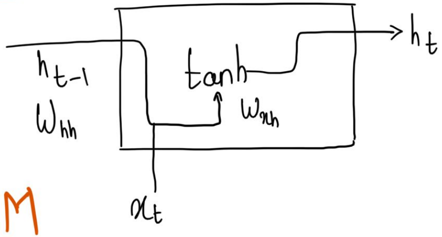
LSTM

B.Tech. Data Science, NMIMS

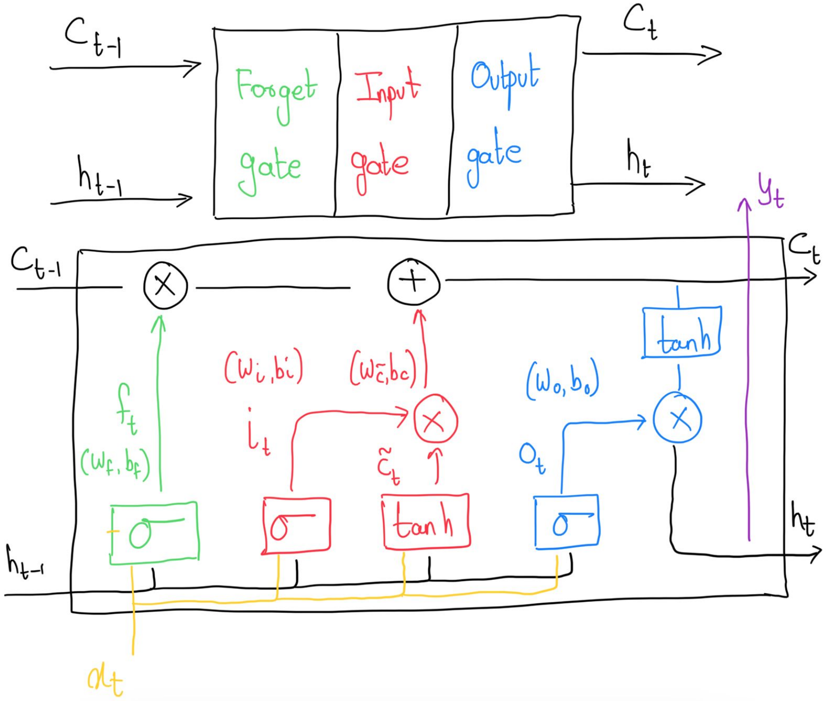
By,

Bilal Hungund, Data Scientist, Halliburton

RNN



LSTM



RNN

$$h_t = \tanh(W_{hh}\, h_{t-1} + W_{xh}\, x_t)$$

$$\hat{y}_t = W_{hy} \times h_t$$

LSTM

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

$$C_t = C_{t-1} * f_t + i_t * \tilde{C}_t$$

$$h_t = O_t * \tanh(C_t)$$

- Cell state
    - Information pass through the path.
- Why sigmoid?
    - Sigmoid can output 0 or 1, it can be used to forget or remember the information.
- Why tanh?
    - To overcome the problem of vanishing gradients.
    - Tanh second derivative can sustain for a long range before going to zero.
- Forget Gate
    - It tells the information to throw away from the cell state.
    - 0 completely forget or 1 to keep all information
- Input Gate
    - It tells that what new information are going to store in the cell state.
    - Sigmoid layer decides which values are updated.
    - Tanh layer gives weights to the values to be added to the state. Candidate to get the memory vector for the current timestamp t.
- Output Gate
    - It is used to provide the activation function output.
    - Sigmoid decides which cell part for output.
    - It returns update hidden state value.

Timeline of LSTM

1991: Sepp Hochreiter analyzed the vanishing gradient problem and developed principles of the method in his German diploma thesis[1] advised by Jürgen Schmidhuber.

1995: "Long Short-Term Memory (LSTM)" is published in a technical report by Sepp Hochreiter and Jürgen Schmidhuber.[2]

1999: Felix Gers and his advisor Jürgen Schmidhuber and Fred Cummins introduced the forget gate (also called "keep gate") into the LSTM architecture,[3] enabling the LSTM to reset its own state. [4]

2004: First successful application of LSTM to speech by Schmidhuber's student Alex Graves et al.[5]

2005: First publication (Graves and Schmidhuber) of LSTM with full backpropagation through time and of bi-directional LSTM. [6]

2014: Kyunghyun Cho et al. put forward a simplified variant of the forget gate LSTM called Gated recurrent unit (GRU).[7]

References

- RNN Paper: https://onlinelibrary.wiley.com/doi/epdf/10.1207/s15516709cog1402_1
- http://www.bioinf.jku.at/publications/older/3804.pdf
- http://www.bioinf.jku.at/publications/older/2604.pdf
- Gers, Felix; Schmidhuber, Jürgen; Cummins, Fred (1999). "Learning to forget: Continual prediction with LSTM". 9th International Conference on Artificial Neural Networks: ICANN '99. Vol. 1999. pp. 850–855
- Klaus Greff; Rupesh Kumar Srivastava; Jan Koutník; Bas R. Steunebrink; Jürgen Schmidhuber (2015). "LSTM: A Search Space Odyssey". IEEE Transactions on Neural Networks and Learning Systems. 28 (10): 2222–2232. arXiv:1503.04069
- Graves, Alex; Beringer, Nicole; Eck, Douglas; Schmidhuber, Juergen (2004). Biologically Plausible Speech Recognition with LSTM Neural Nets. Workshop on Biologically Inspired Approaches to Advanced Information Technology, Bio-ADIT 2004, Lausanne, Switzerland. pp. 175–184.

References

- Graves, A.; Schmidhuber, J. (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". Neural Networks. 18 (5–6): 602–610.
- Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation"