Diagnostic classification and symbolic guidance to understand and improve recurrent neural networks

Dieuwke Hupkes and Willem Zuidema
ILLC, University of Amsterdam

The task

To understand how **recurrent networks** can implement and learn deeply hierarchical **compositional structures**, we train them on a toy task that involves computing the meaning of expressions with a hierarchical compositional semantics.

Arithmetic language

(-two minus
((two plus eight) minus
(six plus one))

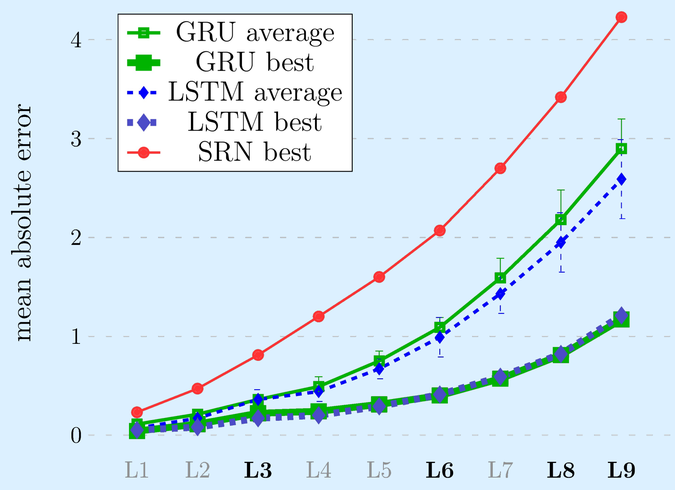
Sentences consist of digits and operators.
Brackets indicate compositional structure.

GRU
LSTM

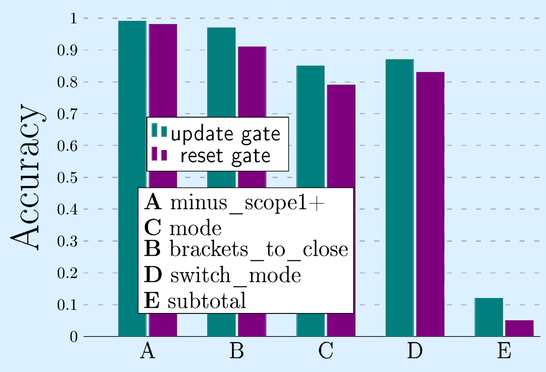
-5

The solution of the arithmetic expression defines its meaning.

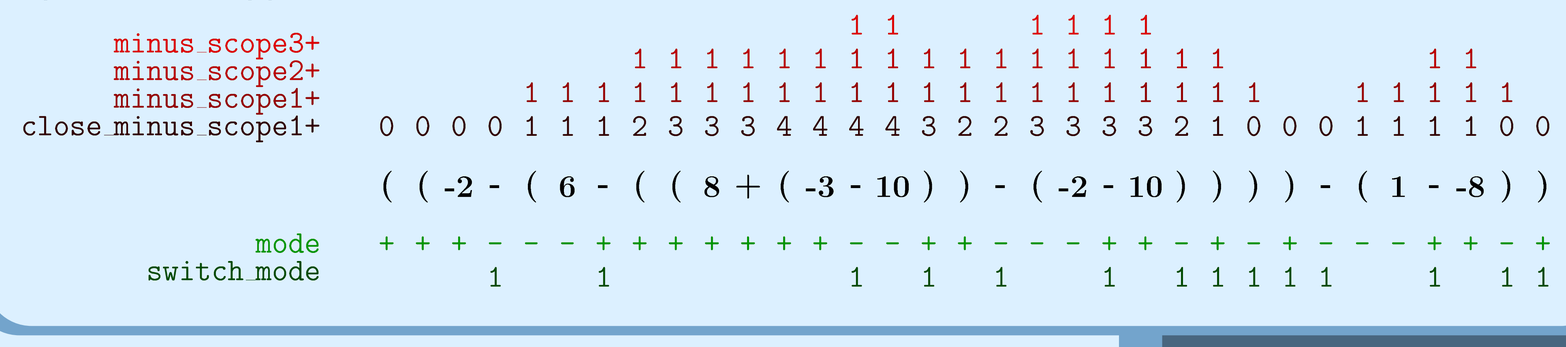
Network performance



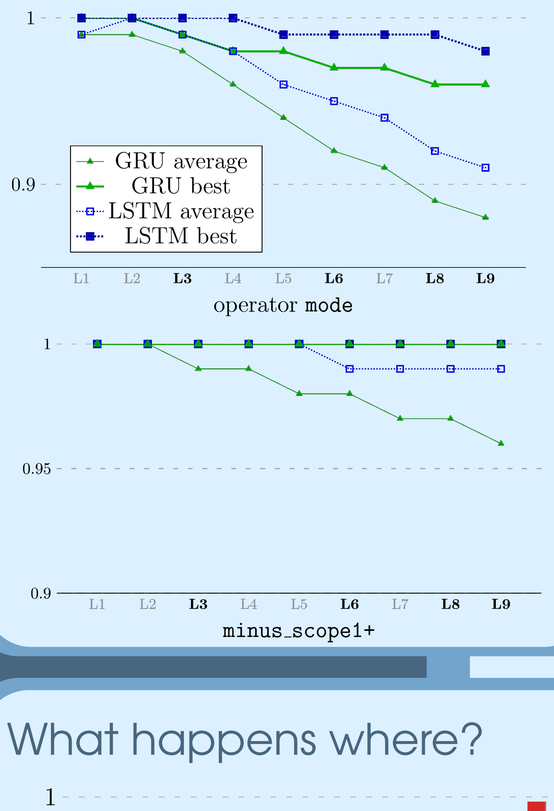
What do the gates do?



Symbolic hypotheses

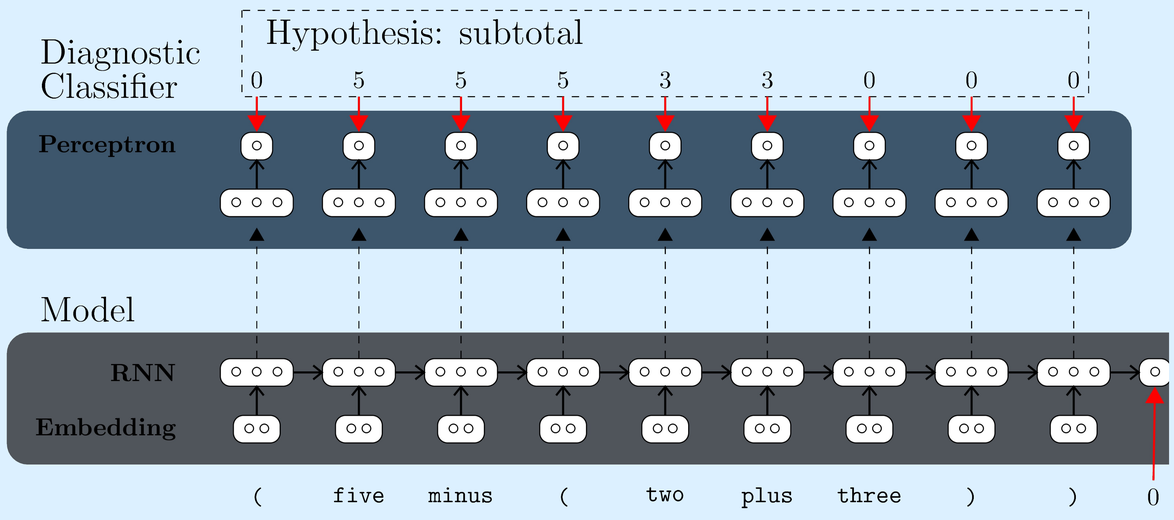


Subtract or add?

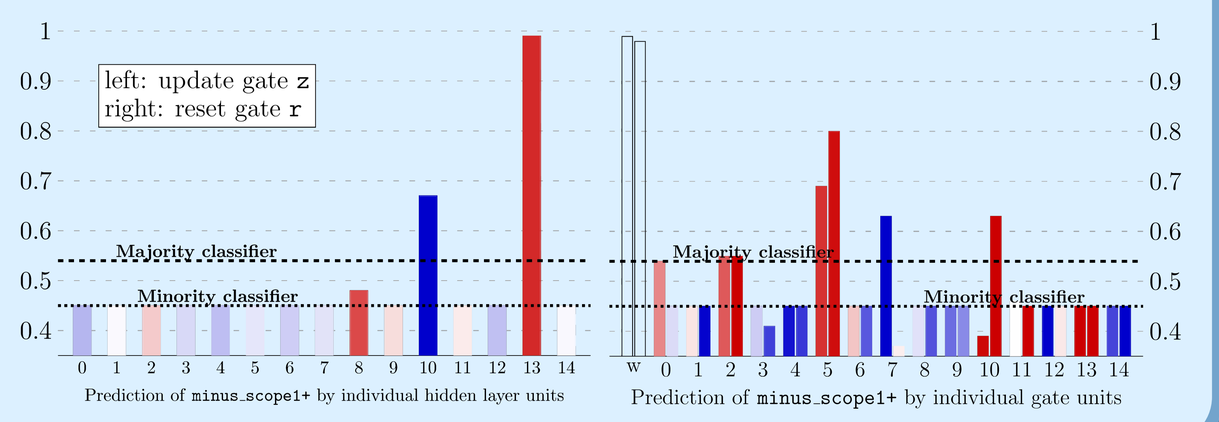


Diagnostic Classification

By training ***diagnostic classifiers*** - *simple* additional neural models - to predict *hypotheses* from the hidden state of trained networks, we investigate their internal dynamics.



What happens where?



Symbolic Guidance

Using the findings of ***diagnostic classification*** to provide additional symbolically motivated targets to the model at training time.