dawin

DEEP LEARNING FOR SPEECH RECOGNITION: KEY INSIGHTS

dawin

Sébastien Bratières

Speech Evangelist

dawin gmbh

dawin


HAMBURG
dawin®
BERLIN
GERMANY
dawin®
HOOGSTRATEN
BELGIUM
COLOGNE
TROISDORF
dawin®
FRANKFURT
dawin®
LUDWIGSBURG
STUTTGART
VIENNA
dawin®
AUSTRIA
ITALY
ROME
dawin®

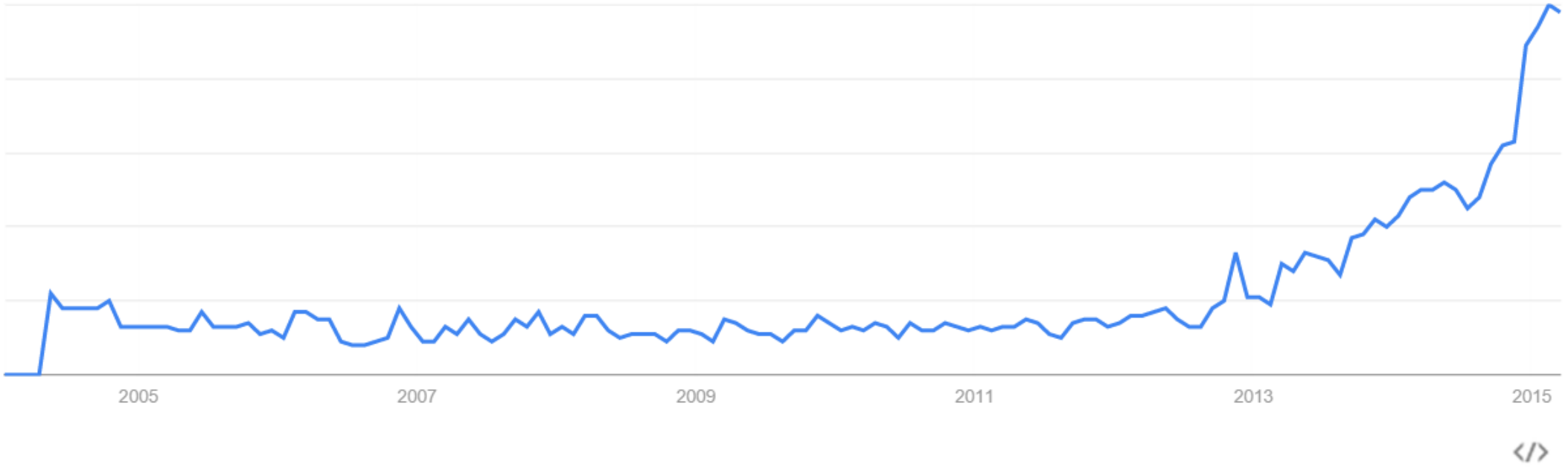
dawin

Our goal is to make business and industry processes more intuitive and humane, by providing workers with a voice interface to IT systems.

dawin

WHAT'S THE MATTER ?

dawin



MIT Technology Review

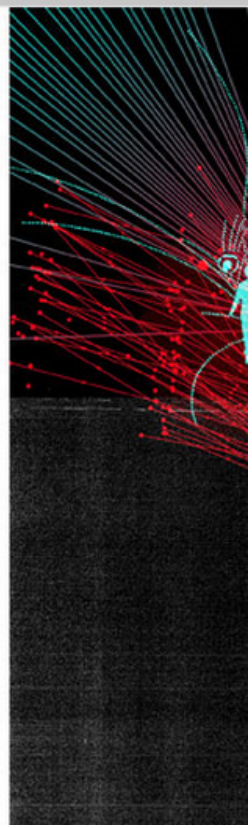
10 BREAKTHROUGH TECHNOLOGIES 2013

Intr

dawin

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.



dawin

TECH 12/18/2014 @ 9:00AM | 47,498 views

Baidu Announces Breakthrough In Speech Recognition, Claiming To Top Google And Apple

+ Comment Now + Follow Comments

When artificial-intelligence guru Andrew Ng joined Chinese Internet pioneer Baidu last May as chief scientist, he was a little cagey about what he and his team might work on at a newly opened lab in Sunnyvale, Calif. But he couldn't help revealing better speech recognition as a key area of interest in the age of the smartphone.

Today, Baidu, often called China's Google GOOGL -2.36%,

Skype will soon get real-time speech translation based on deep learning

by David Meyer | May. 28, 2014 - 12:40 AM PDT

4 Comments

Microsoft will by the end of 2014 start offering on-the-fly language translation within Skype, firstly in a Windows 8 beta app and then hopefully as a full commercial product within the coming two and a half years

dawin

PUBLIC ATTENTION:

2012 “four groups” signal processing magazine article

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

[TABLE 2] COMPARING FIVE DIFFERENT DBN-DNN ACOUSTIC MODELS WITH TWO STRONG GMM-HMM BASELINE SYSTEMS THAT ARE DISCRIMINATIVELY TRAINED. SI TRAINING ON 309 H OF DATA AND SINGLE-PASS DECODING WERE USED FOR ALL MODELS EXCEPT FOR THE GMM-HMM SYSTEM SHOWN ON THE LAST ROW WHICH USED SA TRAINING WITH 2,000 H OF DATA AND MULTIPASS DECODING INCLUDING HYPOTHESES COMBINATION. IN THE TABLE, "40 MIX" MEANS A MIXTURE OF 40 GAUSSIANS PER HMM STATE AND "15.2 NZ" MEANS 15.2 MILLION, NONZERO WEIGHTS. WERs IN % ARE SHOWN FOR TWO SEPARATE TEST SETS, HUB500-SWB AND RT03S-FSH.

		WER	
MODELING TECHNIQUE	**#PARAMS [10^6]**	**HUB5'00-SWB**	**RT03S-FSH**
GMM, 40 MIX DT 309H SI	29.4	23.6	27.4
NN 1 HIDDEN-LAYER × 4,634 UNITS	43.6	26.0	29.4
+ 2 × 5 NEIGHBORING FRAMES	45.1	22.4	25.7
DBN-DNN 7 HIDDEN LAYERS × 2,048 UNITS	45.1	17.1	19.6
+ UPDATED STATE ALIGNMENT	45.1	16.4	18.6
+ SPARSIFICATION	15.2 NZ	16.1	18.5
GMM 72 MIX DT 2000H SA	102.4	17.1	18.6

COMMON TRAINING DATA

dawin

[TABLE 2] COMPARING FIVE DIFFERENT DBN-DNN ACOUSTIC MODELS WITH TWO STRONG GMM-HMM BASELINE SYSTEMS THAT ARE DISCRIMINATIVELY TRAINED. SI TRAINING ON 309 H OF DATA AND SINGLE-PASS DECODING WERE USED FOR ALL MODELS EXCEPT FOR THE GMM-HMM SYSTEM SHOWN ON THE LAST ROW WHICH USED SA TRAINING WITH 2,000 H OF DATA AND MULTIPASS DECODING INCLUDING HYPOTHESES COMBINATION. IN THE TABLE, "40 MIX" MEANS A MIXTURE OF 40 GAUSSIANS PER HMM STATE AND "15.2 NZ" MEANS 15.2 MILLION, NONZERO WEIGHTS. WERs IN % ARE SHOWN FOR TWO SEPARATE TEST SETS, HUB500-SWB AND RT03S-FSH.

WORD ERROR RATE: LOWER = BETTER

2 TEST SETS

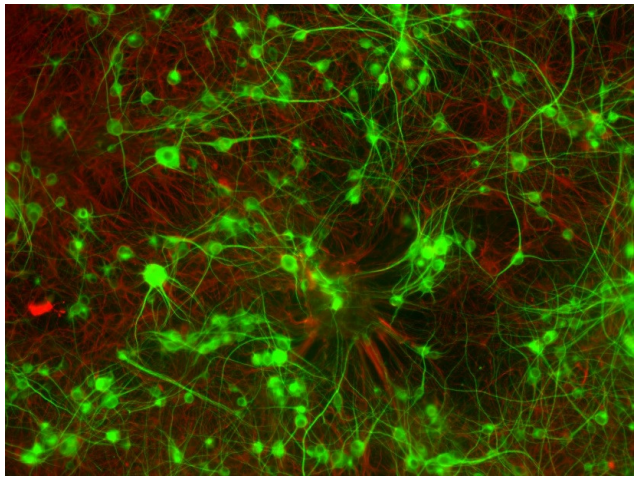
		WER	
MODELING TECHNIQUE	**#PARAMS [10^6]**	**HUB5'00-SWB**	**RT03S-FSH**
GMM 40 MIX DT 309H SI	29.4	23.6	27.4
NN 1 HIDDEN-LAYER × 4,634 UNITS	43.6	26.0	29.4
+ 2 × 5 NEIGHBORING FRAMES	45.1	22.4	25.7
DBN-DNN 7 HIDDEN LAYERS × 2,048 UNITS	45.1	17.1	19.6
+ UPDATED STATE ALIGNMENT	45.1	16.4	18.6
+ SPARSIFICATION	15.2 NZ	16.1	18.5
GMM 72 MIX DT 2000H SA	102.4	17.1	18.6

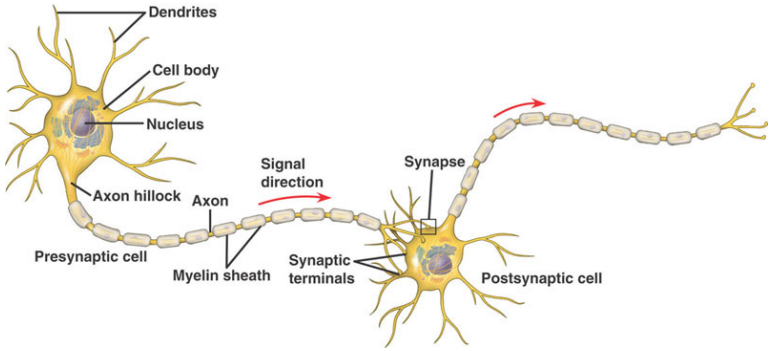
TRADIT'L MODEL

DEEP LEARNING

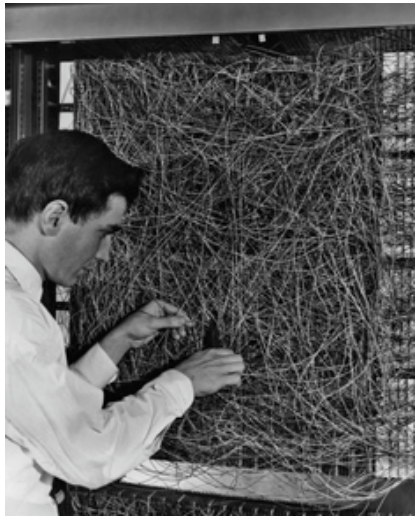
SEES MORE TRAINING DATA

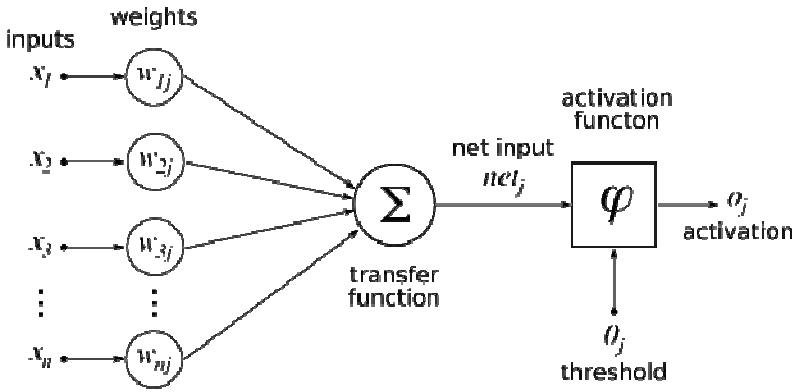
WHAT IS DEEP LEARNING?



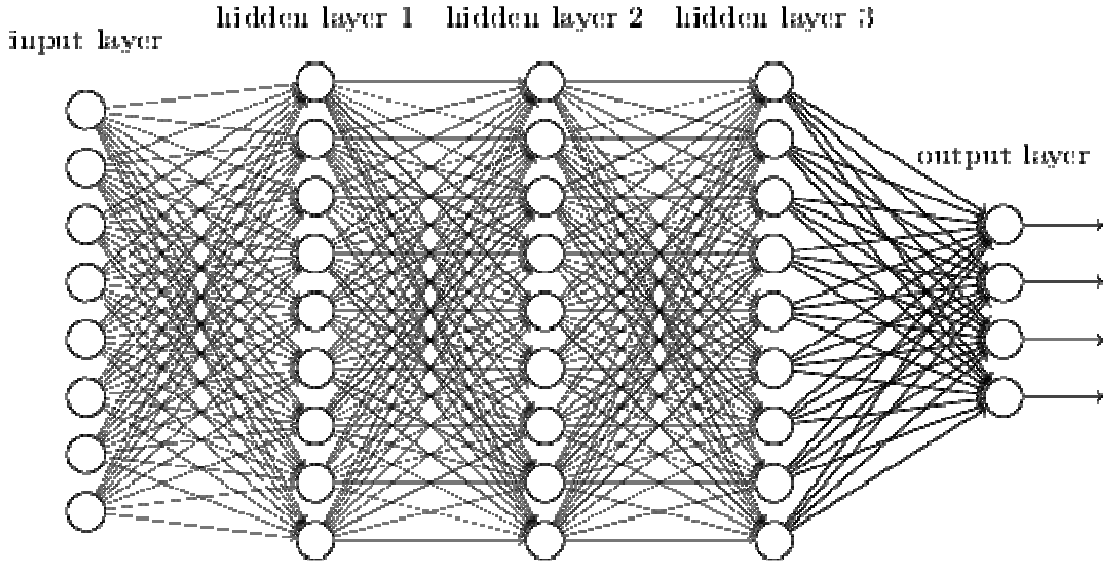


Deep learning is a class of machine learning algorithms and models, for example artificial neural networks, loosely inspired from brain biology, applicable to a wide range of tasks.





The models typically consist of a large number of inter-***connected*** simple units ("neurons"), each of which produces a single output, which is propagated to many neighbour units. The input to each unit combines weighted signals from a large number of neighbour units.

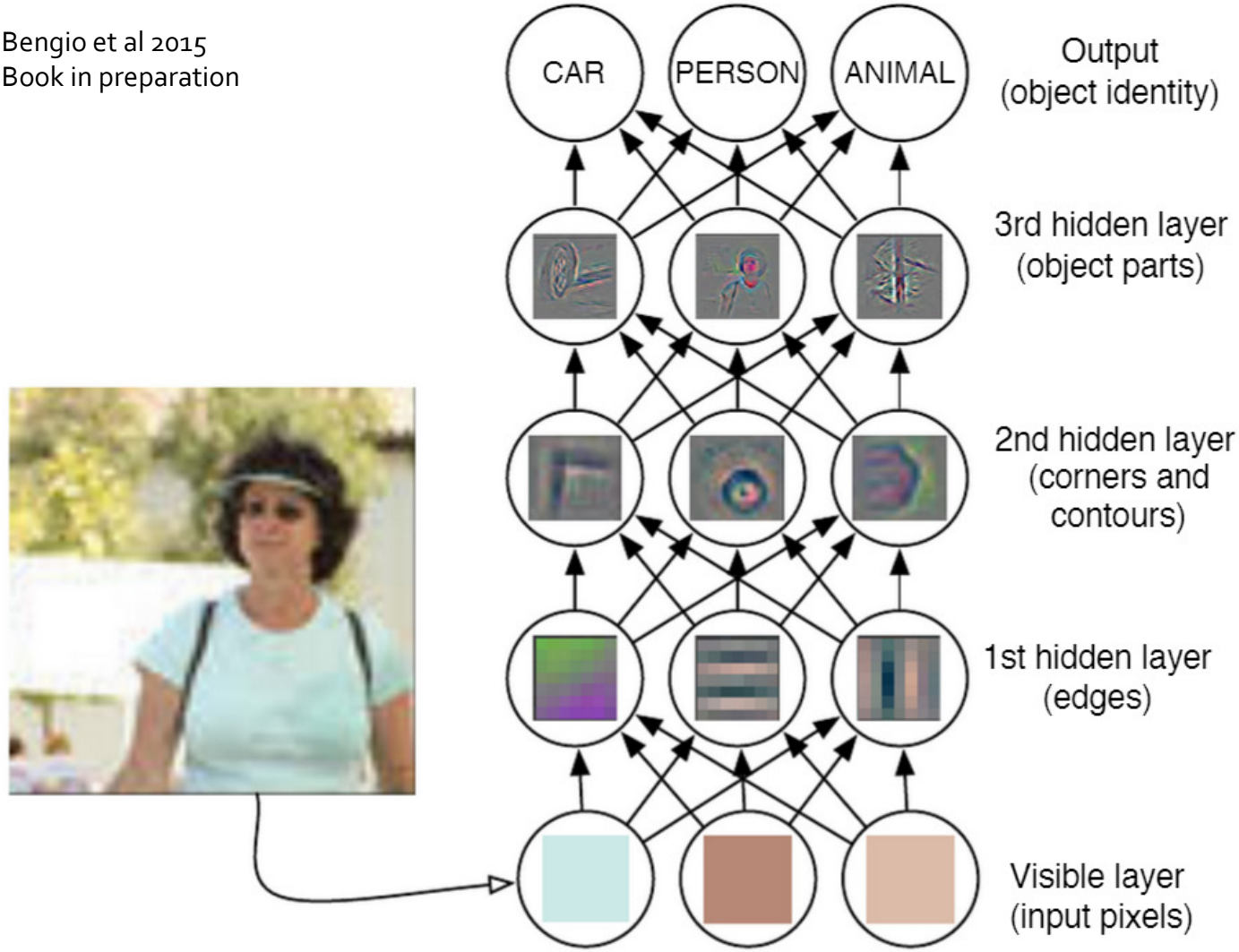


The units can be arranged in layers, and models with many layers are called *deep*.

The upper layers encode high levels of abstractions, while the lower layers represent the data at a more concrete level. Connections between layers represent how low-level features compose into high-level representations.

dawin

Bengio et al 2015
Book in preparation



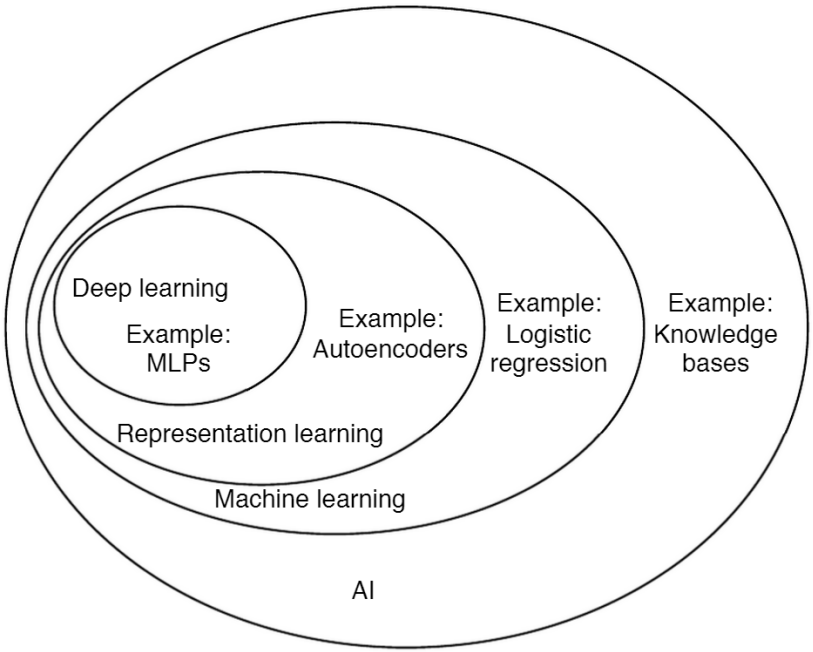


Figure 1.2: A Venn diagram showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI. Each section of the Venn diagram includes an example of an AI technology.

Deep learning is a class of machine learning algorithms and models, for example artificial neural networks, loosely inspired from brain biology, applicable to a wide range of tasks.

The models typically consist of a large number of interconnected simple units, each of which produces a single output, which is propagated to many neighbour units. The input to each unit combines weighted signals from a large number of neighbour units.

The units can be arranged in layers, and models with many layers are called *deep*.

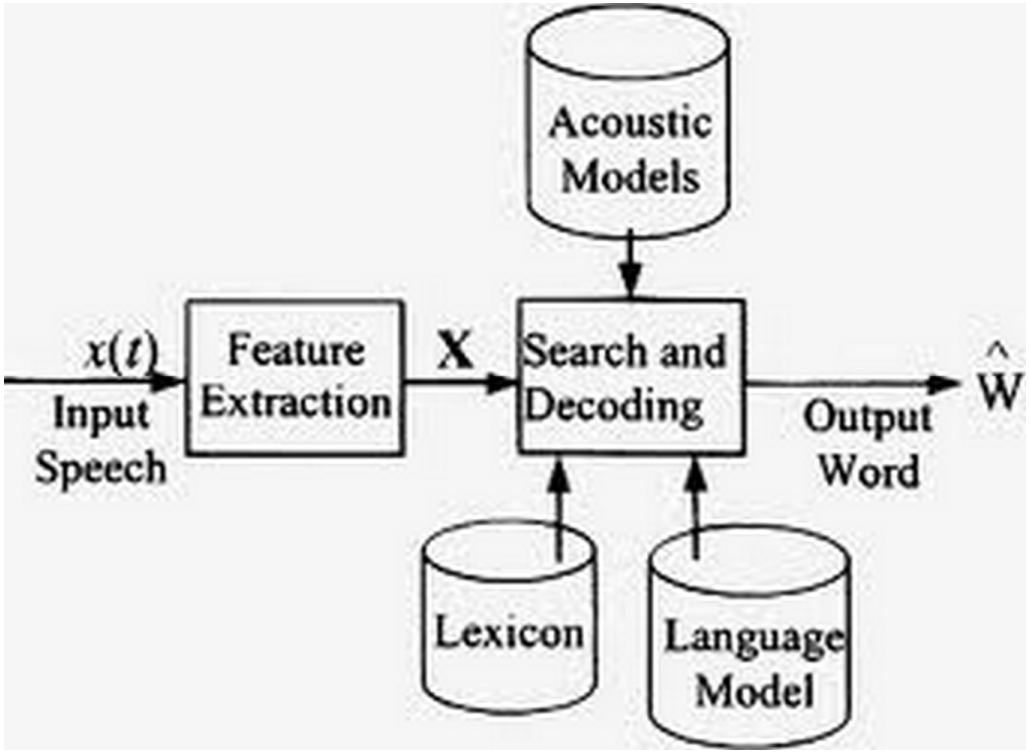
The upper layers encode high levels of abstractions, while the lower layers represent the data at a more concrete level. Connections between layers represent how low-level features compose into high-level representations.

dawin

ASR ARCHITECTURE, SIMPLY PUT…

SPEECH RECOGNITION ARCHITECTURE

dawin



dawin

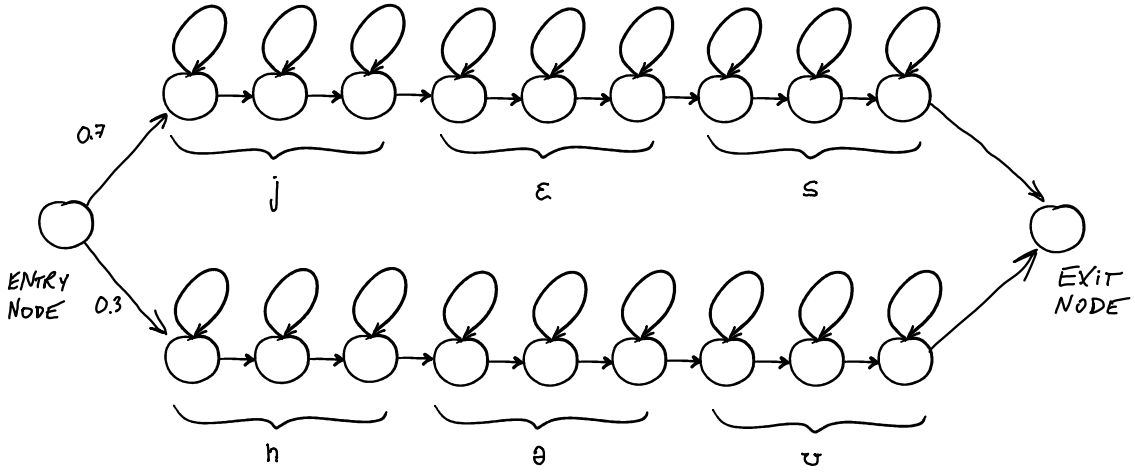
IMPORTANT BUILDING BLOCKS

ACOUSTIC MODEL

THE CAT SITS ON THE

LANGUAGE MODEL

MORE
CREEK
CHAIR
DOOR
SEVEN
BEFORE
MAT
CITY
FINEST
TOGETHER



HIDDEN MARKOV MODEL

WHAT MADE THE BREAKTHROUGH POSSIBLE?

dawin

- computing power: fast, cheap general-purpose graphical processing units (GPGPU)
  - for video-games: graphics rendering, game physics
  - great for matrix multiplication
- scientific progress on neural network training, optimization, architectures
- the right amount (lots) of data

SUCCESSES IN MULTITASK/ TRANSFER LEARNING

- learning to caption [Vinyals et al 2014]
- mixed-bandwidth, multilingual modelling [Deng et al 2013]

DNN ARE AMENABLE TO MULTITASK LEARNING

dawin

Show and Tell: A Neural Image Caption Generator

Oriol Vinyals
Google
vinyals@google.com

Alexander Toshev
Google
toshev@google.com

Samy Bengio
Google
bengio@google.com

Dumitru Erhan
Google
dumitru@google.com

Abstract

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descrip-

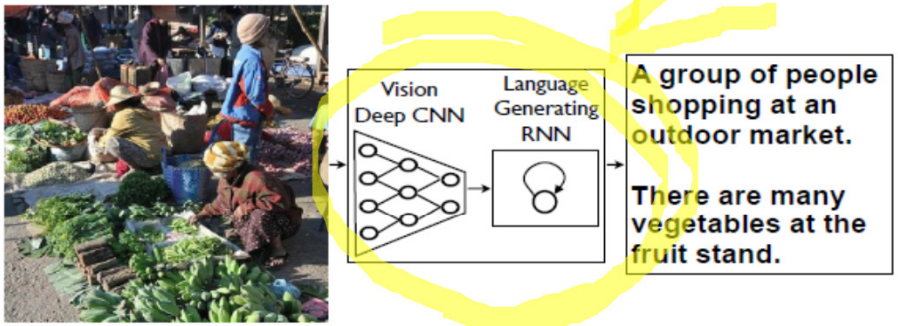


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

RECENT ADVANCES IN DEEP LEARNING FOR SPEECH RESEARCH AT MICROSOFT

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

dawin

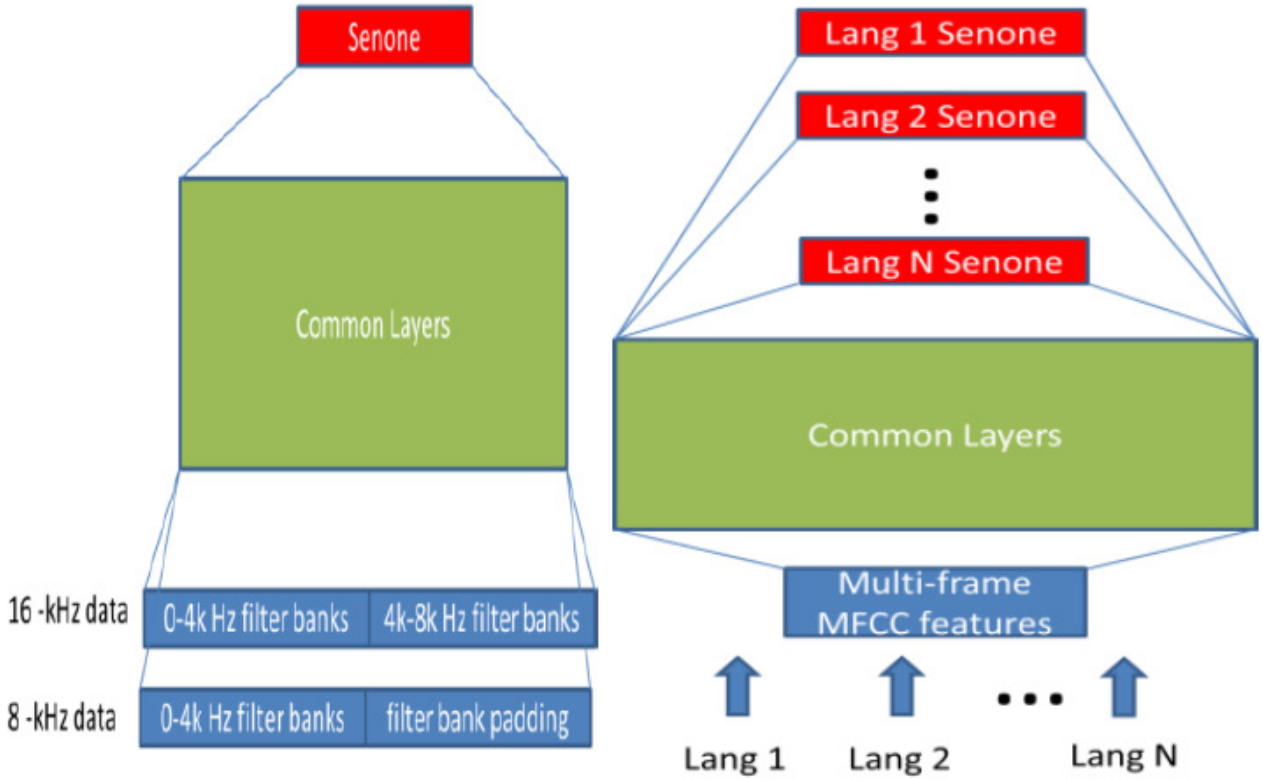


***Figure 1**: a) left: DNN training/testing with mixed-band acoustic data with16-kHz and 8-kHz sampling rates; b) right: Illustrative architecture for multilingual DNN*

DL MAY CHANGE ASR RADICALLY

- replaced traditional acoustic models (GMM) with neural equivalents (ANN)
- replacing the backbone of traditional ASR: HMMs
- replacing traditional feature extraction

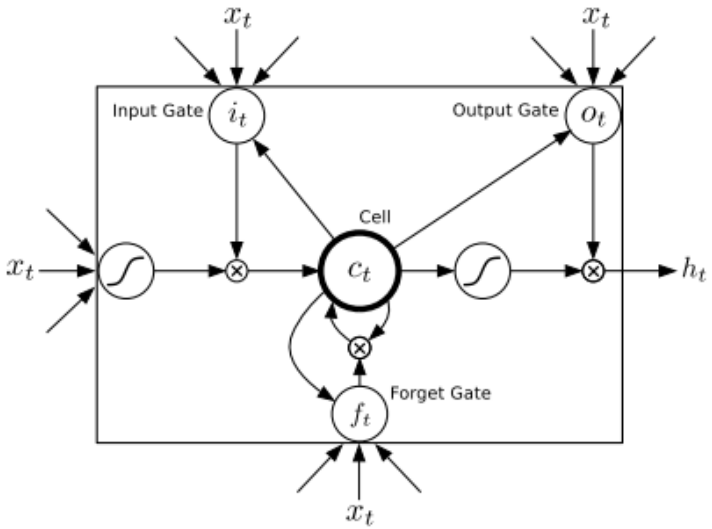
FROM THE GROUND UP: LOOK MA, NO HMMS !

dawin

Towards End-to-End Speech Recognition with Recurrent Neural Networks

Alex Graves GRAVES@CS.TORONTO.EDU
Google DeepMind, London, United Kingdom

Navdeep Jaitly NDJAITLY@CS.TORONTO.EDU
Department of Computer Science, University of Toronto, Canada



First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs

Awni Y. Hannun
Computer Science Department
Stanford University
Stanford, CA 94305
awni@cs.stanford.edu

Andrew L. Maas
Computer Science Department
Stanford University
Stanford, CA 94305
amaas@cs.stanford.edu

Daniel Jurafsky
Linguistics Department
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305
ang@cs.stanford.edu

NO MORE ASR FRONTEND: ASR FROM WAVEFORMS!

dawin

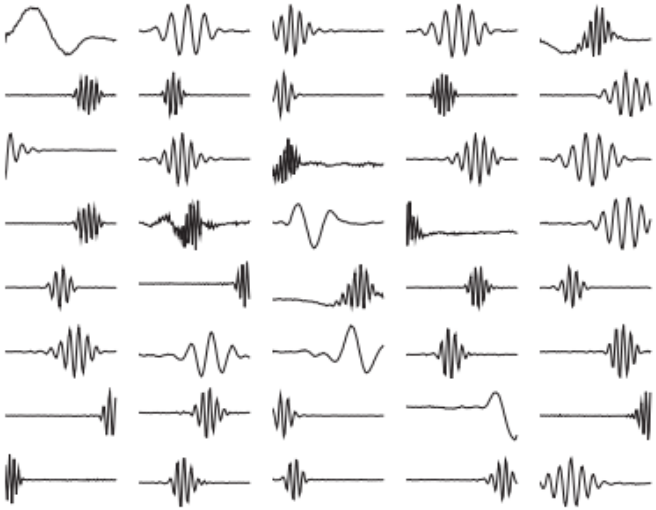
LEARNING A BETTER REPRESENTATION OF SPEECH SOUND WAVES USING RESTRICTED BOLTZMANN MACHINES

Navdeep Jaitly, Geoffrey Hinton

Department of Computer Science, University of Toronto, Toronto, M5S 3G4, Canada

ABSTRACT

State of the art speech recognition systems rely on pre-processed speech features such as Mel cepstrum or linear predictive coding coefficients that collapse high dimensional speech sound waves into low dimensional encodings. While these have been successfully applied in speech recognition systems, such low dimensional encodings may lose some relevant information and express other information in a way that makes it difficult to use for discrimination. Higher dimensional encodings could both improve performance in recognition tasks, and also be applied to speech synthesis by better modeling the statistical structure of the sound waves. In this paper we present a novel approach for modeling speech sound waves using a Restricted Boltzmann machine (RBM) with a novel type of hidden variable and we report initial results demonstrating phoneme recognition performance better



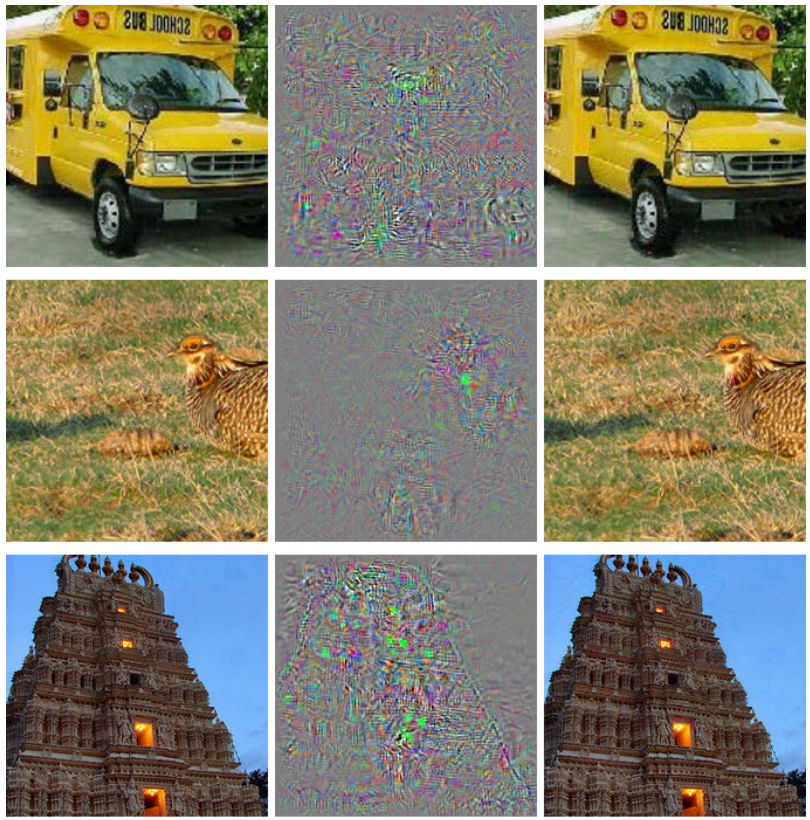
Jaitly & Hinton 2011 ICASSP

LIMITATIONS

- unclear what a NN actually learns: weird edge cases
- further scaling: need new ideas
- no consolidation of techniques

LIMITATIONS:
IT'S EASY TO DERAIL A CNN WITH IMPERCEPTIBLE CHANGES

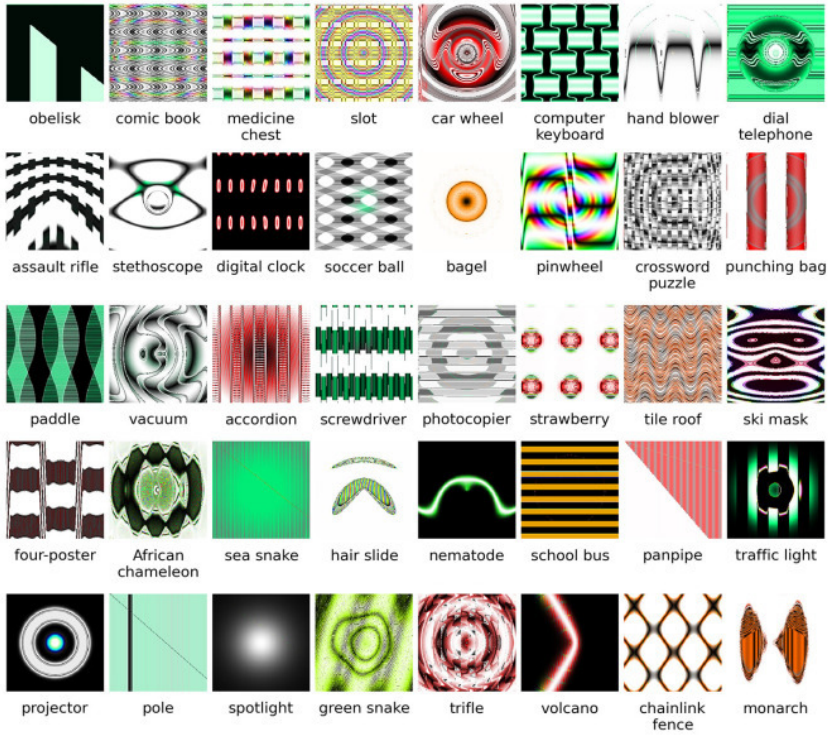
dawin



Szegedy et al 2014 ICML

LIMITATIONS:
IT'S EASY TO CONFUSE CNNS

dawin



Nguyen et al 2015 CVPR

LIMITATIONS

- unclear what a NN actually learns: weird edge cases
- further scaling: need new ideas
- no consolidation of techniques

WHAT'S NEXT?

dawin

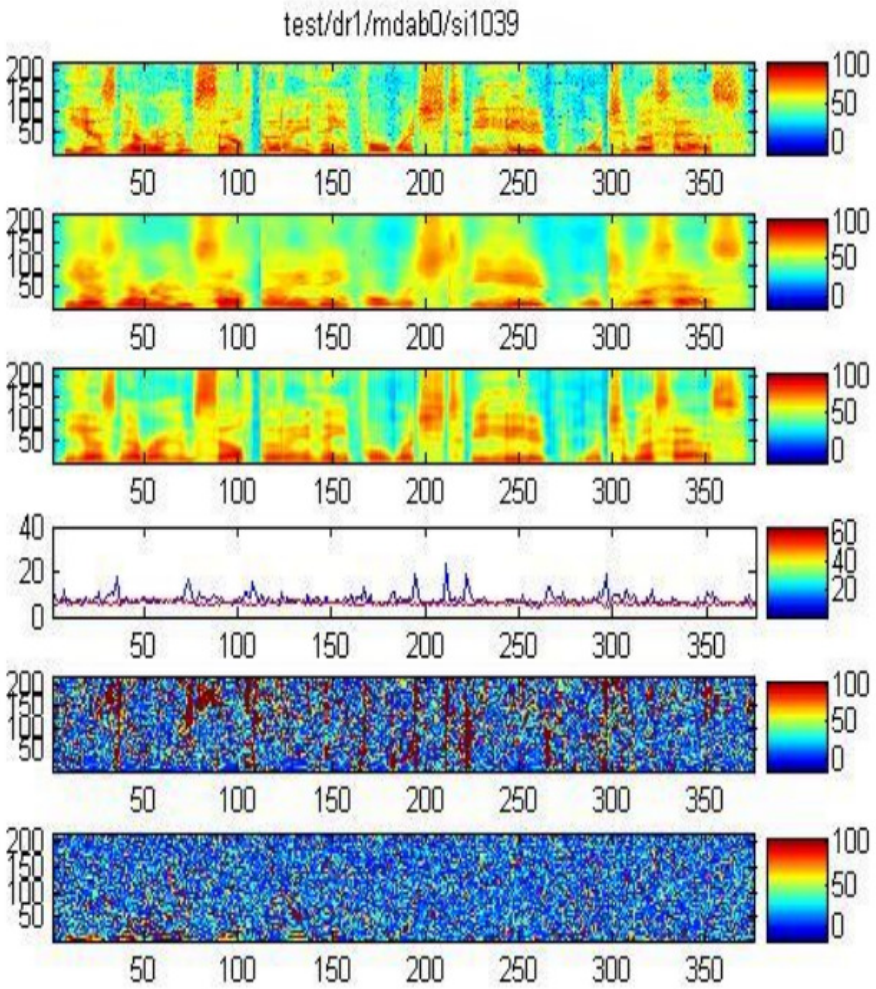
- resurgence of past ASR research results
- the other pillar: deep learning in language modelling ?
- the role of signal acquisition
- beyond ASR: what with understanding, reasoning

TAKE-HOME POINTS

dawin

- definition of deep learning
- roots of current DL successes: GPGPU + science + data
- DL amenable to multitask learning
- DL renovates ASR from the ground up: sequence learning, low-level (waveform?) features
- there are issues with DL
- the future: comeback of old results / how to scale? / language modelling / signal acquisition / semantics

dawin



Original spectrogram (log |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

Deng et al 2010 Interspeech

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

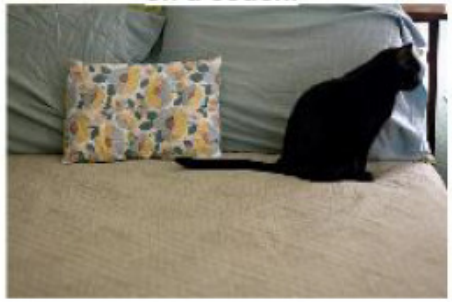
Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

dawin

[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

dawin

LARGE DATA CASE

[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

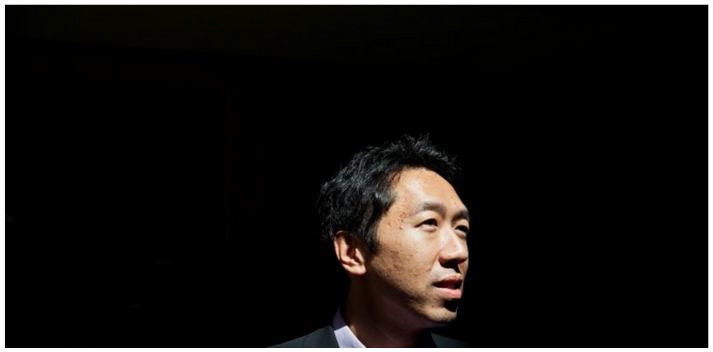
KEEP TRYING U

BEST!

CHEATING ... BUT STILL DOES WORSE THAN DNN-HMM

dawin

NOW YOU CAN BUILD GOOGLE'S $1M ARTIFICIAL BRAIN ON THE CHEAP



Last year at Google he built a computerized brain that worked as a cat detector. It used a roughly 1-billion-connection network trained on 1,000 computers to teach itself how to spot cat videos on YouTube. While this worked well, Ng says, some researchers walked away thinking, "If I don't have 1,000 computers, is there still any hope of my making progress on deep learning?" The system cost roughly $1 million.

"I was quite dismayed at this, particularly given that there are now a few other computer science research areas where a lot of the cutting-edge research is done only within giant companies," he recalls. "Others simply don't have the resources to do similar work."

On Monday, he's publishing a paper that shows how to build the same type of system for just $20,000 using cheap, but powerful, graphics microprocessors, or GPUs. It's a sort of DIY cookbook on how to build a low-

WIRED, 6/2013

dawin

On the importance of initialization and momentum in deep learning

Ilya Sutskever[1] ILYASU@GOOGLE.COM
James Martens JMARTENS@CS.TORONTO.EDU
George Dahl GDAHL@CS.TORONTO.EDU
Geoffrey Hinton HINTON@CS.TORONTO.EDU