LSTM RNN [Long Short Term Memory RNN]

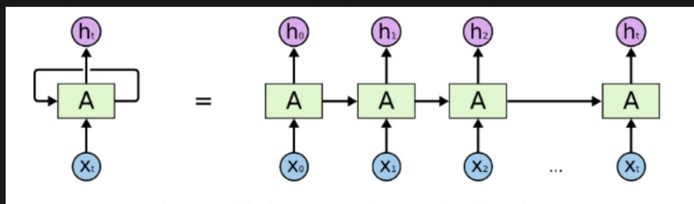
It's better to Watch StatQuest video on YT.

RNN ⟶ Long Term Dependencies ⟶ Vanishing Gradient Problem

Continue to Training Part for LSTM training

1. RNN ⟶ Problem? ✓
2. Why LSTM RNN? ✓ Basic Representation.
3. How LSTM RNN works
   - ⟶ Long Term Memory ✓
   - ⟶ Short Term Memory ✓
4. LSTM Architecture
5. Working of LSTM RNN

Problems With RNN → Long Term Dependency



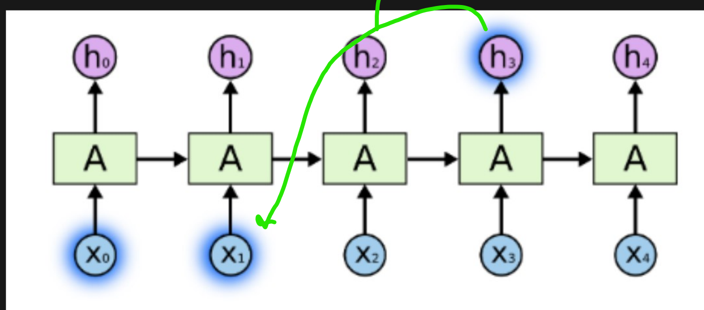
An unrolled recurrent neural network.

Vanishing Gradient Problem

Task :

Next Word In a Sentence

gap is less



The color of the sky is blue

further context

Gap is less

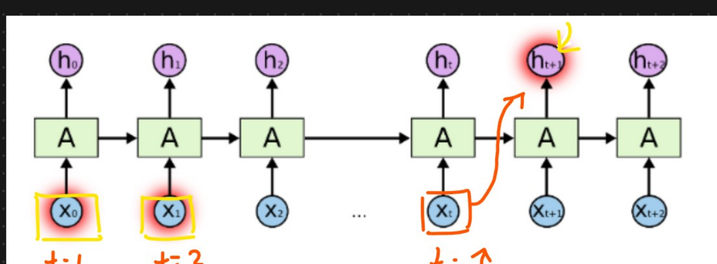
Huge gap O/p ← Context

I grew up in $\boxed{\text{India}}$ .... I speak fluent ______ ↳ context

Name of language

⇓

further Context



t=1 t=2 t=

≈0 ↳ ≈0

$\boxed{0-0.25}$

$\boxed{0-1}$

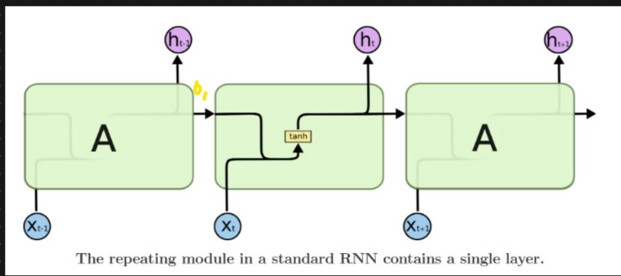
huge Gap → Long Term Dependency

RNN → Long Term Dependency → Vanishing Gradient Problem

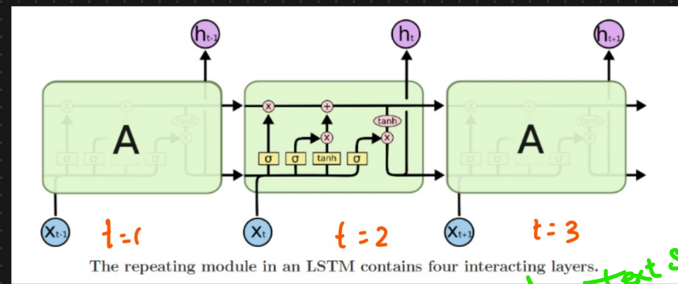
↳ Chain Rule → ≈0.

Basic Representation of RNN And LSTM RNN

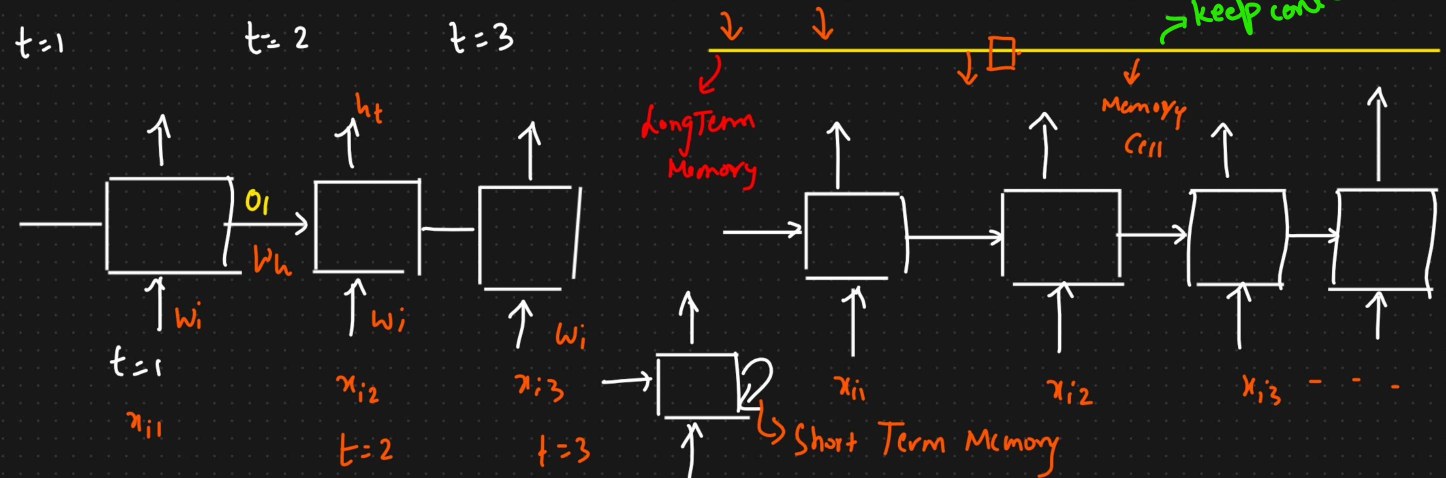
LSTM RNN



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



The repeating module in an LSTM contains four interacting layers.

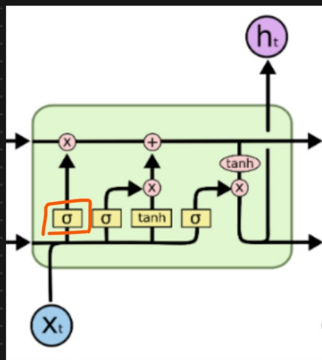


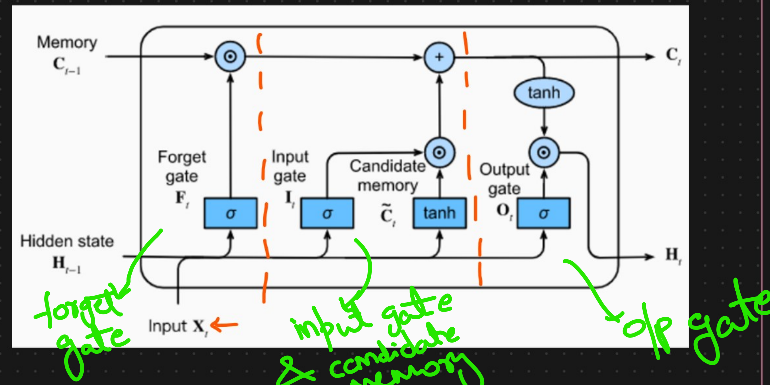
LSTM RNN → Long Term Memory
→ Short Term Memory

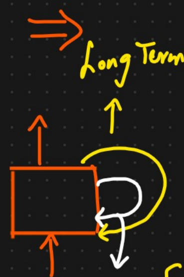
Conveyance Belt : Luggages

LSTM Architecture

LSTM RNN

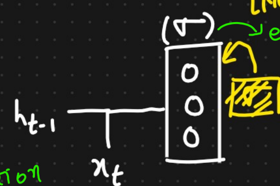






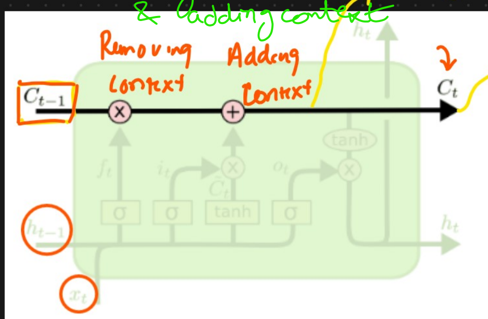
Basic Architecture

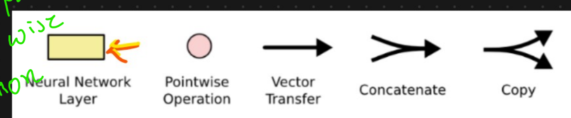
{Neural N/w Layer}.



C_t :- memory state after removing context & adding context

Memory cell → Long Term Memory





V_1 [1 2 3] ← tanh

V_2 [4 5 6]

⊗ ⇒ [4 10 18]

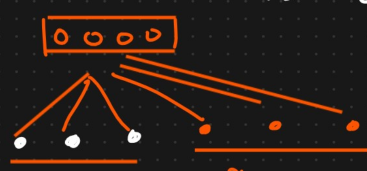
⊕ ⇒ [5 7 9]

tanh ⇒ [tanh(1) tanh(2) tanh(3)]

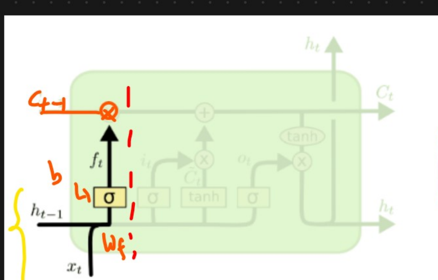
Combining 2 vectors

$h_{t-1} = [1 \quad 2 \quad 3]$

$x_t = [2 \quad 3 \quad 4]$

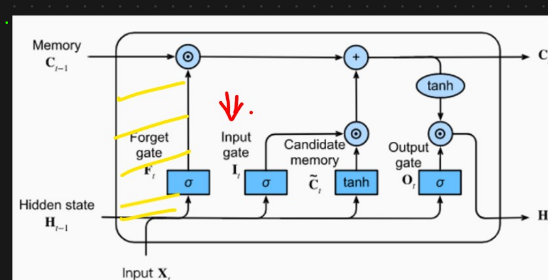


Duplicate copy



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

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



Forget Gate

h_{t-1} = Hidden state of previous time stamp

x_t = Word passed as i/p in the current time stamp

Text	Next Word
x_{11} x_{12} x_{13} x_{14}	y_{15}

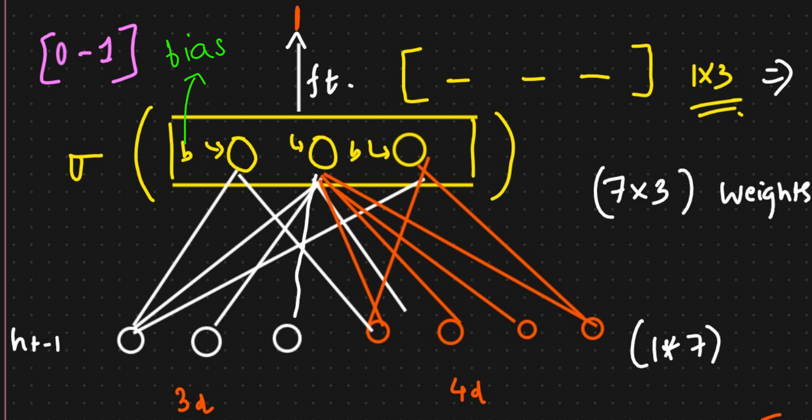
[0 2 4 1] [4 5 1 2] - - - - -

x_t x_{t+1}

Memory cell

h_{t-1} [1 2 4] C_{t-1} [4 2 1] ⇐

3d 3d



$(1*7) \times (7*3)$

(1×3)

$t=2$

$ft = [0 \quad 0 \quad 0]$ ← Complete Sentence Context changed

① $C_{t-1} = [6 \quad 8 \quad 9] \otimes [0 \quad 0 \quad 0]$

$= [0 \quad 0 \quad 0]$ ← Removing all the previous Context

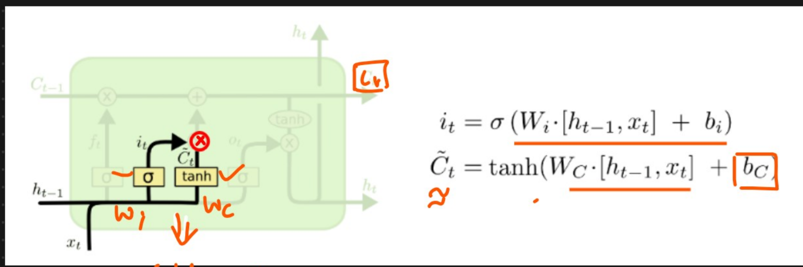
$ft = [1 \quad 1 \quad 1]$

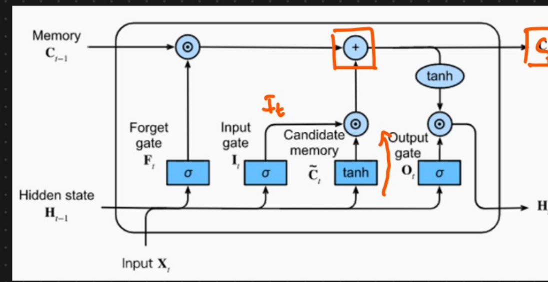
② $C_{t-1} = [6 \quad 8 \quad 9] \otimes [1 \quad 1 \quad 1] = [6 \quad 8 \quad 9]$

③ $C_{t-1} = [\underline{6} \quad 8 \quad \underline{9}] \otimes [0.5, 1, 0.5] = [3 \quad 8 \quad 4.5]$

<u>Conclusion</u> : Based on the context → Forget gate will let go some information or will not let go some Info {Forgetting}.

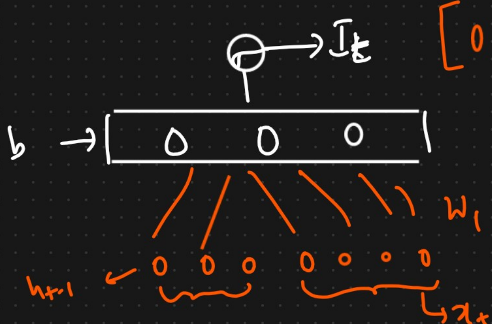
② Input Gate And Candidate Memory

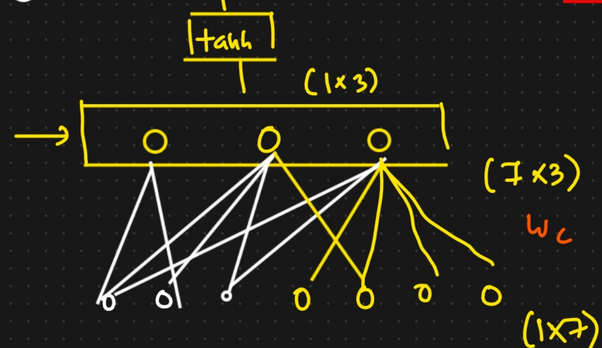




$I_t = [2 \ \boxed{4} \ 6]$ —— ⊛ —— C_t $[0 \ 2 \ 0]$ ⇒ I/p Gate

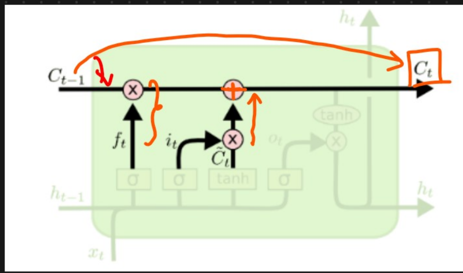
$[0 \ 8 \ 0]$



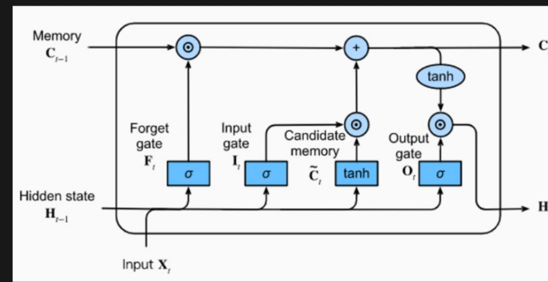


Context = If any information needed to be added it the memory

C_{t-1} → The information will be added



$$C_t = \underbrace{f_t * C_{t-1}}_{\text{Removing or Forgetting Some info}} + \underbrace{i_t * \tilde{C}_t}_{\text{Candidate Memory.}}$$



I stay in $\boxed{\text{India}}$ - - - - -

and I speak →

$\boxed{\text{English}}$ $\boxed{\text{Hindi}}$

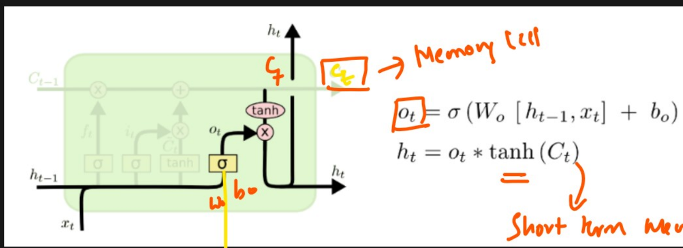
Forget Gate

I/p Gate ⊛ Candidate memory

+

$C_{t-1} \Rightarrow C_t$.

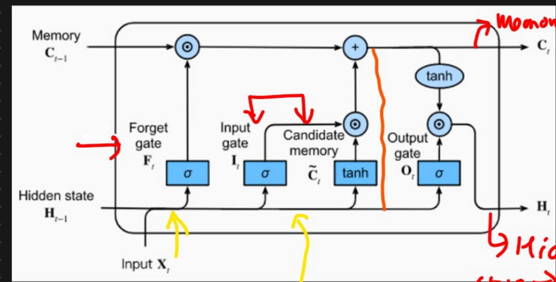
Output gate LSTM RNN



C_t → Memory cell

Short term memory.

W bo → Neural Network



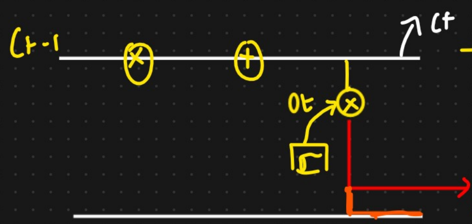
Memory

Hidden state → ST Memory.

Forget

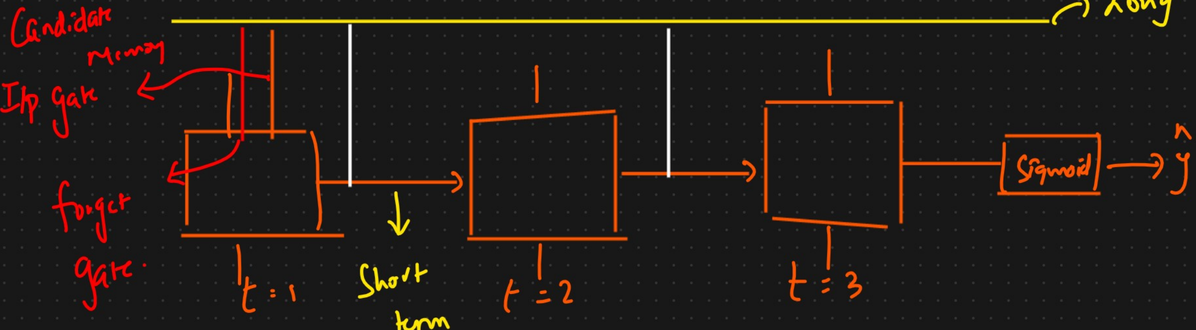
Add Info → Context

Some Info → context → Memory cell



Ct → Long Term Memory

ht ⇒ hidden state → Short term Memory }



→ Long Term

Loss ↓↓.

[Wi, Wc, Wo] → Updating ← Back Propogation

GRU RNN ⇒ LSTM Variant

Training Data With LSTM RNN

{Training} Text Paragraph

I Went to Restaurant and order burger

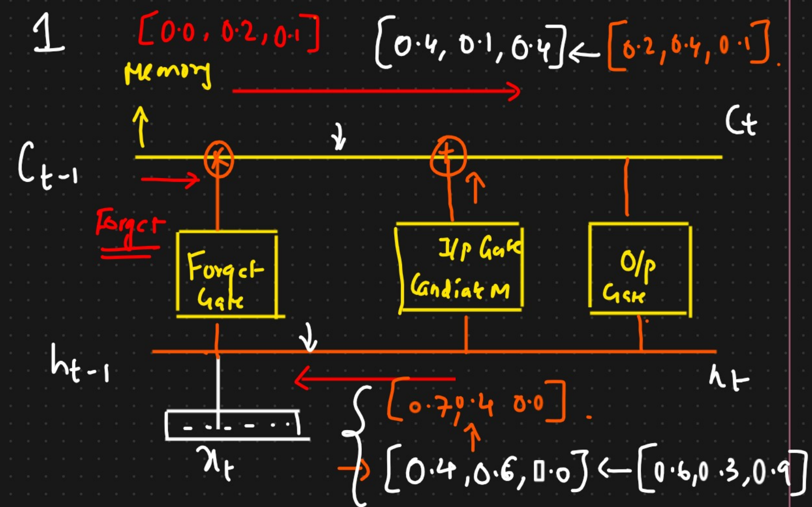
The burger looked tasty and crispy

→ But burger is not good for health

→ It has lot of fats, cholestrol

→ But this burger was made with Whey protein and only vegetables were used, so it was good.

O/P [good/bad] Forget gate I/P gate O/P gate



Word → Vectors → Embedding Layer

Word2Vec [3 dimension - vector]

→ [[Good] [Bad] [Healthy]] ← Black Box

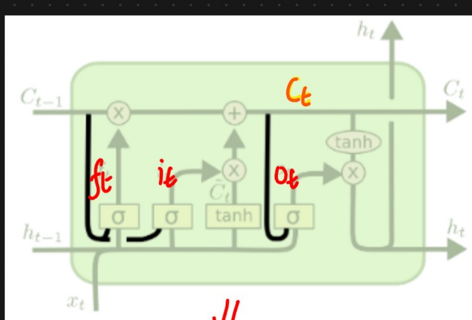
Tasty [0.9 0.0 0.1] ← 3 d.

Variants of LSTM RNN

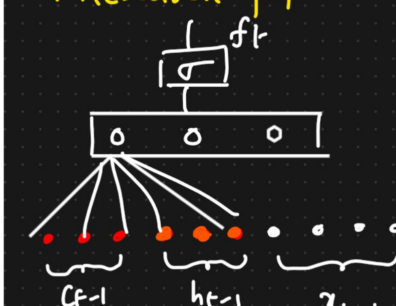
LSTM RNN [1970-80]

LSTM Variants Introduced By Gers & Schmidhuber [2000]

↳ Research paper



$$f_t = \sigma\left(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [C_t, h_{t-1}, x_t] + b_o\right)$$

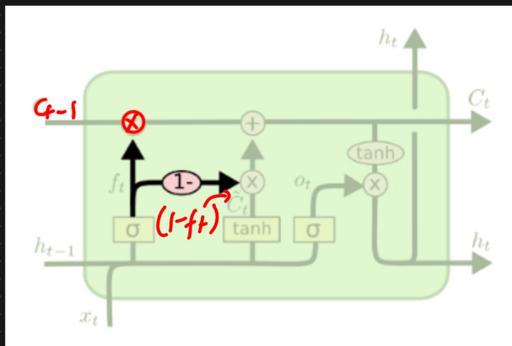


focus

Connections → From Memory cell to forget gate, i/p gate, o/p gate ⇒ Peephole Connections

Peephole Connections : We let the gate layers look at the cell state

Another variation → Coupling Forget And I/p Gates



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

We add new values only when we forget → gate.

Goal: We only forget when we're going to i/p something in its place.

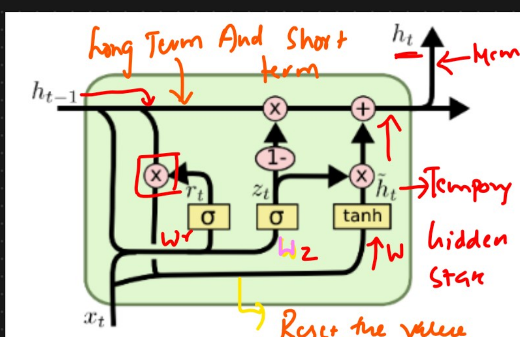
⇓

We only i/p new values to the state when we forget something older.

Instead of separately deciding what to forget and what we should add new Info, we make this decision together.

GRU → Gated Recurrent Unit [Cho, et al [2014]]

1980 → LSTM
2000 – variants
2014 → GRU



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

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

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

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

Context Context

⇓ Coupled.

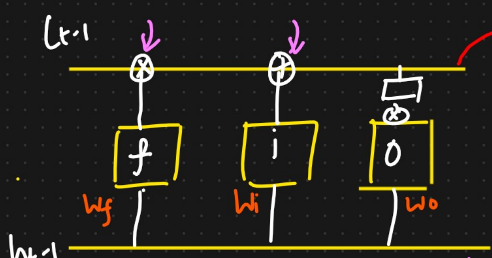
z_t ⇒ Update Gate ←

r_t ⇒ Reset Gate ←

$\tilde{h}_t$ ⇒ Temporary hidden state.

bias

[W_f, W_i, W_o] these Trainable parameters ↓ these parameters increase will complexity



→ Long Term Memory

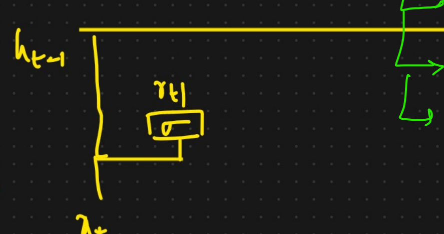
→ Short Term Memory

time

LSTM : Complex Architecture → forget / i/p + Candidate memory / o/p

⇒ LSTM complex
↳ 3 gates
↳ More no. of parameters
↳ Training time ↑es
↳ Very complex

Training Time ↑↑



z_t,

Reset gate = r_t =

→ Resetting some info from h_{t-1} ⇒ Memory → LTM + STM

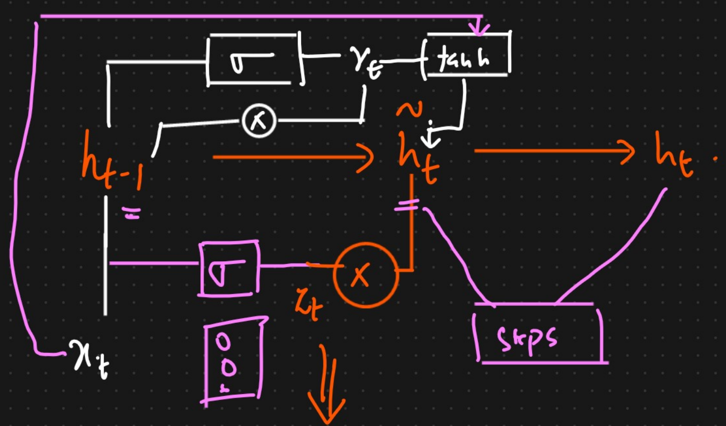
h_{t-1} [0.6 0.5 0.3 0.9]

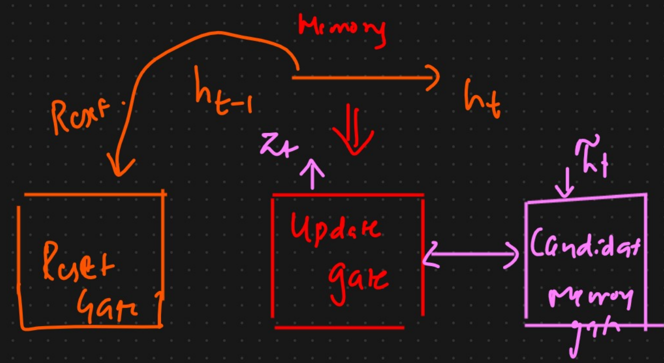
⊗ ←

r_t [0.2 0.4 0.8 0.2]

⇓ ⇓ ⇓ ⇓

x_t ⟶ [0.12 0.20 0.24 0.18] ← Resetting → Context





What Context Info needs to be Added

⇓

Candidate hidden state. [current Context]

⇓

Imp → Add Info

z_t ⇒ Less Info