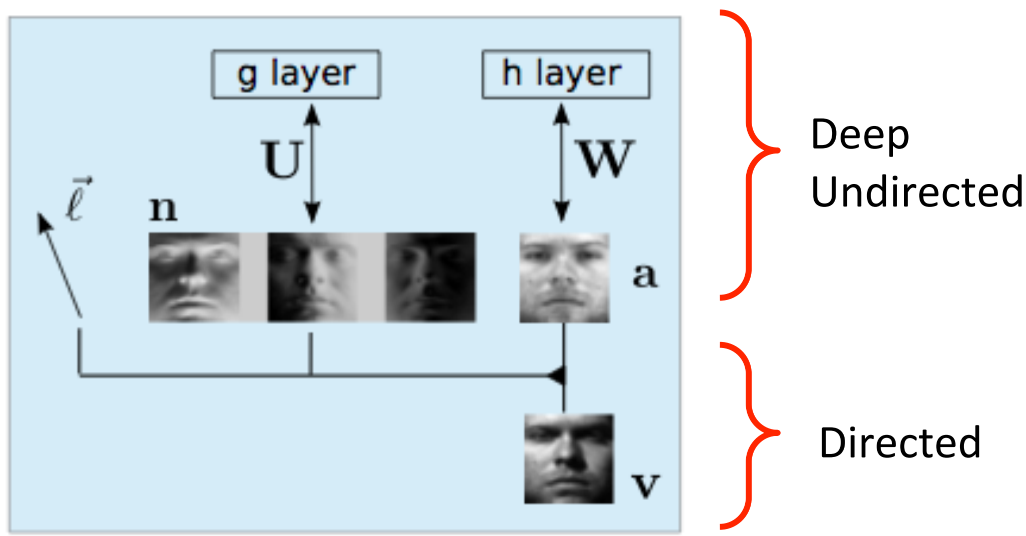
So far we have considered directed + undirected models.





Low-level features: replace GIST, SIFT

Deep Lambertian Networks



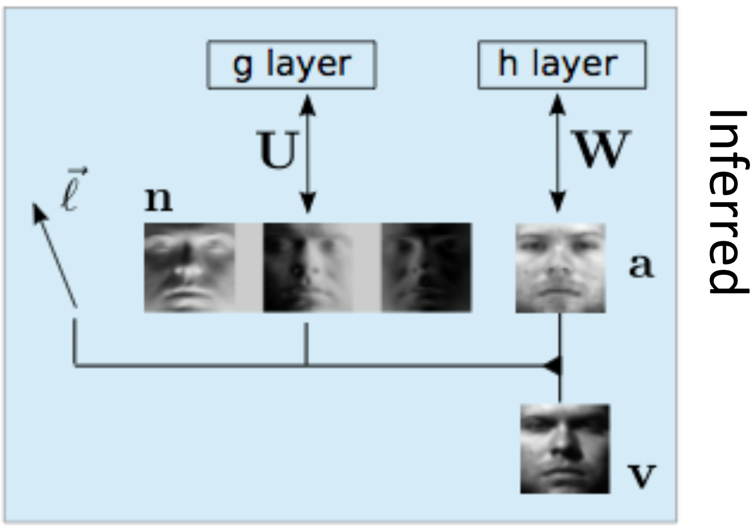
Combines the elegant properties of the Lambertian model with the Gaussian RBMs (and Deep Belief Nets, Deep Boltzmann Machines).

Tang et. al., ICML 2012

Deep Lambertian Networks

Model Specifics

Deep Lambertian Net



Observed

Image albedo — Surface Normals — Light source

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

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

$$\begin{aligned} P(\mathbf{a}) &\sim GRBM(\mathbf{a}), \\ P(\mathbf{N}) &\sim GRBM(\mathbf{N}), \\ P(l) &\sim \mathcal{N}(\mu, \mathbf{\Lambda}) \end{aligned}$$

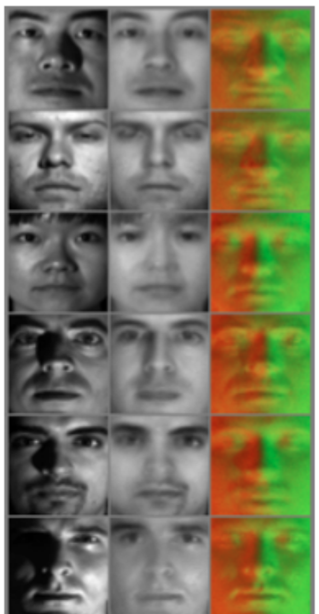
Inference: Gibbs sampler.

Learning: Stochastic Approximation

Deep Lambertian Networks

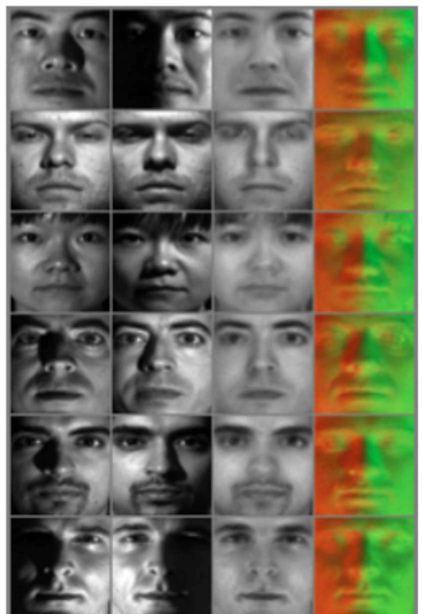
Yale B Extended Database

One Test Image



(a) One test image.

Two Test Images



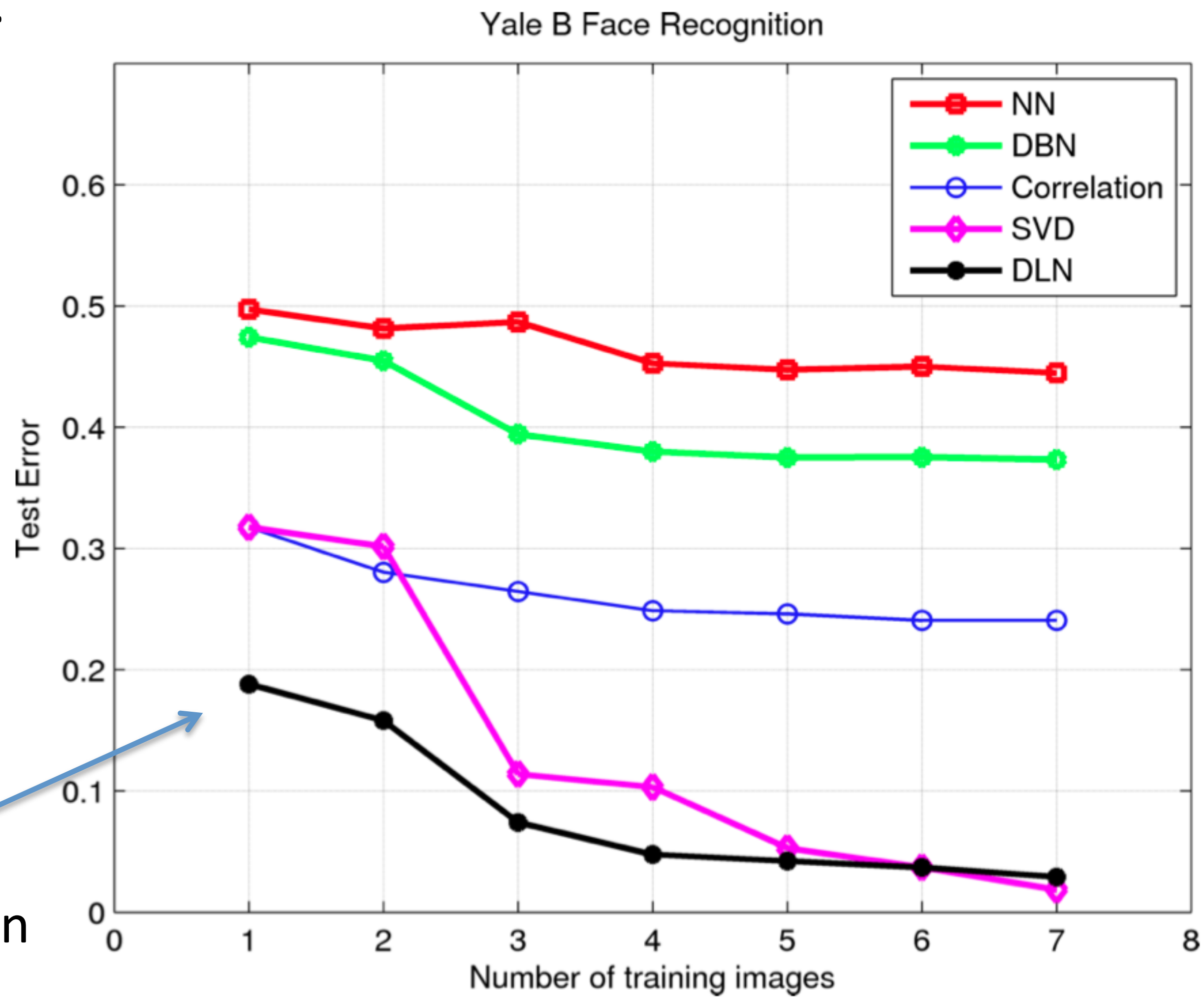
(b) Two test images.

Face Relighting

(c) Face Relighting.

Deep Lambertian Networks

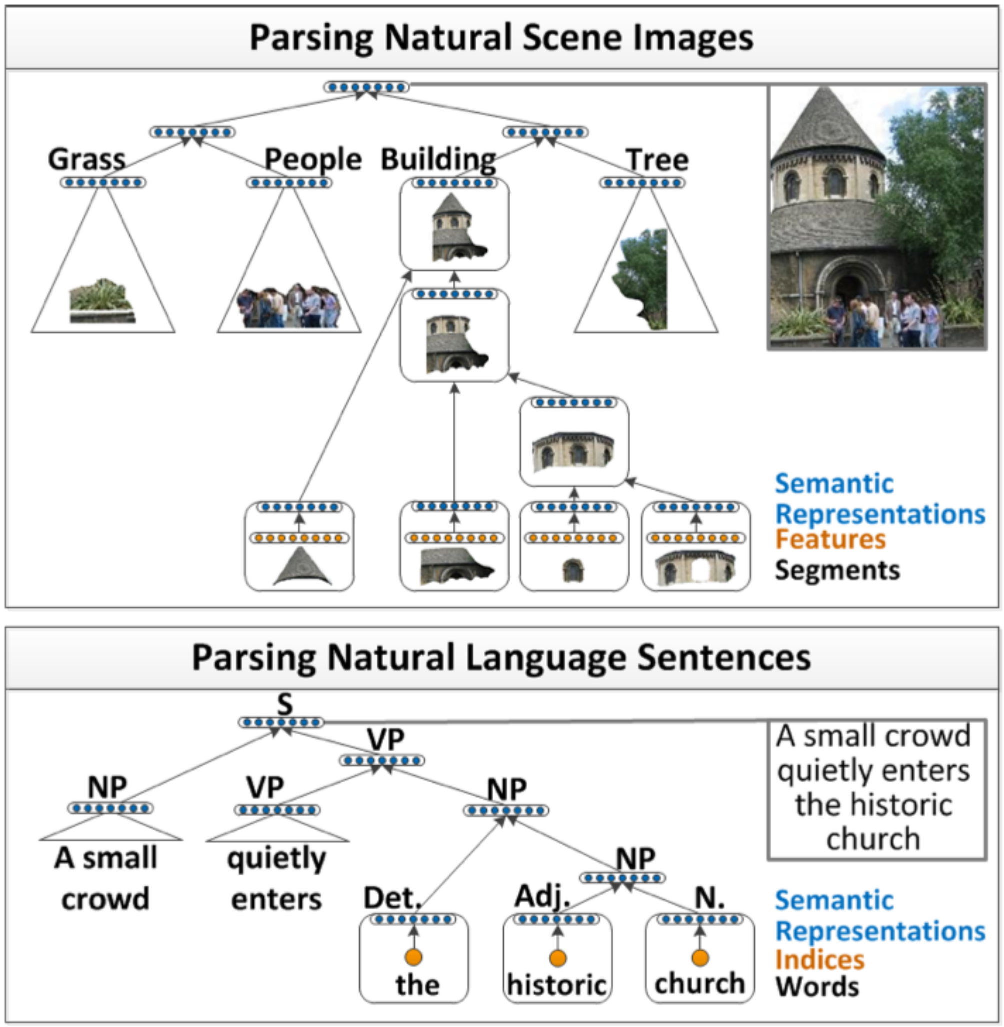
Recognition as function of the number of training images for 10 test subjects.

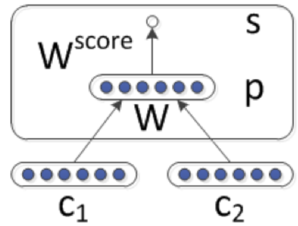


One-Shot Recognition

Recursive Neural Networks

Recursive structure learning





$$s = W^{score} p$$
$$p = f(W[c_1; c_2] + b)$$

Local recursive networks are making predictions whether to merge the two inputs as well as predicting the label.

Use Max-Margin Estimation.

Socher et. al., ICML 2011

Recursive Neural Networks

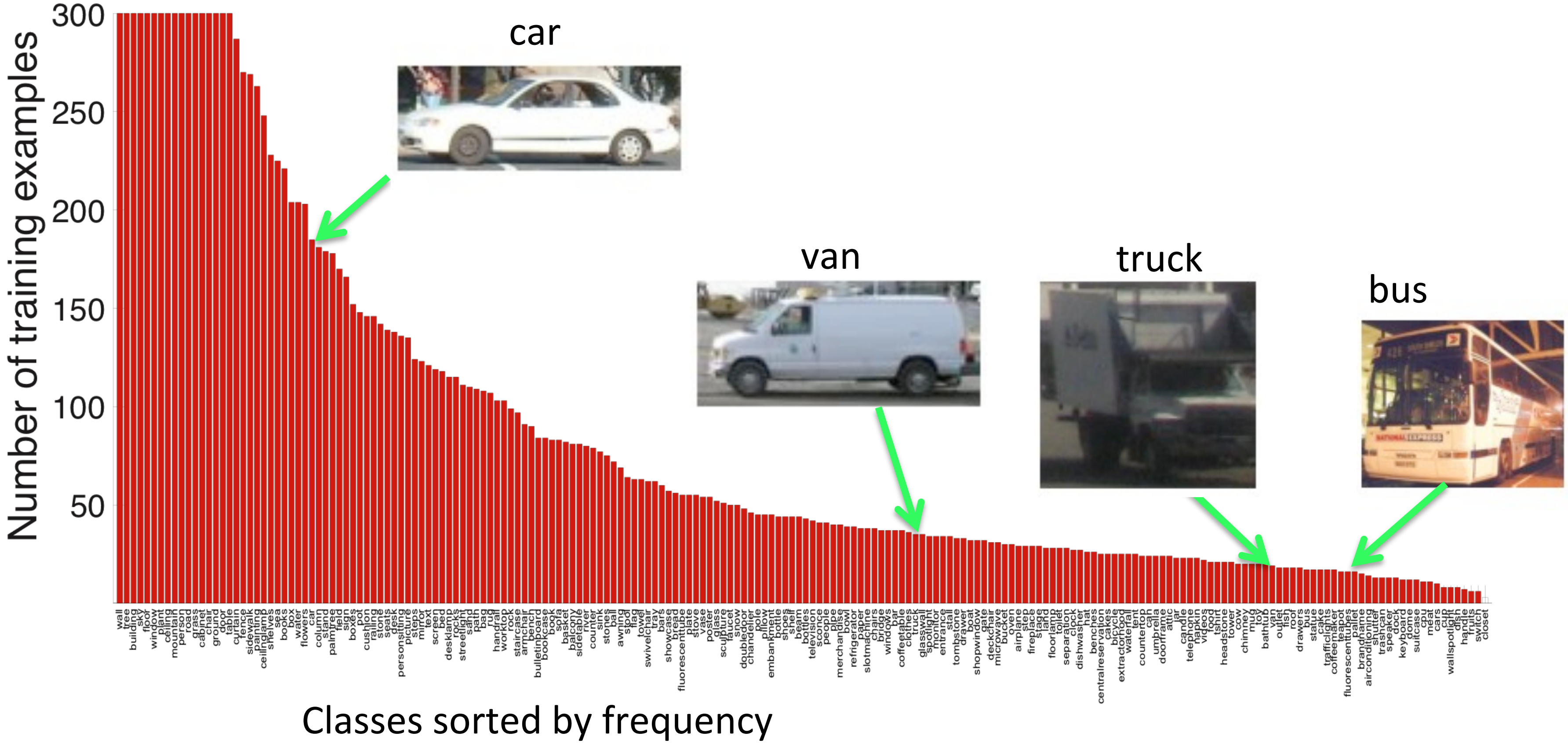
Recursive structure learning

Method and Semantic Pixel Accuracy in	%
Pixel CRF, Gould et al.(2009)	74.3
Log. Regr. on Superpixel Features	75.9
Region-based energy, Gould et al.(2009)	76.4
Local Labeling,TL(2010)	76.9
Superpixel MRF,TL(2010)	77.5
Simultaneous MRF,TL(2010)	77.5
RNN (our method)	**78.1**

Socher et. al., ICML 2011

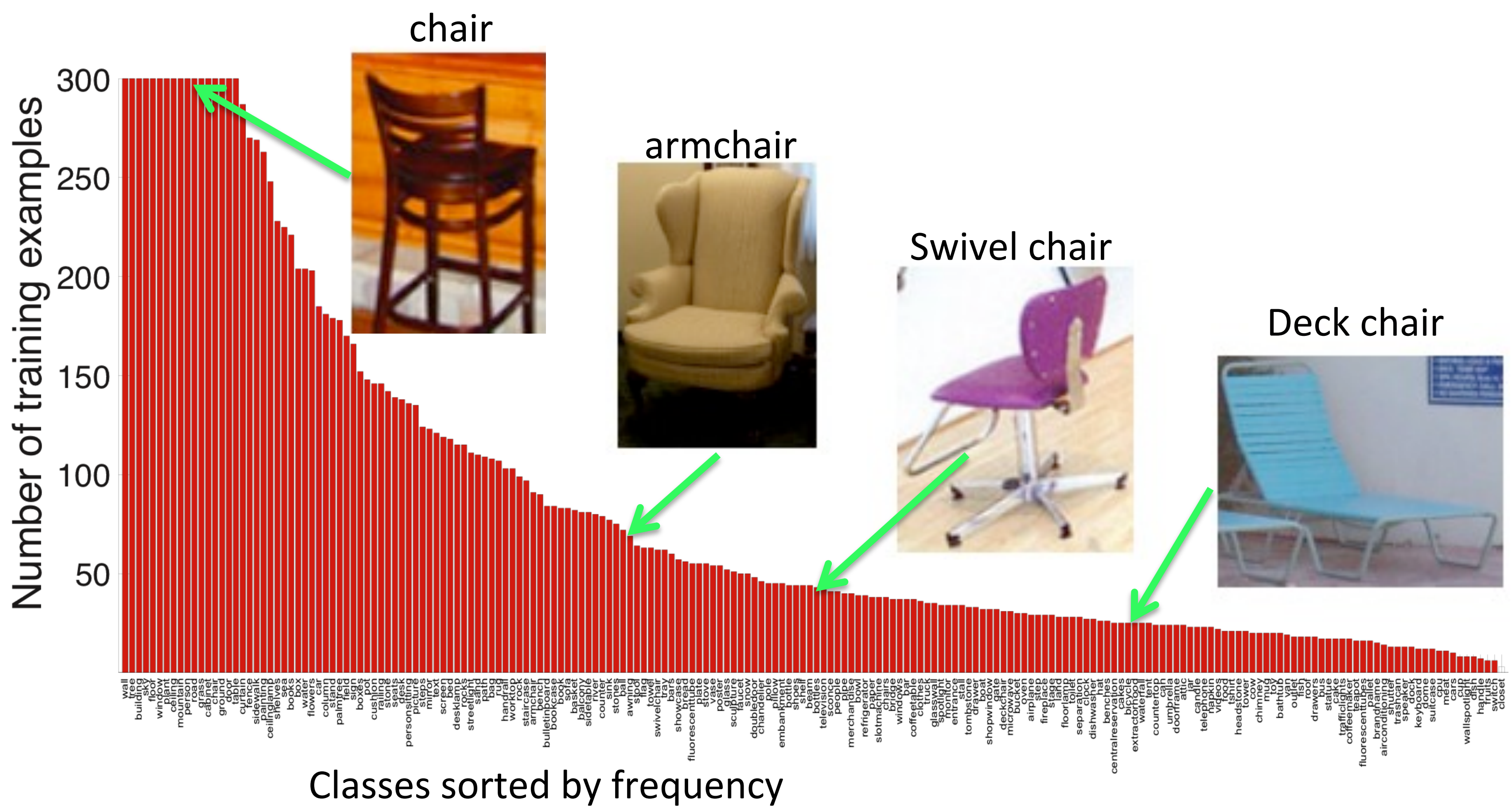
Learning from Few Examples

SUN database



Rare objects are similar to frequent objects

Learning from Few Examples



Generative Model of Classifier Parameters

Many state-of-the-art object detection systems use sophisticated models, based on multiple parts with separate appearance and shape components.

$$y = \beta^\top \Phi(\mathbf{x})$$

Image Specific: concatenation of the HOG feature pyramid at multiple scales.

Felzenszwalb, McAllester & Ramanan, 2008

Detect objects by testing sub-windows and scoring corresponding test patches with a linear function.

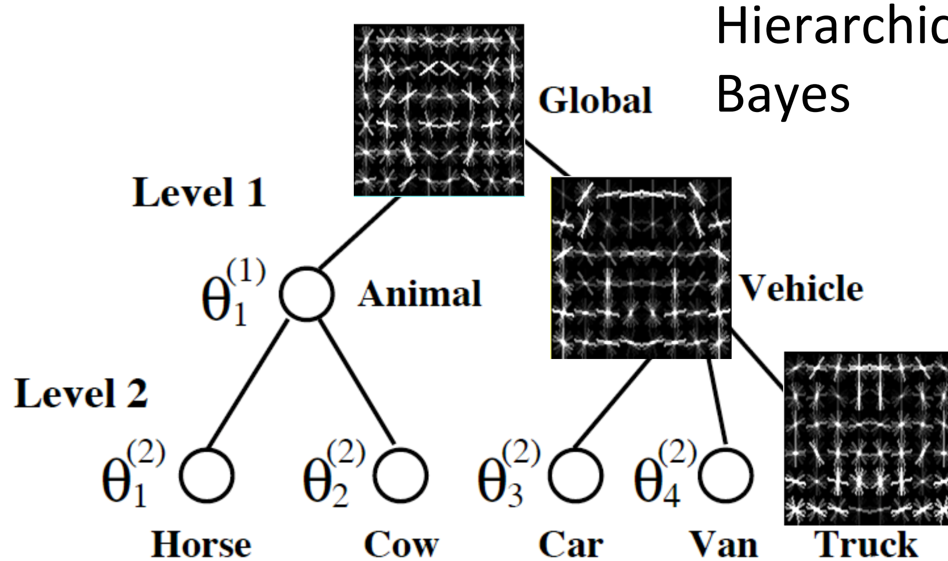
Define hierarchical prior over parameters of discriminative model and learn the hierarchy.

Generative Model of Classifier Parameters

By learning hierarchical structure, we can improve the current state-of-the-art.

Sun Dataset: 32,855 examples of 200 categories

Hierarchical Bayes

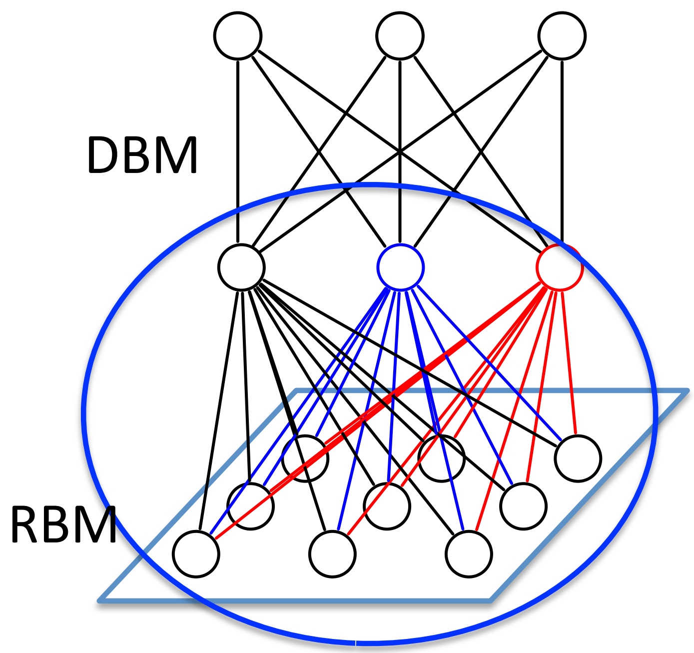


185 ex 27 ex 12 ex

Hierarchical Model

Single Class

Talk Roadmap



- Unsupervised Feature Learning
  - Restricted Boltzmann Machines
  - Deep Belief Networks
  - Deep Boltzmann Machines
- Transfer Learning with Deep Models
- Multimodal Learning

Multi-Modal Input

Learning systems that combine multiple input domains

Images

Text & Language

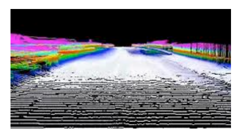
WIKIPEDIA
The Free Encyclopedia

REUTERS

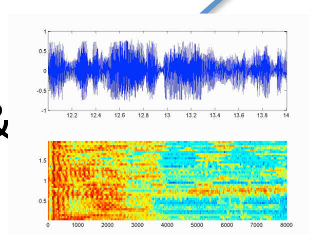
Associated Press

Video

Laser scans



Speech & Audio



Time series data

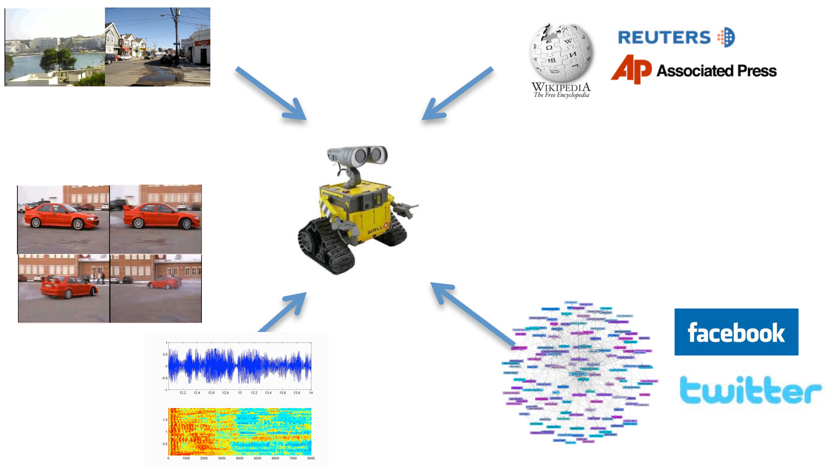
Develop learning systems that come closer to displaying human like intelligence

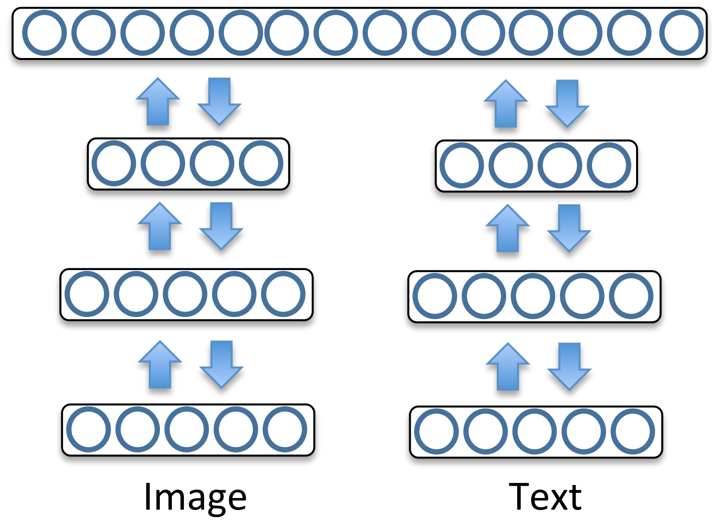
One of Key Challenges:

Inference

Multi-Modal Input

Learning systems that combine multiple input domains





More robust perception.

Ngiam et.al., ICML 2011 used deep autoencoders (video + speech)

- Guillaumin, Verbeek, and Schmid, CVPR 2011
- Huiskes, Thomee, and Lew, Multimedia Information Retrieval, 2010
- Xing, Yan, and Hauptmann, UAI 2005.

Training Data

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

camera, jahdakine, lightpainting, reflection doublepaneglass wowiekazowie

sandbanks, lake, lakeontario, sunset, walking, beach, purple, sky, water, clouds, overtheexcellence

top20butterflies

<no text>

mickikrimmel, mickipedia, headshot

Samples from the MIR Flickr Dataset - Creative Commons License

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

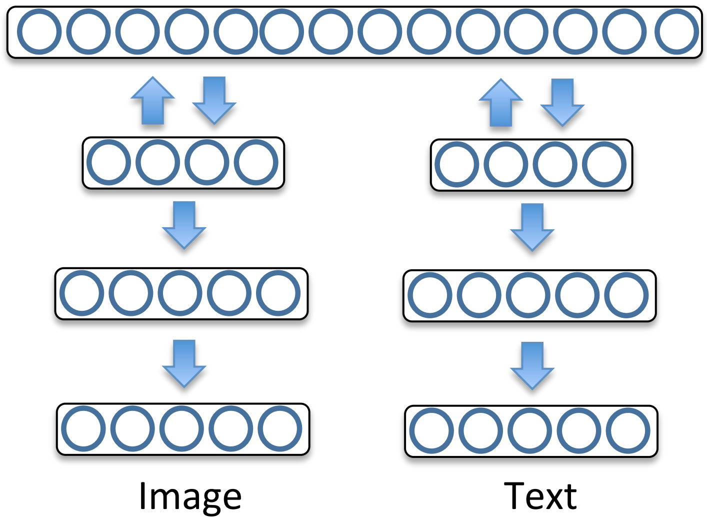
beach, sea, surf,
strand, shore,
wave, seascape,
sand, ocean, waves

→

Multi-Modal Deep Belief Net

Gaussian RBM

Dense



Replicated Softmax

Sparse counts

Multi-Modal Deep Belief Net

- Flickr Data - 1 Million images along with text tags, 25K annotated

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

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

Recognition Results

- Multimodal Inputs (images + text), 38 classes.

Learning Algorithm	Mean Average Precision
Image-text SVM	0.475
Image-text LDA	0.492
Multimodal DBN	0.566

- Unimodal Inputs (images only).

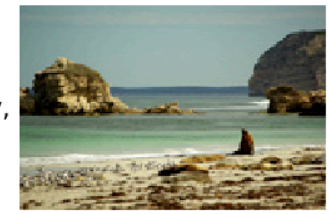
Learning Algorithm	Mean Average Precision
Image-SVM	0.375
Image-LDA	0.315
Image DBN	0.413

Pattern Completion

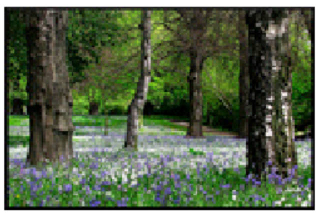
Given a test image, we generate associated text – achieve far better classification results.

landscape, scenery, hills,landscapes, scenic, land, canyon, roadtrip, place, tourism

portrait, black, white, girl, expression, lady, look, blonde, eyes, gorgeous



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

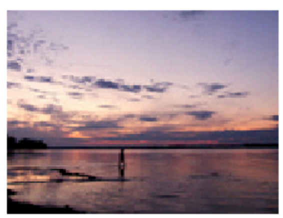


woods, breathtaking, hills, scenery, alone, mist, fields, bush, branches

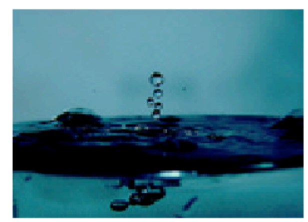
sky, clouds landscape, hills, scenery, horizon, fields, landscapes, scenic, sun

night, city urban, cityscape traffic, notte, skyline, lights, streets, skyscraper

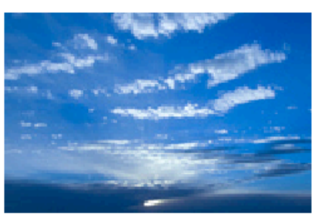
car, engine, auto, supercar, ferrari, fast, gt, jason, parking, automobile



sunset, twilight, strand, wave, breathtaking, horizon, shore, seascape, surf, scenery

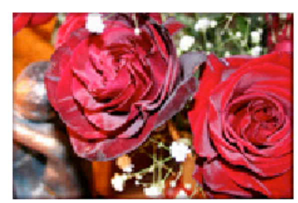


sky, blue, clouds, horizon, céu, twilight, azul, bleu, wave, sunset



sky, clouds, blue, horizon, céu, sunset, hills, twilight, bluesky, breathtaking

structure, facade, place, landmark, industry, skyscraper, tripod, royal, parking, 1910s



red, rouge, rosso, rot, catchycolors, gift, shiny, rojo, vivid, soft

Thank you

Code for learning RBMs, DBNs, and DBMs is available at:
http://www.mit.edu/~rsalakhu/