Deep Learning for AI

from Machine Perception to Machine Cognition

Li Deng

Chief Scientist of AI,

Microsoft Applications/Services Group (ASG) &

MSR Deep Learning Technology Center (DLTC)

A Plenary Presentation at IEEE-ICASSP, March 24, 2016

Thanks go to many colleagues at DLTC & MSR, collaborating universities,
and at Microsoft's engineering groups (ASG+)

WIKIPEDIA
The Free Encyclopedia

Deep learning

From Wikipedia, the free encyclopedia

Definition

Deep learning is a class of machine learning algorithms that[1](pp199–200)

- use a cascade of **many layers of nonlinear processing** .
- are part of the broader machine learning field of learning representations of data facilitating **end-to-end optimization**.
- learn multiple levels of representations that correspond to **hierarchies of concept abstraction**
- …, …

WIKIPEDIA
The Free Encyclopedia

Artificial intelligence

From Wikipedia, the free encyclopedia

Artificial intelligence (**AI**) is the intelligence exhibited by machines or software. It is also the name of the academic field of study on how to create computers and computer software that are capable of intelligent behavior.

Artificial general intelligence

From Wikipedia, the free encyclopedia

Artificial general intelligence (**AGI**) is the intelligence of a (hypothetical) machine that could successfully perform any intellectual task that a human being can. It is a primary goal of artificial intelligence research and an important topic for science fiction writers and futurists. Artificial general intelligence is also referred to as "**strong AI**"…

AI/(A)GI & Deep Learning: the main thesis

AI/GI = **machine perception** (speech, image, video, gesture, touch...)

+ machine cognition (natural language, *reasoning, attention, memory*/learning, knowledge, decision making, action, interaction/conversation, …)

GI: AI that is flexible, general, adaptive, learning from 1st principles

Deep Learning + Reinforcement/Unsupervised Learning →AI/GI

AI/GI & Deep Learning: how AlphaGo fits

AI/GI = machine perception (speech, image, video, gesture, touch...)

+ machine cognition (natural language, *reasoning, attention, memory*/learning, knowledge, **decision making**, action, interaction

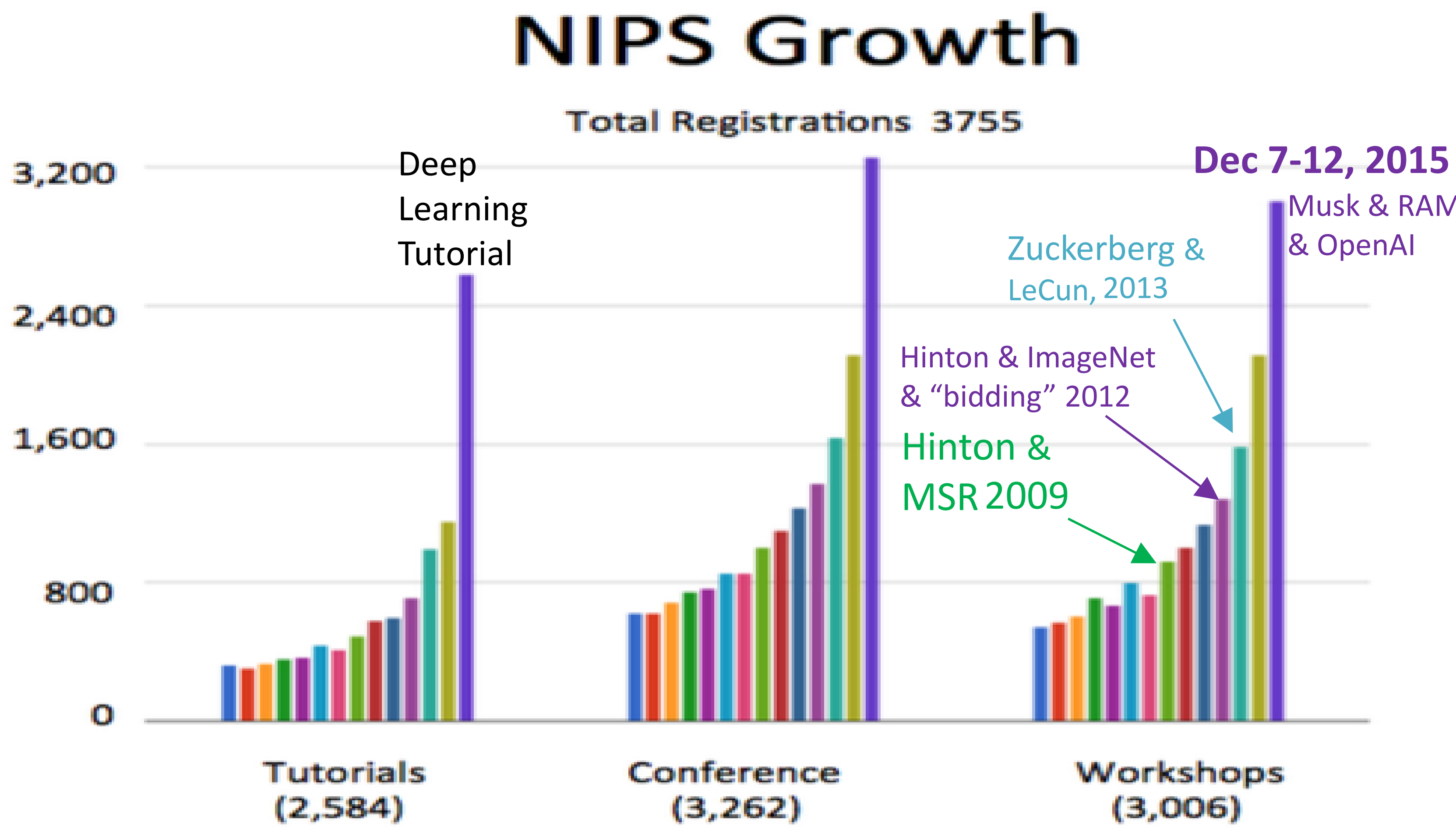
AGI: AI tha… tive, learning from 1^{st} principles

Deep Lear… supervised Learning →AI/AGI

Outline

- Deep learning for machine **perception**
  - **Speech**
  - Image
- Deep learning for machine **cognition**
  - Semantic modeling
  - Natural language
  - Multimodality
  - Reasoning, attention, memory (RAM)
  - Knowledge representation/management/exploitation
  - Optimal decision making (by deep reinforcement learning)
- Three hot areas/challenges of deep learning & AI research

Deep learning Research: centered at NIPS (Neural Information Processing Systems)



Neural Information Processing Systems Foundation

NIPS Home
Overview
Conference Videos
Workshop Videos
Program Highlights
Tutorials
Conference Sessions
Workshops
Publication Models
Demonstrations
Mini Symposia
Accepted Papers
Dates
Committees

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 b
dominated by a "shallow" architecture --- hidden Markov models (HMMs). Sig
technological success has been achieved using complex and carefully engine
of HMMs. The next generation of the technology requires solutions to remaini
challenges under diversified deployment environments. These challenges, not
addressed in the past, arise from the many types of variability present in the
generation process. Overcoming these challenges is likely to require "deep" a
with efficient learning algorithms. For speech recognition and related sequent
recognition applications, some attempts have been made in the past to devel
computational architectures that are "deeper" than conventional HMMs, such

The Universal Translator *..comes true!*

Deep learning technology enabled speech-to-speech translation

The New York Times

Scientists See Promise in Deep-Learning Programs

John Markoff
November 23, 2012

Tianjin, China, October, 25, 2012

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

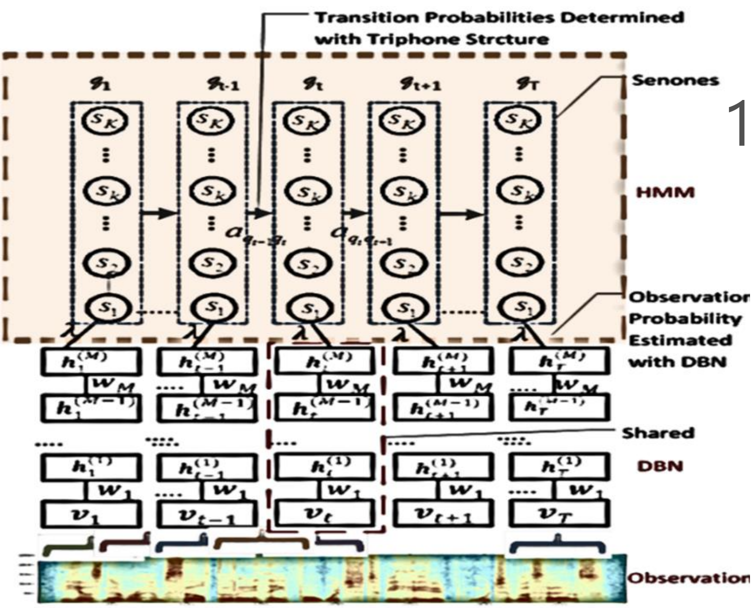
Microsoft Research

Deep belief networks for phone recognition, NIPS, December 2009; 2012

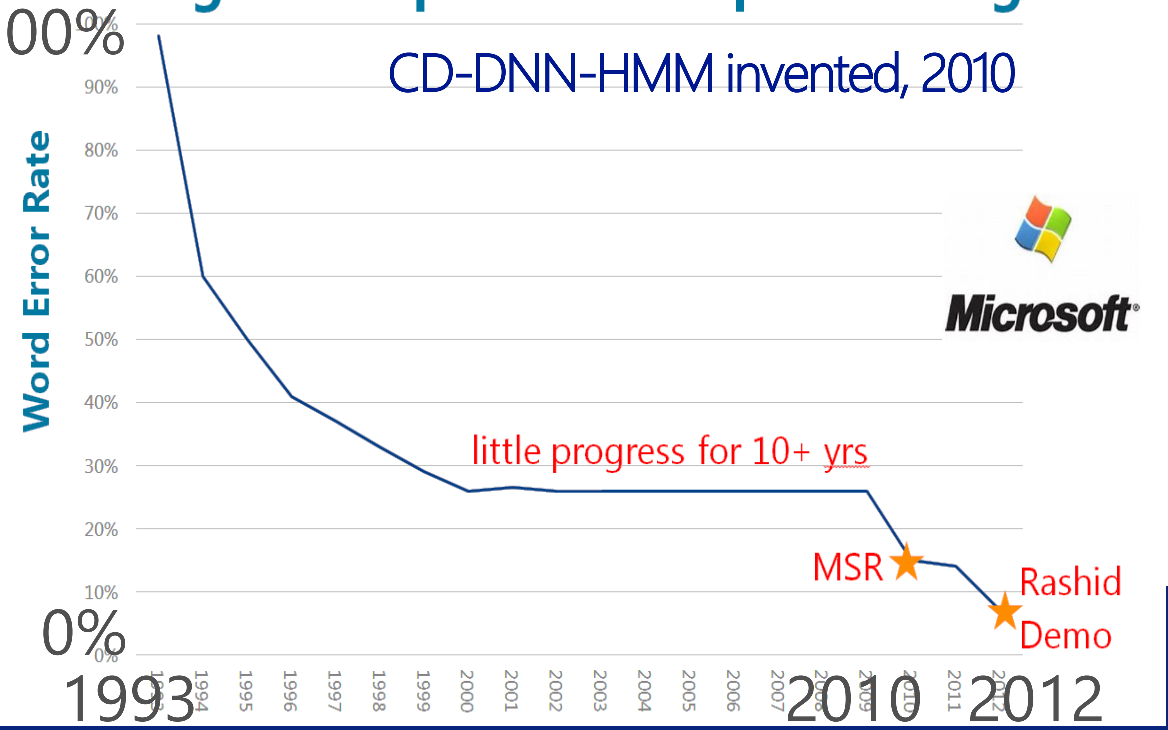
Investigation of full-sequence training of DBNs for speech recognition., Interspeech, Sept 2010
Binary coding of speech spectrograms using a deep auto-encoder, Interspeech, Sept 2010
Roles of Pre-Training & Fine-Tuning in CD-DBN-HMMs for Real-World ASR, NIPS, Dec. 2010
Large Vocabulary Continuous Speech Recognition With CD-DNN-HMMS, ICASSP, April 2011
Conversational Speech Transcription Using Contxt-Dependent DNN, Interspeech, *Aug. 2011*

Making deep belief networks effective for LVCSR, ASRU, Dec. 2011
Application of Pretrained DNNs to Large Vocabulary Speech Recognition., ICASSP, 2012
【胡郁】讯飞超脑 2.0 是怎样炼成的？2011, 2015

Later years with rapid progress,



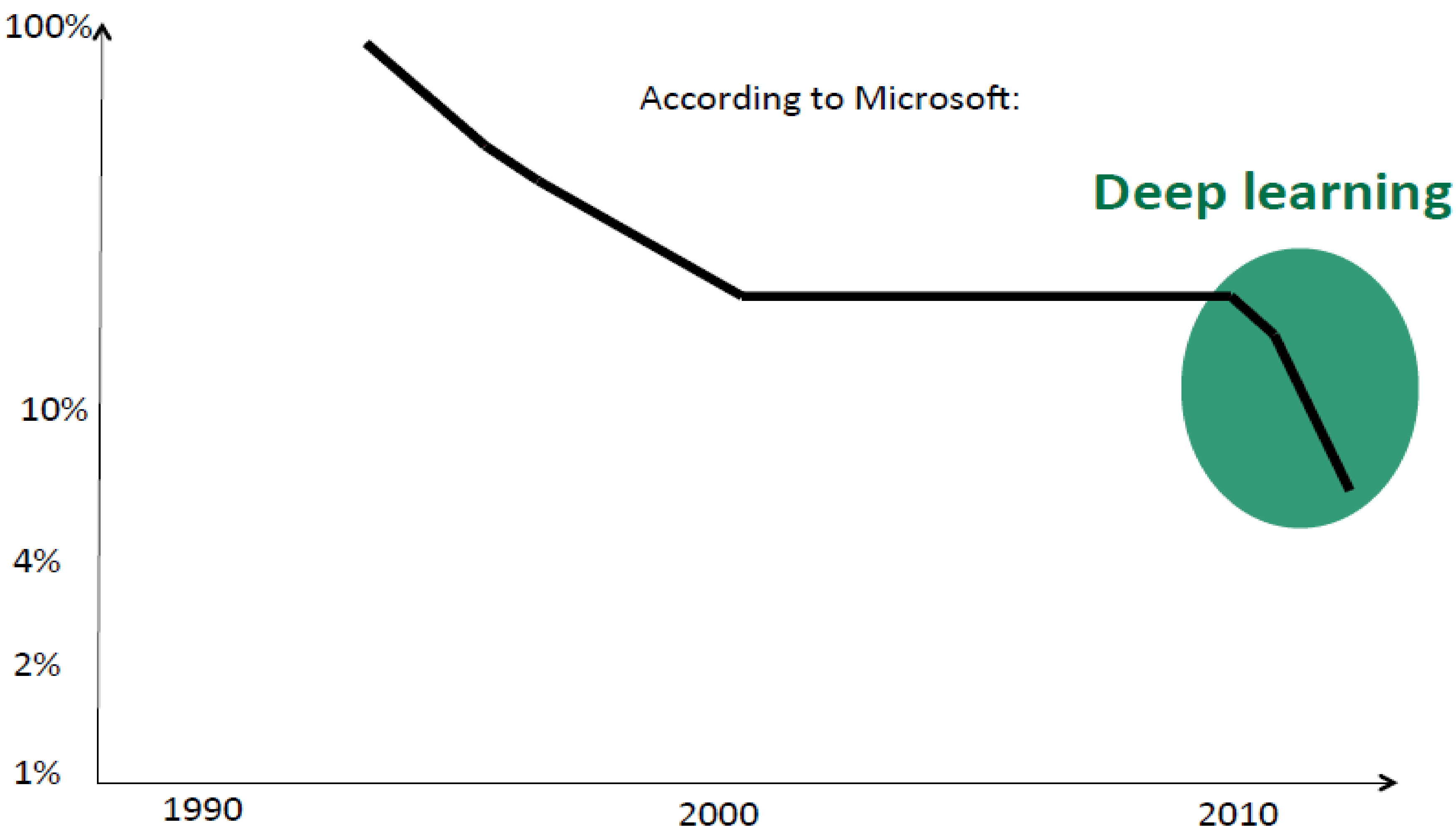
Progress of spontaneous speech recognition



Microsoft Research

Slide from **Y. Bengio**, Y. LeCun, G. Hinton (NIPS tutorial, Dec. 2015)

2010-2012: Breakthrough in speech recognition → in Androids by 2012

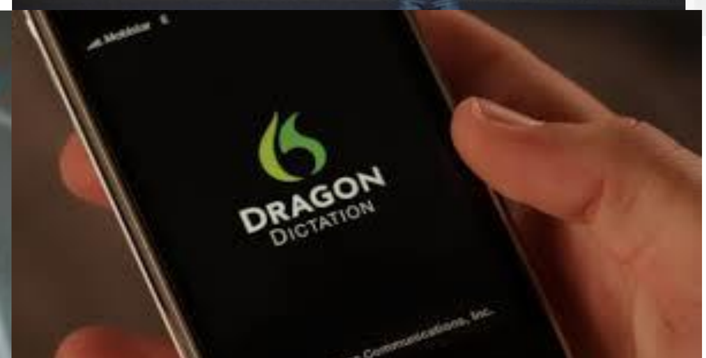


Across-the-Board Deployment of DNN in Speech Industry

(+ in university labs & DARPA programs) **(2012-2014)**

Skype to get 'real-time' translator

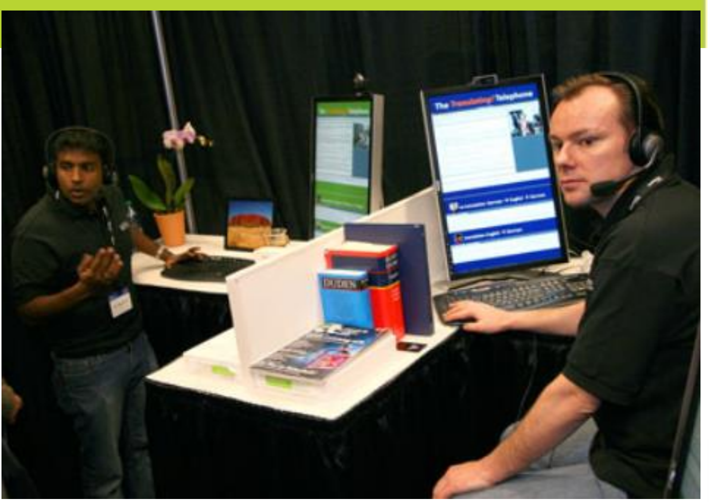
Analysts say the translation feature could have wide ranging applications



Enabling Cross-Lingual Conversations in Real Time

Microsoft Research
May 27, 2014 5:58 PM PT

View milestones on the path to Skype Translator
#speech2speech

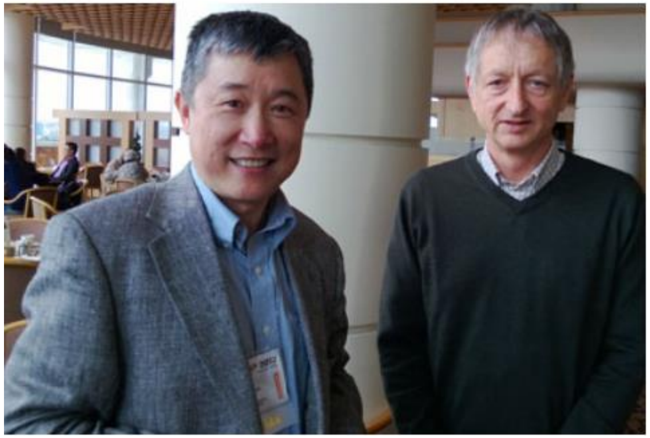


Microsoft Research

WIRED

ROBERT MCMILLAN BUSINESS 12.17.14 1:19 PM

HOW SKYPE USED AI TO BUILD ITS AMAZING NEW LANGUAGE TRANSLATOR



CNN

Taking a cue from science fiction, Microsoft demos 'universal translator'

By Jacopo Prisco, for CNN
Updated 12:35 PM ET, Thu October 16, 2014

In the academic world

nature International weekly journal of science

Deep learning

Yann LeCun, Yoshua Bengio & Geoffrey Hinton

Affiliations | Corresponding author

Nature **521**, 436–444 (28 May 2015) | doi:10.1038/nature14539

Received 25 February 2015 | Accepted 01 May 2015 | Published online

[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]

"This joint paper (2012) from the major speech recognition laboratories details the first major industrial application of deep learning."

Deep Neural Networks for Acoustic Modeling in Speech Recognition

[The shared views of four research groups]

State-of-the-Art Speech Recognition **Today**

(& **tomorrow** --- roles of unsupervised learning)

ASR: Neural Network Architectures at Google

Single Channel:

- **LSTM** acoustic model trained with connectionist temporal classification (**CTC**)
- Results on a **2,000-hr English Voice Search** task show an 11% relative improvement
- Papers: [H. Sak et al - ICASSP 2015, Interspeech 2015, A. Senior et al - ASRU 2015]

Model	WER
LSTM w/ conventional modeling	**14.0**
LSTM w/ CTC	**12.9%**

(Sainath, Senior, Sak, Vinyals)

Multi-Channel:

- Multi-channel raw-waveform input for each channel
- Initial network layers factored to do spatial and spectral filtering
- Output passed to a CLDNN acoustic model, entire network trained jointly
- Results on a 2,000-hr English Voice Search task show more than 10% relative improvement
- Papers: [T. N. Sainath et al - ASRU 2015, ICASSP 2016]

Model	WER
raw-waveform, 1ch	**19.2**
delay+sum, 8 channel	**18.7**
MVDR, 8 channel	**18.8**
factored raw-waveform, 2ch	**17.1**

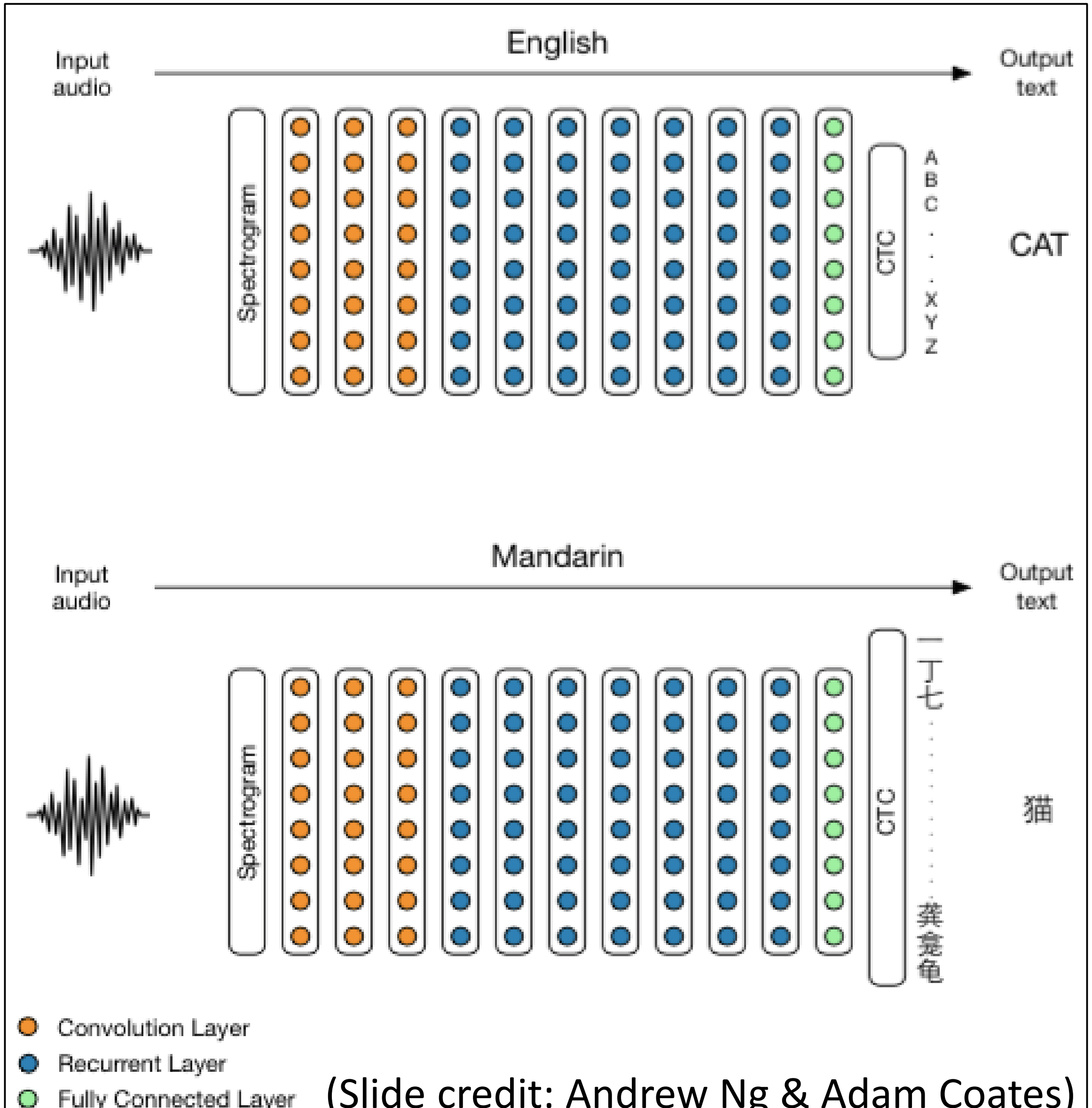
Google

(Slide credit: Tara Sainath & Andrew Senior)

Baidu's Deep Speech 2

End-to-End DL System for Mandarin and English

Bai du Research



(Slide credit: Andrew Ng & Adam Coates)

Paper: bit.ly/deepspeech2

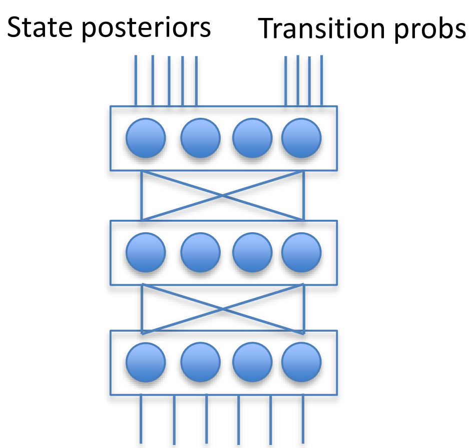
- Human-level Mandarin recognition on short queries:
  - DeepSpeech: **3.7% - 5.7% CER**
  - Humans: 4% - 9.7% CER
- Trained on 12,000 hours of conversational, read, mixed speech.
- 9 layer RNN with CTC cost:
  - 2D invariant convolution
  - 7 recurrent layers
  - Fully connected output
- Trained with SGD on heavily-optimized HPC system. "SortaGrad" curriculum learning.
- "Batch Dispatch" framework for low-latency production deployment.

Learning transition probabilities in DNN-HMM ASR

DNN outputs include not only state posterior outputs but also HMM transition probabilities

Real-time reduction of 16%

WER reduction of 10%



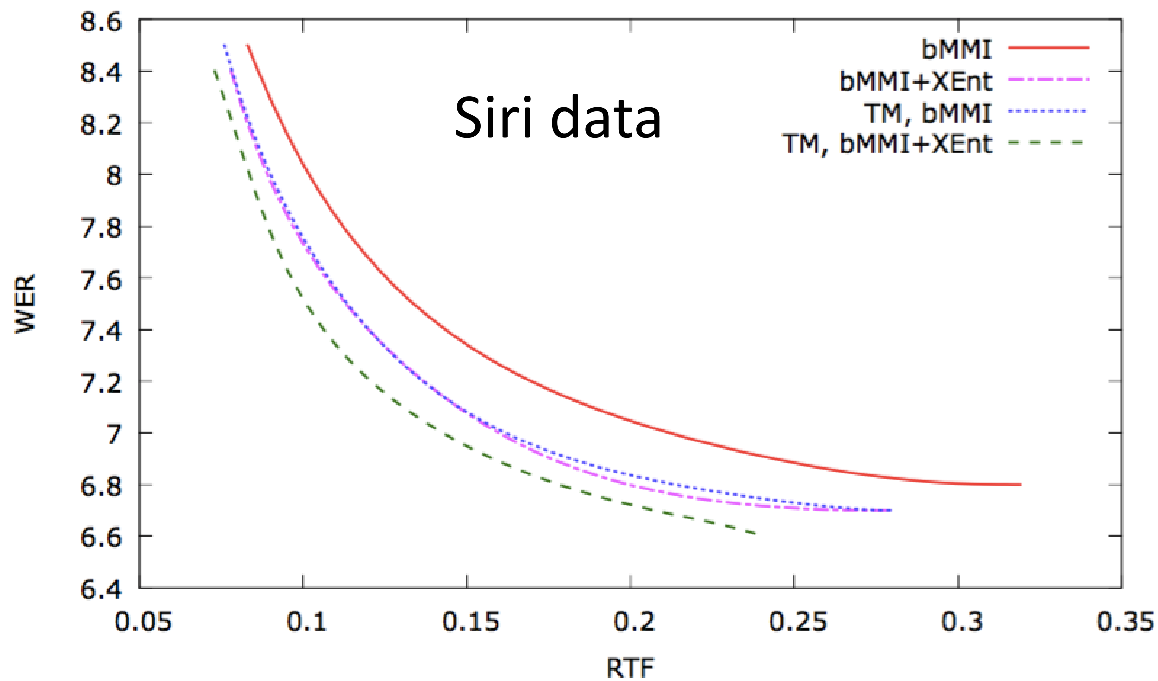


Figure 2: *WER vs. RTF (dev set)*

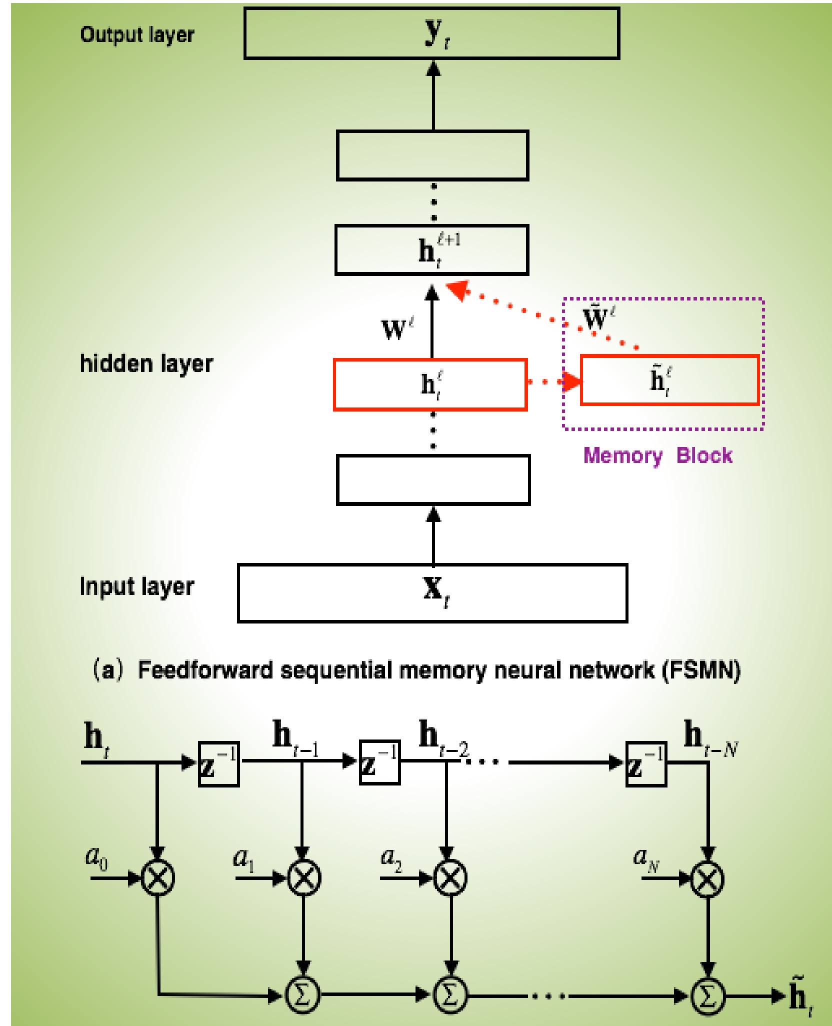
Matthias Paulik, "Improvements to the Pruning Behavior of DNN Acoustic Models". Interspeech 2015

(Slide: Alex Acero)

FSMN-based LVCSR System

- ❑ Feed-forward Sequential Memory Network(FSMN)
- ❑ Results on 10,000 hours Mandarin short message dictation task
  - ➢ 8 hidden layers
  - ➢ Memory block with -/+ 15 frames
  - ➢ CTC training criteria
- ❑ Comparable results to DBLSTM with smaller model size
- ❑ Training costs only 1 day using 16 GPUs and ASGD algorithm

Model	#Para.(M)	CER (%)
ReLU DNN	40	6.40
LSTM	27.5	5.25
BLSTM	45	4.67
FSMN	19.8	4.61



(a) Feedforward sequential memory neural network (FSMN)

(b) Memory block in FSMN as FIR filter

Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, Yu Hu. “Feedforward Sequential Memory Networks: A New Structure to Learn Long-term Dependency ”. arXiv:1512.08031, 2015.

(slide credit: Cong Liu & Yu Hu)

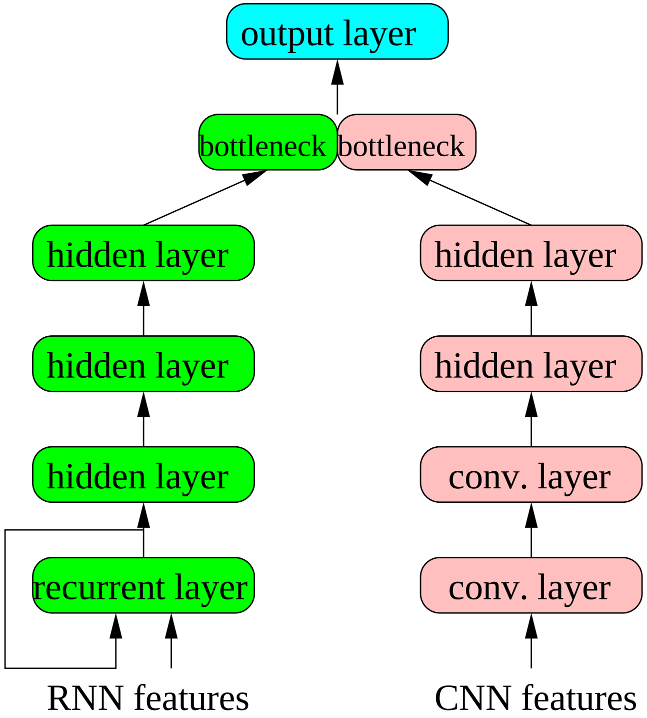
English Conversational Telephone Speech Recognition*

Key ingredients:

- Joint RNN/CNN acoustic model trained on 2000 hours of publicly available audio
- Maxout activations
- Exponential and NN language models

WER Results on Switchboard Hub5-2000:

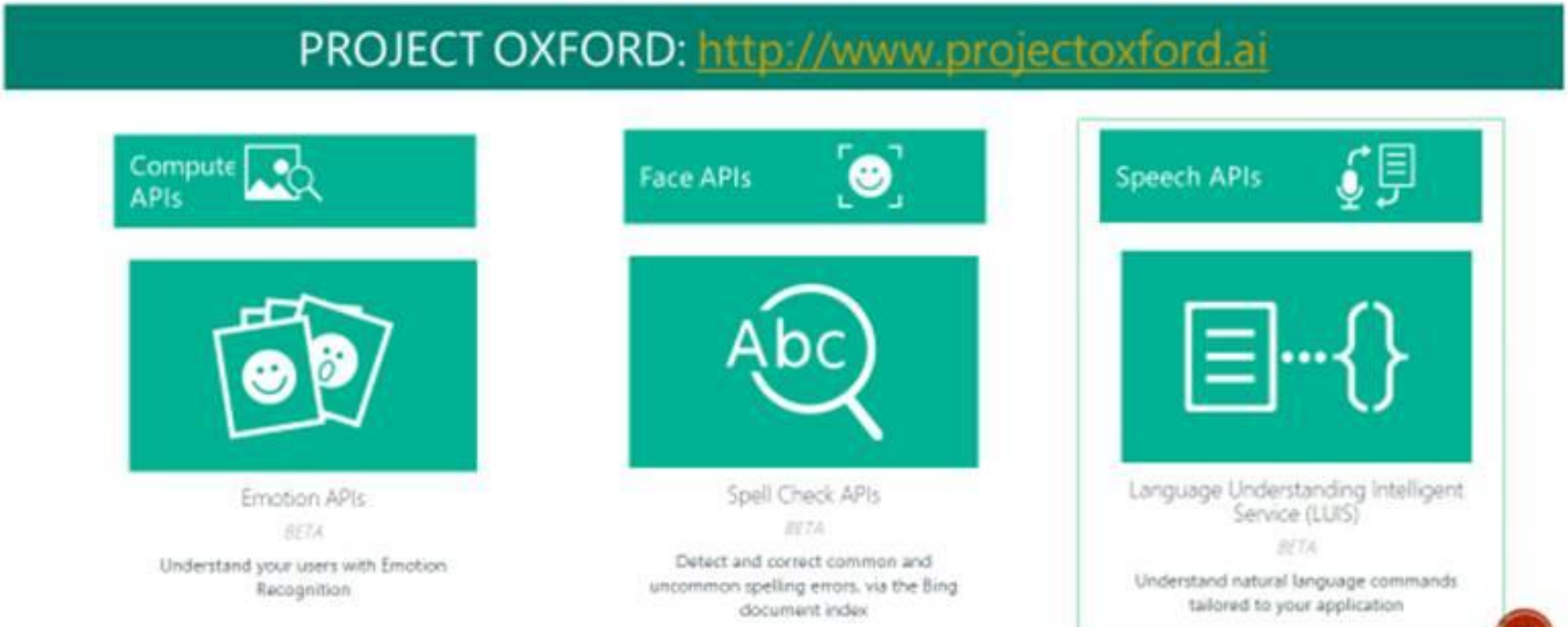
Model	WER SWB	WER CH
CNN	10.4	17.9
RNN	9.9	16.3
Joint RNN/CNN	9.3	15.6
+ LM rescoring	**8.0%**	14.1



*Saon *et al.* "The IBM 2015 English Conversational Telephone Speech Recognition System", Interspeech 2015.

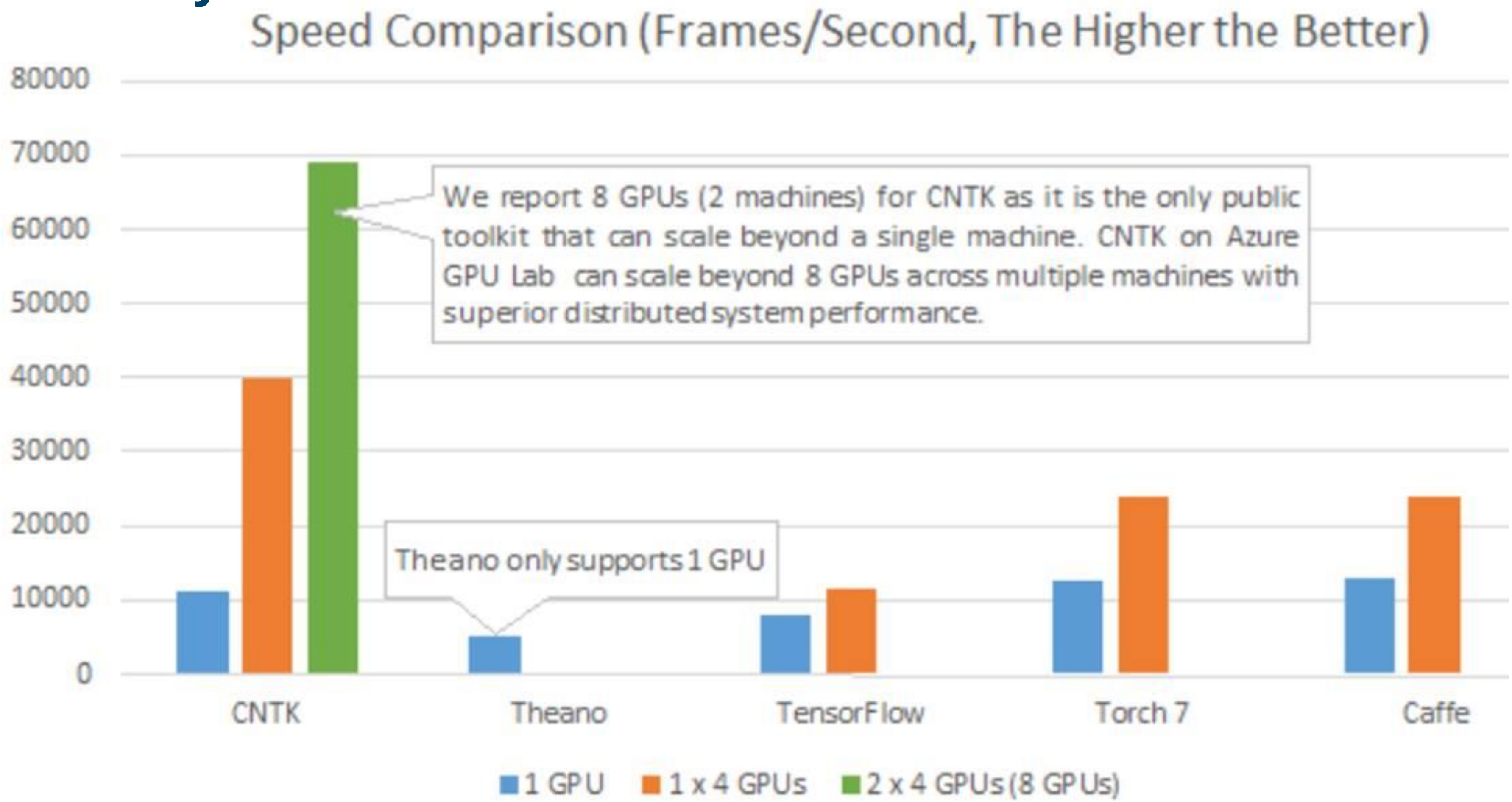
(Slide credit: G. Saon & B. Kingsbury)

MICROSOFT PROJECT OXFORD SERVICES



(Slide credit: Xuedong Huang)

CNTK/Phily



*Google updated that TensorFlow can now scale to support multiple machines recently; comparisons have not been made yet

- Recent Research at MS (ICASSP-2016):

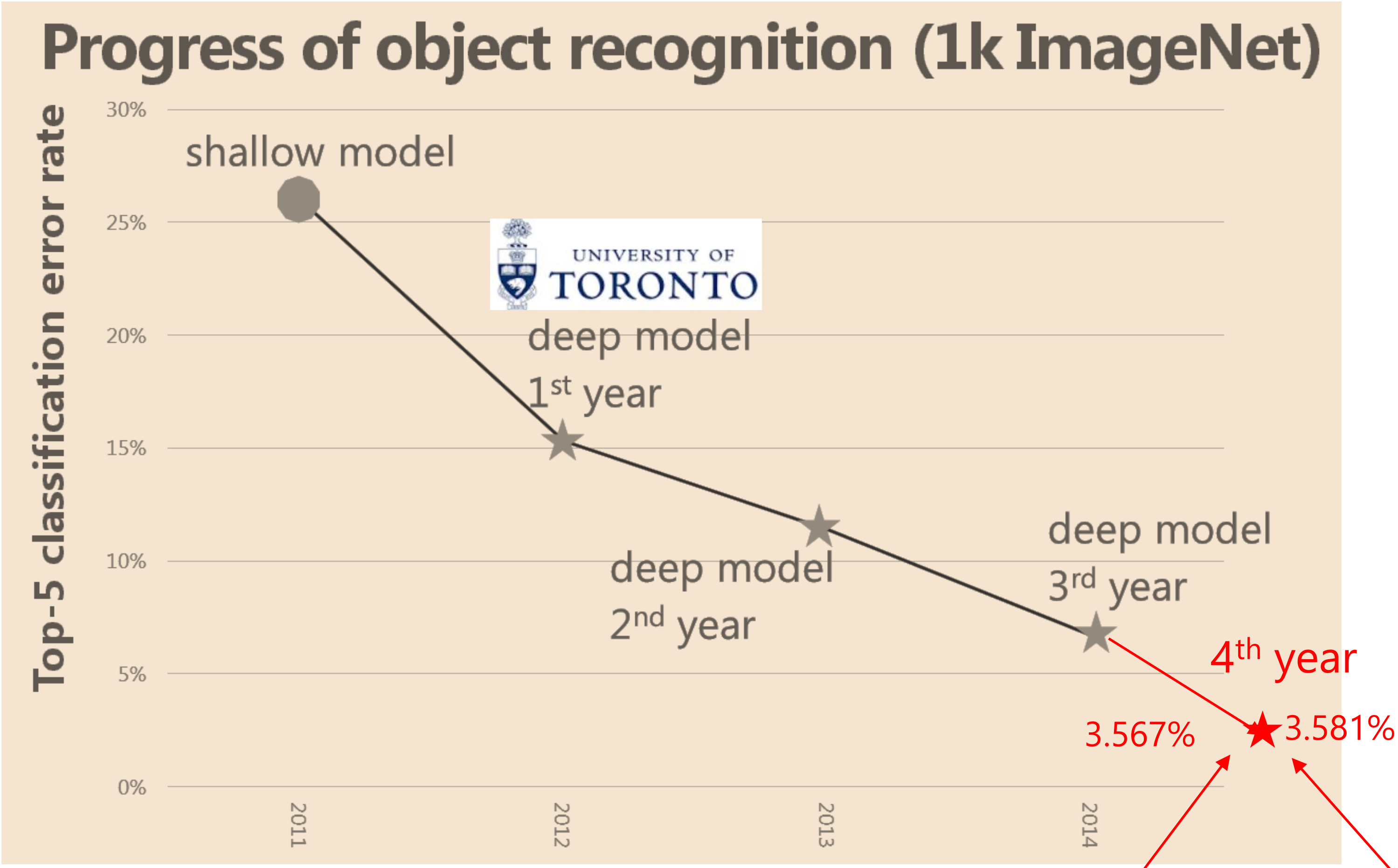
-“SCALABLE TRAINING OF DEEP LEARNING MACHINES BY INCREMENTAL BLOCK TRAINING WITH INTRA-BLOCK PARALLEL OPTIMIZATION AND BLOCKWISE MODEL-UPDATE FILTERING”

-“HIGHWAY LSTM RNNs FOR DISTANCE SPEECH RECOGNITION”

-”SELF-STABILIZED DEEP NEURAL NETWORKS”

Deep Learning also Shattered Image Recognition (since 2012)

IMAGENET Competition



Announced Dec.2015

2012 - 2015

Super-deep: 152 layers

Microsoft Research

Google

WIRED

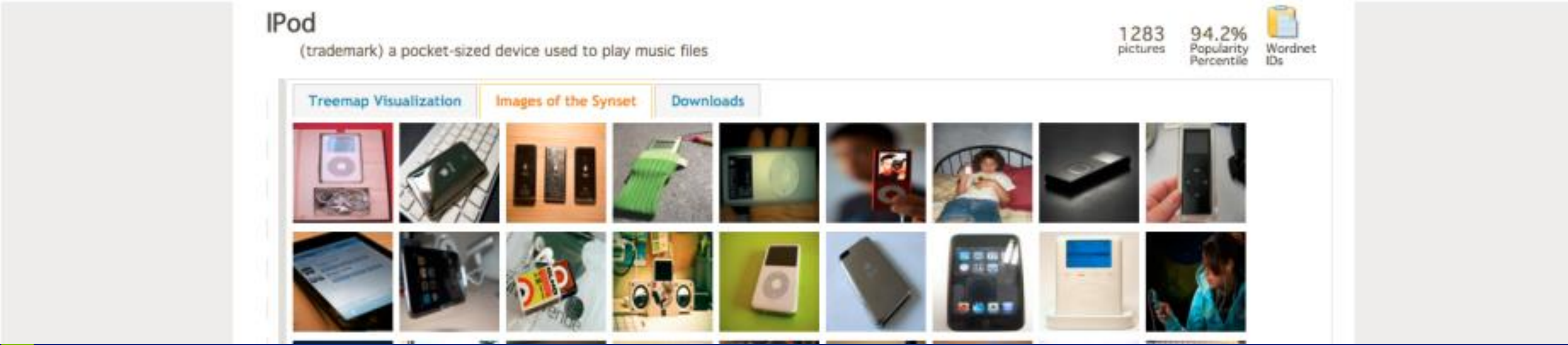
CADE METZ BUSINESS 01.14.16 7:00 AM

MICROSOFT NEURAL NET SHOWS DEEP LEARNING CAN GET WAY DEEPER

Microsoft beats Google, Intel, Tencent, and Qualcomm in image recognition competition

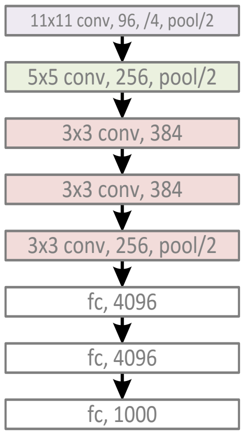
JORDAN NOVET DECEMBER 10, 2015 4:14 PM

TAGS: ARTIFICIAL INTELLIGENCE, DEEP LEARNING, IBM, IMAGE RECOGNITION, IMAGENET, MICROSOFT, MICROSOFT RESEARCH, NVIDIA, SOFTLAYER

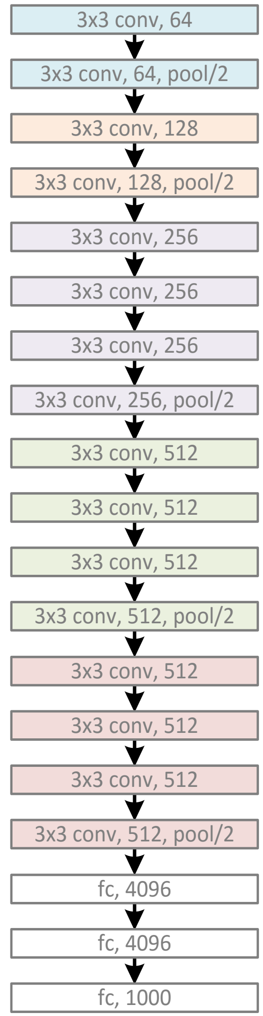


Depth is of crucial importance

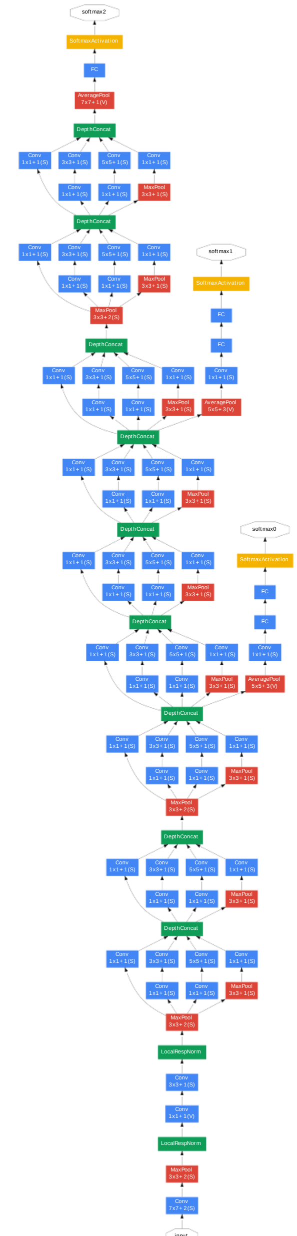
AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



GoogleNet, 22 layers
(ILSVRC 2014)

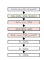


ILSVRC (Large Scale Visual Recognition Challenge)

(slide credit: Jian Sun, MSR)

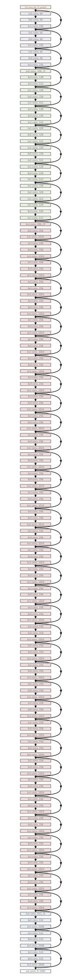
Depth is of crucial importance

AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)

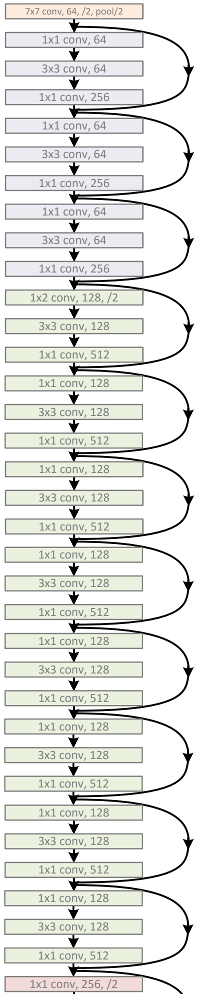


ILSVRC (Large Scale Visual Recognition Challenge)

(slide credit: Jian Sun, MSR)

Depth is of crucial importance

ResNet, 152 layers

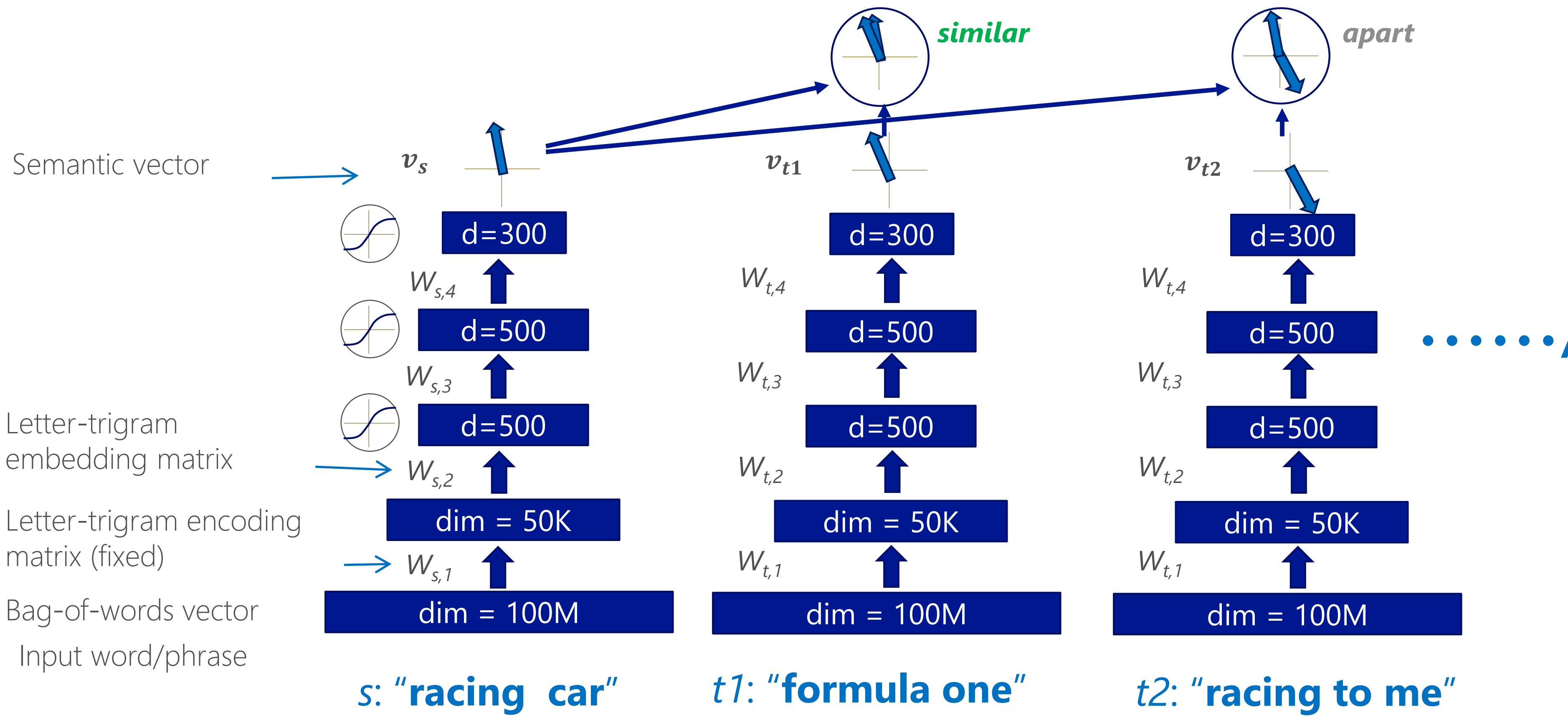


(slide credit: Jian Sun, MSR)

Outline

- Deep learning for machine **perception**
    - Speech
    - Image
- Deep learning for machine **cognition**
    - Semantic modeling
    - Natural language
    - Multimodality
    - Reasoning, attention, memory (RAM)
    - Knowledge representation/management/exploitation
    - Optimal decision making (by deep reinforcement learning)
- Three hot areas/challenges of deep learning & AI research

Deep Semantic Model for Symbol Embedding

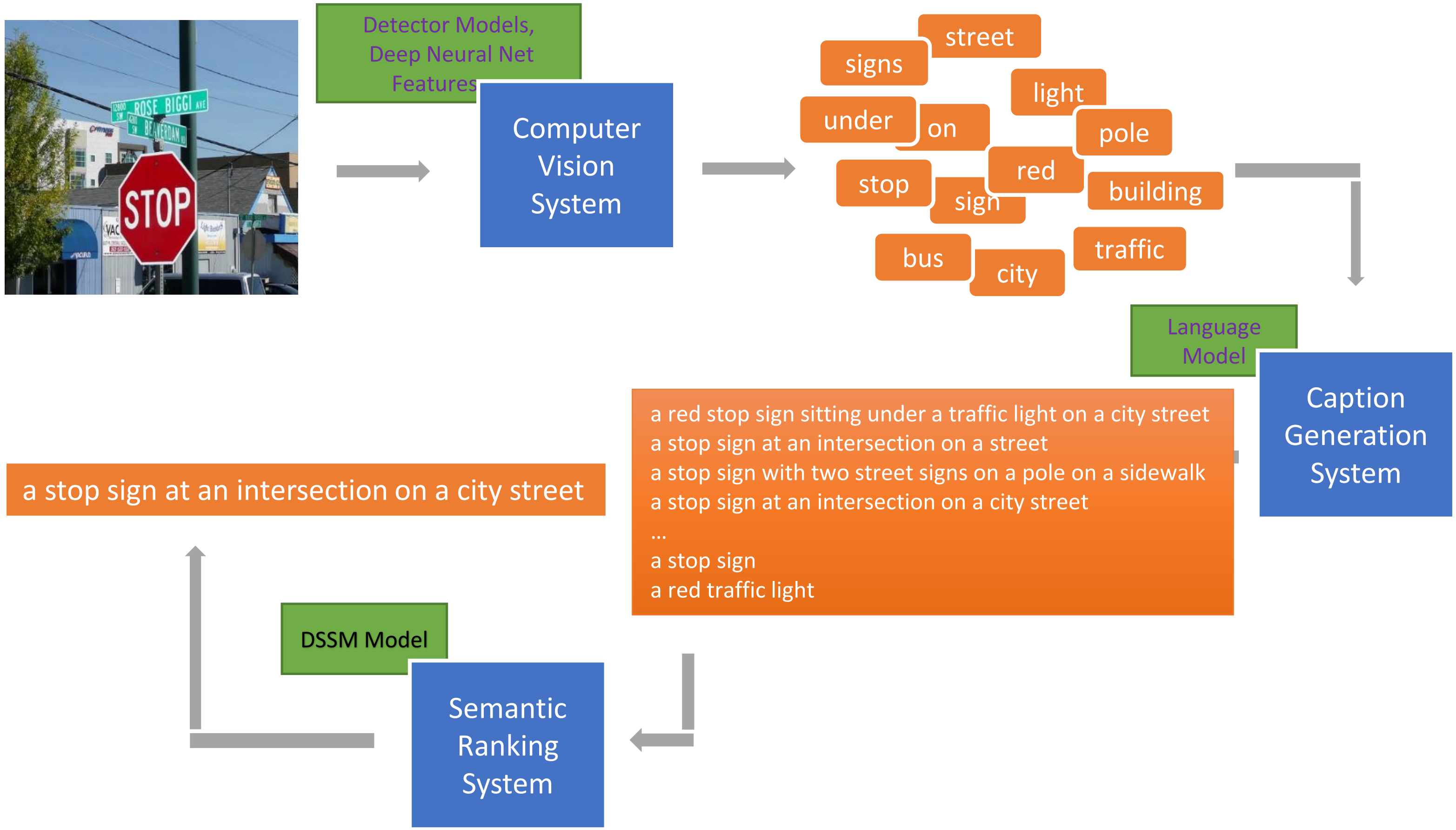


Huang, P., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. Learning deep structured semantic models for web search using clickthrough data. In ACM-CIKM, 2013.

Many applications of Deep Semantic Modeling:
Learning semantic relationship between "*Source*" and "*Target*"

Tasks	*Source*	*Target*
Word semantic embedding	*context*	*word*
Web search	*search query*	*web documents*
Query intent detection	*Search query*	*Use intent*
Question answering	*pattern / mention* (in NL)	*relation / entity* (in knowledge base)
Machine translation	*sentence in language a*	*translated sentences in language b*
Query auto-suggestion	*Search query*	*Suggested query*
Query auto-completion	*Partial search query*	*Completed query*
Apps recommendation	*User profile*	*recommended Apps*
Distillation of survey feedbacks	*Feedbacks in text*	*Relevant feedbacks*
Automatic image captioning	***image***	***text caption***
Image retrieval	*text query*	*images*
Natural user interface	*command (text / speech / gesture)*	*actions*
Ads selection	*search query*	*ad keywords*
Ads click prediction	*search query*	*ad documents*
Email analysis: people prediction	*Email content*	*Recipients, senders*
Email search	*Search query*	*Email content*
Email declutering	*Email contents*	*Email contents in similar threads*
Knowledge-base construction	*entity from source*	*entity fitting desired relationship*
Contextual entity search	*key phrase / context*	*entity / its corresponding page*
Automatic highlighting	*documents in reading*	*key phrases to be highlighted*
Text summarization	*long text*	*summarized short text*

Automatic image captioning (MSR system)



Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "**From captions to visual concepts and back**," CVPR, 2015

A	a woman in a kitchen preparing food
B	woman working on counter near kitchen sink preparing a meal

Machine:	a woman in a kitchen preparing food
Human:	woman working on counter near kitchen sink preparing a meal

COCO Challenge Results (CVPR-2015, Boston)

Tied for 1st prize

	M1	M2	M3	M4	M5
Human[5]	0.638	0.675	4.836	3.428	0.352
MSR[8]	0.268	0.322	4.137	2.662	0.234
Google[4]	0.273	0.317	4.107	2.742	0.233
MSR Captivator[9]	0.250	0.301	4.149	2.565	0.233
Montreal/Toronto [10]	0.262	0.272	3.932	2.832	0.197
Berkeley LRCN[2]	0.246	0.268	3.924	2.786	0.204
Nearest Neighbor [11]	0.216	0.255	3.801	2.716	0.196

M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	Percentage of captions that are similar to human description.

Deep Learning for Machine Cognition

--- Deep reinforcement learning

--- “Optimal” actions: control and business decision making

nature International weekly journal of science

LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

The theory of reinforcement learning provides a normative account[1], deeply rooted in psychological[2] and neuroscientific[3] perspectives on animal behaviour, of how agents may optimize their control of an environment. To use reinforcement learning successfully in situations approaching real-world complexity, however, agents are confronted

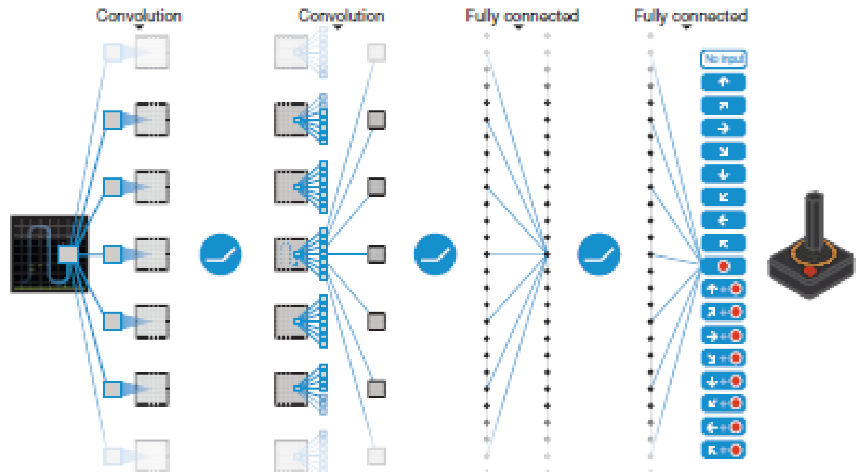
agent is to select actions in a fashion that maximizes cumulative future reward. More formally, we use a deep convolutional neural network to approximate the optimal action-value function

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \middle| s_t = s,\ a_t = a,\ \pi\right],$$

Reinforcement learning from "non-working" to "working", due to Deep Learning (much like DNN for speech)

Deep Q-Network (DQN)

mapping raw screen pixels



to predictions of final score for each of 18 joystick actions

- Input layer: image vector of s
- Output layer: a single output Q-value for each action a, $Q(s, a, \theta)$
- DNN parameters: θ

Reinforcement Learning
--- optimizing long-term values

	Short-term	Long-term
Playing the Breakout game		
Optimizing Business Decision Making	*Maximize immediate reward*	*Optimize life-time revenue, service usages, and customer satisfaction*

Self play to improve skills

nature International weekly journal of science

ARTICLE PREVIEW

view full access options ▸

NATURE | ARTICLE

日本語要約

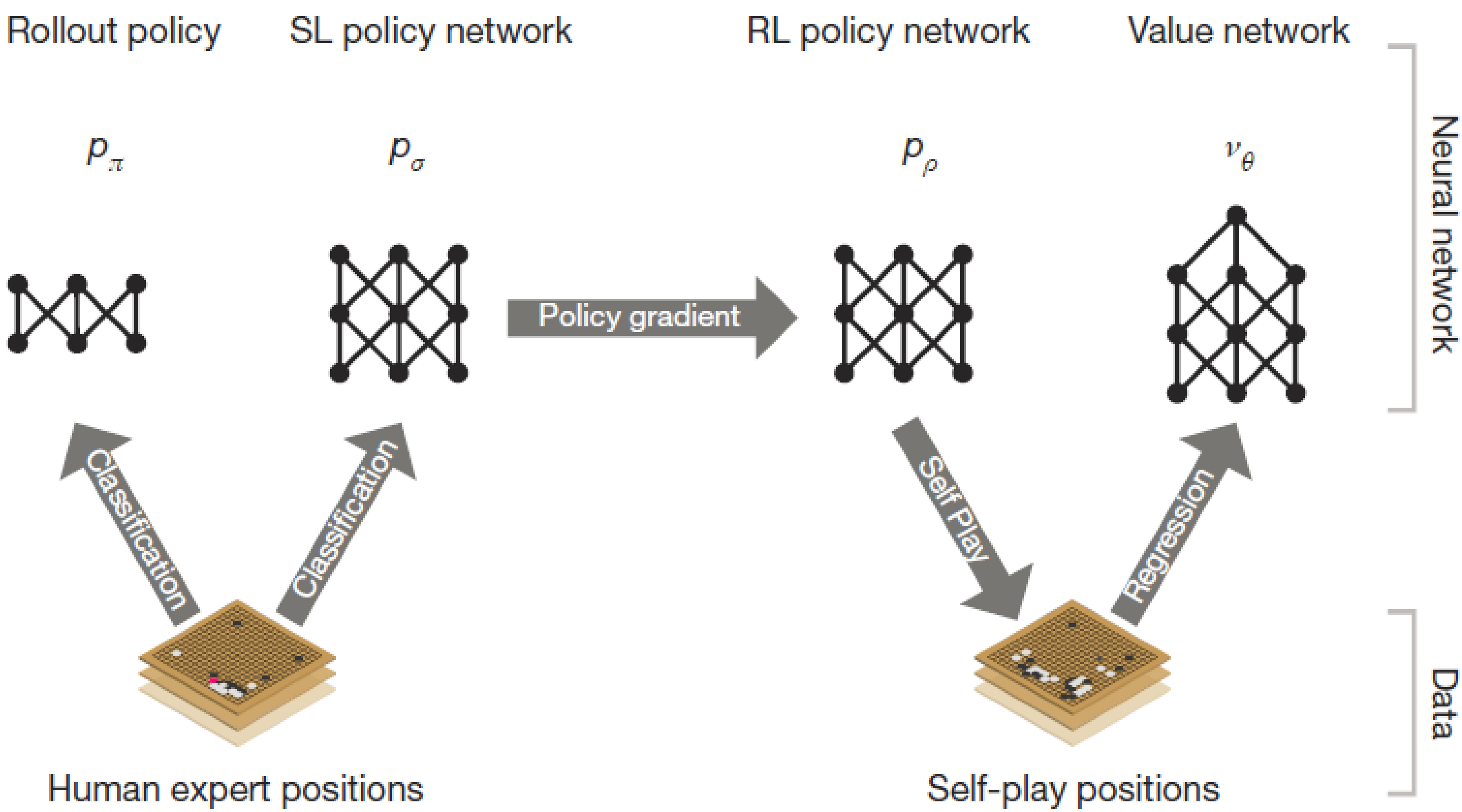
Mastering the game of Go with deep neural networks and tree search

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis

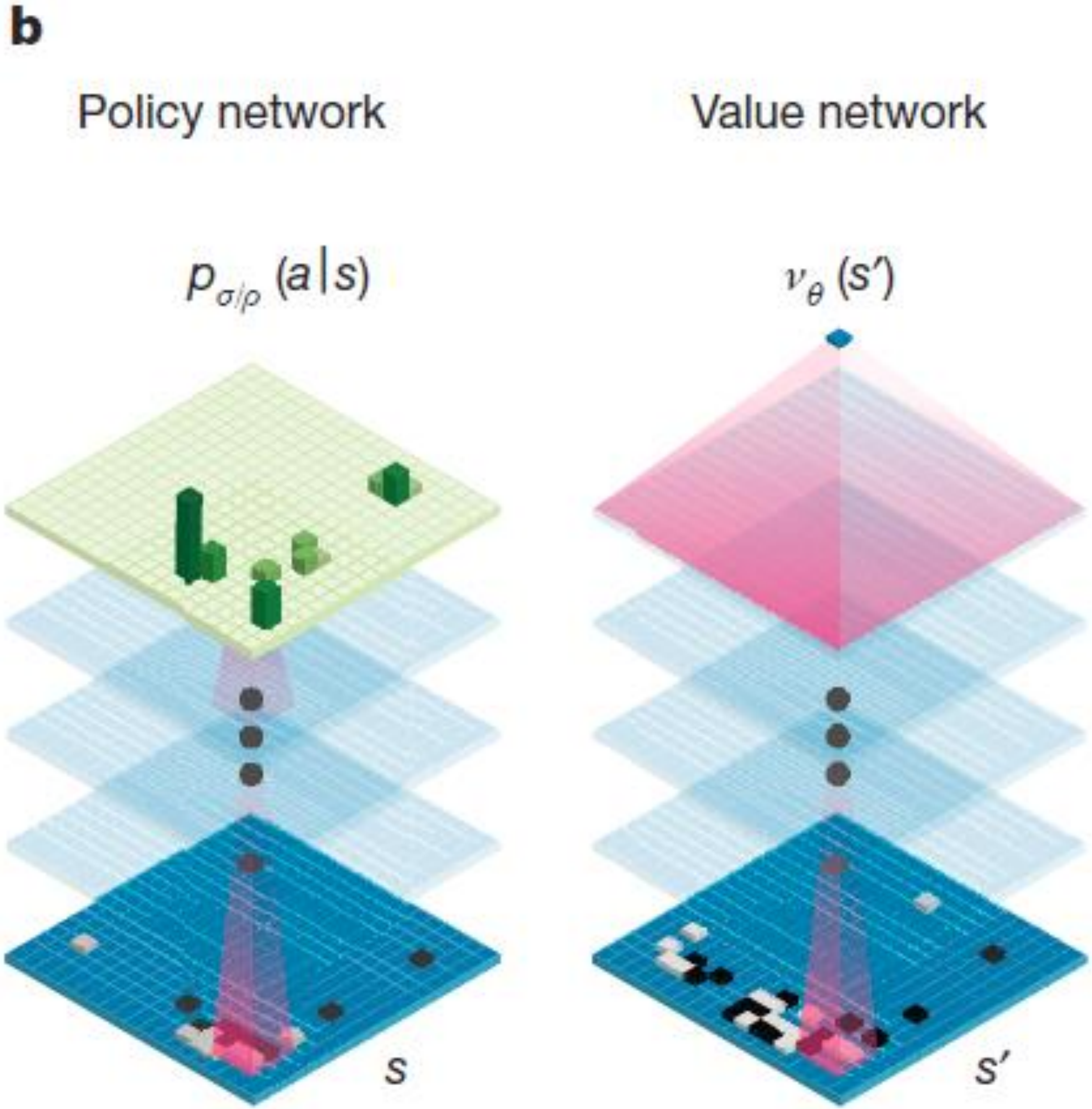
Affiliations | Contributions | Corresponding authors

Nature **529**, 484–489 (28 January 2016) | doi:10.1038/nature16961
Received 11 November 2015 | Accepted 05 January 2016 | Published online 27 January 2016

DNN learning pipeline in AlphaGo



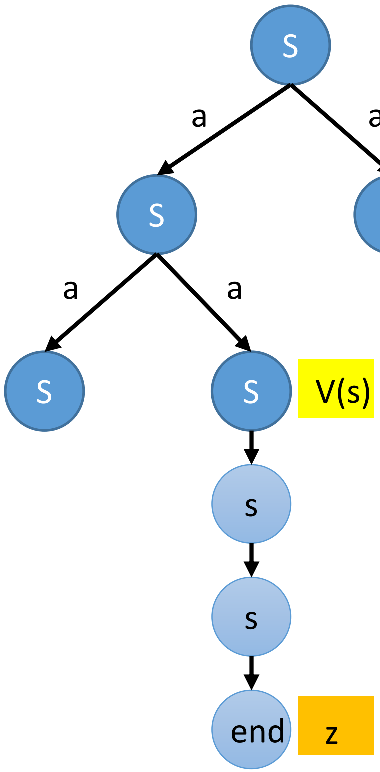
DNN architecture used in AlphaGo



Analysis of four DNNs in AlphaGo

DNNs	**Properties**	**Architecture**	**Additional details**
$\pi_{SL}(a\|s)$	Slow, accurate stochastic supervised learning policy, trained on 30M (s,a) pairs	13 layer network; alternating ConvNets and rectifier non-linearities; output dist. over all legal moves	Evaluation time: 3 ms Accuracy vs. corpus: 57% Train time: 3 weeks
$\tilde{\pi}_{SL}(a\|s)$	Fast, less accurate stochastic SL policy, trained on 30M (s,a) pairs	Linear softmax of small pattern features	Evaluation time: 2 us Accuracy vs. corpus: 24%
$\pi_{RL}(a\|s)$	Stochastic RL policy, trained by self-play	Same as π_{SL}	Win vs. π_{SL} : 80%
$V(s)$	Value function: % chance of π_{RL} winning by starting in state *s*	Same as π_{SL} , but with one output (% chance of winning)	15K less computation than evaluating π_{RL} with roll-outs

Monte Carlo Tree Search in AlphaGo



$$\pi(s) = \mathrm{argmax}_a \, Q(s,a)$$

$$Q(s,a) = \underbrace{Q'(s,a)}_{\text{Roll-out estimate}} + \underbrace{u(s,a)}_{\text{Exploration bonus}}$$

Mixture weight

$$Q'(s,a) = \frac{1}{N(s,a)} \sum_i \left[(1-\lambda) V\left(s_L^i\right) + \lambda z_L^i\right]$$

# of times action *a* taken in state *s*

Value function computed in advance

Win/loss result of 1 roll-out with $\tilde{\pi}_{SL}(a|s)$

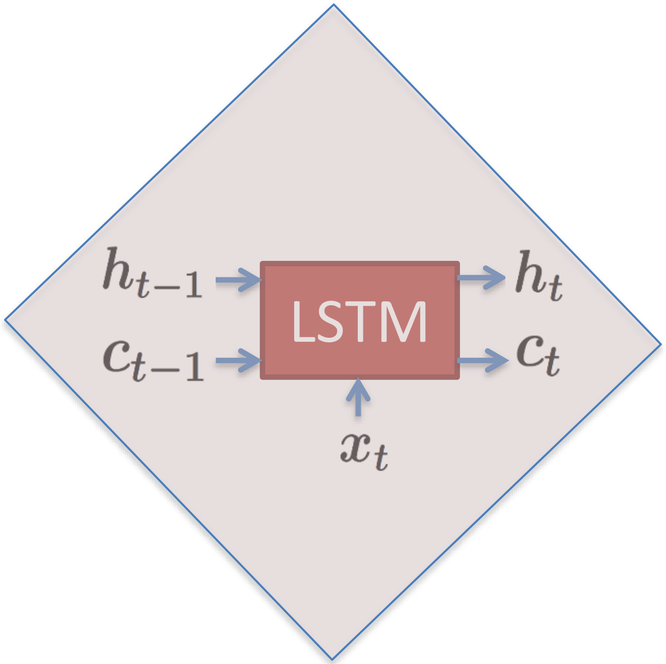
$$u(s,a) = c \cdot \pi_{SL}\,(a|s) \frac{\sqrt{\sum_b N(s,b)}}{1 + N(s,a)}$$

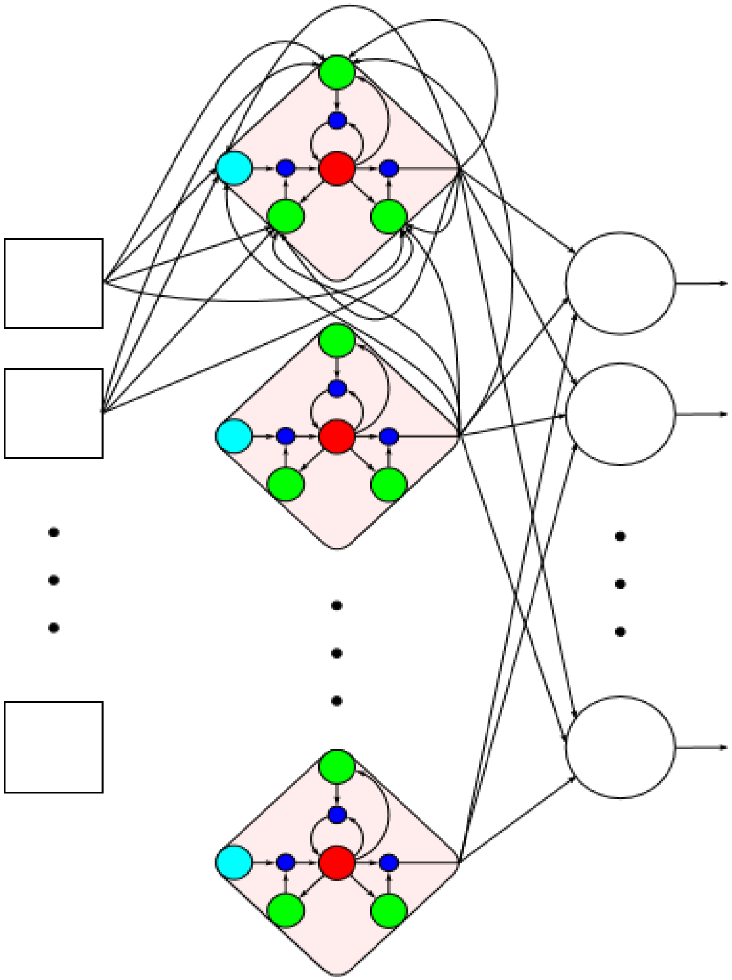
- Think of this MCTS component as a highly efficient "decoder", a concept familiar to ASR
- -> A* search and fast match in speech recognition literature during 80's-90's
- This is tree search (GO-specific), not graph search (A*)
- Speech is a relatively simple signal → sequential beam search sufficient, no need for A* or tree
- Key innovation in AlphaGO: "scores" in MCTS computed by DNNs with RL

Deep Learning for Machine Cognition

--- Memory & attention (applied to machine translation)

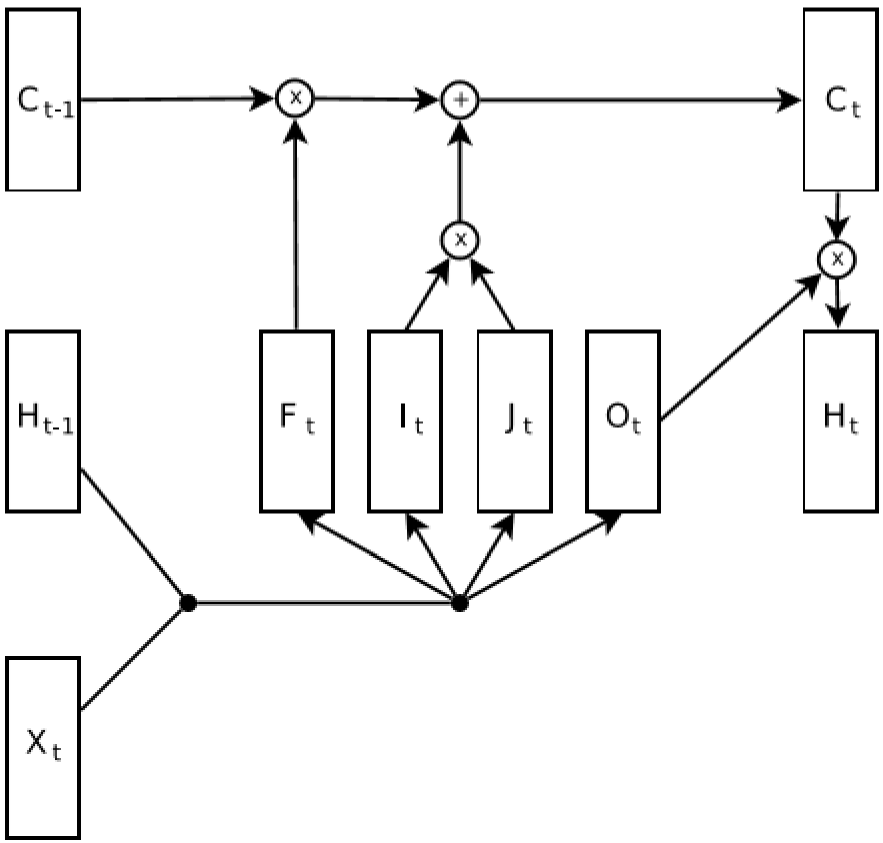
Long Short-Term Memory RNN





(Hochreiter & Schmidhuber, 1997)

LSTM cell unfolding over time



$$
\begin{aligned}
i_t &= \tanh(W_{\mathrm{xi}} x_t + W_{\mathrm{hi}} h_{t-1} + b_{\mathrm{i}}) \\
j_t &= \mathrm{sigm}(W_{\mathrm{xj}} x_t + W_{\mathrm{hj}} h_{t-1} + b_{\mathrm{j}}) \\
f_t &= \mathrm{sigm}(W_{\mathrm{xf}} x_t + W_{\mathrm{hf}} h_{t-1} + b_{\mathrm{f}}) \\
o_t &= \tanh(W_{\mathrm{xo}} x_t + W_{\mathrm{ho}} h_{t-1} + b_{\mathrm{o}}) \\
c_t &= c_{t-1} \odot f_t + i_t \odot j_t \\
h_t &= \tanh(c_t) \odot o_t
\end{aligned}
$$

Figure 1. The LSTM architecture. The value of the cell is increased by $i_t \odot j_t$, where $\odot$ is element-wise product. The LSTM's output is typically taken to be h_t, and c_t is not exposed. The forget gate f_t allows the LSTM to easily reset the value of the cell.

(Jozefowics, Zarembe, Sutskever, ICML 2015)

Gated Recurrent Unit (GRU)

(simpler than LSTM; no output gates)

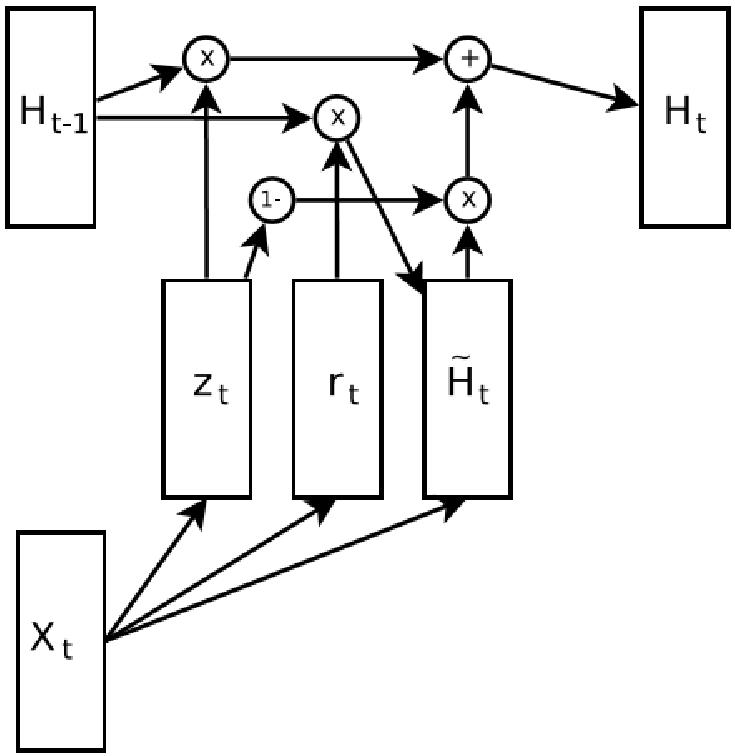


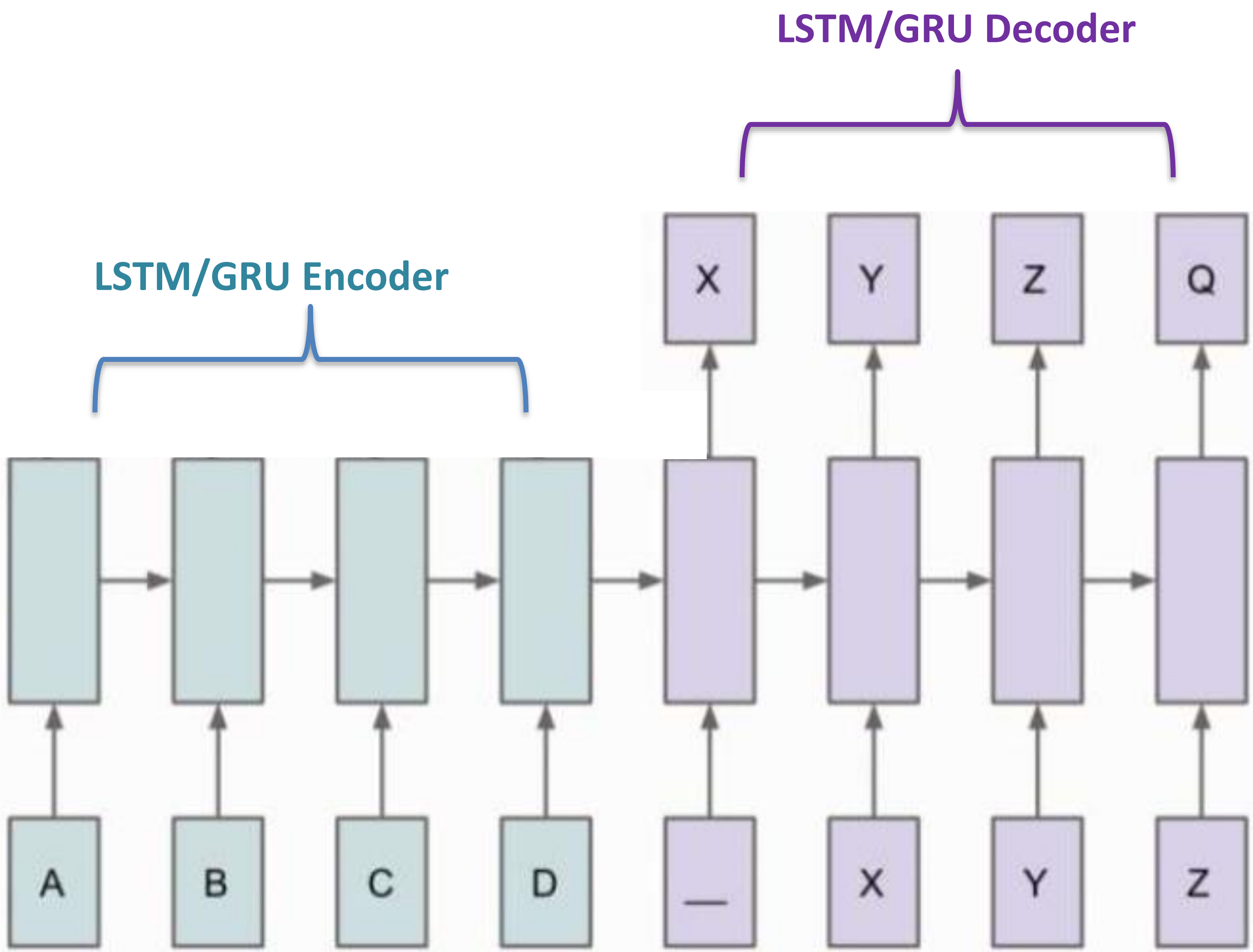
Figure 2. The Gated Recurrent Unit. Like the LSTM, it is hard to tell, at a glance, which part of the GRU is essential for its functioning.

$$
\begin{aligned}
r_t &= \mathrm{sigm}\left(W_{\mathrm{xr}} x_t + W_{\mathrm{hr}} h_{t-1} + b_{\mathrm{r}}\right) \\
z_t &= \mathrm{sigm}(W_{\mathrm{xz}} x_t + W_{\mathrm{hz}} h_{t-1} + b_{\mathrm{z}}) \\
\tilde{h}_t &= \tanh(W_{\mathrm{xh}} x_t + W_{\mathrm{hh}}(r_t \odot h_{t-1}) + b_{\mathrm{h}}) \\
h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t
\end{aligned}
$$

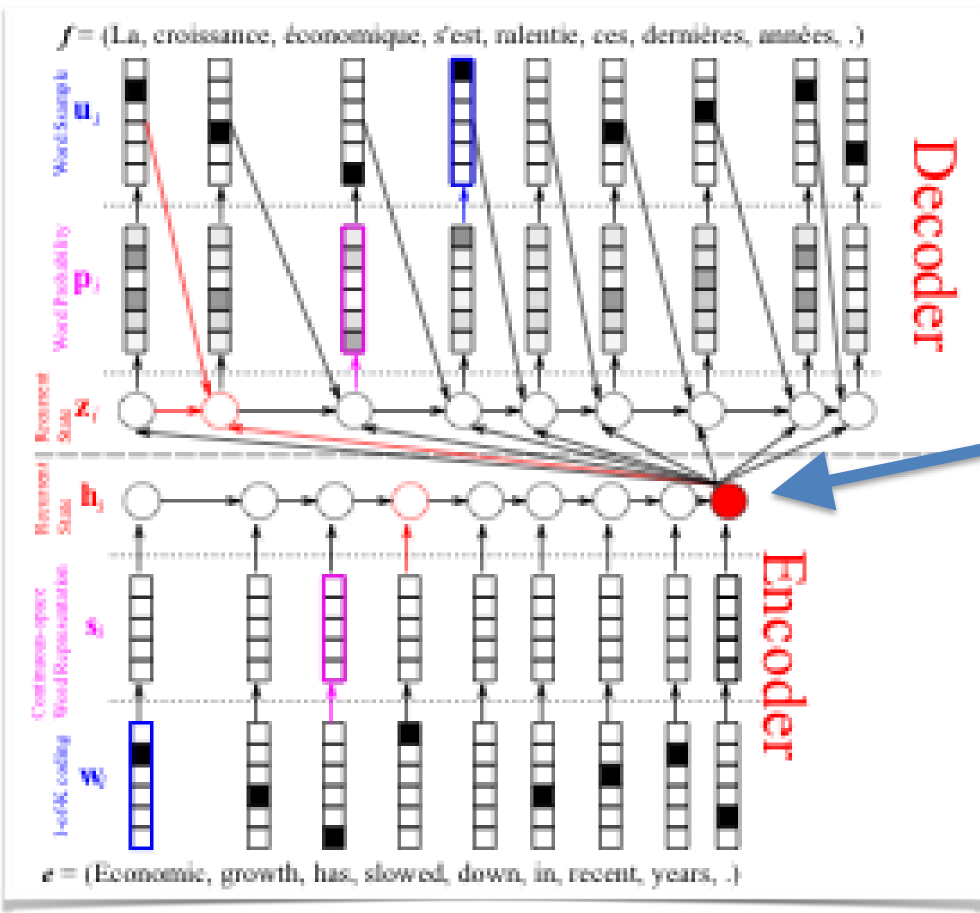
(Jozefowics, Zarembe, Sutskever, ICML 2015; Google
Kumar et al., arXiv, July, 2015; Metamind)

Seq-2-Seq Learning (Neural Machine Translation)

[Sutskever, Vinyals, Le, NIPS, 2014]



Neural Machine Translation



"Thought vector"

(Forcada&Ñeco, 1997;
Castaño&Casacuberta, 1997;
Kalchbrenner&Blunsom, 2013;
Sutskever et al., 2014;
Cho et al., 2014)

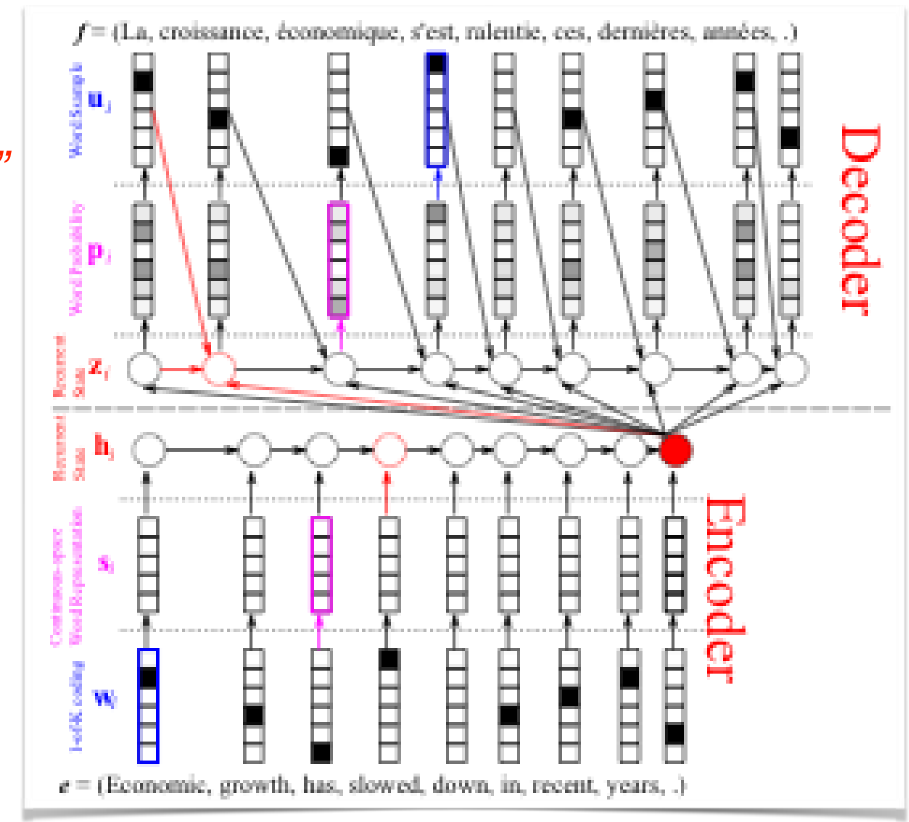
(slide credit: Kyunghyun Cho, 2016)

Neural Machine Translation

- This model replying "thought vector" does not perform well
- Especially for long source sentences
- *Because:*

"You can't cram the meaning of a whole %&!$# sentence into a single $&!# vector!"*

Ray Mooney



(modified from: Kyunghyun Cho, 2016)

Neural Machine Translation with Attention

Attention-based Model

- Encoder: Bidirectional RNN
  - A set of *annotation* vectors

    $\{h_1, h_2, \ldots, h_T\}$
- Attention-based Decoder

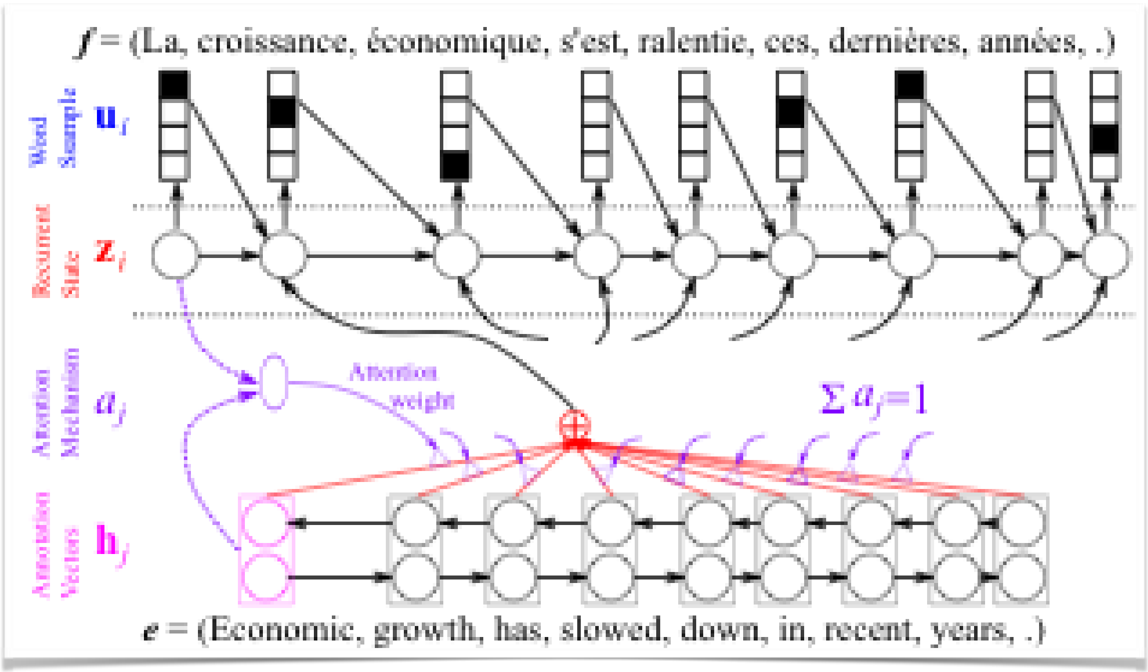
  (1) Compute attention weights

  $\alpha_{t',t} \propto \exp(e(z_{t'-1}, u_{t'-1}, h_t))$

  (2) Weighted-sum of the annotation vectors

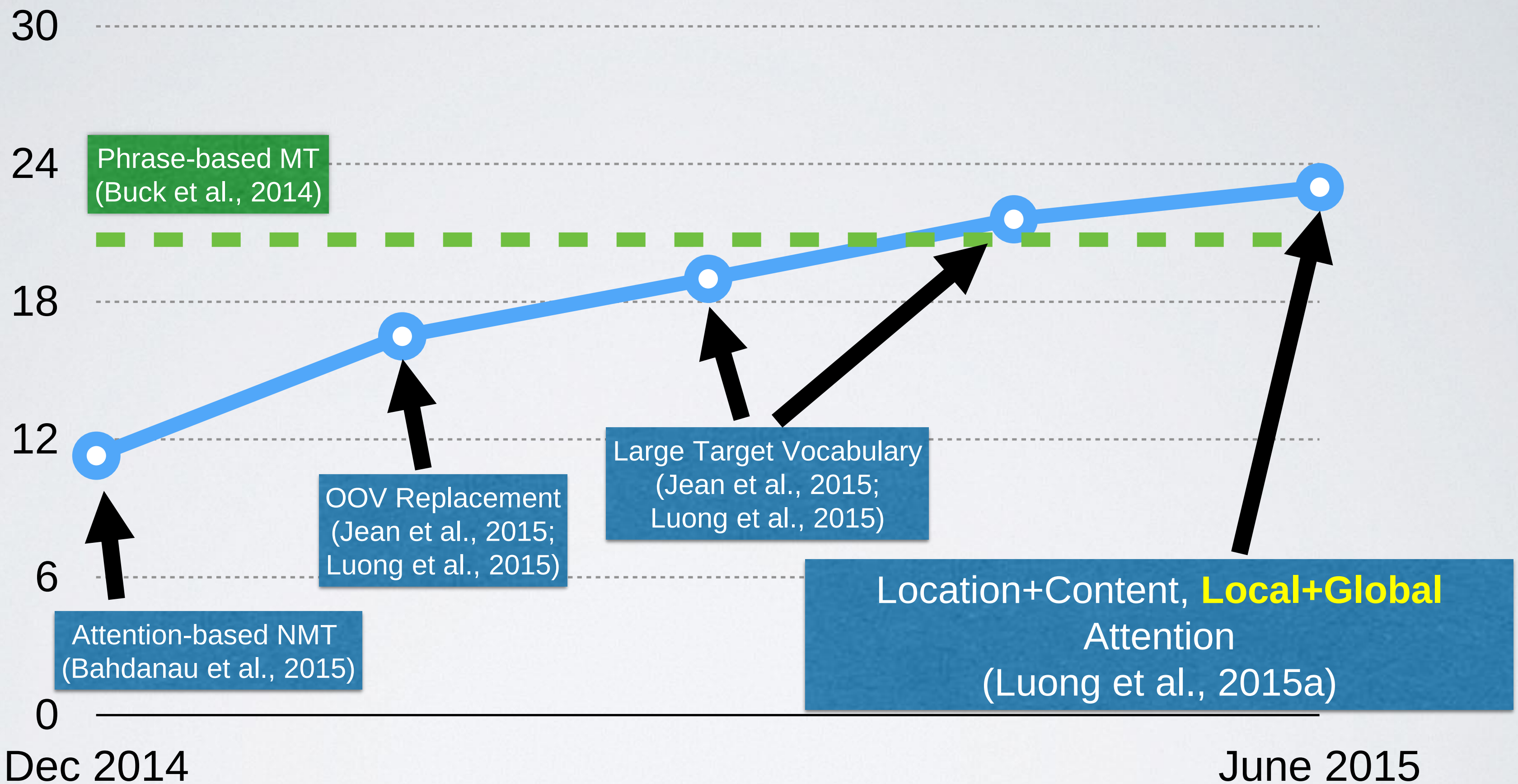
  $c_{t'} = \sum_{t=1}^{T} \alpha_{t',t} h_t$

  (3) Use $c_{t'}$ to replace "though vector" h_T



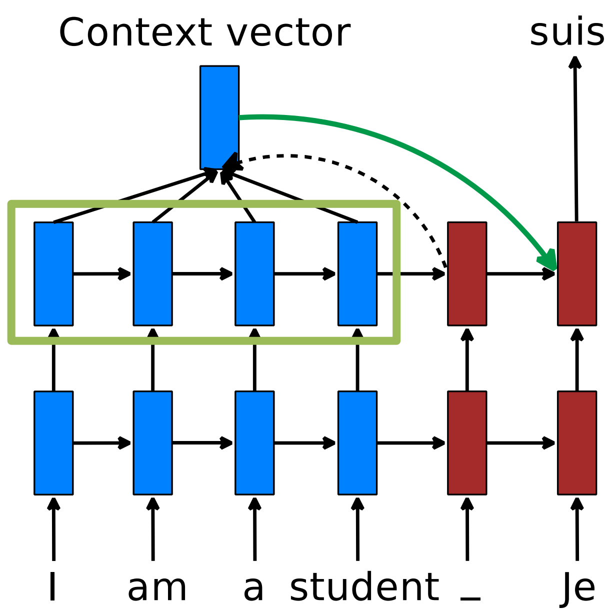
(modified from: Kyunghyun Cho, 2016)

BENCHMARK: WMT'14 EN-DE

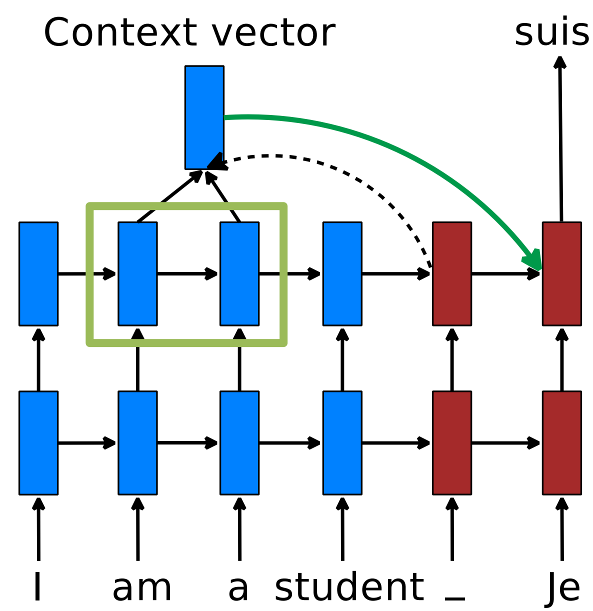


(modified from: Kyunghyun Cho)

Models for Global & Local Attention



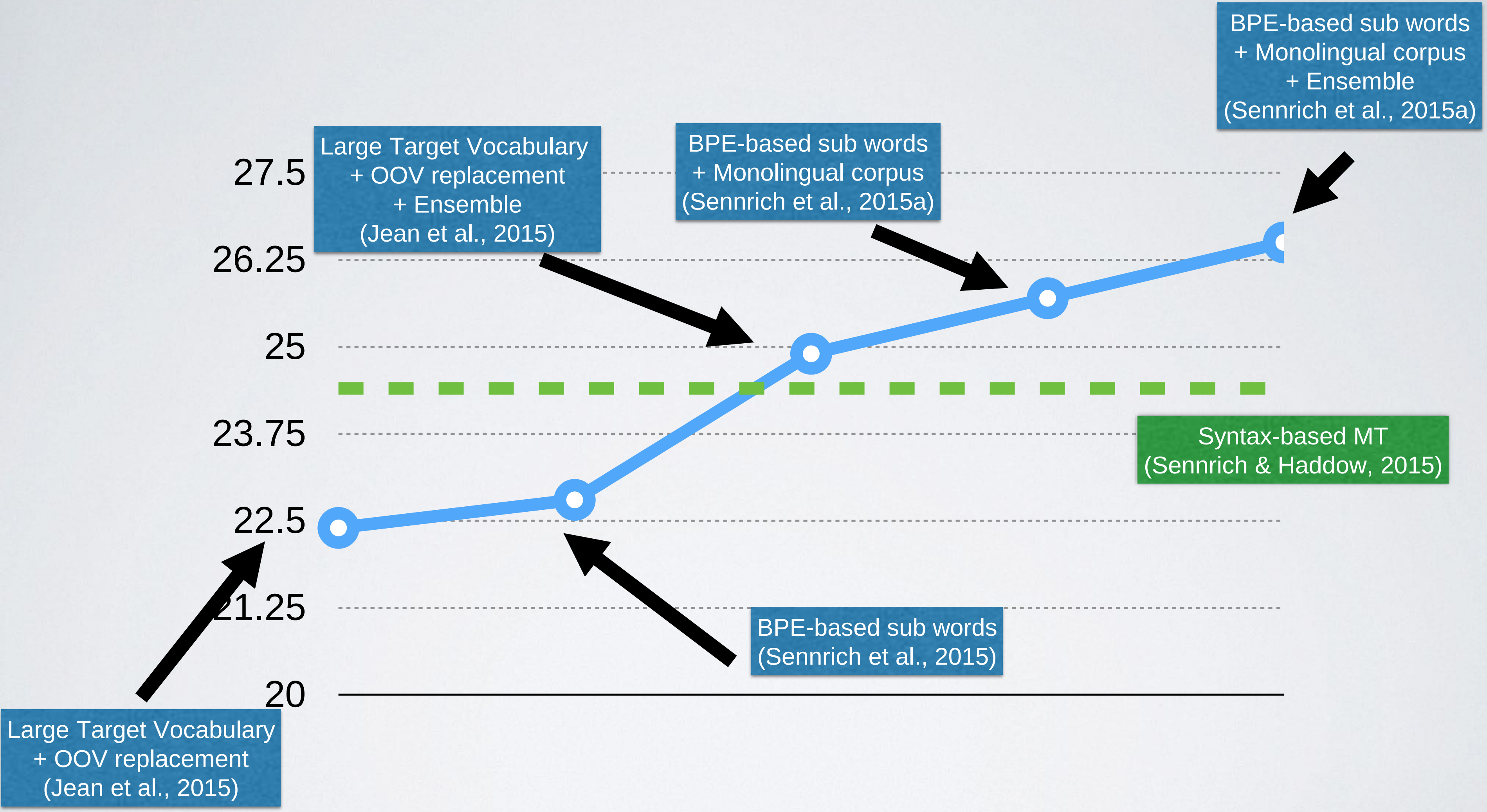
Global: ***all*** source states.



Local: ***subset*** of source states.

(Luong et al., 2015)

BENCHMARK: WMT'15 EN-DE



(modified from: Kyunghyun Cho)

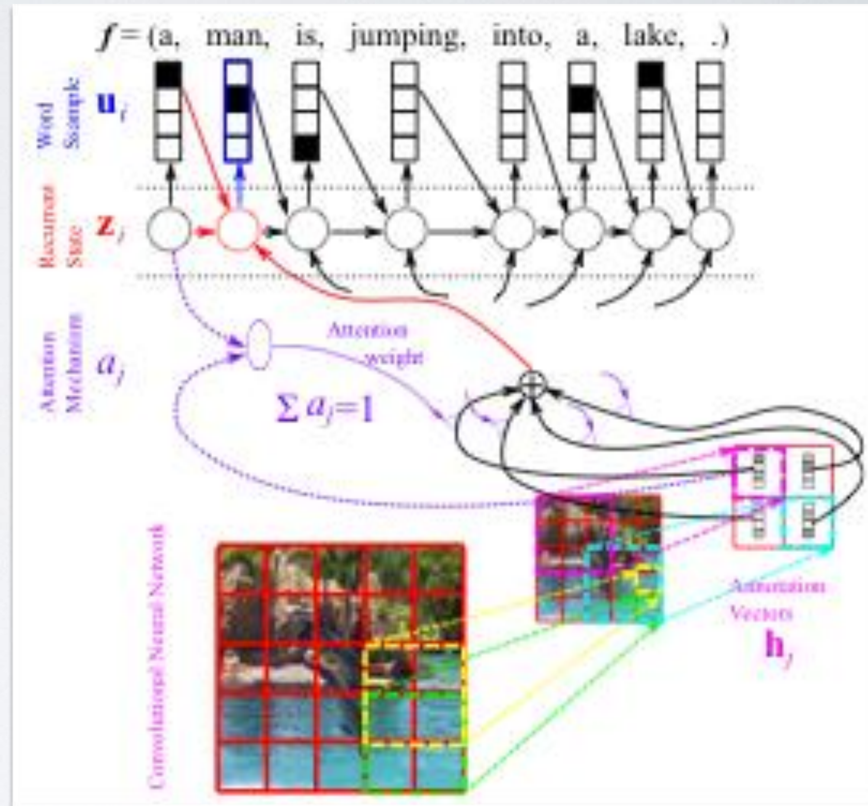
Same Attention Model applied to Image Captioning

Topics: Beyond Natural Languages
— Image Caption Generation

- *Conditional* language modelling

$p(\text{Two}, \text{dolphins}, \text{are}, \text{diving} \,|$  $) = ?$

- Encoder: convolutional network
  - Pretrained as a classifier or autoencoder
- Decoder: recurrent neural network
  - RNN Language model
  - With attention mechanism (Xu et al., 2015)

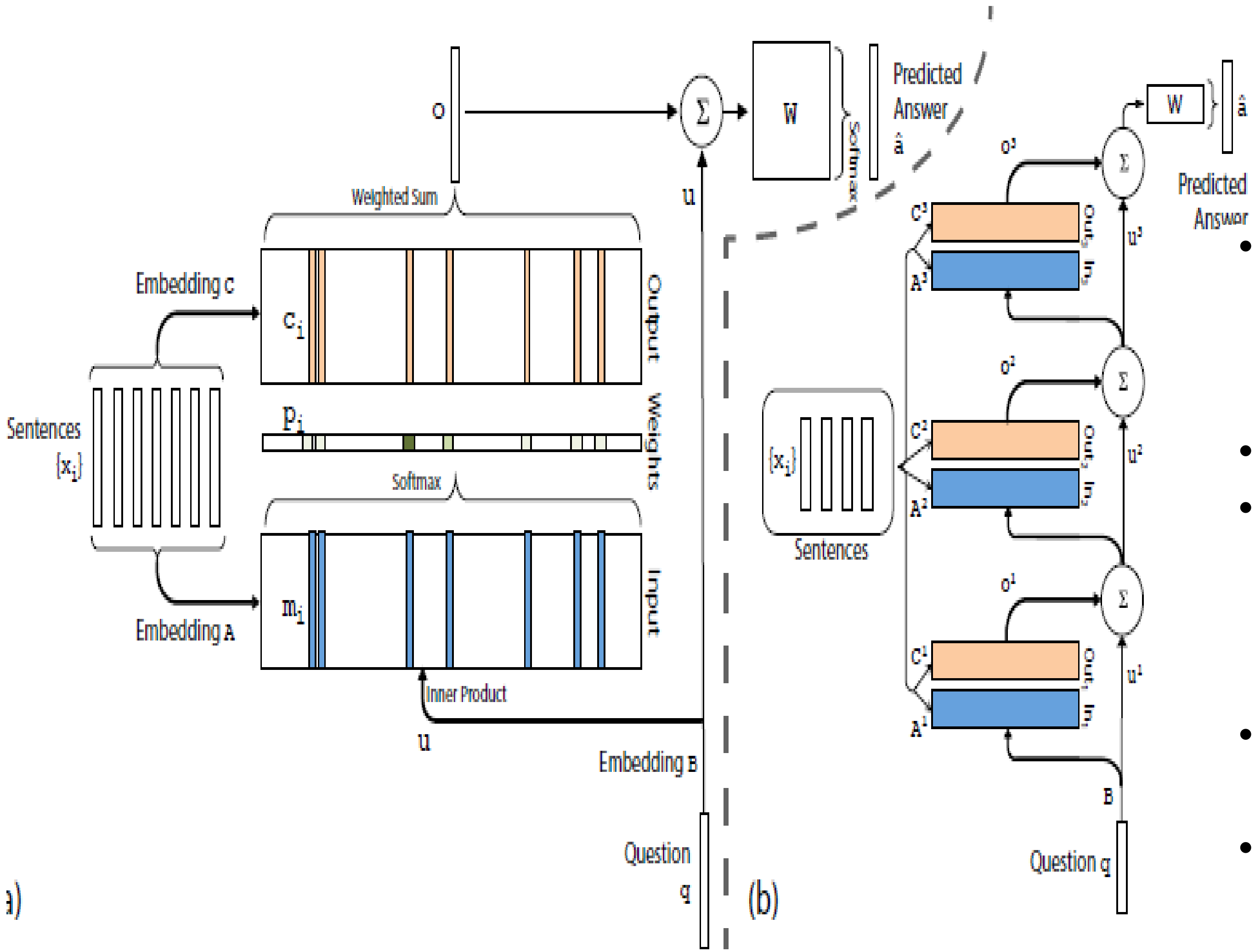


Deep Learning for Machine Cognition

--- Neural reasoning: memory network

--- Better neural reasoning: Tensor Product Representations (TPR) with structured knowledge representation

Memory Networks for Reasoning



- Rather than placing "attention" to part of a sentence, it can be placed to cognitive space with many sentences
- This allows "reasoning"
- Embedding input

$$m_i = Ax_i$$
$$c_i = Cx_i$$
$$u = Bq$$

- Attention over memories

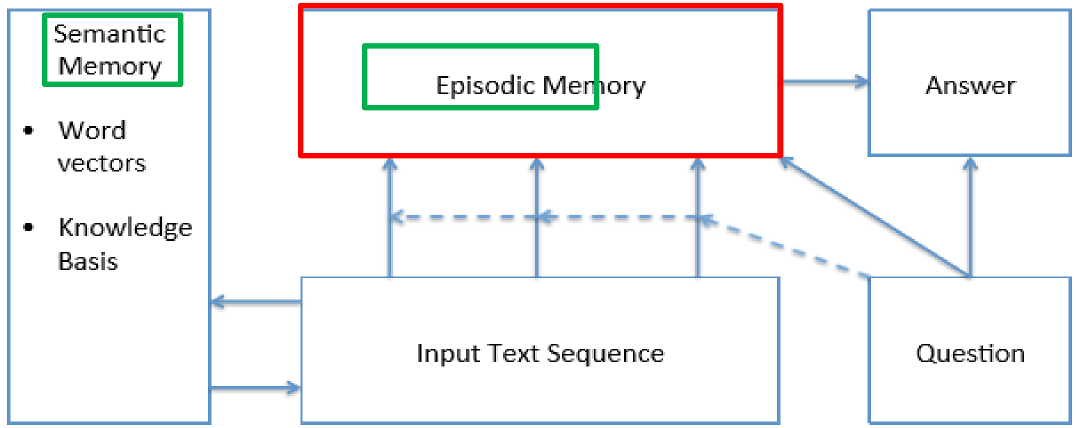
$$p_i = \mathrm{softmax}(u^T m_i)$$

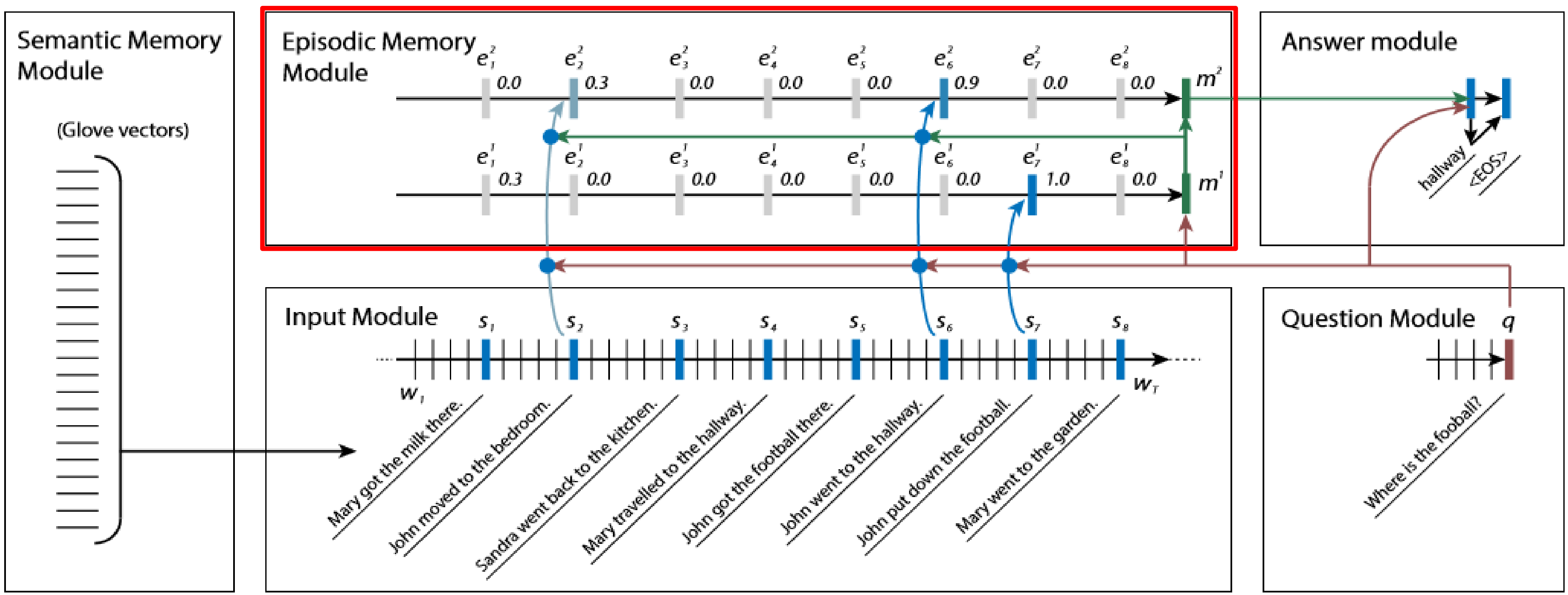
- Generating the final answer

$$o = \sum_i p_i c_i$$

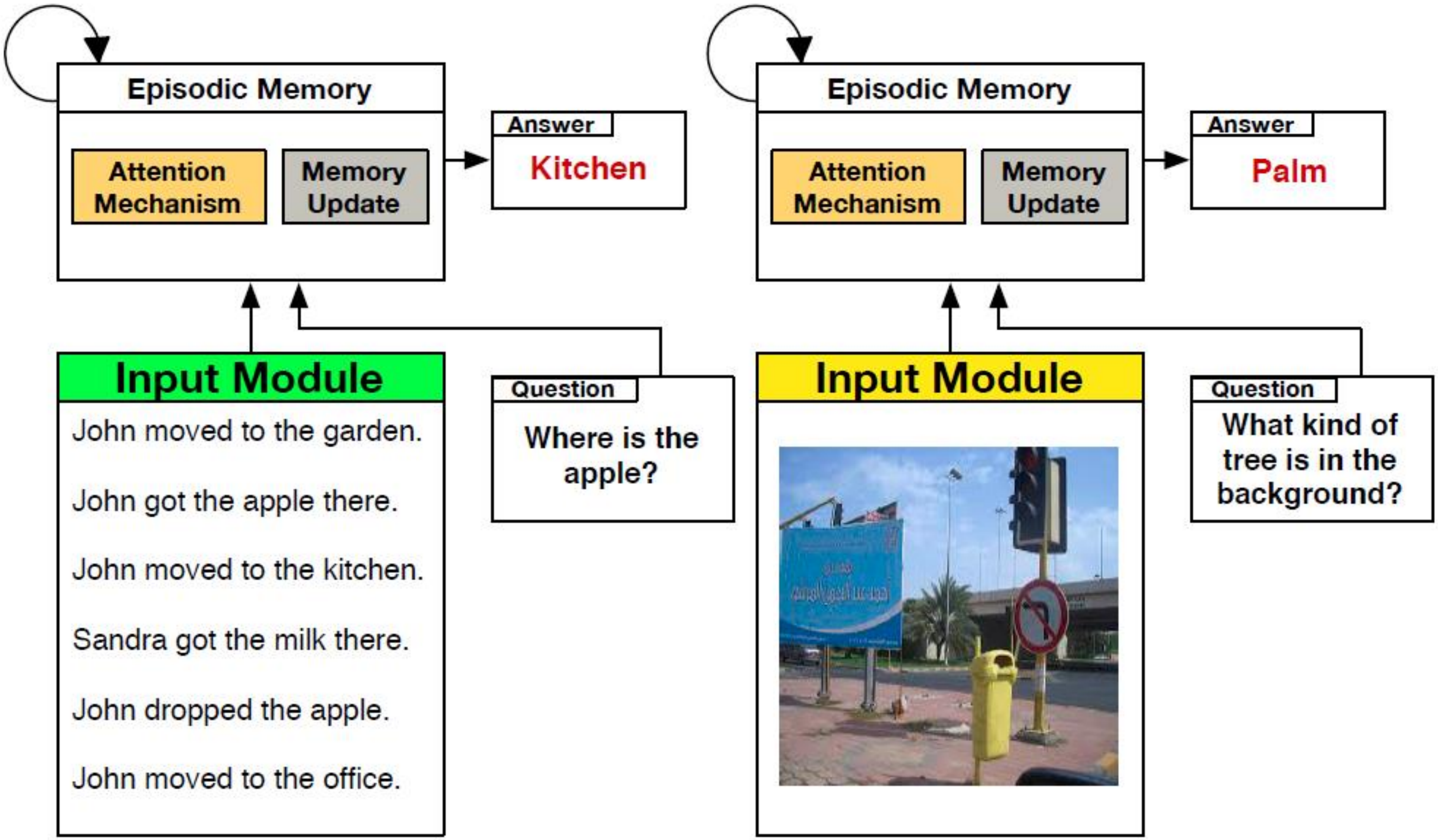
$$a = \mathrm{softmax}(W(o + u))$$

[Sukhbaatar, Szlm, Weston, Fergus: "End-to-end memory networks," NIPS, 2015]





[Kumar, Irsoy, …Socher: "Ask me anything: **Dynamic Memory Networks** for NLP," NIPS, 2015]



(a) Text Question-Answering (b) Visual Question-Answering

[Xiong, Merity, Socher: "Dynamic Memory Networks for visual & textual question answering,"ArXiv, Mar 4, 2016]
Reported in New York Times, Mar 6, 2016

TPR: Neural Representation of Structure

- Structured embedding vectors via tensor-product rep (TPR)

symbolic semantic parse tree (complex relation)

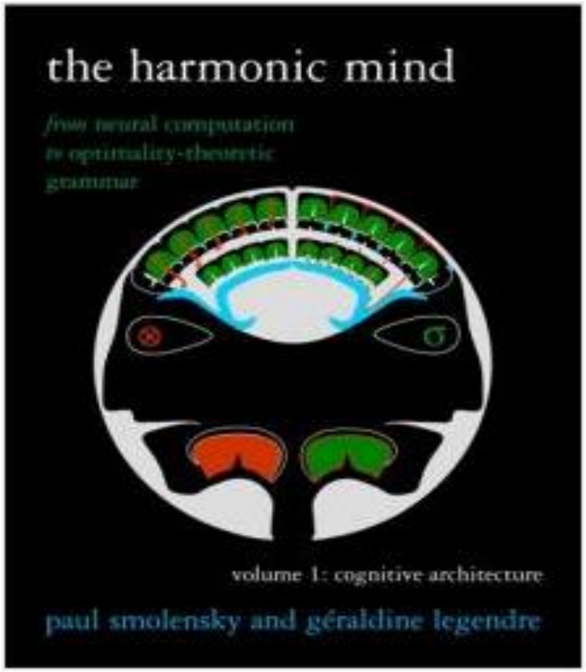
Then, reasoning in symbolic-space (traditional AI) can be beautifully carried out in the continuous-space in human cognitive and neural-net terms

Paul Smolensky & G. Legendre:

The Harmonic Mind, MIT Press, 2006

From Neural Computation to Optimality-Theoretic Grammar

Volume I: Cognitive Architecture; Volume 2: Linguistic Implications



Outline

- Deep learning for machine **perception**
  - **Speech**
  - Image
- Deep learning for machine **cognition**
  - Semantic modeling
  - Natural language
  - Multimodality
  - Reasoning, attention, memory (RAM)
  - Knowledge representation/management/exploitation
  - Optimal decision making (by deep reinforcement learning)
- Three hot areas/challenges of deep learning & AI research

Challenges for Future Research

1. Structured embedding for better reasoning: integrate symbolic/neural representations

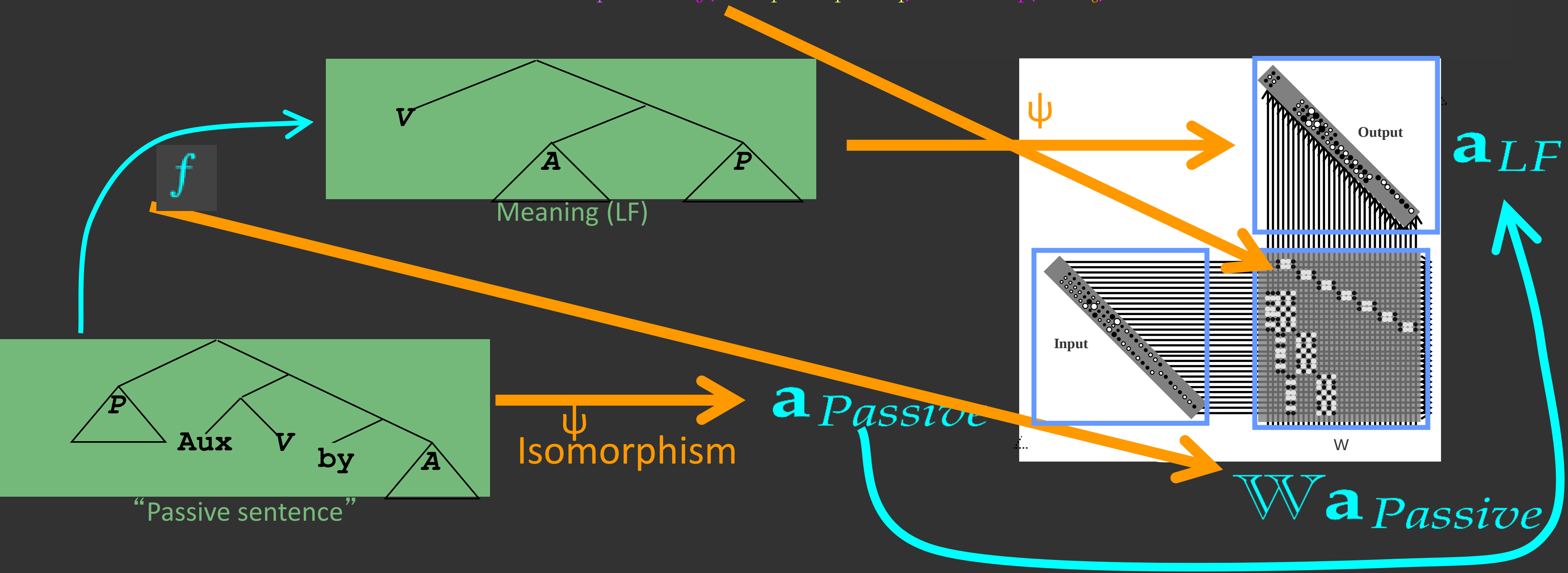
2. Integrate deep discriminative & generative/Bayesian models

3. Deep Unsupervised Learning

Few leaders are admired by George Bush $\xrightarrow{f}$ admire(George Bush, few leaders)

$$f(s) = \mathbf{cons}(\mathbf{ex}_1(\mathbf{ex}_0(\mathbf{ex}_1(s))), \mathbf{cons}(\mathbf{ex}_1(\mathbf{ex}_1(\mathbf{ex}_1(s))), \mathbf{ex}_0(s)))$$

$$W = W_{\mathrm{cons}_0}[W_{\mathrm{ex}_1} W_{\mathrm{ex}_0} W_{\mathrm{ex}_1}] + W_{\mathrm{cons}_1}[W_{\mathrm{cons}_0}(W_{\mathrm{ex}_1} W_{\mathrm{ex}_1} W_{\mathrm{ex}_1}) + W_{\mathrm{cons}_1}(W_{\mathrm{ex}_0})]$$



Slide from Paul Smolensky, 2015

Recurrent NN vs. Dynamic System

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}; \mathbf{W}_{hh}, \mathbf{W}_{xh}, \mathbf{x}_t)$$
$$\mathbf{y}_t = g(\mathbf{h}_t; \mathbf{W}_{hy})$$

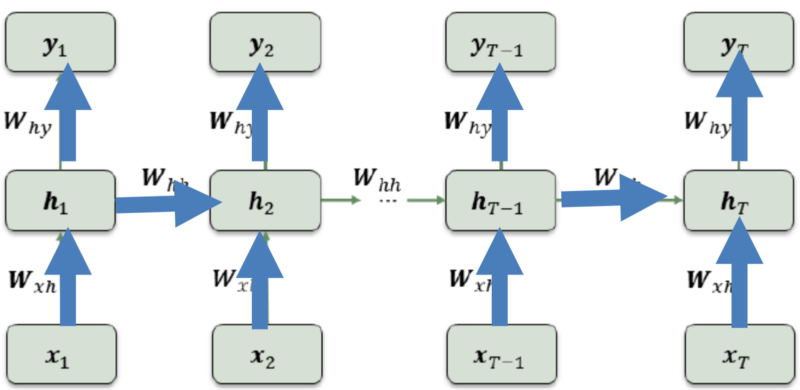
$$\mathbf{h}_t = q(\mathbf{h}_{t-1}; \mathbf{W}_{l_t}, \mathbf{t}_{l_t})$$
$$\mathbf{x}_t = r(\mathbf{h}_t, \Omega_{l_t})$$

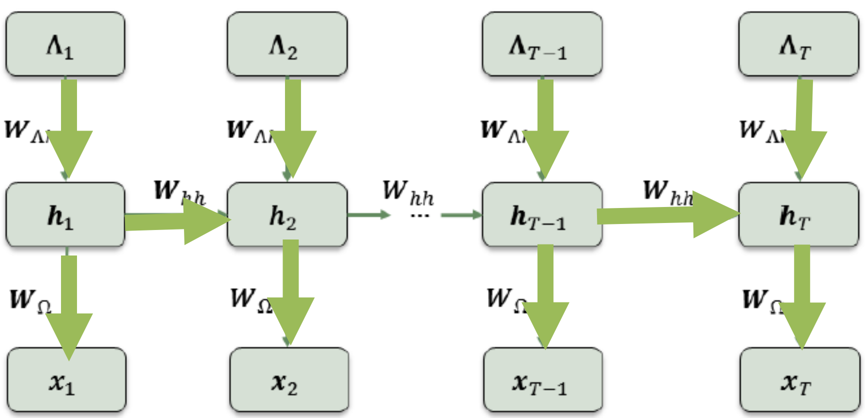
Parameterization:

- W_{hh}, W_{hy}, W_{xh}: all unstructured regular matrices

Parameterization:

- W_{hh} =M(ɣ$_l$); sparse system matrix
- W_{Ω} =(Ω_l); Gaussian-mix params; MLP
- $\Lambda = \boldsymbol{t}_l$

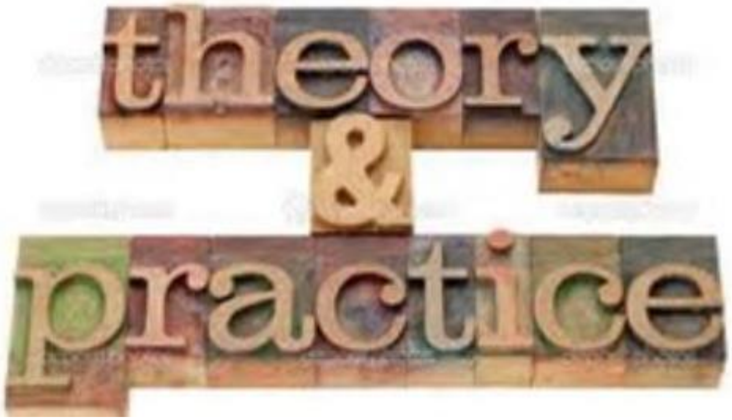




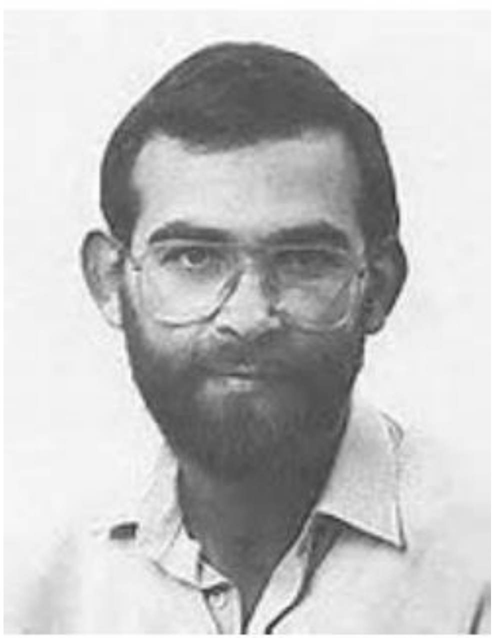
	Deep Discriminative NN	Deep Generative (Bayesian)
Structure	Graphical; info flow: **bottom-up**	Graphical; info flow: **top-down**
Incorp constraints & domain knowledge	Harder; less fine-grained	**Easier; more fine grained**
Semi/unsupervised	Hard or impossible	*Easier, at least possible*
Interpretability	Harder	**Easy** (generative "story" on data and hidden variables)
Representation	**Distributed**	Localist (mostly); can be distributed also
Inference/decode	Easy	Harder (but note **recent progress**)
Scalability/compute	**Easier (regular computes/GPU)**	Harder (but note **recent progress**)
Incorp. uncertainty	Hard	**Easy**
Empirical goal	Classification, feature learning, …	Classification (via Bayes rule), latent variable inference…
Terminology	Neurons, activation/gate functions, weights …	Random vars, stochastic "neurons", potential function, parameters …
Learning algorithm	A single, unchallenged, algorithm -- BackProp	A major focus of open research, many algorithms, & more to come
Evaluation	On a black-box score – end performance	On almost every intermediate quantity
Implementation	Hard (but increasingly easier)	Standardized but insights needed
Experiments	Massive, real data	Modest, often simulated data
Parameterization	Dense matrices	Sparse (often PDFs); can be dense

Deep Unsupervised Learning

- Unsupervised learning (UL) has recently been a very hot topic in deep learning
- Need to have a task to ground UL --- e.g. help improve prediction
- Examples of speech recognition and image captioning:
  - 3000 hrs of paired acoustics (X) & word label (Y)
  - How can we exploit 300,000+ hrs of speech acoustics with no paired labels?
- 4 sources of knowledge
  - Strong structure prior of “labels” Y (sequences)
  - Strong structure prior of input data X (conventional UL)
  - Dependency of x on y (generative modeling for embedding knowledge)
  - Dependency of y on x (state of the art systems w. supervised learning)



"In theory there is no difference between theory and practice, but in practice there is"



[Jan L. A. van de Snepscheut]

End
(of Chapter 1)

Thank you!
Q/A


深度学习是人工智能领域最近20年里最受瞩目的研究方向，近年来显著推动了语音、图像、自然语言理解、机器翻译，甚至是控制等众多技术方向的发展。本书原著作者微软研究院的邓力博士和俞栋博士是语音识别和深度学习方面的先驱之一，对于深度学习的进展有丰富的实践经验和深刻理解。这个学科处于快速进展之际，本书对当前的进展进行全景式系统性的梳理无疑是很有意义的，因为毕竟对于每一位读者，从这几年浩如烟海的论文中准确把握可以沉淀下来的进展是不容易的。谢磊教授受邓力博士之约在百忙之中对这本书进行翻译，对于深度学习在中国的发展具有重大意义。邓力博士和谢磊教授都是我所熟知的学者和好友。我相信，本书作为他们这次合作的成果，对于有志于了解和学习深度学习的中国读者会有极大的帮助。
—— 余凯
地平线机器人技术 创始人/CEO
前百度研究院常务副院长、深度学习实验室主任
地址：北京市百万庄大街22号
邮政编码：100037
电话服务
服务咨询热线：010-88361066
读者购书热线：010-68326294
010-88379203
网络服务
机工官网：www.cmpbook.com
机工官博：weibo.com/cmp1952
金书网：www.golden-book.com
教育服务网：www.cmpedu.com
封面无防伪标均为盗版
ISBN 978-7-111-52906-4
策划编辑◎王康
定价：24.00元
深度学习 方法及应用
谢磊 译
机械工业出版社
CHINA MACHINE PRESS
大/数/据/丛/书
深度学习
方法及应用
[美] 邓力（Li Deng） 俞栋（Dong Yu） 著
谢磊 译
机械工业出版社
CHINA MACHINE PRESS

Broadview®
www.broadview.com.cn

博文视点 · IT出版旗舰品牌
技术凝聚实力 · 专业创新出版

解析深度学习
语音识别实践

语音识别在最近几年里取得了长足的进展，这些进展主要源于在语音识别中引入了深度学习技术。本书作者俞栋博士和邓力博士长期致力于语音识别技术的研究，有丰富的理论和实践经验。他们最先将深度学习技术与传统语音识别技术相结合，成功地大幅降低了大词汇量语音识别系统的错误率。本书译者俞凯博士和钱彦旻博士也是语音识别领域的专家，还是这一进展的积极推动者。他们合作的这本中文译本是第一部系统性地介绍基于深度学习的语音识别技术的专著。在本书中，俞栋博士和邓力博士以第一手资料详细介绍了这一技术发生的背景、发展过程、理论根据、关键技术细节，以及思维方式。本书对所有从事语音识别研究或想了解语音识别技术最新进展和发展方向的读者都是很好的参考书。对想将深度学习技术应用到诸如视觉和文本处理等其他领域的读者，本书也有很强的借鉴意义。

微软公司杰出工程师兼首席语音科学家　黄学东博士

Xuedong Huang

博文视点Broadview

新浪微博 weibo.com
@博文视点Broadview

本书将由电子工业出版社于2016年5月出版，欢迎有兴趣的读者联系我们采购，先睹为快！
联系人：杜老师
联系电话：010-8825 4888
邮箱：duca@phei.com.cn
网址：www.phei.com.cn

策划编辑：刘　皎
责任编辑：李利健
封面设计：吴海燕

Broadview®
www.broadview.com.cn

Signals and Communication Technology
Automatic Speech Recognition
A Deep Learning Approach

解析深度学习
语音识别实践

【美】俞栋 邓力 著
俞凯 钱彦旻 等译

Tensor Product Rep for reasoning

- Facebook's reasoning task

Category 4. Two Argument Relation
01: The office is north of the kitchen.
02: The garden is south of the kitchen.
03: What is north of the kitchen? office 1

01: The kitchen is west of the garden.
02: The hallway is west of the kitchen.
03: What is the garden east of? kitchen 1

Category 17: Positional Reasoning
01: The triangle is above the pink rectangle.
02: The blue square is to the left of the triangle.
03: Is the pink rectangle to the right of the blue square? yes 1 2

01: The red sphere is below the yellow square.
02: The red sphere is above the blue square.
03: Is the blue square below the yellow square? yes 2 1

Category 19: Path Finding
01: The bedroom is south of the hallway.
02: The bathroom is east of the office.
03: The kitchen is west of the garden.
04: The garden is south of the office.
05: The office is south of the bedroom.
06: How do you go from the garden to the bedroom? n,n 4 5

arXiv.org > cs > arXiv:1511.06426

Computer Science > Computation and Language

Reasoning in Vector Space: An Exploratory Study of Question Answering

Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, Li Deng, Paul Smolensky

(Submitted on 19 Nov 2015) Accepted to ICLR, May 2016

Structured Knowledge Representation & Reasoning via TPR

#	Statements/Questions	Relational Translations/Answers	Encodings/Clues	
1	Mary went to the kitchen.	*Mary belongs to the kitchen (from nowhere).*	mk^T	$m(k \circ n)^T$
3	Mary got the football there.	*The football belongs to Mary.*	fm^T	fm^T
4	Mary travelled to the garden.	*Mary belongs to the garden (from the kitchen).*	mg^T	$m(g \circ k)^T$
5	Where is the football?	garden	3, 4	
9	Mary dropped the football.	*The football belongs to where Mary belongs to.*	fg^T	fg^T
10	Mary journeyed to the kitchen.	*Mary belongs to the kitchen (from the garden).*	mk^T	$m(k \circ g)^T$
11	Where is the football?	garden	9, 4	

- Given *containee-container* relationship
- Encode all entities (e.g., actors (*m*ary), objects (*f*ootball), and locations (*n*owhere, *k*itchen, *g*arden)) by vectors
- Encode each statement by a matrix via **binding** (tensor product of two vectors), mk^T
- Reasoning (transitivity) by **matrix multiplication**, $(fm^T) \cdot (mg^T) = f(m^T \cdot m)g^T = fg^T$
- Generate answer (e.g., where is the *f*ootball in #5) via **unbinding** (inner product)
  a. Left-multiply by f^T all statements prior to the current time. (Yields $f^T \cdot mk^T$, $f^T \cdot fg^T$)
  b. Pick the most recent container where 2-norms of the multiplications in (a) are approximately 1.0. (Yields g^T .)

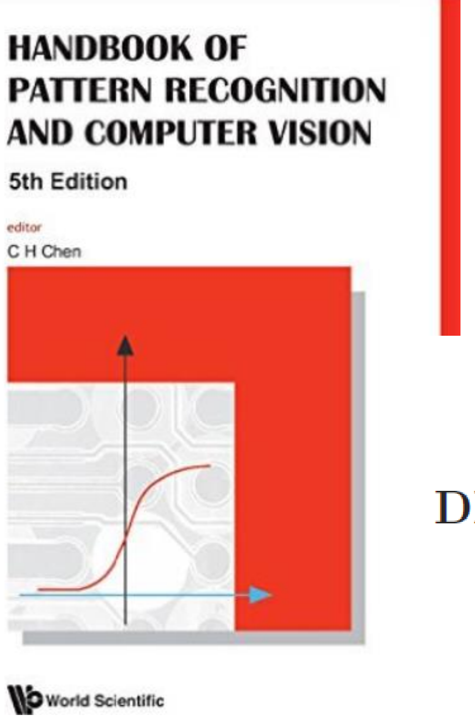
TPR Results on FB's bAbI task

Type	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Accuracy	100%	100%	100%	100%	99.3%	100%	96.9%	96.5%	100%	99%
Model	MNN	MNN	MNN	MNN	DMN	MNN	DMN	DMN	DMN	SSVM
Type	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**	**C20**
Accuracy	100%	100%	100%	100%	100%	100%	72%	95%	36%	100%
Model	MNN	MNN	MNN	DMN	MNN	MNN	Multitask	MNN	MNN	MNN

Table 4: Best accuracies for each category and the best model which achieved the best accuracy. MNN indicates Strongly-Supervised MemNN trained with the clue numbers, and DMN indicates Dynamic MemNN, and SSVM indicates Structured SVM with the coreference resolution and SRL features. Multitask indicates multitask training.

Type	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Training	100%	100%	100%	100%	99.8%	100%	100%	100%	100%	100%
Test	100%	100%	100%	100%	99.8%	100%	100%	100%	100%	100%
Type	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**	**C20**
Training	100%	100%	100%	100%	100%	99.4%	100%	100%	100%	100%
Test	100%	100%	100%	100%	100%	99.5%	100%	100%	100%	100%

Table 5: Accuracies on training and test data on our models. We achieve near-perfect accuracies in almost every category including positional reasoning and path finding.



CHAPTER 1.2

DEEP DISCRIMINATIVE AND GENERATIVE MODELS FOR PATTERN RECOGNITION

Li Deng[1] and Navdeep Jaitly[2]

[1] *Microsoft Research, One Microsoft Way, Redmond, WA 98052*
[2] *Google Research, 1600 Amphitheatre Parkway, Mountain View, CA 94043*
E-mails: deng@microsoft.com; ndjaitly@google.com

In this chapter we describe deep generative and discriminative models as they have been applied to speech recognition and related pattern recognition problems. The former models describe the distribution of data or the joint distribution of data and the corresponding targets, whereas the latter models describe the distribution of targets conditioned on data. Both models are characterized as being 'deep' as they use layers of latent or hidden variables. Understanding