Long Short-Term Memory (LSTM) Networks

Prof. Dr. Eng. Nícolas de Araújo Moreira

nicolas.araujom@gmail.com

Teleinformatics Engineering Department
Universidade Federal do Ceará

25 de outubro de 2022

UNIVERSIDADE FEDERAL DO CEARÁ

Sumário

UNIVERSIDADE FEDERAL DO CEARÁ

Recurrent Neural Networks (RNN)

- **Recurrent Neural Networks:** they are networks with loops in them, allowing information to persist. A Recurrent Neural Network, at its almost fundamental level, is simply a type of Densely Connected Neural Network with a key difference: the output of the hidden layer in a recurrent neural network is fed back into itself. The hidden layer outputs are passed through a conceptual delay block, allowing to model time or sequence-dependent data.

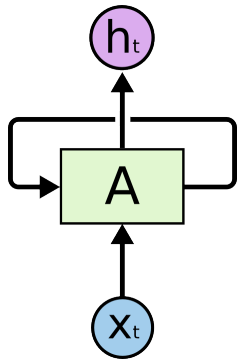
Recurrent Neural Networks (RNN)

- Consider the following representation of a recurrent neural network:

$$\mathbf{h}_g = \sigma(\mathbf{U}\mathbf{x}_t + \mathbf{V}\mathbf{h}_{t-1}) \tag{1}$$

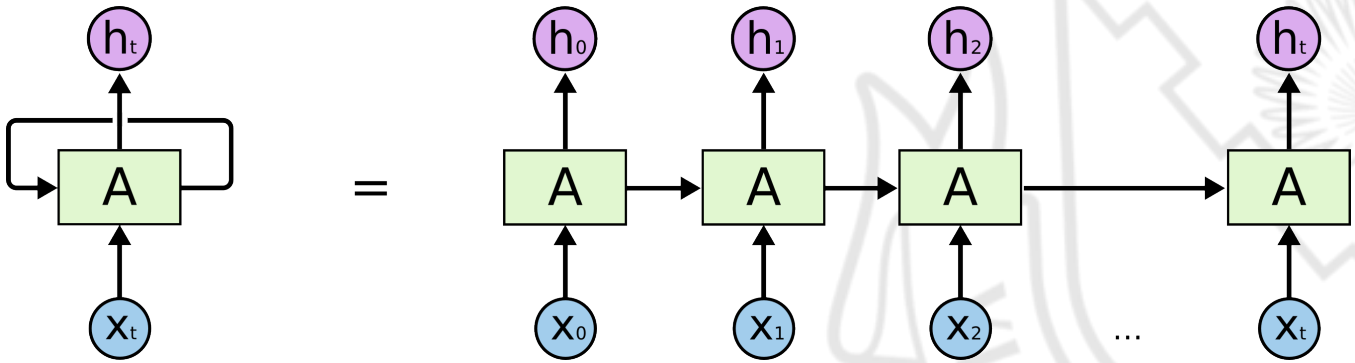
where **U** and **V** are the weight matrices connecting the inputs and the recurrent outputs respectively.

  - Vanishing gradient problem.



Recurrent Neural Networks (RNN)

- Can be though of as a multiple copies of same network, each passing a message to successor:

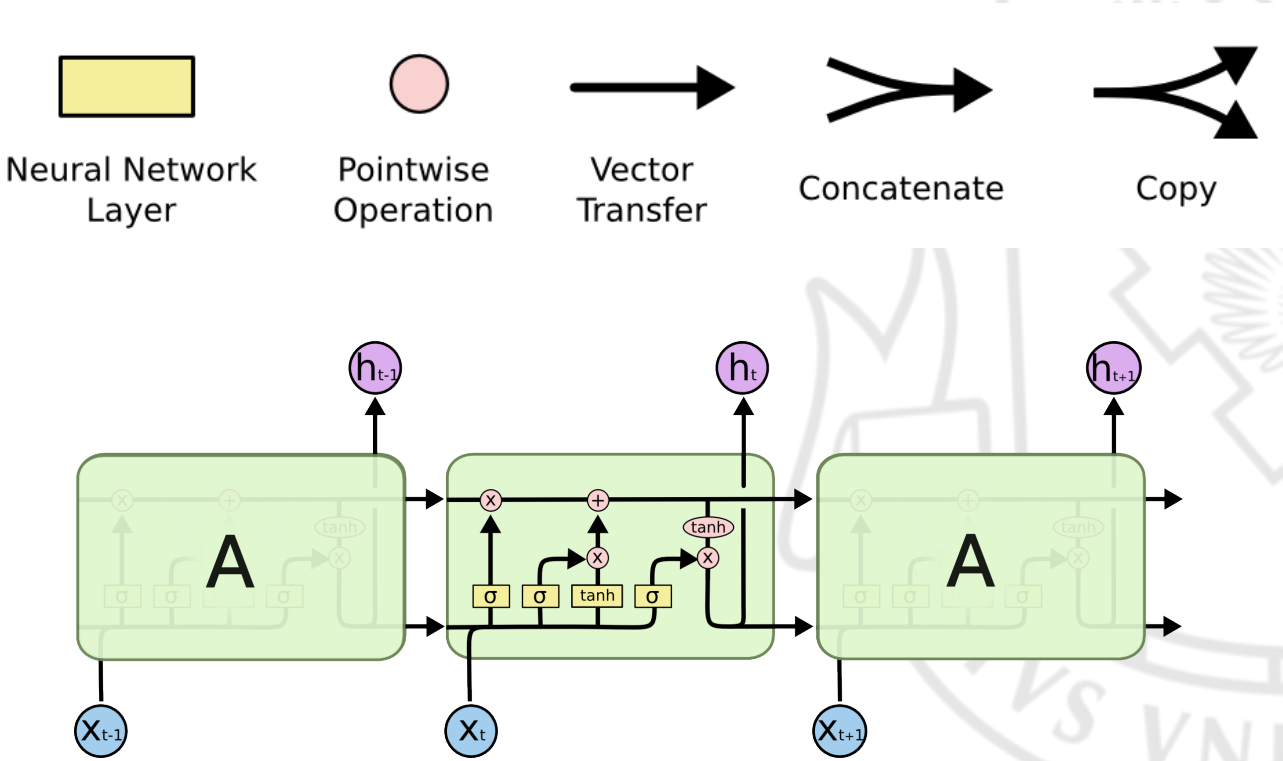


Long Short Term Memory (LSTM) Networks - Introduction and Overview

- Long Short-Term Memory (LSTM) Network is a type of recurrent neural network used in deep learning.
- The time dependence and effects of previous inputs are controlled by an interesting concept called "forget gate"which determines which states are remembered or forgotten. Two other gates are the "input gate"and "output gate"are also featured in LSTM cells.
- Application: Time-series analysis.

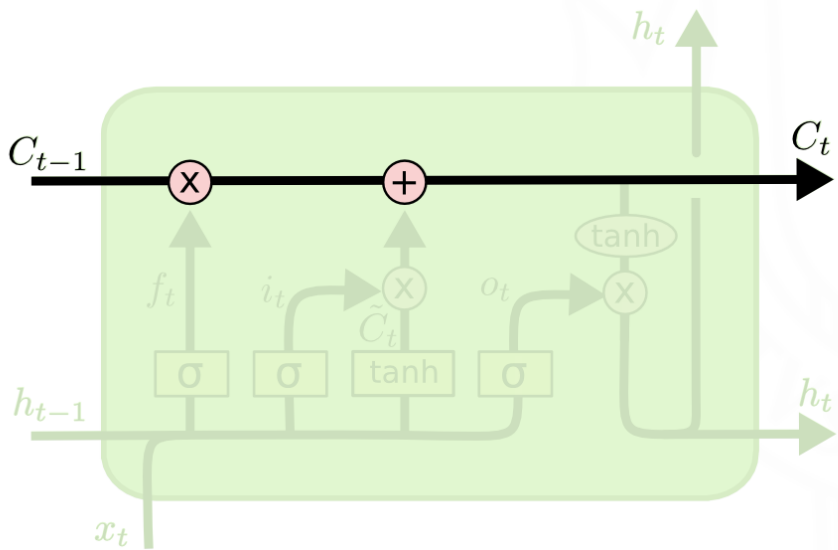
UNIVERSIDADE FEDERAL DO CEARÁ

Long Short Term Memory (LSTM) Networks - Introduction and Overview



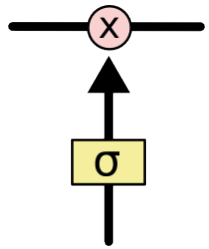
Long Short Term Memory (LSTM) Networks - Gates

- The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates;
- The data flow is from left to right in the diagram below, with te current input x_t and the previous cell output h_{t-1} concatenated together and entering the top "data rail".



Long Short Term Memory (LSTM) Networks - Gates

- Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation;
- The sigmoid layer outputs numbers between zero and one, describing how much f each component should be let through. A value of zero means "let nothing through", while a value of one means "let everything through".



Long Short Term Memory (LSTM) Networks - Gates

- An LSTM module (or cell) has five essential components which allows it to model both long-term data:
  - **Cell state (c_t):** Represents the internal memory of the cell which stores both short term memory and long-term memories;
  - **Hidden state (c_t):** This is output state information calculated w.r.t. current input, previous hidden state and current cell input which you eventually use to predict data. Additionally, the hidden state can decide to only retrieve the short or long-term or both types of memory store in the cell state to make the next prediction;
  - **Input gate (i_t):** Conditionally decides which values from the input to update the memory state, or, in other words, conditionally decides what information to forget, or, in other words, decides how much information from current input flows to the cell state;
  - **Forget gate (f_t):** Decides how much information from the current input and the previous cell state flows into the current cell state;

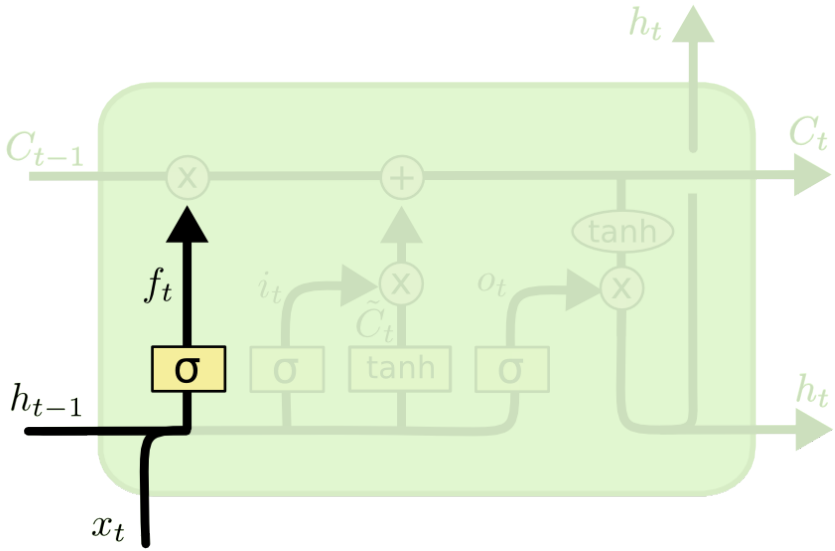
Long Short Term Memory (LSTM) Networks - Gates

- **Output gate (o_t):** Conditionally decides what to output according to the input ant the memory of block, or, in other words, decides how much information from the current cell state flows into the hidden state, so that if need LSTM can only pick the long-term memories or short-term memories and long-term memories.

UNIVERSIDADE FEDERAL DO CEARÁ

Long Short Term Memory (LSTM) Networks - Overview of Data Flow

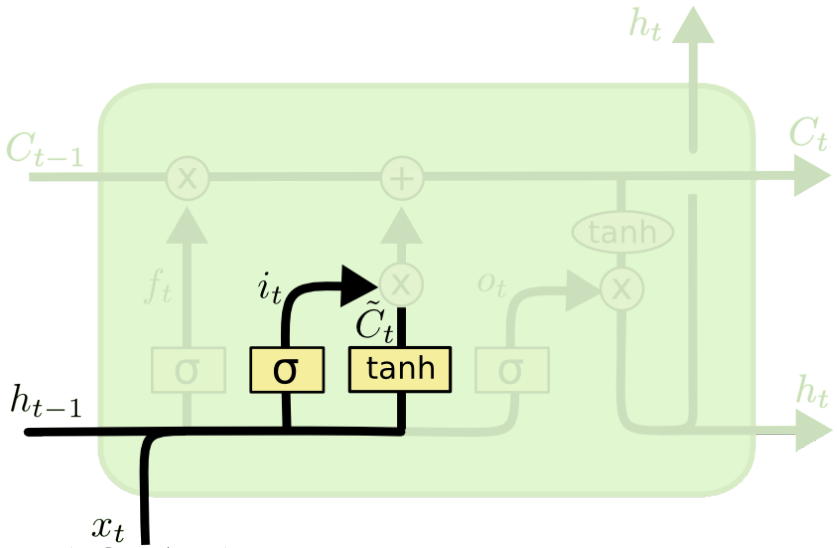
- The first step in our LSTM is to decide what information we are going to throw away from the cell state.



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

Long Short Term Memory (LSTM) Networks - Overview of Data Flow

- The next step is to decide what new information we are going to store:
  - **Sigmoid layer:** which values we'll update;
  - **Tanh layer:** Creates a vector of new candidates $\tilde{C}_t$ to be added to state.



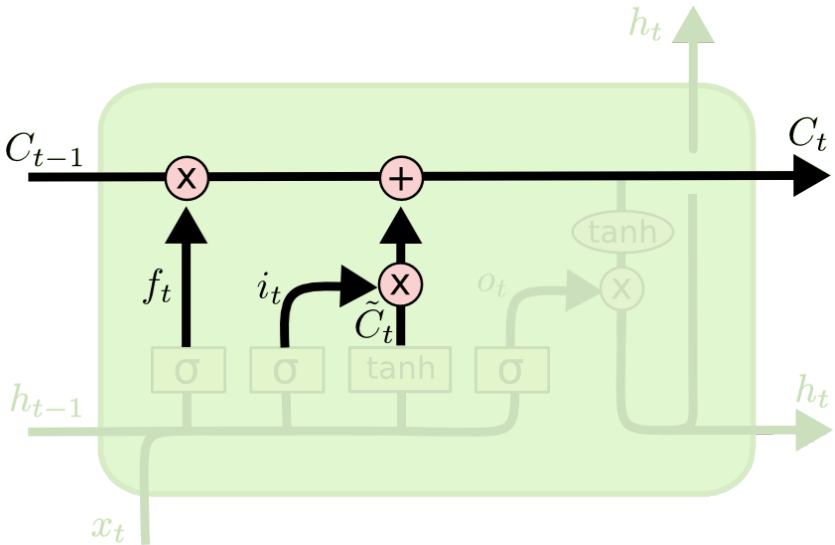
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

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

UNIVERSIDADE FEDERAL DO CEARÁ

Long Short Term Memory (LSTM) Networks - Overview of Data Flow

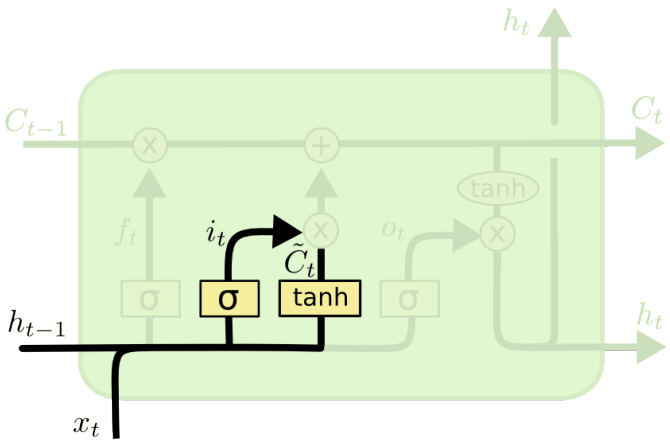
- We update the old cell state, C_{t-1}, into the new cell state C_t;
- We multiply the old state by f_t, forgetting the things we decided to forget earlier. Then we add $i_t \tilde{C}_t$. This is the new candidate values.



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

Long Short Term Memory (LSTM) Networks - Overview of Data Flow

- Finally, we need to decide what we are going to output. This output will be based on our cell state, but will be a filtered version. First, we run a sigmoid layer which decides what parts of the cell state we are going to output. Then, we put the cell state through tanh (to push the values to be between -1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.



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

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

Long Short Term Memory (LSTM) Networks - Input Gate

- The input is squashed between -1 and 1 using $\tanh$ activation function. This can be expressed by:

$$g = \tanh(b^g + x_t U^g + h_{t-1} V^g) \tag{2}$$

where U^g and V^g denote the weights for the input and previous cell output, respectively, and b^g the input bias. Note that the exponents g are not raised a power, but rather signify that these are the input weights and bias values. This squashed input is then multiplied element-wise by the output of the input gate. The input gate is basically a hidden layer of sigmoid activated nodes, with weighted x_t and h_{t-1} input values, which outputs values of between 0 and 1 and when multiplied element-wise by the input determines which inputs are switched on and off. In other words, it is a kind of input filter or gate. The expression for the input gate is:

$$i = \sigma(b^i + x_t U^i + h_{t-1} V^i) \tag{3}$$

Long Short Term Memory (LSTM) Networks - Input Gate

- The output of the input stage of the LSTM cell can be expressed below, where the $\circ$ operator express element-wise multiplication: $g \circ i$. As is possible to observe, the input gate output i acts as the weights for the squashed input g.

Long Short Term Memory (LSTM) Networks - Internal State and the Forget Gate

- s_t denotes which is the inner state of the LSTM cell. This state is delayed by one-time step and is ultimately added to the $g \circ i$ input to provide an internal recurrence loop to learn the relationship between inputs separated by time. There is a forget gate here - this forget gate is again a sigmoid activated set of nodes which is element-wise multiplied bu s_{t-1} to determine which should be remembered (i.e. forget gate output close to 1) and which should be forgotten (i.e. forget gate output close to 0). This allows the LSTM cell to learn appropriate context. The forget gate is expressed by:

$$f = \sigma(b^f + x_t U^f + h_{t-1} V^f) \tag{4}$$

The output of the element-wise product of the previous state and the forget gate is expressed as $s_{t-1} \circ f$. Again, the forget gate output acts as weights for the internal state. The output from this stage s_t is expressed by:

$$s_t = s_{t-1} \circ f + g \circ i \tag{5}$$

Long Short Term Memory (LSTM) Networks - Output Gate

- The output gate is the final stage of the LSTM cell. The output gate has two components - another tanh squashing function and an output sigmoid gating function. The output sigmoid gating function, like the other gating functions in the cell, is multiplied by the squashed state s_t to determine which values of the state are output from the cell. The output gate is expressed as:

$$o = \sigma(b^o + x_t U^o + h_{t-1} V^o) \tag{6}$$

So, the final output of the cell can be expressed as:

$$h_t = tanh(s_t) \circ o \tag{7}$$

Long Short Term Memory (LSTM) Networks - Variations

- We only forget when we are going to input something in its place. We only input new values to the state when we forget something older.

$$C_t = f_t C_{t-1} + (1 - f_t)\tilde{C}_t \tag{8}$$

Long Short Term Memory (LSTM) Networks - Window Method

- **Window Method:** Recent time steps are used to make the prediction for the next time step. For example, given the current time t we want to predict the value at the next time in the sequence $t + 1$, we can use the current time t, as well as the two prior times ($t - 1$ and $t - 2$) as input variables.

Thank you for your attention. Questions?

UNIVERSIDADE FEDERAL DO CEARÁ

Contact:

nicolas.araujom@gmail.com

UNIVERSIDADE FEDERAL DO CEARÁ