

Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 27, 2012

Today:

- Artificial neural networks
- Backpropagation
- Cognitive modeling
- Deep belief networks

Reading:

- Mitchell: Chapter 4
- Bishop: Chapter 5

Optional:

- Le: ICML 2012 paper (on website)

Logistic Regression is linear classifier

Assumes the following functional form for $P(Y|X)$:

$$P(Y = 1|X) = \frac{1}{1 + \exp(-(w_0 + \sum_i w_i X_i))}$$

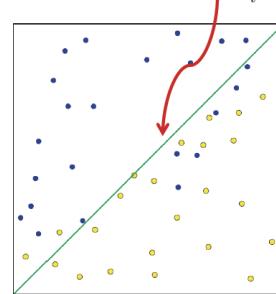
$$w_0 + \sum_i w_i X_i = 0$$

Decision boundary:

$$P(Y = 0|X) \stackrel{0}{\underset{1}{\gtrless}} P(Y = 1|X)$$

$$0 \stackrel{0}{\underset{1}{\gtrless}} w_0 + \sum_i w_i X_i$$

(Linear Decision Boundary)



[slide: Aarti Singh]

Artificial Neural Networks to learn $f: X \rightarrow Y$

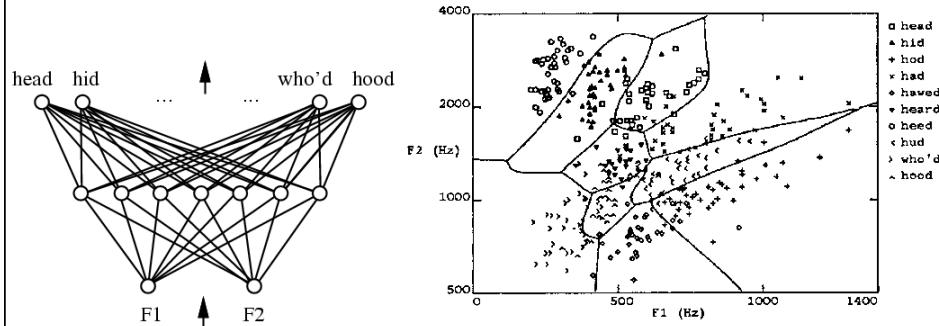
- f might be non-linear function
- X (vector of) continuous and/or discrete vars
- Y (vector of) continuous and/or discrete vars

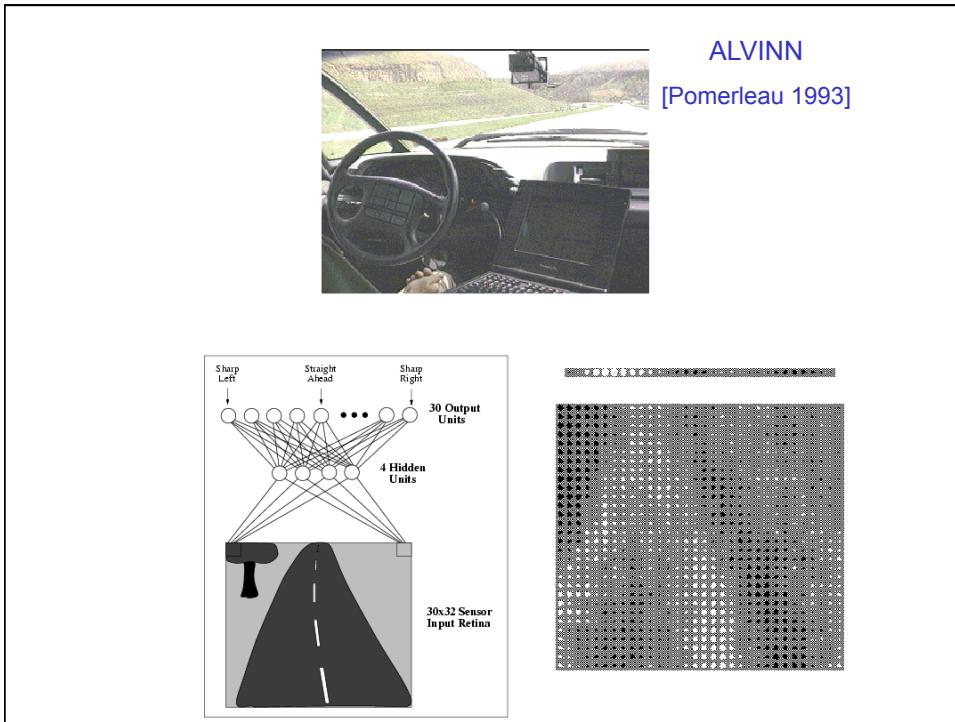
- Represent f by network of logistic units
- Each unit is a logistic function

$$\text{unit output} = \frac{1}{1 + \exp(w_0 + \sum_i w_i x_i)}$$

- MLE: train weights of all units to minimize sum of squared errors of predicted network outputs
- MAP: train to minimize sum of squared errors plus weight magnitudes

Multilayer Networks of Sigmoid Units





Connectionist Models

Consider humans:

- Neuron switching time $\sim .001$ second
- Number of neurons $\sim 10^{10}$
- Connections per neuron $\sim 10^{4-5}$
- Scene recognition time $\sim .1$ second
- 100 inference steps doesn't seem like enough
 \rightarrow much parallel computation

Properties of artificial neural nets (ANN's):

- Many neuron-like threshold switching units
- Many weighted interconnections among units
- Highly parallel, distributed process

Sigmoid Unit

$\sigma(x)$ is the sigmoid function

$$\frac{1}{1 + e^{-x}}$$

Nice property: $\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$

We can derive gradient decent rules to train

- One sigmoid unit
- *Multilayer networks* of sigmoid units → Backpropagation

M(C)LE Training for Neural Networks

- Consider regression problem $f: X \rightarrow Y$, for scalar Y

$$y = f(x) + \varepsilon \quad \begin{matrix} \text{assume noise } N(0, \sigma_\varepsilon), \text{ iid} \\ \text{deterministic} \end{matrix}$$

- Let's maximize the conditional data likelihood

$$W \leftarrow \arg \max_W \ln \prod_l P(Y^l | X^l, W)$$

$$W \leftarrow \arg \min_W \sum_l (y^l - \hat{f}(x^l))^2$$

↑
Learned
neural network

MAP Training for Neural Networks

- Consider regression problem $f: X \rightarrow Y$, for scalar Y

$$y = f(x) + \varepsilon \quad \text{noise } N(0, \sigma_\varepsilon)$$

deterministic

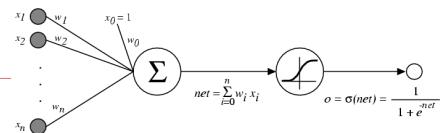
$$\text{Gaussian } P(W) = N(0, \sigma I)$$

$$W \leftarrow \arg \max_W \ln P(W) \prod_l P(Y^l | X^l, W)$$

$$W \leftarrow \arg \min_W \left[c \sum_i w_i^2 \right] + \left[\sum_l (y^l - \hat{f}(x^l))^2 \right]$$

$$\ln P(W) \Leftrightarrow c \sum_i w_i^2$$

Error Gradient for a Sigmoid Unit



$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 \\ &= \frac{1}{2} \sum_d \frac{\partial}{\partial w_i} (t_d - o_d)^2 \\ &= \frac{1}{2} \sum_d 2(t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d) \\ &= \sum_d (t_d - o_d) \left(-\frac{\partial o_d}{\partial w_i} \right) \\ &= -\sum_d (t_d - o_d) \frac{\partial o_d}{\partial net_d} \frac{\partial net_d}{\partial w_i} \end{aligned}$$

x_d = input
 t_d = target output
 o_d = observed unit output
 w_i = weight i

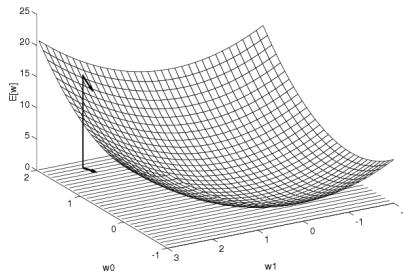
But we know:

$$\begin{aligned} \frac{\partial o_d}{\partial net_d} &= \frac{\partial \sigma(net_d)}{\partial net_d} = o_d(1 - o_d) \\ \frac{\partial net_d}{\partial w_i} &= \frac{\partial (\vec{w} \cdot \vec{x}_d)}{\partial w_i} = x_{i,d} \end{aligned}$$

So:

$$\frac{\partial E}{\partial w_i} = -\sum_{d \in D} (t_d - o_d) o_d (1 - o_d) x_{i,d}$$

Gradient Descent



Gradient

$$\nabla E[\vec{w}] \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \dots, \frac{\partial E}{\partial w_n} \right]$$

Training rule:

$$\Delta \vec{w} = -\eta \nabla E[\vec{w}]$$

i.e.,

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

Backpropagation Algorithm (MLE)

Initialize all weights to small random numbers.
Until satisfied, Do

- For each training example, Do
 1. Input the training example to the network and compute the network outputs
 2. For each output unit k

$$\delta_k \leftarrow o_k(1 - o_k)(t_k - o_k)$$

3. For each hidden unit h

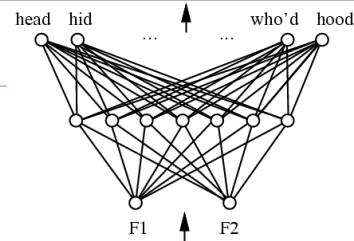
$$\delta_h \leftarrow o_h(1 - o_h) \sum_{k \in \text{outputs}} w_{h,k} \delta_k$$

4. Update each network weight $w_{i,j}$

$$w_{i,j} \leftarrow w_{i,j} + \Delta w_{i,j}$$

where

$$\Delta w_{i,j} = \eta \delta_j x_i$$

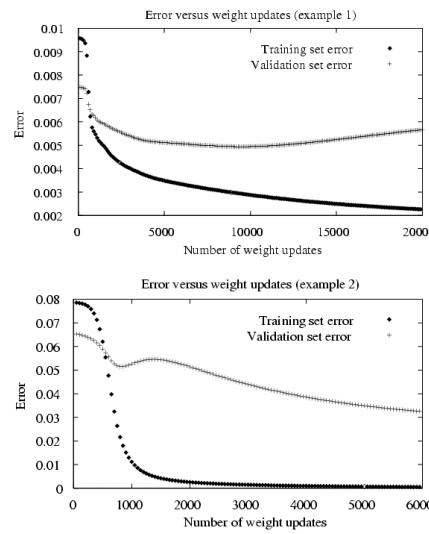


x_d = input
 t_d = target output
 o_d = observed unit output
 w_{ij} = wt from i to j

More on Backpropagation

- Gradient descent over entire *network* weight vector
- Easily generalized to arbitrary directed graphs
- Will find a local, not necessarily global error minimum
 - In practice, often works well (can run multiple times)
- Often include weight *momentum* α
$$\Delta w_{i,j}(n) = \eta \delta_j x_{i,j} + \alpha \Delta w_{i,j}(n-1)$$
- Minimizes error over *training* examples
 - Will it generalize well to subsequent examples?
- Training can take thousands of iterations → slow!
- Using network after training is very fast

Overfitting in ANNs



Dealing with Overfitting

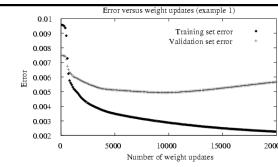
Our learning algorithm involves a parameter

n =number of gradient descent iterations

How do we choose n to optimize future error?

(note: similar issue for logistic regression, decision trees, ...)

e.g. the n that minimizes error rate of neural net over future data



Dealing with Overfitting

Our learning algorithm involves a parameter

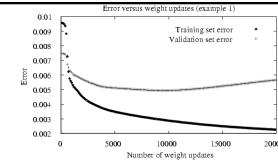
n =number of gradient descent iterations

How do we choose n to optimize future error?

- Separate available data into training and validation set
- Use training to perform gradient descent
- $n \leftarrow$ number of iterations that optimizes validation set error

→ gives *unbiased estimate of optimal n*

(but a biased estimate of true error)



K-Fold Cross Validation

Idea: train multiple times, leaving out a disjoint subset of data each time for test. Average the test set accuracies.

```
Partition data into K disjoint subsets
For k=1 to K
    testData = kth subset
    h ← classifier trained* on all data except for testData
    accuracy(k) = accuracy of h on testData
end
FinalAccuracy = mean of the K recorded testset accuracies
```

* might withhold some of this to choose number of gradient decent steps

Leave-One-Out Cross Validation

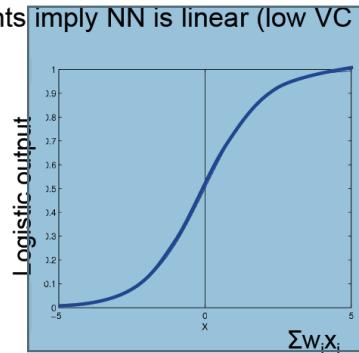
This is just k-fold cross validation leaving out one example each iteration

```
Partition data into K disjoint subsets, each containing one example
For k=1 to K
    testData = kth subset
    h ← classifier trained* on all data except for testData
    accuracy(k) = accuracy of h on testData
end
FinalAccuracy = mean of the K recorded testset accuracies
```

* might withhold some of this to choose number of gradient decent steps

Dealing with Overfitting

- Cross-validation
- Regularization – small weights imply NN is linear (low VC dimension)



- Control number of hidden units – low complexity

[slide: Aarti Singh]

Expressive Capabilities of ANNs

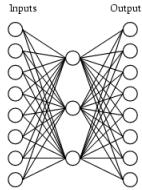
Boolean functions:

- Every boolean function can be represented by network with single hidden layer
- but might require exponential (in number of inputs) hidden units

Continuous functions:

- Every bounded continuous function can be approximated with arbitrarily small error, by network with one hidden layer [Cybenko 1989; Hornik et al. 1989]
- Any function can be approximated to arbitrary accuracy by a network with two hidden layers [Cybenko 1988].

Learning Hidden Layer Representations



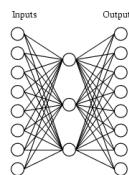
A target function:

Input	Output
10000000	→ 10000000
01000000	→ 01000000
00100000	→ 00100000
00010000	→ 00010000
00001000	→ 00001000
00000100	→ 00000100
00000010	→ 00000010
00000001	→ 00000001

Can this be learned??

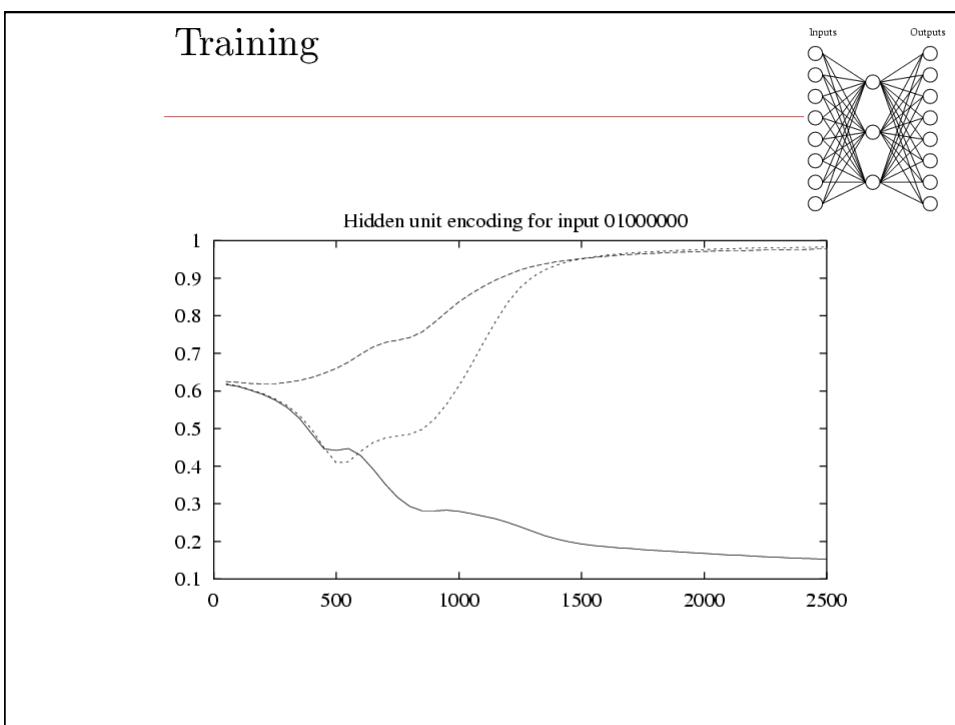
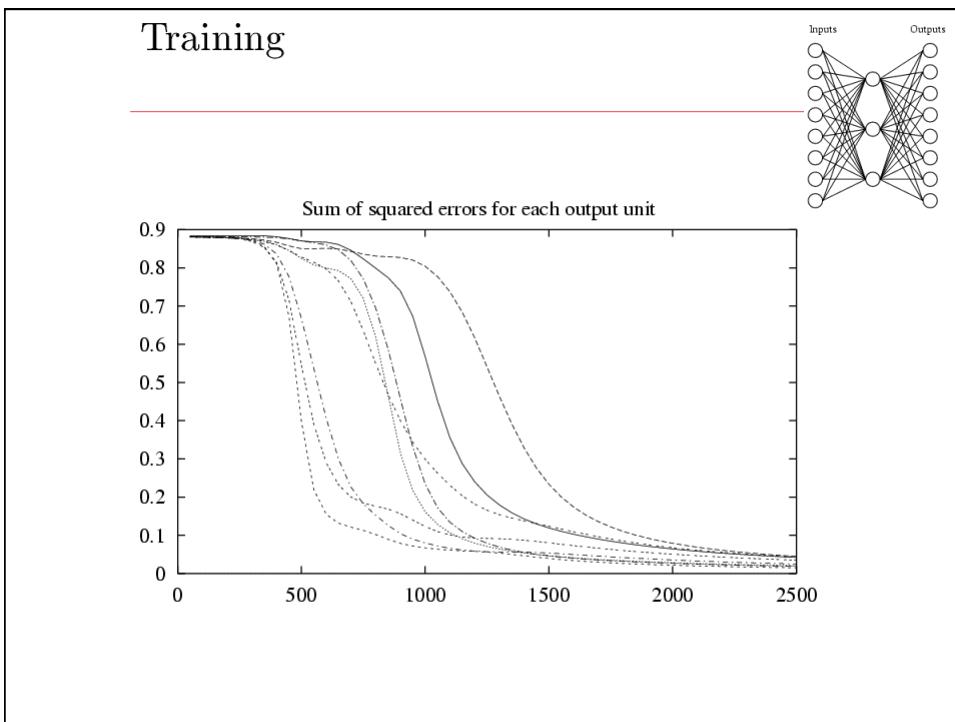
Learning Hidden Layer Representations

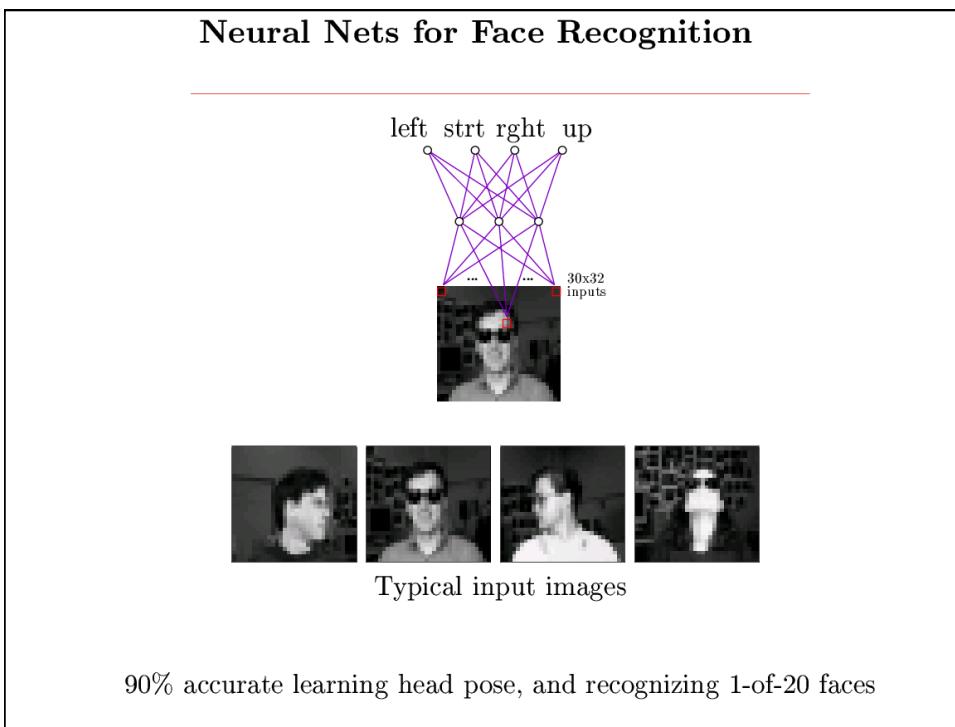
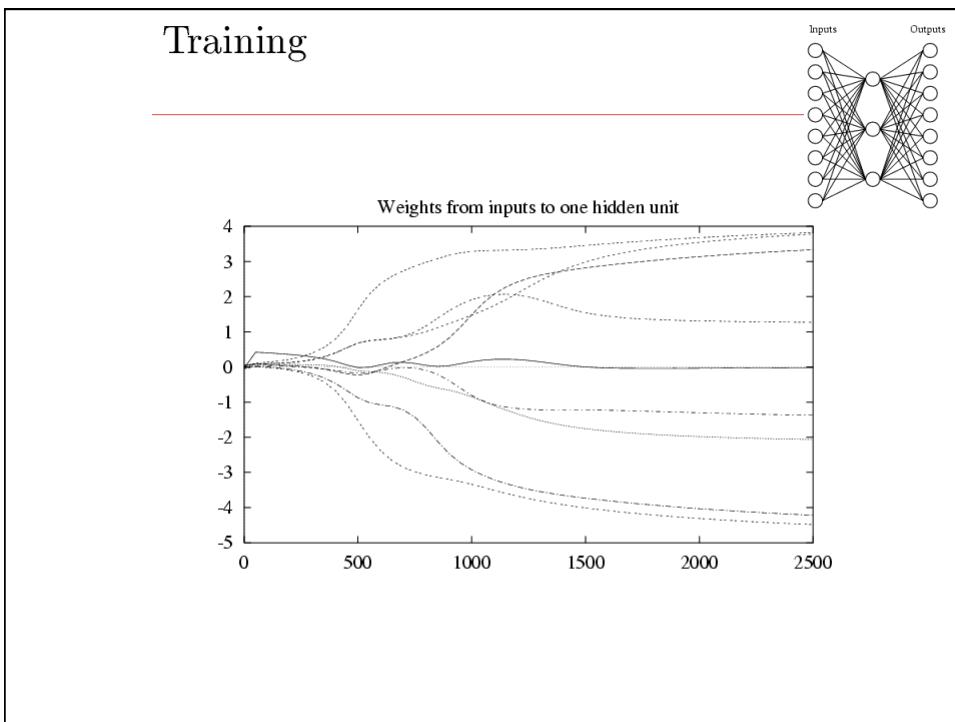
A network:

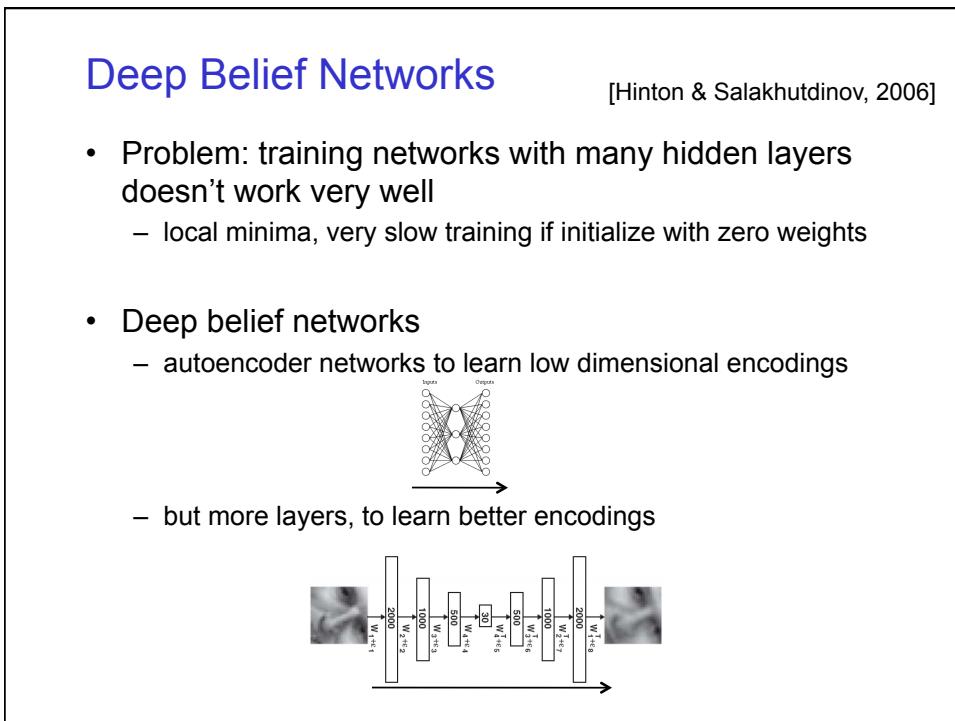
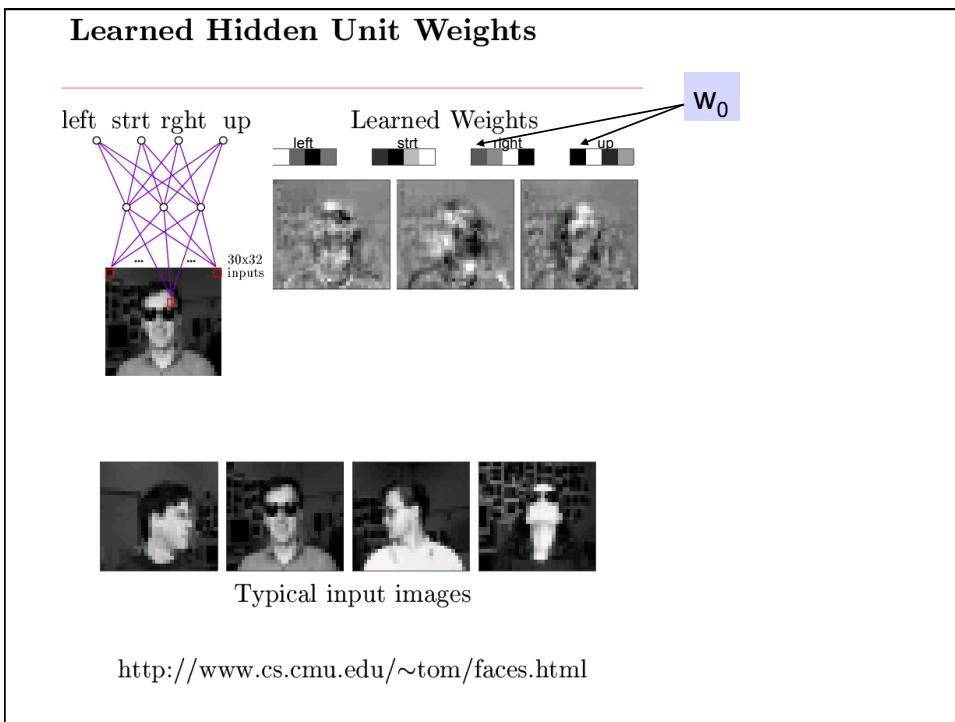


Learned hidden layer representation:

Input	Hidden Values	Output
10000000	→ .89 .04 .08	→ 10000000
01000000	→ .01 .11 .88	→ 01000000
00100000	→ .01 .97 .27	→ 00100000
00010000	→ .99 .97 .71	→ 00010000
00001000	→ .03 .05 .02	→ 00001000
00000100	→ .22 .99 .99	→ 00000100
00000010	→ .80 .01 .98	→ 00000010
00000001	→ .60 .94 .01	→ 00000001

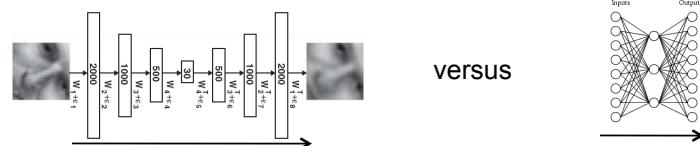
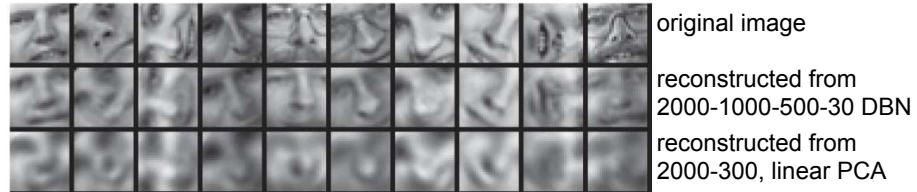






Deep Belief Networks

[Hinton & Salakhutdinov, 2006]

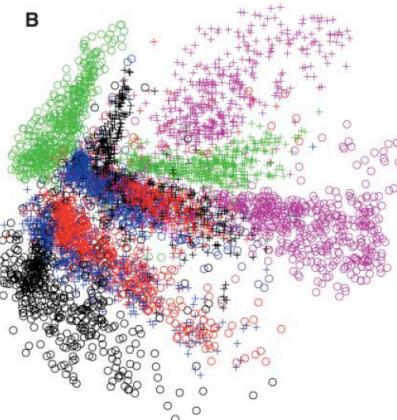
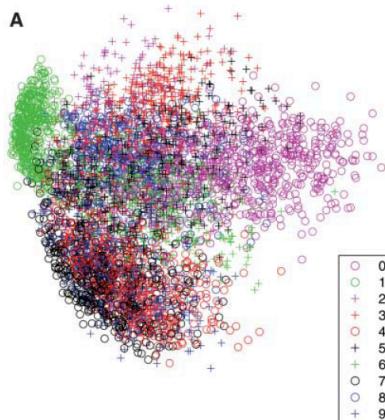


Encoding of digit images in two dimensions

[Hinton & Salakhutdinov, 2006]

784-2 linear encoding (PCA)

784-1000-500-250-2 DBNet



[Hinton & Salakhutdinov, 2006]

Deep Belief Networks: Training

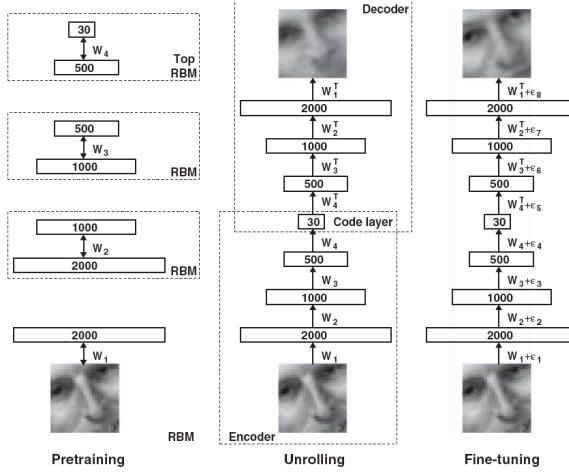
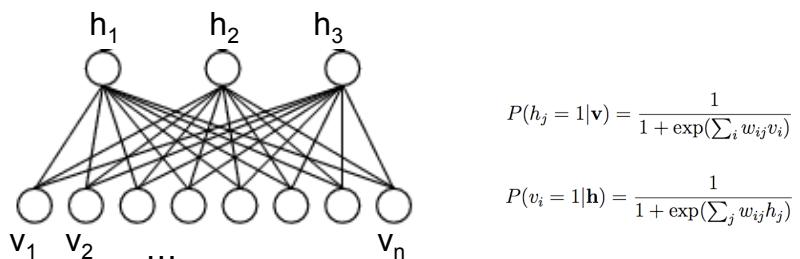


Fig. 1. Pretraining consists of learning a stack of restricted Boltzmann machines (RBMs), each having only one layer of feature detectors. The learned feature activations of one RBM are used as the "data" for training the next RBM in the stack. After the pretraining, the RBMs are "unrolled" to create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

Restricted Boltzmann Machine

- Bipartite graph, logistic activation
- Inference: fill in any nodes, estimate other nodes
- consider v_i, h_j are boolean variables



Very Large Scale Use of DBN's

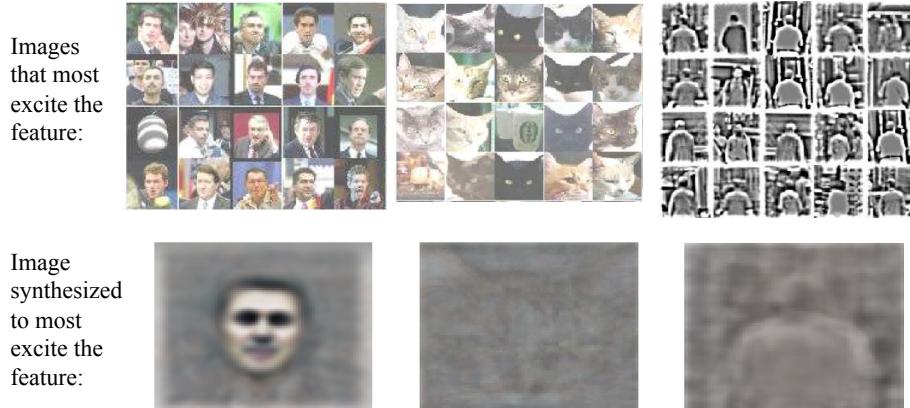
[Quoc Le, et al., *ICML*, 2012]

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

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

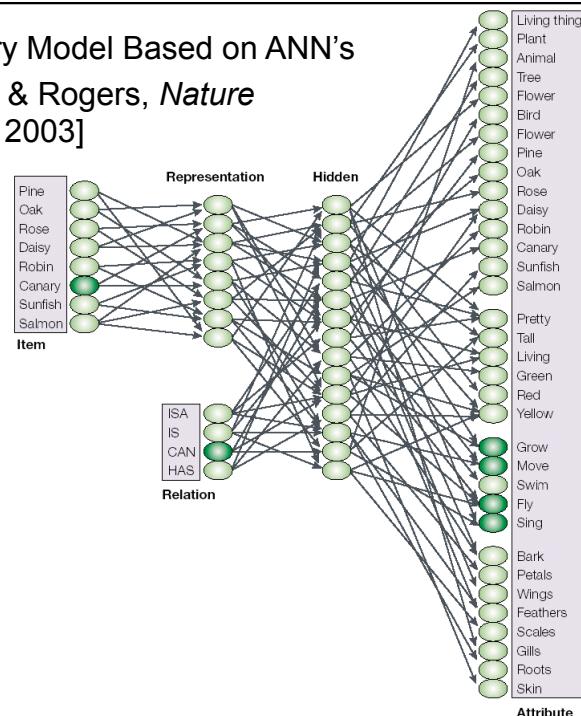
Learned network: 3 multi-stage layers, 1.15 billion parameters

Achieves 15.8% accuracy classifying 1 of 20k categories in ImageNet data



Semantic Memory Model Based on ANN's

[McClelland & Rogers, *Nature*, 2003]



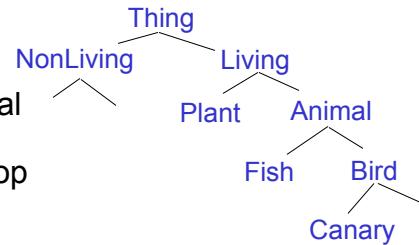
No hierarchy given.

Train with assertions,
e.g., Can(Canary,Fly)

Humans act as though they have a hierarchical memory organization

1. Victims of Semantic Dementia progressively lose knowledge of objects
But they lose specific details first, general properties later, suggesting hierarchical memory organization

2. Children appear to learn general categories and properties first, following the same hierarchy, top down*.



Question: What learning mechanism could produce this emergent hierarchy?

* some debate remains on this.

Memory deterioration follows semantic hierarchy

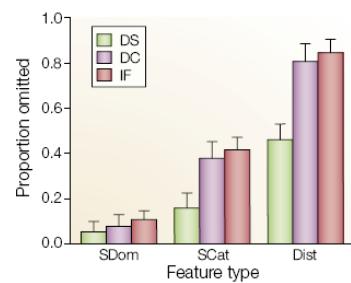
[McClelland & Rogers, *Nature* 2003]

a

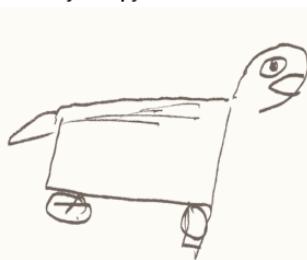
Picture naming responses for JL

Item	Sept. 91	March 92	March 93
Bird	+	+	Animal
Chicken	+	+	Animal
Duck	+	Bird	Dog
Swan	+	Bird	Animal
Eagle	Duck	Bird	Horse
Ostrich	Swan	Bird	Animal
Peacock	Duck	Bird	Vehicle
Penguin	Duck	Bird	Part of animal
Rooster	Chicken	Chicken	Dog

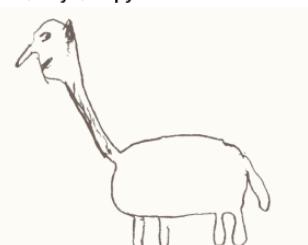
b

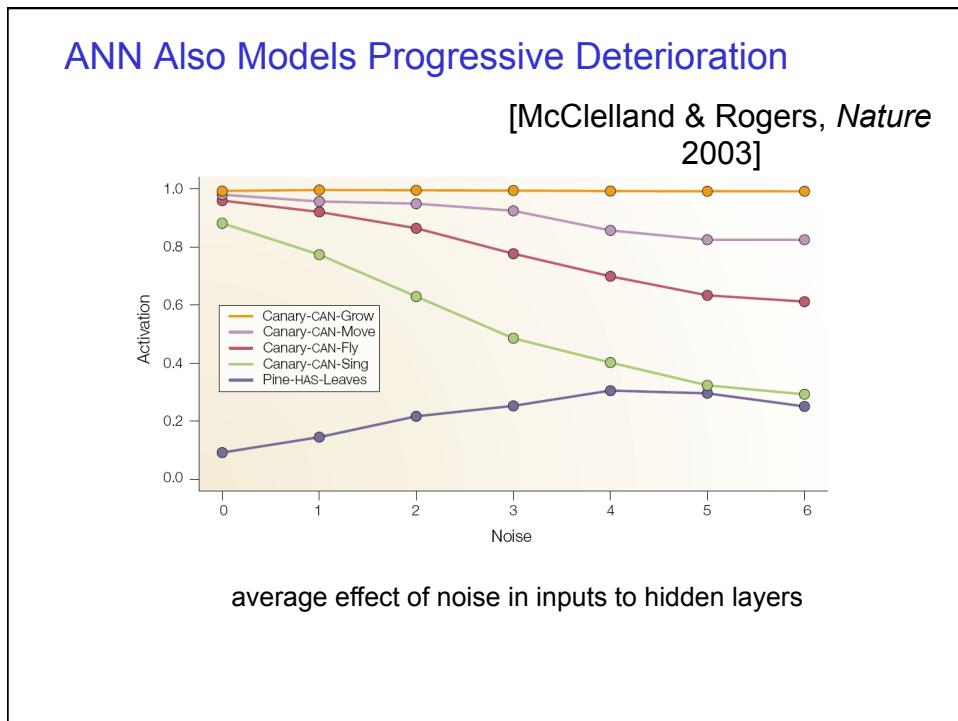
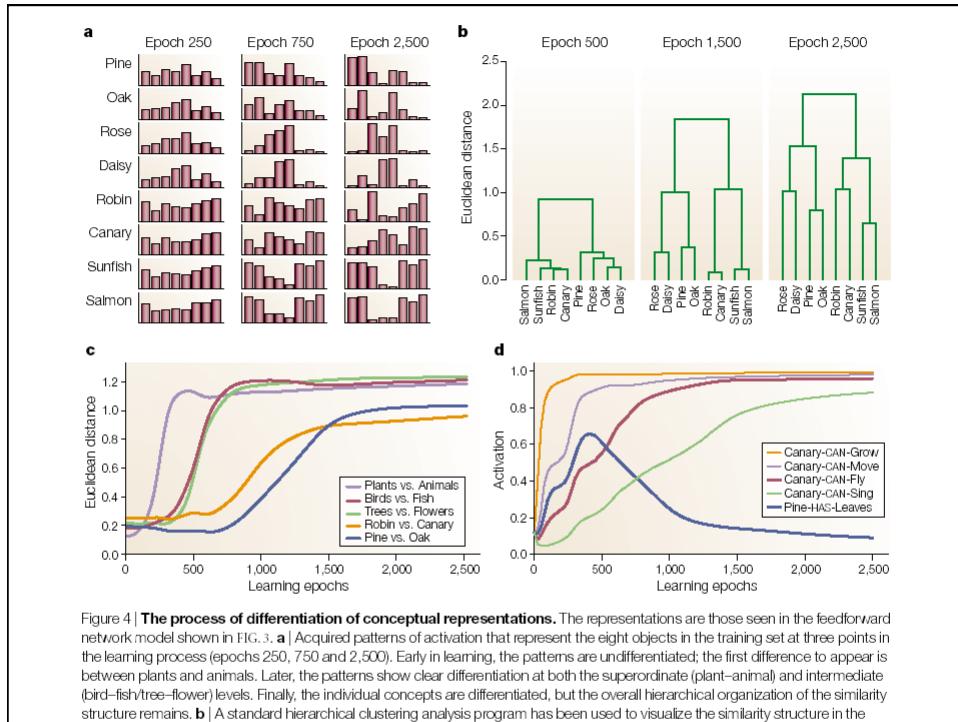


c IF's delayed copy of a camel



d DC's delayed copy of a swan





What you should know: Artificial Neural Networks

- Highly non-linear regression/classification
- Vector-valued inputs and outputs
- Potentially millions of parameters to estimate
- Hidden layers learn intermediate representations
- Actively used to model distributed computation in brain
- Gradient descent, local minima problems
- Overfitting and how to deal with it
- Many extensions