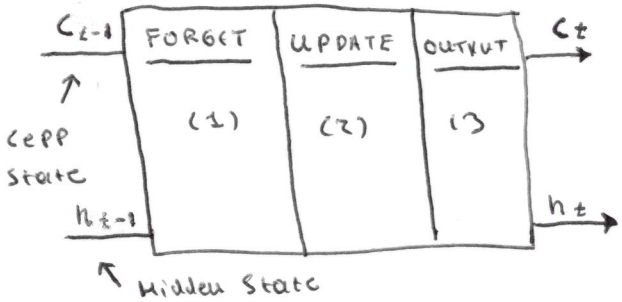
LSTMs:

High Level Overview:

LSTMs have essentially a 3 step process:



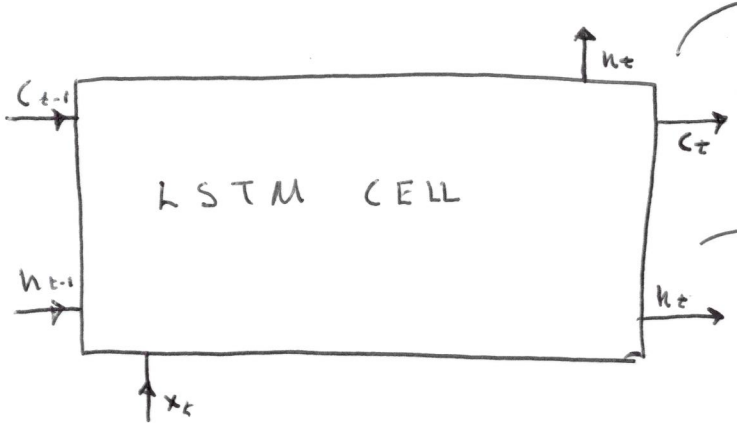
(1): Forget irrelevant parts of the previous state: For example, if we are modelling a sentence and we see a new subject, we might want to forget things about the old subject because we know that future words will be conjugated according to the new subject.

(2): Selectively update state values: Update the cell state to reflect information according to the new input. According to the previous example we update the cell state according to the new subject, i.e its gender, whether it is in plural or singular.

(3): Output certain parts of the cell state: For example if we have just seen a new subject we should predict information about the corresponding new verb as it's tense.

- Each of these steps is implemented by a set of Logic Gates:

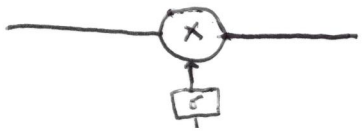
- LSTMs repeating modules contain interacting layers that control information flow.



A cell state is maintained making it easy for information to flow within the cell

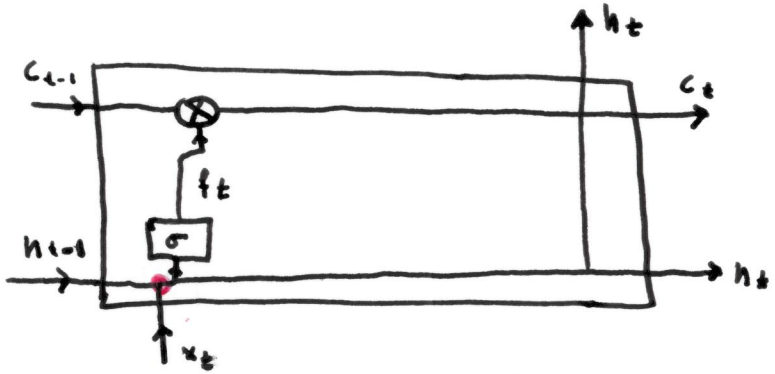
An output gate is utilized to output parts of the cell state.

- Information is added or removed to the cell state through the structure called gates.



Gates optionally let information through, via a sigmoid neural net layer and pointwise multiplication.

LSTMs: Forget Irrelevant Information [FORGET]



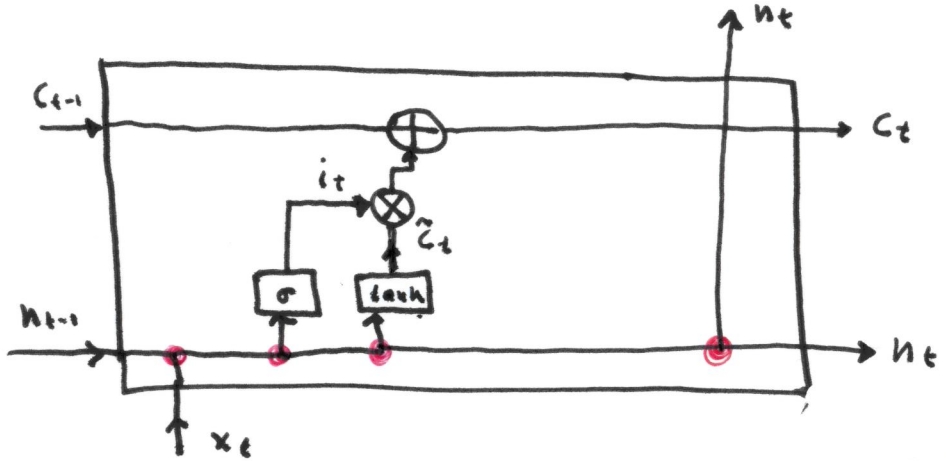
$$f_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_f\right)$$

- Use previous cell output and input
- Sigmoid value within the [0,1] range.

0: Indicates to completely forget.
1: Indicates to completely remember.

"Forget the gender pronoun of previous subject in the sentence".

LSTMs : Identify new information to be stored



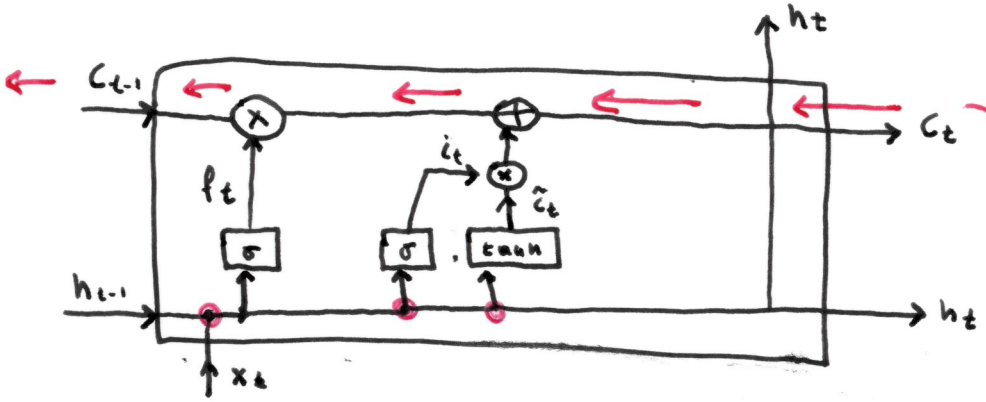
$$i_t = \sigma\left(W_i\,[h_{t-1}, x_t] + b_i\right)$$

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

- Sigmoid layer decides which values to update.
- Tanh layer generates a new vector of "candidate values" that could be added to the state.

"Add gender of new subject to replace the old subject".

LSTMs: Update the cell state



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

→ LSTM GRADIENT FLOW

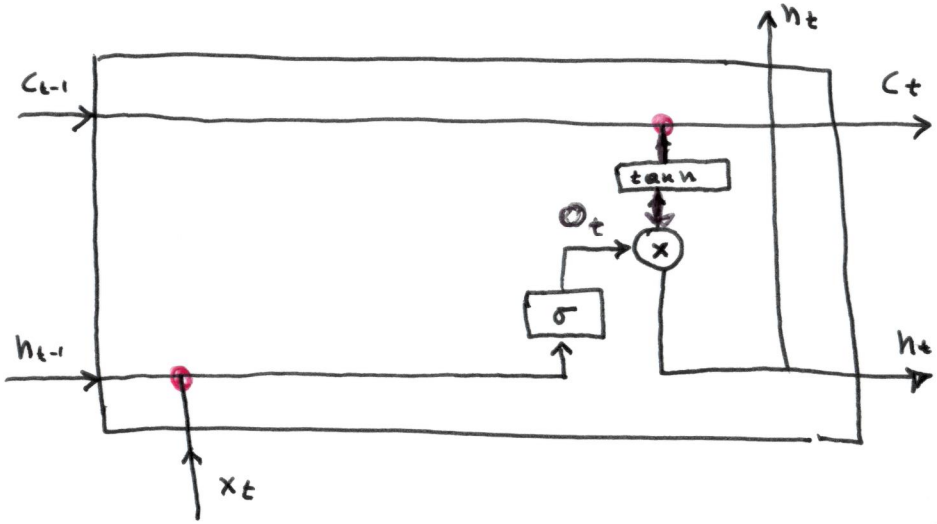
- Apply forget operation to previous internal cell state: $f_t \circledast C_{t-1}$

- Add new candidate values, scaled by how much we decide to update: $i_t \circledast \tilde{C}_t$

"Actually drop all previous information and add new information about subject's gender".

- Backpropagation from C_t to C_{t-1} requires only elementwise multiplication!
- No matrix-multiplication is required → AVOID VANISHING GRADIENT. PROBLEM.

LSTMs: Output a Filtered Version of the Cell State



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

- Sigmoid Layer decides which parts of the state to output.
- Tanh Layer squashes values within $[-1, +1]$
- Output filtered version of the cell state $O_t \circledast \tanh(C_t)$

"Having seen the new subject, may output information relating to a verb".