Interpretation of Neural Networks

Dieuwke Hupkes

Institute for Logic, Language and Computation
University of Amsterdam

May 14, 2019

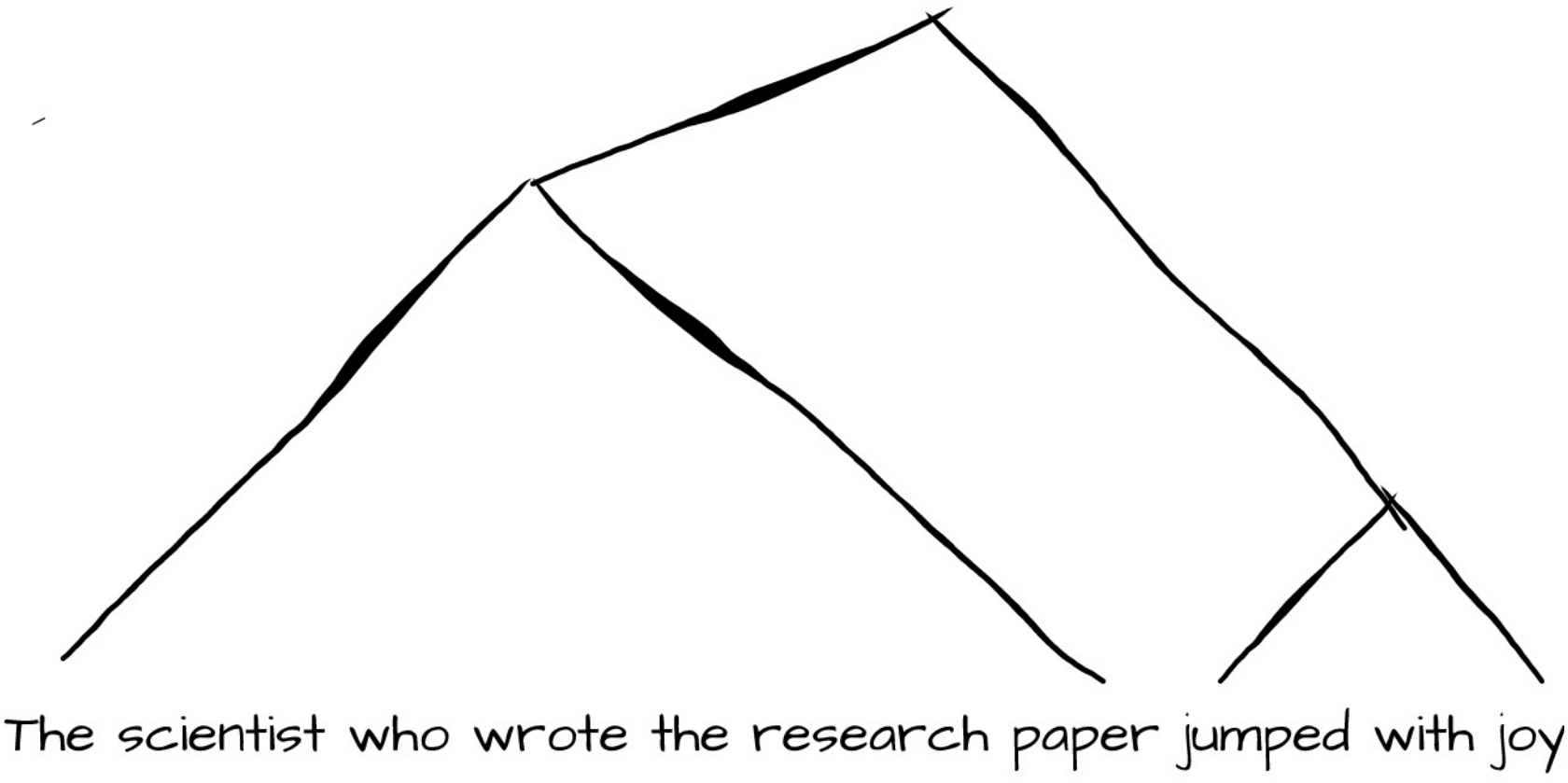
Outline

1. Can RNNs represent hierarchy?
   - The arithmetic language
   - Diagnostic classifiers
2. What do neural language models learn?
   - The subject verb agreement task
   - Neuron ablation studies
   - Temporal generalisation matrix and interventions
   - Contextual Decomposition

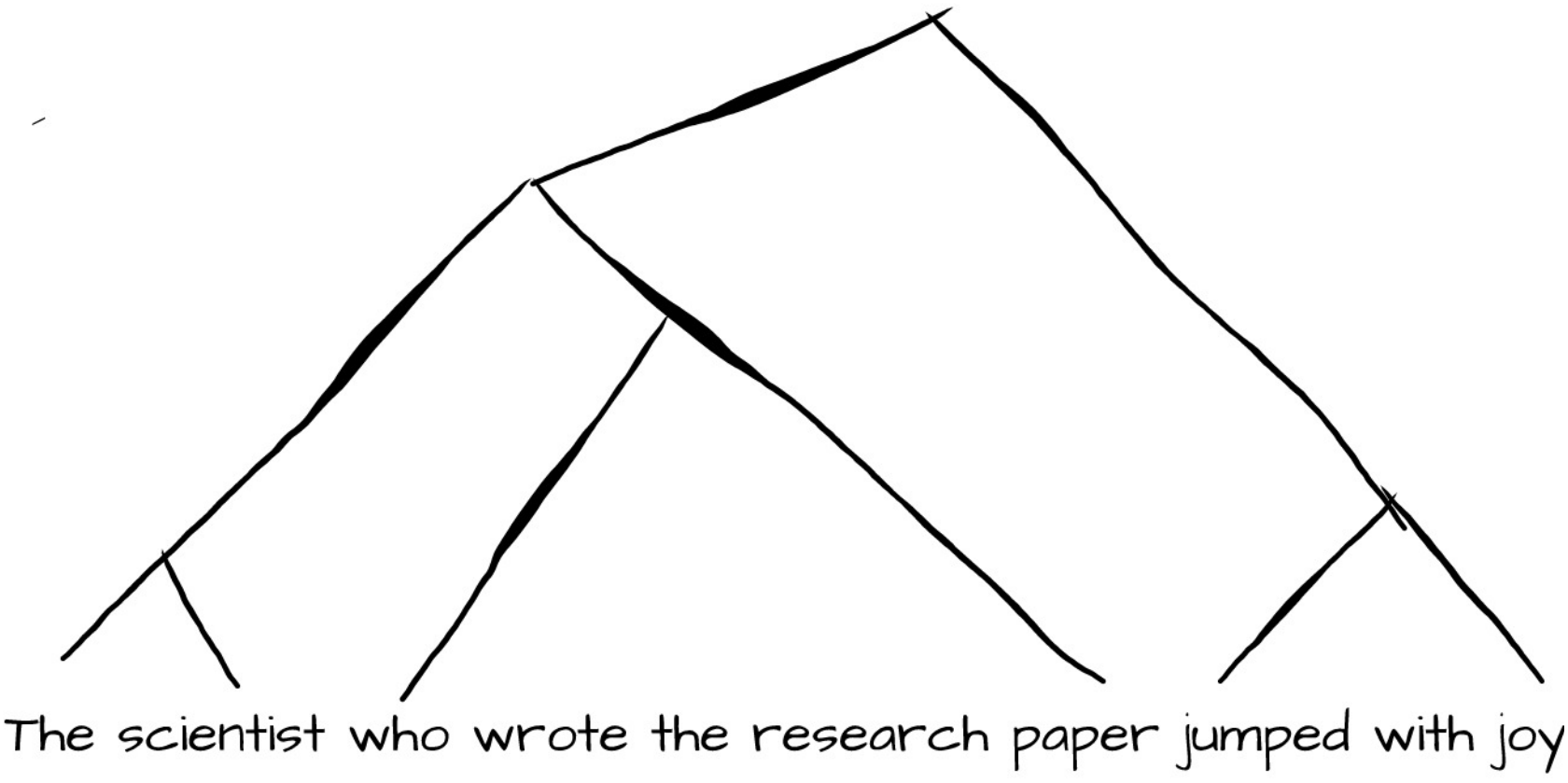
Hierarchical compositionality

The scientist who wrote the research paper jumped with joy

Hierarchical compositionality

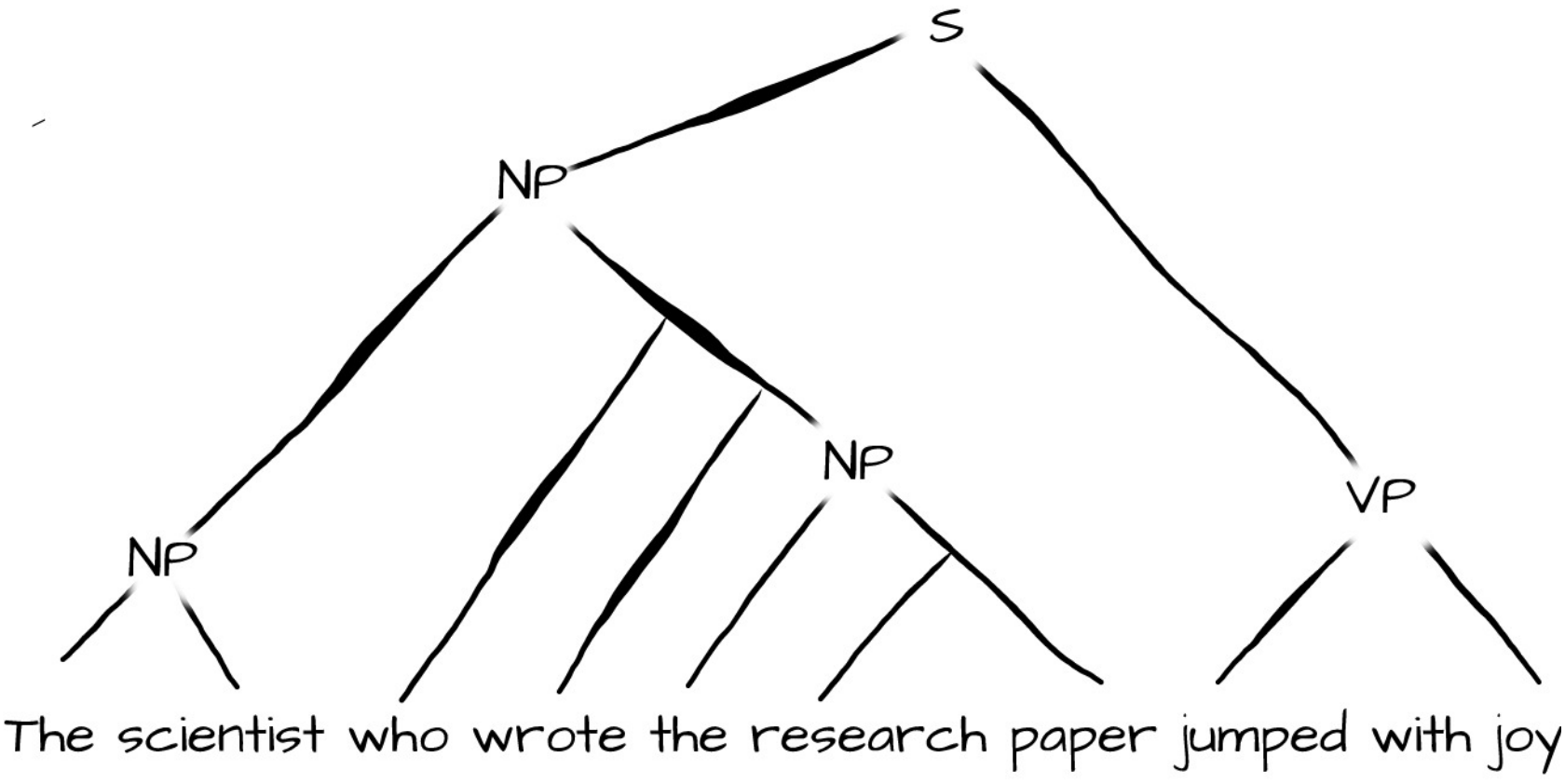


Hierarchical compositionality



Hierarchical compositionality

Hierarchical compositionality

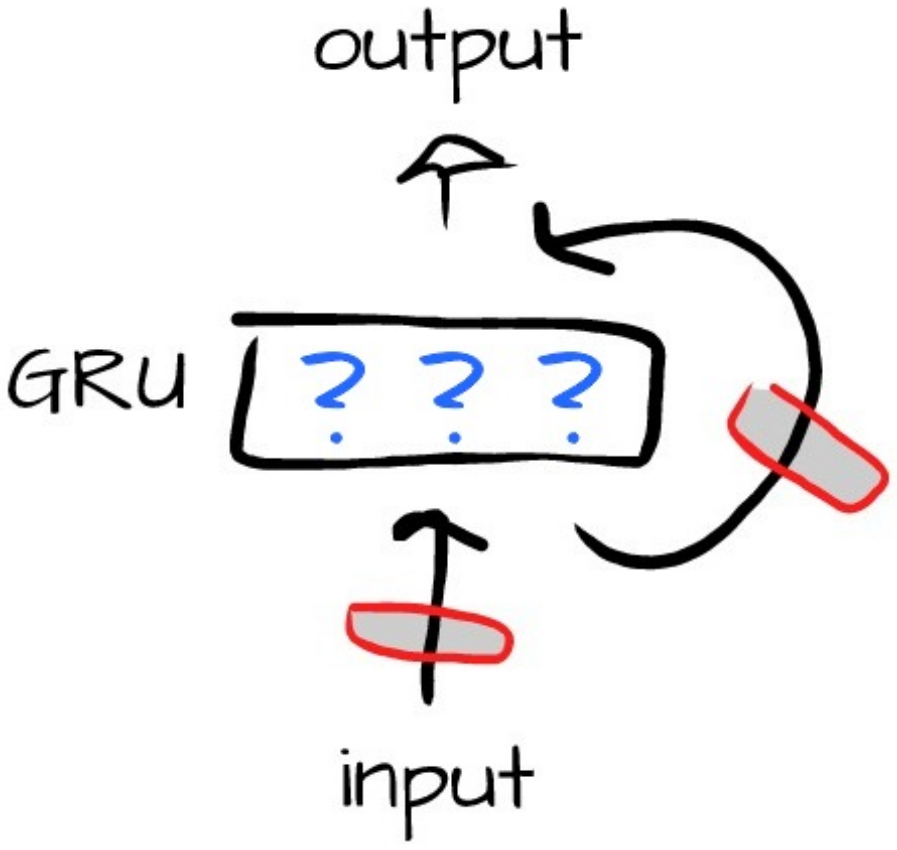


Recurrent Neural Networks

How do recurrent neural networks process such hierarchically compositional structures?

Recurrent Neural Networks

How do recurrent neural networks process such hierarchically compositional structures?

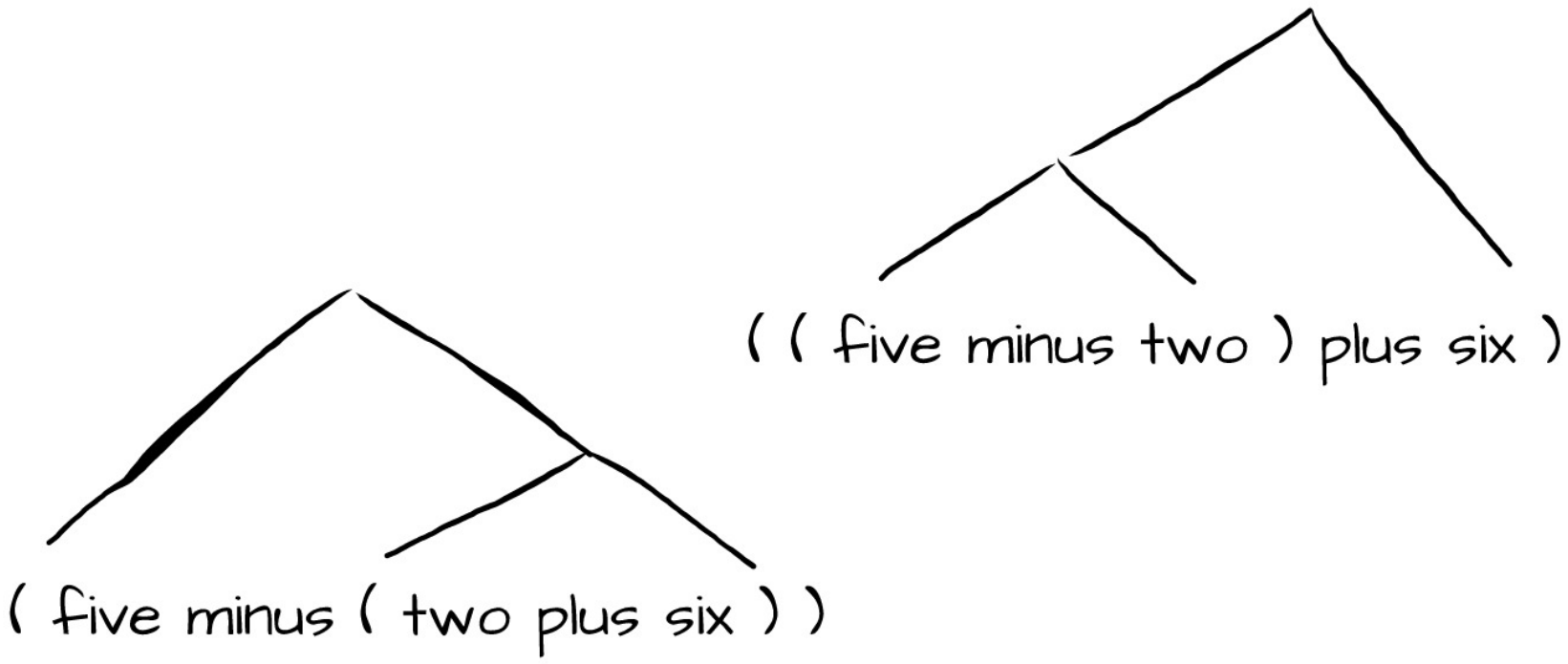


Arithmetic Language

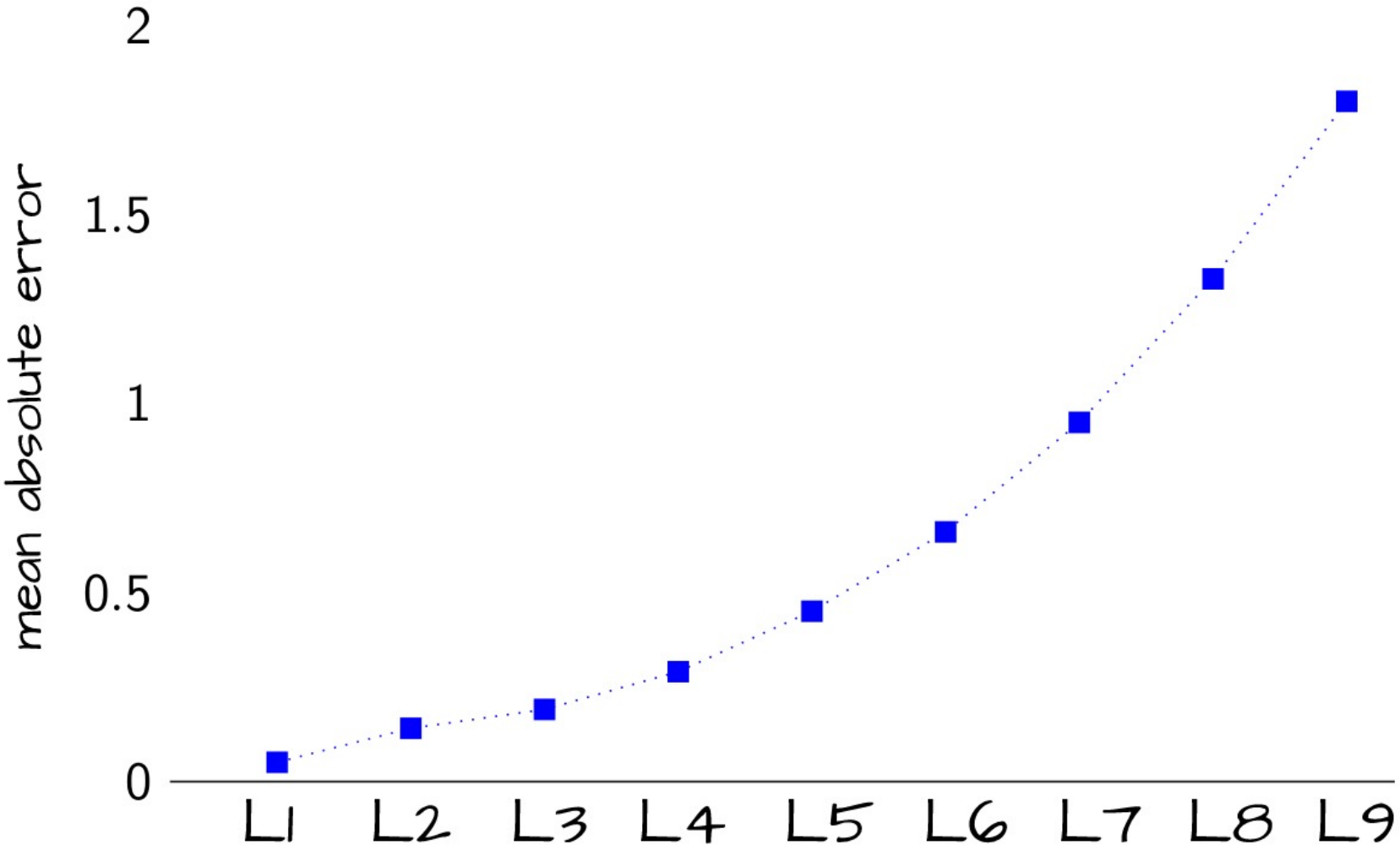
((five minus two) plus six)

(five minus (two plus six))

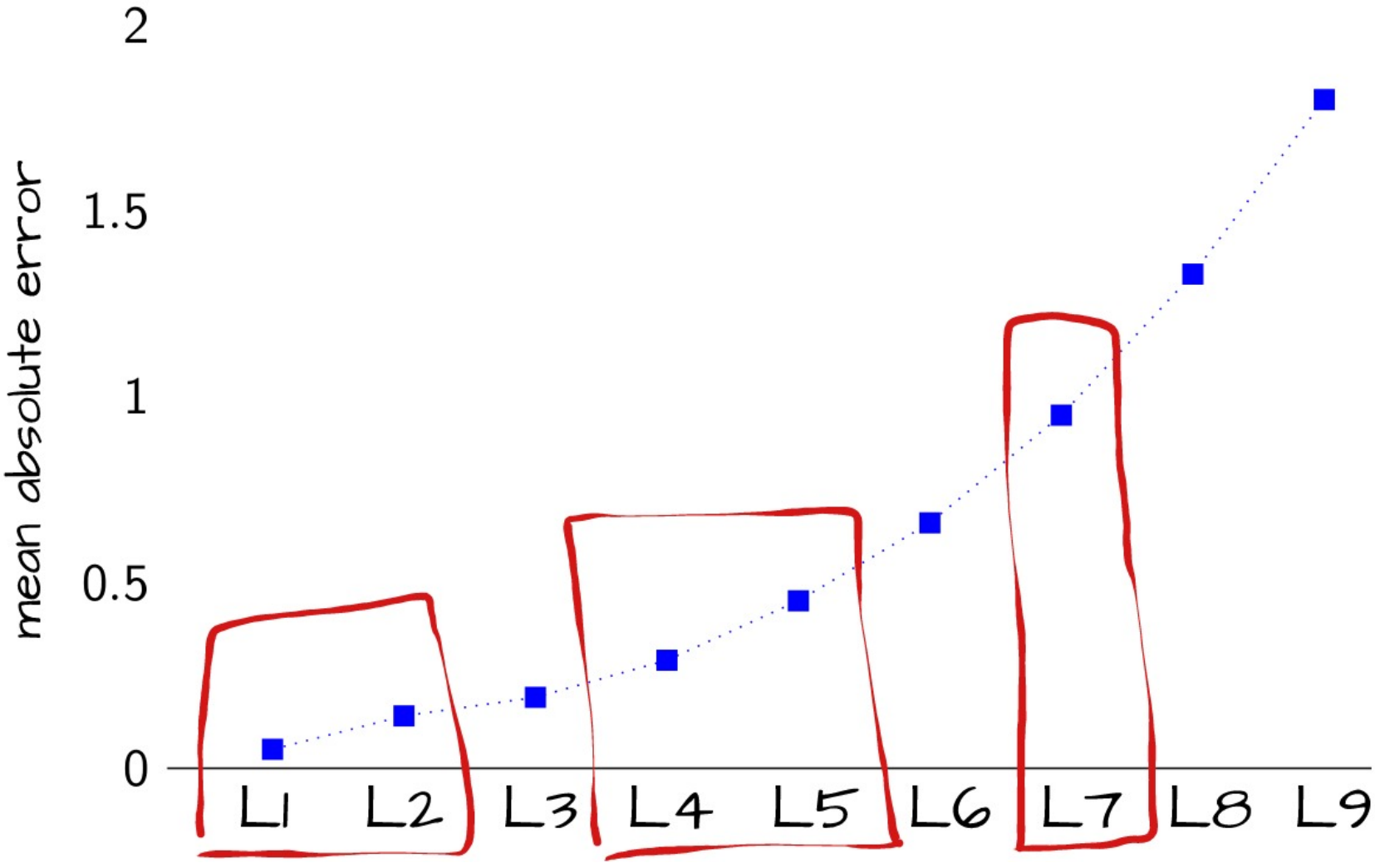
Arithmetic Language



Results



Results

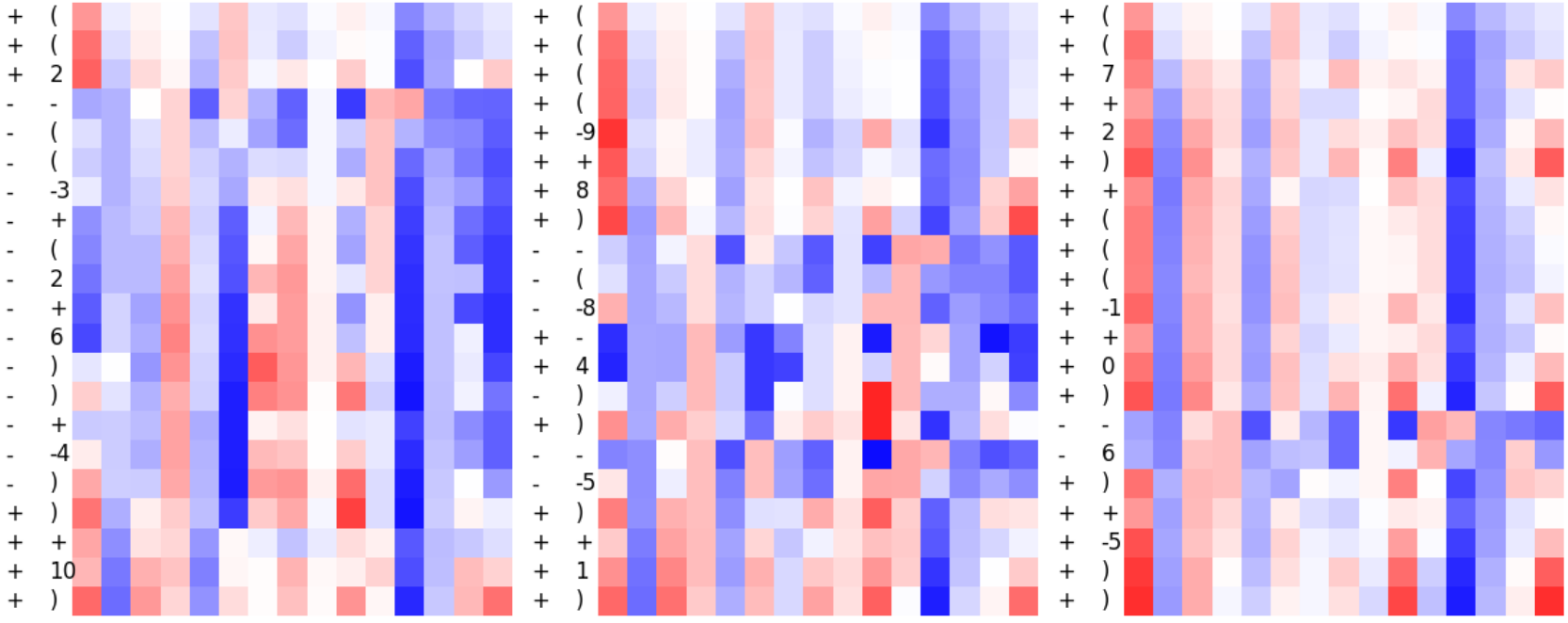


Looking inside

What does the network do?

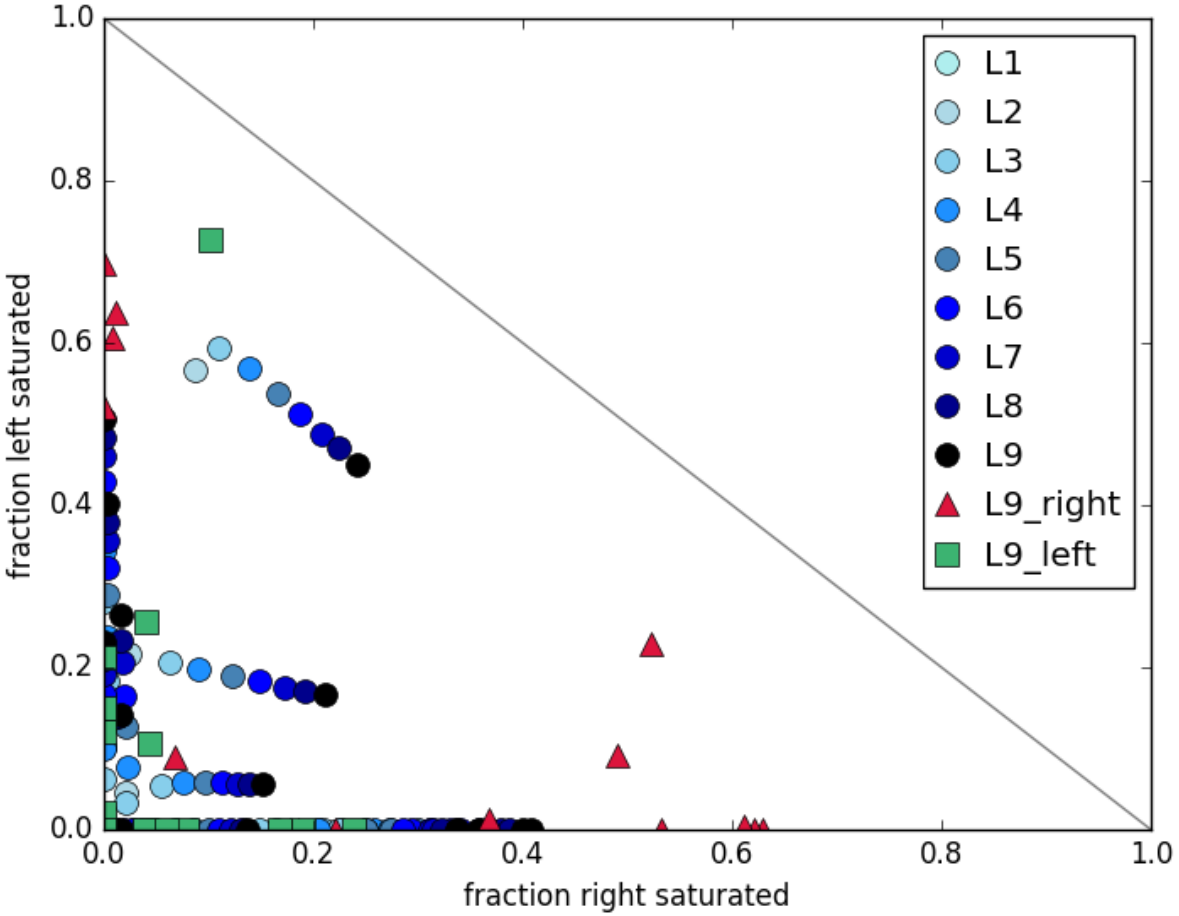
Looking inside

Plotting activation values



Looking inside

Update gate



Karpathy et al. (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))

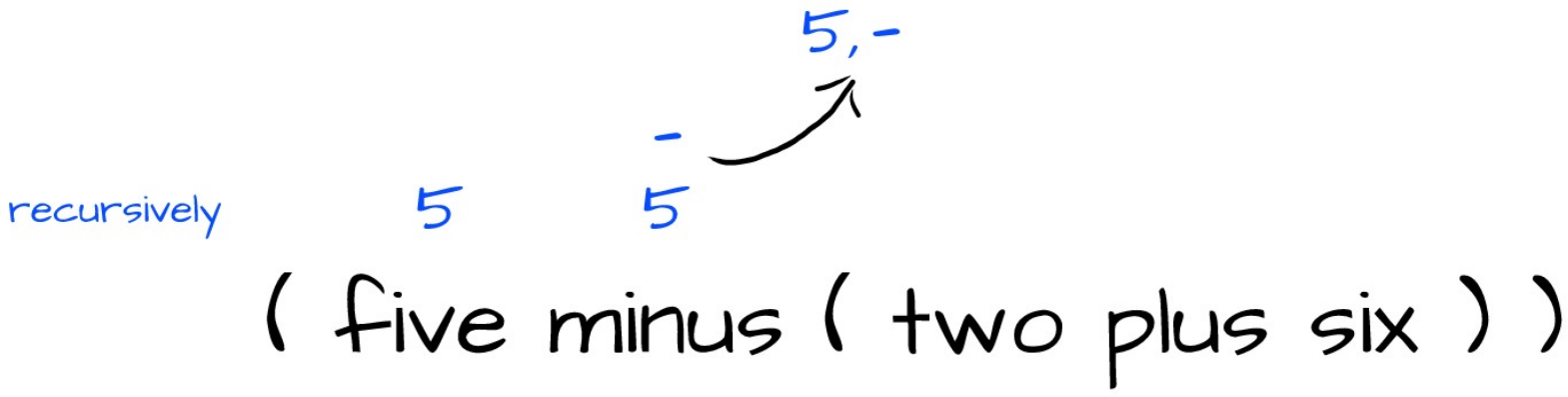
Symbolic solutions

recursively

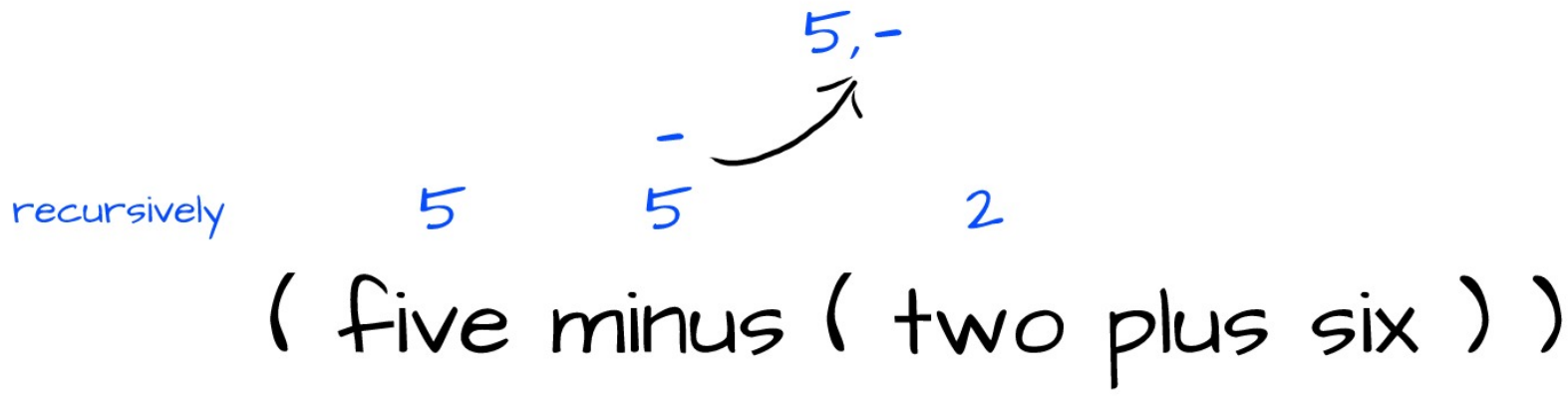
5 - 5

(five minus (two plus six))

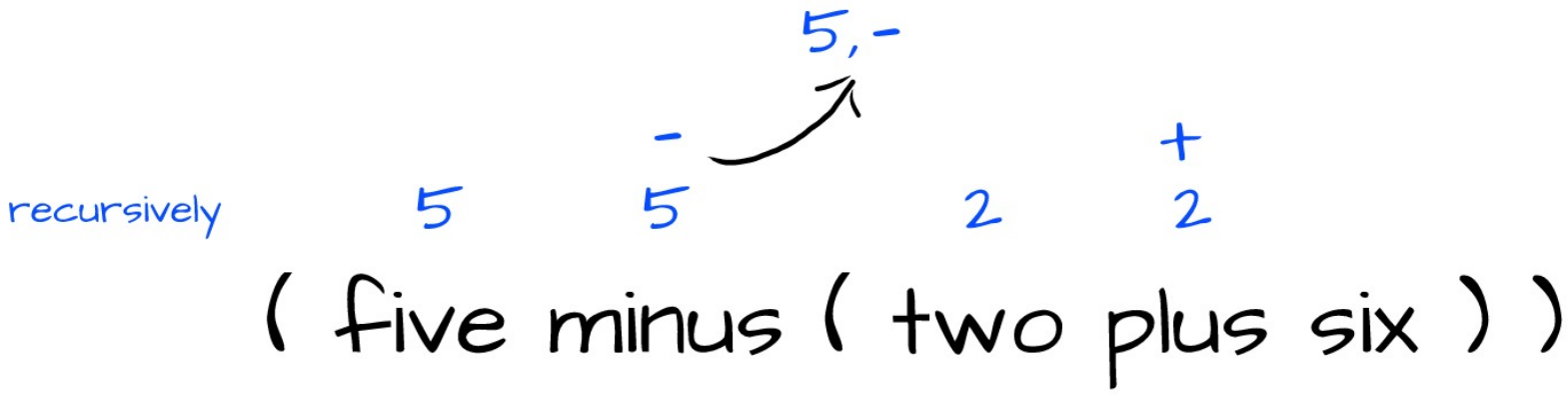
Symbolic solutions



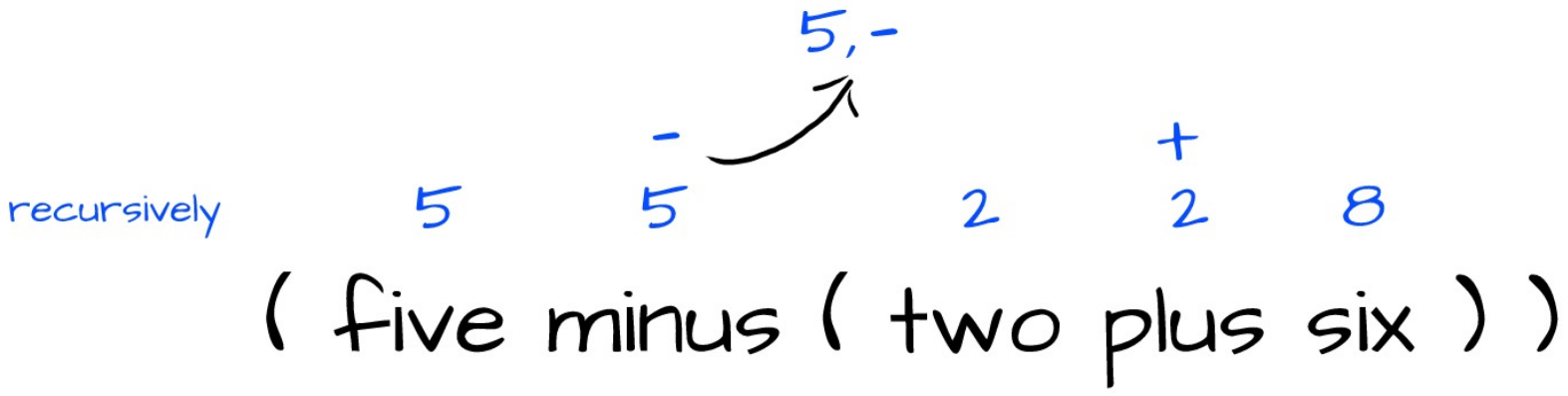
Symbolic solutions



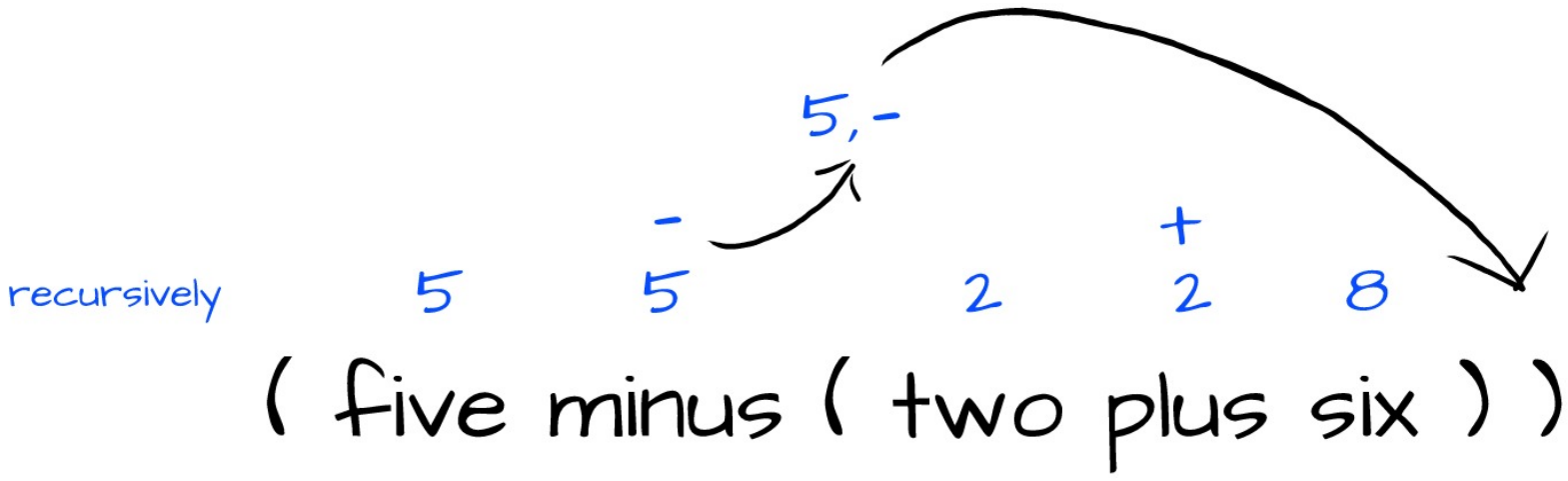
Symbolic solutions



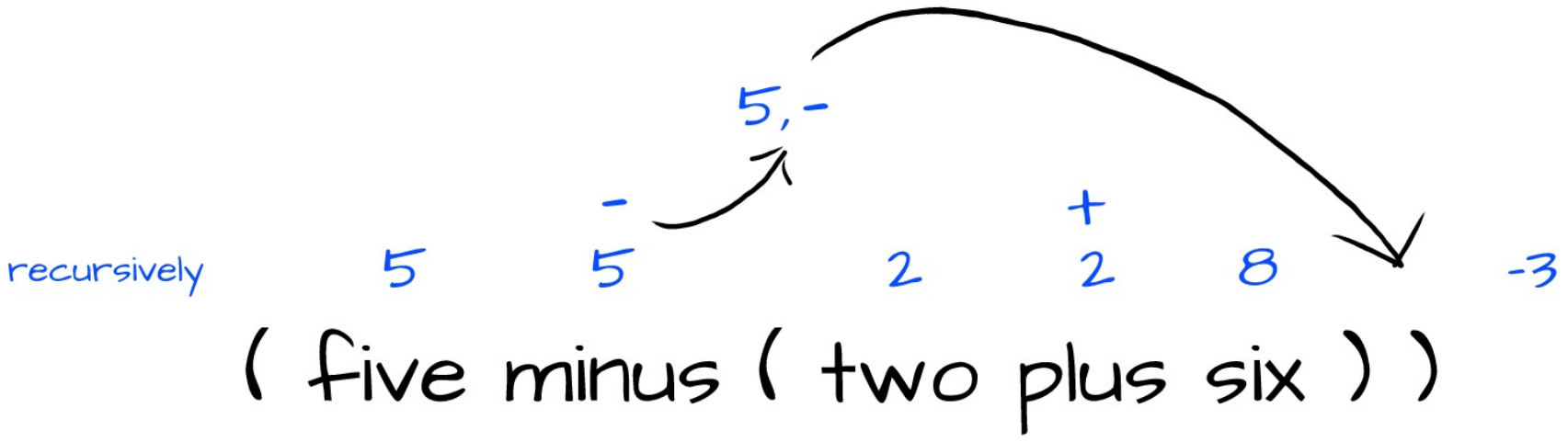
Symbolic solutions



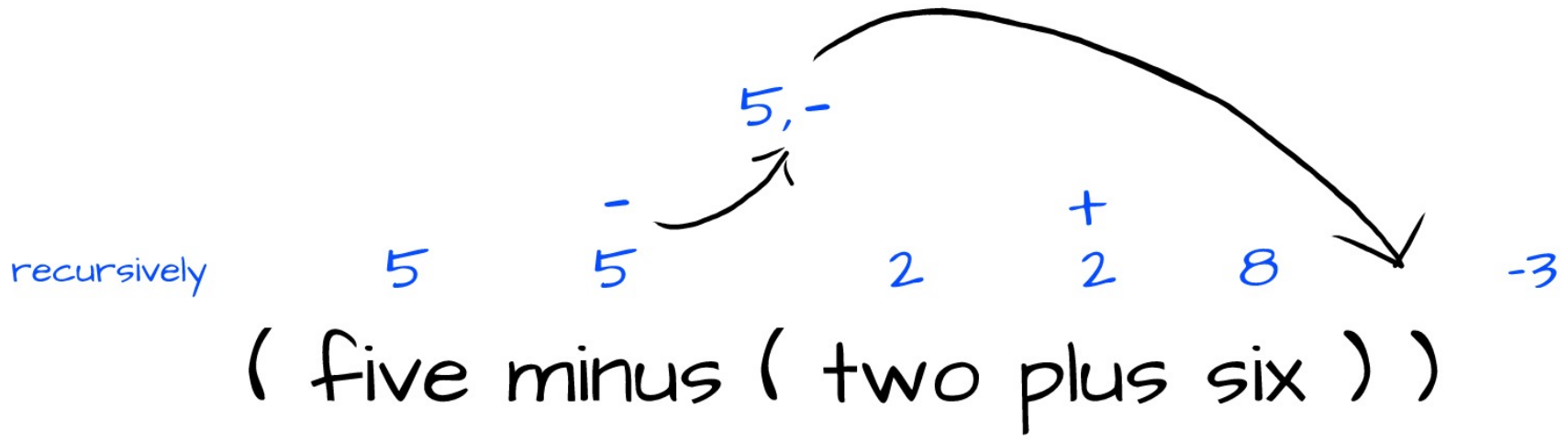
Symbolic solutions



Symbolic solutions

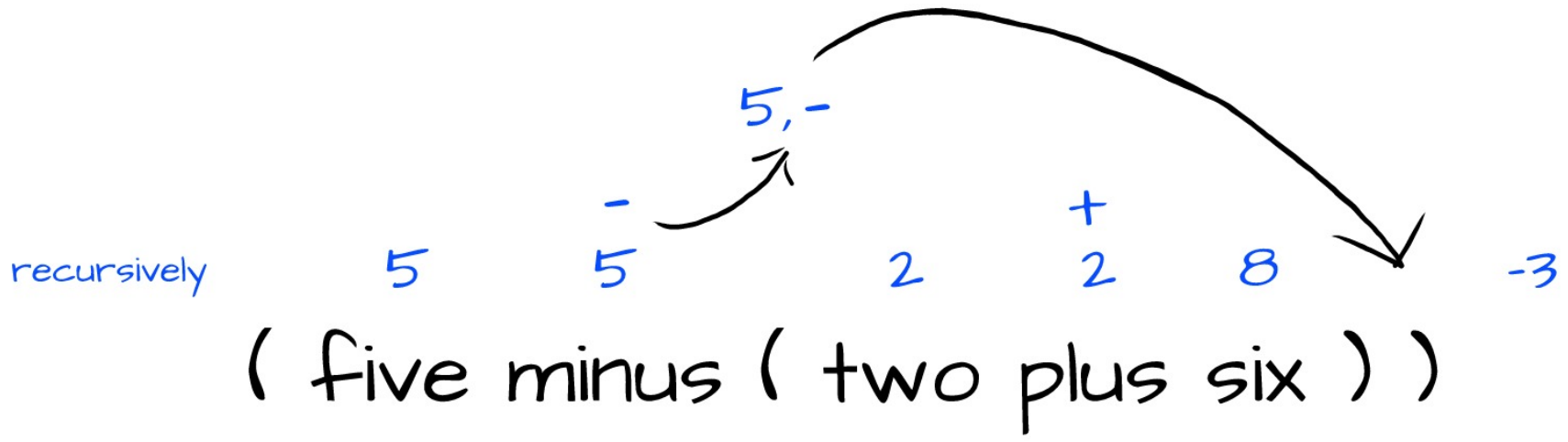


Symbolic solutions

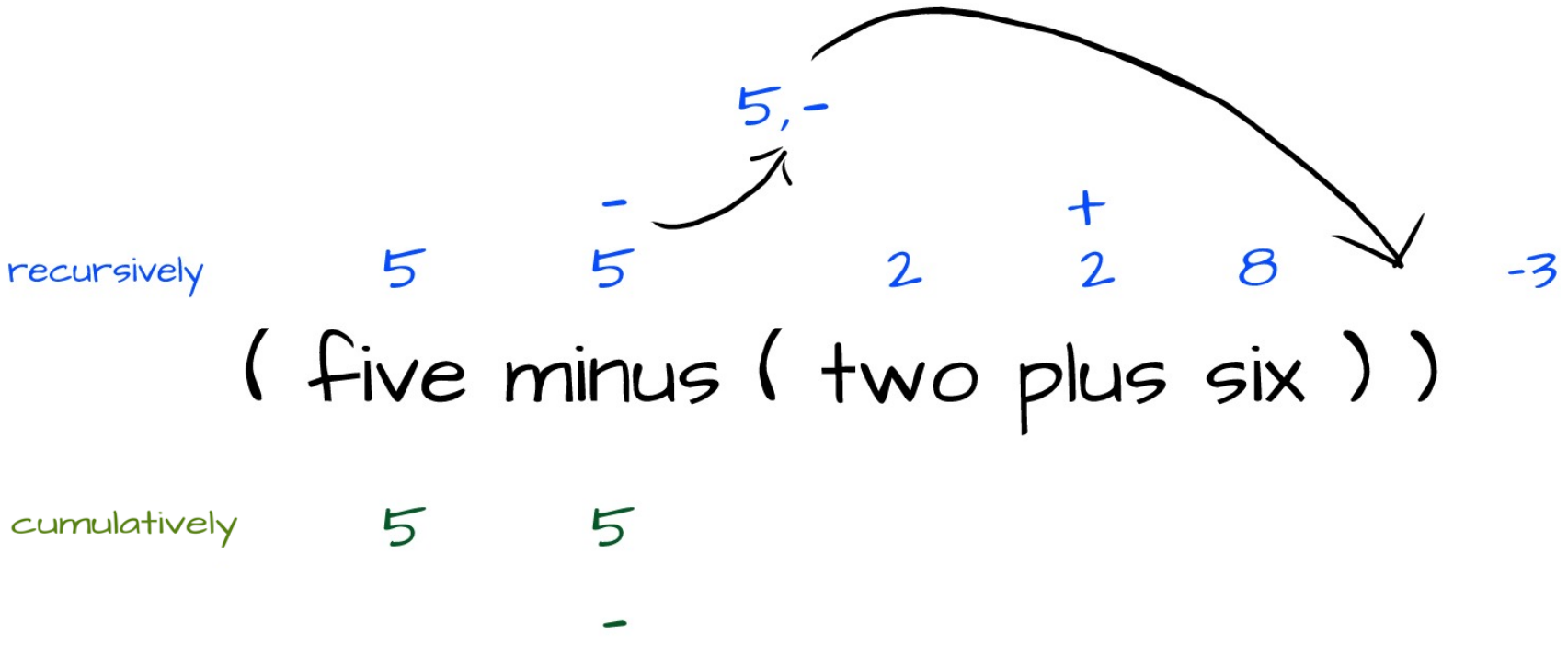


cumulatively

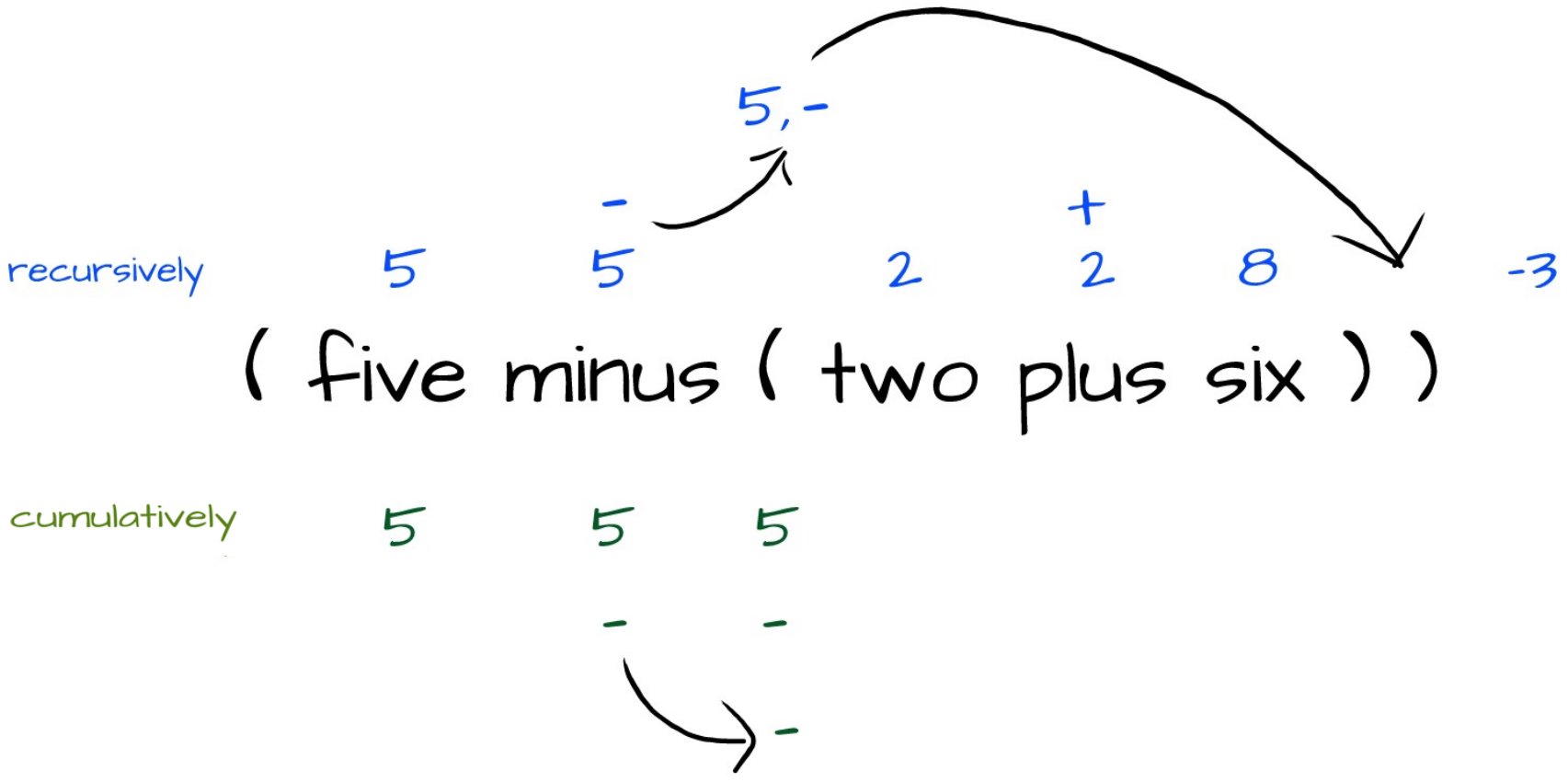
Symbolic solutions



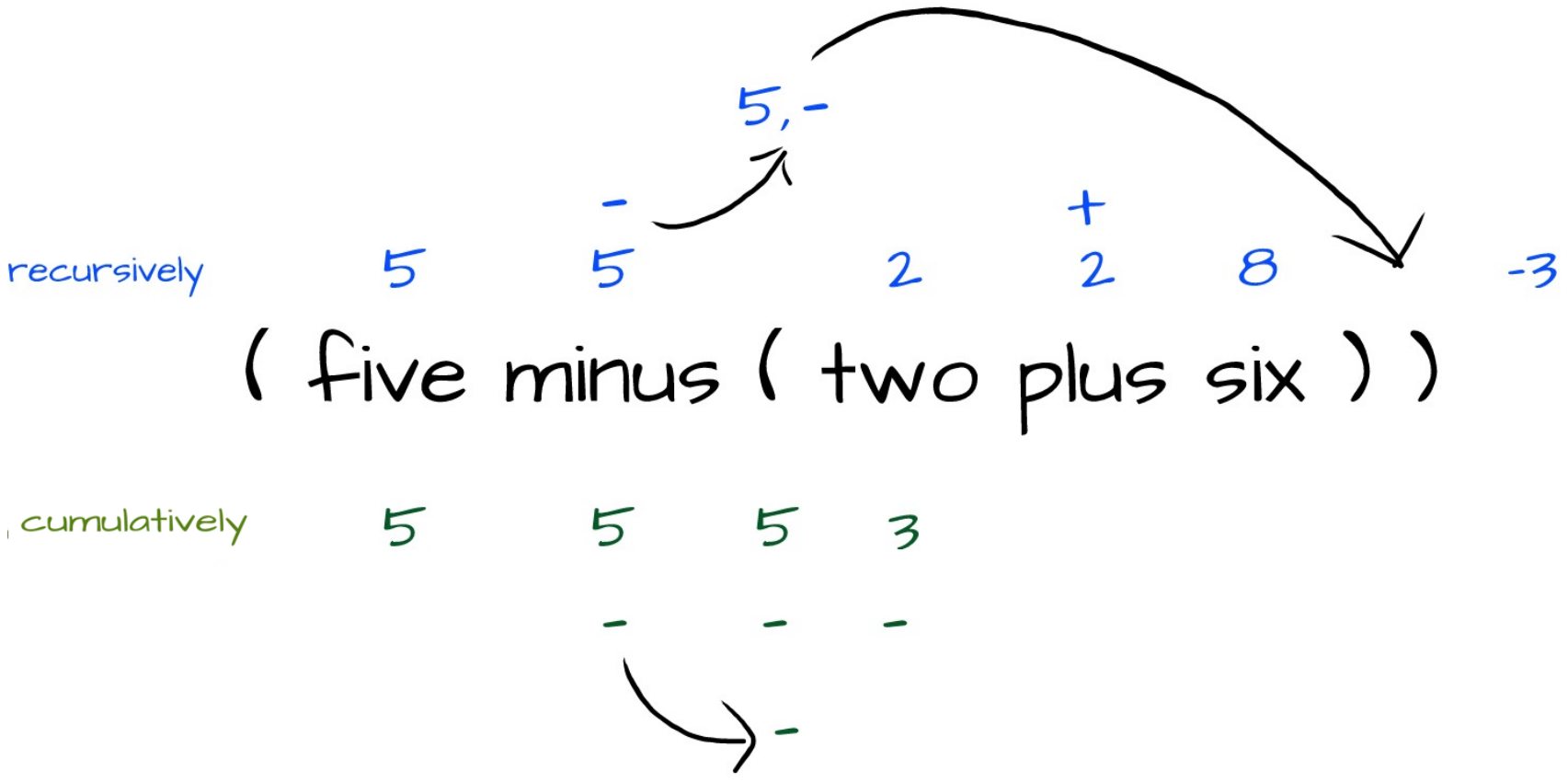
Symbolic solutions



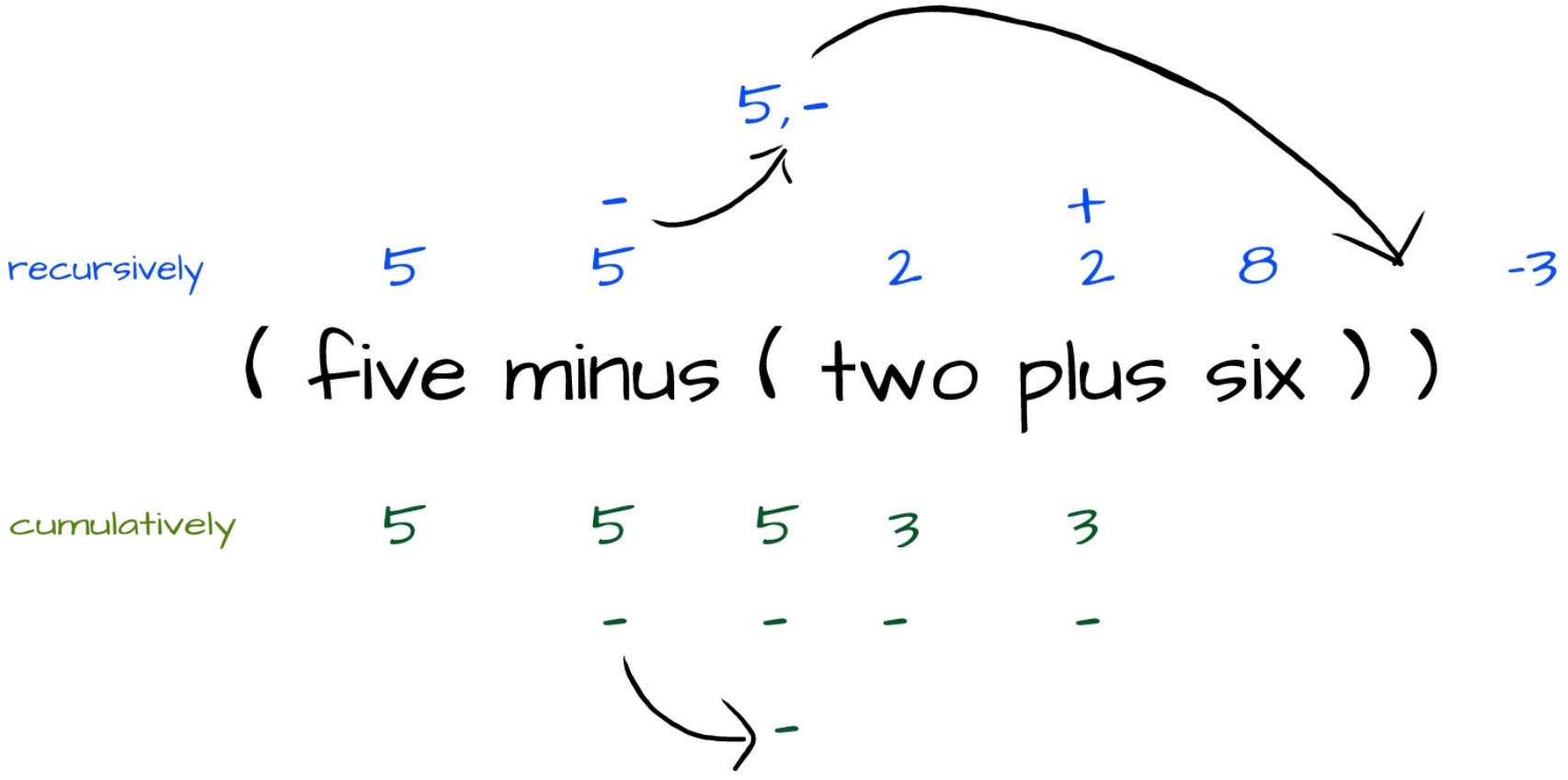
Symbolic solutions



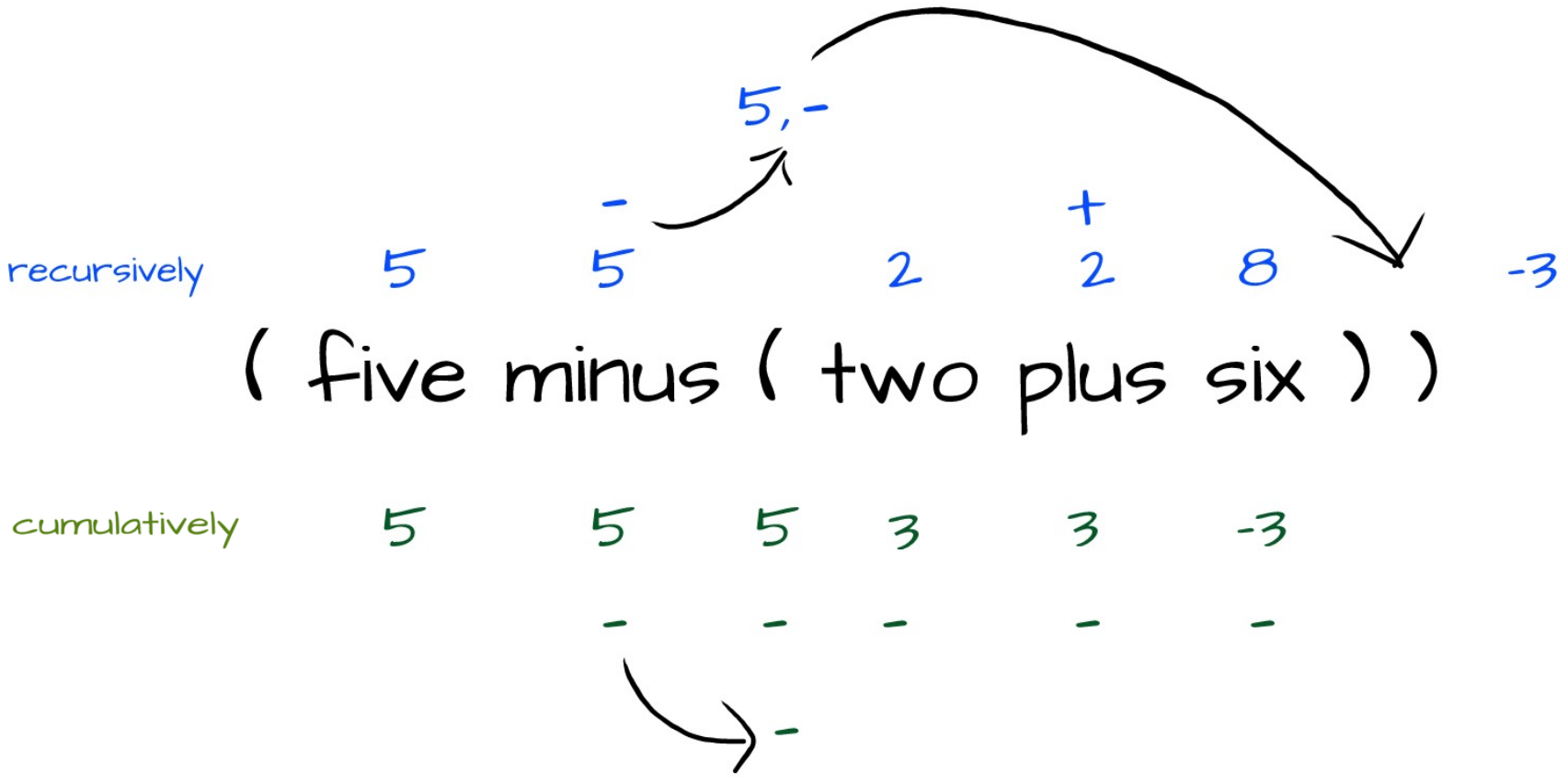
Symbolic solutions



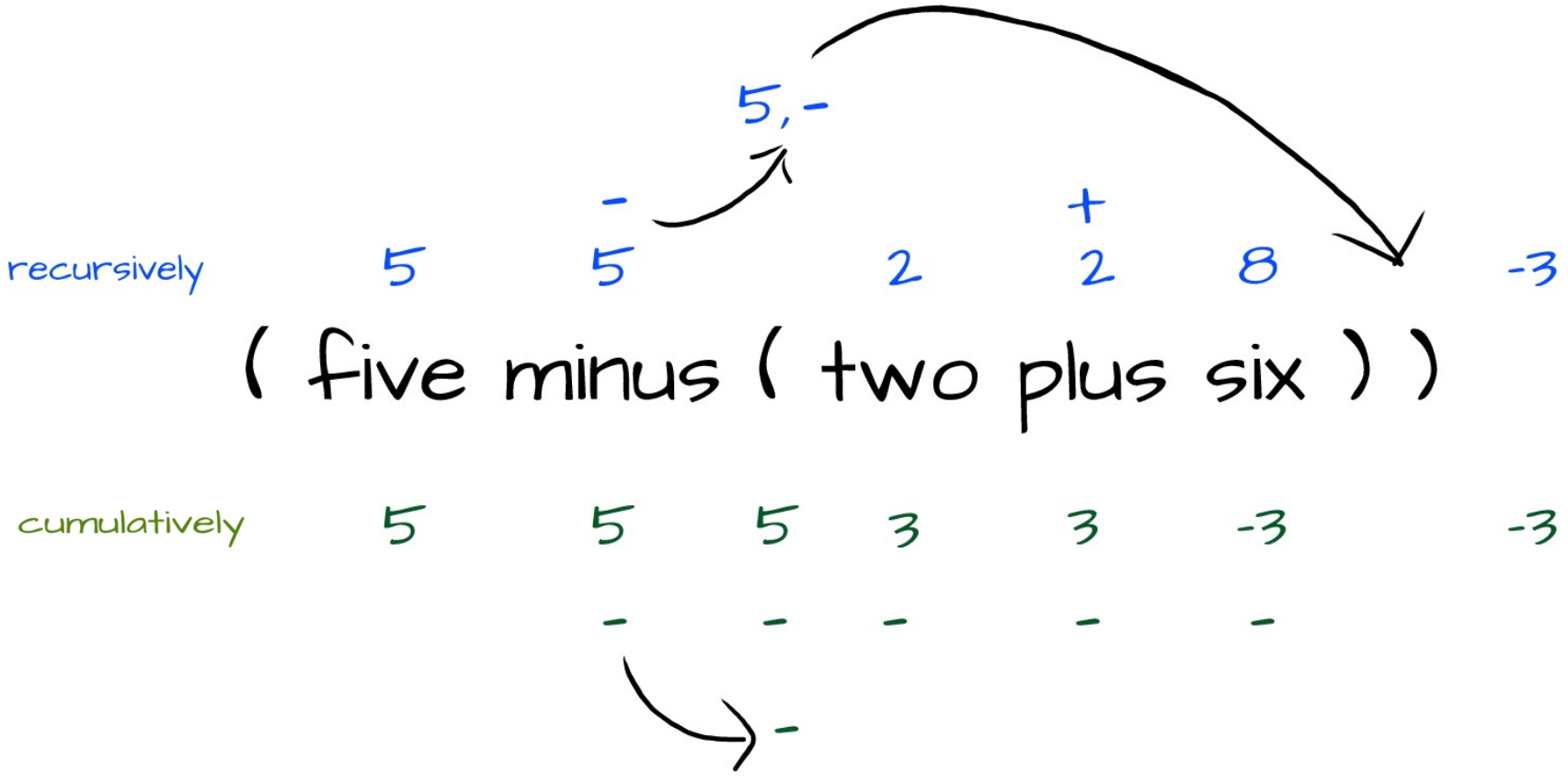
Symbolic solutions



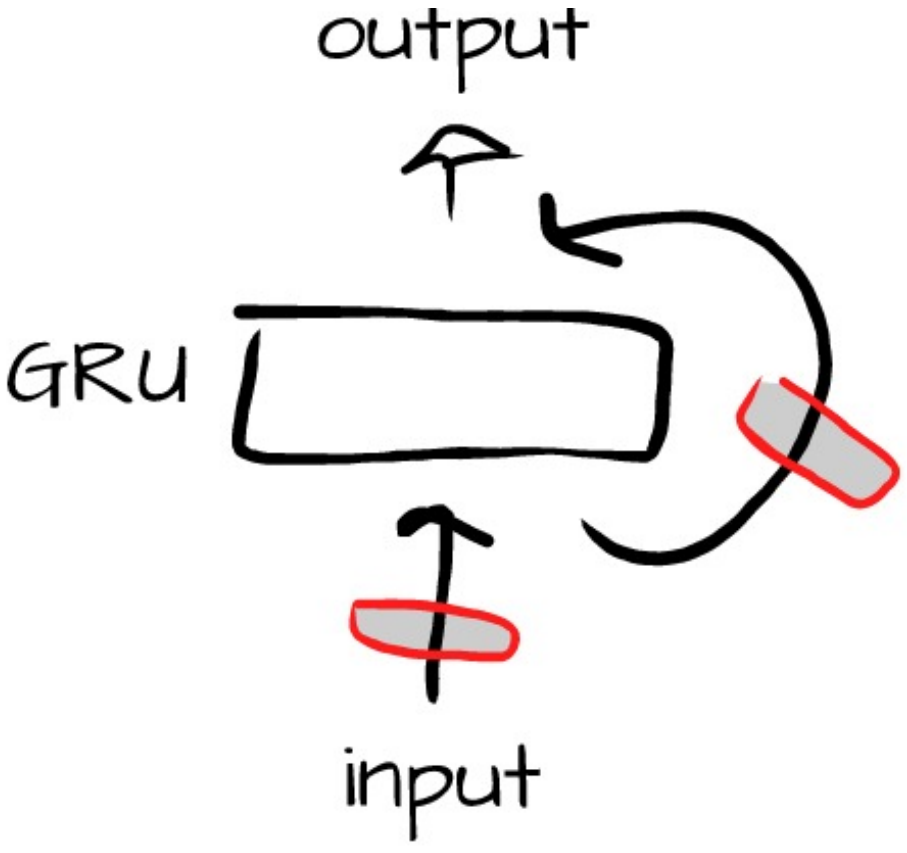
Symbolic solutions



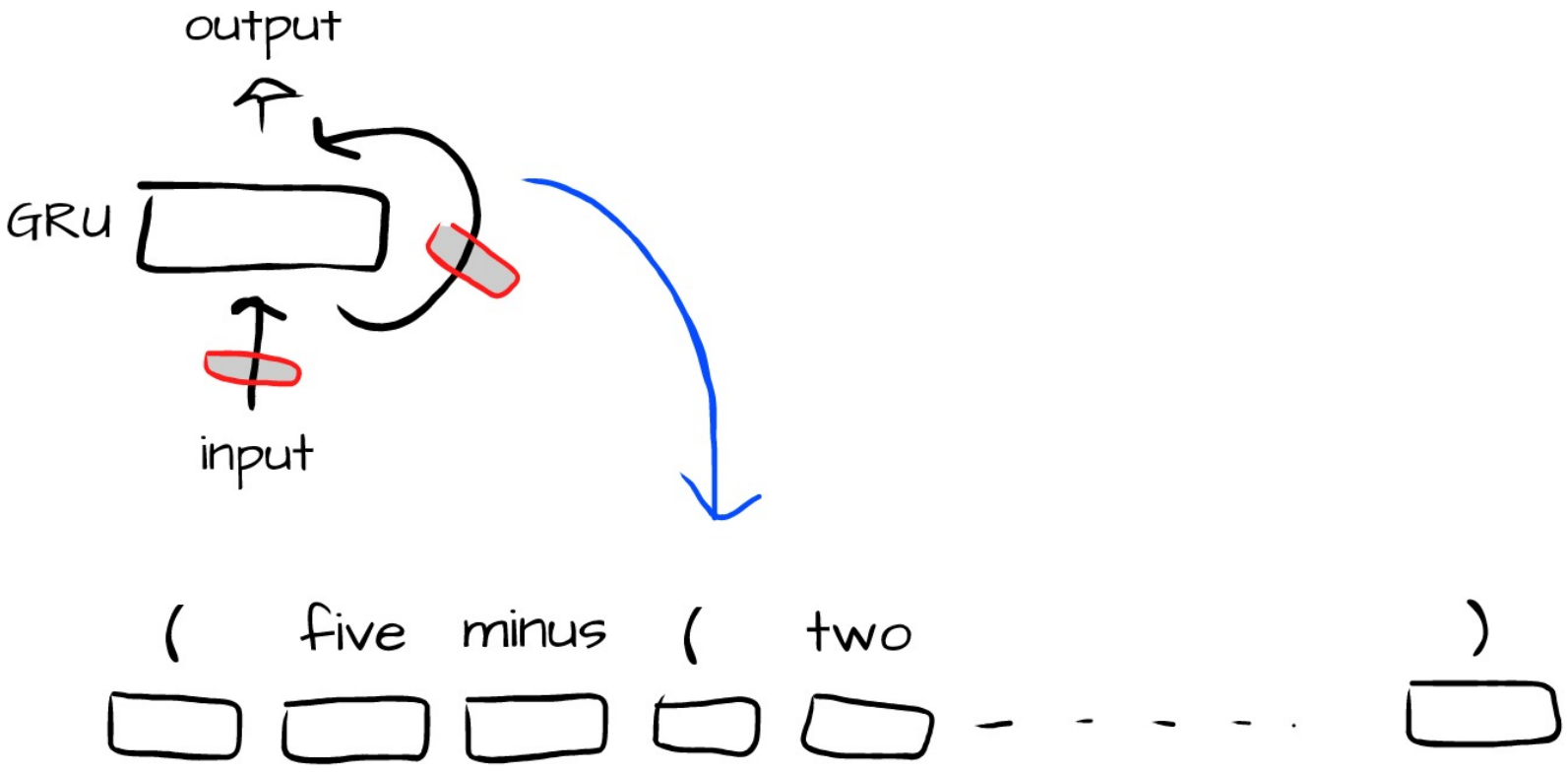
Symbolic solutions



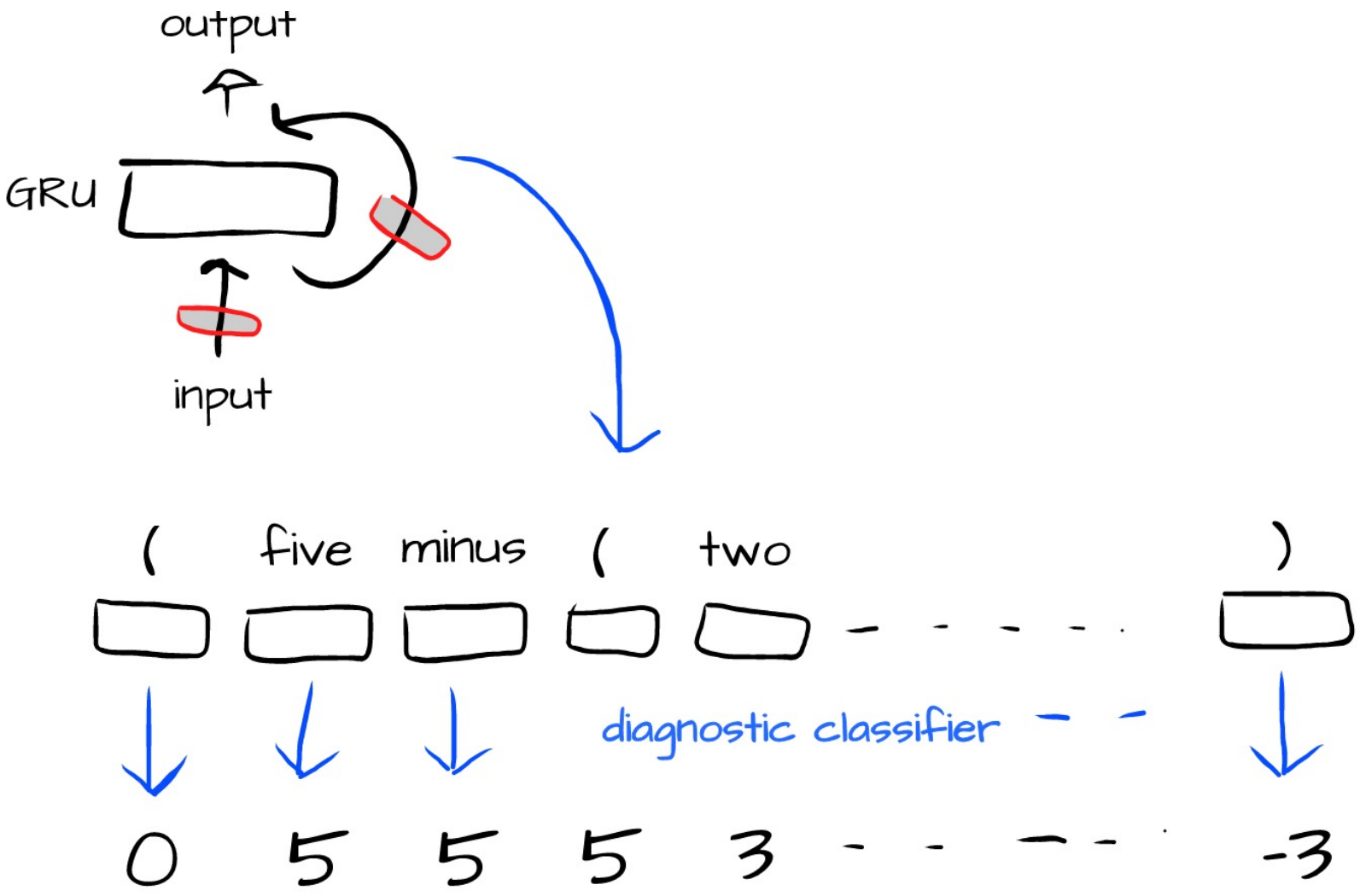
Diagnostic Classifier



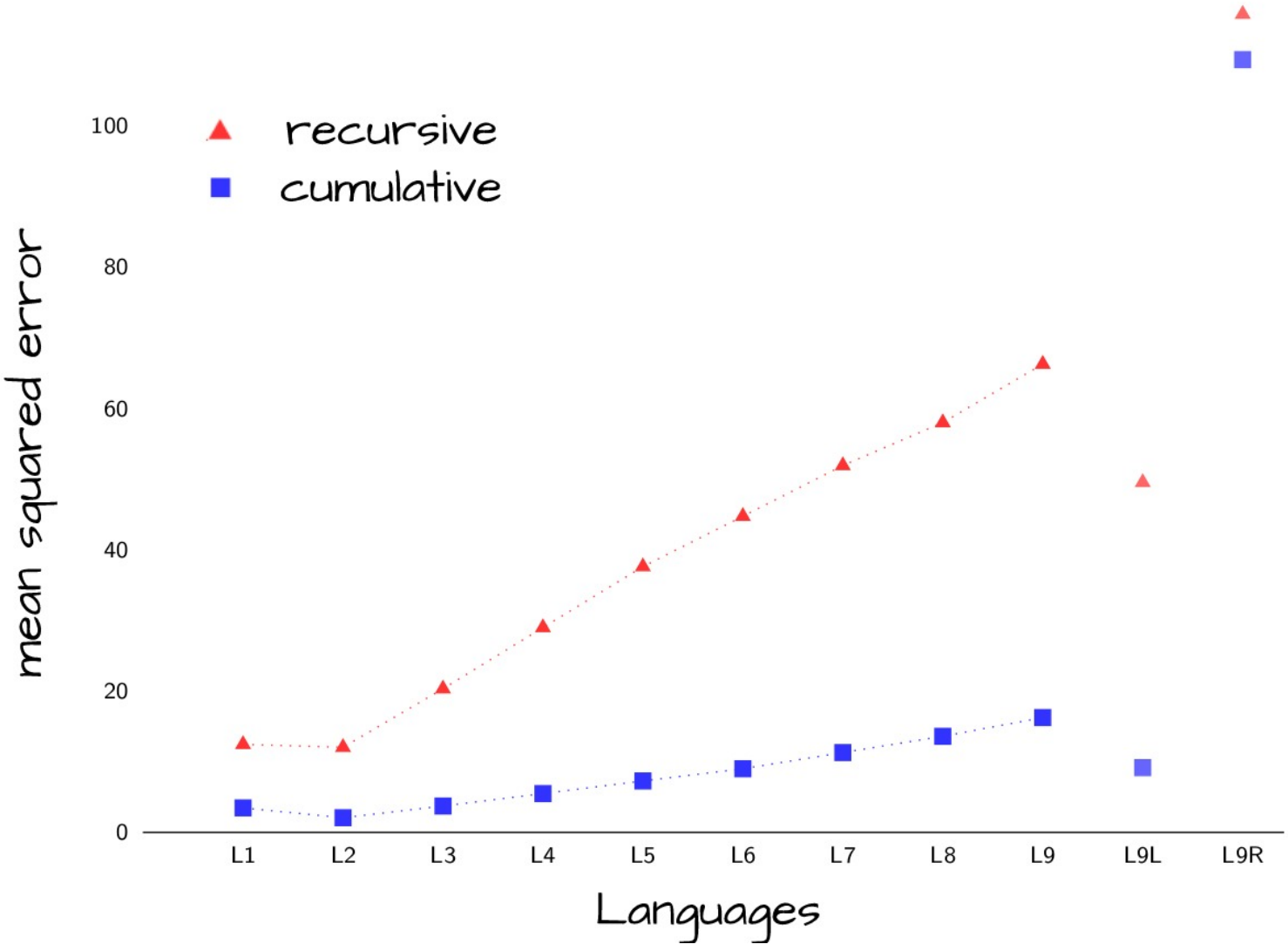
Diagnostic Classifier



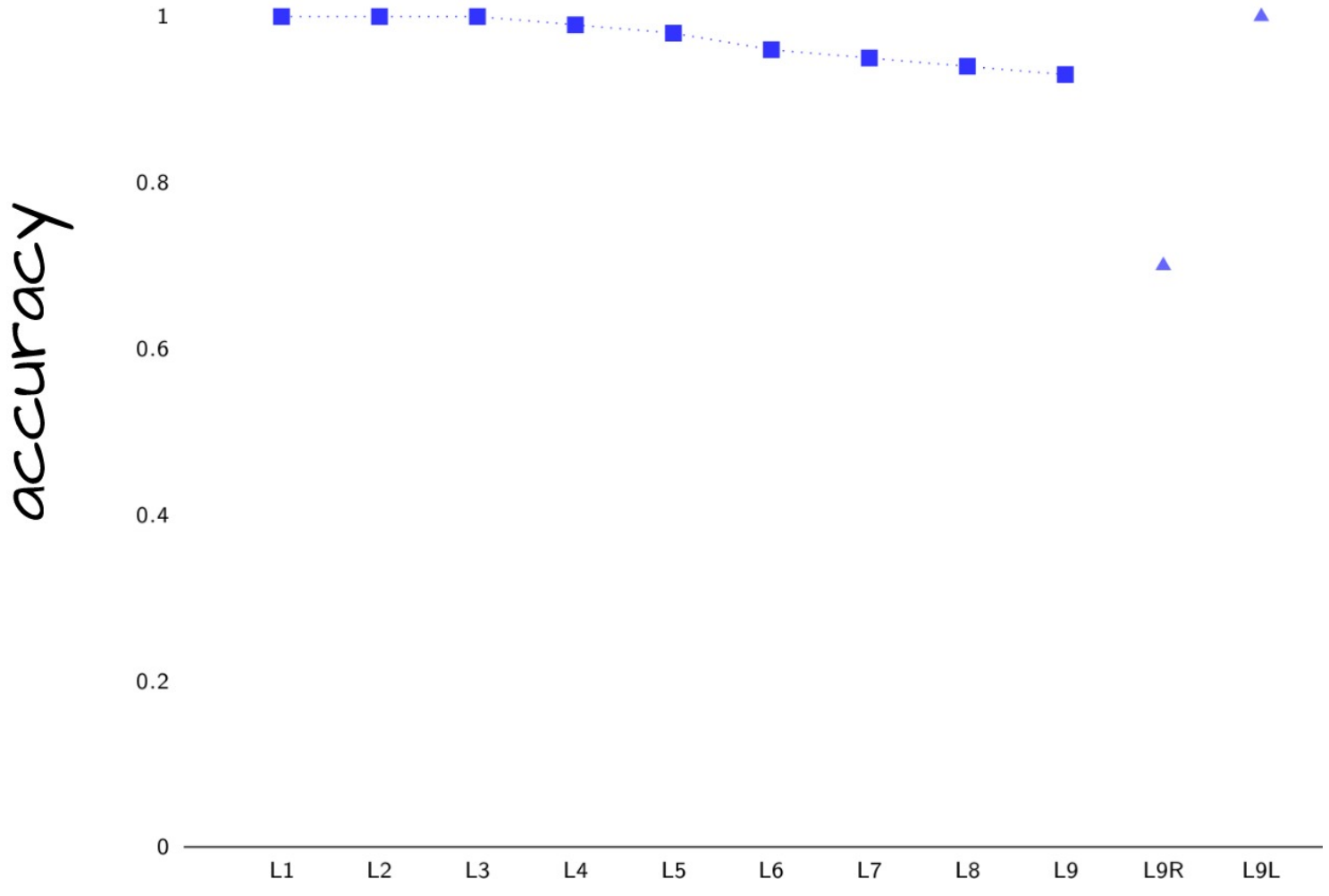
Diagnostic Classifier



Intermediate results



Cumulative strategy, operation mode



Discussion

Some intermediate conclusions:

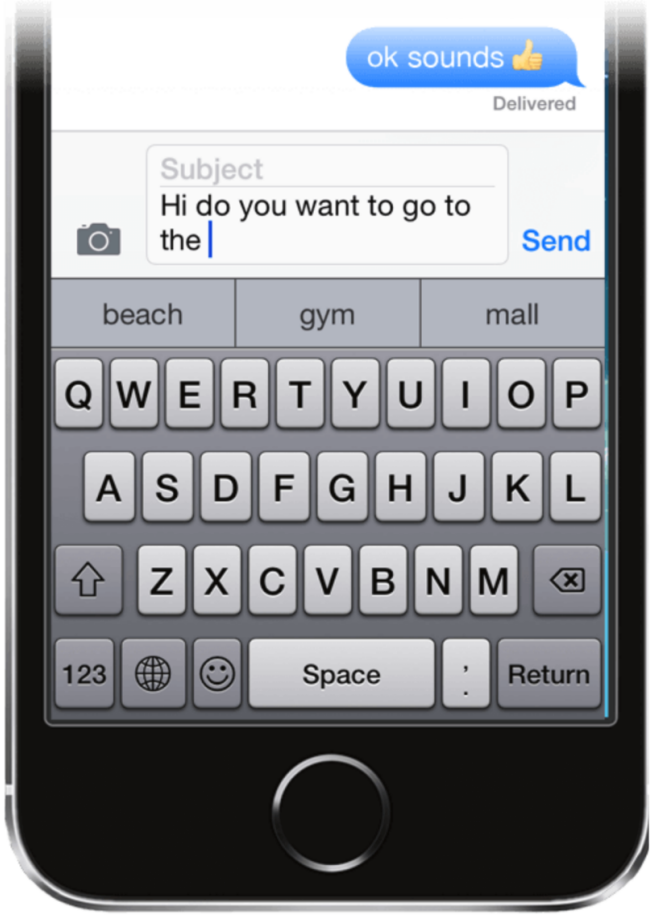
- GRU models seem fairly able to compute the meaning of sequences with hierarchical structure
- With diagnostic classification we can narrow down which strategy they are following

Discussion

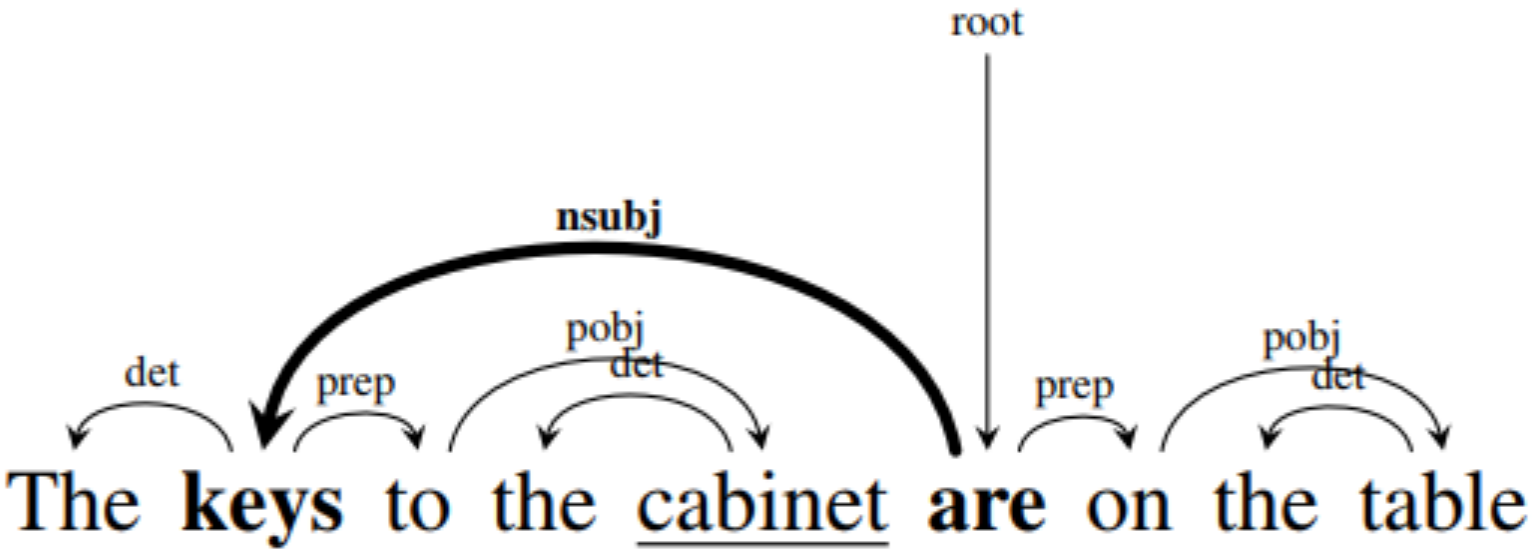
Some other possibilities:

- Further fine-grained analysis of the strategy models are using, and comparison with other recurrent cells (Hupkes et al., 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)

Language Modelling

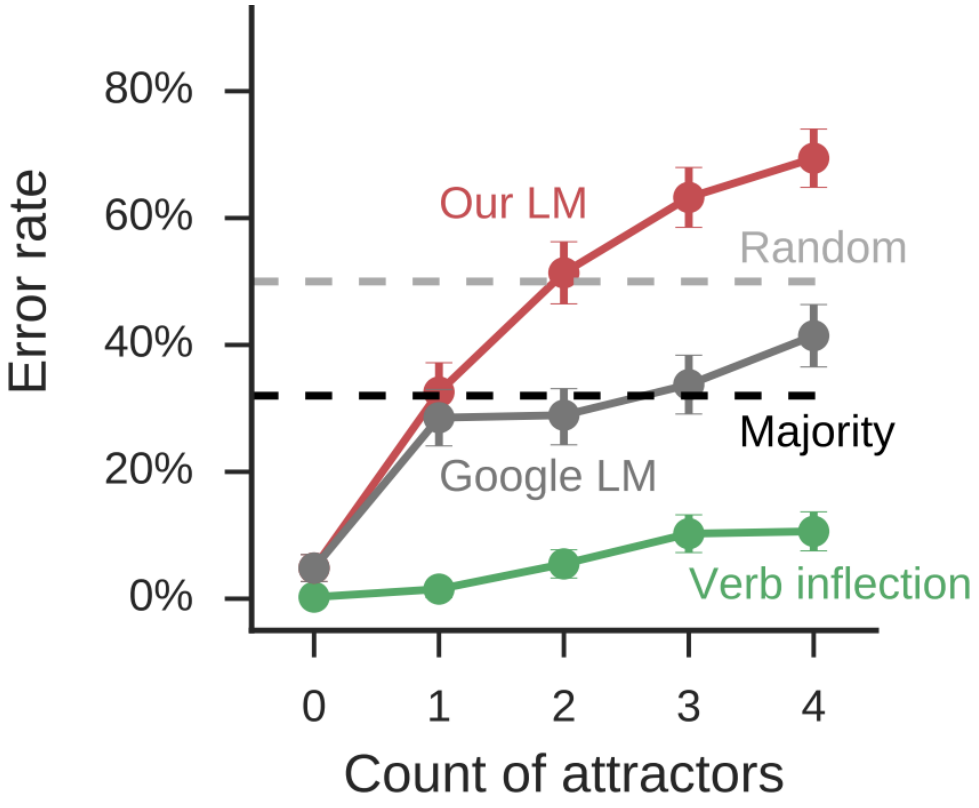


Subject-Verb Agreement



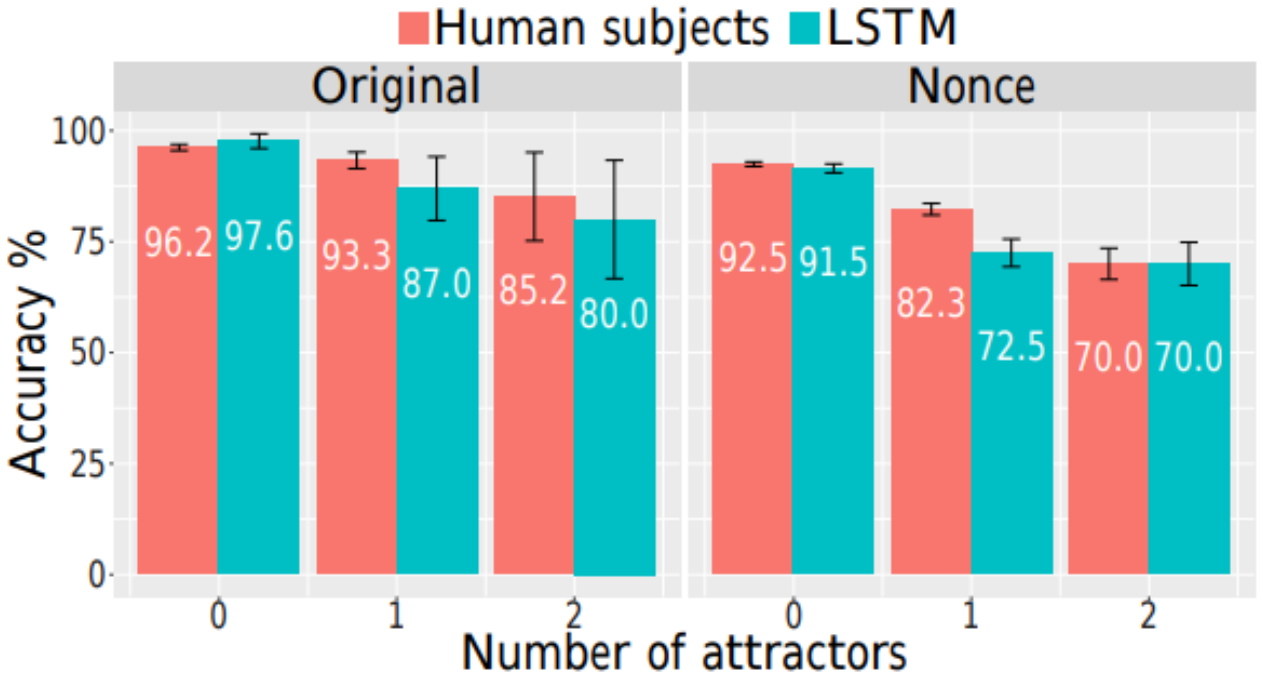
Linzen et al. (2016)

Results



Linzen et al. (2016)

Results 2



Gulordava et al. (2018)

Other linguistic questions

- Negative polarity items (Jumelet and Hupkes, 2018)

Other linguistic questions

- Negative polarity items (Jumelet and Hupkes, 2018)
- Filler-gap dependencies (Wilcox et al., 2018)

Other linguistic questions

- Negative polarity items (Jumelet and Hupkes, 2018)
- Filler-gap dependencies (Wilcox et al., 2018)
- Reflexive anaphora (Marvin and Linzen, 2018; Futrell et al., 2018)

Other linguistic questions

- Negative polarity items (Jumelet and Hupkes, 2018)
- Filler-gap dependencies (Wilcox et al., 2018)
- Reflexive anaphora (Marvin and Linzen, 2018; Futrell et al., 2018)
- And many more. . .

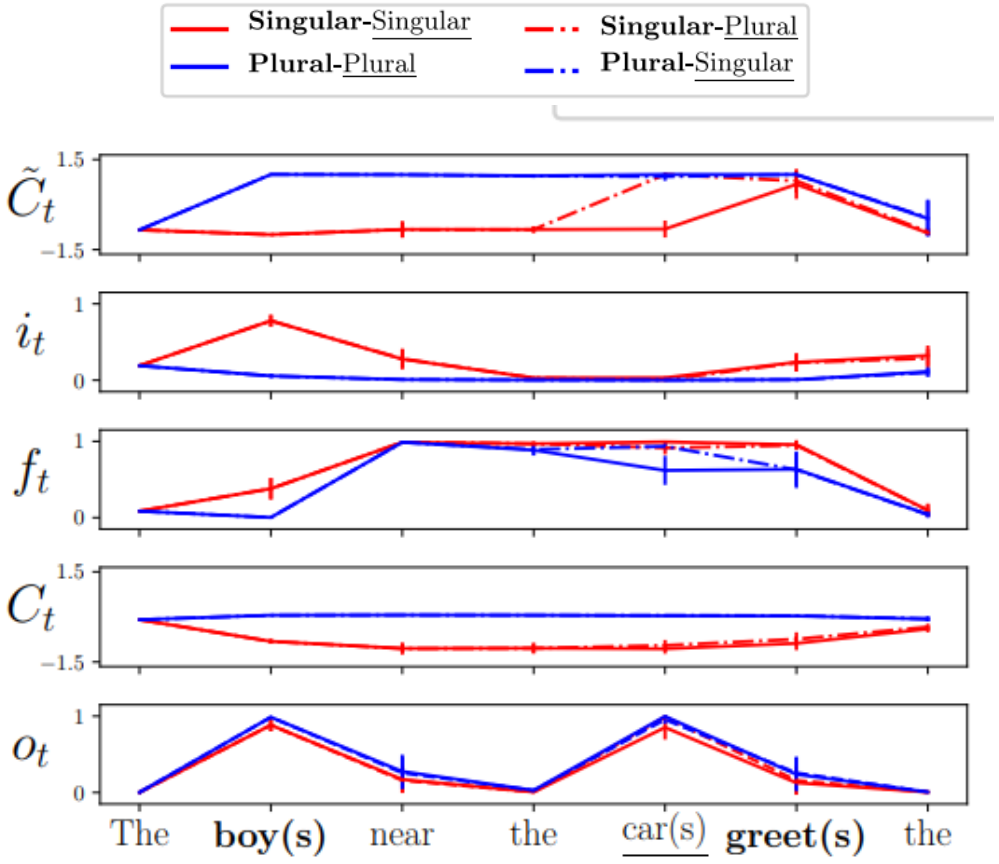
Other linguistic questions

- Negative polarity items (Jumelet and Hupkes, 2018)
- Filler-gap dependencies (Wilcox et al., 2018)
- Reflexive anaphora (Marvin and Linzen, 2018; Futrell et al., 2018)
- And many more...

But *how* do they do this?

But how do they do this?

Ablation studies

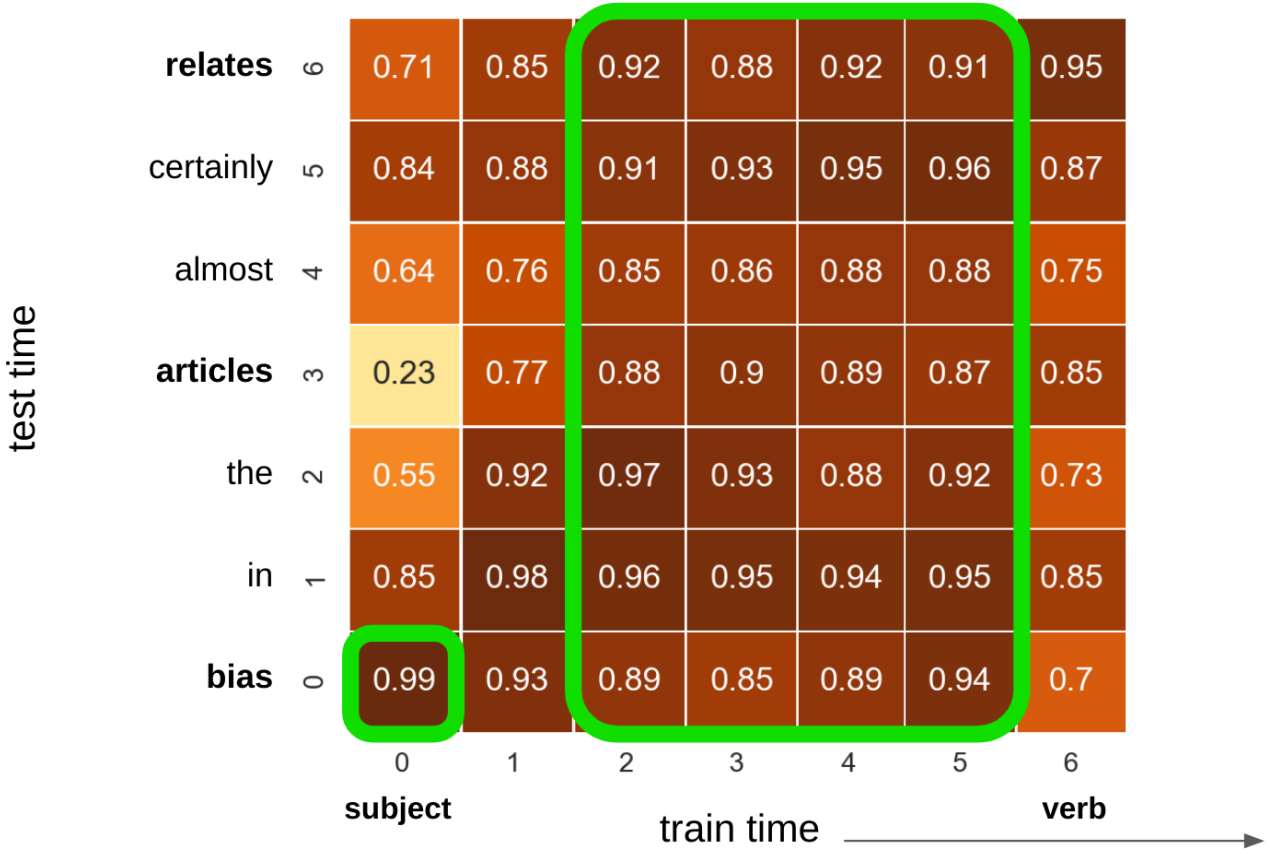


(a) 988 (singular)

Lakretz et al. (2019)

But how do they do this?

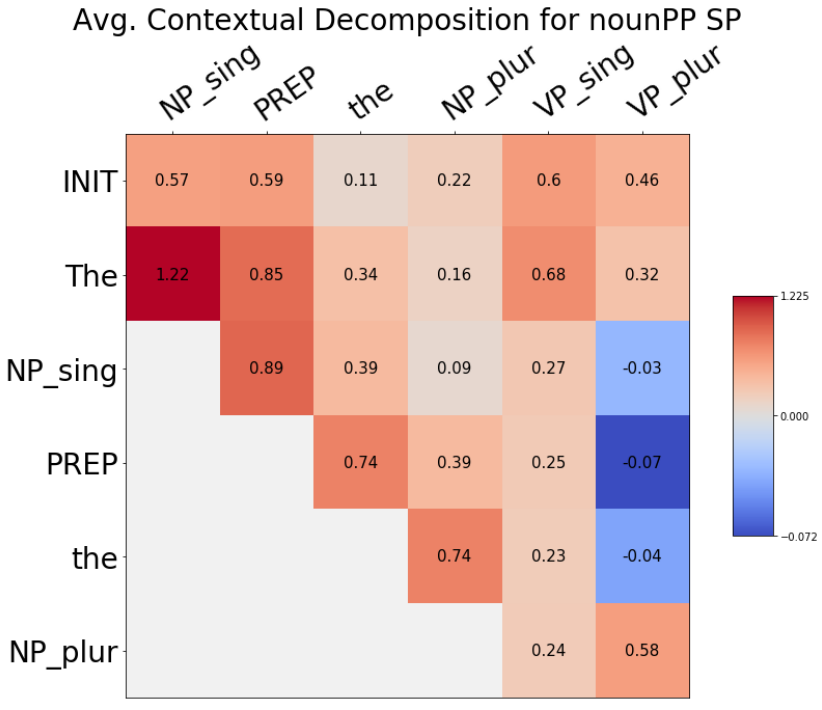
Temporal generalisation matrix

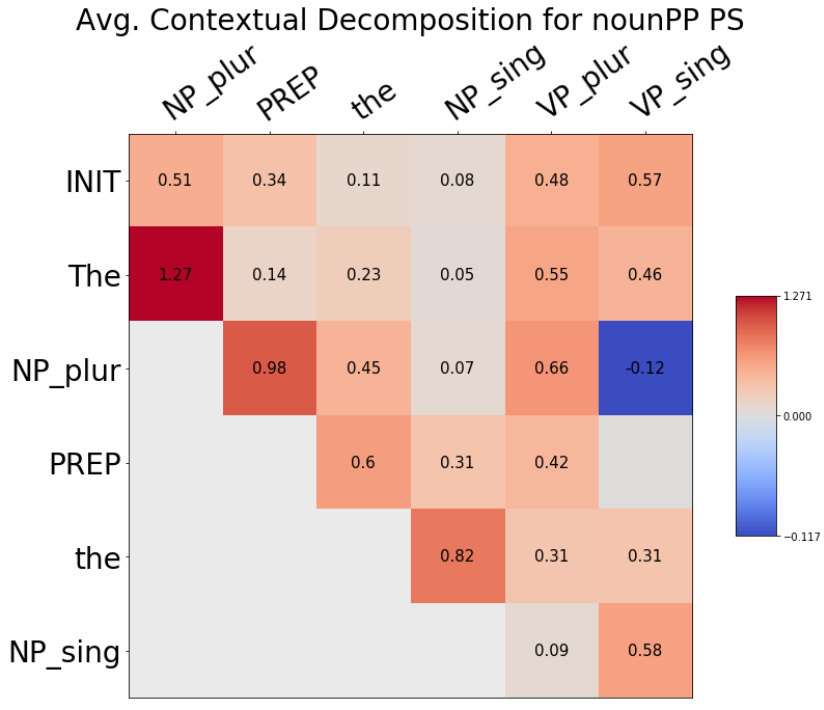


Giulianelli et al. (2018)

But how do they do this?

Contextual Decomposition





(Ongoing work with Jaap Jumelet)

June Projects

- Grammar in use: analysing emergent languages in referential games
- The Syntactic Awareness of Transformer Language Models
- Exploring Language Understanding with Modern Neural Architecture Search Methods
- Irregular world for regular language

References I

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*, 2018. URL `http://arxiv.org/abs/1809.01329`.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 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*, pages 240–248, 2018. doi: arXiv:1808.08079v1. URL `http://arxiv.org/abs/1808.08079`.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*, volume 1, pages 1195–1205, 2018.

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.

References II

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61: 907–926, 2018.

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*, pages 222–231, 2018.

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

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in LSTM language models. *arXiv preprint arXiv:1903.07435*, 2019. URL `http://arxiv.org/abs/1903.07435`.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. ISSN 2307-387X. URL `https://transacl.org/ojs/index.php/tacl/article/view/972`.

What Do you Learn From Context? Probing For Sentence Structure In Contextualized Word Representations - ICLR 2019

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das and Ellie Pavlick

Presented By
Karan Malhotra

Polysemy

“It’s a **good** ten miles to the next gas station.”

Motivation

- To understand **where** contextual representations improve over conventional embeddings.
- Is this information primarily syntactic in nature, or do the representations also encode higher-level semantic relationships? Is this information local, or do the encoders also capture long-range structure?
- **What** do contextual representations encode that conventional word embeddings do not?

Objective of the Paper

To probe **word-level contextual representations** from four recent models and investigate how they encode sentence structure across a range of syntactic, semantic, local, and long-range phenomena.

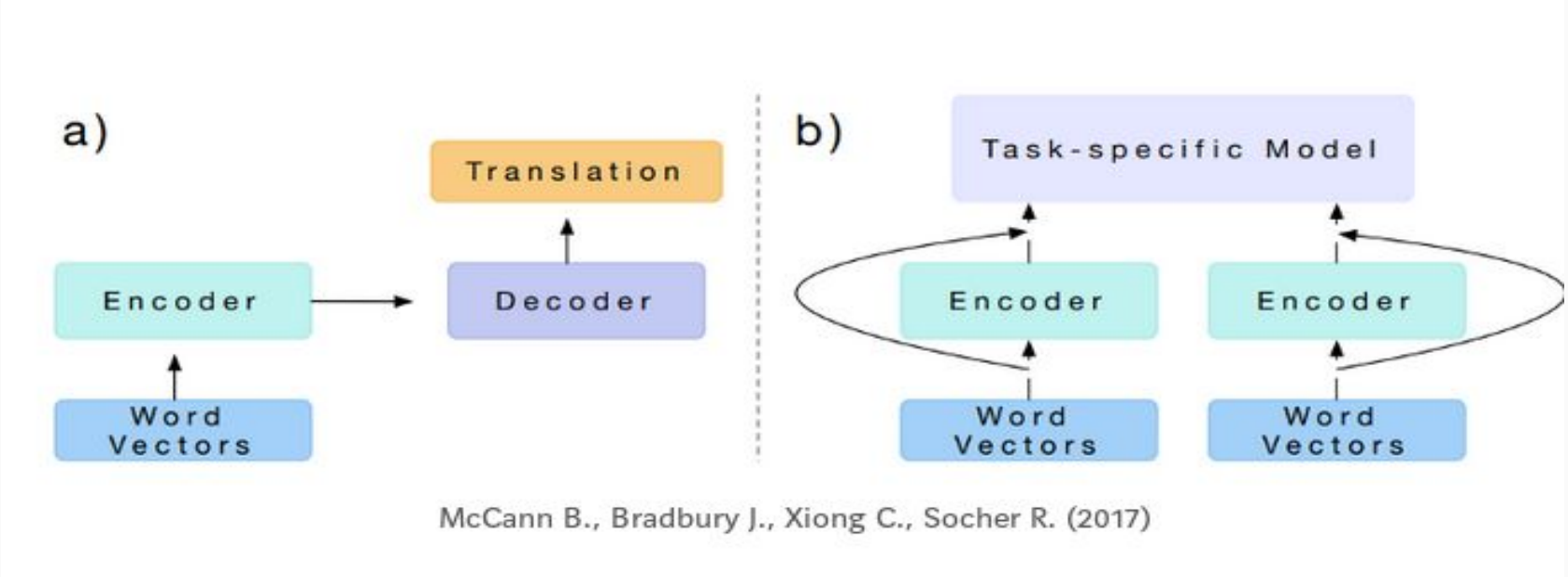
Models Probed : -

1. CoVe
2. ELMo
3. OpenAI GPT
4. BERT

A (very) Quick Recap on the Models

CoVe

- Leverages Machine Translation to build Contextualized Word Vectors (CoVe).



McCann B., Bradbury J., Xiong C., Socher R. (2017)

CoVe

$$\text{CoVe}(w) = \text{MT-LSTM}(\text{GloVe}(w))$$

w: Word
GloVe(w): Word Vectors after applying GloVe Word Embeddings
MT-LSTM(GloVe(w)): The vector learned from MT-LSTM architecture.

McCann B., Bradbury J., Xiong C., Socher R. (2017)

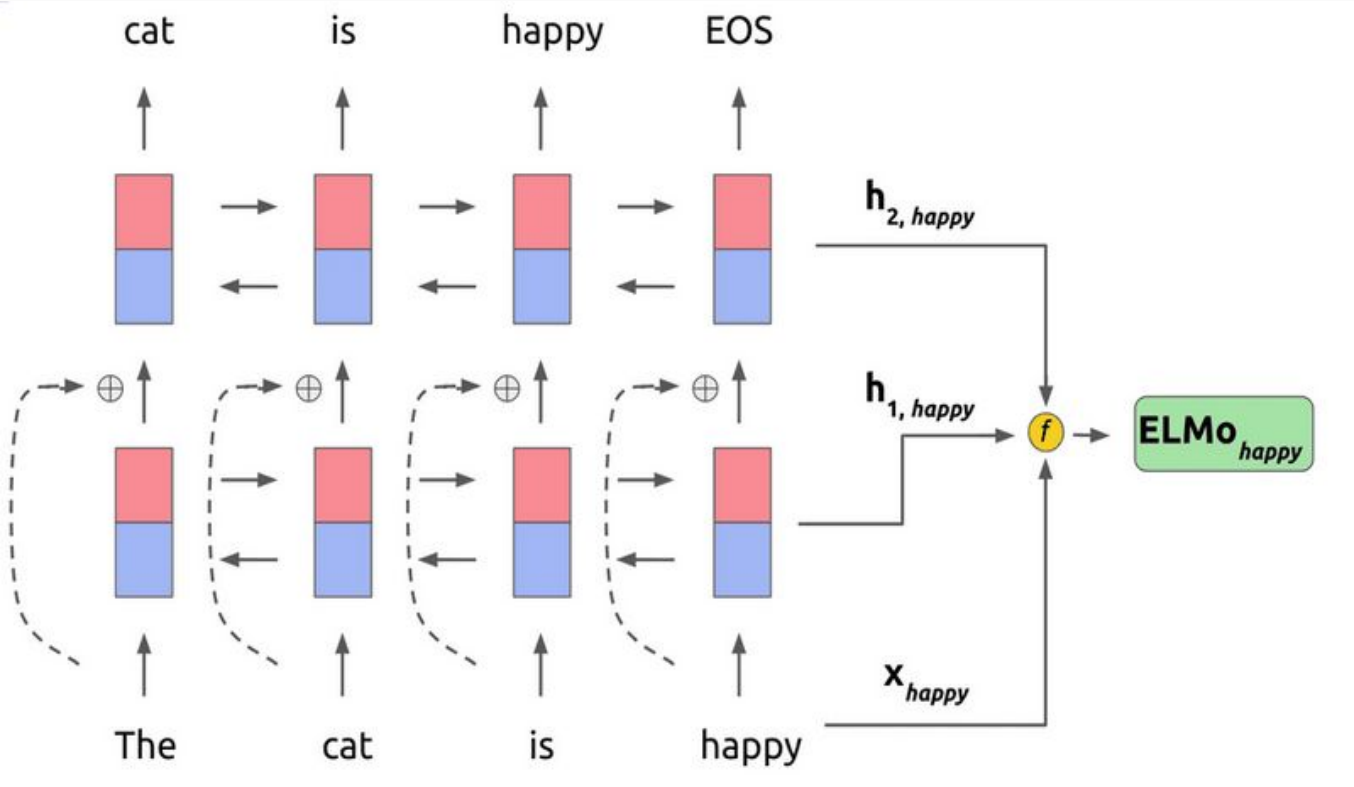
$$\tilde{w} = [\text{GloVe}(w); \text{CoVe}(w)]$$

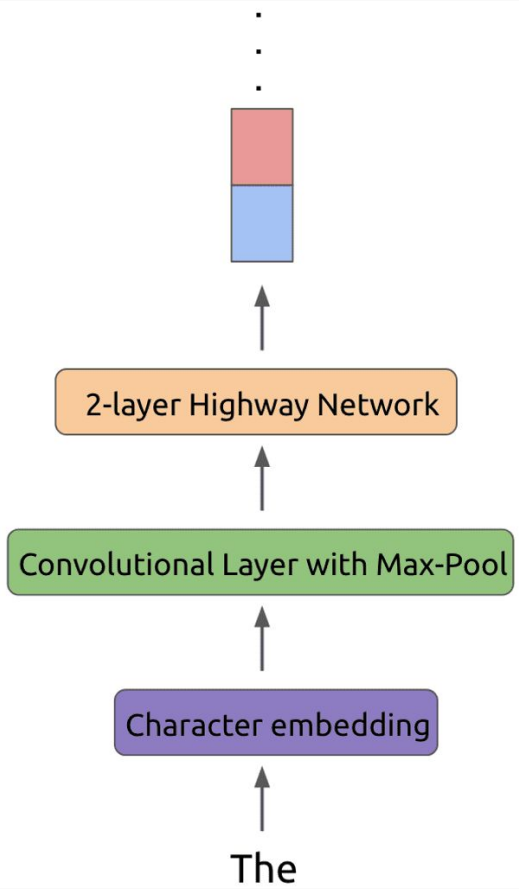
McCann B., Bradbury J., Xiong C., Socher R. (2017)

ELMo (Embeddings from Language Model)

ELMo (Embeddings from Language Model)

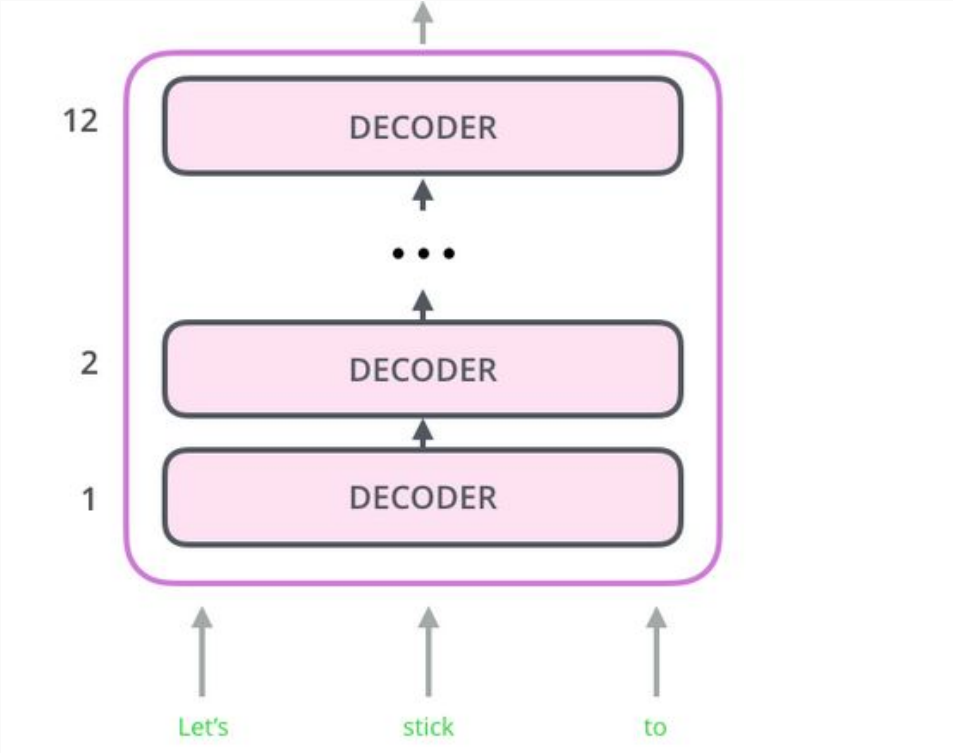
- Leverages a sophisticated Neural Language Model





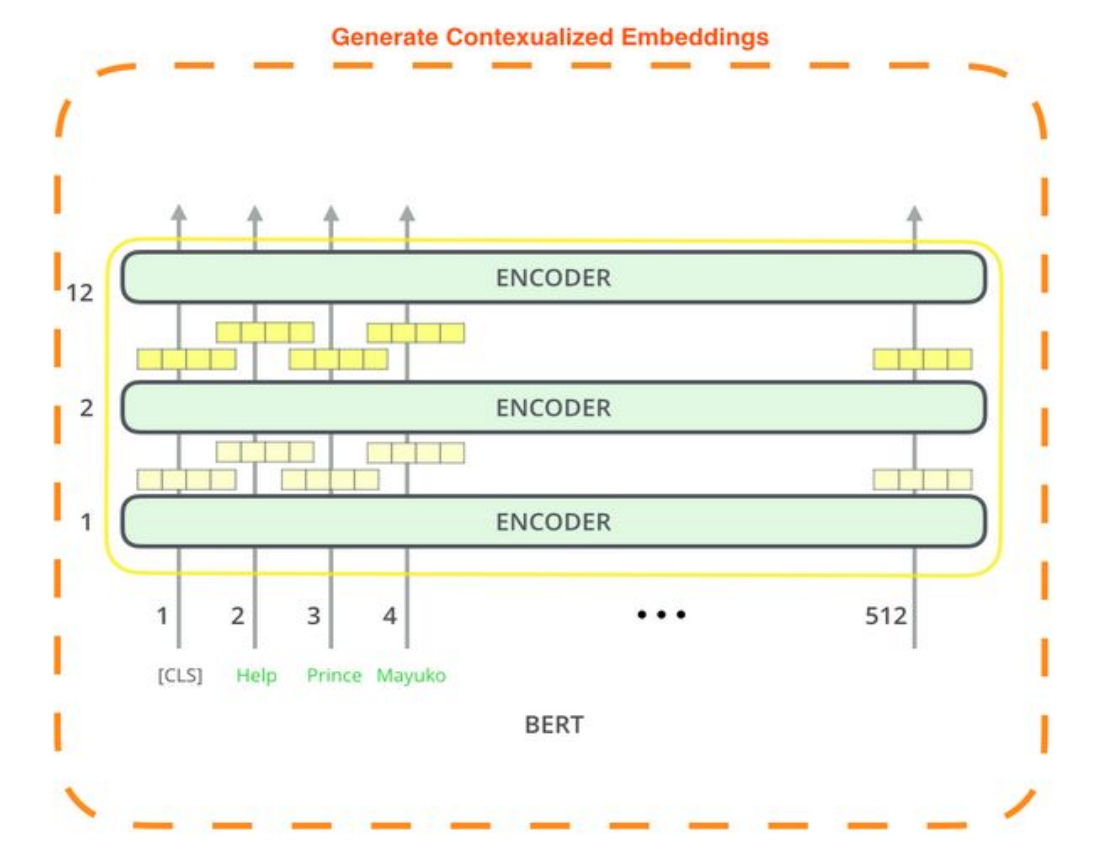
OpenAI GPT

- The model stacks twelve decoder layers (of transformer) and it is trained on Language Modelling task.



Note : The paper being presented mentioned ‘encoder’ instead of ‘decoder’.

BERT



- GPT + Bi-directional = BERT (very vague definition)

Back To The Paper

Tasks

Experiments are conducted on eight core NLP labeling tasks : -

- **Part-of-speech tagging** : Syntactic task of assigning tags such as noun, verb, adjective etc. to individual tokens.
- **Constituent labeling** : Task is to predict a label for a span of tokens within the phrase-structure parse of the sentence: e.g. noun phrase, verb phrase, etc.
- **Dependency labeling :** Dependency labeling seeks to predict the functional relationships of one token relative to another: a subject-object relationship, etc.
- **Named entity labeling :** Task of predicting the category of an entity referred to by a given span,e.g. does the entity refer to a person, a location, an organization, etc.
- **Semantic role labeling (SRL)** : Given a predicate and argument-pair, the task is to predict the role that argument fills. For e.g. given a sentence like “Mary pushed John”, SRL is concerned with identifying “Mary” as the pusher and “John” as the pushee.
- **Coreference :** The task of determining whether two spans of tokens refer to the same entity (or event) i.e pronoun resolution.
- **Semantic proto-role (SPR) :** It is the task of annotating fine-grained, non-exclusive semantic attributes, such as change of state or awareness, over predicate-argument pairs.E.g. given the sentence “Mary pushed John”, SPR is concerned with identifying attributes such as awareness (whether the pusher is aware that they are doing the pushing).
- **Relation Classification (Rel.):** It is the task of predicting the real-world relation that holds between two entities . Eg - “Mary is walking to work”. Relationship between “Mary” and “Work” : Entity-Destination.

Datasets

1. **OntoNotes 5.0 corpus**
   - POS Tagging
   - Constituent Labelling
   - Named entity labeling
   - Semantic role labeling (SRL)
   - Coreference

2. **English Web Treebank portion of Universal Dependencies**
   - Dependency labeling

3. **SPR1** (derived from Penn Treebank) and **SPR2** (derived from English Web Treebank)

4. **Semeval 2010 Task 8 dataset**
   - Relation Classification

5. **Definite Pronoun Resolution dataset**
   - A **challenge** coreference dataset based on “Winograd schema”. Requires subtle semantic inference to resolve correctly.

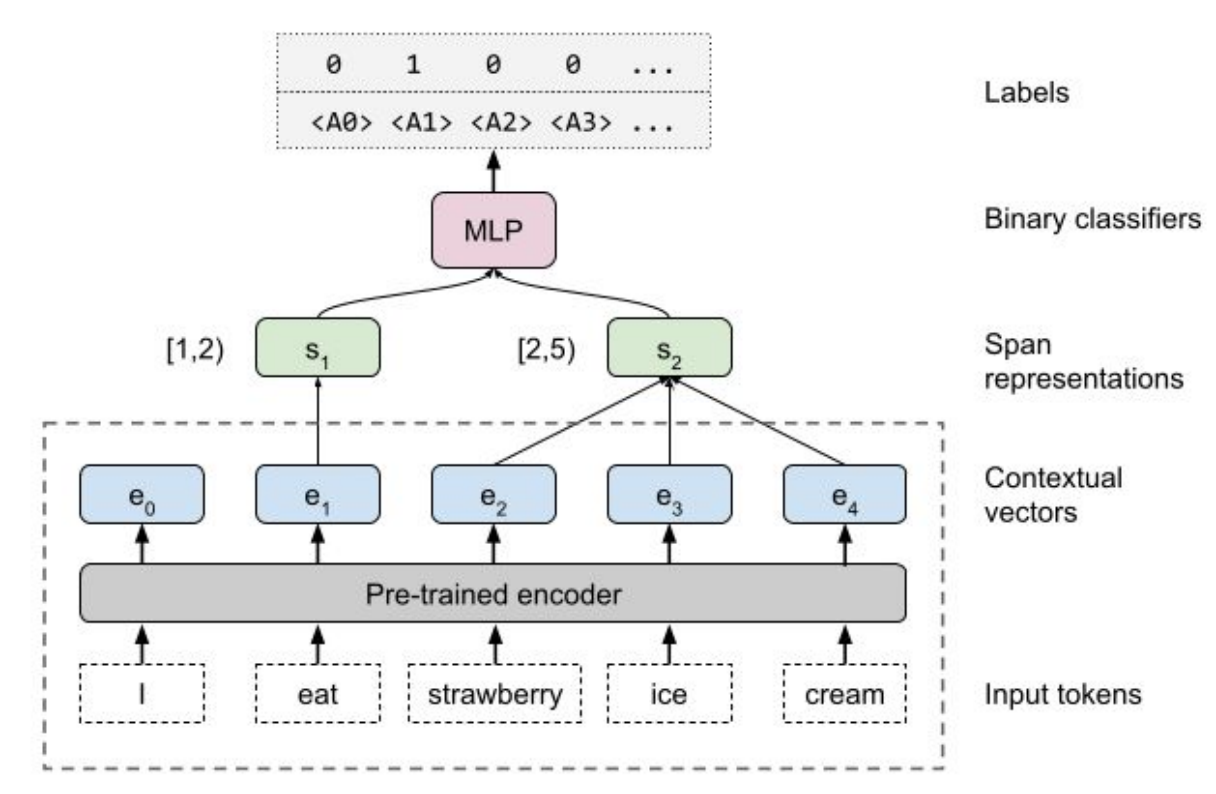
Examples From Paper

POS	The important thing about Disney is that it is a global $[\text{brand}]_1$. $\rightarrow$ NN (Noun)
Constit.	The important thing about Disney is that it $[\text{is a global brand}]_1$. $\rightarrow$ VP (Verb Phrase)
Depend.	$[\text{Atmosphere}]_1$ is always $[\text{fun}]_2$ $\rightarrow$ nsubj (nominal subject)
Entities	The important thing about $[\text{Disney}]_1$ is that it is a global brand. $\rightarrow$ Organization
SRL	$[\text{The important thing about Disney}]_2$ $[\text{is}]_1$ that it is a global brand. $\rightarrow$ Arg1 (Agent)
SPR	$[\text{It}]_1$ $[\text{endorsed}]_2$ the White House strategy... $\rightarrow$ {awareness, existed_after, ...}
Coref.$^{\text{O}}$	The important thing about $[\text{Disney}]_1$ is that $[\text{it}]_2$ is a global brand. $\rightarrow$ True
Coref.$^{\text{W}}$	$[\text{Characters}]_2$ entertain audiences because $[\text{they}]_1$ want people to be happy. $\rightarrow$ True Characters entertain $[\text{audiences}]_2$ because $[\text{they}]_1$ want people to be happy. $\rightarrow$ False
Rel.	The $[\text{burst}]_1$ has been caused by water hammer $[\text{pressure}]_2$. $\rightarrow$ Cause-Effect(e_2, e_1)

Table 1: Example sentence, spans, and target label for each task. O = OntoNotes, W = Winograd.

Probing Model Architecture

Note : Model is trained to predict Multi-Label Target.



- Projection layer is used since span inputs have different dimensions based on the model being probed.
- Self Attention Pooling is used to compute s2.
- The only information model accesses about the rest of the sentence is provided by the contextualised embeddings within the given spans.
- Span representations are concatenated and fed into a two layer MLP for classification.

Experiments

Research Question : **What do contextual representations encode that conventional word embeddings do not?**

The experiments are designed to investigate how the models capture linguistic information.

- **Lexical Baselines** : The authors train a version of the probing model directly on the most closely context-independent word representations.

  CoVe - Glove Embeddings.
  ELMo - Activations of the character CNN layer.
  GPT and BERT - Subword embeddings.

  Factors out access to surrounding words.

- **Randomized ELMo :** All weights above lexical layer (layer 0) are replaced with random orthonormal matrices - To investigate the impact of architecture of ELMo.

Experiments

- Word-Level CNN

  - Lexical Baseline + fixed-width CNN layer.
  - Considers presence of nearby words.
  - Comparison with word level CNN indicates contribution of long-range context to performance of encoder.

Before I show you Results

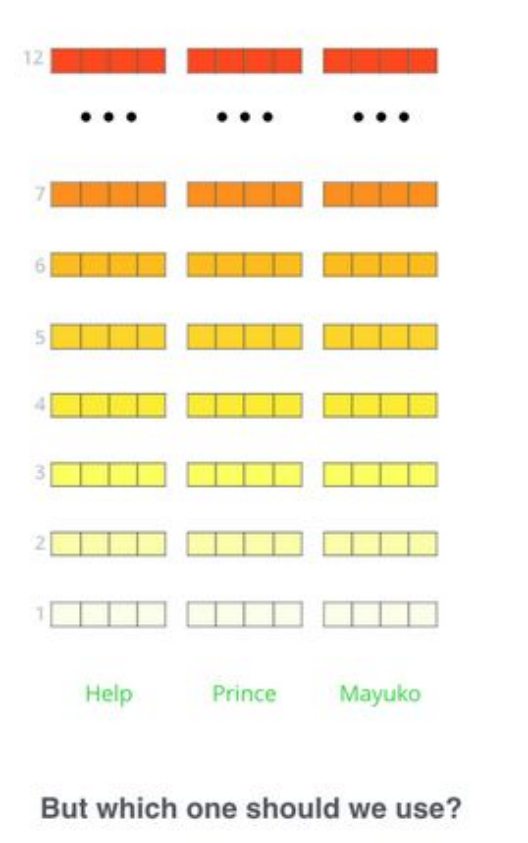
Metric Used : Binary F1 score (Harmonic mean of precision and recall).

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Q) Which layer's activation to use as contextual embedding for BERT and GPT?

The paper uses two methods : -

1. **Cat -** The activations of last layer are concatenated with the subword embeddings.
2. **Mix -** Linear combination of layer activations (including embedding).



Results

	CoVe			ELMo			GPT		
	Lex.	Full	Abs. Δ	Lex.	Full	Abs. Δ	Lex.	cat	mix
Part-of-Speech	85.7	94.0	8.4	90.4	**96.7**	6.3	88.2	94.9	95.0
Constituents	56.1	81.6	25.4	69.1	**84.6**	15.4	65.1	81.3	**84.6**
Dependencies	75.0	83.6	8.6	80.4	**93.9**	13.6	77.7	92.1	**94.1**
Entities	88.4	90.3	1.9	92.0	**95.6**	3.5	88.6	92.9	92.5
SRL (all)	59.7	80.4	20.7	74.1	**90.1**	16.0	67.7	86.0	89.7
Core roles	*56.2*	*81.0*	*24.7*	*73.6*	***92.6***	*19.0*	*65.1*	*88.0*	*92.0*
Non-core roles	*67.7*	*78.8*	*11.1*	*75.4*	***84.1***	*8.8*	*73.9*	*81.3*	***84.1***
OntoNotes coref.	72.9	79.2	6.3	75.3	84.0	8.7	71.8	83.6	**86.3**
SPR1	73.7	77.1	3.4	80.1	**84.8**	4.7	79.2	83.5	83.1
SPR2	76.6	80.2	3.6	82.1	83.1	1.0	82.2	**83.8**	**83.5**
Winograd coref.	52.1	**54.3**	2.2	**54.3**	**53.5**	-0.8	51.7	52.6	**53.8**
Rel. (SemEval)	51.0	60.6	9.6	55.7	77.8	22.1	58.2	**81.3**	**81.0**
Macro Average	69.1	78.1	9.0	75.4	**84.4**	9.1	73.0	83.2	**84.4**

	BERT-base				BERT-large				
	F1 Score			Abs. Δ	F1 Score			Abs. Δ	
Lex.	cat	mix	ELMo	Lex.	cat	mix	(base)	ELMo
88.4	**97.0**	96.7	0.0	88.1	96.5	**96.9**	0.2	0.2
68.4	83.7	86.7	2.1	69.0	80.1	**87.0**	0.4	2.5
80.1	93.0	95.1	1.1	80.2	91.5	**95.4**	0.3	1.4
90.9	96.1	96.2	0.6	91.8	96.2	**96.5**	0.3	0.9
75.4	89.4	91.3	1.2	76.5	88.2	**92.3**	1.0	2.2
74.9	*91.4*	*93.6*	*1.0*	*76.3*	*89.9*	***94.6***	*1.0*	*2.0*
76.4	*84.7*	*85.9*	*1.8*	*76.9*	*84.1*	***86.9***	*1.0*	*2.8*
74.9	88.7	90.2	6.3	75.7	89.6	**91.4**	1.2	7.4
79.2	84.7	**86.1**	1.3	79.6	85.1	**85.8**	-0.3	1.0
81.7	83.0	**83.8**	0.7	81.6	83.2	**84.1**	0.3	1.0
54.3	53.6	54.9	1.4	53.0	53.8	**61.4**	6.5	7.8
57.4	78.3	82.0	4.2	56.2	77.6	**82.4**	0.5	4.6
75.1	84.8	86.3	1.9	75.2	84.2	**87.3**	1.0	2.9

Comparison of representation models and their respective lexical baselines. Numbers reported are micro-averaged F1 score on respective test sets. **Lex** denotes the lexical baseline for each model, and **bold** denotes the best performance on each task. Lines in italics are subsets of the targets from a parent task; these are omitted in the macro average. 95% confidence intervals (normal approximation) are approximately ±3 (±6 with BERT-large) for Winograd, ±1 for SPR1 and SPR2, and ±0.5 or smaller for all other tasks

Analysis

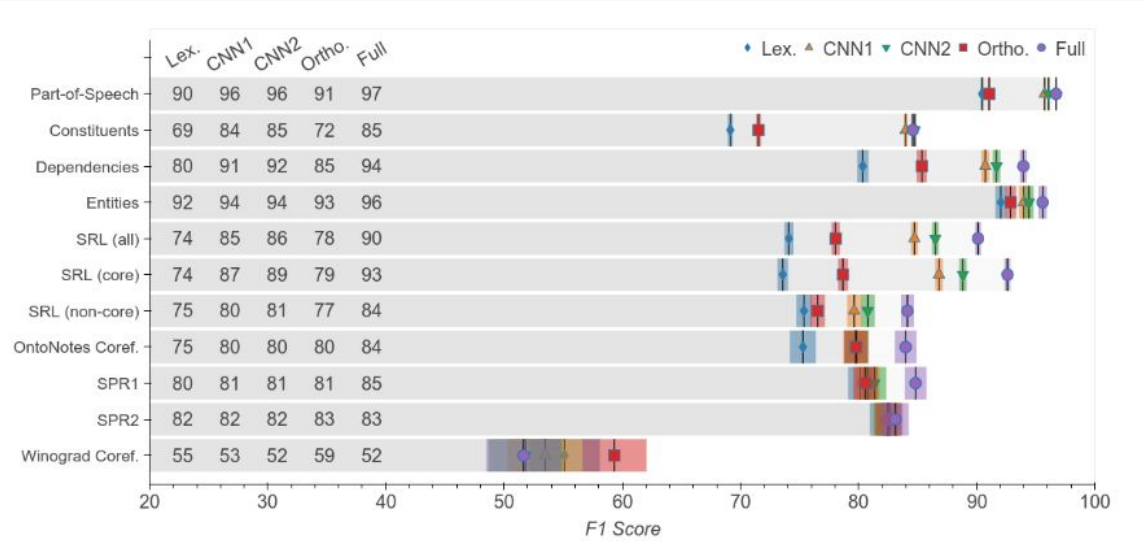
- ELMo and GPT (mix) have comparable performance with GPT higher on relation classification and OntoNotes coreference.
- As expected, both ELMo and GPT outperform CoVe except for Winograd coreference.
- By using character level CNN and subword embeddings in ELMo and GPT respectively, the models benefit by encoding morphological information.
- (mix) is better than (cat) - In agreement with Peters et al (2018) - most relevant information is contained in intermediate layers.
- BERT-base > GPT , BERT-Large > BERT-Base

Analysis (2)

Comparison with Lexical Baselines

- In brief, the authors try to convey that contextualized embeddings offer higher improvements on tasks which are related to syntax in comparison to semantics.
- Large gains on syntactic tasks such as dependency labelling and constituency labelling in comparison to semantic tasks such as SPR and Winograd coreference.
- Note : SRL(core) is an exception. The authors attribute the increase in performance to better labelling of core roles which are closely tied to syntax.
- Another exception: Relation Classification - Semantic tasks but shows high performance with contextual embedding. Authors attribute this to poor performance of lexical priors (embeddings) and presence of keywords like “caused” that suggest “cause-effect” relation and makes classification easy.
- SPR requires higher-level semantic properties, and the improvement is small.

Analysis (3)



Orthonormal ELMo = ELMo with random weights. Shows improvement from lexical baseline.

- However the learned weights account for over 70% of the improvements on full ELMo.

Q)How much information is carried over long distances?

- CNN1 (Kernel width 3) closes 72% (macro avg) of the gap between lexical and full ELMO.
- CNN2 (kernel width 5) closes 79% of the gap.
- On syntactic tasks such as constituent labelling, POS the performance of CNN2 is close enough with Full ELMo -> Local information is very relevant for syntactic tasks.
- On the contrary for semantic tasks, such as coreference, the gap is larger -> ELMo encodes long-range information.

Analysis (4)

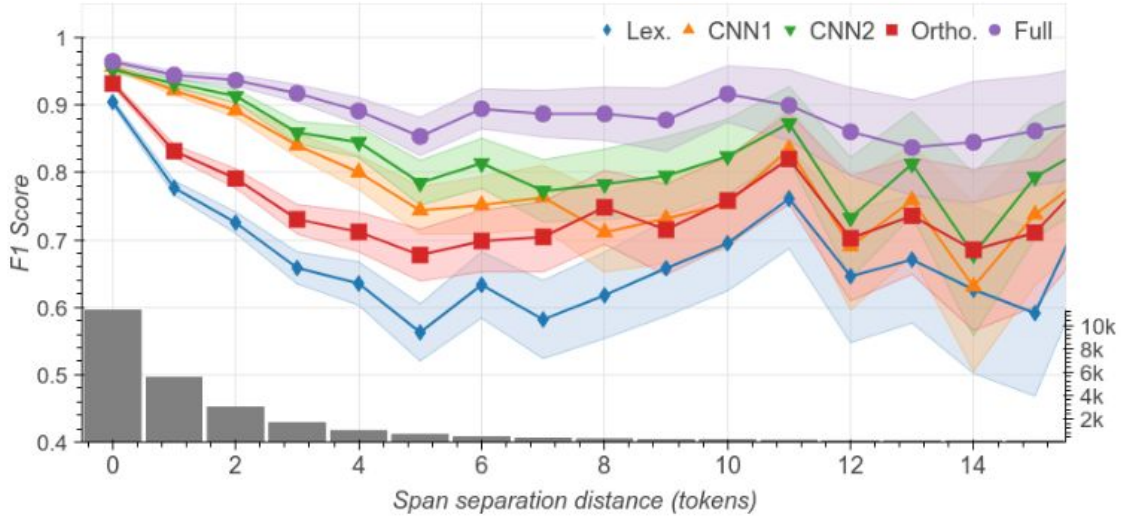


Figure 3: Dependency labeling F1 score as a function of separating distance between the two spans. Distance 0 denotes adjacent tokens. Colored bands are 95% confidence intervals (normal approximation). Bars on the bottom show the number of targets (in the development set) with that distance.

- ELMo indeed encodes long-range dependencies.

Conclusions

1. Contextualized embeddings encode syntax more than higher-level semantics.
2. Contextualized representation encode long-range information.

My Opinion

Pros

1. Well and concisely written paper with helpful appendix and analysis.
2. Builds on related token-probing work.
3. Vast set of tasks and experiments covering syntax, semantics and range of context.
4. Eradicates effect of training genre for GPT vs ELMO to make results comparable.

Cons

1. Authors should have talked about performance on Winograd coreference i.e Why CovE outperforms ELMo and GPT. Also why does ortho ELMo outperform full ELMo?

   - Comment by authors on openreview : https://openreview.net/forum?id=SJzSgnRcKX&noteId=HklxVExYpQ - “Dataset size is small and hence results on Winograd **are not significant**”

2. To prove that contextualized embeddings offer less improvement for semantic tasks, the authors could have added more semantic related tasks for eg - word sense disambiguation or metaphor processing.

Future Research

1. More semantic related tasks.
2. Further Investigations like removing top few encoders of BERT, changing dimensions of LSTMs in ELMo etc.
3. Visualizing the activations of network.
   Demo by OpenAI for Visual Data : https://distill.pub/2018/building-blocks/

Questions Are Welcome!

References III

Rebecca Marvin and Tal Linzen. Targeted Syntactic Evaluation of Language Models. In *EMNLP*, pages 1192–1202. Association for Computational Linguistics, 2018. URL `http://arxiv.org/abs/1808.09031`.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 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*, pages 211–221, 2018.