Cognitive Plausibility of Deep Language Models

Day 4 & 5

Lisa Beinborn
VU Amsterdam

Nora Hollenstein
U Copenhagen

Willem Zuidema
Uv Amsterdam



Interpretability of Neural Language Models

Willem Zuidema
@AmsterdamNLP

Plan

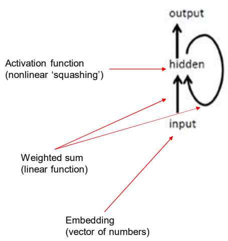
1. LSTM language models
   - recap & deep dive - key feature: trainable gates
   - interpretability case study I: integrated gradients
   - interpretability case study II: diagnostic classifiers (probes)

2. Transformer language models
   - recap & deep dive - key features: trainable attention & content-adressability
   - interpretability case study III: attention tracking
   - interpretability case study IV: representational similarity

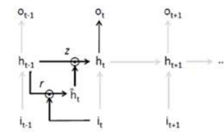
3. The bigger picture: The Linguistics of Deep Learning

Background I: Gating in Recurrent Networks

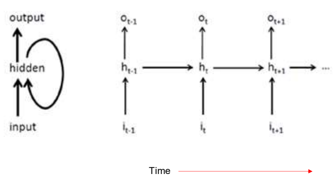
Recurrent network



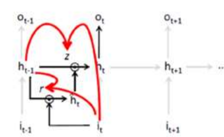
Add reset gate



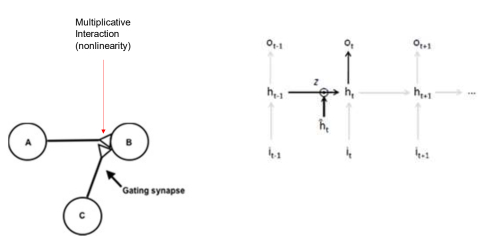
Unfold in time



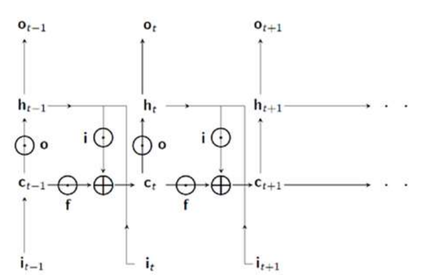
Make gates trainable: GRU (Cho et al., 2015)



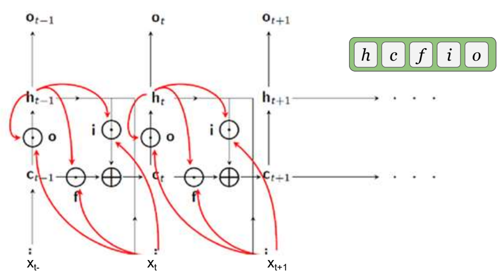
Add update gate



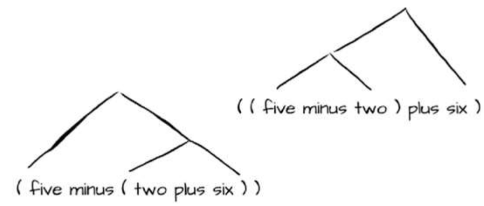
Add memory cell & forget gate



Long-short term memory (Hochreiter & Schmidhuber 1997)



Arithmetic Language

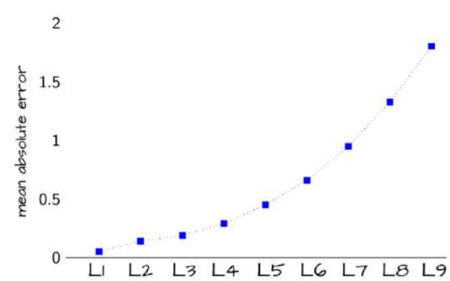


(Veldhoen et al. 2016; Hupkes, Veldhoen & Zuidema, 2018, JAIR)

Can recurrent networks learn languages with combinatorial and hierarchical structure?

Contextfreeness	**Palindromes**	**Arithmetics**
A^nB^n, $n \geq 1$	(ww^r), $w \in$ {A,B,C,D}*	(Expr Op Expr)
AAAABBBB	DABBBBAD	(1 + 3) - (7 - 9)
AABB	BB	(2 - 7)
AAAAAABBBBBB	ACCCCCCCCA	(9 - 3) - ((4 + 2) - (5 + 6))
…	…	…
Negative exs:	**Negative exs:**	**Targets:**
AABBB	AABBCCBBA	6; -5; 11
ABABABAB	DDBCADDBCA	

Results



(Veldhoen et al. 2016; Hupkes, Veldhoen & Zuidema, 2018, JAIR)

How do you demonstrate that a neural network has really learned the intended generalization?

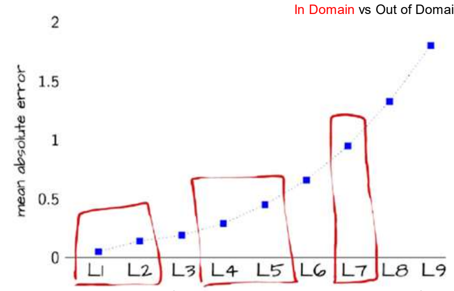
Behavioral tests:

- Generate examples, divide randomly in train set and test set
- Accuracy of test set := 'in domain' generalization

- Leave specific categories of examples out from train set
- Accuracy on various test sets := 'out of domain' generalization

Results

In Domain vs Out of Domain generalization



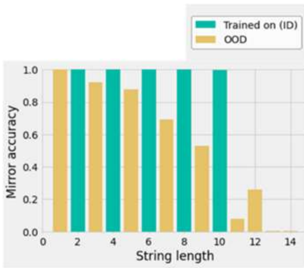
(Veldhoen et al. 2016; Hupkes, Veldhoen & Zuidema, 2018, JAIR)

Palindrome language

Strings of the form: `gbhdd#ddhbg`
Alphabet: 8 symbols (`a-h`)
Lengths train set: 5,9,13,17,21

`gbhdd#ddhbg`
`abbhffga#agffhbba`
`aaaaaaaaaa#aaaaaaaaaa`

LSTM, 1-layer, Dim=170



(Jumulet & Zuidema, in prep)

Case study 1: Attribution methods / Integrated Gradients

Which words are important to generate each prediction?

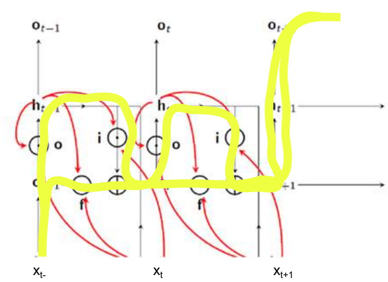
How do you demonstrate that a neural network has really learned the intended generalization?

Behavioral tests:
- Generate examples, divide randomly in train set and test set
- Accuracy of test set := 'in domain' generalization

- Leave specific categories of examples out from train set
- Accuracy on various test sets := 'out of domain' generalization

What level of accuracy do we consider convincing proof?

Tracking gradients



∂**h** / ∂L
∂**c** / ∂L
∂**f** / ∂L
∂**i** / ∂L
∂**x** / ∂L

How do you demonstrate that a neural network has really learned the intended generalization?

Behavioral tests:
- Generate examples, divide randomly in train set and test set
- Accuracy of test set := 'in domain' generalization

- Leave specific categories of examples out from train set
- Accuracy on various test sets := 'out of domain' generalization

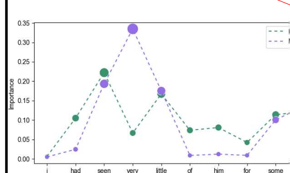
What level of accuracy do we consider convincing proof?

Can we look inside and characterize the learned solution? → **Interpretability**

Saliency 1:= length of the gradient

Beinborn & Hollenstein, "Relative Importance in Sentence Processing", ACL' 21,

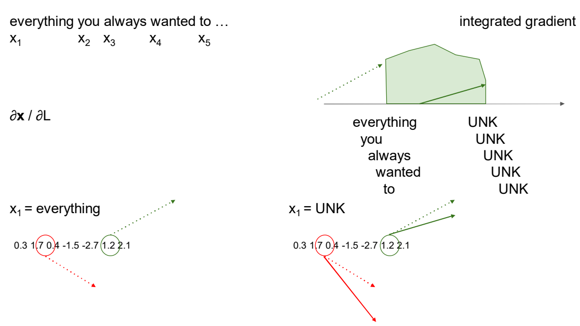
∂**x** / ∂L



We iterate over each token vector $\mathbf{x}_i$ in our input sequence $x_1, x_2, \ldots x_n$. Let $\mathbf{X}_i$ be the input matrix with $\mathbf{x}_i$ being masked. The saliency s_{ij} for input token $\mathbf{x}_j$ for the prediction of the correct token $\mathbf{t}_i$ is then calculated as the Euclidean norm of the gradient of the logit for x_i.

$$s_{ij} = \|\nabla_{\mathbf{x}_j} f_{t_i}(\mathbf{X}_i)\|_2 \qquad (1)$$

The saliency vector $\mathbf{s}_i$ indicates the relevance of each token for the correct prediction of the masked token t_i.[1] The saliency scores are normalized by dividing by the maximum. We determine the rel-



Saliency methods

- Gradient-based
  - Gradient norm
  - Integrated Gradients
  - DeepLIFT
  - Layer-wise relevance propagation
- Shapley-based
  - Pure Shapley
  - SHAP
  - Contextual Decomposition
  - Generalized Contextual Decomposition
- Omission scores
- ….

Saliency 2:= Integrated Gradients

The integrated gradient along the i^{th} dimension for an input x and baseline x' is defined as follows. Here, $\frac{\partial F(x)}{\partial x_i}$ is the gradient of $F(x)$ along the i^{th} dimension.

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha \qquad (1)$$

→ Competitive "attribution method": a method to attribute importance/ assign credit about a decision to input elements.

Lessons case study 1

Positives

- Attribution methods such as Integrated Gradients can help us assign importance scores to the input
- Applied to the palindrome network, they can help confirm that the network has learned to pay attention to the mirror symbol at each generation step
- The methods are *data-driven*

Negatives

- Attribution methods are *local*: they assign credit for each specific instance, but a *global* understanding of the learned function still requires much work

$b^+a^+c^+g^+b^+d^+c^+h^+\#h^-c^-d^-b^-g^-c^-a^-b^-$

Token	Token Contributions
a+	b+
c+	b+ a+
g+	b+ a+ c+
b+	b+ a+ c+ g+
d+	b+ a+ c+ g+ b+
c+	b+ a+ c+ g+ b+ d+
h+	b+ a+ c+ g+ b+ d+ c+
#	b+ a+ c+ g+ b+ d+ c+ h+

$b^+a^+c^+g^+b^+d^+c^+h^+\#h^-c^-d^-b^-g^-c^-a^-b^-$

Token	Token Contributions
h-	b+ a+ c+ g+ b+ d+ c+ h+ #
c-	b+ a+ c+ g+ b+ d+ c+ h+ # h-
d-	b+ a+ c+ g+ b+ d+ c+ h+ # h- c-
b-	b+ a+ c+ g+ b+ d+ c+ h+ # h- c- d-
g-	b+ a+ c+ g+ b+ d+ c+ h+ # h- c- d- b-
c-	b+ a+ c+ g+ b+ d+ c+ h+ # h- c- d- b- g-
a-	b+ a+ c+ g+ b+ d+ c+ h+ # h- c- d- b- g- c-
b-	b+ a+ c+ g+ b+ d+ c+ h+ # h- c- d- b- g- c- a-

(expected gradients: average IG over multiple baselines)

(Jumulet & Zuidema, in prep)

a.k.a. "probing classifiers"

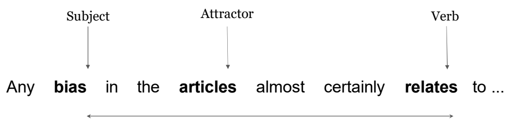
Case study 2: Diagnostic Classification

(Veldhoen et al. 2016; Hupkes, Veldhoen & Zuidema, 2018, JAIR)

Hypothesis-driven, global interpretability methods

Giulianelli, Harding, Mohnert, Hupkes & Zuidema, 2018
Best Paper award BlackboxNLP @EMNLP

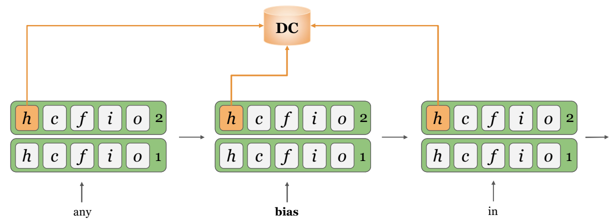
Subject-verb agreement



Linzen et al. 2016

Gulordava et al. 2018

Diagnostic Classification



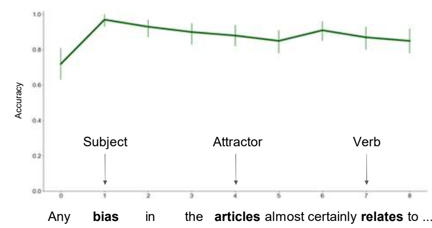
Experimental Setup

- Pretrained Neural Language Model from (Gulordava et al. 2018) with 2 LSTM-layers, with 650 hidden units each
- Wikipedia dependency dataset (Linzen et al. 2016)
- Extract activations for components $\mathbf{h_t}, \mathbf{c_t}, \mathbf{f_t}, \mathbf{i_t}, \mathbf{o_t}$ during forward pass of the LSTM

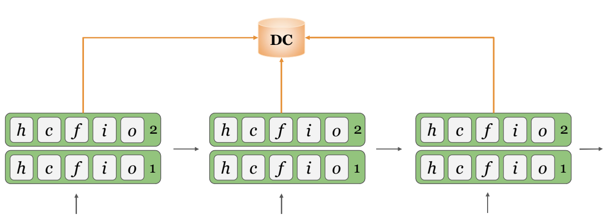
(Giulianelli, Harding, Mohnert, Hupkes & Zuidema, 2018)

Diagnostic Classification to Predict Number

Train: 1000 sentences, context size 5, at least 1 word before subject and at least 1 words after verb.
Test: Two sets of circa 100 sentences with 1 agreement attractor, according to correct/wrong number prediction.

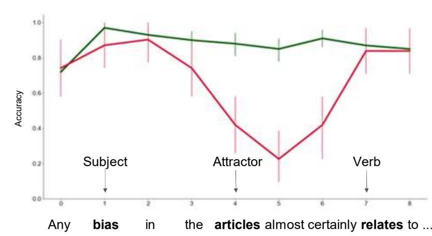


Diagnostic Classification

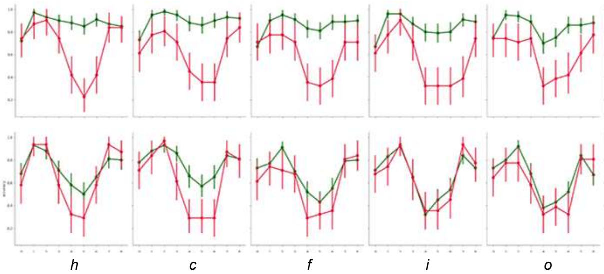


Diagnostic Classification to Predict Number

Train: 1000 sentences, context size 5, at least 1 word before subject and at least 1 words after verb.
Test: Two sets of circa 100 sentences with 1 agreement attractor, according to correct/wrong number prediction.



Diagnostic Classification to Predict Number



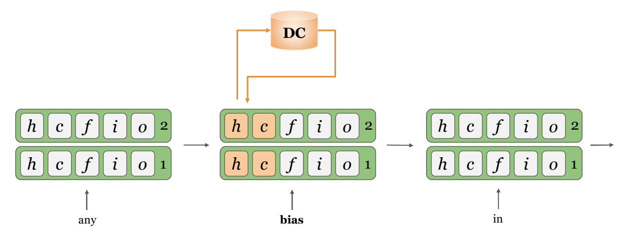
Diagnostic Classification to Predict Number

test time		0 subject	1	2	3	4	5	6 verb
relates	6	0.71	0.85	0.92	0.88	0.92	0.91	0.95
certainly	5	0.84	0.88	0.91	0.93	0.95	0.96	0.87
almost	4	0.64	0.76	0.85	0.86	0.88	0.88	0.75
articles	3	0.23	0.77	0.88	0.9	0.89	0.87	0.85
the	2	0.55	0.92	0.97	0.93	0.88	0.92	0.73
in	1	0.85	0.98	0.96	0.95	0.94	0.95	0.85
bias	0	0.99	0.93	0.89	0.85	0.89	0.94	0.7

train time

How is number agreement information processed across timesteps?

Influencing language models with diagnostic classifiers



Characterizing the dynamics of mental representations: the temporal generalization method

J-R. King[1,2,3] and S. Dehaene[1,2,4,5]

[1] Cognitive Neuroimaging Unit, Institut National de la Santé et de la Recherche Médicale, U992, F-91191 Gif/Yvette, France
[2] NeuroSpin Center, Institute of BioImaging Commissariat à l'Énergie Atomique, F-91191 Gif/Yvette, France
[3] Institut du Cerveau et de la Moelle Épinière Research Center, Institut National de la Santé et de la Recherche Médicale, U975 Paris, France
[4] Université Paris 11, Orsay, France
[5] Collège de France, F-75005 Paris, France

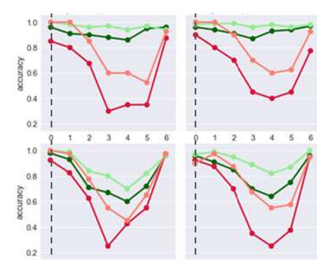
Parsing a cognitive task into a sequence of operations is a central problem in cognitive neuroscience. We argue that a major advance is now possible owing to the

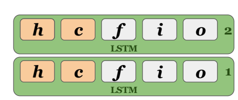
series of steps from those operating as a continuous flow or 'cascade' of overlapping stages [5].

More recently, the advent of brain-imaging techniques

Trends in Cognitive Science, 2014

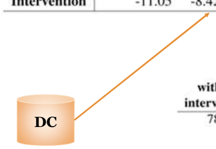
Diagnostic Classification to Predict Number





Influencing language models with diagnostic classifiers

	An	official	estimate	issued	in	2003	suggests	suggest
Original		-11.05	-8.426	-8.472	-1.243	-3.951	-5.753	**-5.6979**
Intervention		-11.05	-8.426	-8.472	-1.268	-3.97	**-5.691**	-6.4361



without intervention	with intervention
78.0	85.4

Plan

1. LSTM language models
- recap & deep dive - key feature: trainable gates
- interpretability case study I: integrated gradients
- interpretability case study II: diagnostic classifiers (probes)

2. Transformer language models
- recap & deep dive - key features: trainable attention & content-adressability
- interpretability case study III: attention tracking
- interpretability case study IV: representational similarity

3. The bigger picture: The Linguistics of Deep Learning

Lessons case study 2

Positives
- Diagnostic Classifiers allow us to track the dynamics of subject-verb agreement in an LSTM-based language model
- Temporal Generalization Method shows the LSTM represents number information in at least two different ways
- An intervention study allows us to go beyond correlation, but shows a causal role for the representations we identified

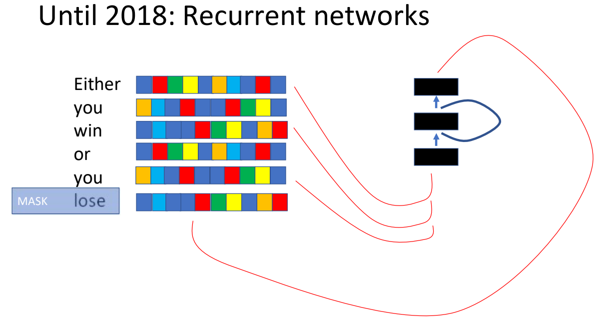
Negatives
- Diagnostic classification is hypothesis-driven: we need good hypotheses first -- no complete understanding of the learned function reached yet
- Classifiers (especially nonlinear) may yield false positives, and false negatives

Background II:
Attention in Transformers

Questions?

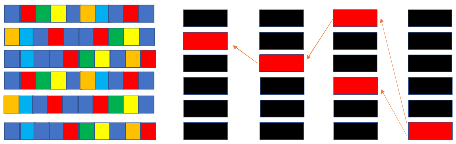
45

Until 2018: Recurrent networks

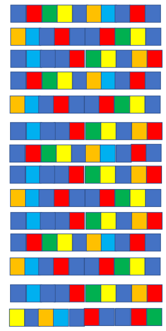


Since 2018: Transformers

Either
you
win
or
you
MASK lose

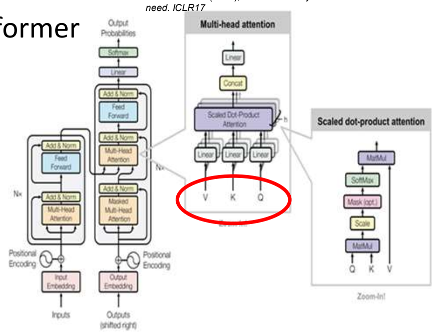


The
animal
didn
’
t
cross
the
street
because
it
was
too
tire
d

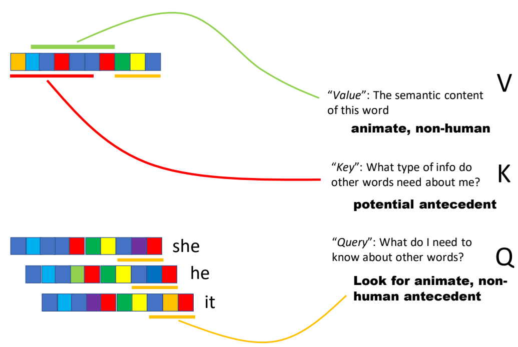


The Transformer

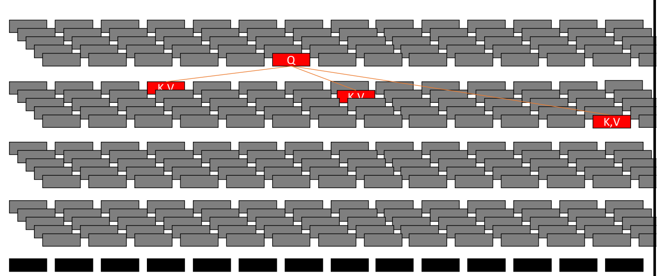
Vaswani et al. (2017), *Attention is all you need. ICLR17*



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



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



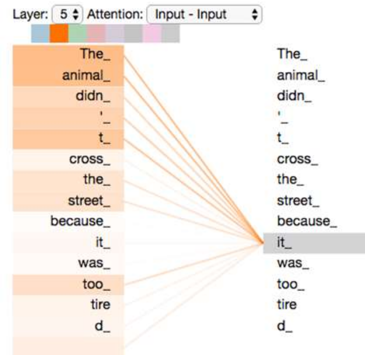
e big brown fox jumped over the fence while the evil hunter in his green

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

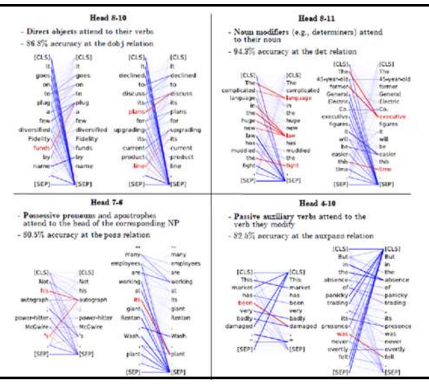
Attention tracking

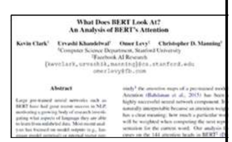


http://jalammar.github.io/illustrated-transformer/

Extremely large language models

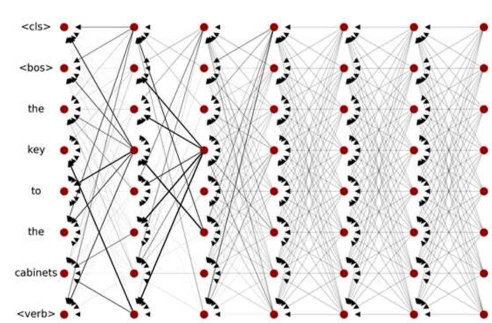
- Bert, GPT3: “Transformer” architecture (Vaswani et al. 2017; 3d most cited paper across academic fields in 2020)
- Extremely large deep learning model (GPT3: 178B parameters)
- Trained on enormous dataset (GPT3: 300M words, extracted from 1B word CommonCrawl + a number of custom datasets)
- Trained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



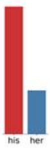


Clark et al. (2019)
BlackboxNLP @ACL2019

Case study 3: Attention Tracking & Attention Flow



Abnar & Zuidema, ACL2020



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

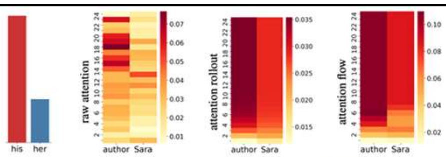
Abnar & Zuidema, ACL2020

Case study 4: Representational Similarity & Stability Analysis

How similar are representations learned by different models, and how similar are they to representations in the brain?

(Abnar, Beinborn, Choenni & Zuidema, 2019)

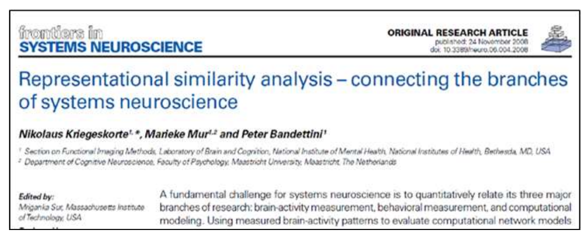
BlackboxNLP @ACL2019



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

- Attention rollout
- Attention flow

Abnar & Zuidema, ACL2020


frontiers in SYSTEMS NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 24 November 2008
doi: 10.3389/neuro.06.004.2008

Representational similarity analysis – connecting the branches of systems neuroscience

Nikolaus Kriegeskorte[1,], Marieke Mur[1,2] and Peter Bandettini[1]*

[1] *Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA*
[2] *Department of Cognitive Neuroscience, Faculty of Psychology, Maastricht University, Maastricht, The Netherlands*

Edited by:
Mriganka Sur, Massachusetts Institute of Technology, USA

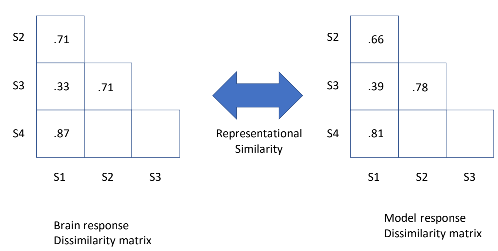
A fundamental challenge for systems neuroscience is to quantitatively relate its three major branches of research: brain-activity measurement, behavioral measurement, and computational modeling. Using measured brain-activity patterns to evaluate computational network models

Lessons case study 3

Positives
- Attention is the central mechanism in Transformer, and often turns out to correspond to linguistic functions
- Attention Roll-out & Flow take into account the structure of the attention network and approximate the effective attention at each point
- Attention Flow corresponds to a well-known attribution method: Shapley values

Negatives
- Attention is a local method; characterizing the globally learned function requires much extra work
- Attention is only one component; may yield false positives, and false negatives

Representational Similarity Analysis

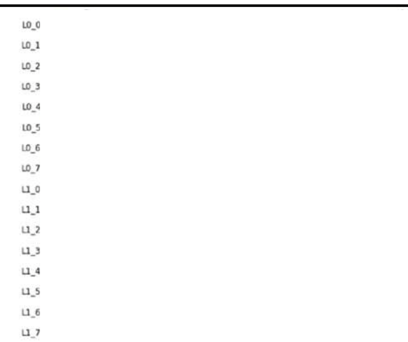


How similar are representations learned by different deep learning models of language, and how similar are they to the brain?

(Abnar, Beinborn, Choenni & Zuidema, 2019)

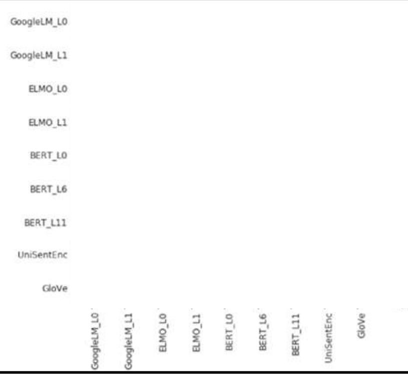
BlackboxNLP @ACL2019

Representational Stability Analysis

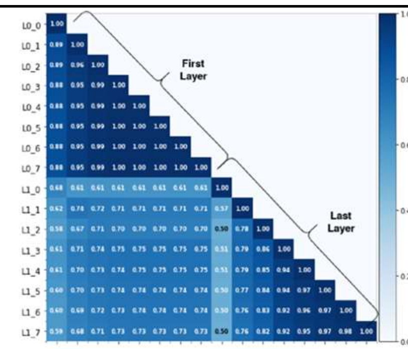


(a) GoogleLM

RSA across models

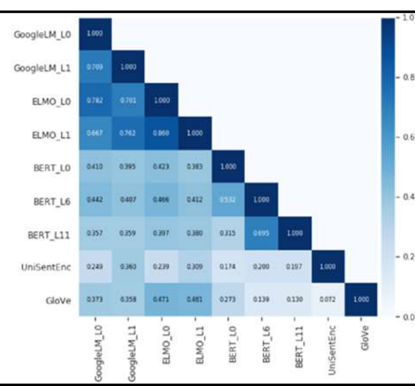


Representational Stability Analysis

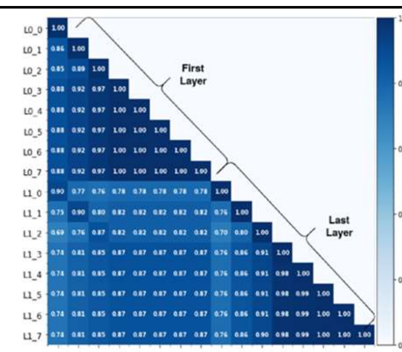


(a) GoogleLM

RSA across models

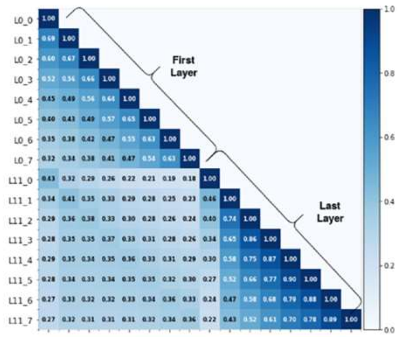


Representational Stability Analysis



(a) ELMO

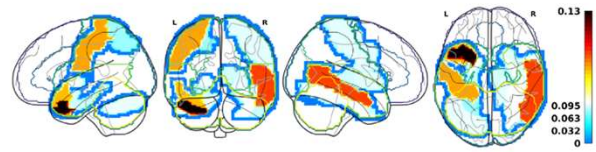
Representational Stability Analysis



(a) BERT

Representational Similarity Analysis

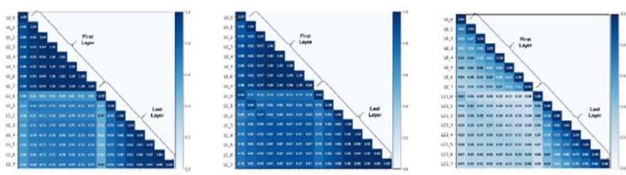
great differences between brain areas in their similarity to the models



http://projects.illc.uva.nl/LaCo/clclab/

Representational Stability Analysis

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



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

Lessons case study 4

Positives

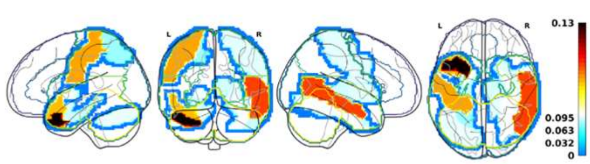
- Representational Similarity Analysis allows us to compare the representational spaces built up by models/brains with radically different architectures
- RSA is a (fairly) global interpretability method, that allows us to characterize the globally learned function
- Representation Stability Analysis allows us to characterize differences in sensitivity to manipulations of different components of the model

Negatives

- If we only compare uninterpretable models with uninterpretable brains, what have we really learned? → need to apply interpretability methods to the models

Representational Similarity Analysis

great differences between brain areas in their similarity to the models



http://projects.illc.uva.nl/LaCo/clclab/

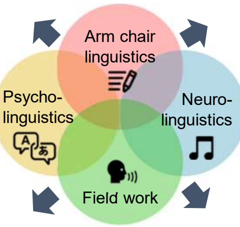
Lessons from all 4 case studies

- All state-of-the-art models in NLP are based on deep learning
- Presents us with the blackbox problem, making it difficult to:
  - Generate *explanations* to users and *justify* decisions based on the systems
  - Allow users to *interact* with the learned solutions and *adapt* them to their needs
  - Use prior knowledge to *augment* machine learned solutions
- Diagnostic classification is a way to test specific hypotheses on what information is represented; should be applied with as much rigor as model testing in (cognitive) neuroscience

Lessons from all 4 case studies (ctd.)

- Representational Similarity Analysis is a way to compare models across paradigms, and test the sensitivity of the learned representations to parameter choices
- Data- & hypothesis-driven methods, local & global methods are complementary, and can often be used together
- There is no silver bullet: the excellent performance of current models is found away from the easily interpretable points in hypothesis space
- We need to systematically apply the ever increasing toolbox of interpretability tools and see how far we get!

The linguistics of deep learning

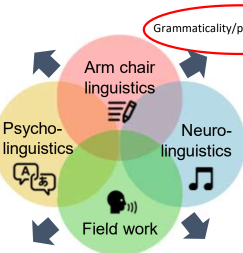


Homework

- Practice with Representation Similarity Analysis using this notebook: https://colab.research.google.com/drive/1LtEwlaqXVEQXCVVTNaTGt0ZIiatkRL3M?usp=sharing
- Practice with Word Embeddings and Dissimilarity Matrices here: https://colab.research.google.com/drive/1JeHTIZ9qi0LRHZurBerwHXOD06nrQb_2?usp=sharing

The linguistics of deep learning

Grammaticality/preference tests



The bigger picture: The virtuous interaction between deep learning, cognitive neuroscience & linguistics

Subject-verb agreement

Subject — Attractor — Verb

Any **bias** in the **articles** almost certainly **relates** / **relate** to …

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

Bias

The author talked to Sara about MASK book

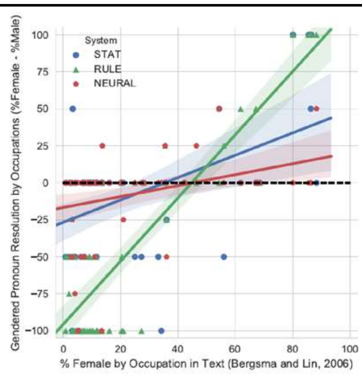
Mary convinced John of MASK love

Abnar & Zuidema, ACL2020

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



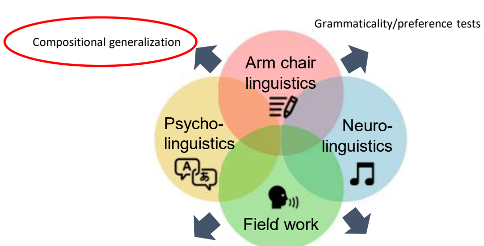
-> neural language model amplify bias already present in the training data

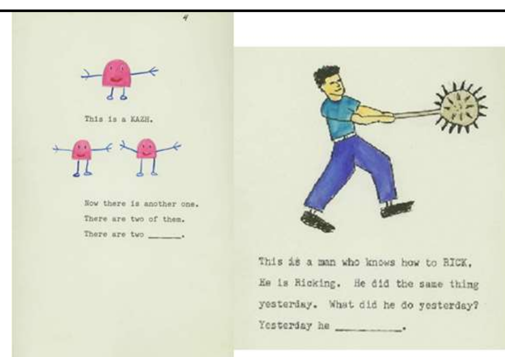
Gender Bias in Coreference Resolution, Rudinger, Naradowsky, Leonard, & Van Durme (2018, NAACL)

Wug test

Jean Berko Gleason (1958)

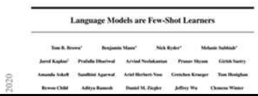
The linguistics of deep learning





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

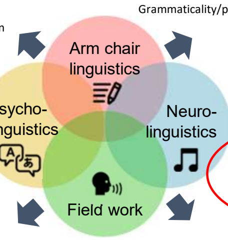
- Zero-shot generalization: generalizing to a new pattern without having seen single example of the target pattern
- One-shot generalization: generalizing to new examples based only on a single example of the target pattern
- Few-shot generalization: generalizing to new examples based only on a handful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


Language Models are Few-Shot Learners

The linguistics of deep learning

Compositional generalization

Grammaticality/preference tests



Interpretability methods
- Attention tracking
- Diagnostic probes
- Attribution methods
- Interpretability by design

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:**

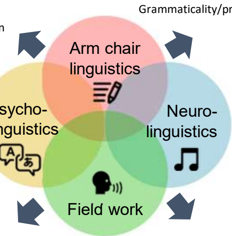
[GPT-3 continuation] One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

The linguistics of deep learning

Compositional generalization

Grammaticality/preference tests

- Classical NLP-tasks: sentiment analysis, translation, QA, …
- Challenge sets/Test batteries
- Adversarial testing



Interpretability methods
- Attention tracking
- Diagnostic probes
- Attribution methods
- Interpretability by design

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

Compositional generalization

`all Germans love all Italians`

implies

`some Germans love some Romans`

Zero-shot generalization to:

`all Germans love all Italians`

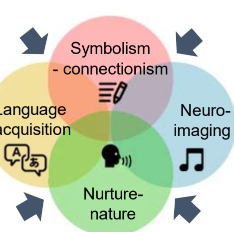
implies

`some Germans love some Venetians`

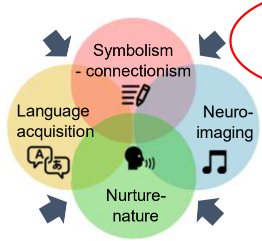
(but not from “all French hate Parisians” to “all French detest Parisians”

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

Deep learning contributing to linguistics



Deep learning contributing to linguistics



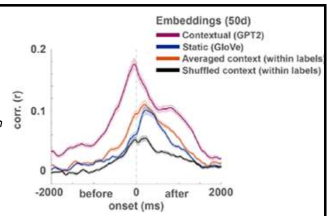
Encoding/decoding models
- fMRI
- EEG
- MEG
- ECoG

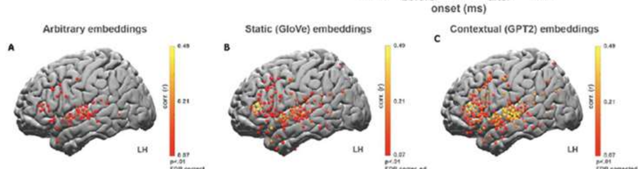
Compositionality revisited

- 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

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.





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.

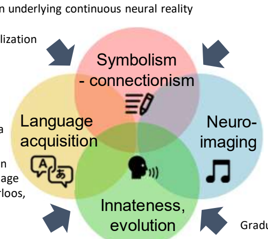
Deep learning contributing to linguistics

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)



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)

- Emergence of structure sensitive units (Lakretz, Kruszewksi, Desbordes, Hupkes, Dehaene & Baroni, 2019)

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

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.

Linguistic relevance

- Aren't neural networks just a lower level of description, relevant for neuroscientists but uninterpretable and irrelevant for linguists?
- Much theory and methodology in linguistics is based on the idea that the symbols, rules, relations, and hierarchical are cognitive and neural primitives.
- We tend to think about neural networks as, at best, approximating the symbolic 'truth'. But what if it is the other way around?
- Symbolic models remain extremely useful for characterizing structure in language.
- But if our symbolic models are really approximations of the underlying neural reality, maybe it not so surprising that 'all grammar leak'.
- Multiple models can coexist as they fulfill different roles!

Suggested Readings

(1) Afra Alishahi, Grzegorz Chrupała, Tal Linzen (2019), Analyzing and Interpreting Neural Networks for NLP: A Report on the First BlackboxNLP Workshop, https://arxiv.org/abs/1904.04063
(2) Yonatan Belinkov, James Glass (2019), Analysis Methods in Neural Language Processing: A Survey, https://arxiv.org/abs/1812.08951
(3) Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, Willem Zuidema (2018), Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information, https://arxiv.org/abs/1808.08079
(4) Samira Abnar, Lisa Beinborn, Rochelle Choenni, Willem Zuidema (2019), Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains, https://arxiv.org/abs/1906.01539
(5) The AllenAI NLP Guide has a chapter on Interpretability, with a useful section on Attribution Methods ("Saliency methods") including Integrated Gradients: https://guide.allennlp.org/interpret

106

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/

Collaborators
& Contact

http://projects.illc.uva.nl/LaCo/clclab/


Lisa Beinborn


Samira Abnar

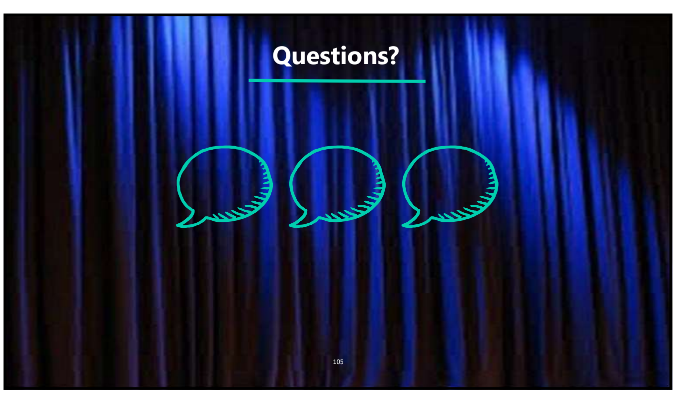

Mario Giulianelli


Jack Harding


Florian Mohnert


Dieuwke Hupkes


Willem Zuidema
zuidema@uva.nl



Extra slides

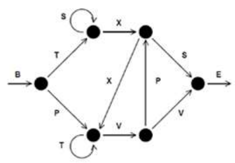


Figure 3: *Transition diagram for the Reber grammar.*

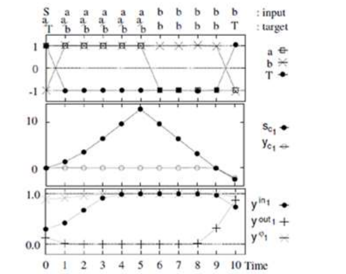


Figure 3: CFL a^nb^n ($n = 5$): Test run with network solutions. Top: Network output y_k. Middle: Cell state s_c and cell output y_c. Bottom: Activations of the gates (input gate y_{in}, forget gate y_φ and output gate y_{out}).

(Hochreiter & Schmidhuber 1997) (Gers & Schmidhuber 2001)

Sources of error

- True function: Y = eval(X),
  - with Y the target,
  - X an expression of the form (Expr Op Expr)
  - Expr ∈ {-9,-8, …, 8, 9} ∪ {(Expr Op Expr)}
  - Op ∈ {+, -}
- Structural error type 1: Not discovering the recursive nature of the problem
  - Because not all lengths are seen at training, OOD generalization will fail catastrophically
- Structural error type 2: errors in learning meaning of brackets
  - Because scope only matters for the minus operation, error will depend on number and depth of minuses, and increase significantly with length of the expression
- Approximation error type 1: errors in learning exact meaning of Expr's
  - Because all operations are linear, total error will grow linearly with length of expression
- Approximation error type 2: errors in learning meaning of Op's
  - Deviations from linearity will yield exponentially increasing error with length of expression

Gradient = partial derivative in a specific point

$$\nabla f(p) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(p) \\ \vdots \\ \frac{\partial f}{\partial x_n}(p) \end{bmatrix}.$$