Impact of Deep Learning

- Speech Recognition
- Computer Vision
- Recommender Systems
- Language Understanding
- Drug Discovery and Medical Image Analysis

[Courtesy of R. Salakhutdinov]

[Hinton & Salakhutdinov, 2006]

Deep Belief Networks: Training

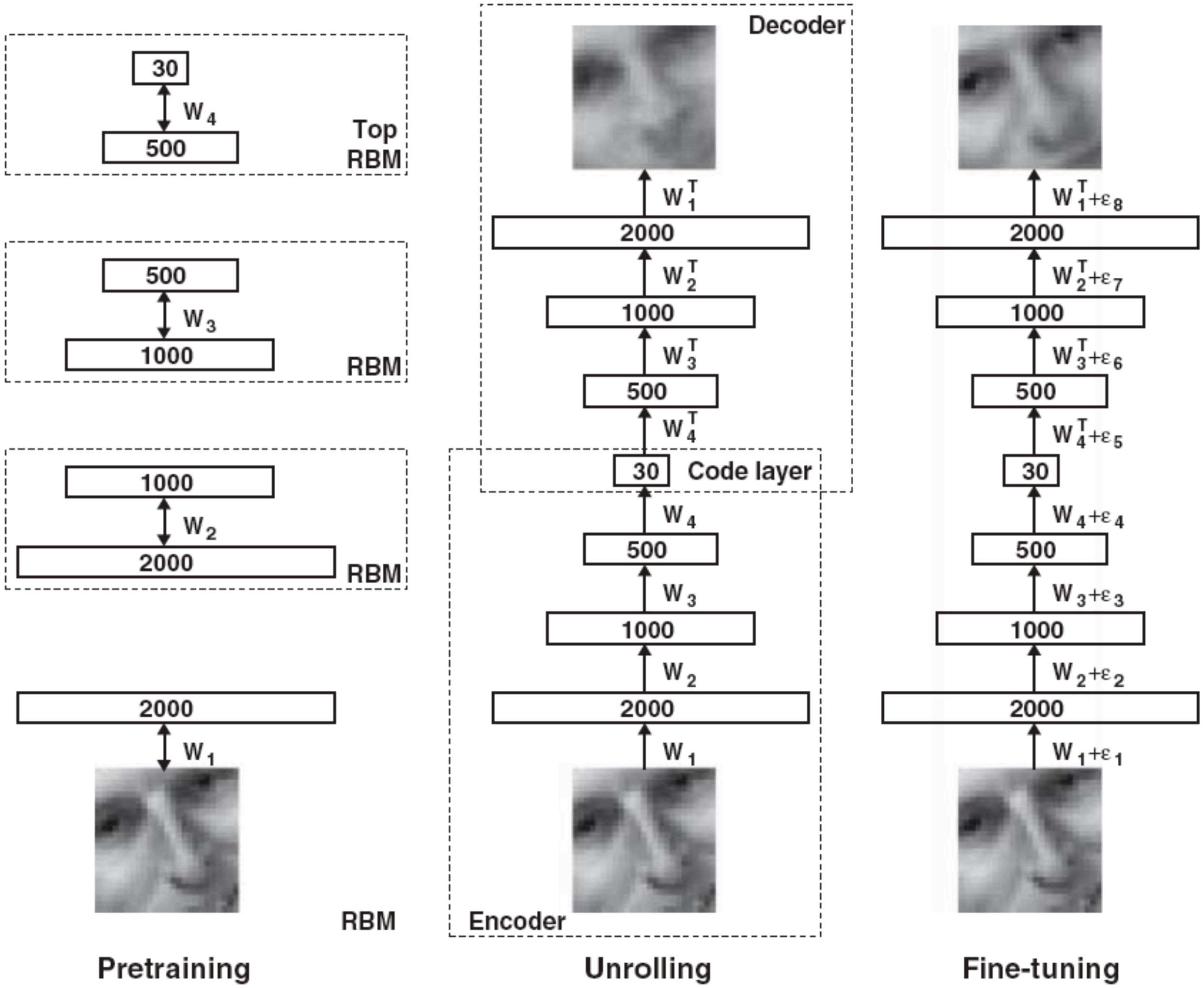


Fig. 1. Pretraining consists of learning a stack of restricted Boltzmann machines (RBMs), each having only one layer of feature detectors. The learned feature activations of one RBM are used as the "data" for training the next RBM in the stack. After the pretraining, the RBMs are "unrolled" to create a deep autoencoder, which is then fine-tuned using backpropagation of error derivatives.

Very Large Scale Use of DBN's

[Quoc Le, et al., *ICML*, 2012]

Data: 10 million 200x200 unlabeled images, sampled from YouTube

Training: use 1000 machines (16000 cores) for 1 week

Learned network: 3 multi-stage layers, 1.15 billion parameters

Achieves 15.8% (was 9.5%) accuracy classifying 1 of 20k ImageNet items

Real images that most excite the feature:

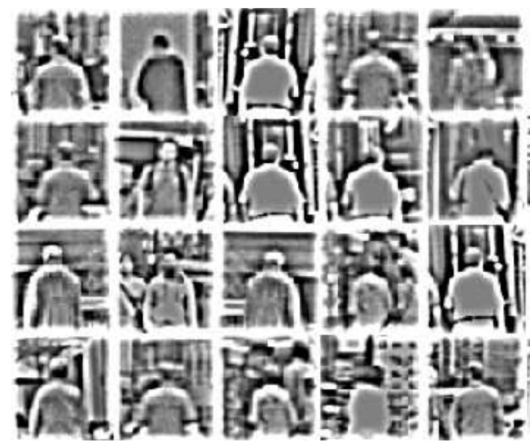
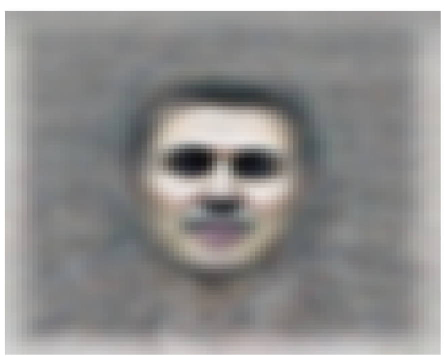
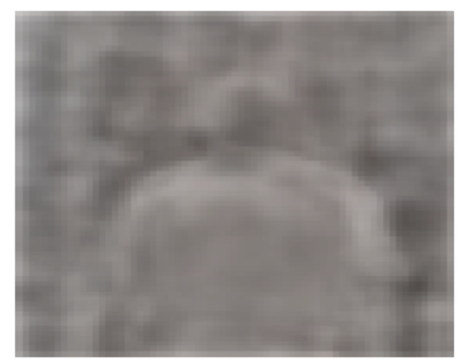


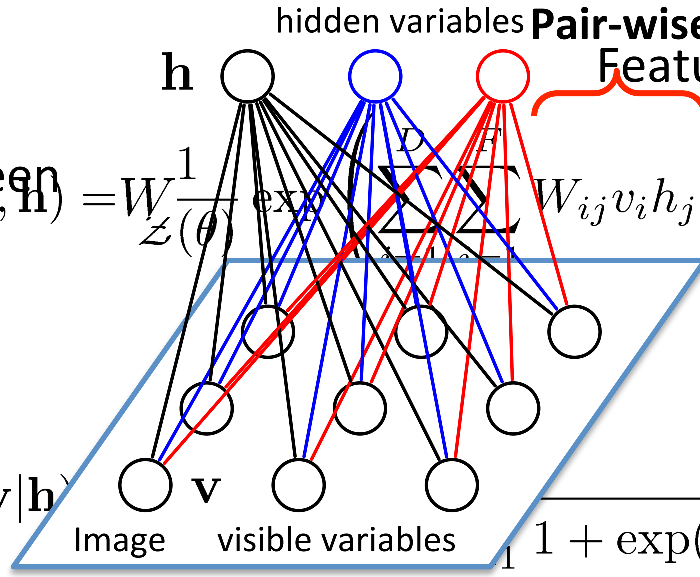
Image synthesized to most excite the feature:





Restricted Boltzmann Machines

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



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

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

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

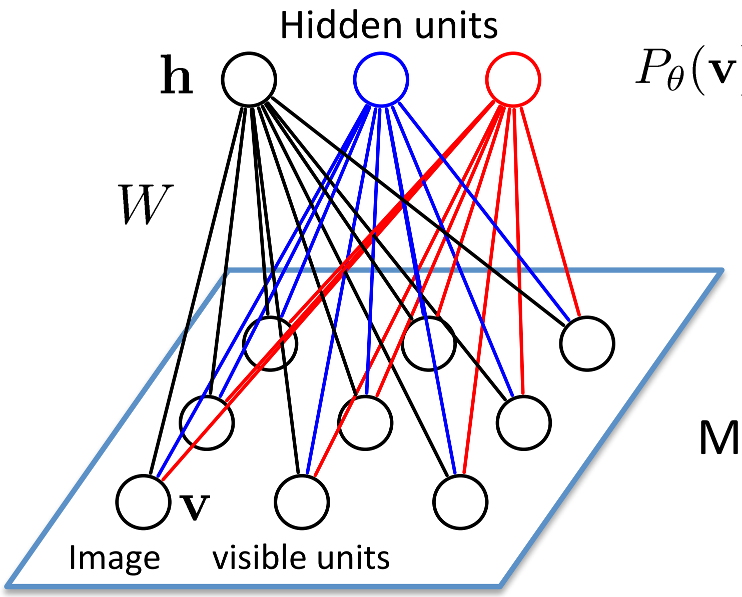
RBM is a Markov Random Field with:

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

[Courtesy, R. Salakhutdinov]

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

Model Learning



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

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

Maximize log-likelihood objective:

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

Derivative of the log-likelihood:

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

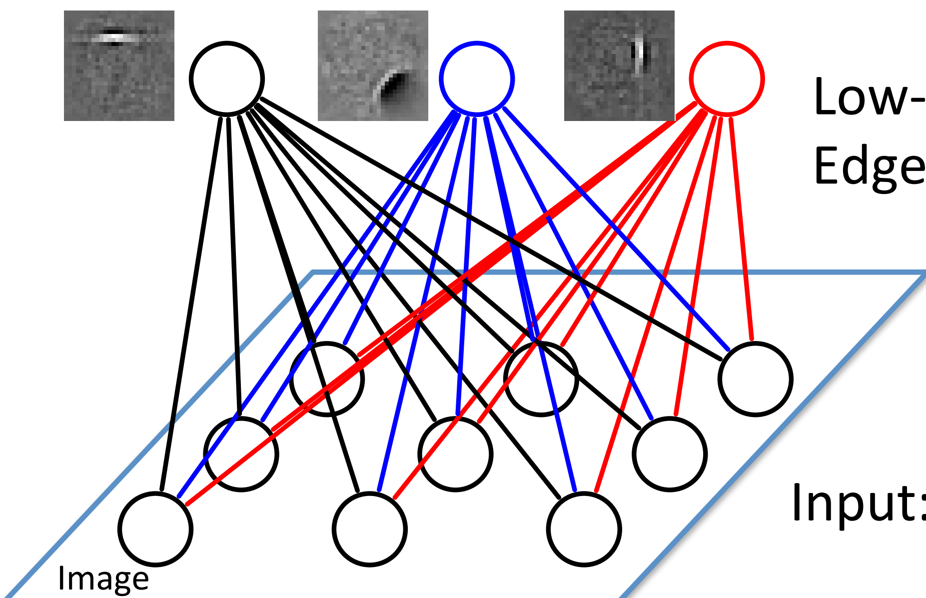
$$= \mathrm{E}_{P_{data}}[v_i h_j] - \mathrm{E}_{P_\theta}[v_i h_j]$$

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

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

[Courtesy, R. Salakhutdinov]

Deep Boltzmann Machines



Low-level features:
Edges

Built from **unlabeled** inputs.

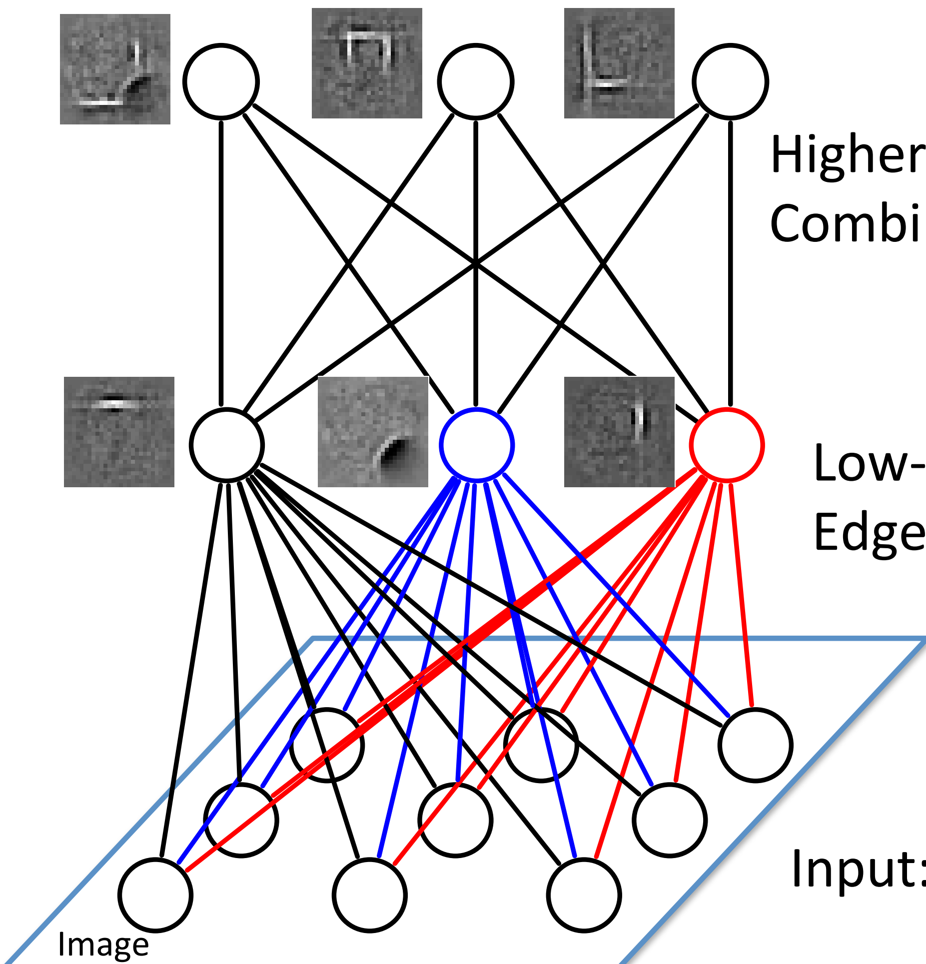
Input: Pixels

[Courtesy, R. Salakhutdinov]

(Salakhutdinov & Hinton, Neural Computation 2012)

Deep Boltzmann Machines

Learn simpler representations, then compose more complex ones



Higher-level features: Combination of edges

Low-level features: Edges

Built from **unlabeled** inputs.

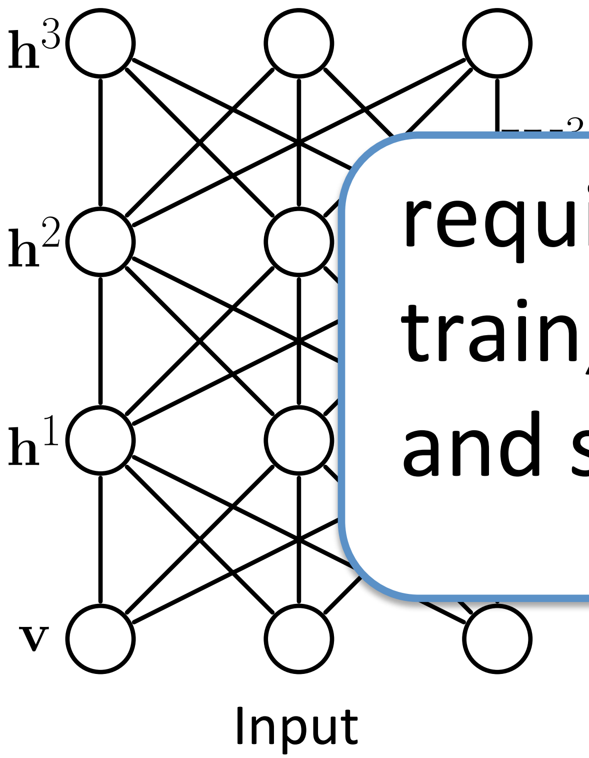
Input: Pixels

[Courtesy, R. Salakhutdinov]

(Salakhutdinov 2008, Salakhutdinov & Hinton 2012)

Model Formulation

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

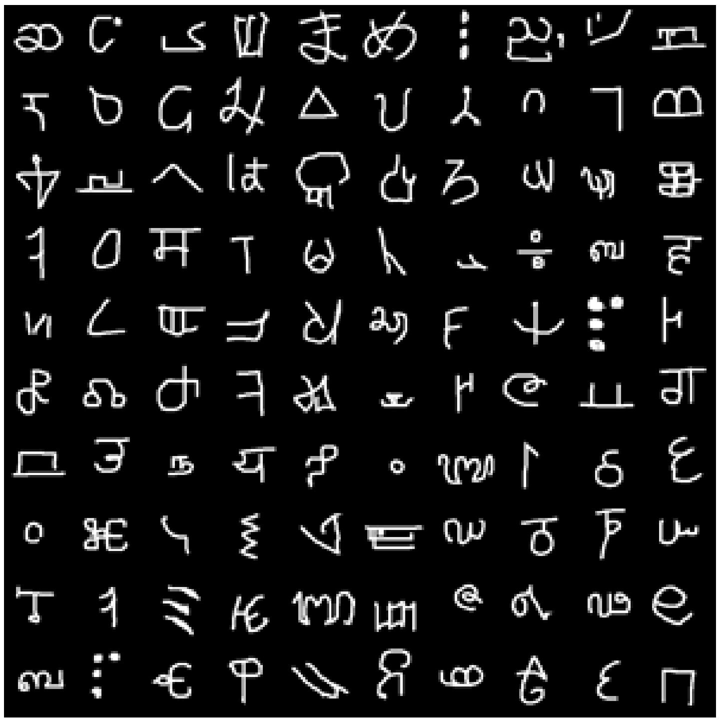


requires approximate inference to train, but it can be done…
and scales to millions of examples

[Courtesy, R. Salakhutdinov]

Samples Generated by the Model

Training Data



Model-Generated Samples

[Courtesy, R. Salakhutdinov]

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

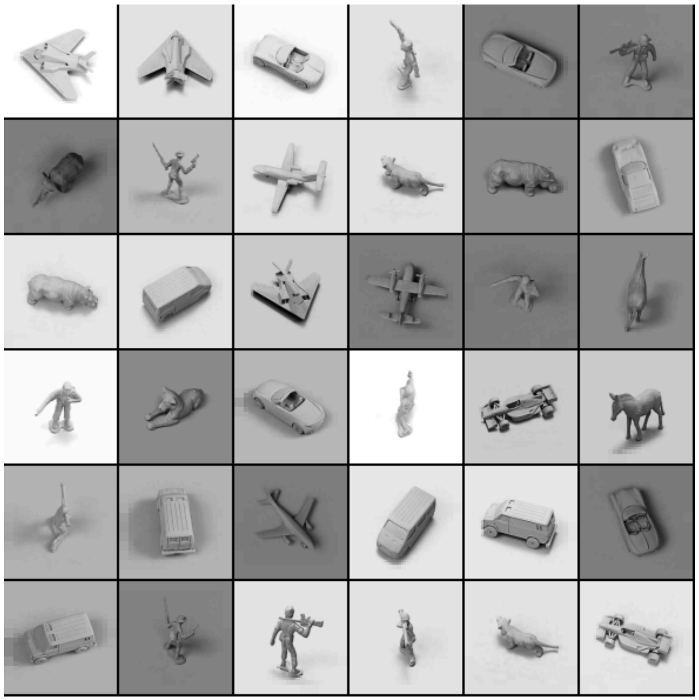
Learning Algorithm	Error
Logistic regression	22.14%
K-NN	18.92%
Neural Net	14.62%
SVM (Larochelle et.al. 2009)	9.70%
Deep Autoencoder (Bengio et. al. 2007)	10.05%
Deep Belief Net (Larochelle et. al. 2009)	9.68%
DBM	**8.40%**

Permutation-invariant version.

[Courtesy, R. Salakhutdinov]

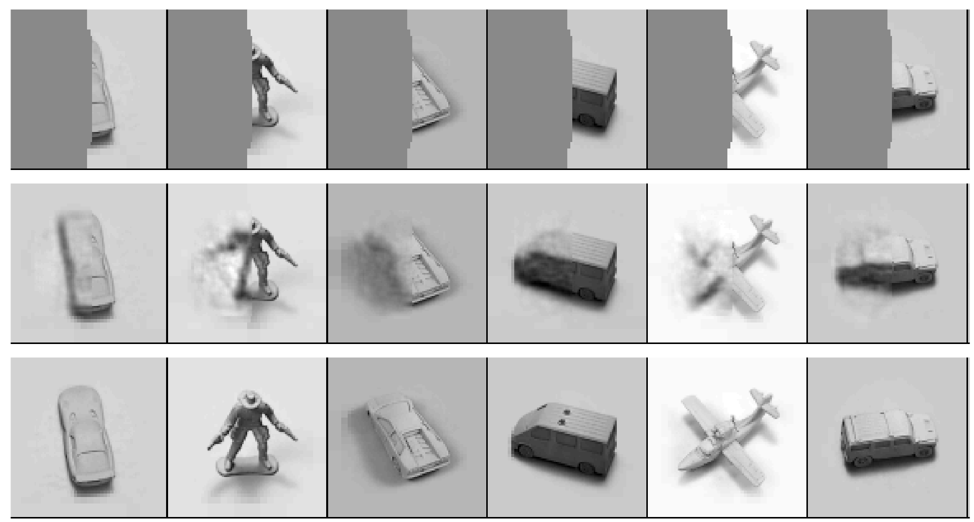
3-D object Recognition

NORB Dataset: 24,000 examples



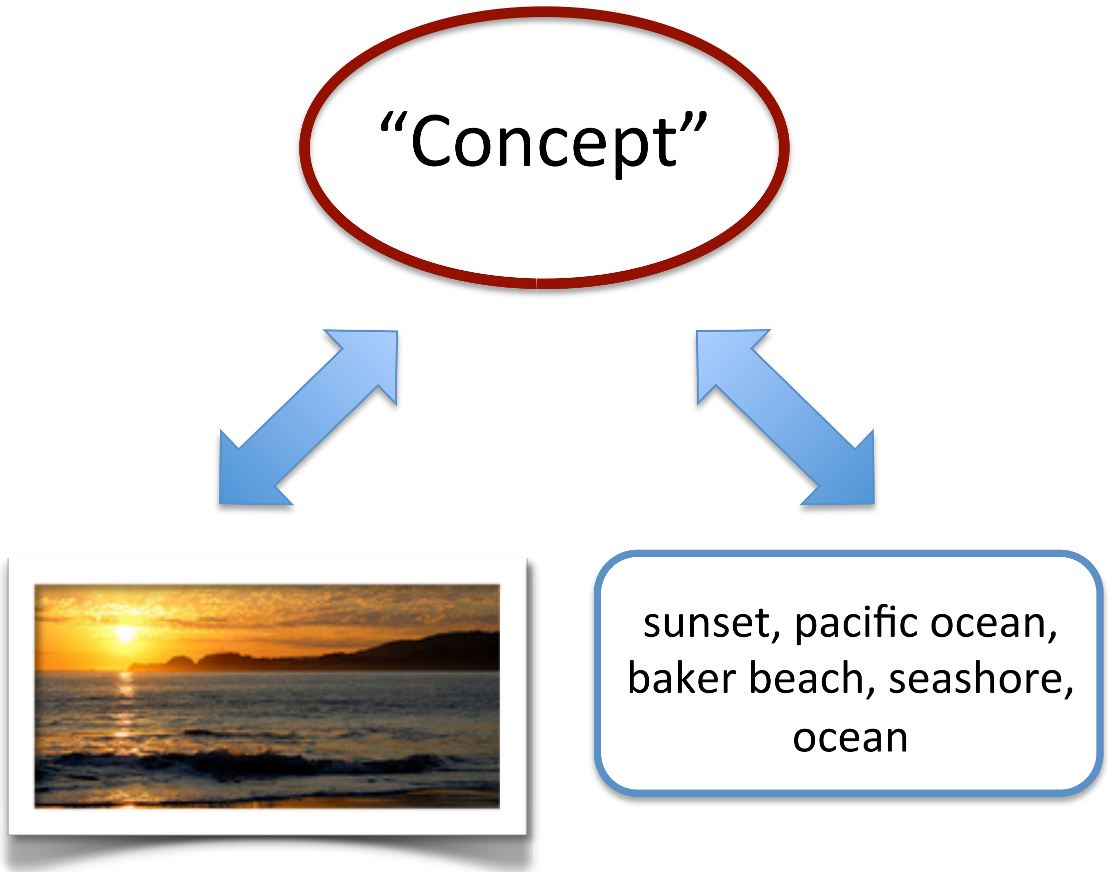
Learning Algorithm	Error
Logistic regression	22.5%
K-NN (LeCun 2004)	18.92%
SVM (Bengio & LeCun 2007)	11.6%
Deep Belief Net (Nair & Hinton 2009)	9.0%
DBM	**7.2%**

Pattern Completion



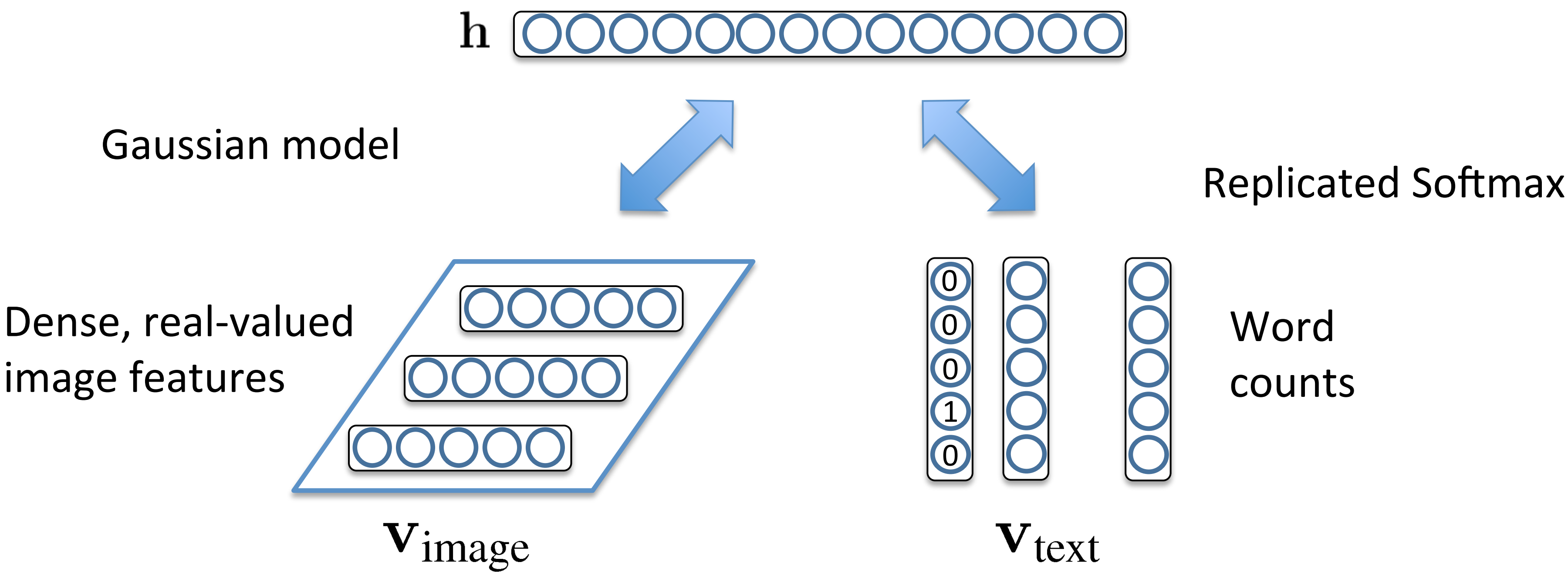
[Courtesy, R. Salakhutdinov]

Learning Shared Representations Across Sensory Modalities



[Courtesy, R. Salakhutdinov]

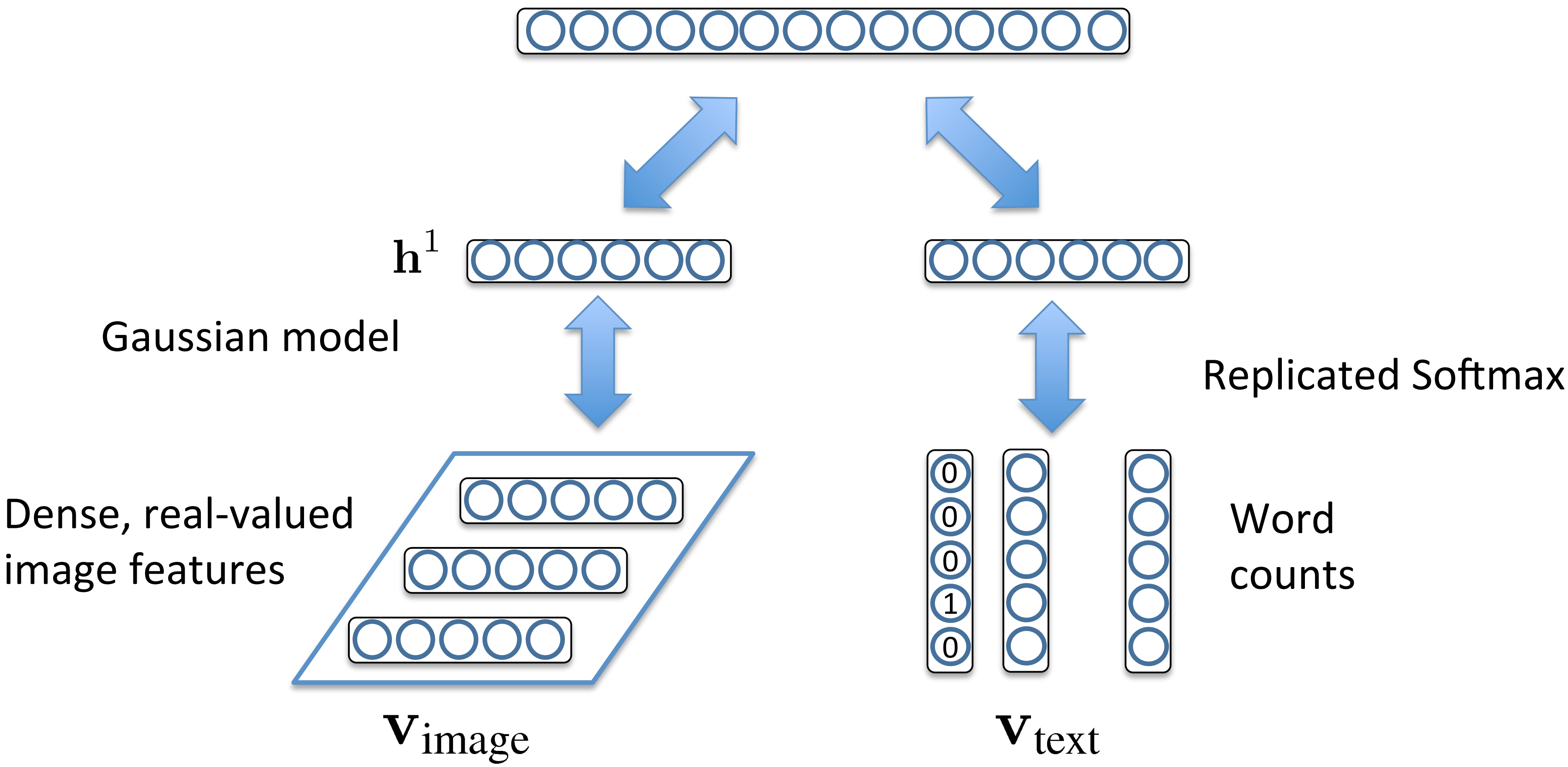
Multimodal DBM



[Courtesy, R. Salakhutdinov]

(Srivastava & Salakhutdinov, NIPS 2012, JMLR 2014)

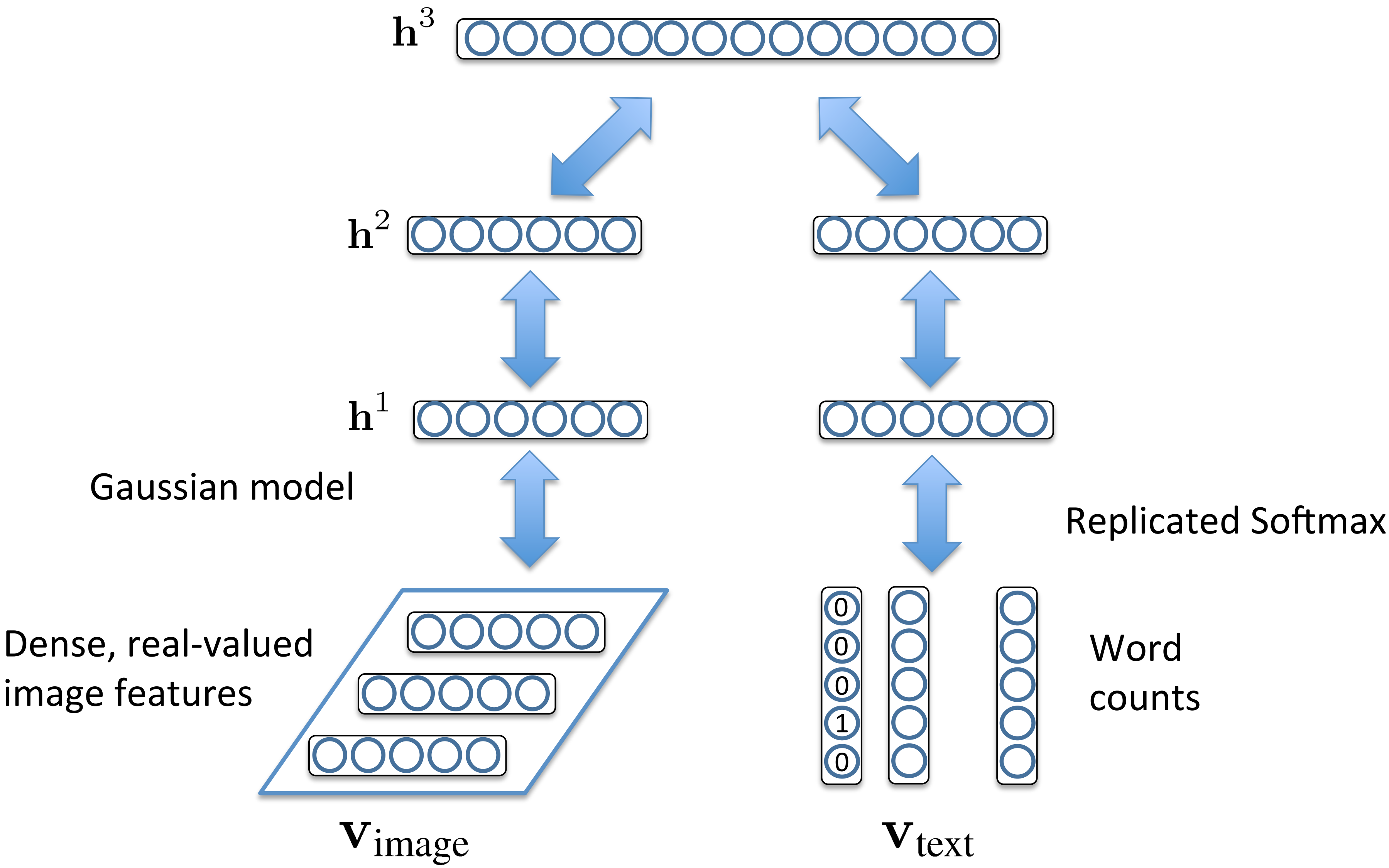
Multimodal DBM



(Srivastava & Salakhutdinov, NIPS 2012, JMLR 2014)

[Courtesy, R. Salakhutdinov]

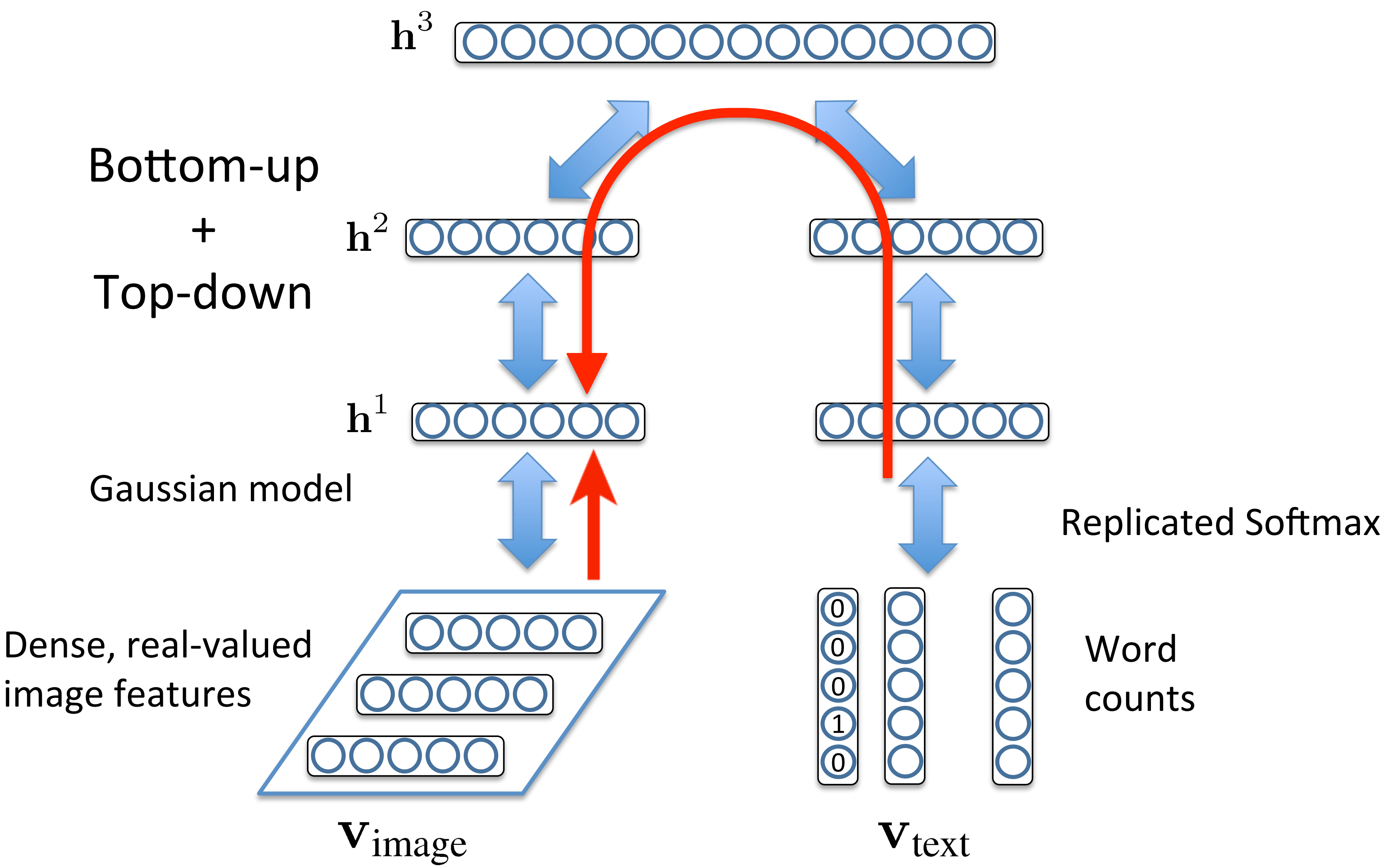
Multimodal DBM



[Courtesy, R. Salakhutdinov]

(Srivastava & Salakhutdinov, NIPS 2012, JMLR 2014)

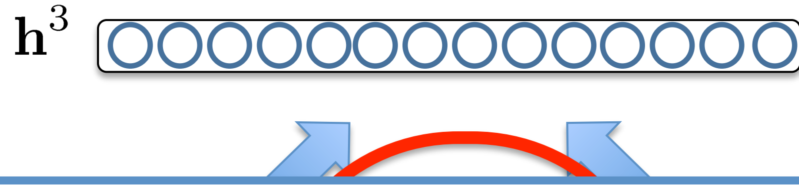
Multimodal DBM



(Srivastava & Salakhutdinov, NIPS 2012, JMLR 2014)

[Courtesy, R. Salakhutdinov]

Multimodal DBM



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

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

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

[Courtesy, R. Salakhutdinov]

(Srivastava & Salakhutdinov, NIPS 2012, JMLR 2014)

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

[Courtesy, R. Salakhutdinov]

Text Generated from Images

Given | Generated

portrait, women, army, soldier, mother, postcard, soldiers

obama, barackobama, election, politics, president, hope, change, sanfrancisco, convention, rally

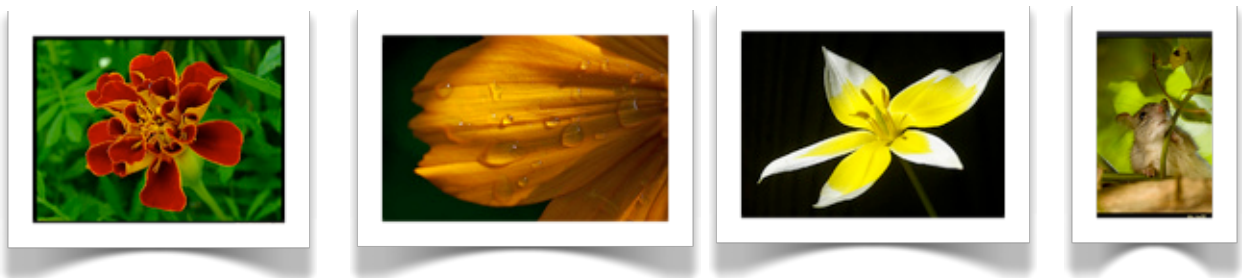
water, glass, beer, bottle, drink, wine, bubbles, splash, drops, drop

Images Selected from Text

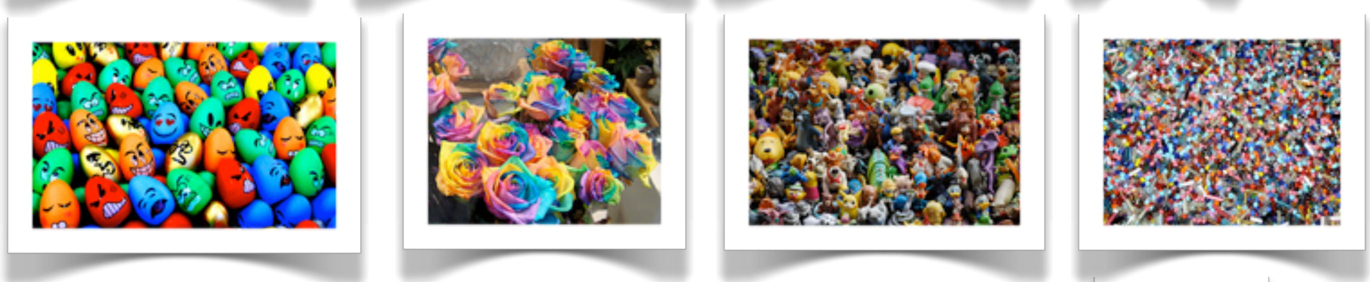
Given | Retrieved

water, red, sunset

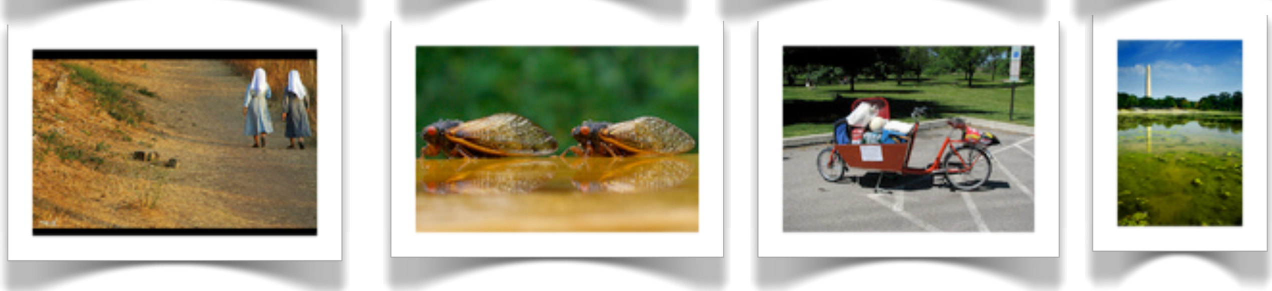
nature, flower, red, green



blue, green, yellow, colors



chocolate, cake



[Courtesy, R. Salakhutdinov]

Summary

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

Text & image retrieval / Object recognition

WIKIPEDIA
The Free Encyclopedia

REUTERS

AP Associated Press

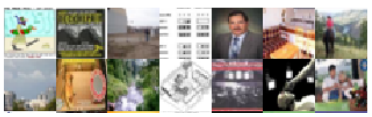
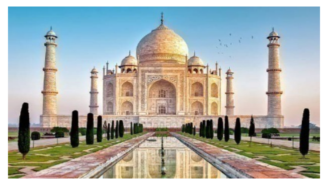
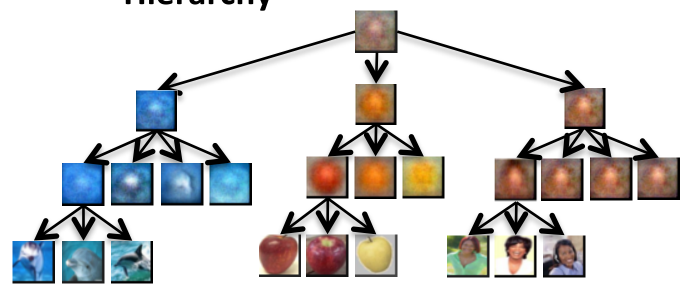


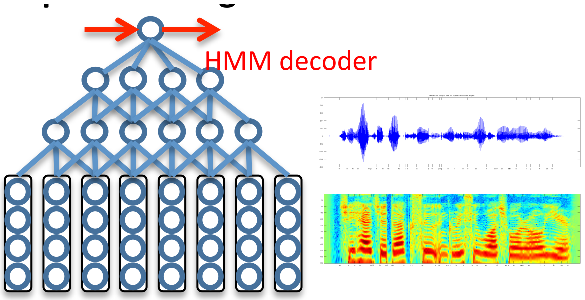
Image Tagging



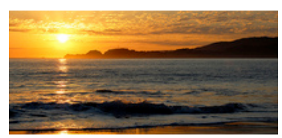
mosque, tower, building, cathedral, dome, castle

Learning a Category Hierarchy

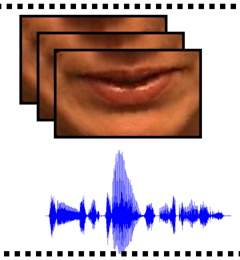




Multimodal Data



sunset, pacific ocean, beach, seashore



Caption Generation

a car is parked in the middle of nowhere .

a ferry boat on a marina with a group of people .

a little boy with a bunch of friends on the street .

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

[Courtesy, R. Salakhutdinov]