Learning compositionally through attentive guidance

Dieuwke Hupkes

Institute for Logic, Language and Computation
University of Amsterdam

June 12, 2018
Structures in language

Dieuwke Hupkes

Learning compositionally through attentive guidance
Neural networks

Learning compositionally through attentive guidance
The successes of neural networks

They work very well:

- Machine translation
- Syntactic parsing
- Semantic role labelling
- Language modelling
The downside of neural networks

They are not useful as explanatory models of language
We don't know how they relate to linguistic theories of language
We don't know how to improve them (other than by applying engineering tricks)
Actually, we don't even have any idea what they encode

Dieuwke Hupkes

Learning compositionally through attentive guidance
The downside of neural networks

- They are not useful as explanatory models of language
- We don’t know how they relate to linguistic theories of language
- We don’t know how to improve them (other than by applying engineering tricks)
The downside of neural networks

- They are not useful as explanatory models of language
- We don’t know how they relate to linguistic theories of language
- We don’t know how to improve them (other than by applying engineering tricks)
- Actually, we don’t even have any idea what they encode
What do we do?

- We wait for the engineers to solve it
- We try to increase our understanding of what these networks are encoding
- We try to find new ways to make them behave more human-like
Gulordava et al. (2018)
Learning compositionally through attentive guidance

Hupkes et al. (2018b)
On a behaviour level

<table>
<thead>
<tr>
<th>Action</th>
<th>Corresponding Action Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>jump</td>
<td>JUMP</td>
</tr>
<tr>
<td>jump left</td>
<td>LTURN JUMP</td>
</tr>
<tr>
<td>jump around right</td>
<td>RTURN JUMP RTURN JUMP RTURN JUMP</td>
</tr>
<tr>
<td>turn left twice</td>
<td>LTURN LTURN</td>
</tr>
<tr>
<td>jump thrice</td>
<td>JUMP JUMP</td>
</tr>
<tr>
<td>jump opposite left and walk thrice</td>
<td>LTURN LTURN JUMP WALK WALK WALK</td>
</tr>
<tr>
<td>jump opposite left after walk around left</td>
<td>LTURN WALK LTURN WALK LTURN WALK LTURN LTURN JUMP</td>
</tr>
</tbody>
</table>

Figure 1: Examples of SCAN commands (left) and the corresponding action sequences (right).

Lake and Baroni (2017)
Behavioually

Random train/test split results

Length split results

Accuracy on new commands (%)

Percent of commands used for training

Accuracy on new commands (%)

Ground-truth action sequence length
Thus?

- Networks can pick up on interesting (hierarchical) patterns
- We have some methods to look inside networks
- Networks are powerfull generalisation machines
- But: they don’t do this in a human understandable way
Thus?

Pattern matching goes a long way
Neural networks and natural language processing

Understanding neural networks
Attentive Guidance

Discussion
References

Lookup tables

\[
\begin{array}{cccc}
  t_1 & t_2 & \ldots & t_N \\
  00 \mapsto 10 & 00 \mapsto 00 & \ldots & \ldots \\
  01 \mapsto 11 & 01 \mapsto 10 & \ldots & \ldots \\
  10 \mapsto 00 & 10 \mapsto 01 & \ldots & \ldots \\
  11 \mapsto 01 & 11 \mapsto 11 & \ldots & \ldots \\
\end{array}
\]

\[
\text{Liška et al. (2018)}
\]

Dieuwke Hupkes
Learning compositionally through attentive guidance
**Experimental setup**

- **Data:** 8 randomly generated 3-bit atomic tasks and corresponding 64 composed tasks.
- **Training** on all atomic tasks and 6 out of 8 inputs of composed tasks, and test on 2 held-out inputs (totaling 128 test compositions).
How do neural networks do?

![Bar chart showing the proportion of trained models with correct zero-shot responses. The chart indicates a peak at random baseline mode with a tail extending towards the monolith.]

**Proportion of trained models**

**Correct zero-shot responses (%)**

- The chart visually represents the performance of neural networks in zero-shot learning scenarios, highlighting the effectiveness of attentive guidance in achieving higher accuracy compared to random baseline models.
Conclusion

1. Some RNNs find a generalising solution
2. Most networks do not exhibit systematic compositionality
Attentive Guidance

Hupkes et al. (2018a)
Hupkes et al. (2018a)

An important part of the training will consist in the teacher’s pointing to the objects, directing the child’s attention to them, and at the same time uttering a word; for instance, the word ”slab” as he points to that shape.

*Philosophical Investigations*

*L. Wittgenstein*
Supervise the attention mask of the network to match a compositional readout of the input.
Neural networks and natural language processing

Understanding neural networks

Attentive Guidance

Discussion

References

Intuition and implementation

Implementation

Lookup tables

Symbol Rewriting

Dieuwke Hupkes

Learning compositionally through attentive guidance

Encoder

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.2 & 0.5 & 0.4 \\
0.3 & 0.4 & 0.3 \\
0.1 & 0.2 & 0.7
\end{bmatrix}
\]

\[
D(a, \hat{a}) = \sum_i a_i \log \hat{a}_i
\]

\[
\hat{a}_{i,t} = \frac{MLP(e_i, d_t)}{\sum_j \hat{a}_{j,t}}
\]

Decoder

\[
c_t = \sum_i \hat{a}_{i,t} e_i
\]


Data

- **Training** 6 out of 8 inputs in 28 compositions unseen: $t_1$ $t_2$ 110
- **Heldout inputs** 2 out of 8 inputs in 28 compositions unseen: e.g. $t_1$ $t_2$ 010
- **Heldout compositions** 8 entirely unseen compositions: $t_1$ $t_3$
- **Heldout tables** compositions with one of the two heldout tables: e.g. $t_7$ $t_1$ 000
- **New compositions** compositions between the two heldout tables: e.g. $t_7$ $t_8$ 000
Accuracies

![Bar chart showing accuracies for different tasks.

- **heldout inputs**
- **heldout compositions**
- **heldout tables**
- **new compositions**

Legend:
- Green: Guided
- Pink: Vanilla

The bar chart illustrates the comparison between Guided and Vanilla models across different tasks, with error bars indicating variability.
Overfitting
\[ \mathcal{L}: X = \{A, B\}, \]
\[ Y_A = \{a_1, a_2, a_3\}, \quad Y_B = \{b_1, b_2, b_3\}. \]
\[ a_1 \rightarrow a_{11}|a_{12}, \quad a_2 \rightarrow a_{21}|a_{22}, \quad a_3 \rightarrow a_{31}|a_{32} \]
\[ b_1 \rightarrow b_{11}|b_{12}, \quad b_2 \rightarrow b_{21}|b_{22}, \quad b_3 \rightarrow b_{31}|b_{32} \]

**Input**  
**Valid output for** \( \mathcal{L} \)

\[ AAB \quad a_{21}a_{32}a_{12}a_{11}a_{22}a_{32}b_{13}b_{21}b_{32} \]

Results

![Bar chart showing sequence accuracy for different conditions: Guided, Baseline (64x64), Baseline (32x256). Conditions include Standard, Repeat, Short, and Long. The y-axis represents sequence accuracy, and the x-axis represents conditions. The chart illustrates the performance comparison between the guided and baseline conditions.]
What’s next?

- Relaxing the need of guidance
What’s next?

- Relaxing the need of guidance

Picture of Mathijs Mul
What’s next?

- Relaxing the need of guidance
- Designing architectures that have compositional biases built in
What’s next?

- Relaxing the need of guidance
- Designing architectures that have compositional biases built in
- Finding other tasks: What would be a convincing proof of compositionality?
Acknowledgments

Anand Kumar Singh  Kris Korrel
Elia Bruni  Germàn Kruszewski


