

LSTM - Long Short-Term Memory

Understanding Through a Complete Example

Watch LSTM Process a Sentence

Sentence: "The cat was hungry. The dog was sleeping."

Word	Forget	Input	Output	Memory
"The"	0.9	0.3	0.2	[article]
"cat"	0.8	0.9	0.8	[cat]
"was"	0.9	0.7	0.9	[cat, was]
"hungry"	0.8	0.8	0.7	[cat, hungry]
"."	0.1	0.4	0.3	[end]
"The"	0.1	0.8	0.2	[article]
"dog"	0.7	0.9	0.9	[dog]
"was"	0.9	0.8	0.9	[dog, was]

Notice:

- Three mysterious numbers per word
- Memory changes as we read
- "cat" disappears at period
- "dog" appears as new subject

Intuition: The Magic

How does LSTM know to:

- Forget "cat" at the period?
- Remember "dog" as new subject?
- Keep the right information?

Let's find out...

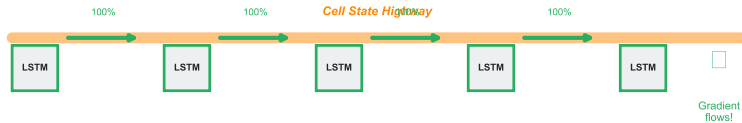
Why Do We Need This?

The Vanishing Gradient Problem

Standard RNN:



LSTM:



Key: LSTM uses addition (cell state) instead of multiplication (RNN hidden state)

RNN Problem:

LSTM Solution:

LSTM - Long Short-Term Memory

Gate 1: Forget - What to Erase?

Forget Gate: What to Erase?

Example: "The cat was hungry. The dog ..."

Inputs:

h_{t-1} : Previous output

x_t : Current word ("dog")

Forget Gate

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

Output: 0 to 1

Decision:

"cat" info  **10%** Forget! (new subject)

"hungry" info  **20%** Forget! (not relevant)

Lower values (close to 0) = FORGET
Higher values (close to 1) = KEEP

Intuition: When you see "dog", forget information about "cat"

Remember Our Table?

Row 5: "." had **Forget** = 0.1

What This Means:

Formula:

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

How It Decides:

Gate 2: Input - What to Add?

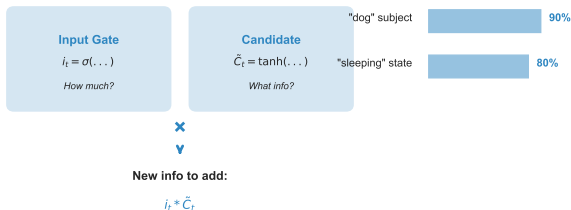
Input Gate: What to Remember?

Example: "The dog was sleeping ..."

Inputs:

h_{t-1} : Previous output

x_t : Current word ("sleeping")



Intuition: Remember "dog is sleeping" for future predictions

Remember Our Table?

Row 7: "dog" had **Input** = 0.9

What This Means:

Formulas (Two Parts):

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

Gate 3: Output - What to Share?

Output Gate: What to Output?

Example: "The dog was sleeping and ..." → predict next word

Cell State:

Contains: dog, sleeping, etc.

Question: *What's relevant NOW?*

Output Gate

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

How much to output?

Final Output:

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



To next layer / prediction

Decision:

"dog" info  **90%** Output! (subject)

"sleeping" info  **70%** Output! (state)

old context  **10%** Hide (not needed)

Intuition: Only share relevant parts of memory for current prediction

Remember Our Table?

Row 8: "was" had **Output = 0.9**

What This Means:

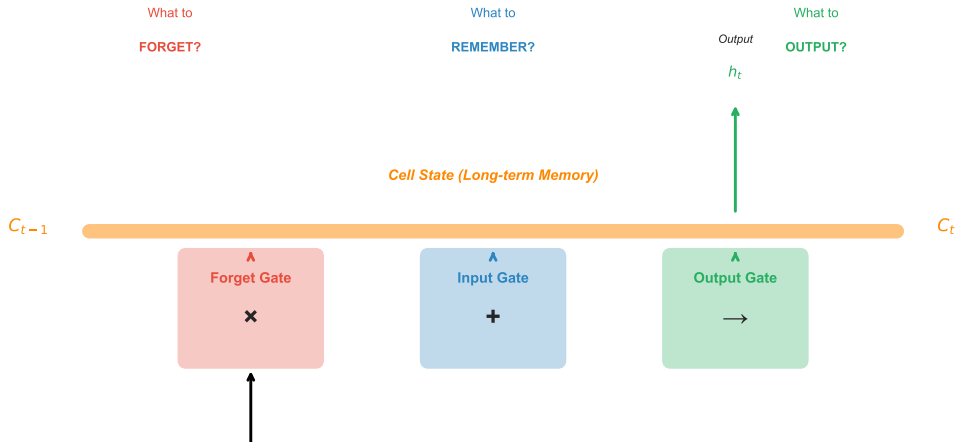
Formulas:

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

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

The Big Picture: Three Gates + Cell State

LSTM Architecture: Three Smart Gates



Now Let's Look Again - You Understand It!

Sentence: "The cat was hungry. The dog was sleeping."

Word	Forget	Input	Output	Memory
"The"	0.9 (keep)	0.3 (small add)	0.2 (hide)	[article]
"cat"	0.8 (keep)	0.9 (add subject!)	0.8 (show)	[cat]
"was"	0.9 (keep)	0.7 (add verb)	0.9 (need it!)	[cat, was]
"hungry"	0.8 (keep)	0.8 (add state)	0.7 (show)	[cat, hungry]
"."	0.1 (ERASE!)	0.4 (ending)	0.3 (hide)	[end]
"The"	0.1 (clear old)	0.8 (new start)	0.2 (hide)	[article]
"dog"	0.7 (keep)	0.9 (NEW subject!)	0.9 (show)	[dog]
"was"	0.9 (keep)	0.8 (add verb)	0.9 (USE dog!)	[dog, was]

Checkpoint: The Critical Transition

Watch rows 4-7: "hungry" → "." → "The" → "dog"

Memory evolves: [cat, hungry] → **FORGET** → [end] → **ADD** → [dog]

This is what RNNs cannot do! LSTM uses gates to control memory intelligently.

All Three Gates:

Forget Gate:

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

Input Gate:

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

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

Output Gate:

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

Cell State Update:

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

Output:

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

In Our Example:

- x_t = current word embedding
- h_{t-1} = previous output
- C_t = cell state (memory)
- σ = sigmoid (0 to 1)
- \tanh = tanh (-1 to 1)
- \odot = element-wise multiply

Summary: LSTM in Practice

What We Learned:

- 1 LSTM uses three gates to control memory
- 2 Forget gate: what to erase (0.1 at "." = erase "cat")
- 3 Input gate: what to add (0.9 at "dog" = add subject)
- 4 Output gate: what to use (0.9 at "was" = use "dog")
- 5 Cell state flows information forward

When to Use LSTM:

- Long sequences (100+ words)
- Long-term dependencies
- Context matters
- Grammar and structure

Real World: Applications

Where LSTMs Excel:

- Machine Translation (Google Translate)
- Speech Recognition (Siri, Alexa)
- Text Generation (ChatGPT foundations)
- Video Analysis
- Music Generation
- Handwriting Recognition

Key Takeaway:

The sentence example showed you *exactly* how LSTM gates work in practice. That's the real magic!

Questions?