Long Short Term Memory (LSTM)

Selective Read, Write and Forget information

The state s_i of an RNN records information from all previous time steps.

At each new time step, the old information is changed by the new information.

After some time it is impossible to extract the original information.

eg. The movie was long but awesome.

Consider solving a problem on a white board

$$q = (xy + z)px$$

1. Read d
2. Read e
3. Write $f = dxe$

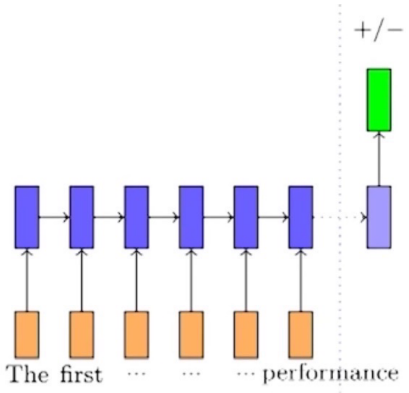
$x = 1$
$y = 3$
$z = 5$
$p = 2$

$c = 3$

$d = 8$
$e = 2$

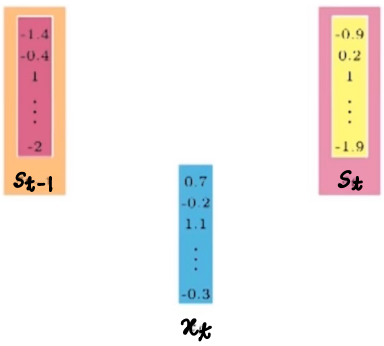
Read
Write
Forget

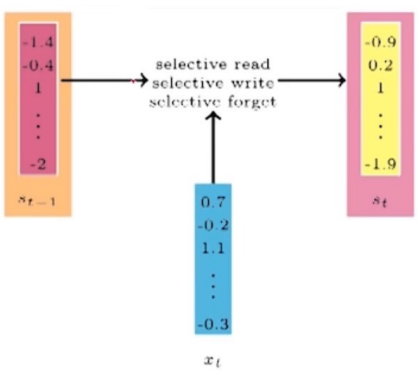
$f = \boxed{16}$



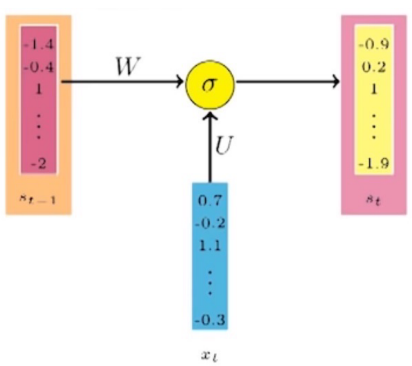
Review: The first half of the movie was dry but the second half really picked up pace. The lead actor delivered an amazing performance

1. Selective forget : the info that is no more useful
   eg. words that come before "but"
2. Selectively read : info added by sentiment bearing words.
   eg. amazing, awesome, bad, ugly
3. Selectively write : Combines previous info. with new info.





Selective Writing



In standard RNN,

$$s_t = \sigma(Ux_t + Ws_{t-1} + b)$$

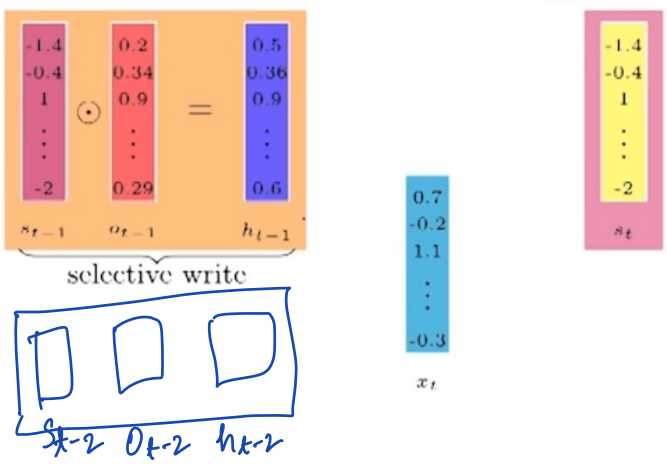
Instead of passing the whole s_{t-1}, pass portions of it.

output gate

$$\begin{bmatrix} -1.4 \\ -0.4 \\ 1 \\ \vdots \\ -2 \end{bmatrix} \odot \begin{bmatrix} 0.1 \\ 0.9 \\ 0.3 \\ \vdots \\ 0.5 \end{bmatrix} = \begin{bmatrix} -0.14 \\ -0.36 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

s_{t-1} Hadamard product O_{t-1} h_{t-1}

We introduce o_{t-1} (output gate / write gate) which decides what portion of s_{t-1} to write.



o_{t-1}: We need to learn o_{t-1}.

- Introduce parameters.

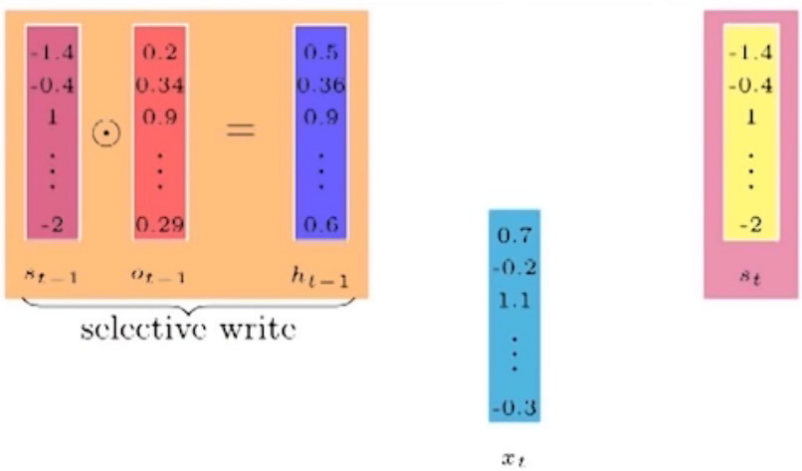
s_{t-2} o_{t-2} h_{t-2}

$$o_{t-1} = \sigma(W_o h_{t-2} + U_o x_{t-1} + b_o)$$

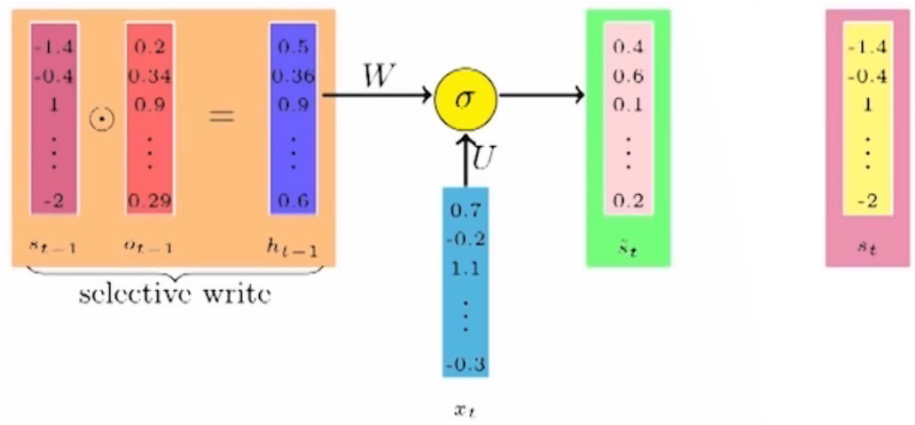
$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

$$h_{t-1} = o_{t-1} \odot s_{t-1}$$

Sigmoid ensures that the elements of o_{t-1} are between 0 and 1.



Selective Read

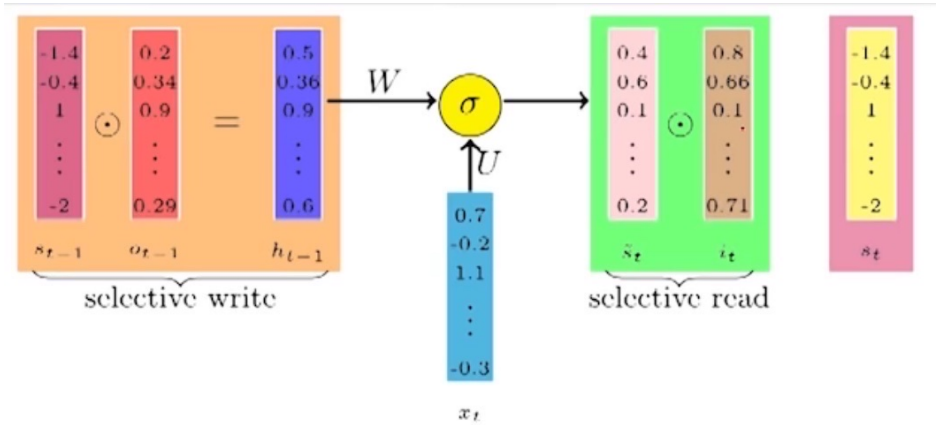


In standard RNN,

$$\tilde{s}_t = \sigma(W h_{t-1} + U x_t + b)$$

However we want to selectively read. Therefore introduce a new gate i_t (input gate)

$\tilde{s}_t \odot i_t$: selectively reading

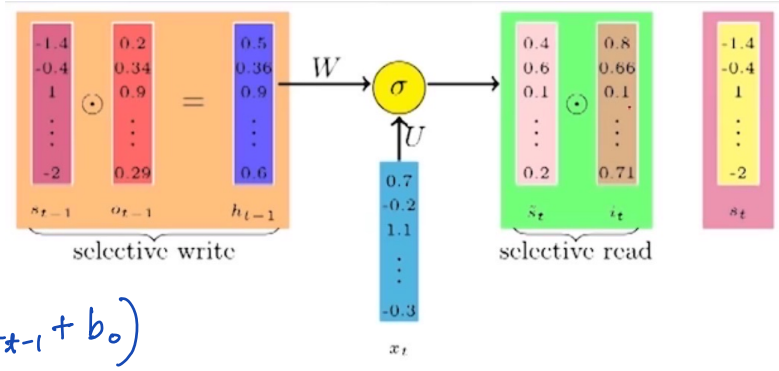


$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$

So far,

Previous state : s_{t-1}

Current word : x_t



Output gate :

$$o_{t-1} = \sigma\left(W_o h_{t-2} + U_o x_{t-1} + b_o\right)$$

Selectively write : $o_{t-1} \odot s_{t-1} = h_{t-1}$

Current temporary state : $\tilde{s}_t = f\left(W h_{t-1} + U x_t + b\right)$

Input gate : $i_t = \sigma\left(W_i h_{t-1} + U_i x_t + b_i\right)$

Selectively read : $\tilde{s}_t \odot i_t$

Selective Forget

$\tilde{s}_t \odot i_t$: new information

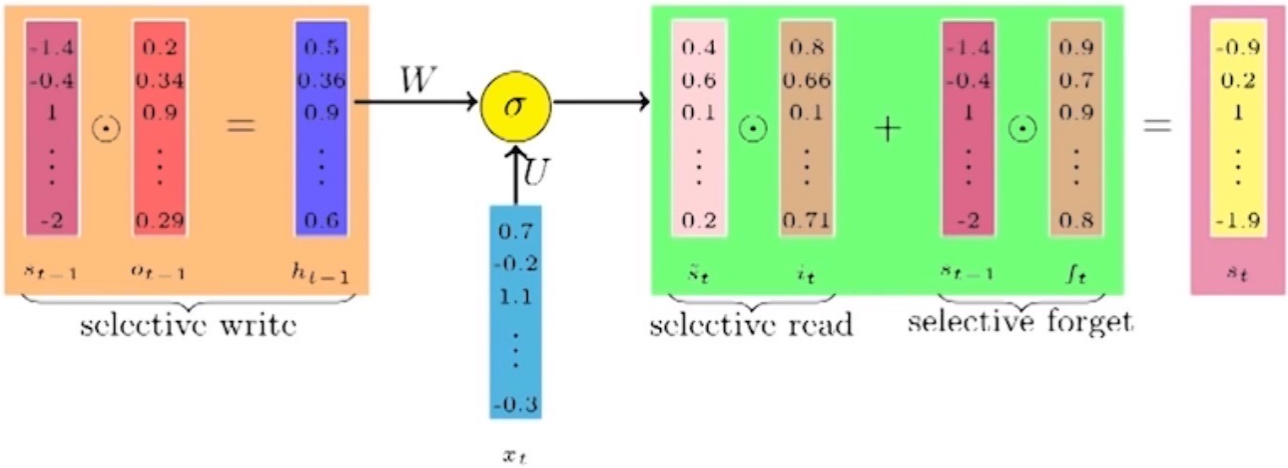
s_{t-1} : old information

A simple way (without forgetting) to combine these would be :

$$s_{t-1} + \tilde{s}_t \odot i_t$$

Forget gate $f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right)$

$$s_t = \underbrace{f_t \odot s_{t-1}}_{\text{selective forgetting}} + \tilde{s}_t \odot i_t$$



LSTM

1. Output gate

$$o_t = \sigma\left(W_o h_{t-1} + U_o x_t + b_o\right)$$

2. Input gate

$$i_t = \sigma\left(W_i h_{t-1} + U_i x_t + b_i\right)$$

3. Forget gate

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right)$$

$$h_{t-1} = o_{t-1} \odot s_{t-1}$$

$$\tilde{s}_t = f(W h_{t-1} + U x_t + b)$$

$$s_t = f_t \odot s_{t-1} + i_t \odot \tilde{s}_t$$

$$h_t = o_t \odot s_t$$

LSTM

Selective forget — Selective write — Selective Read

GRU (Gated Recurrent Units)

Output gate

$$o_t = \sigma(W_o s_{t-1} + U_o x_t + b_o)$$

Input gate

$$i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i)$$

$$\tilde{s}_t = \sigma\left(W(o_t \odot s_{t-1}) + U x_t + b \right)$$

Selective write

$$\tilde{s}_t = (1 - i_t) \odot s_{t-1} + i_t \odot \tilde{s}_t$$

selective forget

Selective reading