The compositionality of neural networks: integrating symbolism and connectionism

Dieuwke Hupkes

Institute for Logic, Language and Computation
University of Amsterdam

May 6, 2019
The appropriateness of neural models

- “Modern approaches [...] do not explicitly formulate and execute compositional paths” (Johnson et al., 2017)
The appropriateness of neural models

- “Modern approaches [...] do not explicitly formulate and execute compositional paths” (Johnson et al., 2017)
- “Neural network models lack the ability to extract systematic rules” (Lake and Baroni, 2018)
The appropriateness of neural models

- “Modern approaches [...] do not explicitly formulate and execute compositional paths” (Johnson et al., 2017)
- “Neural network models lack the ability to extract systematic rules” (Lake and Baroni, 2018)
- “They do not learn in a compositional way” (Liška et al., 2018)
The appropriateness of neural models

▶ “Modern approaches [...] do not explicitly formulate and execute compositional paths” (Johnson et al., 2017)
▶ “Neural network models lack the ability to extract systematic rules” (Lake and Baroni, 2018)
▶ “They do not learn in a compositional way” (Liška et al., 2018)
▶ “[...] neural networks are essentially very large correlation engines that hone in on any statistical, potentially spurious pattern” (Hudson and Manning, 2018)
The appropriateness of neural models

- “Modern approaches [...] do not explicitly formulate and execute compositional paths” (Johnson et al., 2017)
- “Neural network models lack the ability to extract systematic rules” (Lake and Baroni, 2018)
- “They do not learn in a compositional way” (Liška et al., 2018)
- “[...] neural networks are essentially very large correlation engines that hone in on any statistical, potentially spurious pattern” (Hudson and Manning, 2018)
- Neural networks are data-hungry because they don’t develop re-usable representations (almost everyone)
What is compositionality

The principle of compositionality

_The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined._

Partee (1995)
What is compositionality

What does it mean that neural networks are not compositional?

- They find different parts than we’d like them to
- They find different rules than we’d like them to
- They find other aspects of the data more salient
- They cannot represent hierarchy
- They favour memorising sequences over learning rules
- They are not getting the right signal from the data
- . . .
The appropriateness of neural models

Our approach: “dissect” compositionality:
  ▶ Do models find the right parts and rules?
The appropriateness of neural models

Our approach: “dissect” compositionality:

- Do models find the right parts and rules?
- Do models use the parts and rules they find systematically?
The appropriateness of neural models

Our approach: “dissect” compositionality:
- Do models find the right parts and rules?
- Do models use the parts and rules they find systematically
- Do models use the parts and rules they find productively
The appropriateness of neural models

Our approach: “dissect” compositionality:

- Do models find the right parts and rules?
- Do models use the parts and rules they find systematically
- Do models use the parts and rules they find productively
- Do models compute locally consistent representations?
The appropriateness of neural models

Our approach: “dissect” compositionality:

▶ Do models find the right parts and rules?
▶ Do models use the parts and rules they find systematically
▶ Do models use the parts and rules they find productively
▶ Do models compute locally consistent representations?
▶ Do models allow substitution of synonyms?
The appropriateness of neural models

Our approach: “dissect” compositionality:

➤ Do models find the right parts and rules?
➤ Do models use the parts and rules they find systematically
➤ Do models use the parts and rules they find productively
➤ Do models compute locally consistent representations?
➤ Do models allow substitution of synonyms?
➤ Do models prefer rules or exceptions?
The rest of the team

Mathijs Mul

Verna Dankers

Elia Bruni
Data

PCFG SET

**Unary functions**: reverse, swap, copy, ...

**Binary functions**: prepend, append, remove_first, ...

**Characters**: A, B, C, ...
Data

PCFG SET

Unary functions: reverse, swap, copy, ...
Binary functions: prepend, append, remove_first, ...
Characters: A, B, C, ...

reverse A B C
Unary functions: reverse, swap, copy, ...
Binary functions: prepend, append, remove_first, ...
Characters: A, B, C, ...

reverse A B C ⇒ C B A
Unary functions: reverse, swap, copy, ...
Binary functions: prepend, append, remove_first, ...
Characters: A, B, C, ...

reverse A B C \[\Rightarrow\] C B A
append C B A , D E
**Unary functions:** reverse, swap, copy, ...
**Binary functions:** prepend, append, remove_first, ...
**Characters:** A, B, C, ...

reverse A B C  \[\Rightarrow\]  C B A
append C B A , D E  \[\Rightarrow\]  C B A D E
 Unary functions: reverse, swap, copy, ...
 Binary functions: prepend, append, remove_first, ...
 Characters: A, B, C, ...

reverse A B C ⇒ C B A
append C B A , D E ⇒ C B A D E
append reverse A B C , copy D E ⇒ C B A D E
Unary functions: reverse, swap, copy, ...
Binary functions: prepend, append, remove_first, ...
Characters: A, B, C, ...

append reverse A B C , copy D E ⇒ C B A D E
Figure: Distribution of sentence depth and length in the PCFG SET and WMT2017 data.
Models

1. **LSTMS2S** Recurrent encoder-decoder model with attention
2. **ConvS2S** Convolutional encoder and decoder with multistep attention
3. **Transformer** Fully attention based model
## Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>LSTMS2S</th>
<th>ConvS2S</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCFG SET*</td>
<td>0.77 ± 0.01</td>
<td>0.84 ± 0.01</td>
<td>0.93 ± 0.01</td>
</tr>
</tbody>
</table>
Can models systematically recombine unseen pairs of functions?
## Results

### Systematicity

<table>
<thead>
<tr>
<th>Experiment</th>
<th>LSTMS2S</th>
<th>ConvS2S</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCFG SET*</td>
<td>0.77 ± 0.01</td>
<td>0.84 ± 0.01</td>
<td>0.93 ± 0.01</td>
</tr>
<tr>
<td>Systematicity*</td>
<td>0.51 ± 0.03</td>
<td>0.55 ± 0.01</td>
<td>0.70 ± 0.01</td>
</tr>
</tbody>
</table>
Localism

Do models build representations incrementally?

\[
\text{append reverse } A B C , \text{ copy } D E \equiv \text{append } C B A , D E
\]
## Results

### Localism

<table>
<thead>
<tr>
<th>Experiment</th>
<th>LSTMS2S</th>
<th>ConvS2S</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCFG SET*</td>
<td>0.77 ± 0.01</td>
<td>0.84 ± 0.01</td>
<td>0.93 ± 0.01</td>
</tr>
<tr>
<td>Systematicity*</td>
<td>0.51 ± 0.03</td>
<td>0.55 ± 0.01</td>
<td>0.70 ± 0.01</td>
</tr>
<tr>
<td>Localism†</td>
<td>0.45 ± 0.01</td>
<td>0.57 ± 0.04</td>
<td>0.56 ± 0.03</td>
</tr>
</tbody>
</table>
Results

Generality of representations

(a) LSTM2S  (b) Conv2S  (c) Transformer
Overgeneralisation

Do models overgeneralise during training?
## Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>LSTMS2S</th>
<th>ConvS2S</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCFG SET*</td>
<td>0.77 ± 0.01</td>
<td>0.84 ± 0.01</td>
<td>0.93 ± 0.01</td>
</tr>
<tr>
<td>Systematicity*</td>
<td>0.51 ± 0.03</td>
<td>0.55 ± 0.01</td>
<td>0.70 ± 0.01</td>
</tr>
<tr>
<td>Localism†</td>
<td>0.45 ± 0.01</td>
<td>0.57 ± 0.04</td>
<td>0.56 ± 0.03</td>
</tr>
<tr>
<td>Overgeneralisation*</td>
<td>0.73 ± 0.18</td>
<td>0.78 ± 0.12</td>
<td>0.84 ± 0.02</td>
</tr>
</tbody>
</table>
Overgeneralisation profile

LSTM2S

ConvS2S

Transformer
Overgeneralisation

Different exception rates

Overgeneralisation profiles for exceptions occurring 0.01%, 0.05%, 0.1% and 0.5%

(a) LSTM2S  (b) Conv2S  (c) Transformer
The rest of the team

Mathijs Mul  Verna Dankers  Elia Bruni
References


