Diagnostic Classifiers: Revealing how Neural Networks process Hierarchical Structure

Sara Veldhoen, Dieuwke Hupkes, Willem Zuidema
ILLC, University of Amsterdam

We analyse how recursive and recurrent neural networks perform a task that involves hierarchical compositional semantics. The solution of the recursive network can be understood through visual inspection. The solution of the recurrent network by training diagnostic classifiers: models that predict features from the hidden representations.

Arithmetic Language

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

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

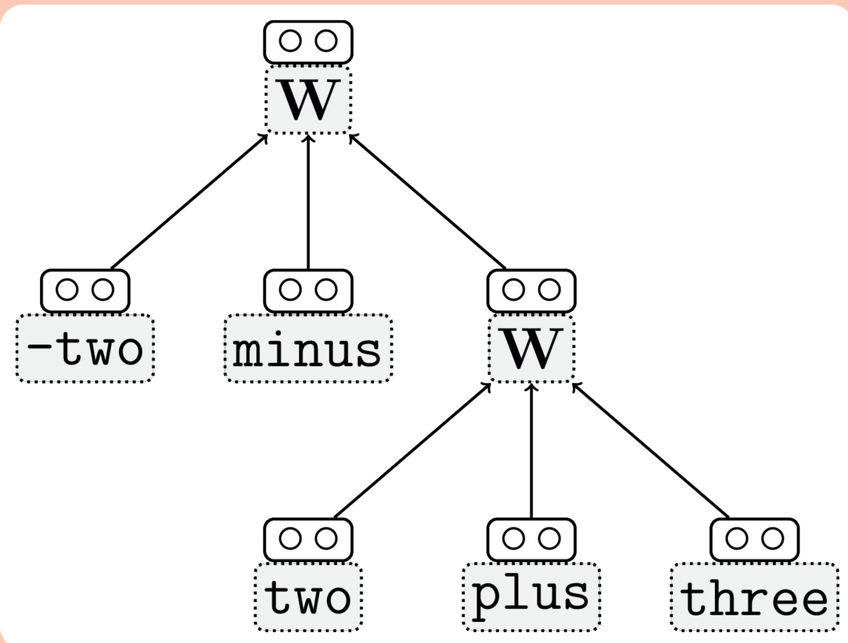
-14

Training signal

The solution to the arithmetic expression defines its semantics.

Recursive Neural Network (TreeRNN)

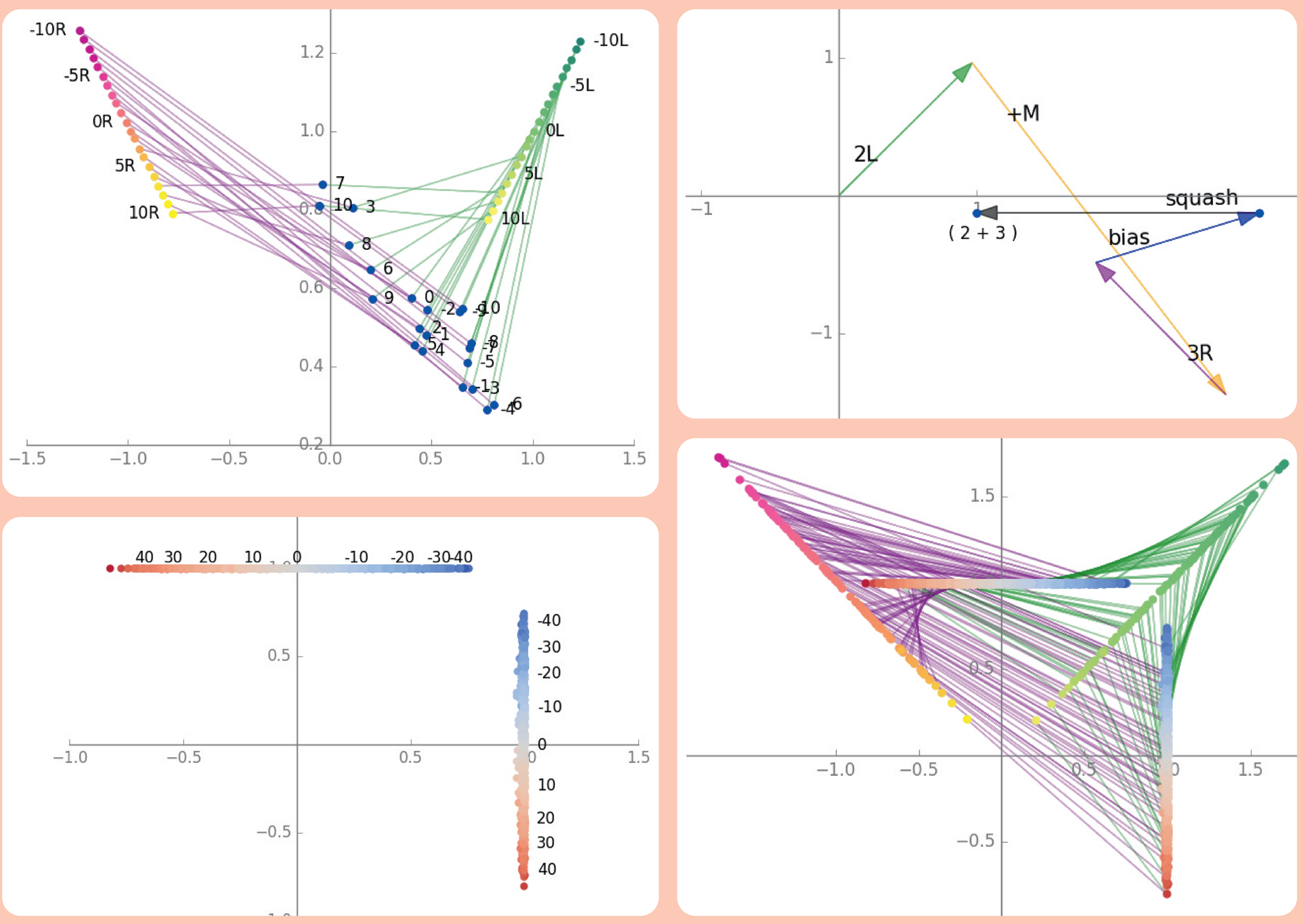
The architecture of the TreeRNN [1] mirrors the syntactic structure of the sentence. A single composition function is applied recursively to compute a vectorial parent representation from concatenated children representations:



$$\mathbf{p} = \tanh(\mathbf{W}_L \cdot \mathbf{x}_1 + \mathbf{W}_M \cdot \mathbf{x}_2 + \mathbf{W}_R \cdot \mathbf{x}_3 + \mathbf{b})$$

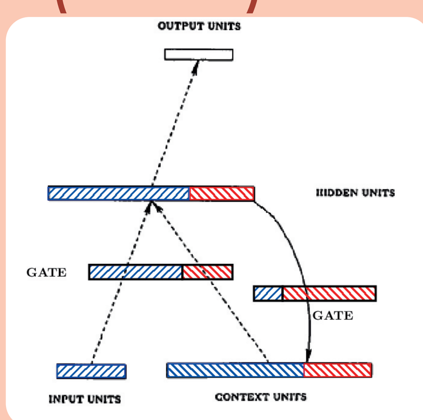
Project - Sum - Squash

By plotting the 2D representations we get full understanding of the learned solution. Each child representation is projected by either W_L, W_M or W_R. The projections are summed together with the bias. The result is squashed through tanh.



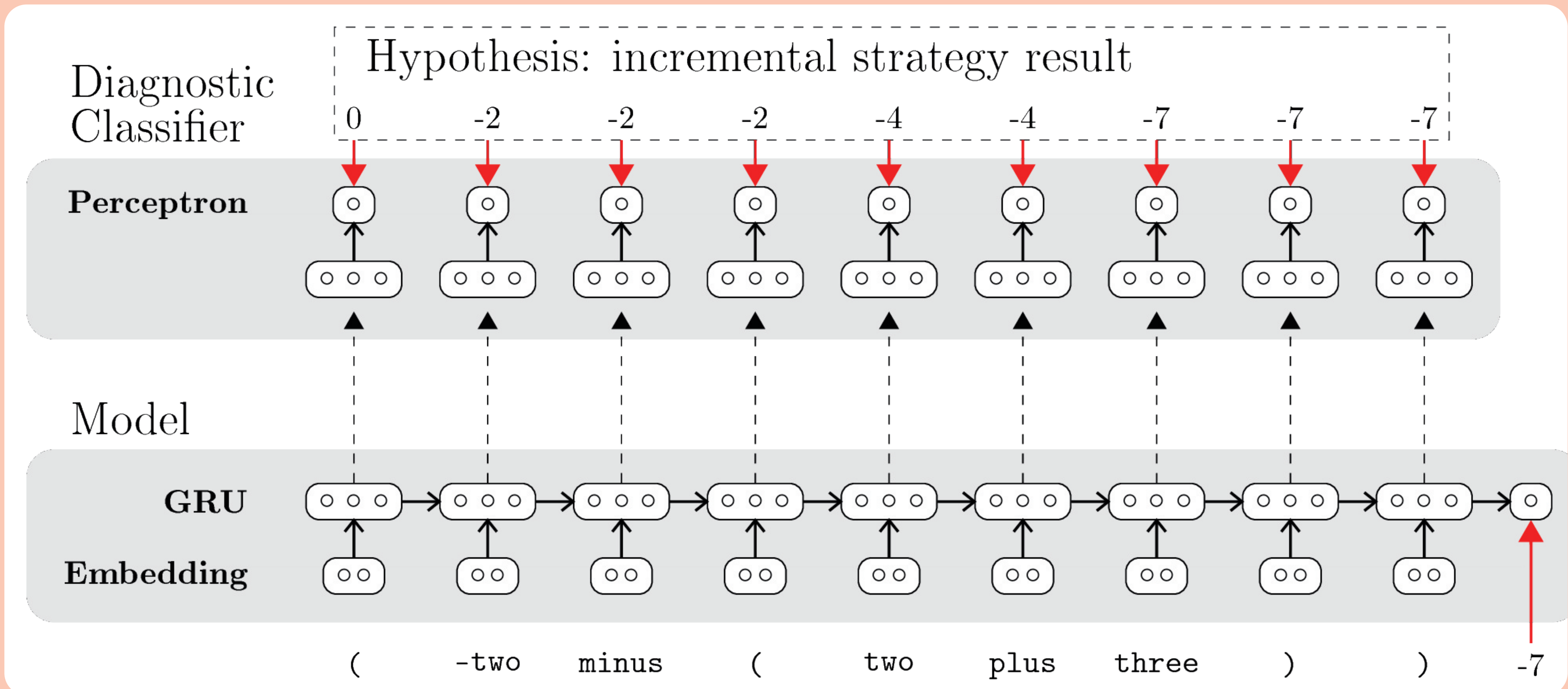
Gated Recurrent Neural Network (GRU)

The GRU [2] processes input sequences incrementally, employing gates to moderate the information flow. Brackets are inputted as words indicating the compositional structure of input sequences.



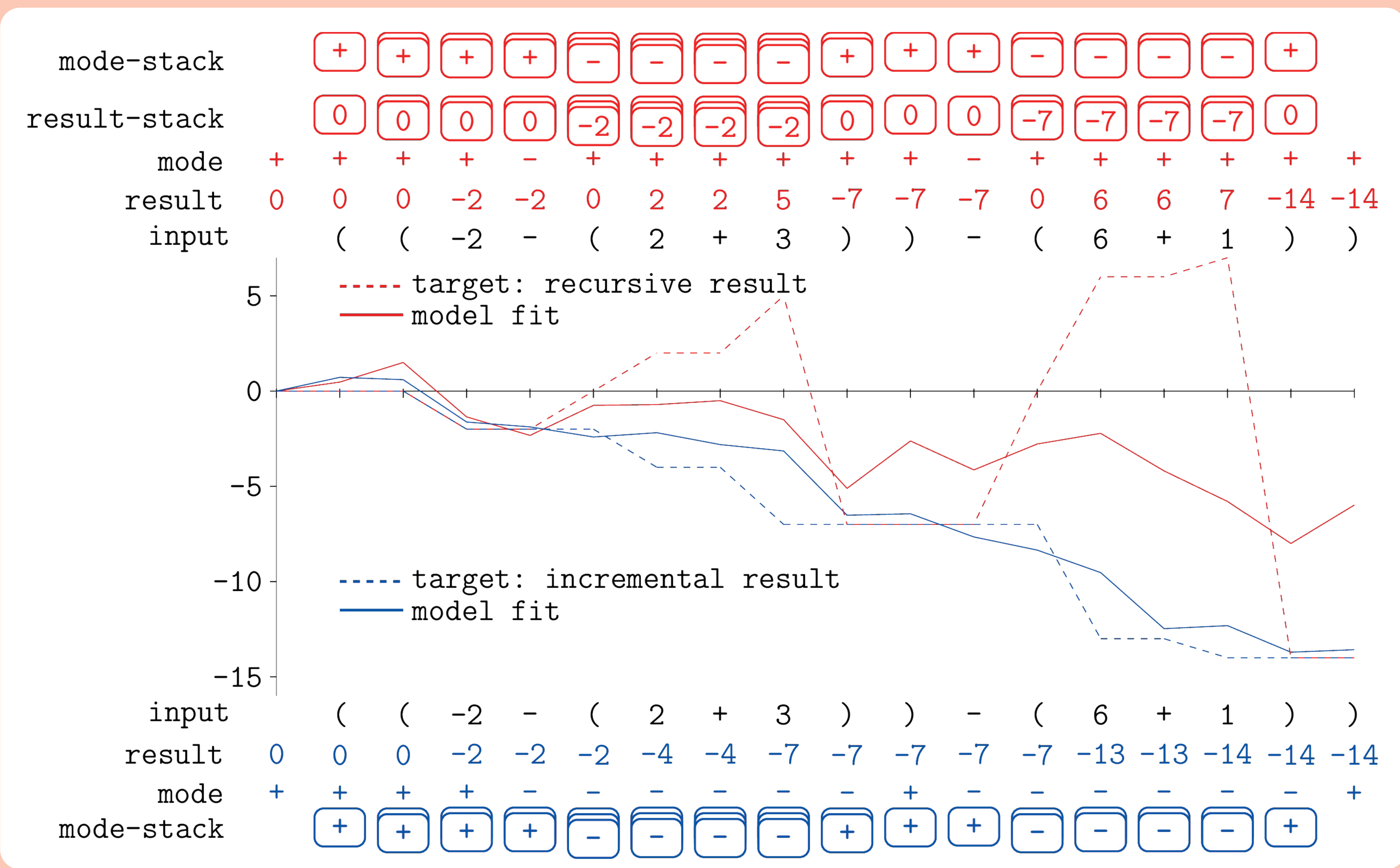
Diagnostic Classifiers

We train additional neural networks to predict hypothesised features from the hidden representations.



Trajectories

We train diagnostic classifiers to predict the intermediate resuls of two different symbolic strategies. We find that the network does not compute the outcome of the expression in a linearised recursive fashion, but integrates the digits incrementally without employing a 'result-stack' of numeric values.



References

[1] Socher et al. Learning continuous phrase representations and syntactic parsing with recursive neural networks. NIPS-2010. DL and Unsup Feat. Learning Workshop, pages 1-9, 2010.
[2] Cho et al. On the properties of neural machine translation: encoder-decoder approaches. SSST-8. 2014.
[3] Elman. Finding structure in time. Cognitive Science. 1990. 14, 2.