What do they learn? Neural networks, compositionality and interpretability

Dieuwke Hupkes

Institute for Logic, Language and Computation
University of Amsterdam

Computational Cognition
October 1, 2019
Hierarchical Compositionality
The scientist who wrote the research paper jumped with joy.
Hierarchical Compositionality

The scientist who wrote the research paper jumped with joy
The scientist who wrote the research paper jumped with joy
The scientist who wrote the research paper jumped with joy
The scientist who wrote the research paper jumped with joy
Symbolic structure and the brain

But our brains do not have any explicit means to represent rules and symbols, so how is language represented?

The scientist who wrote the research paper jumped with joy
But our brains do not have any explicit means to represent rules and symbols, so how is language represented?
Recurrent Neural Networks
Simple Recurrent Network

(output)

(input)

(Elman 1990)
Gated recurrent neural networks

(Cho et al. 2014; Chung et al. 2015)
Gated recurrent neural networks
How can hierarchical compositionality be processed \textit{incrementally}, in \textit{linear time}, by a recurrent artificial neural network?
This talk
Two questions

1. Can recurrent neural networks represent hierarchical structure?
Two questions

1. Can recurrent neural networks represent hierarchical structure?
   - In a clean setting, using *artificial languages*
   - In a noisy setting, dealing with *natural language*
Two questions

1. Can recurrent neural networks represent hierarchical structure?
   - In a clean setting, using *artificial languages*
   - In a noisy setting, dealing with *natural language*

2. How do we understand if and how they can?
Two questions

1. Can recurrent neural networks represent hierarchical structure?
   - In a clean setting, using *artificial languages*
   - In a noisy setting, dealing with *natural language*

2. How do we understand if and how they can?
   - Based on their *behaviour*
   - Based on their *representations*
Artificial Language
Arithmetic Language

\[
( ( \text{five minus two} ) \text{ plus six} )
\]

\[
( \text{five minus ( two plus six )} )
\]

(Veldhoen, Hupkes, and Zuidema 2016; Hupkes, Veldhoen, and Zuidema 2018)
Arithmetic Language

(Veldhoen, Hupkes, and Zuidema 2016; Hupkes, Veldhoen, and Zuidema 2018)

\[
( ( \text{five minus two} ) \text{ plus six} )
\]

\[
( \text{five minus } ( \text{two plus six} ) )
\]
Can a gated recurrent network learn this language?
Can a gated recurrent network learn this language?
What does the network do?
Looking inside
Plotting activation values
Looking inside
Update gate

(Karpathy, Johnson, and Fei-Fei 2015)
Symbolic solutions

\((\text{five minus (two plus six)})\)
Symbolic solutions

recursively

\((\text{five minus } (\text{two plus six} ))\)
Symbolic solutions

\[ 5 \]

\[
\text{recursively} \quad (\text{five minus (two plus six)})
\]
Symbolic solutions

\[
\text{recursively } \quad 5 - 5 = (\text{five minus (two plus six)})
\]
Symbolic solutions

(recursively) \[ 5 - 5 \]

\[ ( \text{five minus (two plus six)} ) \]
Symbolic solutions

\[
\text{recursively}
\frac{5}{5} \rightarrow \left( \frac{5}{5} - 2 \right)
\]

\[
(five \ minus \ (two \ plus \ six))
\]
Symbolic solutions

reductively

\[ (\text{five minus} (\text{two plus six})) \]
Symbolic solutions

recursively

( five minus ( two plus six ) )
Symbolic solutions

(recursively) \( 5 - 5 \quad 2 + 2 + 8 \)

( five minus ( two plus six ) )
Symbolic solutions

( five minus ( two plus six ) )

recursively
Recursive solution: \(5 - 5 + 2 + 2 - 3\)

Cumulative solution: \(\left(\text{five minus} \left(\text{two plus six}\right)\right)\)
Symbolic solutions

Recursive: $5 - 5 + 2 + 2 + 8 - 3$

Cumulative: $5$

Expression: $(\text{five minus (two plus six)})$
Symbolic solutions

( five minus ( two plus six ) )

Recursively:

\[ 5 - 5 - 2 + 2 - 8 \]

Cumulatively:

\[ 5 - 5 - \]

Dieuwke Hupkes (ILLC)
Computational Cognition October 1, 2019
Symbolic solutions

\[ ( \text{five minus (two plus six)} ) \]

- recursively
- cumulatively
Symbolic solutions

(recursively) \( 5 - 5 + 2 + 2 - 8 = -3 \)

(cumulatively) \( 5 - 5 - 5 - 3 \)
Symbolic solutions

recursively

\[
( \text{five minus ( two plus six ) } )
\]

cumulatively

\[
5 \quad 5 \quad 5 \quad 3 \quad 3
\]
Symbolic solutions

Recursively: \((5 - (5 - (2 + 2 + 8) - 3))\)

Cumulatively: \(5 - 5 + 5 - 3 + 3 - 3\)

(five minus (two plus six))
Symbolic solutions

\[ (\text{five minus (two plus six)}) \]

- Recursively:
  \[ \text{5} - \text{5} + \text{2} + \text{2} + \text{8} = \text{-3} \]

- Cumulatively:
  \[ \text{5} - \text{5} - \text{3} - \text{3} - \text{3} = \text{-3} \]
Diagnostic Classifier

output

GRU

input
Diagnostic Classifier

\[ \text{GRU} \]

output

input

\[
\begin{array}{c}
( \text{five} \text{ minus} \text{ two} )
\end{array}
\]

\[
\begin{array}{ccccccc}
0 & 5 & 5 & 5 & 3 & - & - & - & - & - & -3
\end{array}
\]

\[
\text{diagnostic classifier}
\]
Intermediate results

mean squared error

Languages

recursive
cumulative
Some intermediate conclusions:

- GRU models seem fairly able to compute the meaning of sequences with hierarchical structure
- With diagnostic classification we can narrow down which strategy they are following
Some other possibilities:

- Further fine-grained analysis of the strategy models are using, and comparison with other recurrent cells (Hupkes, Veldhoen, and Zuidema 2018)
- Understand by masking DC weights whether information is represented in a distributive or local way (Hupkes and Zuidema 2017)
- Locating important neurons (Lakretz et al. 2019)
- Changing the behaviour of models (Giulianelli et al. 2018)
Natural Language
Language Modelling
The **scientist** who wrote the research paper **jumps** with joy
The **scientist** who wrote the research paper **jumps** with joy

The **scientists** who wrote the research paper **jump** with joy
The number agreement task

The **scientist** who wrote the research paper . . .

(Linzen, Dupoux, and Goldberg 2016)
(Gulordava et al. 2018)
Results 2

(Gulordava et al. 2018)
Other linguistic questions

- Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)

- Filler-gap dependencies (Wilcox et al. 2018; Wilcox et al. 2019)

- Reflexive anaphora (Marvin and Linzen 2018; Futrell et al. 2018)

- Garden path sentences (Futrell et al. 2018; Van Schijndel and Linzen 2018; Futrell et al. 2019)

And many more...
Other linguistic questions

- Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)
- Filler-gap dependencies (Wilcox et al. 2018; Wilcox et al. 2019)
Other linguistic questions

- Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)
- Filler-gap dependencies (Wilcox et al. 2018; Wilcox et al. 2019)
- Reflexive anaphora (Marvin and Linzen 2018; Futrell et al. 2018)
Other linguistic questions

- Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)
- Filler-gap dependencies (Wilcox et al. 2018; Wilcox et al. 2019)
- Reflexive anaphora (Marvin and Linzen 2018; Futrell et al. 2018)
- Garden path sentences (Futrell et al. 2018; Van Schijndel and Linzen 2018; Futrell et al. 2019)
Other linguistic questions

- Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)
- Filler-gap dependencies (Wilcox et al. 2018; Wilcox et al. 2019)
- Reflexive anaphora (Marvin and Linzen 2018; Futrell et al. 2018)
- Garden path sentences (Futrell et al. 2018; Van Schijndel and Linzen 2018; Futrell et al. 2019)
- And many more...
Other linguistic questions

- Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)
- Filler-gap dependencies (Wilcox et al. 2018; Wilcox et al. 2019)
- Reflexive anaphora (Marvin and Linzen 2018; Futrell et al. 2018)
- Garden path sentences (Futrell et al. 2018; Van Schijndel and Linzen 2018; Futrell et al. 2019)
- And many more...

But how do they do this?
Diagnostic classification 2
Any bias in the articles almost certainly relates to ... (Giulianelli et al. 2018)
Diagnostic Classification

All sentences, h

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

(Giulianelli et al. 2018)
Diagnostic Classification
All sentences, all components

(Giulianelli et al. 2018)
Temporal generalisation matrix

(Giulianelli et al. 2018)
Other techniques

What else can we do?
### Ablation studies

<table>
<thead>
<tr>
<th>NA task</th>
<th>C</th>
<th>Ablated 776</th>
<th>Ablated 988</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Adv</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>2Adv</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>99.9</td>
</tr>
<tr>
<td>CoAdv</td>
<td>S</td>
<td>- 82</td>
<td>-</td>
<td>98.7</td>
</tr>
<tr>
<td>namePP</td>
<td>SS</td>
<td>-</td>
<td>-</td>
<td>99.3</td>
</tr>
<tr>
<td>nounPP</td>
<td>SS</td>
<td>-</td>
<td>-</td>
<td>99.2</td>
</tr>
<tr>
<td>nounPP</td>
<td>SP</td>
<td>- 54.2</td>
<td>-</td>
<td>87.2</td>
</tr>
<tr>
<td>nounPPAdv</td>
<td>SS</td>
<td>-</td>
<td>-</td>
<td>99.5</td>
</tr>
<tr>
<td>nounPPAdv</td>
<td>SP</td>
<td>- 54.0</td>
<td>-</td>
<td>91.2</td>
</tr>
<tr>
<td>Simple</td>
<td>P</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Adv</td>
<td>P</td>
<td>-</td>
<td>-</td>
<td>99.6</td>
</tr>
<tr>
<td>2Adv</td>
<td>P</td>
<td>-</td>
<td>-</td>
<td>99.3</td>
</tr>
<tr>
<td>CoAdv</td>
<td>P</td>
<td>79.2</td>
<td>-</td>
<td>99.3</td>
</tr>
<tr>
<td>namePP</td>
<td>PS</td>
<td>39.9</td>
<td>-</td>
<td>68.9</td>
</tr>
<tr>
<td>nounPP</td>
<td>PS</td>
<td>48.0</td>
<td>-</td>
<td>92.0</td>
</tr>
<tr>
<td>nounPP</td>
<td>PP</td>
<td>78.3</td>
<td>-</td>
<td>99.0</td>
</tr>
<tr>
<td>nounPPAdv</td>
<td>PS</td>
<td>63.7</td>
<td>-</td>
<td>99.2</td>
</tr>
<tr>
<td>nounPPAdv</td>
<td>PP</td>
<td>-</td>
<td>-</td>
<td>99.8</td>
</tr>
<tr>
<td>Linzen</td>
<td>-</td>
<td>75.3</td>
<td>-</td>
<td>93.9</td>
</tr>
</tbody>
</table>

- A designated *singular* and *plural* unit encode numerosity over long distances.
- For shorter distances, this is encoded in a more distributed fashion.

(Lakretz et al. 2019)
Ablation studies

Lakretz et al. 2019

(a) 988 (singular)

Lakretz et al. 2019
## Contextual Decomposition

(Jumelet, Hupkes, and Zuidema 2019)

<table>
<thead>
<tr>
<th>Decomposed token</th>
<th>doctor</th>
<th>near</th>
<th>the</th>
<th>dogs</th>
<th>knows</th>
<th>know</th>
</tr>
</thead>
<tbody>
<tr>
<td>INIT</td>
<td>0.23</td>
<td>0.14</td>
<td>-0.03</td>
<td>0.56</td>
<td>0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>The</td>
<td>0.43</td>
<td>0.13</td>
<td>-0.01</td>
<td>0.48</td>
<td>0.34</td>
<td>0.07</td>
</tr>
<tr>
<td>doctor</td>
<td>-0.07</td>
<td>0.00</td>
<td>0.59</td>
<td>0.52</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>near</td>
<td>0.16</td>
<td>-0.10</td>
<td>0.05</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>-0.09</td>
<td>0.08</td>
<td>-0.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dogs</td>
<td>0.20</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Jumelet, Hupkes, and Zuidema 2019)
Conclusions
Conclusions

- We can study black box neural networks with behavioural experiments
Conclusions

- We can study black box neural networks with behavioural experiments.
- But we have also quite some techniques available to study their representations.

Neural networks seem quite capable of modelling hierarchical structure, even if the data they deal with is messy.

I'm looking forward to the next step(s): reconnecting all these findings with human language!

Dieuwke Hupkes (ILLC)
Conclusions

- We can study black box neural networks with behavioural experiments
- But we have also quite some techniques available to study their representations
  - Diagnostic Classification
  - Ablation studies
  - Contextual Decomposition
  - Some others I didn’t discuss

Neural networks seem quite capable of modelling hierarchical structure, even if the data they deal with is messy.

I’m looking forward to the next step(s): reconnecting all these findings with human language!
Conclusions

- We can study black box neural networks with behavioural experiments
- But we have also quite some techniques available to study their representations
  - Diagnostic Classification
  - Ablation studies
  - Contextual Decomposition
  - Some others I didn’t discuss
- Neural networks seem quite capable of modelling hierarchical structure, even if the data they deal with is messy
Conclusions

- We can study black box neural networks with behavioural experiments.
- But we have also quite some techniques available to study their representations:
  - Diagnostic Classification
  - Ablation studies
  - Contextual Decomposition
  - Some others I didn’t discuss
- Neural networks seem quite capable of modelling hierarchical structure, even if the data they deal with is messy.
- I’m looking forward to the next step(s): reconnecting all these findings with human language!
Thanks to my collaborators

Willem Zuidema
Germàn Kruszewski
Mario Giulianelli

Marco Baroni
Yair Lakretz
Florian Mohnert

Jaap Jumelet
Sara Veldhoen
Jack Harding


References V


Interventions
Diagnostic interventions
Diagnostic interventions

![Graphs showing accuracy over timesteps for LSTM models with different configurations.]

- LSTM models with different configurations show variations in accuracy over timesteps.
## Diagnostic interventions, results

<table>
<thead>
<tr>
<th>Original Intervention</th>
<th>An official estimate</th>
<th>issued in 2003</th>
<th>suggests</th>
<th>suggest</th>
</tr>
</thead>
<tbody>
<tr>
<td>-11.05</td>
<td>-8.426</td>
<td>-8.472</td>
<td>-1.243</td>
<td>-3.951</td>
</tr>
<tr>
<td>DC</td>
<td>78.0</td>
<td></td>
<td></td>
<td>-5.753</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-5.691</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-6.4361</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intervention</th>
<th>without intervention</th>
<th>with intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85.4</td>
<td></td>
</tr>
</tbody>
</table>
The keys to the kabinet left of the door (are/is) on the table.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Accuracy with intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>78.1</td>
<td>85.4</td>
</tr>
<tr>
<td>Nonce</td>
<td>70.7</td>
<td>75.6</td>
</tr>
</tbody>
</table>

(Giulianelli et al. 2018)
Gated Recurrent Neural Networks
$h_t = \tanh(Wx_t + Uh_{t-1} + b)$

(Elman 1990)
Gated recurrent neural networks

\[ h_t = \tanh(Wx_t + Uh_{t-1} + b) \]
\[ r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \]

(Cho et al. 2014; Chung et al. 2015)
Gated recurrent neural networks

\[ \tilde{h}_t = \tanh(Wx_t + U(r \odot h_{t-1}) + b) \]

\[ r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \]

(Cho et al. 2014; Chung et al. 2015)
Gated recurrent neural networks

\[ \tilde{h}_t = \tanh(Wx_t + U(r \odot h_{t-1}) + b) \]

\[ r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \]

\[ z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \]

(Cho et al. 2014; Chung et al. 2015)
Gated recurrent neural networks

\[ \tilde{h}_t = \tanh(Wx_t + U(r \odot h_{t-1}) + b) \]

\[ r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \]

\[ z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \]

\[ h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \]

(Cho et al. 2014; Chung et al. 2015)