Deep Learning in Speech and Language Processing

Dong Wang
CSLT, RIIT, Tsinghua Univ.

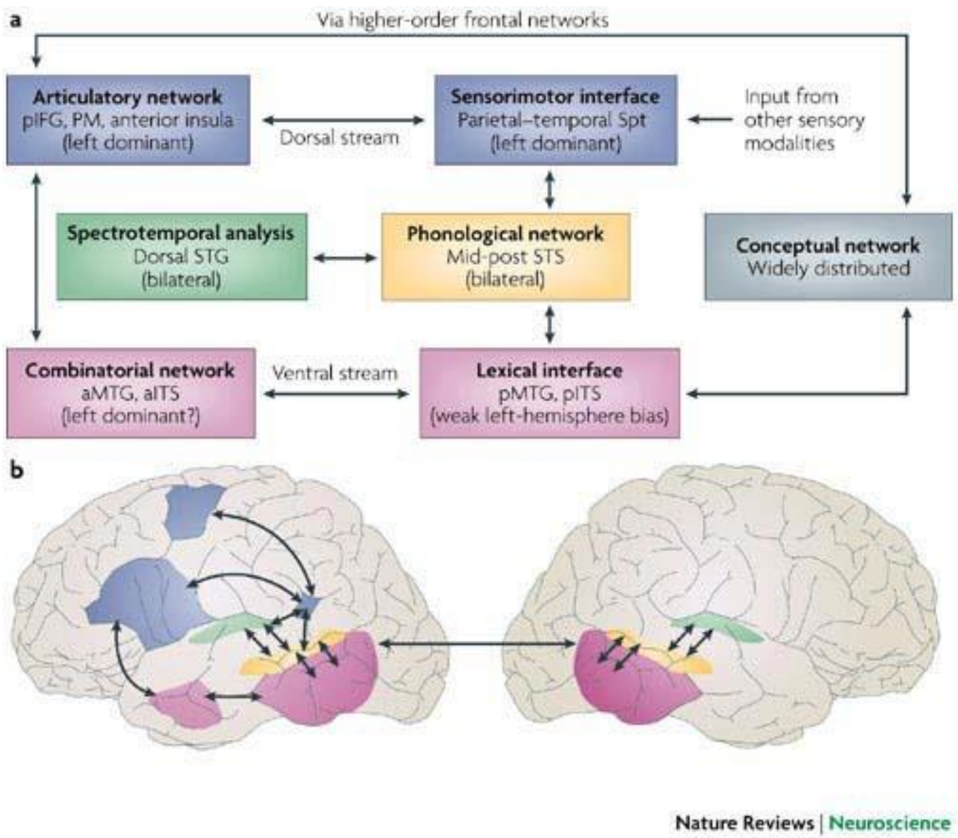
2014-10-10

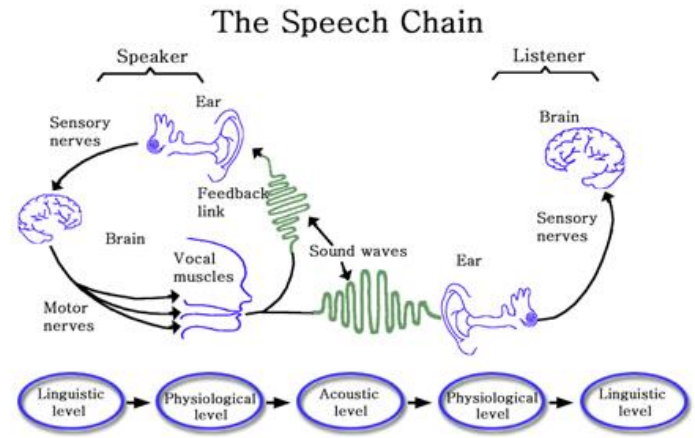
Contents

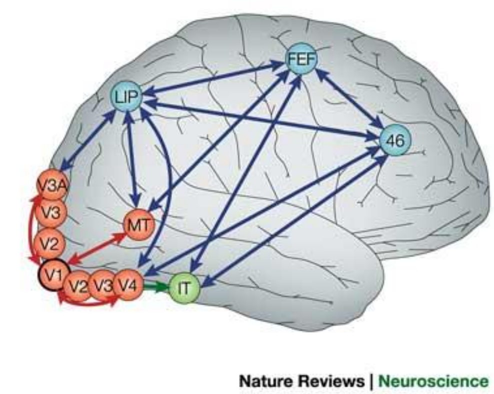
Contents

- **Introduction to deep learning**
- Deep learning in speech processing
- Deep learning in language processing

Our brain is hierarchical…



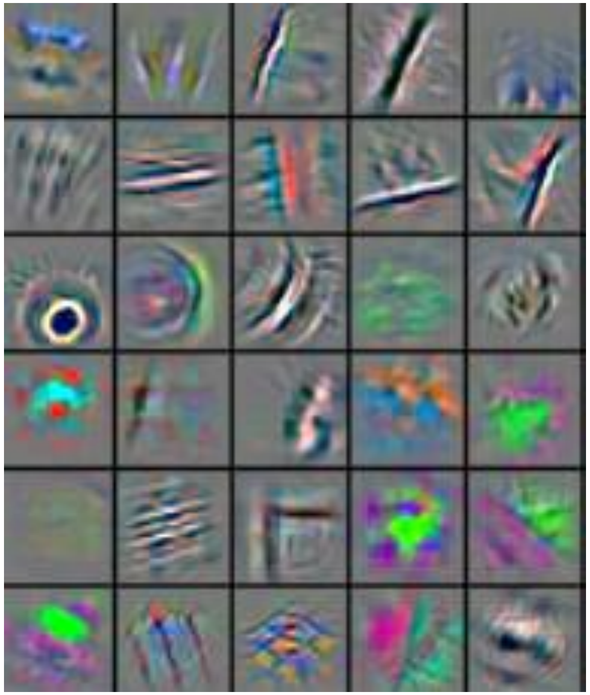




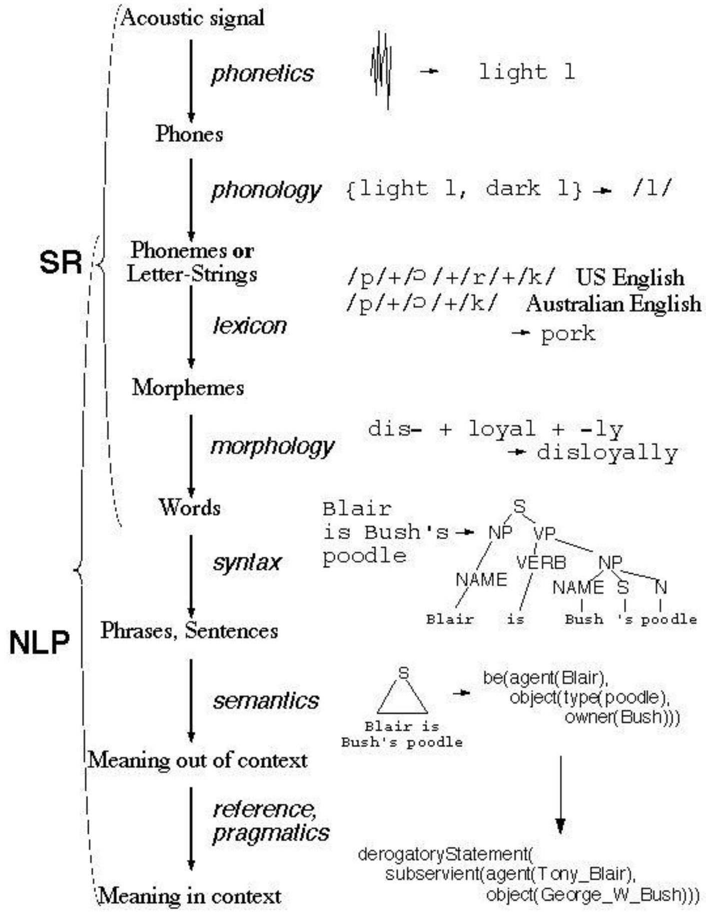
- Primary visual cortex and visual awareness, Frank Tong, Nature Reviews Neuroscience 4, 219-229 (March 2003)
- MIT opencoursewave, Syllabus of Laboratory on the Physiology, Acoustics, and Perception of Speech
- The cortical organization of speech processing, Gregory Hickok & David Poeppel, Nature Reviews Neuroscience 8, 393-402 (May 2007)

Hierarchical processing in CV

Low-level feature → middle-level representation → high-level object →

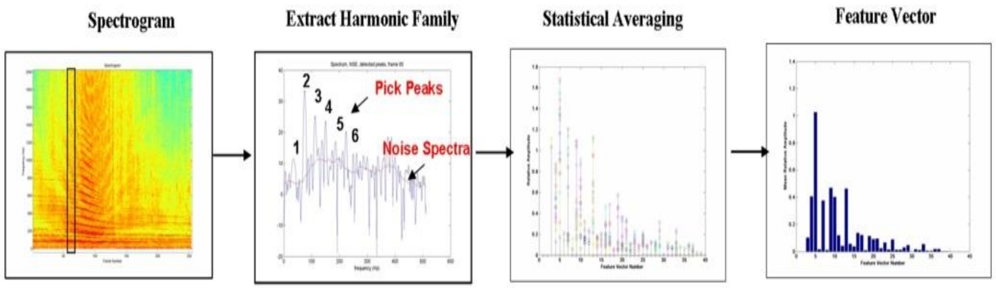


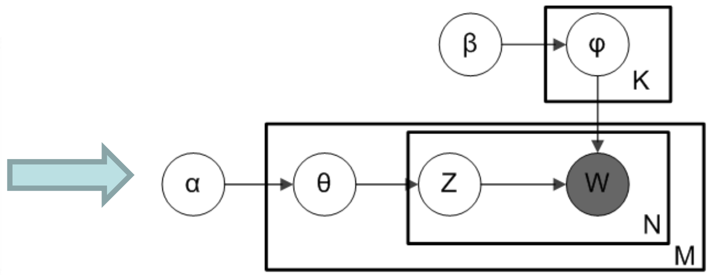
Hierarchical processing in speech & language

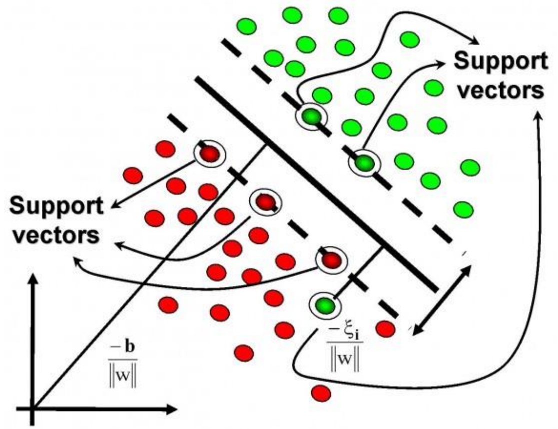


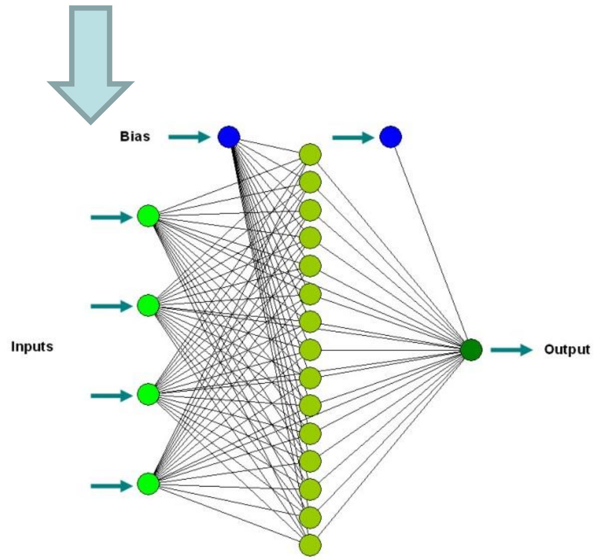
http://www.cse.unsw.edu.au/~billw/cs9414/notes/nlp/nlp-intro/nlp-intro-2007.html

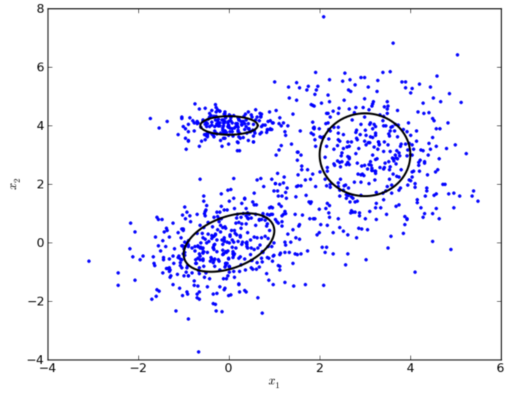
Most of the popular models are shallow







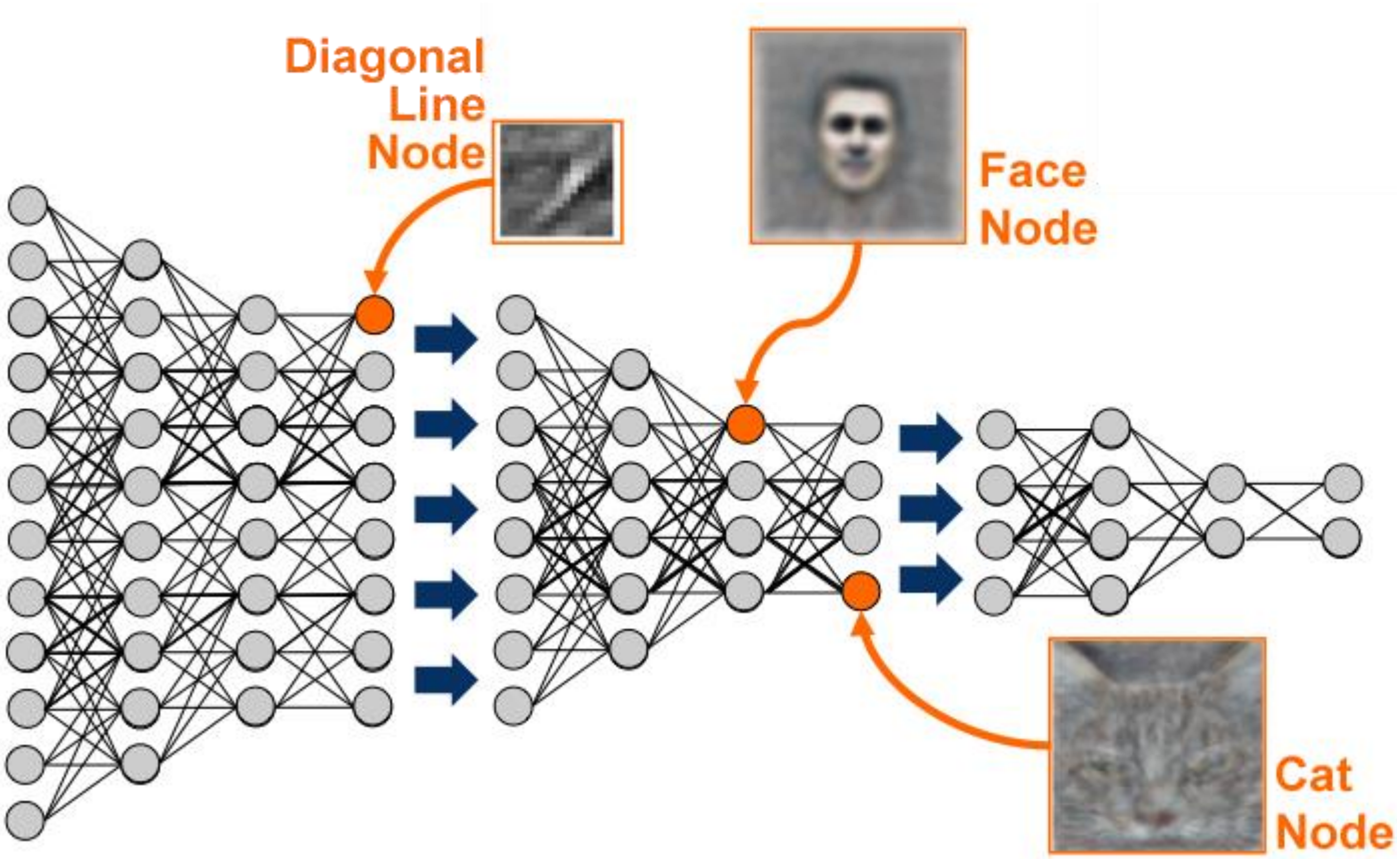




***** Detecting, Tracking, and Identifying Airborne Threats with Netted Sensor Fence, Weiqun Shi, Gus Arabadjis, Brett Bishop, Peter Hill, Rich Plasse and John Yoder

* Neural network topology. Topology of multilayer full feedforward neural network for the estimation of lipase-catalyzed synthesis of palm-based wax ester. Basri *et al. BMC Biotechnology* 2007 **7**:53

Now make the models deep

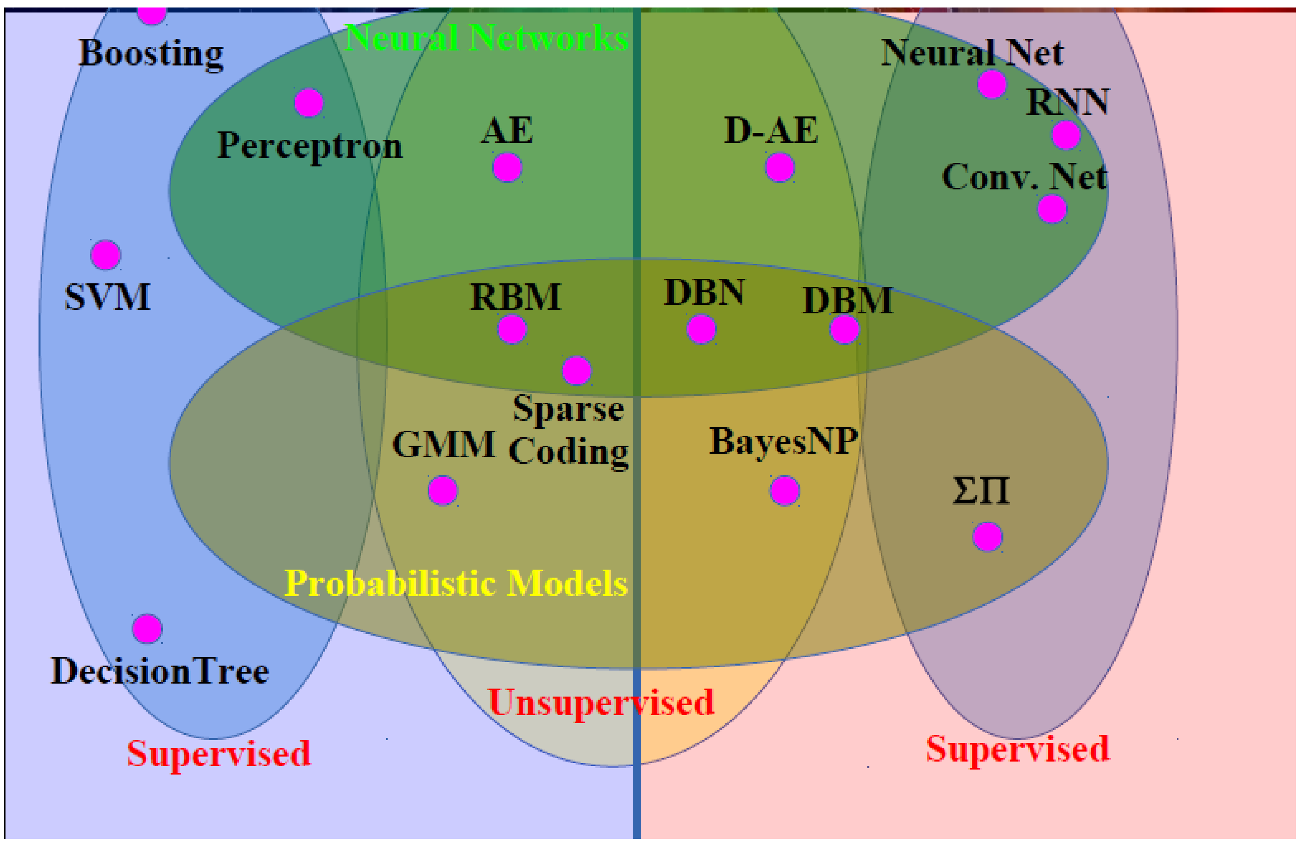


http://theanalyticsstore.com/deep-learning/

What models are deep?

- Most of the current ML models are shallow
  - LR, SVM, Neural perceptron, matrix factorization, GMM, HMM, DT, GP, RBM …
- Neural networks with 1 hidden layer is not deep
- NN with more than 2 hidden layers are deep
- Other deep models: HLDA, HDP

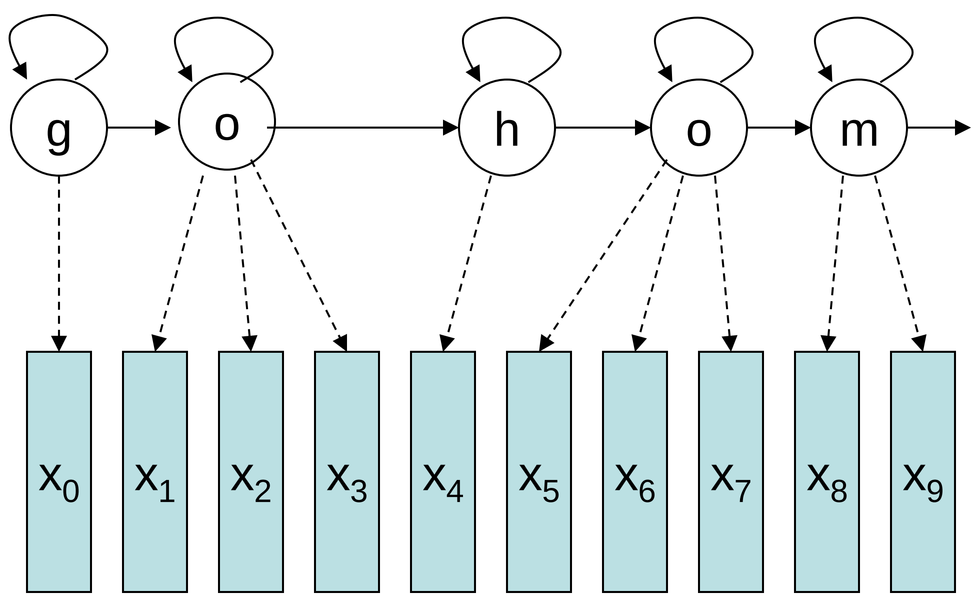
What are deep models



Yann LeCun, Marc'Aurelio Ranzato, Deep Learning Tutorial, ICML, Atlanta, 2013-06-16

Deep learning and graphical models

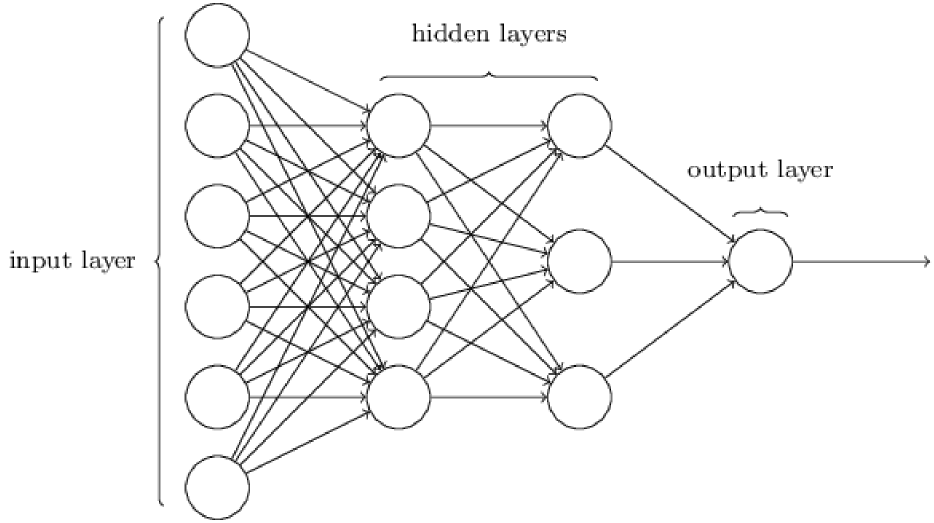
- Graphical models can be deep, but most of them are not. Deep models can be probabilistic, but not necessary.
- A vertex (random variable) can be inferred from a deep structure, e.g., HMM+DNN hybrid model



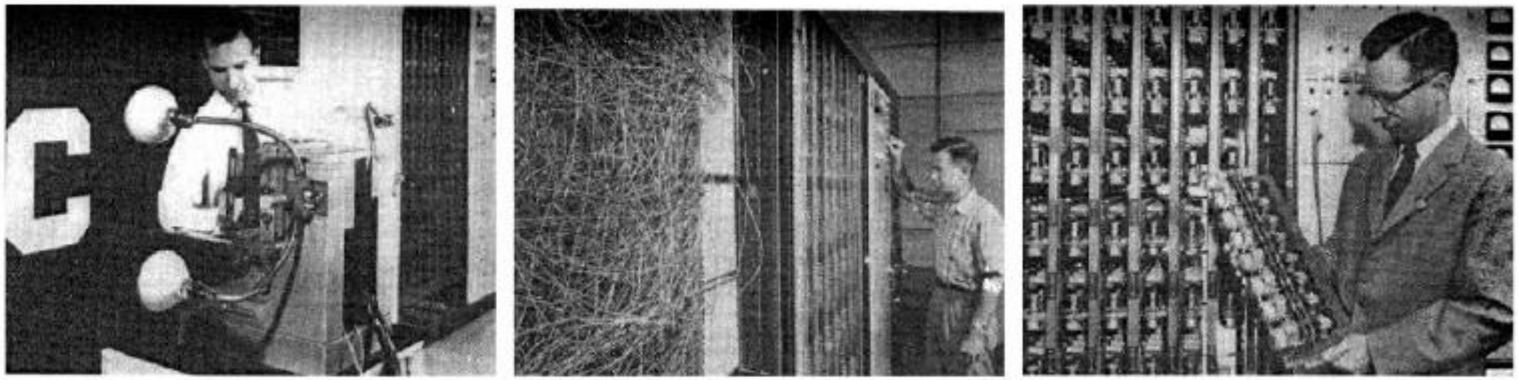
Deep neural networks (DNN)

- Deep structure can be any model with more than 2 hidden layers (stacked RBMs, hierarchical Bayesian models).
- Simple models, such as NN, are always preferred!

$$y=F(W^{K} . F (W^{K-1}. F (\ldots F(W^{0}.X)\ldots)))$$



Some story of neural networks

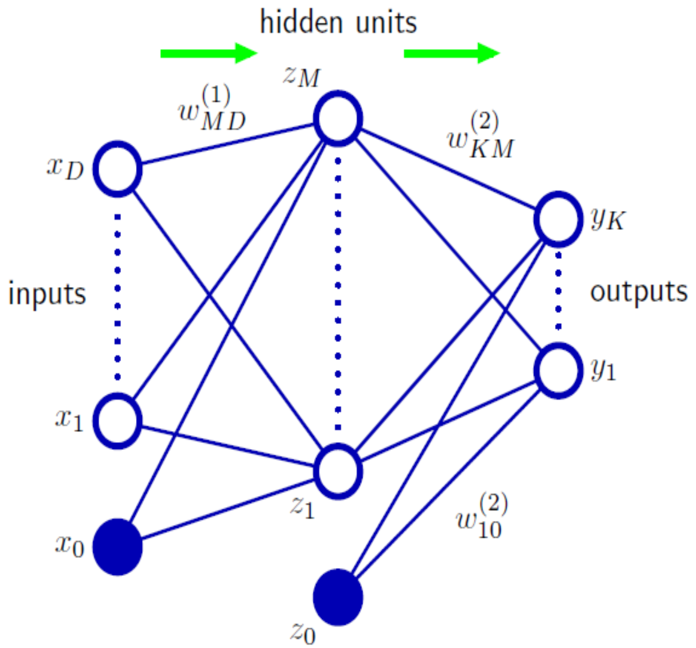


Frank Rosenblatt (11 July 1928 – 11 July 1971) was a New York City born psychologist who completed the Perceptron, or MARK 1, computer at Cornell University in 1960. This was the first computer that could learn new skills by trial and error, using a type of neural network that simulates human thought processes.

1969 Marvin Minsky and Seymour Papert published the book *Perceptrons. The first generation of NN was ended.*

Renaissance in the mid-1980s

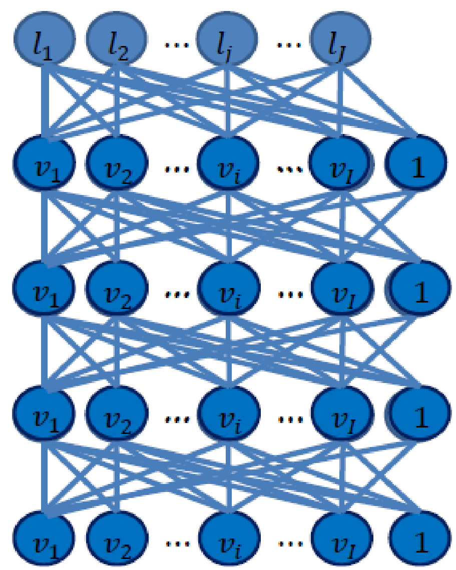
- Multiple layer perceptron (MLP) became popular, with the standard BP training.
- With an appropriate active function, and sufficient hidden units, a 1-hidden-layer MLP can approximate any continuous function to arbitrary accuracy.
- Many interests were invoked, but limited success, such as in speech recognition.
- Became a standard tool in ML but unpopular since mid of 90's.



Google books Ngram Viewer
raph these case-sensitive comma-separated phrases: neural networ
etween 1940 and 2008 from the corpus English with smoothing of 0 .
Search lots of books
Share
Tweet
neural network
0.00025%
0.0002%
0.00015%
0.0001%
0.00005%
0.00%
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005

Reactive after deep

- Deep NNs were found to be much more powerful than the shallow counterpart.
- Quickly became hot in many fields
  - Speech recognition
  - Speech synthesis
  - Image processing
  - NLP
  - …



Some activities in deep learning

- 2008 NIPS deep learning workshop
- 2009 NIPS workshop on deep learning for speech recognition and related applications
- 2009 ICML workshop on learning feature hierarchies
- 2011 ICML workshop on learning architectures, representations, and optimization for speech and visual information processing
- 2012 ICASSP tutorial on deep learning for signal and information processing
- 2012 ICML workshop on representation learning
- 2012 special section on deep learning for speech and language processing in IEEE transactions on audio speech and language processing
- 2010,2011,2012 NIPS workshops on deep learning and unsupervised feature learning

Some activities in deep learning

- 2013 NIPS workshops on deep learning and on output representation learning
- 2013 special issue on learning deep architectures in IEEE transactions on pattern analysis and machine intelligence
- 2013 international conference on learning representations
- 2013 ICML workshop on representation learning challenges
- 2013 ICML workshop on deep learning for audio, speech, and language processing
- 2013 ICASSP special session on new types of deep neural network learning for speech recognition and related applications
- 2014 ICASSP special session on deep learning for music
- 2014 ICML workshop on Deep Learning Models for Emerging Big Data Applications
- 2014 ICML workshop on Knowledge-Powered Deep Learning for Text Mining

Deep Networks

- **A confusion: why deeper, not fatter?**

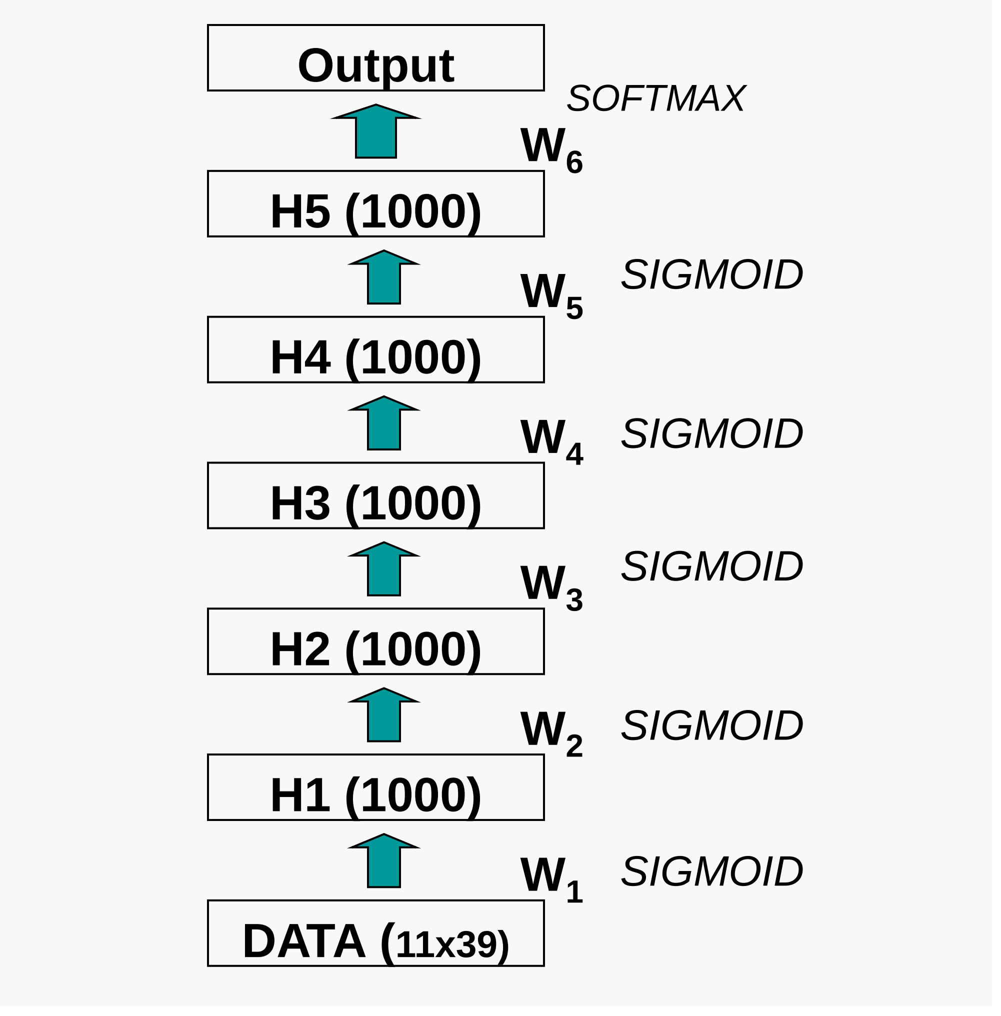
  - ***Deep architectures*** **can be very useful** in order to learn the complicated functions that represent hierarchical abstractions (like in vision, speech, language)

  - ***Insufficient depth*** can require more computational elements and produce worse performances than architectures whose depth matches to the task

Deep Networks

- **Problems with the deep?**
  - In many cases, deep nets are **hard to optimize**
  - Standard **back-propagation can get trapped into poor local minima** when training deep networks
  - The reactive of NN, or DNN, is largely attributed to the better initialization approach started by Goeffery Hinton

Deep by stacking

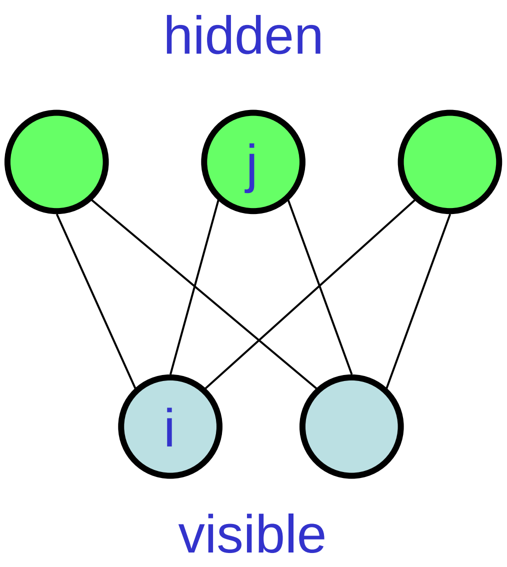


Hinton, G. E. (2002) **Training Products of Experts by Minimizing Contrastive Divergence.** Neural Computation, 14, pp 1771-1800.

Hinton, G. E., Osindero, S. and Teh, Y. **A fast learning algorithm for deep belief nets.** Neural Computation 18, pp 1527-1554.

RBM as a stochastic generative model

- A RBM is a bidirectional fully connected network:
  - one layer of visible input units.
  - one layer of binary stochastic hidden units
  - No connections between hidden units.
- In an RBM, the input units generate the hidden units, and the hidden units generate (reproduce) the input values
- RBM is a undirected graph, and is a generative model.



$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{i,j} h_j$$

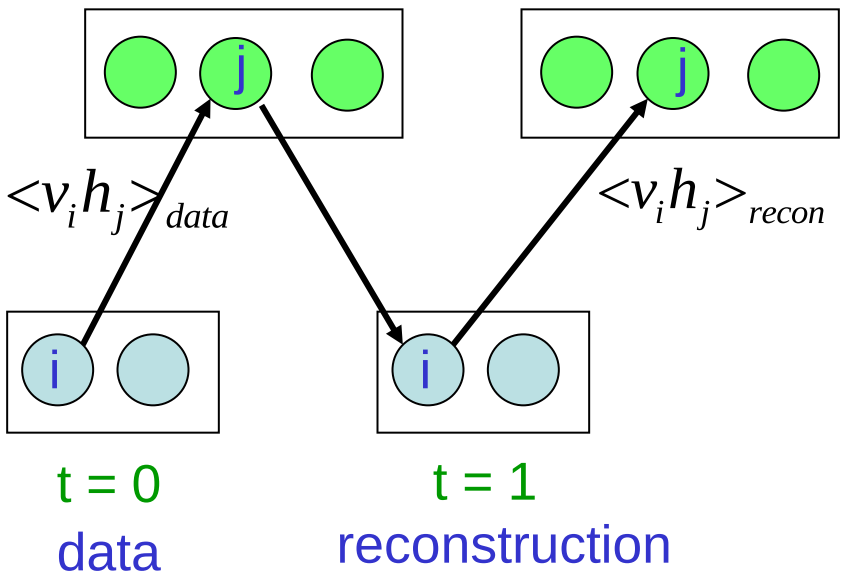
$$P(v) = \frac{1}{Z} \sum_h e^{-E(v,h)}$$

$$P(v|h) = \prod_{i=1}^{m} P(v_i|h) \qquad P(h|v) = \prod_{j=1}^{n} P(h_j|v)$$

http://www.cs.toronto.edu/~hinton/adi/index.htm

Contrastive Divergence for RBM

(Hinton, 2002)

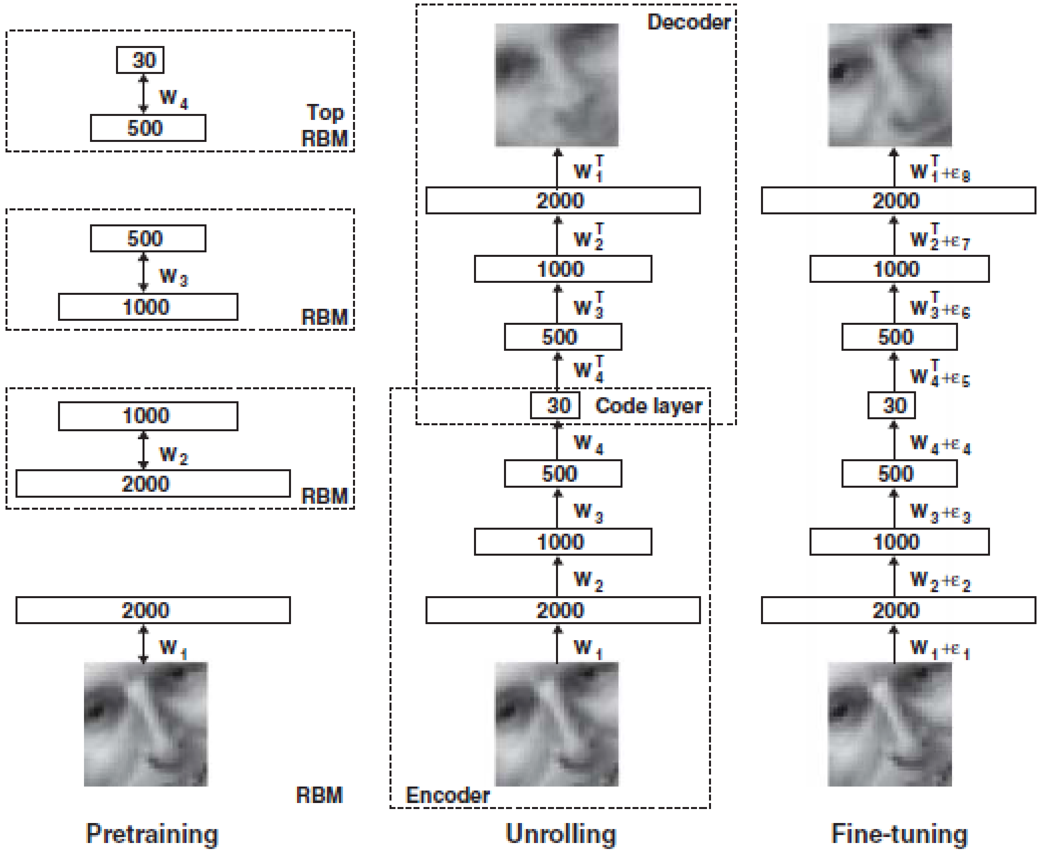


1. Start by setting a training vector on the visible units.
2. Update all the hidden units in parallel by sampling
3. Update all the visible units in parallel to get a “reconstruction”.
4. Update the hidden units again.
5. Update weights with the following rule, that attempts to minimize the way the model distorts the data:

$$\Delta w_{ij} = \varepsilon \left(<v_i h_j>_{data} - <v_i h_j>_{recon} \right)$$

G. Hinton, A Practical Guide to Training Restricted Boltzmann Machines – August 2010

A training recipe



- Hinton, G. E. and Salakhutdinov, R. R, **Reducing the dimensionality of data with neural networks.** Science, Vol. 313. no. 5786, pp. 504 - 507, 28 July 2006.

That is just the beginning

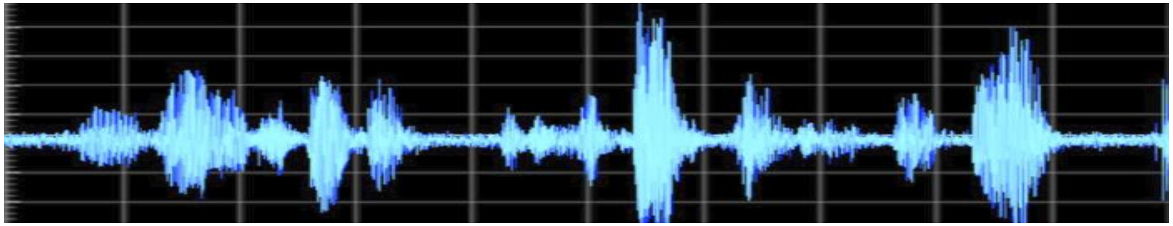
- We now know that the RBM pre-training is not very necessary (scale, discriminative stacking, boosting…).
- We now know that there are rich deep structures that can deliver significant performance gains (CNN, RNN, echo net, BN, DSN, reLU, maxout, pnorm…).
- We now know that there are multitude of approaches for DNN optimizations that boost performance and robustness (DT, drop out, noisy training, sparse, BN tuning, SVD…).
- We now know that many interesting ways to conduct adaptation (linear transform, discriminative transform, i-vector, KL,…)
- We now know that DNN is suitable to learn multi-conditions, multi-tasks.
- ….

Contents

- Introduction to deep learning
- **Deep learning in speech processing**
- Deep learning in language processing

Why speech enjoys deep learning?

- Speech signals are rather raw, and we need to extract informative features.
- Speech signals are full of and noise variations, a robust feature learning is required.
- Speech signals are naturally hierarchical.



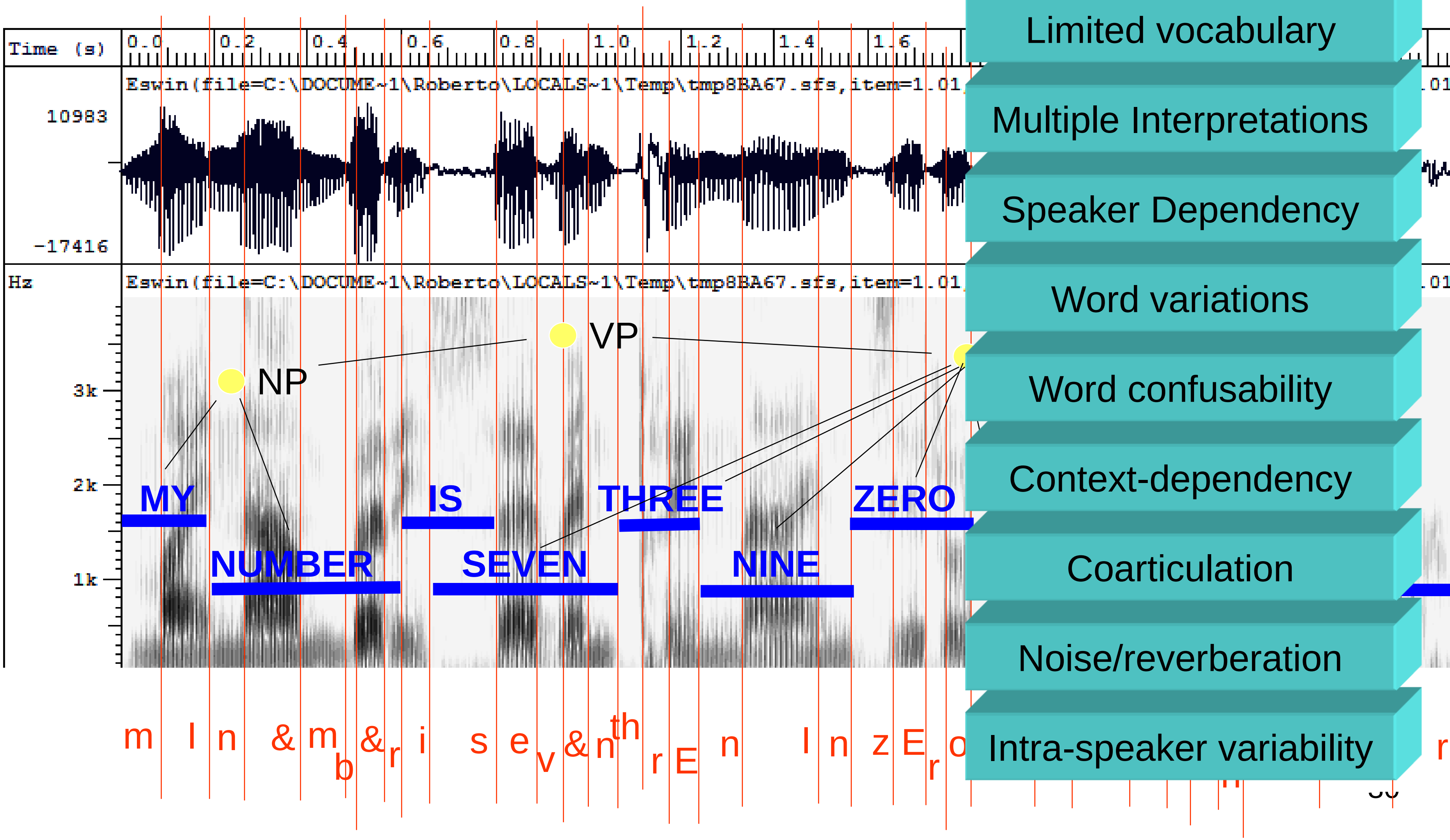
DNN in speech processing

- Speech recognition: over 30% WER reduction!
- Speech synthesis: still young but highly promising.
- Speaker recognition: state-of-the-art by DNN.
- Music genre classification, music recommendation.

DNN in speech recognition

- DNN has become the-state-of-the-art for speech recognition since 2010.
- Alost all the current recognition systems are based on DNN, and it is found that this model is more powerful, more stable, and less saturated with a large amount of data, when compared with conventional approaches.

What is speech recognition

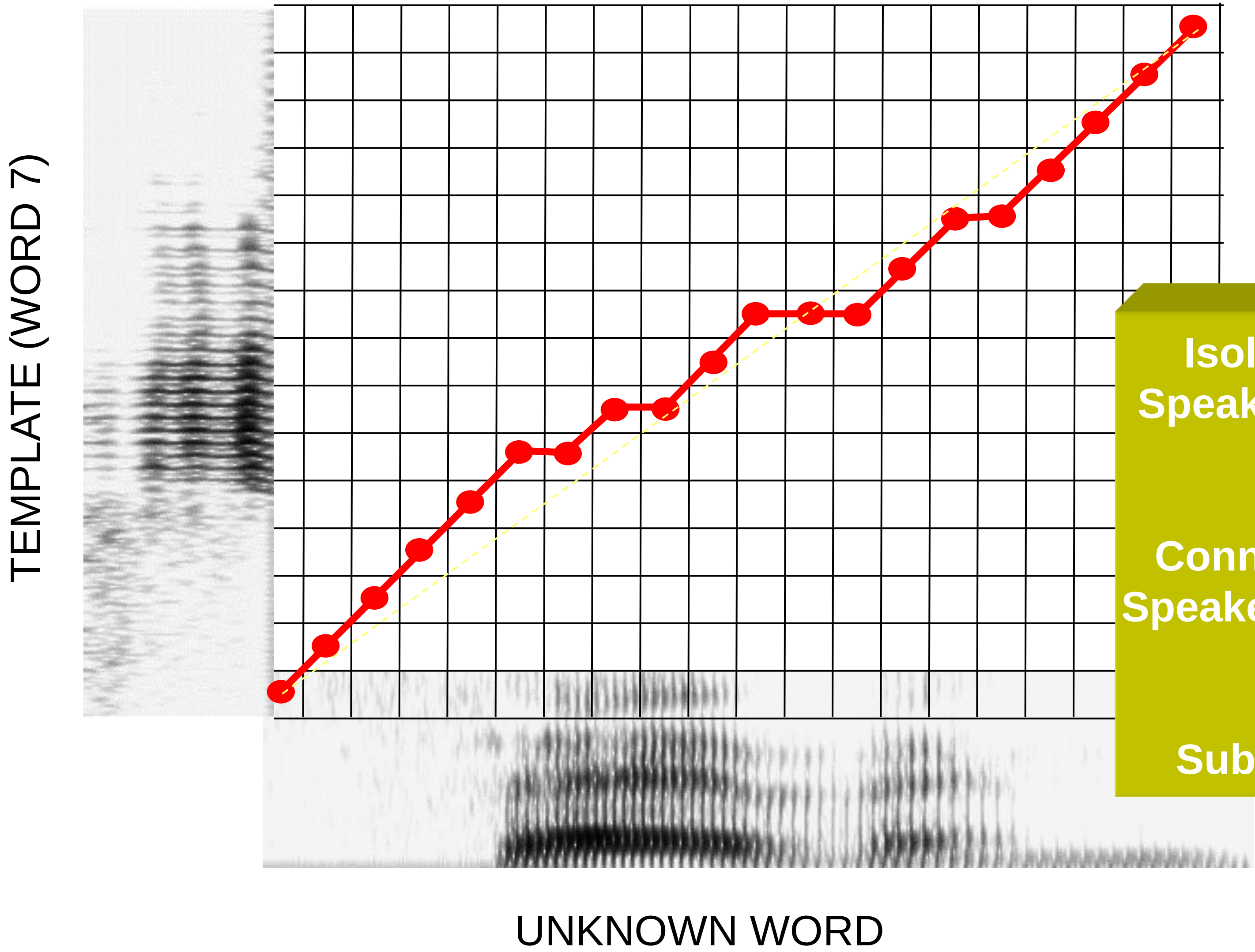


- Ellipses and Anaphors
- Limited vocabulary
- Multiple Interpretations
- Speaker Dependency
- Word variations
- Word confusability
- Context-dependency
- Coarticulation
- Noise/reverberation
- Intra-speaker variability

Key issues

- How to represent speech (**ok**)
- How to model acoustic/speaker/environment variability (**ok**)
- How to model lexical/linguistic knowledge (**ok**)
- How to model high-level knowledge (**not solved yet**)
- How to train the models (**depends**)
- How to search (**ok**)
- How to solve practical problems, e.g., spontaneous, hesitation, overlap, noise, new words, context and discourse… (hard and complex, can we solve?)

1970s – Dynamic Time Warping
The Brute Force of the Engineering Approach



T.K. Vyntsyuk (1968)
H. Sakoe,
S. Chiba (1970)

Isolated Words
Speaker Dependent

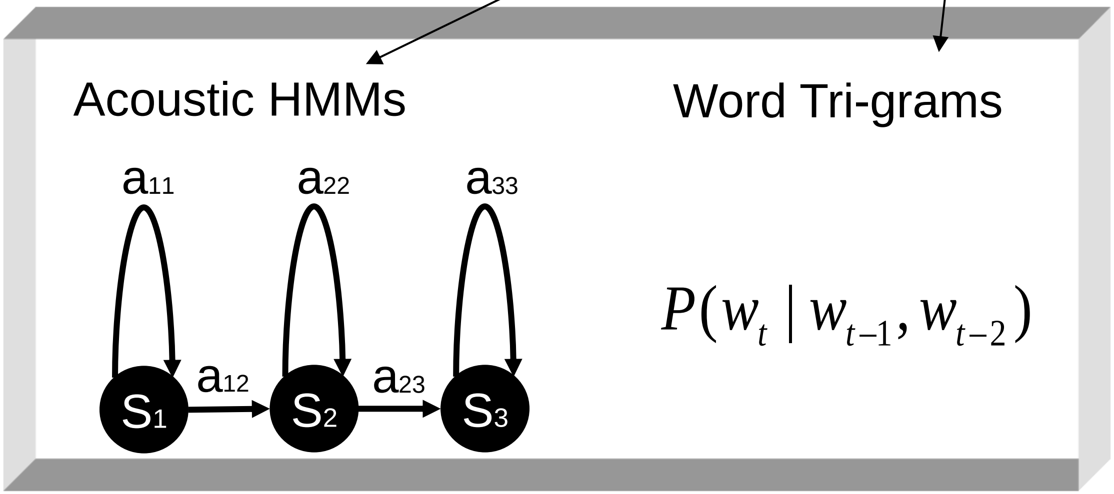
Connected Words
Speaker Independent

Sub-Word Units

1980s -- The statistical approach

- Based on work on Hidden Markov Models done by Leonard Baum at IDA, Princeton in the late 1960s
- Purely statistical approach pursued by Fred Jelinek and Jim Baker, IBM T.J.Watson Research
- Foundations of modern speech recognition engines
- Tasks: discrete recognition

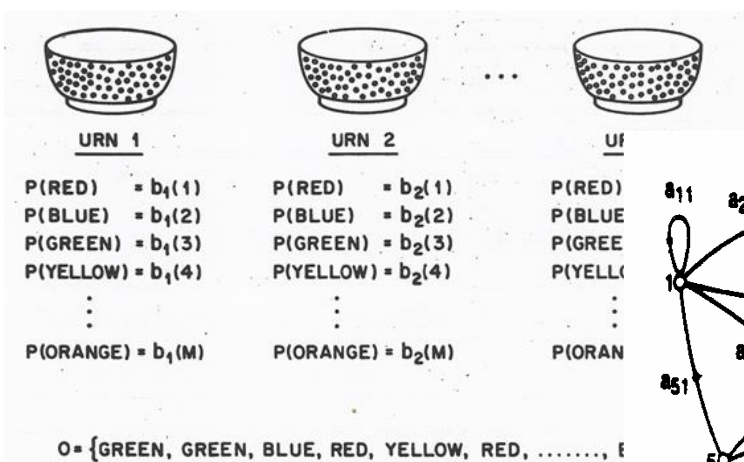
$$\hat{W} = \arg\max_{W} P(A \mid W) P(W)$$

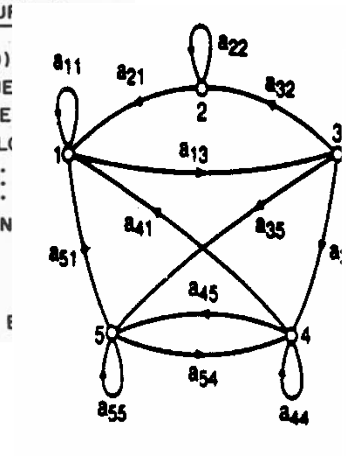


$$P(w_t \mid w_{t-1}, w_{t-2})$$

1990s – Statistical approach becomes ubiquitous

- Lawrence Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,* Proceeding of the IEEE, Vol. 77, No. 2, February 1989.
- *Key words: HMM, GMM, CSAR, MFCC, Viterbi, decision tree, n-gram, adaptation …*
- *Task: very large vocabulary ASR*





Markov Assumption:

$$P[q_t = j | q_{t-1} = i, q_{t-2} = k, \dots] = P[q_t = j | q_{t-1} = i]$$

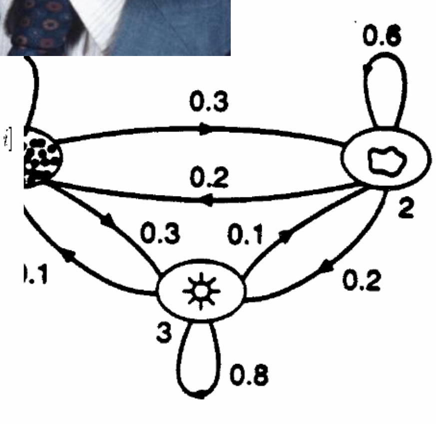
Set

$$a_{ij} = P[q_t = j | q_{t-1} = i] \quad 1 \le i, j \le N$$

Such that

$$a_{ij} \ge 0 \quad \forall i, j$$

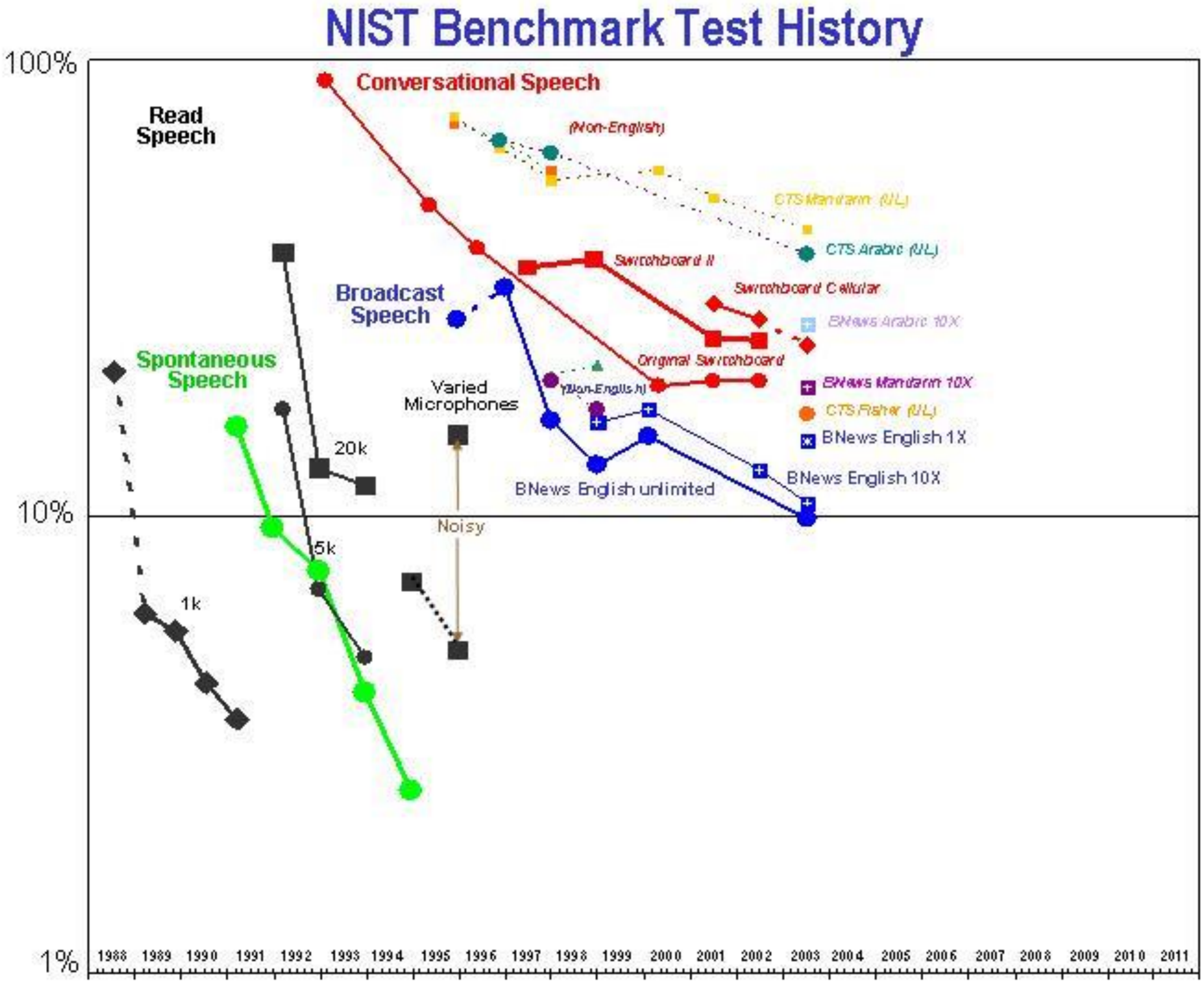
$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$$



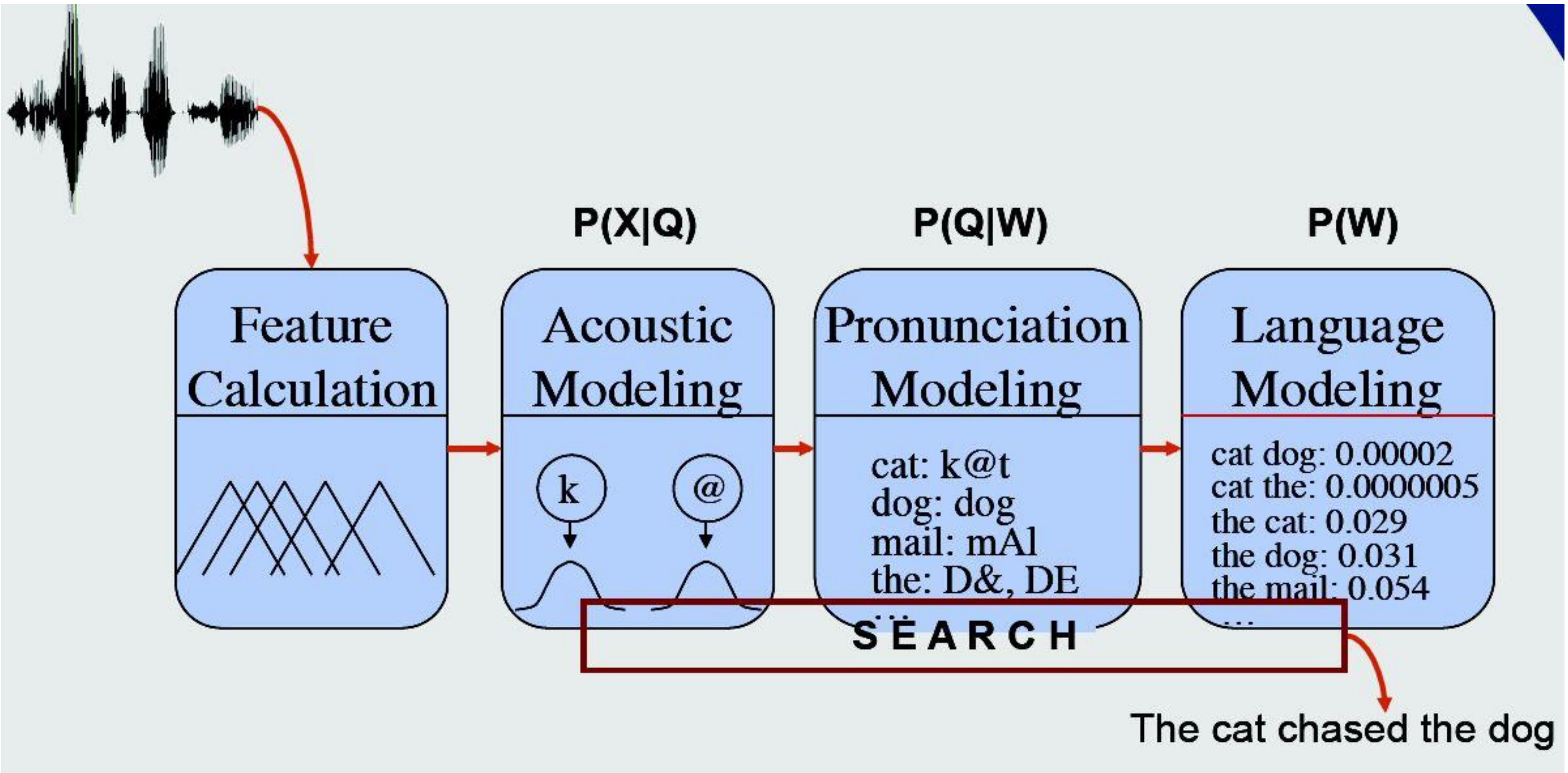
21th Century: Machine learning

- Research: Various ML models are tried on almost all the components
  - DT, LDA, Bayesian, FA, unsupervised learning
  - FST-based decoding
- Trends: cross-domain, knowledge integration
- Task: spontaneous ASR, multilingual ASR, ASR on web

Performance review

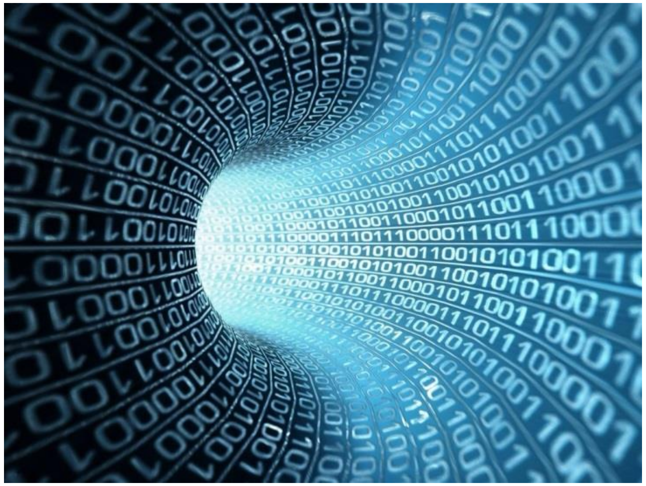


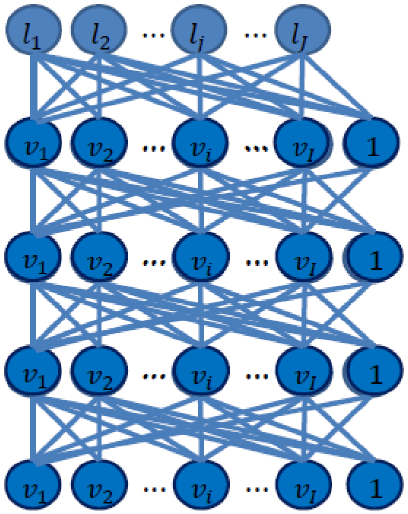
Pre state-of-the-art



2010s: never such hot …

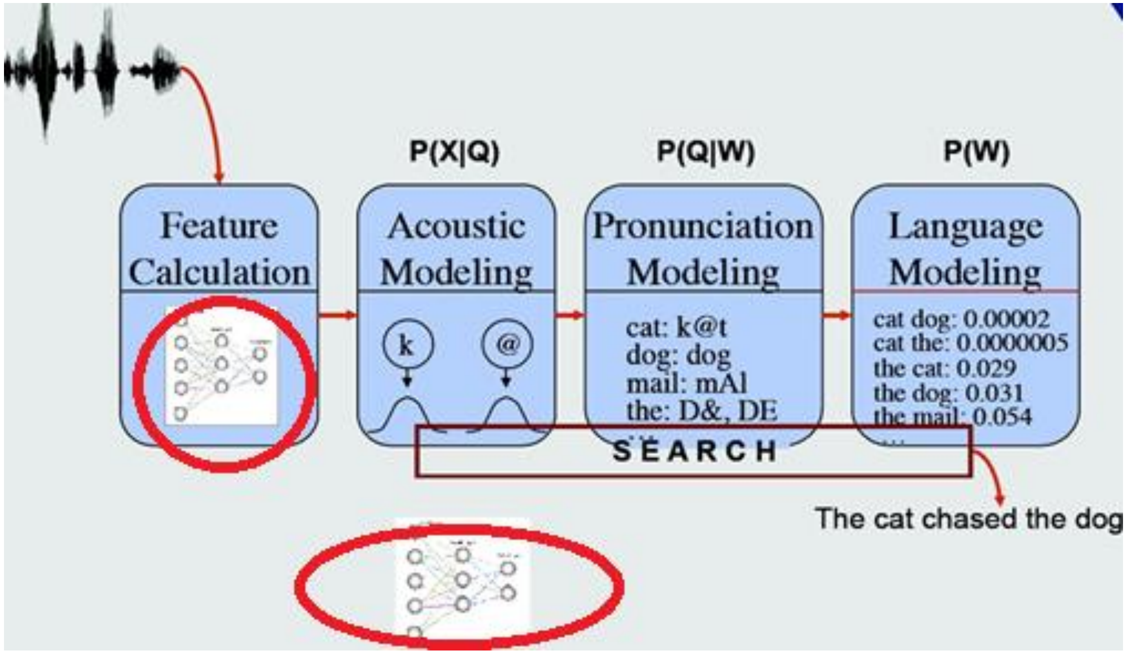
- ✓ Large volume data available
- ✓ Deep learning is changing everything. AM, LM…
- ✓ ASR is going to practice. Siri, google, baidu, Tencent…





Neural networks in ASR

- Replace GMMs to compute likelihood
- Replace MFCCs to generate posterior or bottleneck features



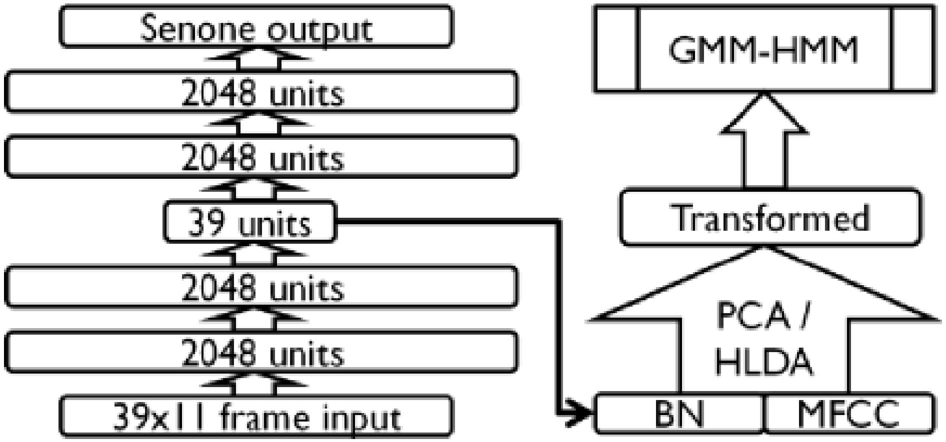
DNN-HMM hybrid approach

State of the art results show that DNN-HMM outperforms CDHMM both on phone recognition and on LVCSR (Icassp 2011, Interspeech 2011, IEEE trans. on ASLP 2012):

Paper	Site	Corpus	Results
A. Mohamed, T. Sainath, G, Dahl, B. Ramabhadran, G. Hinton, M. Picheny, Deep Belief Networks Using Discriminative Features for Phone Recognition, *ICASSP-2011, Dallas, Texas*	Toronto University IBM Watson	TIMIT	PER: 19.3% *best state of art result on TIMIT*
G.E. Dahl, Dong Yu, Li Deng, A. Acero, Context-Dependent Pre-trained Deep Neural Networks foe Large Vocabulary Speech Recognition, *IEEE trans. on Audio, Speech and Language proc.,* to appear in 2012	Toronto University Microsoft	Bing mobile voice search task	23.2% Sentence Error Reduction over GMM-HMMs
F. Seide, Gang Li, Dong Yu, Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, *Interspeech 2011, Florence, Italy*	Microsoft	Phone call transcriptions (Fisher-Switchboard)	33% Recognition Error Reduction over discriminatively trained GMM-HMMs

DNN features

The idea is to insert a bottleneck into a 5 layers DNN, and use the activations of this bottleneck layer as input of GMM-HMM, alone or coupled with conventional MFCC features, after a PCA/HLDA.



Acoustic Model	Test SER%
GMM-HMM MPE	36.2
DNN-HMM	30.4
BN + MFCC MPE	32.2

The results reported in that paper (Windows Live Search for Mobile corpus) shows that the best result is obtained by using directly the DBN-HMM model (using DBN outputs in the decoder), but a good improvement is obtained also using bottleneck features generated by DBN (together with standard MFCC) to train standard GMM-HMM.

Dong Yu and Michael L. Seltzer *“Improved Bottleneck Features Using Pretrained Deep Neural Networks”,* Interspeech 2011

Why DNNs work now?

- Large volume of data enables large scale models
- Powerful computing methods enable large scale training
- Smart designs: context dependence considered, states instead of phones …

Current research frontier (1)

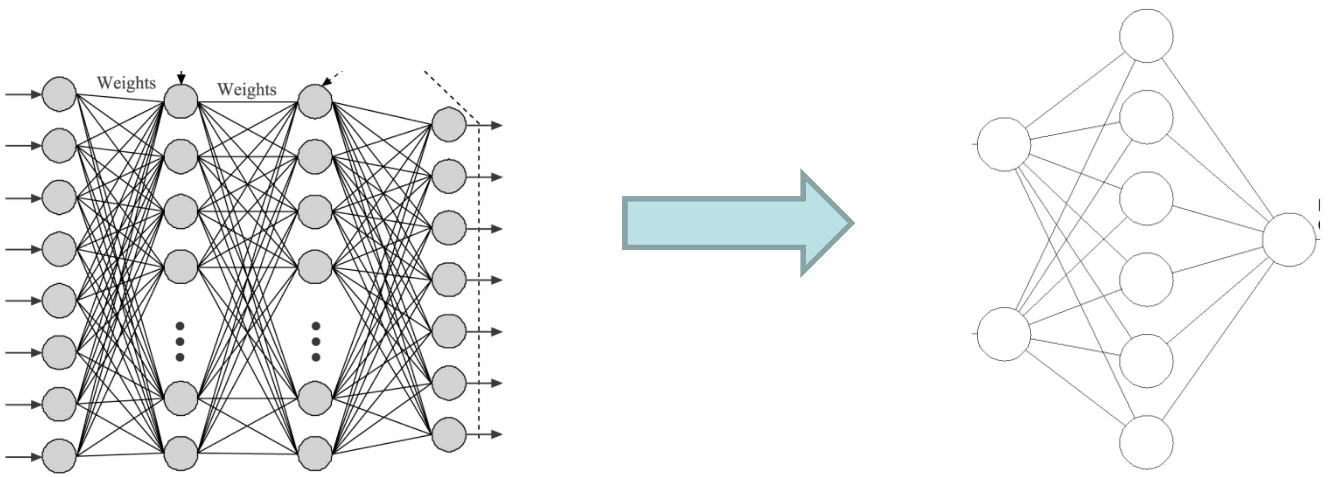
- New optimization approaches
  - Online SGD, GPU training, DistBelief, Hessian-free…
  - Rectify activation, drop out, max-out
- New structures
  - Recurrent DNN, convolutive DNN, sparse networks, deep convex networks, subspace methods…

Current research frontier (2)

- New adaptation methods
  - fDLR, LDA, adaptation layers
- New domains:
  - Language modeling, voice activity detection, confidence estimation
- New applications:
  - Multilingual modeling, multichannel modeling

Our research 1: sparse DNN

- The current DNN learning assumes dense structures, which is a fairly blind way.
- The relationships in nature are complex, but not so much. We target to such succinct learning.
- In other words, we are working on sparse learning.



Problems of dense DNN

- Dense DNNs are not preferable.
  - Dense DNNs lead to unnecessary computing/memory costs in training/prediction.
  - Dense DNNs are sensitive to noise.
  - Dense DNNs tend to be over-fitted.
  - Dense DNNs hide interesting causal relationships.
- By sparse learning, terse and robust models can be expected.

How to make DNNs sparse

- L0 norm: pruning, drop off
- L1 norm: by introducing L1 penalty, drive units/weights to zeroes.
- L1 + L2 norm: smoothed version of L1

$$E'(w) = E(w) + \lambda||w|| + \rho||w||^2$$

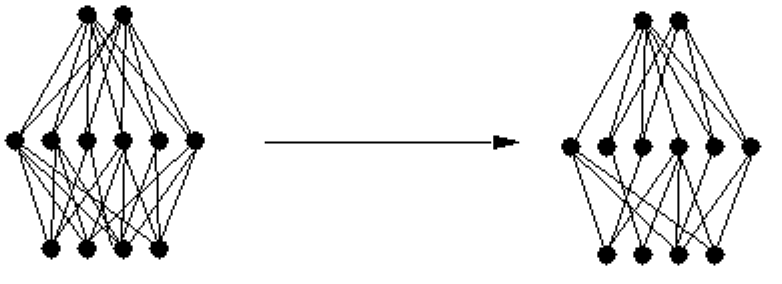
Second order pruning (1)

- Brain optimal damage (OBD) prunes neurons based on Hessian.

$$\delta E = E(w + \delta w) - E(w). \qquad \delta E \approx \delta w \frac{\partial E}{\partial w} + \frac{1}{2} \delta w^T H \delta w$$

$$H = \begin{pmatrix} \frac{\partial E^2}{\partial w_1 \partial w_1} & \cdots & \frac{\partial E^2}{\partial w_1 \partial w_K} \\ \cdots & \cdots & \cdots \\ \frac{\partial E^2}{\partial w_K \partial w_1} & \cdots & \frac{\partial E^2}{\partial w_K \partial w_K} \end{pmatrix}.$$

$$h_{k,k} = \frac{\partial^2 E}{\partial {w_{i,j}^m}^2} = \frac{\partial^2 E}{\partial (y_i^m)^2} (a_j^m)^2$$



Second order pruning (2)

- Make it work for DNN
  - Non-negative normalization
  - Hessian BP

	CA%			
	Magnitude		OBD	
Pruning	-	+	-	+
0%	47.63	-	47.63	-
50%	45.02	46.70	46.55	47.43
75%	33.87	46.83	41.13	47.06
88%	12.84	44.66	27.41	45.34
94%	1.31	41.81	14.43	42.92

	WER%	
Pruning	Magnitude	OBD
0%	24.04	
50%	24.01	24.14
75%	24.36	24.32
88%	25.41	25.09
94%	27.64	26.49

Our research 2: Noisy training

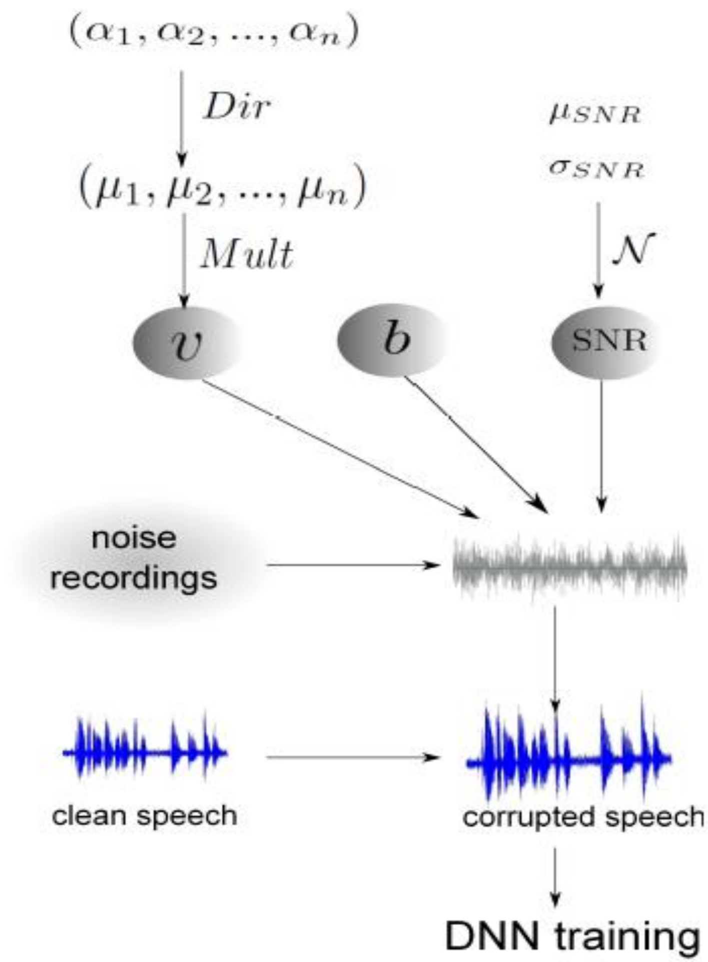
- We found corrupting input features by random noise leads to significant performance improvement for both clean and noise conditions.
- Noise training is equivalent to a L2 norm.

$$E(\theta) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\{\mathbf{y}^{(n)} ln f_k(\mathbf{x}^{(n)})\}$$

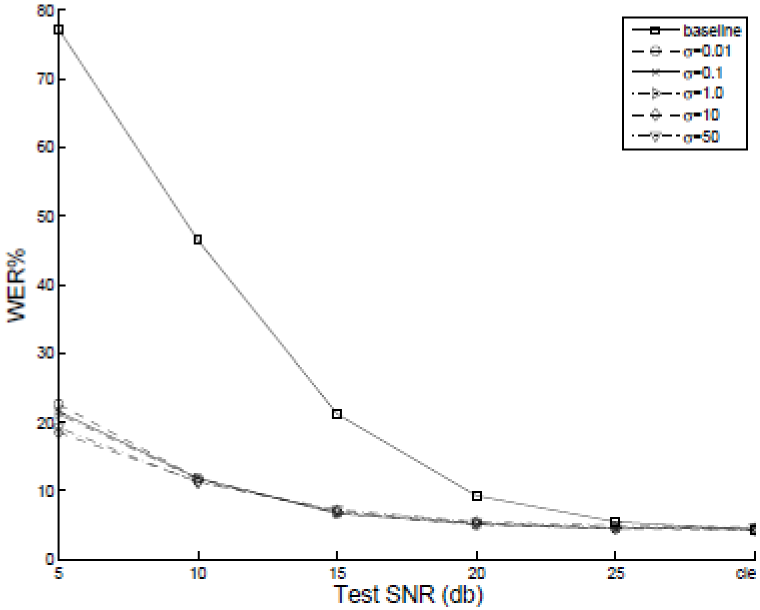
$$\Downarrow$$

$$E_v(\theta) \approx E(\theta) + \frac{\epsilon}{2} \bigtriangledown^2 E(\theta, 0).$$

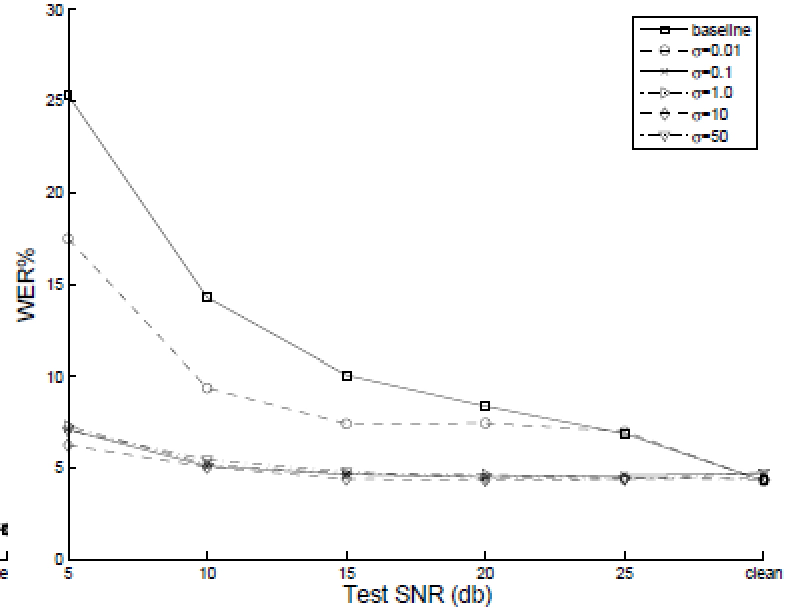
Sampling approach for noisy training



Results with noisy training



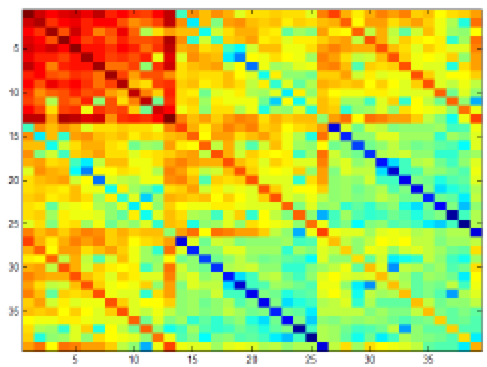
(a) White noise



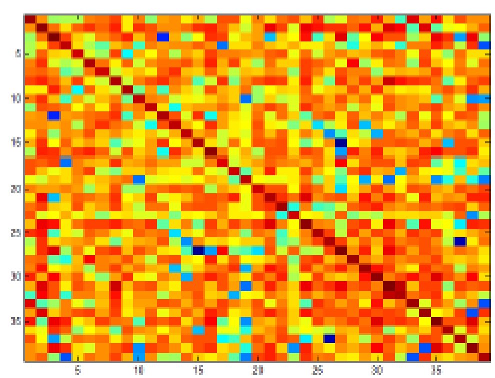
(b) Cafeteria noise

Our research 3: subspace modeling for DNN-BN features

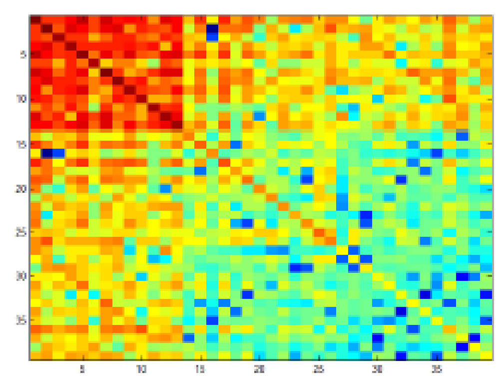
- DNN-BN features are highly correlated. Using subspace model can compensate for this correlation.



(a) MFCC



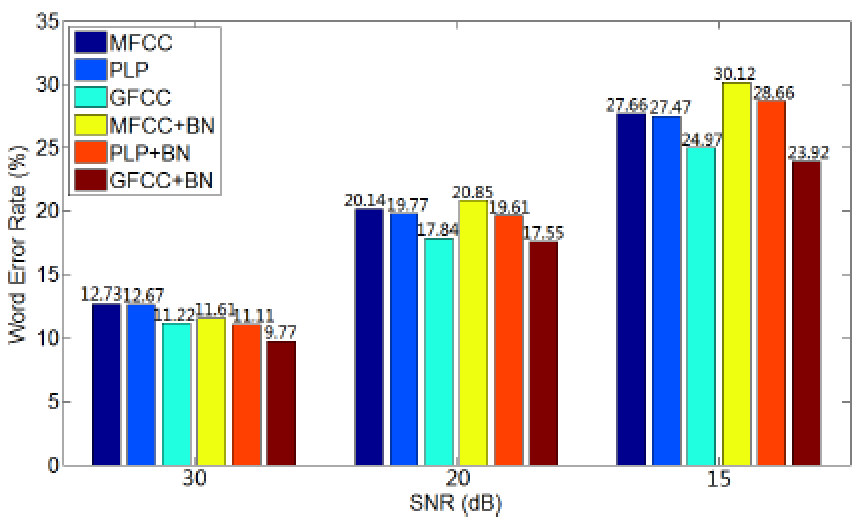
(b) MFCC-BN



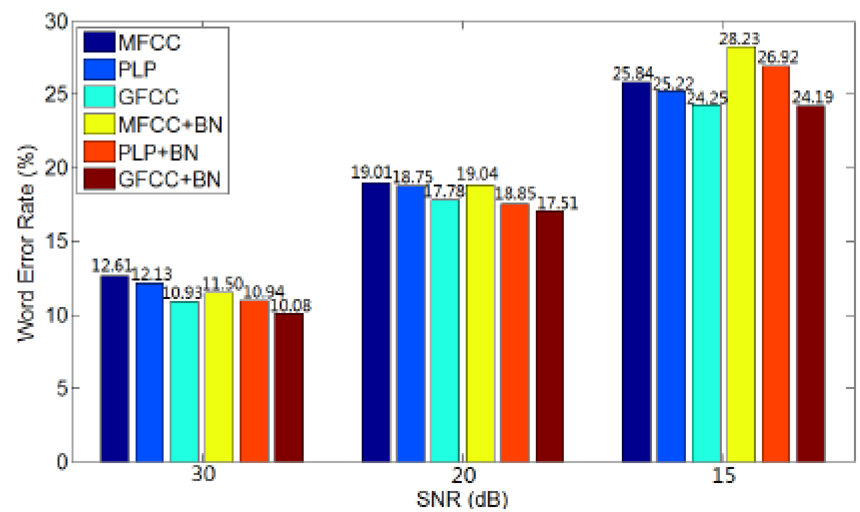
(c) MFCC-BN + KLT

Our research 4: robust features for NN

- NN is highly sensitive to pattern distribution changes.
- Using robust features, such that those are related to human ears, can improve performance.



(a) in white noise



(b) in babble noise

Contents

- Introduction to deep learning
- Deep learning in speech processing
- **Deep learning in language processing**

Language is hierarchical as well…

- Human languages are naturally hierarchical, including phones, words, phrases, syntactic trunks, discourses, etc.
- Traditional natural language processing (NLP) studies different tasks with different hand-crafted features plus specific discriminative or sequential models (SVM, CRF), e.g., word segment, POS tagging, semantic labelling.

A QA example

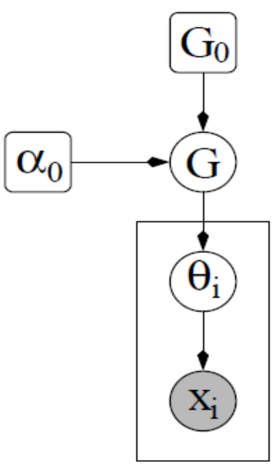
- A: I want to charge my mobile
  - C1: steps to charge phone
  - C2: danger to charge mobiles
  - C3: adapter to change a mobile
- The goal is to match the sematic meaning, however we need start from word analysis, and conduct truncation, NER, fuzzy match…

An Ideal way to NLP

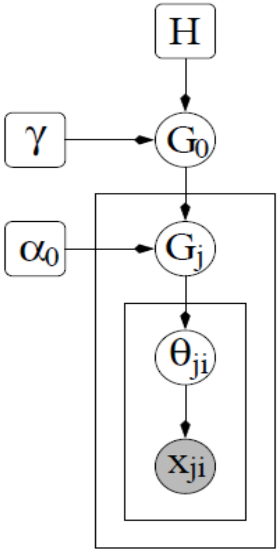
- If we had a deep model to learn the knowledge in different layers automatically, that would be perfect.
- Difficulties
  - We need a reasonable representation of knowledge in different granularity
  - We need a reasonable model to propagate information bottom-up
  - We need a framework to involve human knowledge in different layers
- Certainly not solved yet…

Deep Bayesian networks

- A bayesian network can be deep, and easy to involve human knowledge.
- However, much effort needs to pay for the structure design, and the learning is slow.
- Features still need to be designed by hand.



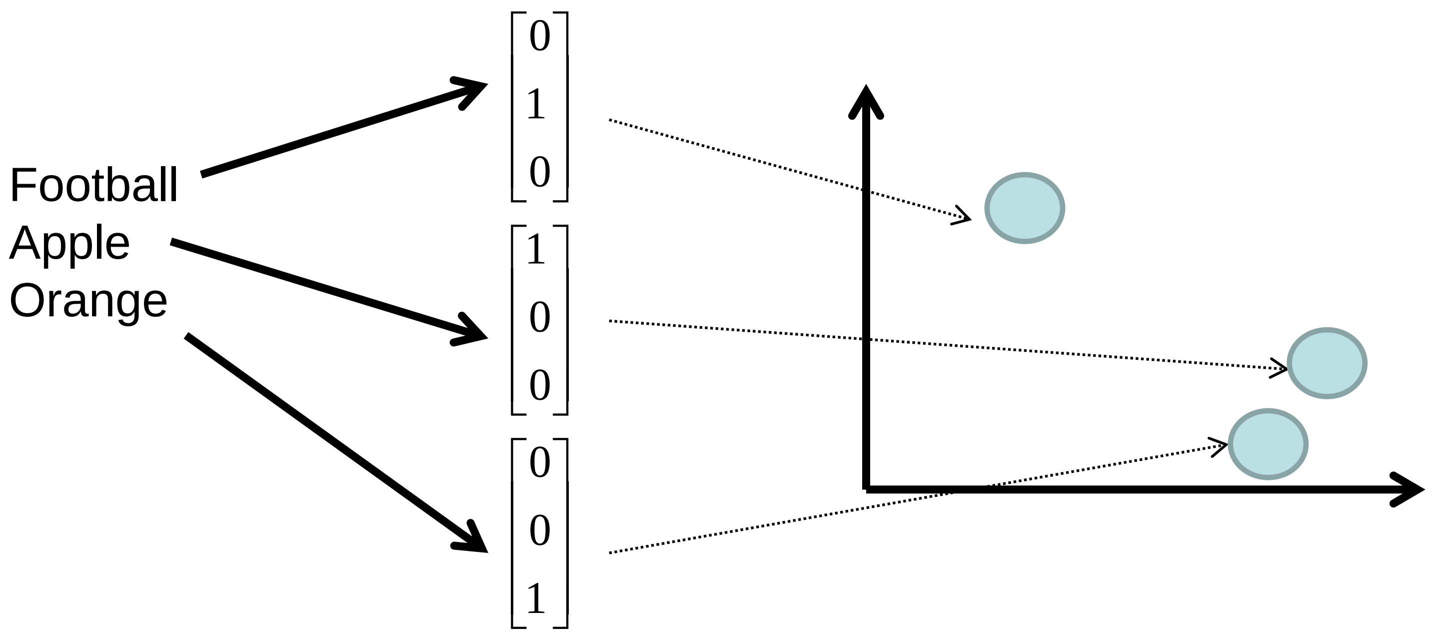
Dirichlet process



Hierarchical Dirichlet process

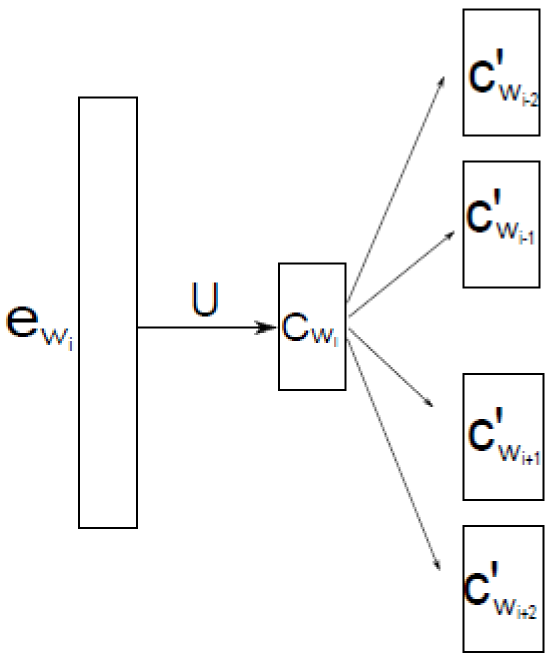
Word vectors: towards unified representation

- Embedding words in a continuous and low dimensional space.
- The ‘semantic meaning’ among words are represented by the distance in the space.



Learn word vectors

- LDA, NNLM, skip-grams



$$\frac{1}{N}\sum_{i=1}^{N}\sum_{-C\leq j\leq C, j\neq 0} logP(w_{i+j}|w_i)$$

$$P(w_{i+j}|w_i) = \frac{\exp(c_{w_{i+j}}c_{w_i})}{\sum_w \exp(c_w c_{w_i})}$$

T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Computation and Language, 2013.

Examples of word vectors

Orange:
yellow 0.685047
purple 0.679007
blue 0.666115
colors 0.610828
pink 0.608364
green 0.606376
white 0.596440
colored 0.596061
pale 0.572422

Football:
soccer 0.797210
rugby 0.748457
basketball 0.747661
baseball 0.688464
teams 0.687909
hockey 0.681951
athletic 0.654311
nhl 0.652300
league 0.644152

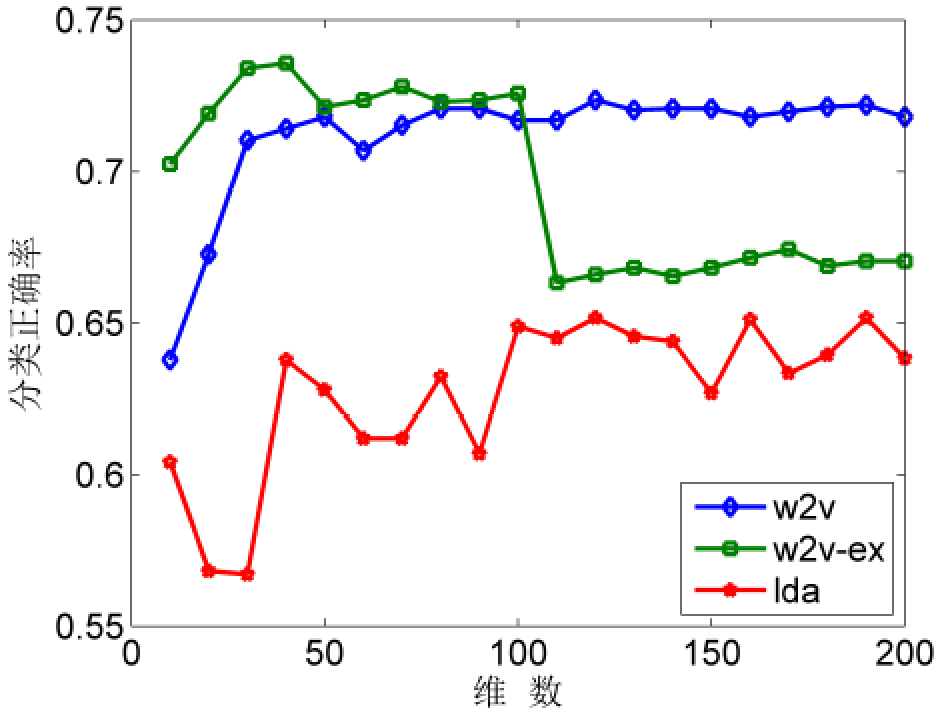
Potentials of word vectors

- Augment semantic meaning in representations, leading to semantic computing
- Easy to conduct inference in the low-dimensional continuous space
- Can be learned in a large corpus, therefore general and robust
- Can be task-specific

Our work (1): word-vector based text classification

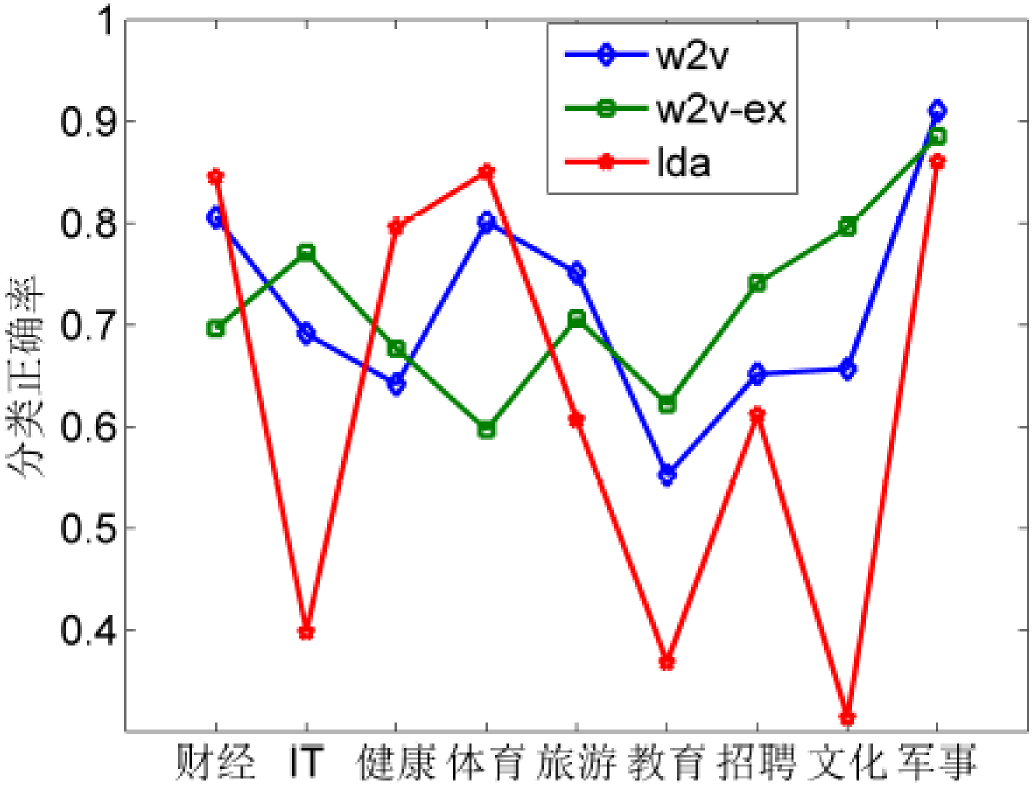
- Use word vectors to classify documents
- Document vectors are average of word vectors
- Naïve Bayesian is choosen as the classifier
- Experiments were conducted on Sogou classification text

Performance with various dimensions



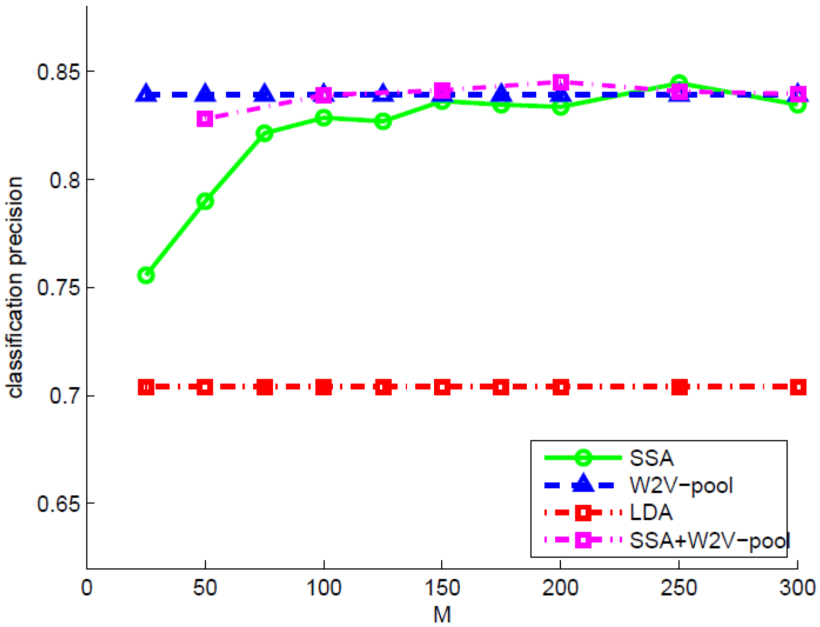
•Rong Liu, Dong Wang, Chao Xing, Text classification based on word vectors, ISCSLP 2014

Performance on different classes



Semantic space allocation model

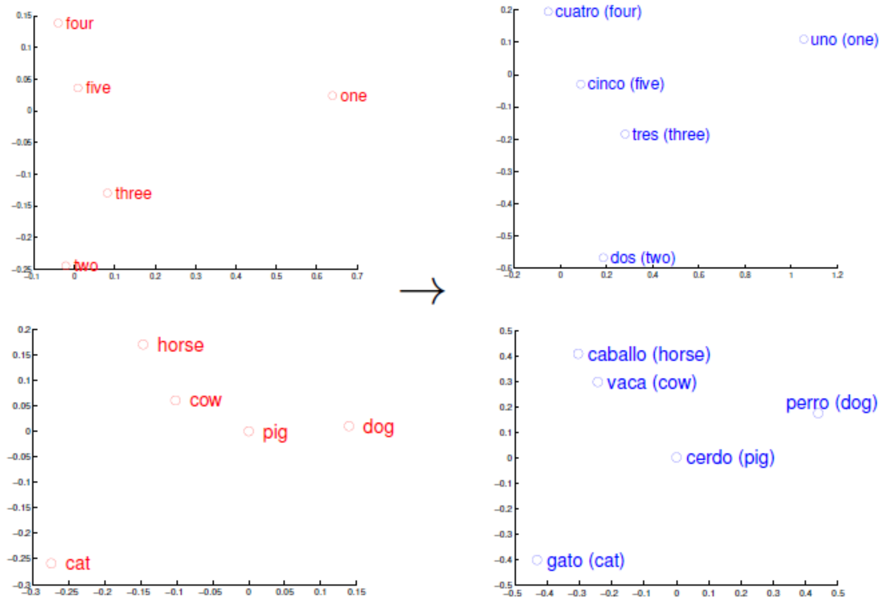
- Cluster word vectors and derive document vectors from the clusters.



Chao Xing, Dong Wang, Xuwei Zhang, Chao Liu, document classification based on i-vector distributions, APSIPA 2014

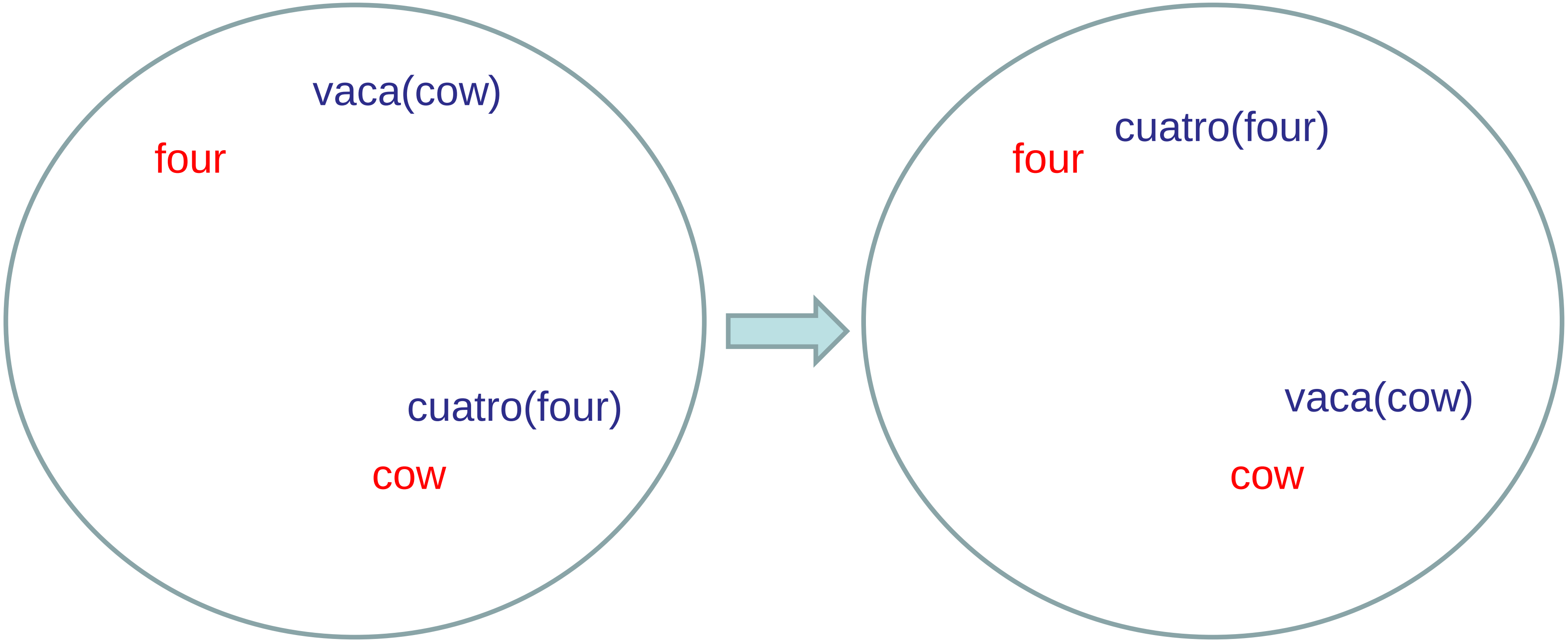
Our work(2): orthogonal vector mapping

- Construct two vector spaces for English and French respectively, then map word pairs of the same meaning.

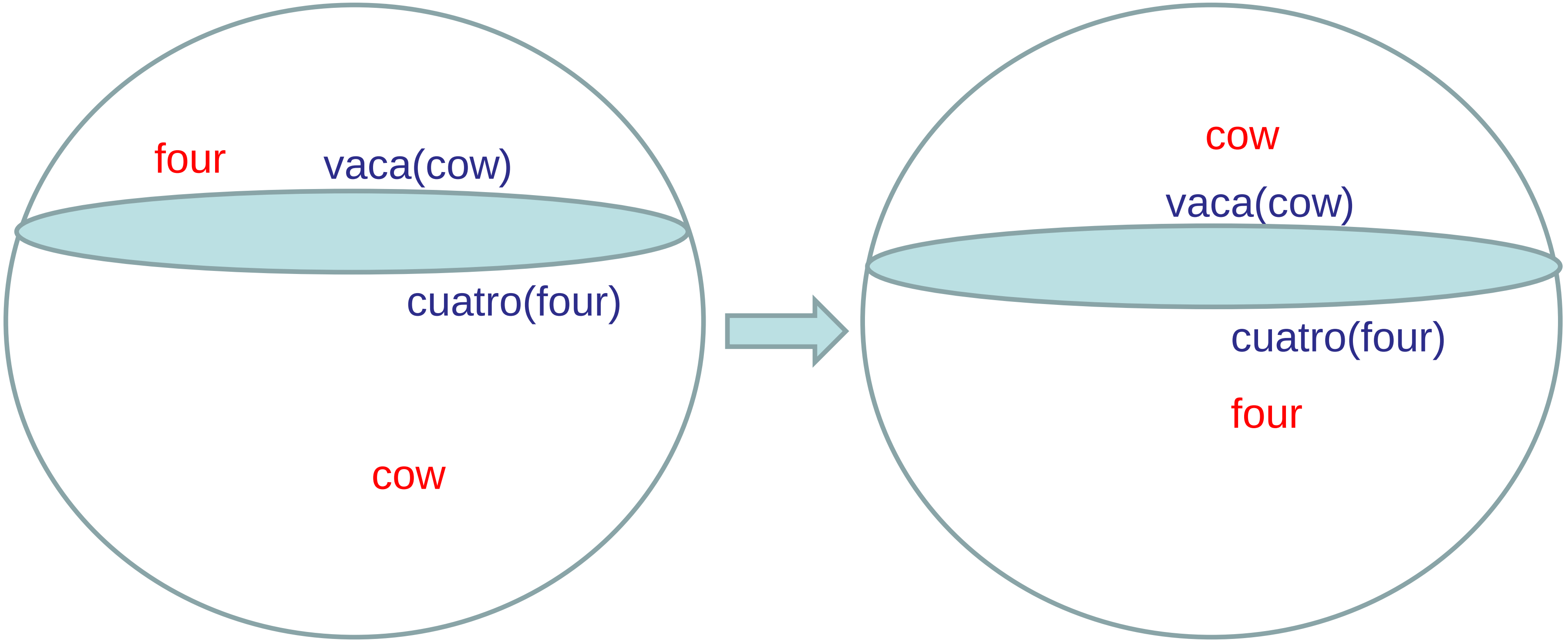


Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation[J].

Orthogonal transform instead of linear transform



Mismatched dimensions



Experimental results

Conclusions

- Deep learning delivers brilliant improvements in a multitude of research fields. For the speech processing and language processing, deep learning plays a role of revolution.
- Still a lot of difficulties exist for deep learning, e.g., quick training, smart structures, knowledge integration, big data…

Acknowledgements

- Thanks to the team.
- Thanks to the audience.
- Thanks to the authors of the slides that I extracted from, particularly Yann LeCun [1] and Deng Li [2],.

Reference

[1] Yann LeCun, Marc'Aurelio Ranzato, Deep Learning Tutorial, ICML, Atlanta, 2013-06-16

[2] Li Deng , Recent Innovations in Speech Technology Ignited by Deep Learning , Tsinghua University, December 14, 2012

[3]Jun Qi, Dong Wang and Javier Tejedor, "Subspace Models for Bottleneck Features", Interspeech 2013.

[4]Jun Qi, Dong Wang, Ji Xu and Javier Tejedor, "Bottleneck Features based on Gammatone Frequency Cepstral Coefficients", Interspeech 2013.

[5] Automatic Speech Recognition, www.informatics.manchester.ac.uk/~harold/LELA300431/

[6] Julia Hirschberg, Automatic Speech Recognition: An Overview. Slides for CS 4706, University of Columbia.

[7] Li Deng, Dong Yu, Deep learning: methods and applications, Foundations and Trends in Signal Processing, vol.7, no.3-4(2013).

- Thanks!