Microsoft Research

Deep Learning

from Speech Analysis/Recognition to Language/Multimodal Processing

Li Deng

Deep Learning Technology Center, Microsoft Research, Redmond, WA. USA

June 21, 2014

A Tutorial at Intern. Conf. Machine Learning (ICML)

Outline

- **Introduction:** Deep learning (DL) & its impact
- **Part I:** A (brief) history of "deep" speech recognition
- **Part II:** DL achievements in speech and vision
- **Part III:** DL challenges: **Language**, mind, & deep intelligence in the big-data world

Tutorial 9: Deep Learning and Its Applications in Signal Processing

(with MSR colleague Dong Yu)

T9 - Deep learning for natural language processing and related applications

Room: Teatrino

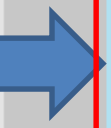
Subject Area: Speech/Audio/Language Processing

(with MSR colleagues Xiaodong He, Jianfeng Gao)

3hrs + 3hrs → 1 hr → 2hrs

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013



Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart. →

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous. →

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child? →

Additive Manufacturing

Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make jet parts. →

Baxter: The Blue-Collar Robot

Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people. →

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain forms long-term memories. Next: testing a prosthetic implant for people suffering from long-term memory loss. →

Smart Watches

The designers of the Pebble watch realized that a mobile phone is more useful if you don't have to take it out of your pocket. →

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nanotechnology just might make it possible. →

Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave – and even help us understand the spread of diseases. →

Supergrids

A new high-power circuit breaker could finally make highly efficient DC power grids practical. →

The New York Times

Scientists See Promise in Deep-Learning Programs

John Markoff

November 23, 2012

Rich Rashid in **Tianjin**, October, 25, 2012

/ MOBILE

Microsoft's Skype "Star Trek" Language Translator Takes on Tower of Babel

May 27, 2014, 5:48 PM PDT

Skype to get 'real-time' translator

Analysts say the translation feature could have wide ranging applications

Remember the universal translator on Star Trek? The gadget that let Kirk and Spock talk

Enabling Cross-Lingual Conversations in Real Time

Microsoft Research
May 27, 2014 5:58 PM PT

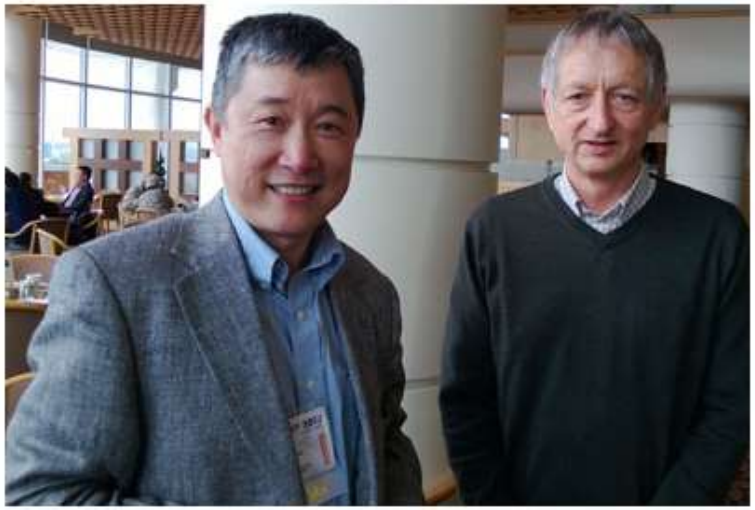
The success of the team's progress to date was on display May 27, in a talk by Microsoft CEO Satya Nadella in Rancho Palos Verdes, Calif., during the Code Conference. During Nadella's conversation with Kara Swisher and Walt Mossberg of the Re/code tech website relating to a new

The path to the Skype Translator gained momentum with an encounter in the autumn of 2010. Seide and colleague Kit Thambiratnam had developed a system they called The Translating! Telephone for live speech-to-text and speech-to-speech trans calls.

e asked
tage. Pall, the
ident of
first time
app, with Pall
rman-
Diana

View milestones on the path to Skype Translator

#speech2speech



Li Deng (*left*) and Geoff Hinton.

A core development that enables Skype translation came from Redmond researcher Li Deng. He invited Geoff Hinton, a professor at the University of Toronto, to visit Redmond in 2009 to work on new neural-network learning methods, based on a couple of seminal papers from Hinton and his collaborators in 2006 that had brought new

DNN Research Improves Bing Voice Search

Inside Microsoft Research 17 Jun 2013 10:00 AM 4

Posted by Rob Knies

We live in a society obsessed with speed. Whether it's download times on a mobile phone or Usain Bolt's time in the 100 meters, the faster the better. We also live during an era when accuracy has become not just preferable but essential. The technological marvels of the 21st century demand it.

Speed=good. Accuracy=good. Put them together, and you've got a leap forward, such as **recent advancements in Bing Voice Search** for **Windows Phone** that enable customers to get faster, more accurate results

Impact of deep learning in speech technology

Microsoft

The Fire Hose

Covering the news of the day at Microsoft

Bing helps you find better images in your searches with 'deep learning'

22 Nov 2013 9:21 AM

MIT Technology Review

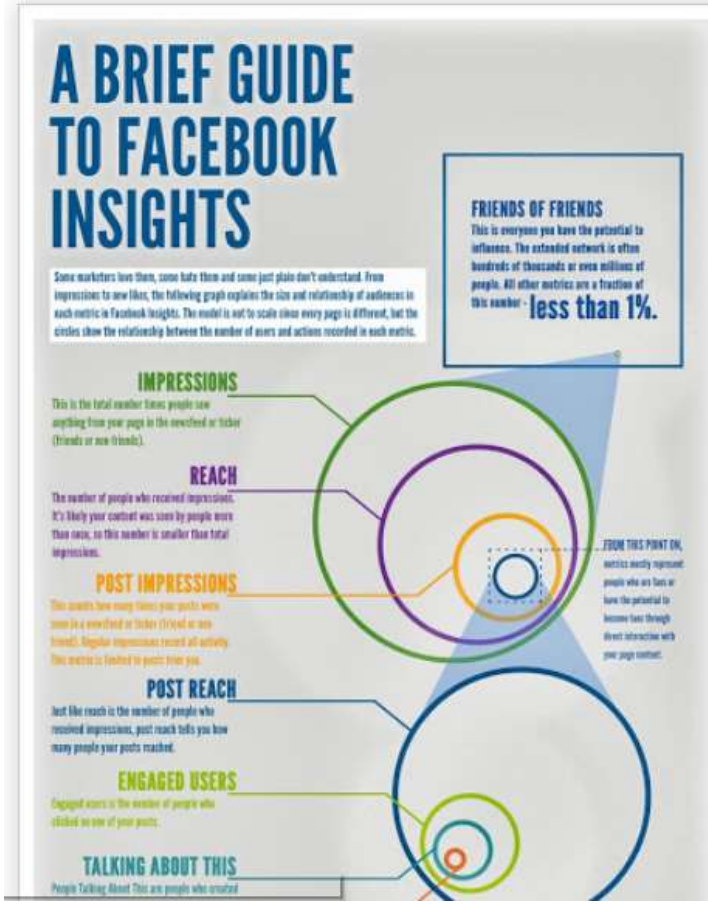
Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

September 20, 2013

A technique called deep learning could help Facebook understand its users and their data better.

By Tom Simonite on September 20, 2013

……Facebook's foray into deep learning sees it following its competitors Google and Microsoft, which have used the approach to impressive effect in the past year. Google has hired and acquired leading talent in the field (see "10 Breakthrough Technologies 2013: Deep Learning"), and last year created software that taught itself to recognize cats and other objects by reviewing stills from YouTube videos. The underlying deep learning technology was later used to slash the error rate of Google's voice recognition services (see "Google's Virtual Brain Goes to Work")….Researchers at Microsoft have used deep learning to build a system that translates speech from English to Mandarin Chinese in real time (see "Microsoft Brings Star Trek's Voice Translator to Life"). Chinese Web giant Baidu also recently established a Silicon Valley research lab to work on deep learning.



MIT Technology Review

BUSINESS NEWS

8 COMMENTS

Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

This week, Google reportedly paid that much to

This is Freescal

BloombergBusinessweek
Technology

Acquisitions

The Race to Buy the Human Brains Behind Deep Learning Machines

By Ashlee Vance | January 27, 2014

intelligence projects. “DeepMind is bona fide in terms of its research capabilities and depth,” says Peter Lee, who heads Microsoft Research.

According to Lee, Microsoft, Facebook (FB), and Google find themselves in a battle for deep learning talent. Microsoft has gone from four full-time deep learning experts to 70 in the past three years. “We would have more if the talent was there to be had,” he says. “Last year, the cost of a top, world-class deep learning expert was about the same as a top NFL quarterback prospect. The cost of that talent is pretty remarkable.”

Deep Learning's Role in the Age of Robots

BY JULIAN GREEN, JETPAC 05.02.14 2:56 PM

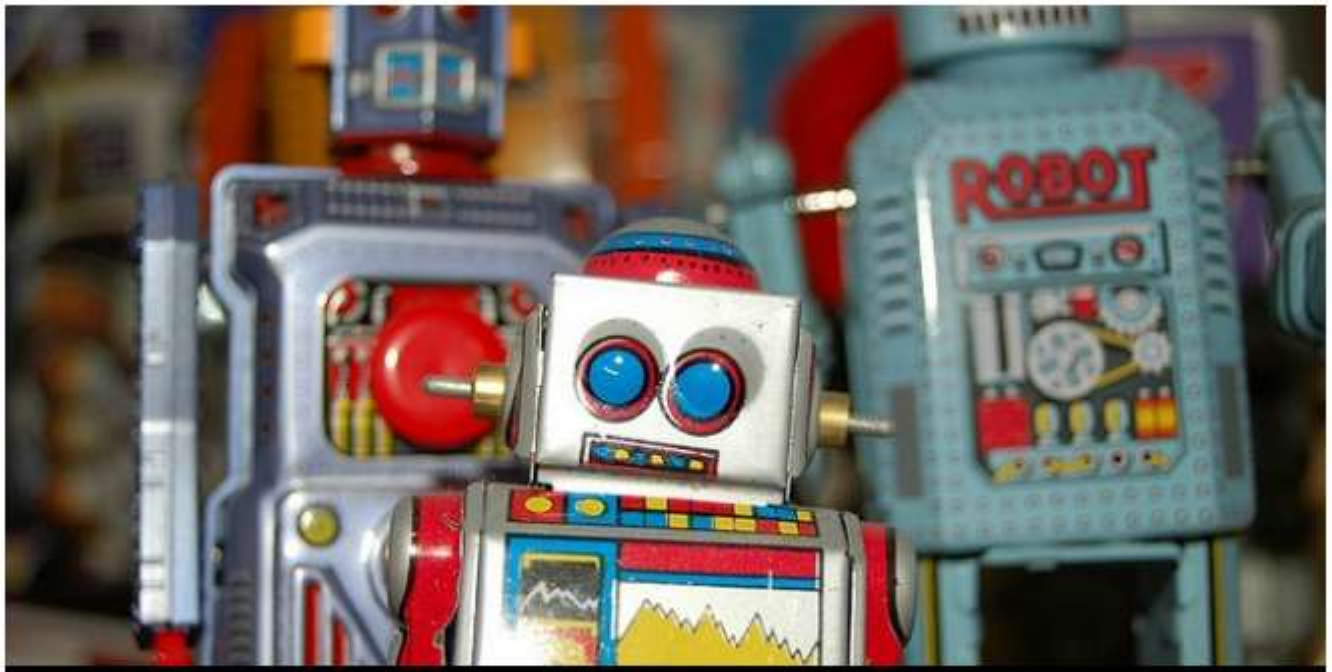


Image: jeffedoe/Flickr

Can robots see as well as humans? That's a question the biggest companies around are trying to answer.

MIT Technology Review

NEWS & ANALYSIS · | FEATURES | VIEWS | MULTIMEDIA | DISCUSSIONS | TOPICS | POPULAR: 10 BREAKTHROUGH TECHNOLOGIES

China's Baidu Bets on **Deep Learning**

MIT Technology Review - 3 days ago

Deep learning makes it possible for machines to process large amounts of date using simulated networks of simple neurons, crudely modeled ...

Baidu snatches Google's **deep-learning** visionary, Andrew ...

VentureBeat - 3 days ago

artificial intelligence / machine-learning / natural language processing

DARPA is working on its own deep-learning project for natural-language processing

by Derrick Harris MAY. 2, 2014 - 10:49 AM PDT

SUMMARY: *The Defense Advanced Research Projects Agency, or DARPA, is building a set of technologies to help it better understand human language so it can analyze speech and text sources and alert analysts of potentially useful information.*

Home
Photo Gallery
Mobile App
Organizing Committee
Plenary Speakers
Call for Papers
Technical Program
Registration
Paper Submission
SP Letters Presentation
Special Sessions
Hotels
Tutorials
Workshops
Supporters
Support And Exhibits
Student Luncheon

Plenary Speakers

Geoffrey E. Hinton

University of Toronto and Google Inc.

View Video

Daphne Koller

Stanford University

Photo Gallery
Mobile App
Organizing Committee
Plenary Speakers
Call for Papers
Technical Program
Registration
Paper Submission
SP Letters Presentation
Special Sessions

Special Sessions

ICASSP 2013 will offer the following special sessions:

Acoustic Event Detection and Scene Analysis
Organized by Mark Plumbley, Dimitris Giannoulis and Mathieu Lagrange

New types of deep neural network learning for speech recognition and related applications
Organized by Li Deng, Geoff Hinton and Brian Kingsbury

artificial intelligence / machine-learning / open source

A startup called Skymind launches, pushing open source deep learning

by Derrick Harris JUN. 2, 2014 - 10:03 AM PDT

SUMMARY: *Skymind is providing commercial support and services for an open source project called deeplearning4j. It's a collection of of approaches to deep learning that mimic those developed by leading researchers, but tuned for enterprise adoption.*

photo: deviantART / rajasegar

Is **Deep Learning**, the 'holy grail' of big data? - CNBC - Video

video.cnbc.com/gallery/?video=3000192292

Aug 22, 2013

Derrick Harris, GigaOM, explains how "**Deep Learning**" computers are able to process and understand ...

Outline

- Introduction: Impact of deep learning (DL)
- **Part I: A (brief) history of "deep" speech recognition**
  - neural nets
  - generative models (background on speech)
  - how DNN made recent inroad into speech recognition
  - roles of academic-industrial collaboration

Microsoft Research

Neural Networks in ASR (before 2009)

- Time-Delay Neural Networks

Waibel, Hanazawa, Hinton, Shikano, Lang. "Phoneme recognition using time-delay neural networks. " IEEE Transactions on Acoustics, Speech and Signal Processing, 1989.

- Recurrent Neural Nets

Bengio. “Artificial Neural Networks and their Application to Sequence Recognition”, 1991

Robinson. "A real-time recurrent error propagation network word recognition system”, ICASSP 1992.

- Hybrid Neural Nets

Morgan, Bourlard, Renals, Cohen, Franco. "Hybrid neural network/hidden Markov model systems for continuous speech recognition,” 1993.

- Neural-Net Nonlinear Prediction

Deng, Hassanein, Elmasry. “Analysis of correlation structure for a neural predictive model with applications to speech recognition,” *Neural Networks*, vol. 7, No. 2, 1994.

- Bidirectional Recurrent Neural Nets

Schuster, Paliwal. "Bidirectional recurrent neural networks," IEEE Trans. Signal Processing, 1997.

- Hierarchical Neural Nets

Fritsch, Finke. "ACID/HNN: Clustering hierarchies of neural networks for context-dependent connectionist acoustic modeling." ICASSP 1998.

- Neural-Net TANDEM

Hermansky, Ellis, Sharma. "Tandem connectionist feature extraction for conventional HMM systems." ICASSP 2000.

Morgan, Zhu, Stolcke, Sonmez, Sivadas, Shinozaki, Ostendorf, Jain, Hermansky, Ellis, Doddington, Chen, Cretin, Bourlard, Athineos, “Pushing the envelope - aside [speech recognition],” IEEE Signal Processing Magazine, vol. 22, no. 5, **2005.**

← **DARPA EARS Program 2001-2004: Novel Approach I**

(Deep) Generative Models in ASR (before 2009)

- **Structured Hidden Trajectory Models**

Deng, Yu, Acero. "Structured speech modeling," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, **2006**.

Deng and Huang. "Challenges in adopting speech recognition," *Communications of the ACM*, vol. 47, no. 1, 2004.

Zhou, Seide, Deng. "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM," ICASSP, 2003.

← DARPA EARS Program 2001-2004: Novel Approach II

- **Segmental Hidden Dynamic Models**

Bridle, Deng, Picone, Richards, Ma, Kamm, Schuster, Pike, Reagan. "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," Final Report for Workshop on Language Engineering, Johns Hopkins U, **1998**.

Deng, Ramsay, Sun. "Production models as a structural basis for automatic speech recognition," Speech Communication, vol. 33, pp. 93–111, 1997.

- **Switching State-Space Models**

Lee, Attias, D, Fieguth. "A Multimodal Variational Approach to Learning and Inference in Switching State Space Models," ICASSP, 2004.

Deng. "Switching Dynamic System Models for Speech Articulation and Acoustics," in Mathematical Foundations of Speech and Language Processing, vol. 138, pp. 115 - 134, Springer, 2003

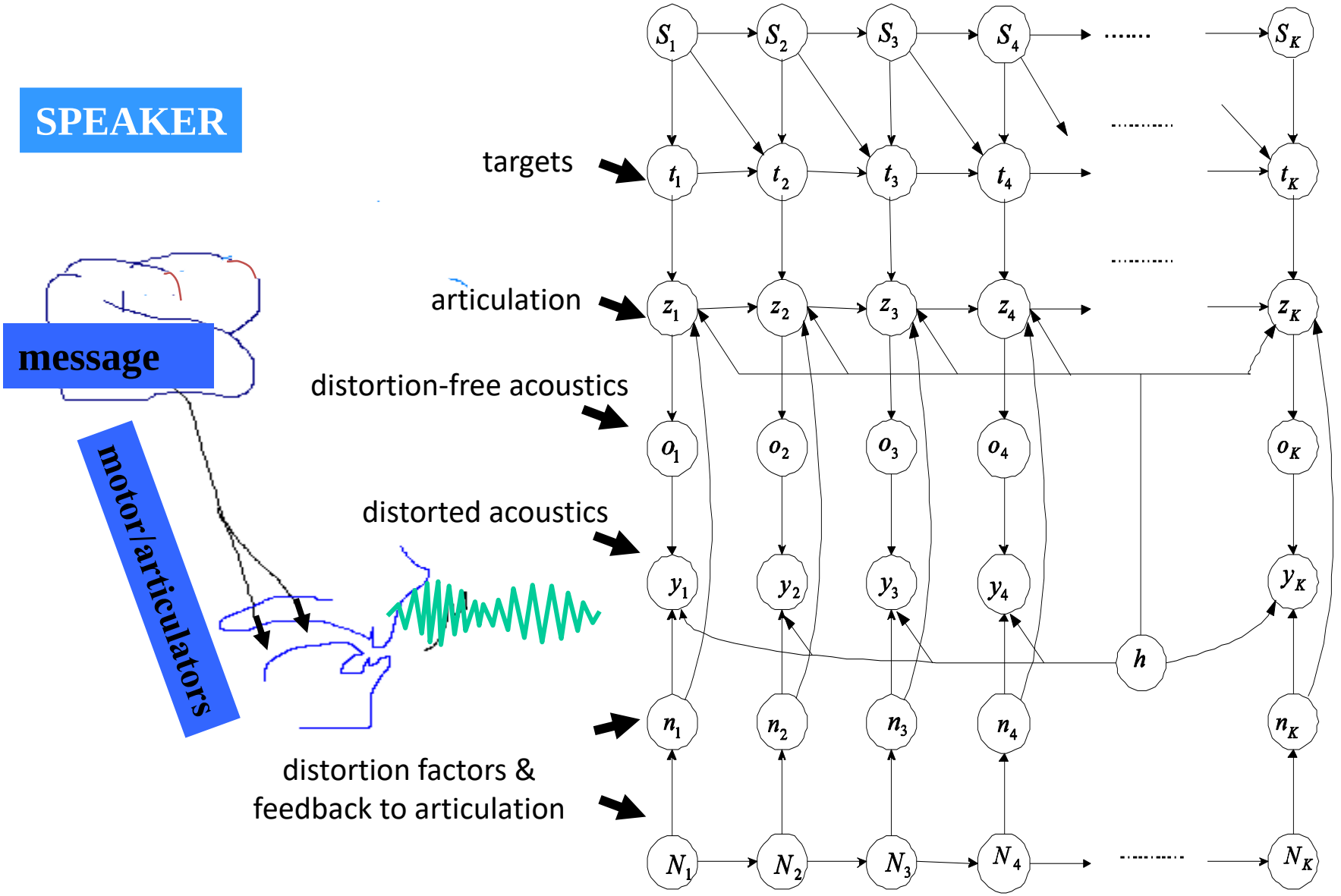
- **Gaussian Mixture Model & Hidden Markov Model (shallow); since 80's**

Rabiner, L. "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, 1989.

Microsoft Research

(Deep) Dynamic Bayesian Net

IIR filter
Nonlinear KF
Variational learning



Microsoft Research

ICASSP-2004

A MULTIMODAL VARIATIONAL APPROACH TO LEARNING AND INFERENCE IN SWITCHING STATE SPACE MODELS

Leo J. Lee[1,2], *Hagai Attias*[2], *Li Deng*[2] *and Paul Fieguth*[3]

University of Waterloo
[1]Electrical & Computer Engineering
[3]Systems Design Engineering
Waterloo, ON, N2L 3G1
Canada

[2]Microsoft Corporation
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6339
USA

ABSTRACT

An important general model for discrete-time signal processing is the switching state space (SSS) model, which generalizes the hidden Markov model and the Gaussian state space model. Inference and parameter estimation in this model are known to be computationally intractable. This paper presents a powerful new approximation to the SSS model. The approximation is based on a variational technique that preserves the multimodal nature of the continuous state posterior distribution. Furthermore, by incorporating a windowing technique, the resulting EM algorithm has complexity that is just linear in the length of the time series. An alternative Viterbi decoding with frame-based likelihood is also presented which is crucial for the speech application that originally motivates this work. Our experiments focus on demonstrating the effectiveness of the algorithm by extensive simulations. A typical example in speech processing is also included to show the potential of this approach for practical applications.

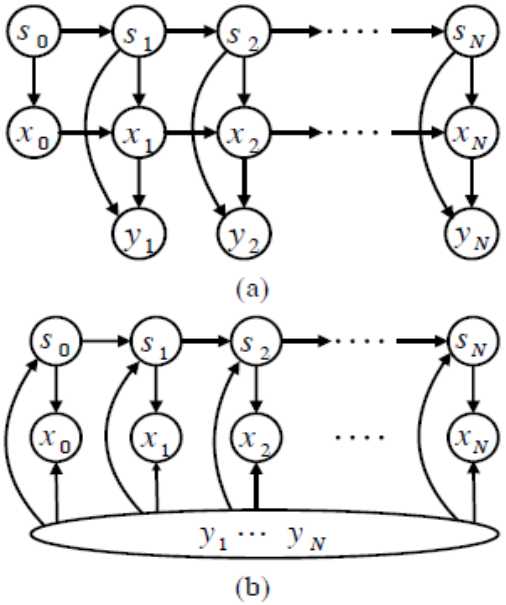


Fig. 1. The model (a) and the variational posterior (b) represented as Bayesian networks.

Structured Speech Modeling

Li Deng, *Fellow, IEEE*, Dong Yu, *Member, IEEE*, and Alex Acero, *Fellow, IEEE*

***Abstract*—Modeling dynamic structure of speech is a novel paradigm in speech recognition research within the generative modeling framework, and it offers a potential to overcome**

make it indistinguishable with human–human verbal interaction, at present, when users interact with any existing speech recog-

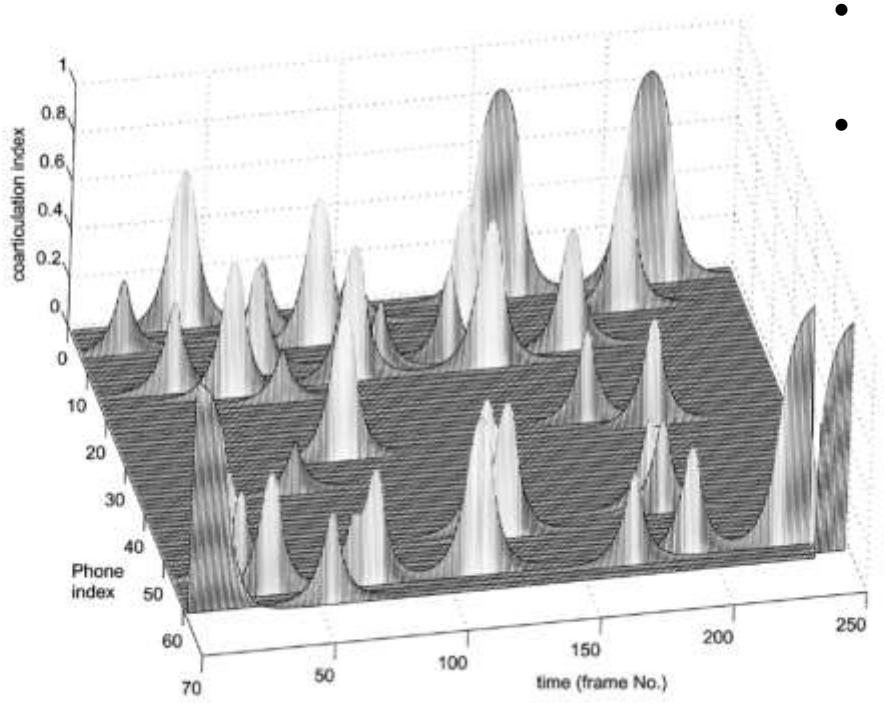


Fig. 1. Illustration of time-varying coarticulatory vectors a_k's for a TIMIT utterance. See text for detailed explanations.

- FIR filter instead of IIR filter
- Vocal tract resonances instead of articulatory variables
- VTRs-to-MFCC nonlinear mapping

Microsoft Research

Excellent Inference Results

- By-product: accurately tracking dynamics of resonances (formants) in the vocal tract
- Best formant tracker (speech analysis); used as basis to create the formant database
- Difficult to decode the full sentences in this generative deep/dynamic model
- With lattices (or huge N-best list), the decoder produces (then) best accuracy in TIMIT

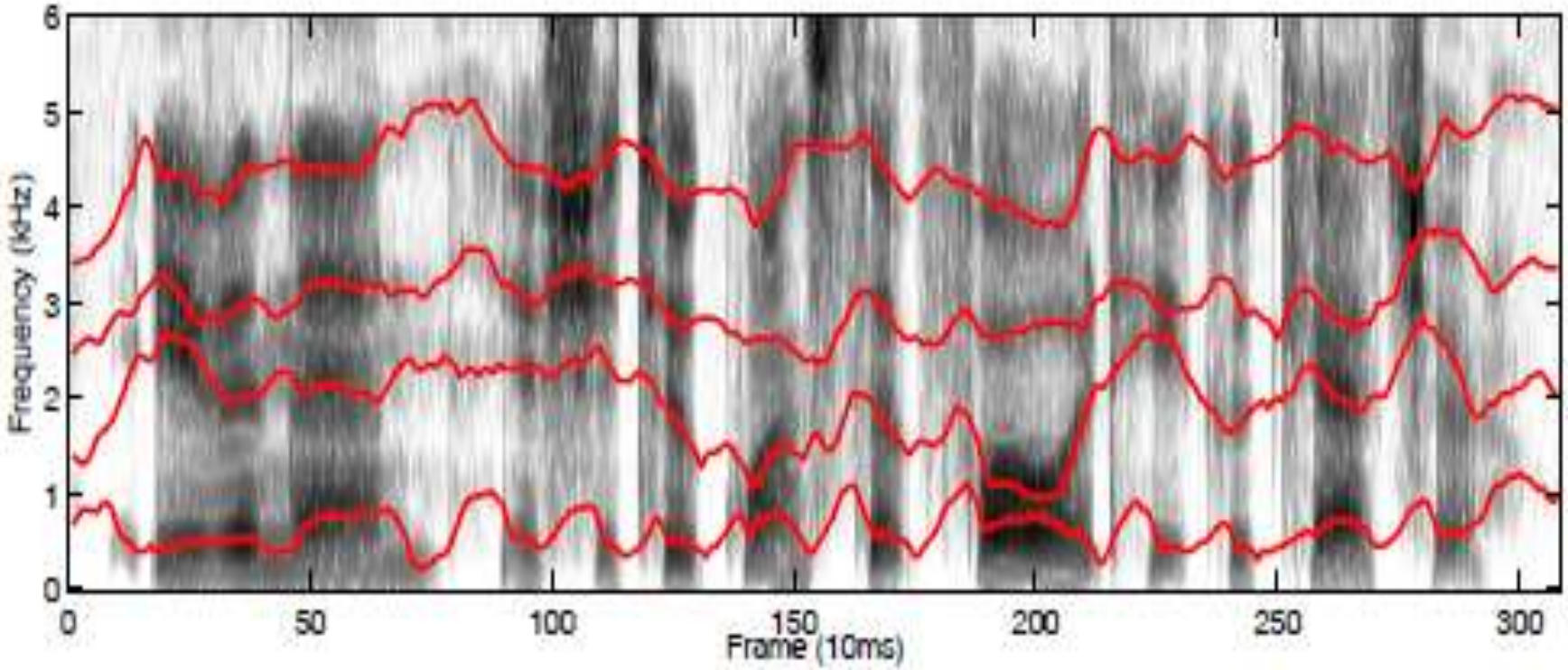


Fig. 5. Tracking VTRs for a speech sentence.

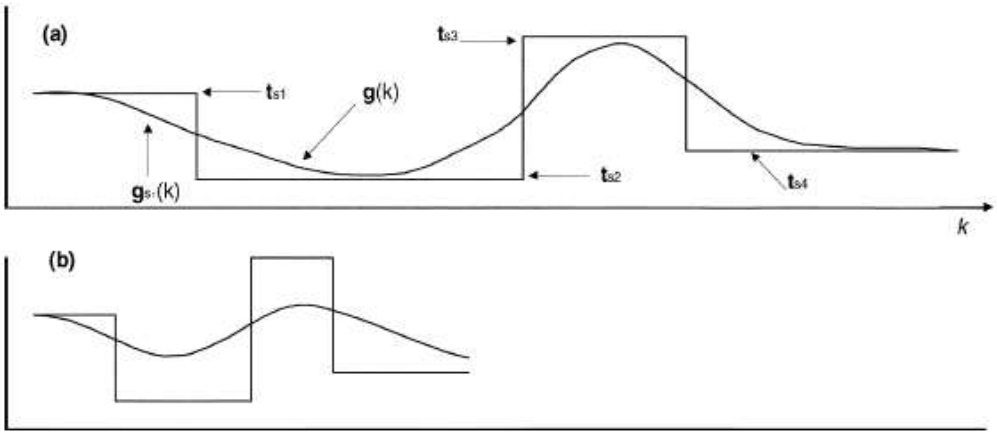


Fig. 1. Illustrations of the various VTR quantities in model Stage-I in an utterance with four phone segments. (a) and (b) are for the same four VTR targets and their filtered results, but the durations of the four segments are shorter in (b) than in (a).

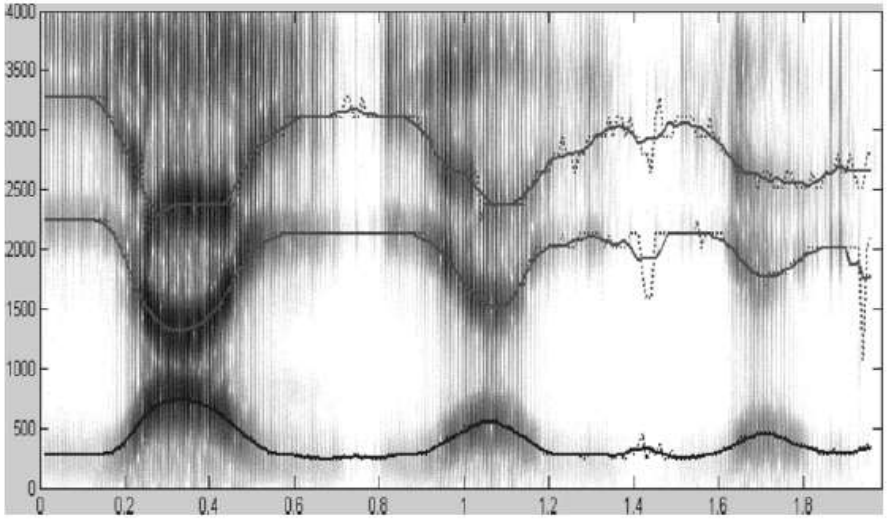


Fig. 2. Spectrogram of three renditions of /iy aa iy/ by one author, with an increasingly higher speaking rate and increasingly lower speaking efforts. The horizontal label is time, and the vertical one is frequency.

TABLE I

PHONETIC RECOGNITION PERFORMANCE COMPARISONS BETWEEN AN YSTEM AND THREE VERSIONS OF THE HTM SYSTEM. HTM-1: N-BEST RING WITH HTM SCORES ONLY; HTM-2: N-BEST RESCORING WITH TED HTM, HMM, AND LM SCORES; HTM-3: LATTICE-CONSTRAINED EARCH WITH WEIGHTED HTM, HMM, AND LM SCORES. IDENTICAL ACOUSTIC FEATURES (FREQUENCY-WARPED LPCCs) ARE USED

	Acc %	Corr %	Sub %	Del %	Ins %
HMM	**71.43**	73.64	17.14	9.22	2.21
HTM-1	**74.31**	77.76	16.23	6.01	3.45
HTM-2	**74.59**	77.73	15.61	6.65	3.14
HTM-3	**75.07**	78.28	15.94	5.78	3.20

TABLE II

COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRECT WITHIN EACH OF FOUR BROAD PHONE CLASSES

	Sonorants	Stops	Fricatives	Closures
Occurrences	3814	889	1252	1578
HMM	64.05	72.10	75.64	88.72
HTM	72.42	76.27	75.74	90.94

corrected many errors:
"short" phones

- Elegant model formulation & knowledge incorporation
- Strong empirical results; 97% TIMIT accuracy with Nbest=1001; fast training
- But very expensive for decoding; could not ship

then Geoff Hinton came along… (2009 at MSR)

- **Kluge 1:** keep the assumption of frame independence (ignore real "dynamics" to speed up decoding) but use bigger time windows
- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets)
- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.
- Well-timed academic-industrial collaboration:
  - ASR industry searching for new solutions when "principled" deep generative approaches could not deliver
  - Academics developed deep learning tools (**DBN**/DNN with hybrid generative/discriminative, 2006)
  - Advent of GPU computing (CUDA library released 2008)

Microsoft Research

(Deep) Generative Models (before 2009)

- **Structured Hidden Trajectory Models**

Deng, Yu, Acero. "Structured speech modeling," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, 2006.

Deng and Huang. "Challenges in adopting speech recognition," *Communications of the ACM*, vol. 47, no. 1, 2004.

Zhou, Seide, Deng. "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM," ICASSP, 2003.

←DARPA EARS Program 2002-2004: Novel Approach II

- **Segmental Hidden Dynamic Models**

Bridle, Deng, Picone, Richards, Ma, Kamm, Schuster, Pike, Reagan. "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," Final Report for 1998 Workshop on Language Engineering, Johns Hopkins U, 1998.

Deng, Ramsay, Sun. "Production models as a structural basis for automatic speech recognition," Speech Communication, vol. 33, pp. 93–111, 1997.

- **Switching State-Space Models (shallow)**

Lee, Attias, Deng, Fieguth. "A Multimodal Variational Approach to Learning and Inference in Switching State Space Models," ICASSP, 2004.

Deng. "Switching Dynamic System Models for Speech Articulation and Acoustics," in Mathematical Foundations of Speech and Language Processing, vol. 138, pp. 115 - 134, Springer, 2003

- **Gaussian Mixture Model & Hidden Markov Model (shallow); 1000's references**

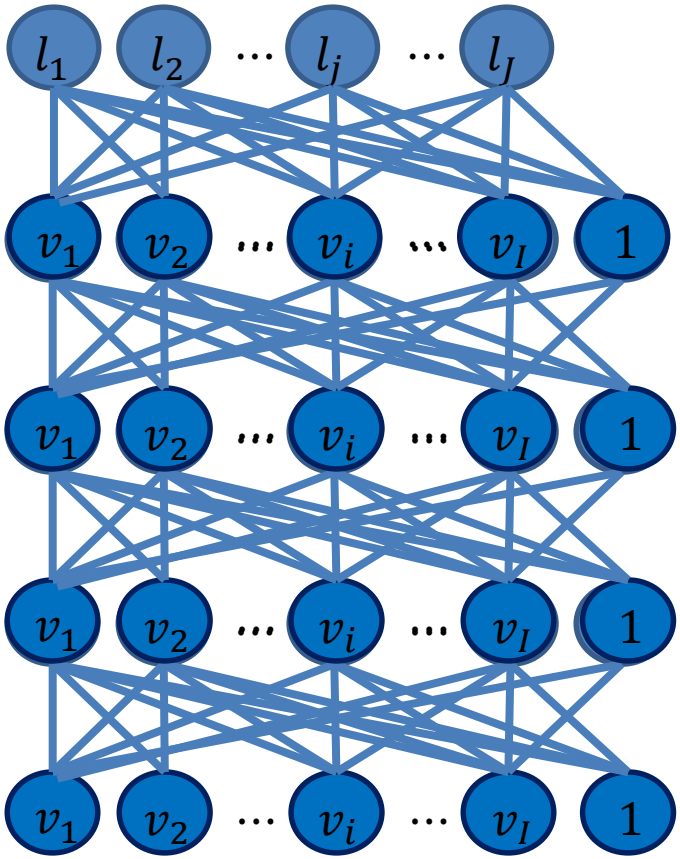
Rabiner, L. "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, 1989.

- **Deep Belief Networks (DBN)**

Hinton, Osindero, Teh. "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, 2006.

Hinton, Salakhutdinov "Reducing the dimensionality of data with neural networks," Science, vol. 313, 2006.

DBN: Layer-by-Layer Unsupervised Learning



- DBN as stacked RBMs
- RBM: $p(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}$

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{b}^{\mathbf{T}}\mathbf{v} - \mathbf{c}^{\mathbf{T}}\mathbf{h} - \mathbf{v}^{\mathbf{T}}\mathbf{W}\mathbf{h}$$

$$p(\mathbf{h}|\mathbf{v}) = \prod_i p(h_i|\mathbf{v})$$

$$p(h_i = 1|\mathbf{v}) = \sigma\left(c_i + \mathbf{v}^{\mathbf{T}}\mathbf{W}_{\boldsymbol{i}}\right)$$

- Pre-train each layer from bottom up by considering each pair of layers as an RBM.
- Jointly fine-tune all layers using back-propagation algorithm→**DNN**

Microsoft Research

DBN/DNN Works Well for TIMIT

(and a stunning discovery at MSR, 2009-2010)

	METHOD	Error rate
	(Shallow) GMM-HMM (1987-2010)	27.3%
	AUGMENTED CONDITIONAL RANDOM FIELDS [26]	26.6%
	RANDOMLY INITIALIZED RECURRENT NEURAL NETS [27]	26.1%
	BAYESIAN TRIPHONE GMM-HMM [28]	25.6%
	Deep/hidden trajectory model (MSR, 2006)	24.8%
DNN	MONOPHONE RANDOMLY INITIALIZED DNNs (SIX LAYERS) [13]	23.4%
DNN	MONOPHONE DBN-DNNs (SIX LAYERS) [13]	22.4%
	MONOPHONE DBN-DNNs WITH MMI TRAINING (MSR, 2010)	22.1%
	TRIPHONE GMM-HMMs DT W/ BMMI (IBM, 2010)	21.7%
	MONOPHONE DBN-DNNs ON FBANK (EIGHT LAYERS) [13]	20.7%
	MONOPHONE MCRBM-DBN-DNNs ON FBANK (FIVE LAYERS) [33]	20.5%
	Deep convolutional nets w. DropOut & Heter. Pooling (MSR, 2012)	**18.7%**

Error pattern A

Error pattern B

IEEE SIGNAL PROCESSING MAGAZINE [6] NOVEMBER 2012

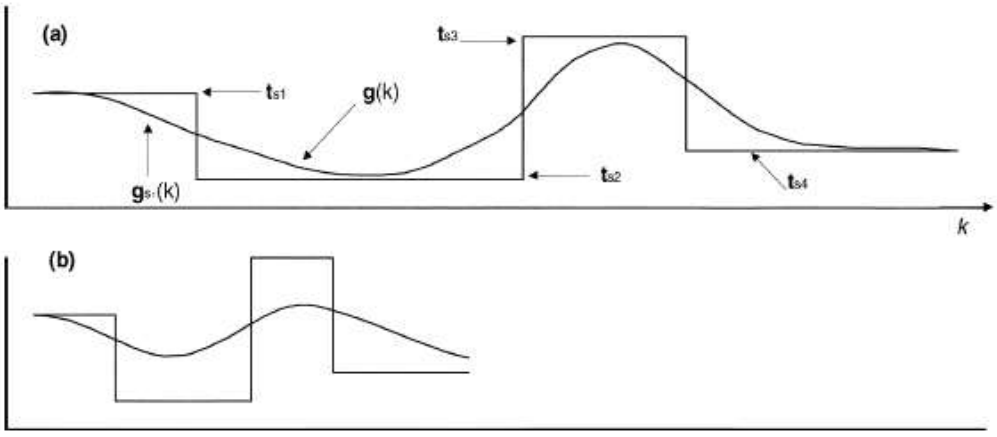


Fig. 1. Illustrations of the various VTR quantities in model Stage-I in an utterance with four phone segments. (a) and (b) are for the same four VTR targets and their filtered results, but the durations of the four segments are shorter in (b) than in (a).

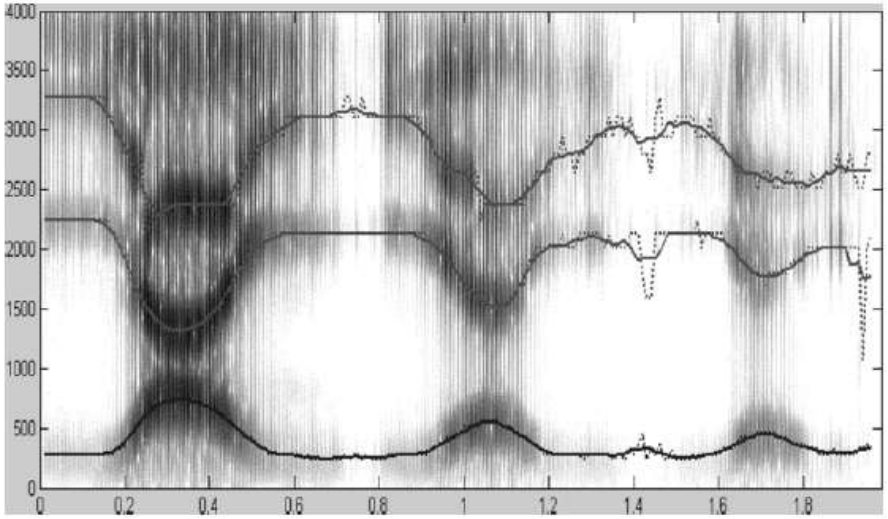


Fig. 2. Spectrogram of three renditions of /iy aa iy/ by one author, with an increasingly higher speaking rate and increasingly lower speaking efforts. The horizontal label is time, and the vertical one is frequency.

TABLE II
COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRECT) WITHIN EACH OF FOUR BROAD PHONE CLASSES

	Sonorants	Stops	Fricatives	Closures
Occurrences	3814	889	1252	1578
HMM	64.05	72.10	75.64	88.72
HTM	72.42	76.27	75.74	90.94

corrected many errors:
"short" phones

Another key discovery at MSR, 2009-2010

- Spectrogram (fbank) features better than cepstra features (MFCCs)

  (on speech analysis & feature coding using deep autoencoders)
- MFCCs dominated speech analysis & recognition: 1982-2012
- Conforming to the basic DL theme: back to raw features (and learn transformations automatically layer by layer)

Microsoft Research

The first use of spectrogram features in speech analysis/coding with deep learning

INTERSPEECH 2010

Binary Coding of Speech Spectrograms Using a Deep Auto-encoder

L. Deng[1], M. Seltzer[1], D. Yu[1], A. Acero[1], A. Mohamed[2], and G. Hinton[2]

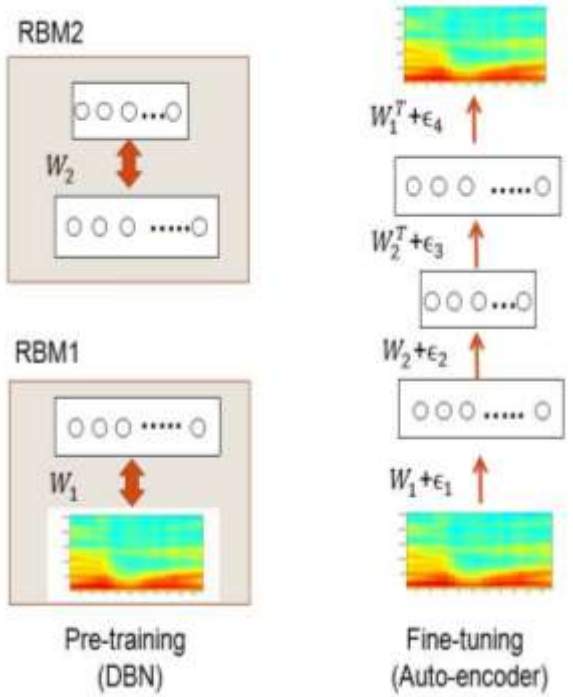
[1] Microsoft Research, One Microsoft Way, Redmond, WA 98052, US
[2] University of Toronto, Toronto, Ontario, Canada

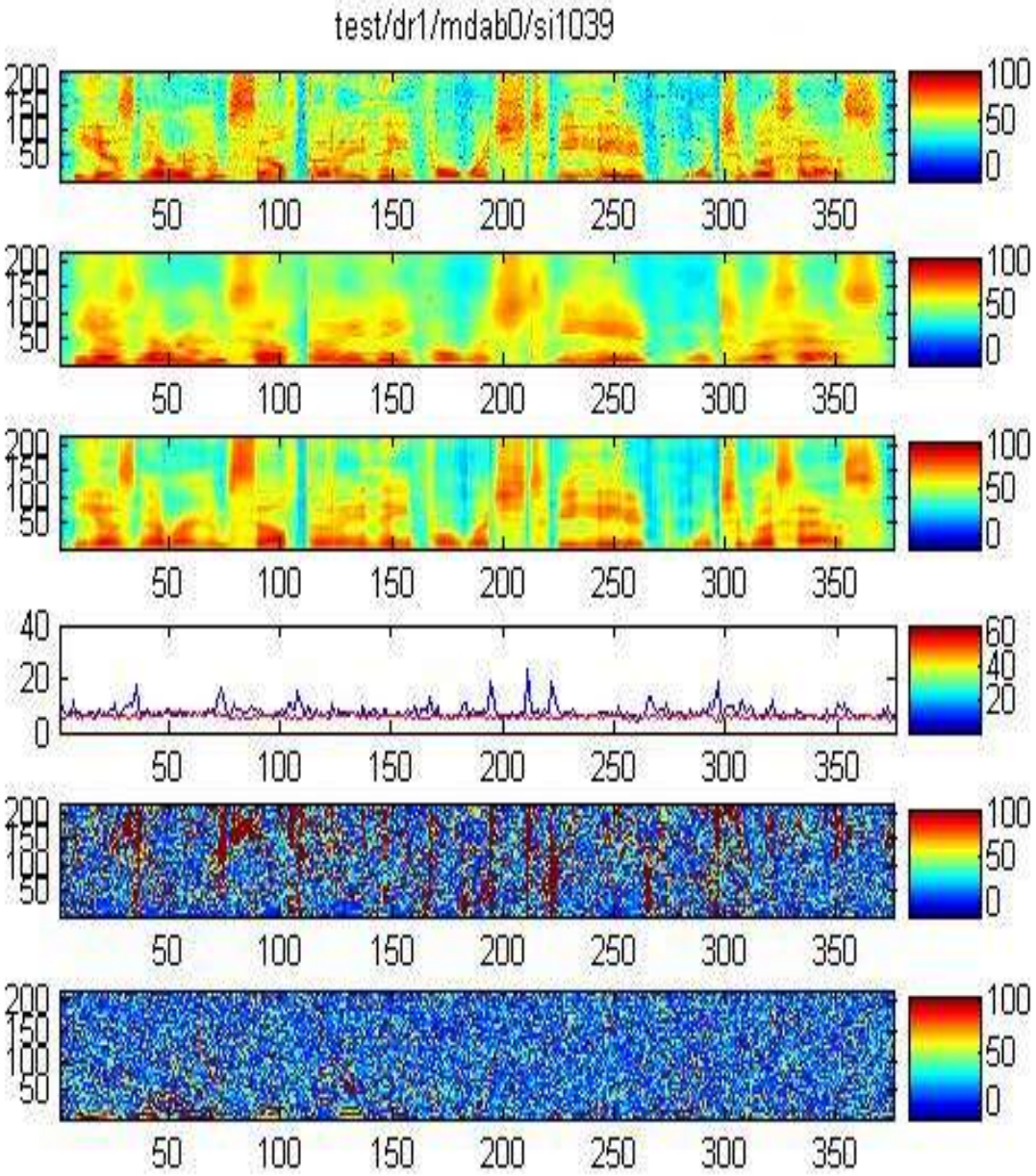
{deng|mseltzer|dongyu|alexac}@microsoft.com; {asamir|hinton}@cs.toronto.edu

Abstract

This paper reports our recent exploration of the layer-by-layer learning strategy for training a multi-layer generative model of patches of speech spectrograms. The top layer of the

The work reported in this paper was inspired by the successful use of deep auto-encoders for dimensionality reduction [8][9] and the extension of this work to the discovery of efficient binary codes in information retrieval [12]. It is also motivated by the potential benefits of using





Original spectrogram (log of |FFT|)

Reconstructed spectrogram from a 312-bit VQ coder

Reconstructed spectrogram from a 312-bit deep autoencoder

Coding errors as a function of time for VQ coder (blue) and autoencoder (red)

VQ coder's error (over time & freq)

Deep autoencoder's error

- No such nice results for MFCCs and other features

Li Deng, Dong Yu, Geoffrey Hinton

Microsoft Research; Microsoft Research; University of Toronto

Deep Learning for Speech Recognition and Related Applications

7:30am - 6:30pm Saturday, December 12, 2009

Location: Hilton: Cheakamus

Abstract: Over the past 25 years or so, speech recognition technology has been dominated by a "shallow" architecture --- hidden Markov models (HMMs). Significant technological success has been achieved using complex and carefully engineered variants of HMMs. The next generation of the technology requires solutions to remaining technical challenges under diversified deployment environments. These challenges, not adequately addressed in the past, arise from the many types of variability present in the speech generation process. Overcoming these challenges is likely to require "deep" architectures with efficient learning algorithms. For speech recognition and related sequential pattern recognition applications, some attempts have been made in the past to develop computational architectures that are "deeper" than conventional HMMs, such as hierarchical HMMs, hierarchical point-process models, hidden dynamic models, and multi-level detection-based architectures, etc. While positive recognition results have been reported, there has been a conspicuous lack of systematic learning techniques and theoretical guidance to facilitate the development of these deep architectures. Further, there has been virtually no effective communication between machine learning researchers and speech recognition researchers who are both advocating the use of deep architecture and learning. One goal of the proposed workshop is to bring together these two groups of researchers to review the progress in both fields and to identify promising and synergistic research directions for potential future cross-fertilization and collaboration.

http://research.microsoft.com/en-us/um/people/dongyu/NIPS2009/

Deep Learning for Speech Recognition and Related Applications

a workshop in conjunction with NIPS 2009

December 12, 2009, Whistler, BC, Canada

Workshop chairs

Li Deng, Microsoft Research, deng [at] microsoft [dot] com
Dong Yu, Microsoft Research, dongyu [at] microsoft [dot] com
Geoff Hinton, University of Toronto, hinton [at] cs [dot] toronto [dot] edu

The first time deep learning shows promise in speech recognition!

Microsoft Research

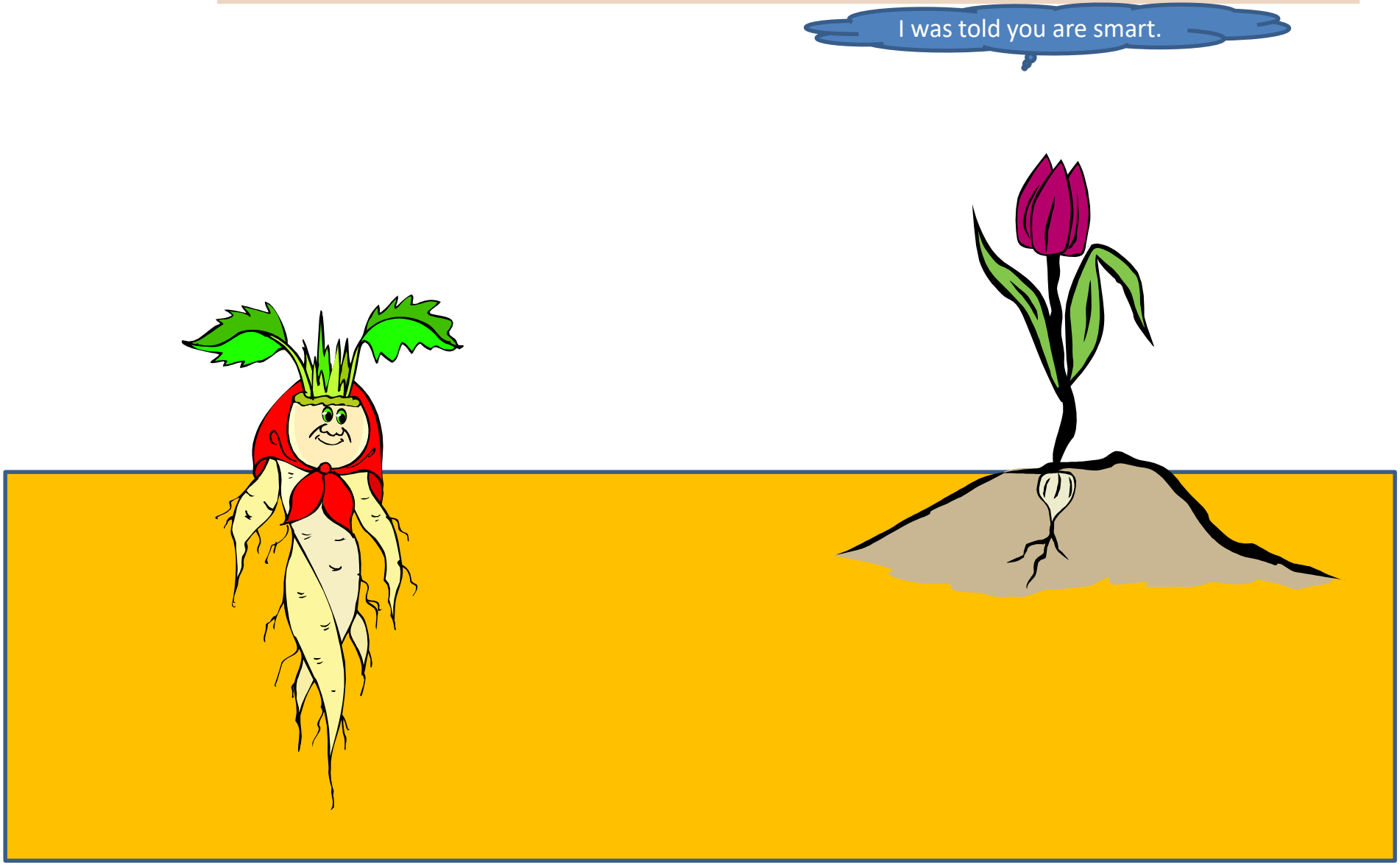
Anecdote: Speechless summary presentation of the NIPS 2009 Workshop on Speech

Deep Learning for Speech Recognition and Related Applications

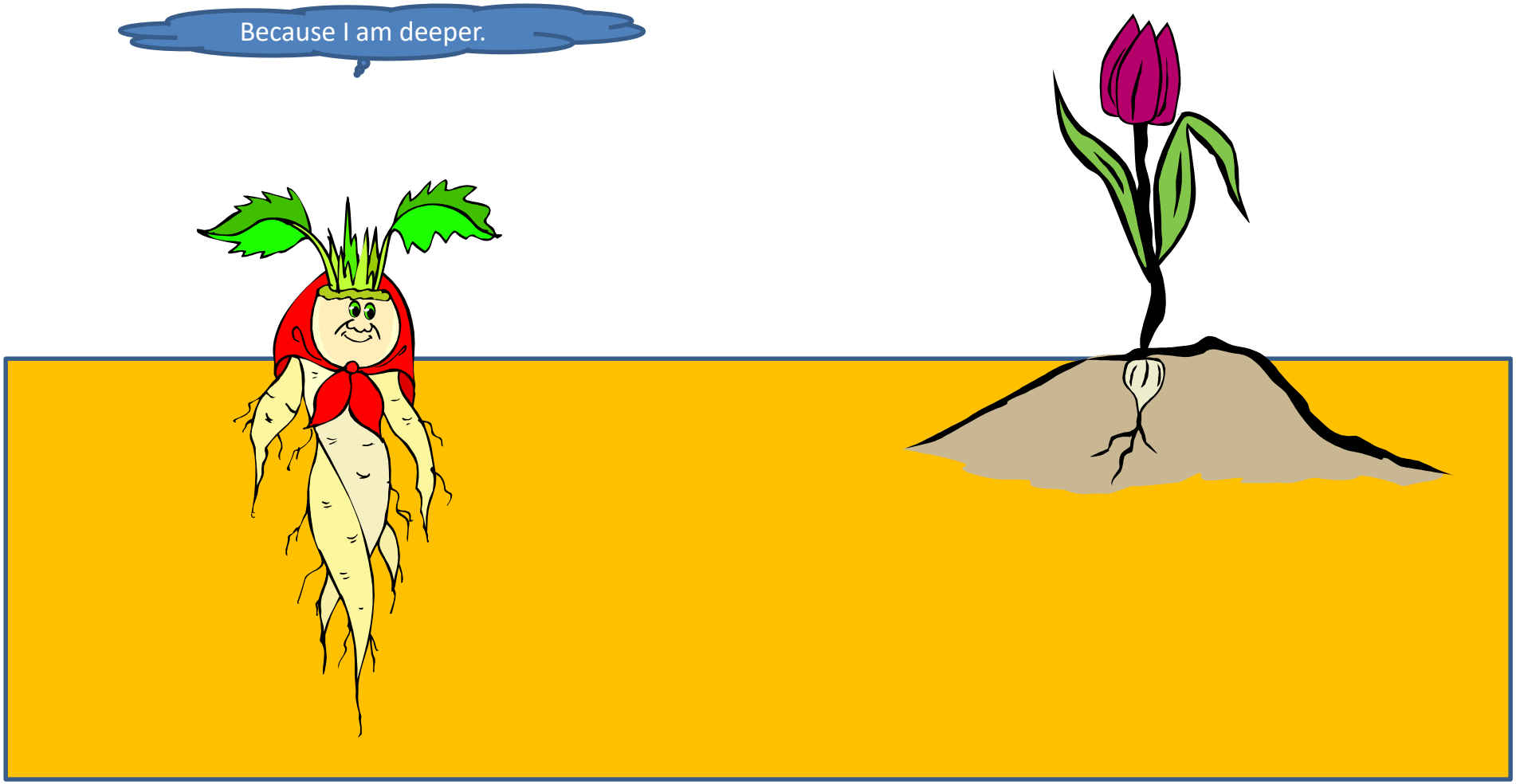
Li Deng, *Dong Yu,* Geoffrey Hinton


They met in year 2009…

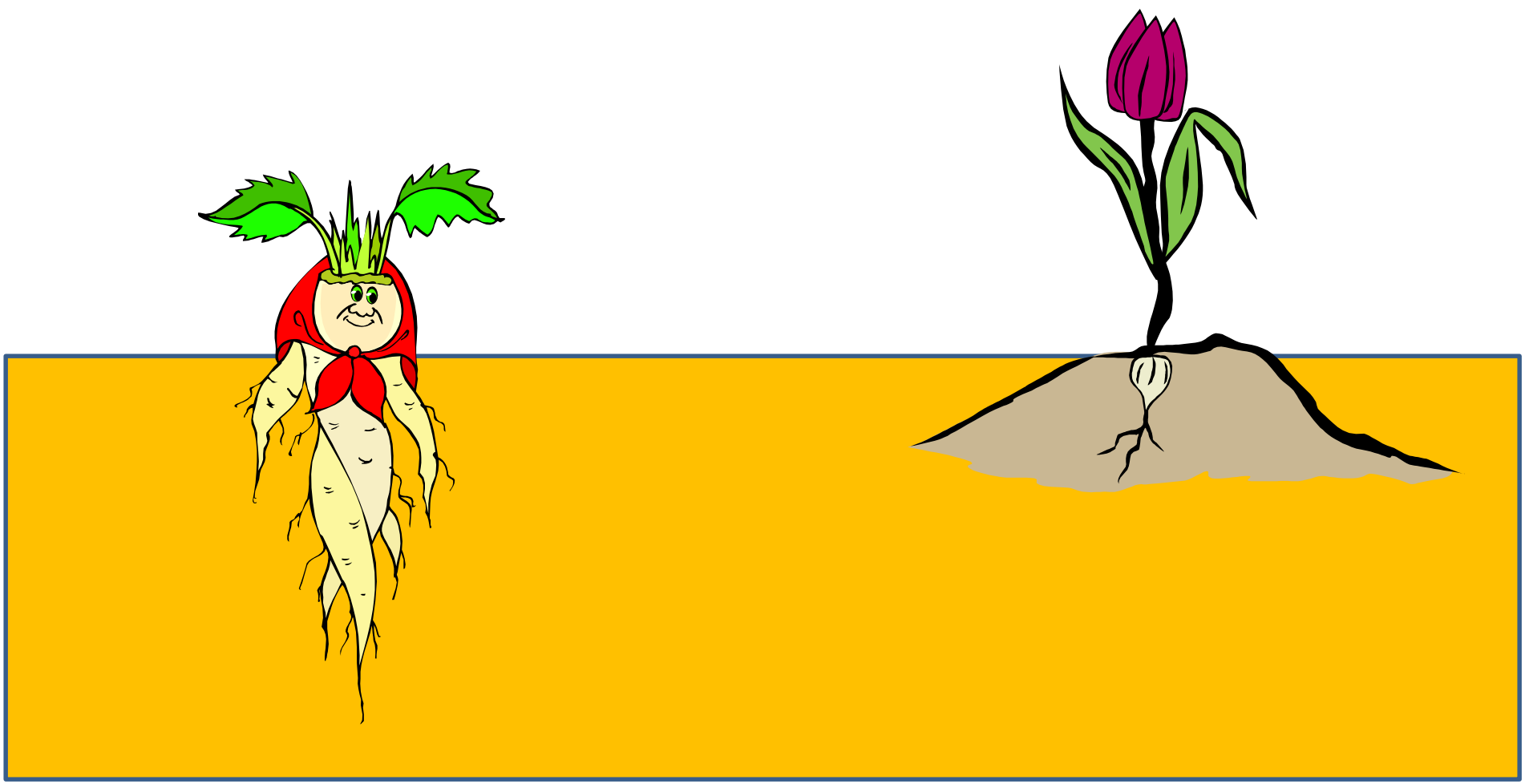
Microsoft Research



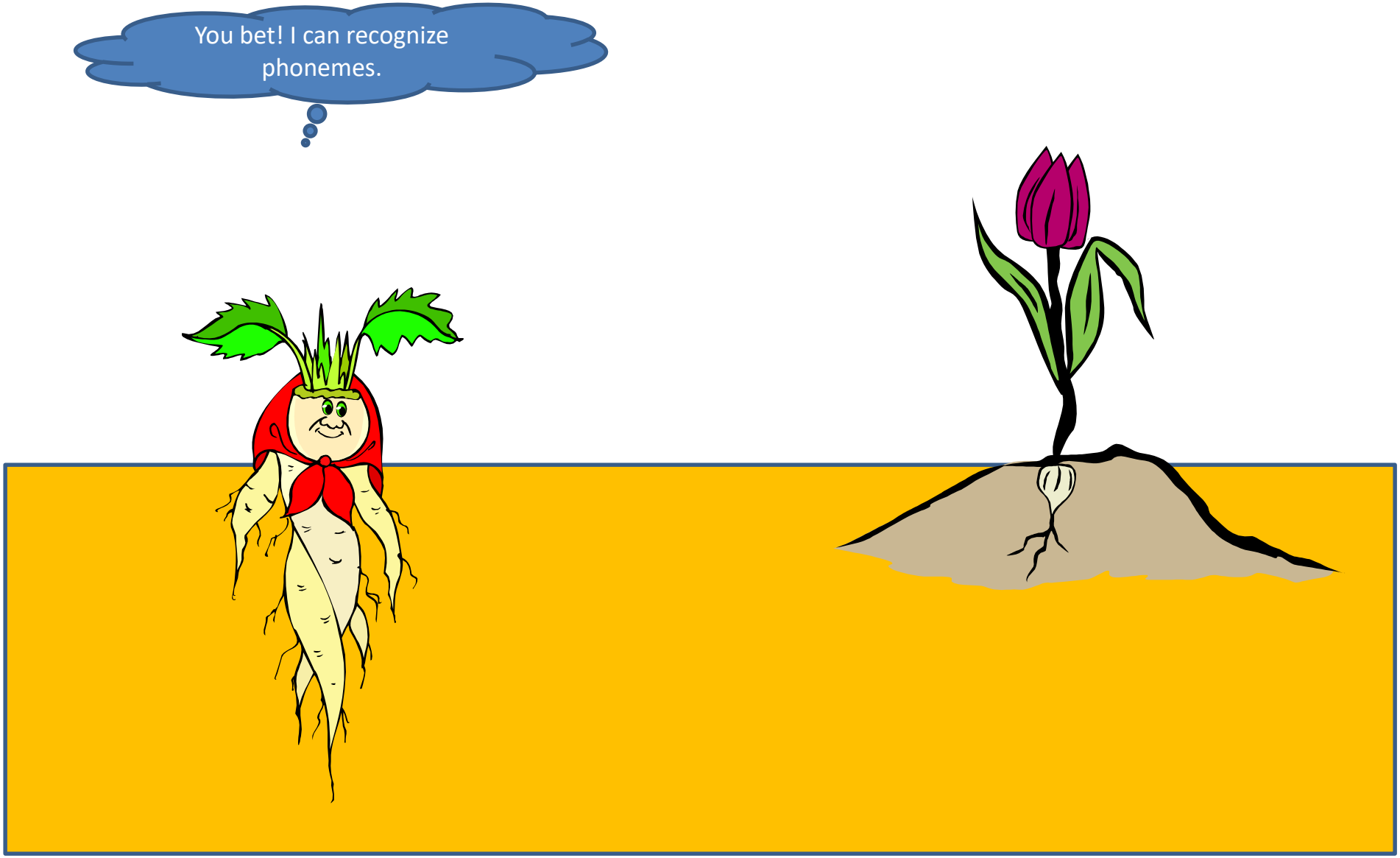
Microsoft Research



Microsoft Research

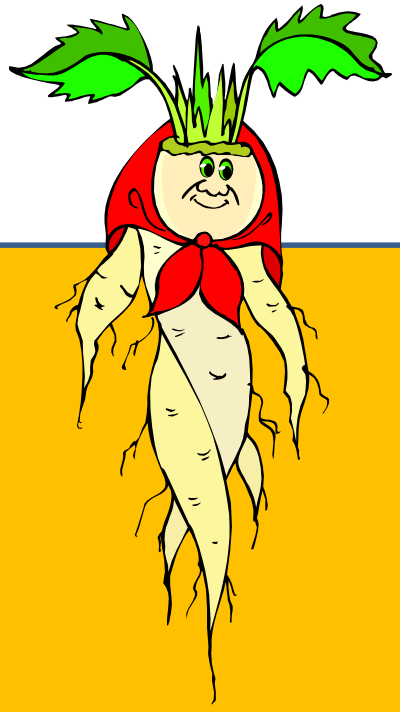


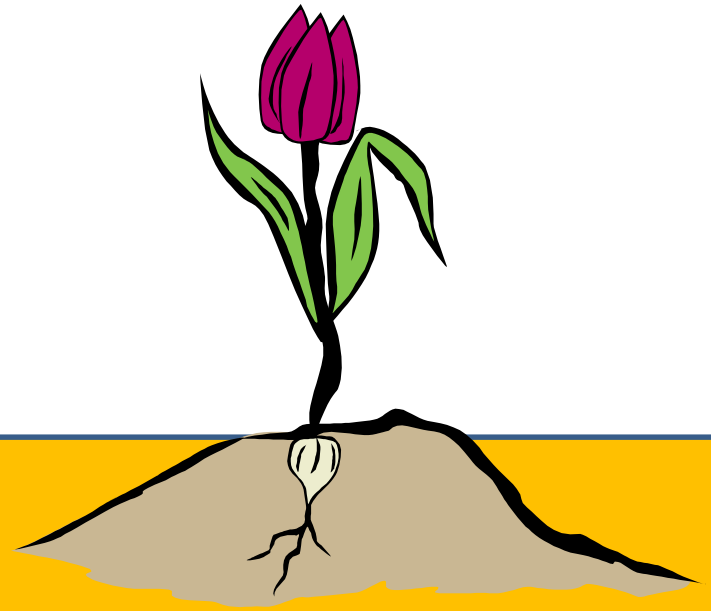
Microsoft Research



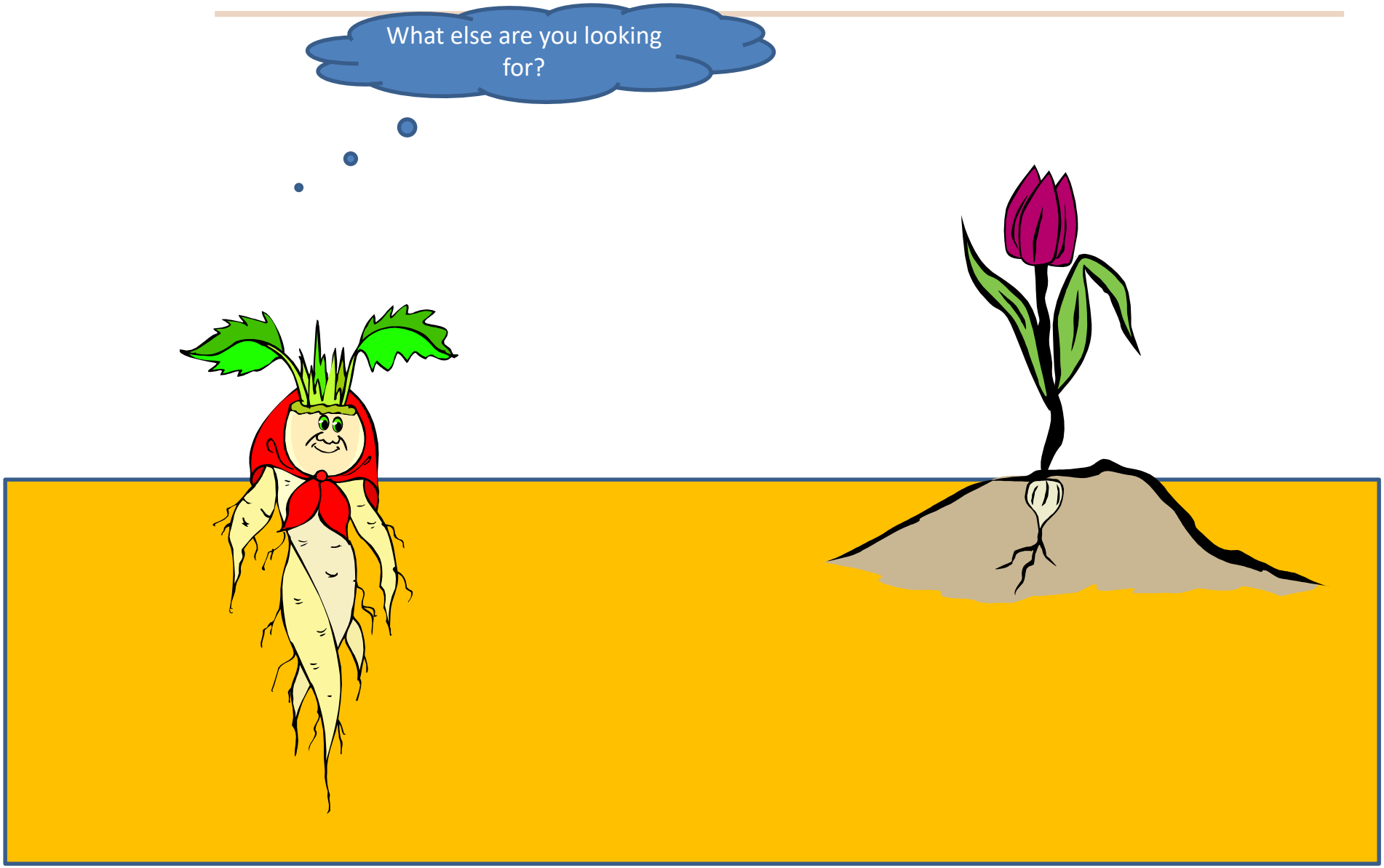
Microsoft Research

That's a nice first step!



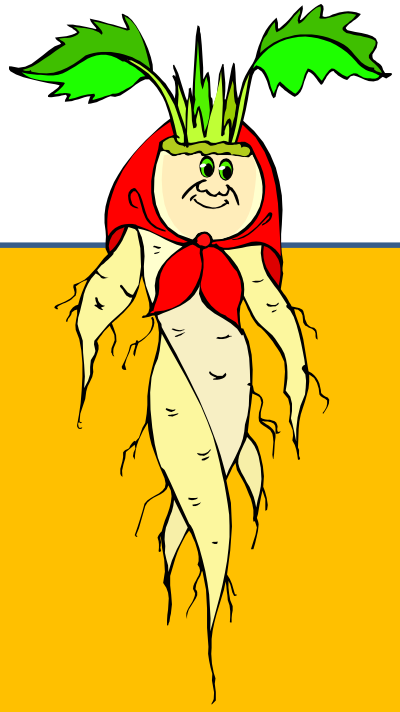


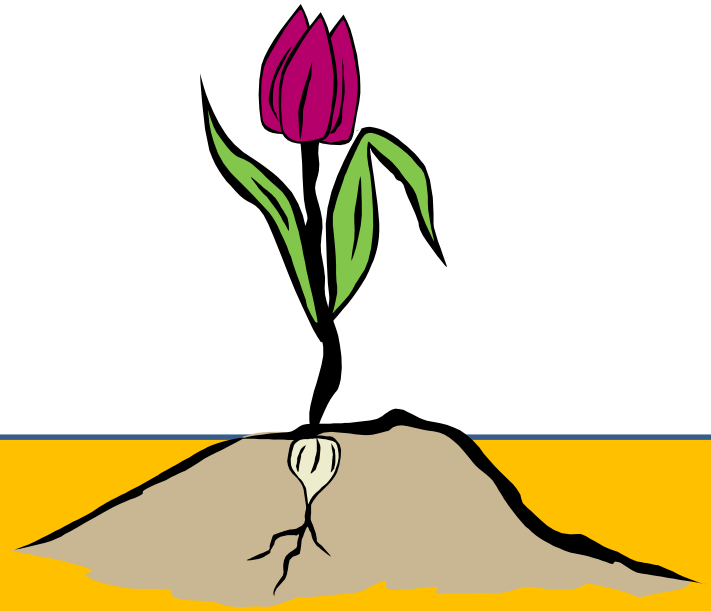
Microsoft Research



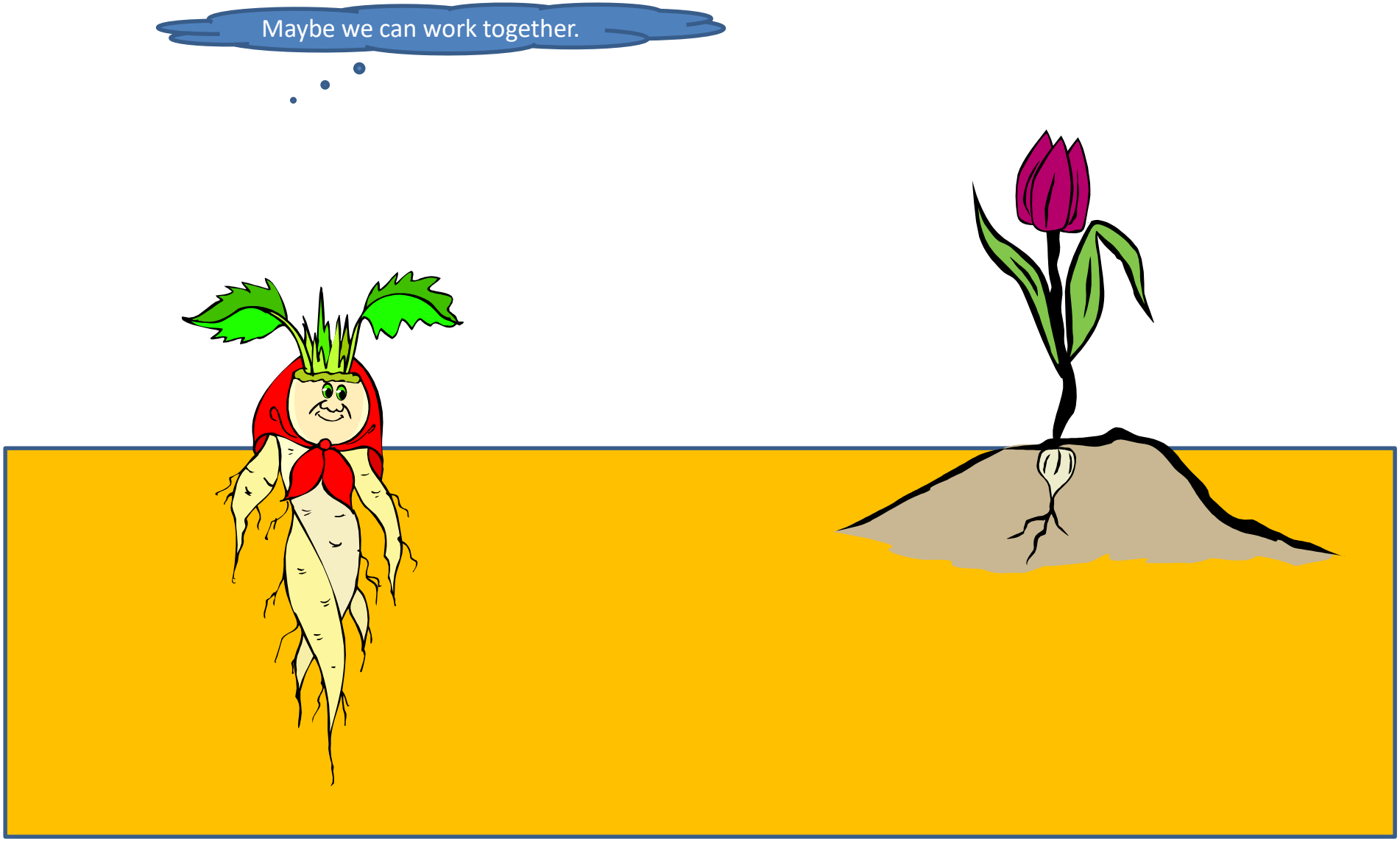
Microsoft Research

Recognizing noisy sentences spoken by unknown people.



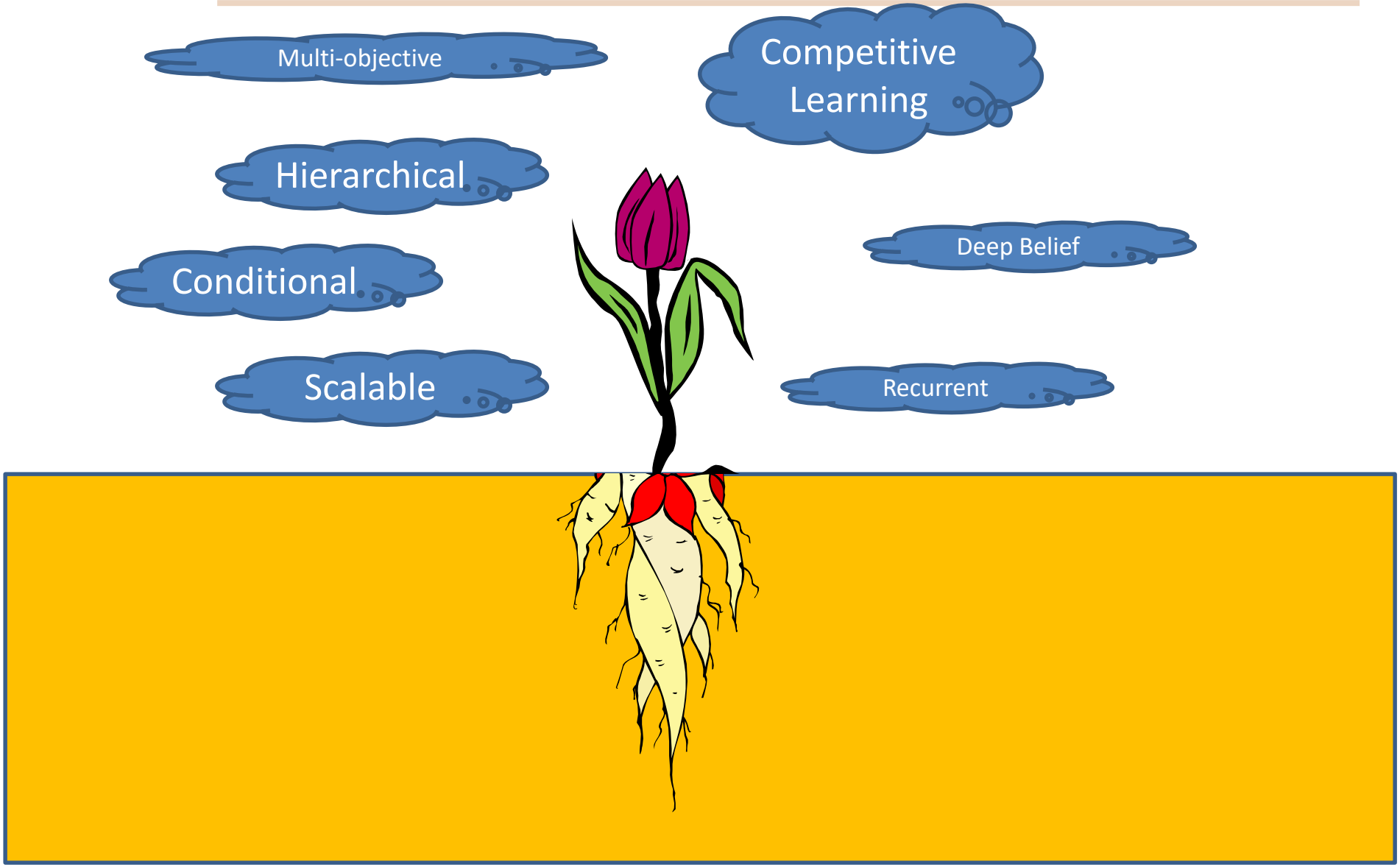


Microsoft Research



Microsoft Research

Deep speech recognizer is born.



Outline

- Introduction: Deep learning (DL) & its impact
- Part I: A (brief) history of "deep" speech recognition
- Part II: DL achievements in speech (and vision)
- Part III: DL Challenges: Language, mind, & deep intelligence

Microsoft Research

Expand DL at Industrial Scale

- Scale DL success to large industrial speech tasks (2010)
  - Grew output neurons from context-independent phone states (100-200) to context-dependent ones (1k-30k) → CD-DNN-HMM for Bing Voice Search tasks
  - Motivated initially by saving huge MSFT investment in the speech decoder software infrastructure (several choices available: **senones**, symbolic articulatory "features", etc)
  - CD-DNN-HMM gives much higher accuracy than CI-DNN-HMM
  - Earlier NNs made use of context only as appended inputs, not coded directly as outputs
- Engineering for large speech systems:
  - Combined expertise in DNN (w. GPU implementation) **and** speech recognition
  - Close collaborations among MSRR/MSRA and academic researchers:

George Dahl, Dong Yu, Li Deng, and Alex Acero, Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing (2013 IEEE SPS Best Paper Award)* , vol. 20, no. 1, pp. 30-42, January 2012.

Frank Seide, Gang Li and Dong Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", Interspeech 2011, pp. 437-440.

Geoffrey Hinton, Li Deng, Dong Yu, G. Dahl, A. Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath George Dahl, and Brian Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, in *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, November 2012

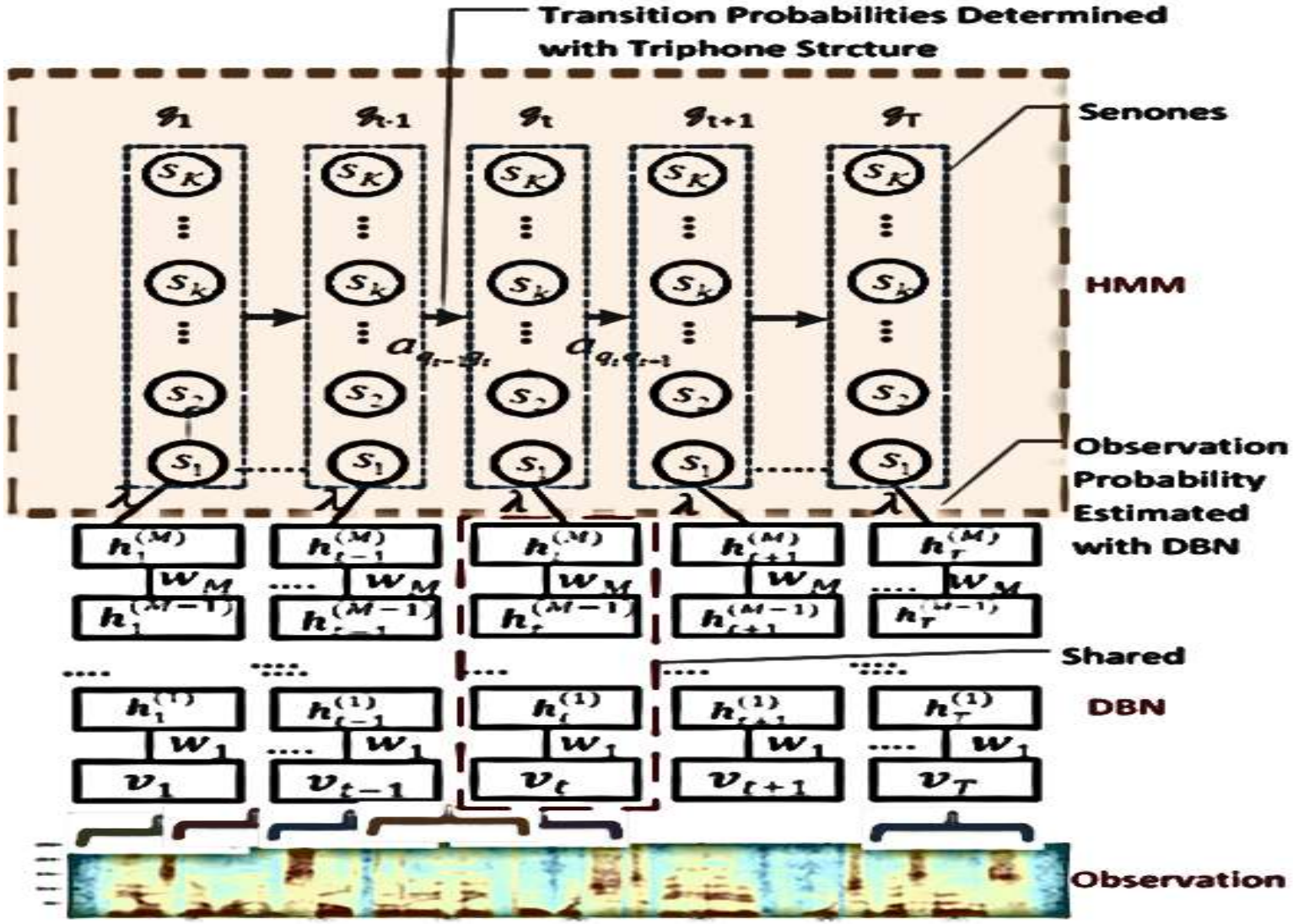
Microsoft Research

What Enabled CD-DNN-HMM?

- Industry knowledge of
  - how to construct very large CD output units in DNN
  - how to make decoding of such huge networks highly efficient using HMM technology
  - how to cut corner in making practical systems
- GPUs are optimized for fast matrix multiplications, major computation in CD-DNN **training**
- Nvidia's CUDA library for GPU computing released in 2008

Microsoft Research

CD-DNN-HMM



Microsoft Research

DNN-HMM vs. GMM-HMM

- **Table:** TIMIT Phone recognition (3 hours of training)

Features	Setup	Error Rates
GMM	Incl. Trajectory Model	24.8%
DNN	5 layers x 2048	23.0%

~10% relative improvement

- **Table:** Voice Search SER (24-48 hours of training)

Features	Setup	Error Rates
GMM	MPE (760 24-mix)	36.2%
DNN	5 layers x 2048	30.1%

~20% relative improvement

- **Table:** Switch Board WER (309 hours training)

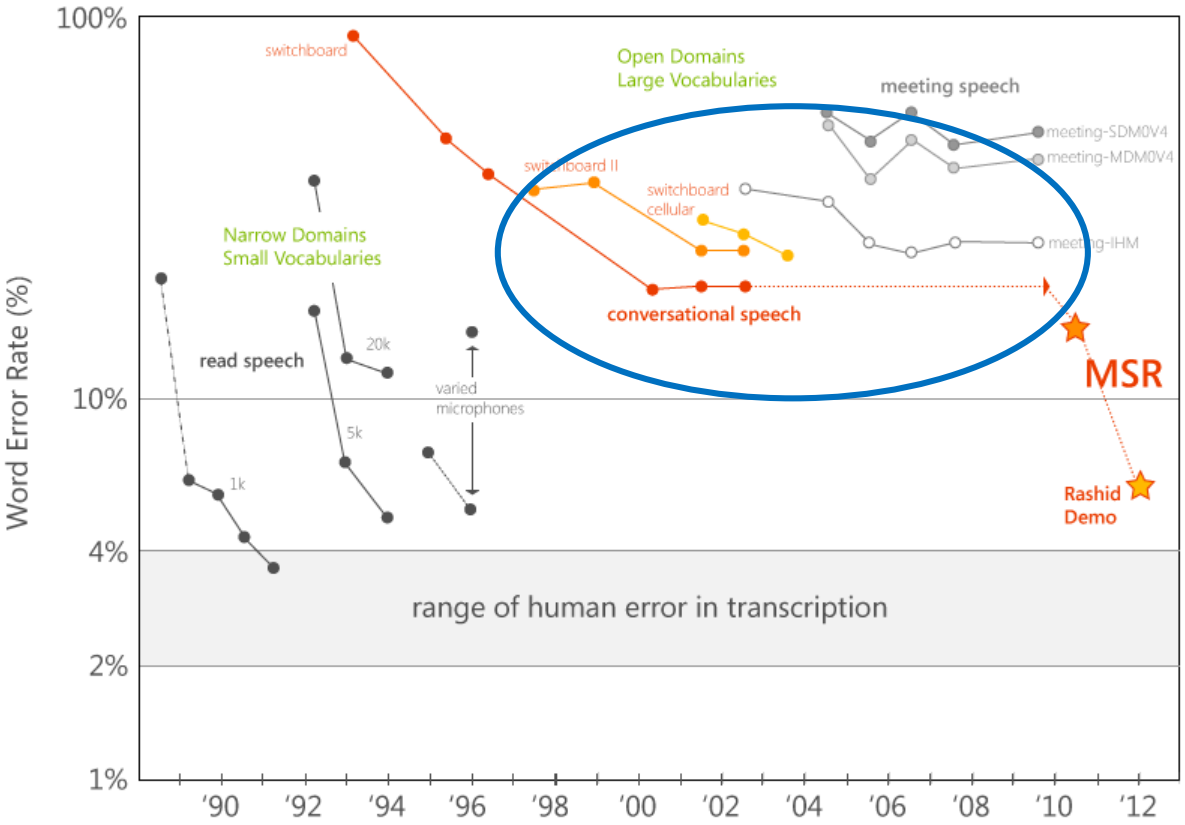
Features	Setup	Error Rates
GMM	BMMI (9K 40-mix)	23.6%
DNN	7 layers x 2048	15.8%

~30% relative improvement

- **Table:** Switch Board WER (2000 hours training)

Features	Setup	Error Rates
GMM	BMMI (18K 72-mix)	21.7%
DNN	7 layers x 2048	14.6%

NIST Evaluations of Automatic Speech Recognition



After no improvement for 10+ years by the research community…

…MSR reduced error from **~23%** to **<15%** (and under 7% for Rick Rashid's demo)!

Across-the-Board Deployment of DNN in ASR Industry

(also in universities; DARPA program)

EARLY DRAFT

Many limitations of early DNNs

- **Kluge 1:** keep the assumption of frame independence (ignore real "dynamics" to speed up decoding) but use bigger time windows
- **Kluge 2:** reverse the direction: instead of "deep generating" speech top-down, do "deep inference" bottom-up (using neural nets)
- **Kluge 3:** don't know how to train this deep neural net? Try DBN to initialize it.

IEEE Signal Processing Magazine, Nov. 2012

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

The shared views of four research groups

New Types of Deep Neural Network & Learning for Speech Recognition+

An Overview

Li Deng, Geoffrey Hinton, Brian Kingsbury

MSR, U. Toronto/Google, IBM

ICASSP Special Session, May 28, 2013

Microsoft

UNIVERSITY OF TORONTO

Google

IBM Research

Five Technical Papers in the Special Session

RECENT ADVANCES IN DEEP LEARNING FOR SPEECH RESEARCH AT MICROSOFT

IMPROVING DEEP NEURAL NETWORKS FOR LVCSR USING RECTIFIED LINEAR UNITS AND DROPOUT

DEEP CONVOLUTIONAL NEURAL NETWORKS FOR LVCSR

IBM Research

MULTILINGUAL ACOUSTIC MODELS USING DISTRIBUTED DEEP NEURAL NETWORKS

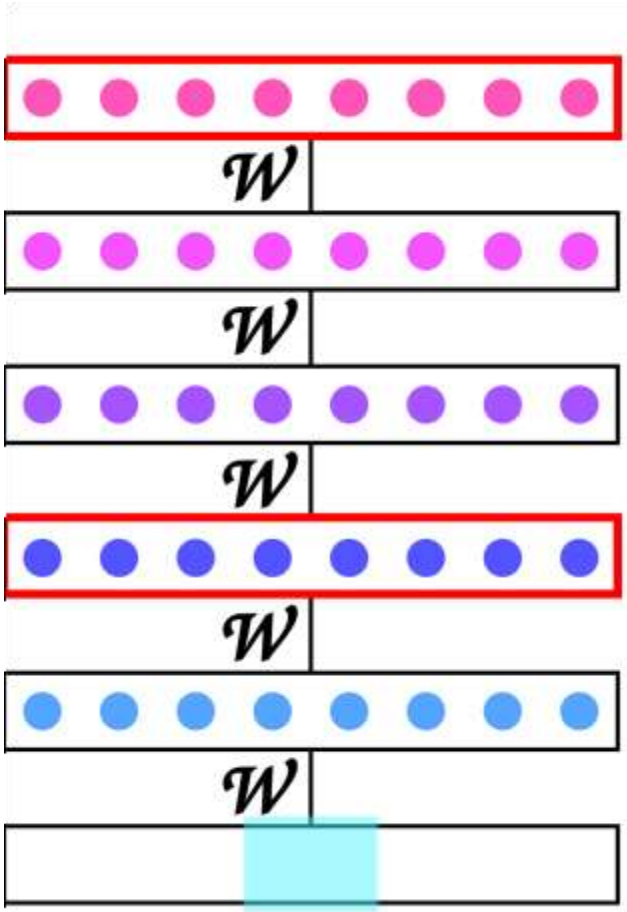
Google

ADVANCES IN OPTIMIZING RECURRENT NETWORKS

Université de Montréal

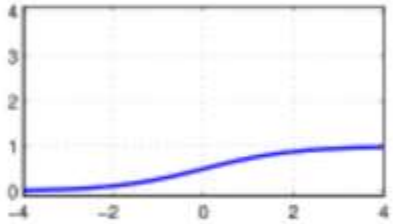
Plus 40+ deep learning papers on ASR sessions at **ICASSP-2014**

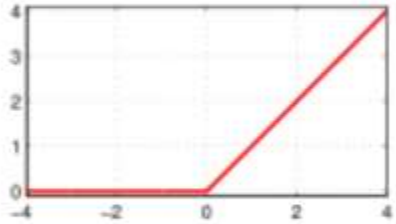
Themes: Nonlinearities



softmax ⟶ NADE, linear SVM,
linear+stacking (DSN)
…

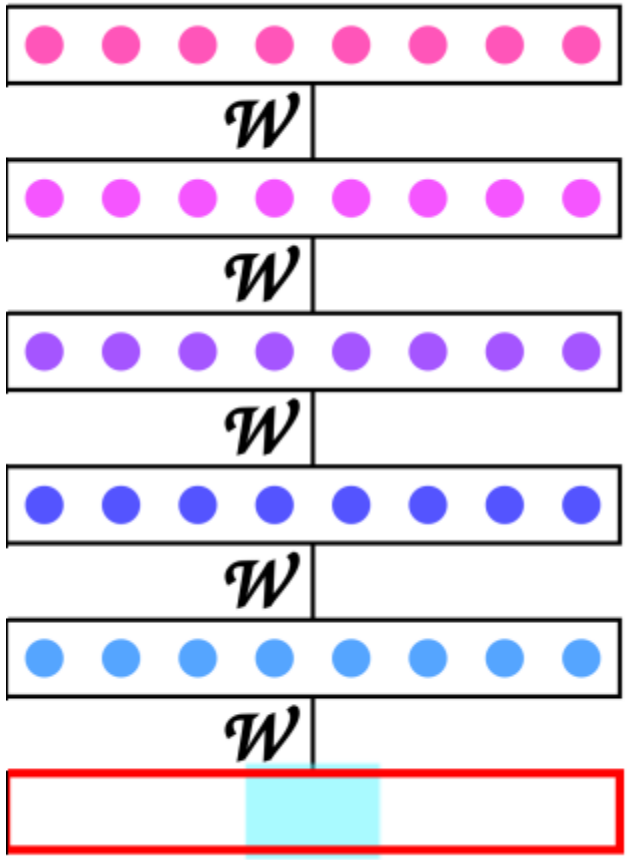
logistic ⟶ ReLU , MaxOut, WTA...





…, …

Themes: Better Inputs



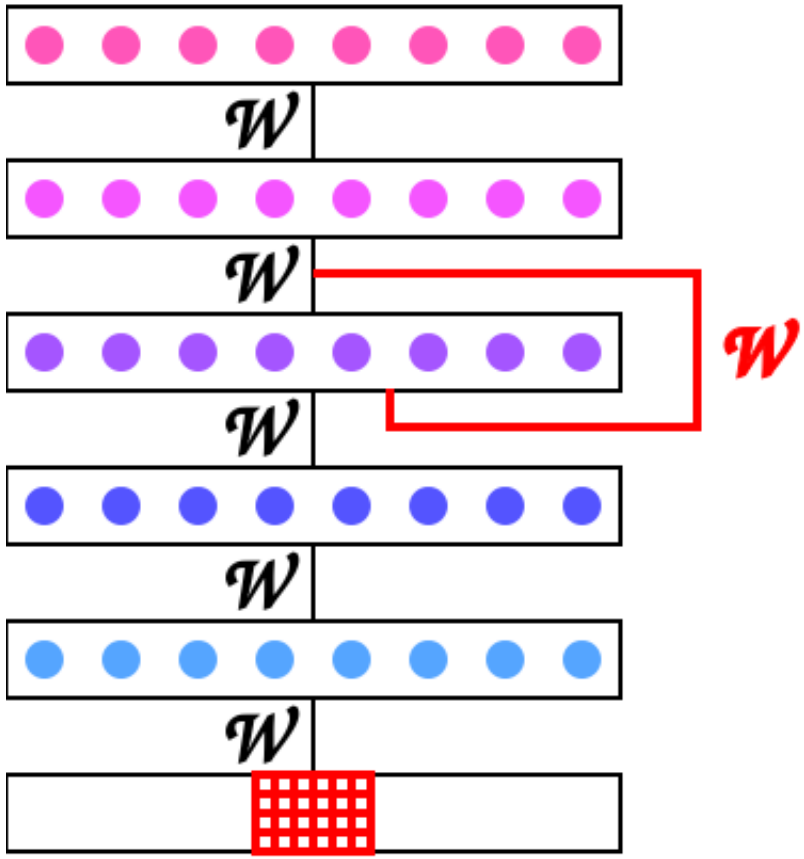
MFCCs ⟶ log Mel spectra (2013)

→ Log linear spectra (2014)

→ Linear spectra?

→ Waveforms?

Themes: Better architectures



W Recurrent NNs

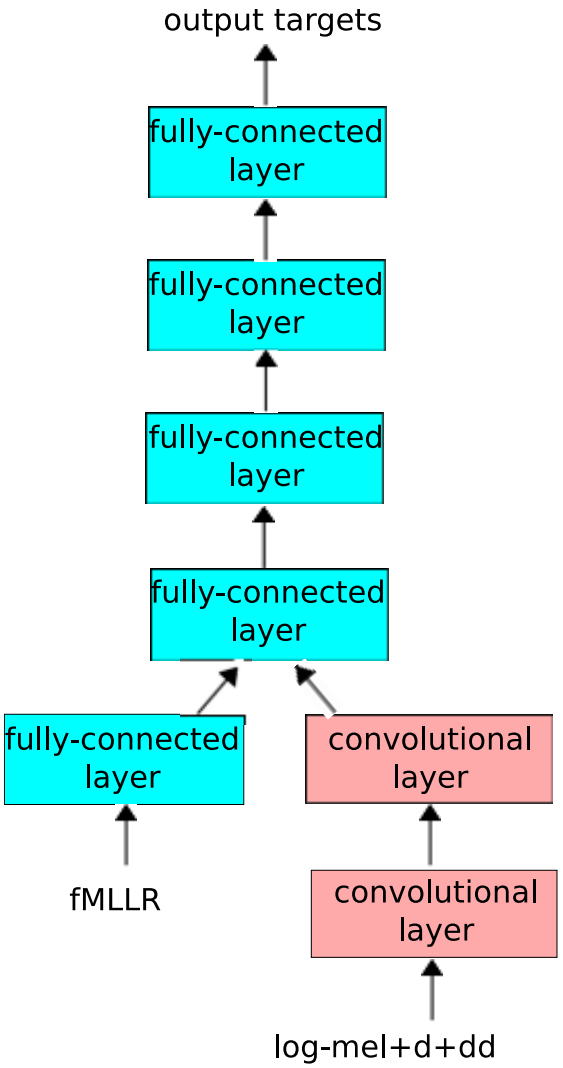
LSTM-RNNs (2014)

Convolutional NNs

Joint CNN/DNN (2014)

Joint CNN/DNN Architecture

- CNNs are good at modeling locally-correlated features while DNNs are good for features which do not have this structure
- fMLLR+ivector features are fed to a fully-connected DNN layer
- log-mel features are fed into a convolutional network
- The entire network is trained jointly
- ReLU provides additional improvements for some tasks
- Papers
  - [H. Soltau, G. Saon and T. N. Sainath – ICASSP 2014]
  - [T.N. Sainath et al, submitted to Special Issue on Deep Learning of Representation, 2014]



Slide provided by Tara Sainath and Brian Kingsbury (June 2014)

Results on 300 hr SWB

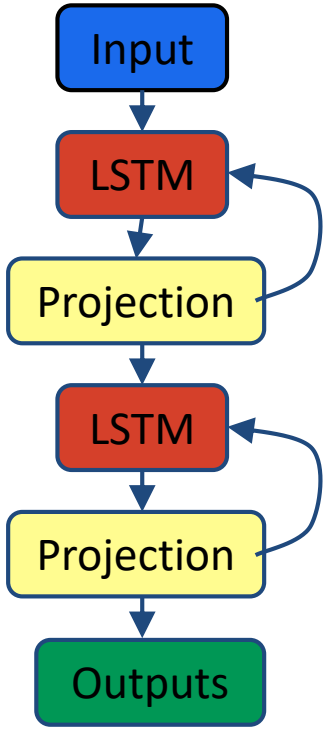
- Networks are cross-entropy and sequence trained
- Including speaker adaptation, multi-scale architecture, sigmoid, we can achieve an additional 10% relative improvement in WER over just CNN alone

Model	feature	Hub5'00
Baseline GMM/HMM	fBMMI	14.5
Hybrid DNN	fMLLR	12.2
Hybrid CNN	Log-mel	11.8
CNN+DNN	Log-mel + (fMLLR+i-vectors)	**10.4**

Slide provided by Tara Sainath and Brian Kingsbury (June 2014)

Google Speech recognition

- Task:
  - Google Now/Voice search / mobile dictation
  - Streaming, real-time recognition in 50 languages
- Model:
  - Deep Projection Long-Short Term Memory Recurrent Neural networks
  - Distributed training with asynchronous gradient descent across hundreds of machines.
  - Cross-entropy objective (truncated backpropagation through time) followed by sequence discriminative training (sMBR).
  - 40-dimensional filterbank energy inputs
  - Predict 14,000 acoustic state posteriors



Slide provided by Andrew Senior, Vincent Vanhoucke, Hasim Sak (June 2014)

Google

LSTM Large vocabulary speech recognition

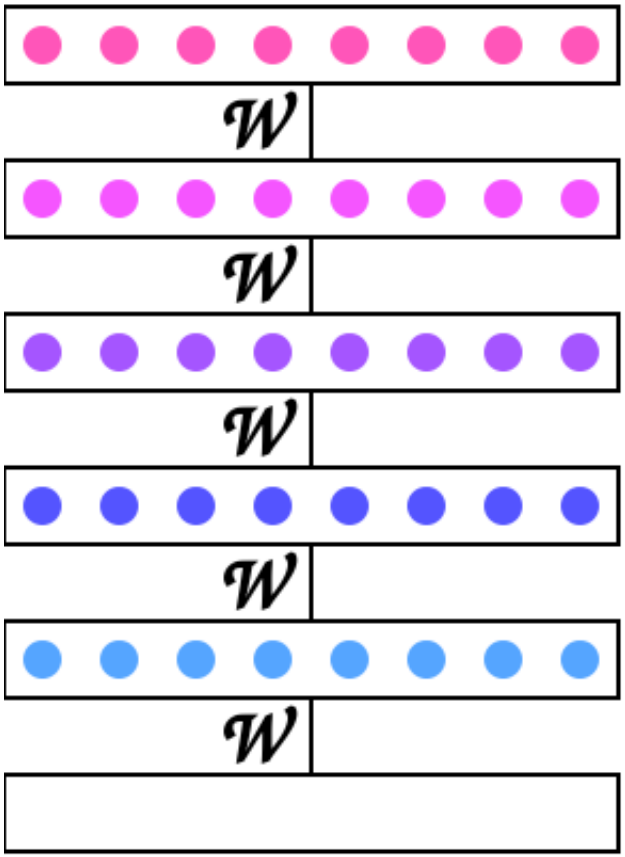
Models	Parameters	Cross-Entropy	sMBR sequence training
ReLU DNN	85M	11.3	10.4
Deep Projection LSTM RNN (2 layer)	13M	10.7	9.7

- ***Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*** **H. Sak, A. Senior, F. Beaufays** to appear in Interspeech 2014
- ***Sequence Discriminative Distributed Training of Long Short-Term Memory Recurrent Neural Networks*** **H. Sak, O. Vinyals, G. Heigold A. Senior, E. McDermott, R. Monga, M. Mao** to appear in Interspeech 2014

Voice search task; Training data: 3M utterances (1900 hrs); models trained on CPU clusters

Slide provided by Andrew Senior, Vincent Vanhoucke, Hasim Sak (June 2014)

Themes: Optimization



Distributed asynchronous SGD

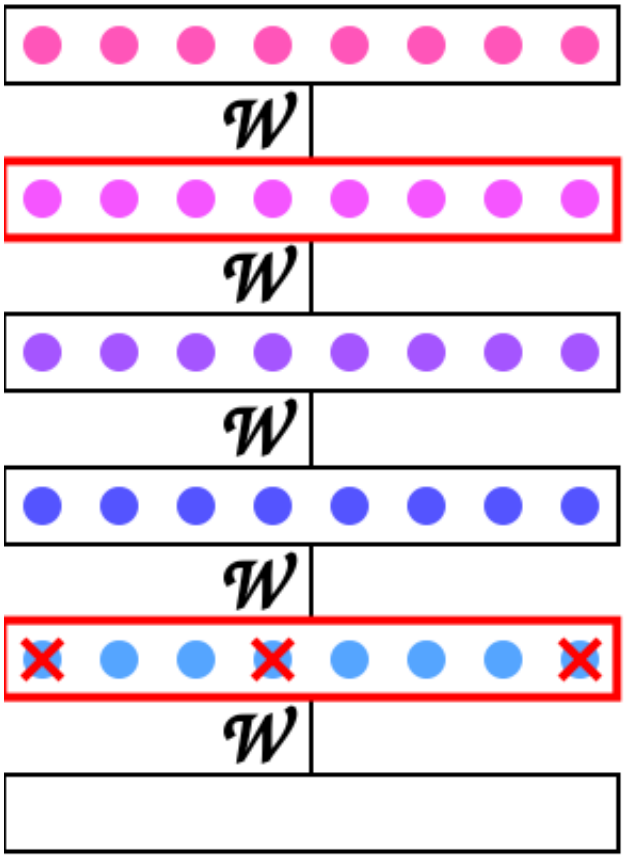
Distributed Hessian-free optimization

Nesterov's accelerated gradient method

Gradient clipping

Primal-dual method (2014)

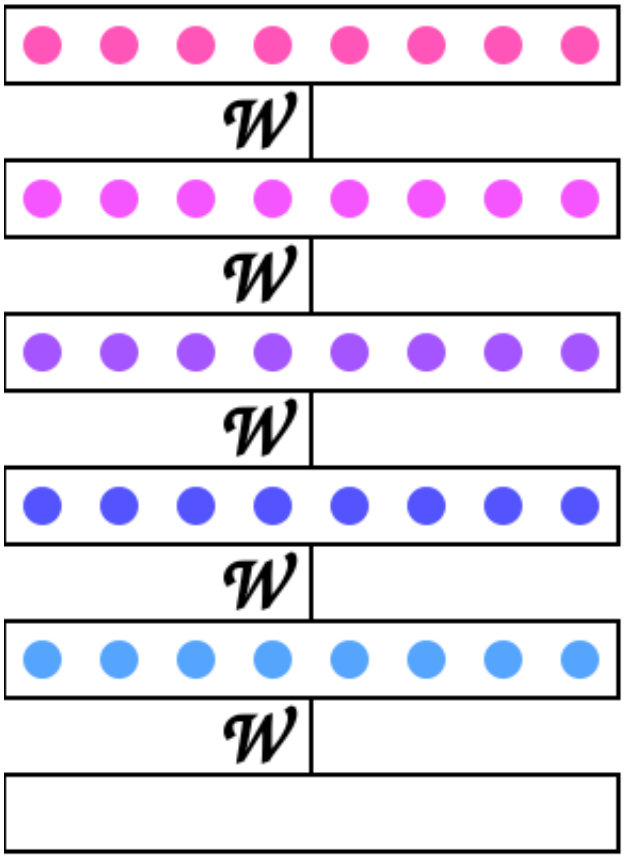
Themes: Regularization



Sparsity in hidden representations

Dropout

Themes: Multi-task Learning



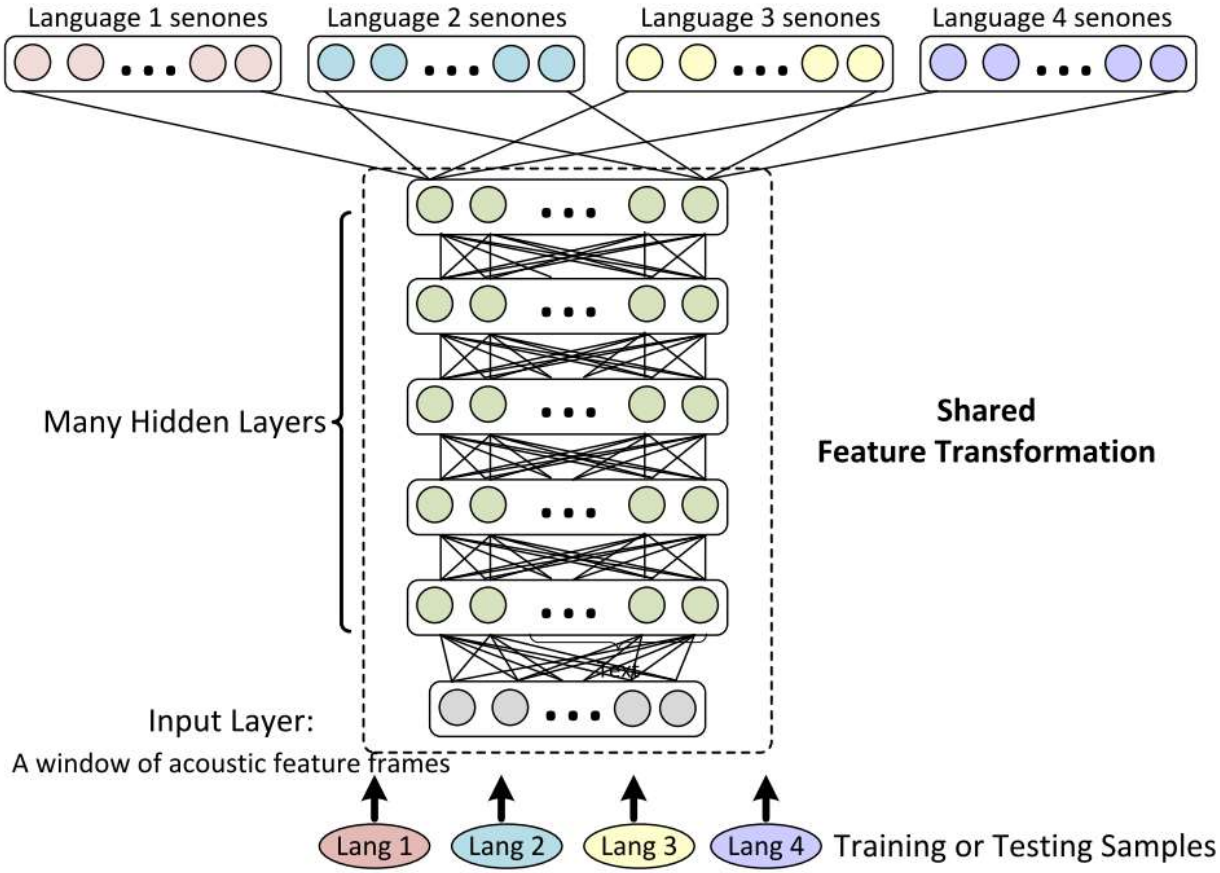
Multi-lingual acoustic modeling

Mixed-bandwidth acoustic modeling

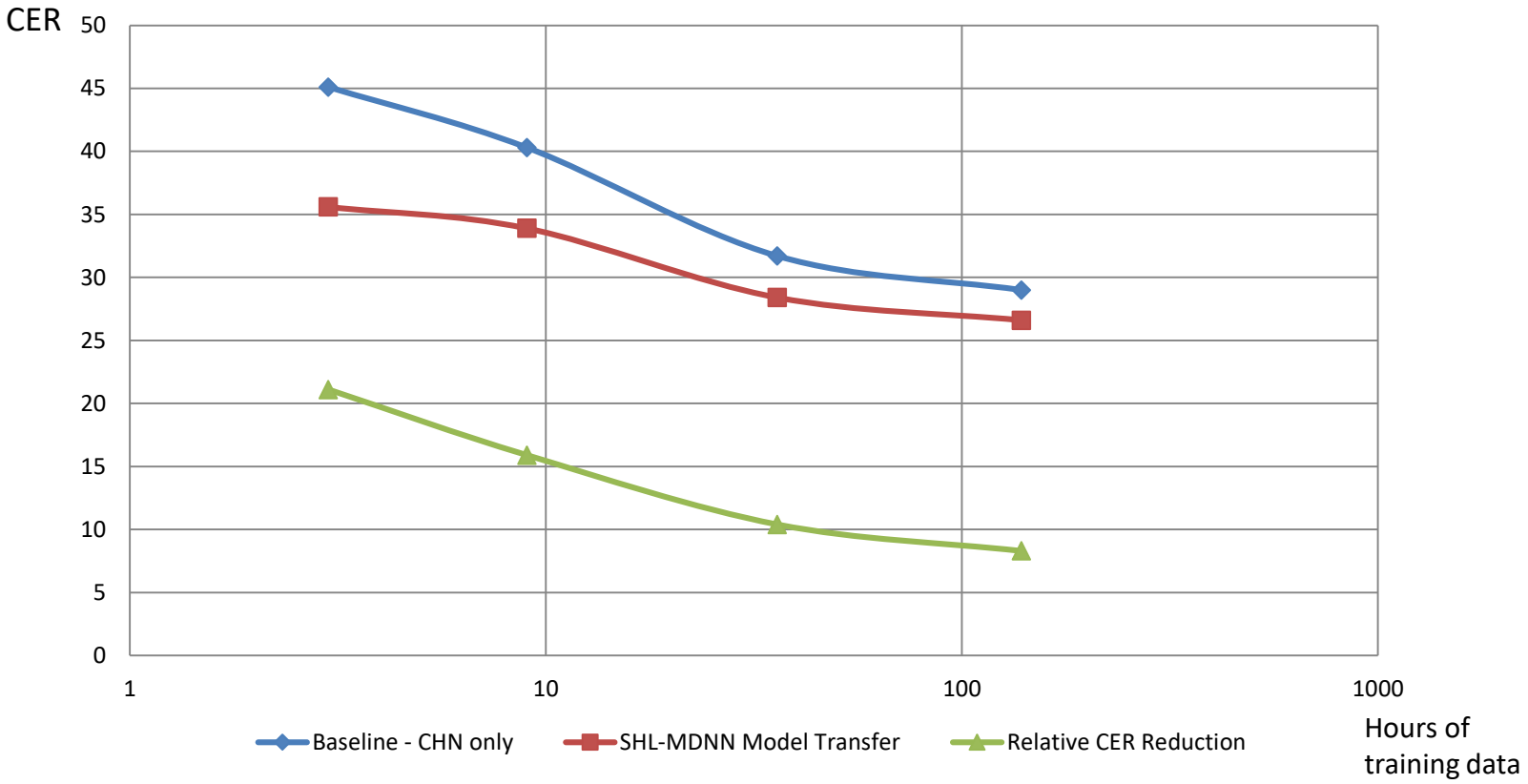
Microsoft

Shared-Hidden-Layer Multi-Lingual DNN

- J.-T. Huang et. al. “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *Proc. ICASSP*, 2013.



Chinese Speech Recognition Character-Error-Rate when Transferring from European Languages



Target language: zh-CN
Non-native source languages: FRA: 138 hours, DEU: 195 hours, ESP: 63 hours, and ITA: 93 hours of speech.

Microsoft

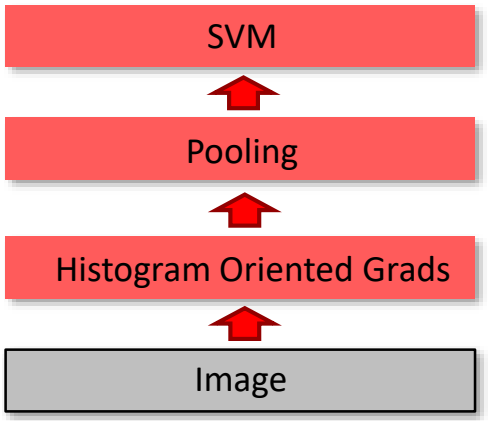
(Slide provided by Jinyu Li)

Achievements of Deep Learning in Object Recognition (Vision)

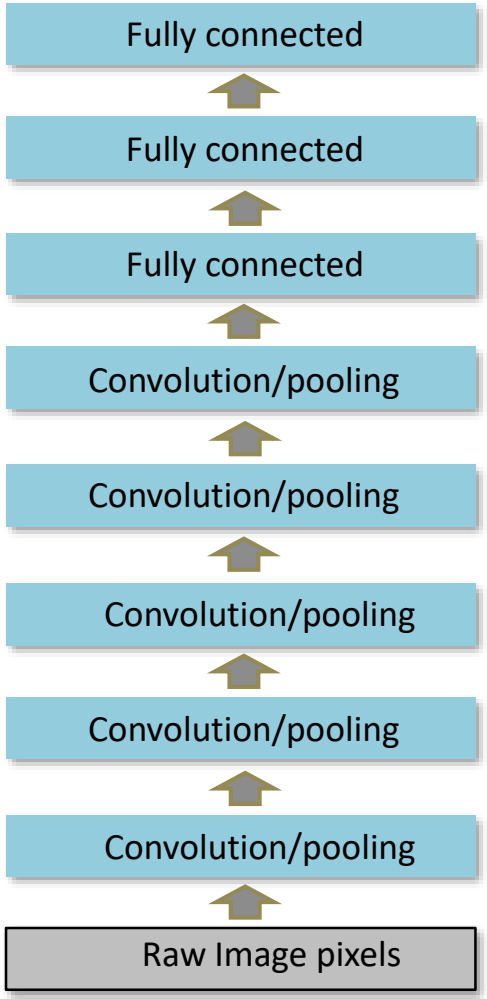
Deep **Convolutional NN** for Image Recognition

CNN: local connections with weight sharing; pooling for translation invariance

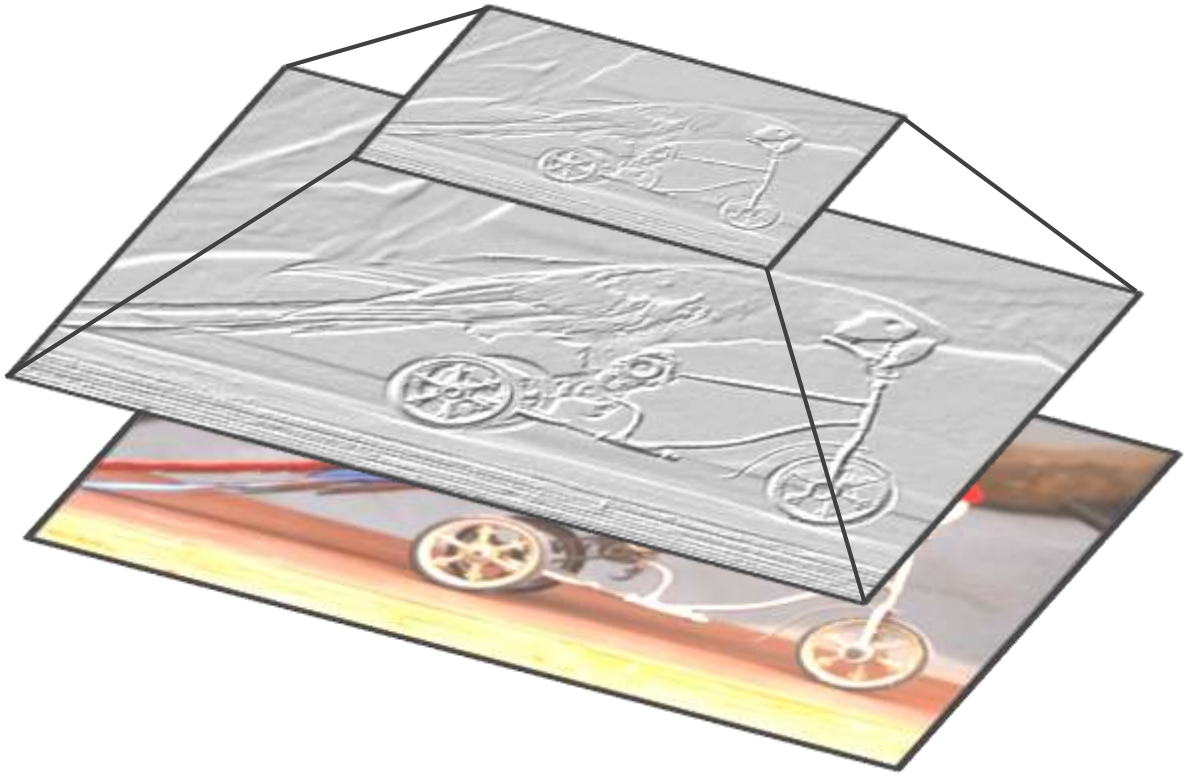
earlier



2012-2013



A Basic Module of the CNN

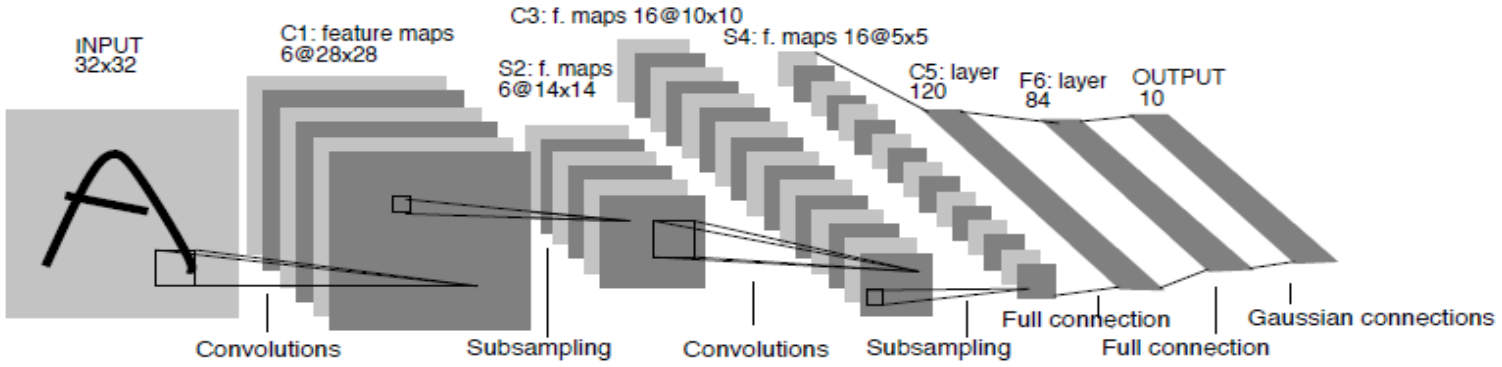


Pooling

Convolution

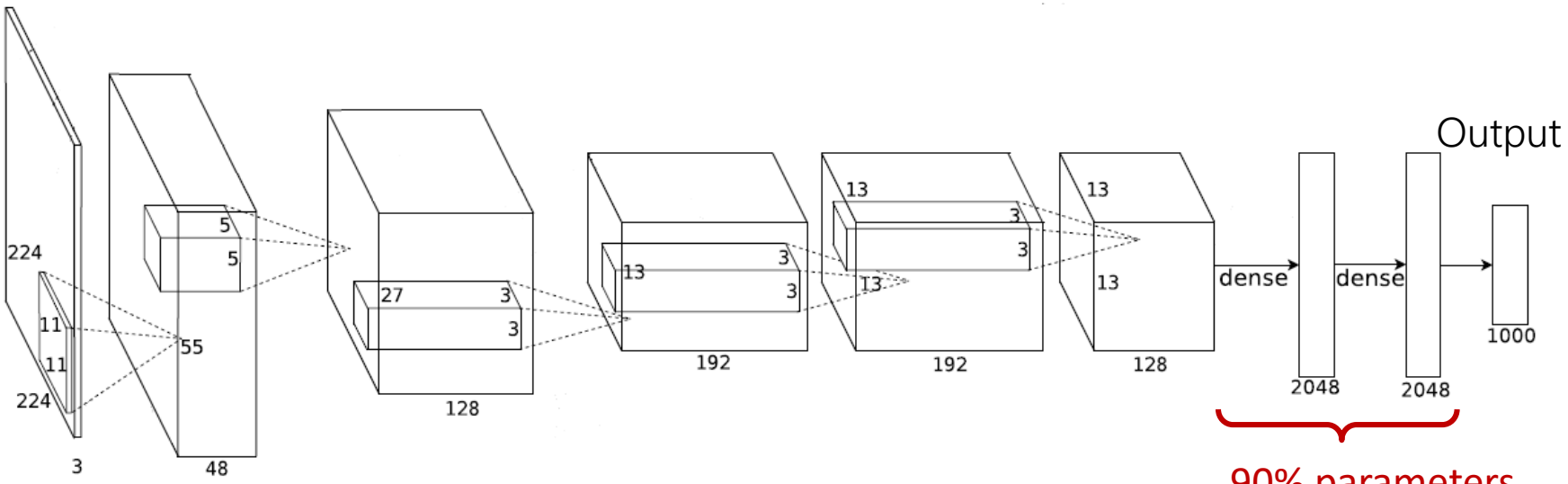
Image

Deep CNN



Image

LeCun et al., 1998

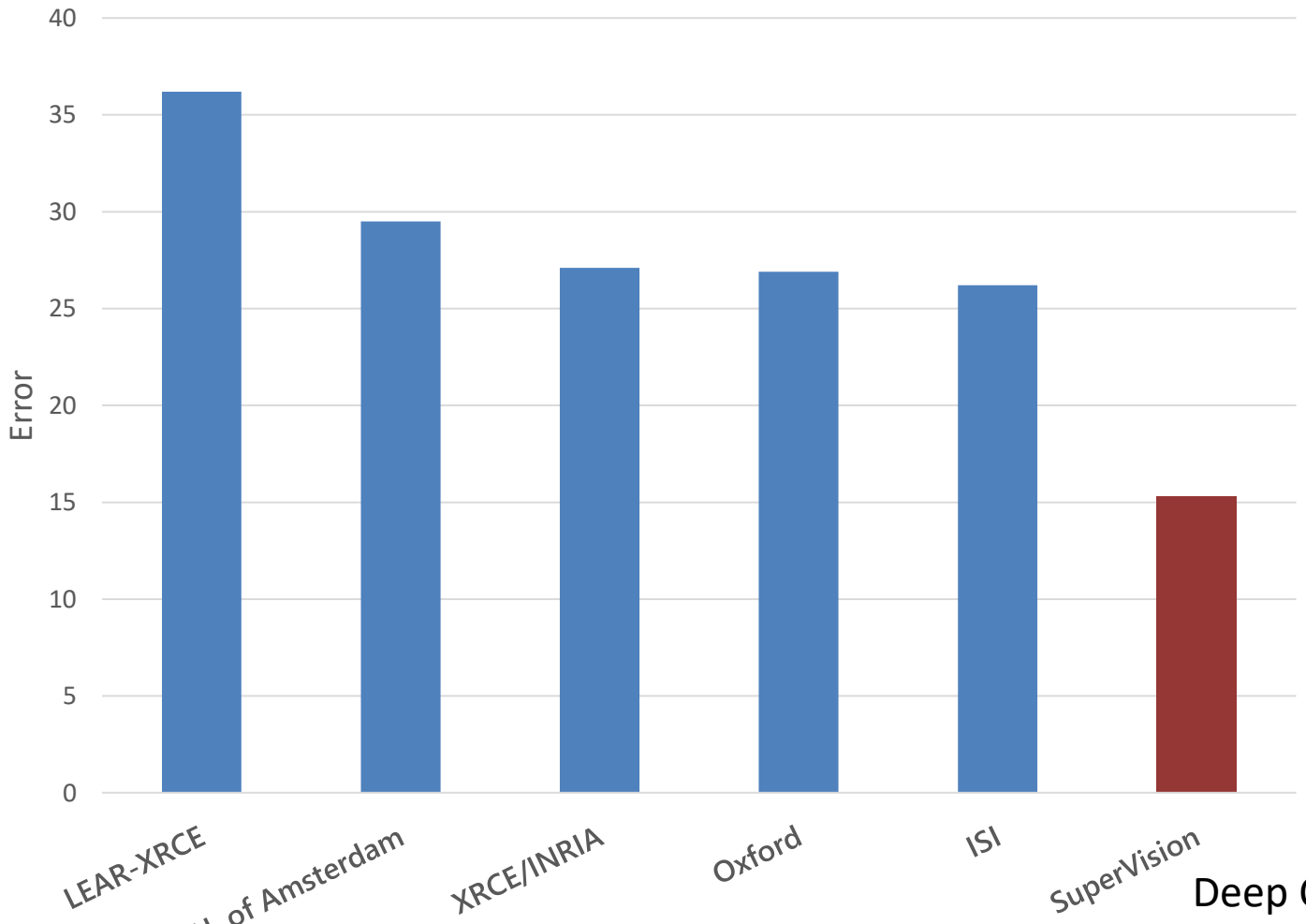


90% parameters

SuperVision, 2012

ImageNet 1K Competition

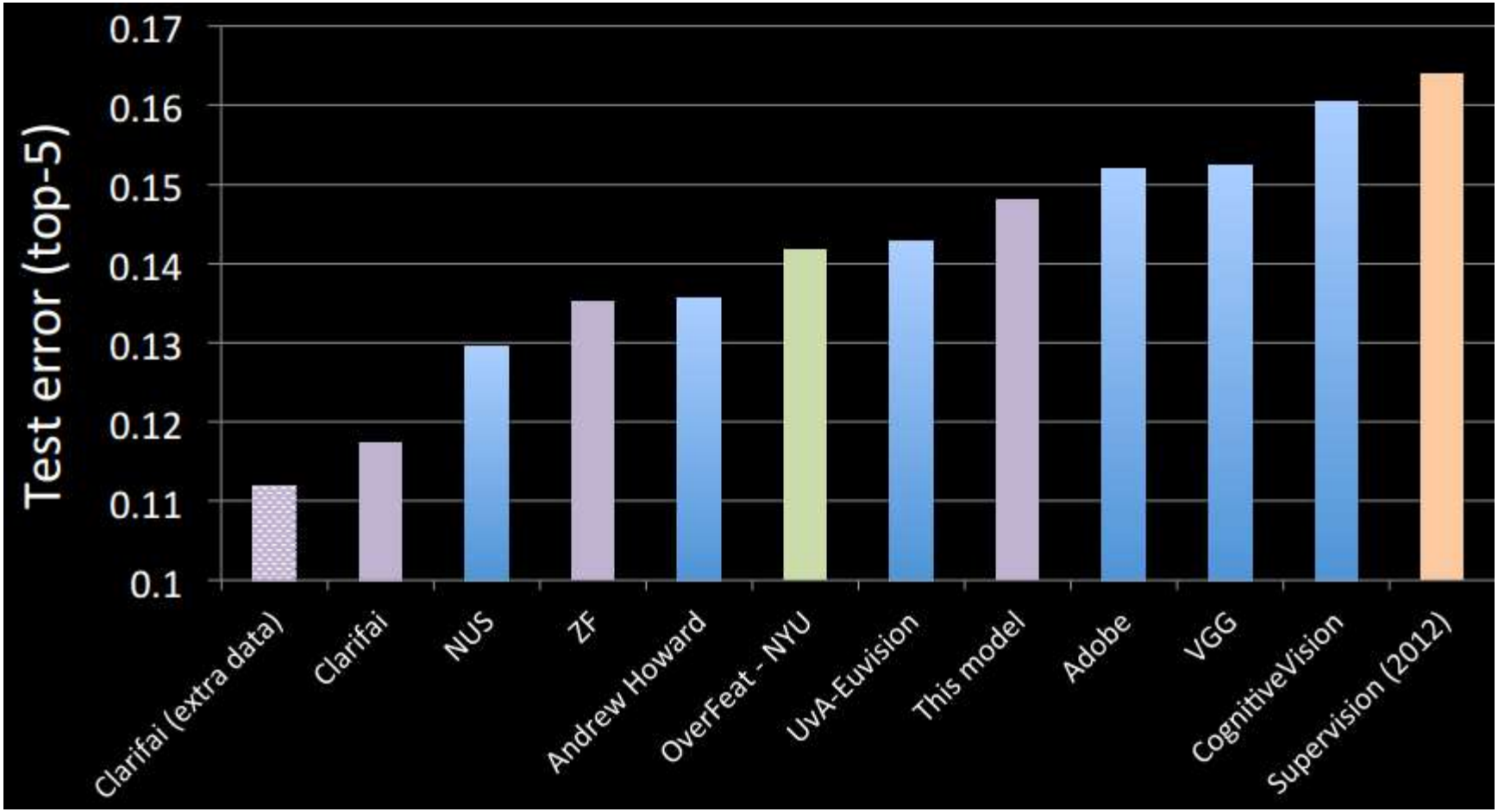
(Fall 2012)



Deep CNN
Univ. Toronto team

Same ImageNet 1K Competition

One year later (Fall 2013)



Summary results of ImageNet ***Large Scale Visual Recognition Challenge 2013 (ILSVRC2013), representing the state-of-the-art performance of object recognition systems.***

Outline

- Introduction: Deep learning (DL) & its impact
- Part I: A (brief) history of "deep" speech recognition
- Part II: DL achievements in speech and vision
- **Part III: DL challenges: Language, mind, & deep intelligence**
  - from DNN to deep semantic modeling (different problems/approaches)
  - **DSSM** developed at MSR (and related models)
  - functional modeling of the brain/mind for deep intelligence

DSSM paper: Huang, He, Gao, Deng, Acero, Heck, "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data," in CIKM, Oct. 2013

Two Key Concepts

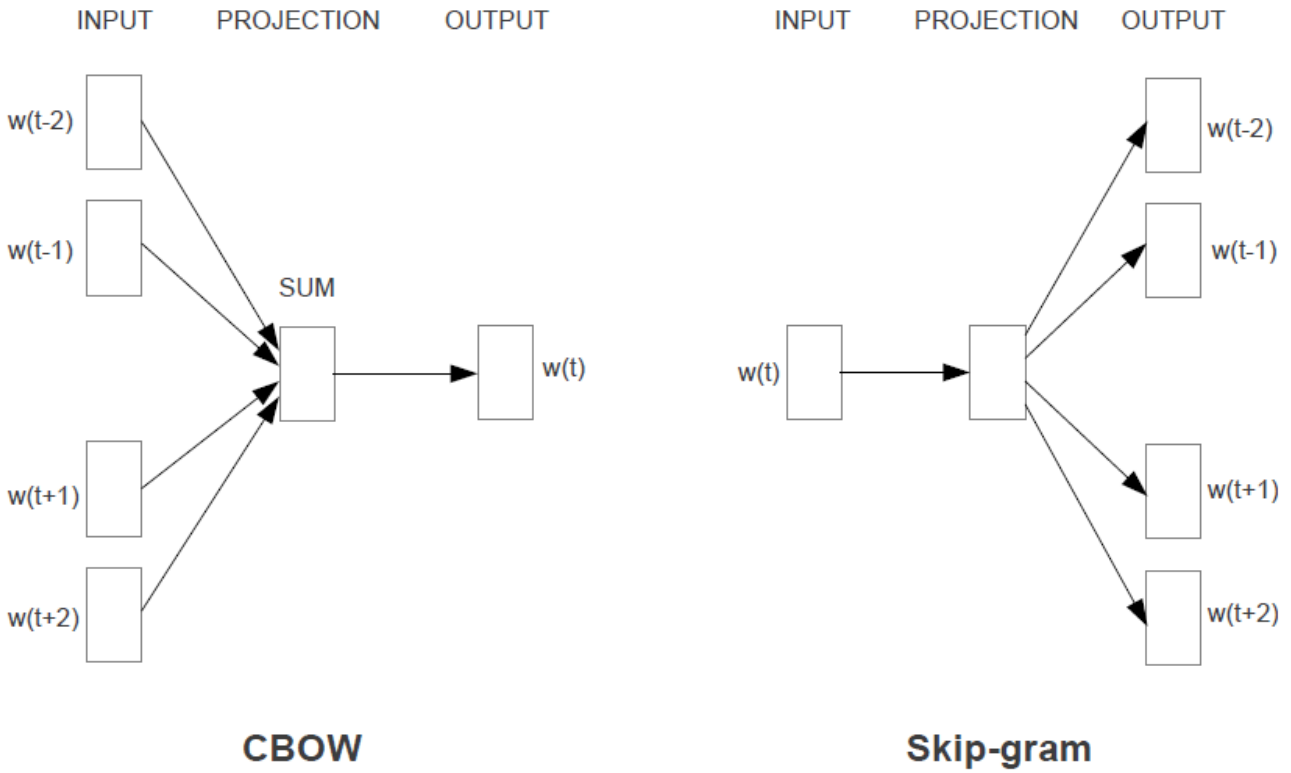
- **Each linguistic (or physical) entity or a simple "relation"**
  ←→ A continuous vector (embedding)

- A collection of such embedded vectors
  ←→ symbolic semantic structure (e.g., trees)

Then, reasoning in symbolic-space (traditional AI) can be beautifully carried out in the continuous-space in human cognitive and neural-net terms

Example of Word Embedding: Word2vec

- Word2vec
  - To derive continuous distributed representations of words via embedding
  - Training data: text corpus consisting of 1-of-V vectors of single words
  - Model file: word vectors expressed as low-dimension embedding vectors)
- **CBOW and Skip-gram**



- Mikolov et al, Efficient Estimation of Word Representations in Vector Space, ICLR 2013.
- Mikolov et al, Distributed Representations of Words and Phrases and their Compositionality, NIPS 2013.
- https://code.google.com/p/word2vec/

Another Example of Word Embedding:

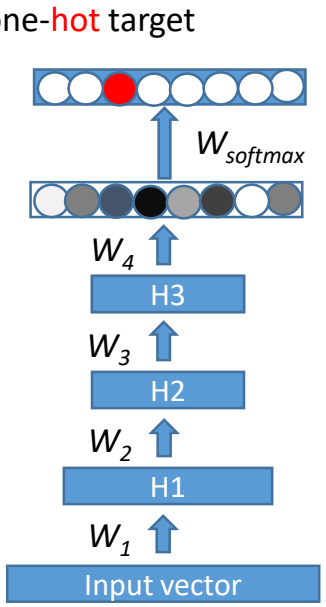
- DSSM --- deriving embedding vectors in a weakly supervised manner

- From DNN to DSSM

Huang, He, Gao, Deng, Acero, Heck, "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data," in CIKM, Oct. 2013

DNN (deep neural net)

- Targets for full supervision available
- Most powerful for classification with tons of labeled data (e.g. speech recognition and ImageNet tasks)
- Target: one-hot vector



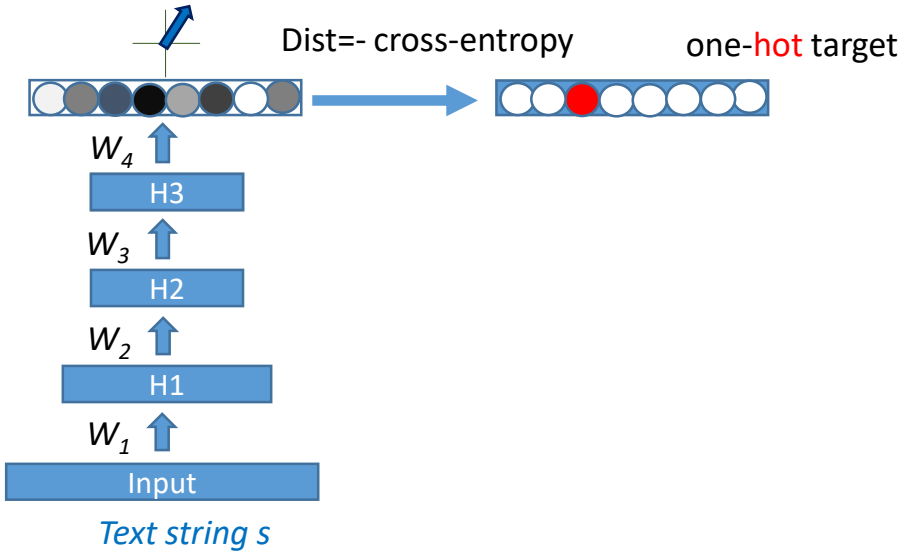
Cross-entropy: (negative) loss fctn or distance for learning

Top-hidden layer: DNN-derived feature vector

Text string or speech/image frame

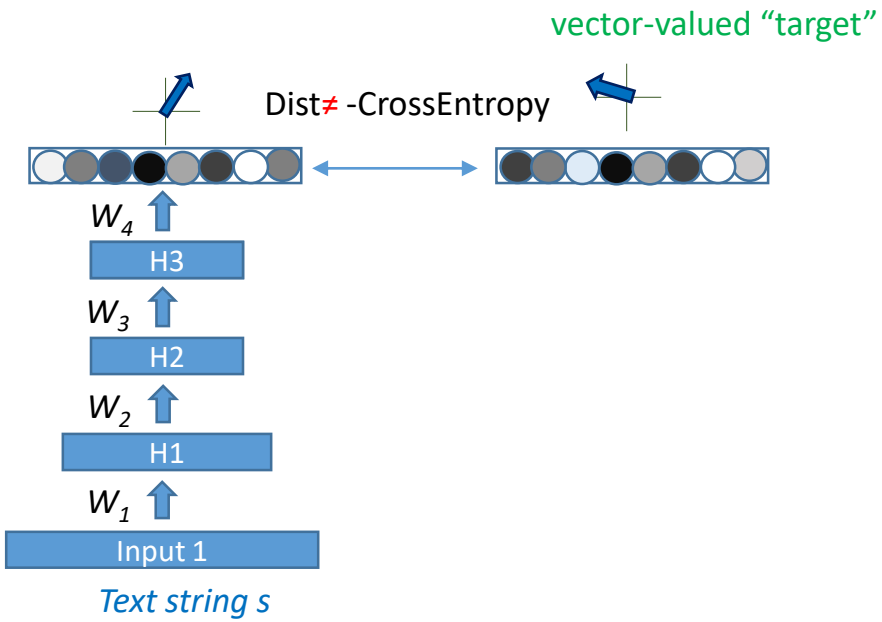
DNN (deep neural net)

- Redrawing:



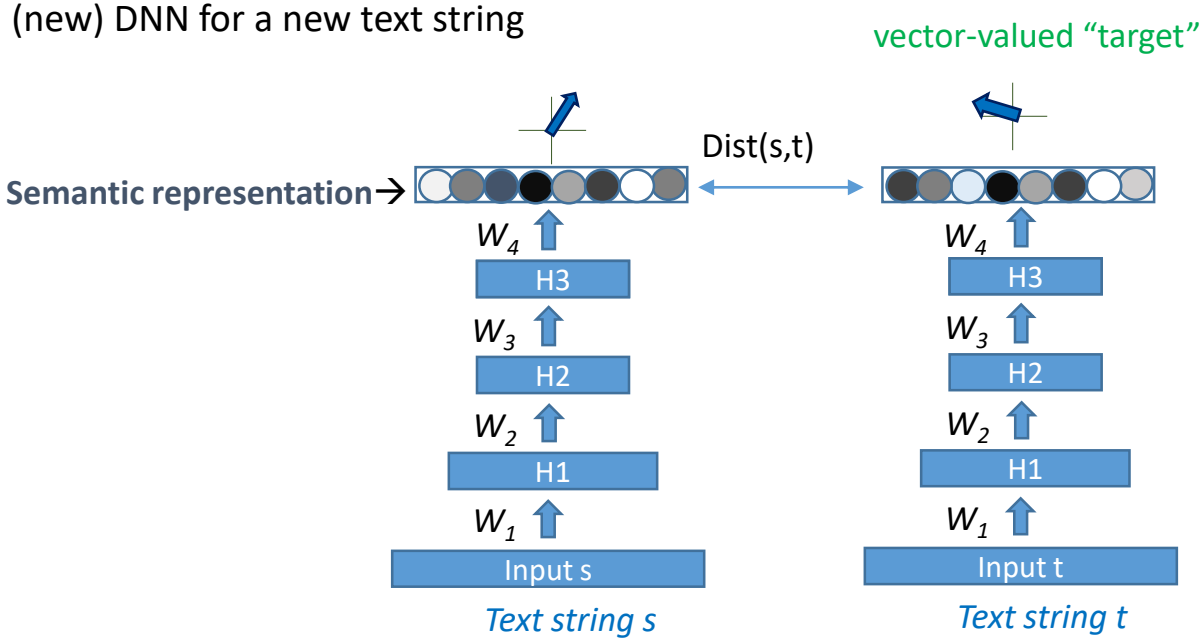
From DNN to DSSM

- DSSM for ranking; not classification
- **Step 1: "**target" from "one-hot"
  to continuous-valued vectors



From DNN to DSSM

- Step 1: "target" from "one-hot" to continuous-valued vectors
- **Step 2:** compute the "target" vector using a (new) DNN for a new text string



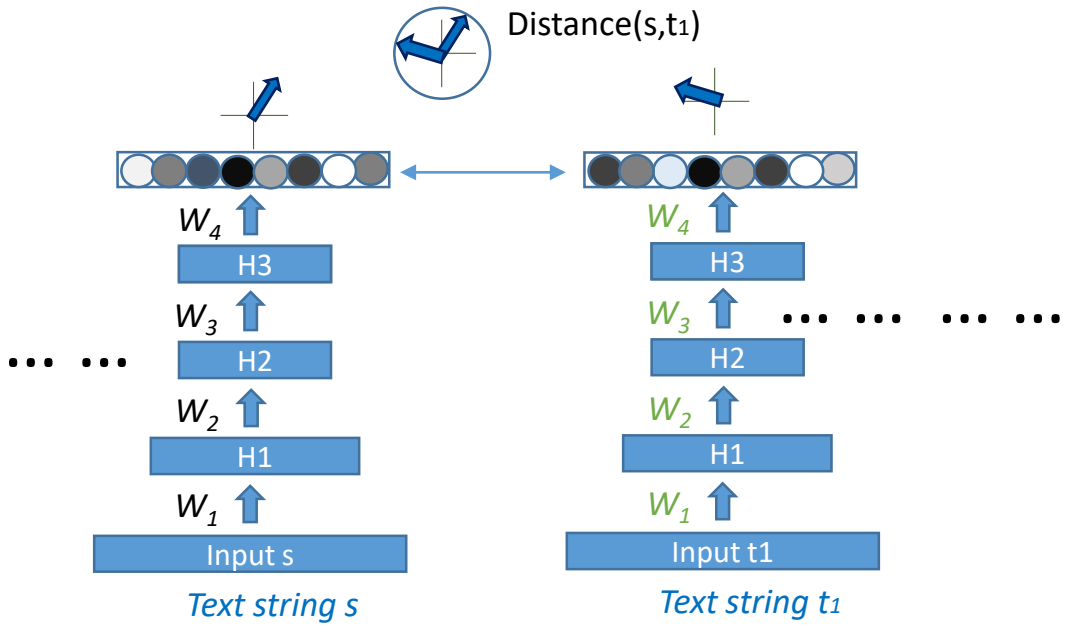
From DNN to DSSM

- Step 1: "target" from "one-hot" to continuous-valued vectors
- **Step 2:** compute the "target" vector using a (new) DNN for a new text string
- **Step 3:** normalize two "semantic" vectors & compute their similarity/distance

Normalized cross product → cosine distance;

Other more general types of "distances"

→ to model relationship



[Huang, He, Gao, Deng, Acero, Heck, "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data," in CIKM, Oct. 2013]

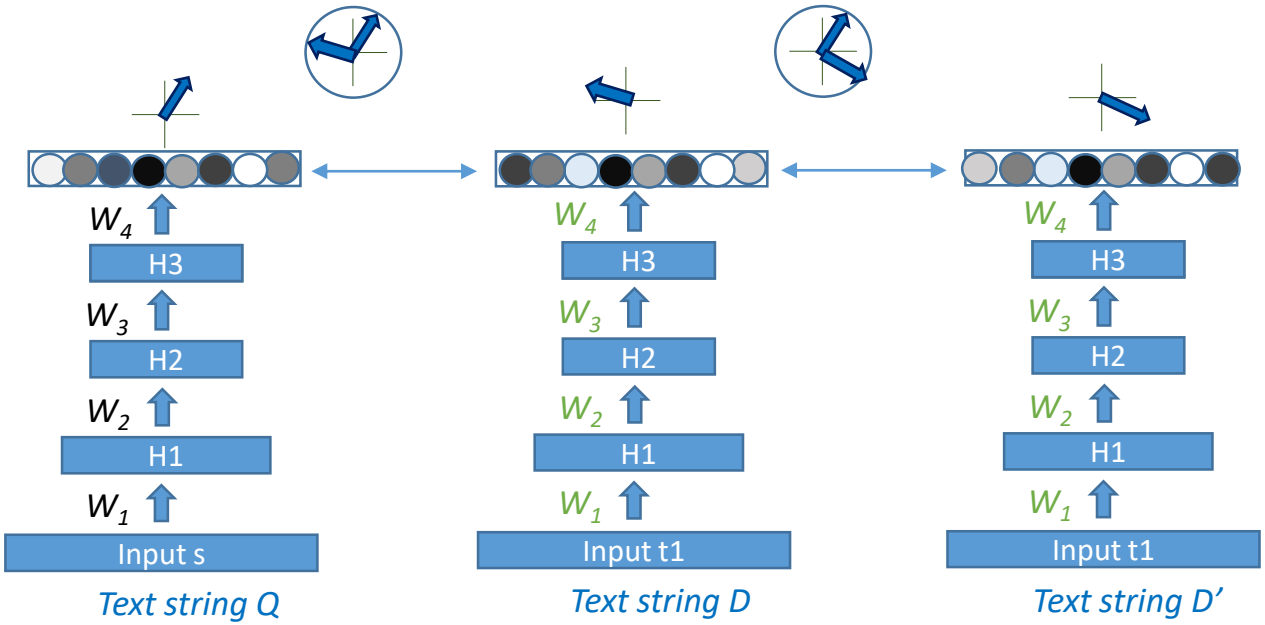
DSSM technical detail:

- Model construction
- Run-time
- Training

DSSM Model Construction

$$R(Q,D) = \cos(y_Q, y_D) = \frac{y_Q^T y_D}{\|y_Q\|\|y_D\|})$$

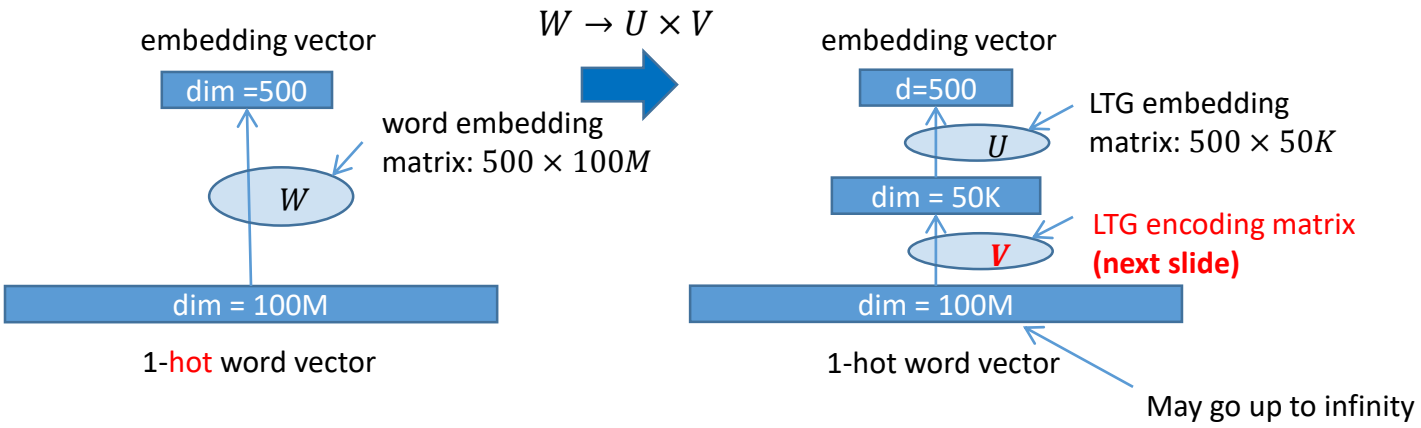
$$R(Q,D') = \cos(y_Q, y_{D'}) = \frac{y_Q^T y_{D'}}{\|y_Q\|\|y_{D'}\|})$$



… … … …

DSSM Model Construction: Additional technical detail

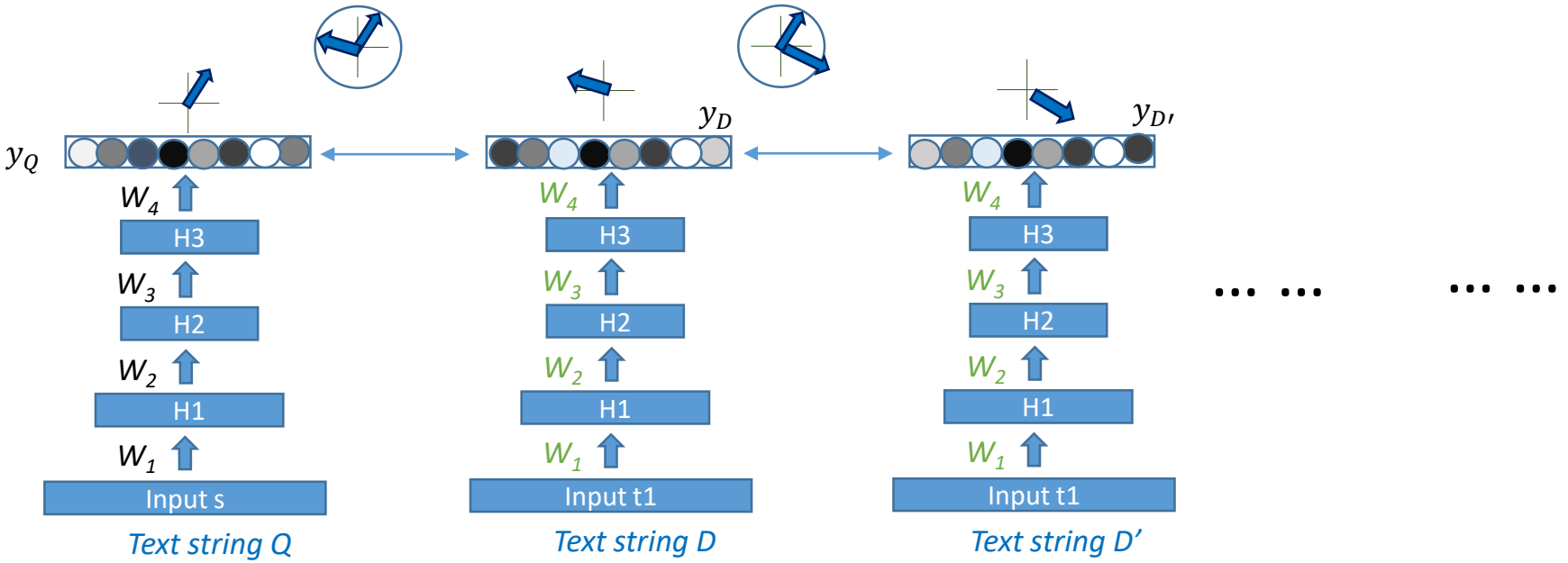
- Sub-word embedding: Learn embedding on sub-word units, such as letter-trigram (LTG)
  - E.g., cat → #cat# → #-c-a, c-a-t, a-t-#
- Solve the problem: almost unbounded variability (word) → bounded variability (sub-word)
  - E.g., there are only ~50K letter-trigrams (37^3)



[Huang, He, Gao, Deng, Acero, Heck, 2013]

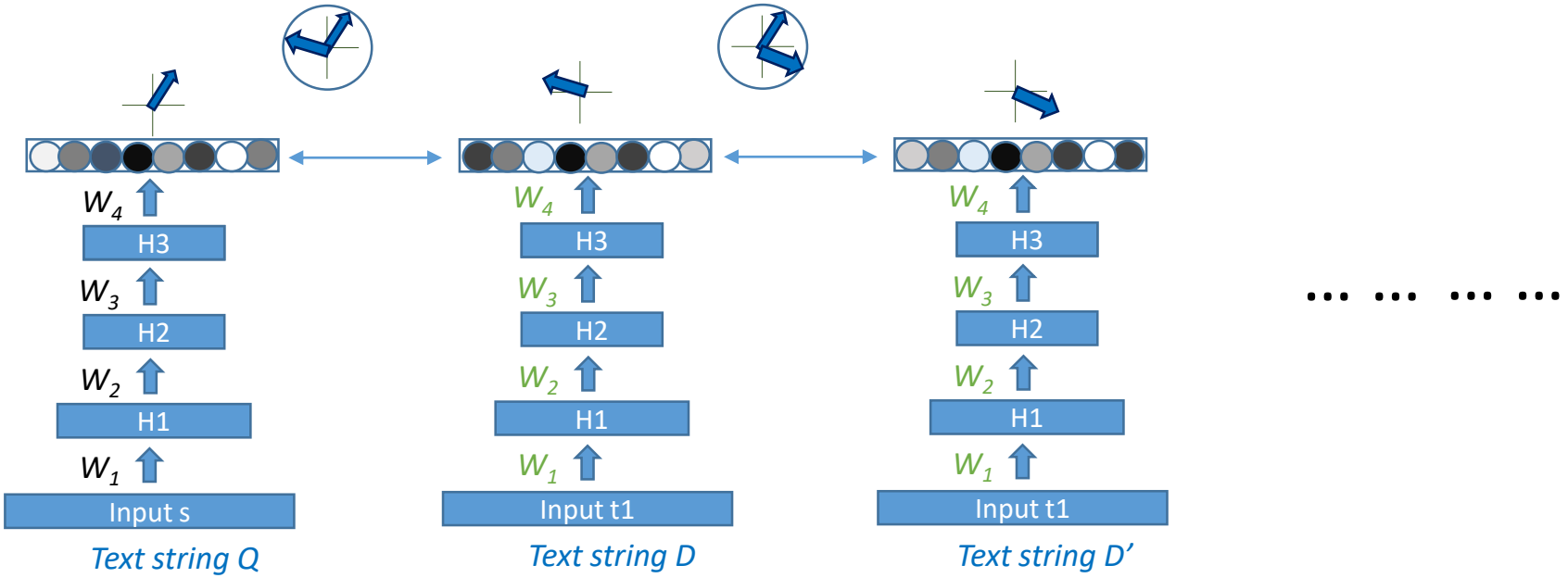
DSSM Run-Time

- Compute query semantic vector y_Q
- Compute semantic vectors for all documents from L1: y_D, $y_{D'}$, , …, …
- Compute similarities $\cos(y_Q, y_D), \cos(y_Q, y_{D'})$, …, …
- Use the similarities for ranking by DSSM alone, OR as a feature for boosting-tree-based full L2 ranker



DSSM Training: How to determine W_i and $W_{i,}$ i=1, 2, 3, 4…

- There is no “one hot” supervision signal available as in labeled speech/image data
- → DSSM cannot be learned using (negative) cross-entropy as the loss function (as in training DNN)
- Then, how to train DSSM? And where to find any “supervision” signal?



DSSM Training: Loss function

- Key insights:
  - "clicked" document (D^+) as "positive" signal
  - randomly sampled "non-clicked" documents (D^-) as "negative" signal
- Loss function for optimization:

$$L(\Lambda) = -\log \prod_{(Q,D^+)} \frac{\exp[\psi R_\Lambda(Q,D^+)]}{\sum_{D' \in \boldsymbol{D}} \exp[\psi R_\Lambda(Q,D')]}$$

$\boldsymbol{D}$: set of N documents in the training instance, including D^+ & (N-1) non-clicked documents D^-; $\boldsymbol{R_\Lambda(Q,D) = \cos(y_Q, y_D) = \frac{y_Q^T y_D}{\|y_Q\| \|y_D\|}}$; ψ: hyper-parameter; and Λ: all DSSM weight matrices to be learned.

DSSM Training: optimization procedure

- Stochastic gradient descent (mini-batch):

$$\Lambda_t = \Lambda_{t-1} - \epsilon_t \frac{\partial L(\Lambda)}{\partial \Lambda}|_{\Lambda=\Lambda_{t-1}}$$

- Λ_t and Λ_{t-1}: DSSM weight matrices at the t^{th} and $t-1^{th}$ iterations
- ϵ_t: learning rate (chosen similar to that for DNN training)

DSSM Training: Gradient computation (1)

- Compute $\frac{\partial L(\Lambda)}{\partial \Lambda}$ efficiently by backpropagation
- Loss function for individual training instance m, (Q_m, D_m^+), is

$$L_m(\Lambda) = -\log P(D_m^+|Q_m)$$

- Then $\frac{\partial L(\Lambda)}{\partial \Lambda} = \sum_{m=1}^{M} \frac{\partial L_m(\Lambda)}{\partial \Lambda}$, and let's compute each term (drop m)
- Rewrite $L(\Lambda) = \log(1 + \sum_j \exp(-\psi \Delta_j))$

  where $\Delta_j = R(Q, D^+) - R(Q, D_j^-)$
- For top DSSM layer N,

$$\frac{\partial L(\Lambda)}{\partial W_N} = \sum_j \alpha_j \frac{\partial \Delta_j}{\partial W_N}$$

where $\alpha_j = \frac{-\psi \exp(-\psi \Delta_j)}{1+\sum_{j'} \exp(-\psi \Delta_{j'})}$, and $\frac{\partial \Delta_j}{\partial W_N} = \frac{\partial R(Q,D^+)}{\partial W_N} - \frac{\partial R\left(Q,D_j^-\right)}{\partial W_N}$

DSSM Training: Gradient computation (2)

- Each derivative for top layer N is

$$\frac{\partial R(Q,D)}{\partial W_N} = \frac{\partial}{\partial W_N}\left[\frac{y_Q^T y_D}{\|y_Q\|\|y_D\|}\right] = \delta_{y_Q}^{(Q,D)} z_{N-1,Q}^T + \delta_{y_D}^{(Q,D)} z_{N-1,D}^T$$

where $z_{i,Q}$ & $z_{i,D}$: outputs of hidden layer i for query Q & document D (*i=N-1*), and

$$\delta_{y_Q}^{(Q,D)} = (1 - y_Q) \circ (1 + y_Q) \circ (bcy_D - acb^3 y_Q)$$
$$\delta_{y_D}^{(Q,D)} = (1 - y_D) \circ (1 + y_D) \circ (bcy_Q - abc^3 y_D)$$
$$a = y_Q^T y_D,\ b = 1/\|y_Q\|, \text{ and } c = 1/\|y_D\|$$

DSSM Training: Gradient computation (3)

- Error backpropagation in each branch of the DNN:

Branch *Q*: $\delta_{i,Q}^{(Q,D)} = \left(1 + z_{i,Q}\right) \circ \left(1 - z_{i,Q}\right) \circ W_i^T \delta_{i+1,Q}^{(Q,D)}$

Branch *D*: $\delta_{i,D}^{(Q,D)} = \left(1 + z_{i,D}\right) \circ \left(1 - z_{i,D}\right) \circ W_i^T \delta_{i+1,D}^{(Q,D)}$

- Then for all DSSM layers W_i, $i = 2, \ldots, N-1$:

$$\frac{\partial L(\Lambda)}{\partial W_i} = \sum_j \alpha_j \frac{\partial \Delta_j}{\partial W_i}$$

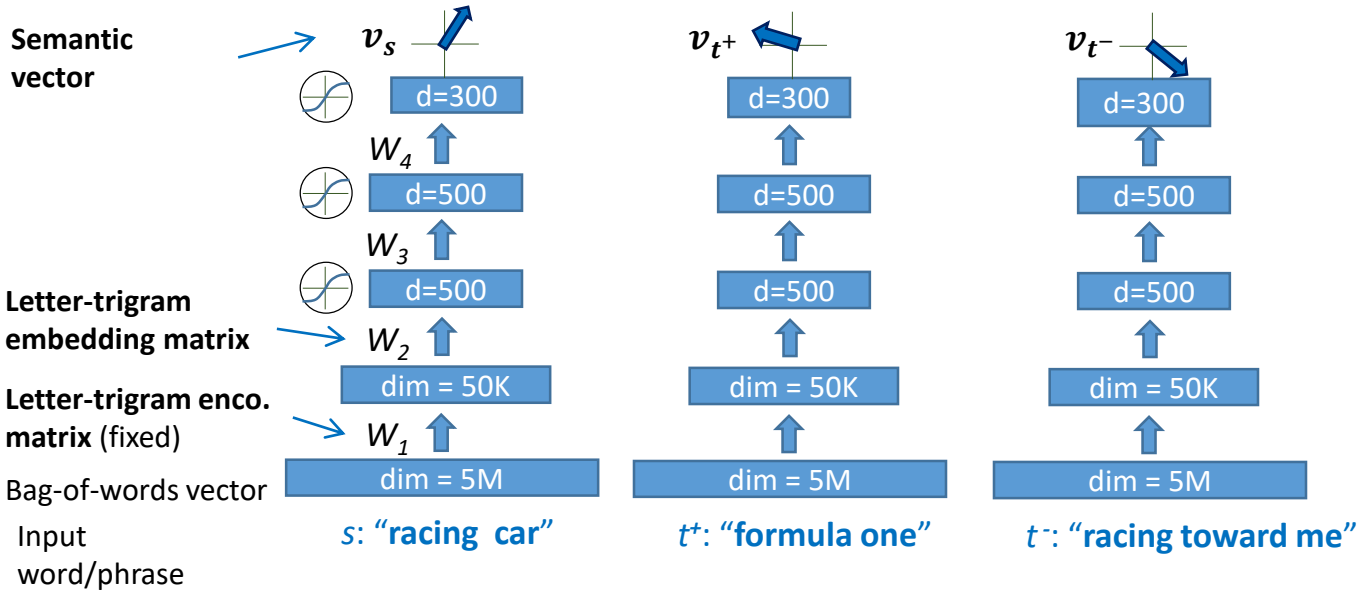
where $\frac{\partial \Delta_j}{\partial W_i} = \left(\delta_{i,Q}^{(Q,D^+)} z_{i-1,Q}^T + \delta_{i,D^+}^{(Q,D^+)} z_{i-1,D^+}^T\right) - \left(\delta_{i,Q}^{(Q,D_j^-)} z_{i-1,Q}^T + \delta_{i,D_j^-}^{(Q,D_j^-)} z_{i-1,D_j^-}^T\right)$

DSSM: Put it all together in animation
--- Training and run-time (before & after training)

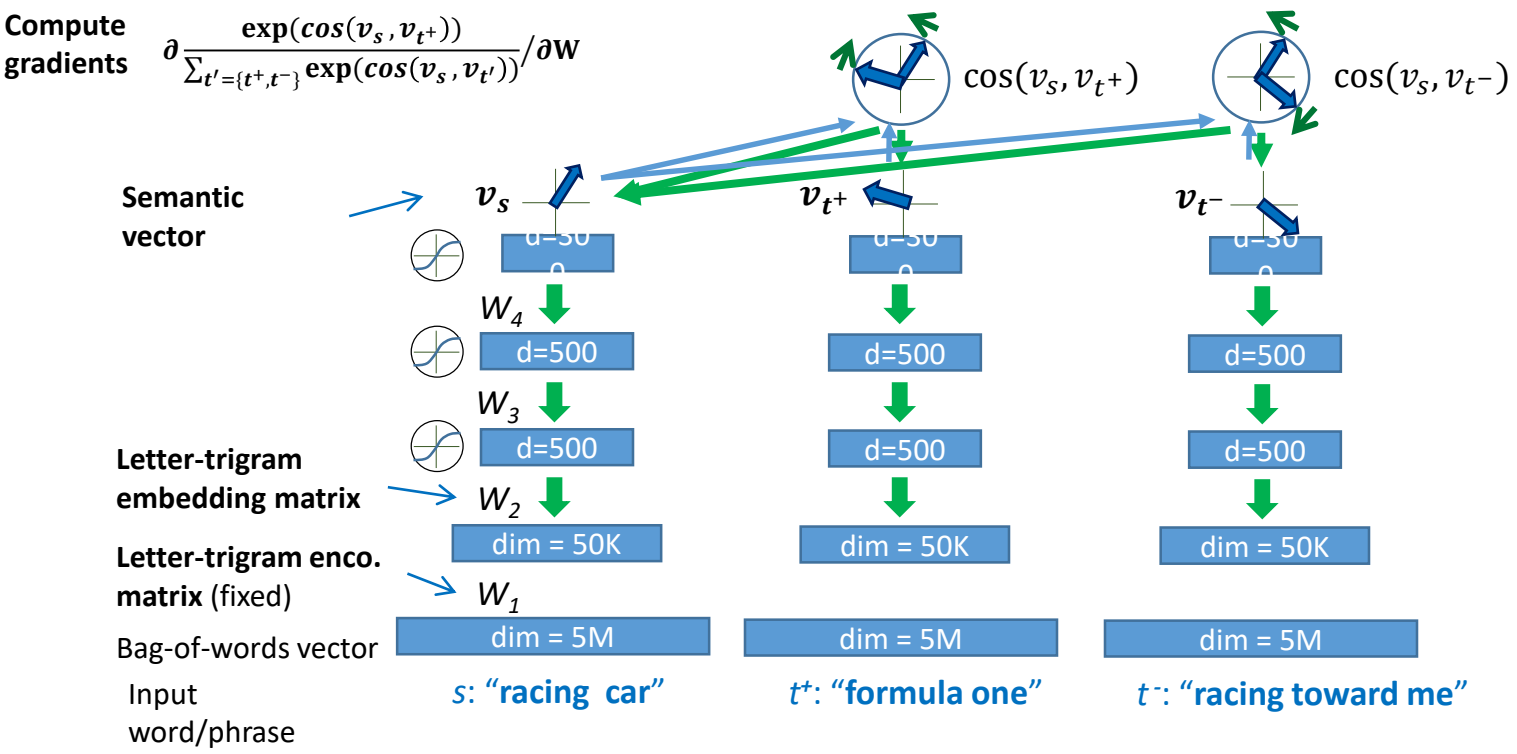
Training: The forward pass

Initialization:

Neural networks are initialized with random weights

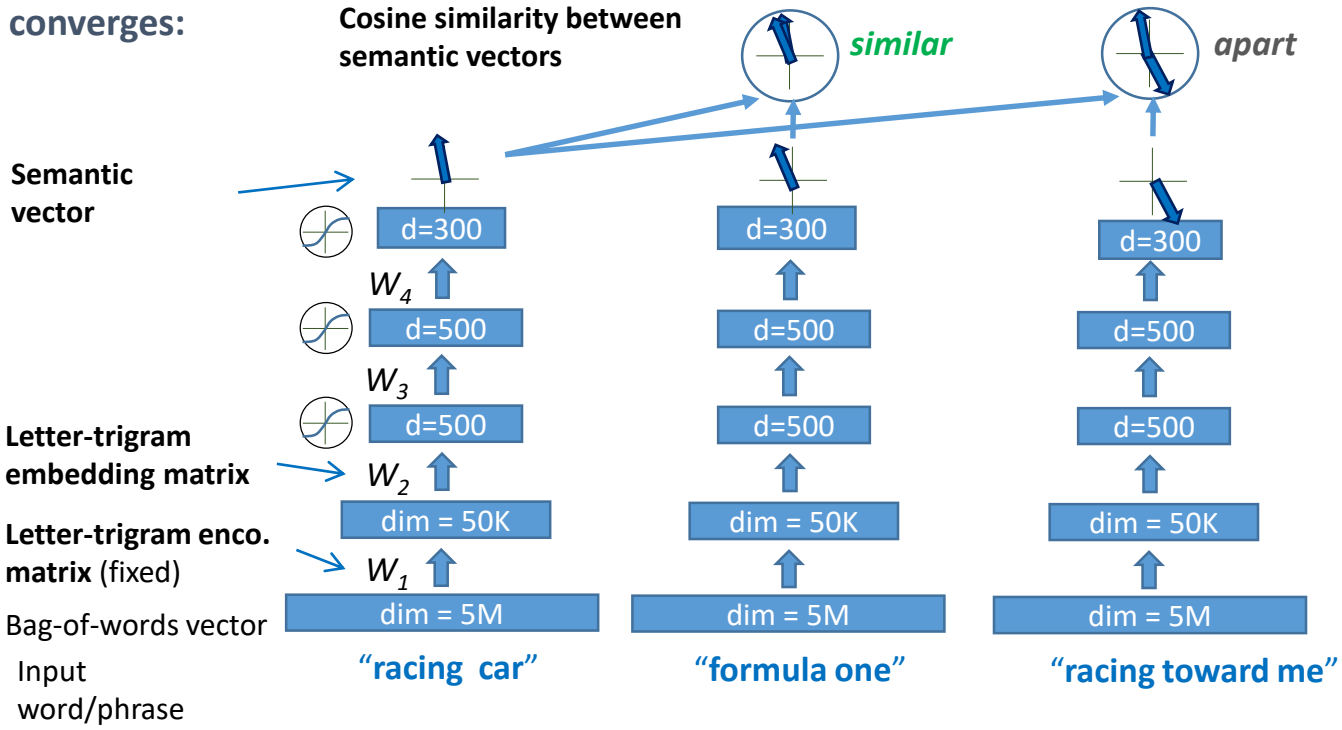


Training: The backward pass (i.e. error backpropagation)



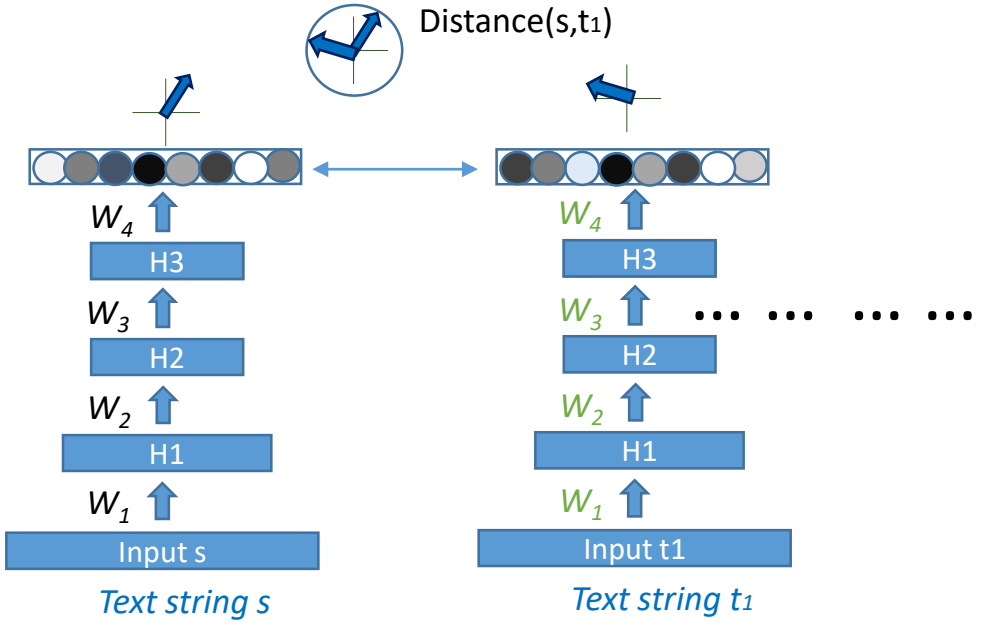
Run-time operation after training

When training converges:



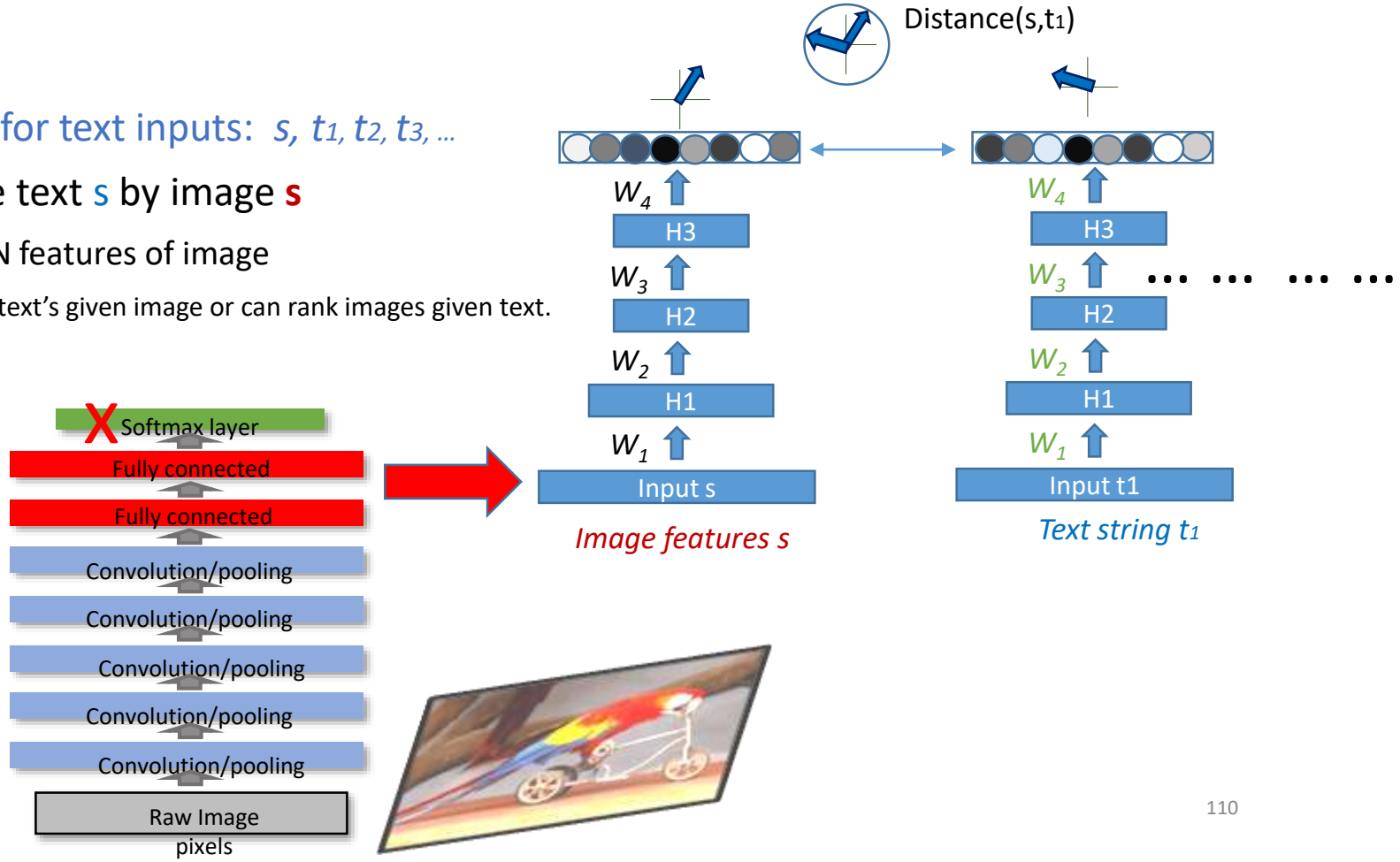
DSSM for **Multi-Modal Learning**

- Recall DSSM for text inputs: *s, t_1, t_2, t_3, …*



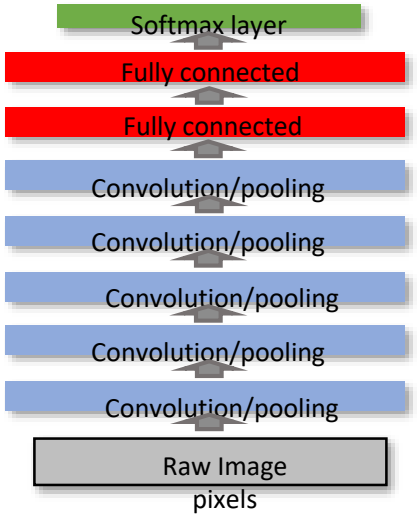
DSSM for **Multi-Modal Learning**

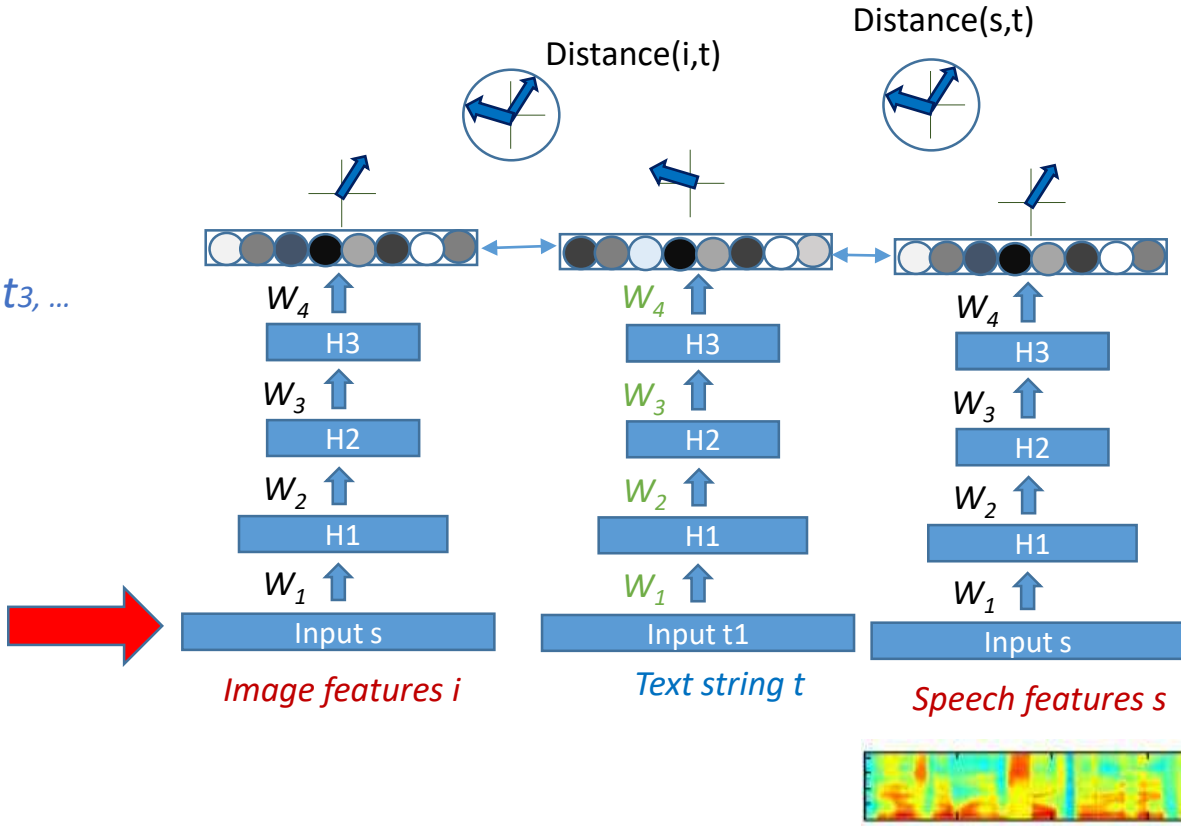
- Recall DSSM for text inputs: *s, t_1, t_2, t_3, …*
- Now: replace text s by image **s**
- Using DNN/CNN features of image
- Can rank/generate text's given image or can rank images given text.

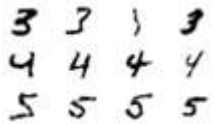


DSSM for **Multi-Modal Learning (text, image, speech)** --- a speech acquisition model through correlation?

- Recall DSSM for text inputs: *s, t_1, t_2, t_3, …*
- Now: replace text s by image **s**
- Using DNN/CNN features of image







Deep Visual Semantic Embedding Model

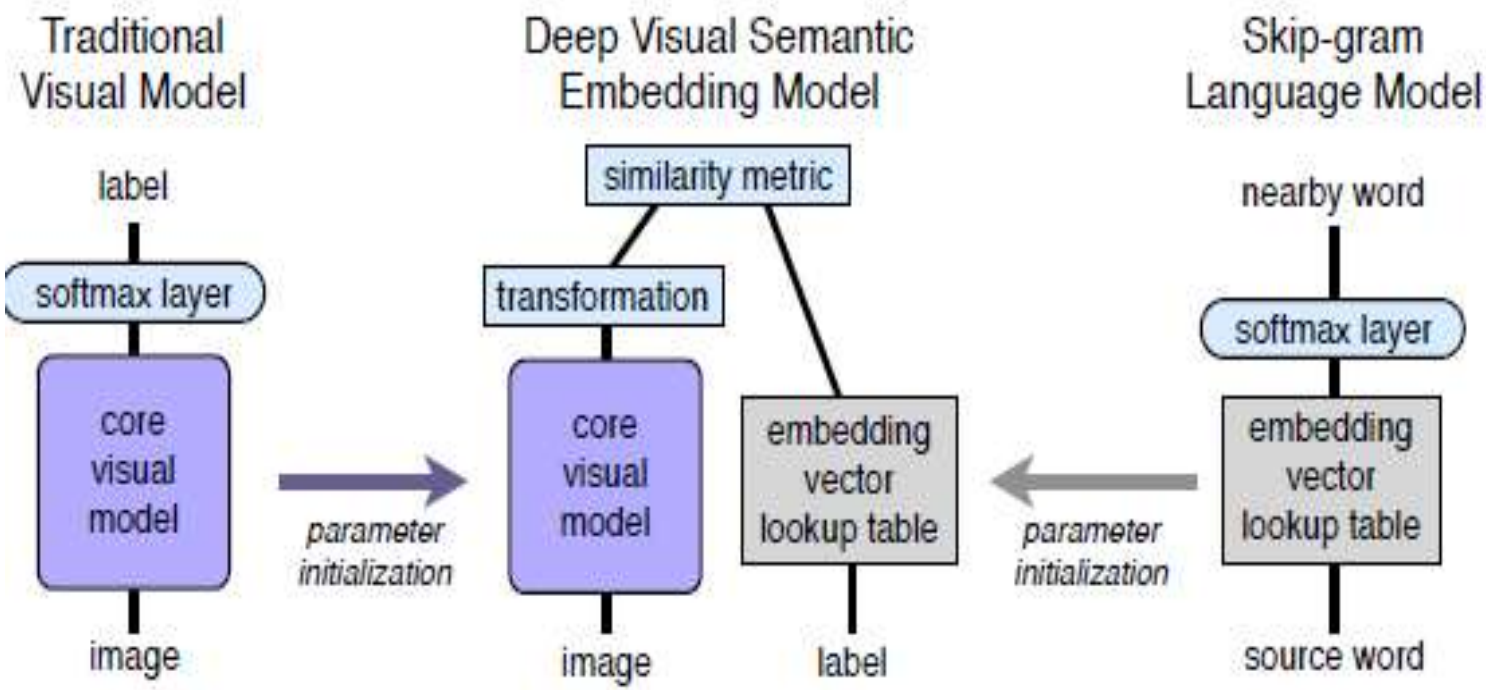
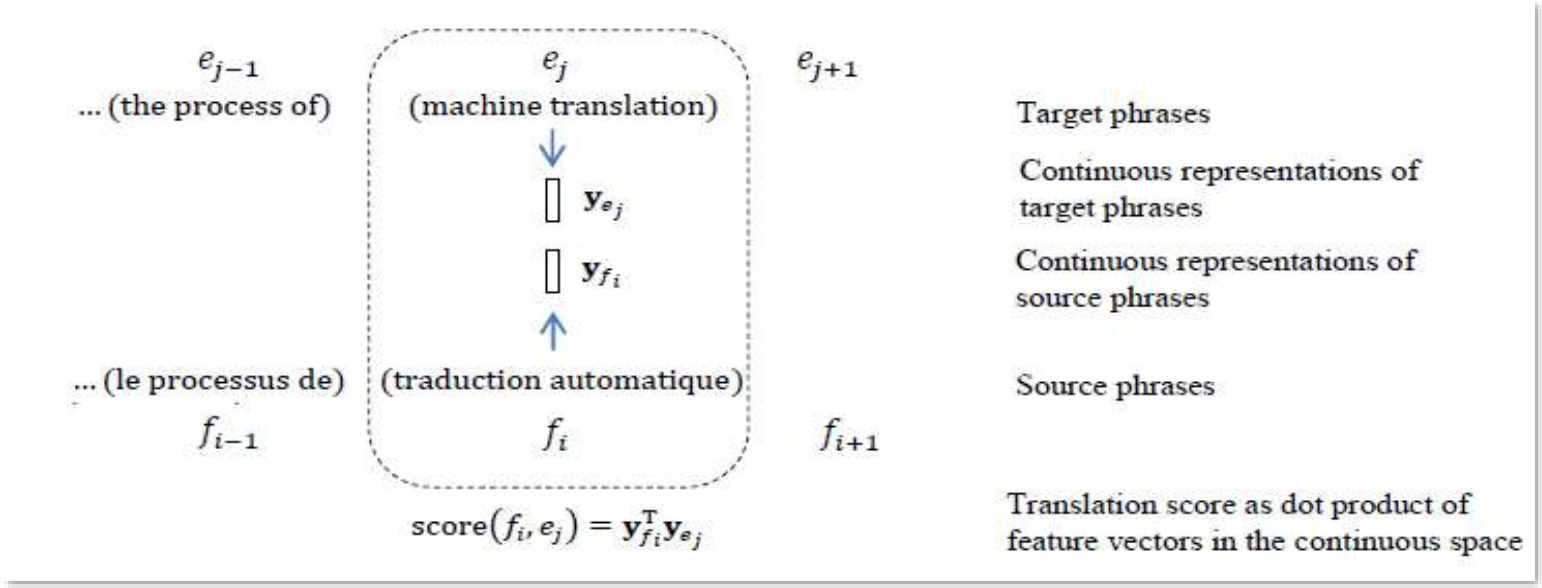


Illustration of the multi-modal DeViSE architecture. The left portion is an image recognition neural network with a softmax output layer. The right portion is a skip-gram text model providing word embedding vectors; The center is the joint deep image-text model of DeViSE, with the two Siamese branches initialized by the image and word embedding models below the softmax layers. The layer labeled "transformation" is responsible for mapping the outputs of the image (left) and text (right) branches into the same semantic space. [after (Frome, et al., 2013), @NIPS].

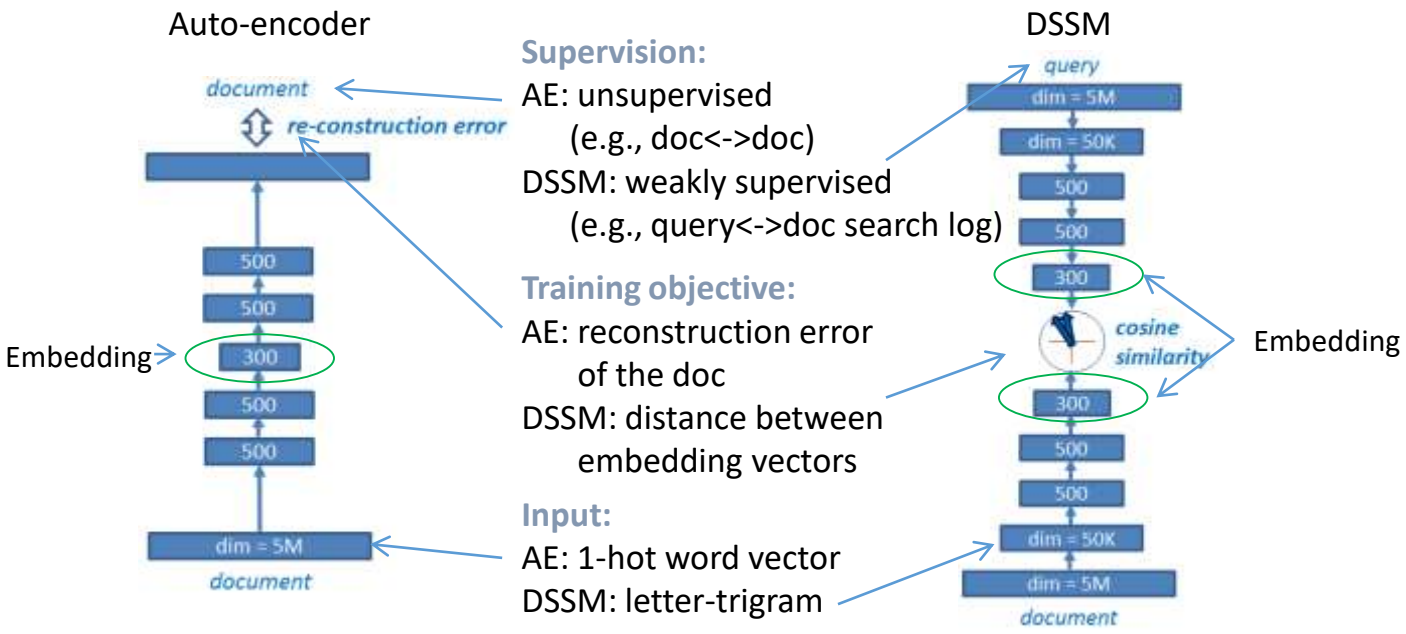
DSSM for **Machine Translation**



- Huge bi-lingual labeled supervision data available (like speech/image)
- Hence, DSSM is naturally fitted for MT
- Source phrase (like Q) and target phrase (like D) are mapped into the same semantic space
- Phrase translation score == similarity between their feature vectors in semantic space

[Gao, He, Yih, Deng, ACL, June 2014]

Deep Auto-Encoder vs. DSSM (& back to unsupervision: DERM)

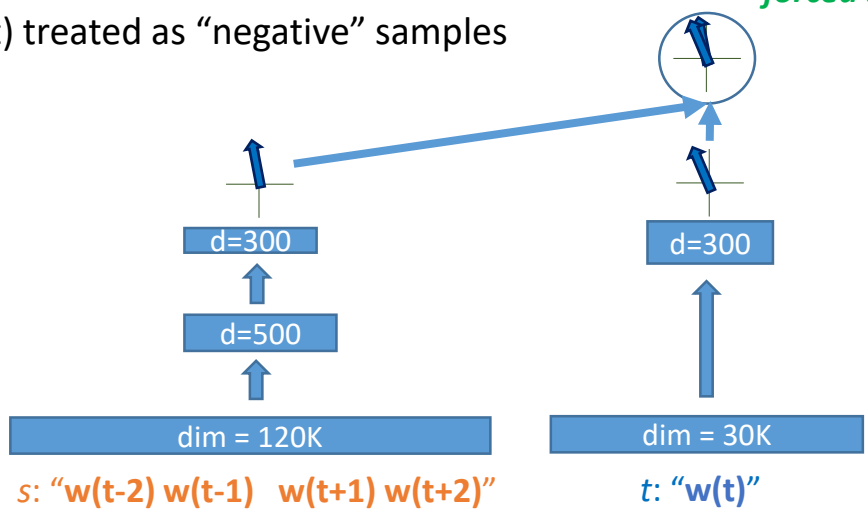


The DSSM can be trained using a variety of weak supervision signals without human labeling effort (e.g. user behavior log data)

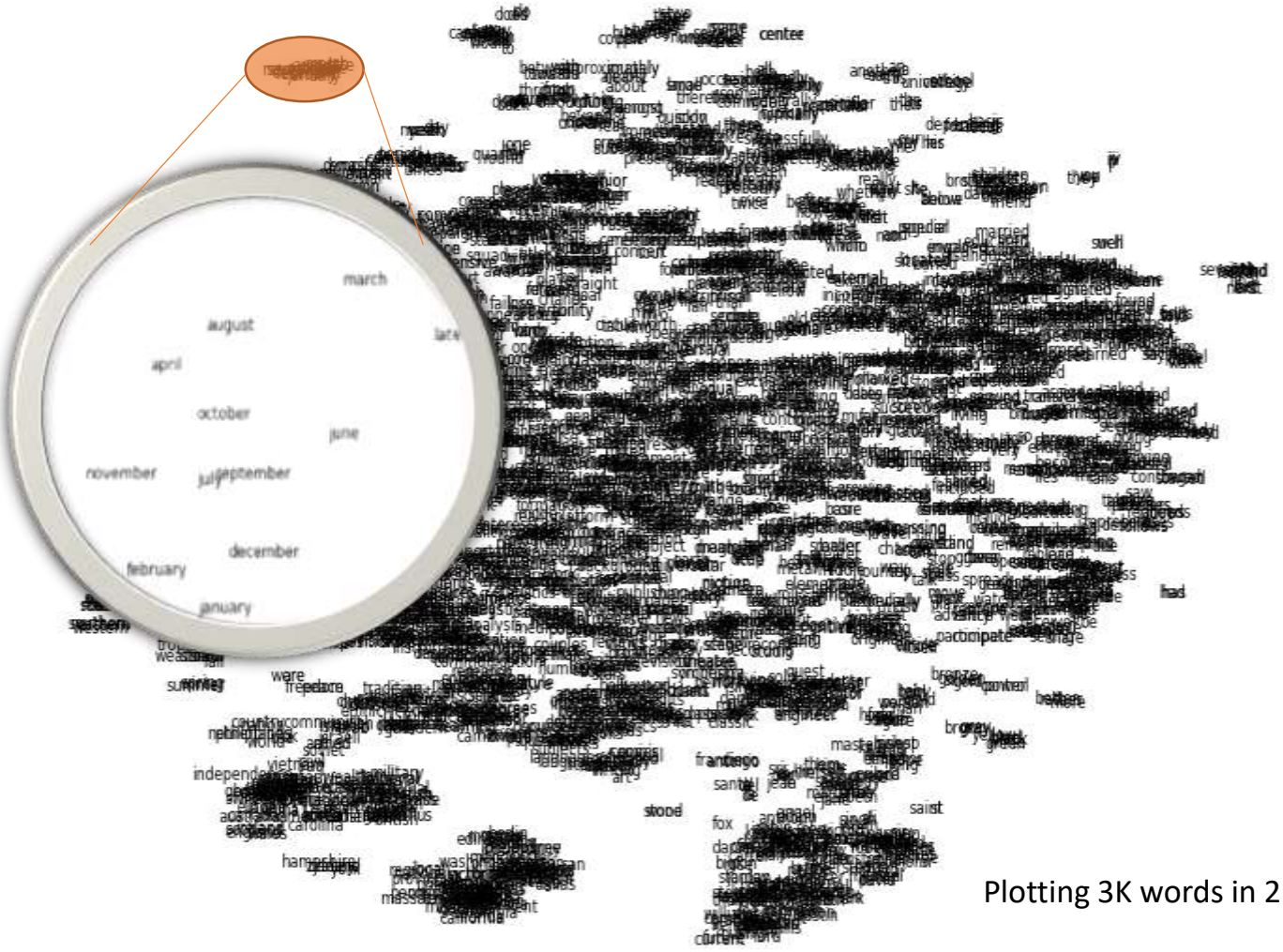
Probing into DSSM

- **Unsupervised** semantic word clustering and analogy
- Learn word embedding by means of its neighbors (context)
  - Construct context <-> word training pair for DSSM
  - Similar words with similar context -> higher cosine similarity
  - Randomly chosen words w(t) treated as "negative" samples

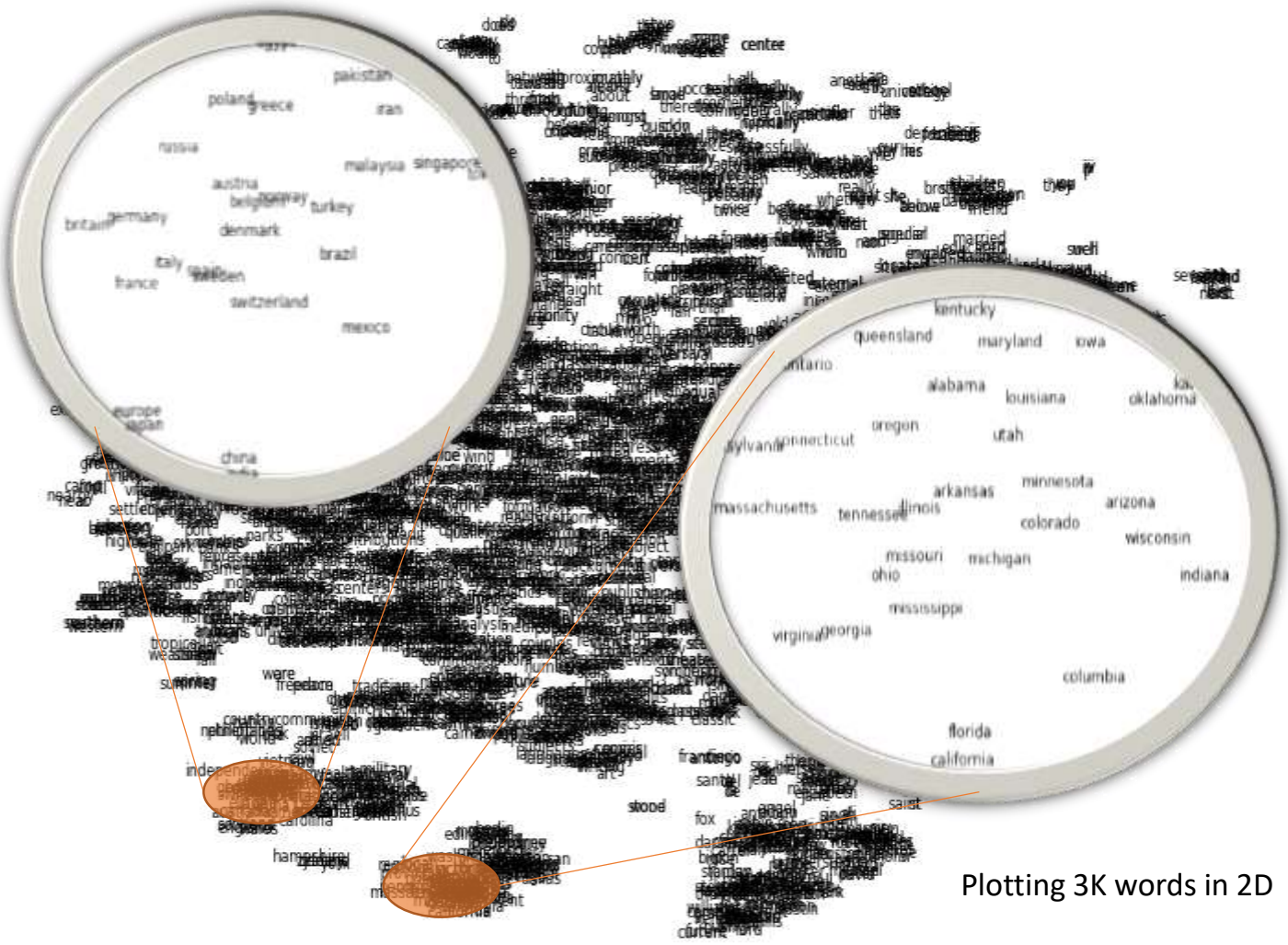
- Training Condition:
  - 30K vocabulary size
  - 10M words from Wikipedia
  - 50-dimentional vector



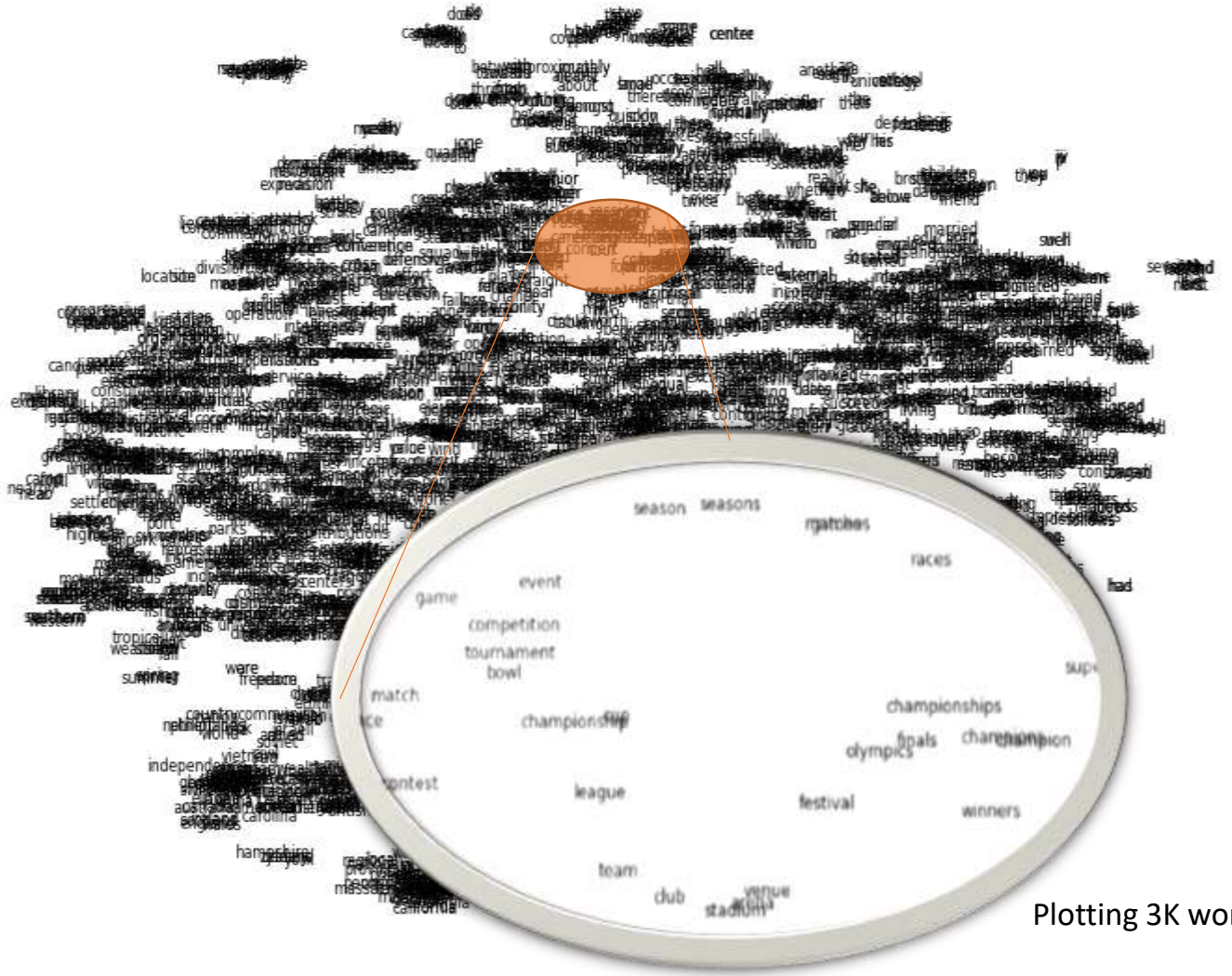
[Song et al. 2014]



Plotting 3K words in 2D



Plotting 3K words in 2D



Plotting 3K words in 2D

Cool results: DSSM for Semantic Word Clustering and Analogy

Semantic clustering examples: top 3 neighbors of each word

king	earl (0.77)	pope (0.77)	lord (0.74)
woman	person (0.79)	girl (0.77)	man (0.76)
france	spain (0.94)	italy (0.93)	belgium (0.88)
rome	constantinople (0.81)	paris (0.79)	moscow (0.77)
winter	summer (0.83)	autumn (0.79)	spring (0.74)
rain	rainfall (0.76)	storm (0.73)	wet (0.72)
car	truck (0.8)	driver (0.73)	motorcycle (0.72)

Semantic analogy examples (following the task in Mikolov et al., 2013)

$w_1 : w_2 = w_3 : ? \quad \Rightarrow \quad V_? = V_3 - V_1 + V_2$

summer : rain = winter : ?	snow (0.79)	rainfall (0.73)	wet (0.71)
italy : rome = france : ?	paris (0.78)	constantinople (0.74)	egypt (0.73)
man : eye = car : ?	motor (0.64)	brake (0.58)	overhead (0.58)
man : woman = king : ?	mary (0.70)	prince (0.70)	queen (0.68)
read : book = listen : ?	sequel (0.65)	tale (0.63)	song (0.60)

[Song et al. 2014]

Many possible applications of DSSM: Learning semantic similarity between *X* and *Y*

Tasks	*X*	*Y*
Web search	***Search query***	***Web documents***
Ad selection	***Search query***	***Ad keywords***
Entity ranking	***Mention (highlighted)***	***Entities***
Recommendation	***Doc in reading***	***Interesting things in doc or other docs***
Machine translation	***Sentence in language A***	***Translations in language B***
Nature User Interface	*Command (text/speech)*	*Action*
Summarization	*Document*	*Summary*
Query rewriting	*Query*	*Rewrite*
Image retrieval	*Text string*	*Images*
…	…	…

Limitations of the DSSM

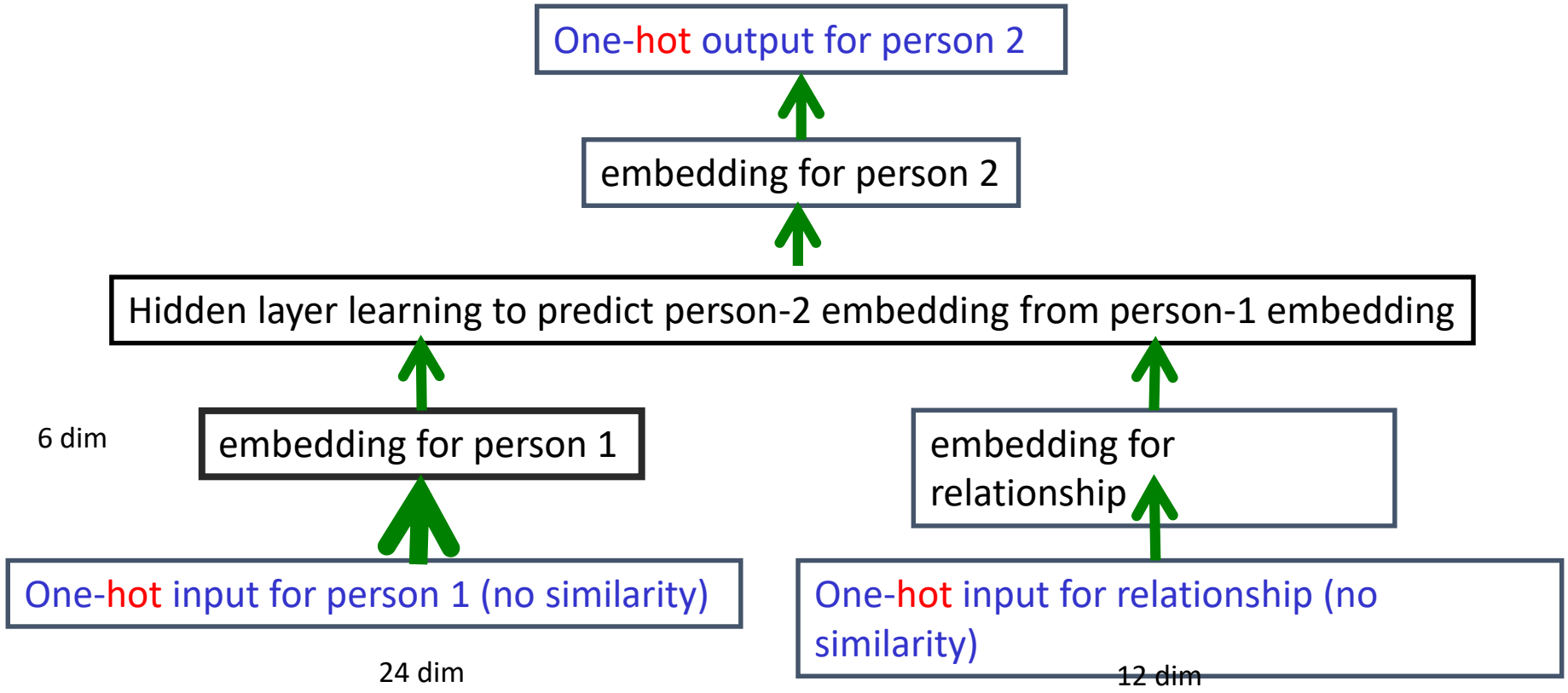
- Requirements of (weakly) supervised signals
- Lack of strong reasoning ability
- Limited reasoning ability of DNN/DSSM equipped with distributed representation comes with inherent problem of overgeneralization
- The classic family-tree example

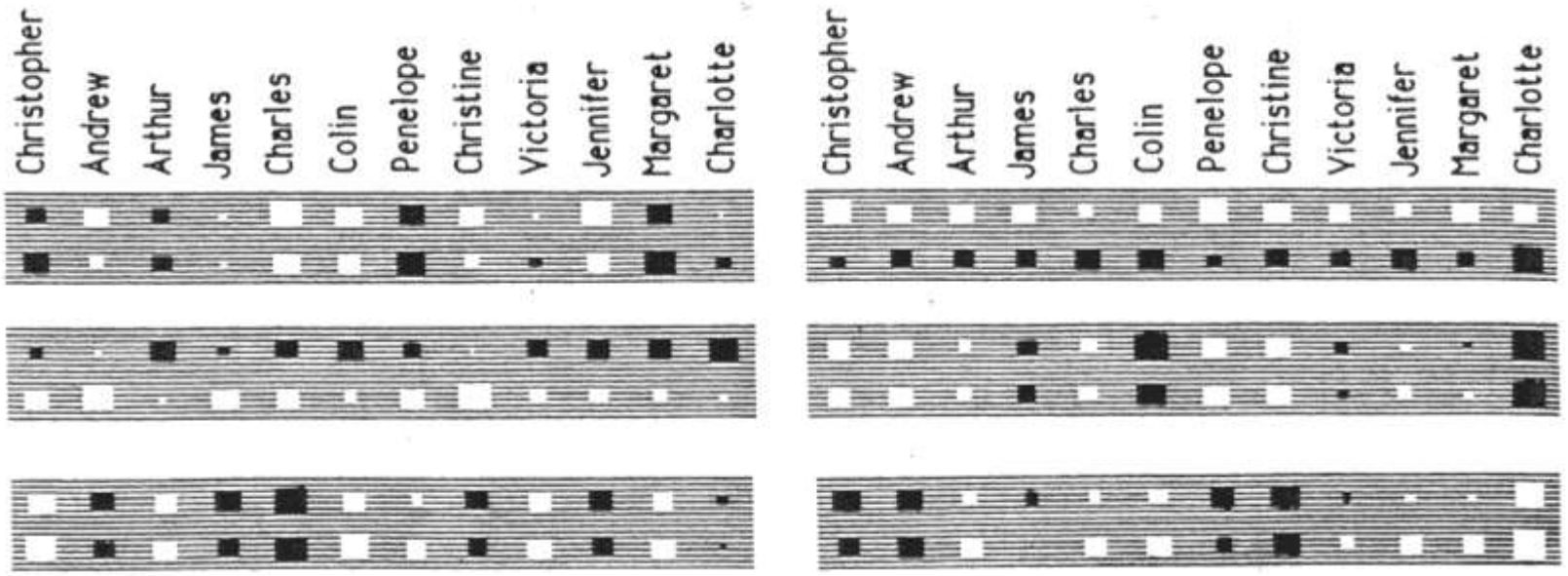
A relational learning task (Hinton, 1990, 2012)

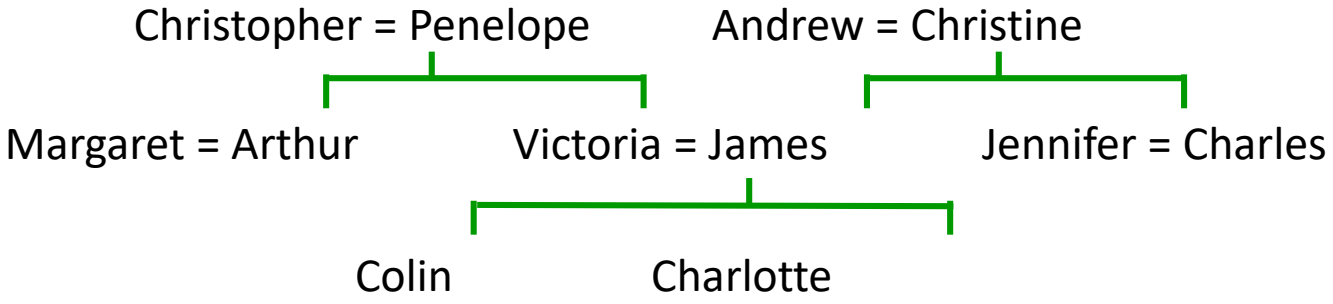
- Given a large set of triples that come from some family trees, figure out the regularities.
  - The obvious way to express the regularities is as symbolic rules

    (x has-mother y) & (y has-husband z) => (x has-father z)
- Finding the symbolic rules involves a difficult search through a very large discrete space of possibilities.
- Can a neural network capture the same knowledge by searching through a continuous space of weights?

DNN for reasoning over family-tree relations







Probing into the 6-by-24 weight matrix to find their meanings →micro-features of hidden units

Learning Micro-Features for Reasoning

- The six hidden units in person-1 embedding learn to represent features of people that are useful for predicting the answer.
  e.g., Nationality, generation, branch of the family tree
- But this DNN over-generates a lot: $2^6 = 64 > 24$.
- This causes 25% errors in predicting person-2
- Smaller codes (dim<6) does not handle noise well, causing more errors
- →Need stronger models than DNN/DSSM for reasoning
  - via *structured* distributed representation & end-to-end learning

Functional Modeling of the Brain/Mind

- For deep intelligence: reasoning, way beyond classification and similarity measure/ranking
- Reasoning at symbolic level (reasonably well understood by AI)
- But how does the brain use neural nets to do symbolic computation?
- Three levels of brain functions:
  - Neural-activity level (e.g., DNN)
  - Vector level (DSSM for entity/concept & multimodal information embedding)
  - Symbolic level (tensor product representation)
- From strongly supervised learning (speech/vision problems)

  to weakly supervised learning (language and multimodal problems)

  to unsupervised learning (reasoning and AI problems)

A Big Picture

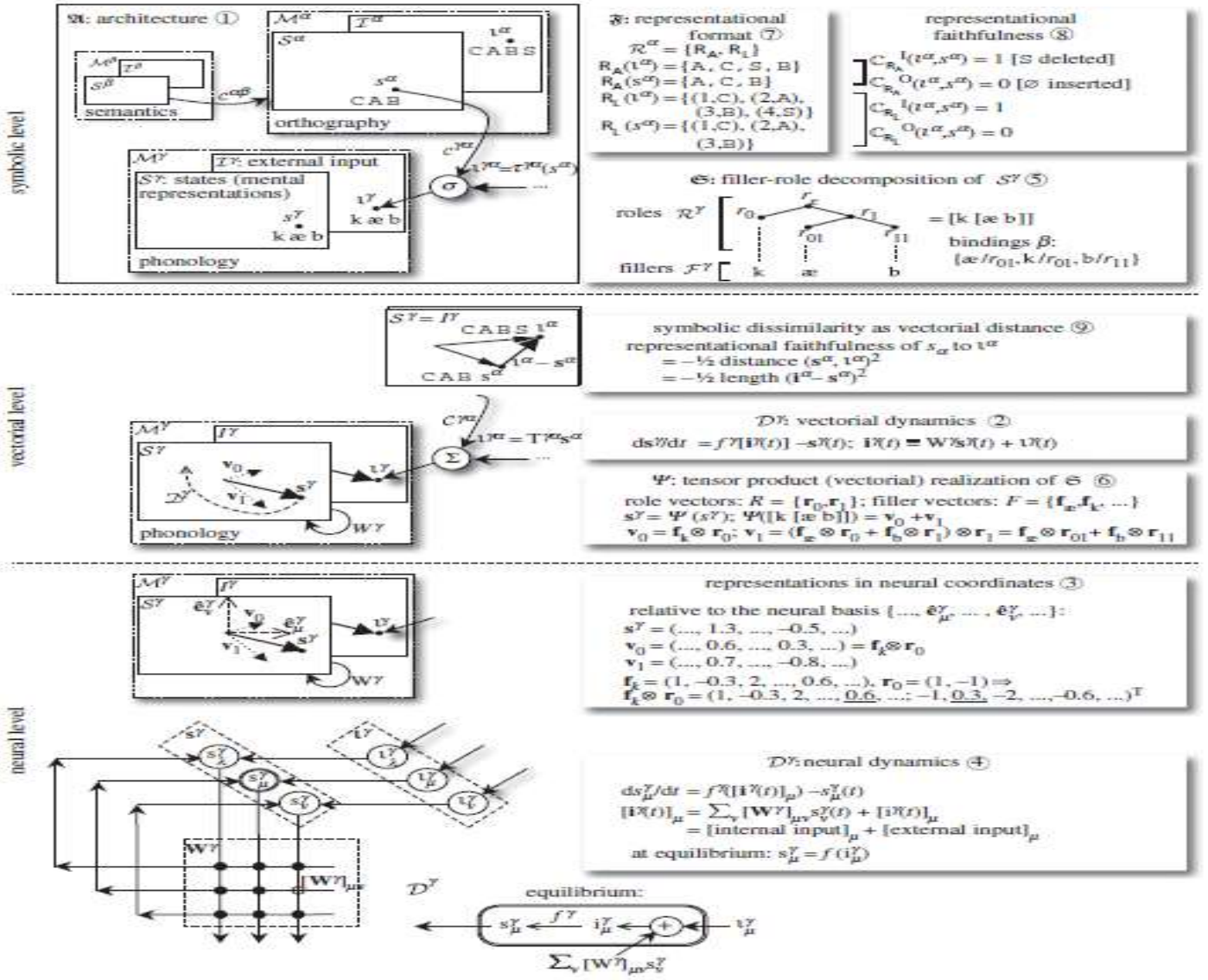


Figure 1. A schematic of the proposed theory for neural computation of symbolic cognitive functions.

Smolensky et al., 1986-2013

Thank You

& backup slides

for Q/A & discussions

DEEP LEARNING:
METHODS AND APPLICATIONS

Li Deng and Dong Yu
Microsoft Research
One Microsoft Way
Redmond, WA 98052

NOW PUBLISHERS, 2014

Data Science 101: Deep Learning Methods and Applications

March 7, 2014 by Daniel Gutierrez Leave a Comment

Microsoft Research, the research arm of the software giant, is a hotbed of data science and machine learning research. Microsoft has the resources to hire the best and brightest researchers from around the globe. A recent publication is available for download (PDF): "Deep Learning: Methods and Applications" by Li Deng and Dong Yu, two prominent researchers in the field.

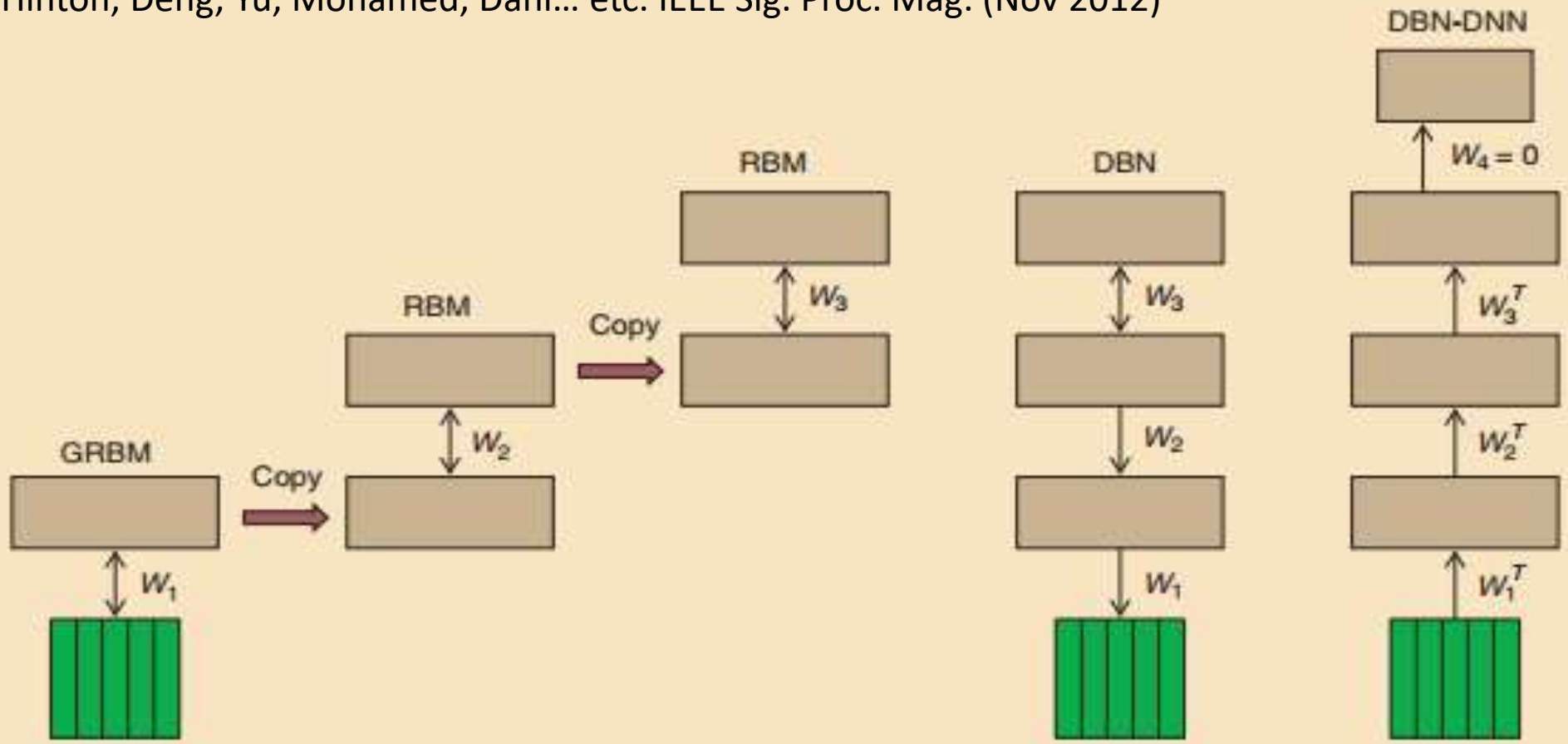
The 134 page book is aimed to provide an overview of general deep

Endorsements

In the past few years, deep learning has rapidly evolved into the de-facto approach for acoustic modeling in automatic speech recognition (ASR), showing tremendous improvement in accuracy, robustness, and cross-language generalizability over conventional approaches. This timely book is written by the pioneers of deep learning innovations and applications to ASR, who, in as early as 2010, first succeeded in large vocabulary speech recognition using deep learning. This was accomplished by a special form of the deep neural net, developed by the authors, perfectly fitting for fast decoding as required by industrial deployment of ASR technology. In addition to recounting this remarkable advance which ignited the industry-scale adoption of deep learning in ASR, this book also provides an overview of a sweeping range of up-to-date deep-learning methodology and its applications to a variety of signal and information processing tasks, including not only ASR but also computer vision, language modeling, text processing, multimodal learning, and information retrieval. This is the first and the most valuable book for *"deep and wide learning"* of deep learning, not to be missed by anyone who wants to know the breadth-taking impact of deep learning in many facets of information processing, especially in ASR, all of vital importance to our modern technological society.

Sadaoki Furui, President of Toyota Technological Institute at Chicago, and Professor of Tokyo Institute of Technology

Hinton, Deng, Yu, Mohamed, Dahl… etc. IEEE Sig. Proc. Mag. (Nov 2012)



First train a stack of three models each of which has one hidden layer. Each model in the stack treats the hidden variables of the previous model as data.

Then compose them into a single Deep Belief Network.

Then add outputs and train the DNN with backprop.

References:

Abdel-Hamid, O., Mohamed, A., Jiang, H., and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recogniti

ABDEL-HAMID, O., DENG, L., AND YU. D. "EXPLORING CONVOLUTIONAL NEURAL NETWORK STRUCTURES AND OPTIMIZATION FOR SPEECH RECOGNITION," INTERSPEECH, 2013.

ABDEL-HAMID, O., DENG, L., YU. D., Jiang, H. "Deep segmental neural networks for speech recognition," Proc. Interspeech, 2013A.

Acero, A., Deng, L., Kristjansson, T., and Zhang, J. "HMM adaptation using vector Taylor series for noisy speech recognition," PROC. INTERSPEECH, 2000.

Alain, G. and Bengio, Y. "What Regularized Autoencoders Learn from the Data Generating Distribution," Proc. International Conference on Learning Repre

Anthes, G. "Deep learning comes of age," Communications of the ACM, Vol. 56 No. 6, pp. 13-15, June 2013.

Arel, I., Rose, C., and Karnowski, T. "Deep Machine Learning - A New Frontier in Artificial Intelligence," IEEE Computational Intelligence Mag., vol. 5, pp. 13

Arisoy E., Sainath, T., Kingsbury, B., Ramabhadran, B. "Deep neural network language models," Proc. HTL-NAACL Workshop, 2012.

Aslan, O., Cheng, H., Schuurmans, D., and Zhang, X. "Convex two-layer modeling," Proc. NIPS, 2013.

Ba, J. and Frey, B. "Adaptive dropout for training deep neural networks," Proc. NIPS, 2013.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., and O'Shaughnessy, D. "Research developments and directions in speech recognition a

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., and O'Shaughnessy, D. "Updated MINS report on speech recognition and understandin

Baldi, P. and Sadowski, P. "Understanding Dropout," Proc. NIPS, 2013.

Battenberg, E., Schmidt, E., and Bello, J. Deep learning for music, special session at ICASSP (http://www.icassp2014.org/special_sections.html#SS8), 2014.

Batternberg, E. and Wessel, D. "Analyzing drum patterns using conditional deep belief networks," Proc. ISMIR, 2012.

Bell, P., Swietojanski, P., and Renals, S. "Multi-level adaptive networks in tandem and hybrid ASR systems", Proc. ICASSP, 2013.

Bengio, Y., Yao, L., Alain, G., and Vincent, P. "Generalized denoising autoencoders as generative models," Proc. NIPS, 2013.

Bengio, Y. "Deep learning of representations: Looking forward," in: Statistical Language and Speech Processing, pp. 1--37, Springer, 2013.

Bengio, Y., Boulanger, N., and Pascanu, R. "Advances in optimizing recurrent networks," Proc. ICASSP, 2013.

Bengio, Y., Courville, A., and Vincent, P. "Representation learning: A review and new perspectives," IEEE Trans. PAMI, vol. 38, pp. 1798-1828, 2013a.

Bengio, Y., Thibodeau-Laufer, E., and Yosinski, J. "Deep generative stochastic networks trainable by backprop," arXiv 1306:1091, 2013b; also accepted to a

Bengio, Y. "Deep Learning of Representations for Unsupervised and Transfer Learning," JMLR Workshop and Conference Proceedings, vol. 27, pp. 17-37, 2

Bengio, Y. "Learning deep architectures for AI," in Foundations and Trends in Machine Learning, Vol. 2, No. 1, 2009, pp. 1-127.

Bengio, Y. "Neural net language models," Scholarpedia, Vol. 3, 2008.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. "Greedy Layer-Wise Training of Deep Networks," Proc. NIPS, 2006.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. "A Neural Probabilistic Language Model," Journal of Machine Learning Research, vol. 3, pp. 1137-115

Bengio, Y. "New Distributed Probabilistic Language Models," Technical Report, University of Montreal, 2002.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. "A neural probabilistic language model," Proc. NIPS, 2000.

Bengio, Y., Simard, P., and Frasconi, P. "Learning Long-Term Dependencies with Gradient Descent is Difficult," IEEE Transactions on Neural Networks, Vol.

Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. "Global Optimization of a Neural Network-Hidden Markov Model Hybrid," IEEE Transactions on Neur

Bengio, Y. Artificial Neural Networks and Their Application to Sequence Recognition, Ph.D. Thesis, McGill University, Montreal, Canada, 1991.

Bergstra, J. and Bengio, Y. "Random search for hyper-parameter optimization," J. Machine Learning Research," Vol. 3, pp. 281-305, 2012.

Biem, A., Katagiri, S., McDermott, E., and Juang, B. "An application of discriminative feature extraction to filter-bank-based speech recognition," IEEE Tran

Bilmes, J. "Dynamic graphical models," IEEE Signal Processing Mag., vol. 33, pp. 29–42, 2010.

Bilmes, J. and Bartels, C. "Graphical model architectures for speech recognition," IEEE Signal Processing Mag., vol. 22, pp. 89–100, 2005.

Bordes, A., Weston, J., Collobert, R., and Bengio, Y. "Learning Structured Embeddings of Knowledge Bases," Proc. AAAI, 2011.