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

- 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

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)
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

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

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

<table>
<thead>
<tr>
<th></th>
<th>animal</th>
<th>stable</th>
<th>village</th>
<th>gallop</th>
<th>jokey</th>
</tr>
</thead>
<tbody>
<tr>
<td>horse</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>run</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

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
Natural Language Processing

Neural networks that learn recursive structures

• Ronan Collobert and Jason Weston (2008): task-independent sequence tagging
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.
Natural Language Processing

Sequence tagging
- Generative models: hidden Markov models
- Conditional models: conditional random fields
  (John Laffery, 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

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
Department of Computer Science
University of Oxford

Phil Blunsom

models the conditional probability $P(f|e)$ of a sentence $f = f_1, \ldots, f_m$ in a target language $F$ being the translation of a sentence $e = e_1, \ldots, e_k$ 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.
Natural Language Processing

Neural Machine Translation

2014: Kyunghyun Cho’s Gated Recurrent Unit
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.
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
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

Glimpse Sensor:

A)

Glimpse Network:

B)
Natural Language Processing

Neural Machine Translation

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

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

14x14 Feature Map

1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation
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)
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.
Neural Machine Translation: convolutional nets instead of RNNs

2016: Nal Kalchbrenner’s ByteNet
Neural Machine Translation: convolutional nets instead of RNNs

2016: Facebook’s ConvS2S

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.
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
Natural Language Processing

Neural Machine Translation

Microsoft (2018): news translation

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
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)
Natural Language Processing

Question-answering

• 2014: James Weston’s memory networks

ICLR 2015

MEMORY NETWORKS

Jason Weston, Sumit Chopra & Antoine Bordes
Facebook AI Research
770 Broadway
New York, USA
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
Discourse Analysis
2015: Richard Socher's dynamic memory networks
Natural Language Processing

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

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
Discourse Analysis
2014: GloVe (an alternative to Word2vec) by Jeffrey Pennington, Richard Socher, Chris Manning (Stanford)
2015: FastText (Mikolov, Facebook)
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)

Skip-Thought Vectors
Natural Language Processing

Discourse Analysis
Next step: capture information about entire sentences

<table>
<thead>
<tr>
<th>Task</th>
<th>Previous SOTA</th>
<th>ELMo +</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>SAN</td>
<td>84.4</td>
</tr>
<tr>
<td>SNLI</td>
<td>Chen et al (2017)</td>
<td>88.6</td>
</tr>
<tr>
<td>SRL</td>
<td>He et al (2017)</td>
<td>81.7</td>
</tr>
<tr>
<td>Coref</td>
<td>Lee et al (2017)</td>
<td>67.2</td>
</tr>
<tr>
<td>NER</td>
<td>Peters et al (2017)</td>
<td>91.9</td>
</tr>
<tr>
<td>Sentiment (5-class)</td>
<td>McCann et al (2017)</td>
<td>53.7</td>
</tr>
</tbody>
</table>

Deep contextualized word representations
Matthew E. Peters¹, Mark Neumann¹, Mohit Iyyer⁹, Matt Gardner¹,
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark*, Kenton Lee*, Luke Zettlemoyer†*
{csquared,kenton1,lsz}@cs.washington.edu

¹Allen Institute for Artificial Intelligence
*Paul G. Allen School of Computer Science & Engineering, University of Washington
Natural Language Processing

Discourse Analysis

TagML (Matthew Peters, 2017): bidirectional language model

Figure 2: Overview of TagLM, our language model augmented sequence tagging architecture.
# Natural Language Processing

## Attention

<table>
<thead>
<tr>
<th>Name</th>
<th>Alignment score function</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive(*)</td>
<td>$\text{score}(s_t, h_i) = v_a^T \tanh(W_a[s_t; h_i])$</td>
<td></td>
</tr>
<tr>
<td>Location-Base</td>
<td>$\alpha_{t,i} = \text{softmax}(W_a s_t)$</td>
<td>Luong2015</td>
</tr>
</tbody>
</table>
| General            | $\text{score}(s_t, h_i) = s_t^T W_a h_i$  
where $W_a$ is a trainable weight matrix in the attention layer. | Luong2015 |
| Dot-Product        | $\text{score}(s_t, h_i) = s_t^T h_i$ | Luong2015 |
| Scaled             | $\text{score}(s_t, h_i) = \frac{s_t^T h_i}{\sqrt{n}}$ | Vaswani2017 |
| Dot-Product(\^)   | Note: very similar to the dot-product attention except for a scaling factor; where $n$ is the dimension of the source hidden state. |          |
| 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

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.
Natural Language Processing

Self-Attention
2016: Jianpeng Cheng, Mirella Lapata
Natural Language Processing

Self-Attention
2016: Ankur Parikh (Google)
2017: Richard Socher (Salesforce)
Natural Language Processing

Self-Attention

2017: Zhouhan Lin (Univ of Montreal)

A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING

Published as a conference paper at ICLR 2017

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
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)
Natural Language Processing

Self-Attention

2017: Google's "transformer"

Abstract

The dominant sequence transduction models are based on simple recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also consist of the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensed 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 25.4 BLEU on the WMT 2014 English-to-German translation task, improving upon 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 28.0 after training for 5 days on eight TPU v3, a small fraction of the training cost 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 and gated recurrent neural networks in particular have been firmly established as state of the art approaches in sequence modeling and 1Equal contribution. Listing order is random. Jakob proposed replacing BERT with attention, and made the effort to evaluate this idea. Ashish, with Elia, designed and implemented the first Transformer model and has built crucial intuition into the aspect of this work. Your proposal scale also contributed attention mechanisms and the hammer-free ablation research, and became the other person involved in nearly every detail. Ashish designed, implemented, tested and evaluated countless model variants in our original codebase and retransform. Ashish also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and optimizations, and Ashish and Elia spent countless long days adjusting various parts of and implementing optimizations, explaining our earlier codebase, greatly improving results and massively accelerating our research.

2017

54

Self-Attention

2017: Google's "transformer"
Natural Language Processing

Self-Attention
2017: Google's "transformer"

$$\text{Attention}(Q, K, V) = \text{softmax}\left( \frac{QK^T}{\sqrt{d_k}} \right)V$$
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)
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\textsuperscript{1,2}\textsuperscript{*} Ross Girshick\textsuperscript{2} Abhinav Gupta\textsuperscript{1} Kaiming He\textsuperscript{2}

\textsuperscript{1}Carnegie Mellon University \textsuperscript{2}Facebook AI Research

Figure 3. Examples of the behavior of a non-local block in res3 computed by a 5-block non-local model trained on Kinetics. These examples
Natural Language Processing

Howard & Ruder: UMLFiT (2018):
Natural Language Processing

Transformer architecture + transfer learning for pre-training:
OpenAI GPT (2018)
Google BERT (2018)

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
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
Natural Language Processing

BERT’s descendants:

• **StructBERT** (Alibaba, 2019)
• Zhenzhong Lan‘s **AIBERT**, a lite BERT (Google, 2019)
• Electra (Stanford, 2020)
• DeBERTa (Microsoft, 2020)


## Natural Language Processing

### GLUE scores (2019)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALBERT-Team Google Language</td>
<td>ALBERT (Ensemble)</td>
<td>89.4</td>
</tr>
<tr>
<td>2</td>
<td>王伟</td>
<td>ALICE v2 large ensemble (Alibaba DAMO NLP)</td>
<td>89.0</td>
</tr>
<tr>
<td>3</td>
<td>Microsoft D365 AI &amp; UMD</td>
<td>FreeLB-RoBERTa (ensemble)</td>
<td>88.8</td>
</tr>
<tr>
<td>4</td>
<td>Facebook AI</td>
<td>RoBERTa</td>
<td>88.5</td>
</tr>
<tr>
<td>5</td>
<td>XLNet Team</td>
<td>XLNet-Large (ensemble)</td>
<td>88.4</td>
</tr>
</tbody>
</table>
Natural Language Processing

SuperGLUE scores (2021)

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
<th>BoolQ</th>
<th>CB</th>
<th>COPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeBERTa / TuringNLRv4</td>
<td>90.3</td>
<td>90.4</td>
<td>95.7/97.6</td>
<td>98.4</td>
</tr>
<tr>
<td>T5 + Meena, Single Model (Meena Team - Google Brain)</td>
<td>90.2</td>
<td>91.3</td>
<td>95.8/97.6</td>
<td>97.4</td>
</tr>
<tr>
<td>SuperGLUE Human Baselines</td>
<td>89.8</td>
<td>89.0</td>
<td>95.8/96.9</td>
<td>100.0</td>
</tr>
<tr>
<td>T5</td>
<td>89.3</td>
<td>91.2</td>
<td>93.9/96.8</td>
<td>94.8</td>
</tr>
<tr>
<td>NEZHA-Plus</td>
<td>86.7</td>
<td>87.8</td>
<td>94.4/96.0</td>
<td>93.5</td>
</tr>
<tr>
<td>PAI Albert</td>
<td>86.1</td>
<td>86.1</td>
<td>92.4/96.4</td>
<td>91.0</td>
</tr>
<tr>
<td>RoBERTa (ensemble)</td>
<td>85.9</td>
<td>88.2</td>
<td>92.5/96.6</td>
<td>90.8</td>
</tr>
<tr>
<td>RoBERTa-mtl-adv</td>
<td>85.7</td>
<td>87.1</td>
<td>92.4/95.6</td>
<td>91.2</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>84.6</td>
<td>87.1</td>
<td>90.5/93.2</td>
<td>90.5</td>
</tr>
</tbody>
</table>

Leaderboard Version: 2.0
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
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)

XLNet: Generalized Autoregressive Pretraining for Language Understanding

Zhilin Yang\textsuperscript{1}, Zihang Dai\textsuperscript{1,2}, Yiming Yang\textsuperscript{1}, Jaime Carbonell\textsuperscript{1}, Ruslan Salakhutdinov\textsuperscript{1}, Quoc V. Le\textsuperscript{2}
\textsuperscript{1}Carnegie Mellon University, \textsuperscript{2}Google Brain

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

![Diagram of XLNet architecture with two-stream self-attention for target-aware representations.](Image)
The Evolved Transformer
David R. So \textsuperscript{1} Chen Liang \textsuperscript{1} Quoc V. Le \textsuperscript{1}

Transformer and Evolved Transformer architecture

\textbf{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 \textit{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

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
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

Eunsol Choi
University of Washington

Published as a conference paper at ICLR 2019.

Wen-tau Yih
Allen Institute for Artificial Intelligence

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)
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)

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)
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

Natural Language Processing

Sentiment Analysis
• Soumith Chintala (2015)
Natural Language Processing

Sentiment Analysis

• 2016: Peter Dodds & Chris Danforth (Univ of Vermont): text-based sentiment analysis
• 2017: Eric Chu & Deb Roy (MIT): visual and audio sentiment analysis

The emotional arcs of stories are dominated by six basic shapes

Audio-Visual Sentiment Analysis for Learning Emotional Arcs in Movies

Eric Chu
MIT Media Lab

Deb Roy
MIT Media Lab

MODELING

IMAGE

A

Sentiment

B

Adjective-noun label / embedding

C

Sentiment

D

Visual emotional arc

E

Audio emotional arc

F

CLUSTERING to find families of visual emotional arcs
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.
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"

Table 2: The performance of FusionNet and competing approaches on SQuAD hidden test set at the time of writing (Oct. 4th, 2017).
Discourse Analysis
2017: Quoc Le’s QANET
Natural Language Processing

Discourse Analysis
2018: Furu Wei's R-Net for reading-comprehension
## Natural Language Processing

### Text Comprehension

<table>
<thead>
<tr>
<th>Model</th>
<th>Published EM</th>
<th>Published F1</th>
<th>LeaderBoard EM</th>
<th>LeaderBoard F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR Baseline (Rajpurkar et al., 2016)</td>
<td>40.4</td>
<td>51.0</td>
<td>40.4</td>
<td>51.0</td>
</tr>
<tr>
<td>Dynamic Chunk Reader (Yu et al., 2016)</td>
<td>62.5</td>
<td>71.0</td>
<td>62.5</td>
<td>71.0</td>
</tr>
<tr>
<td>Match-LSTM with Ans-Ptr (Wang &amp; Jiang, 2016)</td>
<td>64.7</td>
<td>73.7</td>
<td>64.7</td>
<td>73.7</td>
</tr>
<tr>
<td>Multi-Perspective Matching (Wang et al., 2016)</td>
<td>65.5</td>
<td>75.1</td>
<td>70.4</td>
<td>78.8</td>
</tr>
<tr>
<td>Dynamic Coattention Networks (Xiong et al., 2016)</td>
<td>66.2</td>
<td>75.9</td>
<td>66.2</td>
<td>75.9</td>
</tr>
<tr>
<td>FastQA (Weissenborn et al., 2017)</td>
<td>68.4</td>
<td>77.1</td>
<td>68.4</td>
<td>77.1</td>
</tr>
<tr>
<td>BiDAF (Seo et al., 2016)</td>
<td>68.0</td>
<td>77.3</td>
<td>68.0</td>
<td>77.3</td>
</tr>
<tr>
<td>SEDT (Liu et al., 2017a)</td>
<td>68.1</td>
<td>77.5</td>
<td>68.5</td>
<td>78.0</td>
</tr>
<tr>
<td>RaSoR (Lee et al., 2016)</td>
<td>70.8</td>
<td>78.7</td>
<td>69.6</td>
<td>77.7</td>
</tr>
<tr>
<td>FastQAExt (Weissenborn et al., 2017)</td>
<td>70.8</td>
<td>78.9</td>
<td>70.8</td>
<td>78.9</td>
</tr>
<tr>
<td>ReasoNet (Shen et al., 2017b)</td>
<td>69.1</td>
<td>78.9</td>
<td>70.6</td>
<td>79.4</td>
</tr>
<tr>
<td>Document Reader (Chen et al., 2017)</td>
<td>70.0</td>
<td>79.0</td>
<td>70.7</td>
<td>79.4</td>
</tr>
<tr>
<td>Ruminating Reader (Gong &amp; Bowman, 2017)</td>
<td>70.6</td>
<td>79.5</td>
<td>70.6</td>
<td>79.5</td>
</tr>
<tr>
<td>jNet (Zhang et al., 2017)</td>
<td>70.6</td>
<td>79.8</td>
<td>70.6</td>
<td>79.8</td>
</tr>
<tr>
<td>Conductor-net</td>
<td>N/A</td>
<td>72.6</td>
<td>81.4</td>
<td></td>
</tr>
<tr>
<td>Interactive AoA Reader (Cui et al., 2017)</td>
<td>N/A</td>
<td>73.6</td>
<td>81.9</td>
<td></td>
</tr>
<tr>
<td>Reg-RaSoR</td>
<td>N/A</td>
<td>75.8</td>
<td>83.3</td>
<td></td>
</tr>
<tr>
<td>DCN+</td>
<td>N/A</td>
<td>74.9</td>
<td>82.8</td>
<td></td>
</tr>
<tr>
<td>AIR-FusionNet</td>
<td>N/A</td>
<td>76.0</td>
<td>83.9</td>
<td></td>
</tr>
<tr>
<td>R-Net (Wang et al., 2017)</td>
<td>72.3</td>
<td>80.7</td>
<td>76.5</td>
<td>84.3</td>
</tr>
<tr>
<td>BiDAF + Self Attention + ELMo</td>
<td>N/A</td>
<td>77.9</td>
<td><strong>85.3</strong></td>
<td></td>
</tr>
<tr>
<td>Reinforced Mnemonic Reader (Hu et al., 2017)</td>
<td>73.2</td>
<td>81.8</td>
<td>73.2</td>
<td>81.8</td>
</tr>
</tbody>
</table>

**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)
Natural Language Processing

Summarization

• Alexander Rush & Jason Weston at (Facebook, 2015) first to apply seq2seq to summarization
• Read-Again Summarization (Raquel Urtasun & Wenyuan 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

Figure 3: (a) A network diagram for the NNLM decoder with additional encoder element. (b) A network diagram for the attention-based encoder enc3.

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)
Natural Language Processing

Summarization

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

Summarization

IBM’s SummaRunner (Ramesh Nallapati, 2017)

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

Facebook’s AI streamlines sentences while preserving meaning
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

OpenAI

The Guardian

Opinion
AI can write just like me. Brace for the robot apocalypse
Hannah Jane Parkinson
# Natural Language Processing

## Text Generation
**OpenAI’s GPT2 (2019)**

<table>
<thead>
<tr>
<th>DATASET</th>
<th>METRIC</th>
<th>PREVIOUS RECORD</th>
<th>HUMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winograd Schema Challenge</td>
<td>accuracy (+)</td>
<td>70.70%</td>
<td>63.7%</td>
</tr>
<tr>
<td>LAMBADA</td>
<td>accuracy (+)</td>
<td>63.24%</td>
<td>59.23%</td>
</tr>
<tr>
<td>LAMBADA</td>
<td>perplexity (-)</td>
<td>8.6</td>
<td>99</td>
</tr>
<tr>
<td>Children’s Book Test Common Nouns (validation accuracy)</td>
<td>accuracy (+)</td>
<td>93.30%</td>
<td>85.7%</td>
</tr>
<tr>
<td>Children’s Book Test Named Entities</td>
<td>accuracy (+)</td>
<td>89.05%</td>
<td>82.3%</td>
</tr>
</tbody>
</table>

**Reading Comprehension:** answer questions about given passages

![Graph showing performance over number of parameters](image-url)
Natural Language Processing

Question-answering

OpenAI’s GPT2 (2019)

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

Language models after BERT and GPT
Natural Language Processing

Language models after BERT and GPT

Nvidia’s Megatron (2019)

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

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

Mohammad Shoeybi\textsuperscript{1,2}  Mustofa Patwary\textsuperscript{1,2}  Raul Puri\textsuperscript{1,2}  Patrick LeGresley\textsuperscript{2}  Jared Casper\textsuperscript{2}  Bryan Catanzaro\textsuperscript{2}

 Transformer Architecture

Mar 2020

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)

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel  
Google, Mountain View, CA 94043, USA

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

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)

DeBERTa: Decoding-enhanced BERT with Disentangled Attention

Pengcheng He¹, Xiaodong Liu², Jianfeng Gao², Weizhu Chen¹
¹ Microsoft Dynamics 365 AI  ² Microsoft Research
{pcheng, xiaodl, jfgao, wzchen}@microsoft.com

Algorithm 1 Disentangled Attention

Input: Hidden state $H$, relative distance embedding $P$, relative distance matrix $\delta$. Content projection matrix $W_{h,c}$, $W_{q,c}$; position projection matrix $W_{h,p}$, $W_{q,p}$

1: $K_c = HW_{h,c}Q_c$  
2: $V_c = HW_{q,c}Q_c$  
3: $A_{c} = -Q_cK_c^T$  
4: for $i = 0, ..., N - 1$ do
5:     $\bar{A}_{c,i} = Q_c[i,:]K_c$  
6:     end for
7: for $i = 0, ..., N - 1$ do
8:     $A_{mp}[i,j] = A_{mp}(\delta[i,j])$  
9:     end for
10: for $j = 0, ..., N - 1$ do
11:     $\bar{A}_{p,j} = K_c[j,:]*Q_c$  
12:     end for
13: for $j = 0, ..., N - 1$ do
14:     $A_{mp}[j,i] = A_{mp}(\delta[j,i])$  
15:     end for
16: for $i = 0, ..., N - 1$ do
17:     $A = A_{c} - A_{mp} + \bar{A}_{p,j}$  
18:     end for
19: $H_o = \text{softmax}(A)V_o$

Output: $H_o$

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)

Language Models are Few-Shot Learners

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

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 those 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 2015, 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.
A robot wrote this entire article. Are you scared yet, human?

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.
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.

by Gary Marcus and Ernest Davis
Natural Language Processing

The Hype of the 2010s...
What is really knew since 1989?

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

Mar 14, 2022

examine state of 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 knew 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
What is really knew since 1989?

- OpenAI’s DALL-E
What is really knew since 1989?

- OpenAI’s GLIDE: diffusion

Generative models:

GAN: Adversarial training

VAE: maximize variational lower bound

Diffusion models: Gradually add Gaussian noise and then reverse
What is really knew 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
Foundation Models

- Problems with foundation models
  - Computational power

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.
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…)

129
Foundation Models

• Can AI plan for worst-case scenarios?
2010s

- Conversational computing
  - Siri (2011)
  - Google Now (2012)
  - Amazon Alexa (2014)
  - Microsoft XiaoIce (2014)
  - Microsoft Tay (2016)

Stanley Kubrick (1968)
“2001: A Space Odyssey”

(mandatory Hollywood movie for AI presentation!)
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)

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.
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)
Chatbots

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

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
Chatbots

• Google’s Meena chatbot (2020)

Towards a Human-like Open-Domain Chatbot

Daniel Adiwardana  Minh-Thang Luong

Google

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

<table>
<thead>
<tr>
<th>Role</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>Does money buy happiness?</td>
</tr>
<tr>
<td>Bot</td>
<td>Depends how much money you spend on it.</td>
</tr>
<tr>
<td>User</td>
<td>What is the best way to buy happiness?</td>
</tr>
<tr>
<td>Bot</td>
<td>You just have to be a millionaire by your early 20s, then you can be happy.</td>
</tr>
<tr>
<td>User</td>
<td>This is so difficult!</td>
</tr>
<tr>
<td>Bot</td>
<td>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.</td>
</tr>
</tbody>
</table>

Table 5: An interactive example of multi-turn dialogue

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

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

- OpenAI’s ChatGPT (2022)
Chatbots

- OpenAI’s ChatGPT (2022)

**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)

ChatGPT frenzy sweeps China as firms scramble for homegrown options

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

The AI chatbot phenomenon is now making waves in China

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

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

Feb 21, 2023, 12:17pm
NLP Benchmarks

• Surpassing human performance in NLP

Microsoft

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

Microsoft DeBERTa surpasses human performance on the SuperGLUE benchmark

Published date: October 14, 2020

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)

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
<th>BoolQ</th>
<th>CB</th>
<th>COPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeBERTa / TuringNLRv4</td>
<td>90.3</td>
<td>90.4</td>
<td>95.7/97.6</td>
<td>98.4</td>
</tr>
<tr>
<td>T5 + Meena, Single Model (Meena Team - Google Brain)</td>
<td>90.2</td>
<td>91.3</td>
<td>95.8/97.6</td>
<td>97.4</td>
</tr>
<tr>
<td>SuperGLUE Human Baselines</td>
<td>89.8</td>
<td>89.0</td>
<td>95.8/98.9</td>
<td>100.0</td>
</tr>
<tr>
<td>T5</td>
<td>89.3</td>
<td>91.2</td>
<td>93.9/96.8</td>
<td>94.8</td>
</tr>
<tr>
<td>NEZHA-Plus</td>
<td>86.7</td>
<td>87.8</td>
<td>94.4/96.0</td>
<td>93.6</td>
</tr>
<tr>
<td>PAI Albert</td>
<td>86.1</td>
<td>88.1</td>
<td>92.4/96.4</td>
<td>91.8</td>
</tr>
<tr>
<td>RoBERTa (ensemble)</td>
<td>85.9</td>
<td>88.2</td>
<td>92.5/95.6</td>
<td>90.8</td>
</tr>
<tr>
<td>RoBERTa-mlt-adv</td>
<td>85.7</td>
<td>87.1</td>
<td>92.4/95.6</td>
<td>91.2</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>84.6</td>
<td>87.1</td>
<td>90.5/95.2</td>
<td>90.6</td>
</tr>
</tbody>
</table>
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 (Kiela 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
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)
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)
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)
The State of NLP in 2019

Lead-3 = first three sentences of the document

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PGNet (1-ctx)</td>
<td>49.0</td>
<td>43.3</td>
<td>47.5</td>
<td>47.5</td>
<td>45.1</td>
<td>38.6</td>
<td>38.1</td>
<td>44.1</td>
</tr>
<tr>
<td>DrQA (1-ctx)</td>
<td>46.7</td>
<td>53.9</td>
<td>54.1</td>
<td>57.8</td>
<td>59.4</td>
<td>45.0</td>
<td>51.0</td>
<td>52.6</td>
</tr>
<tr>
<td>DrQA + PGNet (1-ctx)</td>
<td>64.2</td>
<td>63.7</td>
<td>67.1</td>
<td>68.3</td>
<td>71.4</td>
<td>57.8</td>
<td>63.1</td>
<td>65.1</td>
</tr>
<tr>
<td>BiDAF++ (3-ctx)</td>
<td>66.5</td>
<td>65.7</td>
<td>70.2</td>
<td>71.6</td>
<td>72.6</td>
<td>60.8</td>
<td>67.1</td>
<td>67.8</td>
</tr>
<tr>
<td>FLOWQA (1-Ans)</td>
<td>73.7</td>
<td>71.6</td>
<td>76.8</td>
<td>79.0</td>
<td>80.2</td>
<td>67.8</td>
<td>76.1</td>
<td>75.0</td>
</tr>
<tr>
<td>Human</td>
<td>90.2</td>
<td>88.4</td>
<td>89.8</td>
<td>88.6</td>
<td>89.9</td>
<td>86.7</td>
<td>88.1</td>
<td>88.8</td>
</tr>
</tbody>
</table>

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

DARPA Challenges

more and more complicated 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)

Deep Belief Networks for phone recognition

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

(b) DBN
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)
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

Figure 1. Long Short-term Memory Cell.

minimise the CTC objective function:

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

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
Speech Recognition

Baidu's Deep Speech 2

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, Zhiquan 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)
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 ⅓ 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)
Next…

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