

Long short-term memory (LSTMs)



LSTMs

Deep Learning A-Z



© SuperDataScience

Long Short-Term Memory

Today:

Deep Learning A-Z



© SuperDataScience

Long Short-Term Memory

Today:

- A bit of history

Long Short-Term Memory

Today:

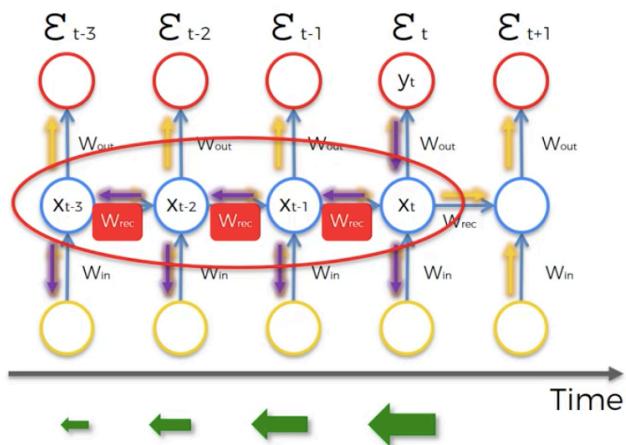
- A bit of history
- LSTM Architecture

Long Short-Term Memory

Today:

- A bit of history
- LSTM Architecture
- Example walkthrough

Long Short-Term Memory



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\partial \mathcal{E}_t}{\partial \theta} \quad (3)$$

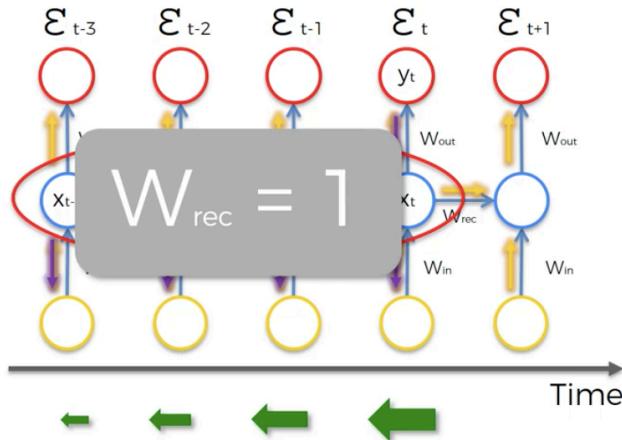
$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \leq k \leq t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_k} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial \mathbf{x}_k}{\partial \theta} \right) \quad (4)$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \geq i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \geq i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \quad (5)$$

$\mathbf{W}_{rec} \sim \text{small}$ Vanishing
 $\mathbf{W}_{rec} \sim \text{large}$ Exploding

Formula Source: Razvan Pascanu et al. (2013)

Long Short-Term Memory



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\partial \mathcal{E}_t}{\partial \theta} \quad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \leq k \leq t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \quad (4)$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{i \geq i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{i \geq i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \quad (5)$$

$W_{rec} < 1$	Vanishing
$W_{rec} > 1$	Exploding

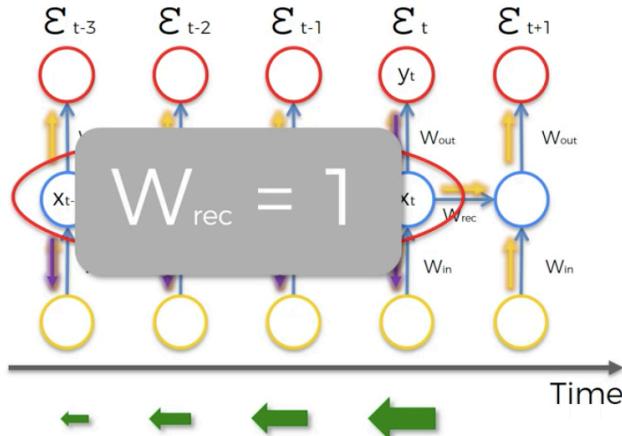
Formula Source: Razvan Pascanu et al. (2013)

Deep Learning A-Z

© SuperDataScience

The first thing that comes to mind for solving this problem and that's exactly what has been done in LSTM.

Long Short-Term Memory



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\partial \mathcal{E}_t}{\partial \theta} \quad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \leq k \leq t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \quad (4)$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{i \geq i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{i \geq i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \quad (5)$$

$W_{rec} < 1$	Vanishing
$W_{rec} > 1$	Exploding

Formula Source: Razvan Pascanu et al. (2013)

Deep Learning A-Z

© SuperDataScience

Long Short-Term Memory

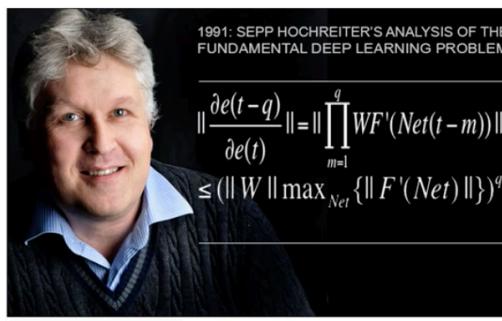
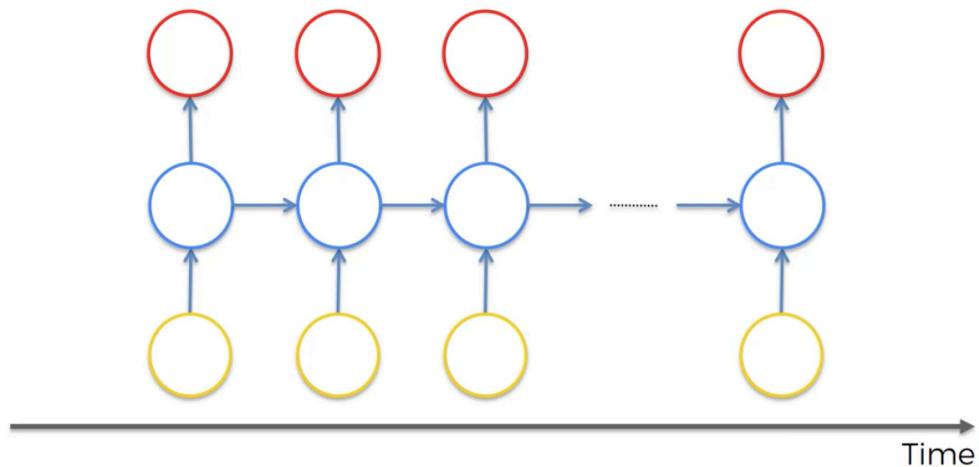


Image Sources: people.idsia.ch, ics.usi.ch

Deep Learning A-Z

© SuperDataScience

Long Short-Term Memory



Deep Learning A-Z

© SuperDataScience

Recurrent neural network on ravelled temporal loop

Long Short-Term Memory

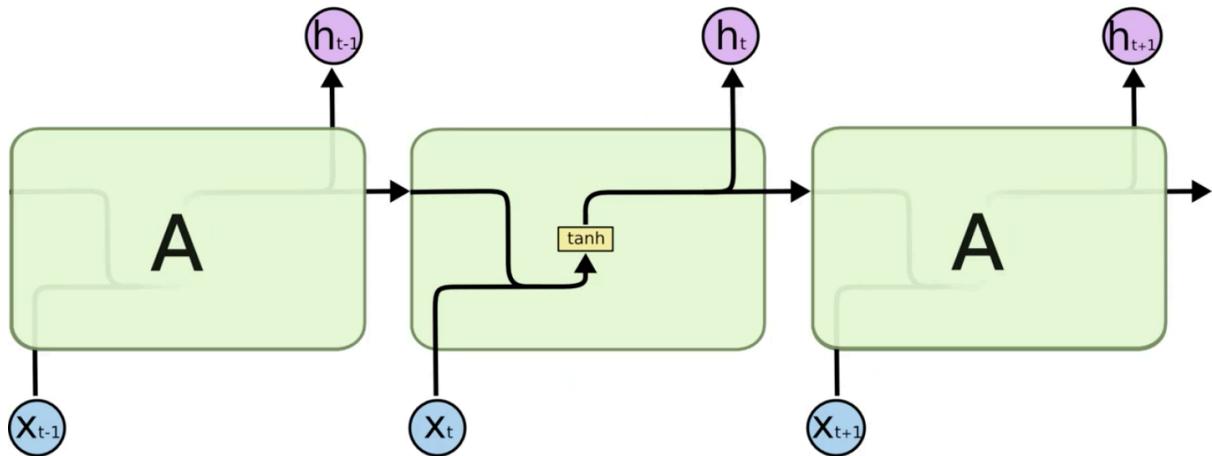


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

This is what it's look like if dig in inside the recurrent neural network.
This is standard RNN.

A screenshot of a blog post titled "Understanding LSTM Networks" by colah. The post is dated August 27, 2015. The main content discusses Recurrent Neural Networks (RNNs), noting that humans have persistence in thinking while traditional neural networks do not. It explains how RNNs address this by having loops. A diagram on the left shows a standard RNN cell 'A' receiving input x_{t-1} and producing output h_t . A smaller diagram at the bottom shows a general RNN structure with a loop from the output of one step back to the input of the next. The right side of the screenshot shows a portion of the same diagram as the one above, illustrating the flow of information through multiple time steps.

Long Short-Term Memory

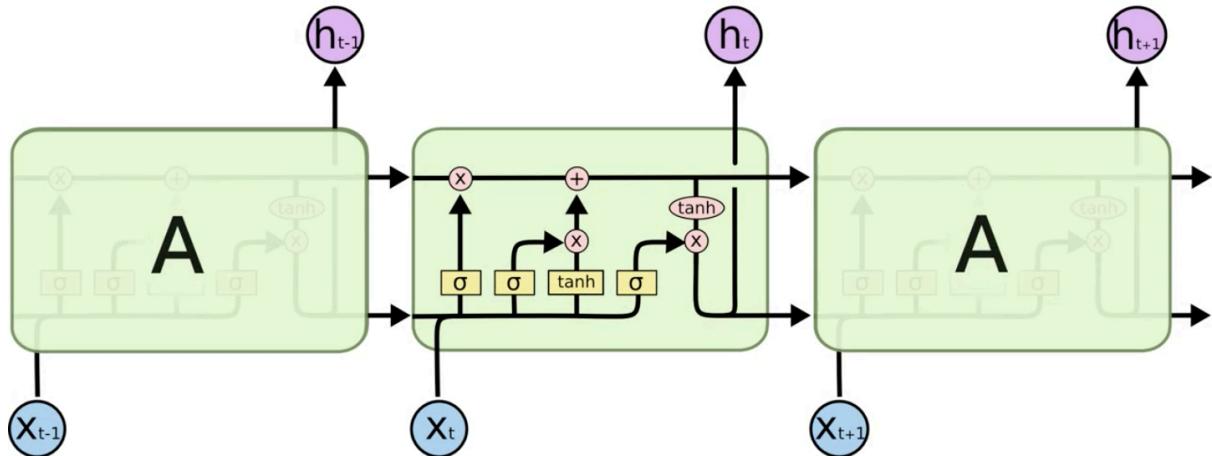


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

This LSRM version.

Long Short-Term Memory

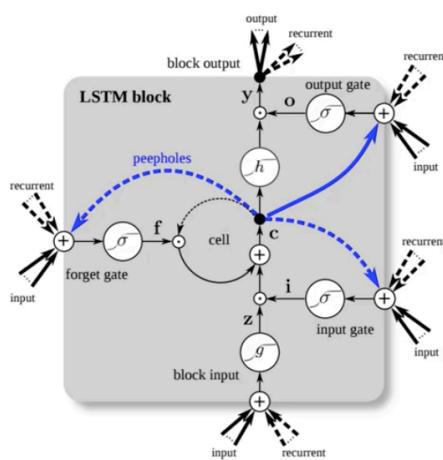


Image Source: arxiv.org/pdf/1503.04069.pdf

Deep Learning A-Z

© SuperDataScience

What usually shown in other sources.

Long Short-Term Memory

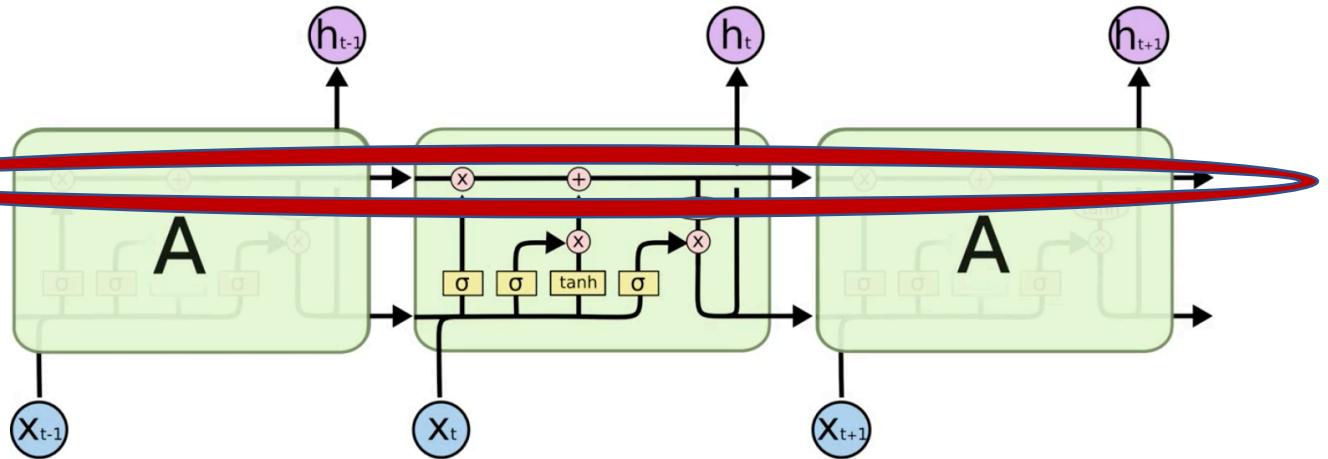


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

When we said W_{rec} is equal to 1, that this line over here. So LSTM has a memory cell (or unofficially memory pipeline) which as it goes through time, it can freely flow through time from left to right or in backpropagation right to left. Sometimes it can be removed (X sign) sometimes it can be added to thing.

Long Short-Term Memory

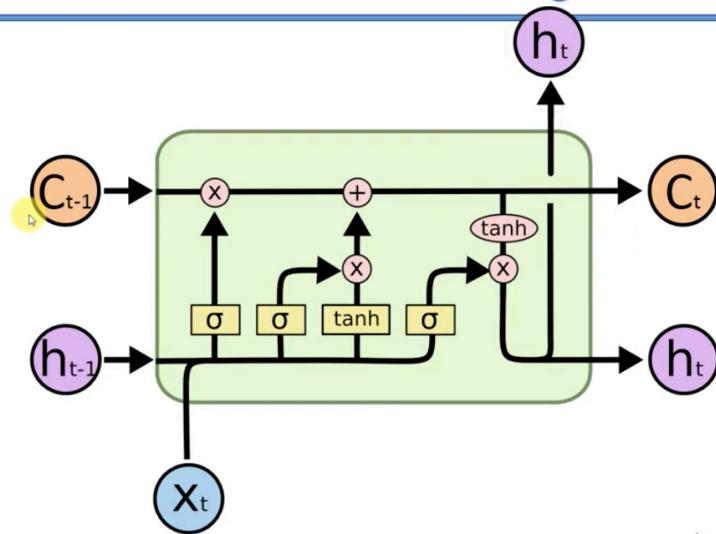


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

C stands for memory (cell). h is our output which is going to world (the upward) or it's going to the next modulo or block (the horizontals). X_t is our input.

This has 3 inputs and 2 outputs (because the h 's are essentially the same).

Reminder: there is not just 5 neurons in here, there is lots of values behind. We refer them as vectors

Long Short-Term Memory

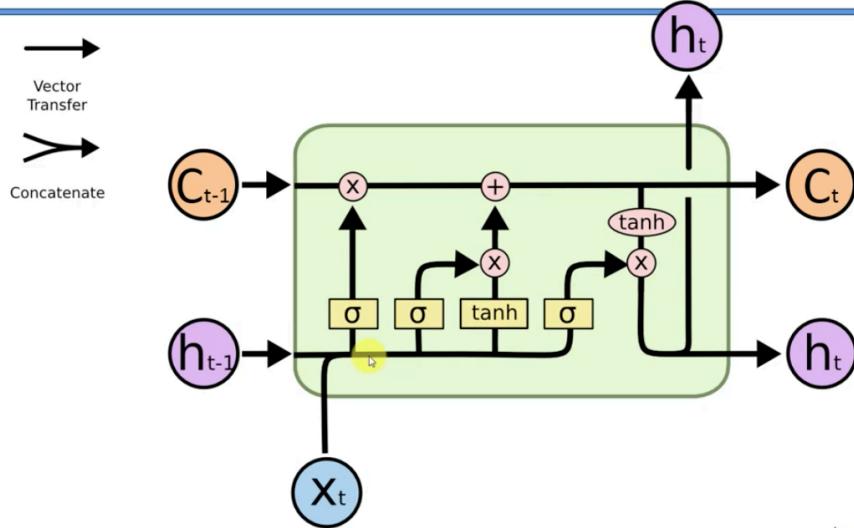


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

The concatenate: combining two lines in parallel

Long Short-Term Memory

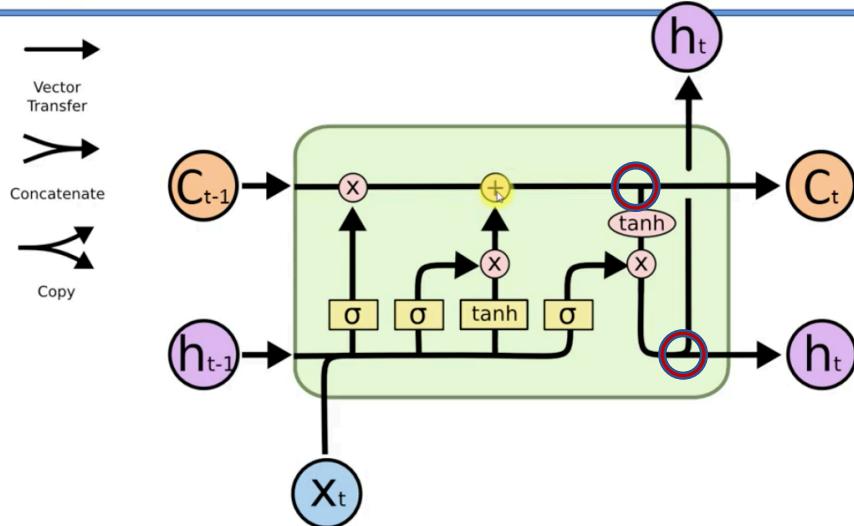


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

Long Short-Term Memory

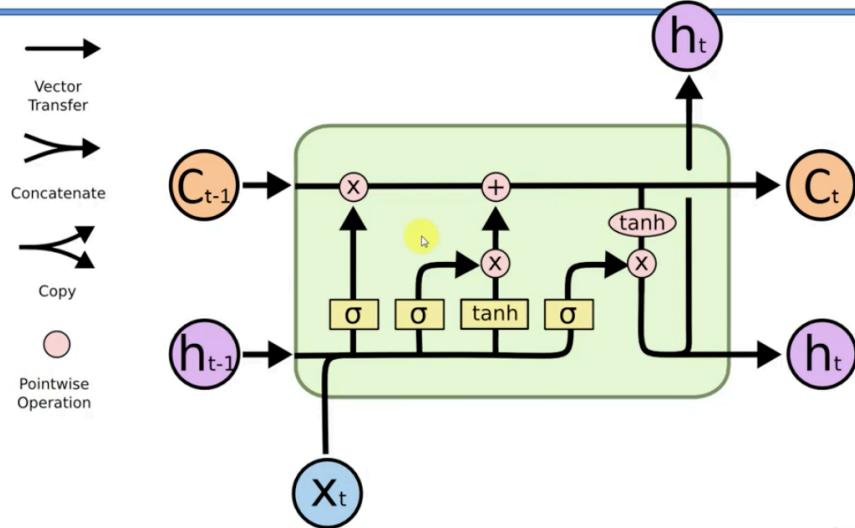


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

X's are Valves. X's from left to right: 1. Forget valve (f) 2. Memory valve (v) 3. Output valve (o)

Long Short-Term Memory



Deep Learning A-Z

© SuperDataScience

This is valve

Long Short-Term Memory

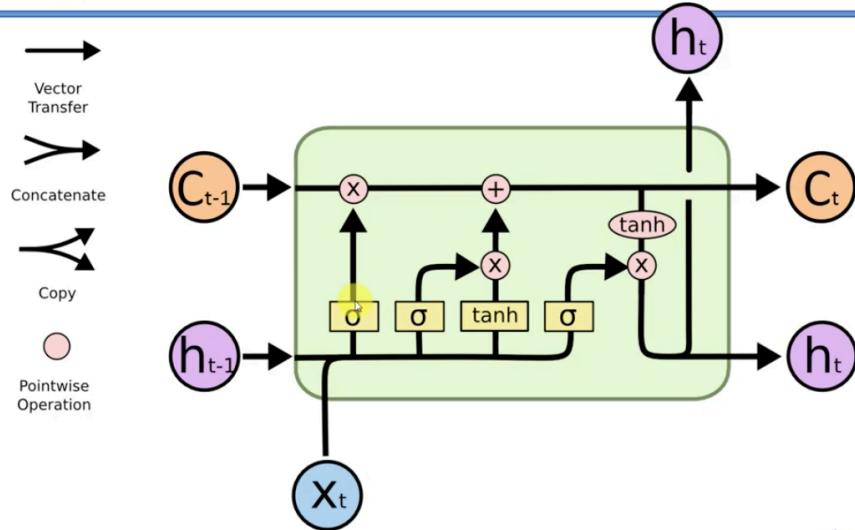


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

If you notice values are controlled by sigma which it stands for sigmoid activation function. The reason for using sigmoid is because it is between 0 and 1. (+) is like a T-shaped pipe.

Long Short-Term Memory



Deep Learning A-Z

© SuperDataScience

T-shaped pipe

Long Short-Term Memory

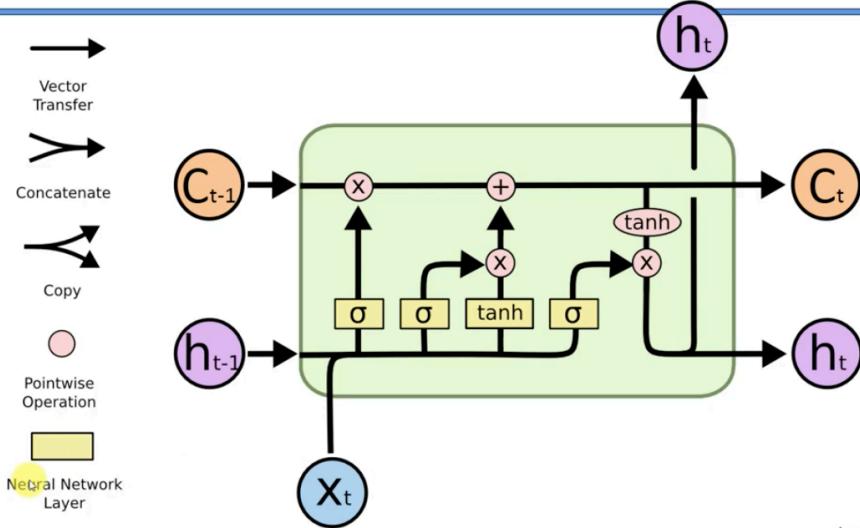


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

Long Short-Term Memory

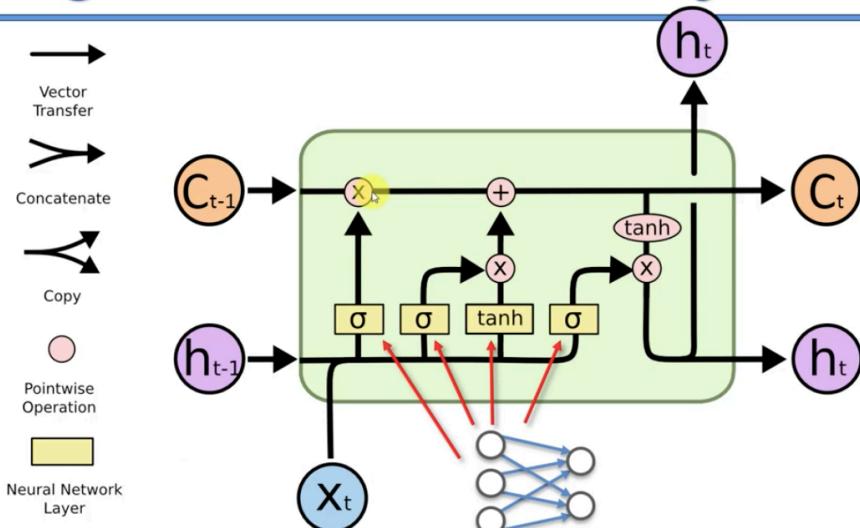


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

Long Short-Term Memory

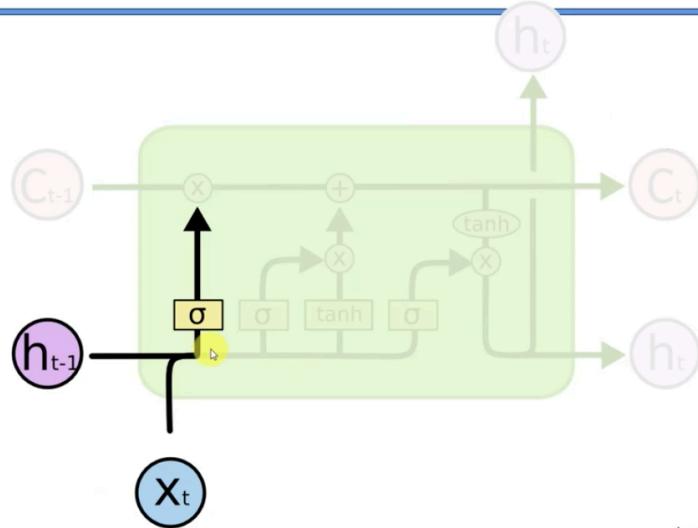


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

In here we got a value coming in and also a value coming from a previous node which together are combined to decide whether the valve (X) should go ahead or not.

Long Short-Term Memory

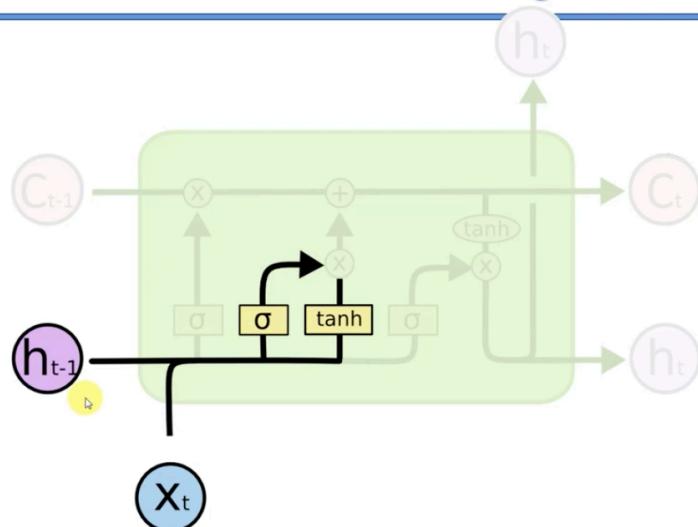


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

These two are combine again to decide first which value pass through and also if that value should pass through or not.

Long Short-Term Memory

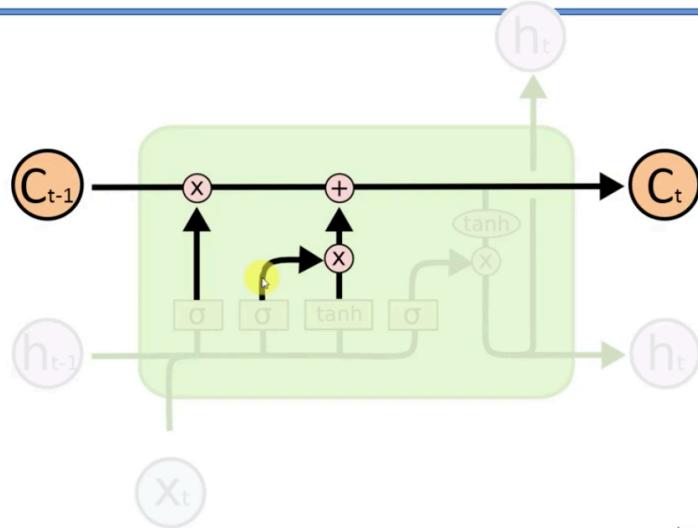


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

Long Short-Term Memory

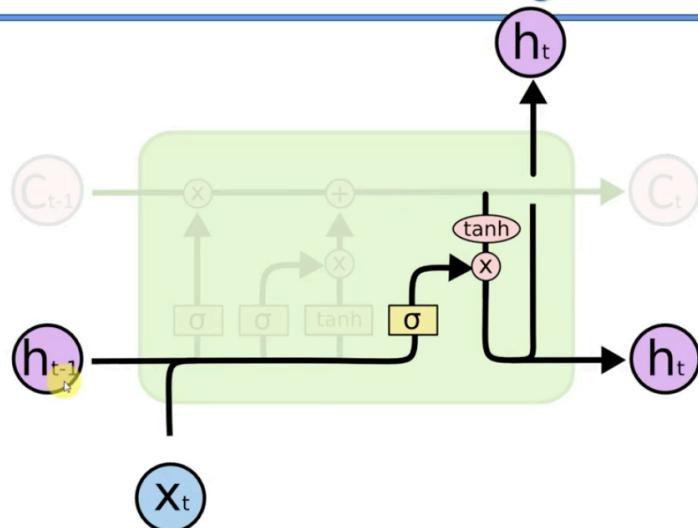


Image Source: colah.github.io

Deep Learning A-Z

© SuperDataScience

And finally, these two are combined to decide what part of memory cell or pipeline is going to be output.

Long Short-Term Memory

The image shows a screenshot of the Google Translate interface. At the top, there's a navigation bar with the Google logo, user profile, and a 'Translate' button. Below this, there are two input fields: one for English and one for Czech. The English input field contains the sentence "I am a boy who likes to learn". The Czech output field shows the translated sentence "Jsem kluk, který rád učit". Both fields have small icons for microphone, keyboard, and other options, and a character count of 29/5000.

Deep Learning A-Z

© SuperDataScience

Long Short-Term Memory

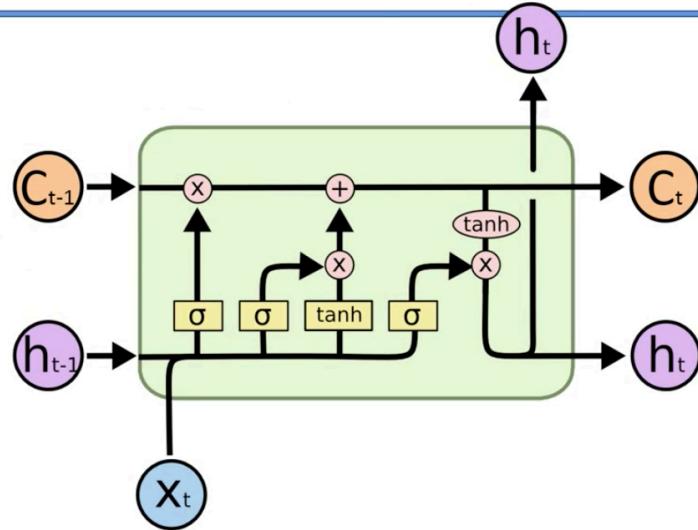
This screenshot is identical to the one above, but with a red underline and a yellow circle highlighting the word "boy" in the English input field. This visual cue likely indicates that the model has focused on or processed this specific word during its translation process.

Deep Learning A-Z

© SuperDataScience

In here if we change the boy to girl, the structure of sentence will change.

Long Short-Term Memory



Deep Learning A-Z

© SuperDataScience

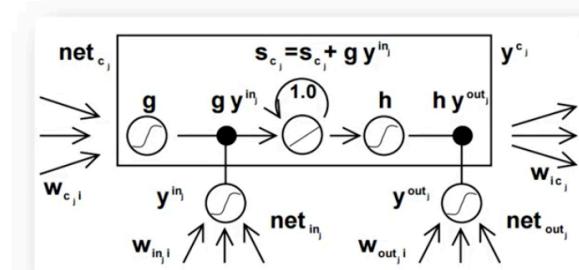
Let's say boy go through the C freely. If we add a girl or a name like Amanda, it comes from X. then we close the first valve we had. Then open the second valve then put the subject on memory cell and also, we extract more information out of it (for example if it's feminine, singular, capitalized, etc.). then it flows through the memory cell and on the third valve, it will extract the information of its gender so it goes as a input to the next modulo to make the necessary changes in there.

Long Short-Term Memory

Additional Reading:

Long Short-Term Memory

By Sepp Hochreiter & Jurgen Schmidhuber (1997)



Link:

<http://www.bioinf.jku.at/publications/older/2604.pdf>

Deep Learning A-Z

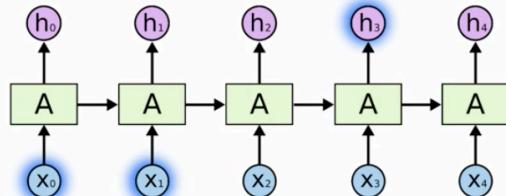
© SuperDataScience

Long Short-Term Memory

Additional Reading:

Understanding LSTM Networks

By Christopher Olah (2015)



Link:

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

Deep Learning A-Z

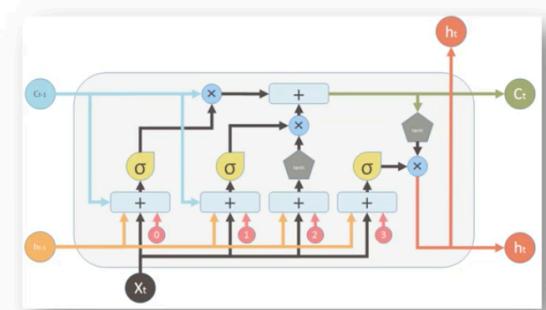
© SuperDataScience

Long Short-Term Memory

Additional Reading:

Understanding LSTM and its diagrams

By Shi Yan (2016)



Link:

<https://medium.com/@shiyans/understanding-lstm-and-its-diagrams-37e2f46f1714>

Deep Learning A-Z

© SuperDataScience
