Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

March 31, 2011

Today:
Learning representations III

- Deep Belief Networks
- ICA
- CCA
  - Neuroscience example
- Latent Dirichlet Allocation

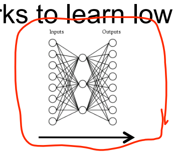
Readings:

-

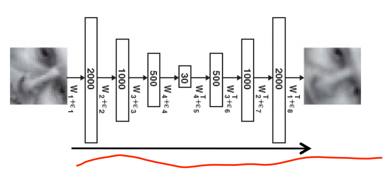
Deep Belief Networks

[Hinton & Salakhutdinov, *Science,* 2006]

- Problem: training networks with many hidden layers doesn't work very well
  - local minima, very slow training if initialize with zero weights
- Deep belief networks
  - autoencoder networks to learn low dimensional encodings

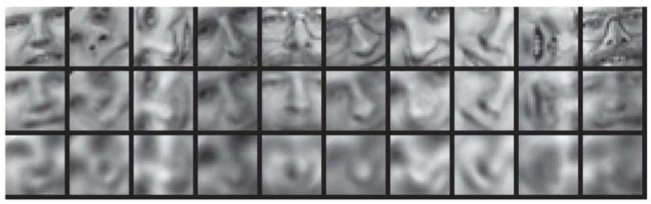


  - but more layers, to learn better encodings



Deep Belief Networks

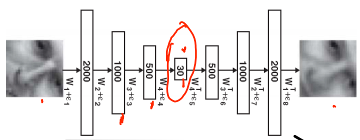
[Hinton & Salakhutdinov, 2006]



original image

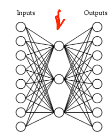
reconstructed from 2000-1000-500-30 DBN

reconstructed from 2000-300, linear PCA



logistic transformations

versus



linear transformations

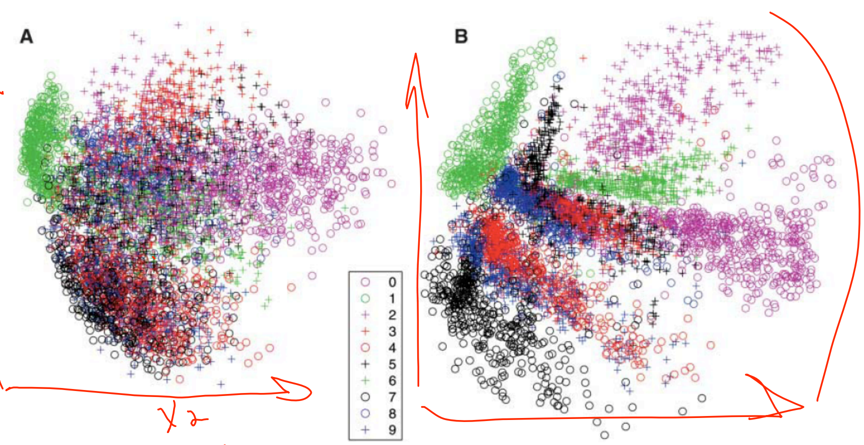
PCA

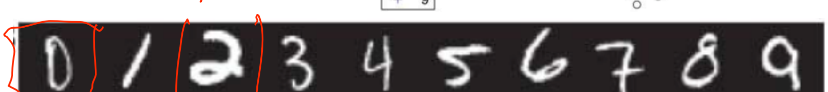
Encoding of digit images in two dimensions

[Hinton & Salakhutdinov, 2006]

784-2 linear encoding (PCA)

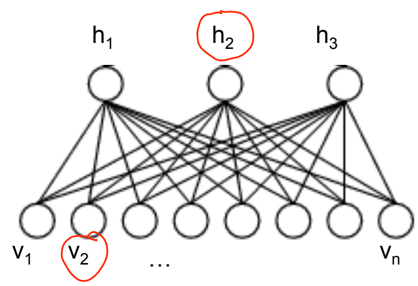
784-1000-500-250-2 DBNet





Restricted Boltzman Machine

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



$$P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp(\sum_i w_{ij} v_i)}$$

$$P(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(\sum_j w_{ij} h_j)}$$

[Hinton & Salakhutdinov, 2006]

Deep Belief Networks: Training

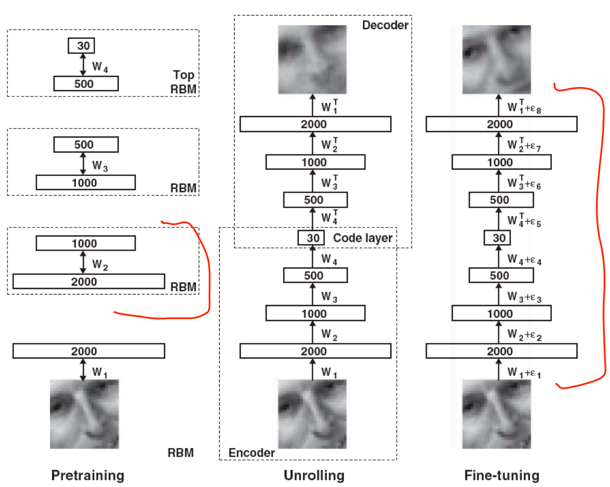


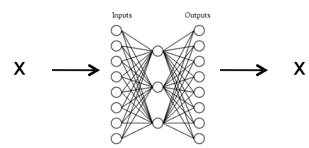
Fig. 1. Pretraining consists of learning a stack of restricted Boltzmann machines (RBMs), each having only one layer of feature detectors. The learned feature activations of one RBM are used as the "data" for training the next RBM in the stack. After the pretraining, the RBMs are "unrolled" to create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

Independent Components Analysis (ICA)

- PCA seeks orthogonal directions $<Y_1 \ldots Y_M>$ in feature space X that minimize reconstruction error

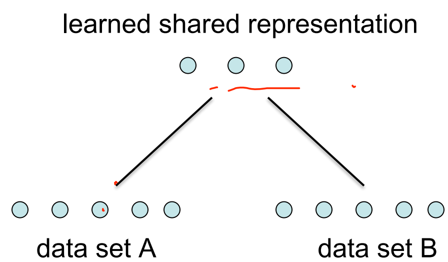
- ICA seeks directions $<Y_1 \ldots Y_M>$ that are most *statistically independent*. I.e., that minimize $I(Y)$, the mutual information between the Y_j :

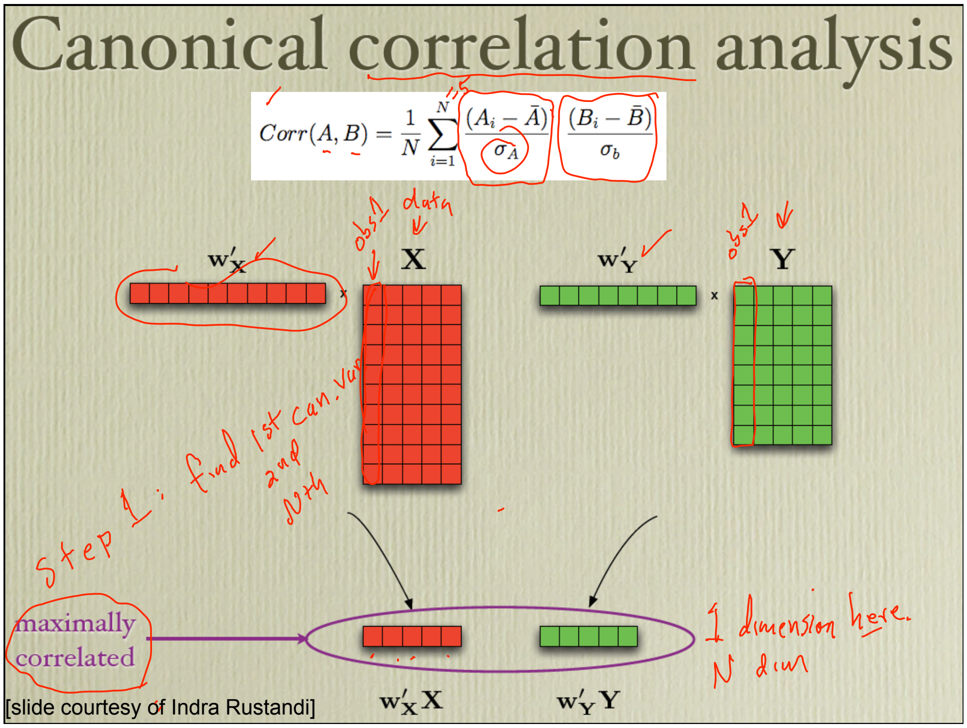
$$I(Y) = \left[\sum_{j=1}^{J} H(Y_j) \right] - H(Y)$$

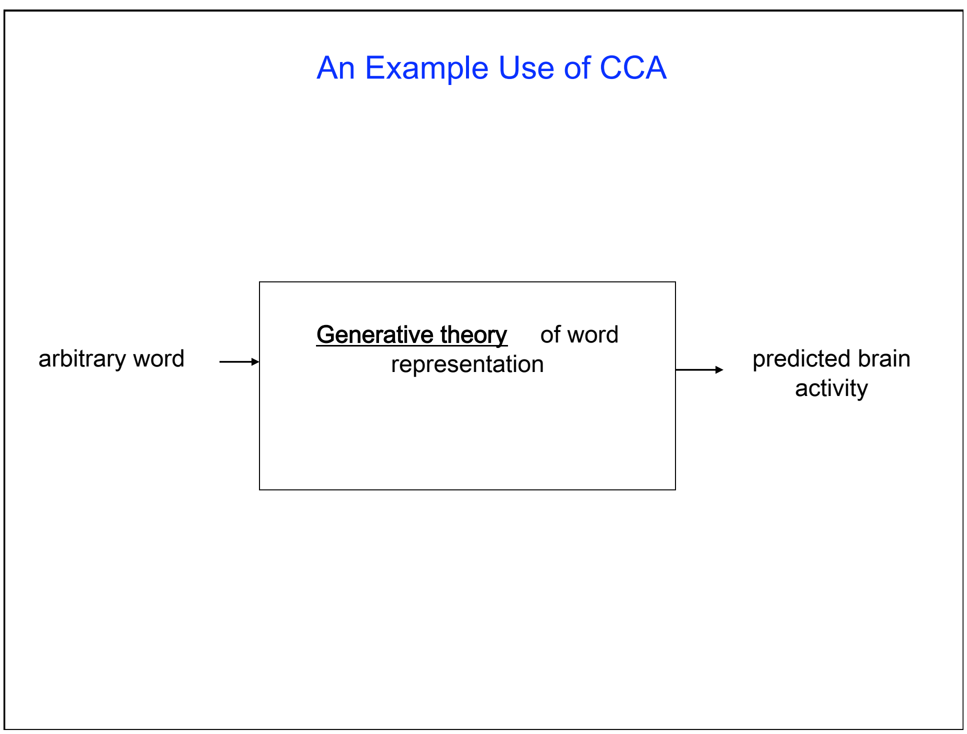


Dimensionality reduction across multiple datasets

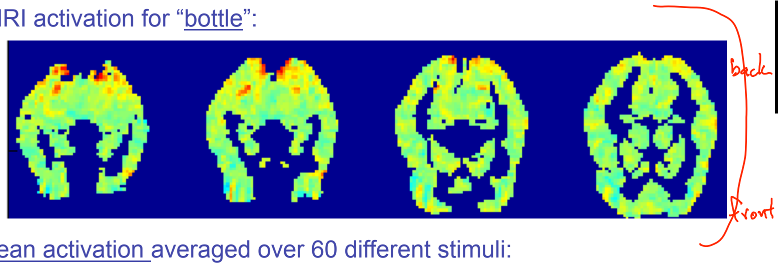
- Given data sets A and B, find linear projections of each into a common lower dimensional space!
  - Generalized SVD: minimize sq reconstruction errors of both
  - Canonical correlation analysis: maximize correlation of A and B in the projected space

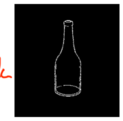



Canonical correlation analysis
$Corr(A,B) = \frac{1}{N}\sum_{i=1}^{N}\frac{(A_i - \bar{A})}{\sigma_A}\frac{(B_i - \bar{B})}{\sigma_b}$
$\mathbf{w}'_\mathbf{X}$
X
$\mathbf{w}'_\mathbf{Y}$
Y
x
x
maximally correlated
[slide courtesy of Indra Rustandi]
$\mathbf{w}'_\mathbf{X}\mathbf{X}$
$\mathbf{w}'_\mathbf{Y}\mathbf{Y}$
data
Step 1: find 1st can. var
2nd
Nth
1 dimension here.
N dim


An Example Use of CCA
arbitrary word
Generative theory of word representation
predicted brain activity

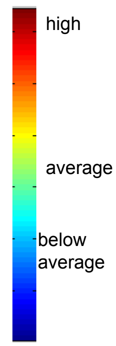
fMRI activation for "bottle":



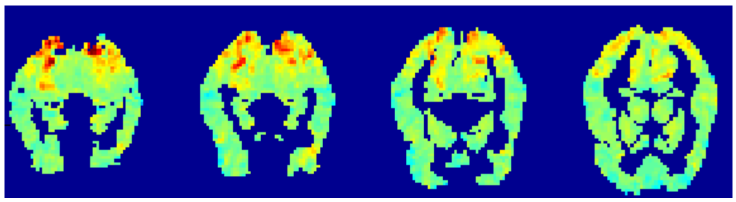


bottle

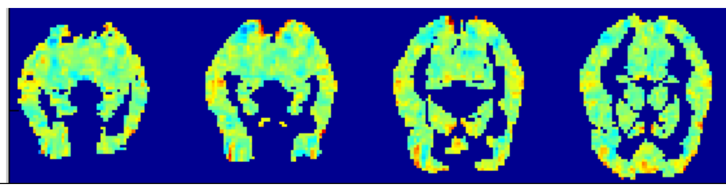
fMRI activation



Mean activation averaged over 60 different stimuli:

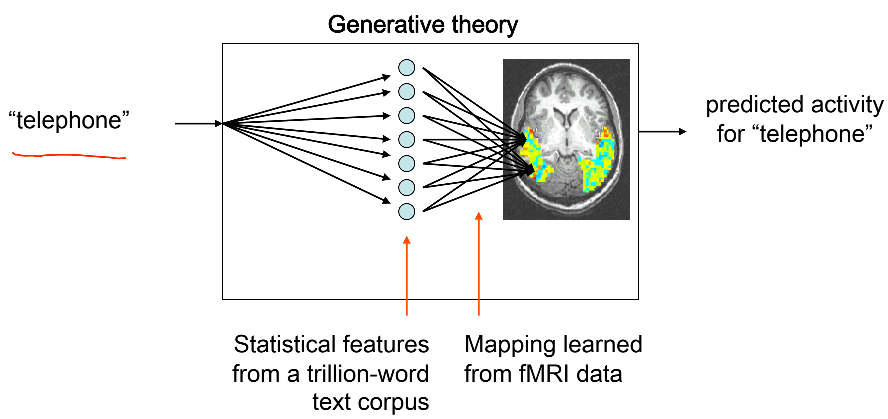


"bottle" minus mean activation:



Idea: Predict neural activity from corpus statistics of stimulus word

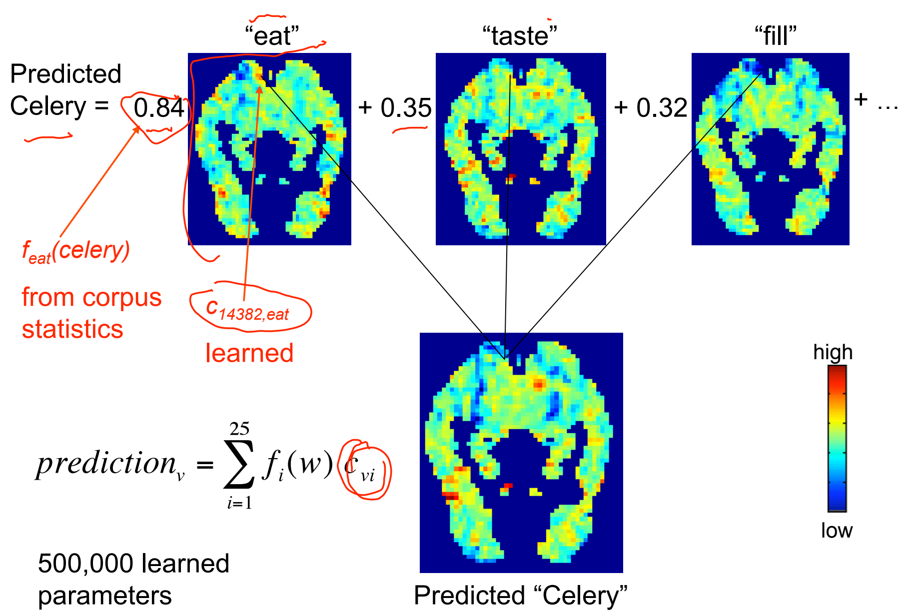
[Mitchell et al., *Science*, 2008]

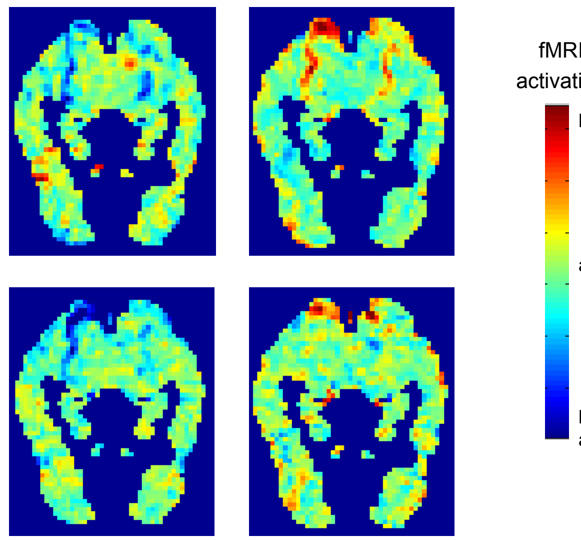


Semantic feature values:
“**celery**”
0.8368, eat
0.3461, taste
0.3153, fill
0.2430, see
0.1145, clean
0.0600, open
0.0586, smell
0.0286, touch
…
…
0.0000, drive
0.0000, wear
0.0000, lift
0.0000, break
0.0000, ride

Semantic feature values:
“**airplane**”
0.8673, ride
0.2891, see
0.2851, say
0.1689, near
0.1228, open
0.0883, hear
0.0771, run
0.0749, lift
…
…
0.0049, smell
0.0010, wear
0.0000, taste
0.0000, rub
0.0000, manipulate

Predicted Activation is Sum of Feature Contributions

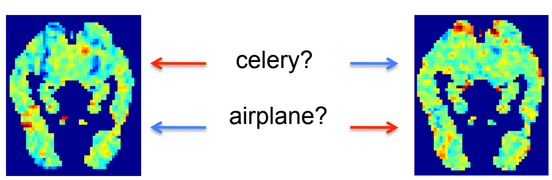




Predicted and observed fMRI images for "celery" and "airplane" after training on 58 other words.

Evaluating the Computational Model

- Train it using 58 of the 60 word stimuli
- Apply it to predict fMRI images for other 2 words
- Test: show it the observed images for the 2 held-out, and make it predict which is which

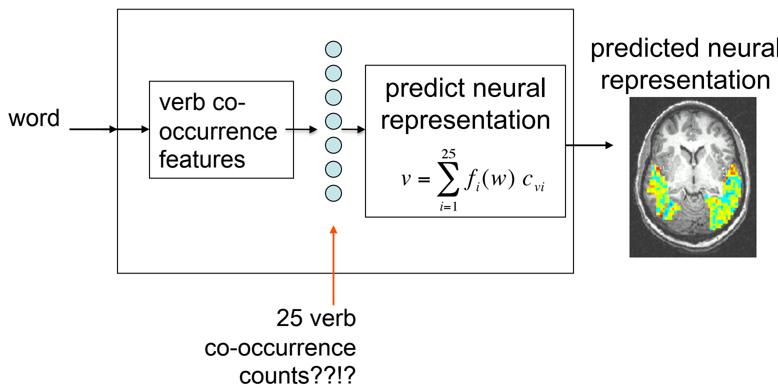


1770 test pairs in leave-2-out:

- Random guessing → 0.50 accuracy
- Accuracy above 0.61 is significant ($p<0.05$)

Mean accuracy over 9 subjects: 0.79

Q4: What are the actual semantic primitives from which neural encodings are composed?

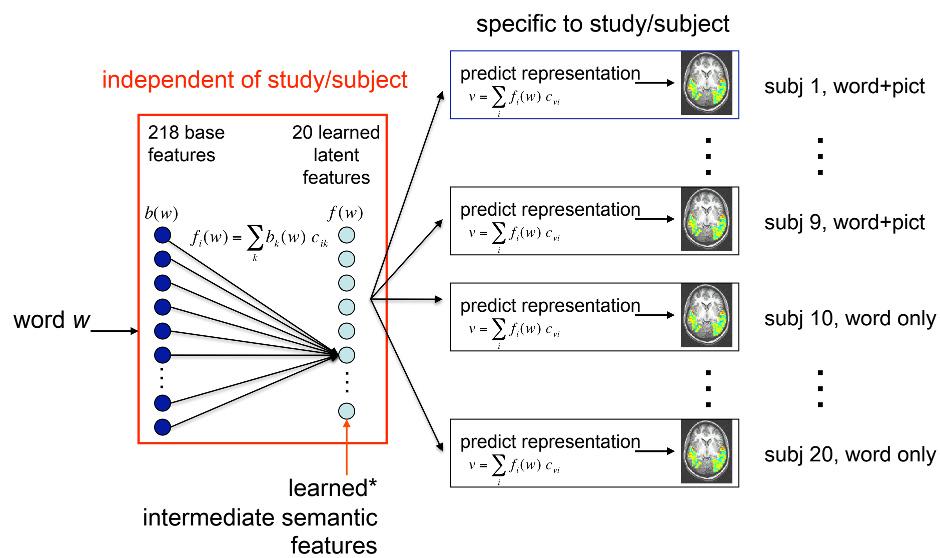


Alternative semantic feature sets

PREDEFINED corpus features	Mean Acc.
25 verb co-occurrences	.79
486 verb co-occurrences	.79
50,000 word co-occurences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78
218 features collected using *Mechanical Turk*	**.83**
20 features discovered from the data**	**.87**

* developed by Dean Pommerleau
** developed by Indra Rustandi

Discovering shared semantic basis [Rustandi et al., 2009]



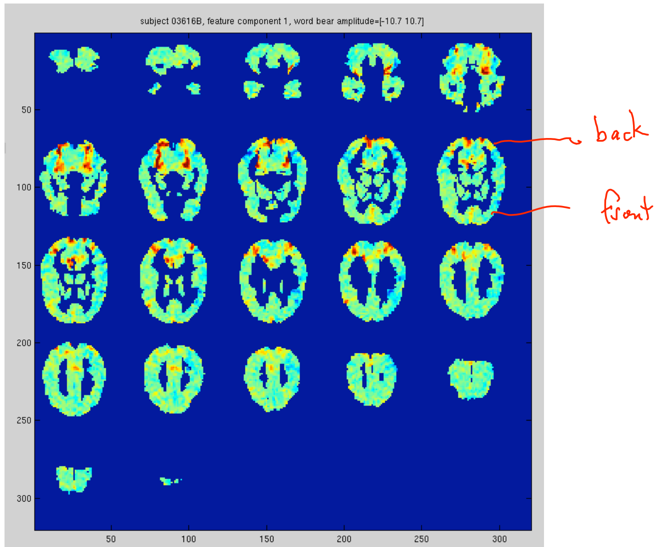
* trained using Canonical Correlation Analysis

Multi-study (WP+WO) Multi-subject (9+11) CCA
Top Stimulus Words

	component 1	component 2	component 3	component 4
most active stimuli	apartment church closet house barn	screwdriver pliers refrigerator knife hammer	telephone butterfly bicycle beetle dog	pants dress glass coat chair

shelter? manipulation? things that touch me?

Subject 1 (Word-Picture stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 1



Subject 1 (Word-ONLY stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 1