Deep Learning

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://bigml.cs.tsinghua.edu.cn/~jun
State Key Lab of Intelligent Technology & Systems
Tsinghua University

June 2, 2015

Why going deep?

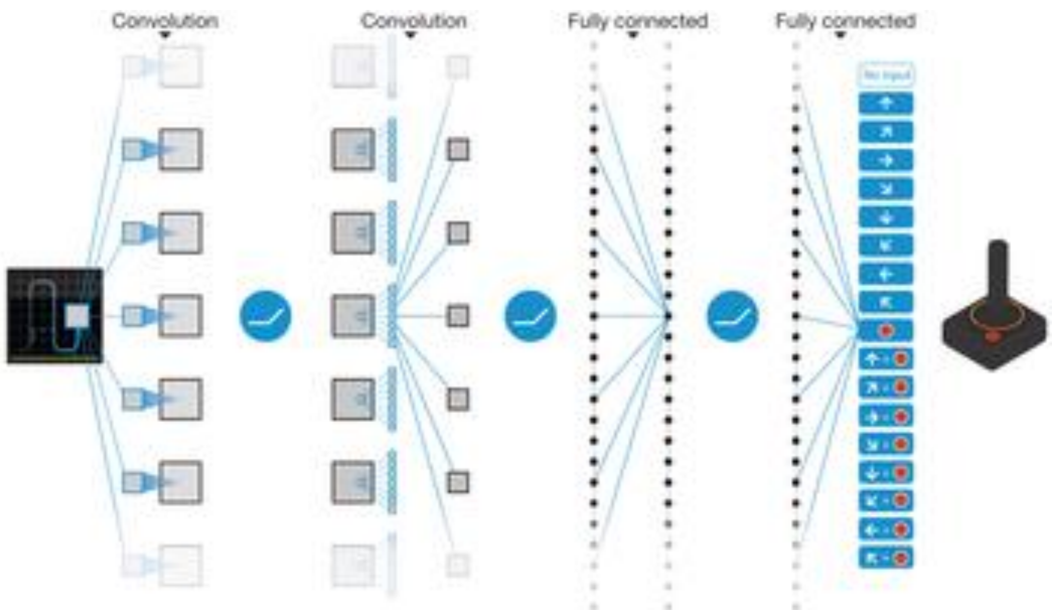
- Data are often high-dimensional.
- There is a huge amount of structure in the data, but the structure is too complicated to be represented by a simple model.
- Insufficient depth can require more computational elements than architectures whose depth matches the task.
- Deep nets provide simpler but more descriptive models of many problems.

Microsoft's speech recognition system

- http://v.youku.com/v_show/id_XNDc0MDY4ODI0.html

Human-Level Control via Deep RL

- Deep Q-network with human-level performance on A



- Minjie will talk more in next lecture

[Mnih et al., Nature 518, 529–533, 2015]

清華大學 Tsinghua University

MIT 10 Breakthrough Tech 2013

http://www.technologyreview.com/featuredstory/513696/deep-learning/

Deep Learning in industry

Driverless car

Face identification

Speech recognition

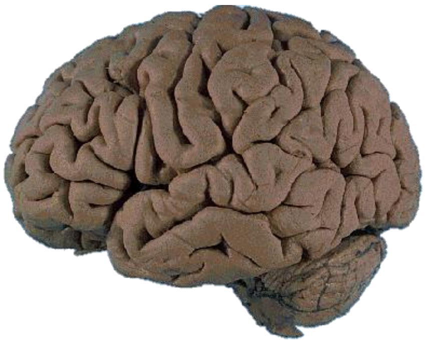
Web search

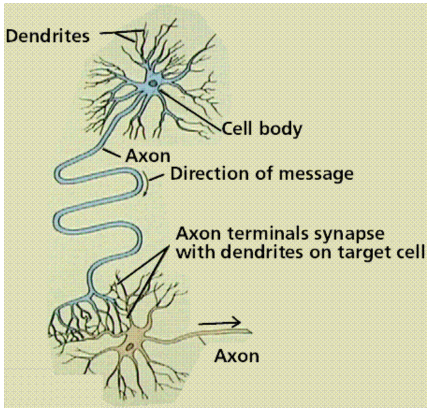
…

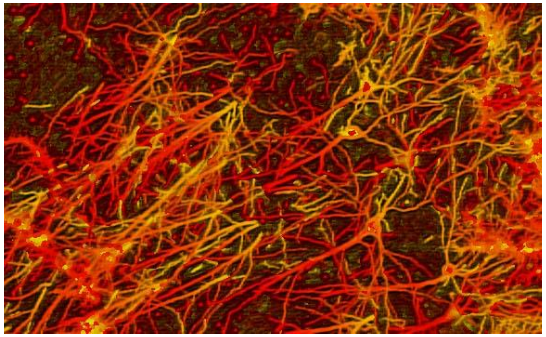
…

Deep Learning Models

How brains seem to do computing?





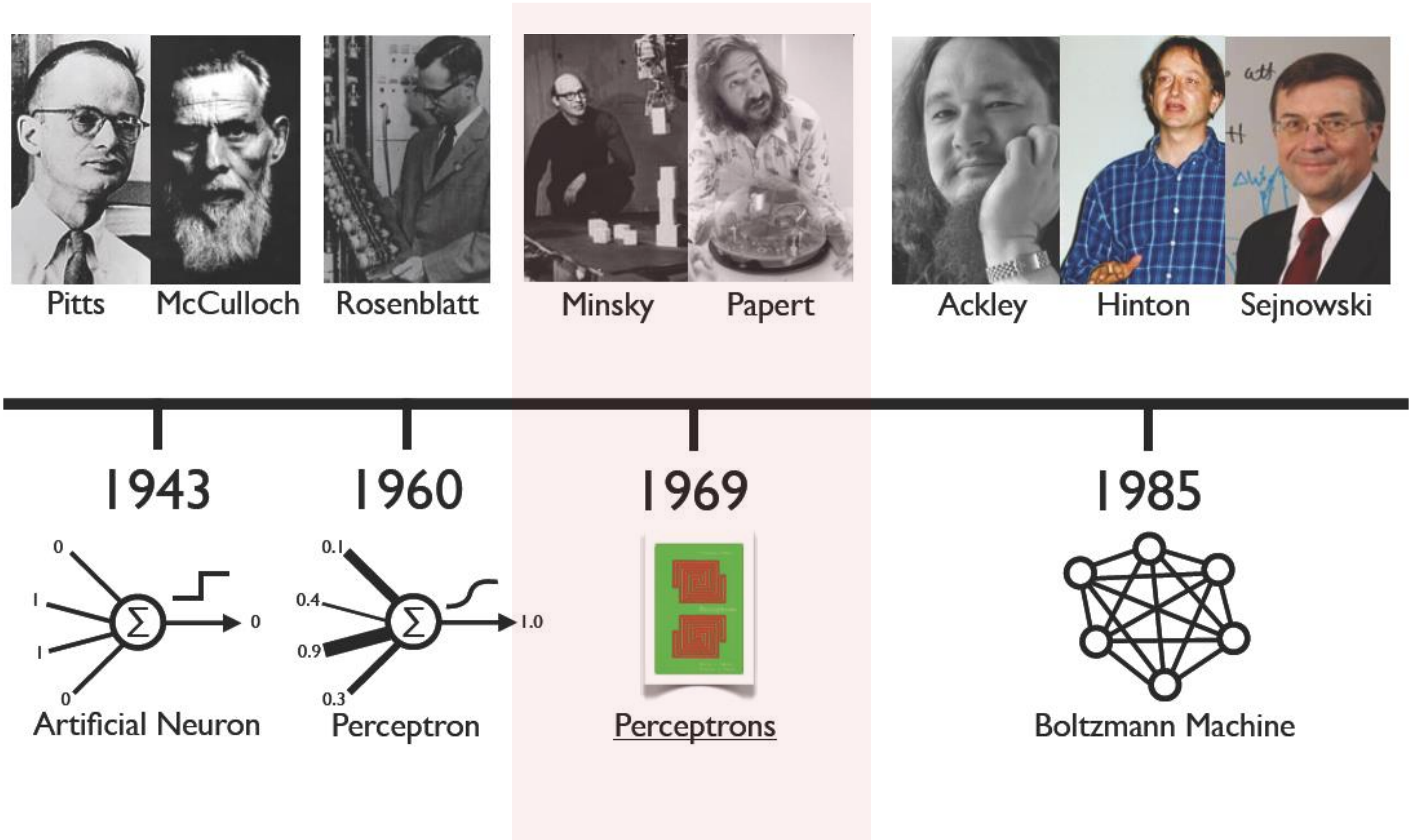


The business end of this is made of lots of these joined in networks like this

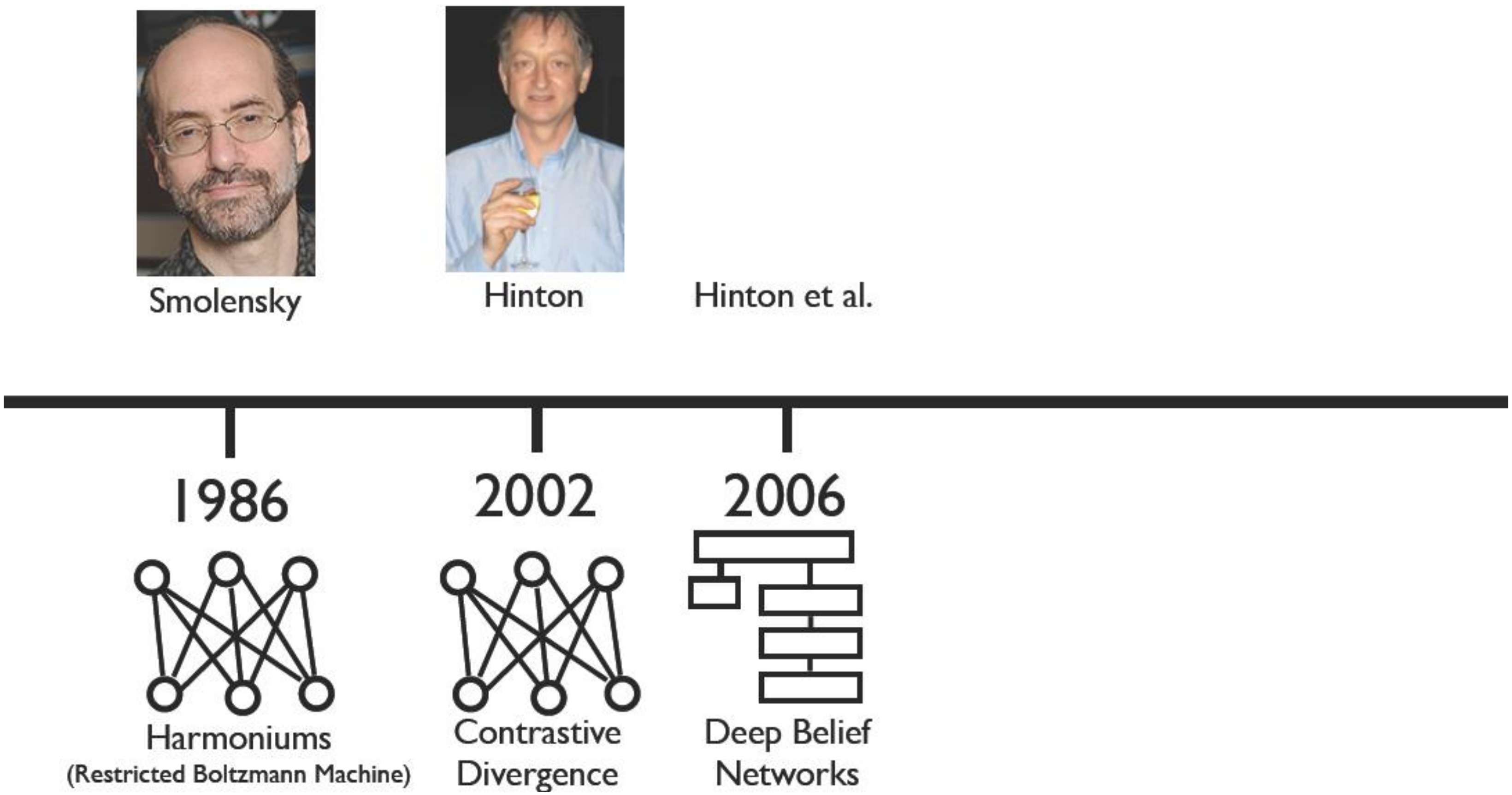
Much of our own "computations" are performed in/by this network

Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes

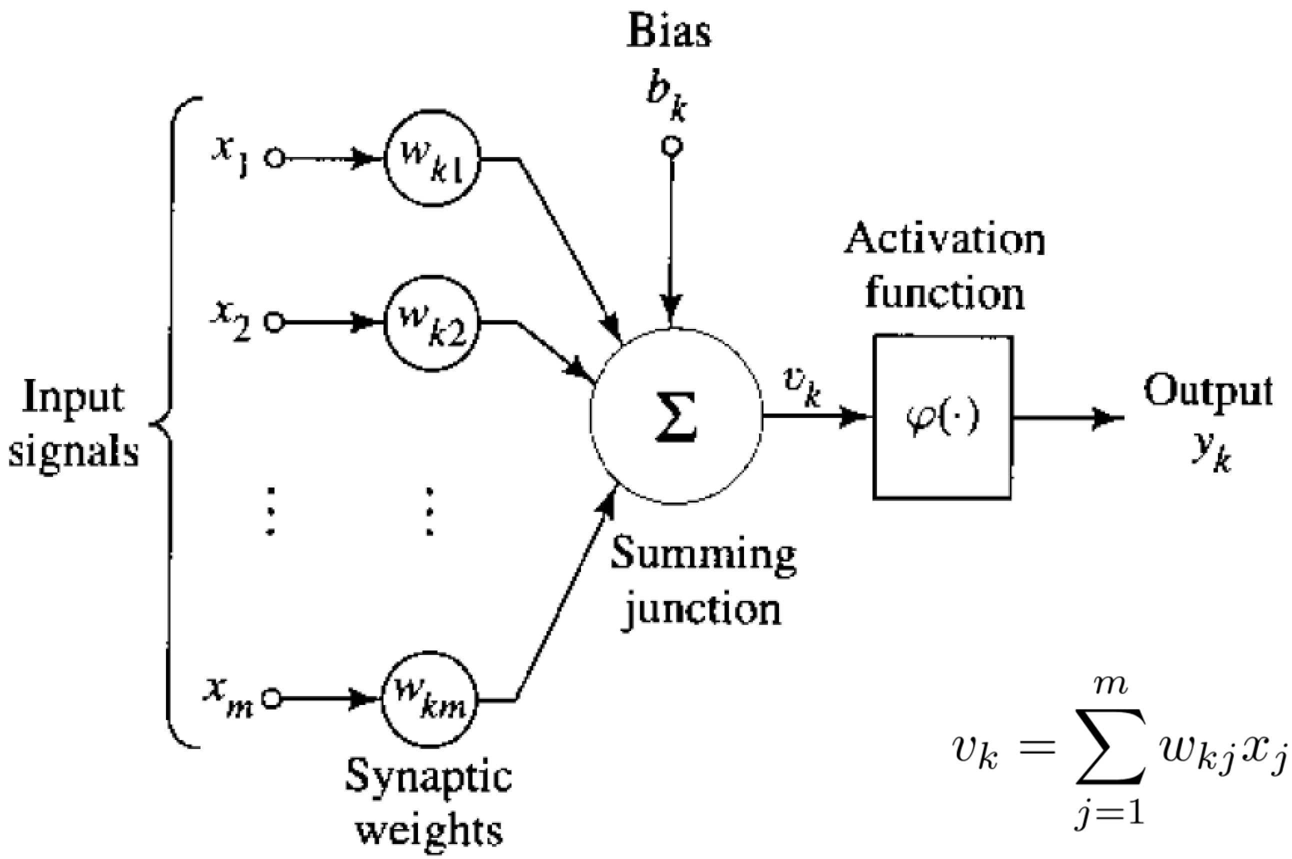
History of neural networks



History of neural networks



Model of a neuron

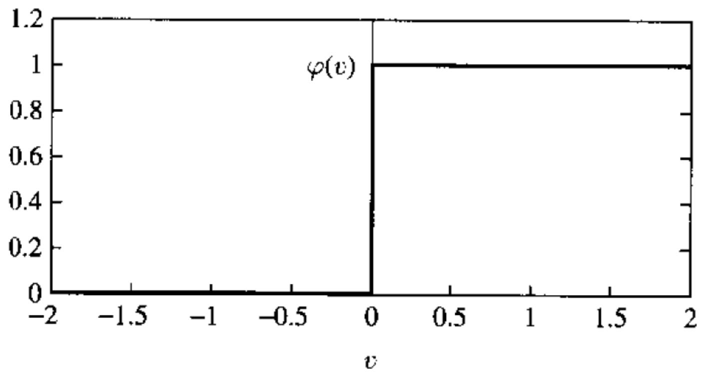


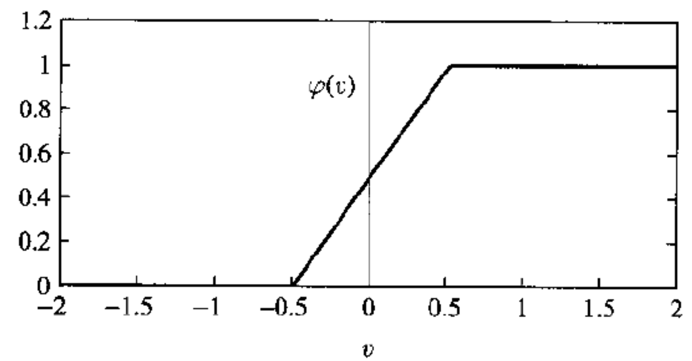
$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k$$

$$y_k = \psi(v_k)$$

Activation function

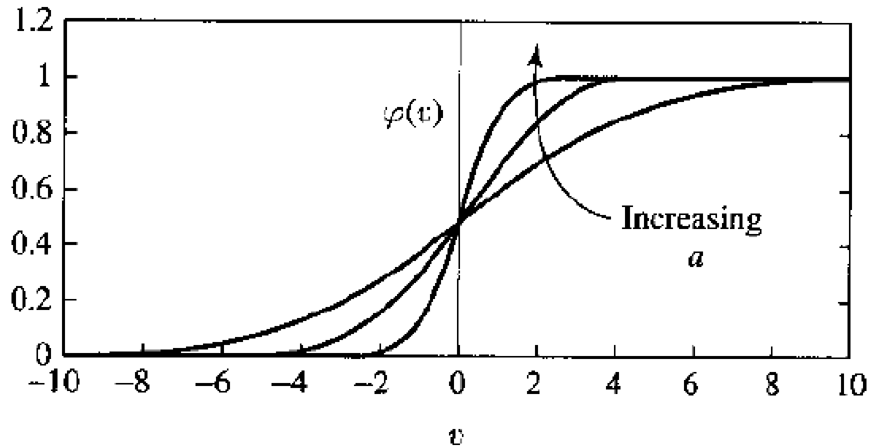
◆ Threshold function & piecewise linear function:





◆ Sigmoid function

$$\psi_\alpha(v) = \frac{1}{1+\exp(-\alpha v)}$$



$a \to \infty$: step function

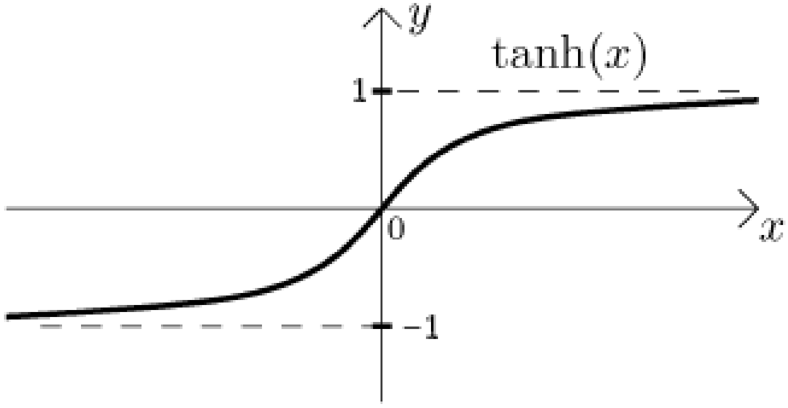
Activation function with negative values

- Threshold function & piecewise linear function:

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

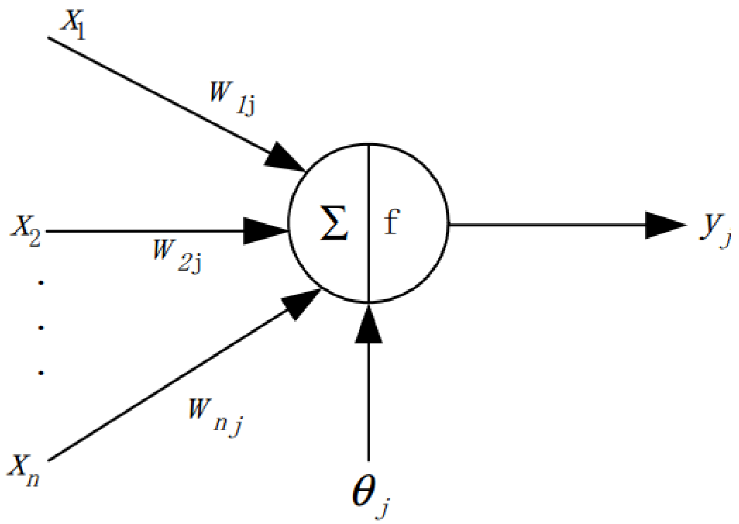
- Hyperbolic tangent function

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$



McCulloch & Pitts's Artificial Neuron

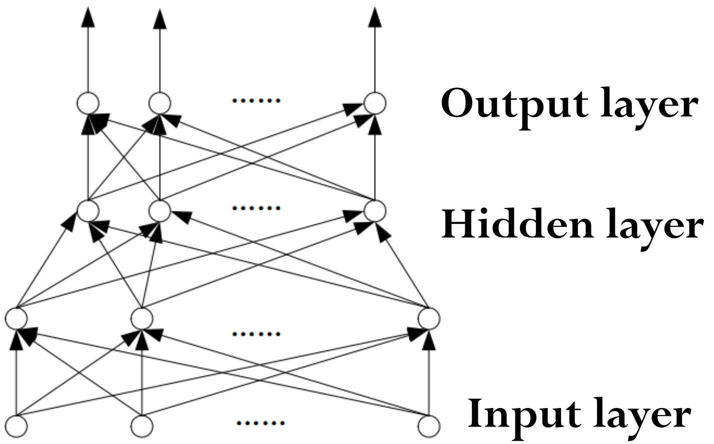
- The first model of artificial neurons in 1943
  - Activation function: a threshold function



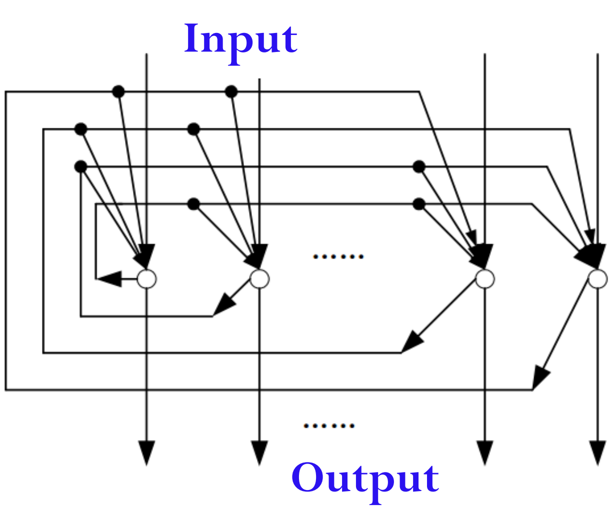
$$y_j = \operatorname{sgn}\left(\sum_i w_{ij} x_i - \theta_j \right)$$

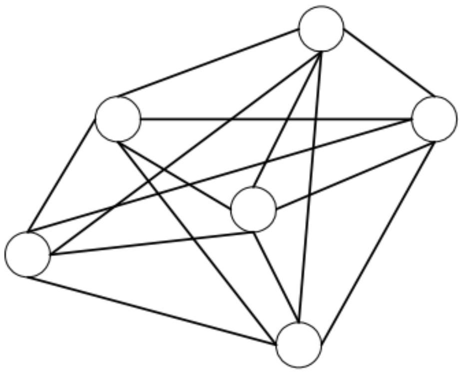
Network Architecture

- Feedforward networks



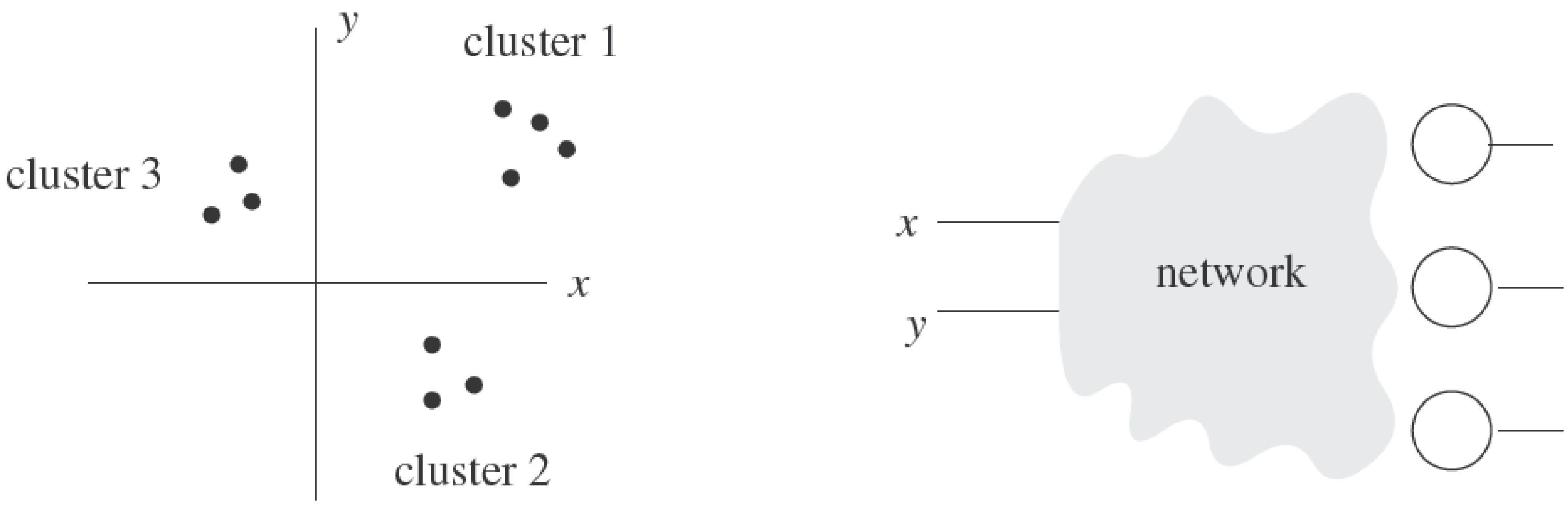
- Recurrent networks





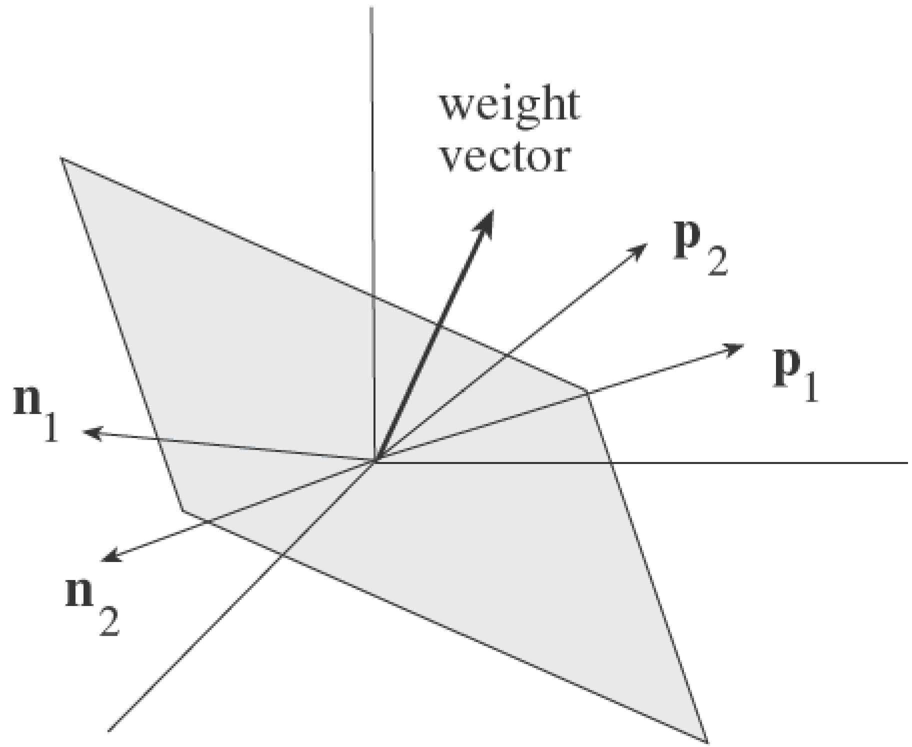
Learning Paradigms

- Unsupervised learning (learning without a teacher)
  - Example: clustering



Learning Paradigms

- Supervised Learning (learning with a teacher)
  - For example, classification: learns a separation plane

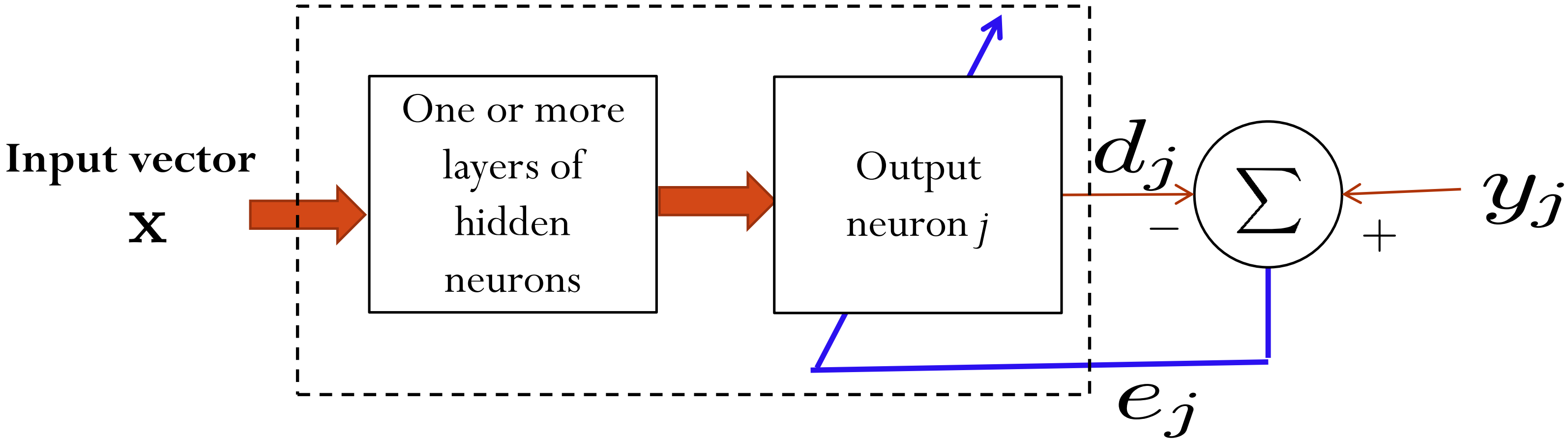


Learning Rules

- Error-correction learning
- Competitive learning
- Hebbian learning
- Boltzmann learning
- Memory-base learning
  - Nearest neighbor, radial-basis function network

Error-correction learning

- The generic paradigm:



- Error signal:

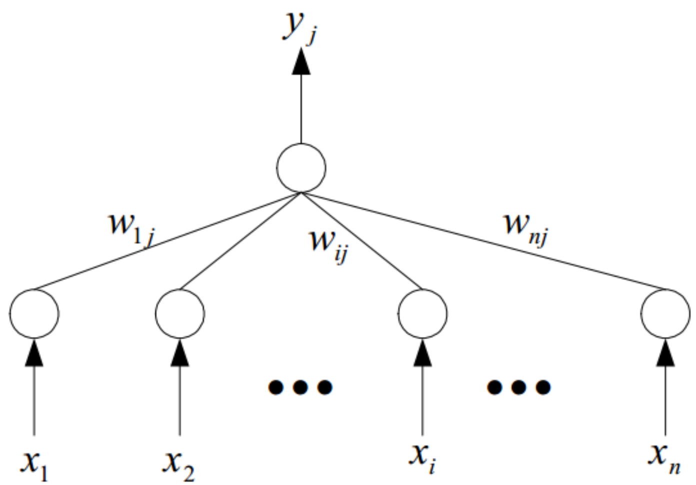
$$e_j = y_j - d_j$$

- Learning objective:

$$\min_{\mathbf{w}} R(\mathbf{w}; \mathbf{x}) := \frac{1}{2} \sum_j e_j^2$$

Example: Perceptron

- One-layer feedforward network based on error-correction learning (no hidden layer):



  - Current output (at iteration t):

$$d_j = (\mathbf{w}_t^j)^\top \mathbf{x}$$

  - Update rule (*exercise?*):

$$\mathbf{w}_{t+1}^j = \mathbf{w}_t^j + \eta(y_j - d_j)\mathbf{x}$$

Perceptron for classification

- Consider a single output neuron
- Binary labels:

$$y \in \{+1, -1\}$$

- Output function:

$$d = \mathrm{sgn}\left(\mathbf{w}_t^\top \mathbf{x}\right)$$

- Apply the error-correction learning rule, we get … (next slide)

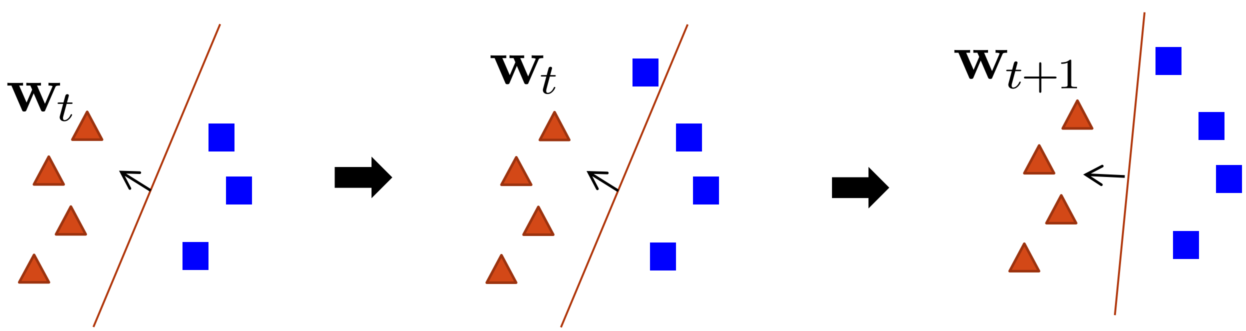
Perceptron for Classification

- Set $\mathbf{w}_1 = 0$ and t=1; scale all examples to have length 1 (doesn't affect which side of the plane they are on)
- Given example x, predict positive *iff*

$$\mathbf{w}_t^\top \mathbf{x} > 0$$

- If a mistake, update as follows
  - Mistake on positive: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta_t \mathbf{x}$
  - Mistake on negative: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{x}$

$t \leftarrow t + 1$



Convergence Theorem

- For linearly separable case, the perceptron algorithm will converge in a finite number of steps

Mistake Bound

- Theorem:
  - *Let $\mathcal{S}$ be a sequence of labeled examples consistent with a linear threshold function* $\mathbf{w}_*^\top \mathbf{x} > 0$ *, where* $\mathbf{w}_*$ *is a unit-length vector.*
  - *The number of mistakes made by the online Perceptron algorithm is at most* $(1/\gamma)^2$*, where*

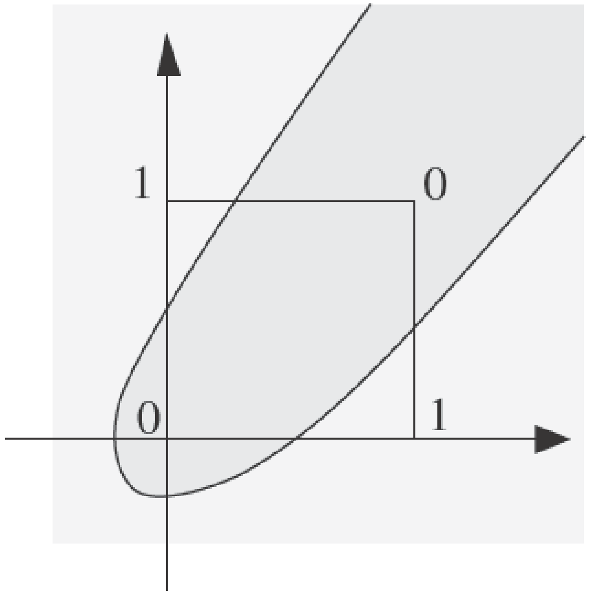
$$\gamma = \min_{\mathbf{x} \in \mathcal{S}} \frac{|\mathbf{w}_*^\top \mathbf{x}|}{\|\mathbf{x}\|}$$

  - *i.e.: if we scale examples to have length 1, then* γ *is the minimum distance of any example to the plane* $\mathbf{w}_*^\top \mathbf{x} = 0$
  - γ *is often called the "margin" of* $\mathbf{w}_*$ *; the quantity* $\frac{\mathbf{w}_*^\top \mathbf{x}}{\|\mathbf{x}\|}$ *is the cosine of the angle between* $\mathbf{x}$ *and* $\mathbf{w}_*$

Deep Nets

- Multi-layer Perceptron
- CNN
- Auto-encoder
- RBM
- Deep belief nets
- Deep recurrent nets

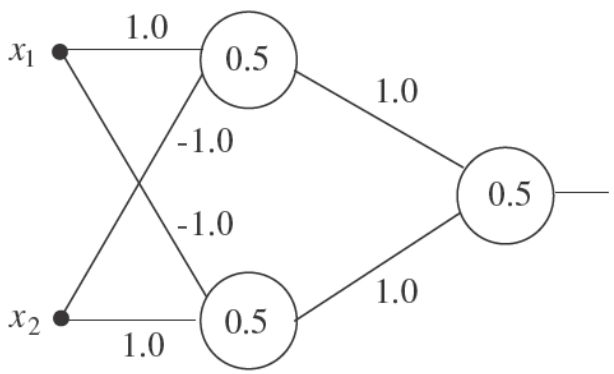
XOR Problem



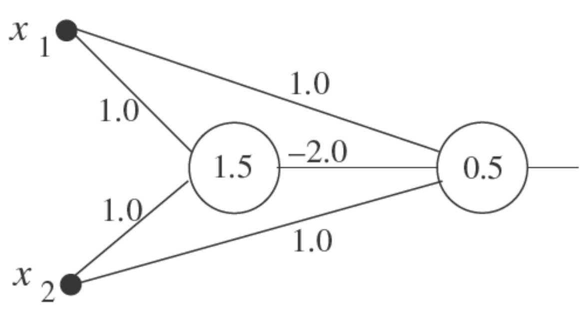
- Single-layer perceptron can’t solve the problem

XOR Problem

- A network with 1-layer of 2 neurons works for XOR:
  - threshold activation function



  - Many alternative networks exist (not layered)



Multilayer Perceptrons

- Computatonal limitations of single-layer Perceptron by Minsky & Papert (1969)

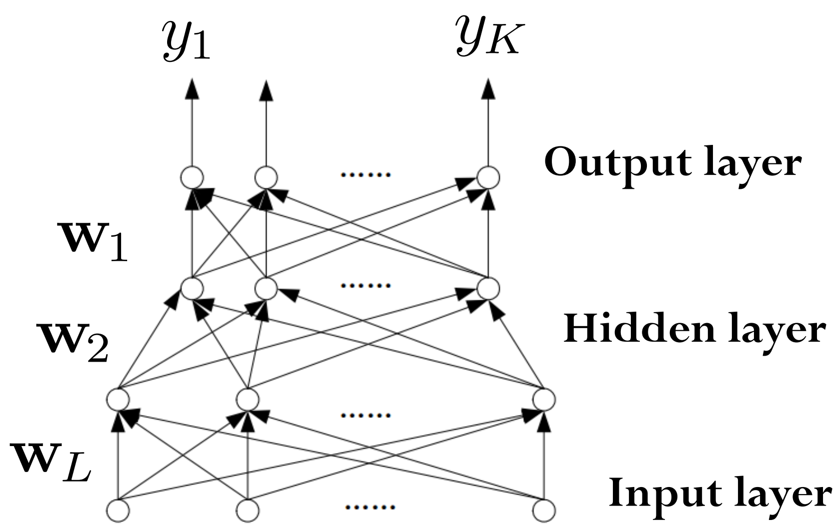
- Multilayer Perceptrons:
  - Multilayer feedforward networks with an error-correction learning algorithm, known as error ***back-propagation***
  - A generalization of single-layer percetron to allow nonlinearity

Backpropagation

- Learning as loss minimization

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2} \sum_j e_j^2(\mathbf{x})$$

$$e_j = y_j - d_j$$



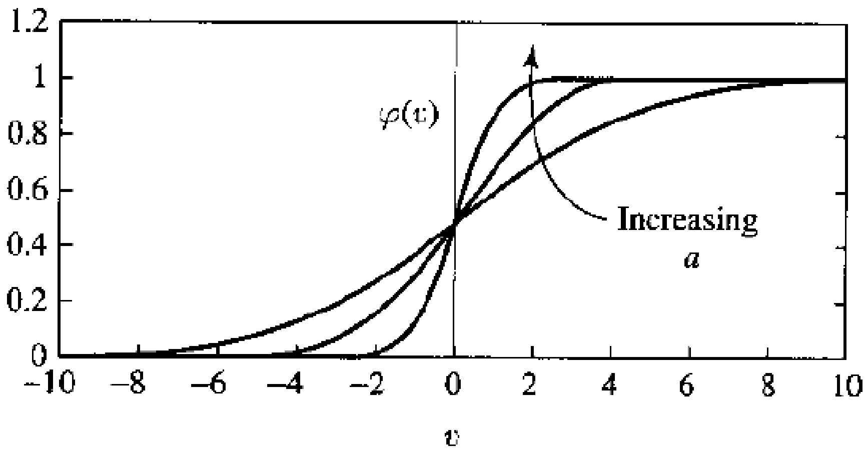
- Learning with gradient descent

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \lambda_t \nabla R(\mathbf{w}; \mathcal{D})$$

Backpropagation

- Step function in perceptrons is non-differentiable
- Differentiable activation functions are needed to calculate gradients, e.g., sigmoid:

$$\psi_\alpha(v) = \frac{1}{1 + \exp(-\alpha v)}$$



Backpropagation

- Derivative of a sigmoid function ($\alpha = 1$)

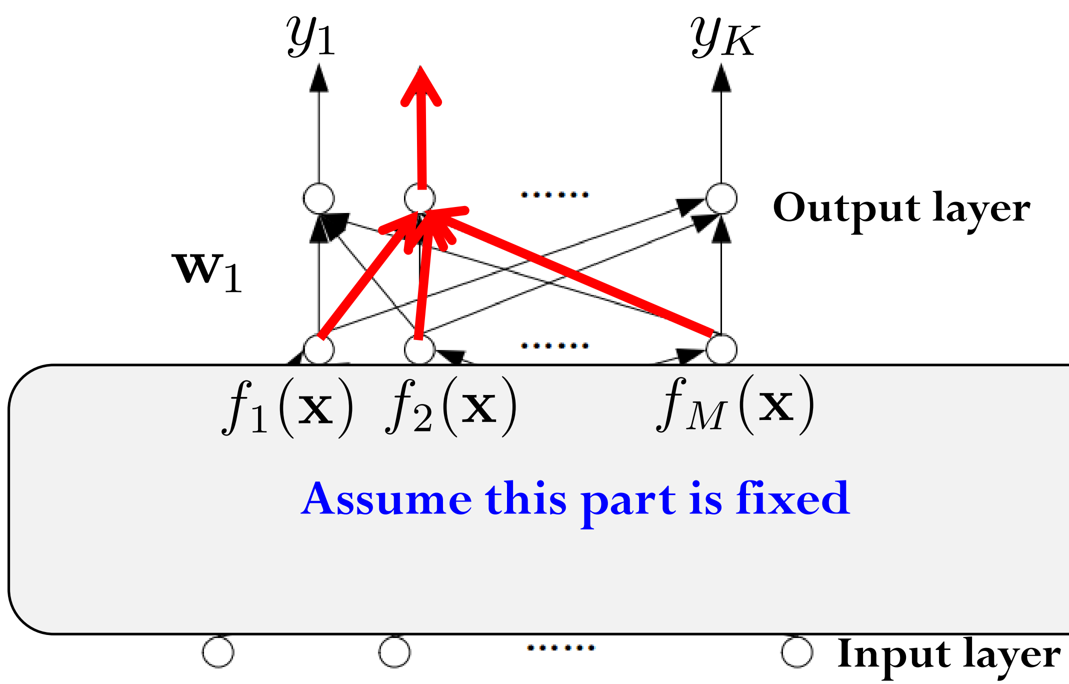
$$\nabla_v \psi(v) = \frac{e^{-v}}{(1+e^{-v})^2} = \psi(v)(1-\psi(v))$$

  - Notice about the small scale of the gradient
  - Gradient vanishing issue

- Many other activation functions examined

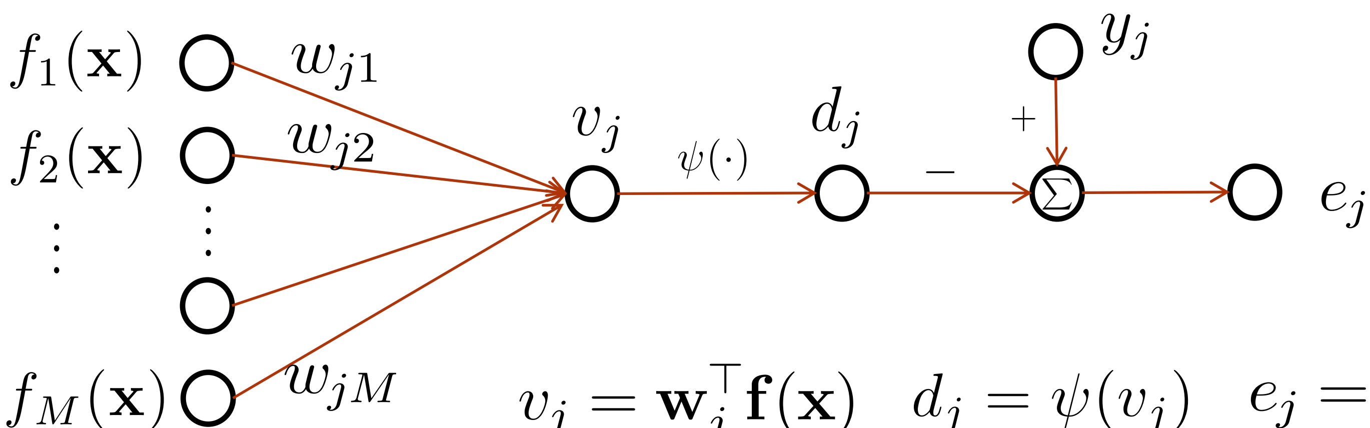
Gradient computation at output layer

- Output neurons are separate:



Gradient computation at output layer

- Signal flow:



$$v_j = \mathbf{w}_j^\top \mathbf{f}(\mathbf{x}) \quad d_j = \psi(v_j) \quad e_j = y_j - d_j$$

$$R_j = \frac{1}{2} e_j^2$$

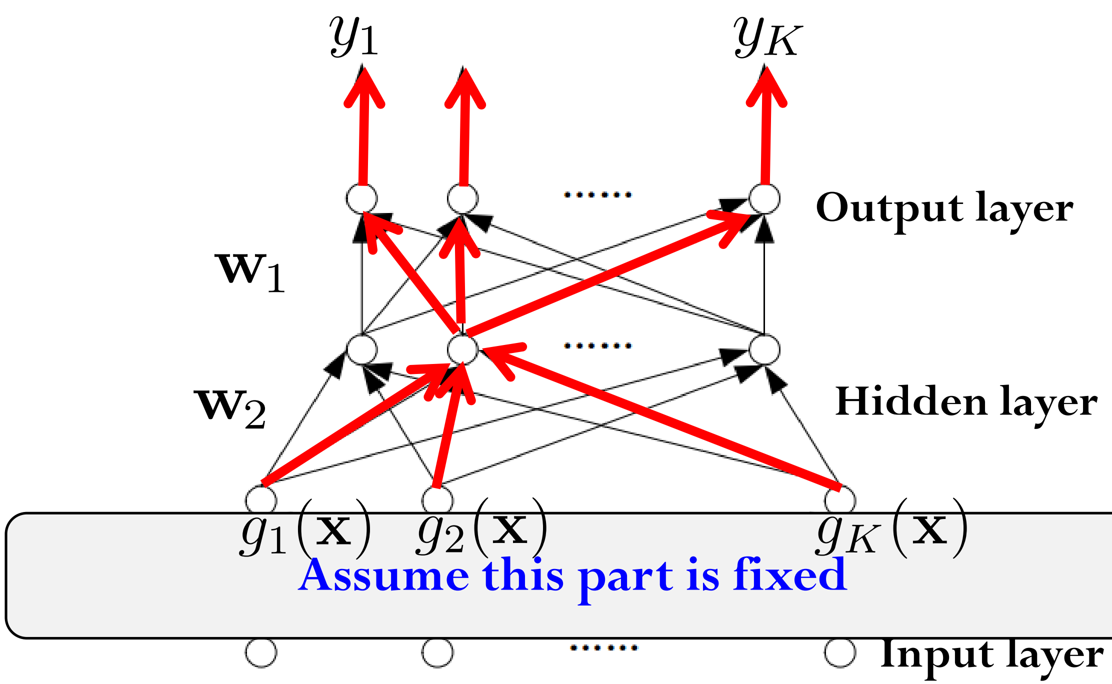
$$R = \frac{1}{2} \sum_j e_j^2$$

$$\begin{aligned}\nabla_{w_{ji}} R &= \frac{\partial R_j}{\partial e_j} \frac{\partial e_j}{\partial d_j} \frac{\partial d_j}{\partial v_j} \frac{\partial v_j}{\partial w_{ji}} \\ &= e_j \cdot (-1) \cdot \psi'(v_j) \cdot f_i(\mathbf{x}) \\ &= -e_j \psi'(v_j) f_i(\mathbf{x})\end{aligned}$$

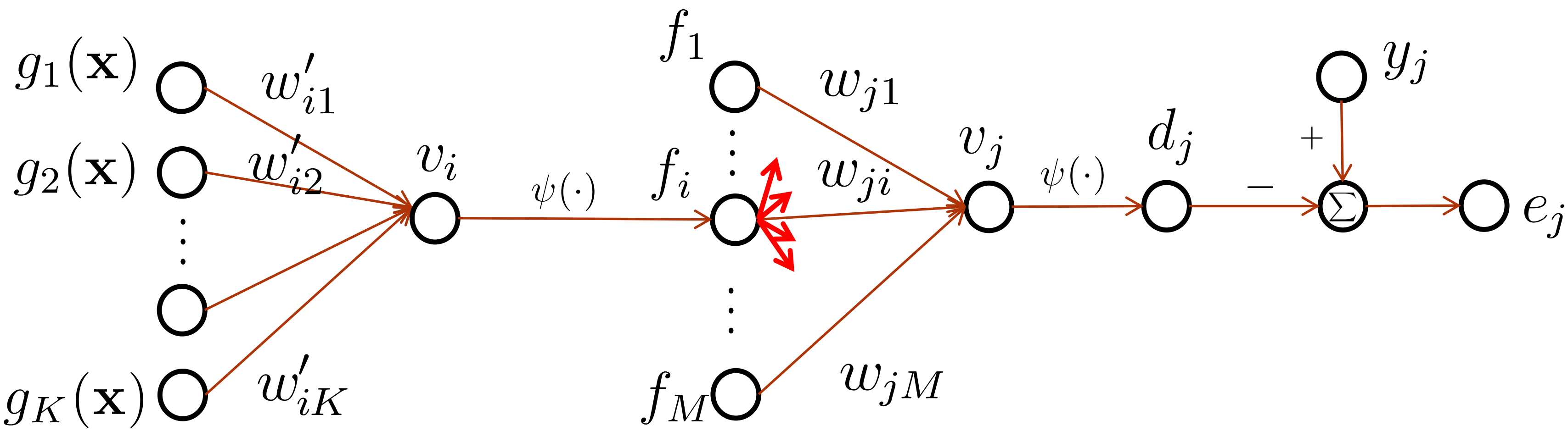
Local gradient: $\delta_j = -\frac{\partial R}{\partial v_j}$

Gradient computation at hidden layer

- Output neurons are NOT separate:



Gradient computation at hidden layer



$v_i = (\mathbf{w}'_i)^\top \mathbf{g} \quad f_i = \psi(v_i) \quad v_j = \mathbf{w}_j^\top \mathbf{f} \quad d_j = \psi(v_j) \quad e_j = y_j - d_j$

$$R_j = \frac{1}{2} e_j^2$$

$$R = \frac{1}{2} \sum_j e_j^2$$

$$\nabla_{w'_{ik}} R = \sum_j \frac{\partial R_j}{\partial e_j} \frac{\partial e_j}{\partial d_j} \frac{\partial d_j}{\partial v_j} \frac{\partial v_j}{\partial f_i} \frac{\partial f_i}{\partial v_i} \frac{\partial v_i}{\partial w'_{ik}}$$

$$= -\sum_j e_j \psi'(v_j) w_{ji} \psi'(v_i) g_k(\mathbf{x})$$

$$= -\sum_j \delta_j w_{ji} \psi'(v_i) g_k(\mathbf{x})$$

Local gradient:

$$\delta_i = -\frac{\partial R}{\partial v_i}$$

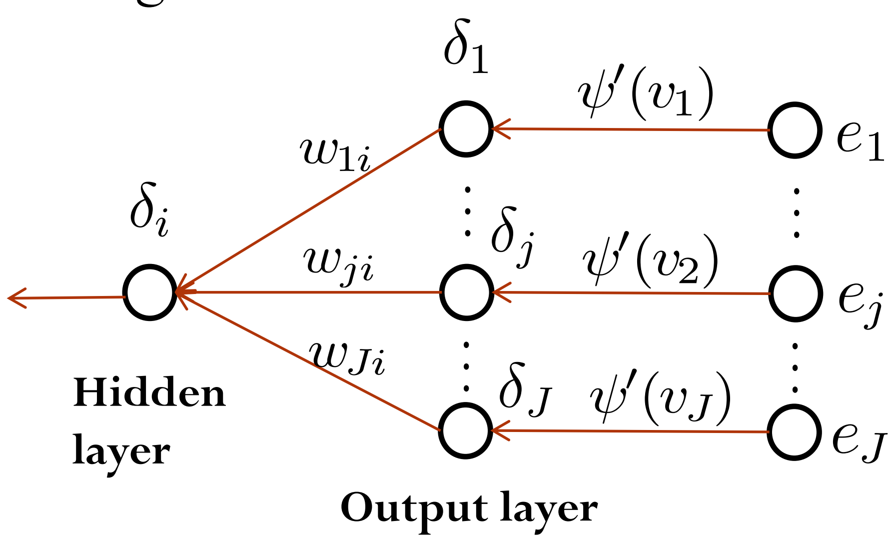
Back-propagation formula

- The update rule of **local gradients**:
  - for hidden neuron *i*:

$$\delta_i = \psi'(v_i) \sum_j \delta_j w_{ji}$$

Only depends on the activation function at hidden neuron i

- Flow of error signal:



Back-propagation formula

- The update rule of weights:
  - Output neuron:

$$\Delta w_{ji} = \lambda \cdot \delta_j \cdot f_i(\mathbf{x})$$

  - Hidden neuron:

$$\Delta w'_{ik} = \lambda \cdot \delta_i \cdot g_k(\mathbf{x})$$

$$\begin{pmatrix} Weight \\ correction \\ \Delta w_{ji} \end{pmatrix} = \begin{pmatrix} learning \\ rate \\ \lambda \end{pmatrix} \cdot \begin{pmatrix} local \\ gradient \\ \delta_j \end{pmatrix} \cdot \begin{pmatrix} input\ signal \\ of\ neuron\ j \\ v_i \end{pmatrix}$$

Two Passes of Computation

- Forward pass
  - Weights fixed
  - Start at the first hidden layer
  - Compute the output of each neuron
  - End at output layer

- Backward pass
  - Start at the output layer
  - Pass error signal backward through the network
  - Compute local gradients

Stopping Criterion

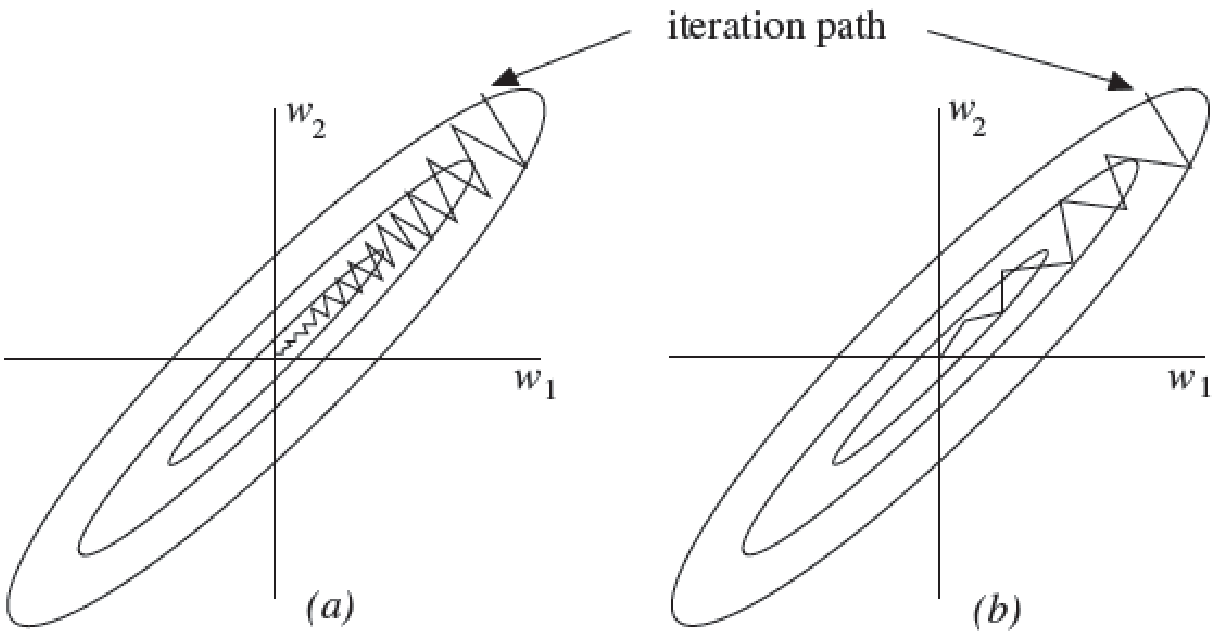
- No general rules

- Some reasonable heuristics:
  - The norm of gradient is small enough
  - The number of iterations is larger than a threshold
  - The training error is stable
  - …

Improve Backpropagation

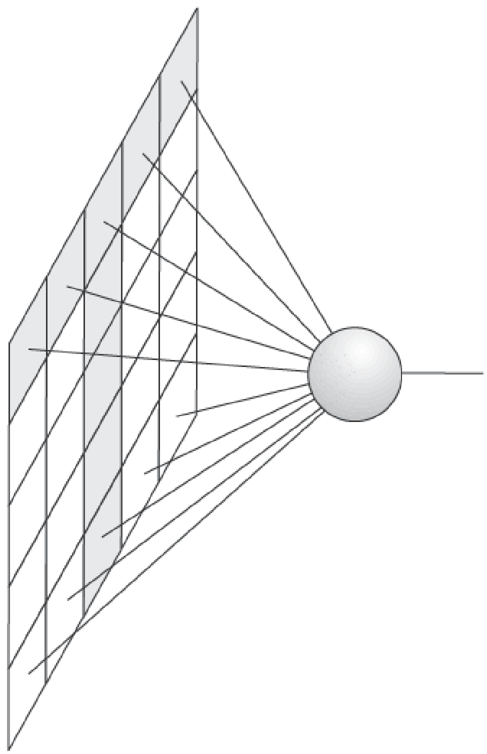
- Many methods exist to improve backpropagation
- E.g., backpropagation with momentum

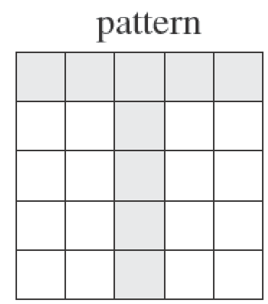
$$\Delta w_{ij}^{t} = -\lambda \frac{\partial R}{\partial w_{ij}} + \alpha \Delta w_{ij}^{t-1}$$



Neurons as Feature Extractor

- Compute the similarity of a pattern to the ideal pattern of a neuron
- Threshold is the minimal similarity required for a pattern
- Reversely, it visualizes the connections of a neuron





weights

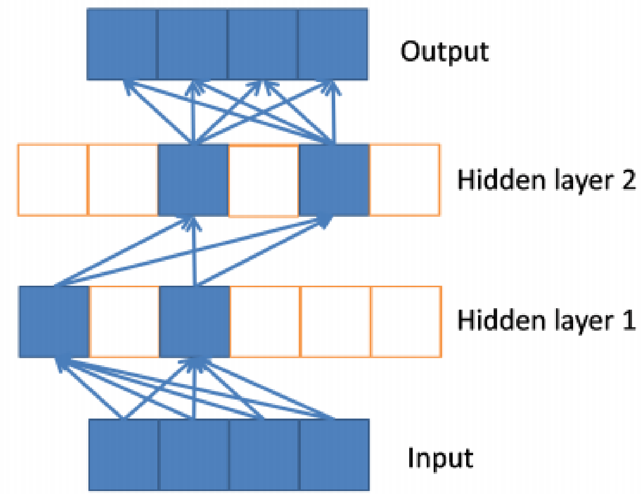
1	1	1	1	1
–1	–1	1	–1	–1
–1	–1	1	–1	–1
–1	–1	1	–1	–1
–1	–1	1	–1	–1

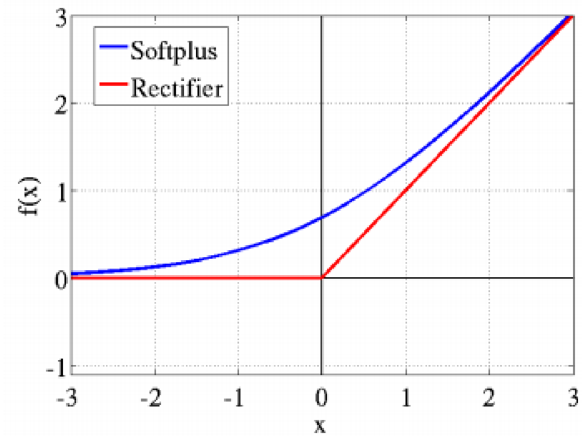
Vanishing gradient problem

- The gradient can decrease exponentially during back-prop
- Solutions:
  - Pre-training + fine tuning
  - Rectifier neurons (sparse gradients)
- Ref:
  - Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. Hochreiter, Bengio, & Frasconi, 2001

Deep Rectifier Nets

- Sparse representations without gradient vanishing





- Non-linearity comes from the path selection
  - Only a subset of neurons are active for a given input
- Can been seen as a model with an exponential number of linear models that share weights

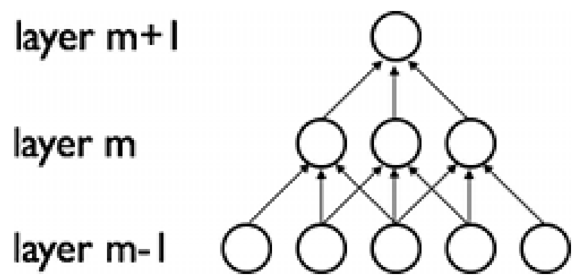
[Deep sparse rectifier neural networks. Glorot, Bordes, & Bengio, 2011]

CNN

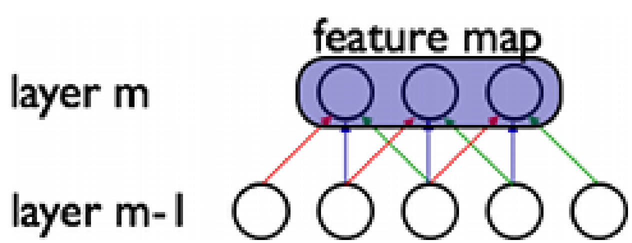
- Hubel and Wiesel's study on annimal's visual cortex:
  - Cells that are sensitive to small sub-regions of the visual field, called a *receptive field*
  - Simple cells respond maximally to specific edge-like patterns within their receptive field. Complex cells have larger receptive fields and are locally invariant to the exact position of the pattern.

Convolutional Neural Networks

- Sparse local connections (spatially contiguous receptive fields)

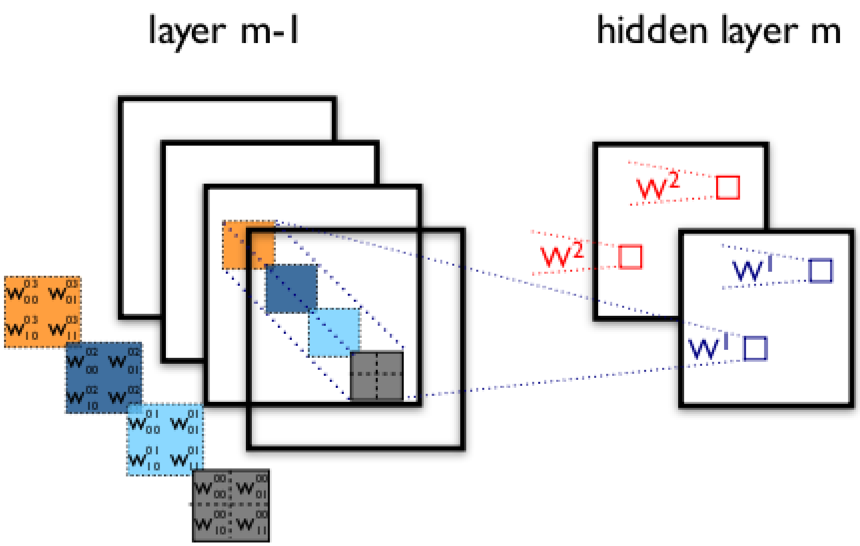


- Shared weights: each filter is replicated across the entire visual field, forming a feature map



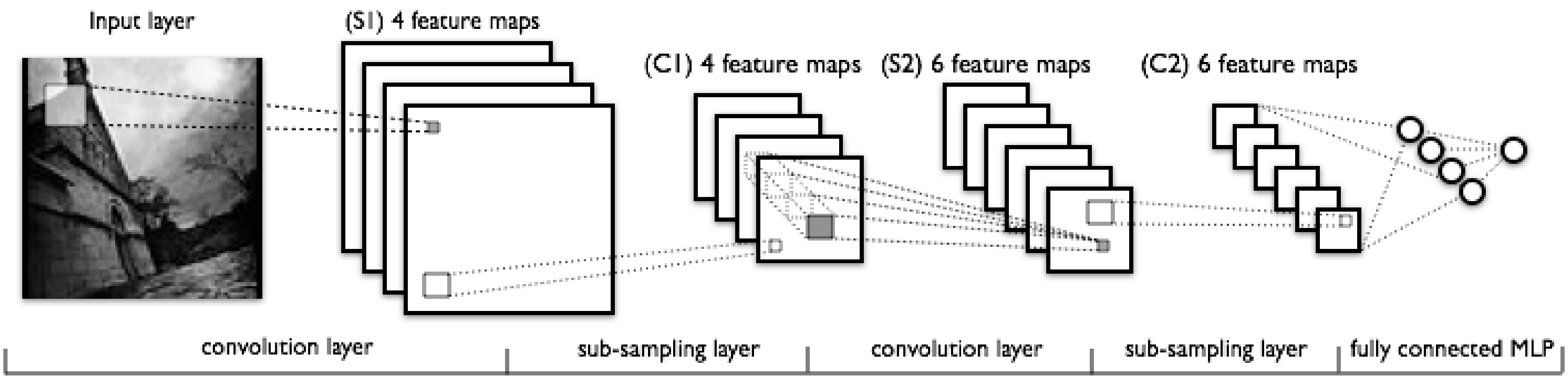
CNN

- Each layer has multiple feature maps



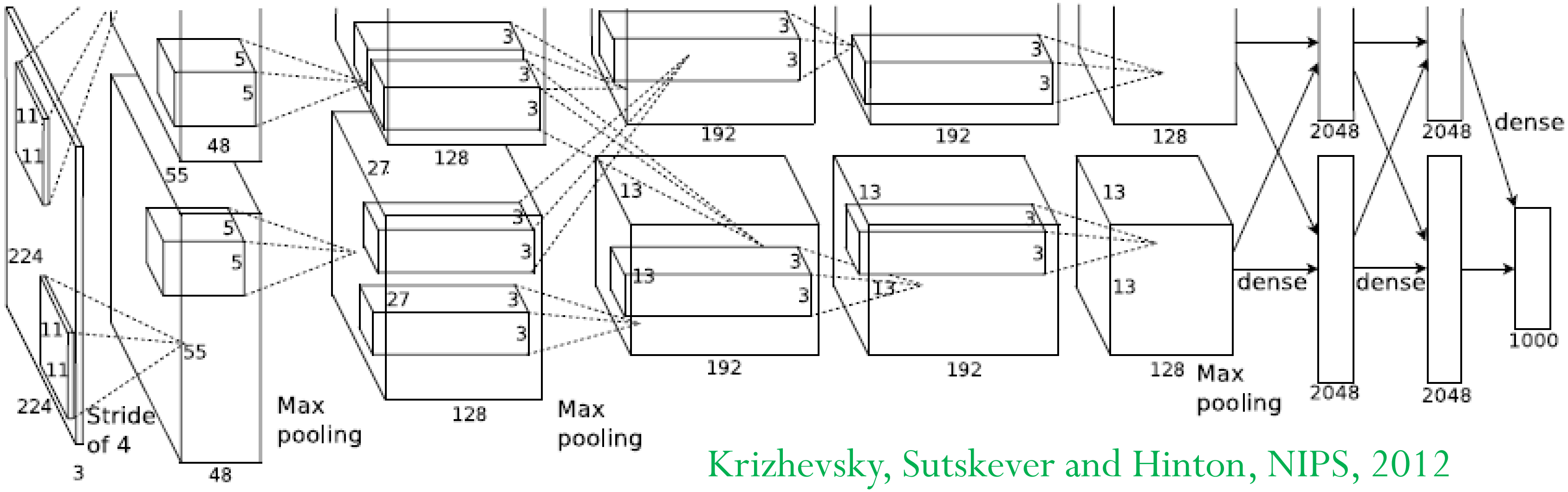
CNN

- The full model



- ***Max-pooling***, a form of non-linear down-sampling.
  - Max-pooling partitions the input image into a set of non-overlapping rectangles and, for each such sub-region, outputs the maximum value.

Example: CNN for image classification



Krizhevsky, Sutskever and Hinton, NIPS, 2012

- Network dimension: 150,528(input)-253,440–186,624–64,896–64,896–43,264–4096–4096–1000(output)
  - In total: 60 million parameters
  - Task: classify 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes
  - Results: state-of-the-art accuracy on ImageNet

IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenge New People Publication About

Issues with CNN

- Computing the activations of a single convolutional filter is much more expensive than with traditional MLPs

- Many tuning parameters
  - # of filters:
    - Model complexity issue (overfitting vs underfitting)
  - Filter shape:
    - the right level of "granularity" in order to create abstractions at the proper scale, given a particular dataset
    - Usually 5x5 for MNIST at 1st layer
  - Max-pooling shape:
    - typical: 2x2; maybe 4x4 for large images

Auto-Encoder

- Encoder: (a distributed code)

$$\mathbf{y} = s(\mathbf{W}\mathbf{x} + \mathbf{b})$$

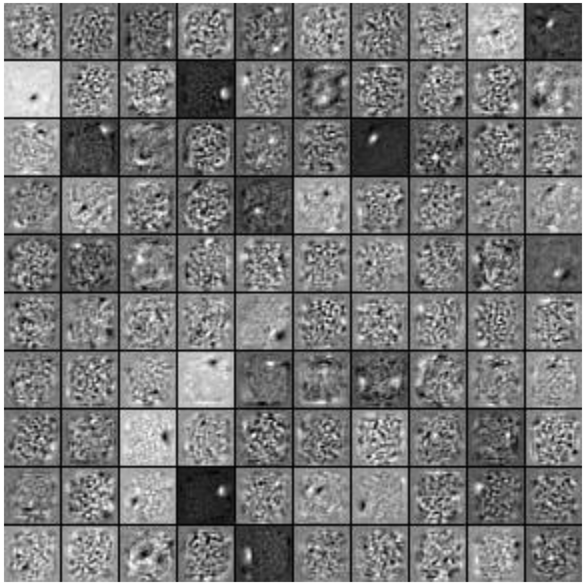
- Decoder:

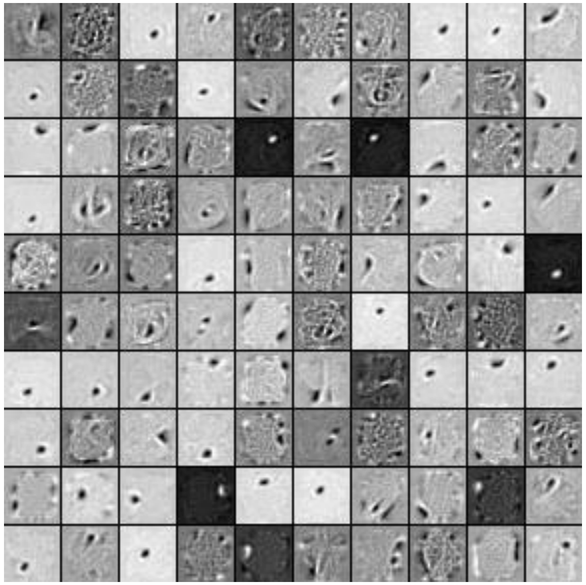
$$\mathbf{z} = s(\mathbf{W}'\mathbf{y} + \mathbf{b}')$$

- Minimize reconstruction error

- Connection to PCA
  - PCA is linear projection, which Auto-Encoder is nonlinear
  - Stacking PCA with nonlinear processing may perform as well (MaYi's work)
- Denoising Auto-Encoder
  - A stochastic version with corrupted noise to discover more robust features
  - E.g., randomly set some inputs to zero

- Left: no noise; right: 30 percent noise



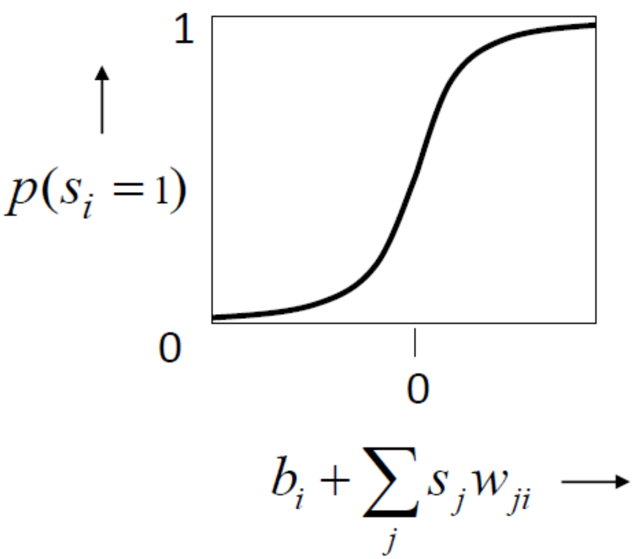


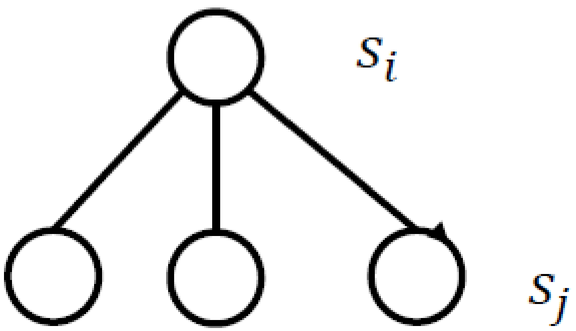
Deep Generative Models

Stochastic Binary Units

- Each unit has a state of 0 or 1
- The probability of turning on is determined by

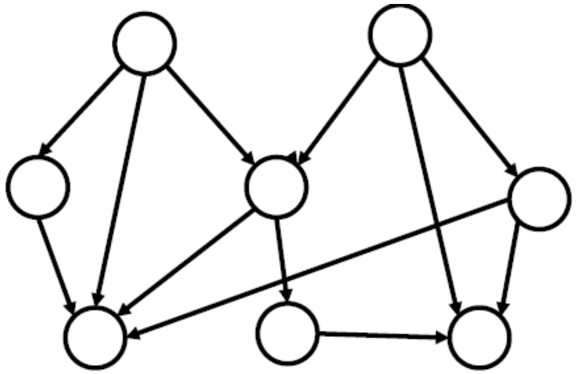
$$p(s_i = 1) = \frac{1}{1 + \exp(-b_i - \sum_j s_j w_{ji})}$$



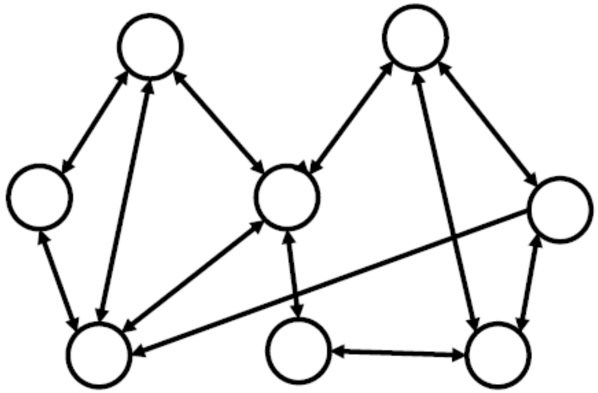


Generative Models

- Directed acyclic graph with stochastic binary units is termed Sigmoid Belief Net (Radford Neal, 1992)

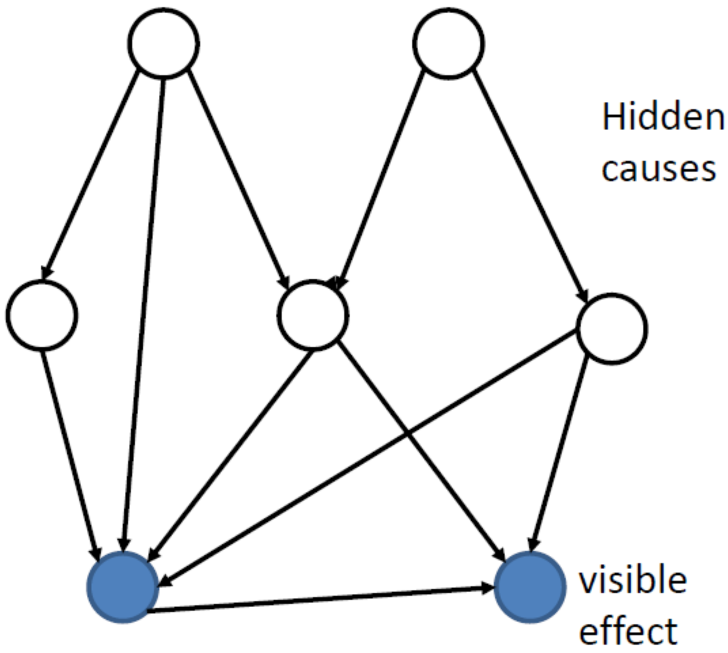


- Undirected graph with stochastic binary units is termed Boltzmann Machine (Hinton & Sejnowski, 1983)



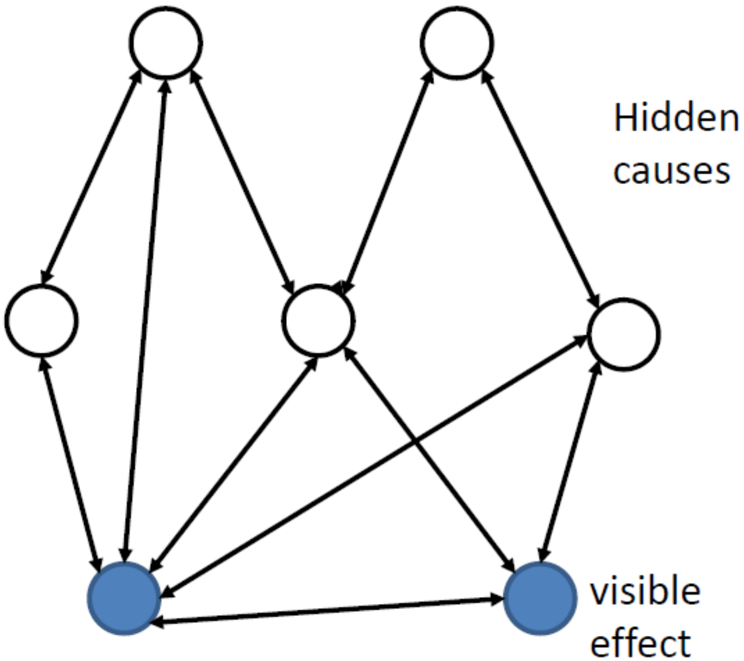
Learning Deep Belief Nets

- Easy to generate an unbiased example at the leaf nodes
- Hard to infer the posterior distribution over all possible configurations of hidden causes – explain away effect!
  - Hard to even get a sample from the posterior



Learning Boltzmann Machine

- Hard to generate an unbiased example for the visible units
- Hard to infer the posterior distribution over all possible configurations of hidden causes
  - Hard to even get a sample from the posterior

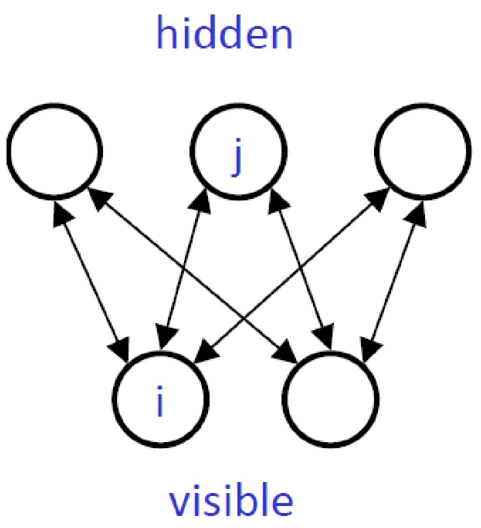


Restricted Boltzmann Machines

- An energy-based model with hidden units

$$P(x) = \sum_h P(x,h) = \sum_h \frac{e^{-E(x,h)}}{Z}.$$

- Graphical structure:



$$E(v,h) = -b'v - c'h - h'Wv$$

- Restrict the connectivity to make learning easier.

Restricted Boltzmann Machines

◈ Factorized conditional distribution over hidden units

$$
\begin{aligned}
p(\boldsymbol{h} \mid \boldsymbol{v}) &= \frac{1}{p(\boldsymbol{v})} \frac{1}{Z} \exp\left\{\boldsymbol{b}^\top \boldsymbol{v} + \boldsymbol{c}^\top \boldsymbol{h} + \boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{h}\right\} \\
&= \frac{1}{Z'} \exp\left\{\boldsymbol{c}^\top \boldsymbol{h} + \boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{h}\right\} \\
&= \frac{1}{Z'} \exp\left\{\sum_{j=1}^{n} c_j h_j + \sum_{j=1}^{n} \boldsymbol{v}^\top \boldsymbol{W}_{:,j} \boldsymbol{h}_j\right\} \\
&= \frac{1}{Z'} \prod_{j=1}^{n} \exp\left\{c_j h_j + \boldsymbol{v}^\top \boldsymbol{W}_{:,j} \boldsymbol{h}_j\right\}
\end{aligned}
$$

Restricted Boltzmann Machines

- For Gibbs sampling
  - Hidden units:

$$P(h_j = 1 \mid \boldsymbol{v}) = \frac{\tilde{P}(h_j = 1 \mid \boldsymbol{v})}{\tilde{P}(h_j = 0 \mid \boldsymbol{v}) + \tilde{P}(h_j = 1 \mid \boldsymbol{v})}$$
$$= \frac{\exp\left\{c_j + \boldsymbol{v}^\top \boldsymbol{W}_{:,j}\right\}}{\exp\{0\} + \exp\left\{c_j + \boldsymbol{v}^\top \boldsymbol{W}_{:,j}\right\}}$$
$$= \text{sigmoid}\left(c_j + \boldsymbol{v}^\top \boldsymbol{W}_{:,j}\right).$$

  - Observed units:

$$P(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_{i=1}^{d} \text{sigmoid}\left(b_i + \boldsymbol{W}_{i,:}\boldsymbol{h}\right)$$

MLE

- Log-likelihood

$$\ell(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{t=1}^{n} \log P(\boldsymbol{v}^{(t)})$$

$$= \sum_{t=1}^{n} \log \sum_{\boldsymbol{h}} \exp\left\{-E(\boldsymbol{v}^{(t)}, \boldsymbol{h})\right\} - n \log Z$$

- Gradient

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} \sum_{t=1}^{n} \log \sum_{\boldsymbol{h}} \exp\left\{-E(\boldsymbol{v}^{(t)}, \boldsymbol{h})\right\} - n \frac{\partial}{\partial \theta} \log \sum_{\boldsymbol{v}, \boldsymbol{h}} \exp\{-E(\boldsymbol{v}, \boldsymbol{h})\}$$

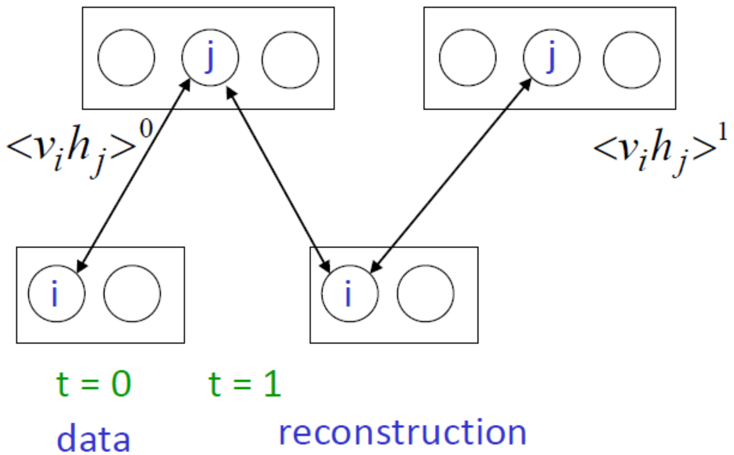
$$= \sum_{t=1}^{n} \frac{\sum_{\boldsymbol{h}} \exp\left\{-E(\boldsymbol{v}^{(t)}, \boldsymbol{h})\right\} \frac{\partial}{\partial \theta} - E(\boldsymbol{v}^{(t)}, \boldsymbol{h})}{\sum_{\boldsymbol{h}} \exp\left\{-E(\boldsymbol{v}^{(t)}, \boldsymbol{h})\right\}}$$

$$- n \frac{\sum_{\boldsymbol{v}, \boldsymbol{h}} \exp\{-E(\boldsymbol{v}, \boldsymbol{h})\} \frac{\partial}{\partial \theta} - E(\boldsymbol{v}, \boldsymbol{h})}{\sum_{\boldsymbol{v}, \boldsymbol{h}} \exp\{-E(\boldsymbol{v}, \boldsymbol{h})\}}$$

$$= \sum_{t=1}^{n} \mathbb{E}_{P(\boldsymbol{h}|\boldsymbol{v}^{(t)})}\left[\frac{\partial}{\partial \theta} - E(\boldsymbol{v}^{(t)}, \boldsymbol{h})\right] - n \mathbb{E}_{P(\boldsymbol{v}, \boldsymbol{h})}\left[\frac{\partial}{\partial \theta} - E(\boldsymbol{v}, \boldsymbol{h})\right]$$

Contrastive Divergence (CD)

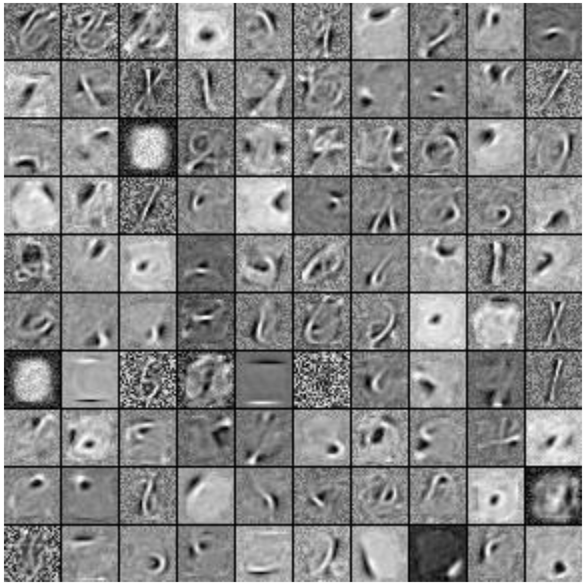
- Gibbs sampling for negative phase
  - Random initialization: v'->h'-> … -> v -> h
  - Slow because of long burn-in period
- Intuition of CD
  - Start from a data closed to the model samples
- CD-k for negative phase
  - Start from empirical data and ran k-steps
  - Typically, k=1: v1->h1->v2->h2



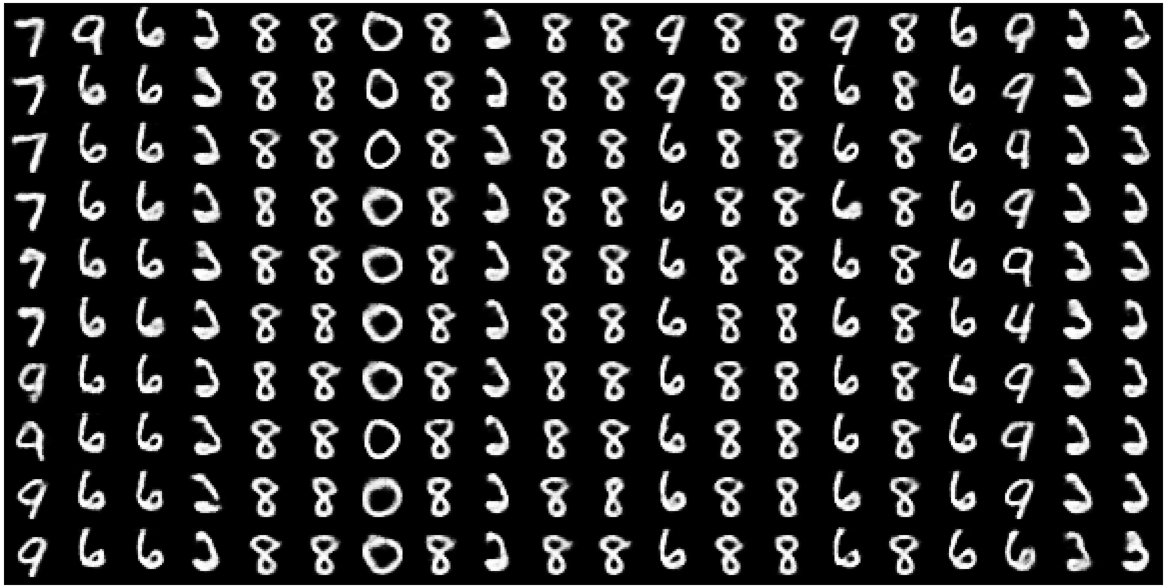
$$\Delta w_{ij} = \varepsilon \,(<v_i h_j>^0 - <v_i h_j>^1)$$

RBM

- Filters



- Samples (RBM is a generative model)

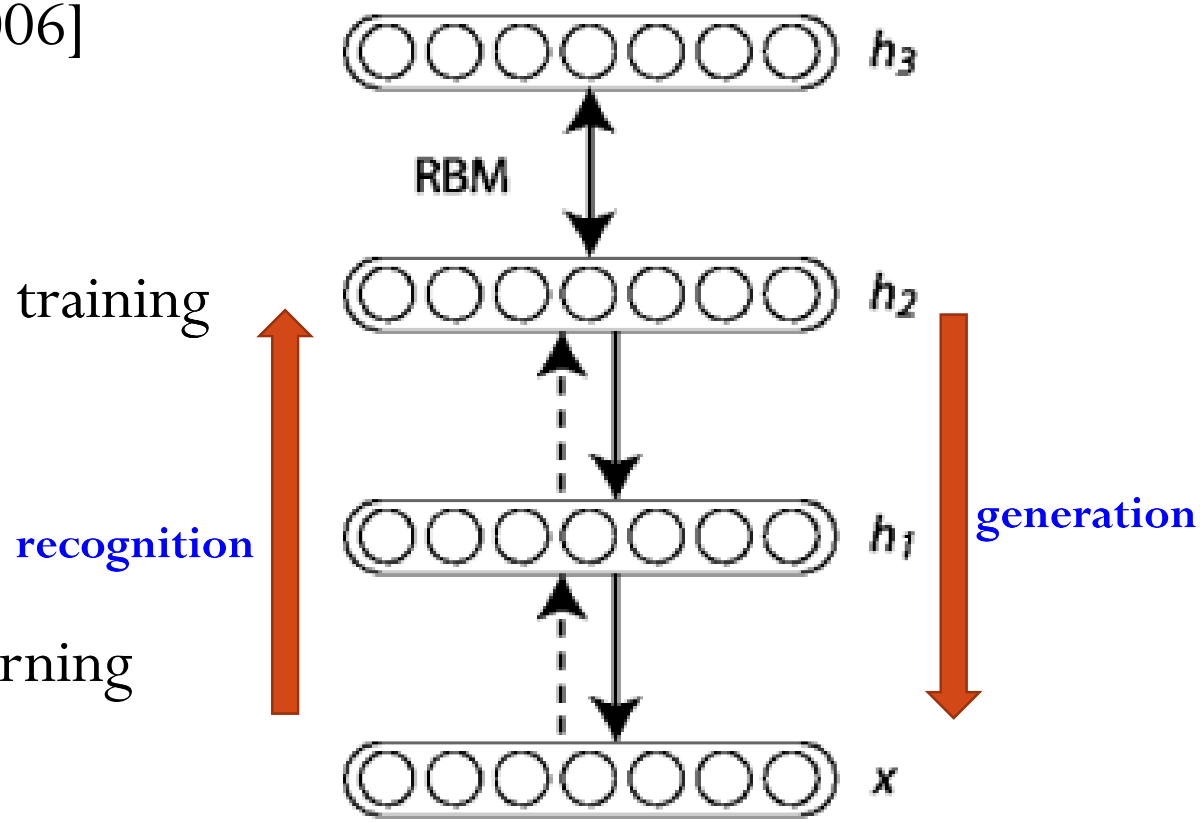


Issues with RBM

- Log-partition function is intractable
- No direct metric for choosing hyper-parameters
- (one hidden layer) Much too simple for modeling high-dimensional and richly structured sensory data

Deep Belief Nets – deep generative model

- [Hinton et al., 2006]
- Stacking RBM
- Greedy layerwise training
- Unsupervised learning
  - No labels
  - MLE



Neural Evidence?

- Our visual systems contain multilayer generative models

- Top-down connections:
  - Generate low-level features of images from high-level representations
  - Visual imagery, dreaming?

- Bottom-up connections:
  - Infer the high-level representations that would have generated an observed set of low-level features

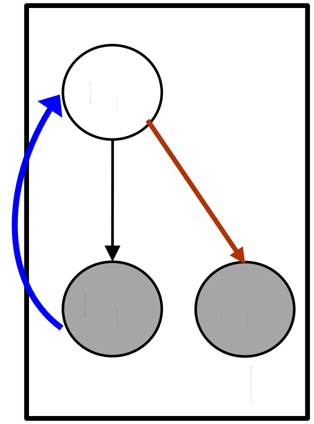
[Hinton, Trends in Cognitive Science, 11(10), 2007]

Recent Advances on DGMs

- Models:
  - Deep belief networks (Salakhutdinov & Hinton, 2009)
  - Autoregressive models (Larochelle & Murray, 2011; Gregor et al., 2014)
  - Stochastic variations of neural networks (Bengio et al., 2014)
  - …
- Applications:
  - Image recognition (Ranzato et al., 2011)
  - Inference of hidden object parts (Lee et al., 2009)
  - Semi-supervised learning (Kingma et al., 2014)
  - Multimodal learning (Srivastava & Salakhutdinov, 2014; Karpathy et al., 2014)
  - …
- Learning algorithms
  - Stochastic variational inference (Kingma & Welling, 2014; Rezende et al., 2014)
  - …

Learning with a Recognition Model

- Characterize the variational distribution with a recognition model



- For example:

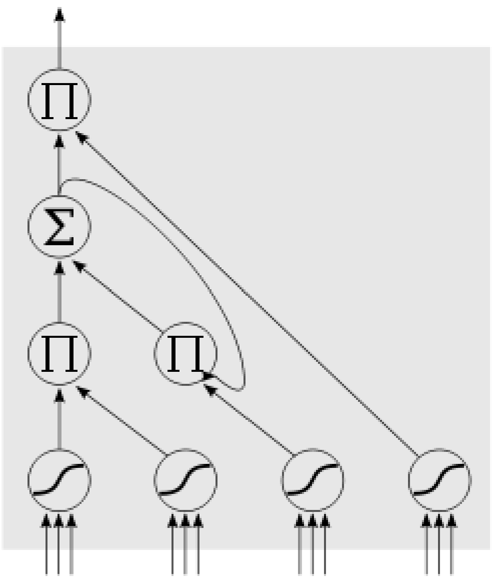
$$q_\phi(\mathbf{z}|\mathbf{x}, y) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}, y; \phi), \boldsymbol{\sigma}^2(\mathbf{x}, y; \phi))$$

  - where both mean and variance are nonlinear function of data by a DNN

Long Short-Term Memory

- A RNN architecture without gradient vanishing issue

- A RNN with LSTM blocks
  - Each block is a "smart" network, determing when to remember, when to continue to remember or forget, and when to output

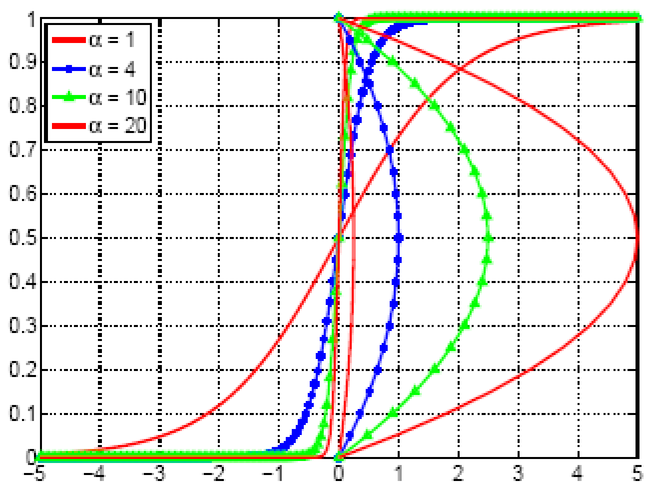


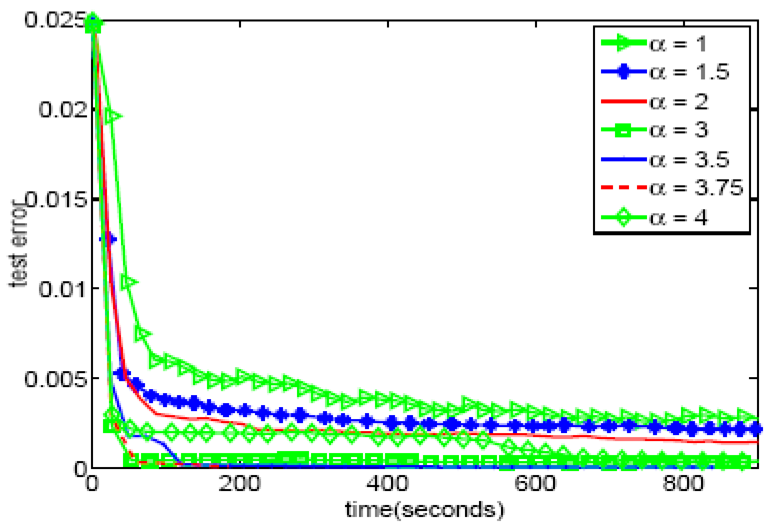
[Graves et al., 2009. A Novel Connectionist System for Improved Unconstrained Handwriting Recognition]

Issues

◈ The sharpness of Gates' activation functions maters!

$$f(x) = \frac{1}{1+e^{-\alpha * x}}$$





[Lv & Zhu, 2014. Revisit Long Short-Term Memory: an Optimization Perspective]

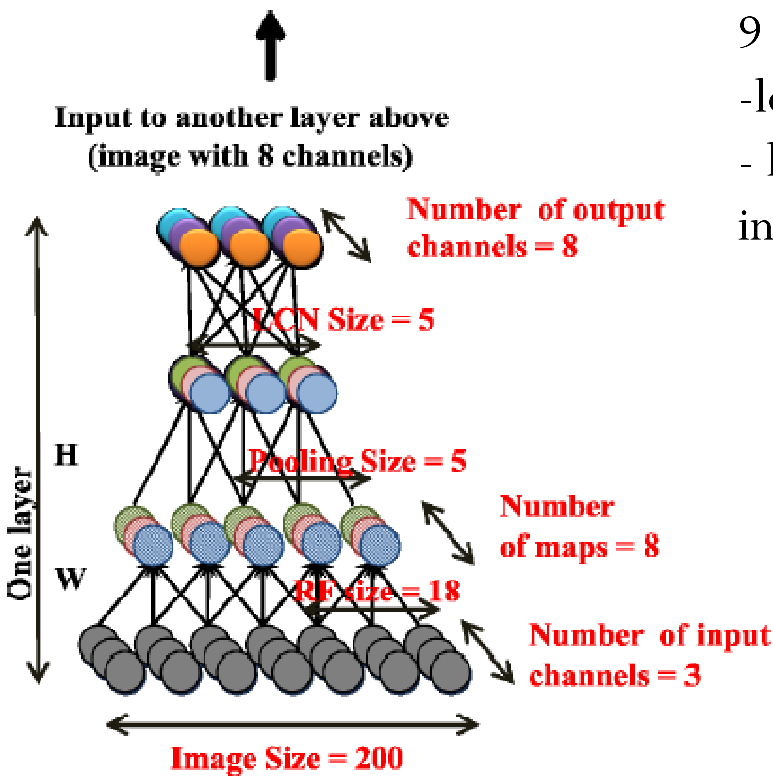
Discussions

Challenges of DL

- Learning
  - Backpropagation is slow and prone to gradient vanishing
  - Issues with non-convex optimization in high-dimensions
- Overfitting
  - Big models are lacking of statistical information to fit
- Interpretation
  - Deep nets are often used as black-box tools for learning and inference

Expensive to train

- "Big Model + Big Data + Big/Super Cluster"



9 layers sparse autoencoder with:
-local receptive fields to scale up;
- local L2 pooling and local contrast normalization for invariant features

- 1B parameters (connections)
- 10M 200x200 images
- train with 1K machines (16K cores) for 3 days

-able to build high-level concepts, e.g., cat faces and human bodies
-15.8% accuracy in recognizing 22K objects (70% relative improvements)

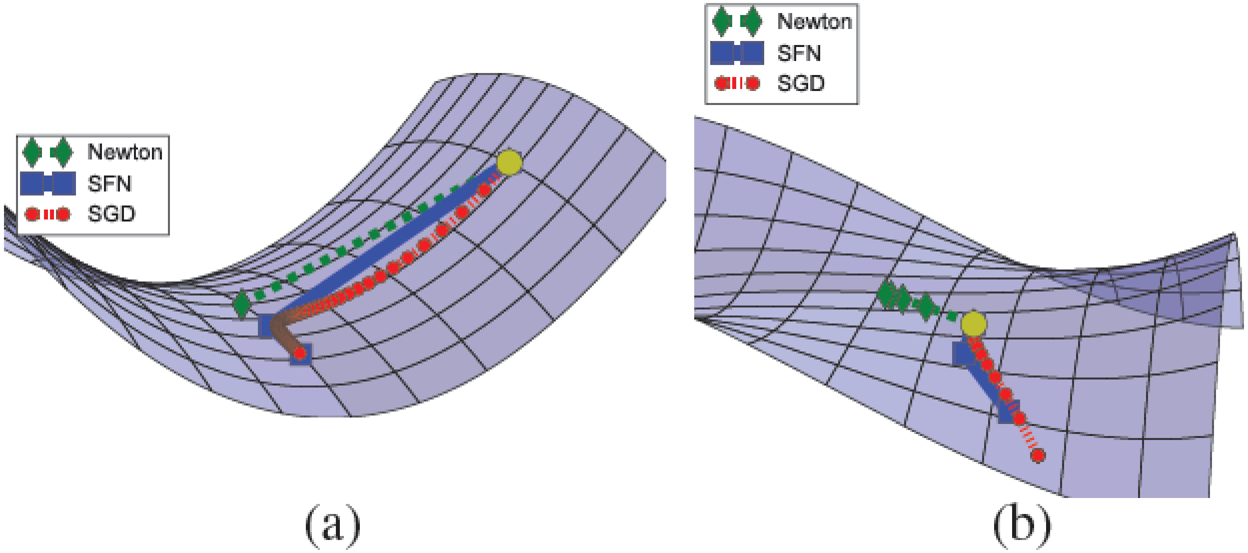
Local Optima vs Saddle Points

- Statistic Physics provide analytical tools

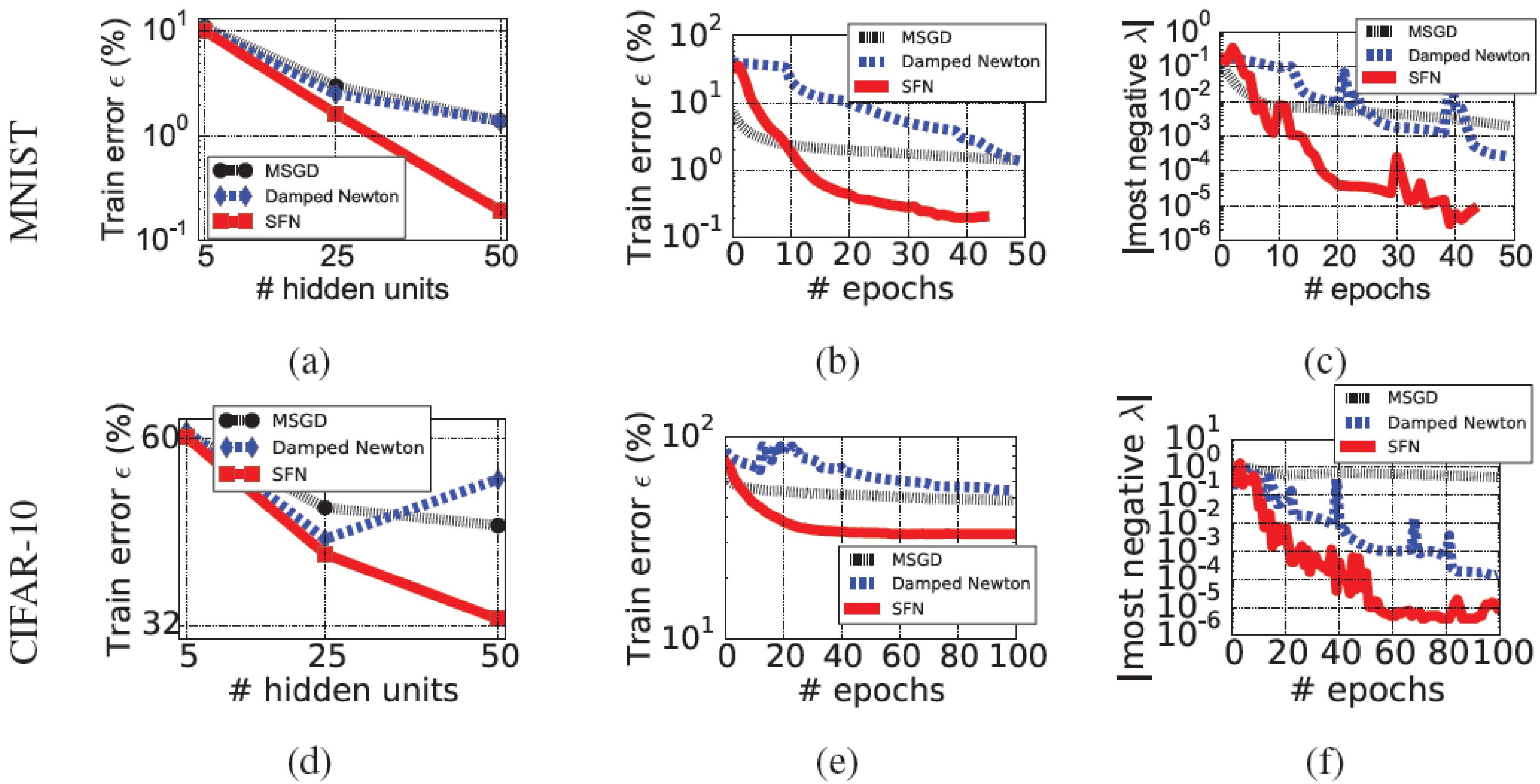
- High-dimensional optimization problem
  - Most critical points are saddle points
  - The likelihood grows exponentially!

Dynamics of Various Opt. Techniques

- SGD:
  - Gradient is accurate, but may suffer from slow steps
- Newton method:
  - Wrong directions when negative curvatures present
  - Saddle points become attractors! (can't escape)
- Saddle-free method:
  - A generalization of Newton's method to escape saddle points (more rapidly than SGD)

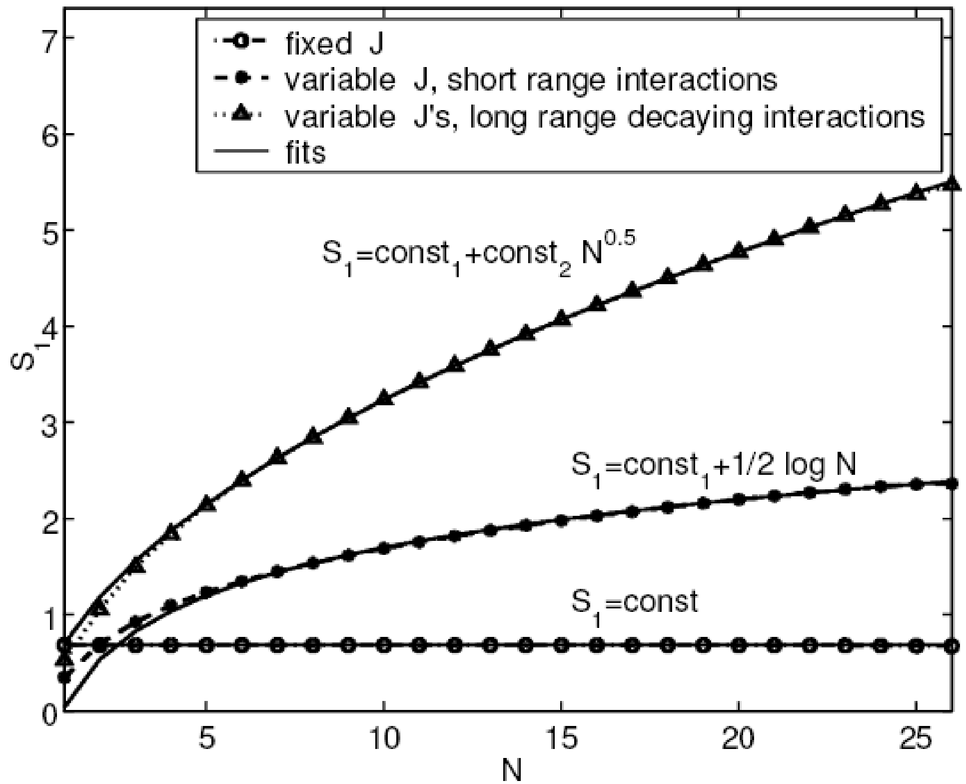


Some Empirical Results



Overfitting in Big Data

- **Predictive information** grows slower than the amount of Shannon entropy (Bialek et al., 2001)



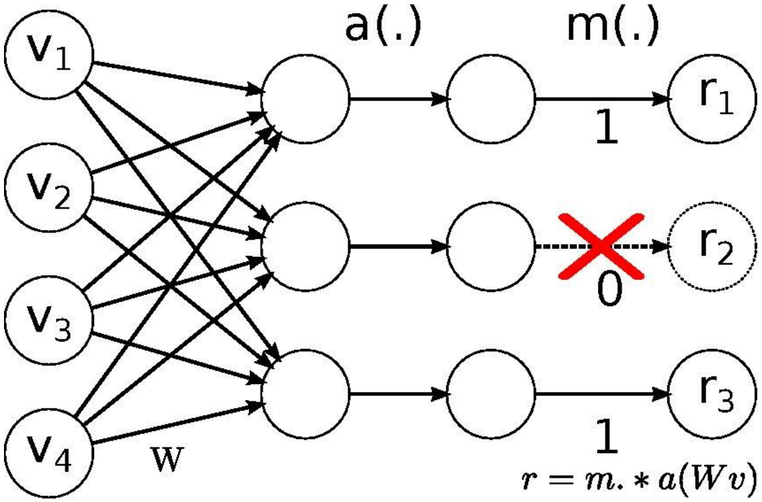
Overfitting in Big Data

- **Predictive information** grows slower than the amount of Shannon entropy (Bialek et al., 2001)

Model capacity grows faster than the amount of predictive information!

Overfitting in DL

- Increasing research attention, e.g., dropout training (Hinton, 2012)

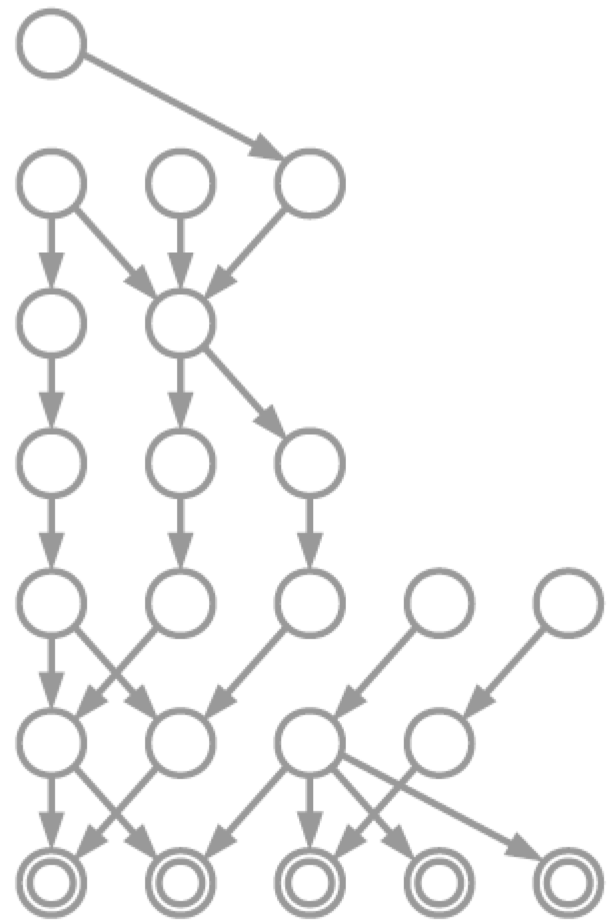


- More theoretical understanding and extensions
  - MCF (van der Maaten et al., 2013); Logistic-loss (Wager et al., 2013); Dropout SVM (Chen, Zhu et al., 2014)

Model Complexity

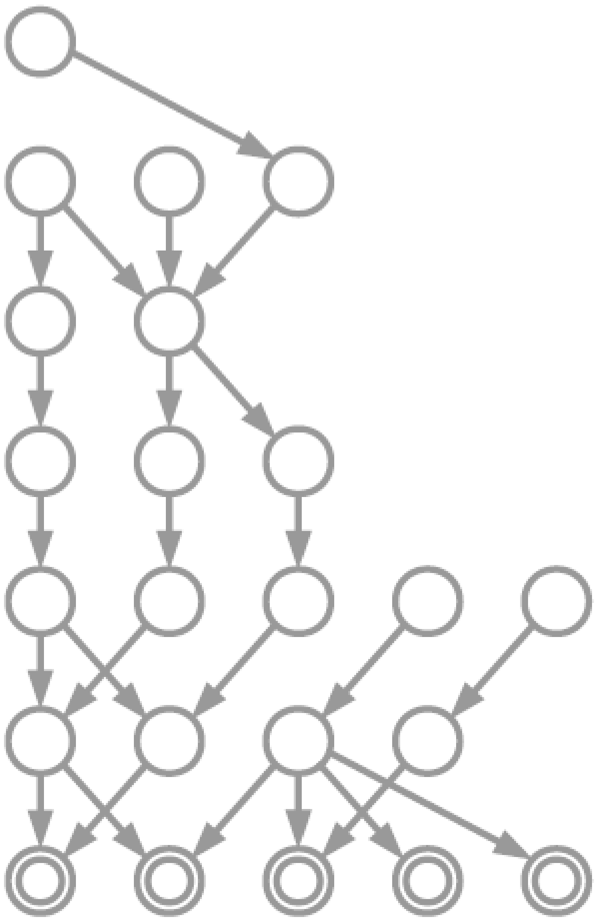
- What do we mean by structure learning in deep GMs?
  - # of layers
  - # of hidden units at each layer
  - The type of each hidden unit (discrete or continuous?)
  - The connection structures (i.e., edges) between hidden units

- Adams et al. presented a structure learning method using nonparametric Bayesian techniques – a cascading IBP (CIBP) process [Admas, Wallach & Ghahramani, 2010]



Structure of Deep Belief Networks

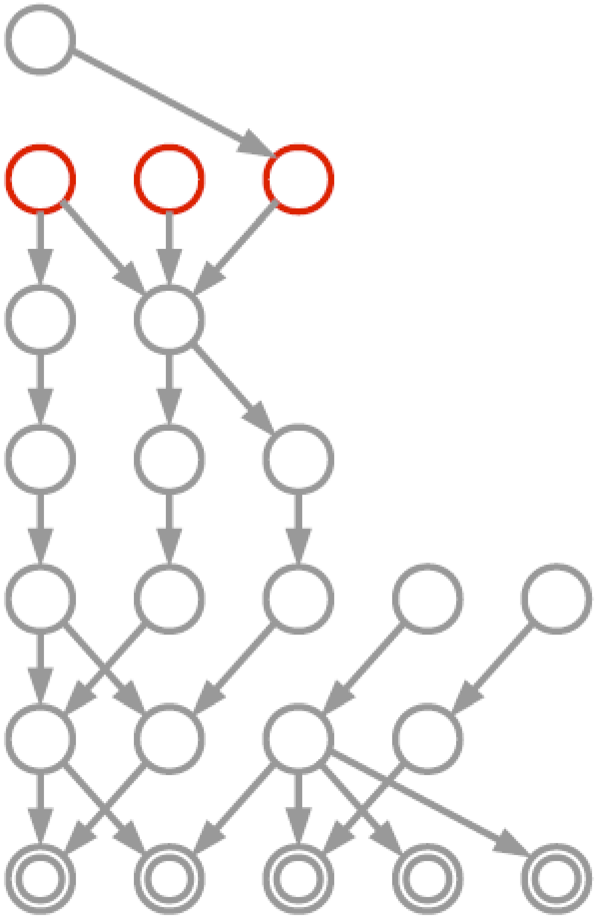
- What do we mean by structure learning in deep GMs?
  - # of layers
  - # of hidden units at each layer
  - The type of each hidden unit (discrete or continuous?)
  - The connection structures (i.e., edges) between hidden units



[Animation by Wallach]

Structure of Deep Belief Networks

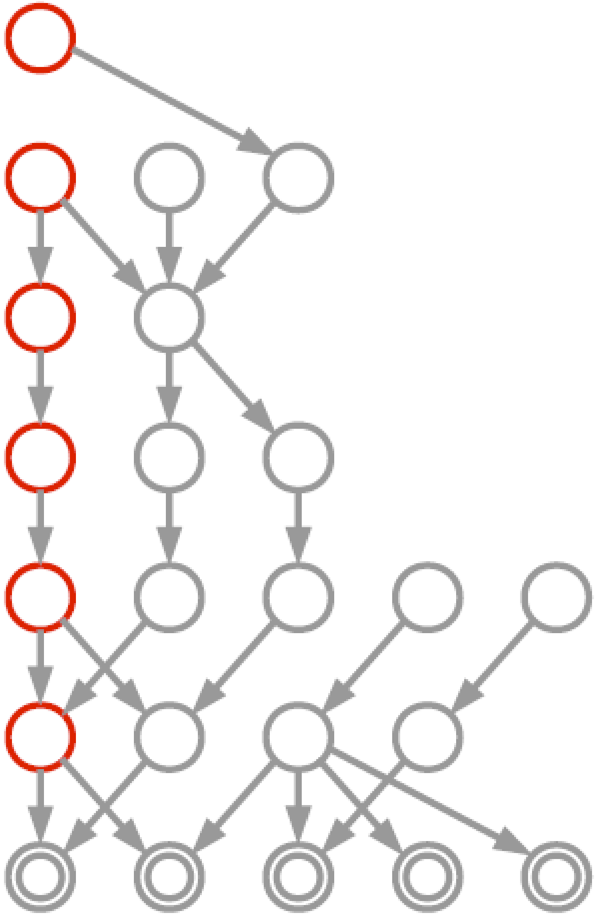
- What do we mean by structure learning in deep GMs?
  - # of hidden units at each layer
  - # of layers
  - The type of each hidden unit (discrete or continuous?)
  - The connection structures (i.e., edges) between hidden units



[Animation by Wallach]

Structure of Deep Belief Networks

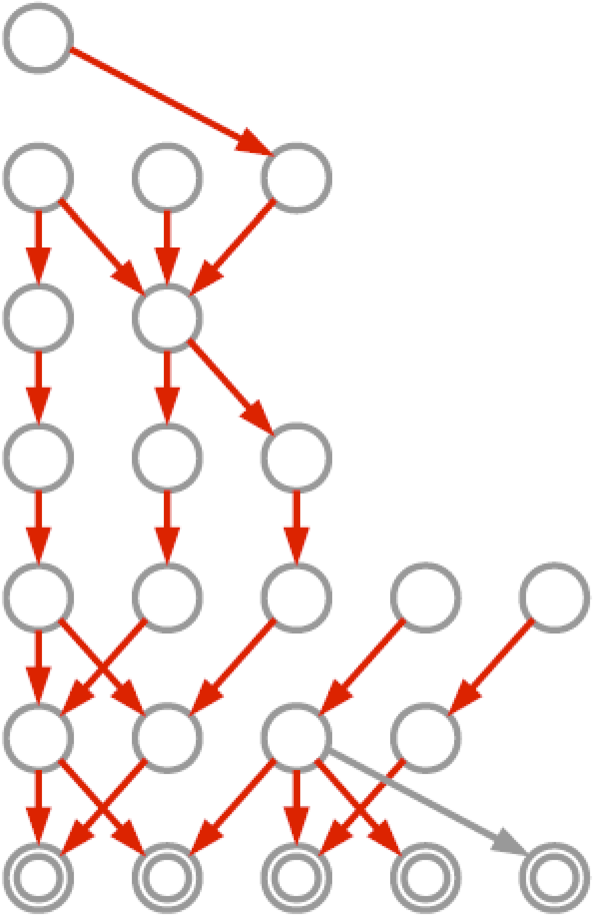
- What do we mean by structure learning in deep GMs?
  - # of hidden units at each layer
  - # of layers
  - The type of each hidden unit (discrete or continuous?)
  - The connection structures (i.e., edges) between hidden units



[Animation by Wallach]

Structure of Deep Belief Networks

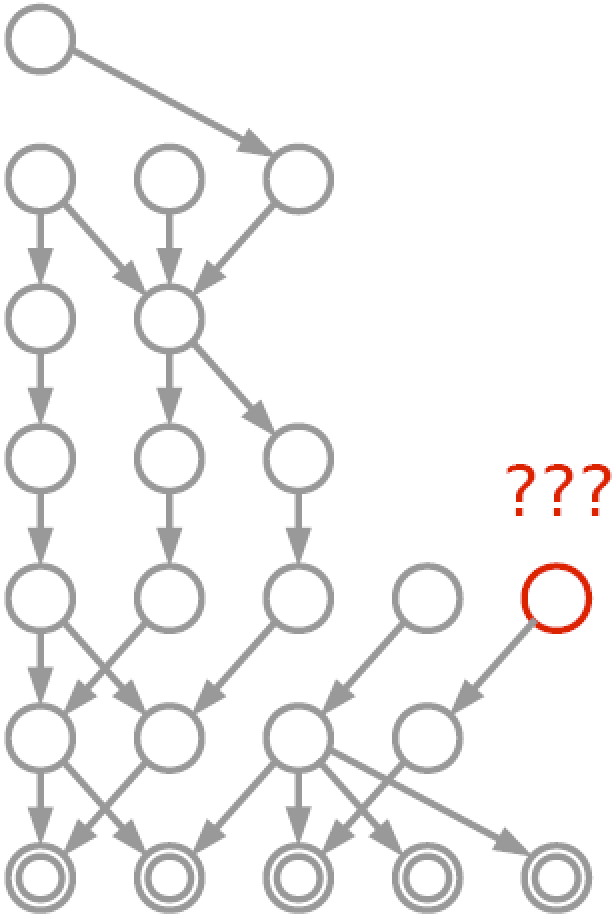
- What do we mean by structure learning in deep GMs?
  - # of layers
  - # of hidden units at each layer
  - The connection structures (i.e., edges) between hidden units
  - The type of each hidden unit (discrete or continuous?)



[Animation by Wallach]

Structure of Deep Belief Networks

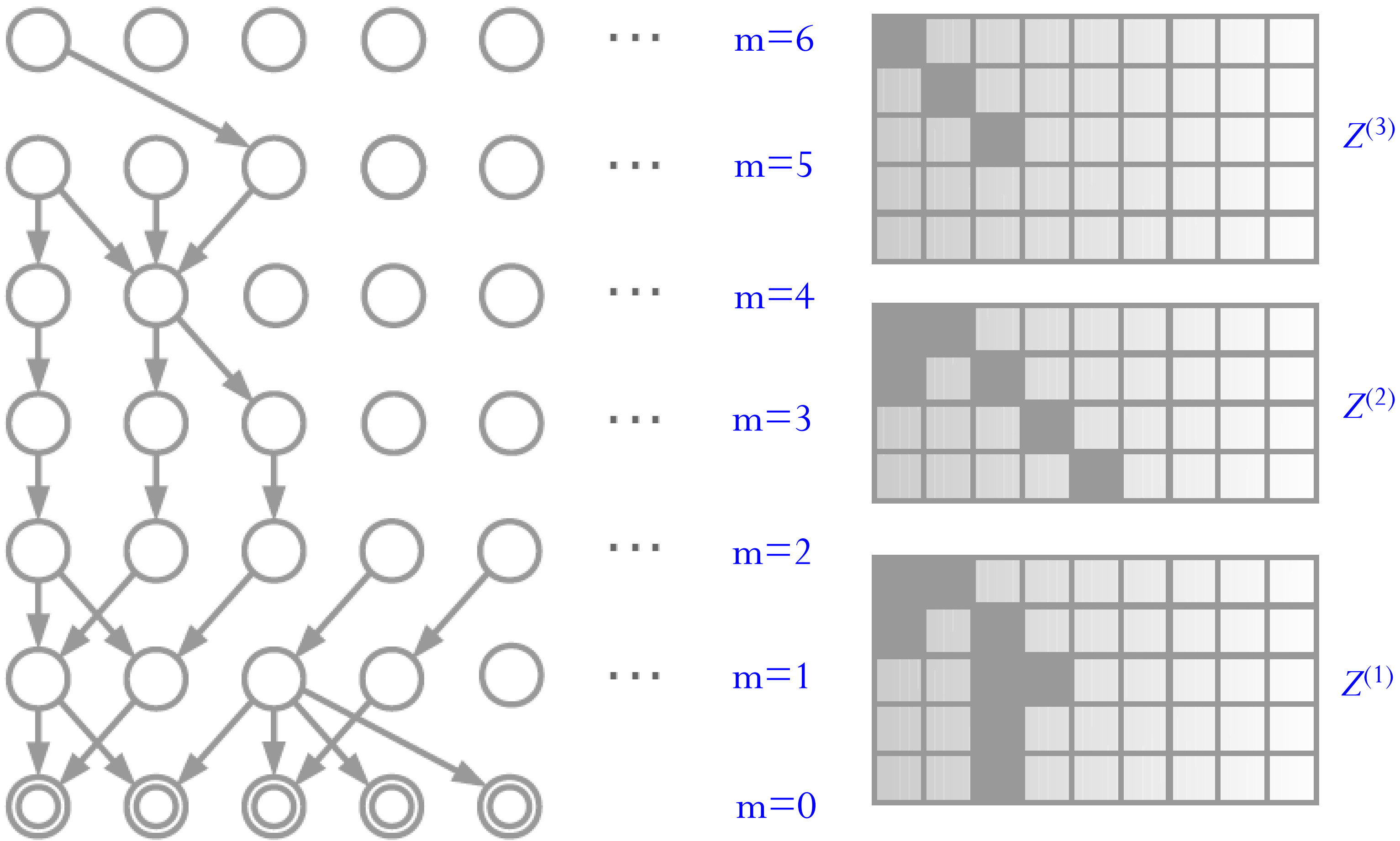
- What do we mean by structure learning in deep GMs?
  - # of layers
  - # of hidden units at each layer
  - The connection structures (i.e., edges) between hidden units
  - The type of each hidden unit (discrete or continuous?)



[Animation by Wallach]

Multi-Layer Belief Networks

- A sequence of binary matrices => deep BNs



The Cascading IBP (CIBP)

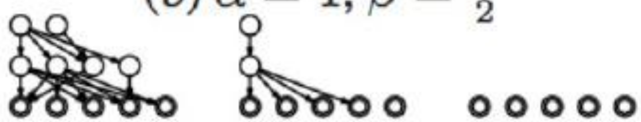
- A stochastic process which results in an infinite sequence of infinite binary matrices
  - Each matrix is exchangeable in both rows and columns
- How do we know the CIBP converges?
  - The number of dishes in one layer depends only on the number of customers in the previous layer
  - Can prove that this Markov chain reaches an absorbing state in finite time with probability one

Samples from CIBP Prior

*** only connected units are shown**

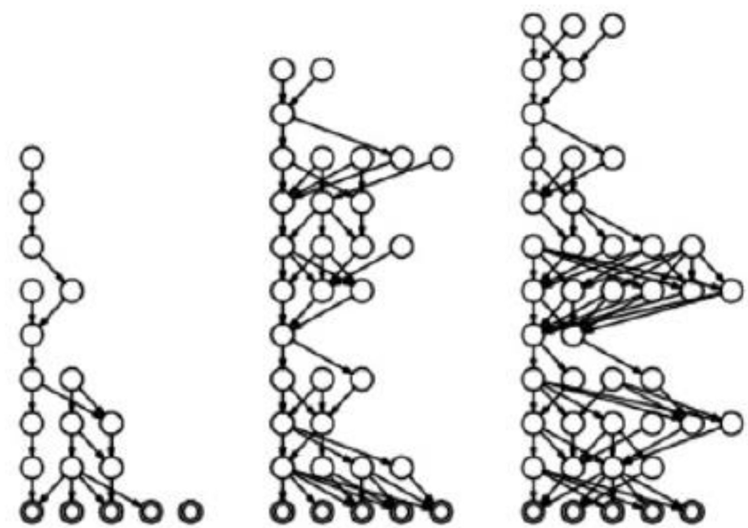
(a) $\alpha = 1, \beta = 1$

(b) $\alpha = 1, \beta = \frac{1}{2}$



(c) $\alpha = \frac{1}{2}, \beta = 1$

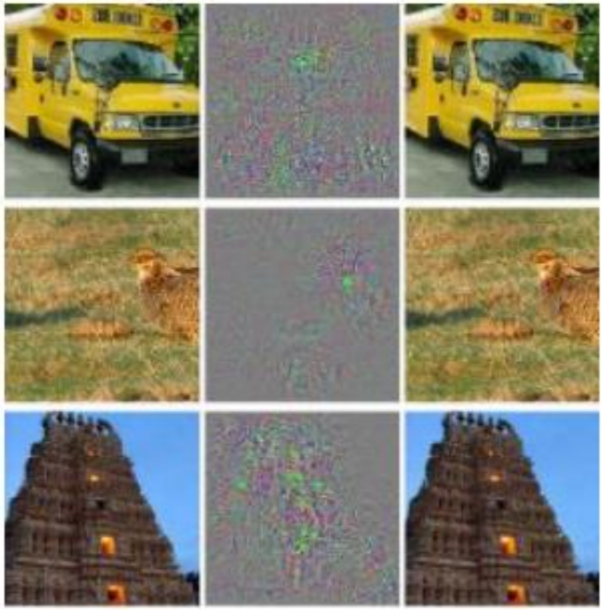
(d) $\alpha = 1, \beta = 2$



(e) $\alpha = \frac{3}{2}, \beta = 1$

Some counter-intuitive properties

- Stability w.r.t small perturbations to inputs
  - Imperceptible non-random perturbation can arbitrarily change the prediction (adversarial examples exist!)



(a)

(b)

10x of differences

[Szegedy et al., Intriguing properties of neural nets, 2013]

Criticisms of DL

- Just a buzzword, or largely a rebranding of neural networks
- Lack of theory
  - gradient descent has been understood for a while
  - DL is often used as black-box
- DL is only part of the larger challenge of building intelligent machines, still lacking of:
  - causal relationships
  - logic inferences
  - integrating abstract knowledge

How can neural science help?

- The current DL models:
  - loosely inspired by the densely interconnected neurons of the brain
  - mimic human learning by changing weights based on experience

- How to improve?
  - Transparent architecture?
    - Attention mechanism?
  - Cheap learning?
    - (partially) replace back-propagation?
  - Others?

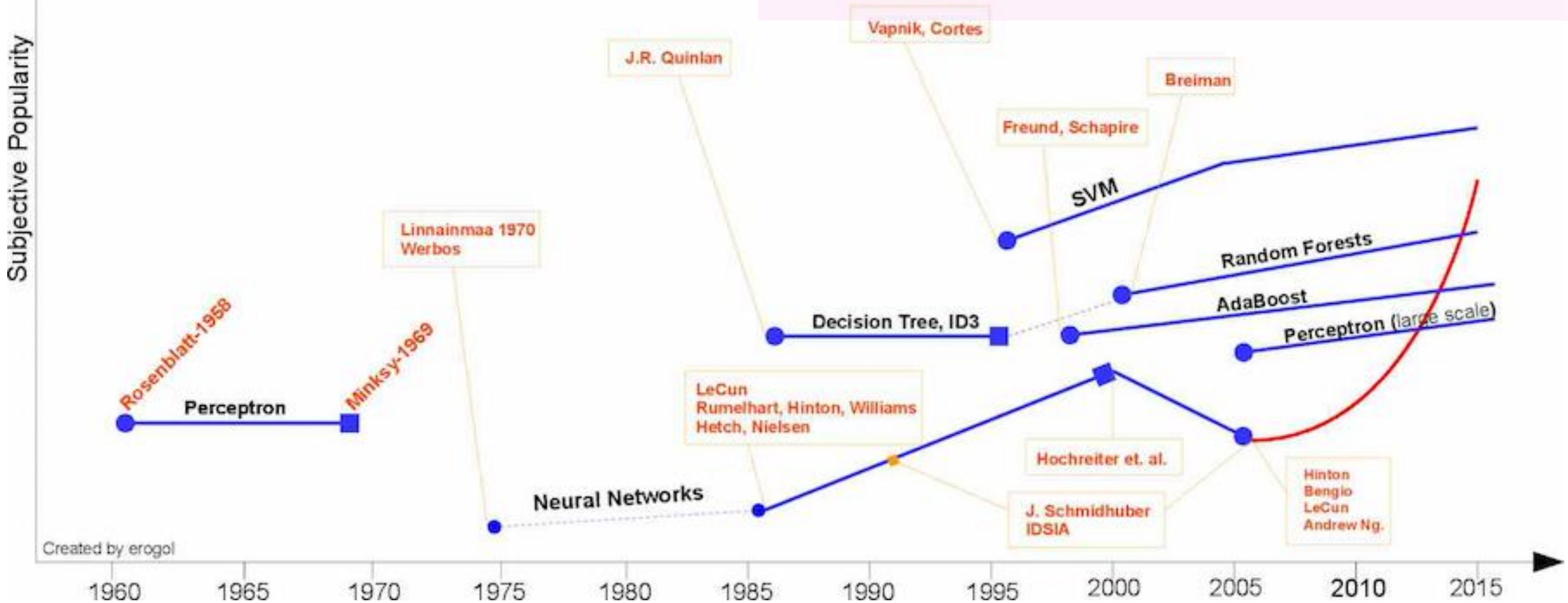
清華大學 Tsinghua University

Will DL make other ML methods obsolete?

Quora 2014/12/23

Yes (2 post, 113 upvotes)
- best predictive power when data sufficient
- DL is far from saturated
- Google et al invests on DL, it is the "richest" AI topic

No (10 posts, 284 upvotes)
- simpler algorithms are just fine in many cases
- methods with domain knowledge works better
- DL is feature learning, needs other methods to work
- DL is not that well developed, a lot of work to be done using more traditional methods
- No free lunch
- a lot like how ANN was viewed in the late 80s



清華大學 Tsinghua University

What are people saying?

- **Yann LeCun**:
  - "AI has gone from failure to failure, with bits of progress. This could be another leapfrog"
- **Jitendra Malik**:
  - in the long term, deep learning may not win the day; … "Over time people will decide what works best in different domains."
  - "Neural nets were always a delicate art to manage. There is some black magic involved"
- **Andrew Ng**:
  - "Deep learning happens to have the property that if you feed it more data it gets better and better,"
  - "Deep-learning algorithms aren't the only ones like that, but they're arguably the best — certainly the easiest. That's why it has huge promise for the future."

[Nature 505, 146–148 (09 January 2014)]

What are people saying?

- **Oren Etzioni**:
  - "It's like when we invented flight" (not using the brain for inspiration)

- **Alternatives**:
  - Logic, knowledge base, grammars?
  - Quantum AI/ML?

[Nature 505, 146–148 (09 January 2014)]

Thank You!