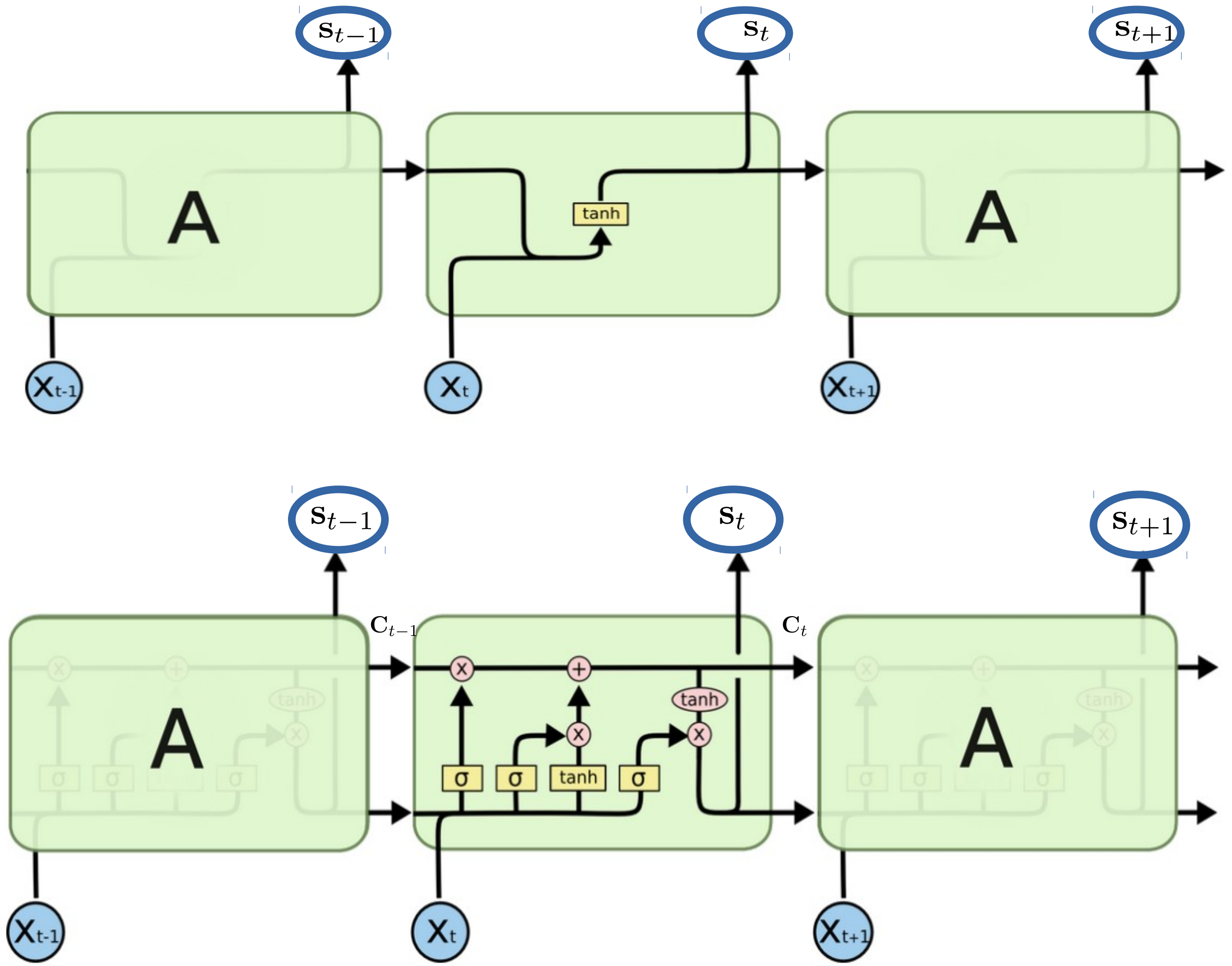
TEK 5040/9040
LSTM Supplement

Narada Warakagoda

Long Short Time Memory (LSTM)

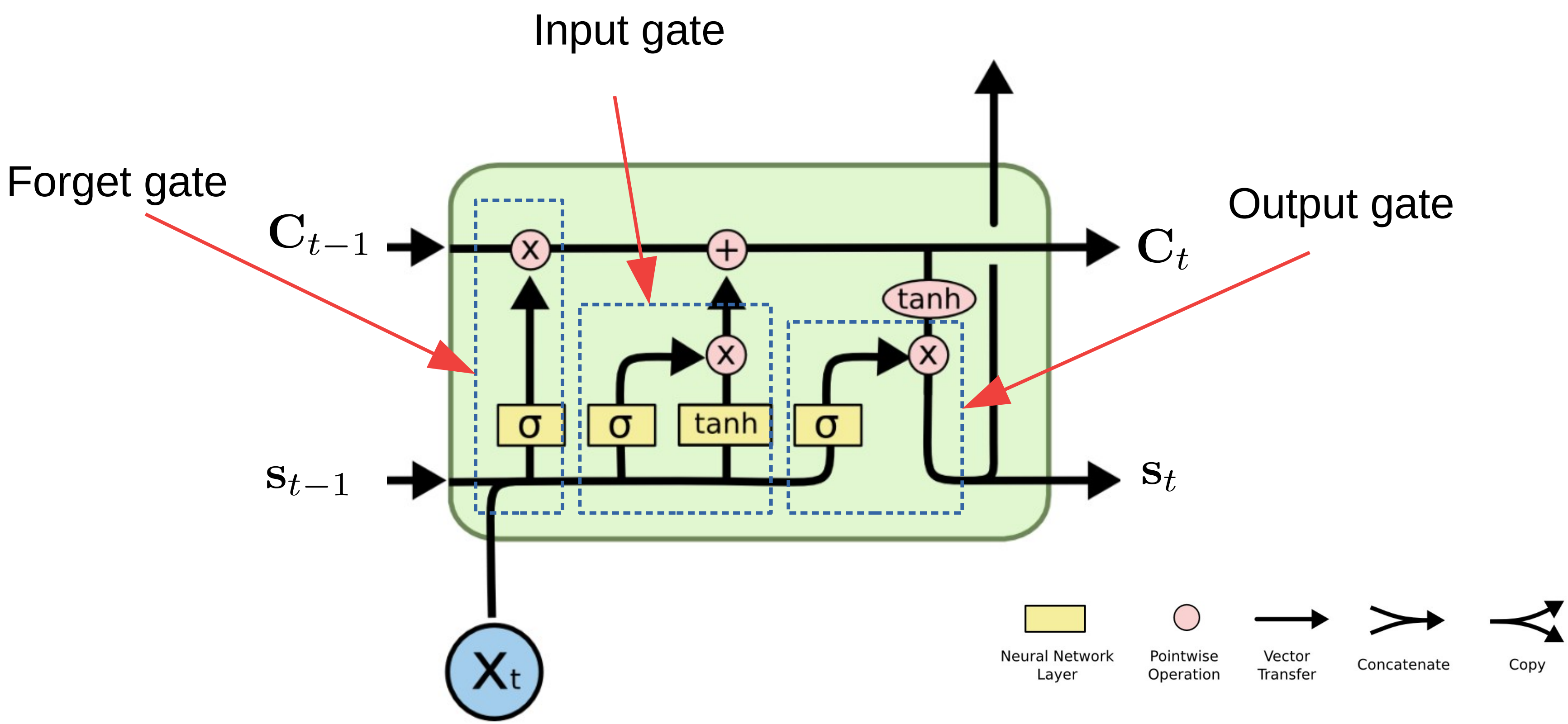
- Plain RNN cells are very hard to train with long sequences
  - Due to gradient explosion or vanishing problem
- LSTM was proposed as a solution to this problem
- Similar solutions such as Gated Recurrent Unit (GRU) also exist.

Plain RNN vs LSTM cells



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

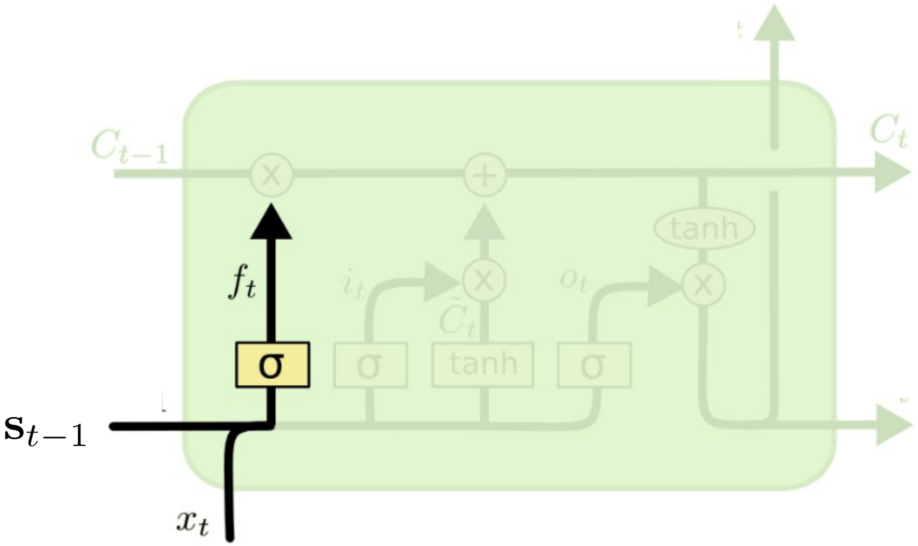
Main components of LSTM



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Forget gate

- Forgets irrelevant information ($\mathbf{f}_t \approx 0$) in the control state
- Let relevant information passes through ($\mathbf{f}_t \approx 1$)

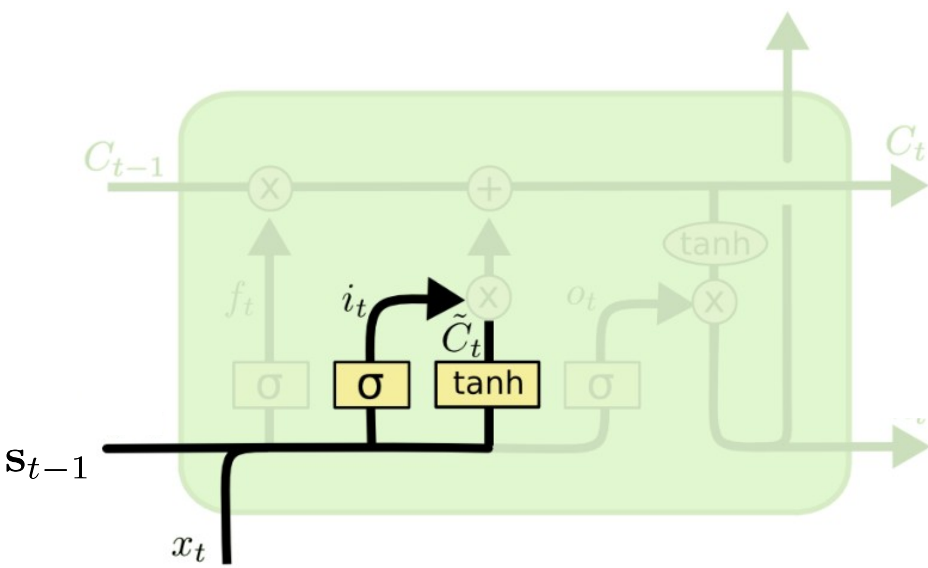


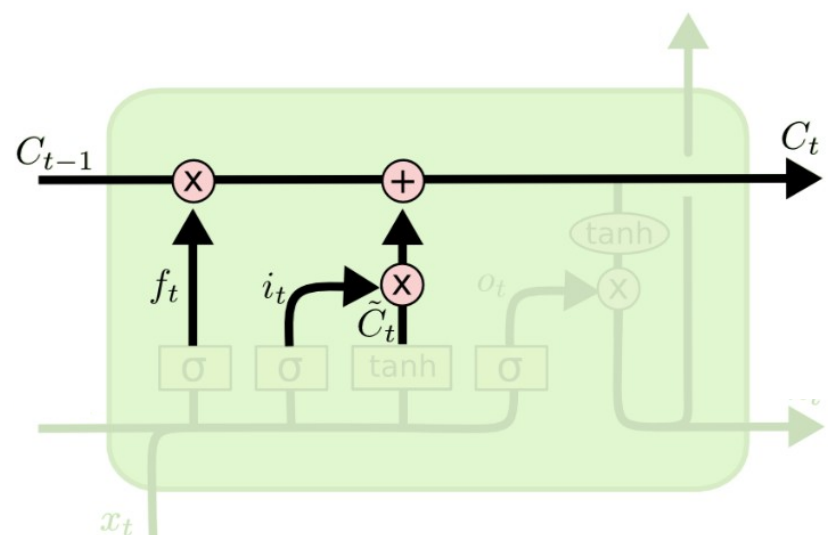
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \cdot [\mathbf{s}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f)$$
$$\mathbf{C}'_{t-1} = \mathbf{C}_{t-1} * \mathbf{f}_t$$

Input gate

- Picks new information to be added to the control state





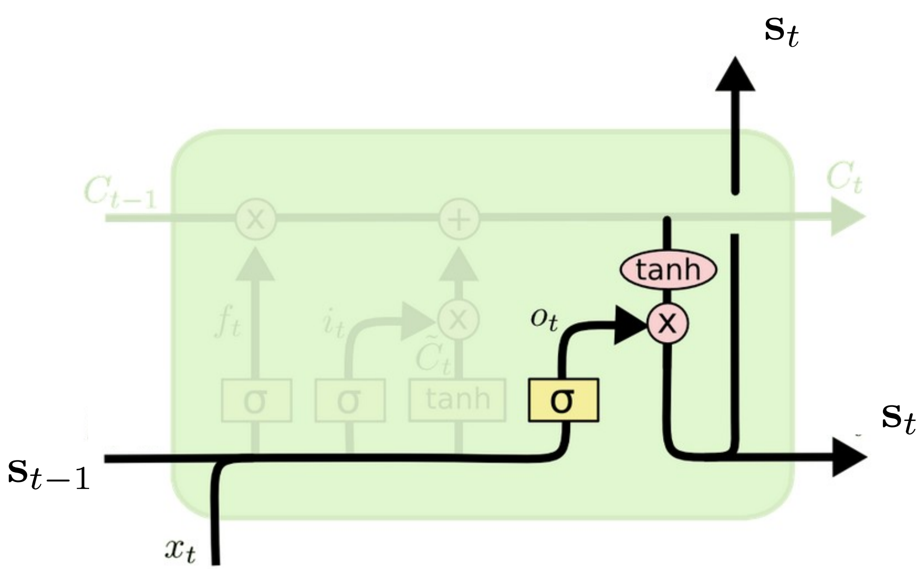
$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot [\mathbf{s}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i)$$
$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_c \cdot [\mathbf{s}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c)$$

$$\begin{aligned}\mathbf{C}_t &= \mathbf{C}'_{t-1} + \mathbf{i}_t * \tilde{\mathbf{C}}_t \\ &= \mathbf{f}_t * \mathbf{C}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{C}}_t\end{aligned}$$

https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Output gate

- Picks information to be transferred from control state to the cell output



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot [\mathbf{s}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o)$$
$$\mathbf{s}_t = \mathbf{o}_t * \tanh(\mathbf{C}_t)$$

LSTM vs plain RNN vanishing gradients

Plain RNN

$$
\begin{aligned}
s_t &= \sigma(w s_{t-1}) \\
\frac{\partial s_t}{\partial s_{t-1}} &= w\sigma'(w s_{t-1}) = w\sigma'(v_{t-1}) \\
\frac{\partial E}{\partial s_{t-1}} &= \frac{\partial E}{\partial s_t}\frac{\partial s_t}{\partial s_{t-1}} \\
&= \frac{\partial E}{\partial s_t}\, w\sigma'(v_{t-1}) \\
\frac{\partial E}{\partial s_{t-N}} &= \frac{\partial E}{\partial s_t}\, w^N \prod_n \sigma'(v_{t-1-n})
\end{aligned}
$$

if $w < 1$, $\frac{\partial E}{\partial s_{t-N}} \to 0$ when N is large due to the factor w^N

LSTM

$$
\begin{aligned}
c_t &= c_{t-1}\sigma(w_f s_{t-1}) + \sigma(w_i s_{t-1})\tanh(w_c s_{t-1}) \\
\frac{\partial c_t}{\partial c_{t-1}} &= A_t + B_t + G_t + H_t(\text{ i.e. several terms}) \\
\frac{\partial E}{\partial c_{t-1}} &= \frac{\partial E}{\partial c_t}\frac{\partial c_t}{\partial c_{t-1}} \\
&= \frac{\partial E}{\partial c_t}(A_t + B_t + C_t + D_t) \\
\frac{\partial E}{\partial c_{t-N}} &= \frac{\partial E}{\partial c_t}(\text{ complicated\_product})
\end{aligned}
$$

$\frac{\partial E}{\partial c_{t-N}}$ does not coverge to 0 as easily