Learning Large-Scale Hierarchical Models

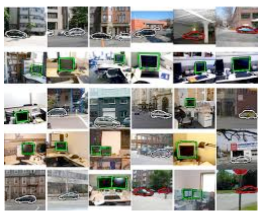
Russ Salakhutdinov

Department of Computer Science and Statistics
University of Toronto

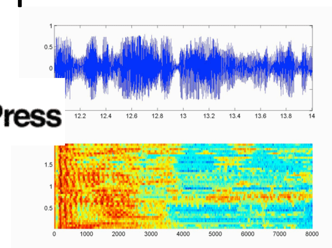
Discovering Structure

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

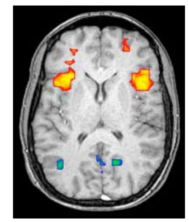
Images & Video



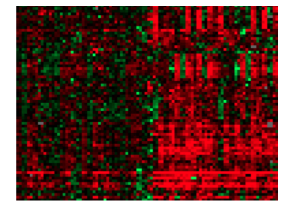
Text & Speech



Medical Imaging

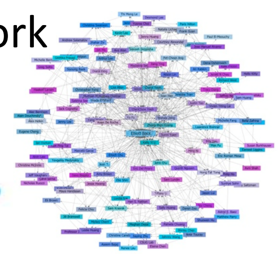


Gene Expression



Product Recommendation

Relational Data/ Social Network

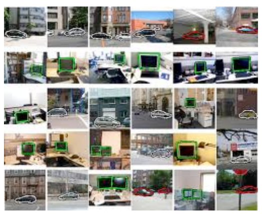


Discovering Structure

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

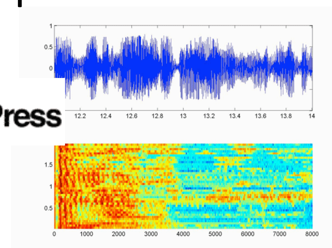
Images & Video

flickr
Google
bing

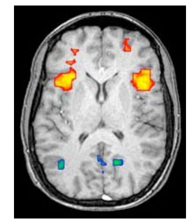


Text & Speech

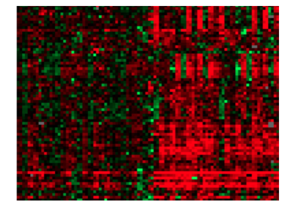
REUTERS
AP Associated Press
Wikipedia
The Free Encyclopedia



Medical Imaging



Gene Expression

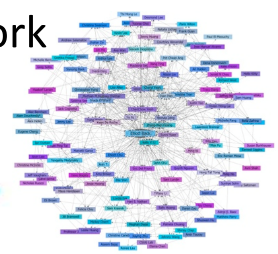


Product Recommendation

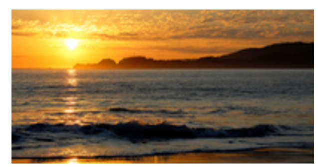
amazon
NETFLIX
ebay

Relational Data/ Social Network

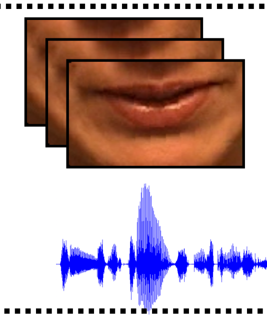
facebook
twitter



Multimodal Data



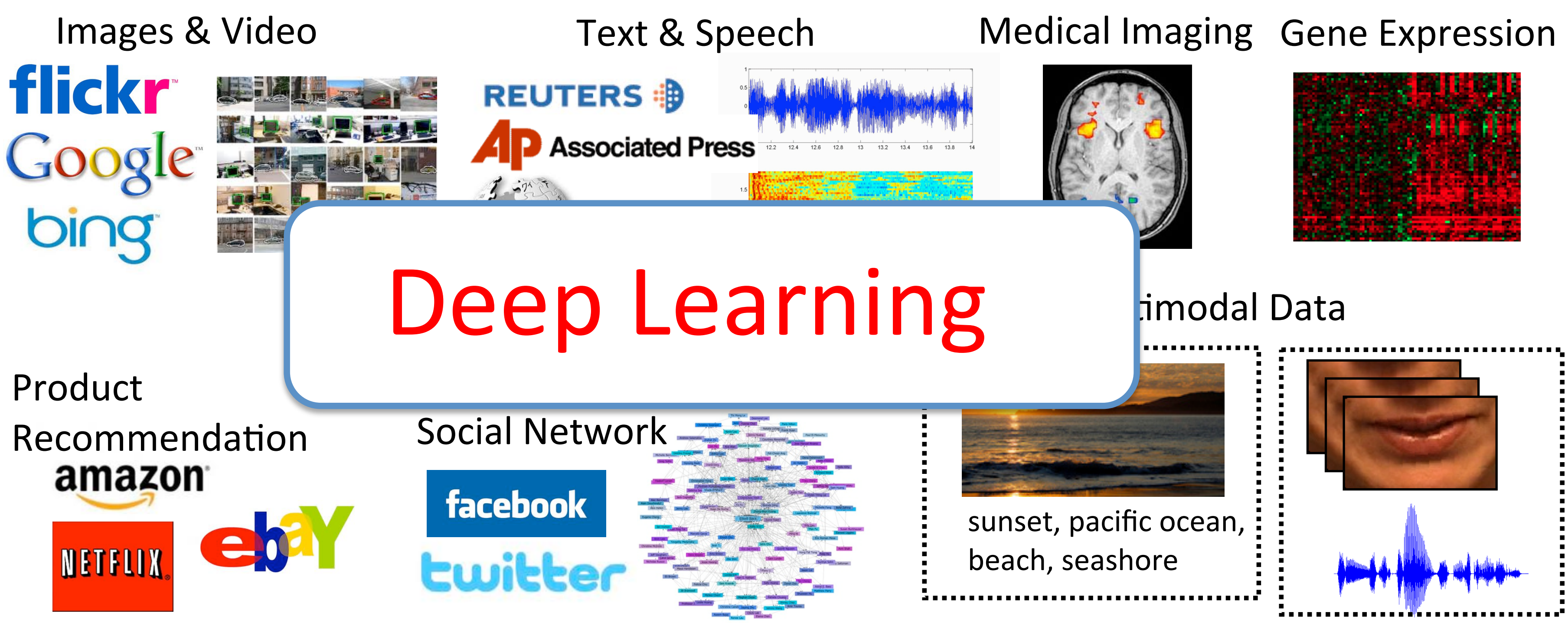
sunset, pacific ocean, beach, seashore



- Discover underlying structure, semantic relations, or invariances from data.
- Robust, adaptive models models that can deal with missing measurements, nonstationary distributions, multimodal data.

Discovering Structure

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



- Discover underlying structure, semantic relations, or invariances from data.
- Robust, adaptive models models that can deal with missing measurements, nonstationary distributions, multimodal data.

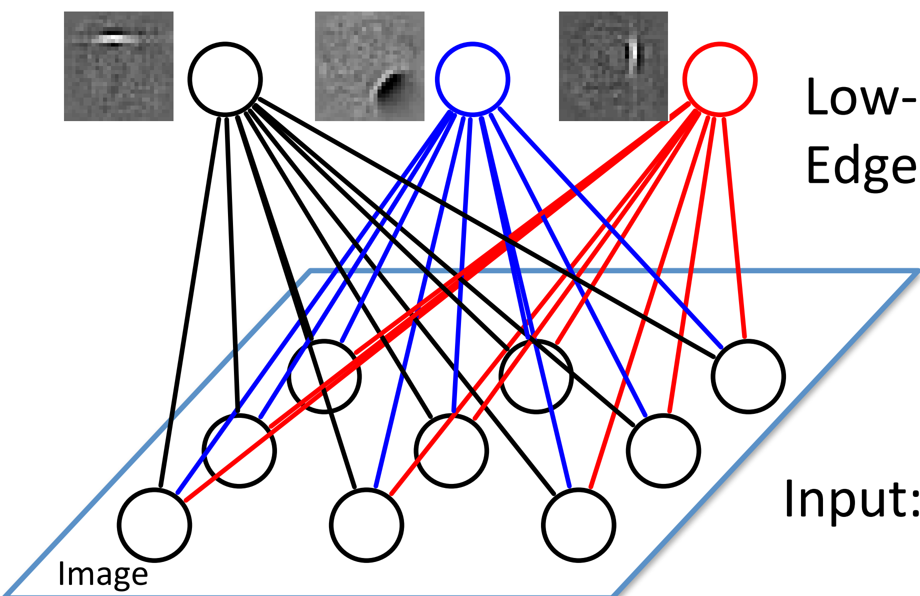
Research Directions

- Deep Learning

- Learning More Structured Models: Transfer and One-Shot Learning

- Multimodal Learning

Example: Deep Boltzmann Machines



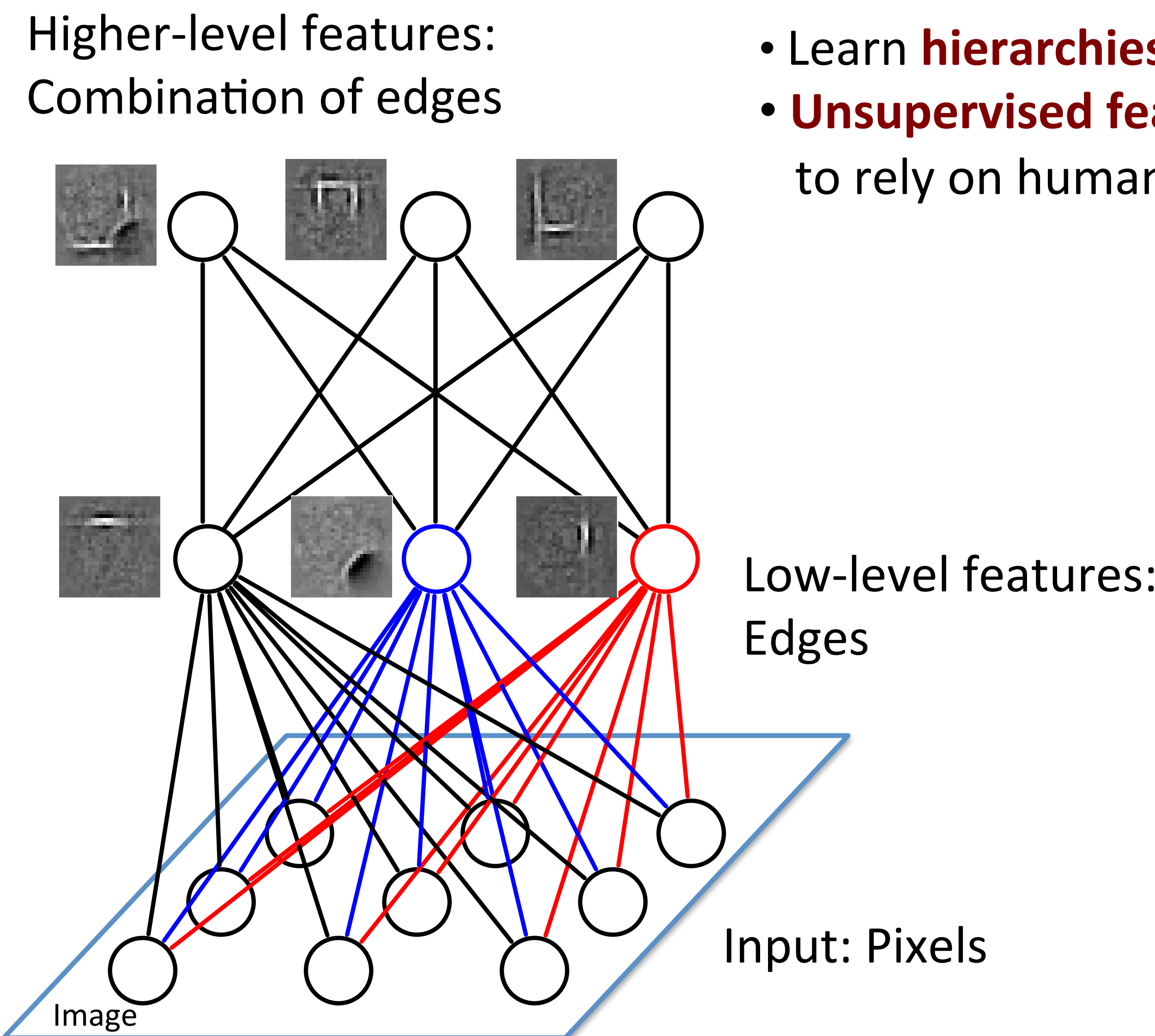
Low-level features:
Edges

Input: Pixels

(Salakhutdinov & Hinton, AIStats 2009, Neural Computation 2012)

Example: Deep Boltzmann Machines

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

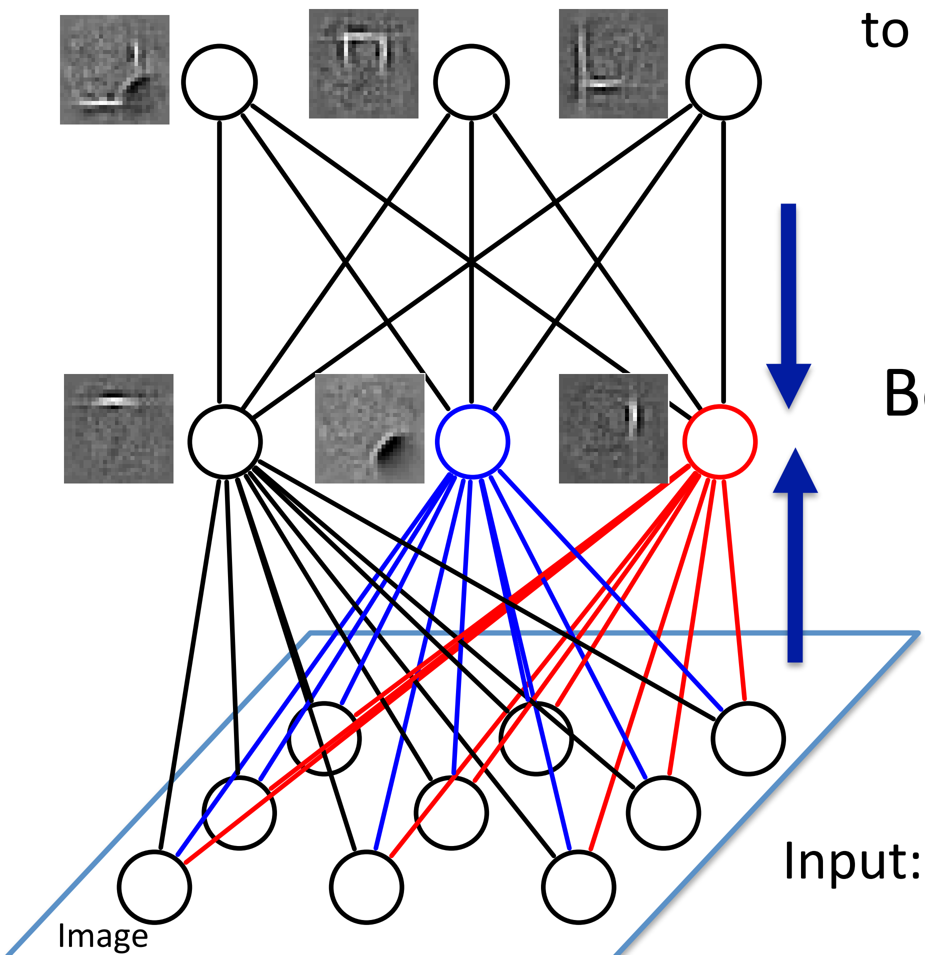


(Salakhutdinov & Hinton, AIStats 2009, Neural Computation 2012)

Example: Deep Boltzmann Machines

Higher-level features:
Combination of edges

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

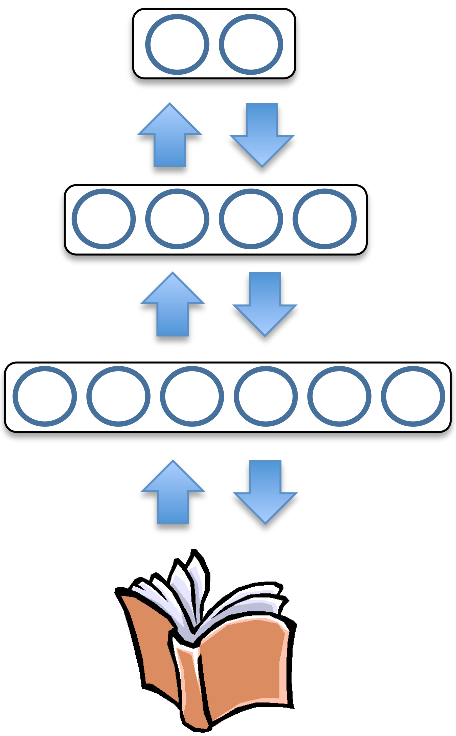


Bottom-up + Top-down

Input: Pixels

(Salakhutdinov & Hinton, AIStats 2009, Neural Computation 2012)

Deep Generative Model

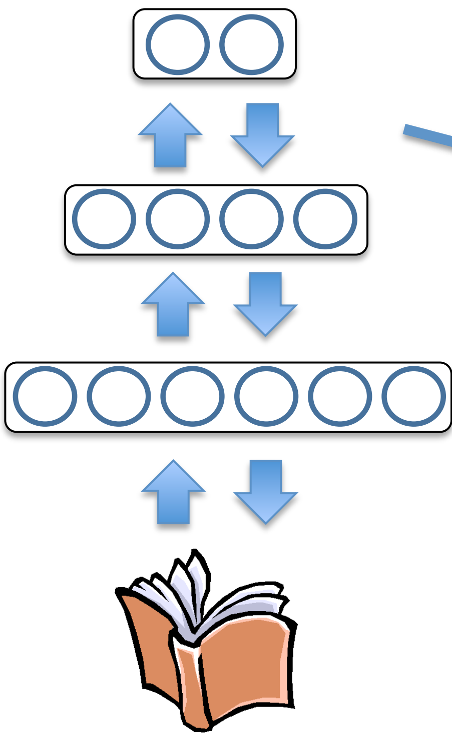


Bag of words

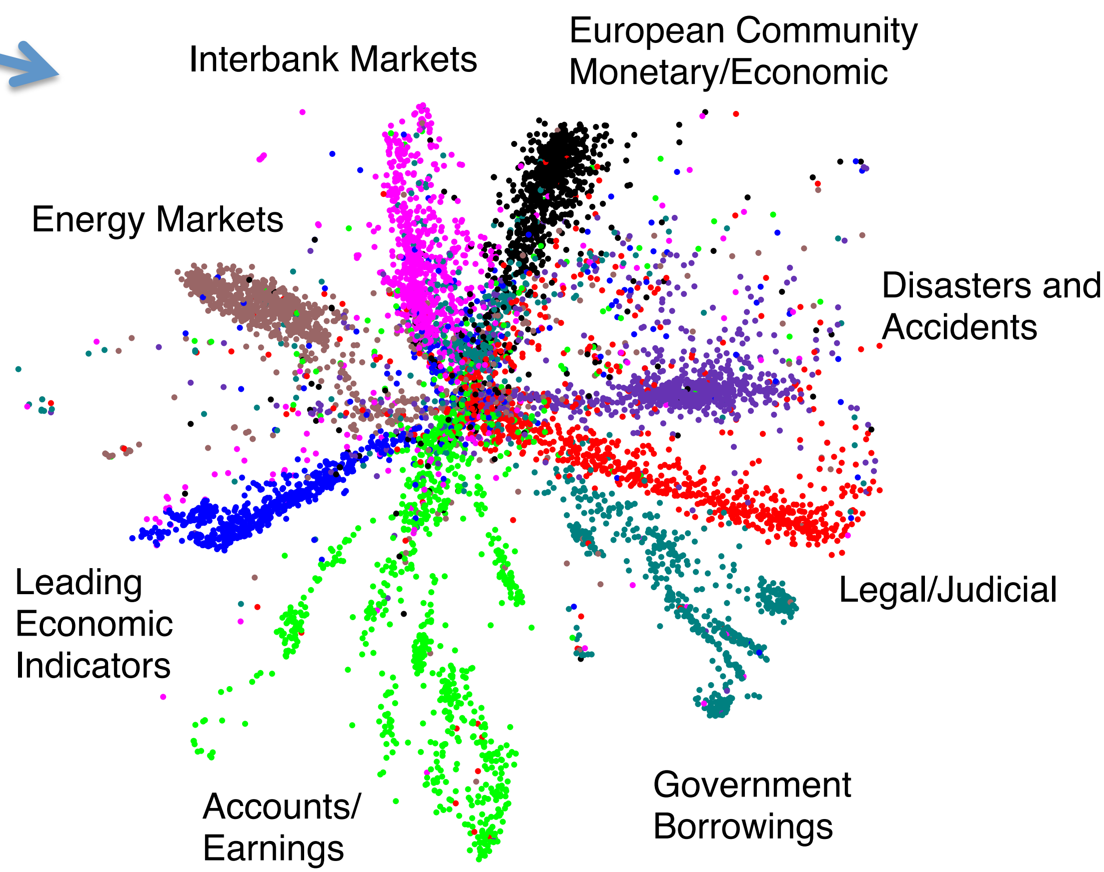
Reuters dataset: 804,414 newswire stories: **unsupervised**

Deep Generative Model

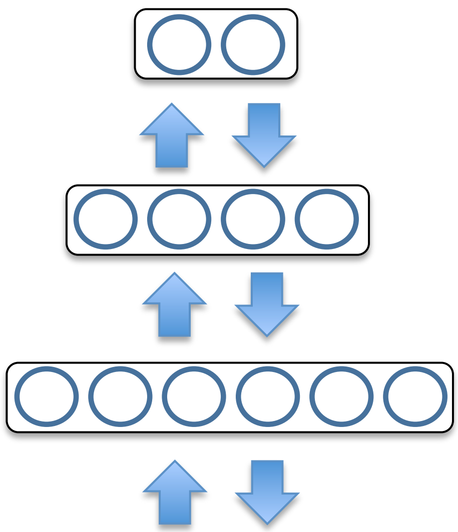
Reuters dataset: 804,414 newswire stories: **unsupervised**



Bag of words



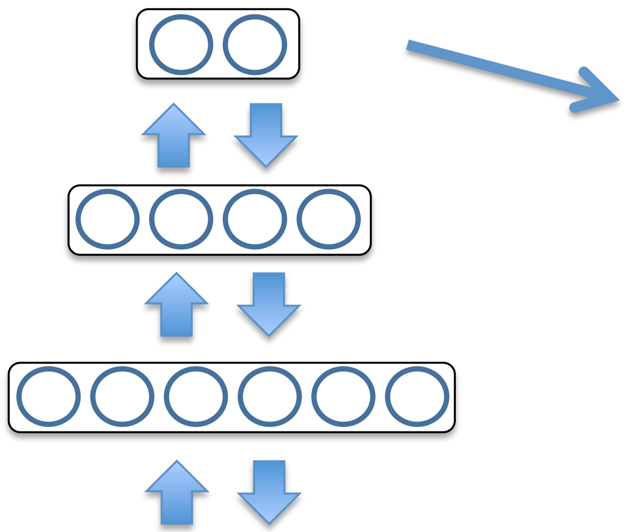
Deep Generative Model



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

(Salakhutdinov et. al. ICML 2007)

Deep Generative Model



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

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

Learned features: ``genre''

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

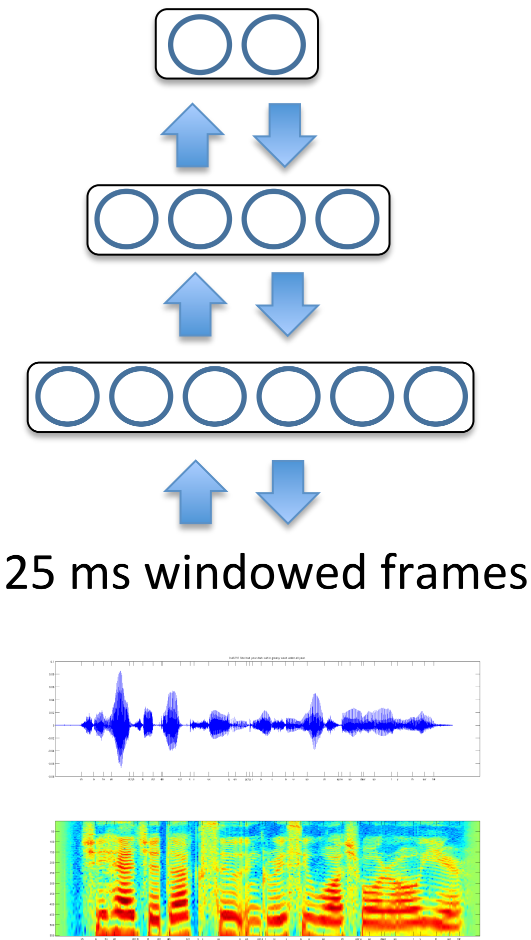
Independence Day
The Day After Tomorrow
Con Air
Men in Black II
Men in Black

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

(Salakhutdinov et. al. ICML 2007)

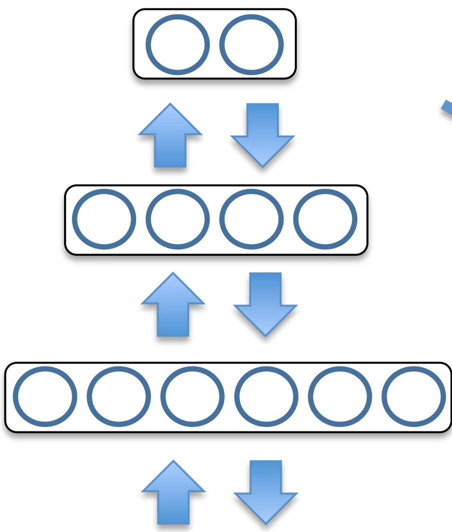
Deep Generative Model



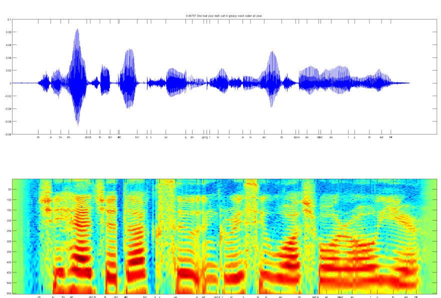
- **Speech Recognition: Spoken Query Detection**: For each keyword, estimate utterance's probability of containing that keyword.

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

Deep Generative Model



25 ms windowed frames



- **Speech Recognition: Spoken Query Detection**: For each keyword, estimate utterance's probability of containing that keyword.

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%

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

Research Directions

- Deep Learning

- Learning More Structured Models: Transfer and One-Shot Learning

- Multimodal Learning

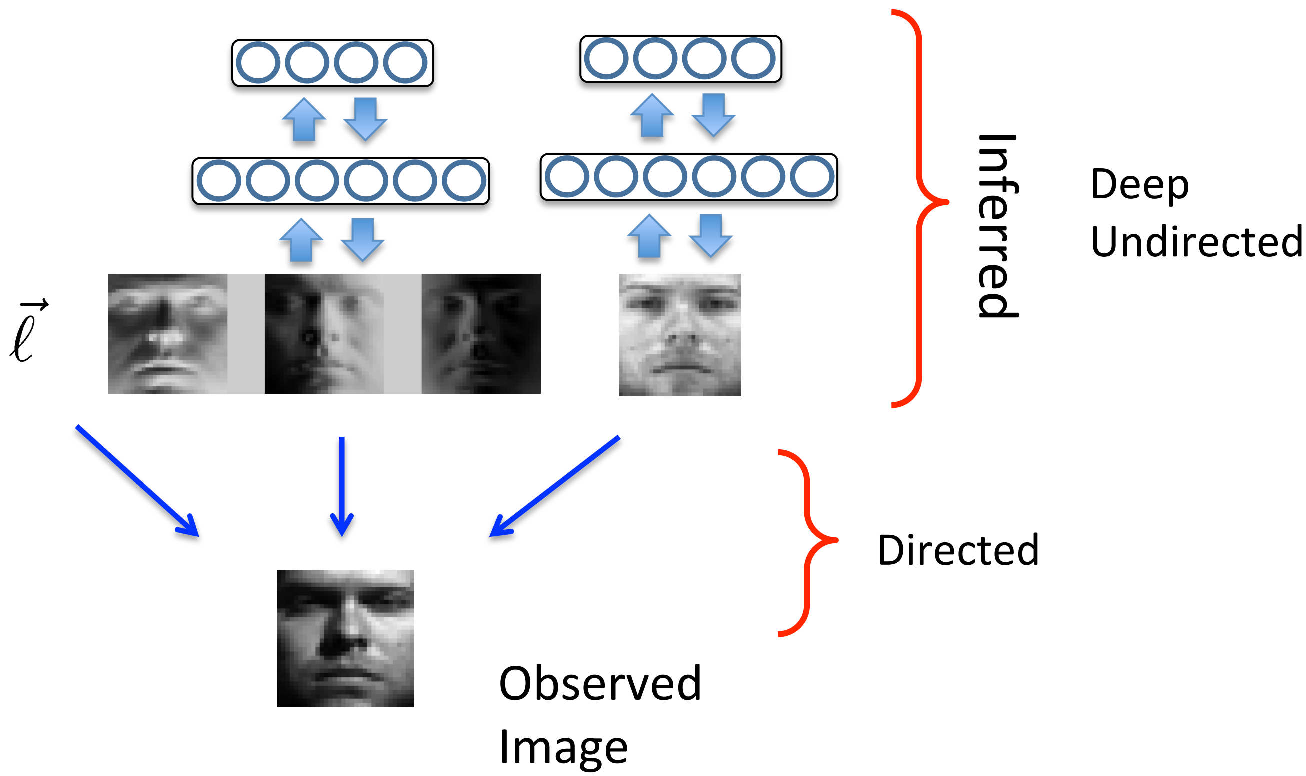
Face Recognition

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.

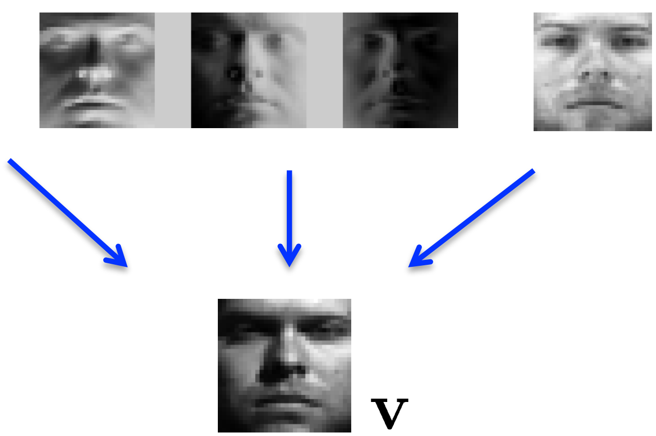


Combines the elegant properties of the Lambertian model with the Gaussian DBM model.

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

Deep Lambertian Model

$\vec{\ell}$



$\mathbf{v}$

Observed Image

Image albedo

Surface normals

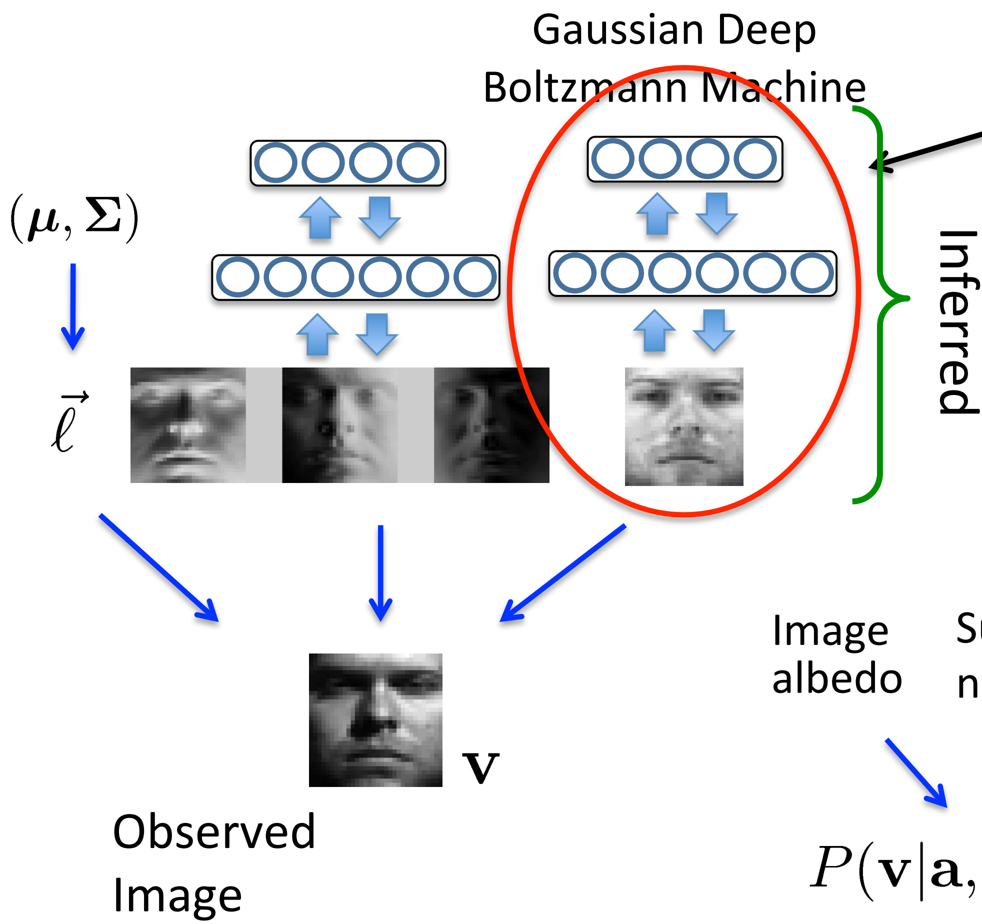
Light source

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

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

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

Deep Lambertian Model



Transfer Learning

Albedo DBM:
Pretrained using
Toronto Face Database

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

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

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

Face Relighting

One Test Image

Observed

Inferred albedo

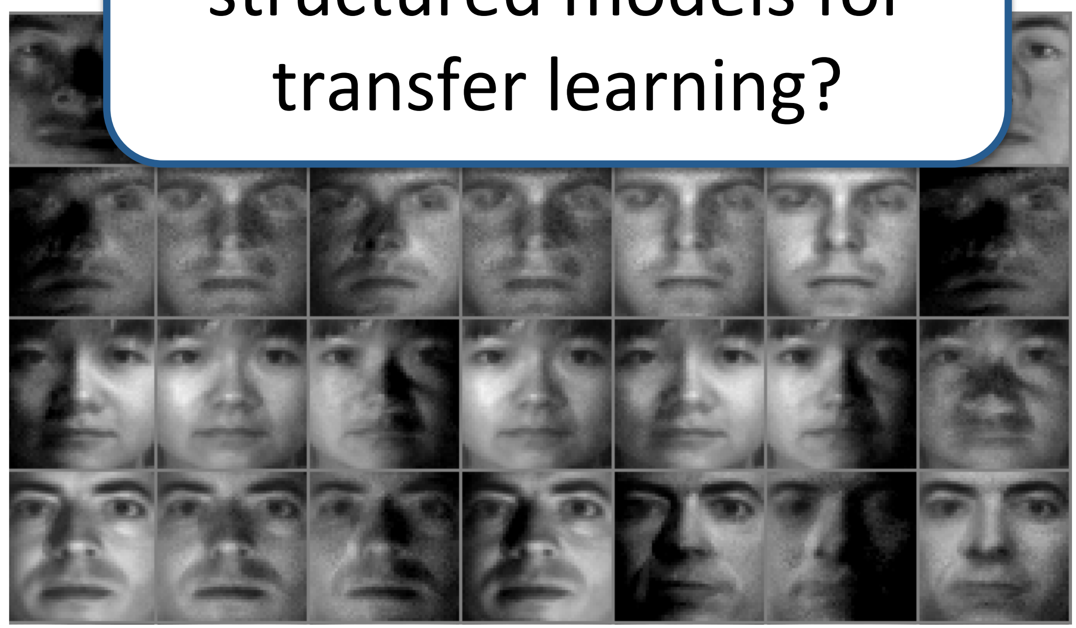
Face Relighting

Face Relighting

One Test Image

Observed Inferred albedo

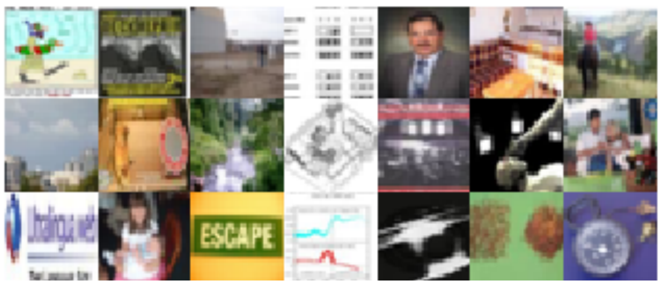
What about building structured models for transfer learning?



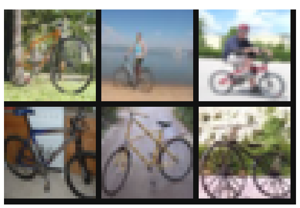
Transfer Learning

Background Knowledge

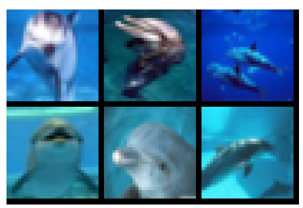
Millions of unlabeled images



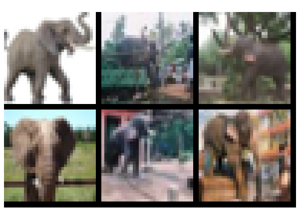
Some labeled images



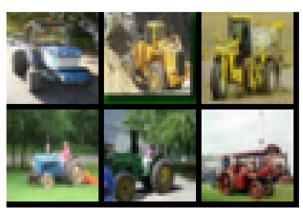
Bicycle



Dolphin



Elephant



Tractor

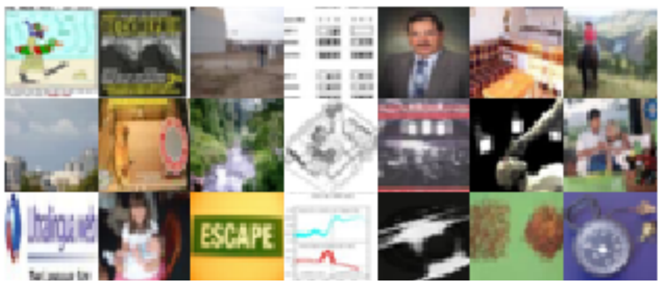
Learn to Transfer Knowledge

Learn novel concept from one example

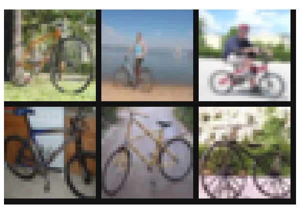
Transfer Learning

Background Knowledge

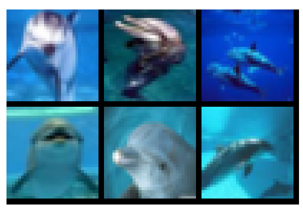
Millions of unlabeled images



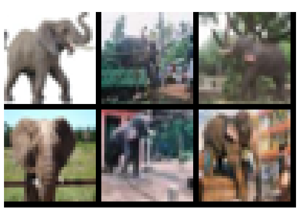
Some labeled images



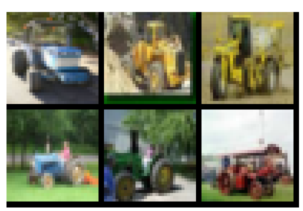
Bicycle



Dolphin



Elephant



Tractor

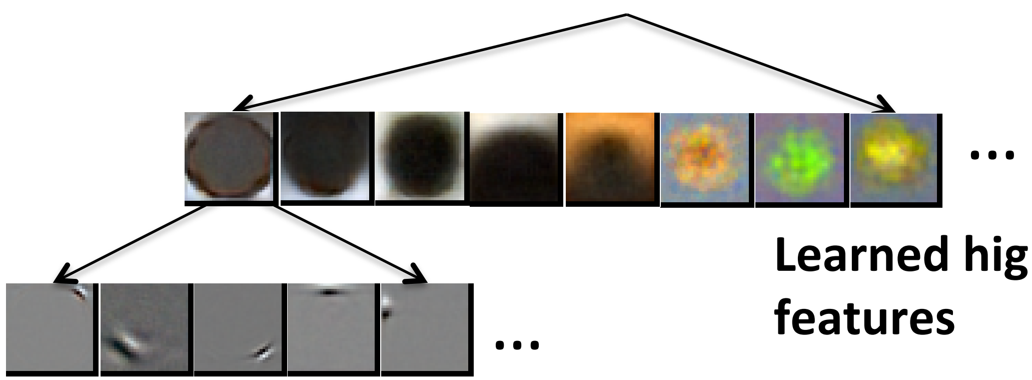
Learn to Transfer Knowledge

Learn novel concept from one example

Test:
What is this?

Learning Category Hierarchy

Learning to share the knowledge across many visual categories.



Learned higher-level features

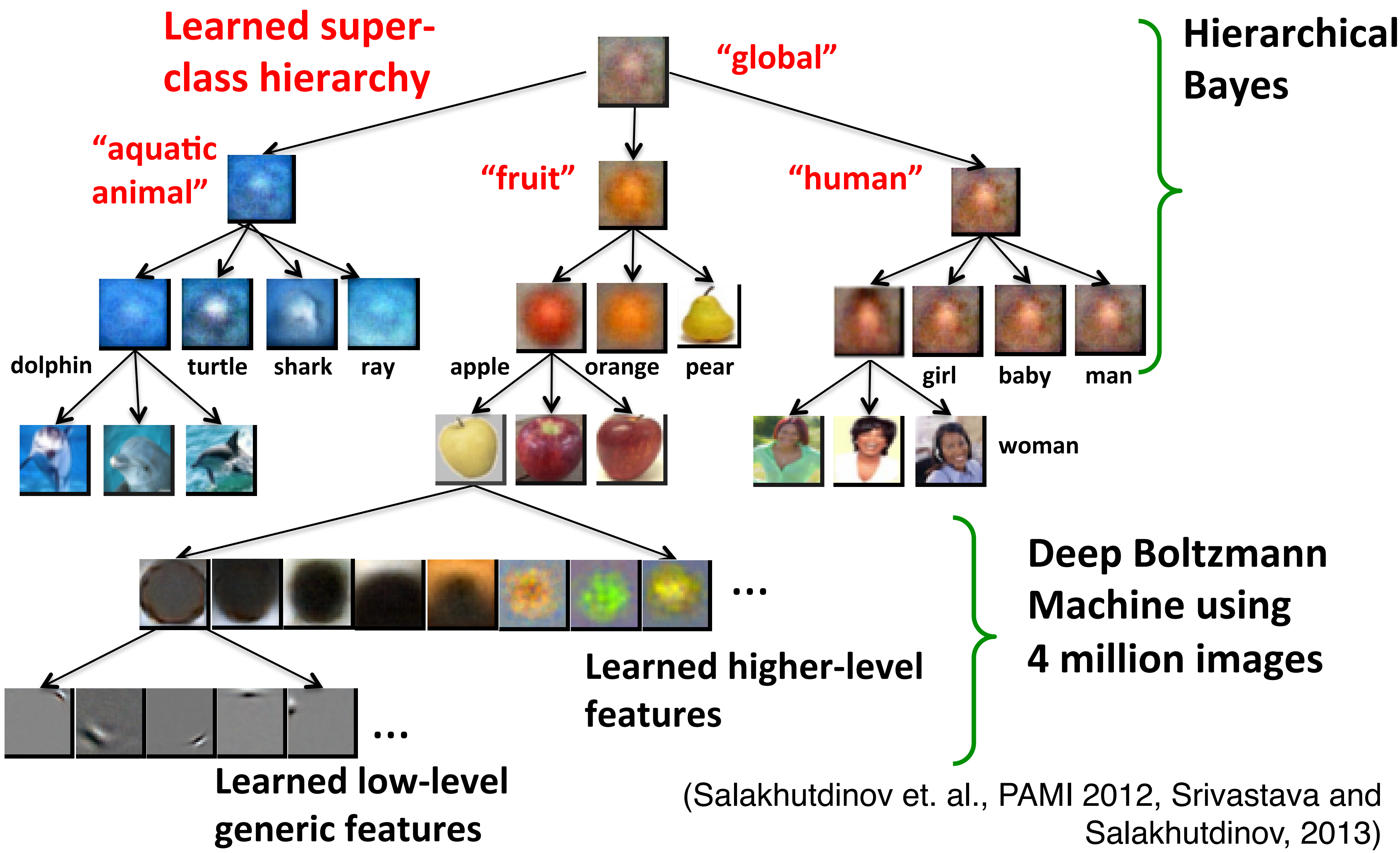
Learned low-level generic features

Deep Boltzmann Machine using 4 million images

(Salakhutdinov et. al., PAMI 2012, Srivastava and Salakhutdinov, 2013)

Learning Category Hierarchy

Learning to share the knowledge across many visual categories.



(Salakhutdinov et. al., PAMI 2012, Srivastava and Salakhutdinov, 2013)

Research Directions

- Deep Learning

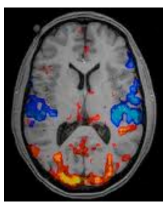
- Learning More Structured Models: Transfer Learning

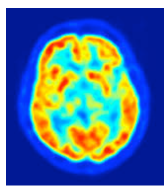
- Multimodal Learning

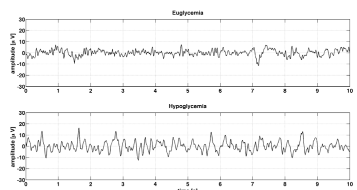
Data – Collection of Modalities

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

- Biomedical Imaging


fMRI


PET Scan

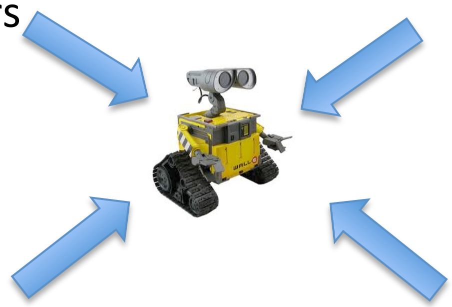

EEG


X-ray

- Robotics applications.

Touch sensors

Motor control



Vision

Audio

Shared Concept

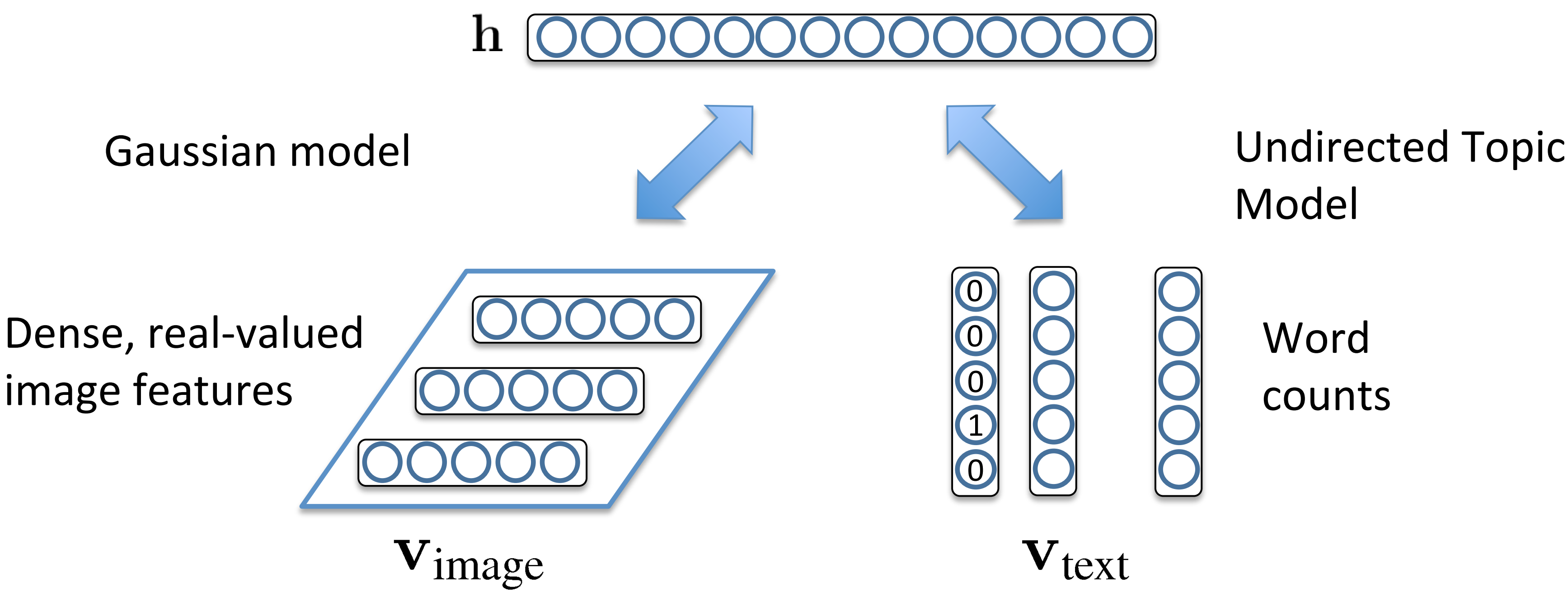
"Modality-free" representation

"Concept"

sunset, pacific ocean, baker beach, seashore, ocean

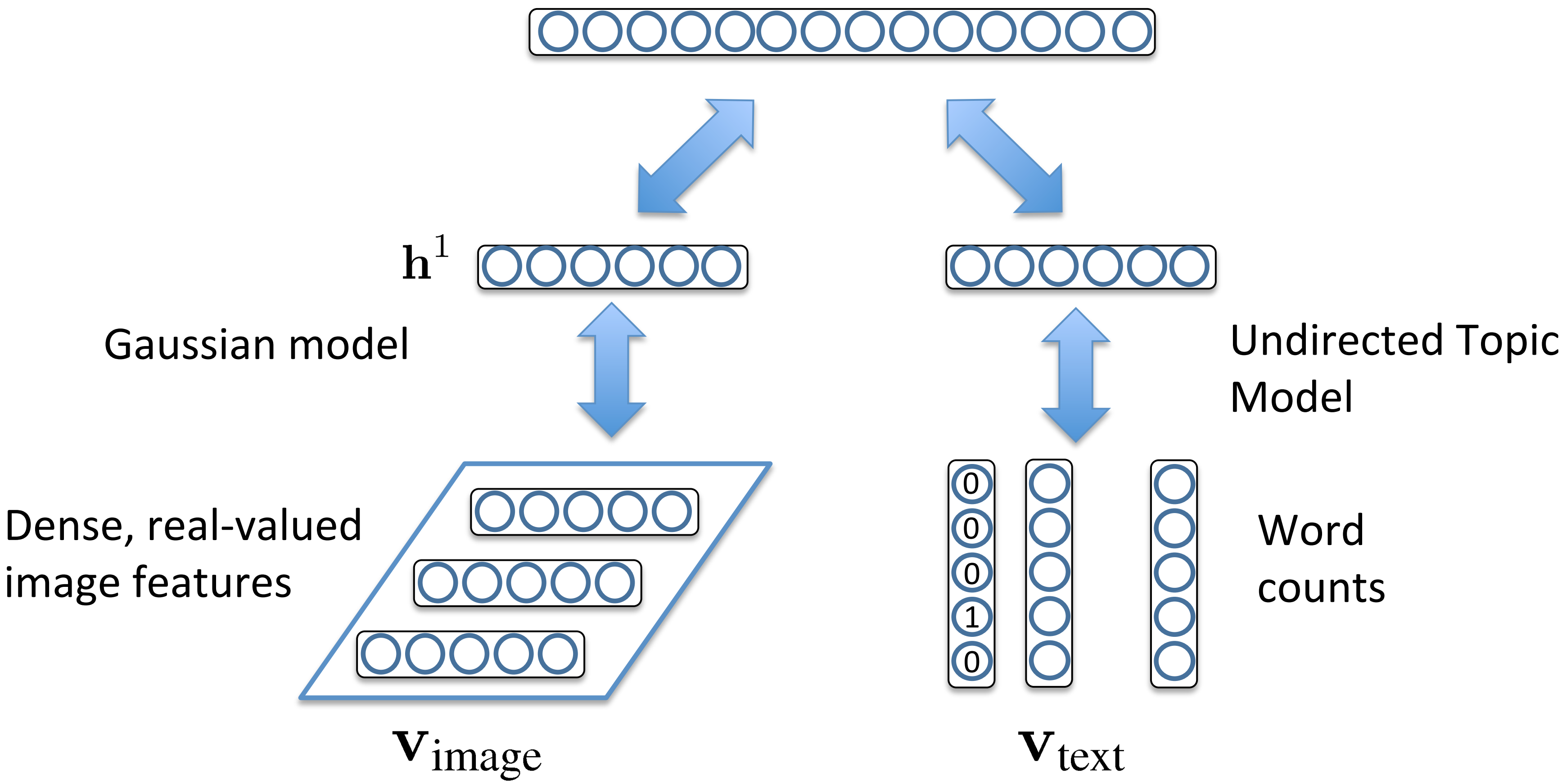
"Modality-full" representation

Multimodal DBM



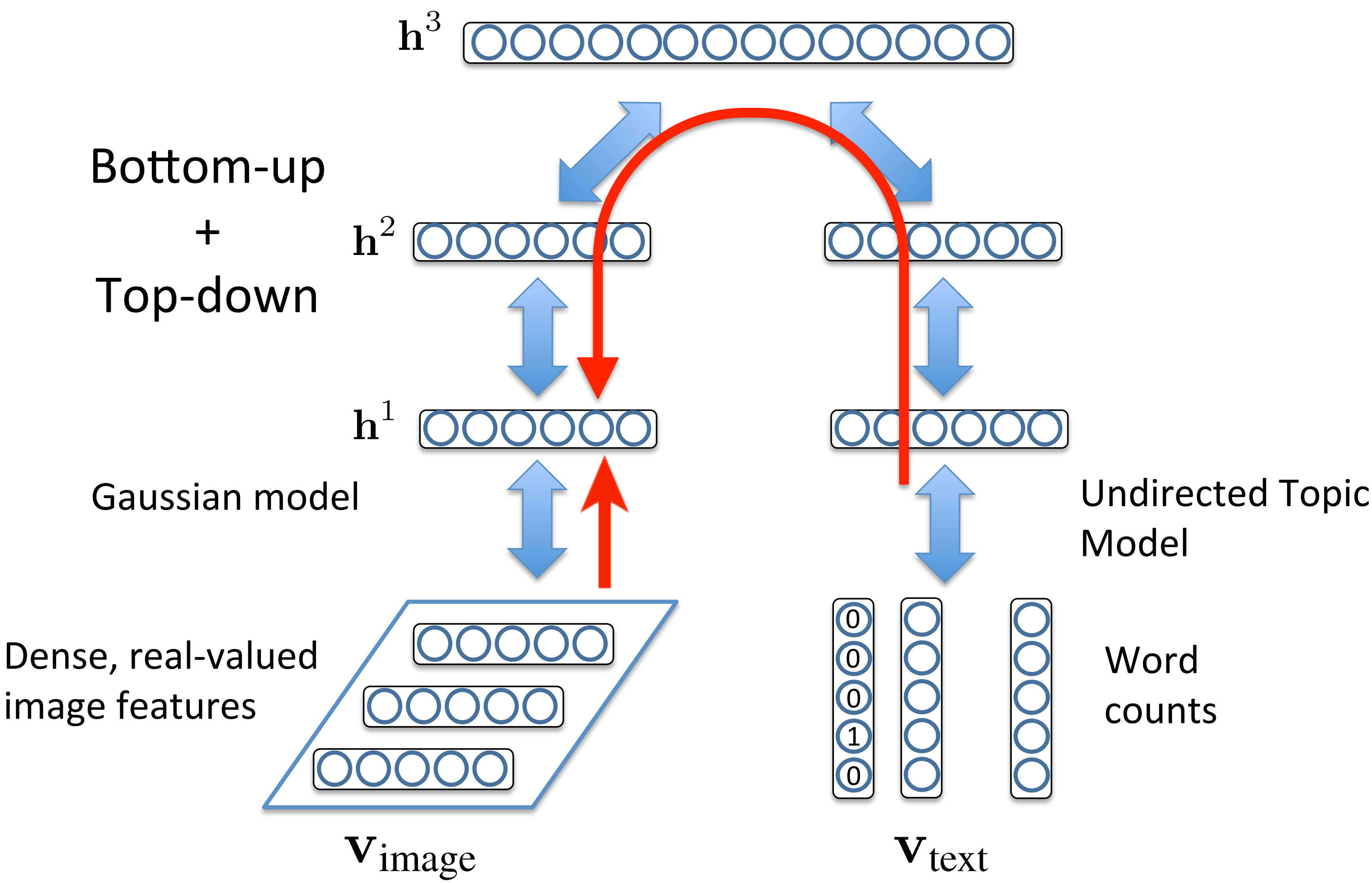
(Srivastava and Salakhutdinov, NIPS 2013)

Multimodal DBM



(Srivastava and Salakhutdinov, NIPS 2013)

Multimodal DBM

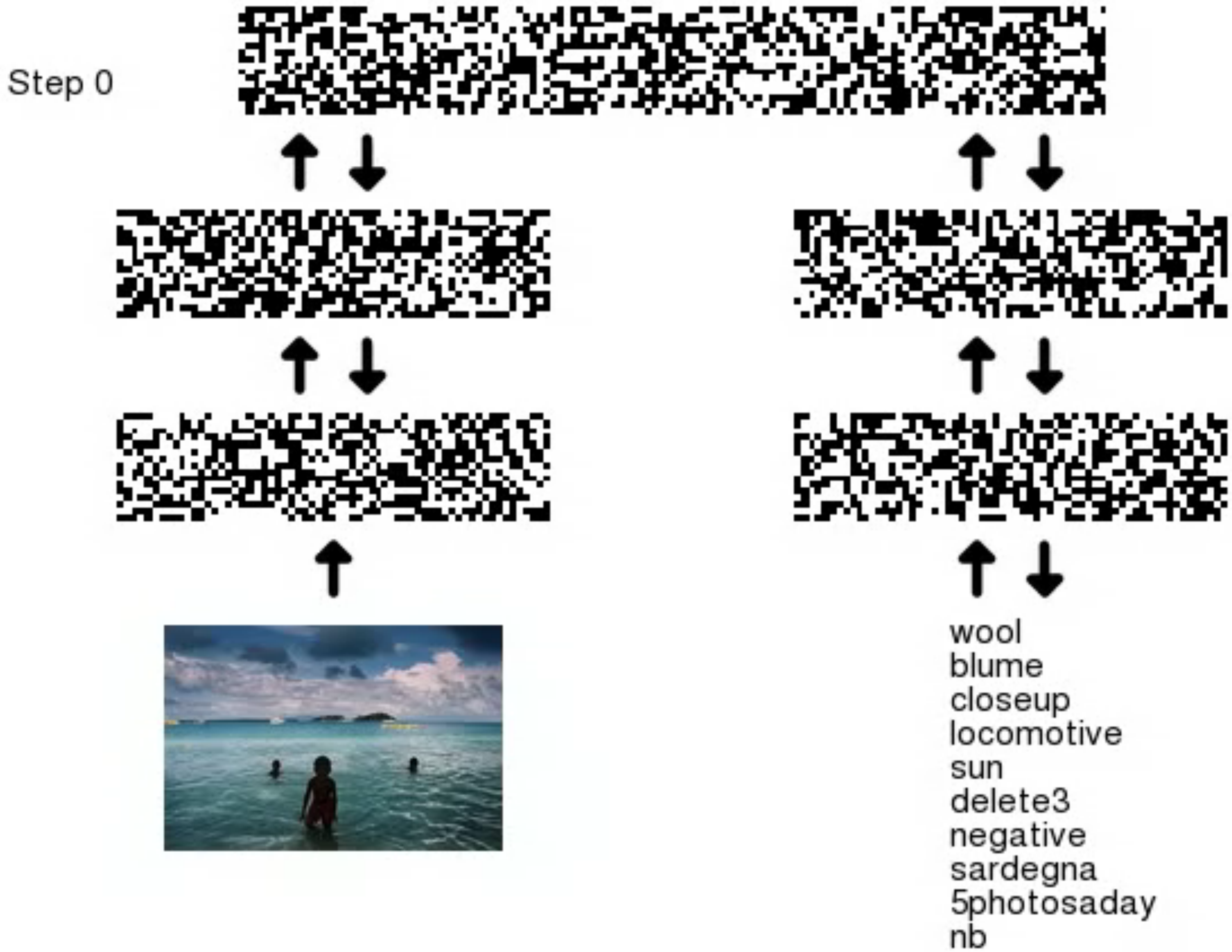


(Srivastava and Salakhutdinov, NIPS 2013)

Text Generated from Images

Image	Given 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

Multimodal DBMs



Samples drawn after every 50 steps of Gibbs updates

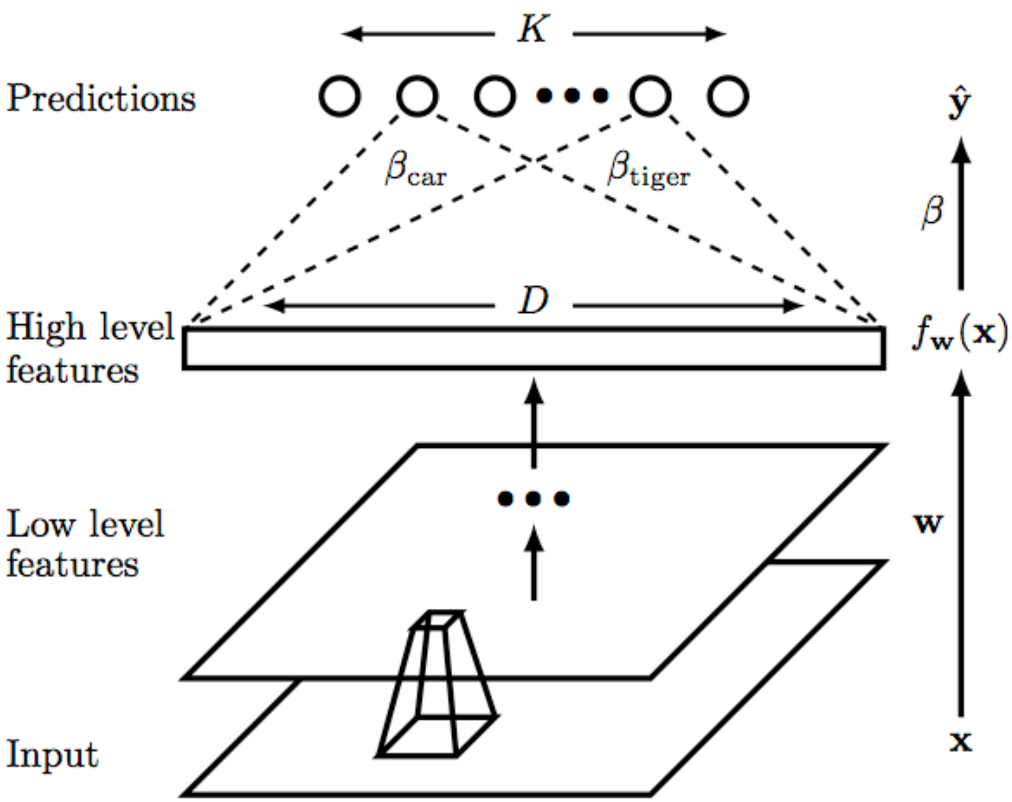
Sample at step 0
wool
blume
closeup
locomotive
sun
delete3
negative
sardegna
5photosaday
nb

Impact of Deep Learning

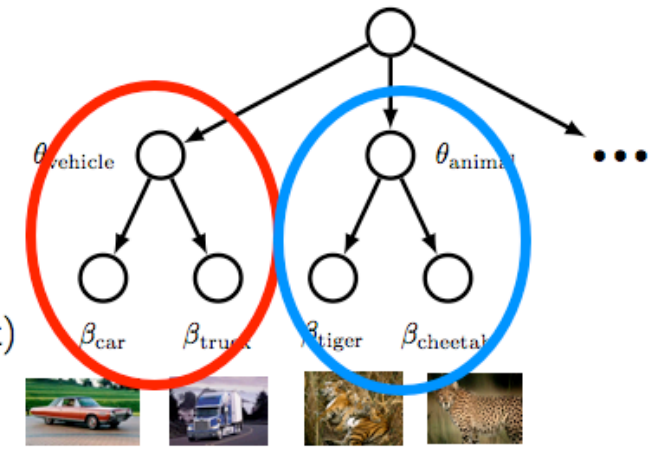
- Speech Recognition
- Computer Vision
- Recommender Systems
- Neuroscience
- Beginning: Drug Discovery and Medical Image Analysis

Thank you

Knowledge Transfer by Shared Priors



Tree-based priors



$$\beta_{\text{truck}} \sim \mathcal{N}(\theta_{\text{vehicle}}, 1/\lambda_1)$$

$$\beta_{\text{car}} \sim \mathcal{N}(\theta_{\text{vehicle}}, 1/\lambda_1)$$

$$\beta_{\text{tiger}} \sim \mathcal{N}(\theta_{\text{animal}}, 1/\lambda_1)$$

$$\theta_{\text{vehicle}} \sim \mathcal{N}(0, 1/\lambda_2)$$

$$\theta_{\text{animal}} \sim \mathcal{N}(0, 1/\lambda_2)$$