Sequence Processing

Moritz Wolter

September 25, 2022

Uni Bonn

Overview

Recurrent neural networks

Applications

Motivation

- Thus far we have never integrated information over time.
- We want the ability to create internal memory.
- Consider the sentence: I live in Paris. I speak ...
- ... French.
- Clearly it is likely for someone in Paris to speak French.
- Memory should help networks taking Paris into account when deciding what language is spoken.

Recurrent neural networks

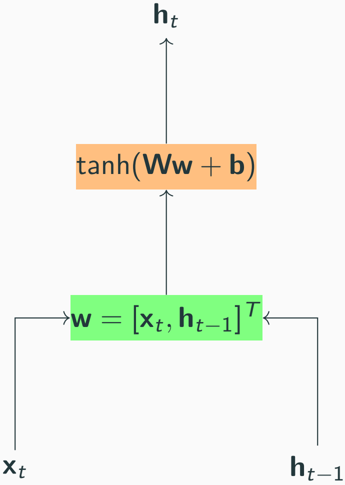
Elman-recurrent neural networks

A simple solution is to add a state to the network and feed this state recurrently back into the network [Elm90],

$$\overline{\mathbf{h}_t} = \mathbf{W}_h \mathbf{h}_t + \mathbf{W}_x \mathbf{x}_t + \mathbf{b}, \tag{1}$$

$$\mathbf{h}_{t+1} = f(\overline{\mathbf{h}_t}). \tag{2}$$

Elman-recurrent neural networks



Unrolling in Time

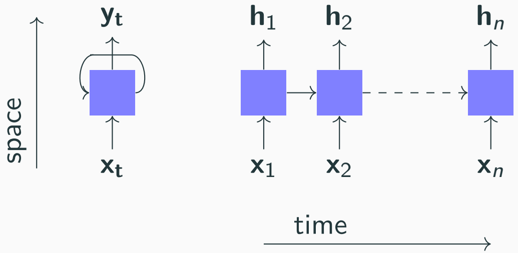


Figure: The rolled (left) cell can be unrolled (right) by considering all inputs it saw during the current gradient computation iteration.

Stability of recurrent connections

For an intuition. Consider a linear network without activations or inputs.

$$\mathbf{h}_{t+1} = \mathbf{W}_h \mathbf{h}_t \tag{3}$$

The evolution of the **h**-sequence is guided by it's largest eigenvalue. If an eigenvalue larger than one exists. The state explodes. If all eigenvalues are smaller than one the state vanishes [GBC16].

Long Short Term Memory (LSTM)

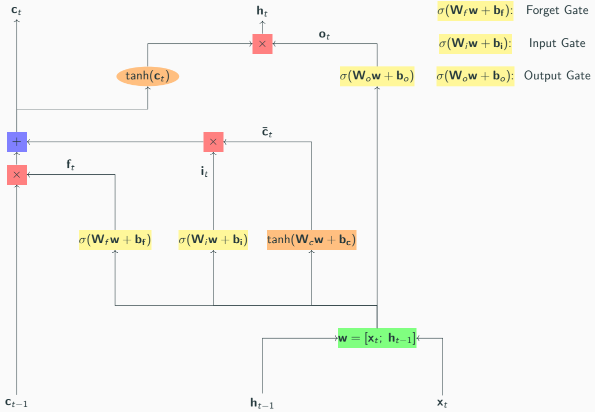


Figure: An LSTM cell as described in[HS97; Gre+16].

Long Short Term Memory (LSTM)

Llike a differentiable memory chip [Gra12] LSTM-memory can store n_h numbers. Gates govern all changes to the cell state. Gate and state equations are defined as [HS97; Gre+16]

$$\overline{\mathbf{z}_t} = \mathbf{W}_z\mathbf{x}_t + \mathbf{R}_z\mathbf{h}_{t-1} + \mathbf{b}_z, \mathbf{z}_t = \tanh(\overline{\mathbf{z}_t}), \tag{4}$$

$$\overline{\mathbf{i}_t} = \mathbf{W}_i\mathbf{x}_t + \mathbf{R}_i\mathbf{h}_{t-1} + \mathbf{p}_i \odot \mathbf{c}_{t-1} + \mathbf{b}_i, \mathbf{i}_t = \sigma(\overline{\mathbf{i}_t}), \tag{5}$$

$$\overline{\mathbf{f}_t} = \mathbf{W}_f\mathbf{x}_t + \mathbf{R}_f\mathbf{h}_{t-1} + \mathbf{p}_f \odot \mathbf{c}_{t-1} + \mathbf{b}_f, \mathbf{f}_t = \sigma(\overline{\mathbf{f}_t}), \tag{6}$$

$$\mathbf{c}_t = \mathbf{z}_t \odot \mathbf{i}_t + \mathbf{c}_{t-1} \odot \mathbf{f}_t, \tag{7}$$

$$\overline{\mathbf{o}_t} = \mathbf{W}_o\mathbf{x}_t + \mathbf{R}_o\mathbf{h}_{t-1} + \mathbf{p}_o \odot \mathbf{c}_t + \mathbf{b}_o, \mathbf{o}_t = \sigma(\overline{\mathbf{o}_t}), \tag{8}$$

$$\mathbf{h}_t = \tanh(\mathbf{c}_t) \odot \mathbf{o}_t. \tag{9}$$

Potential new states $\mathbf{z}_t$ are called block input. $\mathbf{i}$ is called the input gate. The forget gate is $\mathbf{f}$ and $\mathbf{o}$ denotes the output gate. $\mathbf{p} \in \mathbb{R}^{n_h}$ are peephole weights, $\mathbf{W} \in \mathbb{R}^{n_i \times n_h}$ denotes input, $\mathbf{R} \in \mathbb{R}^{n_o \times n_h}$ are the recurrent matrices. $\odot$ indicates element-wise products.

Long Short Term Memory (LSTM)

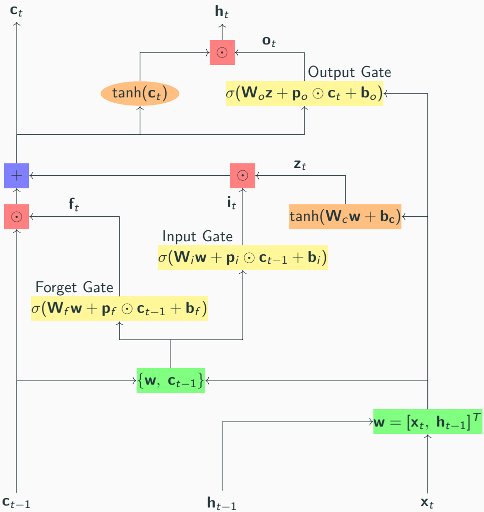
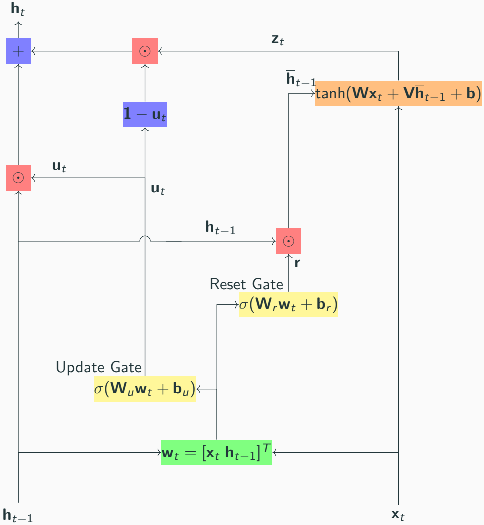


Figure: An LSTM-cell with peephole connections as described in [HS97; Gre+16]

Gated recurrent units



Orthogonal networks

TODO

Summary

TODO

Applications

Language Processing

TODO

Speech Processing

TODO

Time-series forecasting

TODO

Conclusion

TODO

References

[Elm90] Jeffrey L Elman. “Finding structure in time.” In: *Cognitive science* 14.2 (1990), pp. 179–211.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[Gra12] Alex Graves. “Supervised sequence labelling.” In: *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 5–13.

[Gre+16] Klaus Greff, Rupesh K Srivastava, Jan Koutnik, Bas R Steunebrink, and Jürgen Schmidhuber. “LSTM: A search space odyssey.” In: *IEEE transactions on neural networks and learning systems* 28.10 (2016), pp. 2222–2232.