Basic Knowledge of Recurrent Neural Network & Long Short Term Memory Network

Reported by Steve Liu

Summary:

- Basic knowledge of the Tensorflow
- Basic knowledge of the Recurrent Neural Network and LSTM
- Tutorial of the word prediction

What is Recurrent Neural Network?

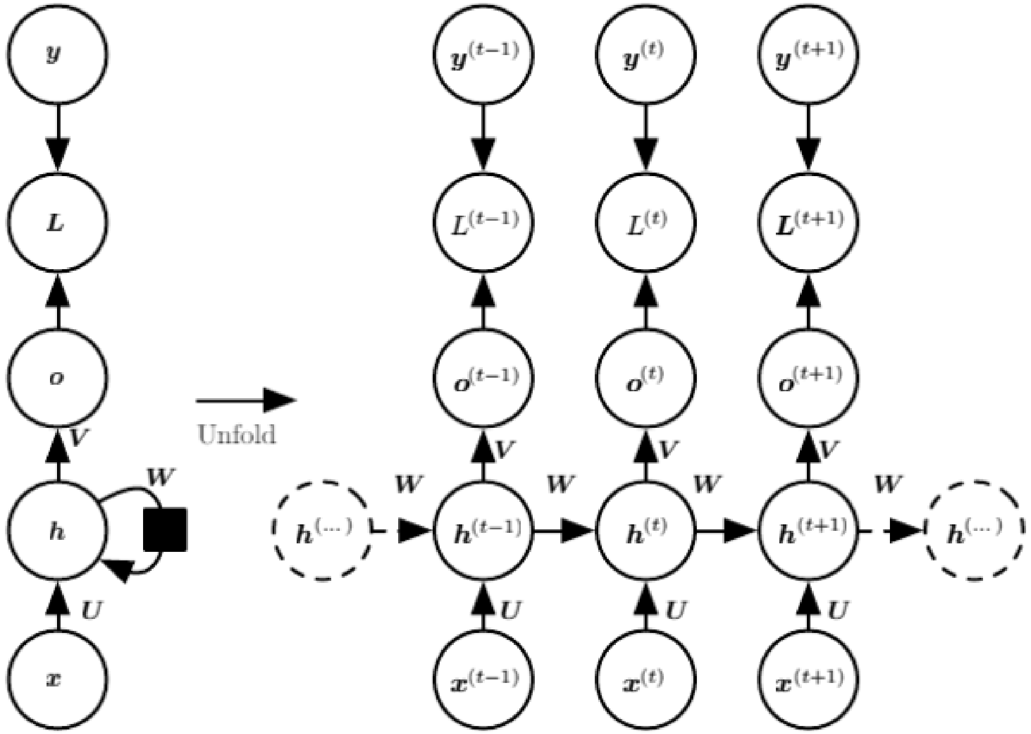
- Recurrent Neural Network in essence is neural networks that employ recurrence, which is basically using information from a previous forward pass over the neural network.

What can Recurrent Neural Network used for?

- To solve the problems that the predictions are based on a time sequence of input data.
- Translation, Natural Language Processing, Composing, Picture Describing, .etc.

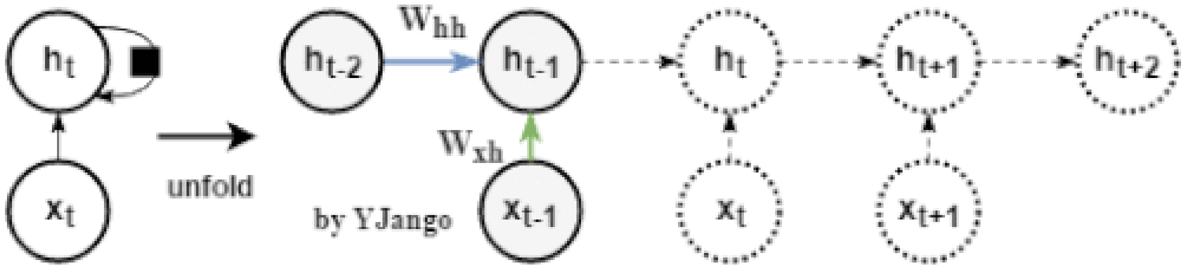
What is the input and the output?

- The input should be a sequence.
- The output of the network can be a sequence or just one.



$$\begin{cases} a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)} \\ h^{(t)} = \tanh(a^{(t)}) \\ o^{(t)} = c + Vh^{(t)} \end{cases}$$

- $x^{(t)}$ is the input sequence
- $h^{(t)}$ is the hidden state
- $o^{(t)}$ is the output



$$h_t = \emptyset(W_{xh} \cdot x_t + W_{hh} \cdot h_{t-1} + b)$$

For easy to understand, we just focus on the blue arrow, then the h_t can be calculated as,

$$h_t = (W_{hh})^t \cdot h_0$$

The eigen decomposition of the Matrix will be,

$$h_t = Q\Lambda^t Q^T \cdot h_0$$

If $\Lambda < 1$, then Λ^t trend to 0; If $\Lambda > 1$, then Λ^t trend to ∞， h_t will be covered

What is the disadvantage of the Standard Recurrent Neural Network?

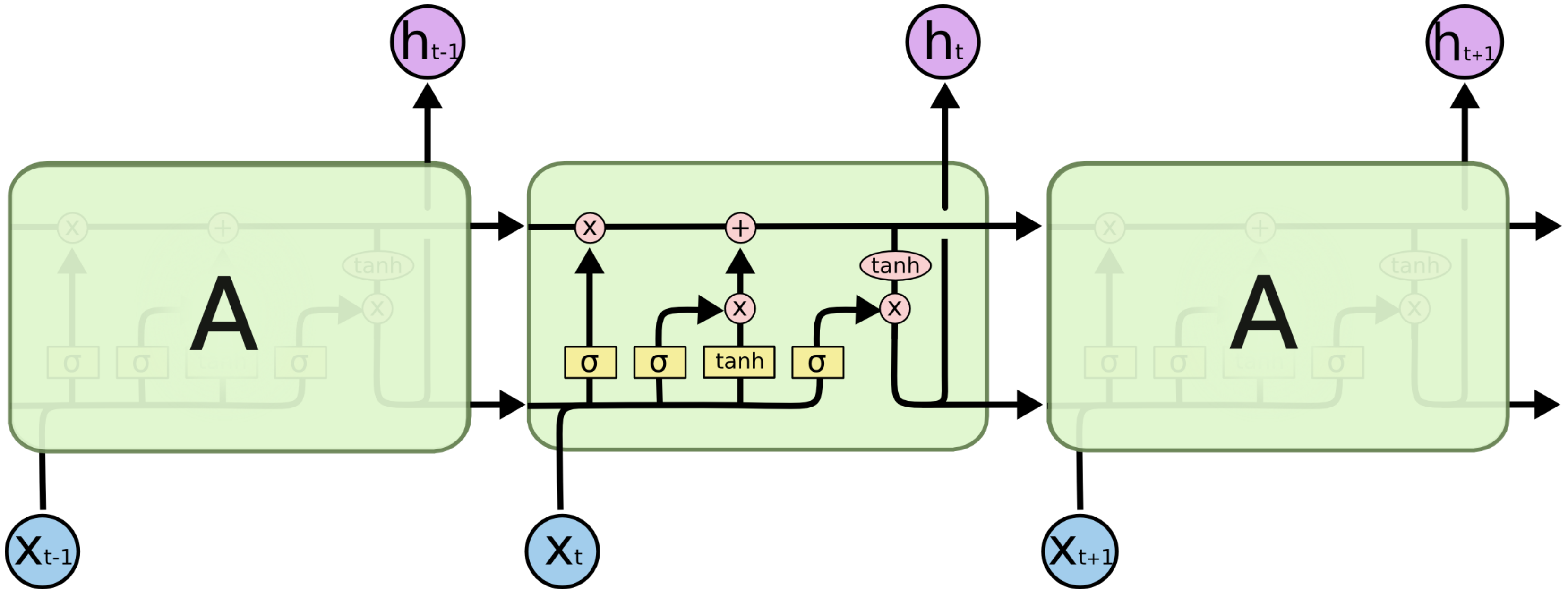
In practice, RNNs are not capable of the “long-term dependencies”, so it can not keep the remote memory’s influence.

What is LSTM?

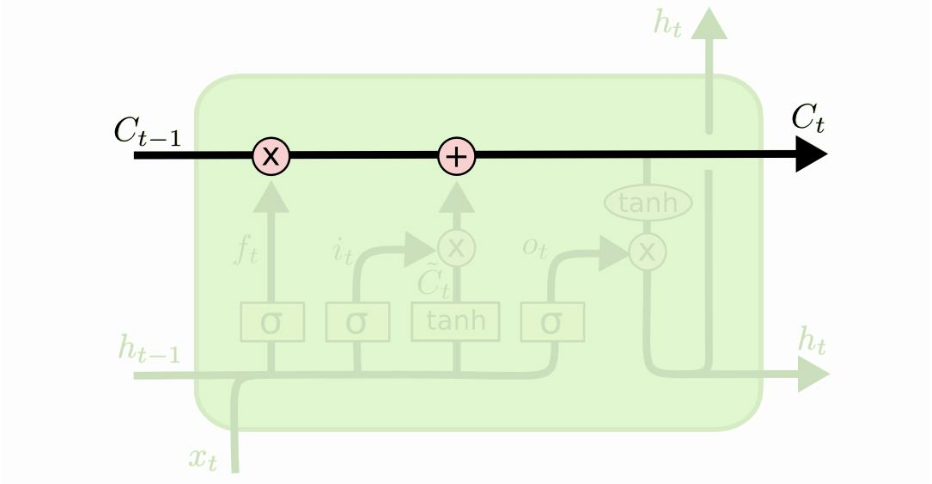
Long Short Term Memory Network is a special RNN, capable of learning the long term dependencies.

Compared with the standard Recurrent Neural Network, what is LSTM’s differences?

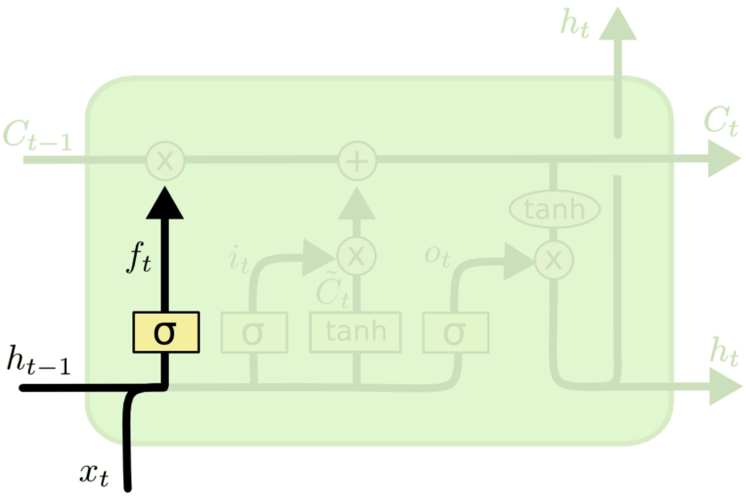
- It can use the long term memory to compute the final output.
- The repeating module in an LSTM contains three gates.



h_{t-1}
h_t
h_{t+1}
A
A
x
+
tanh
x
x
σ
σ
tanh
σ
x_{t-1}
x_t
x_{t+1}

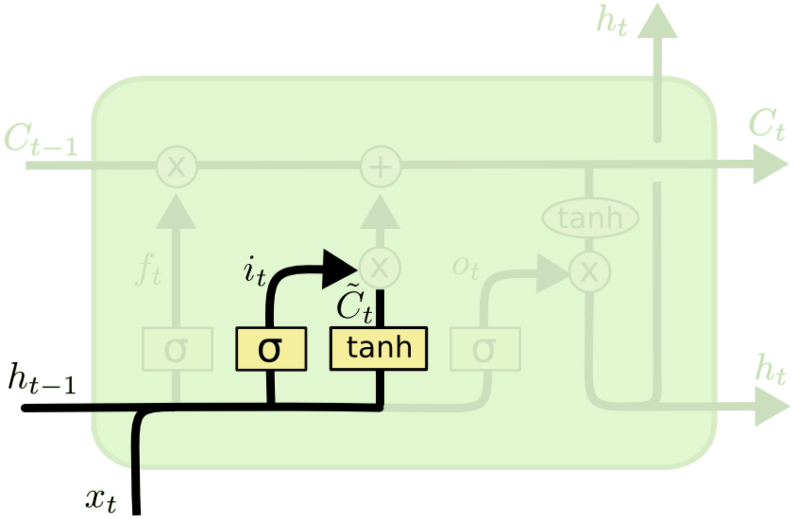


Cell State: It runs straight down the entire chain, with only some minor linear interaction.



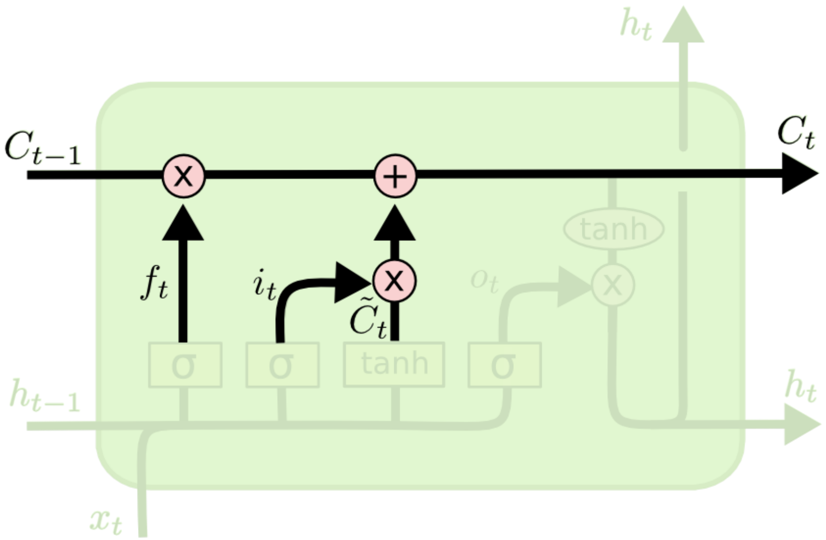
Forget Gate: To decide what information is going to throw away from the cell state.

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



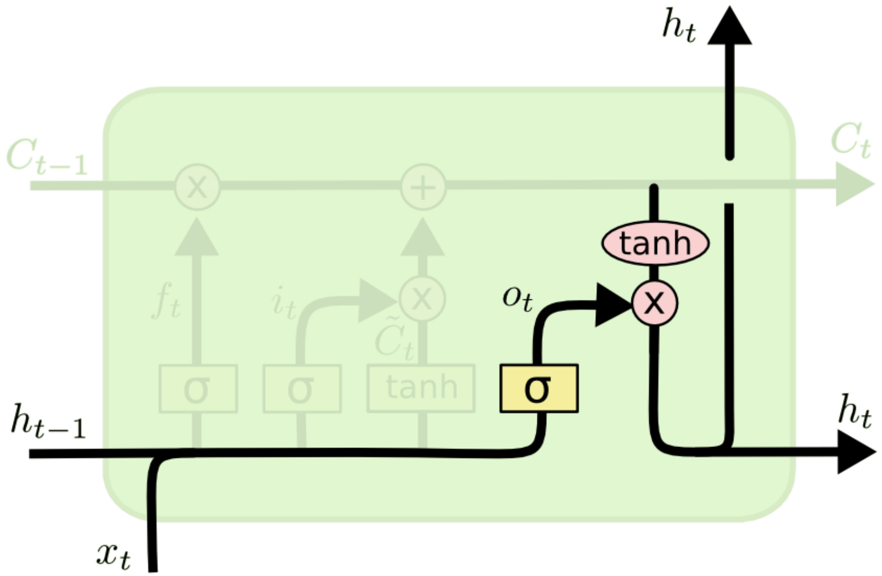
Input Gate: To decide what information is going to store in the cell.

$$i_t = \sigma(W_i[h_t, x_t] + b_i)$$
$$\widetilde{c_t} = tanh(W_C \cdot [h_t, x_t] + b_C)$$



Then we update the cell state

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t}$$



Output Gate: To decide what information is going to output

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