Visualisation and ‘Diagnostic Classifiers’ Reveal how Recurrent and Recursive Neural Networks Process Hierarchical Structure

Dieuwke Hupkes

Institute for Logic, Language and Computation
University of Amsterdam

July 18, 2018
Compositional solutions in Recurrent Neural Networks

Recurrent neural networks are not good at finding systematic/compositional solutions to problems, like humans.
Compositional solutions in Recurrent Neural Networks

Recurrent neural networks are not good at finding systematic/compositional solutions to problems, like humans

- Compositionality is difficult to (directly) evaluate
Compositional solutions in Recurrent Neural Networks

Recurrent neural networks are not good at finding systematic/compositional solutions to problems, like humans

- Compositionality is difficult to (directly) evaluate
- Neural networks are black boxes
<table>
<thead>
<tr>
<th>Name</th>
<th>#digits</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>1</td>
<td>minus three</td>
</tr>
<tr>
<td>L2</td>
<td>2</td>
<td>( five plus seven )</td>
</tr>
<tr>
<td>L3</td>
<td>3</td>
<td>( three - ( one + minus two ) )</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L5R</td>
<td>5</td>
<td>( ( ( nine + six) + seven ) + five ) - seven</td>
</tr>
<tr>
<td>L5L</td>
<td>5</td>
<td>( eight + ( six - ( two - ( ten + nine ) ) ) )</td>
</tr>
</tbody>
</table>
Arithmetic Language
Deep Hierarchical Structure

((five minus two) plus six)

(five minus (two plus six))
( five minus ( two plus six ) )
Arithmetic Language
Symbolic Solutions

recursively

( five minus ( two plus six ) )
Arithmetic Language
Symbolic Solutions

recursively

5

(five minus (two plus six))
Arithmetic Language
Symbolic Solutions

recursively

\( 5 - \frac{5}{\left( \text{five minus } (\text{two plus six}) \right)} \)
Recursive 5 - 5

(five minus (two plus six))
Arithmetic Language
Symbolic Solutions

recursively

\[
\begin{align*}
&5, - \\
\rightarrow & - \\
&5, - 2 \\
\end{align*}
\]

(five minus (two plus six))
(five minus (two plus six))
Arithmetic Language
Symbolic Solutions

Recursively

\[ (5 - (\text{two plus six}) + 2) \]
Arithmetic Language
Symbolic Solutions

Recursively

\( ( \text{five minus} \ ( \text{two plus six} ) ) \)
Arithmetic Language
Symbolic Solutions

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

\( \text{(five minus (two plus six)}) \)
Arithmetic Language
Symbolic Solutions

((five minus (two plus six)))

recursively

5
-5
-

2
+

2

8

-3

cumulatively
Arithmetic Language
Symbolic Solutions

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

recursively \[ 5 \quad 5 \quad 2 \quad 2 \quad 8 \quad -3 \]
cummulatively \[ 5 \]
Arithmetic Language

Symbolic Solutions

(five minus (two plus six))
Arithmetic Language
Symbolic Solutions

\[
(\text{five minus (two plus six)})
\]
Arithmetic Language
Symbolic Solutions

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

Recursively:

\[
5 - 5 - 2 + 2 + 8 - 3
\]

Cumulatively:

\[
5 - 5 - 5 - 3
\]
Arithmetic Language
Symbolic Solutions

( five minus ( two plus six ) )

five
minus
five
- 5
two
+ 2
plus
six
8
recursively
5
5
- 5,-
- 5
- 5
5
+ 2
2
- 3
- 8
- 3
cumulatively
5
5
5
- 3
- 3
-
Arithmetic Language
Symbolic Solutions

( five minus ( two plus six ) )

recursively

cumulatively
Arithmetic Language
Symbolic Solutions

(five minus (two plus six))
How do we study the network?
Diagnostic Classification

GRU

output

input

( five minus ( two

- - - - - ) )
Diagnostic Classification

GRU

Output

Input

(five minus (two
diagnostic classifier

0 5 5 5 3 - - - - - - -3
Recursive or cumulative?

![Graph showing mean squared error for languages categorized as recursive (red triangles) or cumulative (blue squares).](image-url)
• How do you know diagnostic classifiers don’t just pick up noise?  
• (or: shouldn’t you use more complicated diagnostic models?)  
• What do you do when you don’t have a symbolic hypothesis?  
• How does this knowledge help us?
Subject-verb agreement in Language Models

The keys to the kabinet left of the door (are/is) on the table.

Linzen et al., (2016); Gulordava et al., (2018)
Subject-verb agreement in Language Models

The keys to the kabinet left of the door (are/is) on the table.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>78.1</td>
</tr>
<tr>
<td>Nonce</td>
<td>70.7</td>
</tr>
</tbody>
</table>

Hupkes et al (2018), in prep
Subject-verb agreement in Language Models

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>

Hupkes et al (2018), in prep
Thank you

Dieuwke Hupkes (d.hupkes@uva.nl)

My collaborators:
Dr. Willem Zuidema
Jack Harding
Florian Mohnert
Mario Giulianelli
Results

- GRU average
- GRU best
- LSTM average
- LSTM best
- SRN best
Hypotheses

\[
( -2 - ( 6 - ( ( 8 + ( -3 - 10 ) ) - ( -2 - 10 ) ) ) ) - ( 1 - 8 )
\]

| minus_scope3+ | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |
| minus_scope2+ | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |
| minus_scope1+ | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |
| close_minus_scope1+ | 0 0 0 0 1 1 1 2 3 3 4 4 4 4 3 2 2 3 3 3 2 1 0 0 0 1 1 1 1 0 0 |

| mode | + + + - - - + + + + + + - - + - - - + - - + - - - + + - + |
| switch_mode | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |
Hypotheses

\[
0 0 0 0 1 1 1 2 3 3 3 4 4 4 4 3 2 2 3 3 3 3 2 1 0 0 0 1 1 1 1 0 0
\]

\[
( -2 - ( 6 - ( ( 8 + ( -3 - 10 ) ) - ( -2 - 10 ) ) ) ) - ( 1 - -8 ) )
\]

**mode**

- \(+\)
- \(-\)

**switch_mode**

- \(+\)
- \(-\)

**operator**

- \(+\)
- \(-\)

**LSTM best**

- \(+\)
- \(-\)

**LSTM average**

- \(+\)
- \(-\)

**GRU best**

- \(+\)
- \(-\)

**GRU average**

- \(+\)
- \(-\)
Using diagnostic classifier weights

What happens where?

**Majority classifier**

**Minority classifier**

left: update gate $z$
right: reset gate $r$

Prediction of $\text{minus\_scope1+}$ by individual hidden layer units

Prediction of $\text{minus\_scope1+}$ by individual gate units