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

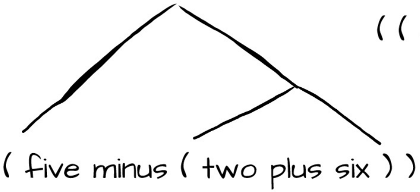
- Compositionality is difficult to (directly) evaluate
- Neural networks are black boxes

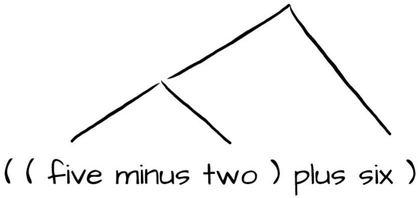
Arithmetic Language

Name	*#digits*	*Example*
L1	1	`minus three`
L2	2	`( five plus seven )`
L3	3	`( three - ( one + minus two ) )`
...		
L5R	5	`( ( ( ( nine + six) + seven ) + five ) - seven )`
L5L	5	`( eight + ( six- ( two - ( ten + nine ) ) ) )`

Arithmetic Language

Deep Hierarchical Structure





(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 $\overset{-}{5}$

(five minus (two plus six))

Arithmetic Language

Symbolic Solutions

5, -

-

recursively 5 5

(five minus (two plus six))

Arithmetic Language

Symbolic Solutions

5, -

-

recursively 5 5 2

(five minus (two plus six))

Arithmetic Language

Symbolic Solutions

5, -

- +

recursively 5 5 2 2

(five minus (two plus six))

Arithmetic Language

Symbolic Solutions

5, -

- +

recursively 5 5 2 2 8

(five minus (two plus six))

Arithmetic Language

Symbolic Solutions

5, -

- +

recursively 5 5 2 2 8

(five minus (two plus six))

Arithmetic Language

Symbolic Solutions

5, -

- +

recursively 5 5 2 2 8 -3

(five minus (two plus six))

Arithmetic Language

Symbolic Solutions

5, -

\-   +

recursively   5   5   2   2   8   -3

(five minus (two plus six))

cummulatively

Arithmetic Language

Symbolic Solutions

5, -

\- +

recursively 5 5 2 2 8 -3

(five minus (two plus six))

cummulatively 5

Arithmetic Language

Symbolic Solutions

5, -

\- +

recursively 5 5 2 2 8 -3

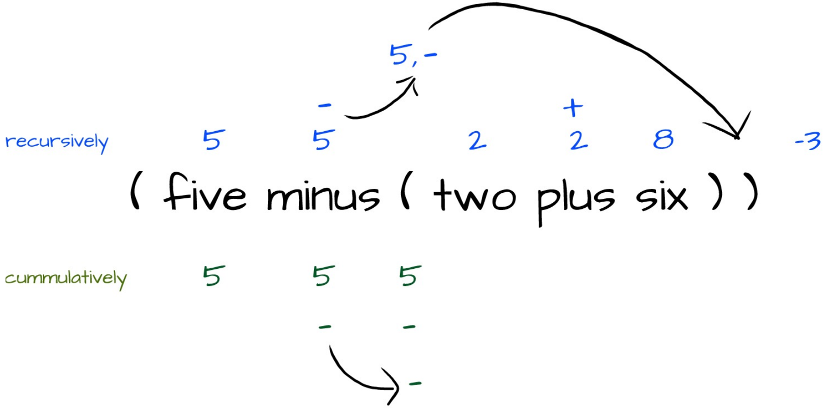
(five minus (two plus six))

cummulatively 5 5

\-

Arithmetic Language

Symbolic Solutions



Arithmetic Language

Symbolic Solutions

5, -

- +

recursively 5 5 2 2 8 -3

(five minus (two plus six))

cummulatively 5 5 5 3

- - -

-

Arithmetic Language

Symbolic Solutions

5, -

- +

recursively 5 5 2 2 8 -3

(five minus (two plus six))

cummulatively 5 5 5 3 3

- - - -

-

Arithmetic Language

Symbolic Solutions

5, -

- +

recursively 5 5 2 2 8 -3

(five minus (two plus six))

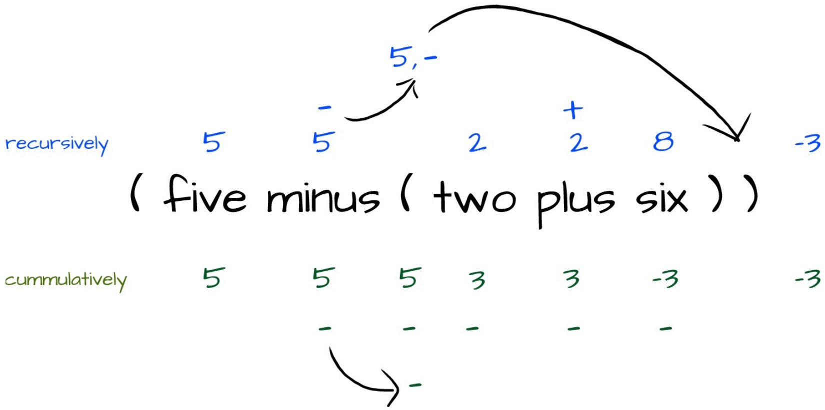
cummulatively 5 5 5 3 3 -3

- - - - -

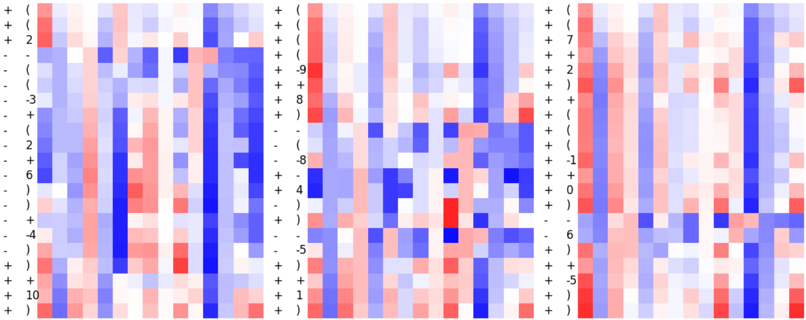
-

Arithmetic Language

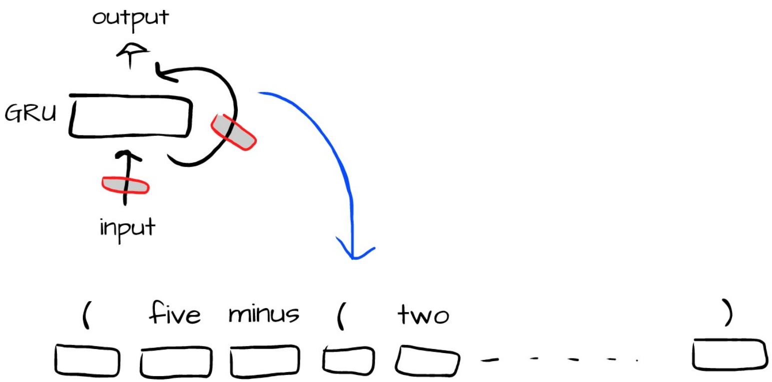
Symbolic Solutions



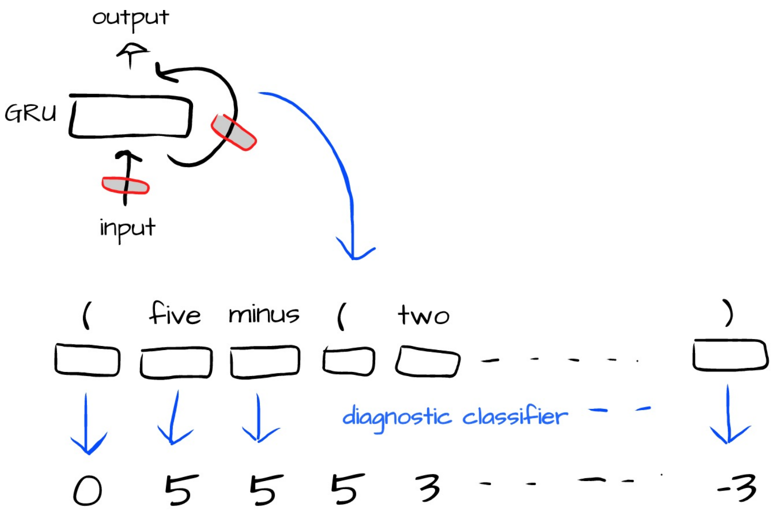
How do we study the network?



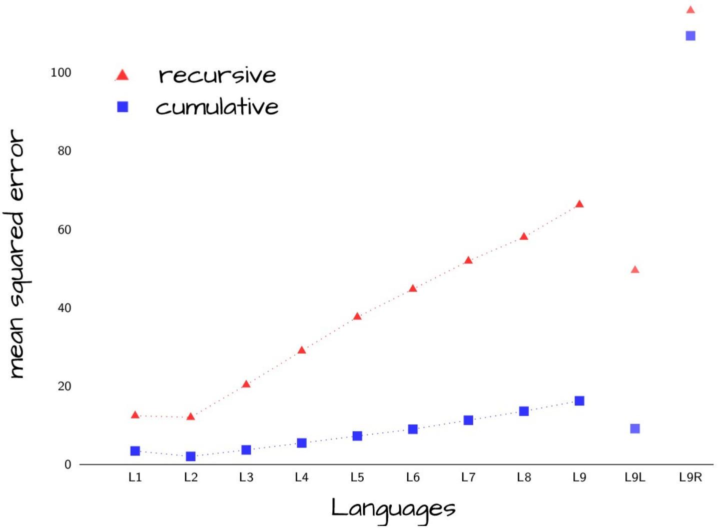
Diagnostic Classification



Diagnostic Classification



Recursive or cumulative?



Critical notes

- 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.

	Accuracy
Original	78.1
Nonce	70.7

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.

	Accuracy	**Accuracy with intervention**
Original	78.1	85.4
Nonce	70.7	75.6

Hupkes et al (2018), in prep

Thank you

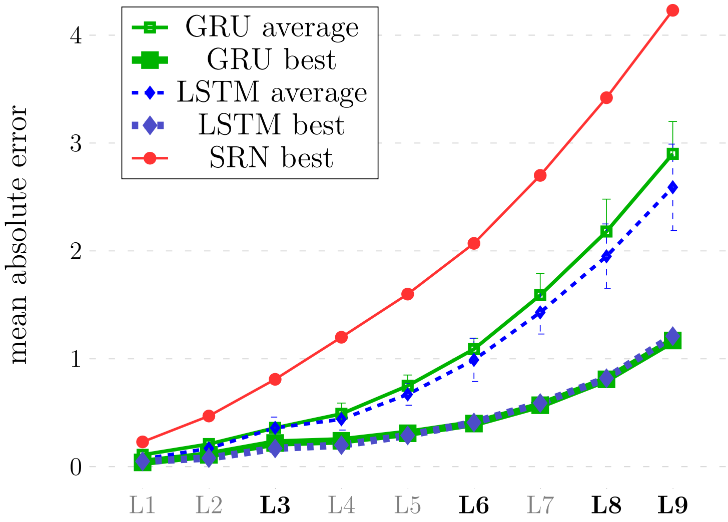
Dieuwke Hupkes (d.hupkes@uva.nl)

My collaborators:

Dr. Willem Zuidema

Jack Harding
Florian Mohnert
Mario Giulianelli

Results

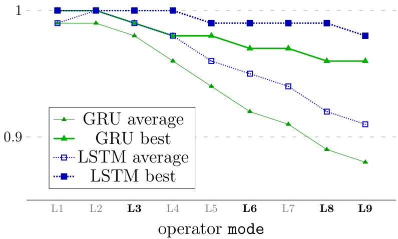

GRU average
GRU best
LSTM average
LSTM best
SRN best
mean absolute error
0
1
2
3
4
L1
L2
L3
L4
L5
L6
L7
L8
L9

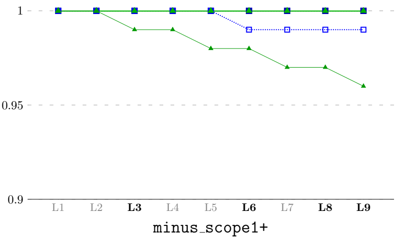
Hypotheses

minus_scope3+														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						1	1		
minus_scope1+					1	1	1	1	1	1	1	1	1	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	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				1			1							1		1		1			1		1	1	1	1	1			1		1	1

Hypotheses

minus_scope3+														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						1	1		
minus_scope1+					1	1	1	1	1	1	1	1	1	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	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				1			1							1		1		1			1		1	1	1	1	1			1		1	1





Using diagnostic classifier weights

What happens where?