

# Understanding LSTM Networks

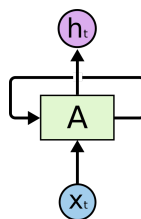
Posted on August 27, 2015

## Recurrent Neural Networks

Humans don't start their thinking from scratch every second. As you read this essay, you understand each word based on your understanding of previous words. You don't throw everything away and start thinking from scratch again. Your thoughts have persistence.

Traditional neural networks can't do this, and it seems like a major shortcoming. For example, imagine you want to classify what kind of event is happening at every point in a movie. It's unclear how a traditional neural network could use its reasoning about previous events in the film to inform later ones.

Recurrent neural networks address this issue. They are networks with loops in them, allowing information to persist.

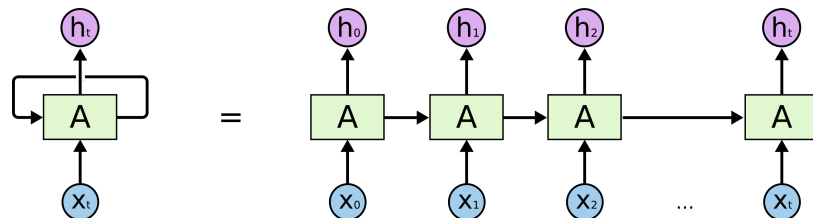


**Recurrent Neural Networks have loops.**

8

In the above diagram, a chunk of neural network,  $A$ , looks at some input  $x_t$  and outputs a value  $h_t$ . A loop allows information to be passed from one step of the network to the next.

These loops make recurrent neural networks seem kind of mysterious. However, if you think a bit more, it turns out that they aren't all that different than a normal neural network. A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor. Consider what happens if we unroll the loop:



**An unrolled recurrent neural network.**

This chain-like nature reveals that recurrent neural networks are intimately related to sequences and lists. They're the natural architecture of neural network to use for such data.

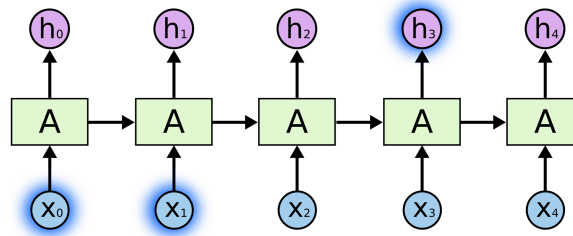
And they certainly are used! In the last few years, there have been incredible success applying RNNs to a variety of problems: speech recognition, language modeling, translation, image captioning... The list goes on. I'll leave discussion of the amazing feats one can achieve with RNNs to Andrej Karpathy's excellent blog post, The Unreasonable Effectiveness of Recurrent Neural Networks (<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>). But they really are pretty amazing.

Essential to these successes is the use of "LSTMs," a very special kind of recurrent neural network which works, for many tasks, much much better than the standard version. Almost all exciting results based on recurrent neural networks are achieved with them. It's these LSTMs that this essay will explore.

## The Problem of Long-Term Dependencies

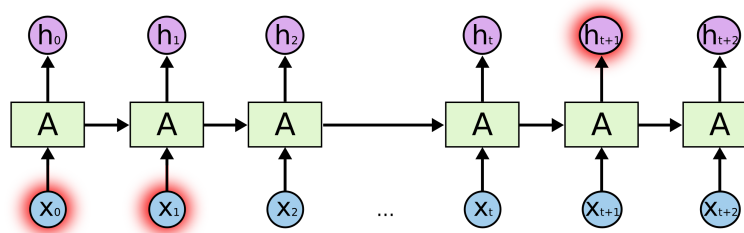
One of the appeals of RNNs is the idea that they might be able to connect previous information to the present task, such as using previous video frames might inform the understanding of the present frame. If RNNs could do this, they'd be extremely useful. But can they? It depends.

Sometimes, we only need to look at recent information to perform the present task. For example, consider a language model trying to predict the next word based on the previous ones. If we are trying to predict the last word in "the clouds are in the sky," we don't need any further context – it's pretty obvious the next word is going to be sky. In such cases, where the gap between the relevant information and the place that it's needed is small, RNNs can learn to use the past information.



But there are also cases where we need more context. Consider trying to predict the last word in the text “I grew up in France... I speak fluent *French*.” Recent information suggests that the next word is probably the name of a language, but if we want to narrow down which language, we need the context of France, from further back. It’s entirely possible for the gap between the relevant information and the point where it is needed to become very large.

Unfortunately, as that gap grows, RNNs become unable to learn to connect the information.



In theory, RNNs are absolutely capable of handling such “long-term dependencies.” A human could carefully pick parameters for them to solve toy problems of this form. Sadly, in practice, RNNs don’t seem to be able to learn them. The problem was explored in depth by Hochreiter (1991) [German] (<http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf>) and Bengio, et al. (1994) (<http://www-dsi.ing.unifi.it/~paolo/ps/tmn-94-gradient.pdf>), who found some pretty fundamental reasons why it might be difficult.

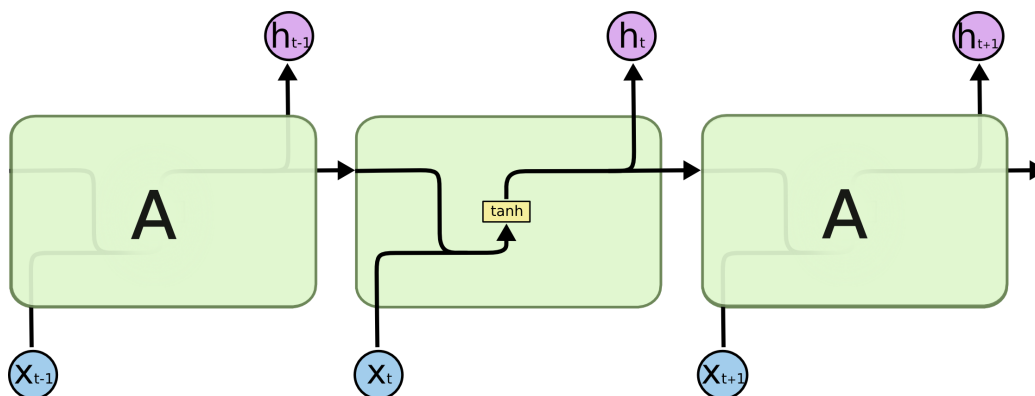
Thankfully, LSTMs don’t have this problem!

## LSTM Networks

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997) (<http://www.bioinf.jku.at/publications/older/2604.pdf>), and were refined and popularized by many people in following work.<sup>1</sup> They work tremendously well on a large variety of problems, and are now widely used.

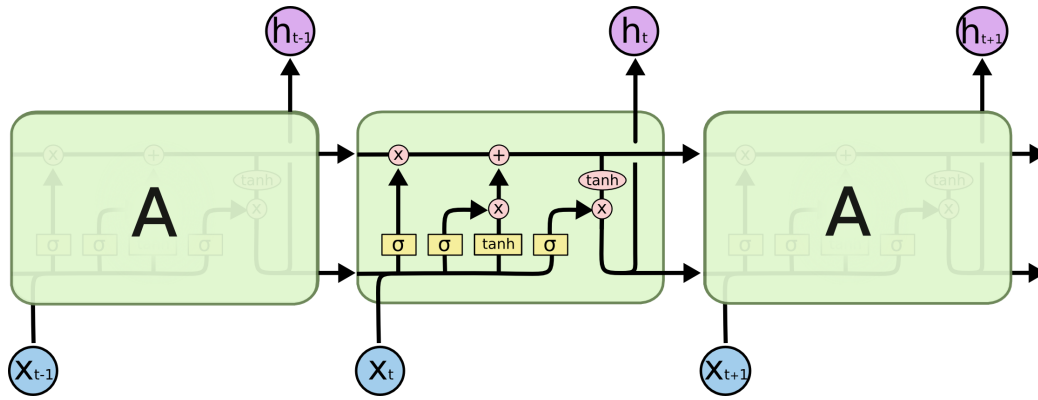
LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.



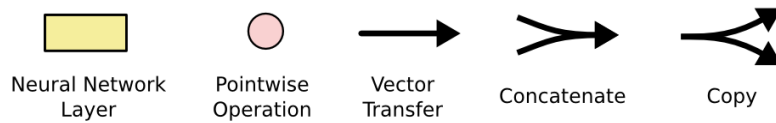
The repeating module in a standard RNN contains a single layer.

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.



The repeating module in an LSTM contains four interacting layers.

Don't worry about the details of what's going on. We'll walk through the LSTM diagram step by step later. For now, let's just try to get comfortable with the notation we'll be using.

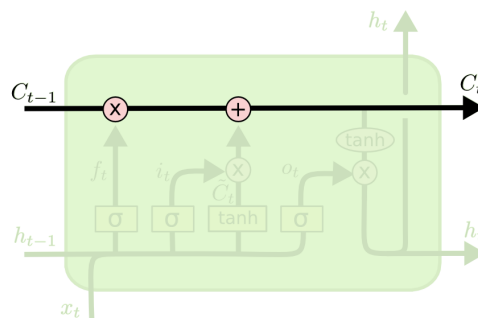


In the above diagram, each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural network layers. Lines merging denote concatenation, while a line forking denote its content being copied and the copies going to different locations.

## The Core Idea Behind LSTMs

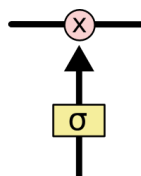
The key to LSTMs is the cell state, the horizontal line running through the top of the diagram.

The cell state is kind of like a conveyor belt. It runs straight down the entire chain, with only some minor linear interactions. It's very easy for information to just flow along it unchanged.



The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called 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.



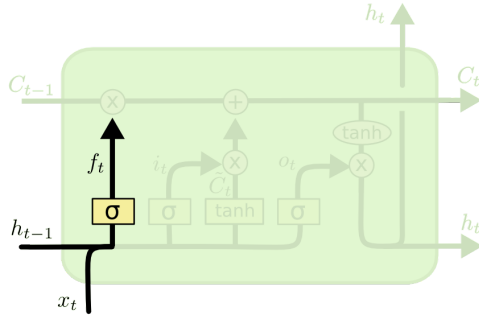
The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through. A value of zero means "let nothing through," while a value of one means "let everything through!"

An LSTM has three of these gates, to protect and control the cell state.

## Step-by-Step LSTM Walk Through

The first step in our LSTM is to decide what information we're going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer." It looks at  $h_{t-1}$  and  $x_t$ , and outputs a number between 0 and 1 for each number in the cell state  $C_{t-1}$ . A 1 represents "completely keep this" while a 0 represents "completely get rid of this."

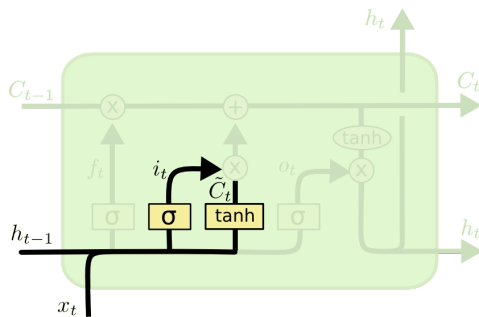
Let's go back to our example of a language model trying to predict the next word based on all the previous ones. In such a problem, the cell state might include the gender of the present subject, so that the correct pronouns can be used. When we see a new subject, we want to forget the gender of the old subject.



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

The next step is to decide what new information we're going to store in the cell state. **This has two parts.** First, a sigmoid layer called the "input gate layer" decides which values we'll update. Next, a tanh layer creates a vector of new candidate values,  $\tilde{C}_t$ , that could be added to the state. In the next step, we'll combine these two to create an update to the state.

In the example of our language model, we'd want to add the gender of the new subject to the cell state, to replace the old one we're forgetting.



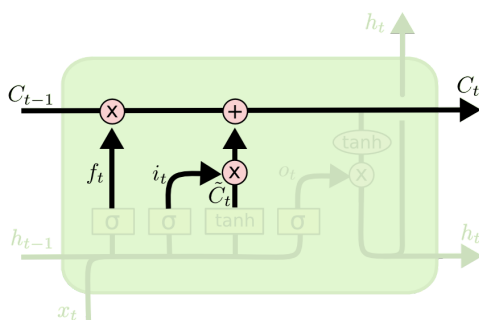
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

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

It's now time to update the old cell state,  $C_{t-1}$ , into the new cell state  $C_t$ . The previous steps already decided what to do, we just need to actually do it.

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, scaled by how much we decided to update each state value.

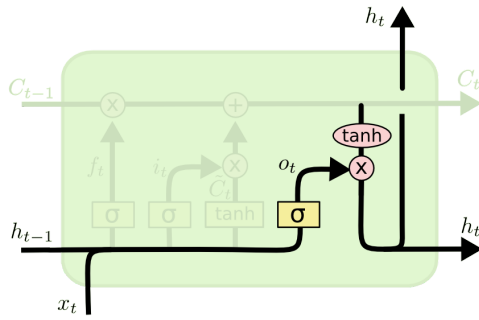
In the case of the language model, this is where we'd actually drop the information about the old subject's gender and add the new information, as we decided in the previous steps.



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

Finally, we need to decide what we're 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're 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.

For the language model example, since it just saw a subject, it might want to output information relevant to a verb, in case that's what is coming next. For example, it might output whether the subject is singular or plural, so that we know what form a verb should be conjugated into if that's what follows next.



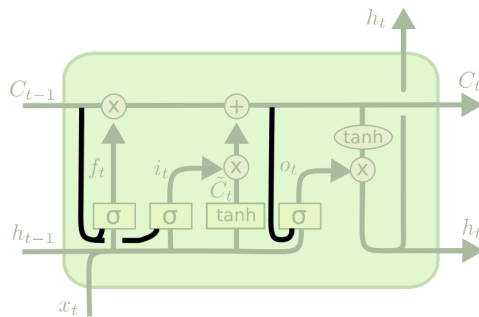
$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$$

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

## Variants on Long Short Term Memory

What I've described so far is a pretty normal LSTM. But not all LSTMs are the same as the above. In fact, it seems like almost every paper involving LSTMs uses a slightly different version. The differences are minor, but it's worth mentioning some of them.

One popular LSTM variant, introduced by Gers & Schmidhuber (2000) (<ftp://ftp.idsia.ch/pub/juergen/TimeCount-IJCNN2000.pdf>), is adding “peephole connections.” This means that we let the gate layers look at the cell state.



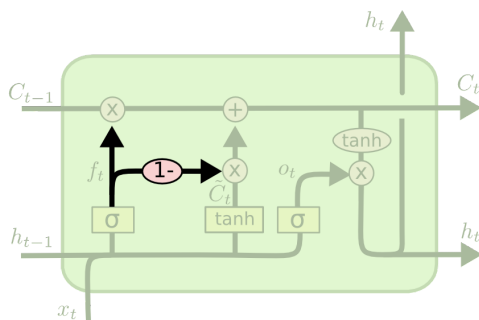
$$f_t = \sigma(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f)$$

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

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

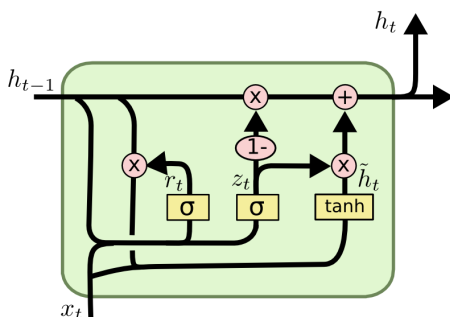
The above diagram adds peepholes to all the gates, but many papers will give some peepholes and not others.

Another variation is to use coupled forget and input gates. Instead of separately deciding what to forget and what we should add new information to, we make those decisions together. We only forget when we're going to input something in its place. We only input new values to the state when we forget something older.



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

A slightly more dramatic variation on the LSTM is the Gated Recurrent Unit, or GRU, introduced by Cho, et al. (2014) (<http://arxiv.org/pdf/1406.1078v3.pdf>). It combines the forget and input gates into a single “update gate.” It also merges the cell state and hidden state, and makes some other changes. The resulting model is simpler than standard LSTM models, and has been growing increasingly popular.



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$



## 402 Comments (/posts/2015-08-Understanding-LSTMs/#disqus\_thread)

Comments Community  Login ▾

 Recommend 357  Tweet  Share

Sort by Best ▾

Join the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS 

Name



**Brandon Rohrer** • 3 years ago



Thank you for the clear explanation! It's a rare that someone both has the ability to write so clearly and takes the time to do so. Your work is a gift to the rest of us.

196   • Reply • Share >



**Enish Paneru** → Brandon Rohrer  
• 10 months ago



I just finished your video on RNN and LSTM minutes ago.  
Your work is a gift as well. Thank you very much for your effort.

4   • Reply • Share >



**Yogi Tri Cahyono** → Brandon Rohrer  
• 2 years ago



true

1   • Reply • Share >



**이인목** → Brandon Rohrer • 2 years ago

absolutly

1   • Reply • Share >



**Karthik Srinivasan** → Brandon Rohrer  
• 3 months ago

Absolutely!

  • Reply • Share >



**Harry** → Brandon Rohrer • a year ago

So true :)

  • Reply • Share >



**Josue Valdez** → Brandon Rohrer  
• a year ago

I like your videos

  • Reply • Share >



**qinlu zhang** → Josue Valdez  
• a year ago



can you post the link of his video

  • Reply • Share >



**Josue Valdez** → qinlu zhang  
• a year ago

https://disq.us/url?url=htt...

8   • Reply • Share >



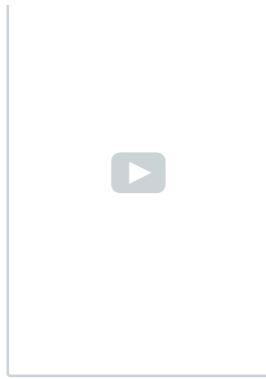
**George William** → Josue Valdez  
• a year ago

can you post the link of his video

  • Reply • Share >



**Josue Valdez** → George William  
• a year ago



1 ^ | v • Reply • Share >



**Dieka Nugraha** • 2 years ago

Thank you for the most informative and clear post about LSTM in the net with superb visualization! about the dimensionality of weight matrices,  $W_f$ ,  $W_i$ ,  $W_o$  have same shape, which is dimension of  $[h_{t-1}, x_t]$  X dimension of  $C_t$  right?

58 ^ | v • Reply • Share >



**chrisolah** Mod → Dieka Nugraha  
• 2 years ago

That's right! :)

1 ^ | v • Reply • Share >



**Faiyaz Lalani** • 2 years ago

Nice post Chris - just wanted to thank you.

49 ^ | v • Reply • Share >



**Long Ye** • a year ago

What is the structure like if there are 128 units in A, no a single unit?

48 ^ | v • Reply • Share >



**rohan chikorde** • 2 years ago

Thank you for the nice and clear explanation. I really appreciate that. As I am new to this model and learning so just had few doubts:

- 1) What is  $b(c)$ ,  $b(i)$ ,  $b(f)$  in the 1st, 2nd, 3rd equations? what "b" stands for?
- 2)  $W(f)$ ,  $W(i)$ ,  $W(o)$  are neural network matrix but how are they created and why are there 3 separate matrix? Please put some light on this.
- 3) In one of the equation, we are using tanh and why not sigmoid? Do you use tanh only for range -1 to 1?

Please explain these doubts. It will help us to understand this model more clearly. Thanks in advance.

48 ^ | v • Reply • Share >



**chrisolah** Mod → rohan chikorde  
• 2 years ago

1)  $b_c$ ,  $b_i$ ,  $b_f$  ... are neural network bias vectors. They are initialized with random numbers, and learned as the network trains.

2)  $W_f$ ,  $W_i$ ,  $W_o$  ... are neural network weight matrices. They are initialized with random numbers, and learned as the network trains.

3) Yes, we use tanh for the range (-1,1)

2 ^ | v • Reply • Share >



**Priyansh Agrawal** → chrisolah  
• 4 months ago

As it was mentioned in this paper :- <http://proceedings.mlr.press...> [An Empirical Exploration of Recurrent Network Architectures] that we cant randomly initialised the  $h$  as it may

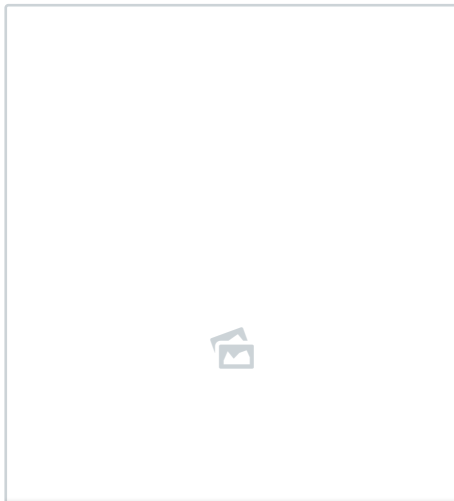


randomly initialized the  $b_f$  as it may get smaller values due to which "vanishing gradient" problem occur. If the bias of the forget gate is not properly initialized, we may erroneously conclude that the LSTM is incapable of learning to solve problems with long-range dependencies, which is not the case since LSTM don't suffer from "vanishing gradient". One can set  $b_f$  to 1 or 2 as suggested in the aforementioned paper.

2 ^ | v • Reply • Share >



Jan Alberts • 2 years ago



see more

35 ^ | v • Reply • Share >



chrisolah Mod → Jan Alberts • 2 years ago

Hi Jan -- that's a great question!

In a normal recurrent neural network, every part of the state is connected to every part of the state.

LSTMs are slightly different. The LSTM cell states don't directly modify themselves. They're carefully protected by design, so that their default behavior is to not change. But every cell does effect the forget gate, input gate, and tanh later. Together, all those layers decide how to change the cell state.

So, all the cell states can change each other, but they do it indirectly, through the special layers we've set up.

1 ^ | v • Reply • Share >



ศิริชัย หลอดสีลา • 2 years ago

ฉันดีใจที่เพื่อนๆติดตามการคิดค้นข้อมูลของจีน และขอบคุณเพื่อนๆที่ช่วยเหลือฉันมาตลอด

30 ^ | v • Reply • Share >



li kai • 4 years ago

Great!Great!Great!Great!Great!Great!Great!Great!G

Much better than Alex Graves's paper!!!!

27 ^ | v • Reply • Share >



Daniel Bigham → li kai • 3 years ago

LOL. I was just reading Alex Grave's paper and started to frown with the lack of explanation on the core LSTM idea. I googled, found this article, and your comment was at the top. Hilarious.

16 ^ | v • Reply • Share >



**Dan Quang** • 4 years ago

Thank you! I've been trying to understand LSTM for a long time! This made it very clear :)

32 ^ | v • Reply • Share >



**chrisolah** Mod → Dan Quang • 4 years ago

I'm glad it helped!

4 ^ | v • Reply • Share >



**Dan Quang** → chrisolah  
• 3 years ago

I have to give a presentation soon to my department explaining LSTMs for my candidacy exam. Do you mind if I use some of these figures to explain LSTMs?

14 ^ | v • Reply • Share >



**chrisolah** Mod → Dan Quang • 3 years ago

Sorry I didn't respond earlier. For you, and anyone else reading, I'm happy for anyone to use these figures with attribution.

21 ^ | v • Reply • Share >



**Morshed Derbali** → Dan Quang • 2 years ago

hello dear ..i'm new in this domain( LSTMs) ..i'm asking you if you permit to pass the presentation to me ..Thanks in advance  
morshed.derbali28@gmail.com

^ | v • Reply • Share >



**Srinidhi Bheesette** → chrisolah  
• 2 years ago

Can these be used for particle pattern recognition?

^ | v • Reply • Share >



**Shimul Hassan Nahid** • 3 years ago

Excellent post ...

19 ^ | v • Reply • Share >



**Alex Sosnovshchenko** • 3 years ago

I made Russian translation of this excellent post, I hope you will not mind.

<http://alexsosn.github.io/m...>

14 ^ | v • Reply • Share >



**bobriakov** → Alex Sosnovshchenko  
• 3 years ago

Спасибо!

2 ^ | v • Reply • Share >



**chrisolah** Mod → Alex Sosnovshchenko  
• 3 years ago

That looks lovely, Alex! Thanks for the translation!

1 ^ | v • Reply • Share >



**Nithin Ts** → Alex Sosnovshchenko  
• 10 months ago

Did you use LSTM to build a model to translate this article ? If so please share the code

^ | v • Reply • Share >



**Anjali** • 4 months ago

I think this is one of the best introductions one can read for LSTMs before reading the papers. Thanks.

5 ^ | v • Reply • Share >



**Steven Tang** • 3 years ago

Great post! I am a little confused by the output  $h_t$  portion of the lstm.

It looks like the dimension of  $C_t$  should be the number of dimensions in  $h_{t-1}$  and  $x_t$  combined.

It also looks like  $h_t$  should be the dimension of  $C_t$ .

Obviously, my understanding can't be correct, or else the dimensionality of  $h_t$  would grow by the dimensions of  $x_t$  at every time step. Is there a mistake in my understanding somewhere? Is the output of  $h_t$  only the dimension of  $h_{t-1}$ , and NOT  $C_t$ ?

Thanks again!



6 ^ | v • Reply • Share >



**Alfredo Canziani** → Steven Tang  
• 3 years ago

$h_t$ ,  $C_t$  and  $\tilde{C}_t$  share the same dimensionality.

The size of  $\tilde{C}_t$  comes directly from the 'height' of  $W_C$  matrix (check  $\tilde{C}_t$  definition, above).  $W_C$ 's 'width' is instead equal to  $\text{size}(h_t) + \text{size}(x_t)$ .

4 ^ | v • Reply • Share >



**Steven Tang** → Alfredo Canziani  
• 3 years ago

Great, thanks for the explanation!

46 ^ | v • Reply • Share >



**Majid al-Dosari** → Alfredo Canziani  
• 3 years ago

i don't get it. everything that's not a weight matrix is a vector of the same dimension. i don't get what you mean by this concatenation thing mathematically  $W_C[h_{t-1}, x_t]$ . it seems unconventional.

do you mean 'wide' matrix times 'tall' vector b/c you put  $h$  and  $x$  on top of each other?? that would equal a tall vector which can't be right. it can't be two vectors side by side b/c then the result would be a matrix.

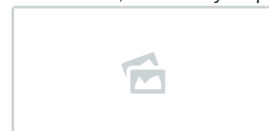
in the literature the weight matrix is separated out for each vector.

1 ^ | v • Reply • Share >



**chrisolah** Mod → Majid al-Dosari • 3 years ago

Here's a figure I made for the post but didn't include in the final version, which may help:



11 ^ | v • Reply • Share >