


361 Comments

colah's blog

 Junior Osho ▾ Recommend 298 Share

Sort by Best ▾



Join the discussion...

**Junior Osho** • a few seconds ago

Thank you for the clear explanation!..

^ | ▾ • Edit • Reply • Share >

**Brandon Rohrer** • 2 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.

169 ^ | ▾ • Reply • Share >

**Enish Paneru** → Brandon Rohrer • 2 months agoI just finished your video on RNN and LSTM minutes ago.
Your work is a gift as well. Thank you very much for your effort.

1 ^ | ▾ • Reply • Share >

**Yogi Tri Cahyono** → Brandon Rohrer • a year ago

true

1 ^ | ▾ • Reply • Share >

**이인묵** → Brandon Rohrer • a year ago

absolutly

1 ^ | ▾ • Reply • Share >

**Huy Hoang** → Brandon Rohrer • 3 months ago

So true :)

^ | ▾ • Reply • Share >

**Josue Valdez** → Brandon Rohrer • 5 months ago

I like your videos

^ | ▾ • Reply • Share >

**qinlu zhang** → Josue Valdez • 4 months ago

can you post the link of his video

^ | ▾ • Reply • Share >

**Josue Valdez** → qinlu zhang • 4 months ago<https://disq.us/url?url=htt...>

^ | ▾ • Reply • Share >

**George William** → Josue Valdez • 4 months ago

can you post the link of his video

^ | v • Reply • Share >



Josue Valdez → George William • 4 months ago

^ | 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?

26 ^ | v • Reply • Share >



chrisolah Mod → Dieka Nugraha • 2 years ago

That's right! :)

1 ^ | v • Reply • Share >



li kai • 3 years ago

Great!Great!Great!Great!Great!Great!Great!Great!Great!Great!

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

26 ^ | 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.

15 ^ | v • Reply • Share >



Long Ye • 8 months ago

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

23 ^ | v • Reply • Share >



Faiyaz Lalani • 2 years ago

Nice post Chris - just wanted to thank you.

23 ^ | 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.

22 ^ | 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 >

**Dan Quang** • 3 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 • 3 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?

17 ^ | v • Reply • Share >

**chrisolah** Mod → Dan Quang • 2 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.

20 ^ | v • Reply • Share >

**Morshed Derbali** → Dan Quang • 10 months ago

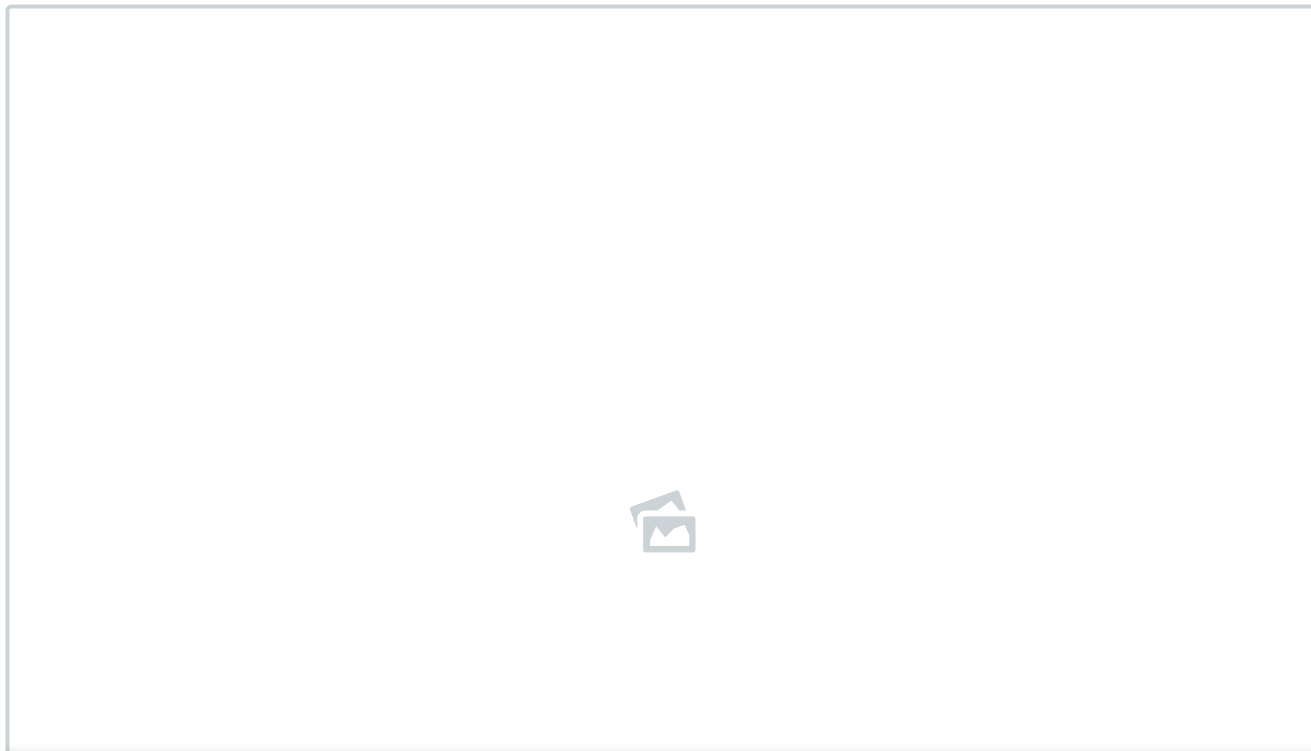
hello dear ..i'm new in this domain(LSTMs) ..i'm asking you if you permit to pass the presentation to me ..Thanks in advance
morshed.derbali28@gmail.com

^ | v • Reply • Share >

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

Can these be used for particle pattern recognition?

^ | v • Reply • Share >

**Jan Alberts** • 2 years ago[see more](#)

20 ^ | 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.

^ | 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 • 2 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 • 2 months ago

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

^ | v • Reply • Share ›

**Shimul Hassan Nahid** • 2 years ago

Excellent post ...

6 ^ | v • Reply • Share ›

**Steven Tang** • 3 years agoGreat 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!

22 ^ | 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.

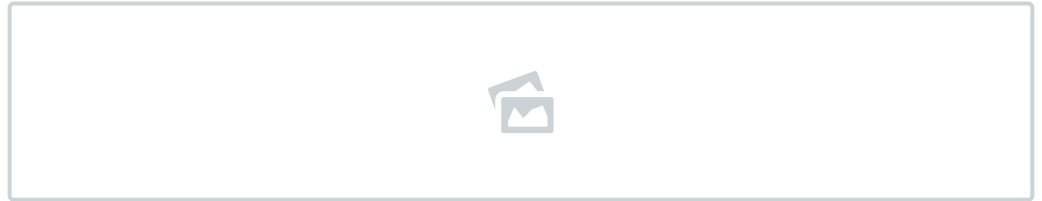
in the literature the weight matrix is separated out for each vector.

1 ^ | v • Reply • Share >



chrisolah Mod → Majid al-Dosari • 2 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 >



hughperkins → chrisolah • 2 years ago

These diagrams look really nice. What are you using to draw them?

17 ^ | v • Reply • Share >



Carl Thomé → hughperkins • 2 years ago

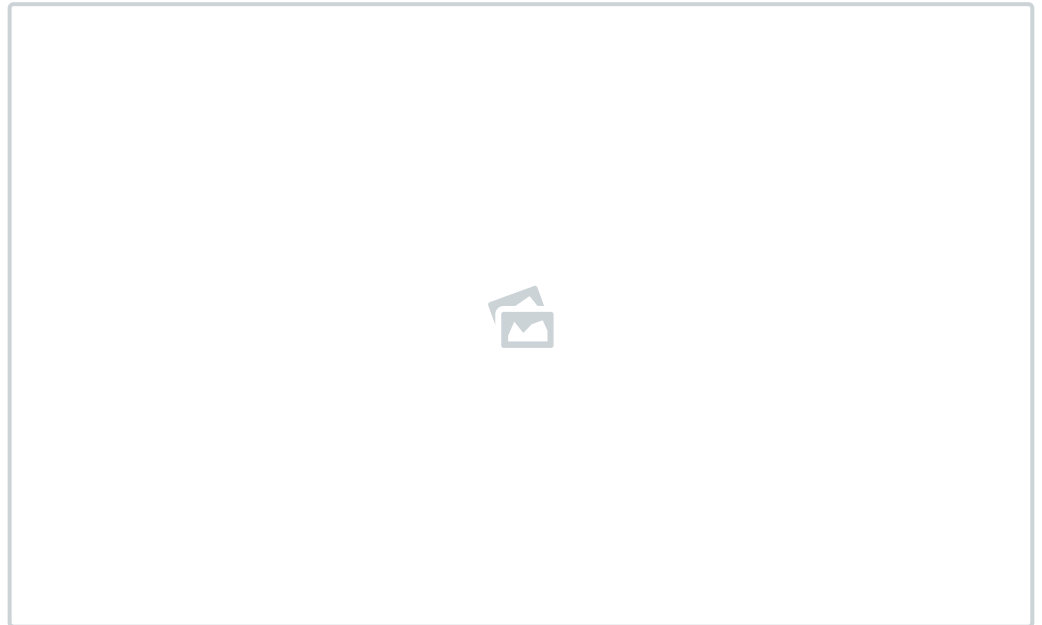
Chris mentioned he did them in Inkscape. Beautiful stuff, indeed!

2 ^ | v • Reply • Share >



Majid al-Dosari → chrisolah • 2 years ago

thx. i made a figure similar to yours actually (might need some more arrows) it's funny how in the literature the focus is usually on a 'cell' but you can't really see the data flow.



4 ^ | v • Reply • Share >



chrisolah Mod → Majid al-Dosari • 2 years ago

Nice! :)

2 ^ | v • Reply • Share >



Artsiom → chrisolah • 2 years ago

Hi Chris,

Could you please tell which software do you use to draw these histograms?

3 ^ | v • Reply • Share ›



Evalds Urtans ➔ chrisolah • 10 months ago

What did you use to draw this diagram?

1 ^ | v • Reply • Share ›



Steve_Ohr ➔ Evalds Urtans • 9 months ago

Not what he used, but Omnigraffle will do this for Macintosh users. It's excellent for diagrams.

^ | v • Reply • Share ›



Evalds Urtans ➔ Steve_Ohr • 9 months ago

Thank you, I will check it out!

1 ^ | v • Reply • Share ›



Steve_Ohr ➔ Evalds Urtans • 9 months ago

You can download it and use it for free for 30 days.

Objects have 'magnets' that lines snap to. There are pre-defined magnet locations (e.g. 2 each side) but if you hold down the 'm' key you can drag them around and change where lines attach. Lines can hop over/under other lines to and have properties like arrows, dashes, that you can give them.

Great program for any kind of diagramming.

1 ^ | v • Reply • Share ›



pluviosilla ➔ chrisolah • 7 months ago

Guess I'm dense, but I still don't see the reason for the extra lines: 2 input lines for x_t , 2 input lines for h_{t-1} , and two input lines for the state. The simpler diagrams show one input line for each of these three. Why does this diagram sport 6 input lines and precisely 6?

^ | v • Reply • Share ›



Chris Anderson ➔ chrisolah • 2 years ago

For that diagram, why is the input and hidden state split in two? I can see why the cell state is. Also, how/why do the yellow layers reduce the 4 inputs to 2 outputs?

^ | v • Reply • Share ›



chrisolah Mod ➔ Chris Anderson • 2 years ago

The input and the hidden state come from different places. (See the original diagram, where they are two different lines that merge.)

The yellow layers have 2 outputs, because they need to output the same size as the cell state.

^ | v • Reply • Share >



Phosphenes ➔ chrisolah • 3 months ago

But how does a simple sigmoid function reduce a 4-dimensional input vector to a 2-dimensional output vector?

^ | v • Reply • Share >



Chris Anderson ➔ chrisolah • 2 years ago

Thanks for the reply.

Sorry, I meant why are they each split in two - why are there two inputs and two hidden states? How do you combine the 4 to produce 2?

^ | v • Reply • Share >

[Load more comments](#)

ALSO ON COLAH'S BLOG

Groups & Group Convolutions

1 comment • 2 years ago



Ryan Gnabasik — You point out something that I think is true for vector spaces as well. In general vector spaces inherit properties from the fields ...

Visualizing Representations: Deep Learning and Human Beings

1 comment • 2 years ago



Andrew Brereton — Have you ever read Blindsight by Peter Watts? You might enjoy it. Great blog btw.

Calculus on Computational Graphs: Backpropagation

1 comment • 3 years ago



atomicthumbs — Dang it, I KNEW I shouldn't have chosen Business Math!

Distill

29 comments • a year ago



chrisolah — Thanks for the link -- that thread is a lovely surprise. We have a bunch of 90% finished Distill articles, and ~10 drafts floating

[Subscribe](#) [Add Disqus to your site](#) [Add Disqus](#) [Disqus' Privacy Policy](#) [Privacy Policy](#) [Privacy Policy](#)