What do they learn? Neural networks, compositionality and interpretability

Dieuwke Hupkes

Institute for Logic, Language and Computation
University of Amsterdam

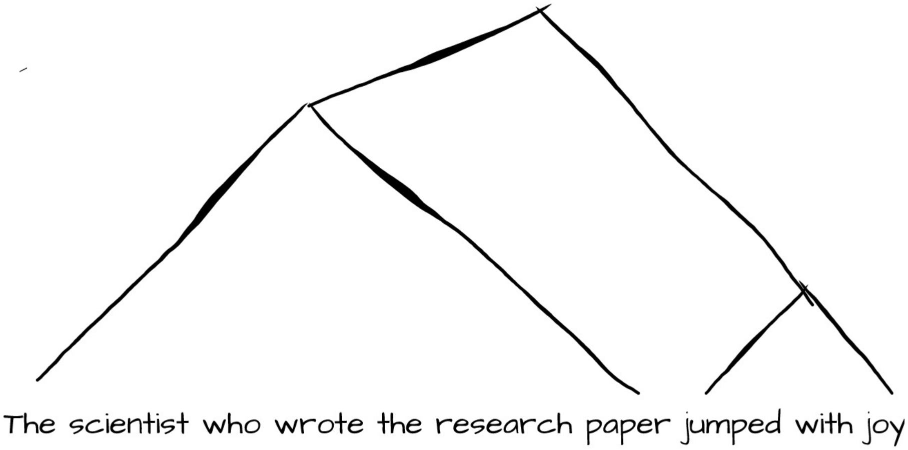
Computational Cognition
October 1, 2019

Hierarchical Compositionality

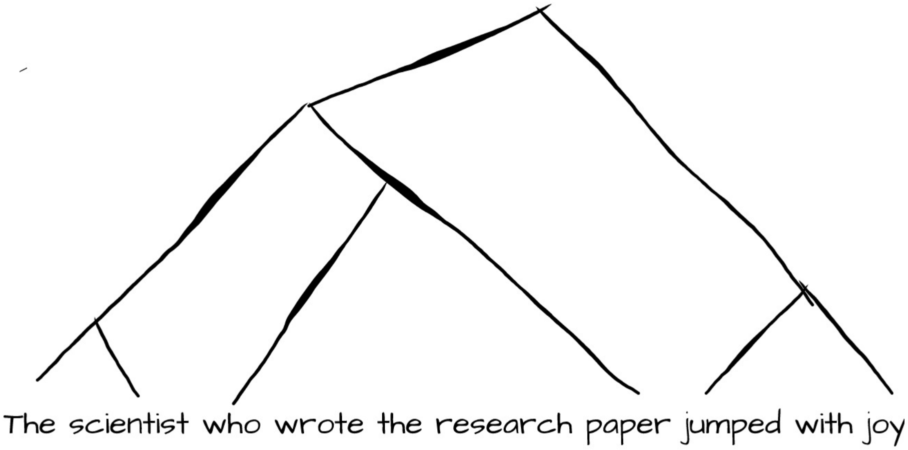
Hierarchical Compositionality

The scientist who wrote the research paper jumped with joy

Hierarchical Compositionality

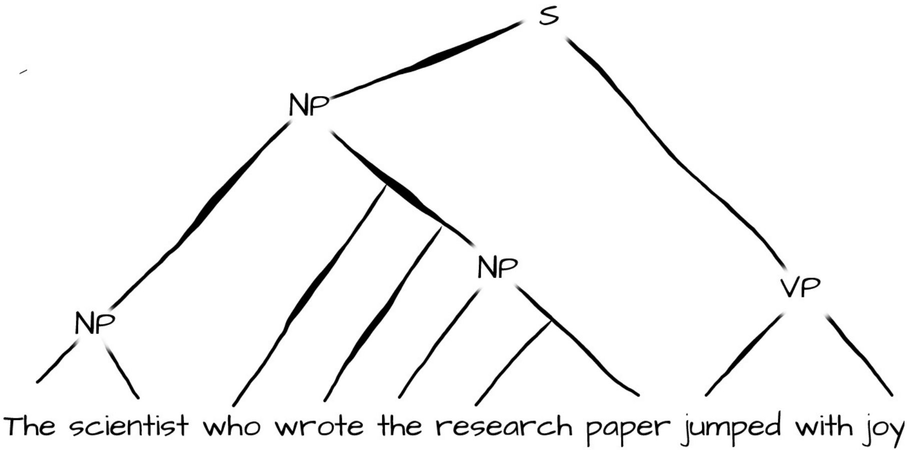


Hierarchical Compositionality



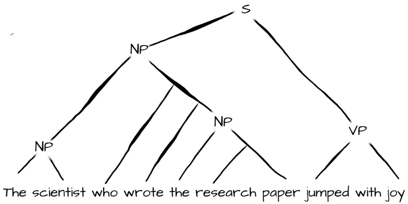
Hierarchical Compositionality

Hierarchical Compositionality



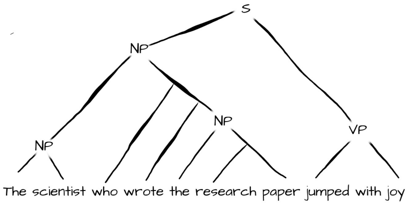
Symbolic structure and the brain

?



Symbolic structure and the brain

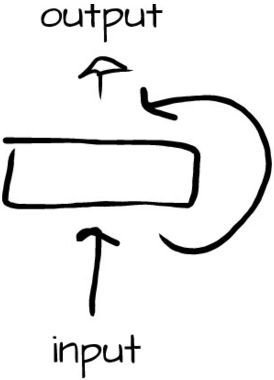
?



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

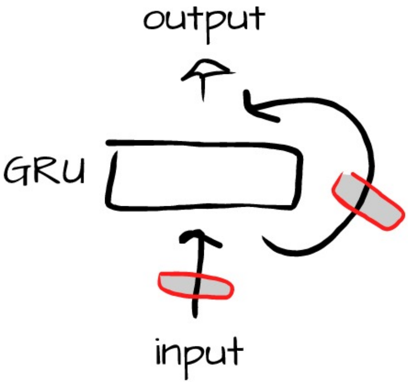
Recurrent Neural Networks

Simple Recurrent Network



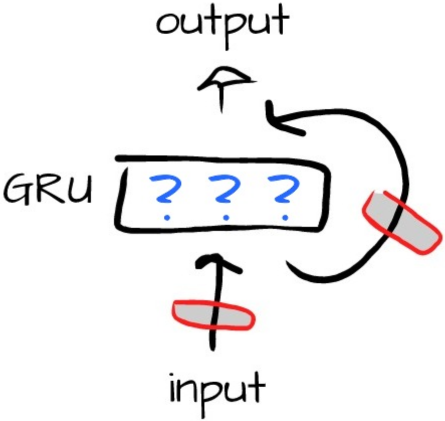
(Elman 1990)

Gated recurrent neural networks



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

Gated recurrent neural networks



Gated recurrent neural networks

- How can hierarchical compositionality be processed **incrementally**, in **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

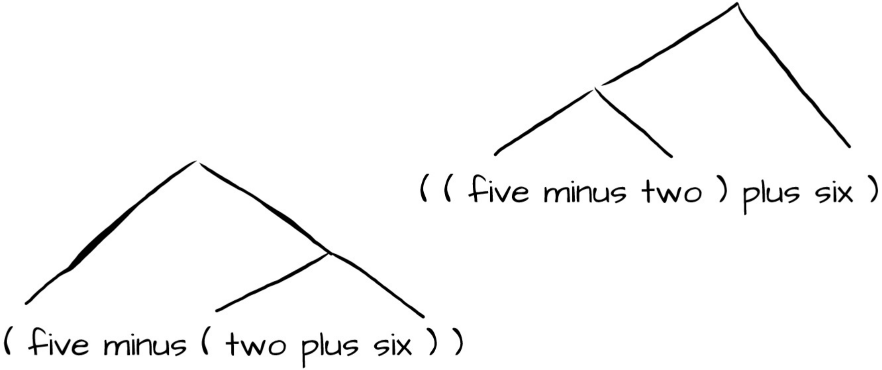
Arithmetic Language

((five minus two) plus six)

(five minus (two plus six))

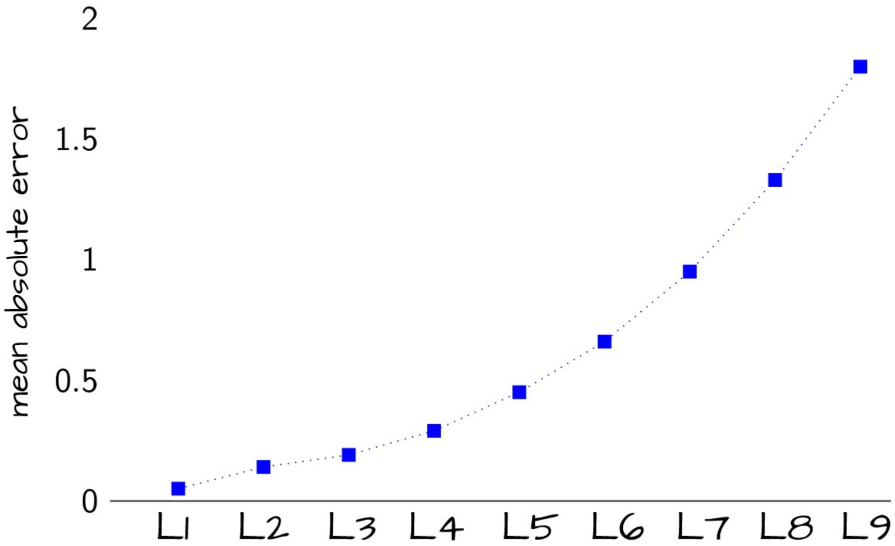
(Veldhoen, Hupkes, and Zuidema 2016; Hupkes, Veldhoen, and Zuidema 2018)

Arithmetic Language

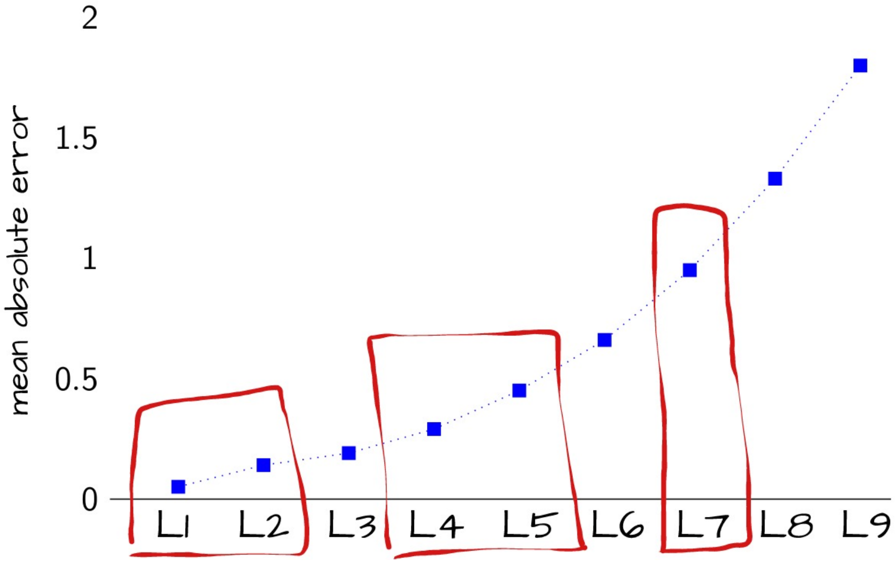


(Veldhoen, Hupkes, and Zuidema 2016; Hupkes, Veldhoen, and Zuidema 2018)

Can a gated recurrent network learn this language?



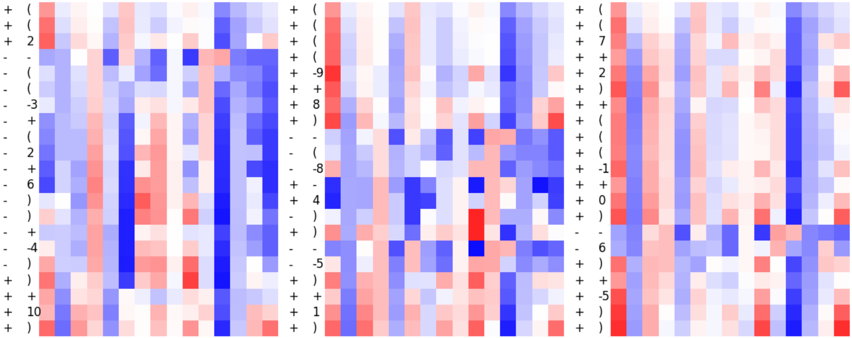
Can a gated recurrent network learn this language?



What does the network do?

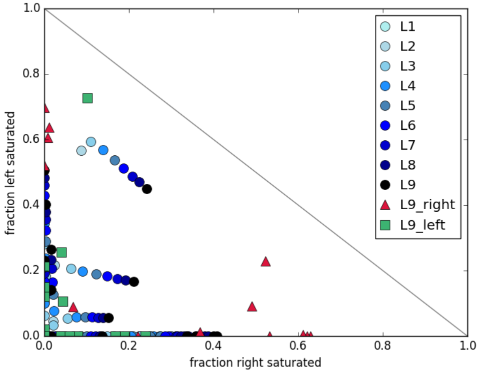
Looking inside

Plotting activation values



Looking inside

Update gate



(Karpathy, Johnson, and Fei-Fei 2015)

Symbolic solutions

(five minus (two plus six))

Symbolic solutions

recursively

(five minus (two plus six))

Symbolic solutions

recursively 5

(five minus (two plus six))

Symbolic solutions

recursively 5 5

$-$

(five minus (two plus six))

Symbolic solutions

5, -

-

recursively 5 5

(five minus (two plus six))

Symbolic solutions

5, -

-

recursively 5 5 2

(five minus (two plus six))

Symbolic solutions

5, -

\- +

recursively 5 5 2 2

(five minus (two plus six))

Symbolic solutions

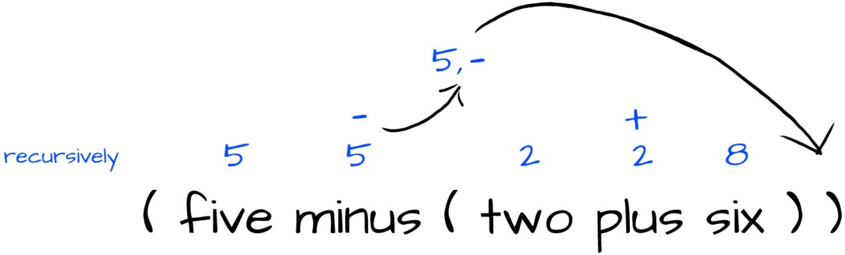
5, -

- +

recursively 5 5 2 2 8

(five minus (two plus six))

Symbolic solutions



Symbolic solutions

5, -

\- +

recursively 5 5 2 2 8 -3

(five minus (two plus six))

Symbolic solutions

5, -

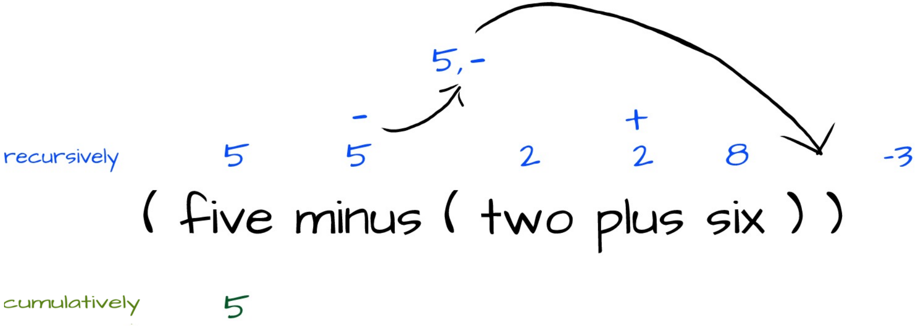
\- +

recursively 5 5 2 2 8 -3

(five minus (two plus six))

cumulatively

Symbolic solutions



Symbolic solutions

5, -

\- +

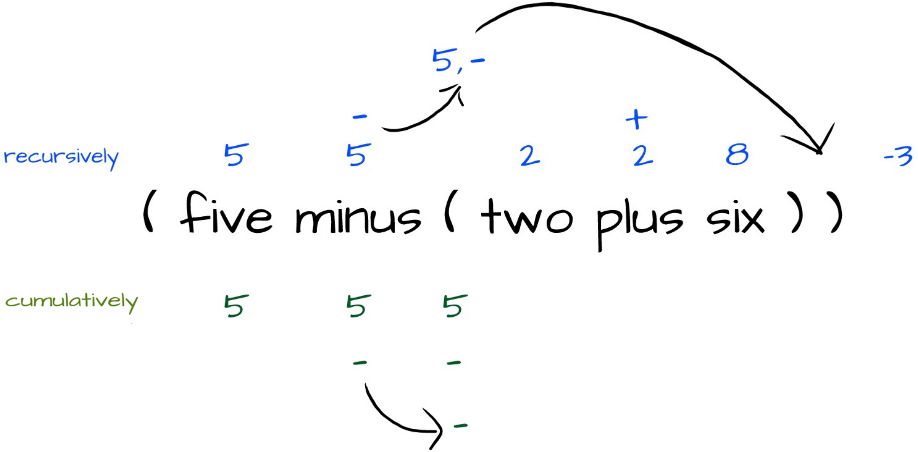
recursively 5 5 2 2 8 -3

(five minus (two plus six))

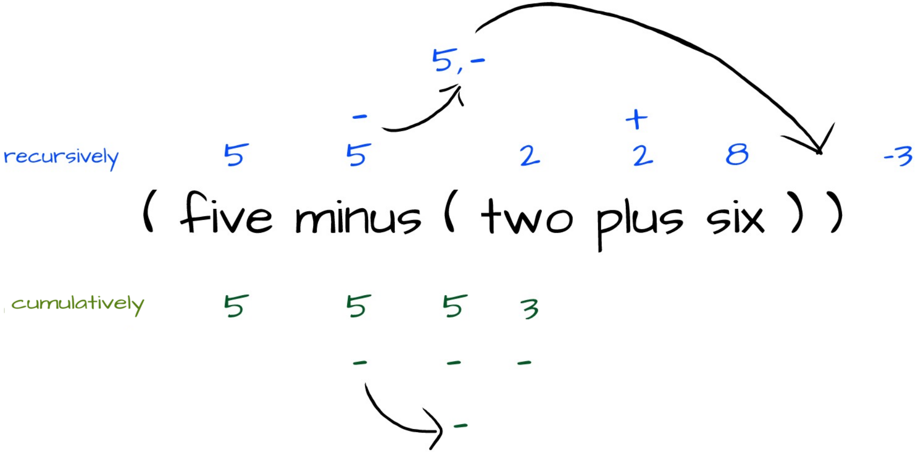
cumulatively 5 5

\-

Symbolic solutions



Symbolic solutions



Symbolic solutions

5, -

- +

recursively 5 5 2 2 8 -3

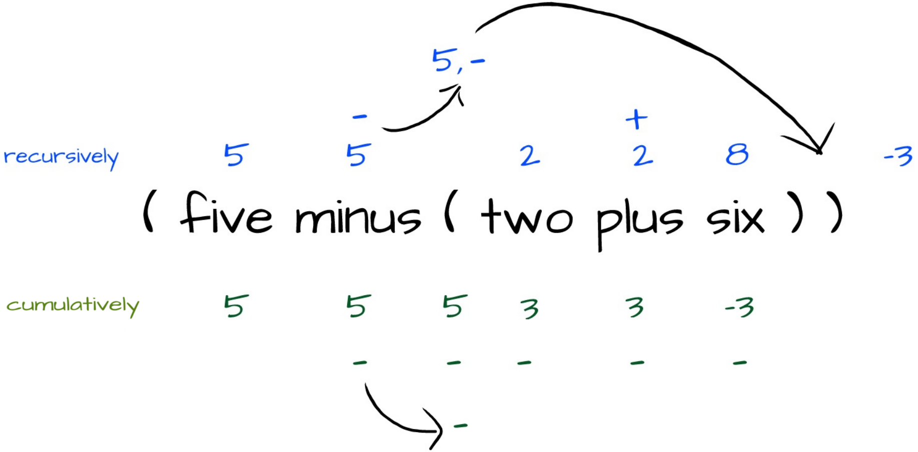
(five minus (two plus six))

cumulatively 5 5 5 3 3

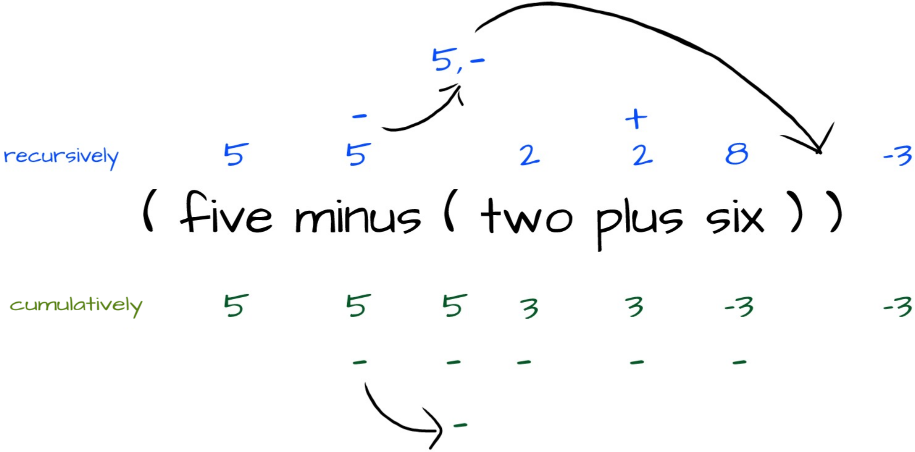
- - - -

-

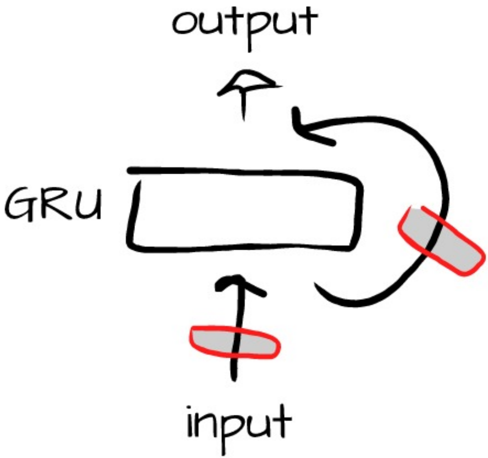
Symbolic solutions



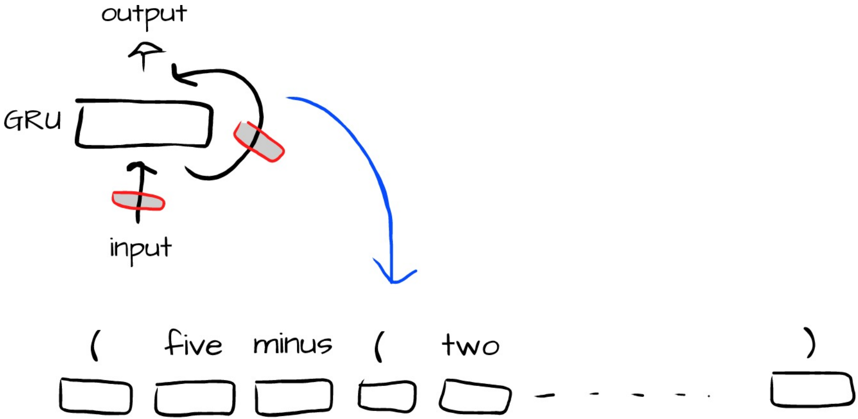
Symbolic solutions



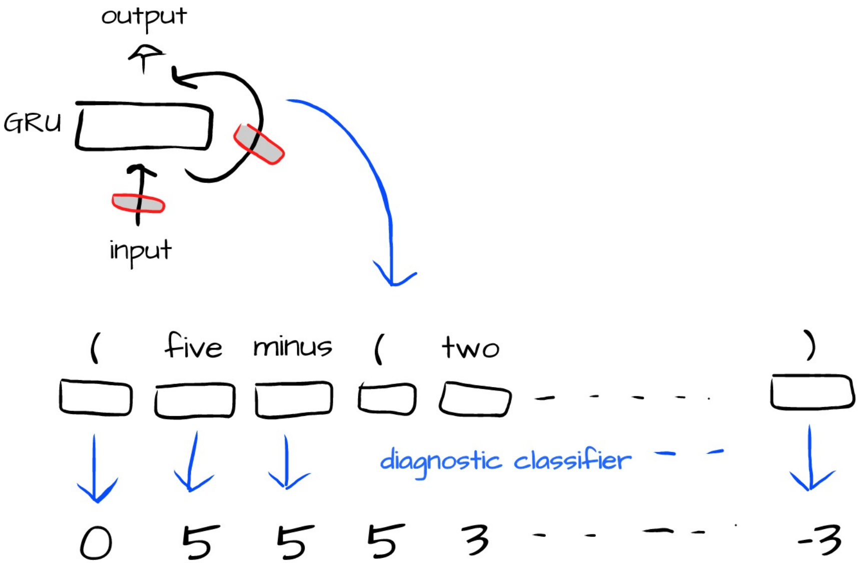
Diagnostic Classifier



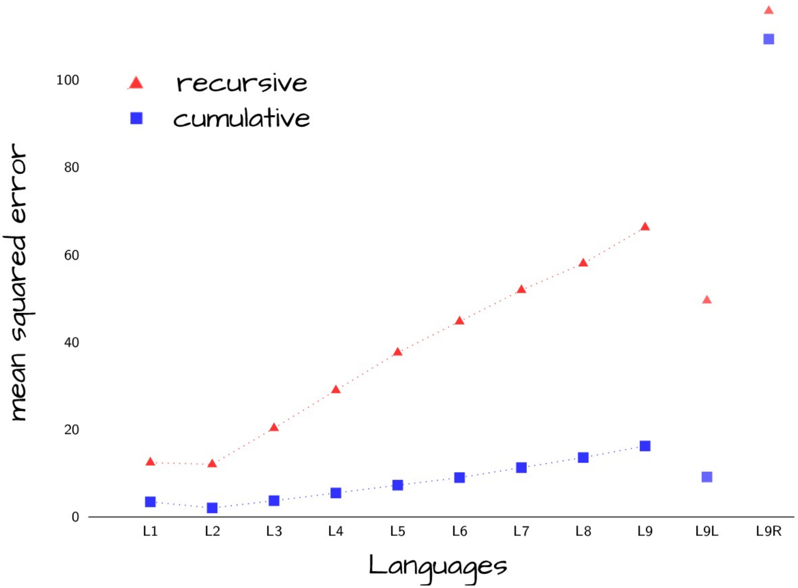
Diagnostic Classifier



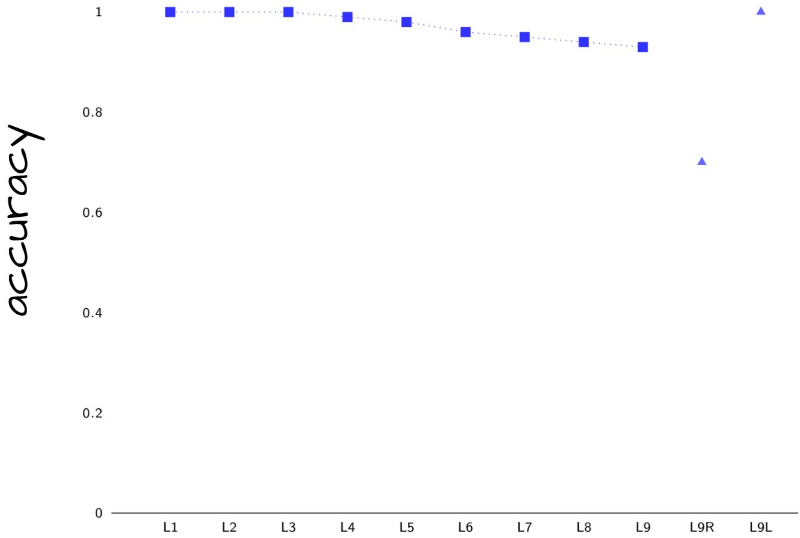
Diagnostic Classifier



Intermediate results



Cumulative strategy, operation mode



Discussion

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

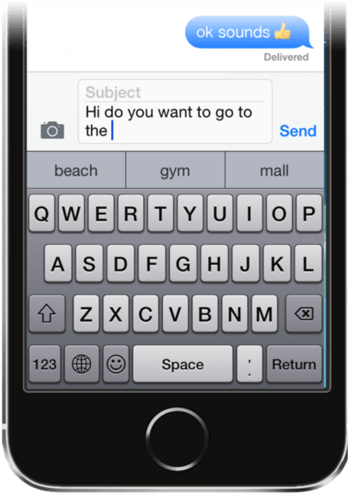
Discussion

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



Subject-verb agreement

The **scientist** who wrote the research paper **jumps** with joy

Subject-verb agreement

The **scientist** who wrote the research paper **jumps** with joy

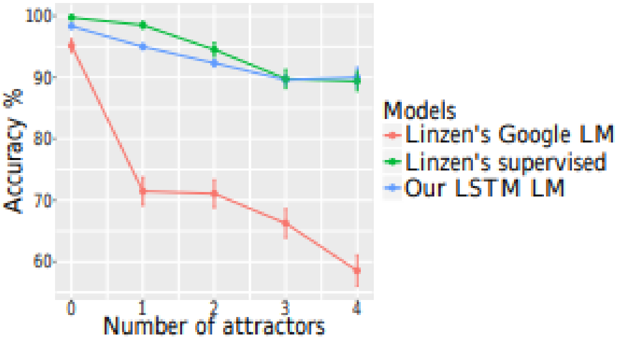
The **scientists** who wrote the research paper **jump** with joy

The number agreement task

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

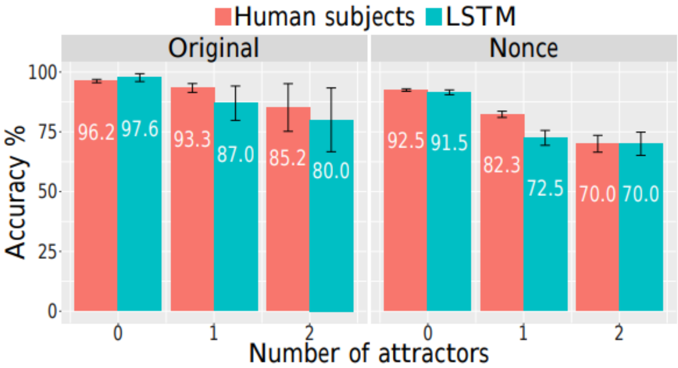
(Linzen, Dupoux, and Goldberg 2016)

Results



(Gulordava et al. 2018)

Results 2



(Gulordava et al. 2018)

Other linguistic questions

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

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

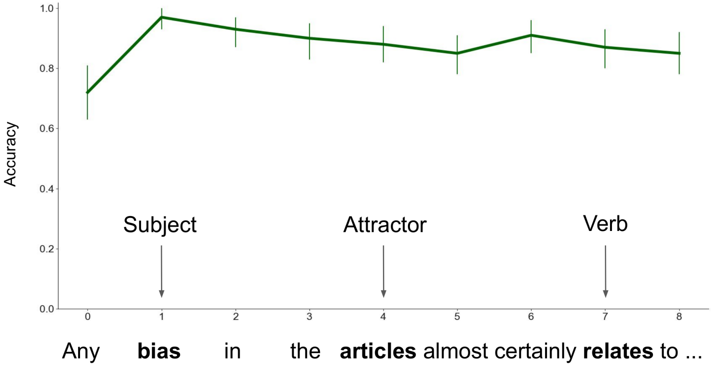
- Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)
- Filler-gap dependencies (Wilcox et al. 2018; Wilcox et al. 2019)
- Reflexive anaphora (Marvin and Linzen 2018; Futrell et al. 2018)
- Garden path sentences (Futrell et al. 2018; Van Schijndel and Linzen 2018; Futrell et al.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

But *how* do they do this?

Diagnostic classification 2

Diagnostic Classification

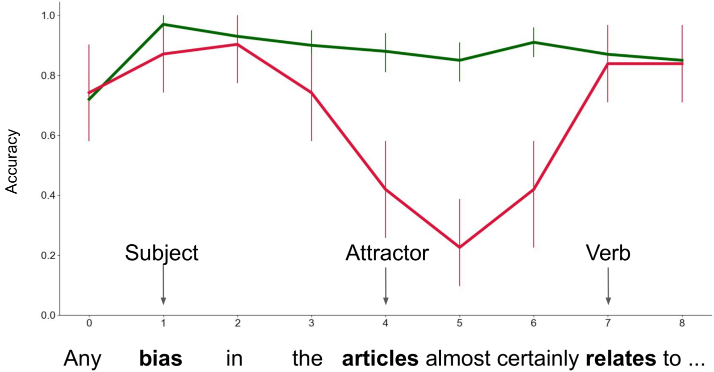
Sentences with correct predictions, h



(Giulianelli et al. 2018)

Diagnostic Classification

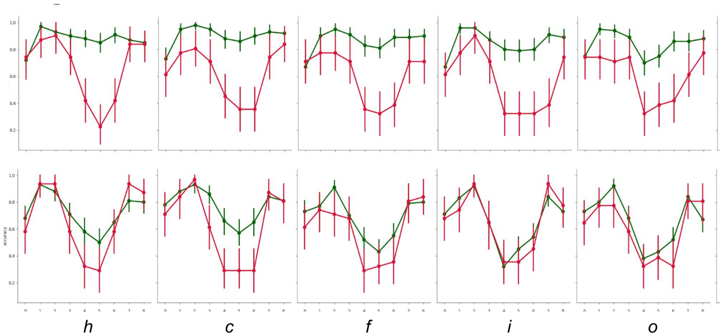
All sentences, h



(Giulianelli et al. 2018)

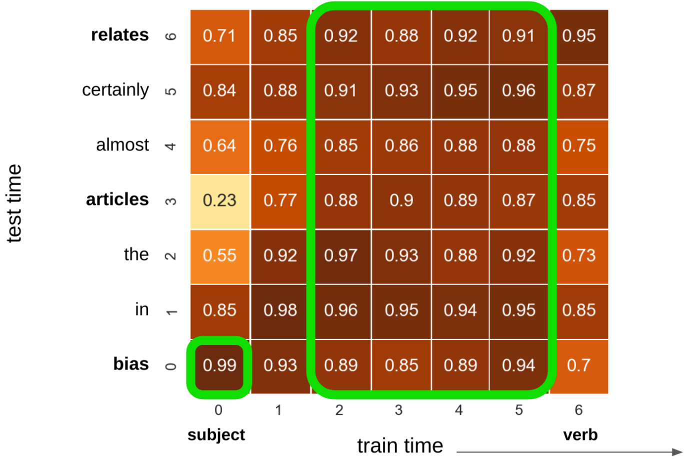
Diagnostic Classification

All sentences, all components



(Giulianelli et al. 2018)

Temporal generalisation matrix



(Giulianelli et al. 2018)

What else can we do?

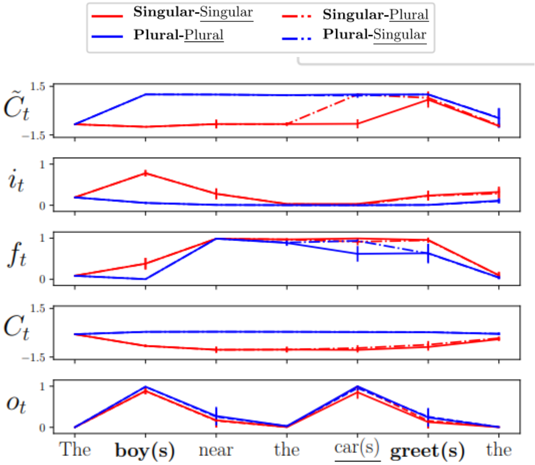
Ablation studies

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

(Lakretz et al. 2019)

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

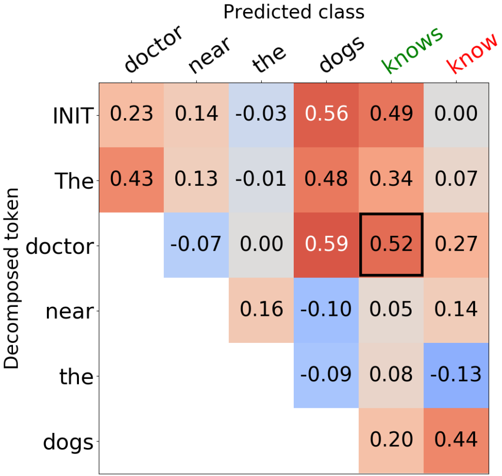
Ablation studies



(a) 988 (singular)

Lakretz et al. 2019

Contextual Decomposition



(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

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

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

Marco Baroni

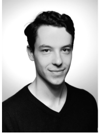
Jaap Jumelet

Germàn Kruszewski

Yair Lakretz

Sara Veldhoen

Mario Giulianelli



Florian Mohnert

Jack Harding

References I

Kyunghyun Cho et al. "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259* (2014).

Junyoung Chung et al. "Gated feedback recurrent neural networks". In: *arXiv preprint arXiv:1502.02367* (2015).

Jeffrey L Elman. "Finding structure in time". In: *Cognitive science* 14.2 (1990), pp. 179–211.

Richard Futrell et al. "RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency". In: *arXiv preprint arXiv:1809.01329* (2018).

References II

Richard Futrell et al. "Neural language models as psycholinguistic subjects: Representations of syntactic state". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 32–42.

Mario Giulianelli et al. "Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 240–248.

Kristina Gulordava et al. "Colorless Green Recurrent Networks Dream Hierarchically". In: *Proceedings of NAACL*. Vol. 1. 2018, pp. 1195–1205.

References III

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. “Visualisation and ‘diagnostic classifiers’ reveal how recurrent and recursive neural networks process hierarchical structure”. In: *Journal of Artificial Intelligence Research* 61 (2018), pp. 907–926.

Dieuwke Hupkes and Willem Zuidema. “Diagnostic classification and symbolic guidance to understand and improve recurrent neural networks”. In: *Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning, NIPS2017*. 2017.

Jaap Jumelet and Dieuwke Hupkes. “Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 222–231.

Jaap Jumelet, Dieuwke Hupkes, and Willem Zuidema. "Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment". In: *Proceedings of CoNLL*. 2019.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. "Visualizing and understanding recurrent networks". In: *arXiv preprint arXiv:1506.02078* (2015).

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535. ISSN: 2307-387X.

Yair Lakretz et al. "The emergence of number and syntax units in LSTM language models". In: *arXiv preprint arXiv:1903.07435* (2019).

References V

Rebecca Marvin and Tal Linzen. "Targeted Syntactic Evaluation of Language Models". In: *EMNLP*. Association for Computational Linguistics, 2018, pp. 1192–1202.

Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. "Diagnostic Classifiers: Revealing how Neural Networks Process Hierarchical Structure". In: *Pre-Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo @ NIPS 2016)*. 2016.

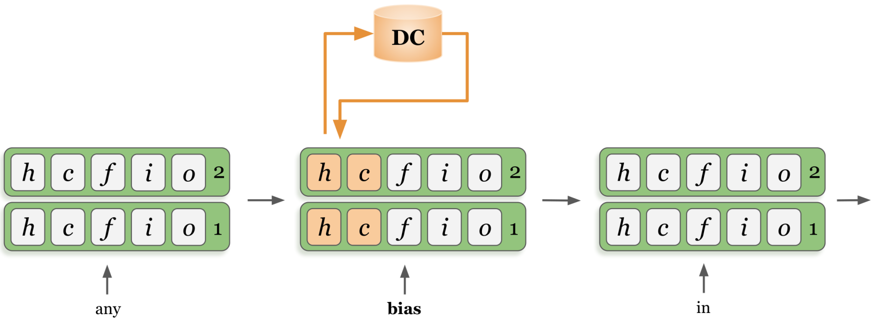
Marten Van Schijndel and Tal Linzen. "Modeling garden path effects without explicit hierarchical syntax.". In: *CogSci*. 2018.

Ethan Wilcox et al. "What do RNN language models learn about filler–gap dependencies?" In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 211–221.

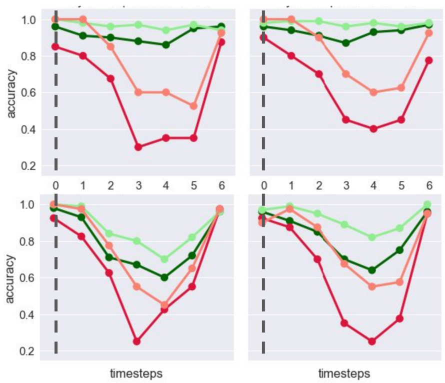
Ethan Wilcox et al. “Structural Supervision Improves Learning of Non-Local Grammatical Dependencies”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 3302–3312.

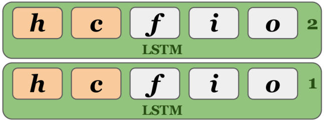
Interventions

Diagnostic interventions



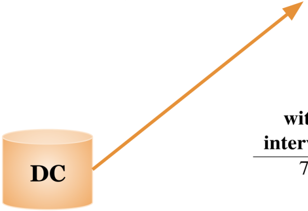
Diagnostic interventions





Diagnostic interventions, results

	An	official	estimate	issued	in	2003	suggests	suggest
Original		-11.05	-8.426	-8.472	-1.243	-3.951	-5.753	**-5.6979**
Intervention		-11.05	-8.426	-8.472	-1.268	-3.97	**-5.691**	-6.4361



without intervention	**with intervention**
78.0	85.4

Subject-verb agreement in Language Models

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

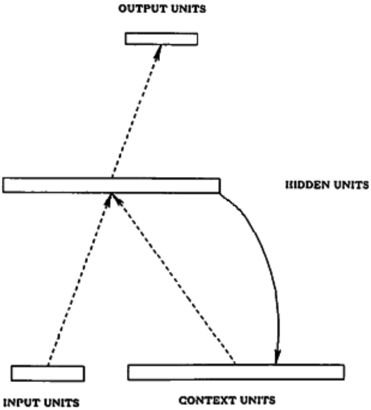
	Accuracy	**Accuracy with intervention**
Original	78.1	85.4
Nonce	70.7	75.6

(Giulianelli et al. 2018)

Gated Recurrent Neural Networks

Simple Recurrent Network

$$\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{\mathbf{t-1}} + \mathbf{b})$$

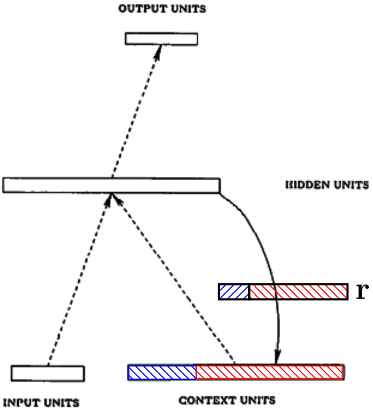


(Elman 1990)

Gated recurrent neural networks

$$\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{\mathbf{t-1}} + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r)$$

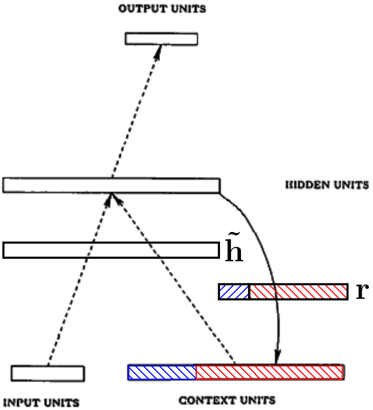


(Cho et al. 2014; Chung et al. 2015)

Gated recurrent neural networks

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r} \odot \mathbf{h}_{t-1}) + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r)$$



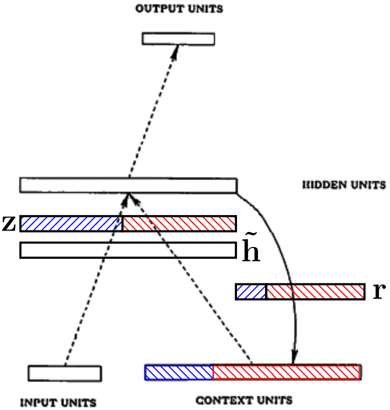
(Cho et al. 2014; Chung et al. 2015)

Gated recurrent neural networks

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r} \odot \mathbf{h}_{t-1}) + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z)$$



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

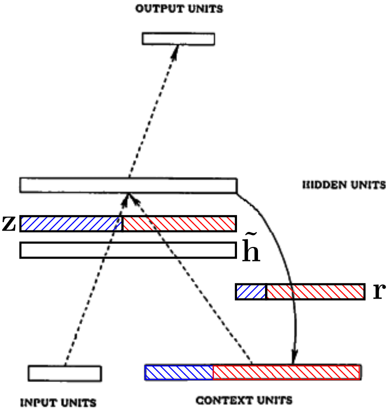
Gated recurrent neural networks

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r} \odot \mathbf{h}_{t-1}) + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z)$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t$$



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