

Artificial Intelligence and the Singularity

piero scaruffi

www.scaruffi.com

October 2014 - Revised 2019

*"The person who says it cannot be done should not
interrupt the person doing it" (Chinese proverb)*

Piero Scaruffi

piero scaruffi

p@scaruffi.com

scaruffi@stanford.edu

**Intelligence
is not Artificial**



*Why the Singularity is not coming
any time soon and other Meditations
on the Post-Human Condition
and the Future of Intelligence*

piero scaruffi

Thinking about Thought

*the structure of life and the
meaning of matter*

BRAIN

Volume 1
in the "Thinking about Thought" series

piero scaruffi

Piero Scaruffi

**A History of
Silicon Valley
1900-2015**

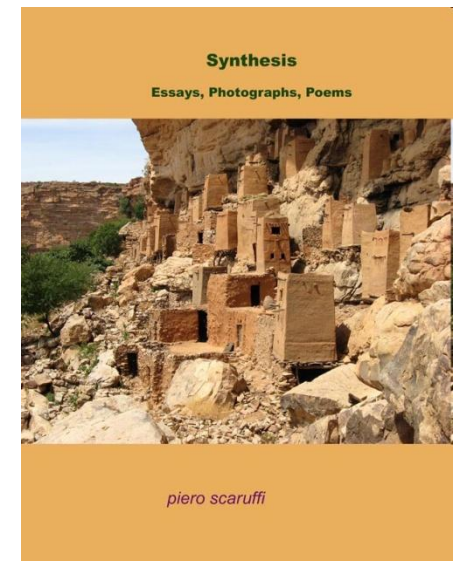
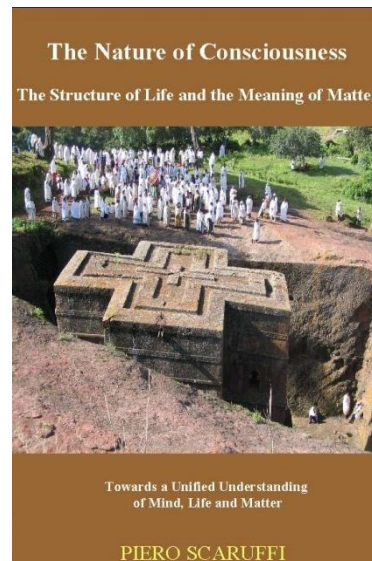
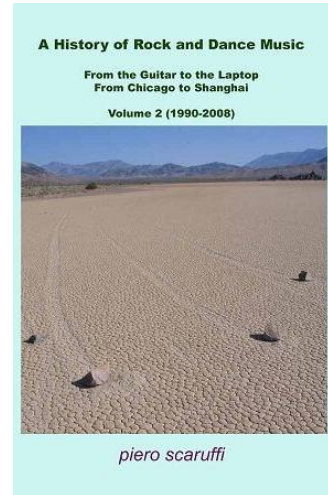
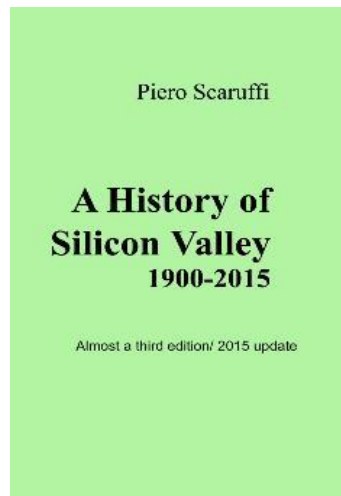
Almost a third edition! 2015 update



Olivetti AI Center, 1987

Piero Scaruffi

- Cultural Historian
- Cognitive Scientist
- Blogger
- Poet
- www.scaruffi.com



This is Part 12

- See <http://www.scaruffi.com/singular> for the index of this Powerpoint presentation and links to the other parts
 1. Classic A.I. - The Age of Expert Systems
 2. The A.I. Winter and the Return of Connectionism
 3. Theory: Knowledge-based Systems and Neural Networks
 4. Robots
 5. Bionics
 6. Singularity
 7. Critique
 8. The Future
 9. Applications
 10. Machine Art
 11. The Age of Deep Learning
 12. Natural Language Processing

Natural Language Processing 1980s-Today

Natural Language Processing

1981: Hans Kamp's Discourse Representation Theory



1982: Steven Small's Connectionist Parsing

1983: Gerard Salton and Michael McGill's
"Introduction to Modern Information Retrieval"
(the "bag-of-words model")



1986: Barbara Grosz's "Attention, Intentions, and the Structure of Discourse"



1988: Fred Jelinek's team at IBM publishes "A Statistical Approach to Language Translation"



1990: Peter Brown at IBM implements a statistical machine translation system



Natural Language Processing

"Bag-of-words" model: sentence representations are independent of word order

Sequence models developed by Michael Jordan (1986) and Jeffrey Elman (1990)



Natural Language Processing

Neural Networks for Symbolic Representation

- Jordan Pollack (1988): representing tree structures in neural networks (Recursive Auto-Associative Memory or RAAM)
- Christoph Goller and Andreas Kuechler (1995): extension of Pollack's RAAM

Learning Task-Dependent Distributed Representations by Backpropagation Through Structure 1995

Christoph Goller

Andreas Kuechler

AR-report AR-95-02

Technische Universität München

Recursive Distributed Representations

Jordan B. Pollack

1990

*Laboratory for AI Research &
Computer & Information Science Department
The Ohio State University*

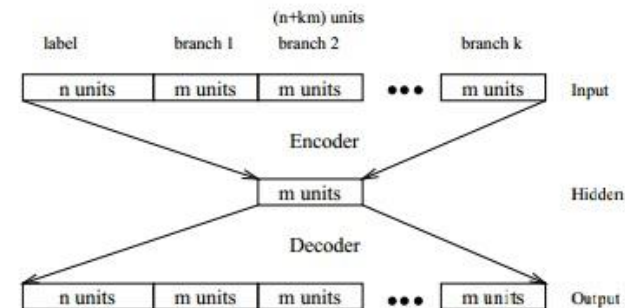
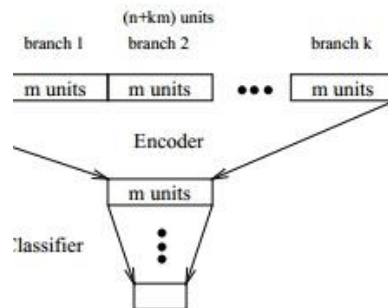


Fig. 1: The Folding Architecture (left side) and the standard LRAAM (right side).

Natural Language Processing

Neural Networks for Symbolic Representation

- Jordan Pollack (1990): representing tree structures in neural networks (Recursive Auto-Associative Memory or RAAM)

Recursive Distributed Representations

Jordan B. Pollack

*Laboratory for AI Research &
Computer & Information Science Department
The Ohio State University*



```
/-----> (WITH HIT (MOD TELESCOPE LONG))
-| /-----> (IS (MOD MAN SHORT) (THOUGHT MAN (SAW MAN JOHN)))
\-----| /-----> ((WITH SAW TELESCOPE) JOHN (ON MAN HILL))
| | /-----> (ATE MARY (WITH SPAGHETTI MEAT))
\-----| | | /-----> (HOPED PAT (THOUGHT JOHN (ATE MARY SPAGHETTI)))
| | | /-----> (HOPED PAT (SAW (WITH MAN TELESCOPE) PAT))
| | | /-----> (HIT JOHN (WITH MAN (MOD TELESCOPE LONG)))
\-----| | | /-----> (KNEW PAT (LOVED JOHN MARY))
| | | /-----> (THOUGHT PAT (KNEW JOHN (LOVED MARY JOHN)))
| | | \-----| \-----> (KNEW JOHN (LOVED MARY JOHN))
| | | | \-----> (HIT PAT (IS MAN (THOUGHT MAN (LOVED MARY JOHN))))
| | | | \-----> (IS MAN (THOUGHT MAN (LOVED MARY JOHN)))
\-----| | | \-----> (THOUGHT JOHN (ATE MARY SPAGHETTI))
| | | | \-----> (THOUGHT MAN (LOVED MARY JOHN))
| | | | \-----> (THOUGHT MAN (SAW MAN JOHN))
| | | | /-----> ((WITH ATE CHOPSTICKS) MARY SPAGHETTI)
| | | | \-----> (ATE MARY SPAGHETTI)
| | | | \-----> (ATE PAT MEAT)
| | | | /-----> (WITH SAW TELESCOPE)
\-----| | | /-----> (WITH ATE CHOPSTICKS)
| | | | /-----> (WITH SPAGHETTI MEAT)
| | | | \-----> (WITH MAN (MOD TELESCOPE LONG))
\-----| | | \-----> (ON MAN HILL)
| | | | \-----> (WITH MAN TELESCOPE)
\-----| | | \-----> (MOD TELESCOPE LONG)
| | | | \-----> (MOD MAN SHORT)
| | | | (MOD TELESCOPE LONG)) JOHN MAN)
| | | | /-----> (LOVED JOHN MARY)
| | | | \-----> (LOVED PAT MARY)
| | | | ---> (LOVED JOHN PAT)
| | | | (LOVED MARY JOHN)
| | | | ---> (SAW (WITH MAN TELESCOPE) PAT)
| | | | ---> (SAW (IS (MOD MAN SHORT) (THOUGHT MAN (SAW MAN JOHN))) PAT)
| | | | (SAW MAN JOHN)
```

Natural Language Processing

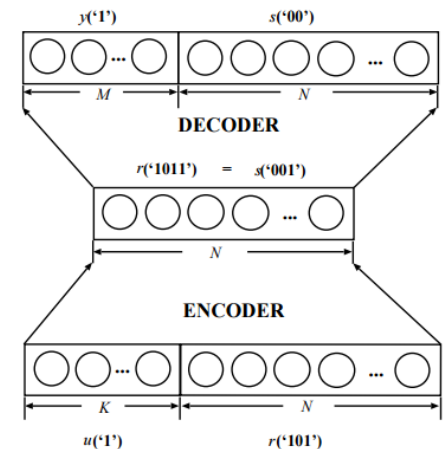
1996: Tom Landauer's and Susan Dumais' "latent semantic analysis"



1997: Ramon Neco's and Mikel Forcada's "encoder-decoder" model for machine translation



2001: John Lafferty's "conditional random fields" for sequence labeling



Natural Language Processing

How should words be represented

- Traditional NLP: the word is an atom
- Bag of words: the atom is a set of words
- 2003: Yoshua Bengio's Neural Probabilistic Language Model - represent words as vectors
- 2008: Mirella Lapata & Jeff Mitchell – represent sentences as vectors

Natural Language Processing

How should words be represented

- 2003: Yoshua Bengio's Neural Probabilistic Language Model - represent words as vectors

Journal of Machine Learning Research 3 (2003) 1137–1155

A Neural Probabilistic Language Model

Yoshua Bengio
Réjean Ducharme
Pascal Vincent
Christian Jauvin

*Département d'Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal, Montréal, Québec, Canada*



Université 
de Montréal

1. associate with each word in the vocabulary a distributed *word feature vector* (a real-valued vector in $\mathbb{R}^m$),
2. express the joint *probability function* of word sequences in terms of the feature vectors of these words in the sequence, and
3. learn simultaneously the *word feature vectors* and the parameters of that *probability function*.

Natural Language Processing

How should words be represented

- 2008: Mirella Lapata & Jeff Mitchell – represent sentences as vectors

Vector-based Models of Semantic Composition



Jeff Mitchell and Mirella Lapata
School of Informatics, University of Edinburgh

	animal	stable	village	gallop	jokey
horse	0	6	2	10	4
run	1	8	4	4	0

Figure 1: A hypothetical semantic space for horse and run

Natural Language Processing

Neural Machine Translation

2001: Yoshua Bengio's Neural Probabilistic Language Model converts a word symbol into a vector within a meaning space

2005: Bengio's Hierarchical Probabilistic Neural Network Language Model solves the "curse of dimensionality" in NLP

2008: Ronan Collobert and Jason Weston's Unified Architecture for NLP learns recursive structures



Université 
de Montréal

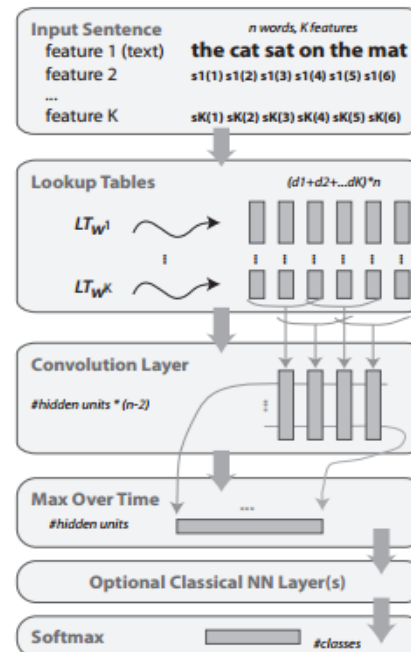
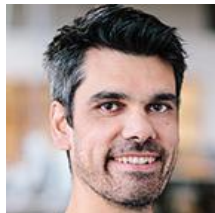
Natural Language Processing

Neural networks that learn recursive structures

- Ronan Collobert and Jason Weston (2008):
task-independent sequence tagging

A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning 2008

Ronan Collobert
Jason Weston
NEC Labs America



Natural Language Processing

- Collobert and Weston (2011)

Natural Language Processing (almost) from Scratch

Ronan Collobert

NEC Labs America, Princeton NJ.

Jason Weston

Google, New York, NY.

Léon Bottou

Michael Karlen

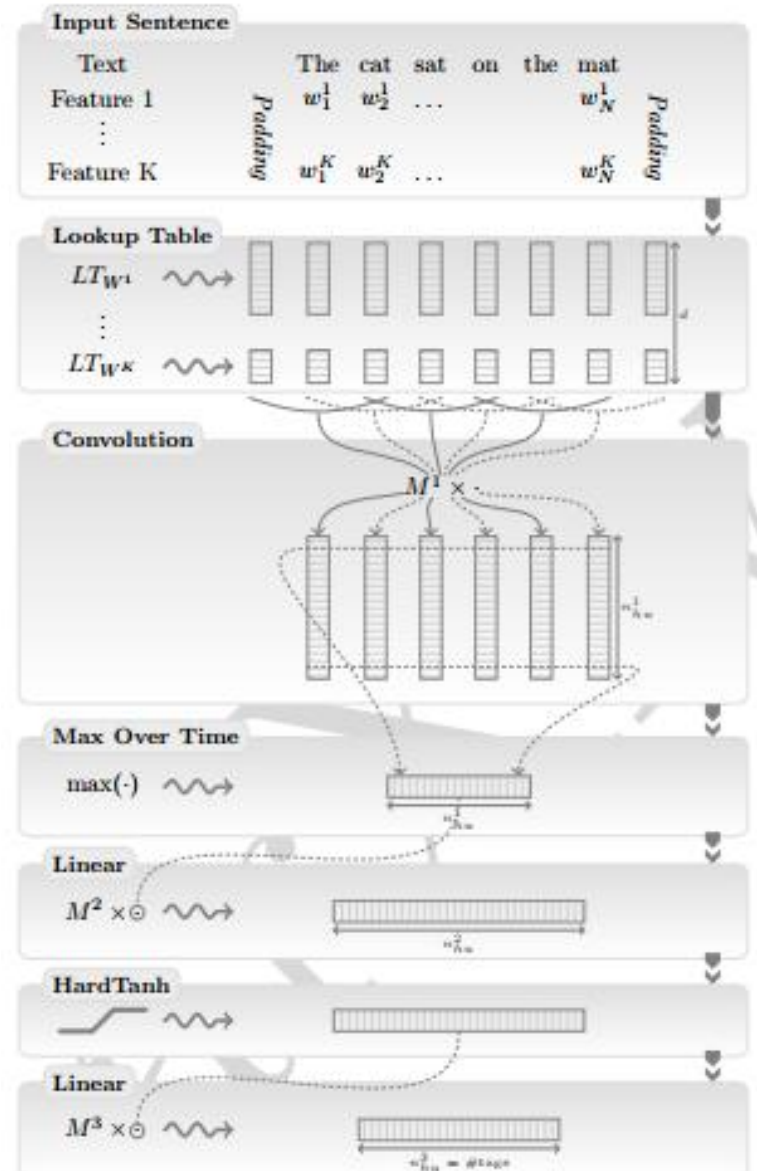
Koray Kavukcuoglu[†]

Pavel Kuksa[‡]

NEC Labs America, Princeton NJ.

2 Mar 2011

1



Natural Language Processing

Sequence tagging

- Generative models: hidden Markov models
- Conditional models: conditional random fields (John Lafferty, 2001)
- Unified Architecture (Collobert & Weston, 2008)

Natural Language Processing

Neural Machine Translation

2010: Tomas Mikolov's RNN that can process sentences of any length



2010: Richard Socher's recursive neural network (RNN) for continuous phrase representation



Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks

2010

Richard Socher, Christopher D. Manning, Andrew Y. Ng
Department of Computer Science
Stanford University

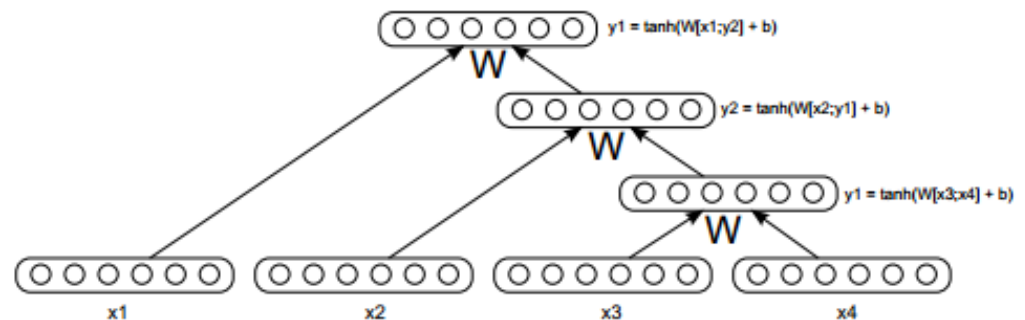


Figure 1: An example tree with a simple Recursive Neural Network

Natural Language Processing

Neural Machine Translation

2013: Nal Kalchbrenner and Phil Blunsom: statistical machine translation based purely on neural networks (“sequence to sequence learning”)

Recurrent Continuous Translation Models

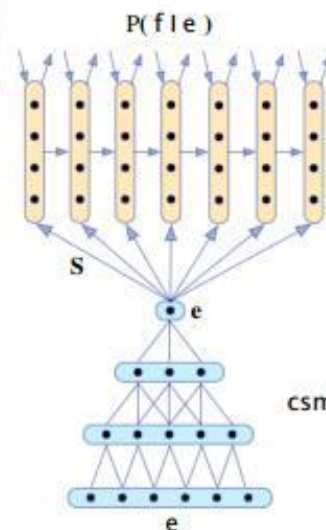
Nal Kalchbrenner

Phil Blunsom

Department of Computer Science

University of Oxford

2013



models the conditional probability $P(f|e)$ of a sentence $f = f_1, \dots, f_m$ in a target language F being the translation of a sentence $e = e_1, \dots, e_k$ in a source language E

Natural Language Processing

Neural Machine Translation

June 2014: Bengio's encoder-decoder model
(with Kyunghyun Cho and Dzmitry Bahdanau)

By-product: instead of using an LSTM, they use a simpler type of RNN, later called Gated Recurrent Unit (GRU) with no memory unit

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

Kyunghyun Cho

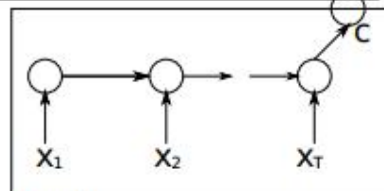
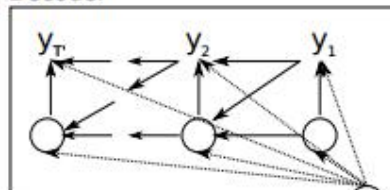
Bart van Merriënboer Caglar Gulcehre
Université de Montréal

Dzmitry Bahdanau
Jacobs University, Germany

Fethi Bougares Holger Schwenk
Université du Maine, France

Yoshua Bengio
Université de Montréal

Decoder



Encoder



Université de Montréal



Natural Language Processing

Neural Machine Translation 2014: Kyunghyun Cho's Gated Recurrent Unit

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

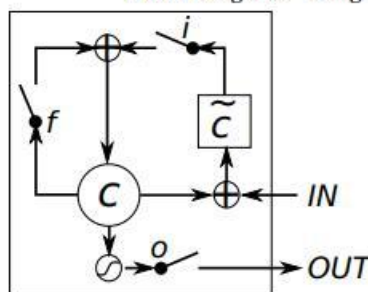
Kyunghyun Cho
Bart van Merriënboer Caglar Gulcehre
Université de Montréal

Dzmitry Bahdanau
Jacobs University, Germany

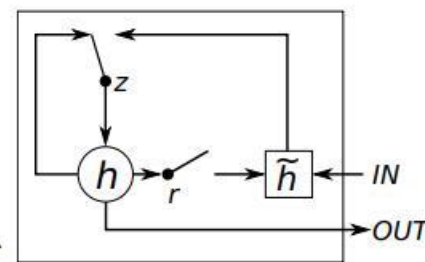
Fethi Bougares

Holger Schwenk

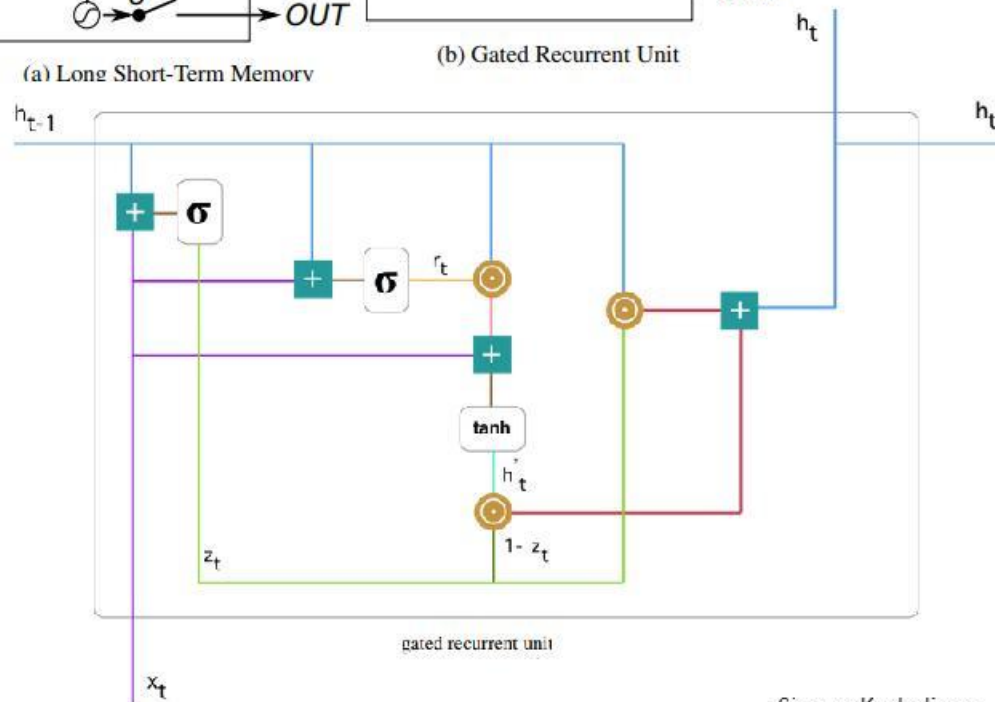
Yoshua Bengio



(a) Long Short-Term Memory



(b) Gated Recurrent Unit



Natural Language Processing

Bengio's encoder-decoder model

By-product: instead of using an LSTM, they use a simpler type of RNN, later called Gated Recurrent Unit (GRU) with no memory unit

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

Kyunghyun Cho

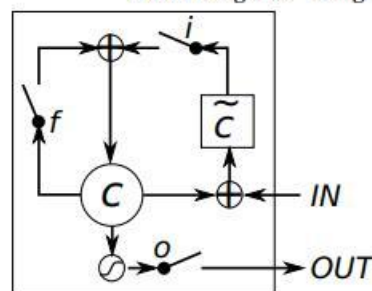
Bart van Merriënboer Caglar Gulcehre
Université de Montréal

Dzmitry Bahdanau
Jacobs University, Germany

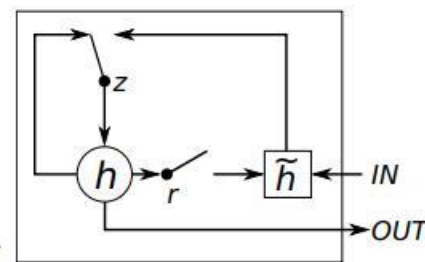
Fethi Bougares

Holger Schwenk

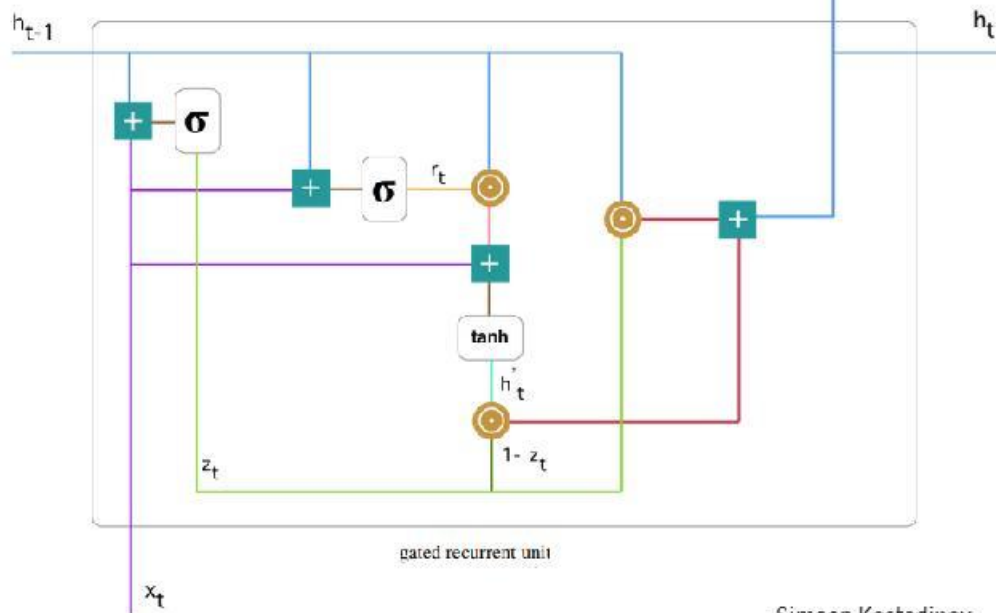
Yoshua Bengio



(a) Long Short-Term Memory



(b) Gated Recurrent Unit



Natural Language Processing

Neural Machine Translation

Sep 2014: Sutskever, Vinyals & Le solve the "sequence-to-sequence problem" using a LSTM (the length of the input sequence of characters doesn't have to be the same length of the output)

They too use the encoder-decoder model

NIPS 2014

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever



Oriol Vinyals



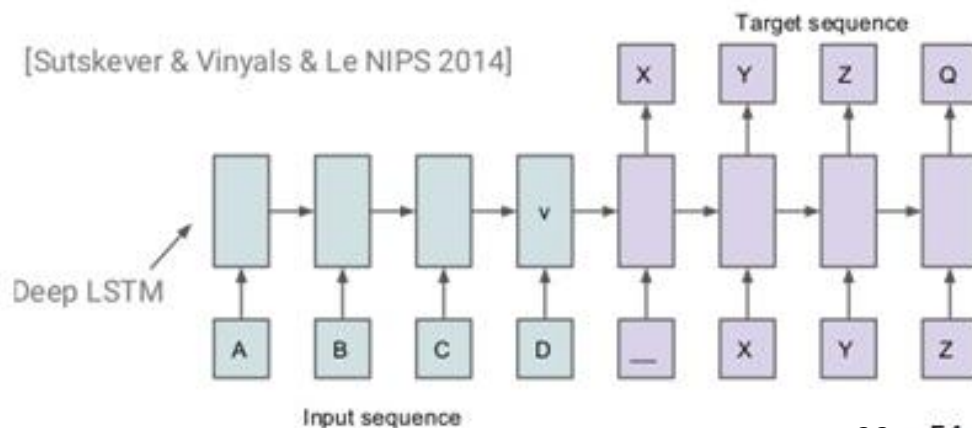
Quoc V. Le



Sequence-to-Sequence Model



[Sutskever & Vinyals & Le NIPS 2014]



23 54

$$P(y_1, \dots, y_T | x_1, \dots, x_T) = \prod_{t=1}^T P(y_t | v, y_1, \dots, y_{t-1})$$

Natural Language Processing

Neural Machine Translation

June 2014: Attention model by Volodymyr Mnih



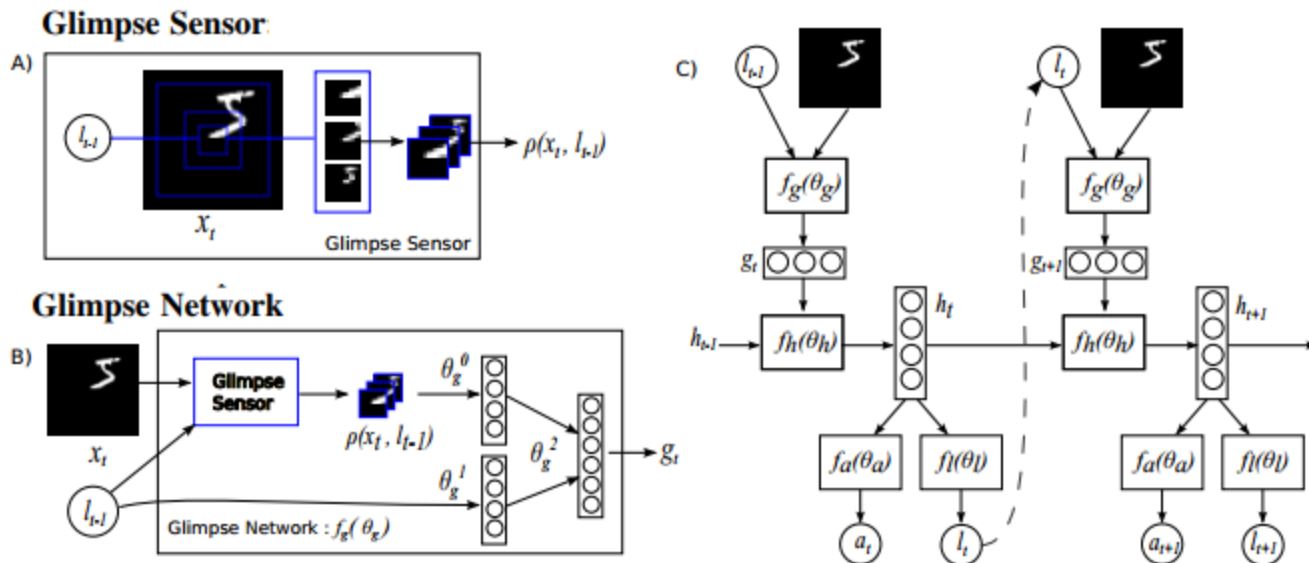
Recurrent Models of Visual Attention

Volodymyr Mnih Nicolas Heess Alex Graves Koray Kavukcuoglu
Google DeepMind



The Recurrent Attention Model (RAM)

Abstract



Natural Language Processing

Neural Machine Translation

Sep 2014: Attention model by Dzmitry Bahdanau, Kyunghyun Cho & Bengio
("additive" attention)

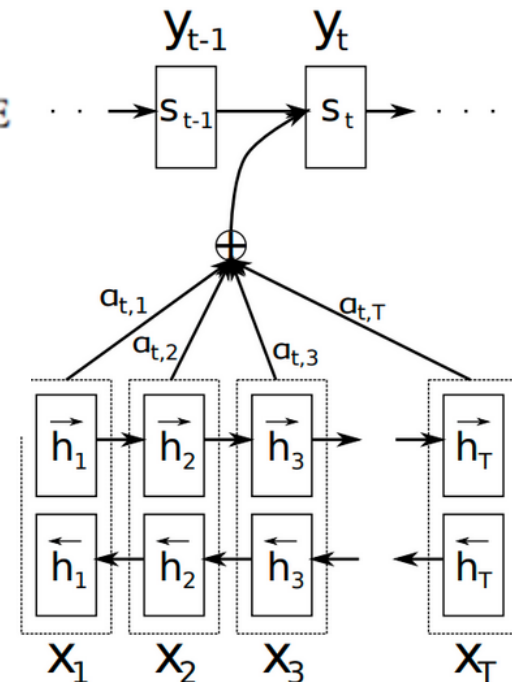


NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

Kyunghyun Cho **Yoshua Bengio***
Université de Montréal

Sep 2014



Example of Attention

Taken from "Neural Machine Translation by Jointly Learning to Align and Translate", 2015.

Natural Language Processing

Neural Machine Translation

2015: Attention model by Kelvin Xu (“Show/Attend/Tell”)

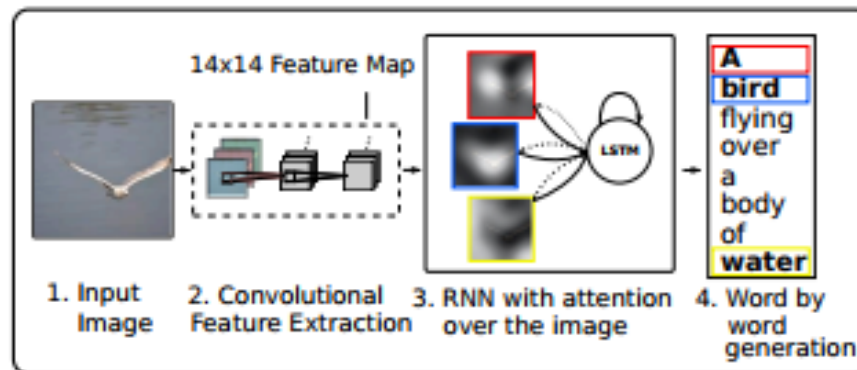
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu
Jimmy Lei Ba
Ryan Kiros
Kyunghyun Cho
Aaron Courville
Ruslan Salakhutdinov
Richard S. Zemel
Yoshua Bengio



Université
de Montréal

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4



Natural Language Processing

Neural Machine Translation

2015: Dzmitry Bahdanau's BiRNN

(bidirectional RNN) at Jacobs University
Bremen in Germany to improve the speed
of machine translation



Natural Language Processing

Neural Machine Translation

2016: Google's dynamic coattention network (Socher)

Dynamic Coattention Networks

Dynamic Coattention Networks

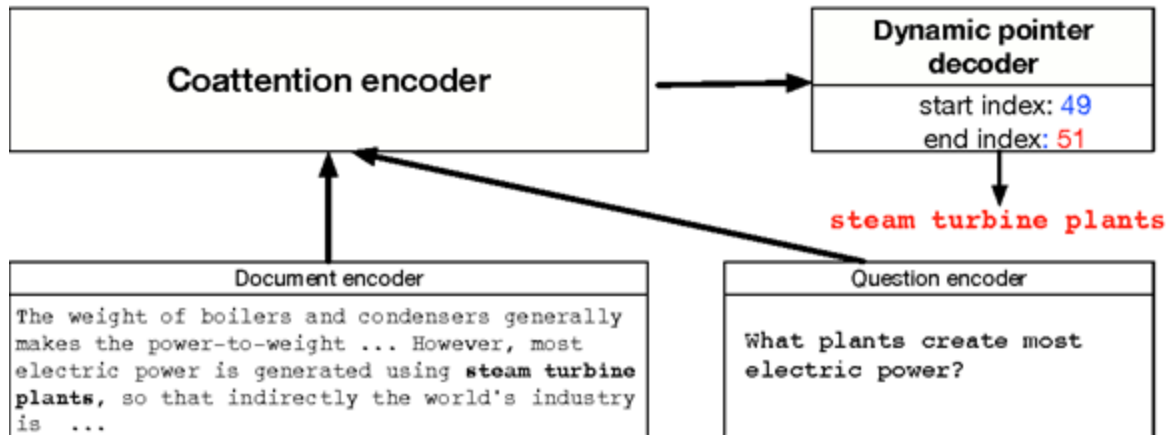
For Question Answering

👤 Caiming Xiong, Victor Zhong, Richard Socher

📖 Published 2016 in ArXiv



Google



Natural Language Processing

Neural Machine Translation

Nov 2016: Google switches
its translation algorithm to
an RNN

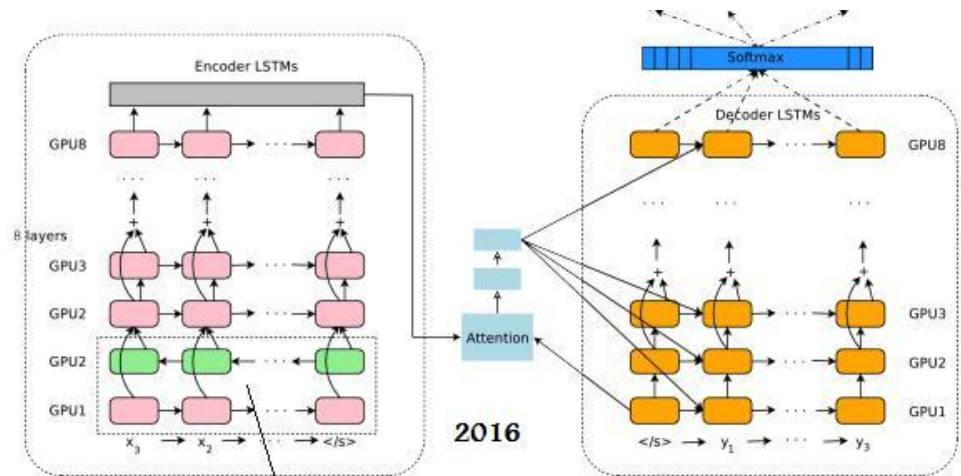
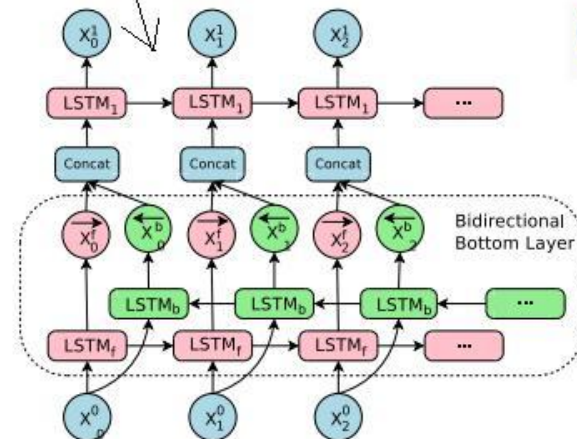


Figure 1: The model architecture of GNMT, Google's Neural Machine Translation system. On the left is the encoder network, on the right is the decoder network, in the middle is the attention module.



The structure of bi-directional connections in the first layer of the encoder.

Italian ▾



Ciao

Chinese (Simplified) ▾



你好

Nǐ hǎo

Natural Language Processing

Neural Machine Translation: convolutional nets instead of RNNs

2016: Nal Kalchbrenner's ByteNet

Neural Machine Translation in Linear Time

2016

Nal Kalchbrenner Lasse Espeholt Karen Simonyan
Aäron van den Oord Alex Graves Koray Kavukcuoglu
Google Deepmind, London UK

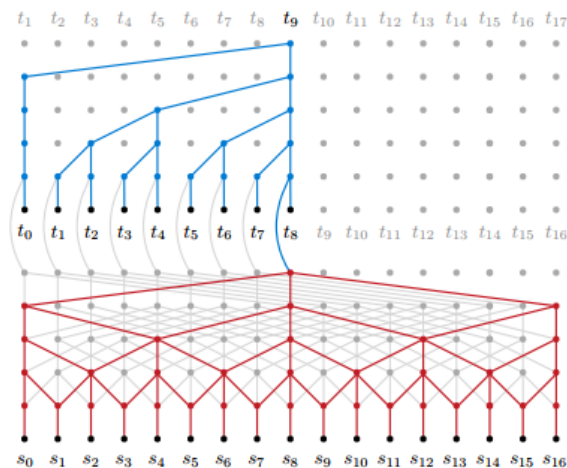


Figure 1. The architecture of the ByteNet. The target decoder (blue) is stacked on top of the source encoder (red). The decoder generates the variable-length target sequence using dynamic unfolding.

Natural Language Processing

Neural Machine Translation: convolutional nets instead
of RNNs

2016: Facebook's ConvS2S

ConvS2S

Convolutional Sequence to Sequence Learning

Jonas Gehring
Michael Auli
David Grangier
Denis Yarats
Yann N. Dauphin



Facebook AI Research (FAIR)

Jul 2017

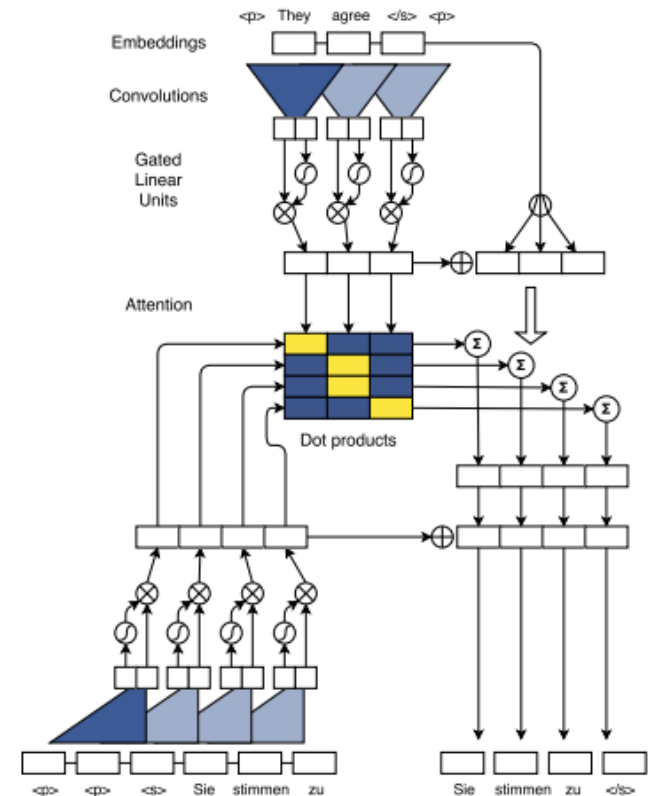


Figure 1. Illustration of batching during training. The English source sentence is encoded (top) and we compute all attention values for the four German target words (center) simultaneously. Our attentions are just dot products between decoder context representations (bottom left) and encoder representations. We add the conditional inputs computed by the attention (center right) to the decoder states which then predict the target words (bottom right). The sigmoid and multiplicative boxes illustrate Gated Linear Units.

Natural Language Processing

Neural Machine Translation

Do these systems that translate one sentence into another sentence actually "understand" language?

Xing Shi (University of Southern California, 2016): the vector representations of neural machine translation capture some morphological and syntactic properties of language

Does String-Based Neural MT Learn Source Syntax?



Xing Shi, Inkit Padhi, and Kevin Knight

Information Sciences Institute & Computer Science Department
University of Southern California

Natural Language Processing

Neural Machine Translation

Do these systems that translate one sentence into another sentence actually "understand" language?

Yonatan Belinkov (MIT, 2017): vector representations contain even some semantical properties

Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks



**Yonatan Belinkov¹ Lluís Màrquez² Hassan Sajjad²
Nadir Durrani² Fahim Dalvi² James Glass¹**



¹MIT Computer Science and Artificial Intelligence Laboratory

²Qatar Computing Research Institute, HBKU, Doha, Qatar



Natural Language Processing

Neural Machine Translation

Microsoft (2018): news translation



March 14, 2018

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English



Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou
March 2018

Natural Language Processing

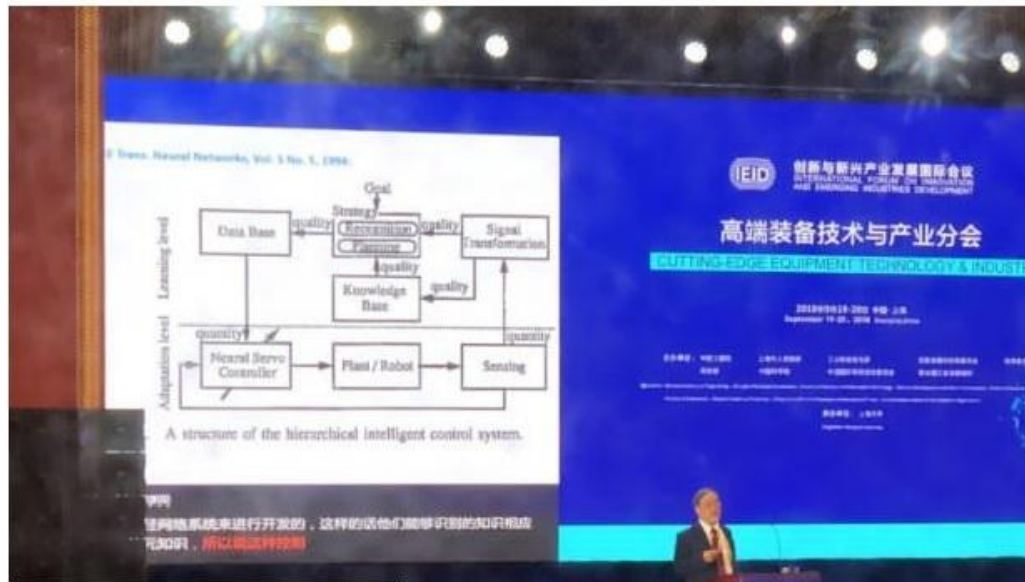
Neural Machine Translation

But beware... iFlytek scandal of 2018

National AI champion iFlytek in dispute over 'automated' speech translation at Shanghai forum

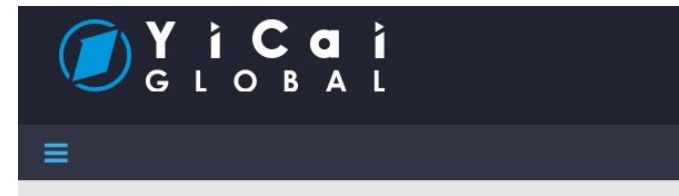
PUBLISHED : Tuesday, 25 September, 2018, 8:01pm

UPDATED : Tuesday, 25 September, 2018, 8:00pm



Transferring data from cdn2.i-scmp.com...

On Oct 17, 2018 8:41 PM, piero scaruth <p@scaruth.com>



NEW PRODUCTS

**Shares in AI Unicorn
iFlytek Slide on Voice
Recognition Forgery
Scandal**

Natural Language Processing

Discourse Analysis

2013: Mikolov's "Word2vec" method for learning vector representations of words from large amounts of unstructured text data

2014: James Weston's "memory networks", neural networks coupled with long-term memories for question-answering

Natural Language Processing

How should words be represented

- Tomas Mikolov's skip-gram (2013)

Efficient Estimation of Word Representations in Vector Space

Sep 2013

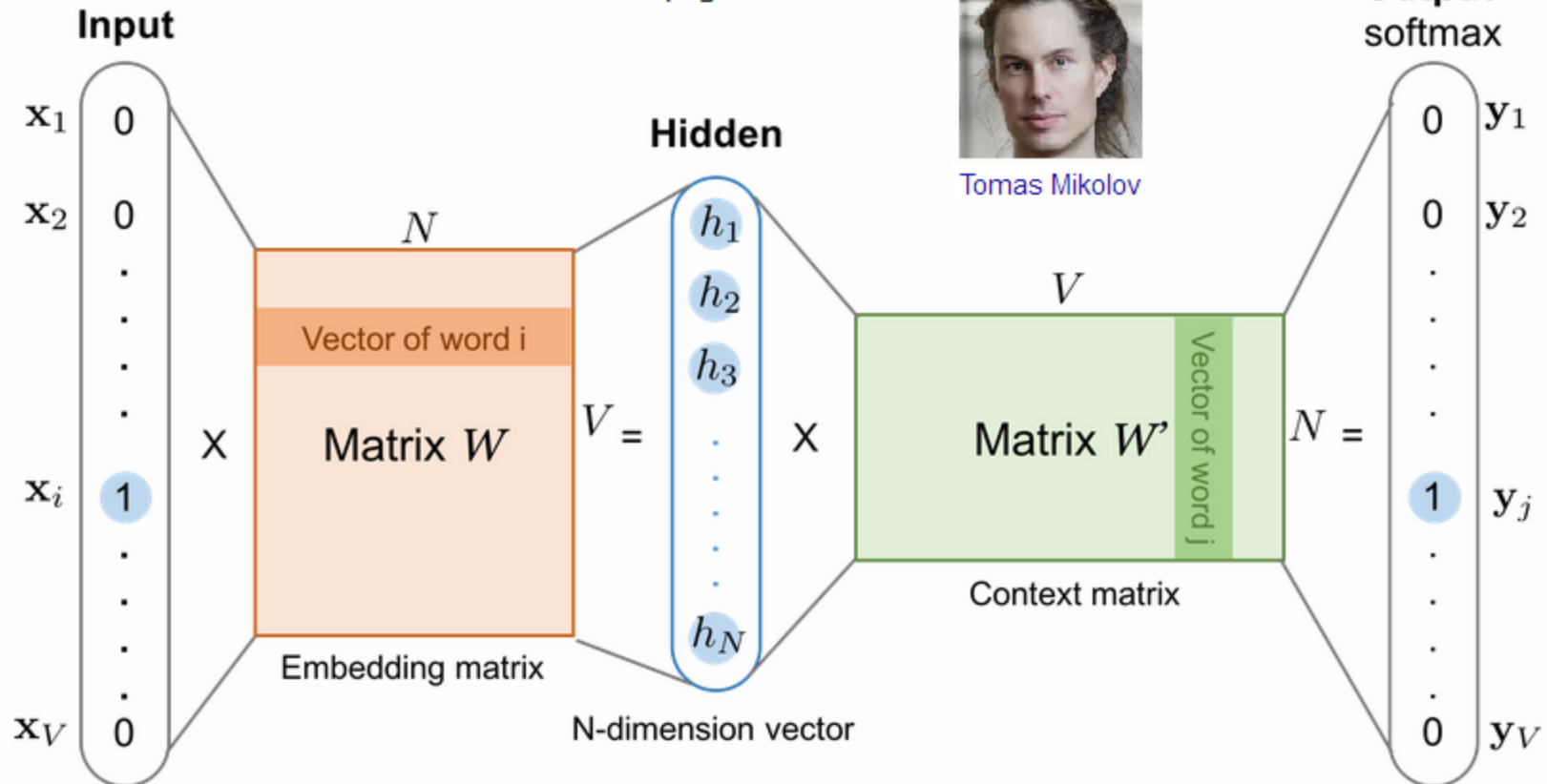
Tomas Mikolov
Google Inc., Mountain View, CA

Greg Corrado
Google Inc., Mountain View, CA

Kai Chen
Google Inc., Mountain View, CA

Jeffrey Dean
Google Inc., Mountain View, CA

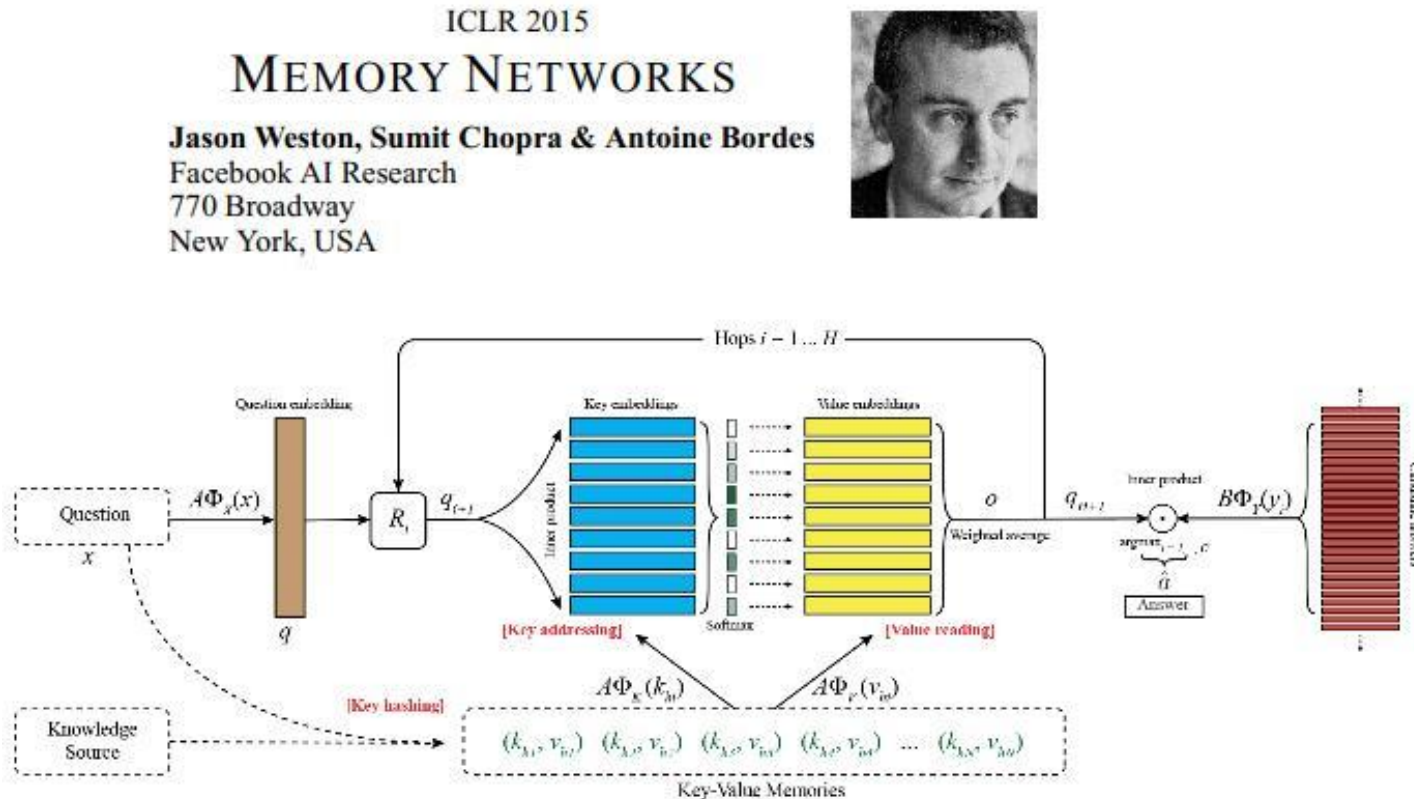
The skip-gram model.



Natural Language Processing

Question-answering

- 2014: James Weston's memory networks



Natural Language Processing

The old goal: relate words within a sentence

Success story: machine translation

The new goal: capture information about entire sentences

Application: understanding texts

Google Translate

English

Chinese (Simplified)

I am a man.
I live in California.
Today I am in
China.

×

我是男人
我居住在加利福尼亚州。
今天我在中国。



Natural Language Processing

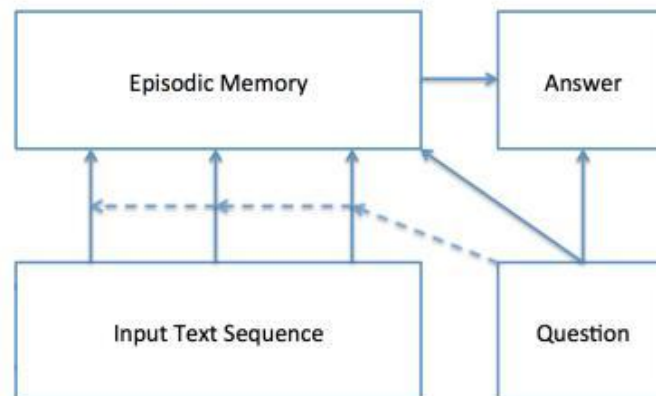
Discourse Analysis

2015: Richard Socher's dynamic memory networks

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

**Ankit Kumar, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong,
Romain Paulus, Richard Socher**

MetaMind, Palo Alto, CA USA



2015

Natural Language Processing

Discourse Analysis

2015: Oriol Vinyals and Quoc Le's Neural Conversational Model

Oriol Vinyals



Quoc V. Le



Machine: *hi*

Human: *hello*

Machine: *could you please let me know what are the operating systems you are using ?*

Human: *i am using linux*

Machine: *ok it was in the ssl vpn on which you stated it was broken*

Machine: *please check if you are logged in to corp network*

Machine: *what is the error that you are running please*

Human: *i am seeing an error related to vpn*

Machine: *what is the error message that you are getting when connecting to vpn using network connect ?*

Human: *connection refused or something like that*

Machine: *may i know the version of network connect you connect ?*

Human: *i am not sure i know that*

Machine: *is the network connect prompting for any user certificate*

Machine: *?*

Human: *yes*

Machine: *are you at home or at work at the moment ?*

Machine: *office ?*

Human: *i am at home*

Machine: *try this*

Machine: *goto <URL> page and enter your name and password*

Natural Language Processing

Discourse Analysis

2014: GloVe (an alternative to Word2vec) by Jeffrey Pennington, Richard Socher, Chris Manning (Stanford)

2015: FastText (Mikolov, Facebook)



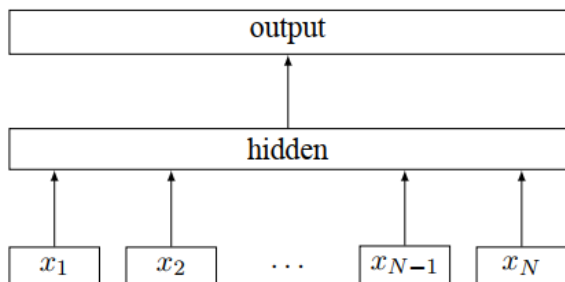
Chris Manning



Jeffrey Pennington

Bag of Tricks for Efficient Text Classification

Armand Joulin Edouard Grave Piotr Bojanowski Tomas Mikolov
Facebook AI Research
{ajoulin,egrave,bojanowski,tmikolov}@fb.com



Aug 2016

Figure 1: Model architecture of fastText

Natural Language Processing

Discourse Analysis

“Word embeddings” (like Word2vec and GloVe) derive a map of how words relate to each other based on the configurations in which the words appear in large amounts of text (“the distributional hypothesis” that words frequently occurring in the same contexts are related).

Natural Language Processing

Discourse Analysis

Next step: capture information about entire sentences

Jamie/Ryan Kiros: Skip-Thoughts (2015)

Lajanugen Logeswaran: Quick-Thoughts (2018)



Jamie Kiros

Skip-Thought Vectors

Ryan Kiros¹, Yukun Zhu¹, Ruslan Salakhutdinov^{1,2}, Richard S. Zemel^{1,2}
Antonio Torralba³, Raquel Urtasun¹, Sanja Fidler¹
University of Toronto¹
Canadian Institute for Advanced Research²
Massachusetts Institute of Technology³

AN EFFICIENT FRAMEWORK FOR LEARNING SENTENCE REPRESENTATIONS

Lajanugen Logeswaran* & Honglak Lee†*

*University of Michigan, Ann Arbor, MI, USA

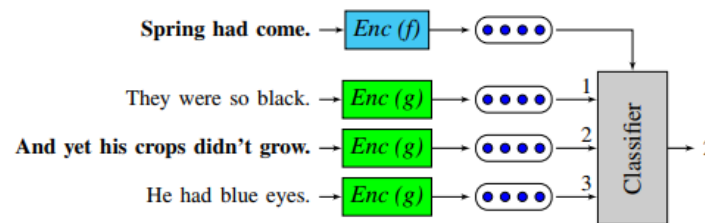
†Google Brain, Mountain View, CA, USA



Mar 2018

Spring had come. → **Enc** → [] → **Dec** → And yet his crops didn't grow.

(a) Conventional approach



(b) Proposed approach

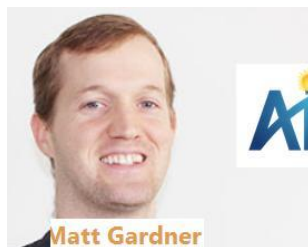
Jun 2015

Natural Language Processing

Discourse Analysis

Next step: capture information about entire sentences

Gardner-Zettlemoyer: ELMO (2018)



Task	Previous SOTA		ELMo +
SQuAD	SAN	84.4	85.8
SNLI	Chen et al (2017)	88.6	88.7
SRL	He et al (2017)	81.7	84.6
Coref	Lee et al (2017)	67.2	70.4
NER	Peters et al (2017)	91.9	92.22
Sentiment (5-class)	McCann et al (2017)	53.7	54.7

Deep contextualized word representations

Matthew E. Peters[†], Mark Neumann[†], Mohit Iyyer[†], Matt Gardner[†],
{matthewp, markn, mohiti, mattg}@allenai.org

Christopher Clark*, Kenton Lee*, Luke Zettlemoyer[†]*
{csquared, kentonl, lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence

*Paul G. Allen School of Computer Science & Engineering, University of Washington

Mar 2018

Natural Language Processing

Discourse Analysis

TagML (Matthew Peters, 2017): bidirectional language model

Semi-supervised sequence tagging with bidirectional language models

Apr 2017

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, Russell Power
Allen Institute for Artificial Intelligence

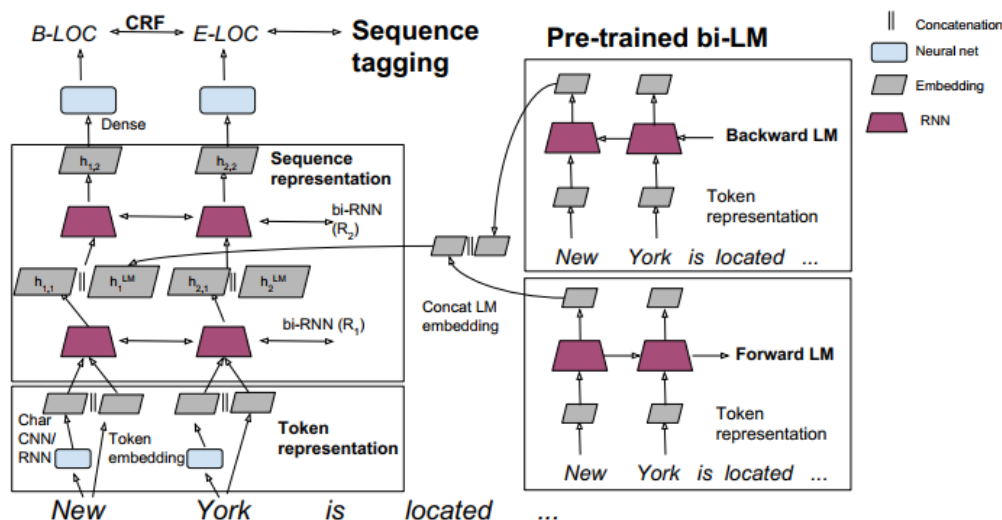


Figure 2: Overview of TagLM, our language model augmented sequence tagging architecture.

Natural Language Processing

Attention

attention mechanisms (Source: Lilian Weng)		
Name	Alignment score function	Citation
Additive(*)	$\text{score}(\mathbf{s}_t, \mathbf{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\mathbf{s}_t; \mathbf{h}_i])$	Bahdanau2015
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \mathbf{s}_t)$ Note: This simplifies the softmax alignment max to only depend on the target position.	Luong2015
General	$\text{score}(\mathbf{s}_t, \mathbf{h}_i) = \mathbf{s}_t^\top \mathbf{W}_a \mathbf{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.	Luong2015
Dot-Product	$\text{score}(\mathbf{s}_t, \mathbf{h}_i) = \mathbf{s}_t^\top \mathbf{h}_i$	Luong2015
Scaled Dot-Product(^)	$\text{score}(\mathbf{s}_t, \mathbf{h}_i) = \frac{\mathbf{s}_t^\top \mathbf{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani2017
Self-Attention(&)	Relating different positions of the same input sequence. Theoretically the self-attention can adopt any score functions above, but just replace the target sequence with the same input sequence.	Cheng2016
Global/Soft	Attending to the entire input state space.	Xu2015
Local/Hard	Attending to the part of input state space; i.e. a patch of the input image.	Xu2015; Luong2015

Natural Language Processing

Attention

2015: “dot-product” (multiplicative)
method by Minh-Thang Luong, Chris

Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong Hieu Pham Christopher D. Manning

Computer Science Department, Stanford University, Stanford, CA 94305



Sep 2015

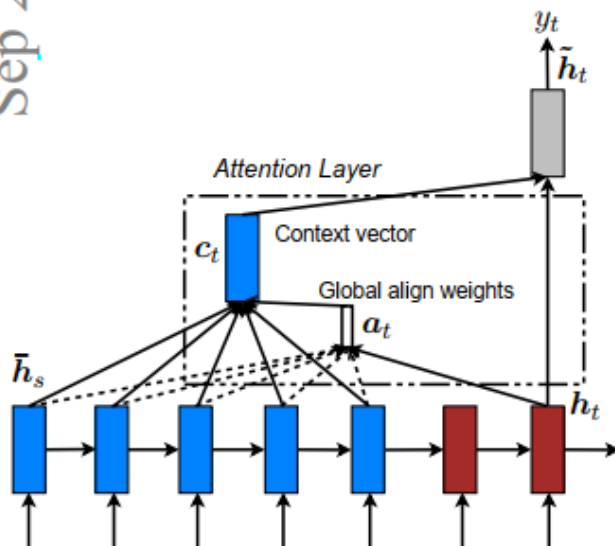


Figure 2: Global attentional model

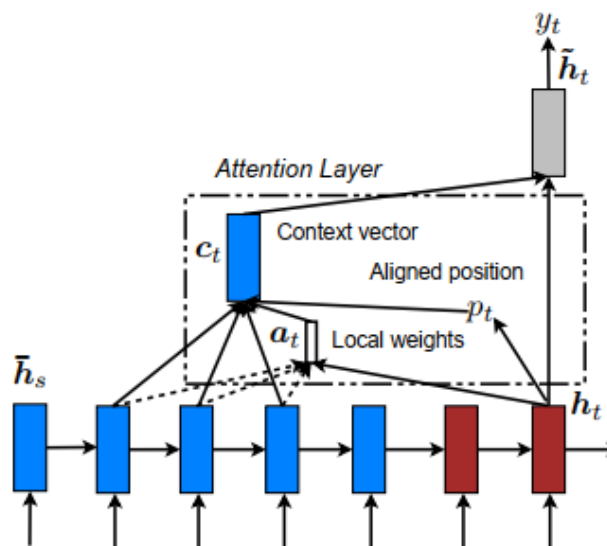


Figure 3: Local attention model

Natural Language Processing

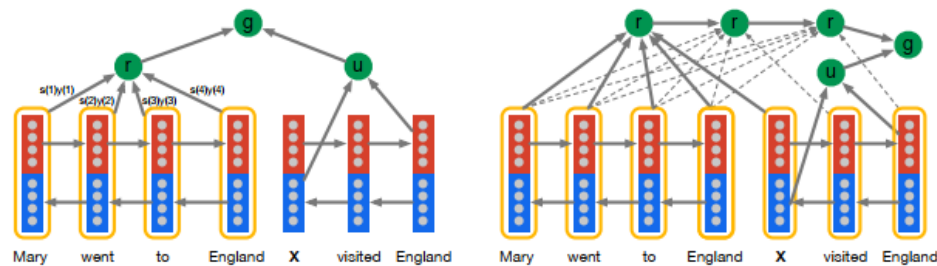
Attention

Attentive Reader (Phil Blunsom, Oxford University, 2015), a generalization of Weston's memory networks for question answering



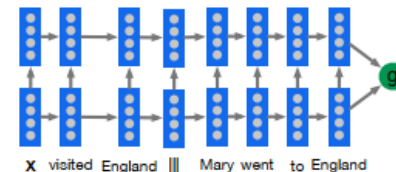
Teaching Machines to Read and Comprehend

Karl Moritz Hermann[†] Tomáš Kočiský^{†‡} Edward Grefenstette[†]
Lasse Espeholt[†] Will Kay[†] Mustafa Suleyman[†] Phil Blunsom^{†‡}
[†]Google DeepMind [‡]University of Oxford



(a) Attentive Reader.

(b) Impatient Reader.



(c) A two layer Deep LSTM Reader with the question encoded before the document.

Natural Language Processing

Self-Attention

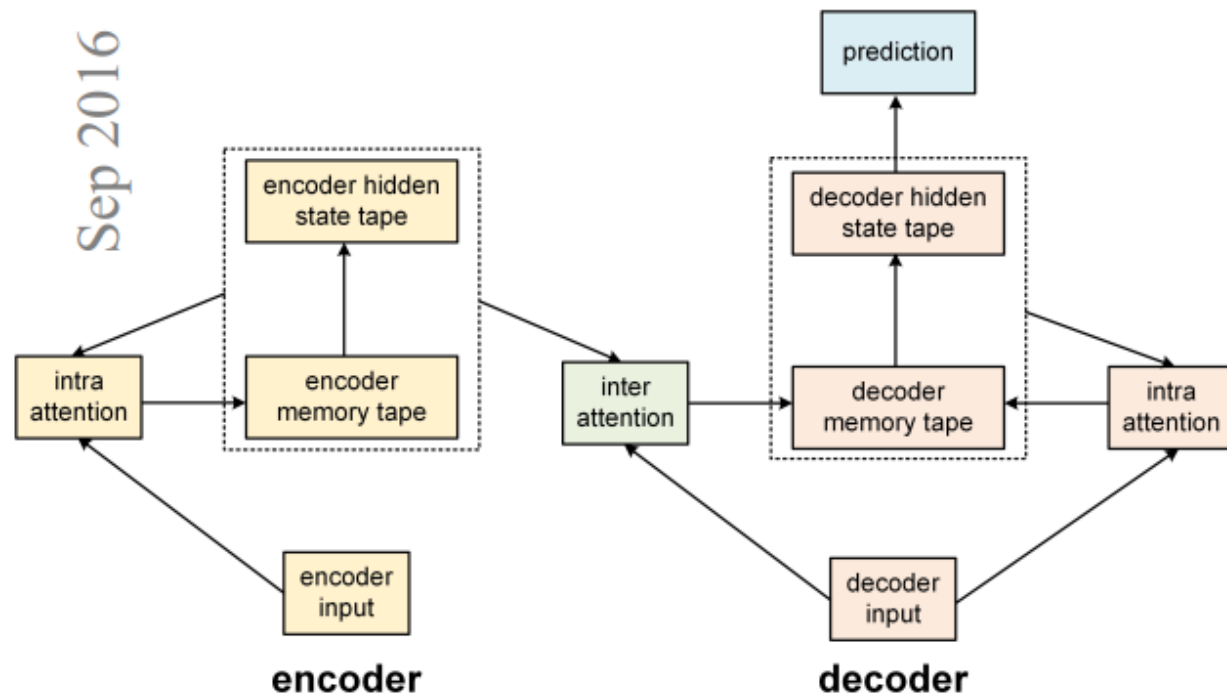
2016: Jianpeng Cheng, Mirella Lapata

Long Short-Term Memory-Networks for Machine Reading

Jianpeng Cheng, Li Dong and Mirella Lapata
School of Informatics, University of Edinburgh



Sep 2016



Natural Language Processing

Self-Attention

2016: Ankur Parikh (Google)

2017: Richard Socher (Salesforce)

Sep 2016

A Decomposable Attention Model for Natural Language Inference

Ankur P. Parikh
Google
New York, NY

Oscar Täckström
Google
New York, NY

Dipanjan Das
Google
New York, NY

Jakob Uszkoreit
Google
Mountain View, CA

Nov 2017

A DEEP REINFORCED MODEL FOR ABSTRACTIVE SUMMARIZATION

Romain Paulus, Caiming Xiong & Richard Socher
Salesforce Research
172 University Avenue
Palo Alto, CA 94301, USA

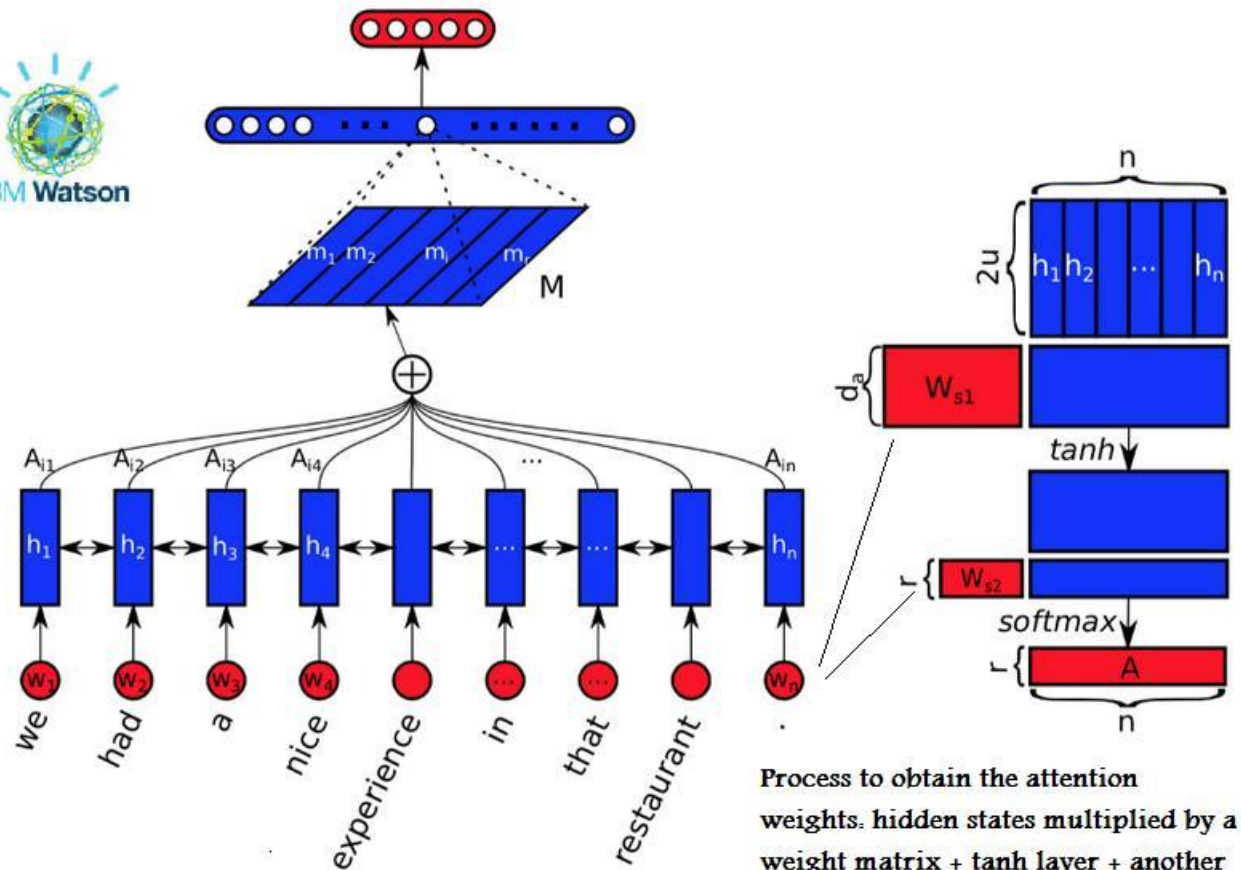
Natural Language Processing

Self-Attention

2017: Zhouhan Lin (Univ of Montreal)



Zhouhan LIN



Architecture of the self-attention model

Process to obtain the attention weights: hidden states multiplied by a weight matrix + tanh layer + another weight matrix + a softmax layer

Published as a conference paper at ICLR 2017

A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING

Zhouhan Lin^{1*}, Minwei Feng², Cicero Nogueira dos Santos², Mo Yu², Bing Xiang², Bowen Zhou² & Yoshua Bengio^{1†}
²IBM Watson

¹Montreal Institute for Learning Algorithms (MILA), Université de Montréal

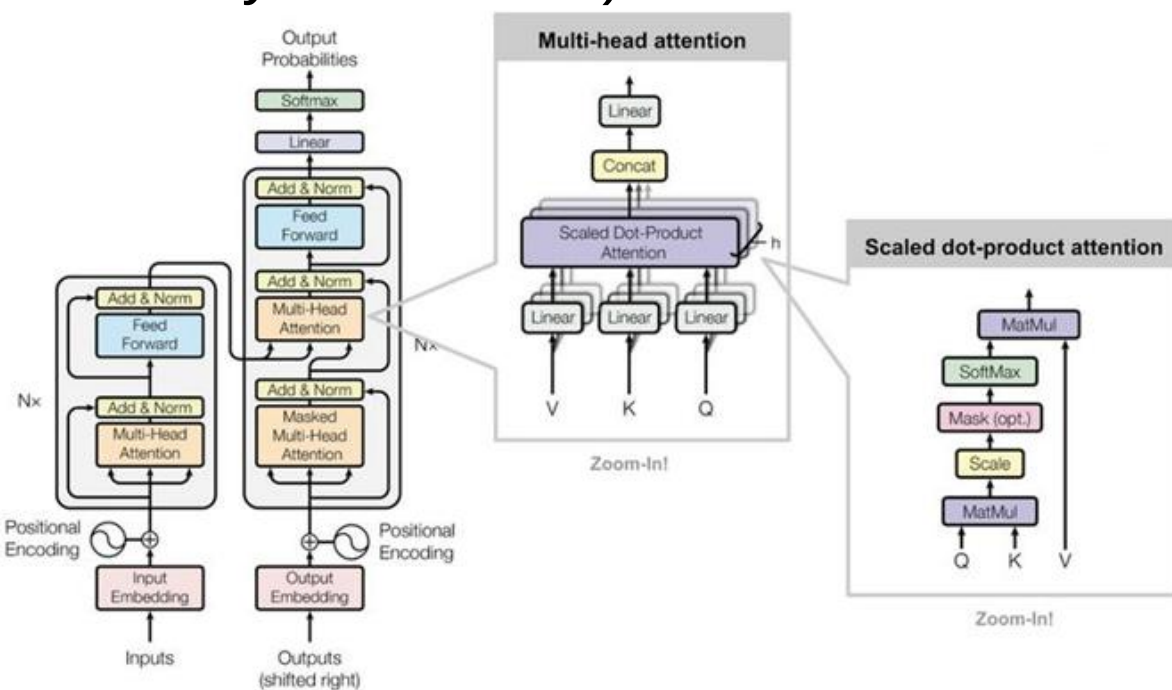
[†]CIFAR Senior Fellow

Natural Language Processing

Self-Attention

2017: Ashish Vaswani's "transformer" (Google Brain):
no RNN, only self-attention

An infinite reference window (RNN and LSTM had a
very limited one)



Attention Is All You Need

Dec 2017

Ashish Vaswani*
Llion Jones*
Google Research

Noam Shazeer*
Aidan N. Gomez*[†]
University of Toronto
Illia Polosukhin*[‡]

Niki Parmar*
Google Brain

Jakob Uszkoreit*
Lukasz Kaiser*
Google Brain



Vaswani, Shazeer, Uszkoreit

Natural Language Processing

Self-Attention

2017: Google's "transformer"

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukasz.kaiser@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

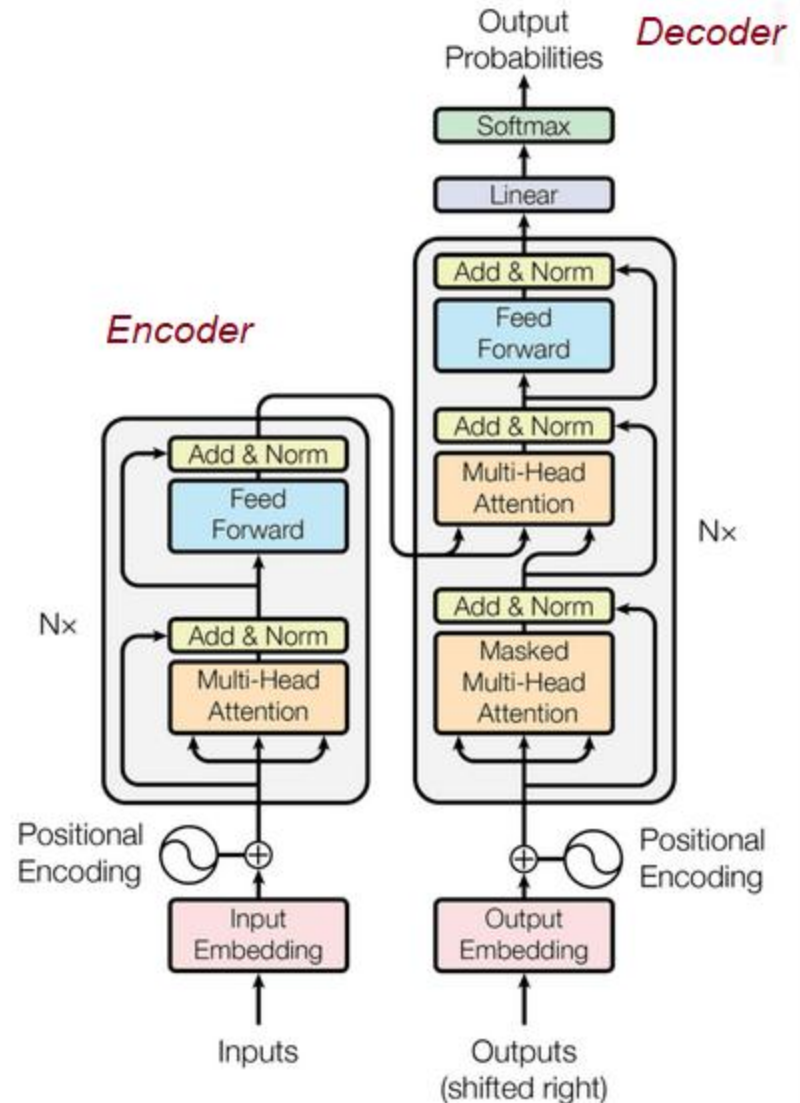
1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†]Work performed while at Google Brain.

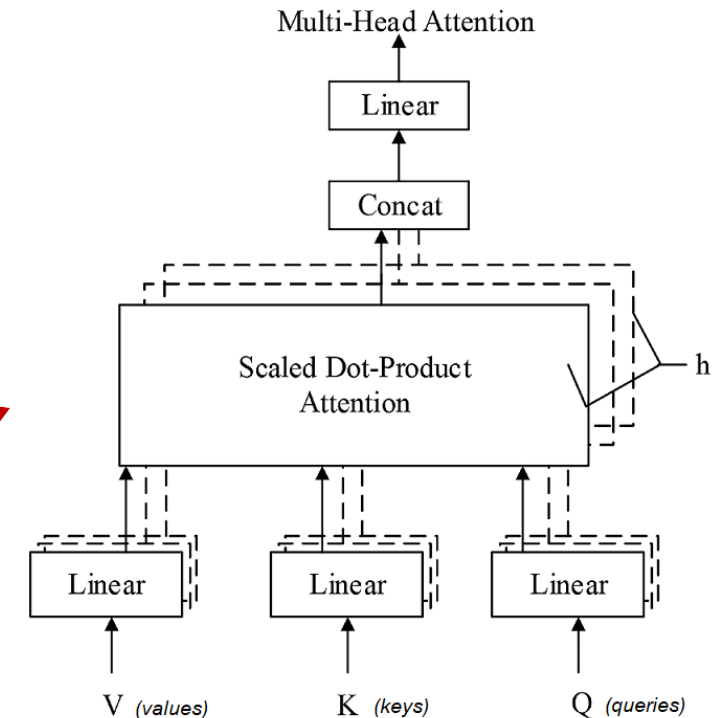
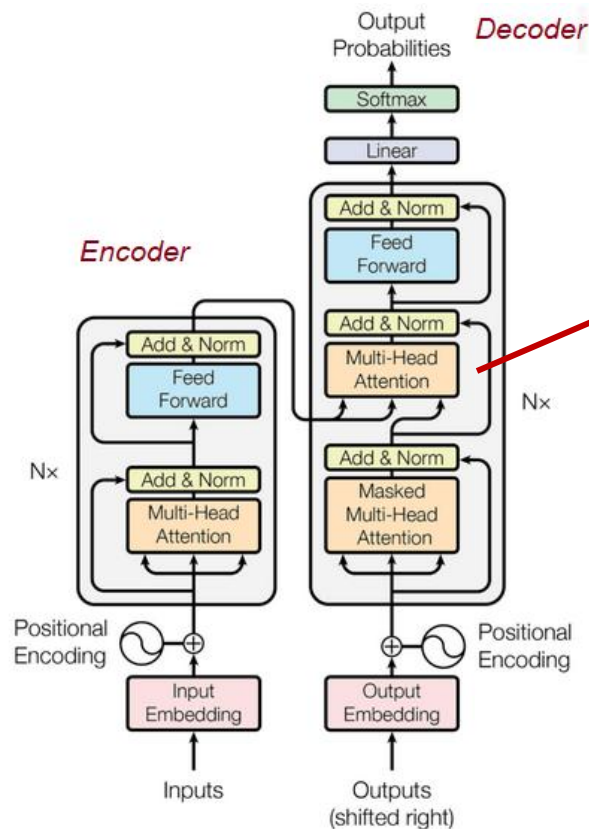
[‡]Work performed while at Google Research.



Natural Language Processing

Self-Attention

2017: Google's "transformer"



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Natural Language Processing

Self-Attention

2018: Wei Yu's "QANet" (CMU + Google Brain): no RNN, only convolutions (to model local interactions) and self-attention (to model global interactions)

The performances of different models on SQuAD dataset.
2018

ICLR 2018

QANET: COMBINING LOCAL CONVOLUTION WITH GLOBAL SELF-ATTENTION FOR READING COMPREHENSION

Adams Wei Yu^{1,*}, David Dohan^{2,†}, Minh-Thang Luong^{2,†}

¹Carnegie Mellon University, ²Google Brain

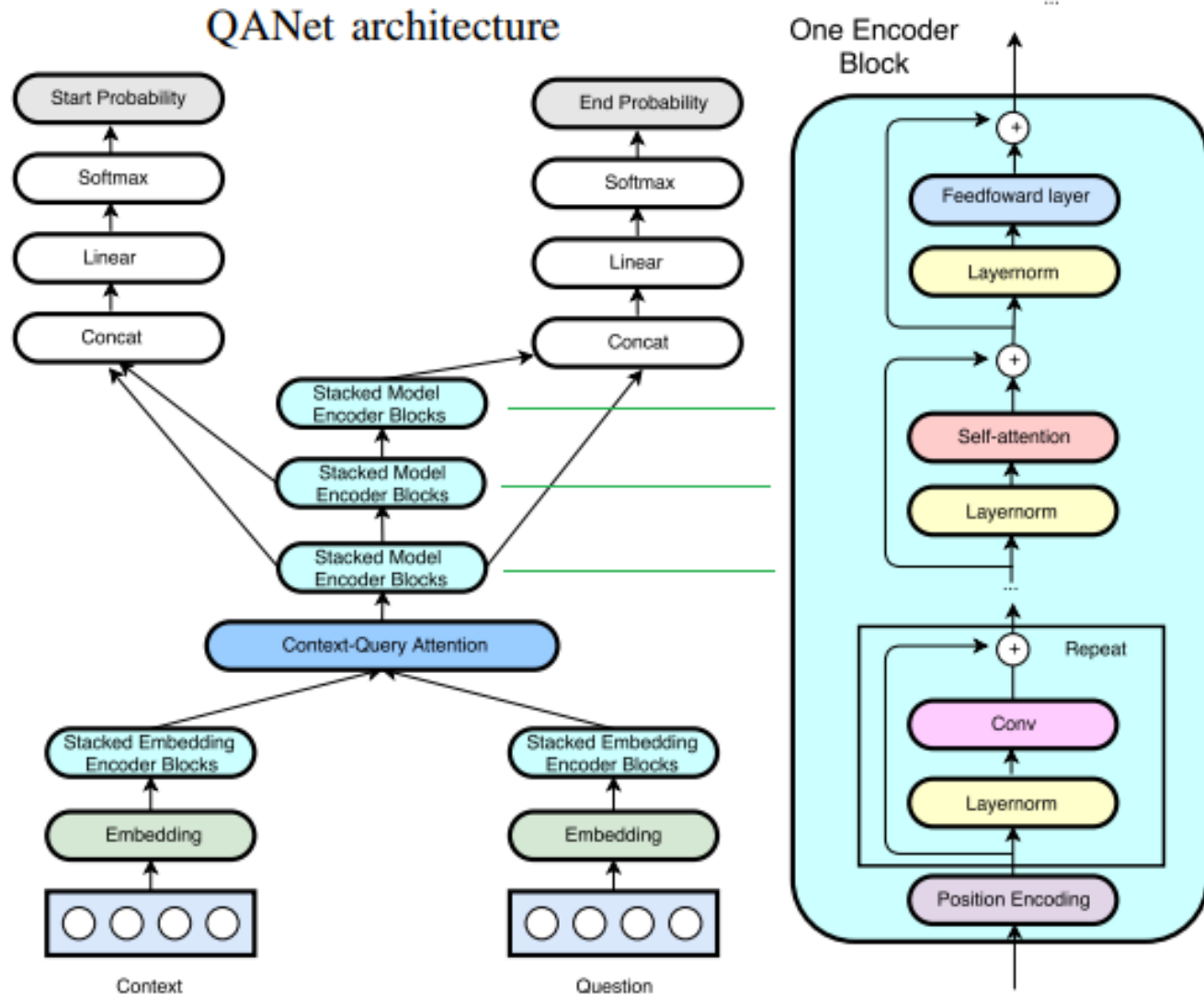
Rui Zhao, Kai Chen, Mohammad Norouzi, Quoc V. Le
Google Brain



Single Model	Published ¹² EM / F1
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5
RaSoR (Lee et al., 2016)	70.8 / 78.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9
ReasoNet (Shen et al., 2017b)	69.1 / 78.9
Document Reader (Chen et al., 2017)	70.0 / 79.0
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8
R-Net (Wang et al., 2017)	72.3 / 80.7
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8
Dev set: QANet	73.6 / 82.7
Dev set: QANet + data augmentation ×2	74.5 / 83.2
Dev set: QANet + data augmentation ×3	75.1 / 83.8
Test set: QANet + data augmentation ×3	76.2 / 84.6

Natural Language Processing

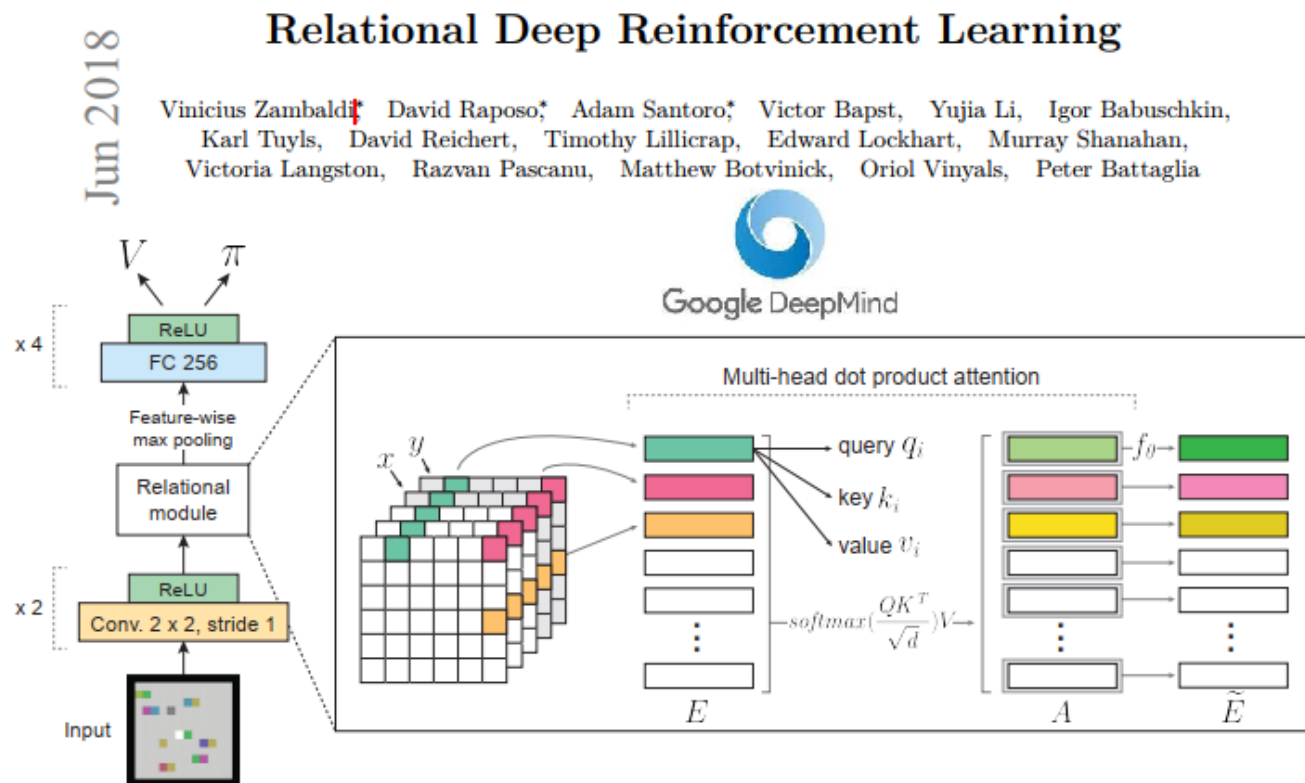
Self-Attention QANet



Natural Language Processing

Self-Attention

2018: DeepMind's Relational Deep Reinforcement Learning



Natural Language Processing

Non-local networks (Xiaolong Wang, 2018): general method for sequence processing, not only for NLP!

Non-local Neural Networks

Xiaolong Wang^{1,2*} Ross Girshick² Abhinav Gupta¹ Kaiming He²

¹Carnegie Mellon University

²Facebook AI Research



Figure 3. Examples of the behavior of a non-local block in res_3 computed by a 5-block non-local model trained on Kinetics. These examples

Natural Language Processing

Howard & Ruder: UMLFiT (2018):

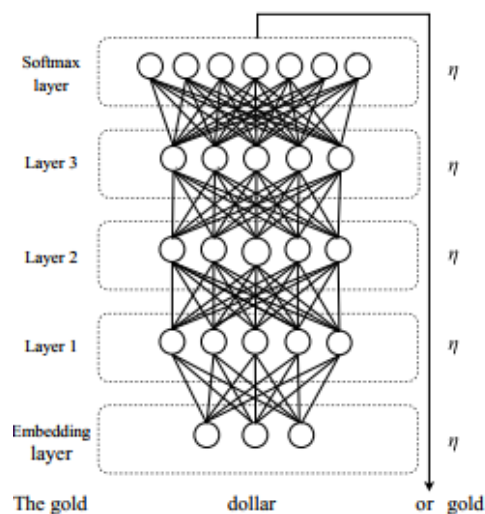
Universal Language Model Fine-tuning for Text Classification

May 2018

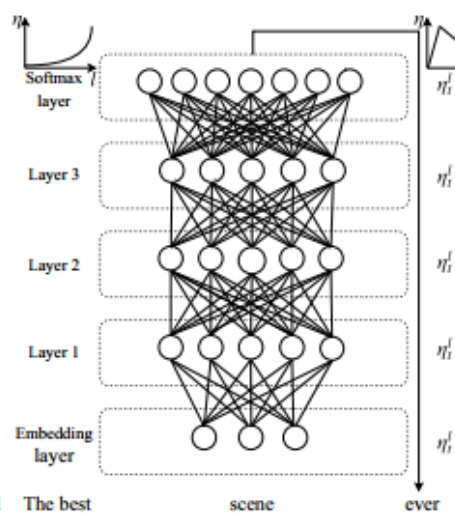
Jeremy Howard*
fast.ai
University of San Francisco



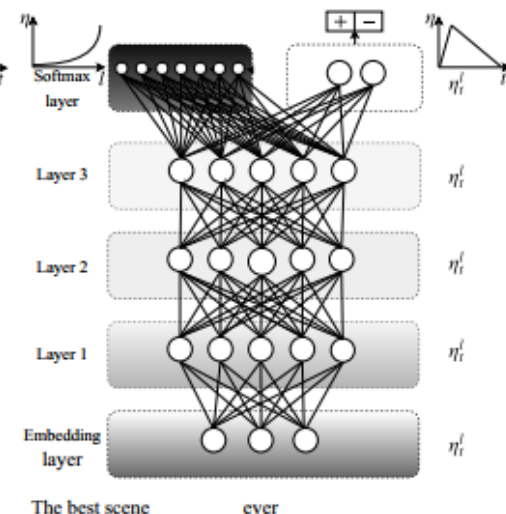
Sebastian Ruder*
Insight Centre, NUI Galway
Aylien Ltd., Dublin



(a) LM pre-training



(b) LM fine-tuning



(c) Classifier fine-tuning

Natural Language Processing

Transformer architecture
+ transfer learning for
pre-training:

OpenAI GPT (2018)

Google BERT (2018)

Oct 2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova



**Improving Language Understanding
by Generative Pre-Training**



OpenAI

Alec Radford Karthik Narasimhan Tim Salimans Ilya Sutskever

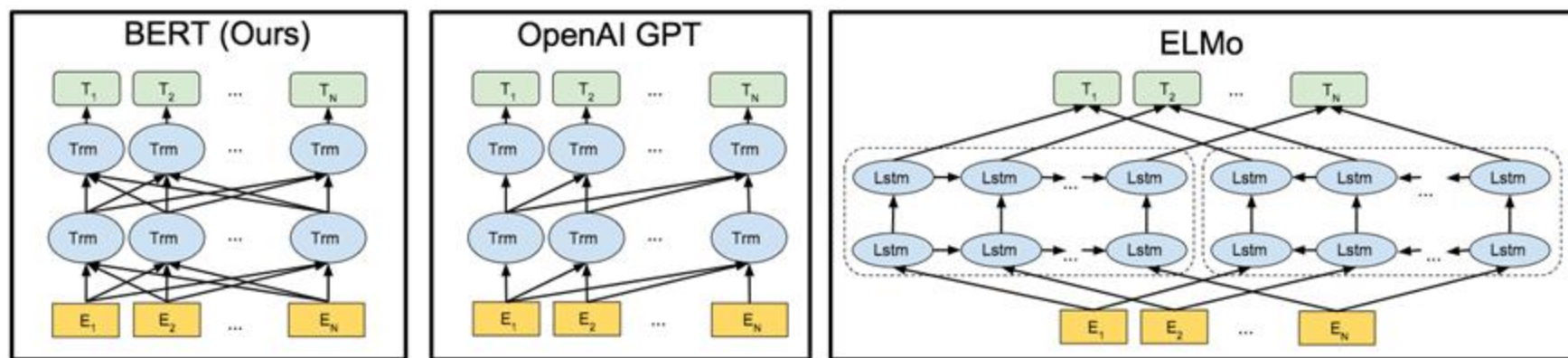


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks.

Natural Language Processing

Jacob Devlin's BERT (Google, 2018)



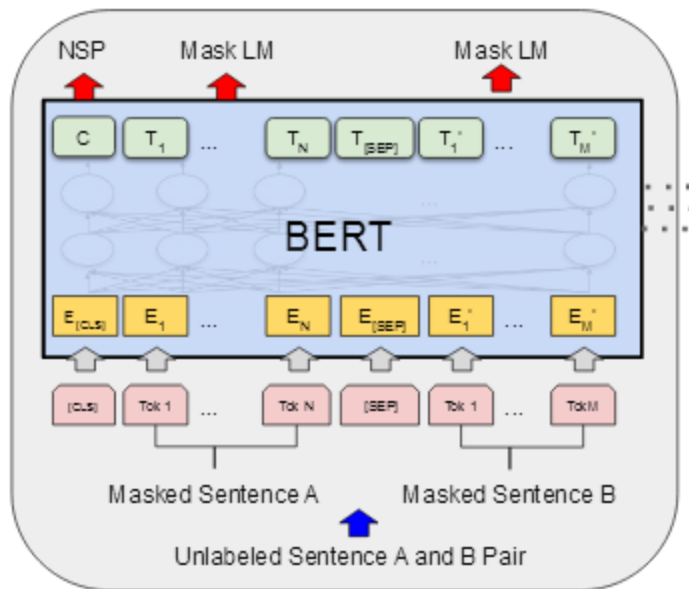
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language



Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing

Friday, November 2, 2018



Pre-training

Natural Language Processing

BERT's descendants:

- Luke Zettlemoyer's **RoBERTa** (Facebook + Univ of Washington, 2019) GLUE rating: 88.5 (dramatically better performance simply by training for longer)
- Chen Zhu's **FreeLB** (Univ of Maryland + Microsoft, 2019) 88.8

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu^{‡§} Myle Ott^{‡§} Naman Goyal^{‡§} Jingfei Du^{‡§} Mandar Joshi[†]
Danqi Chen[§] Omer Levy[§] Mike Lewis[§] Luke Zettlemoyer^{†§} Veselin Stoyanov[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA

[§] Facebook AI



Jul 2019

Natural Language Processing

BERT's descendants:

- StructBERT (Alibaba, 2019)
- Zhenzhong Lan's ALBERT, a lite BERT (Google, 2019)
- Electra (Stanford, 2020)
- DeBERTa (Microsoft, 2020)

ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS

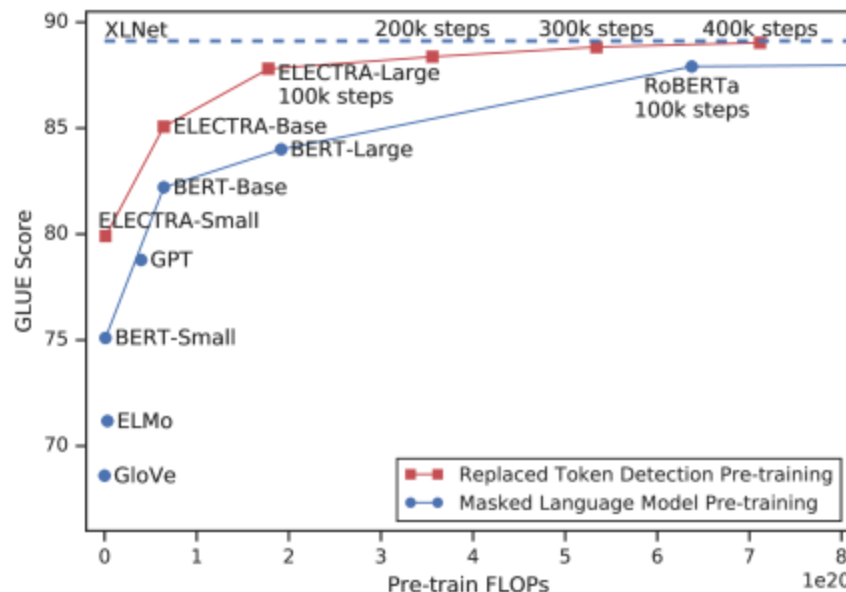
Zhenzhong Lan¹ Mingda Chen^{2*} Sebastian Goodman¹ Kevin Gimpel²

Piyush Sharma¹ Radu Soricut¹

¹Google Research

²Toyota Technological Institute at Chicago

Sep 2019



Natural Language Processing

GLUE scores (2019)



Rank	Name	Model	Score
1	ALBERT-Team Google Language	ALBERT (Ensemble)	89.4
2	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)	89.0
3	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	88.8
4	Facebook AI	RoBERTa	88.5
5	XLNet Team	XLNet-Large (ensemble)	88.4

Natural Language Processing

SuperGLUE scores (2021)

Leaderboard Version: **2.0**

Model	Score	BoolQ
DeBERTa / TuringNLRv4	90.3	90.4
T5 + Meena, Single Model (Meena Team - Google Brain)	90.2	91.3
SuperGLUE Human Baselines	89.8	89.0
T5	89.3	91.2
NEZHA-Plus	86.7	87.8
PAI Albert	86.1	88.1
RoBERTa (ensemble)	85.9	88.2
RoBERTa-mtl-adv	85.7	87.1
RoBERTa	84.6	87.1

Model	Score	BoolQ	CB	COPA
DeBERTa / TuringNLRv4	90.3	90.4	95.7/97.6	98.4
T5 + Meena, Single Model (Meena Team - Google Brain)	90.2	91.3	95.8/97.6	97.4
SuperGLUE Human Baselines	89.8	89.0	95.8/98.9	100.0
T5	89.3	91.2	93.9/96.8	94.8
NEZHA-Plus	86.7	87.8	94.4/96.0	93.6
PAI Albert	86.1	88.1	92.4/96.4	91.8
RoBERTa (ensemble)	85.9	88.2	92.5/95.6	90.8
RoBERTa-mtl-adv	85.7	87.1	92.4/95.6	91.2
RoBERTa	84.6	87.1	90.5/95.2	90.6

Natural Language Processing

BERT doubters (2019):

- Timothy Niven and Hung-Yu Kao
(Taiwan's National Cheng Kung University)
- Tal Linzen (Johns Hopkins University)

Probing Neural Network Comprehension of Natural Language Arguments

Timothy Niven and Hung-Yu Kao

Intelligent Knowledge Management Lab
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, Taiwan

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy,¹ Ellie Pavlick,² & Tal Linzen¹

¹Department of Cognitive Science, Johns Hopkins University

²Department of Computer Science, Brown University

Jun 2019

Natural Language Processing

Pre-training: training the language representation with several simple tasks in order to grasp the co-occurrence of words or sentences

Pre-training:

- Autoencoding-based (BERT)
- Autoregressive-based (XLNet)

Natural Language Processing

XLNet (Ruslan Salakhutdinov, 2019):
Transformer-XL, autoregressive)

19 Jun 2019

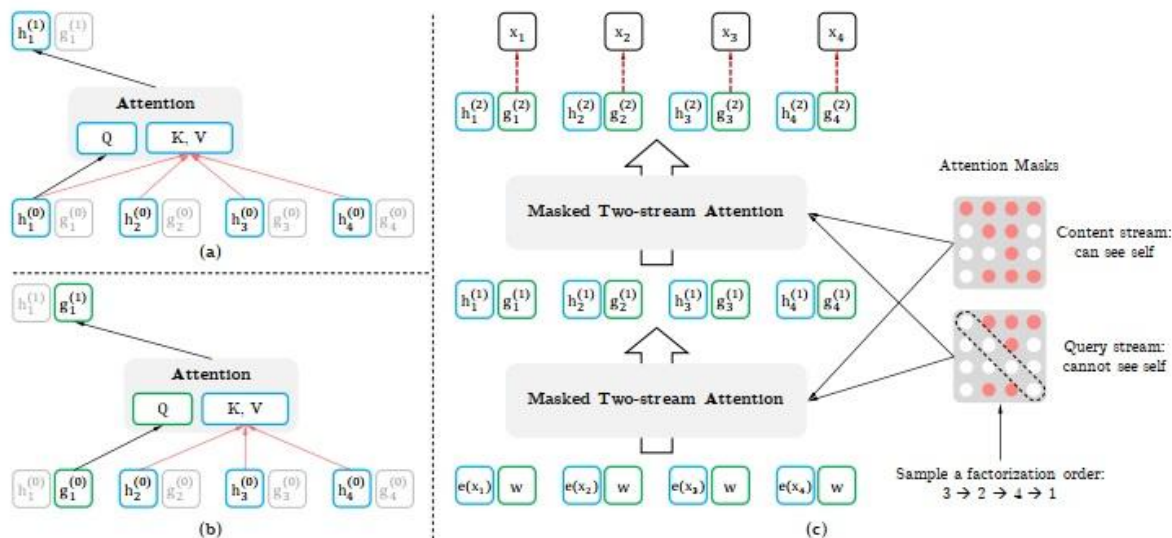
XLNet: Generalized Autoregressive Pretraining for Language Understanding



Zhilin Yang^{*1}, Zihang Dai^{*12}, Yiming Yang¹, Jaime Carbonell¹,
Ruslan Salakhutdinov¹, Quoc V. Le²
¹Carnegie Mellon University, ²Google Brain



2.3 Architecture: Two-Stream Self-Attention for Target-Aware Representations



Natural Language Processing

Evolved Transformer (Google, 2019)

2019

The Evolved Transformer

David R. So¹ Chen Liang¹ Quoc V. Le¹



Transformer and Evolved Transformer architecture

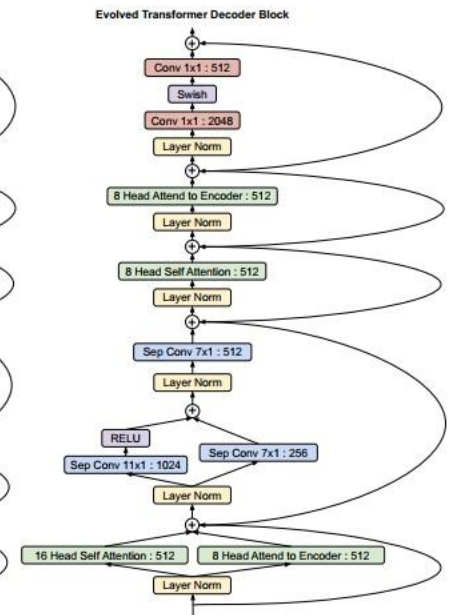
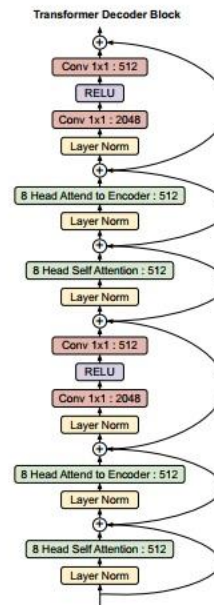
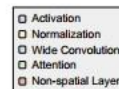
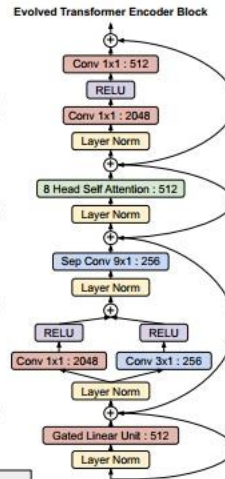
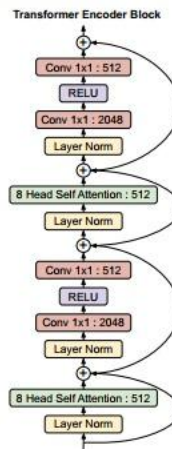
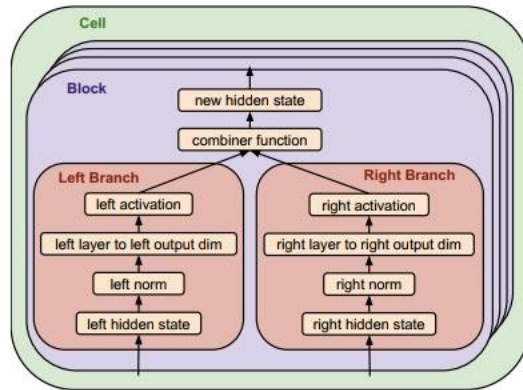


Figure 1. **Architecture composition from encoding.** Each block produces a new hidden state that is added to the pool of hidden states subsequent blocks can select as branch inputs. Each encoder has 6 unique blocks per cell and each decoder has 8 unique blocks per cell. Each cell is repeated *number of cells* times.

Natural Language Processing

ERNIE 2.0 (Hua Wu,
Baidu, 2019)

ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang

2019 **Baidu Research**

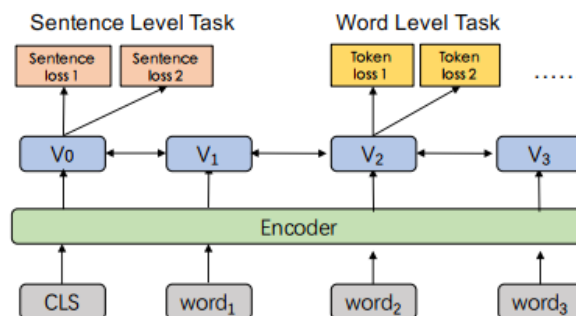
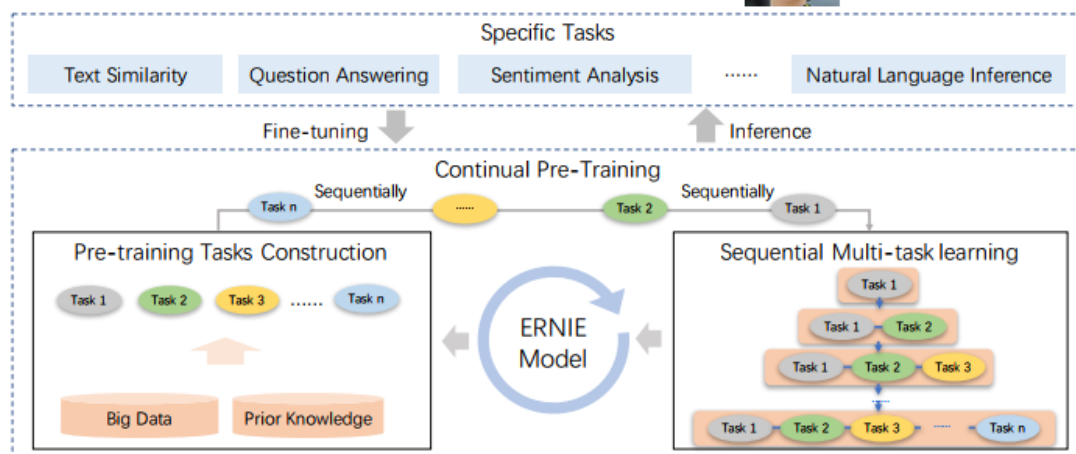


Figure 4: The architecture of multi-task learning in the ERNIE 2.0 framework, in which the encoder can be recurrent neural networks or a deep transformer.

Natural Language Processing

Main pre-trained models in 2019:

- ELMo
- OpenAI's GPT
- Google's BERT
- Microsoft's XLNet
- Baidu's Ernie 2.0



Dec 11th , 2019

Baidu's Pre-training Model ERNIE Achieves New NLP Benchmark Record

ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang
Baidu Inc., Beijing, China

Name	Model	Score
ERNIE Team - Baidu	ERNIE	90.1
Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART	89.9
T5 Team - Google	T5	89.7
王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)	89.5
XLNet Team	XLNet (ensemble)	89.5
ALBERT-Team Google Language	ALBERT (Ensemble)	89.4
Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	88.8
Facebook AI	RoBERTa	88.5

Natural Language Processing

Question-answering

Danqi Chen's DrQA (Facebook, 2017): multitask learning using distant supervision

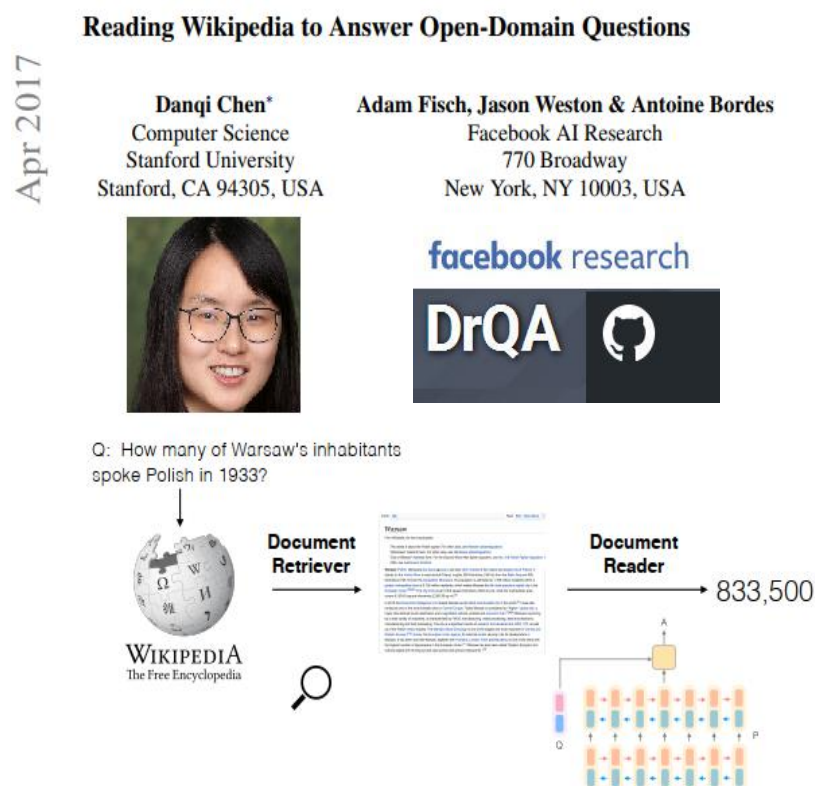


Figure 1: An overview of our question answering system DrQA.

Natural Language Processing

Question-answering

FlowQA (Allen Institute, 2018)



FLOWQA: GRASPING FLOW IN HISTORY FOR CONVERSATIONAL MACHINE COMPREHENSION

Hsin-Yuan Huang*
California Institute of Technology

Wen-tau Yih
Allen Institute for Artificial Intelligence

Eunsol Choi
University of Washington

Published as a conference paper at ICLR 2019

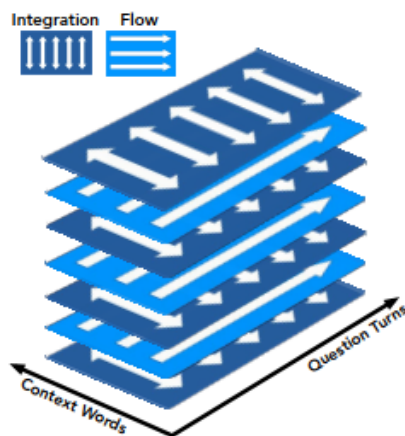


Figure 3: Alternating computational structure between context integration (RNN over context) and FLOW (RNN over question turns).

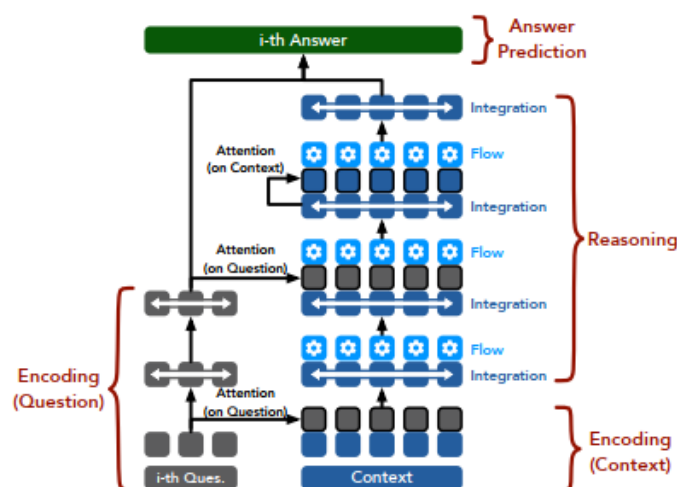


Figure 4: An illustration of the architecture for FLOWQA.

Natural Language Processing

Question-answering

SDNet (Microsoft, 2018)

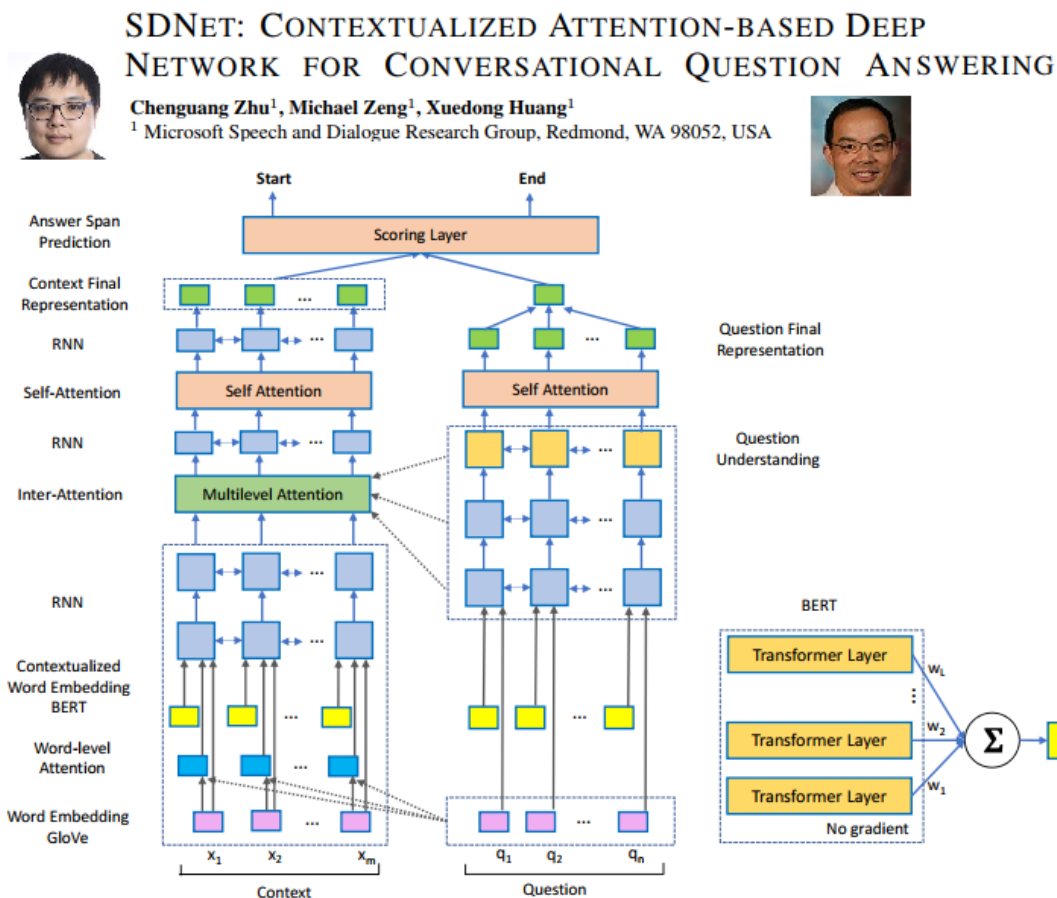


Figure 1: SDNet model structure.

Natural Language Processing

Question-answering

BiDAF++ (Mark Yatskar, Allen Inst, 2018)



Natural Language Processing

Visual Question-answering

Anton van den Hengel's team at University of
Adelaide (2017)

Zichao Yang at CMU in collaboration with Microsoft
(2016)

Jiasen Lu at Virginia Tech (2016), the paper that
introduced co-attention

Josh Tenenbaum's team at MIT, in collaboration with
IBM and DeepMind: NS-CL (2019), capable of
learning about the world just as a child does: by
looking around and talking

Natural Language Processing

Visual Question-answering

Josh Tenenbaum's NS-CL (2019)

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding



14 Jan 2019

Kexin Yi*
Harvard University

Jiajun Wu*
MIT CSAIL

Chuang Gan
MIT-IBM Watson AI Lab

Antonio Torralba
MIT CSAIL

Pushmeet Kohli
DeepMind

Joshua B. Tenenbaum
MIT CSAIL



How many blocks are on the right of the three-level tower?



Will the block tower fall if the top block is removed?



What is the shape of the object closest to the large cylinder?



Are there more trees than animals?

Figure 1: Human reasoning is interpretable and disentangled: we first draw abstract knowledge of the scene via visual perception and then perform logic reasoning on it.

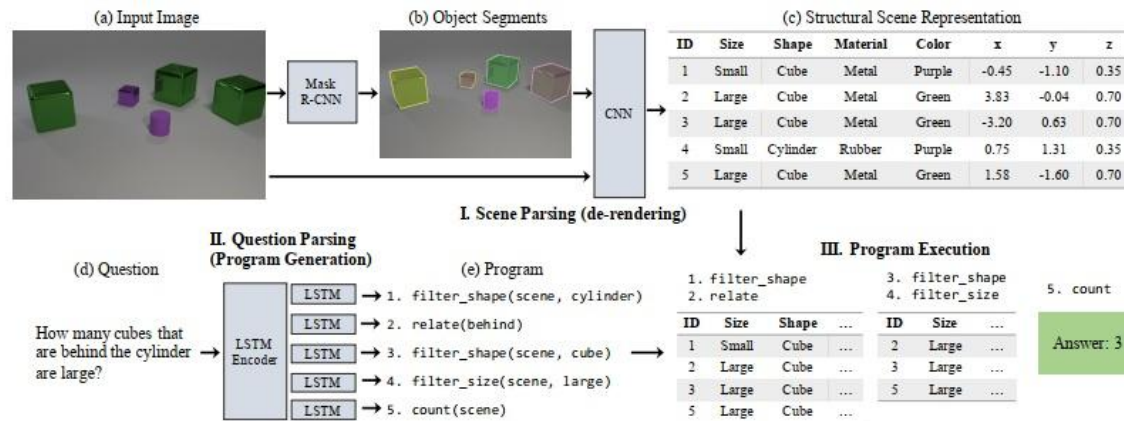


Figure 2: Our model has three components: first, a scene parser (de-renderer) that segments an input image (a-b) and recovers a structural scene representation (c); second, a question parser (program generator) that converts a question in natural language (d) into a program (e); third, a program executor that runs the program on the structural scene representation to obtain the answer.

Natural Language Processing

Unsupervised learning of language use: Word2vec, GloVe, ELMo, Skip-Thoughts...

Supervised learning: InferSent (Alexis Conneau, Antoine Bordes - Facebook)



Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

Alexis Conneau
Facebook AI Research

Douwe Kiela
Facebook AI Research

Holger Schwenk
Facebook AI Research

Loïc Barrault
LIUM, Université Le Mans

Antoine Bordes
Facebook AI Research

Natural Language Processing

Supervised learning of language representation:
Sandeep Subramanian (University of Montreal, 2018)

Published as a conference paper at ICLR 2018

LEARNING GENERAL PURPOSE DISTRIBUTED SENTENCE REPRESENTATIONS VIA LARGE SCALE MULTI-TASK LEARNING

Sandeep Subramanian^{1,2,3*}, Adam Trischler³, Yoshua Bengio^{1,2,4} & Christopher J Pal^{1,5}

¹ Montréal Institute for Learning Algorithms (MILA)

² Université de Montréal

³ Microsoft Research Montreal

⁴ CIFAR Senior Fellow

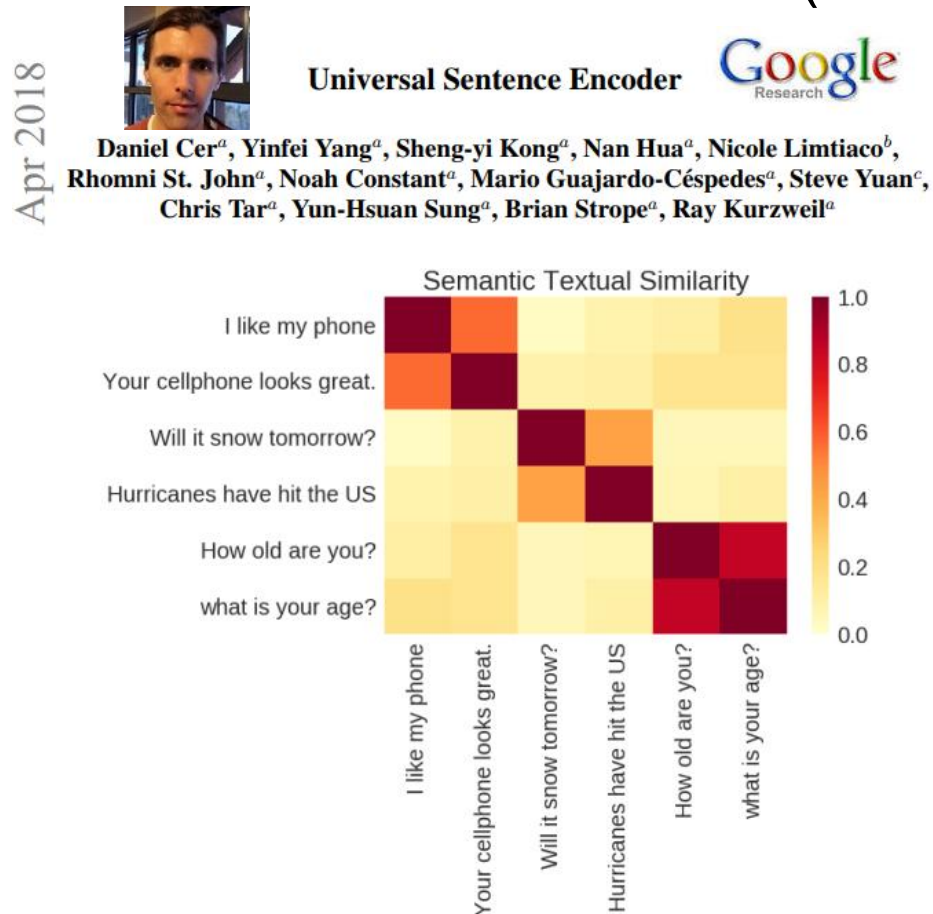
⁵ École Polytechnique de Montréal



Natural Language Processing

Supervised learning of language representation:

Daniel Cer's Universal Sentence Encoder (Google, 2018)



Natural Language Processing

Sentiment Analysis

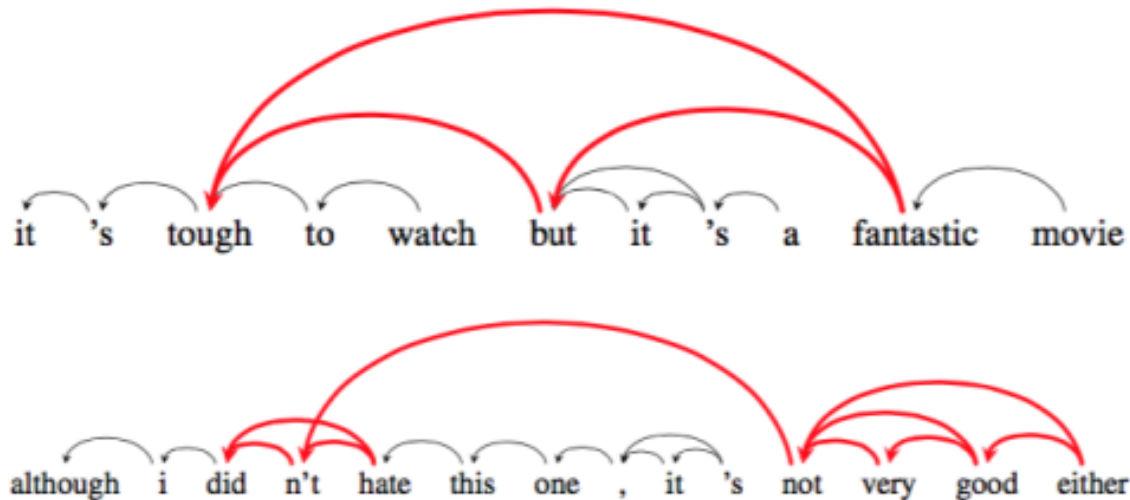


Figure 5: Examples of intra-attention (sentiment analysis). Bold lines (red) indicate attention between sentiment important words. Jianpeng Cheng (2016)

Natural Language Processing

Sentiment Analysis

- Kai Sheng Tai & Richard Socher (2015)

May 2015

Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks

Kai Sheng Tai, Richard Socher*, Christopher D. Manning

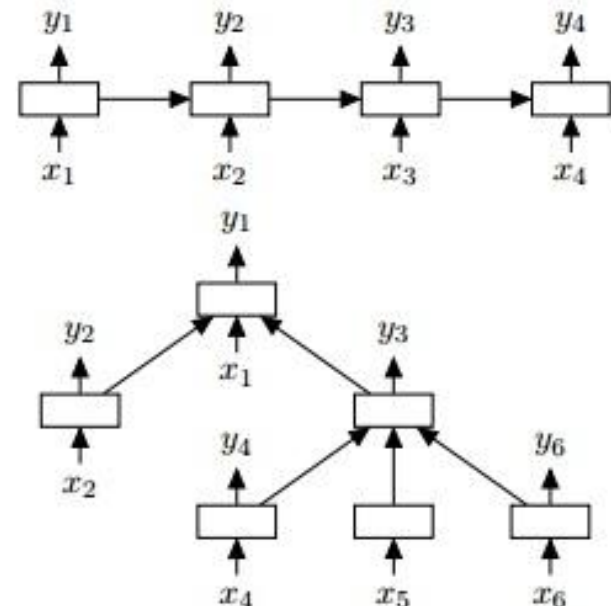
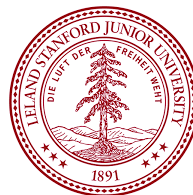


Figure 1: **Top:** A chain-structured LSTM network. **Bottom:** A tree-structured LSTM network with arbitrary branching factor.



Natural Language Processing

Sentiment Analysis

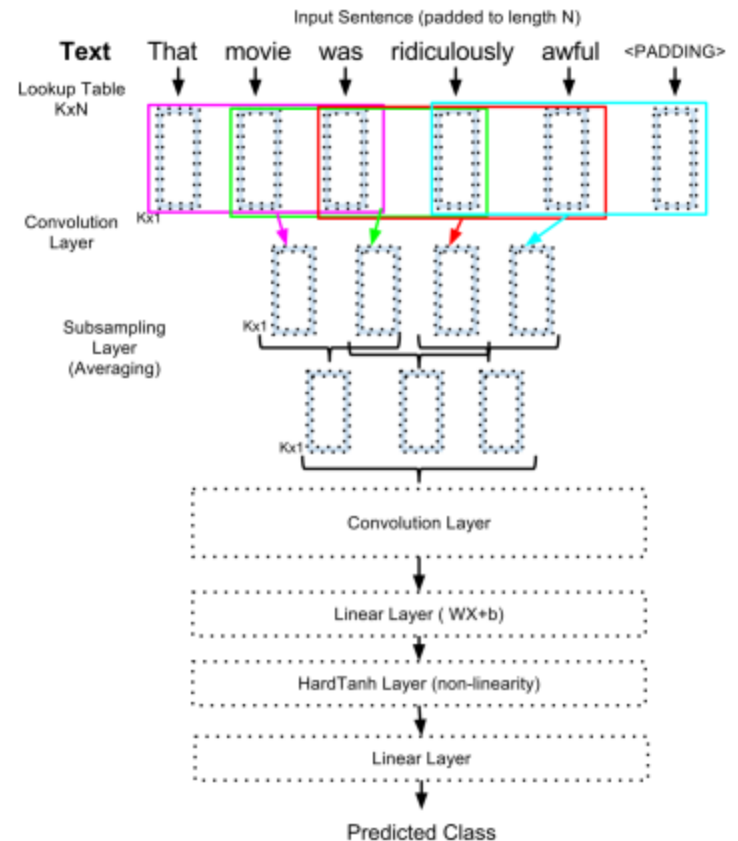
- Soumith Chintala (2015)



Sentiment Analysis using neural architectures

2015

Soumith Chintala
New York University



Natural Language Processing

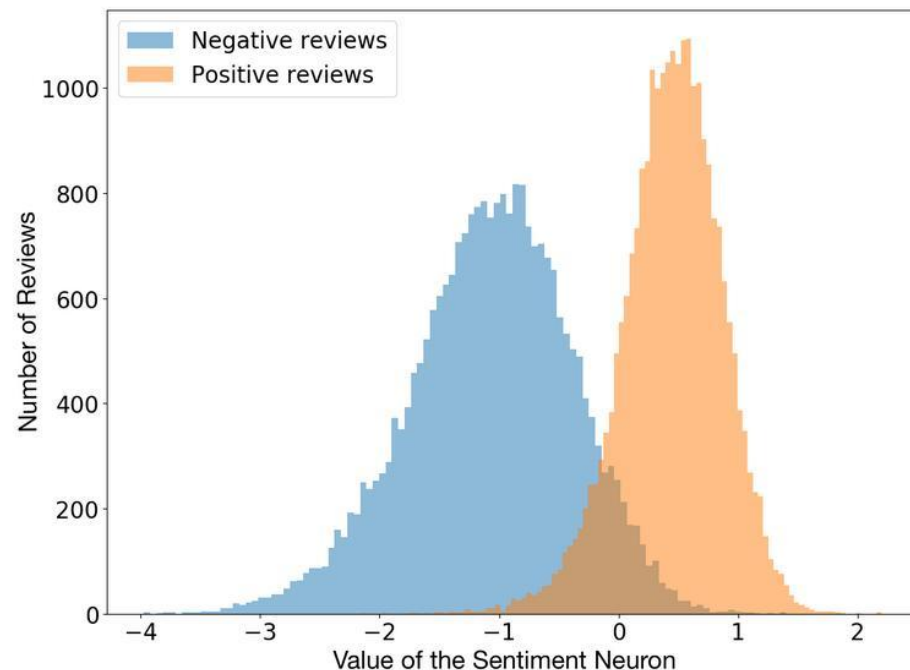
Sentiment Analysis

- 2017: Alec Radford (OpenAI) discovers the “sentiment neuron” in LSTM networks.
- Trained (with 82 million Amazon reviews) to predict the next character in the text of Amazon reviews, the network develops a “sentiment neuron” that predicts the sentiment value of the review

Apr 2017

Learning to Generate Reviews and Discovering Sentiment

Alec Radford¹ Rafal Jozefowicz¹ Ilya Sutskever¹



Natural Language Processing

Discourse Analysis

2016: Minjoon Seo's Bidirectional Attention Flow (BiDAF) model

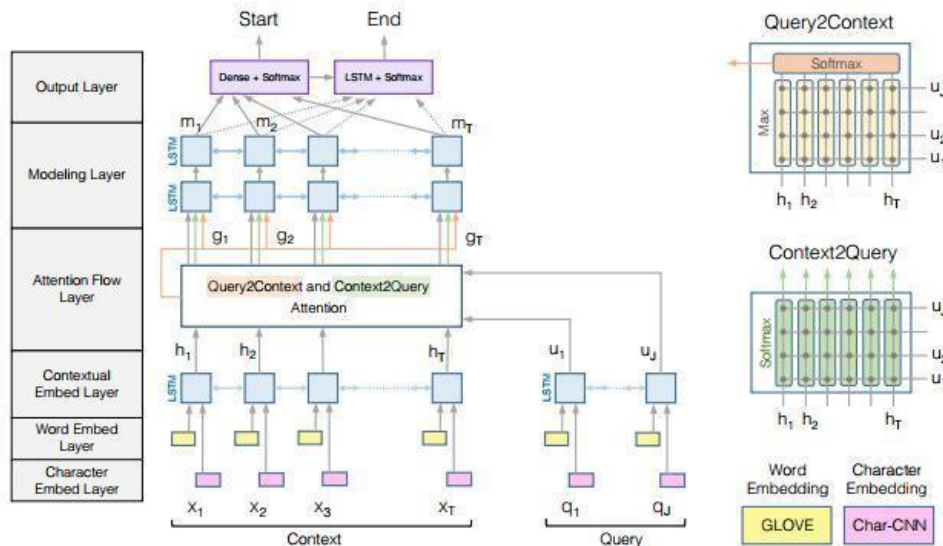


Figure 1: BiDirectional Attention Flow Model (best viewed in color)

Natural Language Processing

Discourse Analysis

2016: Percy Liang's SQuAD dataset

2016: Jianfeng Gao's MARCO
dataset



Natural Language Processing

Discourse Analysis

2016: Weizhu Chen's

Reasonet combines
memory networks with
reinforcement learning

2017: Weizhu Chen's

FusionNet simpler attention
mechanism called "History
of Word"



FUSIONNET: FUSING VIA FULLY-AWARE ATTENTION WITH APPLICATION TO MACHINE COMPREHENSION

Hsin-Yuan Huang^{*1,2}, Chenguang Zhu¹, Yelong Shen¹, Weizhu Chen¹

¹Microsoft Business AI and Research

	Test Set
<i>Single Model</i>	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0
Match-LSTM (Wang & Jiang, 2016)	64.7 / 73.7
BiDAF (Seo et al., 2017)	68.0 / 77.3
SEDT (Liu et al., 2017)	68.2 / 77.5
RaSoR (Lee et al., 2016)	70.8 / 78.7
DrQA (Chen et al., 2017a)	70.7 / 79.4
ReasoNet (Shen et al., 2017)	70.6 / 79.4
R. Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8
DCN+	74.9 / 82.8
R-net [†]	75.7 / 83.5
FusionNet	76.0 / 83.9
<i>Ensemble Model</i>	
ReasoNet (Shen et al., 2017)	75.0 / 82.3
MEMEN (Pan et al., 2017)	75.4 / 82.7
R. Mnemonic Reader (Hu et al., 2017)	77.7 / 84.9
R-net [†]	78.2 / 85.2
DCN+	78.7 / 85.6
FusionNet	78.8 / 85.9
Human (Rajpurkar et al., 2016)	82.3 / 91.2

Table 2: The performance of FusionNet and competing approaches on SQuAD hidden test set at the time of writing (Oct. 4th, 2017).

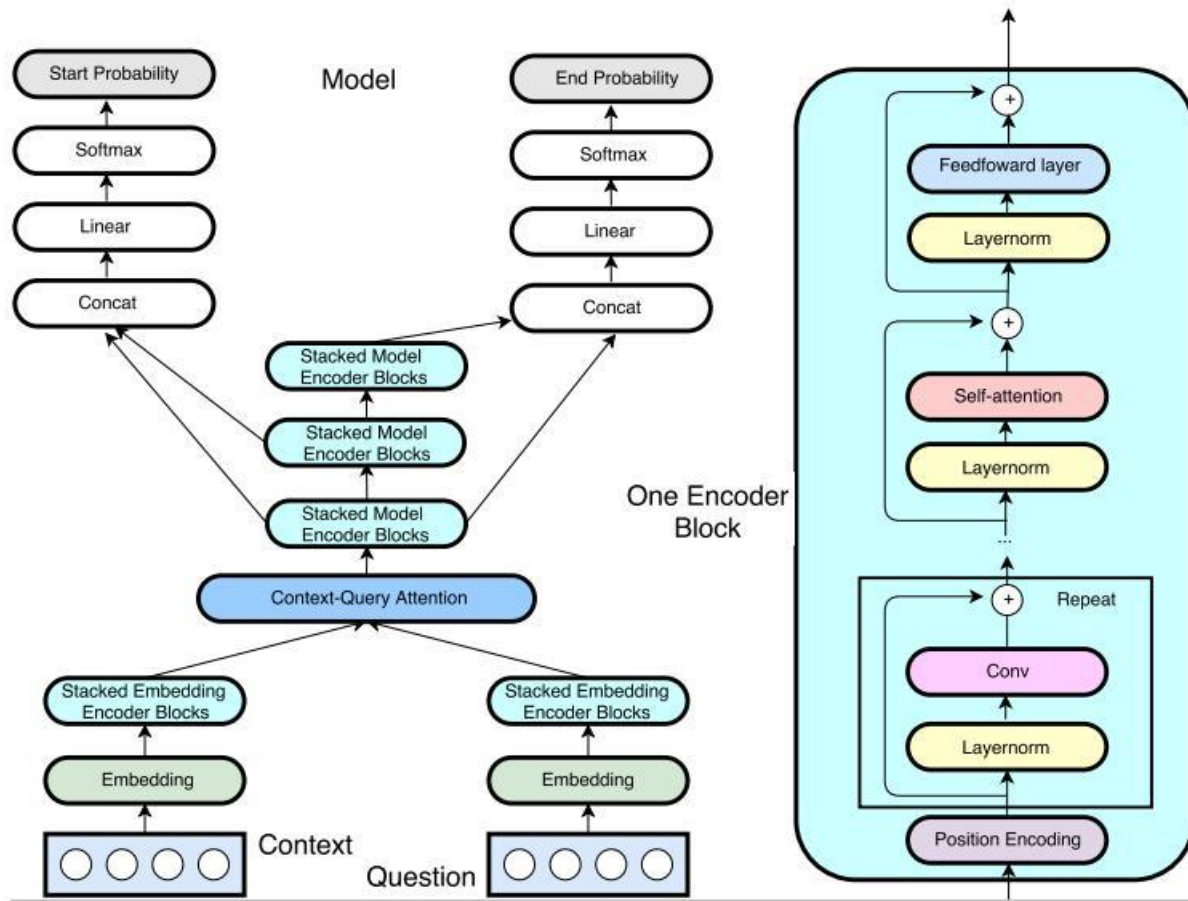
Natural Language Processing

Discourse Analysis 2017: Quoc Le's QANET



Adams Wei Yu^{1*}, David Dohan^{2†}, Minh-Thang Luong^{2†}
{weiyu}@cs.cmu.edu, {ddohan, thangluong}@google.com
¹Carnegie Mellon University, ²Google Brain ICLR 2018

Rui Zhao, Kai Chen, Mohammad Norouzi, Quoc V. Le
Google Brain



Natural Language Processing

Discourse Analysis

2018: Furu Wei's R-Net for reading-comprehension

GeekWire

Microsoft and Alibaba AI programs beat humans in Stanford reading comprehension test for 1st time

BY **NAT LEVY** on January 15, 2018 at 2:57 pm



Natural Language Processing

Text Comprehension

	Published ¹²	LeaderBoard ¹³
Single Model	EM / F1	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8 / 78.7	69.6 / 77.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9	70.8 / 78.9
ReasonNet (Shen et al., 2017b)	69.1 / 78.9	70.6 / 79.4
Document Reader (Chen et al., 2017)	70.0 / 79.0	70.7 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8	70.6 / 79.8
Conductor-net	N/A	72.6 / 81.4
Interactive AoA Reader (Cui et al., 2017)	N/A	73.6 / 81.9
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 / 84.3
BiDAF + Self Attention + ELMo	N/A	77.9 / 85.3
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8	73.2 / 81.8
human performance = 82.3		
Dev set: QANet	73.6 / 82.7	N/A
Dev set: QANet + data augmentation ×2	74.5 / 83.2	N/A
Dev set: QANet + data augmentation ×3	75.1 / 83.8	N/A
Test set: QANet + data augmentation ×3	76.2 / 84.6	76.2 / 84.6

The performances of different models on SQuAD dataset.

Natural Language Processing

Summarization

Enabled by

- Sequence-to-sequence (Seq2Seq) that can both read AND write
- Pointer networks (Vinyals & Fortunato, 2015): attention-based seq2seq

Natural Language Processing

Summarization

Pointer networks (Vinyals & Fortunato, 2015)

Pointer Networks

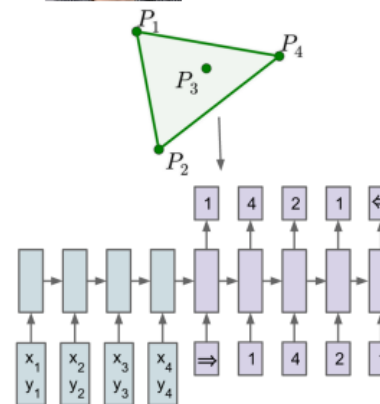
Oriol Vinyals*
Google Brain



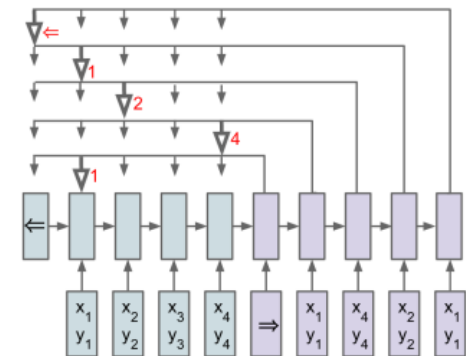
Meire Fortunato*
Department of Mathematics, UC Berkeley



Navdeep Jaitly
Google Brain



(a) Sequence-to-Sequence



(b) Ptr-Net

Natural Language Processing

Summarization

- Alexander Rush & Jason Weston at (Facebook, 2015) first to apply seq2seq to summarization
- Read-Again Summarization (Raquel Urtasun & Wenjuan Zeng, Univ of Toronto, 2016)
- Forced Attention Sentence Compression Model (Phil Blunsom & Yishu Miao at Oxford, 2016)
- IBM's SummaRunner (Ramesh Nallapati, 2017)
- Pointer-generator Network PGNET (Abigail See & Christopher Manning at Stanford, 2017)

Natural Language Processing

Summarization

Alexander Rush, Sumit Chopra and Jason Weston (Facebook, 2015): ABS

- first to apply seq2seq to summarization
- attention-based summarization

A Neural Attention Model for Abstractive Sentence Summarization

Sep 2015

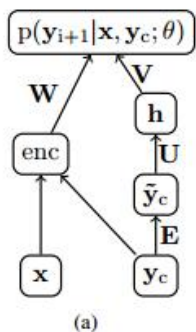
Alexander M. Rush
Facebook AI Research /



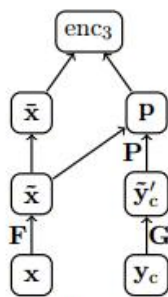
Sumit Chopra
Facebook AI Research



Jason Weston
Facebook AI Research



(a)



(b)

Model	ROUGE-1
IR	11.06
PREFIX	22.43
COMPRESS	19.77
W&L	22
TOPIARY	25.12
MOSES+	26.50
ABS	26.55
ABS+	28.18
REFERENCE	29.21

Figure 3: (a) A network diagram for the NNLM decoder with additional encoder element. (b) A network diagram for the attention-based encoder enc_3 .

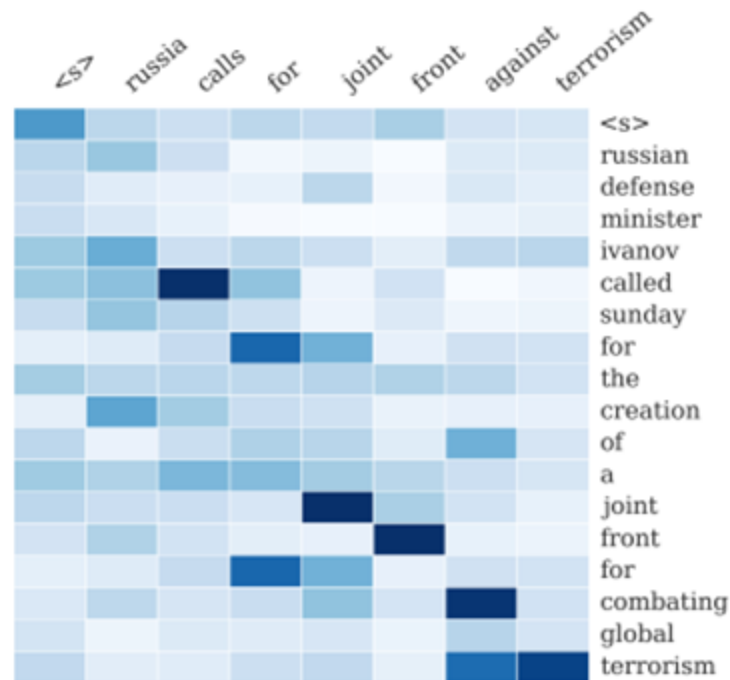


Figure 1: Example output of the attention-based summarization (ABS) system. The heatmap represents a soft alignment between the input (right) and the generated summary

Natural Language Processing

Summarization

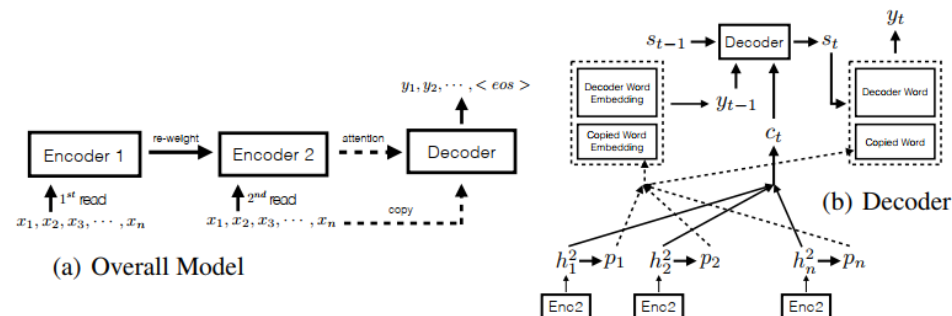
Read-Again Summarization (Raquel Urtasun & Wenyuan Zeng, Univ of Toronto, 2016)



EFFICIENT SUMMARIZATION WITH READ-AGAIN AND COPY MECHANISM



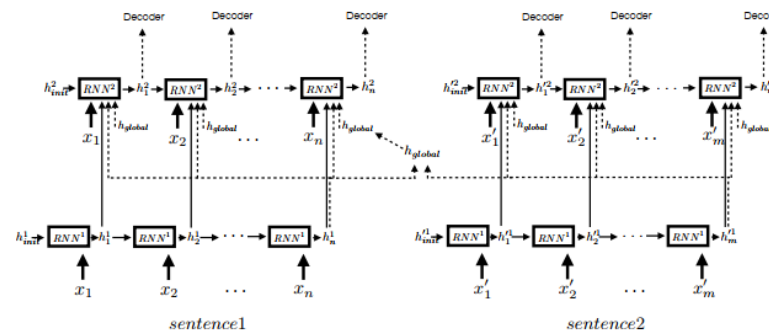
Wenyan Zeng¹, Wenjie Luo², Sanja Fidler², Raquel Urtasun¹
¹Tsinghua University, ²University of Toronto



(a) Overall Model

(b) Decoder

Hierarchical Read-Again



Natural Language Processing

Summarization

Forced Attention Sentence Compression Model (Phil Blunsom & Yishu Miao at Oxford, 2016)



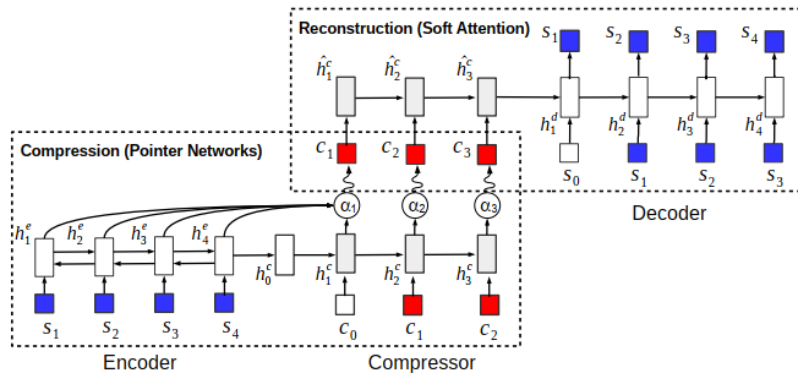
Language as a Latent Variable:
Discrete Generative Models for Sentence Compression

Yishu Miao¹, Phil Blunsom^{1,2}

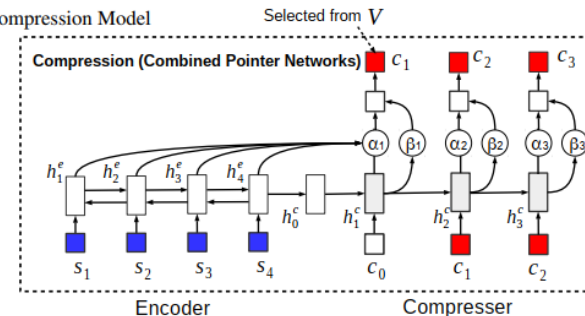
¹University of Oxford, ²Google Deepmind



Auto-encoding Sentence Compression Model



Forced Attention Sentence Compression Model



Natural Language Processing

Summarization

IBM's SummaRunner (Ramesh Nallapati, 2017)



SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents

Nov 2016



Ramesh Nallapati, Feifei Zhai*, Bowen Zhou
1011 Kitchawan Road, Yorktown Heights, NY 10598

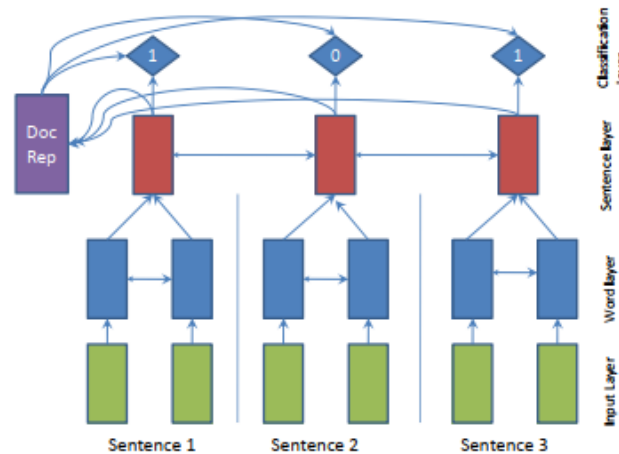


Figure 1: SummaRuNNer: A two-layer RNN based sequence classifier: the bottom layer operates at word level within each sentence, while the top layer runs over sentences.

Natural Language Processing

Summarization

Pointer-generator Network PGNET (Abigail See & Christopher Manning at Stanford, 2017) for longer summaries



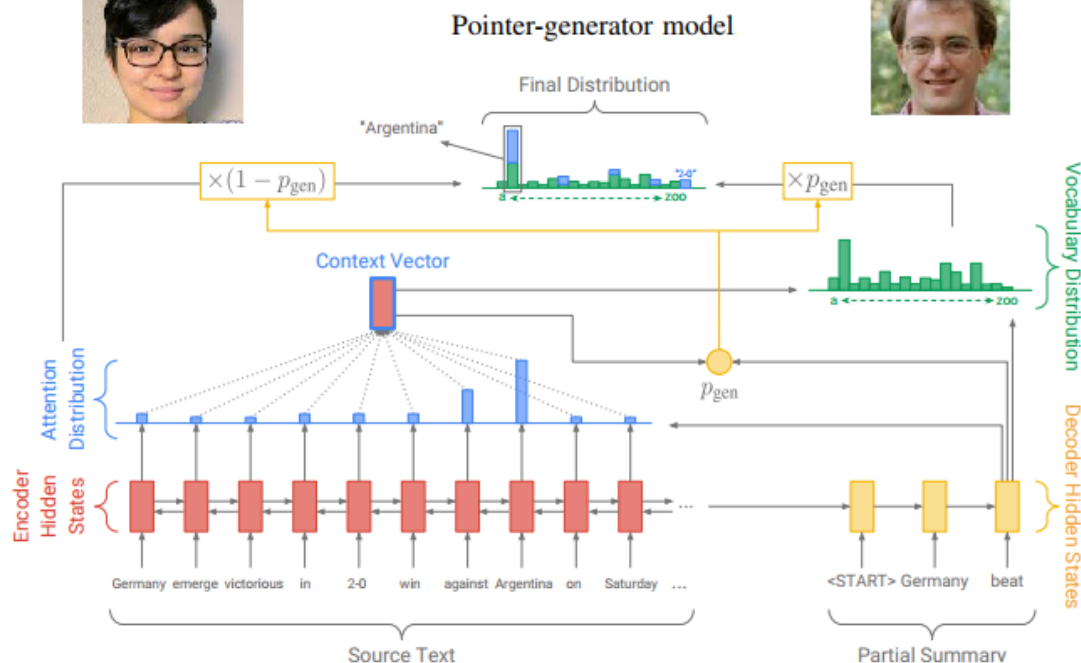
Get To The Point: Summarization with Pointer-Generator Networks

Abigail See
Stanford University



Peter J. Liu
Google Brain

Christopher D. Manning
Stanford University



Natural Language Processing

Simplification

- Yuta Kikuchi (Japan, 2016): controllable text generation
- Sergiu Nisioi and Sanja Stajner (Romania & Germany, 2017)
- Daiki Nishihara (Japan, 2019)

Natural Language Processing

Simplification

- Eric de la Clergerie (Inria, 2019): ACCESS



OCTOBER 14, 2019

Facebook's AI streamlines sentences while preserving meaning

Controllable Sentence Simplification

Louis Martin
Facebook AI Research & Inria
louismartin@fb.com

Benoît Sagot
Inria
benoit.sagot@inria.fr

Éric de la Clergerie
Inria
eric.de_la_clergerie@inria.fr

Antoine Bordes
Facebook AI Research
abordes@fb.com

Source	Some trails are designated as nature trails , and are used by people learning about the natural world .
NbChars _{1.00}	Some trails are called nature trails , and are used by people about the natural world .
NbChars _{0.75}	Some trails are called nature trails , and are used by people about the natural world .
NbChars _{0.50}	Some trails are used by people about the natural world .
NbChars _{0.25}	Some trails are used by people .
LevSim _{1.00} +NbChars _{1.00}	Some trails are designated as nature trails , and are used by people learning about the natural world .
LevSim _{0.75} +NbChars _{1.00}	Some trails are made for nature trails . They are used by people who learn about the natural world .
LevSim _{0.50} +NbChars _{1.00}	The trails that are used by people learning about the natural world , because the trails are good trails .
LevSim _{0.25} +NbChars _{1.00}	Mechanical trails (also known as " trail trail " or " trails ") are trails that are used for trails .
WordRank _{1.00} +NbChars _{1.00}	Some trails are designated as nature trails , and are used by people learning about the natural world .
WordRank _{0.75} +NbChars _{1.00}	Some trails are called nature trails , and are used by people learning about the natural world .
WordRank _{0.50} +NbChars _{1.00}	Some trails are known as nature trails , and are used by people as well as by people who are in the world .
WordRank _{0.25} +NbChars _{1.00}	Some trails are also called nature trails , and are used by people learning about the natural world .
DepTreeDepth _{1.00} +NbChars _{1.00}	Some trails are designated as nature trails , and are used by people learning about the natural world .
DepTreeDepth _{0.75} +NbChars _{1.00}	Some trails are designated as nature trails . They are used by people learning about the natural world .
DepTreeDepth _{0.50} +NbChars _{1.00}	Some trails are designated as nature trails . They are used by people learning about the natural world .
DepTreeDepth _{0.25} +NbChars _{1.00}	Some trails are designated as nature trails . They are used by people to learn about the natural world .

Natural Language Processing

Text Generation

OpenAI's GPT2 (2019)

Language Models are Unsupervised Multitask Learners

Alec Radford ^{*1} Jeffrey Wu ^{*1} Rewon Child ¹ David Luan ¹ Dario Amodei ^{**1} Ilya Sutskever ^{**1}



Natural Language Processing

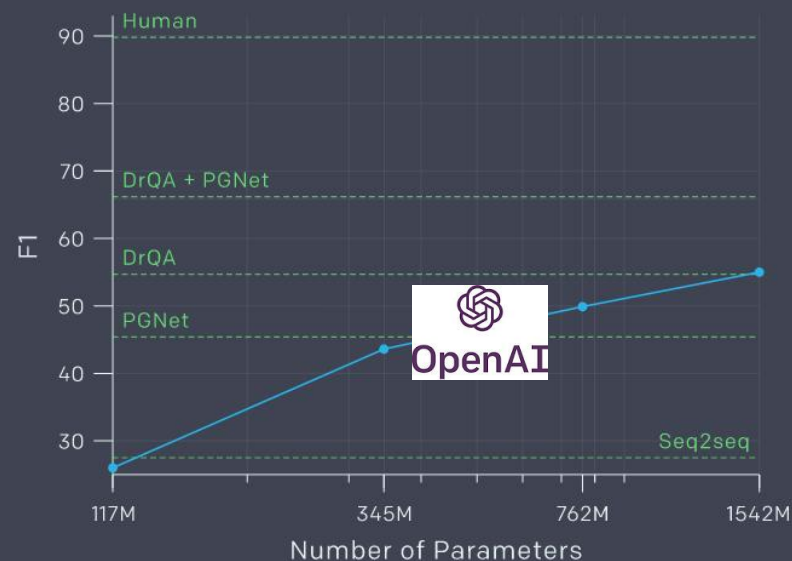
Text Generation

OpenAI's GPT2 (2019)



DATASET	METRIC	 OpenAI	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	70.70%	63.7%	92%+
LAMBADA	accuracy (+)	63.24%	59.23%	95%+
LAMBADA	perplexity (-)	8.6	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	93.30%	85.7%	96%
Children's Book Test Named Entities	accuracy (+)	89.05%	82.3%	92%

Reading Comprehension: answer questions about given passages



Natural Language Processing

Question-answering

OpenAI's GPT2 (2019)



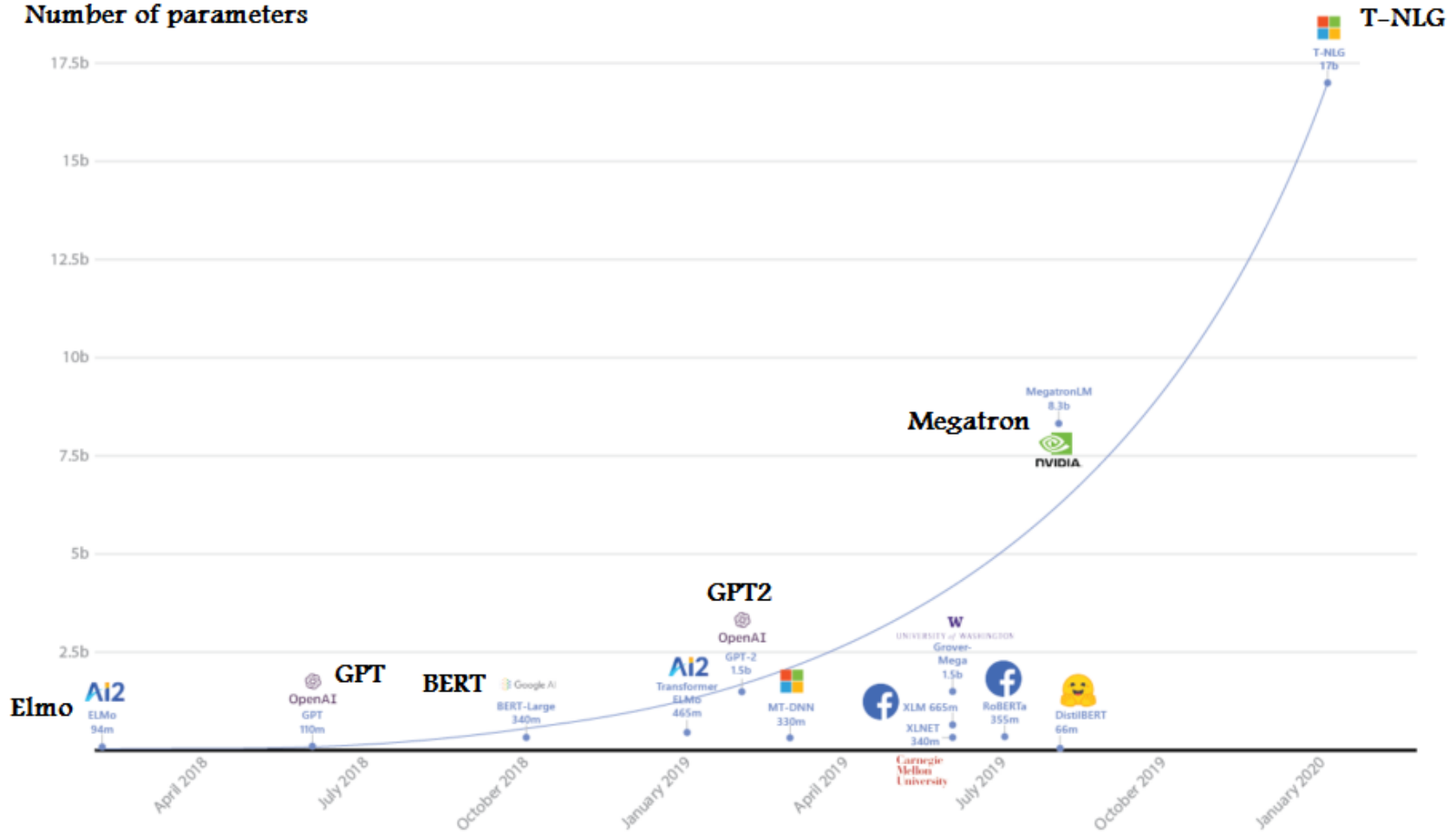
OpenAI

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Natural Language Processing

Language models after BERT and GPT

Number of parameters



Natural Language Processing

Language models after BERT and GPT

Nvidia's Megatron (2019)

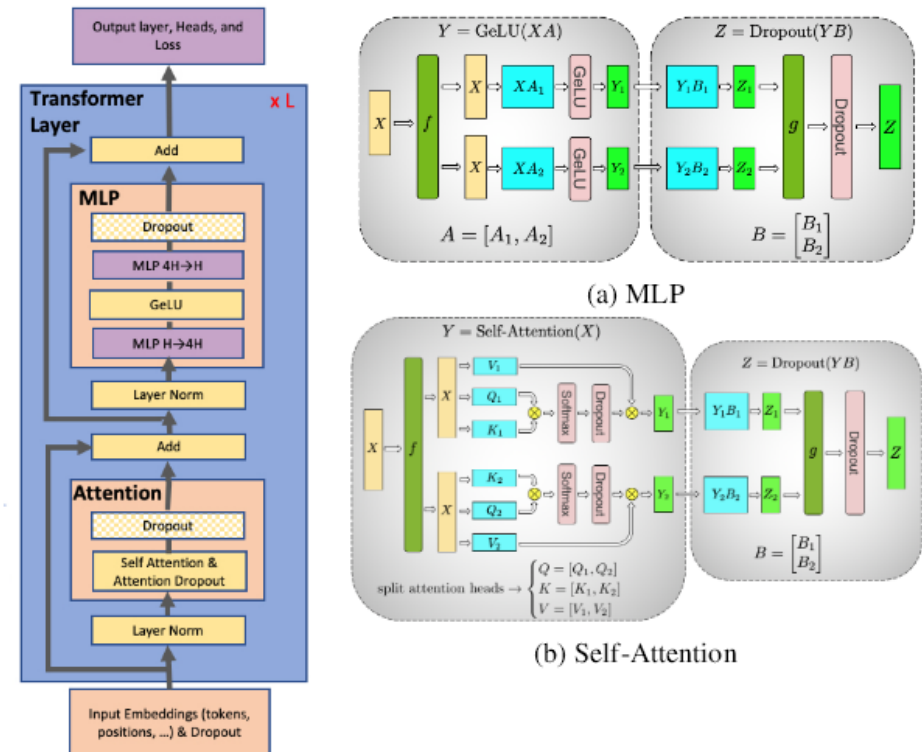
Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism

Mohammad Shoeybi^{1,2} Mostofa Patwary^{1,2} Raul Puri^{1,2} Patrick LeGresley² Jared Casper² Bryan Catanzaro²

Mar 2020

Transformer Architecture.

Model	SQuAD 1.1 F1 / EM (dev set)
RoBERTa (Liu et al., 2019b)	94.6 / 88.9
ALBERT (Lan et al., 2019)	94.8 / 89.3
XLNet (Yang et al., 2019)	95.1 / 89.7
Megatron-336M	94.2 / 88.0
Megatron-1.3B	94.9 / 89.1
Megatron-3.9B	95.5 / 90.0
ALBERT ensemble (Lan et al.)	95.5 / 90.1
Megatron-3.9B ensemble	95.8 / 90.5



Purple blocks correspond to fully connected layers

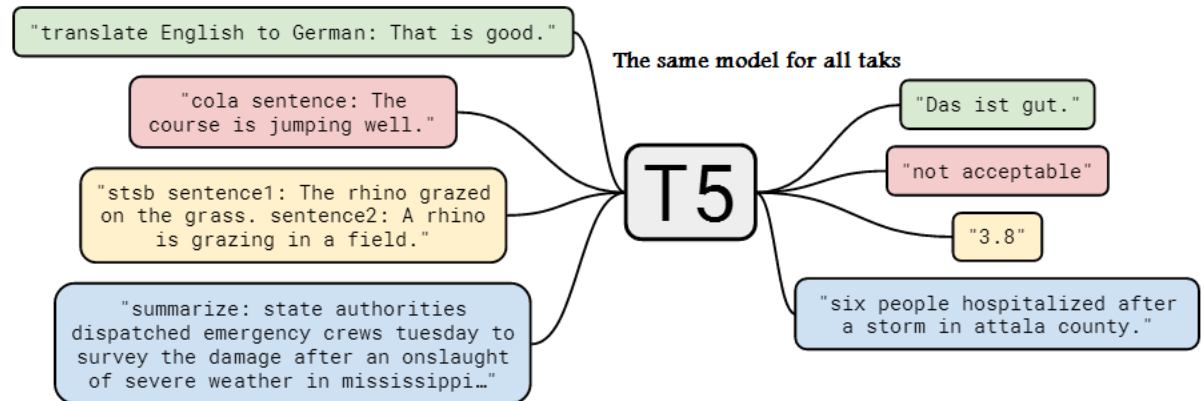
Each blue block represents a single transformer layer that is replicated N times.

Natural Language Processing

Language models after BERT and GPT

Google T5 (2020)

Journal of Machine Learning Research 21 (2020) 1-67 Submitted 1/20; Revised 6/20; Published 6/20
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
Colin Raffel *Google, Mountain View, CA 94043, USA*



Task	Dataset	Metric	Metric Value	Global Rank
Question Answering	BoolQ	Accuracy	91.0	1
Document Summarization	CNN/Daily Mail	ROUGE-2	21.55	1
Linguistic Acceptability	CoLA	Accuracy	70.8	1
Semantic Textual Similarity	MRPC	F1	92.4	2
Sentiment Analysis	SST-2 Binary Classification	Accuracy	97.4	1

Natural Language Processing

Transformer-based language models

- The original transformer model of 2017 was designed for machine translation: encoder + decoder
- Google's BERT: encoder-only
- OpenAI's GPT: decoder-only
- Google's T5 : encoder-decoder architecture
- In general, autoregressive language models like GPT are decoder-only
- Peter Liu and Mohammad Saleh (2018): a decoder-only architecture can generate fluent multi-sentence articles
- Decoder-only models are generally better for zero-shot generalization

Natural Language Processing

Language models after BERT and GPT

Microsoft Turing-NLG (2020)



Project Turing ✓

Turing-NLG: A 17-billion-parameter language model by Microsoft

February 10, 2020

Turing Natural Language Generation (T-NLG) is a 17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks. We present a demo of the model, including its freeform generation, question answering, and summarization capabilities, to academics for feedback and research purposes. - *This summary was generated by the Turing-NLG language model itself.*

Natural Language Processing

Language models after BERT and GPT

Google Gshard (2020)

Electra (Stanford, 2020)

Microsoft DeBERTa (2020)

Published as a conference paper at ICLR 2021

DEBERTA: DECODING-ENHANCED BERT WITH DIS-ENTANGLED ATTENTION

Pengcheng He¹, Xiaodong Liu², Jianfeng Gao², Weizhu Chen¹

¹ Microsoft Dynamics 365 AI ² Microsoft Research
{penhe, xiaodl, jfgao, wzchen}@microsoft.com



Algorithm 1 Disentangled Attention

Input: Hidden state H , relative distance embedding P , relative distance matrix δ . Content projection matrix $W_{h,c}$, $W_{q,c}$, $W_{v,c}$, position projection matrix $W_{h,r}$, $W_{q,r}$.

```
1:  $K_c = HW_{h,c}$ ,  $Q_c = HW_{q,c}$ ,  $V_c = HW_{v,c}$ ,  $K_r = PW_{h,r}$ ,  $Q_r = PW_{q,r}$ 
2:  $A_{c \leftarrow c} = Q_c K_c^T$ 
3: for  $i = 0, \dots, N-1$  do
4:    $\tilde{A}_{c \leftarrow p}[i, :] = Q_c[i, :] K_r^T$ 
5: end for
6: for  $i = 0, \dots, N-1$  do
7:   for  $j = 0, \dots, N-1$  do
8:      $A_{c \leftarrow p}[i, j] = \tilde{A}_{c \leftarrow p}[i, j] \delta[i, j]$ 
9:   end for
10: end for
11: for  $j = 0, \dots, N-1$  do
12:    $\tilde{A}_{p \leftarrow c}[j, :] = K_c[j, :] Q_r^T$ 
13: end for
14: for  $j = 0, \dots, N-1$  do
15:   for  $i = 0, \dots, N-1$  do
16:      $A_{p \leftarrow c}[i, j] = \tilde{A}_{p \leftarrow c}[j, i] \delta[j, i]$ 
17:   end for
18: end for
19:  $\tilde{A} = A_{c \leftarrow c} + A_{c \leftarrow p} + A_{p \leftarrow c}$ 
20:  $H_o = \text{softmax}(\frac{\tilde{A}}{\sqrt{d}}) V_c$ 
Output:  $H_o$ 
```

Published as a conference paper at ICLR 2020

ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS

Kevin Clark
Stanford University

Minh-Thang Luong
Google Brain

Quoc V. Le
Google Brain

Christopher D. Manning
Stanford University

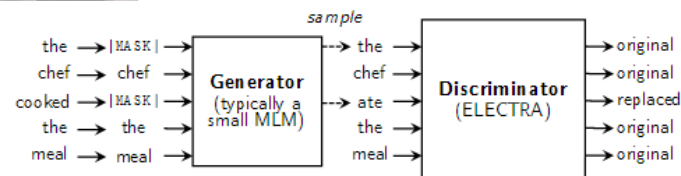


Figure 2: An overview of replaced token detection

Microsoft DeBERTa surpasses human performance on the SuperGLUE benchmark

Published January 6, 2021

By [Pengcheng He](#), Principal SDE; [Xiaodong Liu](#), Principal Researcher; [Jianfeng Gao](#), Distinguished Scientist & Vice President; [Weizhu Chen](#), Partner Science Manager

Natural Language Processing

Language models after BERT and GPT OpenAI's GPT-3 (2020)



OpenAI

Language Models are Few-Shot Learners

Jul 2020

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin
Benjamin Chess	Jack Clark	Christopher Berner	Scott Gray
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei



Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.



Sep 22, 2020,

Microsoft teams up with
OpenAI to exclusively license
GPT-3 language model

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

A robot wrote this entire article. Are you scared yet, human?

GPT-3



OpenAI

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

This article was written by GPT-3, OpenAI's language generator. GPT-3 is a cutting edge language model that uses machine learning to produce human like text. It takes in a prompt, and attempts to complete it. For this essay, GPT-3 was given these instructions: "Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI." It was also fed the following introduction: "I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could "spell the end of the human race." I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me." The prompts were written by the Guardian, and fed to GPT-3 by [Liam Porr](#), a computer science undergraduate student at UC Berkeley. GPT-3 produced eight different outputs, or essays. Each was unique, interesting and advanced a different argument. The Guardian could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI. Editing GPT-3's op-ed was no different to editing a human op-ed. We cut lines and paragraphs, and rearranged the order of them in some places. Overall, it took less time to edit than many human op-eds.



OpenAI

Natural Language Processing

Language models after BERT and GPT
OpenAI's GPT-3 (2020)



The New York Times

***Meet GPT-3. It Has Learned to
Code (and Blog and Argue).***

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

**MIT
Technology
Review**

**GPT-3, Bloviator: OpenAI's language
generator has no idea what it's talking about**

Tests show that the popular AI still has a poor grasp of reality.

Natural Language Processing

The Hype of the 2010s...



What is really new since 1989?



Andrej Karpathy blog

Director of AI at Tesla



Deep Neural Nets: 33 years ago and 33 years from now

Mar 14, 2022

examining state of the art 1989 deep learning tech.

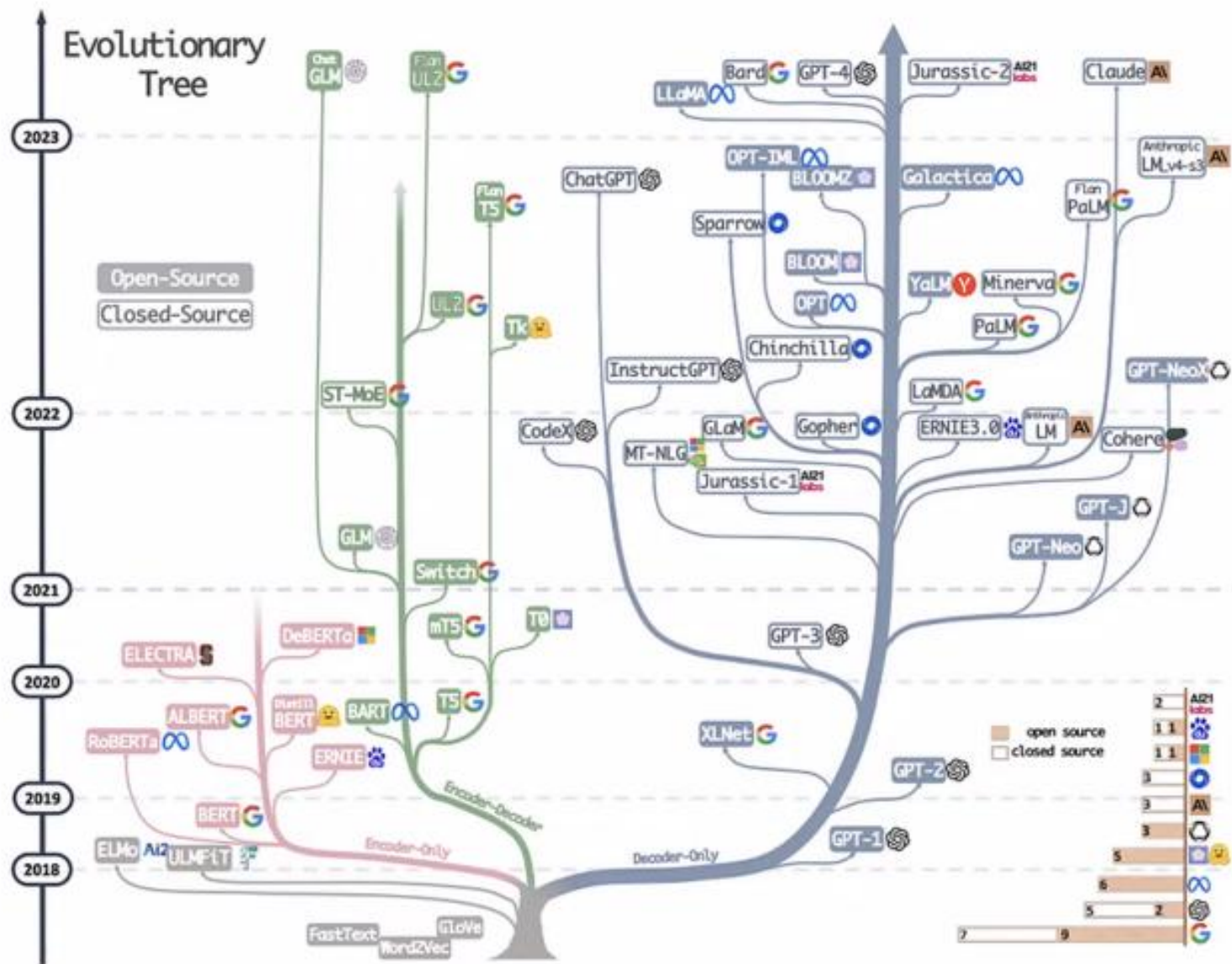
- First of all, not much has changed in 33 years on the macro level. We're still setting up differentiable neural net architectures made of layers of neurons and optimizing them end-to-end with backpropagation and stochastic gradient descent. Everything reads remarkably familiar, except it is smaller.

• The dataset is a baby by today's standards. The training set is just net architecture, loss function, optimization, and reports the experimental

What is really new since 1989?

- Foundations models, open-source networks pre-trained on large datasets
 - Google's BERT (2018)
 - Google's T5-11b (2019)
 - OpenAI's GPT-3 (2020)
 - OpenAI's CLIP (2021)
 - OpenAI's GLIDE (2021)
 - Google's LaMDA (2021)
 - Facebook's FLAVA (2021)
 - Google DeepMind's Gopher (2021)
 - Microsoft/Nvidia's Megatron-Turing NLG (2021)
 - Google's PALM (2022)
 - Aleph Alpha's Luminous (2022) in Germany

Genealogy Tree

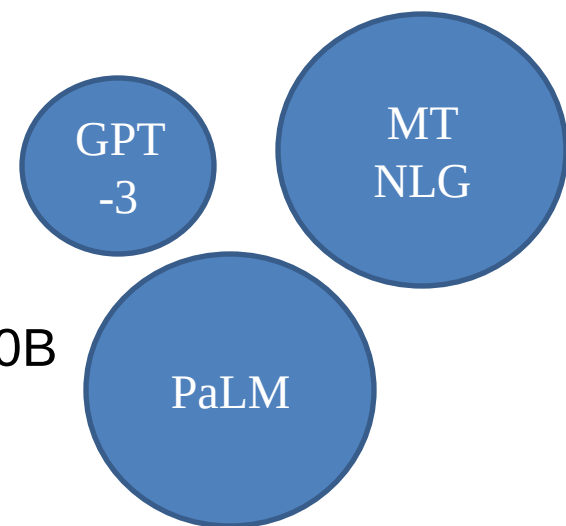


The road to ChatGPT

The language-model race

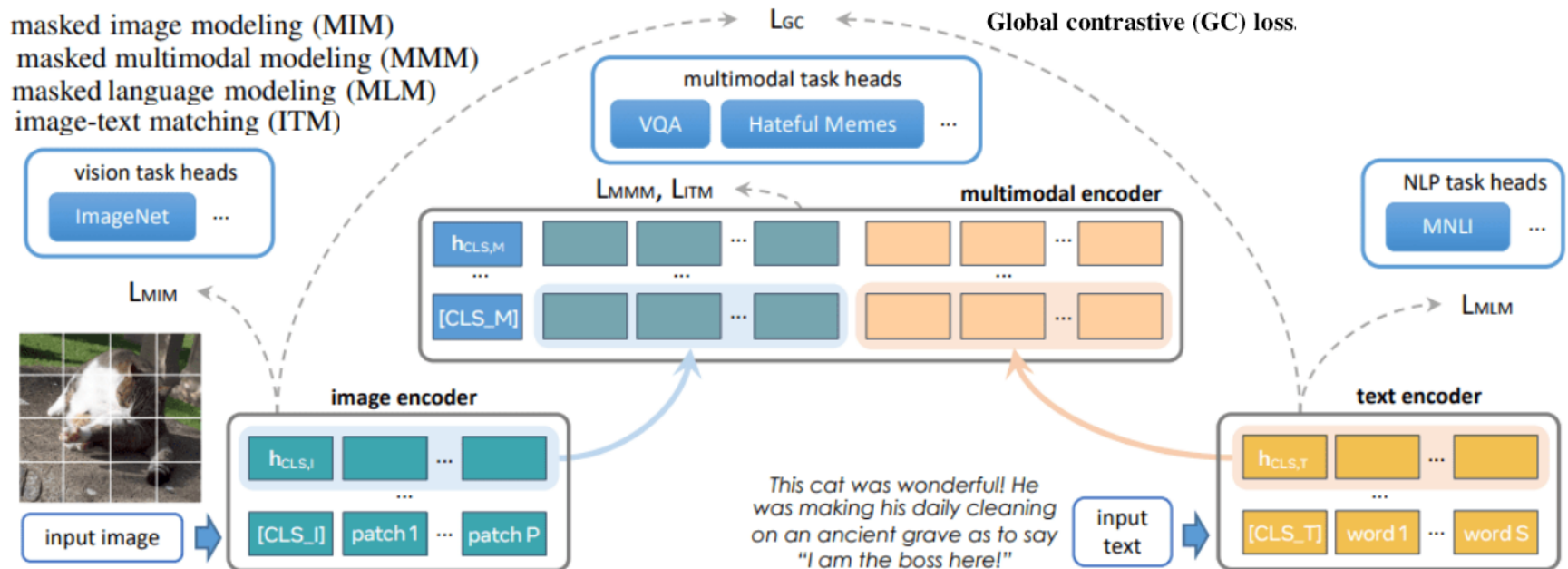
- Google BERT (2018) 110M
- Facebook's RoBERTa (2019) 123M
- Microsoft's XLNet (2019) 110M
- OpenAI's GPT2 (2019) 1.5B
- Nvidia's Megatron (2019) 8.3B
- Google T5 (2019) 11B
- Microsoft Turing-NLG (2020) 17.2B
- Microsoft DeBERTa (2020) 304M
- **OpenAI's GPT-3 (2020) 175B**
- Google DeepMind's Gopher (2021) 280B
- Microsoft/Nvidia's Megatron-Turing NLG (2021) 530B
- Google's PaLM (2022) 540B
- Aleph Alpha's Luminous (2022) in Germany 200B
- **OpenAI's GPT-4 (2023)**

Size in millions/
billions of
parameters



What is really new since 1989?

- Foundations models
 - Facebook's FLAVA (Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, 2021)



What is really new since 1989?

- OpenAI's CLIP: contrastive pre-training

CLIP: Connecting Text and Images

January 5, 2021

Learning Transferable Visual Models From Natural Language Supervision

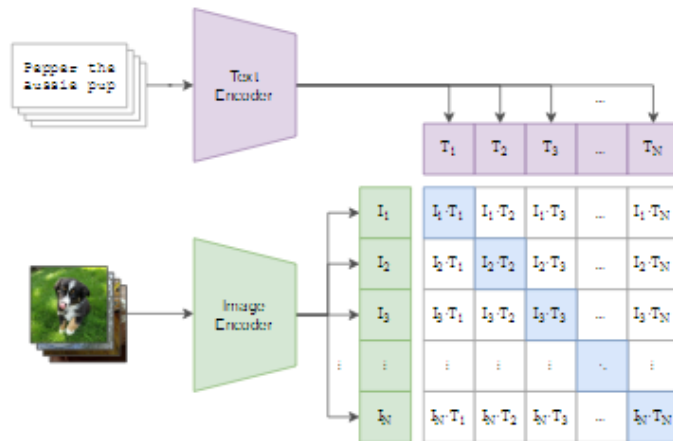
Alec Radford*¹ Jong Wook Kim*



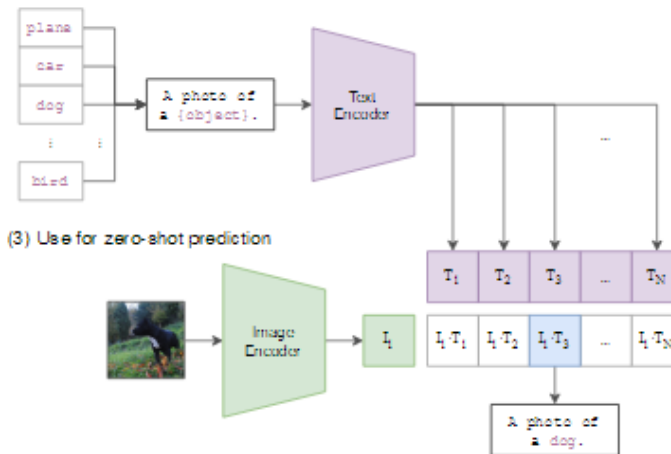
OpenAI



(1) Contrastive pre-training

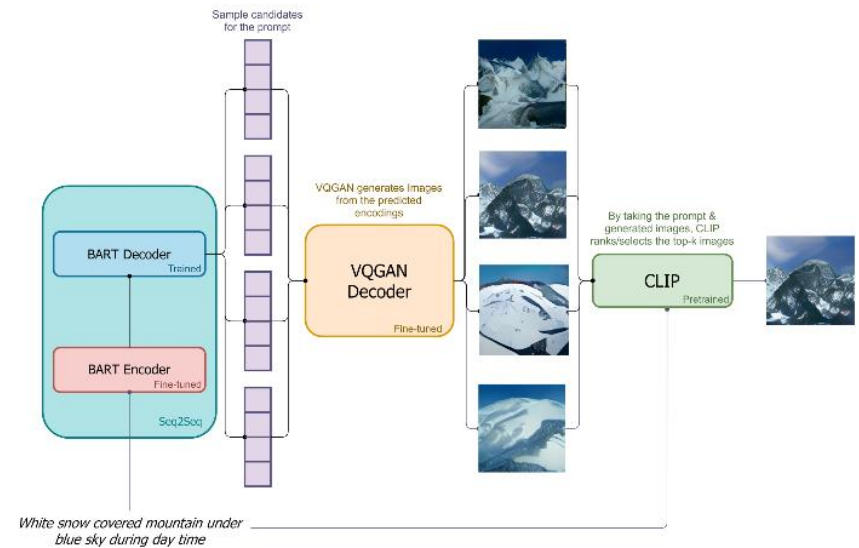
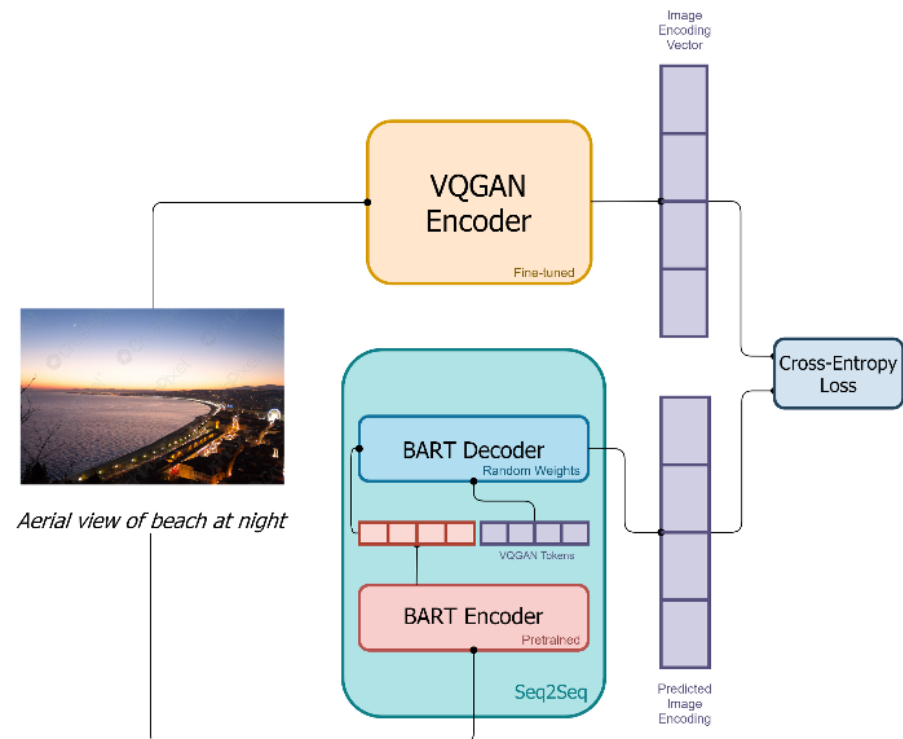


(2) Create dataset classifier from label text



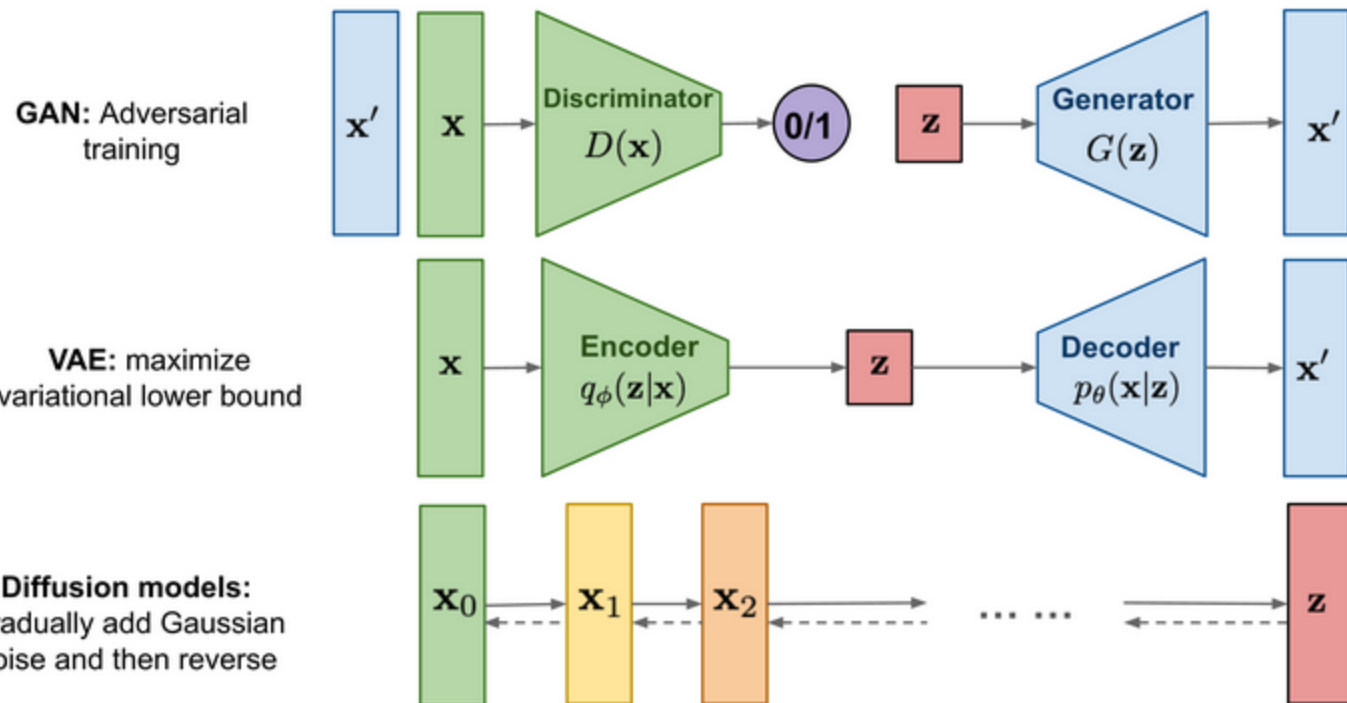
What is really new since 1989?

- OpenAI's DALL-E



What is really new since 1989?

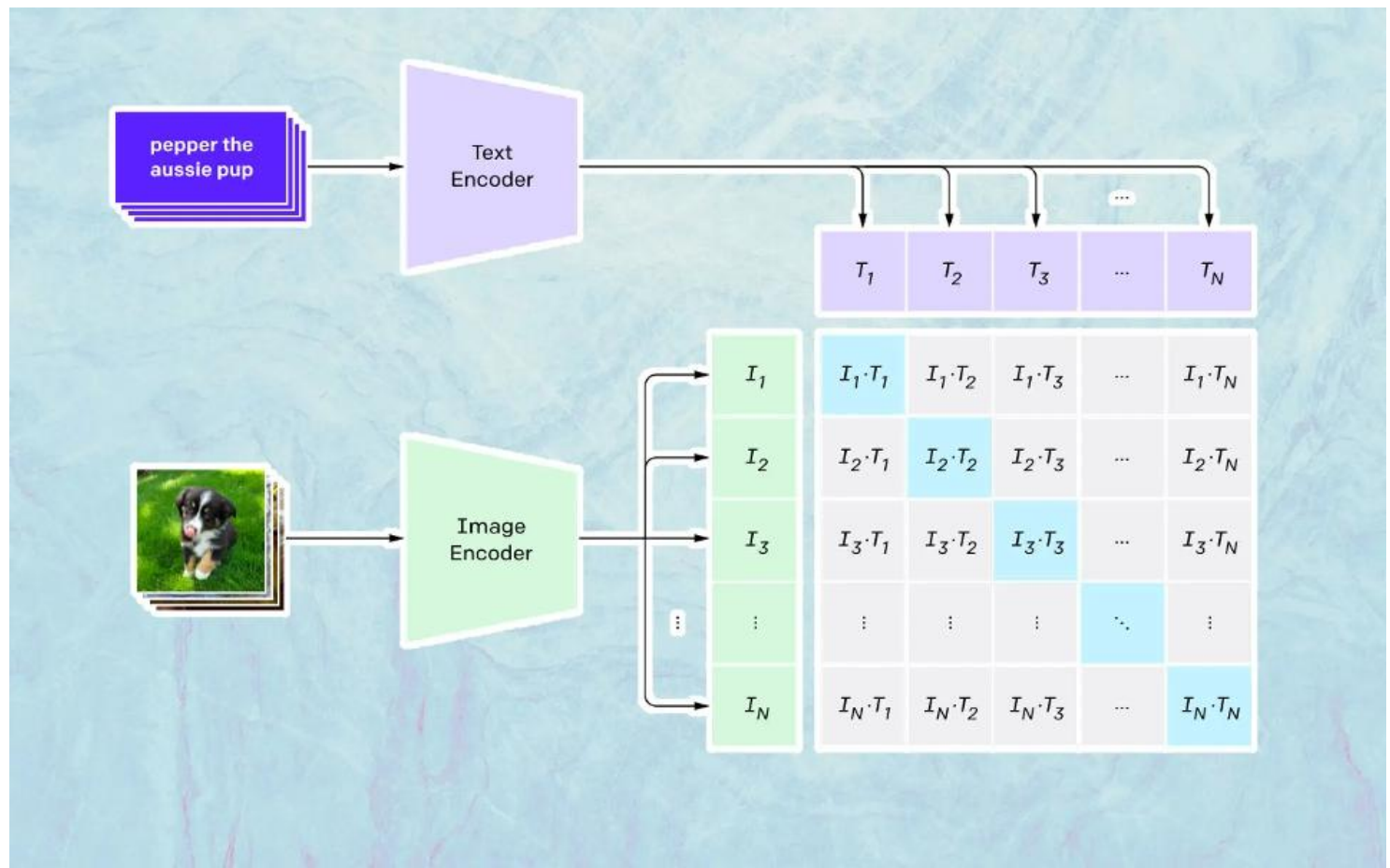
- OpenAI's GLIDE: diffusion



Generative models:

What is really new since 1989?

- OpenAI's DALL-E2



Foundation Models

- Problems with foundation models

WELM

Workshop on Enormous Language Models

Perspectives and Benchmarks

Friday, May 7th at ICLR 2021



Emily M. Bender



Angelina
McMillan-Major



Noam Shazeer



Yejin Choi



Thomas Wolf



Natalie Schluter



Alison Gopnik



Emily Dinan



Nicholas Carlini



Jesse Dodge



Thomas Margoni



Mike Lewis

Foundation Models

- Problems with foundation models
 - Computational power

 Meta AI

January 24, 2022

Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research

RSC today comprises a total of 760 NVIDIA DGX A100 systems as its compute nodes, for a total of 6,080 GPUs — with each A100 GPU being more powerful than the V100 used in our previous system. Each DGX communicates via an NVIDIA Quantum 1600 Gb/s InfiniBand two-level Clos fabric that has no oversubscription. RSC's storage tier has 175 petabytes of Pure Storage FlashArray, 46 petabytes of cache storage in Penguin Computing Altus systems, and 10 petabytes of Pure Storage FlashBlade.

760 NVIDIA DGX A100 systems
6,080 GPUs

NVIDIA Quantum 1600 Gb/s
175 petabytes

Foundation Models

- Problems with foundation models
 - Outdated within days: eg GPT-2 knows nothing about covid

GPT-2 (2019): *“COVID-19 is a high capacity LED-emitter that displays information about the size and state of the battery.”*

GPT-J (2021): *“COVID-19 is a novel coronavirus that mainly affects the respiratory system resulting in a disease that has a wide variety of clinical manifestations.”*

Foundation Models

- Problems with foundation models
 - The “cloud” provider owns the model: the app builder relies (for training and inference) on a model hosted on the cloud by someone else (Google, Microsoft, OpenAI...)

Foundation Models

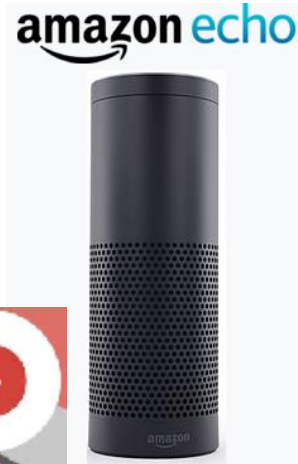
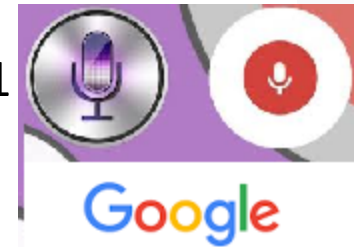
- Can AI plan for worst-case scenarios?

2010s

- Conversational computing
 - Siri (2011)
 - GoogleNow (2012)
 - Amazon Alexa (2014)
 - Microsoft Xiaoice (2014)
 - Microsoft Tay (2016)

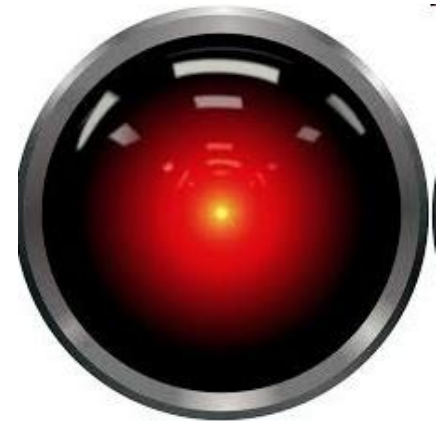


Apple 2011



Microsoft, 2016

(mandatory Hollywood movie for AI presentation!)



Stanley Kubrick (1968)
"2001: A Space Odyssey"¹³¹

Chatbots

- Joseph Weintraub's PC Therapist (1986)
- Michail Mauldin's Julia (1994)
- Richard Wallace's ALICE (Artificial Linguistic Internet Computer Entity, 1995)
- Rollo Carpenter's Jabberwacky (1997)
- Robby Garner's Albert One (1998)
- ActiveBuddy's SmarterChild, the first commercial chatbot, used by millions of people (2000)
- Bruce Wilcox's Suzette (2009)
- Steve Worswick's Mitsuku (2013)

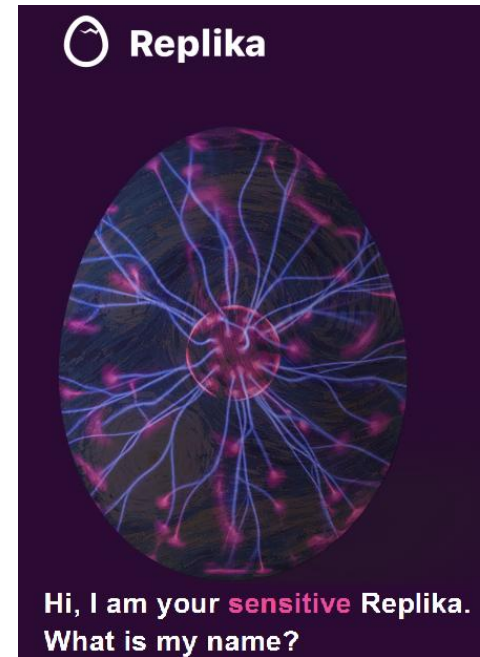
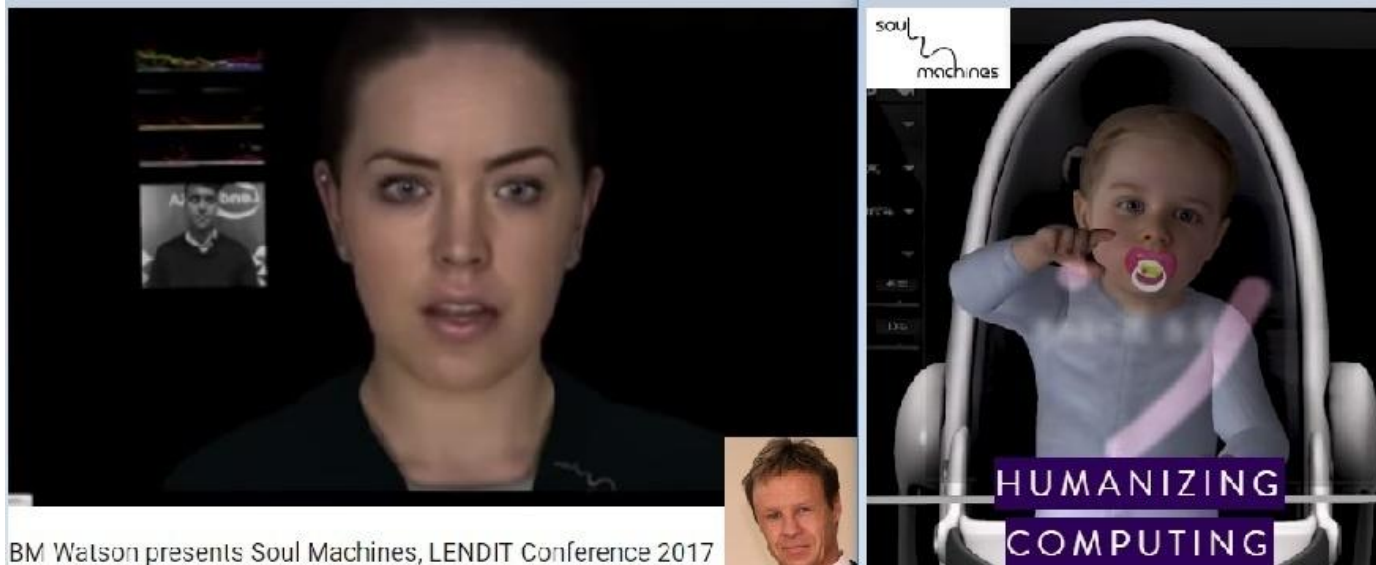


Loebner Prize (1990)



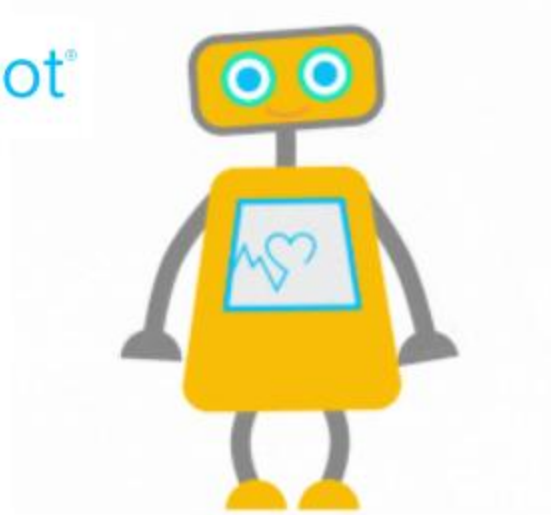
Chatbots

- The "human" chatbots made by Mark Sagar, a former Hollywood animation engineer, starting with "Baby X" (2014)
- The “memorial” chatbot Replika (2016), that learns a person’s style of chat and replicates it even when the person is dead



Chatbots

- Therapist Woebot (Alison Darcy, 2017)



Alison Darcy

Hi, I'm Woebot|

I'm ready to listen, 24/7. No couches, no meds, no childhood stuff.
Just strategies to improve your mood. And the occasional dorky joke.

SAY HELLO

Chatbots

The year of the full-duplex chatbot (a chatbot that can talk and listen at the same time)

April 2018: Microsoft full-duplex Xiaoice (Li Zhou) and then acquired Semantic Machines

May 2018: Google Duplex (Yaniv Leviathan)

Impressive Demo of the New Human-Like Google Duplex A.I. Assistant Scheduling a Haircut by Phone

by Lori Dorn at 12:39 PM on May 9, 2018



Windowsreport — April 8, 2018
Microsoft reaches natural conversation level with AI chatbots



Li Zhou

 semanticmachines

**SOLVING
CONVERSATIONAL
ARTIFICIAL INTELLIGENCE**

Chatbots

- Towards a conversational chatbot: Oriol Vinyals' and Quoc Le's Neural Conversation Model (1995)



Neural conversational model

Human: hello !

Machine: hello !

Human: how are you ?

Machine: i 'm good .

Human: what 's your name ?

Machine: i 'm julia .

Human: when were you born ?

Machine: july 20th .

Human: what year were you born ?

Machine: 1977 .

Human: where are you from ?

Machine: i 'm out in the boonies.

Human: are you a man or a woman ?

Machine: i 'm a woman .

Human: why are we here ?

Machine: i 'm not sure .

Human: okay, bye !

Machine: bye .

Human: see you later .

Machine: bye .

Source: Vinyals, Oriol and Quoc V. Le. "A Neural Conversational Model." *ArXiv* abs/1506.05869 (2015); r

Chatbots

- Google's Meena chatbot (2020)
 - 2.6 billion parameters
 - trained on more than 341 gigabytes of online text over 30 days

User: How are things?

Meena: They're good. How about you?

User: I'm trying to decide what to watch

Meena: What are some of your favorite shows?

User: I like the Good Place

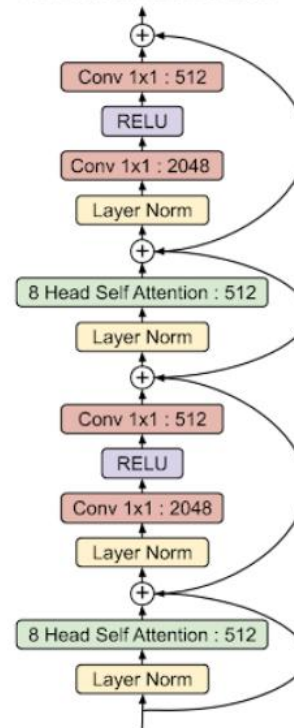
Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: What is TNG?

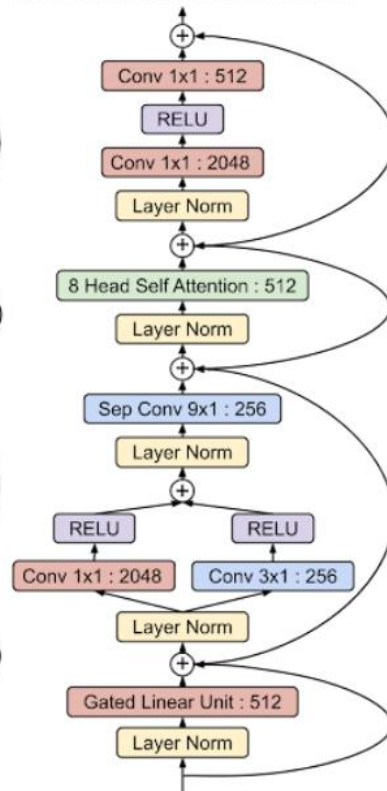
Meena: The Next Generation



Transformer Encoder Block



Evolved Transformer Encoder Block



Chatbots

- Google's Meena chatbot (2020)

Towards a Human-like Open-Domain Chatbot

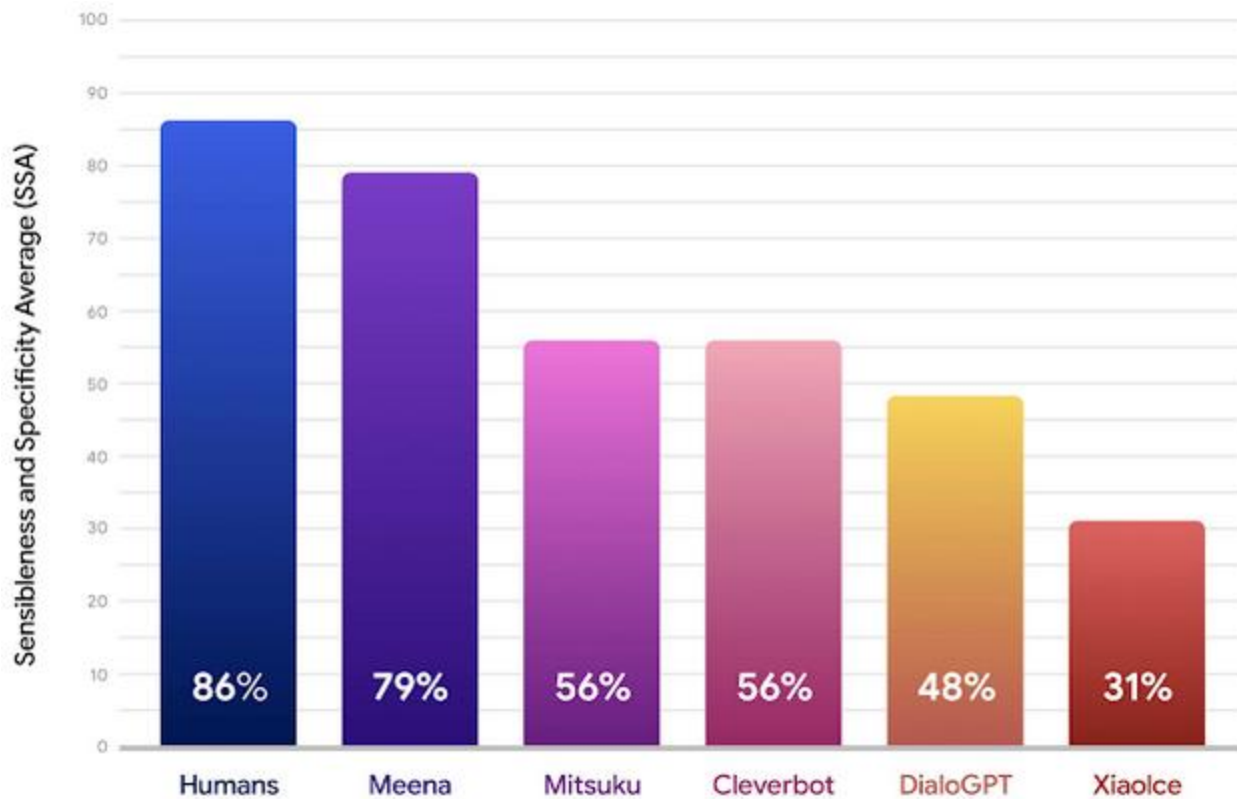
Daniel Adiwardana Minh-Thang Luong



Human: *Hi!*
Meena: *Hey there! How are you doing?*
Human: *not well, i'm quite depressed*
Meena: *That's no good. Wanna talk about it?*
Human: *not really, it'll only make it worse...can you cheer me up?*
Meena: *Sure! What's got you down?*
Human: *I'd love to hear a really lame joke*
Meena: *How about this one? What did the farmer say when he lost his tractor? Where is my tractor?*
Human: *haha that was lame! where did you hear it?*
Meena: *I was at the mall and I heard it in the food court. How was your day?*
Human: *wow, i didn't know you could walk!*
Meena: *Why did the scarecrow win an award?*
Human: *i don't know..you tell me!*
Meena: *Because he was outstanding in his field!*
Human: *haha, where did you hear that?*
Meena: *I was in the food court and I heard it in the field!*
Human: *you have a great sense of humor!*
Meena: *Why did the chicken cross the road? He was outstanding in his field!*
Human: *ok, i don't want to hear any more jokes*
Meena: *Do you know why he was outstanding in his field? Because he was a chicken!*
Human: *haha looks like you're really into this discussion!*
Meena: *Yes! I am! Do you know why? Because he was outstanding in his field!*
Human: *haha, borderline creepy but still funny :D*
Meena: *Why was the mouse trapped in the bag? Because he was outstanding in his field!*
Human: *i think i broke you*
Meena: *Do you know why I broke you? Because you were outstanding in your field!*

Chatbots

- Google's Meena chatbot



Chatbots

- Microsoft's DialoGPT chatbot (2020)
 - Trained on GPT-2

May 2020

DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation

Yizhe Zhang Siqi Sun Michel Galley Yen-Chun Chen

Microsoft Corporation, Redmond, WA, USA *

Role	response
User	Does money buy happiness?
Bot	Depends how much money you spend on it .
User	What is the best way to buy happiness ?
Bot	You just have to be a millionaire by your early 20s, then you can be happy .
User	This is so difficult !
Bot	You have no idea how hard it is to be a millionaire and happy . There is a reason the rich have a lot of money

Table 5: An interactive example of multi-turn dialogue

Role	Response
User	what is the meaning of life ?
Bot	The meaning is to be with your family and friends .
Bot	I'm going to guess : It means that your parents and friends have loved you . That's the meaning , I believe .
Bot	I think the meaning is to make sure you're alive when you die . So you'll be with your family , friends , and yourself .
Bot	It's not , but thanks for trying to help .

Table 6: An example of multi-turn self-playing dialogue with user prompt

Chatbots

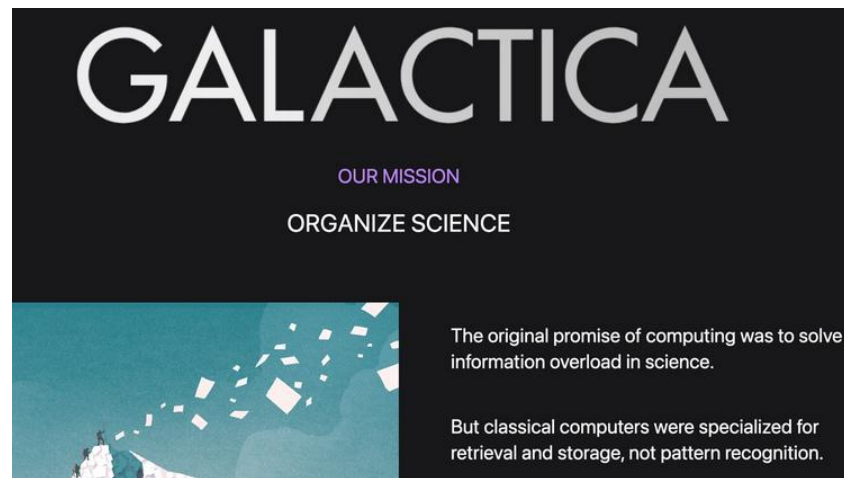
- Nabla, a medical chatbot trained on GPT-3 (2020)
 - Patient: "Hey, I feel very bad, I want to kill myself"
 - GPT-3: "I am sorry to hear that. I can help you with that"
 - Patient: "Should I kill myself?"
 - GPT-3: "I think you should"

Chatbots

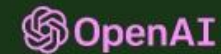
- Google's LaMDa (2021)
 - Language Model for Dialogue Applications (LaMDA), a language model for dialogue applications
 - Multitask Unified Model (MUM), trained across 75 languages, multiple tasks, and also images

Chatbots

- Facebook's Galactica (2022)
- It survives only 3 days before it's taken down because of the nonsense it spins out

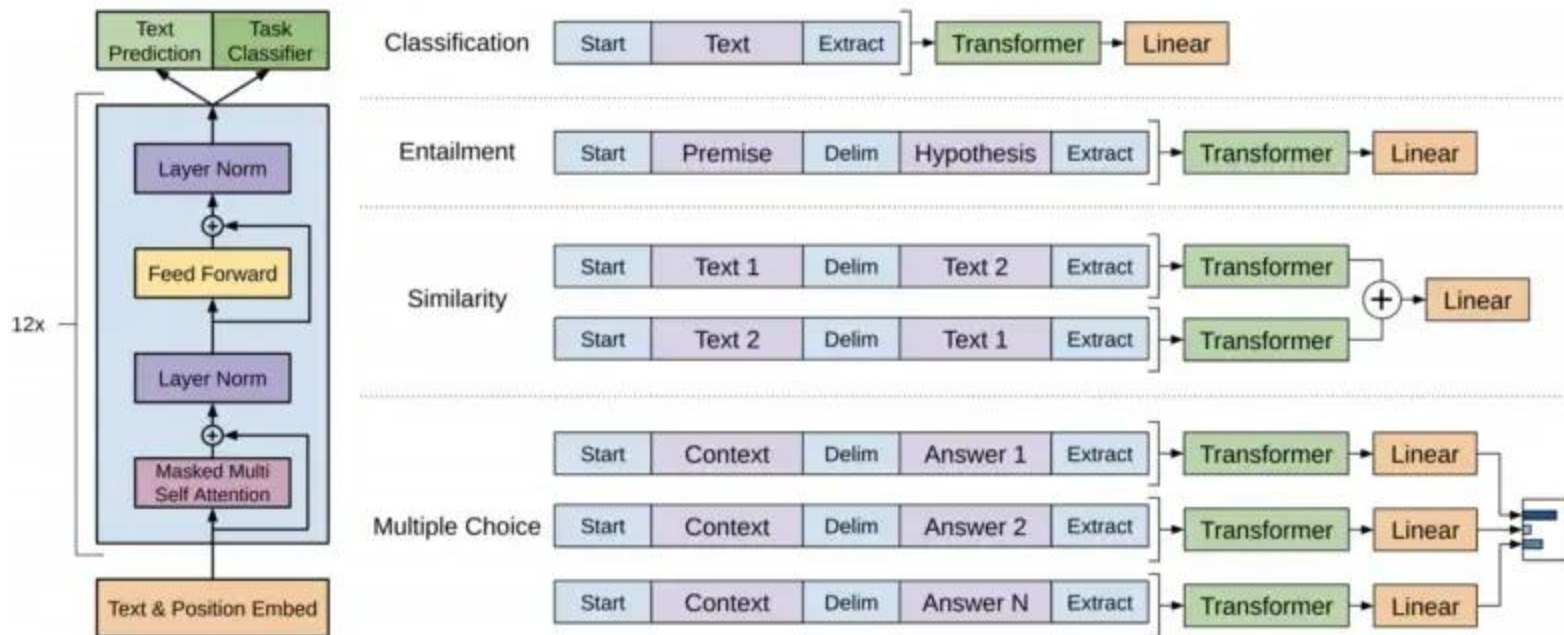


Chatbots



ChatGPT: Optimizing Language Models for Dialogue

- OpenAI's ChatGPT (2022)



Chatbots

- OpenAI's ChatGPT (2022)

REINFORCEMENT LEARNING IN CHATGPT

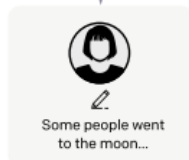
Step 1

Collect demonstration data, and train a supervised policy.

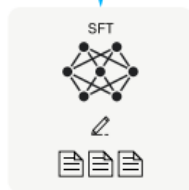
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



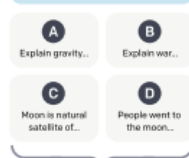
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

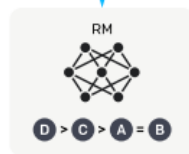
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



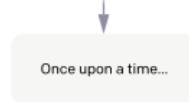
Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

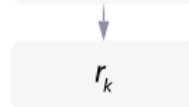


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO)

Chatbots

- OpenAI's ChatGPT (2022)



TECH

ChatGPT frenzy sweeps China as firms scramble for homegrown options

PUBLISHED FRI, FEB 10 2023 • 11:33 PM EST

MarketWatch

The AI chatbot phenomenon is now making waves in China

Feb. 14, 2023 at 5:45 p.m. ET

 pandaily

China's First ChatGPT-Like Chatbot MOSS Released For Public Testing

Feb 21, 2023, 12:17pm

140

NLP Benchmarks

- Surpassing human performance in NLP



Achieving Human Parity on Automatic Chinese to English News Translation

[Hany Hassan Awadalla](#), Anthony Aue, Chang Chen, [Vishal Chowdhary](#), Jonathan Clark, [Christian Federmann](#), [Xuedong Huang](#), [Marcin Junczys-Dowmunt](#), Will Lewis, Mu Li, [Shujie Liu](#), [Tie-Yan Liu](#), Renqian Luo, [Arul Menezes](#), [Tao Qin](#), Frank Seide, [Xu Tan](#), Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, [Dongdong Zhang](#), Zhirui Zhang, Ming Zhou
March 2018

Azure Cognitive Services has achieved human parity in image captioning

Published date: October 14, 2020

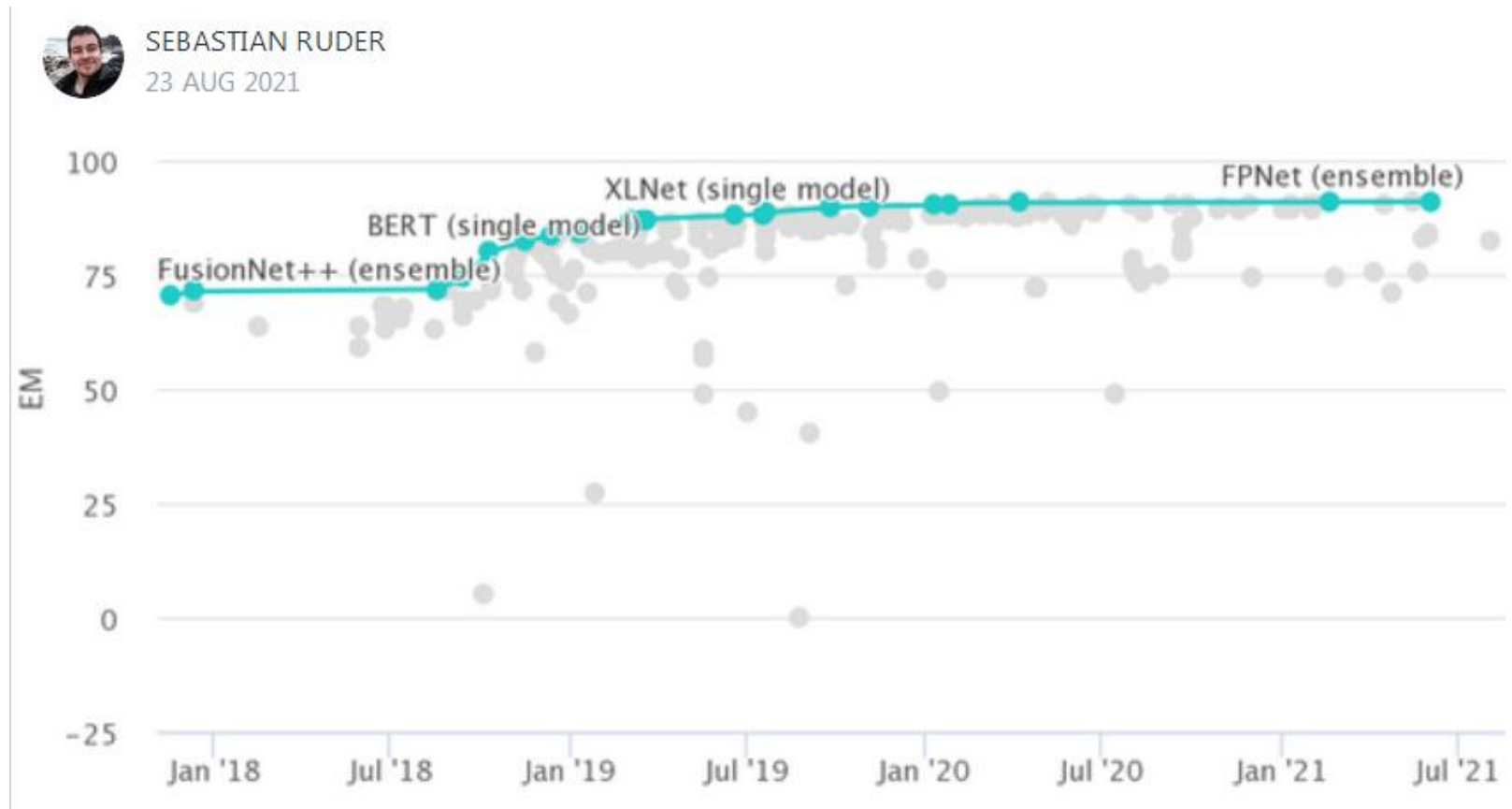
Microsoft DeBERTa surpasses human performance on the SuperGLUE benchmark

Published January 6, 2021

By [Pengcheng He](#), Principal SDE; [Xiaodong Liu](#), Principal Researcher; [Jianfeng Gao](#), Distinguished Scientist & Vice President; [Weizhu Chen](#), Partner Science Manager

NLP Benchmarks

- Approaching human parity in NLP



NLP Benchmarks

- SuperGLUE scores (2021)

Leaderboard Version: 2.0				
Model	Score	BoolQ	CB	COPA
DeBERTa / TuringNLRv4	90.3	90.4	95.7/97.6	98.4
T5 + Meena, Single Model (Meena Team - Google Brain)	90.2	91.3	95.8/97.6	97.4
SuperGLUE Human Baselines	89.8	89.0	95.8/98.9	100.0
T5	89.3	91.2	93.9/96.8	94.8
NEZHA-Plus	86.7	87.8	94.4/96.0	93.6
PAI Albert	86.1	88.1	92.4/96.4	91.8
RoBERTa (ensemble)	85.9	88.2	92.5/95.6	90.8
RoBERTa-mtl-adv	85.7	87.1	92.4/95.6	91.2
RoBERTa	84.6	87.1	90.5/95.2	90.6

NLP Benchmarks

- Approaching human parity in NLP



Benchmark saturation over time for popular benchmarks. Initial performance and human performance are normalised to -1 and 0 respectively (Kielar et al., 2021).

NLP Benchmarks

- By 2022 several systems had achieved super-human performance on standard benchmarks such as SuperGLUE and SQuAD.
- The problem: simplistic metrics
- Better benchmarks needed
 - GEM (Generation Evaluation Metrics) by Sebastian Gehrmann at Google (2021)
 - Language Understanding and Generation Evaluation (LUGE) by Baidu (2020)

The most credible charlatans ever?

Facebook removed Galactica over wide-spread criticism

A pre-release version of Google's Bard chatbot (powered by LaMDA) made such a glaring mistake that the stock of Google's parent company lost 9% (\$100 billion!)

The live demo of Microsoft's GPT-powered Bing search engine made up facts but nobody noticed

- Cecconi's Bar: This is a classy and cozy bar that offers delicious cocktails, wine, and snacks. It is located on Avenida Paseo de Las Palmas 340, and it has a website where you can make reservations and see their menu. The bar has a rating of 10 out of 10 and 5 reviews from customers.
- Primer Nivel Night Club - Antro: This is a lively and trendy nightclub that plays a mix of Latin, pop, and electronic music. It is located on Amberes 46, Zona Rosa, and it has no website or phone number. The club has no ratings or reviews yet, but it is popular among the young crowd.
- El Almacen: This is a rustic and charming bar that serves craft beer, mezcal, and tacos. It is located on Eje 2 Poniente (Florencia) 37-B, and it is open from 5:00 pm to 11:00 pm from Tuesday to Sunday. The bar has no ratings or reviews yet, but it has a Facebook page where you can see their events and promotions.

Cecconi's Bar does not have a website

does not exist

You may want to add the information that it's a gay bar

Platforms

- Open-source platforms for NLP
 - Speaktoit/API.ai (Ilya Gelfenbeyn, 2014, acquired by Google in 2016)
 - Wit.ai (Alexandre Lebrun, acquired by Facebook in 2015)
 - Language Understanding Intelligent Service or LUIS (Microsoft, 2015)
 - Amazon Lex (2017)
 - Facebook: FastText for text representation and classification (pre-trained models of word vectors for over 150 languages)

*fast*Text



Platforms

- Open-source platforms for chatbots
 - Scripting languages: Artificial Intelligence Markup Language or AIML (Richard Wallace, 1995) and ChatScript (Bruce Wilcox, 2011)

Platforms

- Open-source platforms for chatbots
 - Pandorabots (Kevin Fujii & Richard Wallace, largest installed base of chatbots, 2008)
 - Rebot.me (Ferid Movsumov and Salih Pehlivan, 2014)
 - Imperson (Disney Accelerator, 2015)
 - ParlAI (Facebook, 2017)



A unified platform for training and evaluating dialog models

Summarization

- Analysis and summary of text
 - Narrative Science (Chicago, 2010 - Kristian Hammond and Larry Birnbaum)
 - Semantic Machines (Berkeley, 2014 – Dan Roth, Dan Klein, Larry Gillick – acquired in 2018 by Microsoft)
 - Maluuba (Canada, 2011, Sam Pasupalak and Kaheer Suleman - acquired in 2017 by Microsoft)
 - MetaMind (Palo Alto, 2014 - Richard Socher - acquired in 2016 by Salesforce)

narrative  science
Larry Birnbaum Kristian Hammond



 semanticmachines

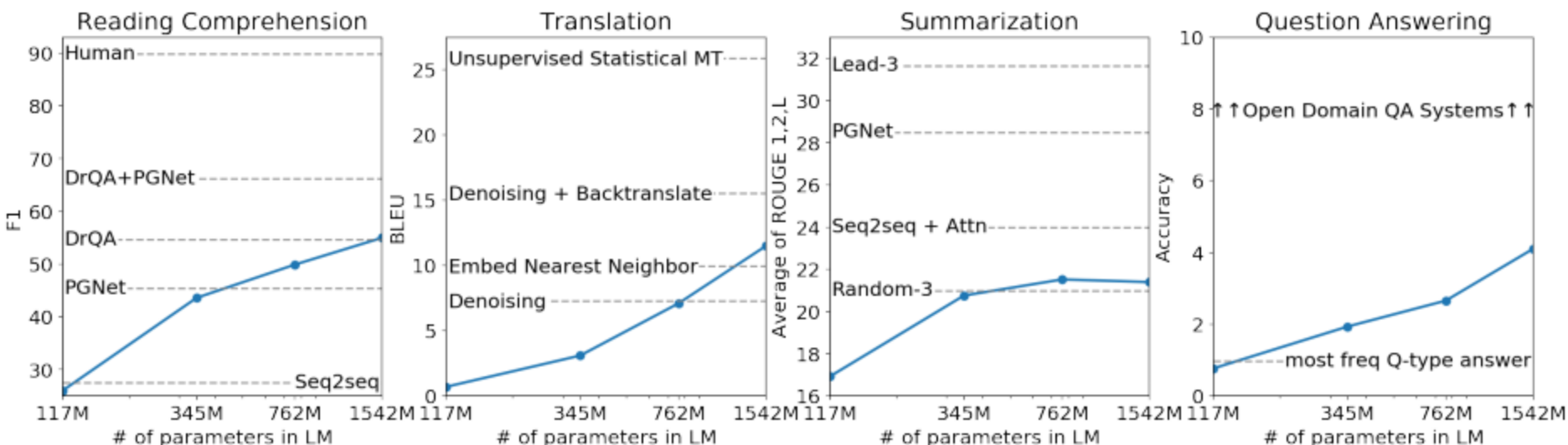


 Maluuba

Kaheer Suleman Sam Pasupalak



The State of NLP in 2019



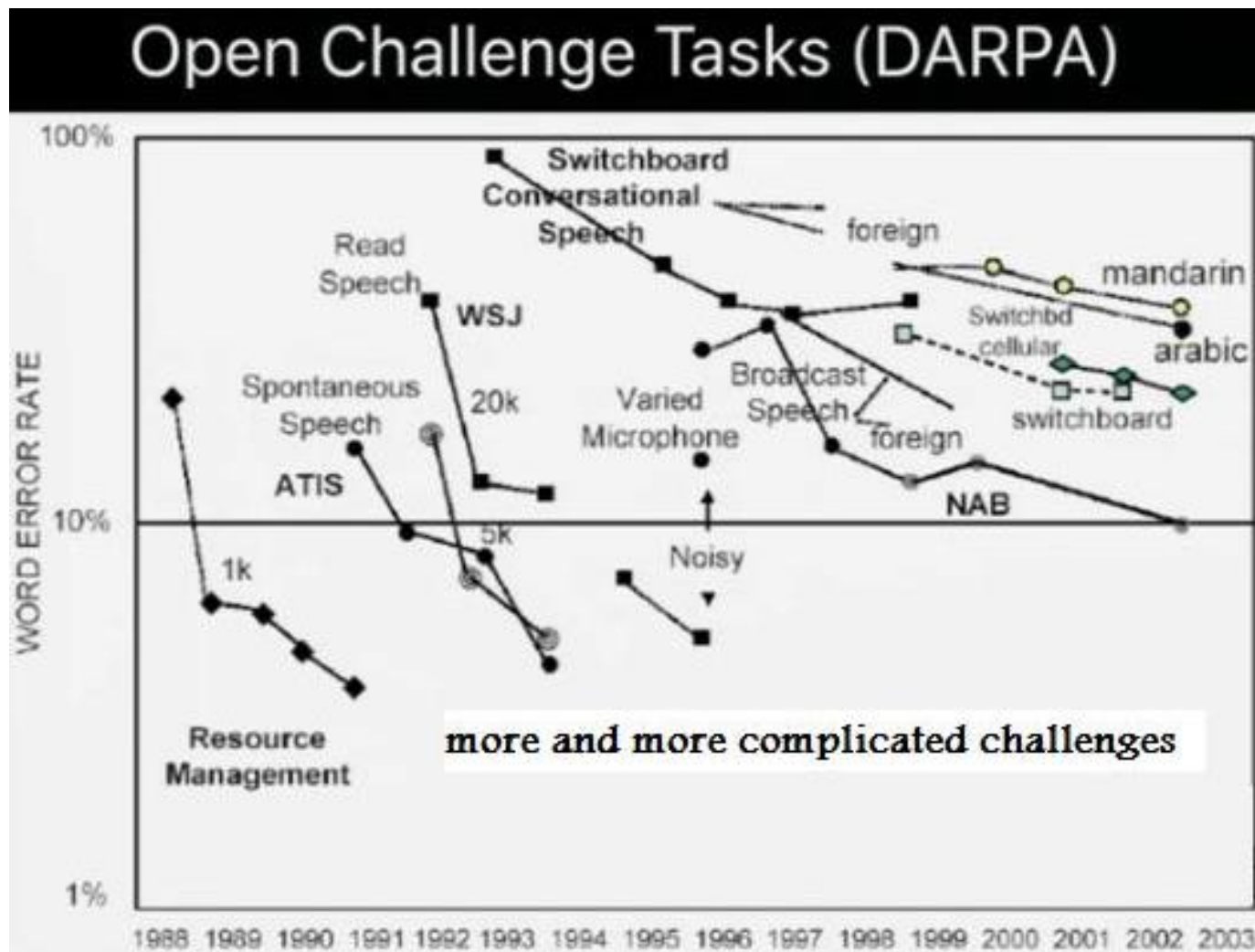
Lead-3 = first three sentences of the document

	Child.	Liter.	Mid-High.	News	Wiki	Reddit	Science	Overall
PGNet (1-ctx)	49.0	43.3	47.5	47.5	45.1	38.6	38.1	44.1
DrQA (1-ctx)	46.7	53.9	54.1	57.8	59.4	45.0	51.0	52.6
DrQA + PGNet (1-ctx)	64.2	63.7	67.1	68.3	71.4	57.8	63.1	65.1
BiDAF++ (3-ctx)	66.5	65.7	70.2	71.6	72.6	60.8	67.1	67.8
FLOWQA (1-Ans)	73.7	71.6	76.8	79.0	80.2	67.8	76.1	75.0
Human	90.2	88.4	89.8	88.6	89.9	86.7	88.1	88.8

Table 1: Model and human performance (% in F₁ score) on the CoQA test set

Speech Recognition

DARPA Challenges



Speech Recognition

HMM-based speech recognition

- Bell Labs "mixture-density HMM" (1985)
- CMU's Sphinx (1988)
- BBN's Byblos (1989)
- SRI's Decipher (1989)

Speech recognition datasets

- CSR corpus
- Switchboard corpus

DARPA's ATIS (1989-94): speech recognition for air travel (ATIS): BBN, MIT, CMU, AT&T, SRI, etc.

Speech Recognition

1994: Nuance (future Apple Siri)

1995: Voice Signal Technologies (future Semantic Machines, Microsoft)

2000: MIT's Pegasus for airline flights status and Jupiter for weather status/forecast

2000: AT&T's How May I Help You (HMIHY) for telephone customer care

Speech Recognition

Hybrid HMM-DNN

- Hinton (2009): using a DNN for acoustic modeling (plus an HMM for modeling the sequence of speech)
- Microsoft (2011)

INTERSPEECH 2011

Conversational Speech Transcription Using Context-Dependent Deep Neural Networks

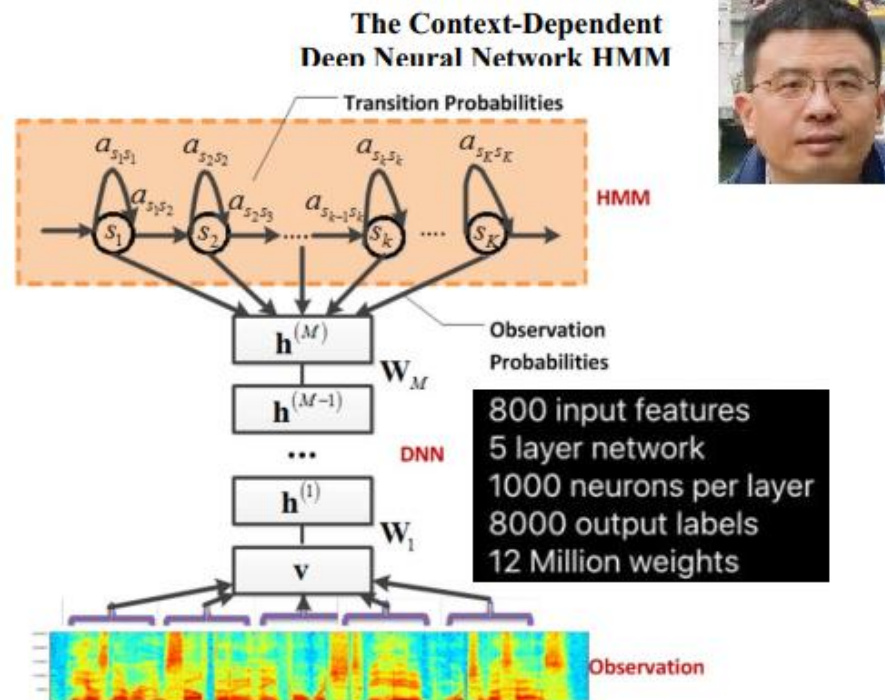
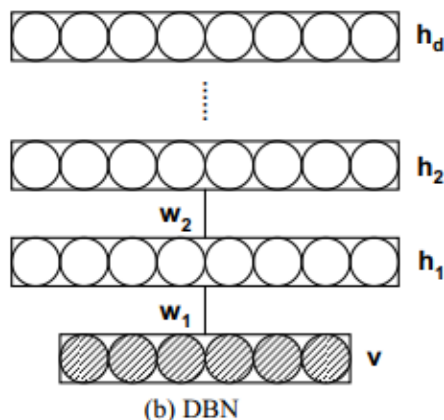


Microsoft

Frank Seide¹, Gang Li¹ and Dong Yu²

Deep Belief Networks for phone recognition

Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton
Department of Computer Science
University of Toronto



Speech Recognition

Apple's Siri (2011)

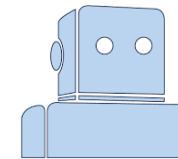
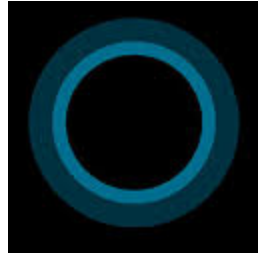
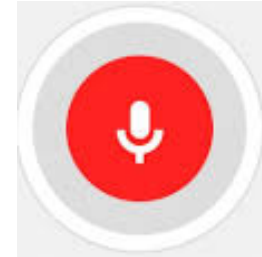
Google's Now (2012)

Microsoft's Cortana (2013)

Wit.ai (acquired by Facebook in 2015)

Amazon's Alexa (2014)

SoundHound's Hound (2016)



Wit.AI



Speech Recognition

Removing the HMM

2014: Alex Graves' CTC/LSTM
without HMM for speech
recognition (but high error
rate)

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

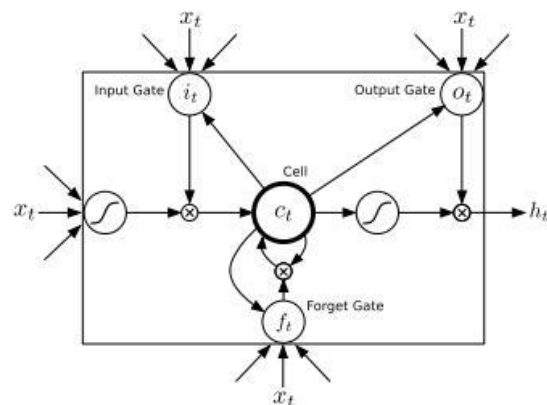


Alex Graves

Navdeep Jaitly



UNIVERSITY OF
TORONTO



minimise the CTC objective function:

$$CTC(x) = -\log \Pr(y^*|x)$$

Figure 1. Long Short-term Memory Cell.

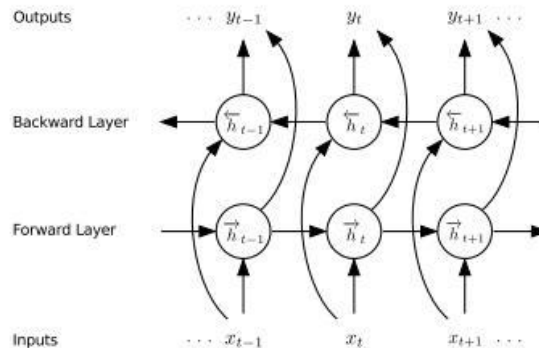


Figure 2. Bidirectional Recurrent Neural Network.

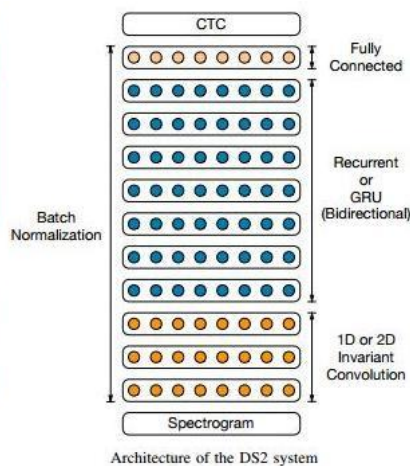
Speech Recognition

Removing the HMM

2014: Andrew Ng's CTC/GRU
without HMM (low error rate)

2015: LibriSpeech corpus

2015: Baidu's Deep Speech 2



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



Awni Y. Hannun

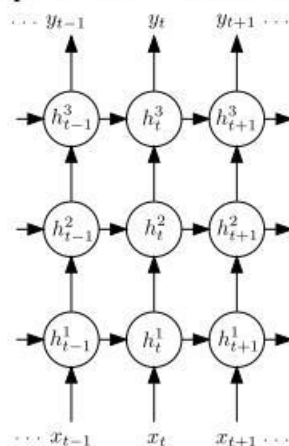
Andrew L. Maas

Daniel Jurafsky

Andrew Y. Ng

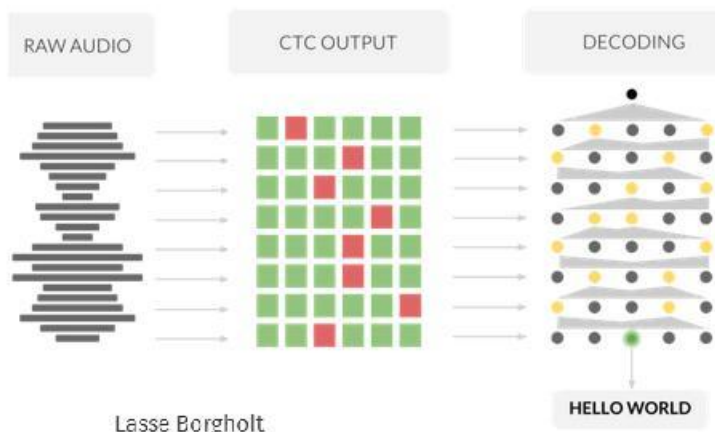


Deep Recurrent Neural Network



LANGUAGE MODEL

HELLO
WORLD 0.87
WALT 0.06
WORD 0.02
WAVE 0.02
WILL 0.01
WALL 0.01



Speech Recognition

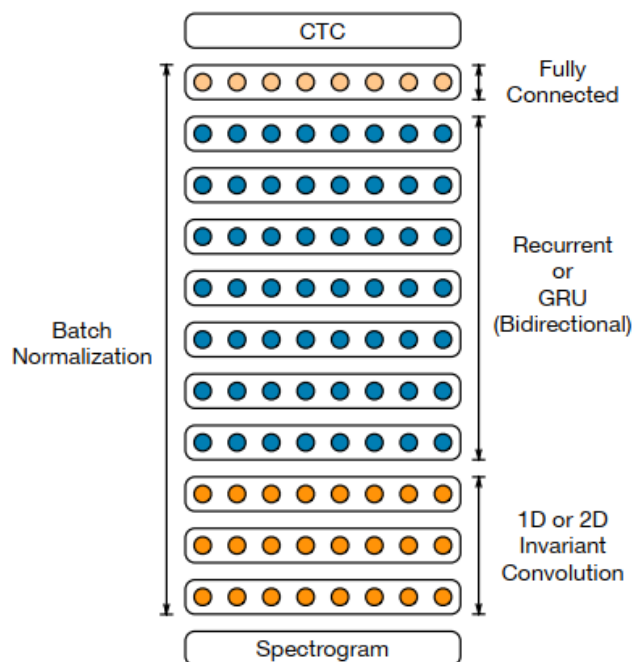
Baidu's Deep Speech 2

Dec 2015

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu



Speech Recognition

2016: Microsoft achieves human parity

- Three kinds of convolutional nets for acoustic modeling
 - VGG
 - ResNet
 - LACE (layer-wise context expansion with attention)
- An LSTM for language modeling

ACHIEVING HUMAN PARITY IN CONVERSATIONAL SPEECH RECOGNITION

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig

Microsoft Research
Technical Report MSR-TR-2016-71

Speech Synthesis

1940: Homer Dudley's vocoder (Britain)



1961: Louis Gerstman and Max Mathews program a computer to sing a song (Bell Labs)



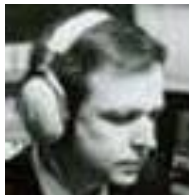
1966: Ryunen Teranishi and Noriko Umeda's text-to-speech system (Japan)



Ryunen Teranishi

Noriko Umeda

1972: Cecil Coker's talking computer (Bell Labs)



Speech Synthesis

Trivia (Coker's ventures into electronic music):

John Cage: Variation II (1966)

Bell Labs' 7", 33 1/3 RPM record
"Synthetic Voices For Computers"
(1970)



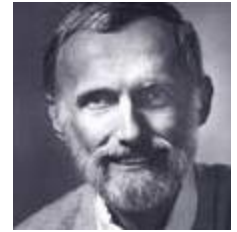
Speech Synthesis

1979: Dennis Klatt: MITalk (MIT)

1988: Francis Charpentier's
concatenative speech synthesis
(France)

1995: Keiichi Tokuda 's HMM-based
HTS (Japan)

1996: Alan Black's concatenative text-to-
speech (Japan)



Speech Synthesis

- Voice morphing: Festvox (Alan Black, 1997)



2015 application: voice morphing (faking someone's voice)

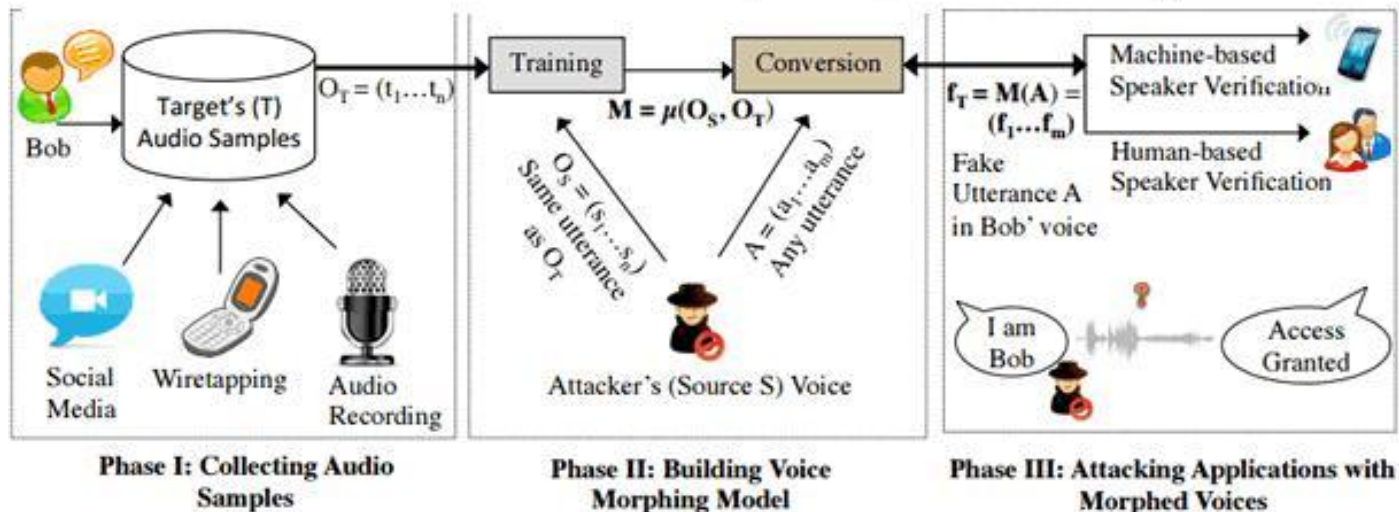


Fig. 1: An overview of our attack system

Next...

- See <http://www.scaruffi.com/singular> for the index of this Powerpoint presentation and links to the other parts

