Introduction to:
Recurrent Neural Networks

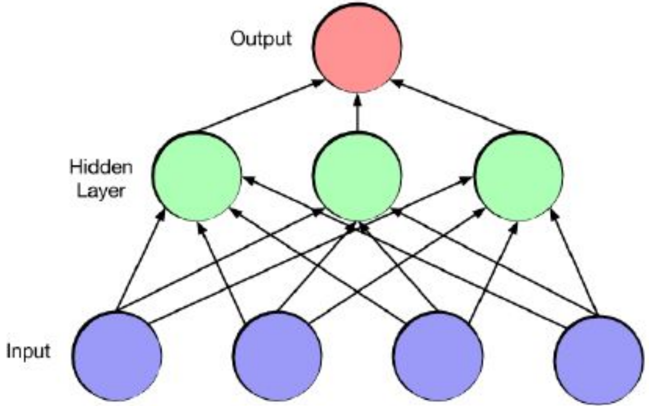
Journal Club - March 5, 2017

Artificial Neural Networks

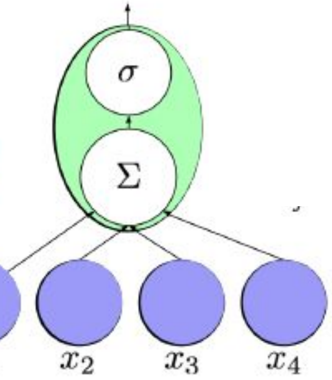
- Network of Artificial Neurons

- Connected Layerwise

- Activation
  - Weighted Sum
  - Nonlinearity



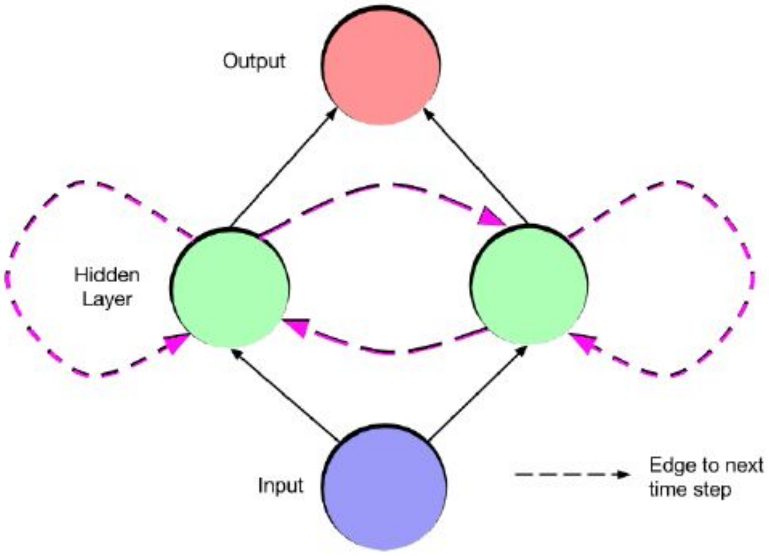
$$\boldsymbol{h} = \sigma(W^{\mathrm{hx}}\boldsymbol{x} + \boldsymbol{b}_h)$$



Recurrent Neural Networks (RNNs)

- Addition of State Information

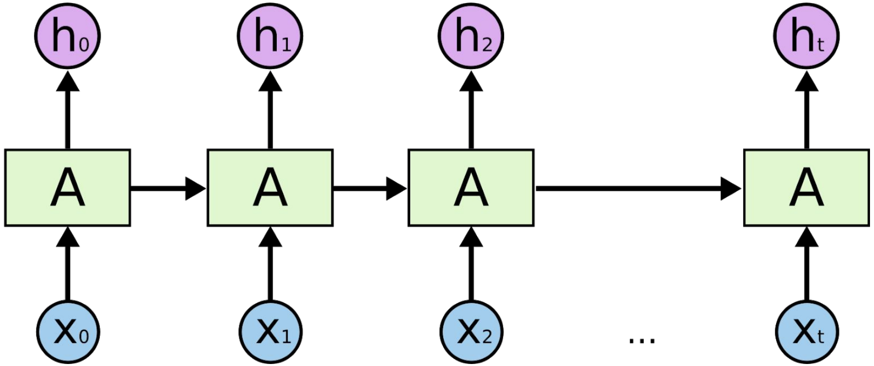
- W_{hh} - Dependence on Previous State



$$\boldsymbol{h}^{(t)} = \sigma(W^{\mathrm{hx}}\boldsymbol{x}^{(t)} + W^{\mathrm{hh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

Recurrent Neural Networks (RNNs)

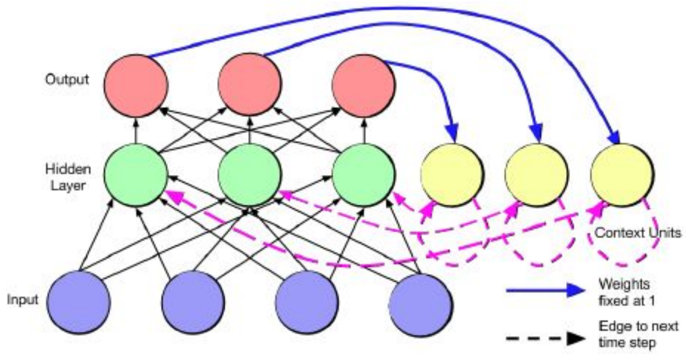
- ANN with copies of itself
- W_{hh} - Dependence on Previous State



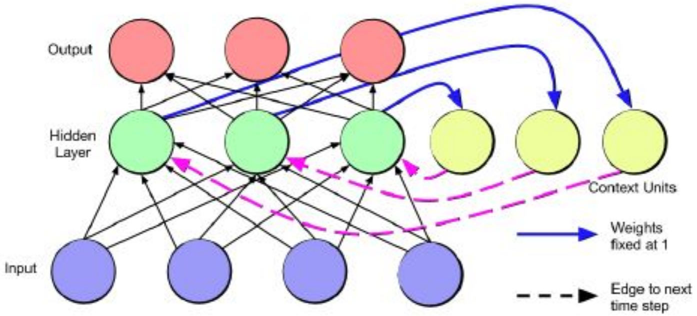
$$\boldsymbol{h}^{(t)} = \sigma(W^{\mathrm{hx}}\boldsymbol{x}^{(t)} + W^{\mathrm{hh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

Early RNN designs

- Jordan et al. (1986)

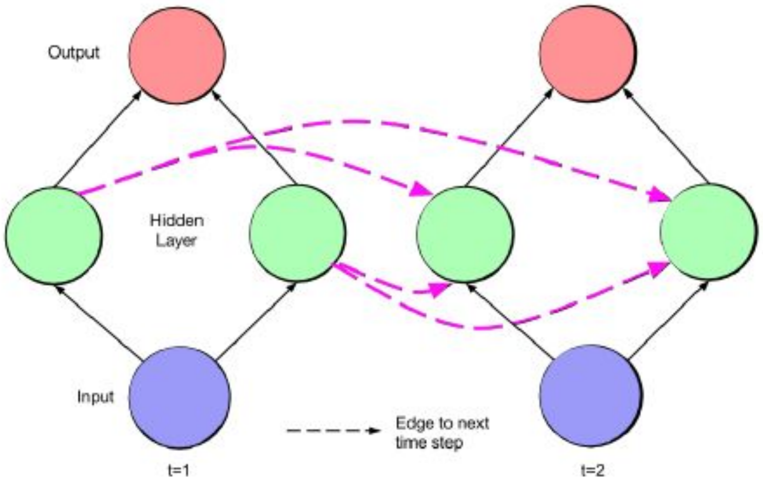


- Elman et al. (1990)



Recurrent Neural Networks (RNNs)

- Addition of State Information
- W_{hh} - Dependence on Previous State



$$\boldsymbol{h}^{(t)} = \sigma(W^{\text{hx}}\boldsymbol{x}^{(t)} + W^{\text{hh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

Recurrent Neural Networks (RNNs)

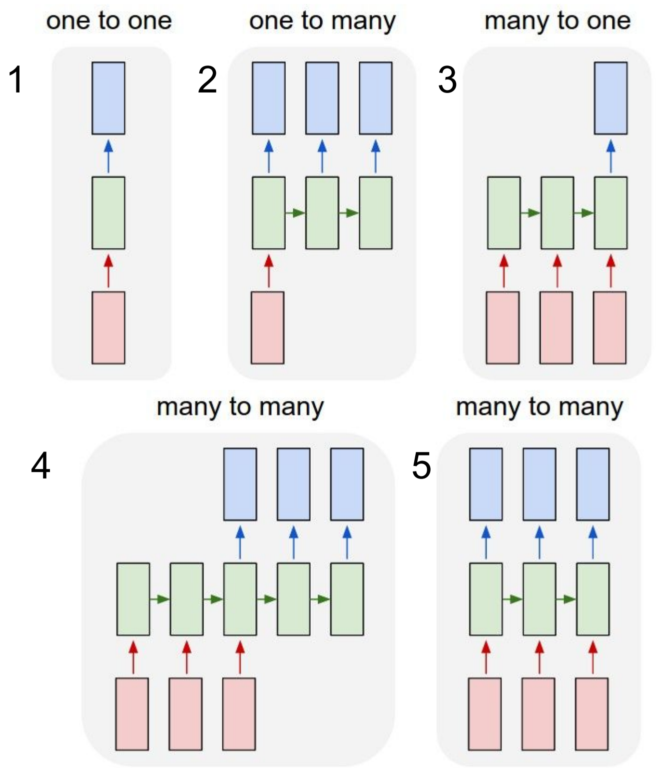
1. Vanilla ANN

2. Single Input, Sequence Output
   (e.g image captioning)

3. Sequence Input, Single Output
   (e.g. sentiment analysis)

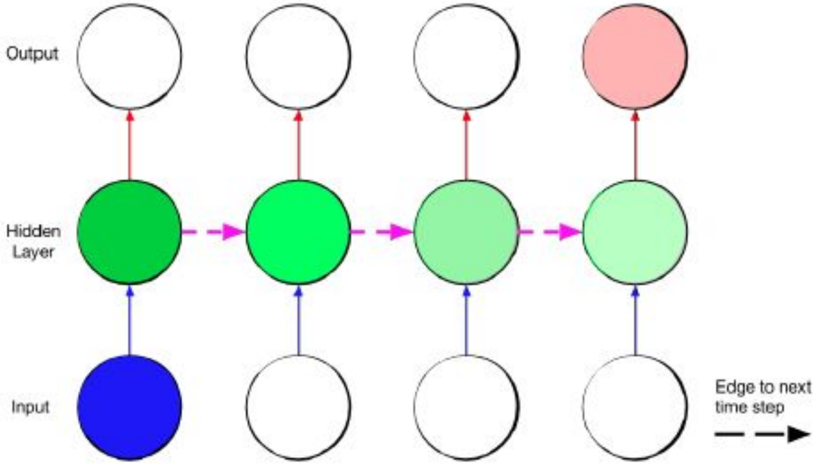
4. Sequence Input, Sequence Output
   (e.g. translator {eng -> fr}

5. Synced Sequence I/O
   (e.g. label each frame in a video)



Challenge

- Decaying/Exploding Gradients



Bidirectional RNNs

Memory Cells

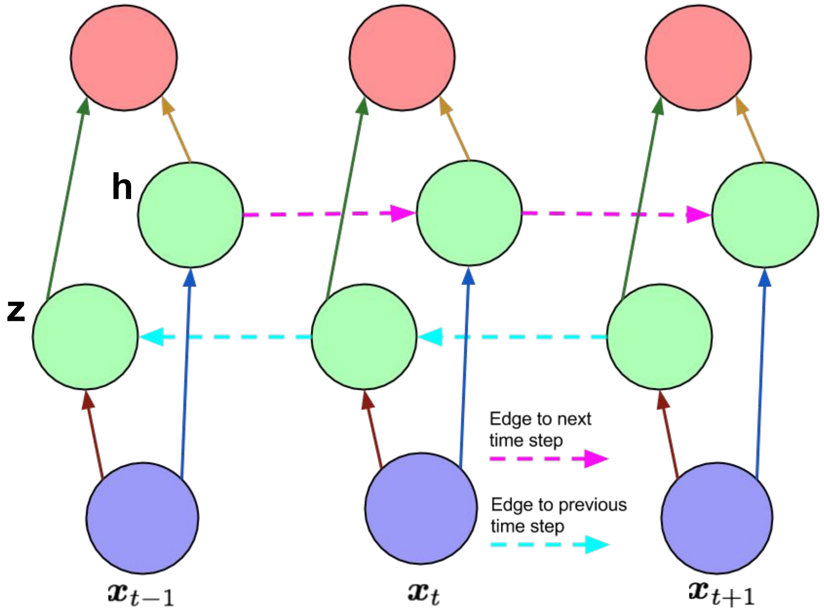
- Two layers of hidden nodes
- Works on spatial sequences
- Must run on finite sequence

$$h^{(t)} = \sigma(W^{\text{hx}} x^{(t)} + W^{\text{hh}} h^{(t-1)} + b_h)$$

$$z^{(t)} = \sigma(W^{\text{zx}} x^{(t)} + W^{\text{zz}} z^{(t+1)} + b_z)$$

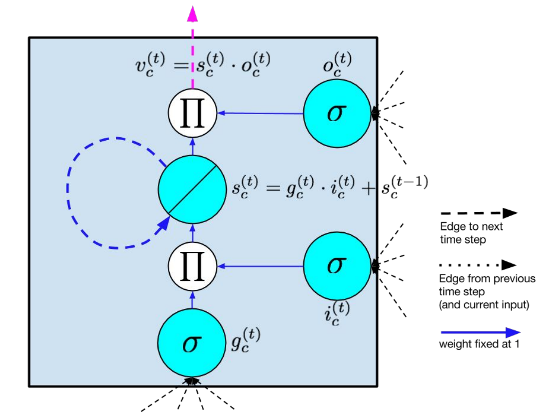
$$\hat{y}^{(t)} = \text{softmax}(W^{\text{yh}} h^{(t)} + W^{\text{yz}} z^{(t)} + b_y)$$

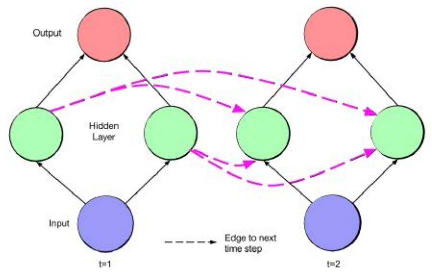
Schuster and Paliwal [1997]



Long Short Term Memory (LSTM)

- Node ->Memory Cells
- Node w. recurrent edge
- Weights (long term memory)
- Secondary Nodes
  (short term memory)



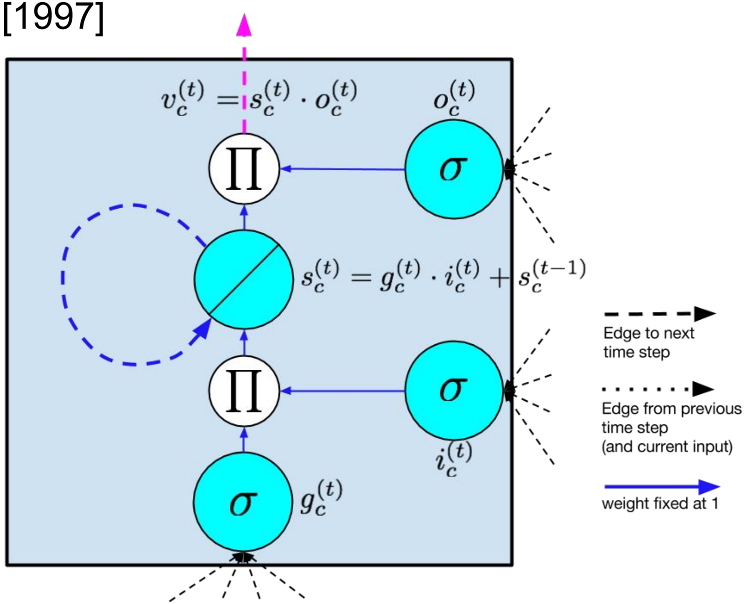


Long Short Term Memory (LSTM)

Memory Cells

- Input Node ($g_c^{(t)}$)
- Input Gate ($i_c^{(t)}$)
- Internal State ($s_c^{(t)}$)
- Output Gate ($o_c^{(t)}$)

Hochreiter and Schmidhuber [1997]



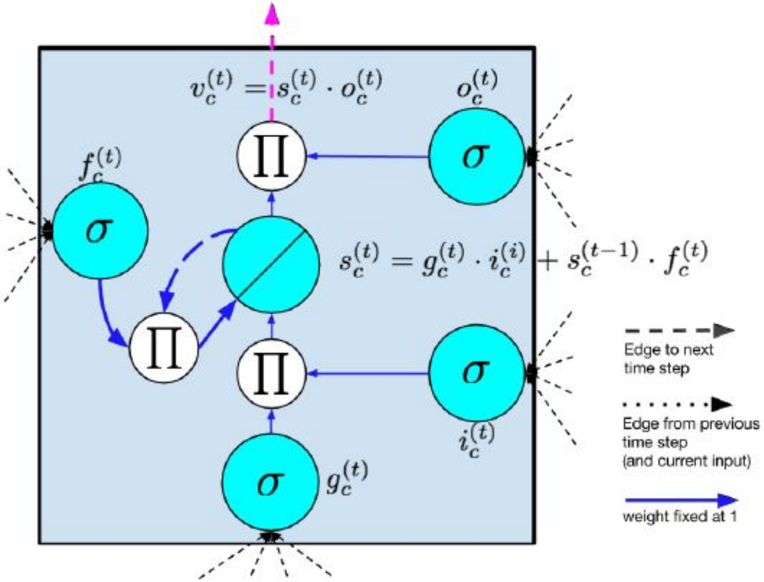
$$\boldsymbol{h}^{(t)} = \sigma(W^{\mathrm{hx}}\boldsymbol{x}^{(t)} + W^{\mathrm{hh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

Long Short Term Memory (LSTM)

Memory Cells

- Input Node ($g_c^{(t)}$)
- Input Gate ($i_c^{(t)}$)
- Internal State ($s_c^{(t)}$)
- Output Gate ($o_c^{(t)}$)
- Forget Gate ($f_c^{(t)}$)

Gers et al. [2000]



Long Short Term Memory (LSTM)

Memory Cells

- Input Node ($g_c^{(t)}$)
- Input Gate ($i_c^{(t)}$)
- Internal State ($s_c^{(t)}$)
- Output Gate ($o_c^{(t)}$)
- Forget Gate ($f_c^{(t)}$)

$$\boldsymbol{g}^{(t)} = \phi(W^{\mathrm{gx}}\boldsymbol{x}^{(t)} + W^{\mathrm{gh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_g)$$

$$\boldsymbol{i}^{(t)} = \sigma(W^{\mathrm{ix}}\boldsymbol{x}^{(t)} + W^{\mathrm{ih}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_i)$$

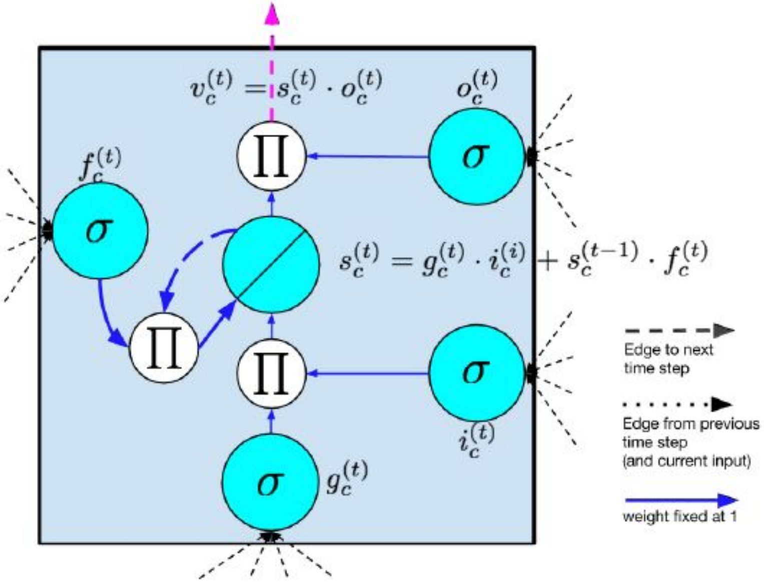
$$\boldsymbol{f}^{(t)} = \sigma(W^{\mathrm{fx}}\boldsymbol{x}^{(t)} + W^{\mathrm{fh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_f)$$

$$\boldsymbol{o}^{(t)} = \sigma(W^{\mathrm{ox}}\boldsymbol{x}^{(t)} + W^{\mathrm{oh}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_o)$$

$$\boldsymbol{s}^{(t)} = \boldsymbol{g}^{(t)} \odot \boldsymbol{i}^{(i)} + \boldsymbol{s}^{(t-1)} \odot \boldsymbol{f}^{(t)}$$

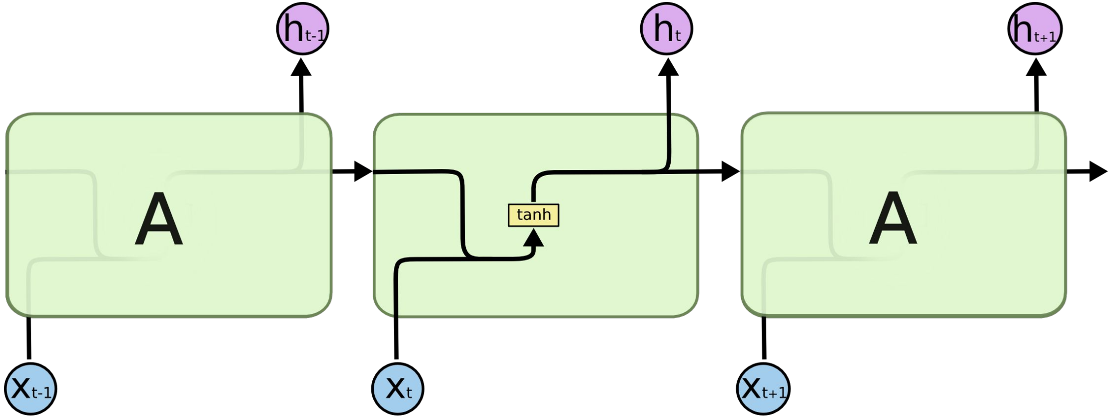
$$\boldsymbol{h}^{(t)} = \phi(\boldsymbol{s}^{(t)}) \odot \boldsymbol{o}^{(t)}.$$

Gers et al. [2000]

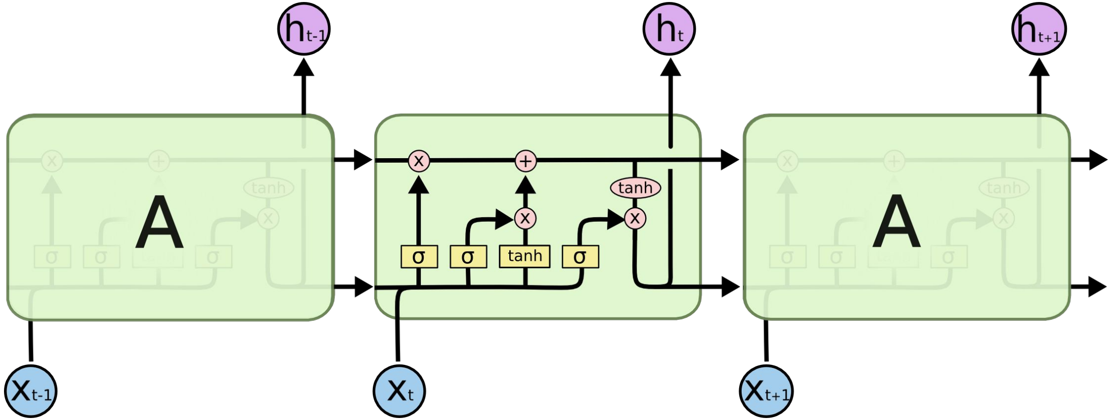


Long Short Term Memory (LSTM)

RNN

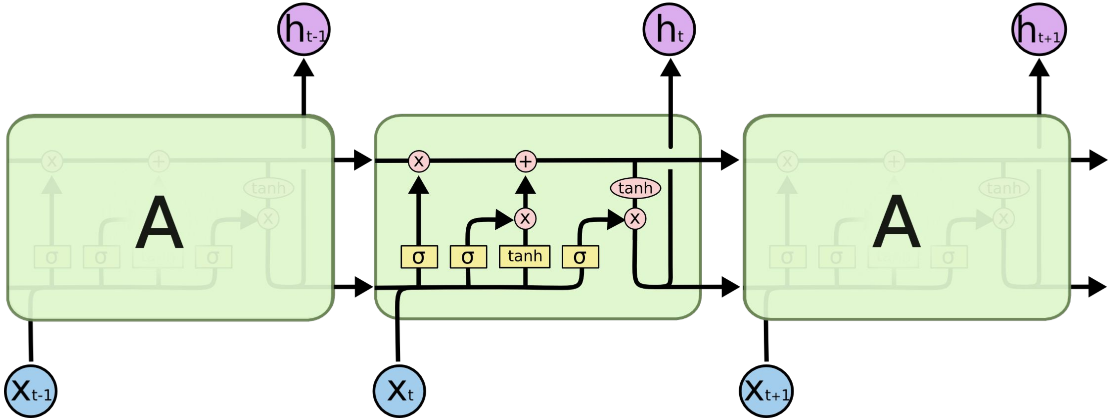


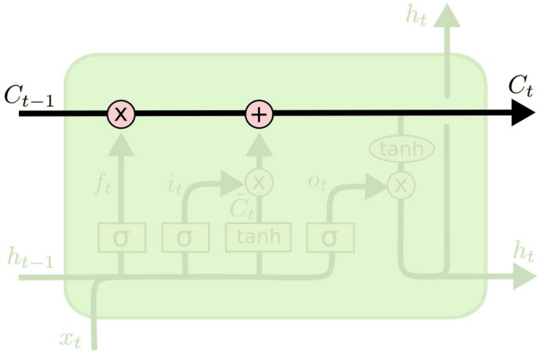
LSTM

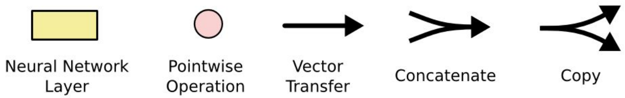


Long Short Term Memory (LSTM)

LSTM

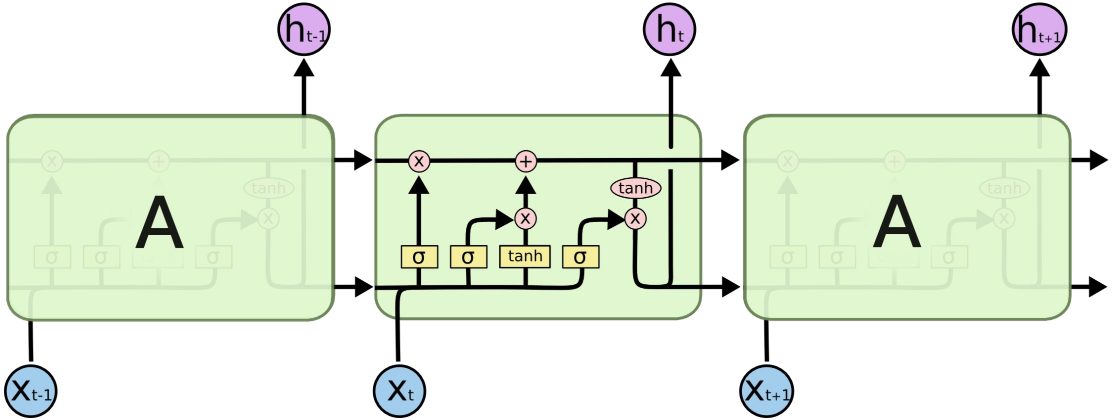


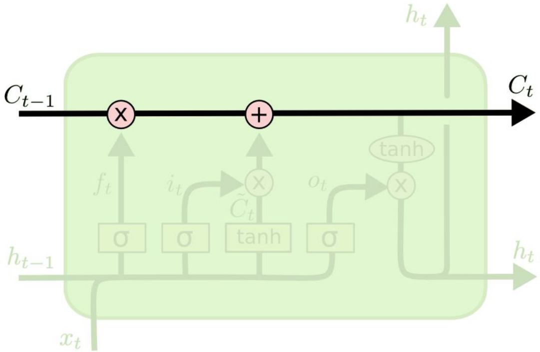


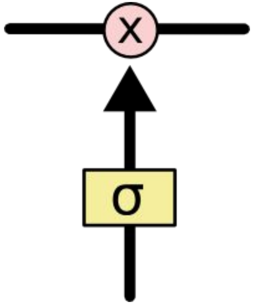


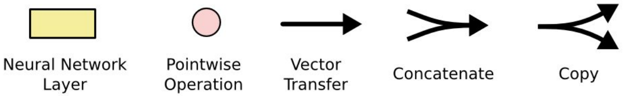
Long Short Term Memory (LSTM)

LSTM





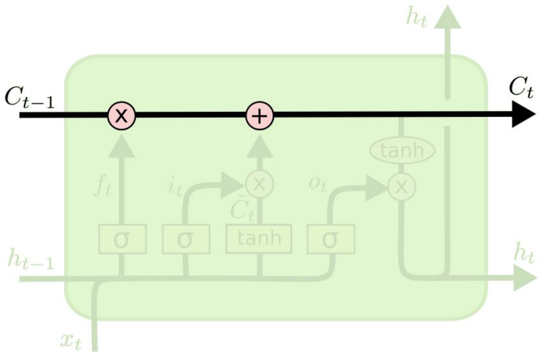


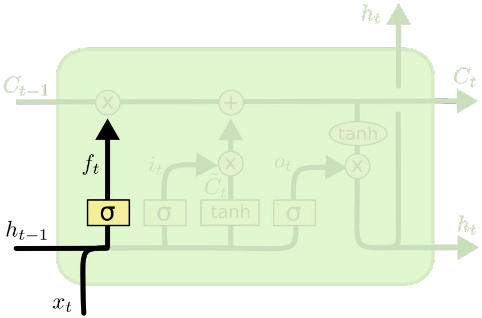


Long Short Term Memory (LSTM)

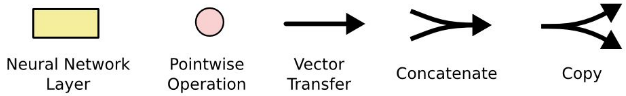
<u>Cell State</u>

What should we forget?





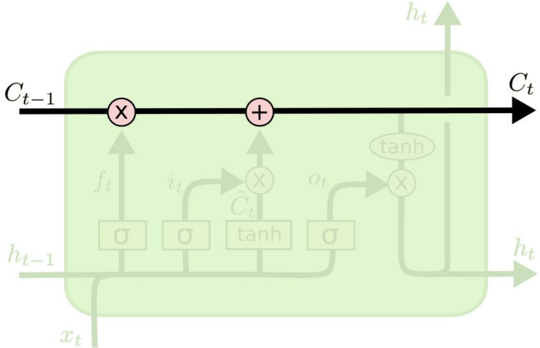
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

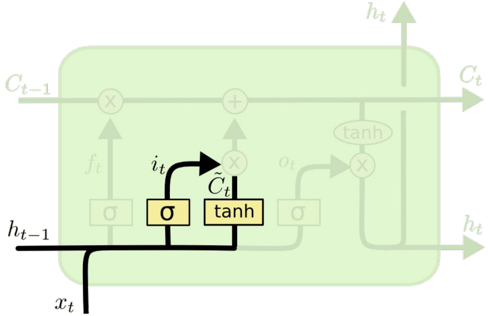


Long Short Term Memory (LSTM)

Cell State

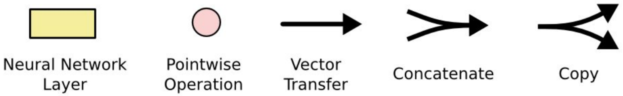
What should we save?





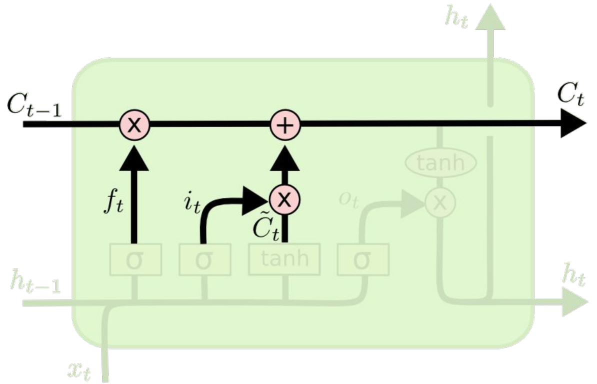
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

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

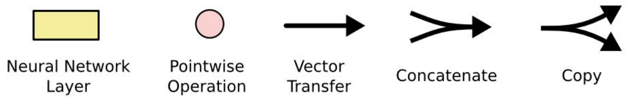


Long Short Term Memory (LSTM)

Update Cell State

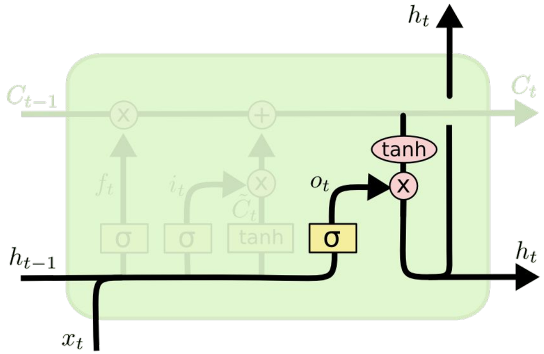


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

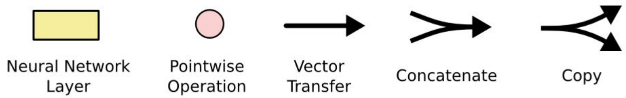


Long Short Term Memory (LSTM)

Hidden State Output



$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$



Resources

Lipton, Zachary C., John Berkowitz, and Charles Elkan. "A critical review of recurrent neural networks for sequence learning." arXiv preprint arXiv:1506.00019 (2015).

Kawakami, Kazuya. Supervised Sequence Labelling with Recurrent Neural Networks. Diss. Ph. D. thesis, Technical University of Munich, 2008.

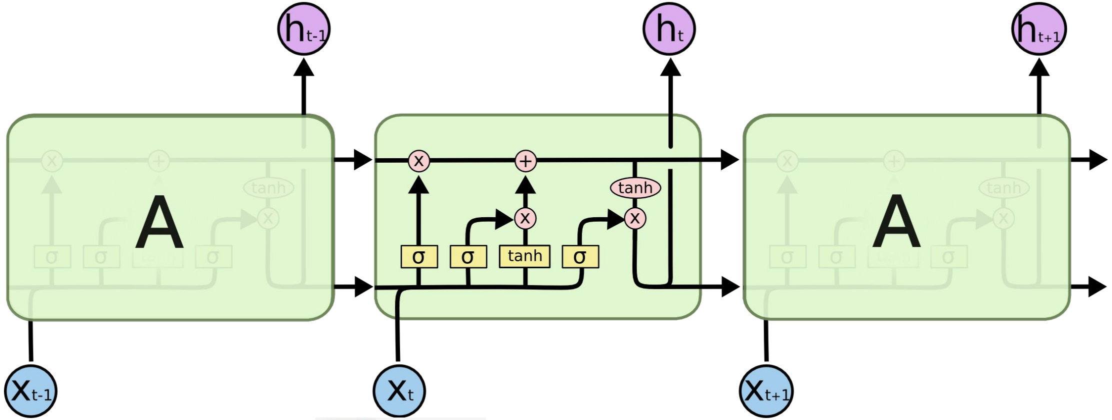
www.youtube.com/sirajraval

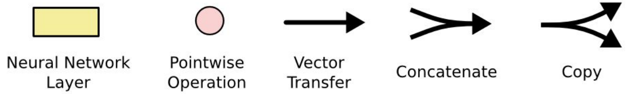
The Unreasonable Effectiveness of Recurrent Neural Networks, Andrej Karpathy

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Example (LSTM)

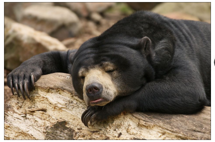
Grumpy Bear Example

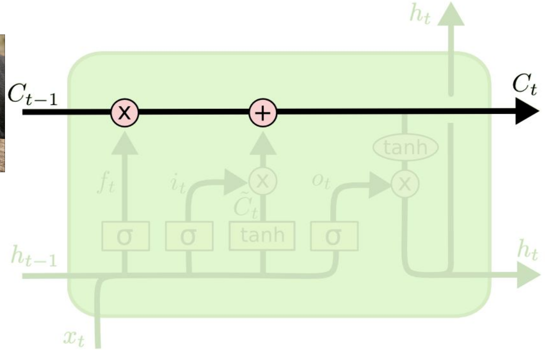


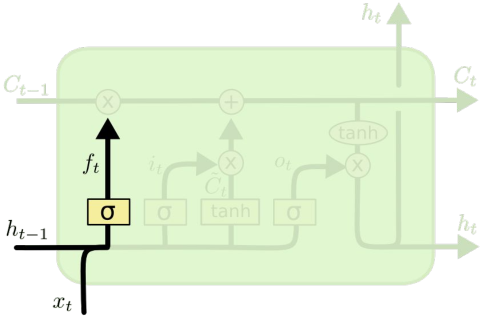


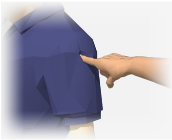
Example (LSTM)

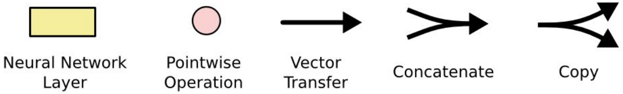
Grumpy Bear Example





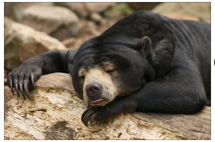


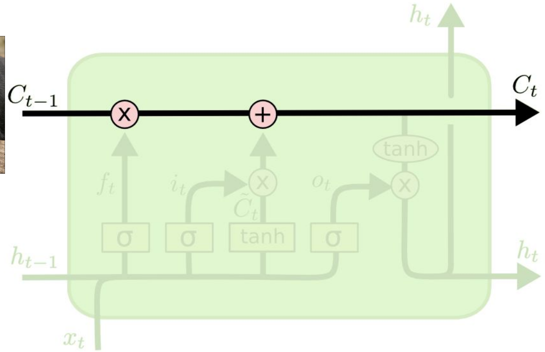




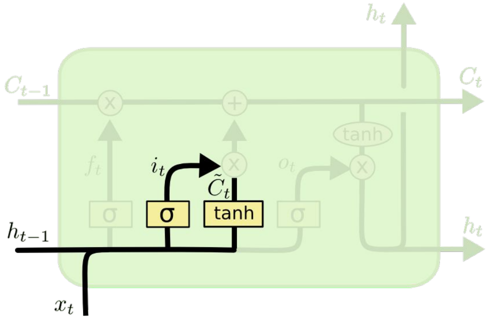
Example (LSTM)

Grumpy Bear Example

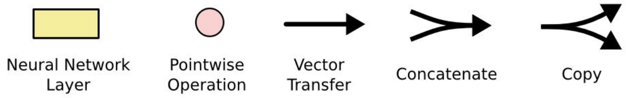




....

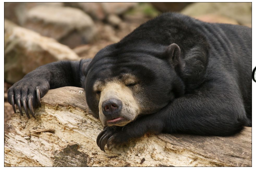


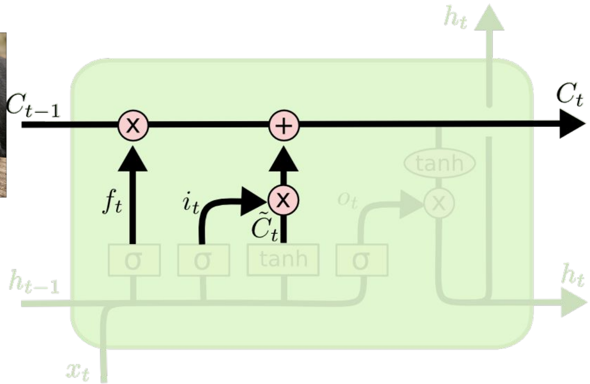
"Wake up bear"



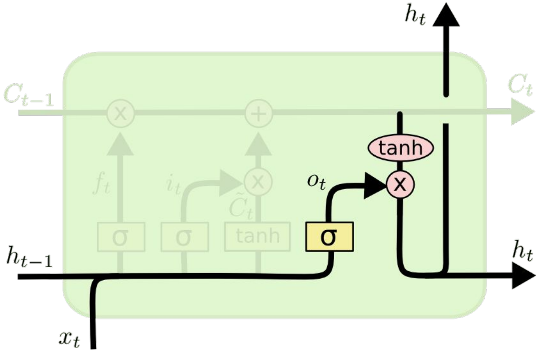
Example (LSTM)

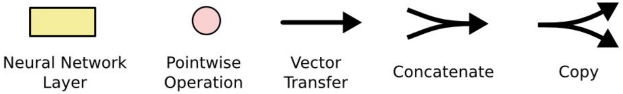
Grumpy Bear Example

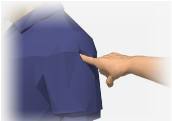




....

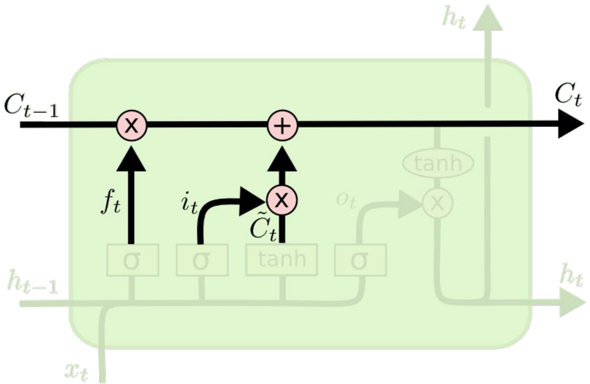






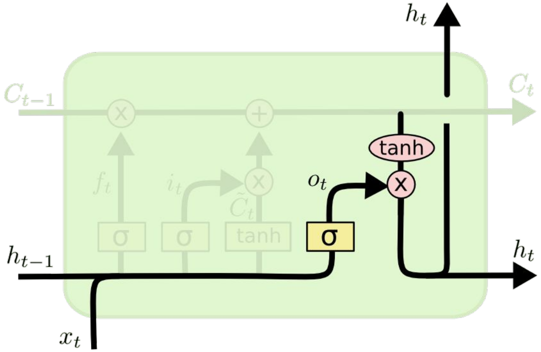
“Wake up bear”

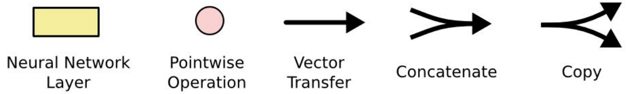
Example (LSTM)



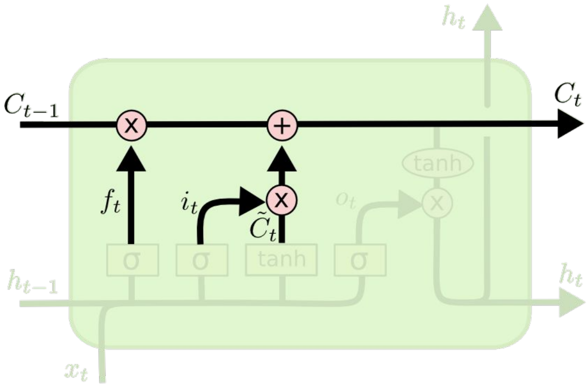
….

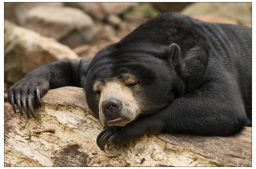
“Wake up lazy bear”





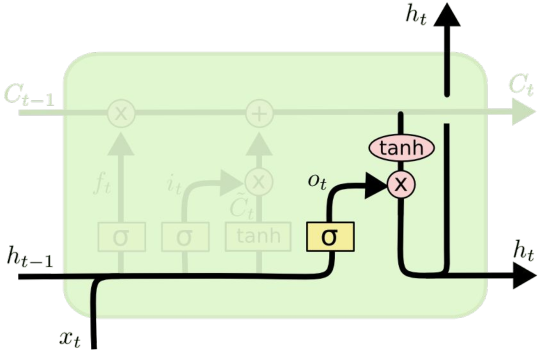
Example (LSTM)

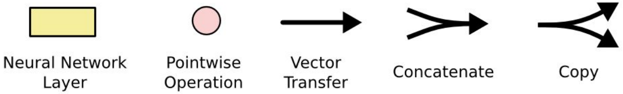




....

....





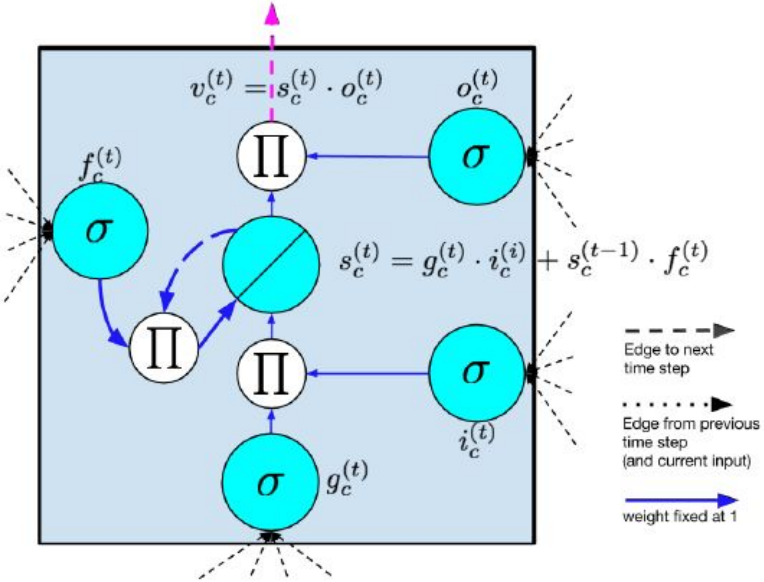
Long Short Term Memory (LSTM)

Memory Cells

- Input Node ($g_c^{(t)}$)
- Input Gate ($i_c^{(t)}$)
- Internal State ($s_c^{(t)}$)
- Output Gate ($o_c^{(t)}$)
- Forget Gate ($f_c^{(t)}$)

$$g^{(t)} = \phi(W^{\mathrm{gx}} x^{(t)} + W^{\mathrm{gh}} h^{(t-1)} + b_g)$$
$$i^{(t)} = \sigma(W^{\mathrm{ix}} x^{(t)} + W^{\mathrm{ih}} h^{(t-1)} + b_i)$$
$$f^{(t)} = \sigma(W^{\mathrm{fx}} x^{(t)} + W^{\mathrm{fh}} h^{(t-1)} + b_f)$$
$$o^{(t)} = \sigma(W^{\mathrm{ox}} x^{(t)} + W^{\mathrm{oh}} h^{(t-1)} + b_o)$$
$$s^{(t)} = g^{(t)} \odot i^{(i)} + s^{(t-1)} \odot f^{(t)}$$
$$h^{(t)} = \phi(s^{(t)}) \odot o^{(t)}.$$

Gers et al. [2000]

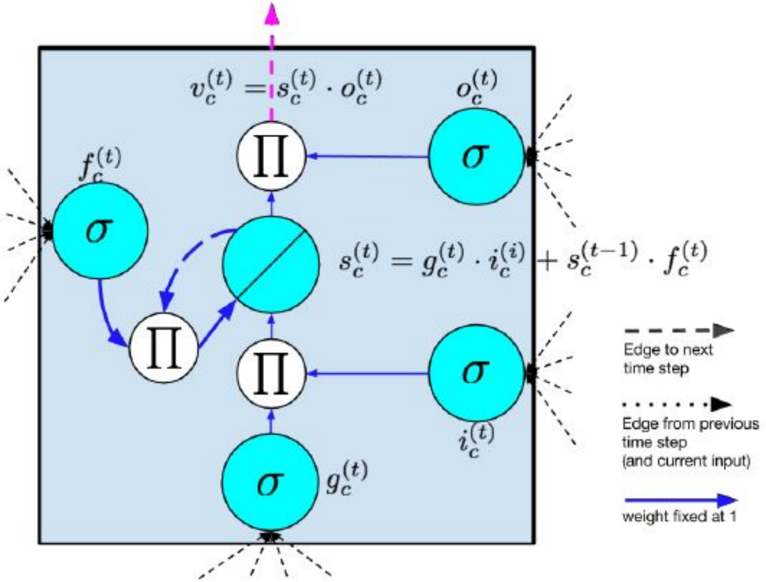


Long Short Term Memory (LSTM)

Memory Cells

- Input Node ($g_c^{(t)}$)
- Input Gate ($i_c^{(t)}$)
- Internal State ($s_c^{(t)}$)
- Output Gate ($o_c^{(t)}$)
- Forget Gate ($f_c^{(t)}$)

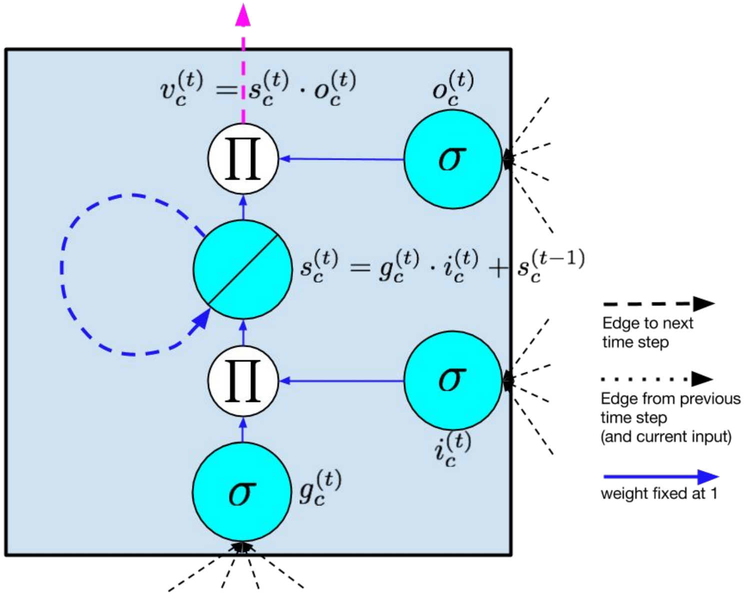
Gers et al. [2000]



Long Short Term Memory (LSTM)

Memory Cells

-



Modern RNN Architectures

- LSTM

- Elman (1990)

Recurrent Neural Networks (RNNs)

- Network of Artificial Neurons
- Connected Layerwise
- Activation
  - Weighted Sum
  - Nonlinearity