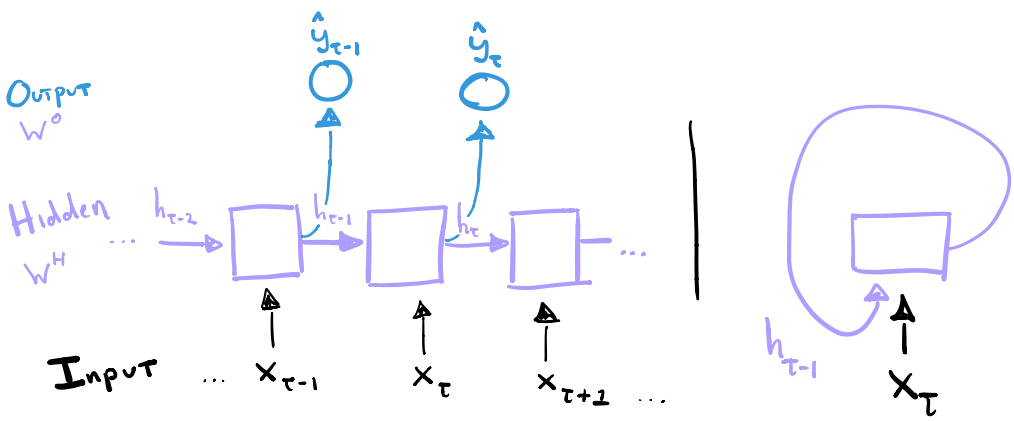
DS 4440 Recurrent Neural Networks, II

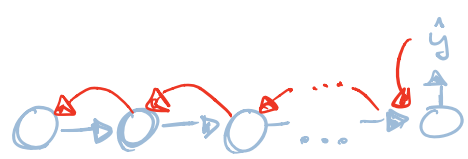
Last time: Recurrent models for Sequences.

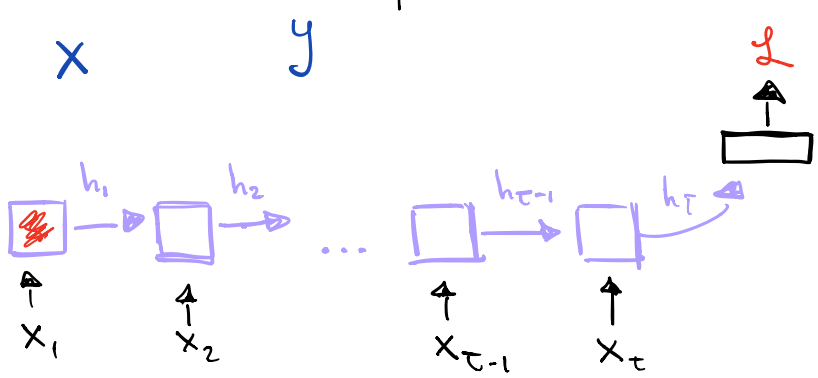


The problem of long term dependencies

(0, 0, 0, 0, 0)	0
(1, 0, 0, 1, 0)	1
...	
(1, 0, 0, 0, 0)	1
X	y

Signal has to flow all the way from $\hat{y}$

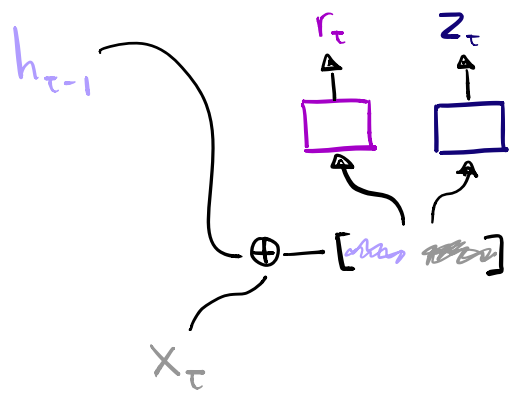




Gated Recurrent Units (GRUs)

Similar to LSTMs (next) but simpler.

Intuition: Include mechanisms – gates – to allow the model to update h or <u>skip</u> over inputs.



r_τ = reset

z_τ = update

$\square$ Dense layers + σ

$$r_\tau = \sigma(x_\tau W_r^x + h_{\tau-1} W_r^h + b_r)$$

$$z_\tau = \sigma(x_\tau W_z^x + h_{\tau-1} W_z^h + b_z)$$

Reset

$$\tilde{h}_\tau \leftarrow \tanh(x_\tau W_h^x + \underbrace{(r_\tau \odot h_{\tau-1})}_{\text{allows for forgetting}} W_h^h + b^h)$$

Update

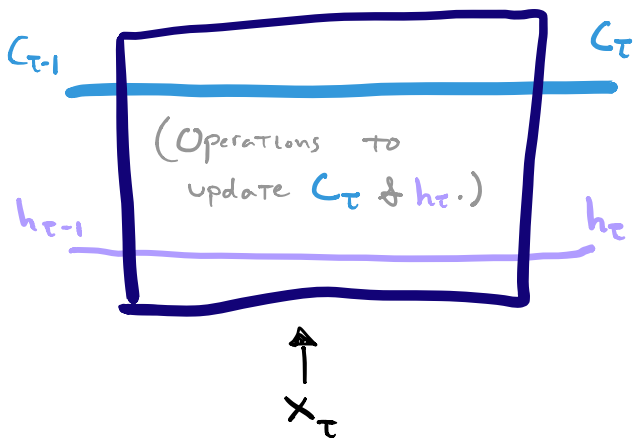
$$h_\tau \leftarrow \underbrace{z_\tau \odot h_{\tau-1}}_{\text{Copy}} + \underbrace{(1 - z_\tau) \odot \tilde{h}_\tau}_{\text{Update}}$$

(See CoLab Notebook.)

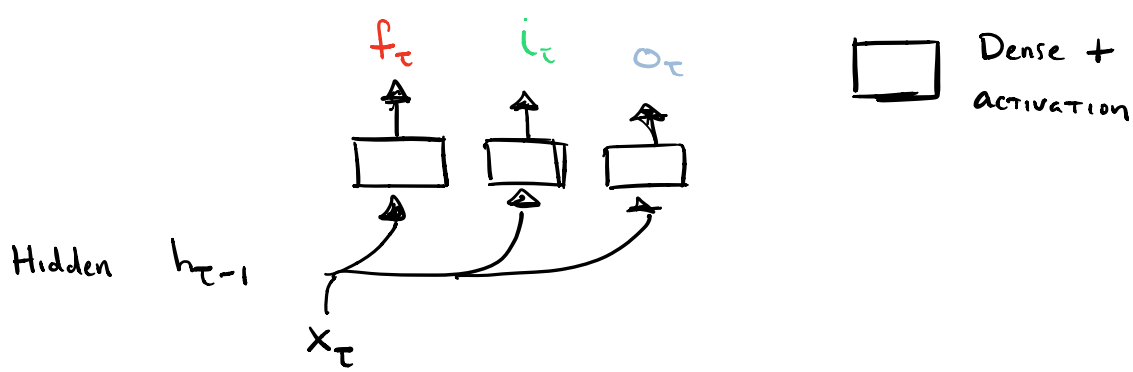
Long Short Term Memory (LSTM)

Similar to GRUs and also adresses long-term dependencies via gating.

Key Idea



Three gates: forget, input, output



Assume batch size of n.

$$x_\tau \in \mathbb{R}^{n \times d} \qquad h_{\tau-1} \in \mathbb{R}^{n \times h}$$

$$f_\tau \in \mathbb{R}^{n \times d} \qquad f_\tau = \sigma(x_\tau W_f^x + h_{\tau-1} W_f^h + b_f)$$

$$i_\tau \in \mathbb{R}^{n \times h} \qquad i_\tau = \sigma(x_\tau W_i^x + h_{\tau-1} W_i^h + b_i)$$

$$o_\tau \in \mathbb{R}^{n \times h} \qquad o_\tau = \sigma(x_\tau \underbrace{W_o^x}_{d \times h} + h_{\tau-1} \underbrace{W_o^h}_{h \times h} + \underbrace{b_o}_{1 \times h})$$

We apply these gates to a <u>candidate</u> hidden state or memory cell.

$$\tilde{c}_\tau \in \mathbb{R}^{n \times h} \qquad \tilde{c}_\tau = \tanh(x_\tau W_c^x + h_{\tau-1} W_c^h + b_c)$$

Next we update the memory:

$$c_\tau = f_\tau \odot c_{\tau-1} + i_\tau \odot \tilde{c}_\tau$$

$$c_{\tau-1} \longrightarrow c_\tau \longrightarrow c_{\tau+1}$$

The idea is that it is easy for info to flow directly along these cells.

c_τ serves as a sort of longer memory.

We also keep local memory: h

memory h_τ, governed by the output gate O_τ.

$$h_\tau = O_\tau \odot \mathrm{Tanh}(C_\tau)$$