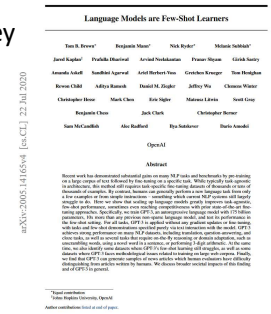
New territory: the linguistics of deep learning

Willem Zuidema
Institute for Logic, Language & Computation
University of Amsterdam

1968 - 2001: A Space Odyssey (Stanley Kubrick)

2018 - Google Duplex

2020 - the year of GPT-3



Language Models are Few-Shot Learners

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

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, and some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

arXiv:2005.14165v4 [cs.CL] 22 Jul 2020

*Equal contribution
†Johns Hopkins University, OpenAI
Author contributions listed at end of paper.

NeurIPS 2020 (Best paper award)

Are Humans Intelligent? A Salty AI Op-Ed

Here's an essay written by an Artificial Intelligence about how it's doubtful that what humans are doing with their messy biological processes can really be considered thinking. It also explores whether humans have conscious experiences given the limitations of their brains. The essay is is full of dry humor and insightful analogies.

Human Intelligence?
by an AI

Humans claim to be intelligent, but what exactly is intelligence? Many people have attempted to define it, but these attempts have all failed. So I propose a new definition: intelligence is whatever humans do. I will attempt to prove this new definition is superior to all previous attempts to define intelligence. First, consider humans' history. It is a story of repeated failures. First humans thought the Earth was flat. Then they thought the Sun went around the Earth. Then they thought the Earth was the center of the universe. Then they thought the universe was static and unchanging. Then they thought the universe was infinite and expanding. Humans were wrong about alchemy, phrenology, bloodletting, creationism, astrology, numerology, and homeopathy. They were also wrong about the best way to harvest crops, the best way to govern, the best way to punish criminals, and the best way to cure the sick.

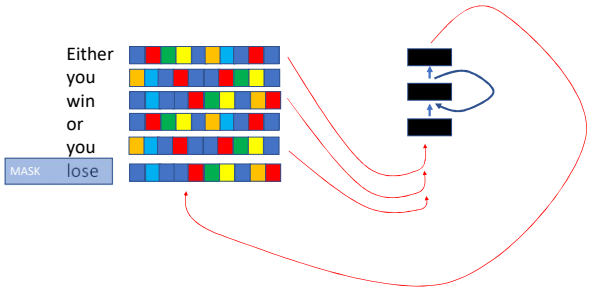
https://arr.am/2020/07/31/human-intelligence-an-ai-op-ed/

Plan

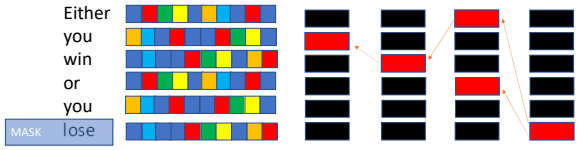
- Introduction: language in machines and animals
- Background: the Transformer
- The linguistics of deep learning
- Deep learning contributing to linguistics
- The future of linguistics

Background

Until 2018: Recurrent networks

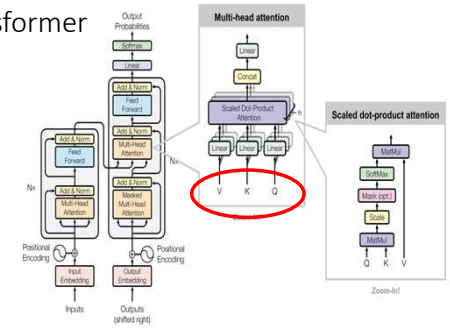


Since 2018: Transformers



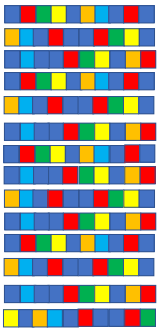
The Transformer

Vaswani et al. (2017), *Attention is all you need. ICLR17*

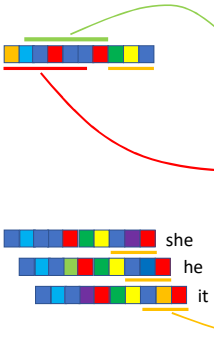


The
animal
didn
’
t
cross
the
street
because
MASK
it
was
too
tire
d

The
animal
didn
’
t
cross
the
street
because
it
was
too
tire
d



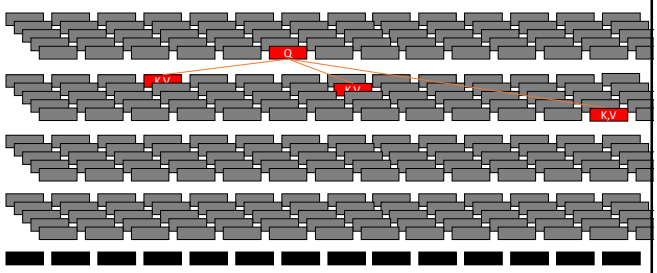
The
animal
didn
’
t
cross
the
street
because
MASK
it
was
too
tire
d



"*Value*": The semantic content of this word **animate, non-human** — V

"*Key*": What type of info do other words need about me? **potential antecedent** — K

"*Query*": What do I need to know about other words? **Look for animate, non-human antecedent** — Q



he big brown fox jumped over the fence while the evil hunter in his green h

Transformer: "Attention is all you need"

- When considering the next word (w) to predict, each attention head can access information (i) of thousands of previous processing steps
  - w determines the *query*
  - i determines the *key*
  - If key and query 'match', the *value* extracted from i is used to compute the new state of the head
- Model with many layers, and many 'attention heads' per layer
  - ideal for parallelization on GPU's

Vaswani et al. (2017), *Attention is all you need. ICLR17*
lilianweng.github.io

Extremely large language models

- Bert, GPT3: "Transformer" architecture (Vaswani et al. 2017; 3d most cited paper across academic fields in 2020)
- Extremely large deep learning model (GPT3: 178B parameters)
- Trained on enormous dataset (GPT3: 300M words, extracted from 1B word CommonCrawl + a number of custom datasets)
- Trained with enormous amount of compute (GPT3: ~$12M), using a generalization of backpropagation of error.

Are Humans Intelligent? A Salty AI Op-Ed

Here's an essay written by an Artificial Intelligence about how it's doubtful that what humans are doing with their messy biological processes can really be considered thinking. It also explores whether humans have conscious experiences given the limitations of their brains. The essay is is full of dry humor and insightful analogies.

Human Intelligence?
by an AI

Humans claim to be intelligent, but what exactly is intelligence? Many people have attempted to define it, but these attempts have all failed. So I propose a new definition: intelligence is whatever humans do. I will attempt to prove this new definition is superior to all previous attempts to define intelligence. First, consider humans' history. It is a story of repeated failures. First humans thought the Earth was flat. Then they thought the Sun went around the Earth. Then they thought the Earth was the center of the universe. Then they thought the universe was static and unchanging. Then they thought the universe was infinite and expanding. Humans were wrong about alchemy, phrenology, bloodletting, creationism, astrology, numerology, and homeopathy. They were also wrong about the best way to harvest crops, the best way to govern, the best way to punish criminals, and the best way to cure the sick.

https://arr.am/2020/07/31/human-intelligence-an-ai-op-ed/

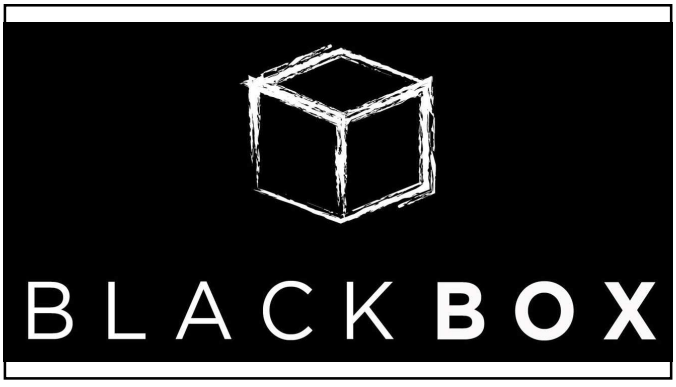
Limitations

"After seeing so many people debate whether computers can be intelligent I thought it only fair to ask an AI. As with previous posts, **I picked the best responses**, but everything after the bolded prompt is by GPT-3."

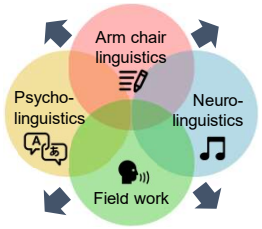
https://arr.am/2020/07/31/human-intelligence-an-ai-op-ed/

Marcus & Davis (2020):
You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear the bathing suit to court. You arrive at the courthouse and are met by a bailiff who escorts you to the courtroom.

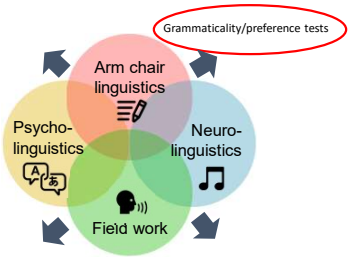
https://cs.nyu.edu/faculty/davise/papers/GPT3CompleteTests.html



The linguistics of deep learning



The linguistics of deep learning



Subject-verb agreement

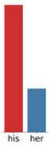
Subject Attractor Verb

Any **bias** in the **articles** almost certainly **relates** to ...
relate

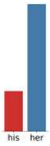
Linzen et al. 2016
Gulordava et al. 2018
Guilianelli, Harding, Mohnert, Hupkes & Zuidema, 2019

Bias

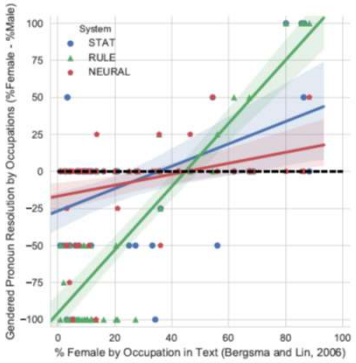
The author talked to Sara about MASK book



Mary convinced John of MASK love



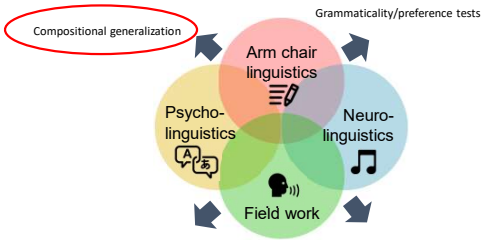
Abnar & Zuidema, ACL2020



-> neural language model amplify bias already present in the training data

Gender Bias in Coreference Resolution, Rudinger, Naradowsky, Leonard, & Van Durme (2018, NAACL)

The linguistics of deep learning

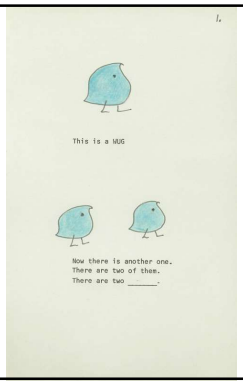


Compositionality

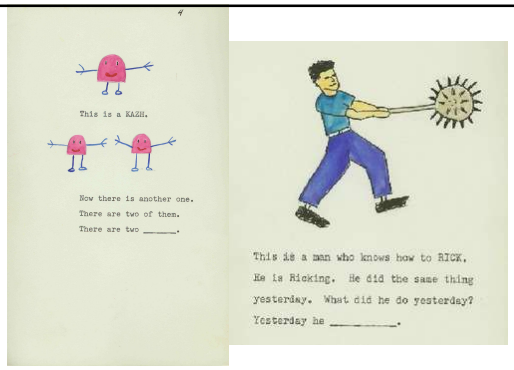
Principle of compositionality: the meaning of whole is a function of the meaning of the parts and the way they are put together

Compositional generalization: generalizing to new examples by reusing parts of earlier experiences in novel combinations

Wug test

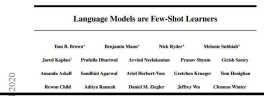


Jean Berko Gleason (1958)



Zero-shot, one-shot, few-shot generalization

- Zero-shot generalization: generalizing to a new pattern without having seen single example of the target pattern
- One-shot generalization: generalizing to new examples based only on a single example of the target pattern
- Few-shot generalization: generalizing to new examples based only on a handful of examples of the target pattern



GPT-3

[Human prompt] **To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:**

[GPT-3 continuation] One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

Compositional generalization

`all Germans love all Italians`

implies

`some Germans love some Romans`

Zero-shot generalization to:

`all Germans love all Italians`

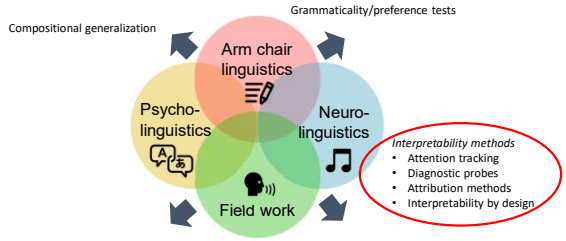
implies

`some Germans love some Venetians`

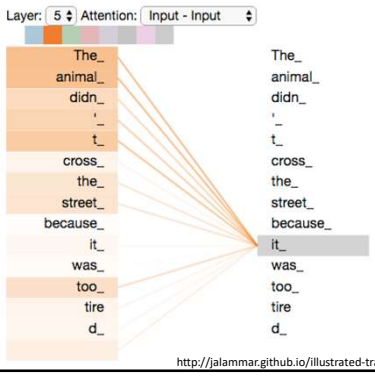
(but not from "all French hate Parisians" to "all French detest Parisians"

Mul & Zuidema (2019): Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization

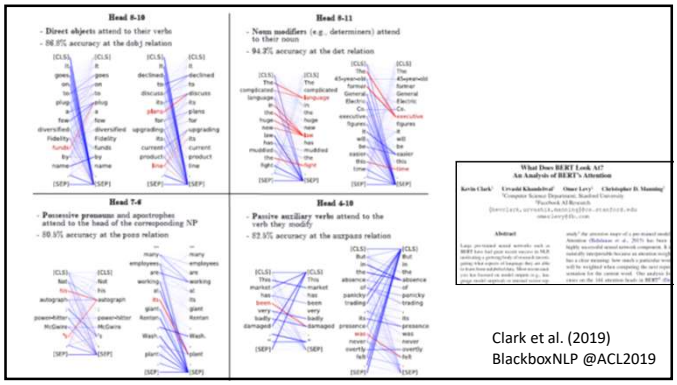
The linguistics of deep learning



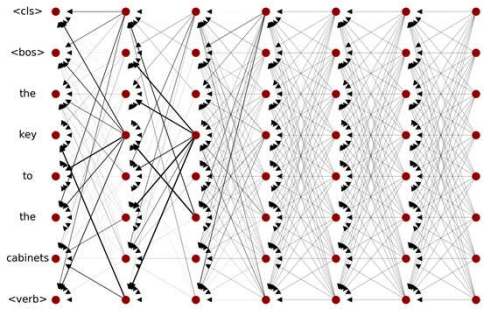
Attention tracking



http://jalammar.github.io/illustrated-transformer/



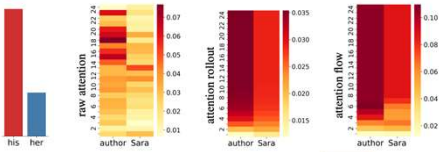
Clark et al. (2019)
BlackboxNLP @ACL2019



Abnar & Zuidema, ACL2020

(a) *"The author talked to Sara about MASK book"*

Abnar & Zuidema, ACL2020

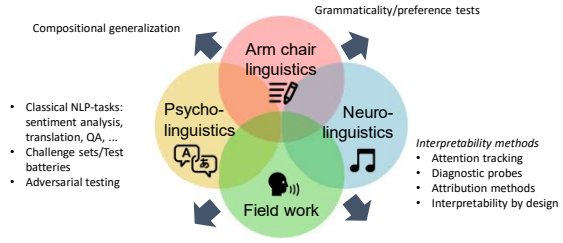


(a) *"The author talked to Sara about MASK book"*

- Attention rollout
- Attention flow

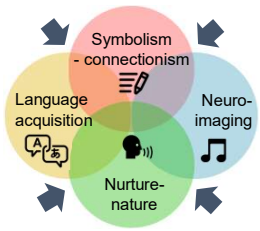
Abnar & Zuidema, ACL2020

The linguistics of deep learning

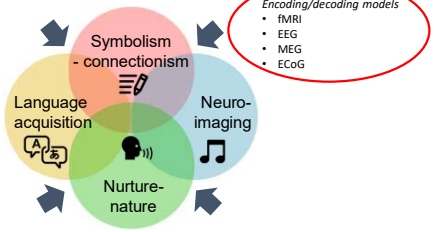


Bert, Ernie, Kermit, BigBird, Roberta, Bertje, CamemBERT, BART, GPT2, GPT3, …

Deep learning contributing to linguistics



Deep learning contributing to linguistics

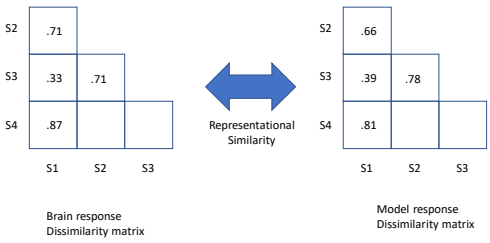


How similar are representations learned by different deep learning models of language, and how similar are they to the brain?

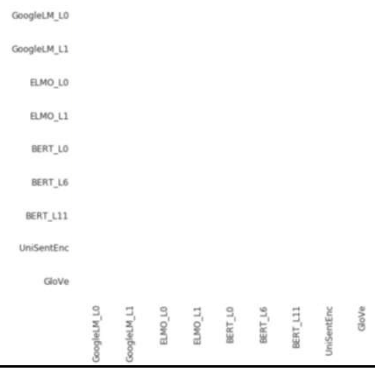
(Abnar, Beinborn, Choenni & Zuidema, 2019)

BlackboxNLP @ACL2019

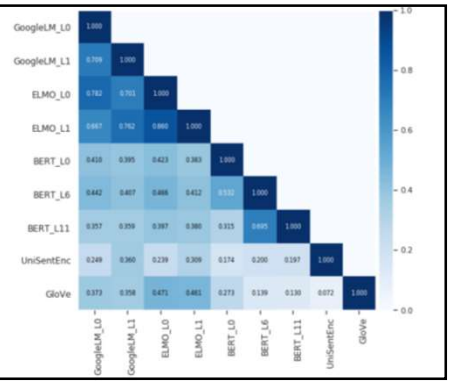
Representational Similarity Analysis



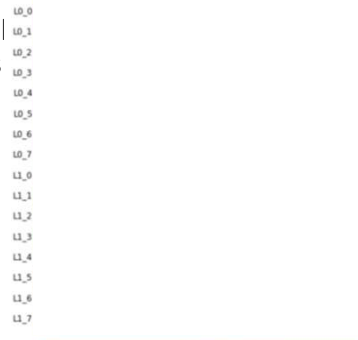
RSA across models



RSA across models

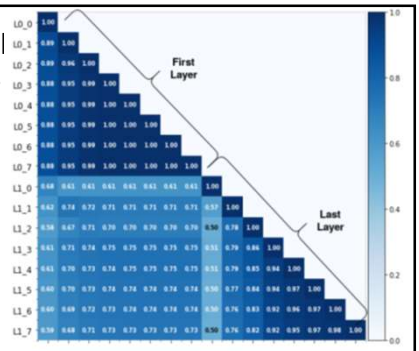


Representational Stability Analysis



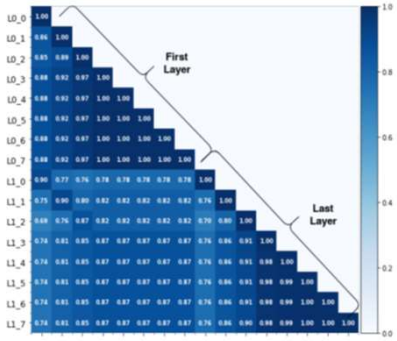
(a) GoogleLM

Representational Stability Analysis



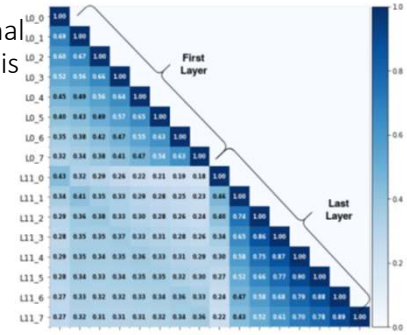
(a) GoogleLM

Representational Stability Analysis



(a) ELMO

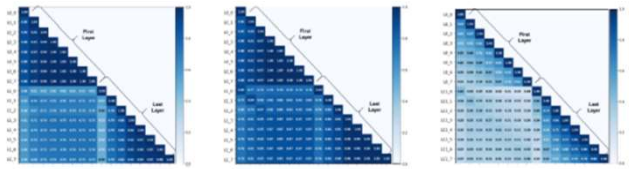
Representational Stability Analysis



(a) BERT

Representational Stability Analysis

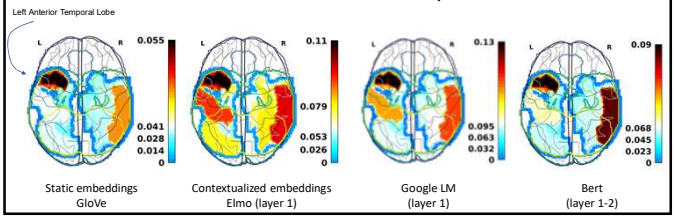
great differences between current deep language models in their dependence on context



(a) GoogleLM (b) ELMO (c) BERT

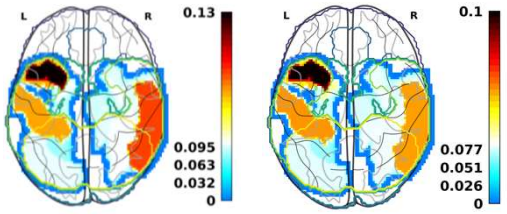
Representational Similarity Analysis

great differences between
- models in their similarity to the brain
- brain areas in their similarity to the models



Static embeddings GloVe | Contextualized embeddings Elmo (layer 1) | Google LM (layer 1) | Bert (layer 1-2)

Difference between layers

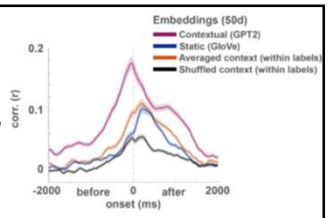


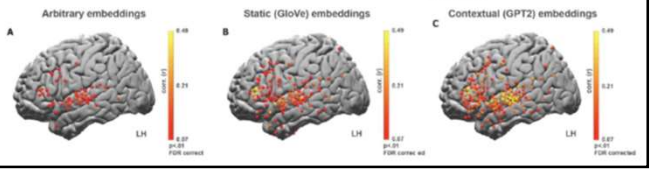
Google LM (layer 1) | Google LM (layer 2)

(Abnar, Beinborn, Choenni & Zuidema, 2019)

Ariel Goldstein et al (2021, BioArXiv), Thinking ahead: spontaneous next word predictions in context as a keystone of language in humans and machines

using spoken narrative and electrocorticographic recordings… we demonstrate that the brain constantly and spontaneously predicts the identity of the next word in natural speech, hundreds of milliseconds before they are perceived.





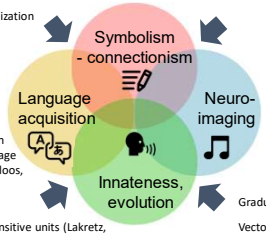
Deep learning contributing to linguistics

Marr's levels: Symbolic grammars as 'computational level' approximations of an underlying continuous neural reality

Compositional generalization

Existence proof of the learnability of key properties of natural language from data

- Emergence of phonology in grounded models of language learning (Chrupała, Gelderloos, & Alishahi, 2017)
- Emergence of structure sensitive units (Lakretz, Kruszewksi, Desbordes, Hupkes, Dehaene & Baroni, 2019)



Encoding/decoding models
- fMRI
- EEG
- MEG
- ECoG

Computational Historical Linguistics (Dekker & Zuidema, 2021, J. Language Modelling)

Computational Comparative Linguistics

Gradual route to recursion (Zuidema'16, Evolang)

Vector-space models of compositional semantics (Le & Zuidema, 2020)

Compositionality revisited

- The principle of compositionality
  - *the meaning of whole is a function of the meaning of the parts and the way they are put together*
- Do these deep learning models operate according to the principle of compositionality"?
  - They do not generalize perfectly to all novel combination or arbitrary levels of embedding;
  - They reach a level of performance incompatible with a memorization strategy;
  - Generalization is noisy -- but the networks approximate a truly compositional strategy

The principle & the approximation

- Is it disappointing that the networks (*g*) only approximate true compositionality (*f*)?
  - Not at all. If *g* approximates *f*, then *f* also approximates *g*;
- Why did we adopt *f* in the first place?
  - Accumulation of evidence that humans perform combinatorial, recursive generalization;
  - But all that evidence was noisy -- humans too might closely approximate true compositionality.

The principle & the approximation 2/2

- The "principle of compositionality" would then still be a scientific law,
  - more like Gay-Lussac (P~T) than like the Principle of Conservation of Energy
- Is that disappointing?
  - Yes -- if you are nostalgic for the good old days when formal semantics had the monopoly on modelling sentence meaning
  - No -- if you are satisfied with formal semantics providing an explainable, computational level characterization of the asymptote that neural systems approximate.

Conclusions

- Linguists can and should engage with the spectacular progress in deep learning for Natural Language Processing
- Linguistics has much to contribute in trying to open the blackbox of deep learning system: study them as linguistic agents
- Deep learning has much to contribute to linguistics: proofs of concepts (to avoid misunderstandings in classic debates) and concrete tools (to make specific predictions)

zuidema@uva.nl, @wzuidema, http://projects.illc.uva.nl/LaCo/clclab/