CS 316: Introduction to Deep Learning

Advanced Language Models – LSTM & GRU

Dr Abdul Samad

Gradients

- Long chain of dependencies for backprop
  - Need to keep a lot of intermediate values in memory
  - Butterfly effect style dependencies
  - Gradients can vanish or diverge (more on this later)
- Clipping to prevent divergence

$$\mathbf{g} \leftarrow \min\left(1, \frac{\theta}{\|\mathbf{g}\|}\right)\mathbf{g}$$

rescales to gradient of size at most θ

Recurrent Neural Networks (with hidden state)

- Hidden State update

$$h_t = f(h_{t-1}, x_{t-1}, w)$$

- Observation update

$$o_t = g(h_t, w)$$

Analysis of gradient in RNN

Consider a simple RNN

$$h_t = f(x_t, h_{t-1}, w_h)$$
$$o_t = g(h_t, w_0)$$

Loss is calculated as

$$L(x_1, \dots, x_T, y_1, \dots, y_T, w_h, w_0) = \frac{1}{T}\sum_{t=1}^{T} l(y_t, 0_t)$$

Analysis of gradient in RNN

$$\frac{\partial L}{\partial w_h} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial l(y_t, 0_t)}{\partial w_h}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial l(y_t, 0_t)}{\partial o_t}\frac{\partial g(h_t, w_0)}{\partial h_t}\frac{\partial h_t}{\partial w_h}$$

Analysis of gradient in RNN

$$\frac{\partial \boldsymbol{h_t}}{\partial \boldsymbol{w_h}} = \frac{\partial \boldsymbol{f(x_t, h_{t-1}, w_h)}}{\partial \boldsymbol{w_h}} + \frac{\partial \boldsymbol{f(x_t, h_{t-1}, w_h)}}{\partial \boldsymbol{h_{t-1}}} \frac{\partial \boldsymbol{h_{t-1}}}{\partial \boldsymbol{w_h}}$$

Assume we have three sequences $\{a_t\}, \{b_t\}, \{c_t\}$ satisfying $a_0 = 0$ and $a_t = b_t + c_t a_{t-1}$ for $t = 1,2, \ldots$

$$\boldsymbol{a_t = b_t + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} c_j \right) b_i}$$

$\boldsymbol{a_t} = \frac{\partial \boldsymbol{h_t}}{\partial \boldsymbol{w_h}}$

$\boldsymbol{b_t} = \frac{\partial \boldsymbol{f(x_t, h_{t-1}, w_h)}}{\partial \boldsymbol{w_h}}$

$\boldsymbol{c_t} = \frac{\partial \boldsymbol{f(x_t, h_{t-1}, w_h)}}{\partial \boldsymbol{h_{t-1}}}$

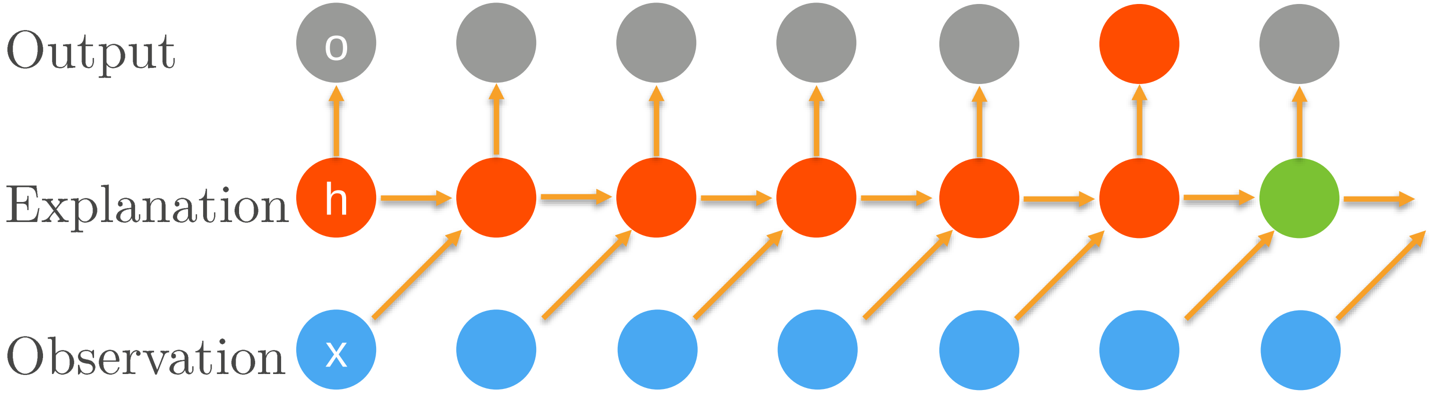
Analysis of gradient of RNN

$$\frac{\partial h_t}{\partial w_h} = \frac{\partial f(x_t, h_{t-1}, w_h)}{\partial w_h} + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} \frac{\partial f(x_j, h_{j-1}, w_h)}{\partial h_{j-1}} \right) \frac{\partial f(x_i, h_{i-1}, w_h)}{\partial w_h}$$

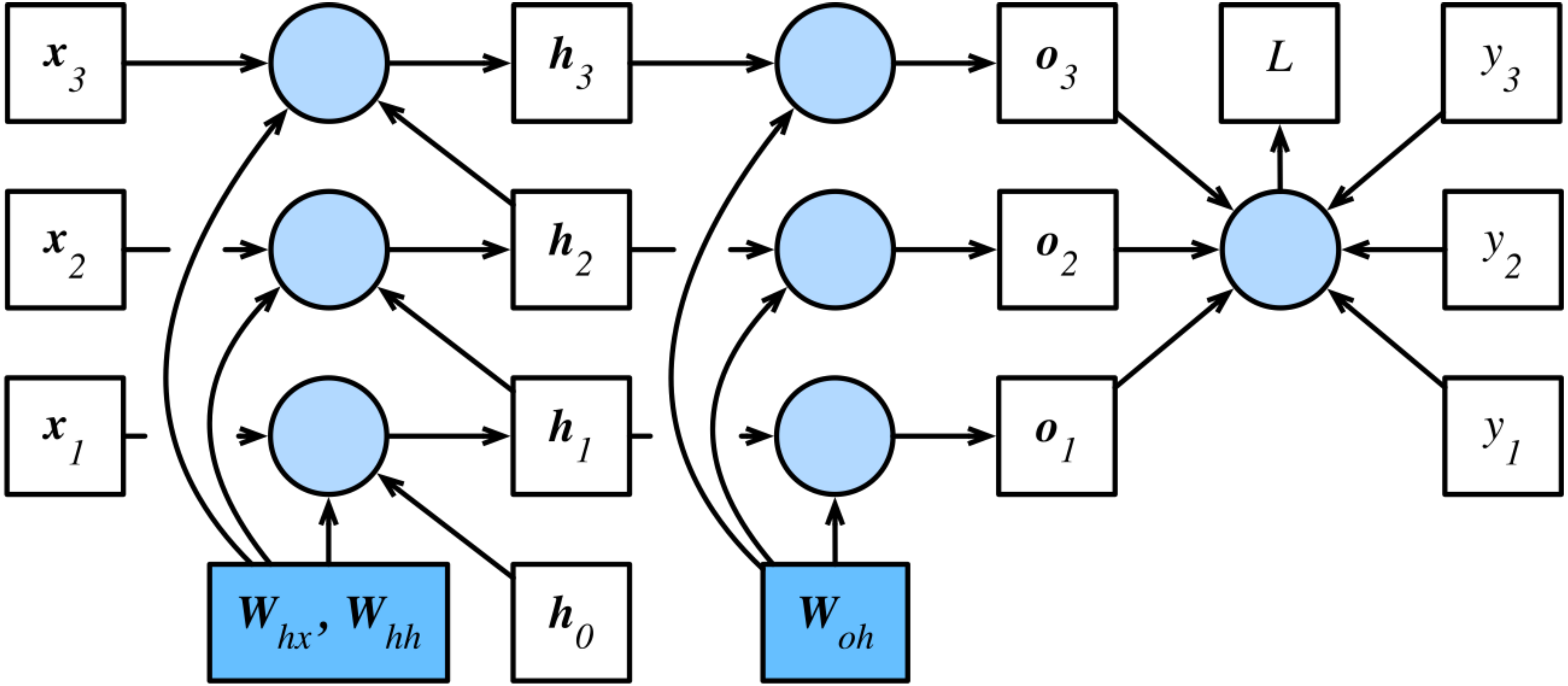
Latent State Gradient $\partial_w h_t$

- Gradient Recursion

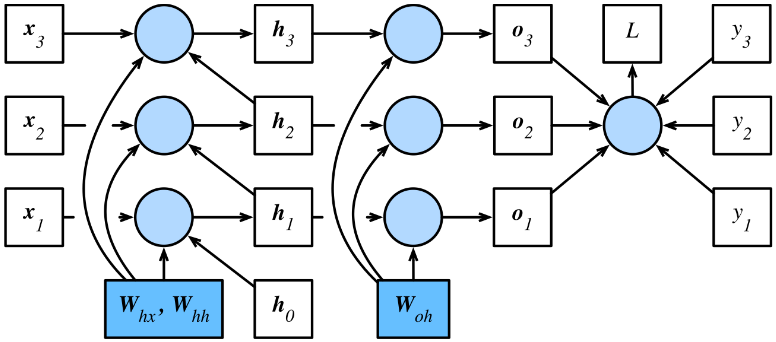
$$\frac{\partial h_t}{\partial w_h} = \frac{\partial f(x_t, h_{t-1}, w_h)}{\partial w_h} + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} \frac{\partial f(x_j, h_{j-1}, w_h)}{\partial h_{j-1}} \right) \frac{\partial f(x_i, h_{i-1}, w_h)}{\partial w_h}$$



Computational Graph



Class Activity



$$\frac{\partial L}{\partial o_t} =$$

$$\frac{\partial L}{\partial w_{qh}} =$$

$$\frac{\partial L}{\partial h_T} =$$

$$\frac{\partial L}{\partial W_{hx}} =$$

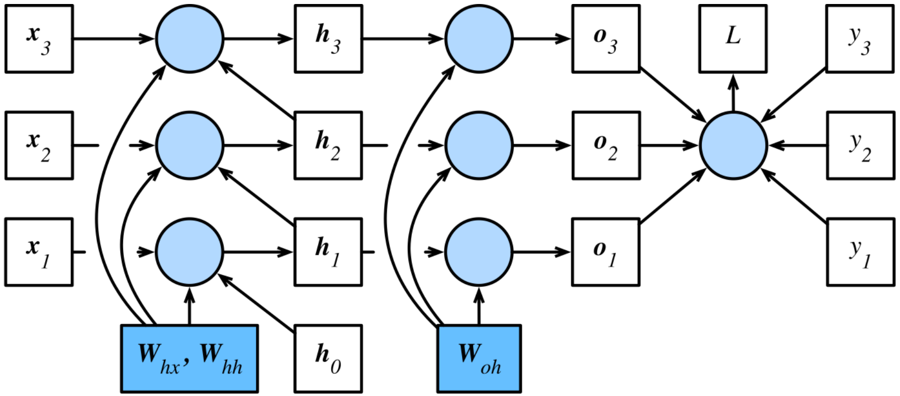
$$\frac{\partial L}{\partial W_{hh}} =$$

Toy Model

$$h_t = W_{hx}x_t + W_{hh}h_{t-1}$$

$$o_t = W_{qh}h_t$$

$$L = \frac{1}{T}\sum_{t=1}^{T} l(y_t, 0_t)$$



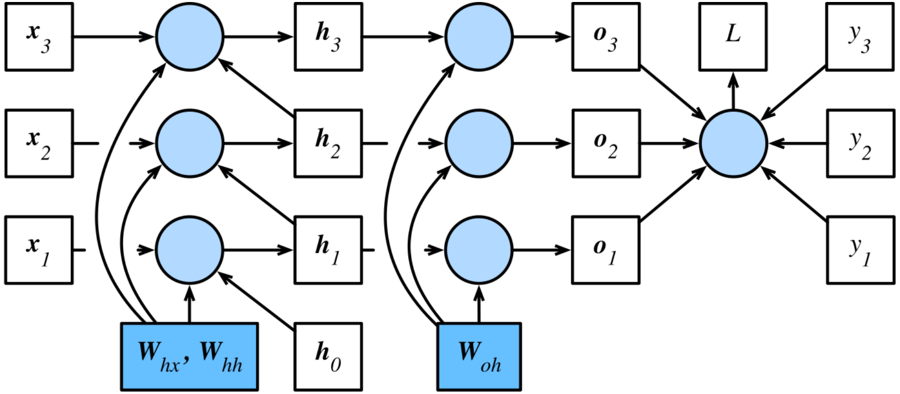
$$\frac{\partial L}{\partial o_t} = \frac{\partial l(o_t, y_t)}{T.\partial o_t} \in R^q$$

Toy Model

$$h_t = W_{hx}x_t + W_{hh}h_{t-1}$$

$$o_t = W_{qh}h_t$$

$$L = \frac{1}{T}\sum_{t=1}^{T} l(y_t, 0_t)$$



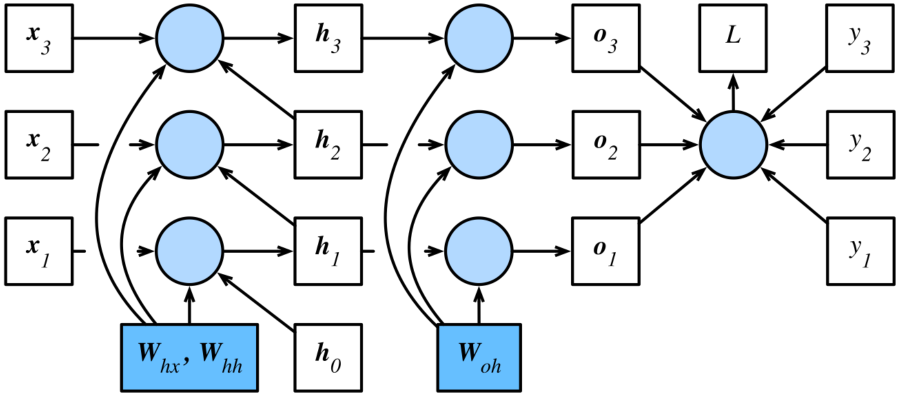
$$\frac{\partial L}{\partial w_{qh}} = \sum_{t=1}^{T} prod\left(\frac{\partial L}{\partial o_t}, \frac{\partial o_t}{\partial w_{qh}}\right)$$

Toy Model

$$h_t = W_{hx}x_t + W_{hh}h_{t-1}$$

$$o_t = W_{qh}h_t$$

$$L = \frac{1}{T}\sum_{t=1}^{T} l(y_t, 0_t)$$



$$\frac{\partial L}{\partial h_T} = prod\left(\frac{\partial L}{\partial o_t}, \frac{\partial o_t}{\partial h_T}\right)$$

$$\frac{\partial L}{\partial h_t} = prod\left(\frac{\partial L}{\partial h_{t+1}}, \frac{\partial h_{t+1}}{\partial h_T}\right) + prod\left(\frac{\partial L}{\partial o_t}, \frac{\partial o_t}{\partial h_T}\right)$$

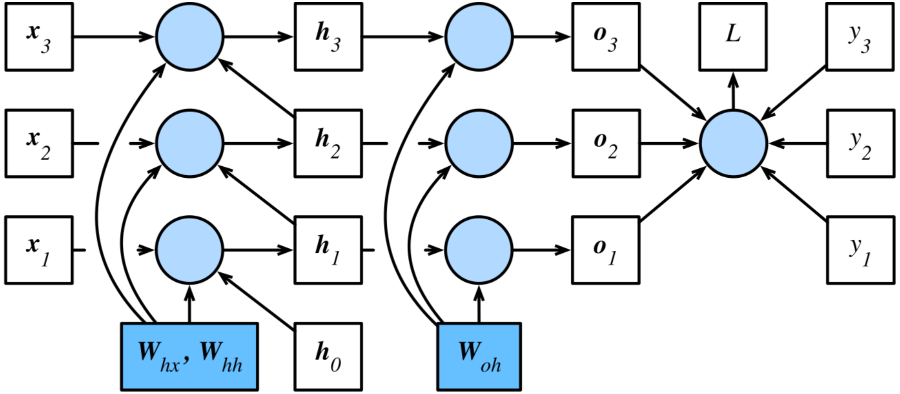
$$\frac{\partial L}{\partial h_t} = \sum_{i=t}^{T} \left(W_{hh}^T\right)^{T-i} W_{qh}^T \frac{\partial L}{\partial o_{T+t-i}} \qquad 1 \le t \le T$$

Toy Model

$$h_t = W_{hx} x_t + W_{hh} h_{t-1}$$

$$o_t = W_{qh} h_t$$

$$L = \frac{1}{T} \sum_{t=1}^{T} l(y_t, 0_t)$$

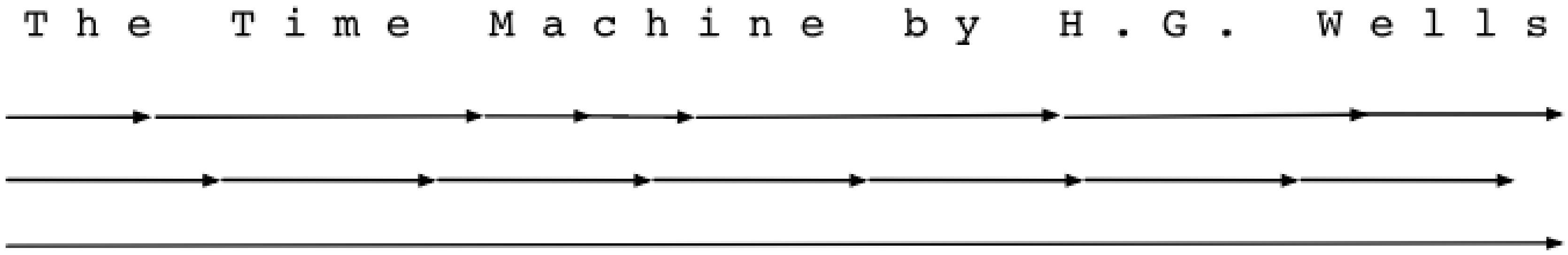


$$\frac{\partial L}{\partial W_{hx}} = \sum_{i=t}^{T} prod\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hx}}\right)$$

$$\frac{\partial L}{\partial W_{hh}} = \sum_{i=t}^{T} prod\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hh}}\right)$$

Truncated BPTT

- Don't truncate (naive strategy, costly and divergent)
- Truncate at fixed intervals
  (standard approach, is approximation but works well)
- Variable length (Tallec and Olivier, 2015)



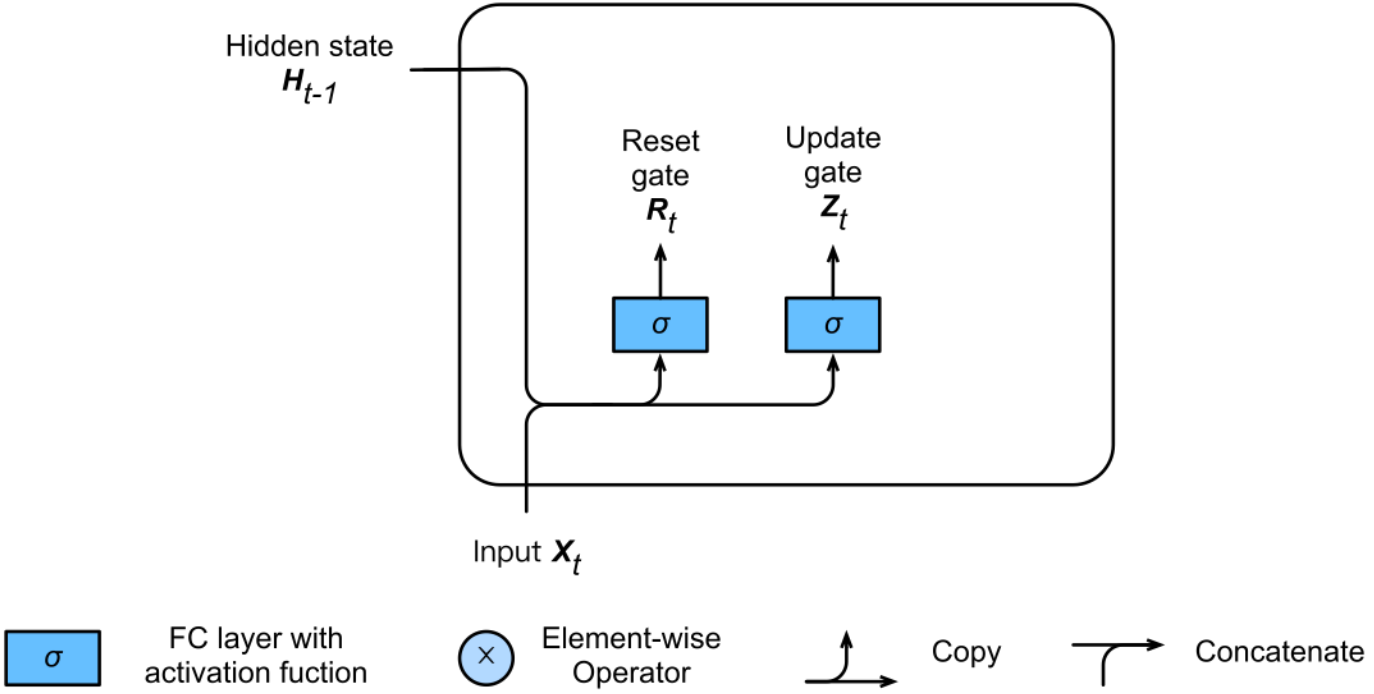
Paying attention to a sequence

- Not all observations are equally relevant

- Only remember the relevant ones
  - Need mechanism to pay attention (update gate)
  - Need mechanism to forget (reset gate)

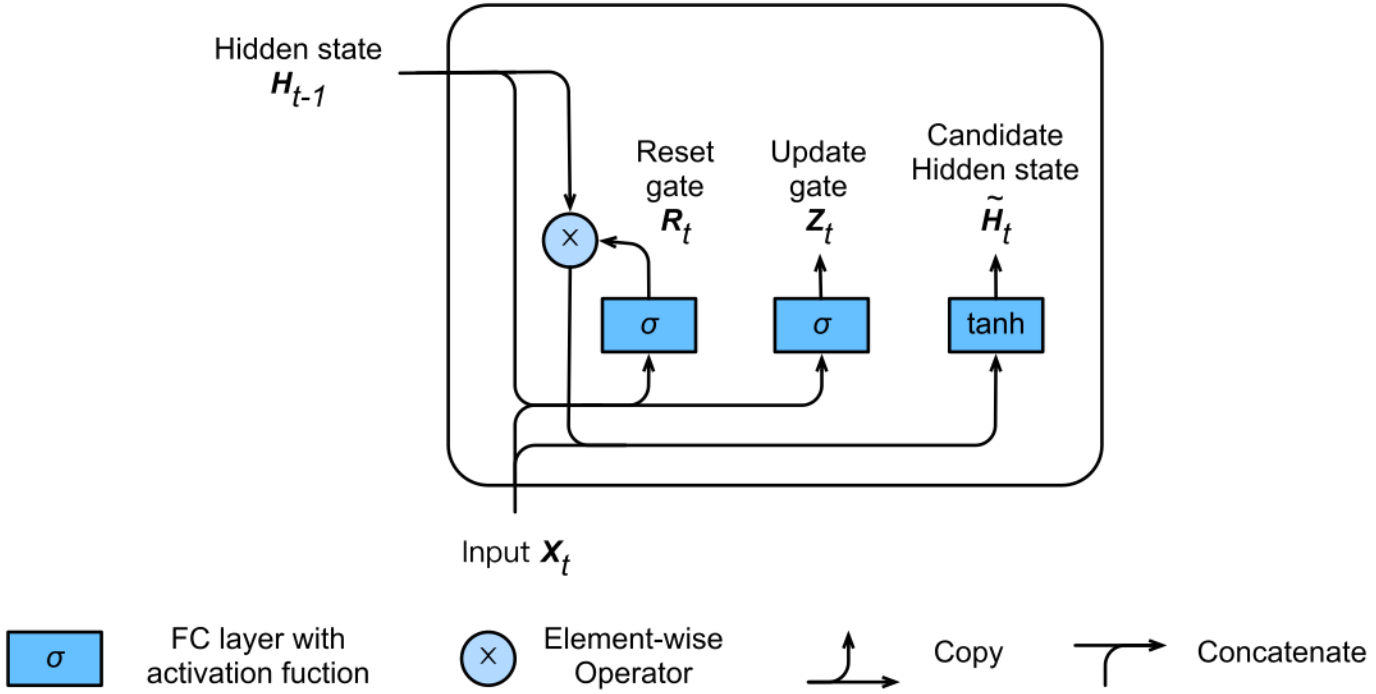
Gating

$$\boldsymbol{R}_t = \sigma(\boldsymbol{X}_t\boldsymbol{W}_{xr} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hr} + \boldsymbol{b}_r),$$
$$\boldsymbol{Z}_t = \sigma(\boldsymbol{X}_t\boldsymbol{W}_{xz} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hz} + \boldsymbol{b}_z)$$



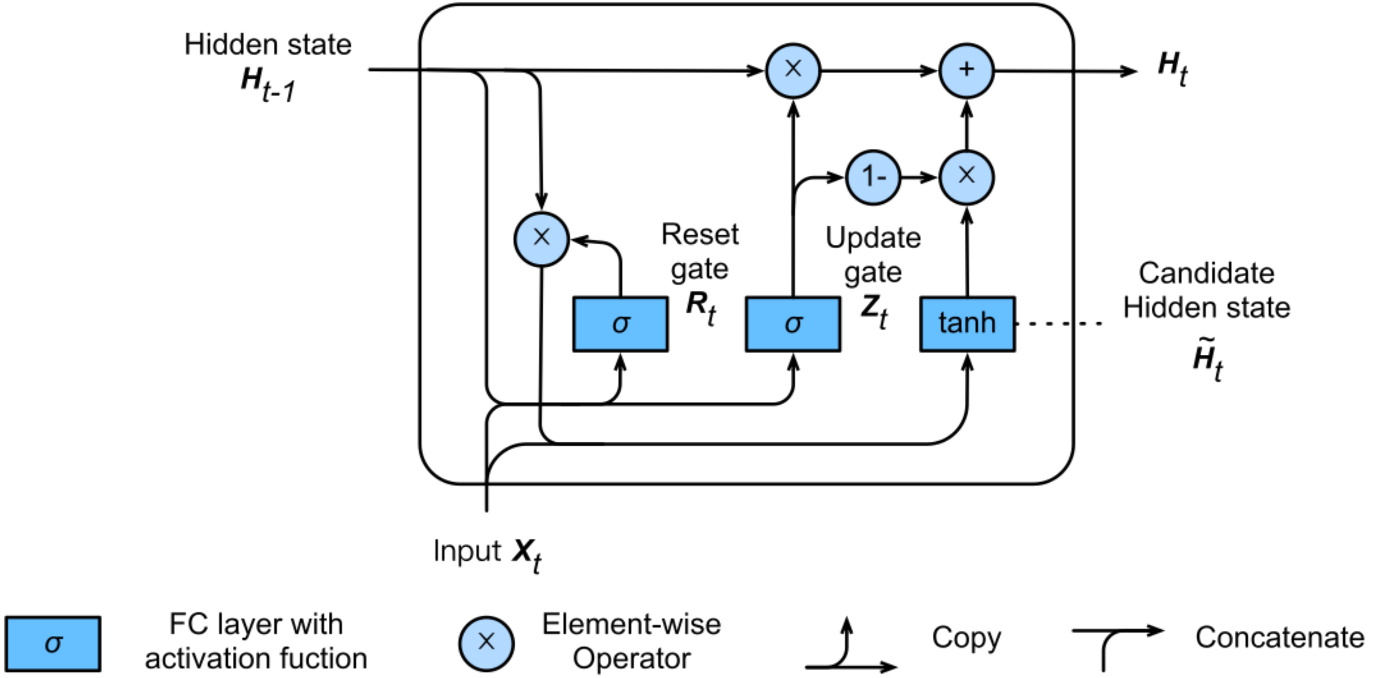
Candidate Hidden State

$$\tilde{\boldsymbol{H}}_t = \tanh(\boldsymbol{X}_t \boldsymbol{W}_{xh} + \left(\boldsymbol{R}_t \odot \boldsymbol{H}_{t-1}\right) \boldsymbol{W}_{hh} + \boldsymbol{b}_h)$$



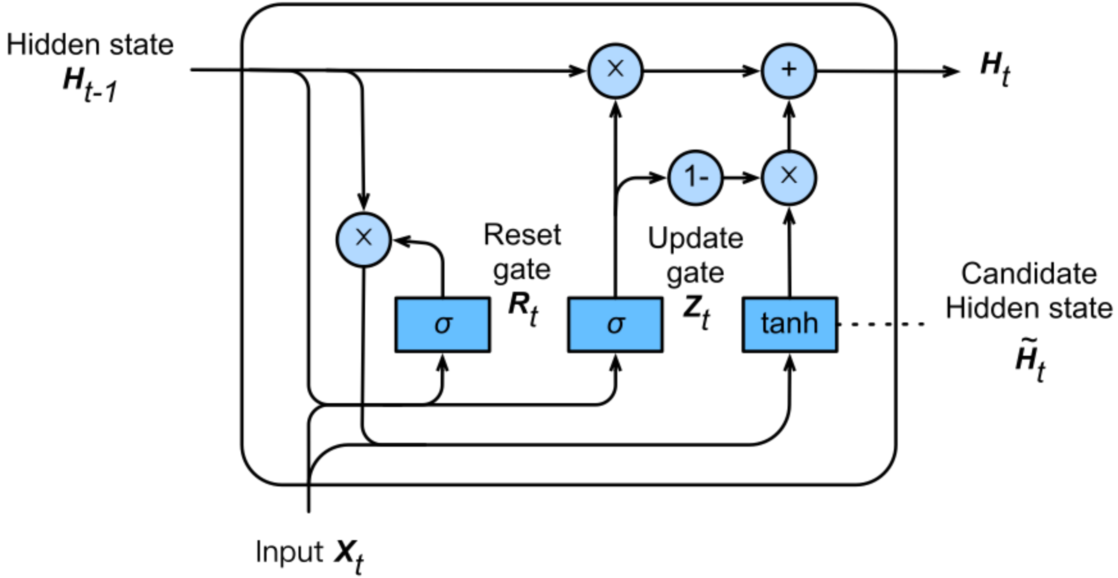
GRU-Hidden State

$$\boldsymbol{H}_t = \mathbf{Z}_t \odot \boldsymbol{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\boldsymbol{H}}_t$$



GRU-Summary

$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r),$$
$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z)$$
$$\tilde{H}_t = \tanh(X_t W_{xh} + \left(R_t \odot H_{t-1}\right) W_{hh} + b_h)$$
$$H_t = Z_t \odot H_{t-1} + (1 - Z_t) \odot \tilde{H}_t$$



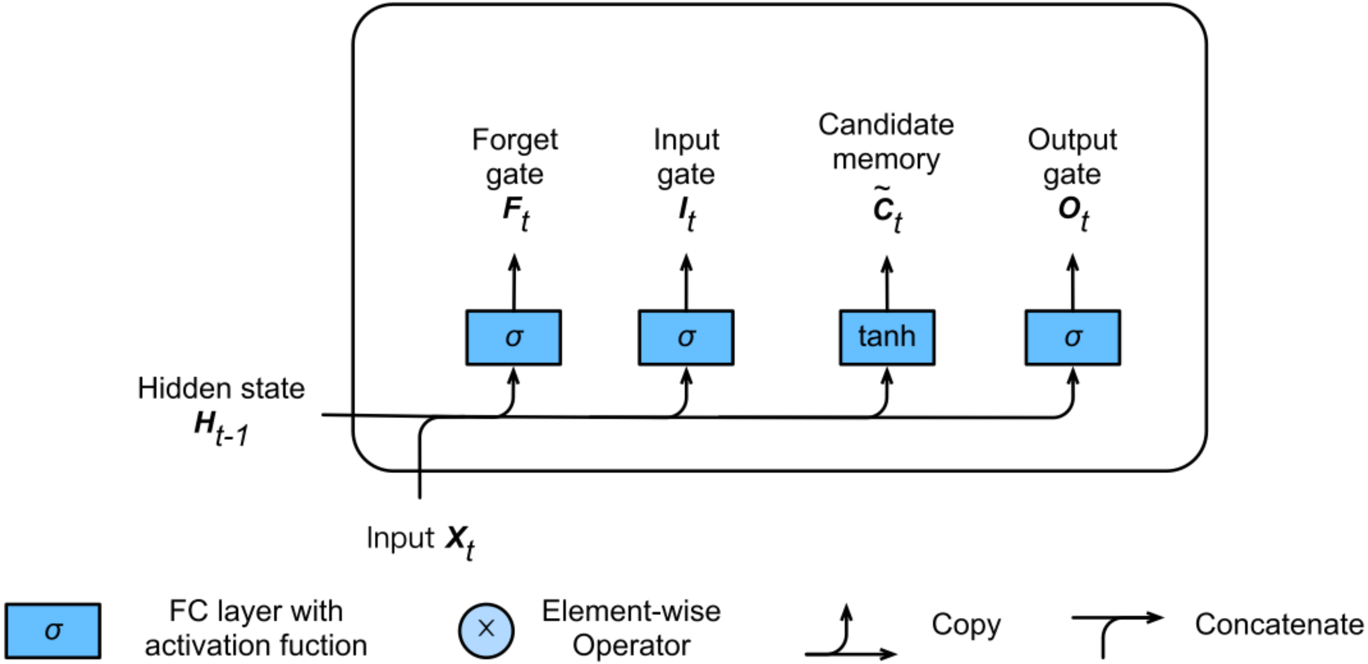
Long Short Term Memory-LSTM

- **Forget gate**
  Shrink values towards zero
- **Input gate**
  Decide whether we should ignore the input data
- **Output gate**
  Decide whether the hidden state is used for the output generated by the LSTM

- **Hidden state** and **Memory cell**

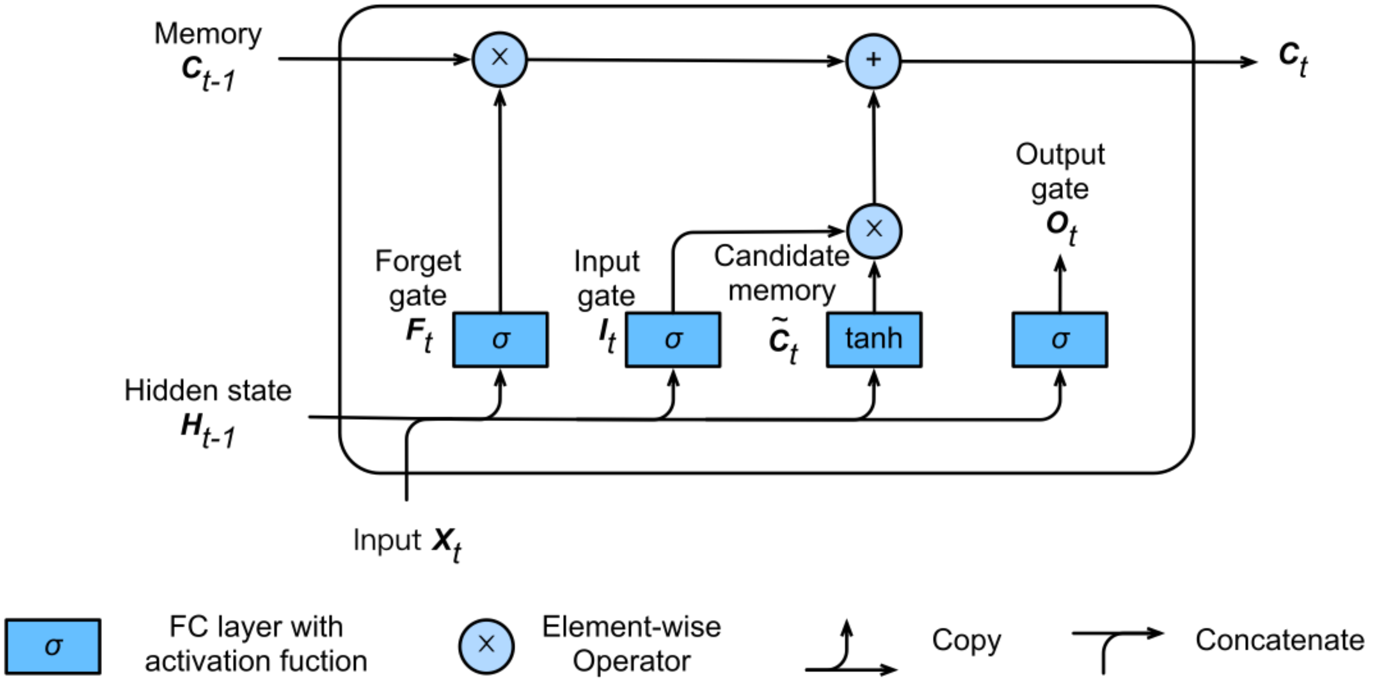
Gates

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$$
$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$$
$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$$



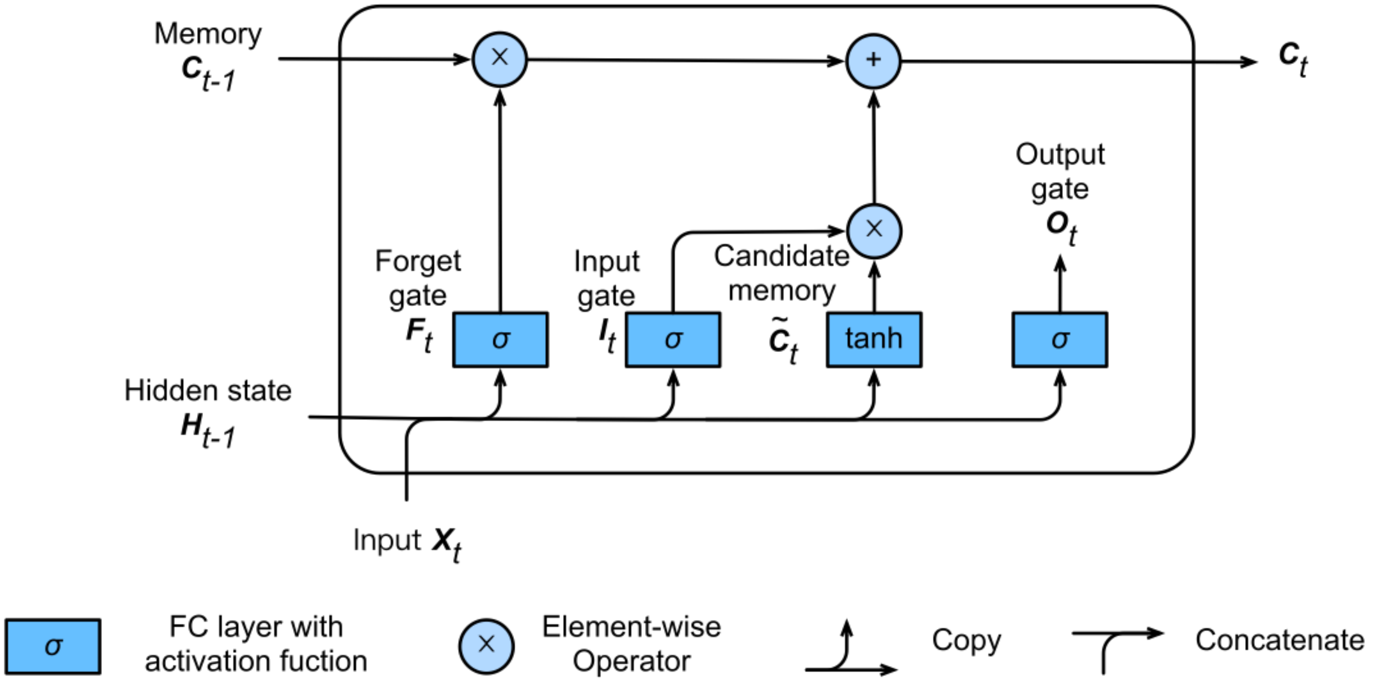
Candidate Memory Cell

$$\tilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$



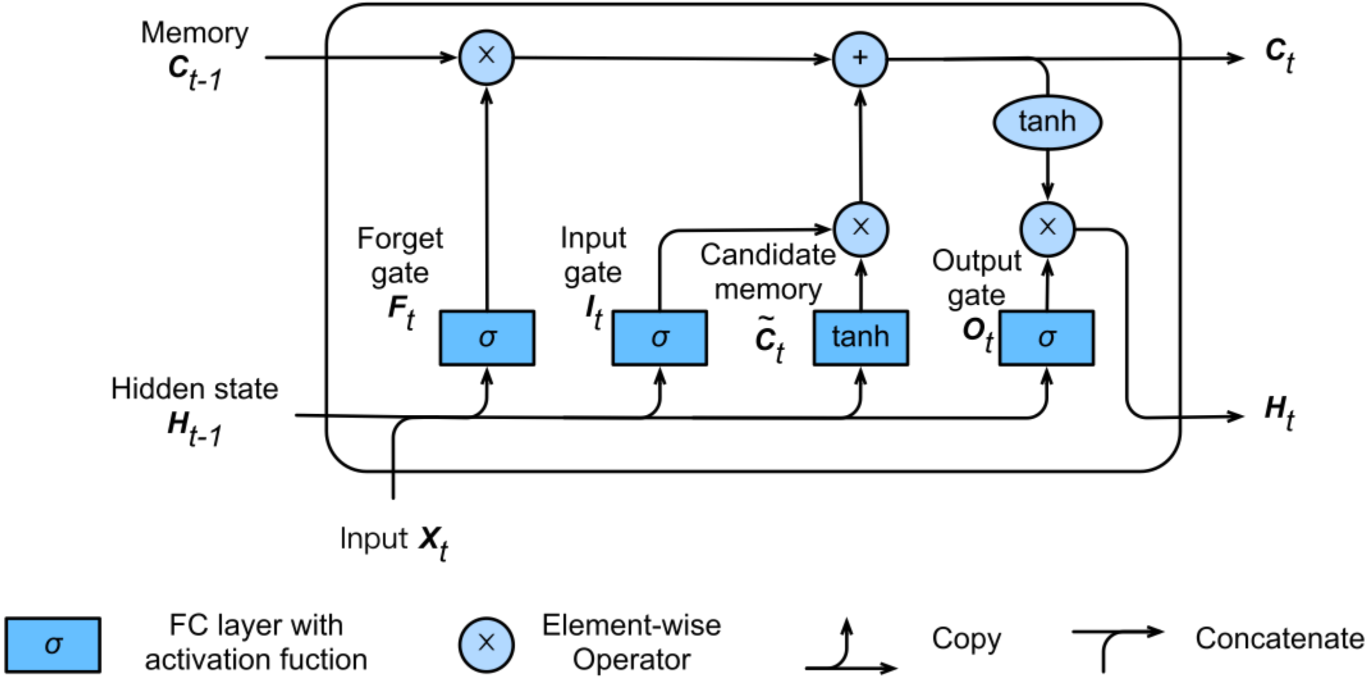
Memory Cell

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

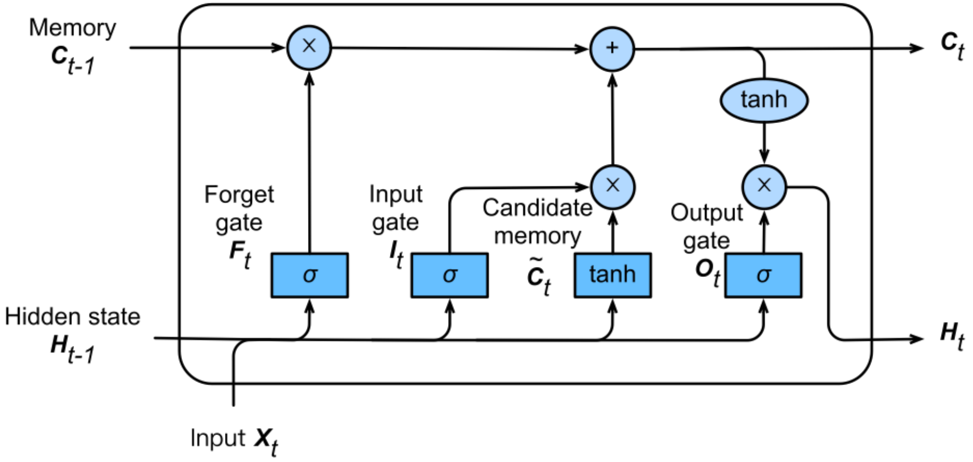


Hidden State / Output

$$\boldsymbol{H}_t = \boldsymbol{O}_t \odot \tanh(\boldsymbol{C}_t)$$



Hidden State / Output



$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$$

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$$

$$\tilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

$$H_t = O_t \odot \tanh(C_t)$$