Gated RNNs

Geena Kim

University of Colorado **Boulder**

How hard is it to train an RNN?

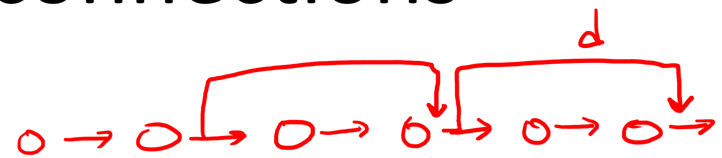
- Slow to train (TBPTT helps)
- RNN can suffer exploding/vanishing gradient
- First or early memory or info get lost through the time step

Remedies

- ReLU activation function
- Truncated BPTT
- Clip gradient
- Use learning rate scheduling
- Add residual connection
- Change architectures- LSTM, GRU

Long-term dependencies

- Skip connections



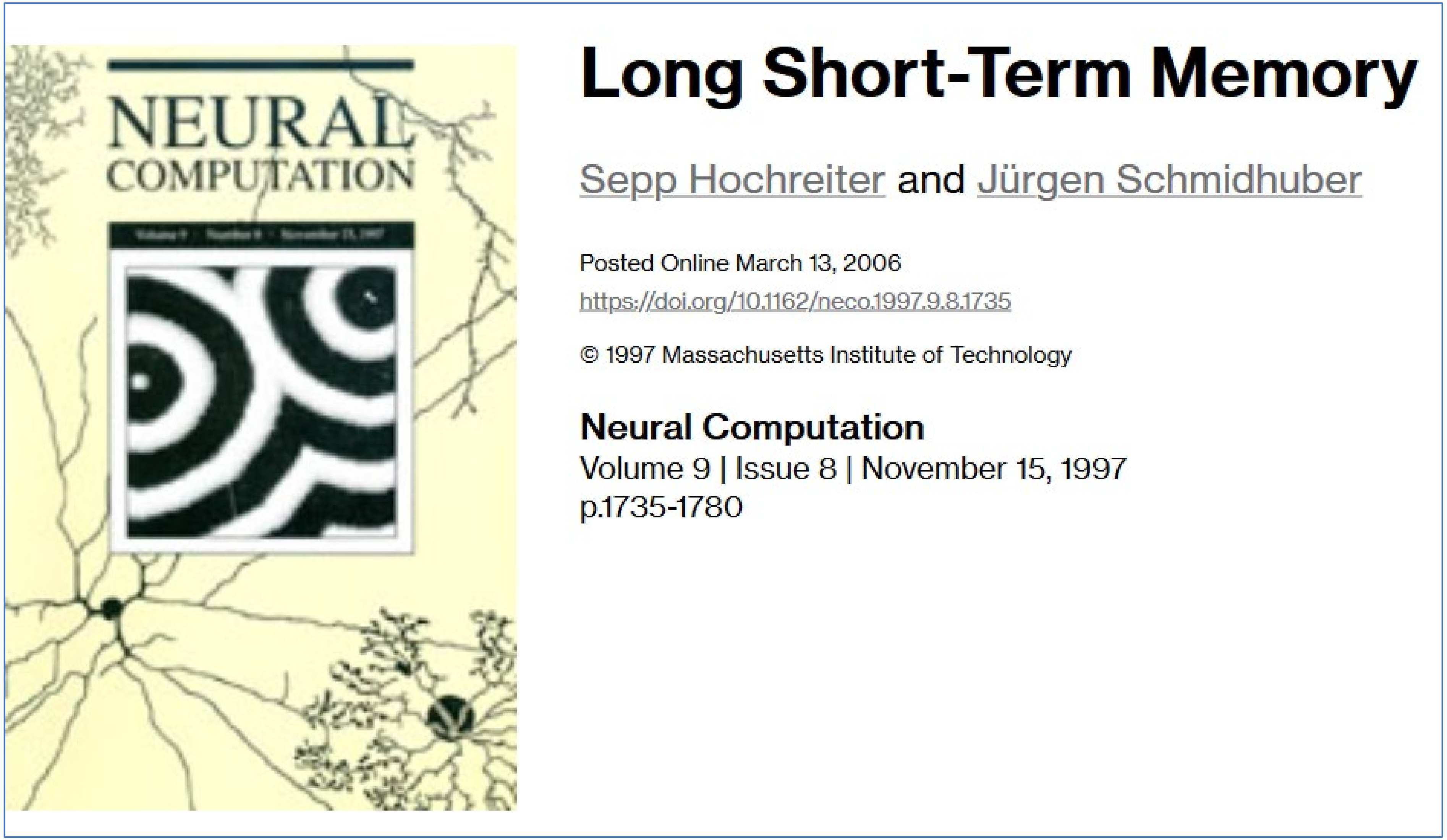
$$|W|^{N} \rightarrow |W|^{N/d}$$

- Leaky units

$$\underset{\sim}{h_t} = \underline{f(Wx_t + Uh_{t-1})}$$

$$\underset{\sim}{h_t} = \alpha h_{t-1} + (1-\alpha) \cdot f(Wx_t + Uh_{t-1})$$

Long Short-Term Memory cell



Long Short-Term Memory

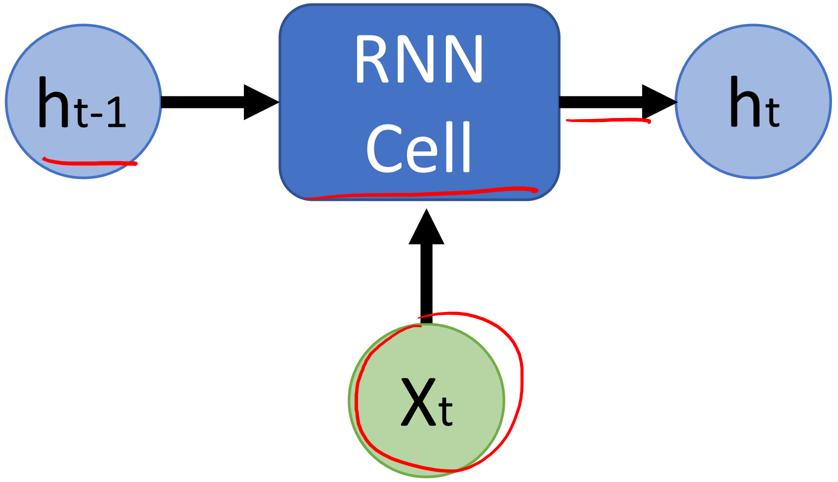
Sepp Hochreiter and Jürgen Schmidhuber

Posted Online March 13, 2006
https://doi.org/10.1162/neco.1997.9.8.1735

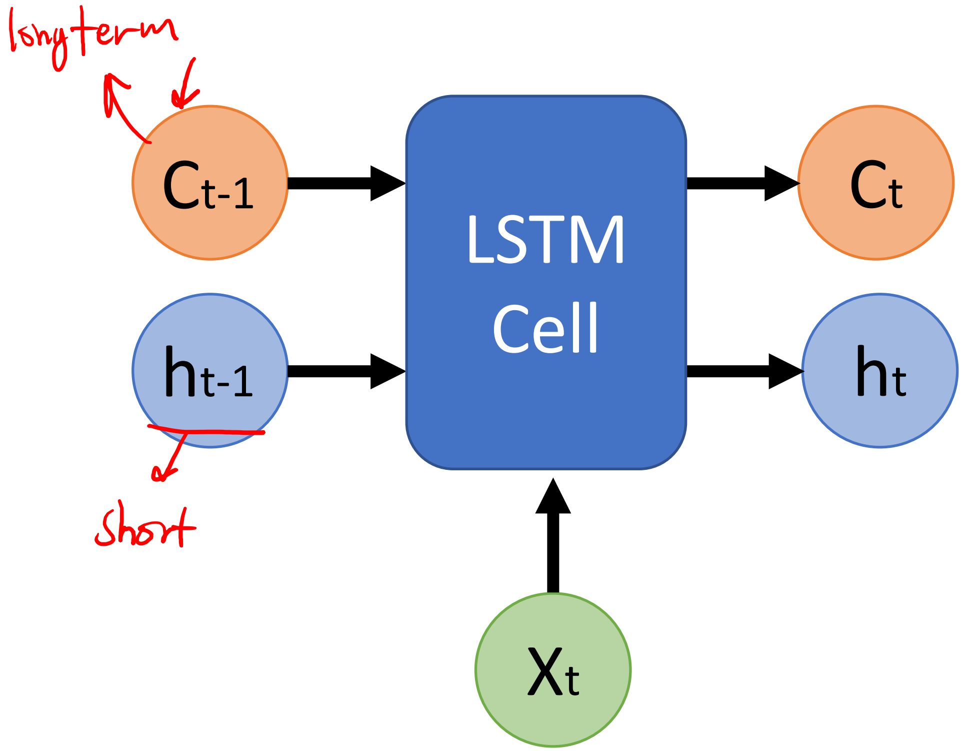
Neural Computation
Volume 9 | Issue 8 | November 15, 1997
p.1735-1780

What is LSTM cell?

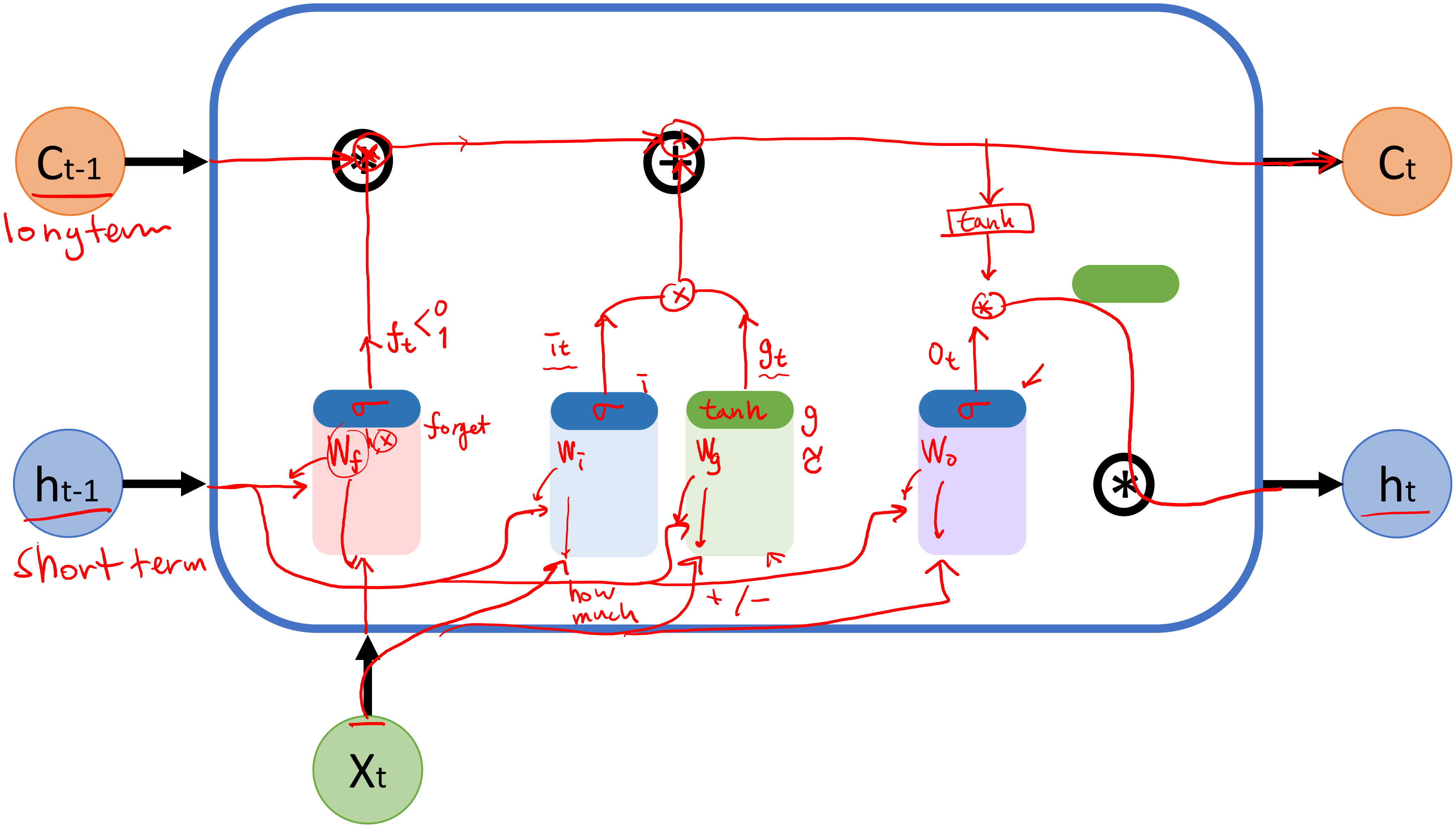
A Vania RNN cell



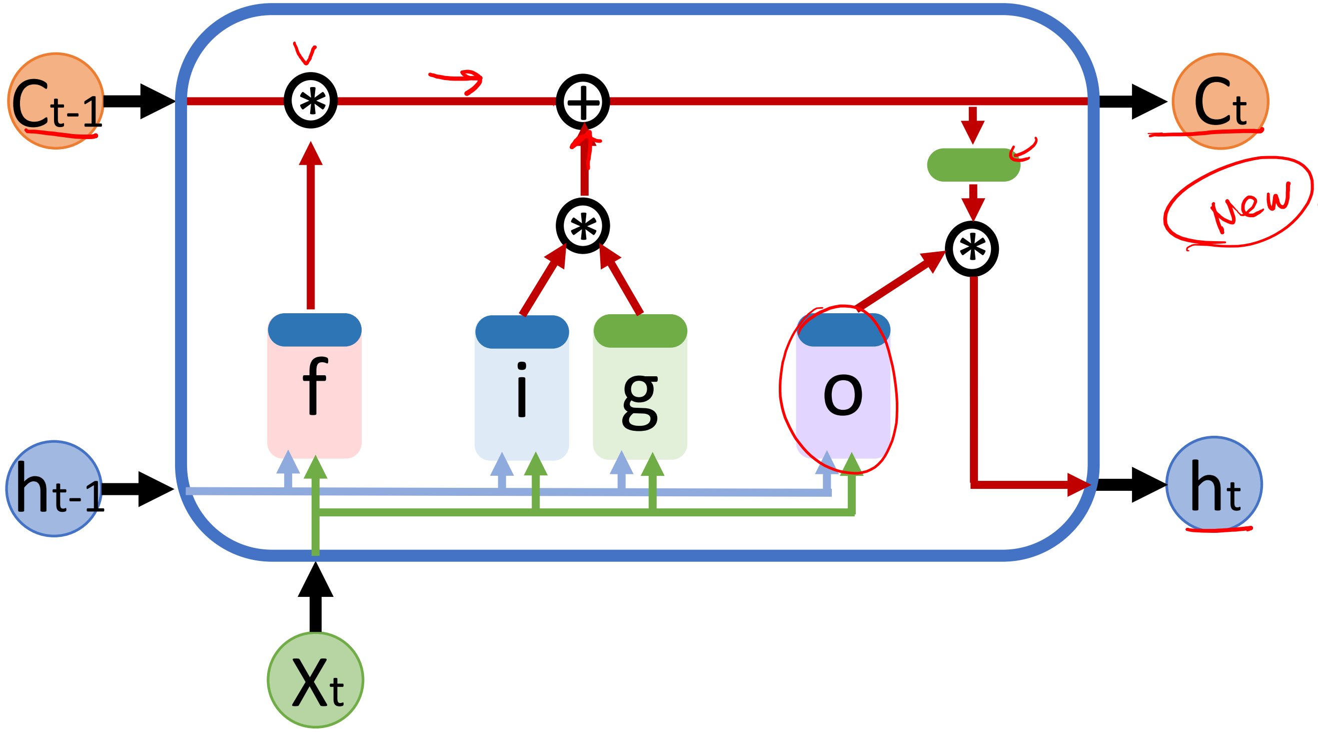
An LSTM cell



Inside the LSTM cell



Inside the LSTM cell

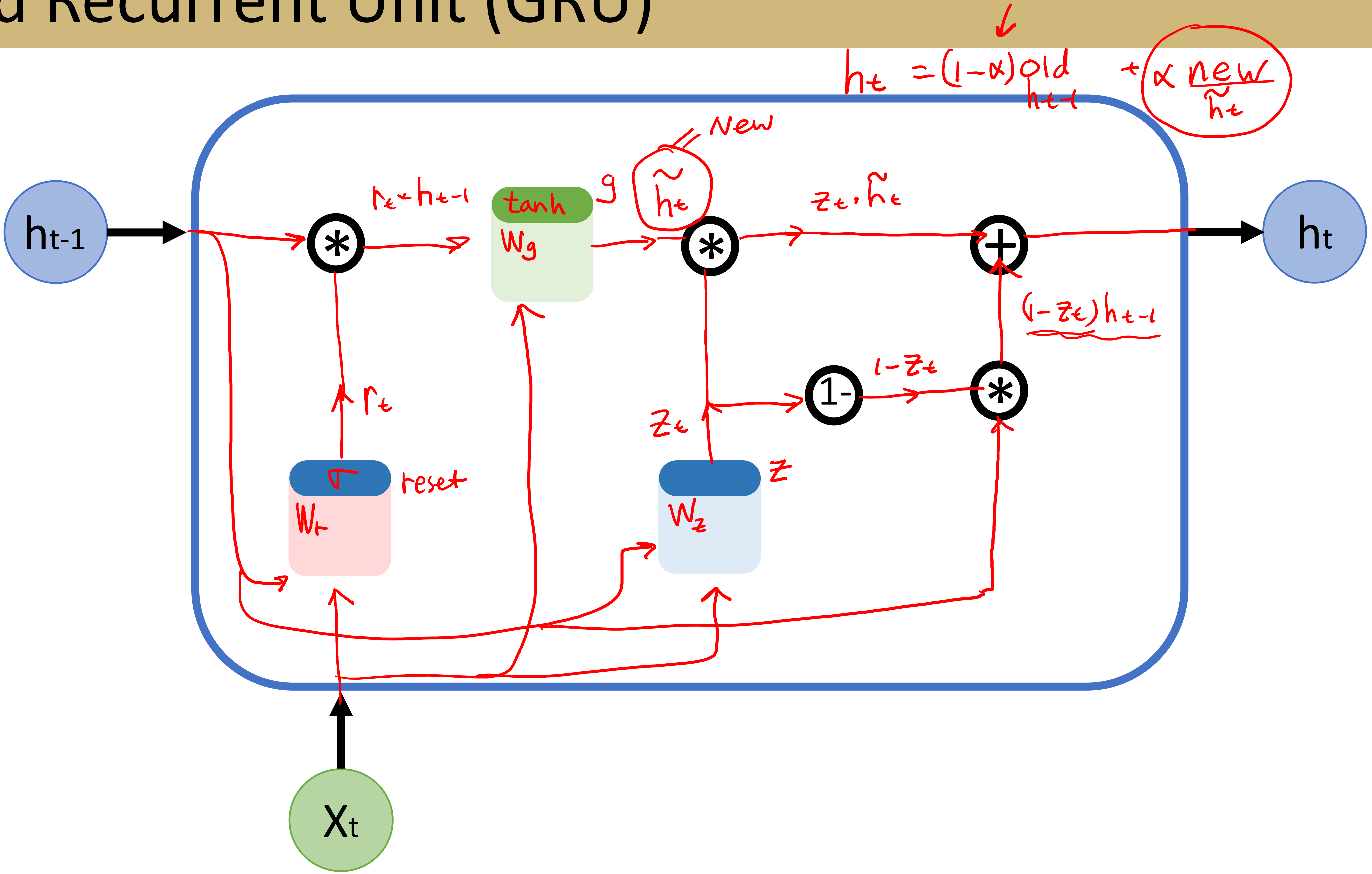


$$f_t = \sigma(W_f \cdot [X_t, h_{t-1}] + b_f)$$
$$i_t = \sigma(W_i \cdot [X_t, h_{t-1}] + b_i)$$
$$g_t = \tanh(W_g \cdot [X_t, h_{t-1}] + b_g)$$
$$o_t = \sigma(W_o \cdot [X_t, h_{t-1}] + b_o)$$

New

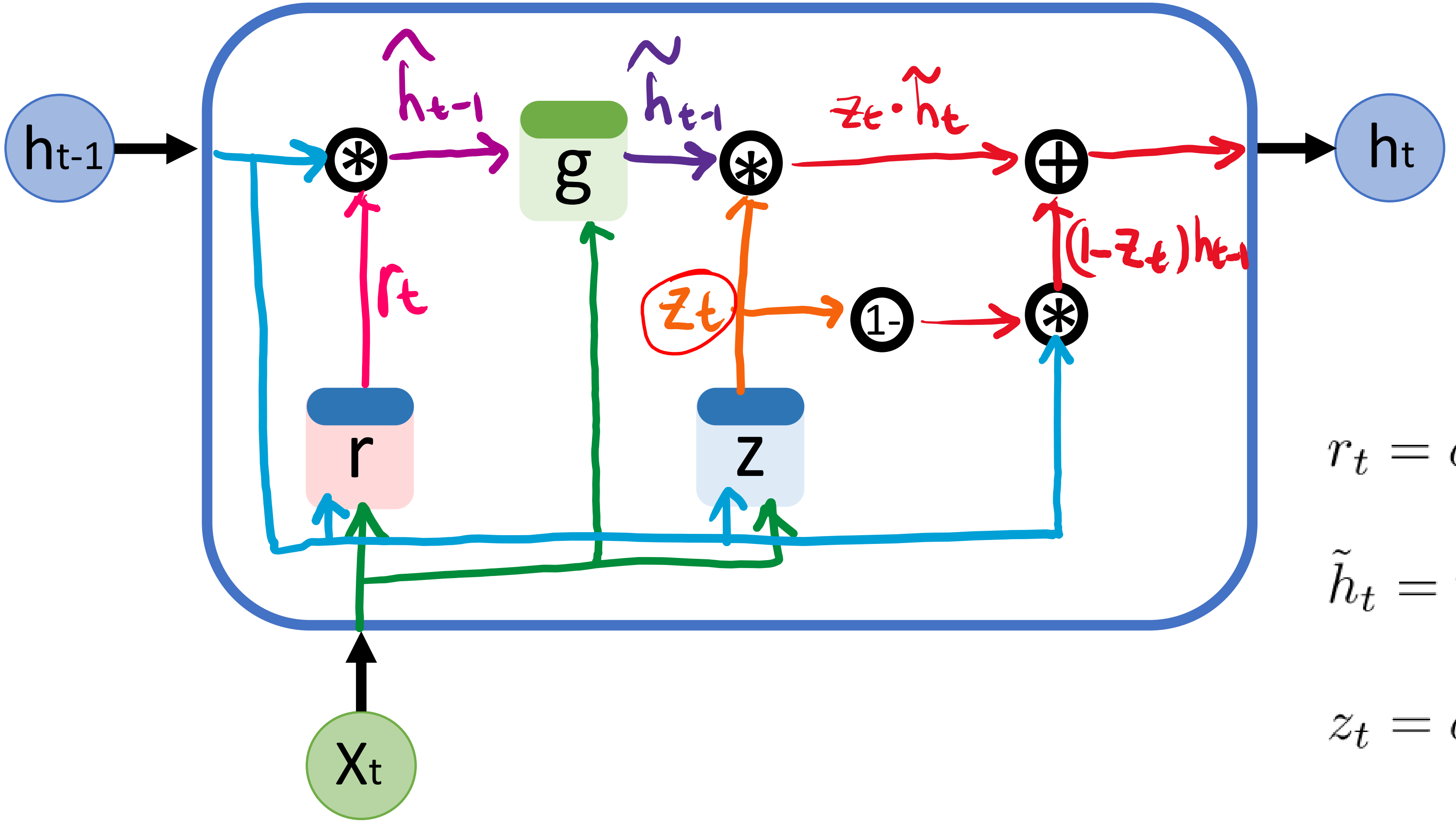
$$c_t = f_t * c_{t-1} + i_t * g_t$$
$$h_t = o_t * \tanh(c_t)$$

Gated Recurrent Unit (GRU)



$$h_t = (1-\alpha)\underset{h_{t-1}}{\text{old}} + \boxed{\alpha \underset{\tilde{h}_t}{\underline{\text{new}}}}$$

Gated Recurrent Unit (GRU)



✓

$$r_t = \sigma(W_r \cdot [X_t, h_{t-1}] + b_r)$$

$$\tilde{h}_t = \tanh(W_g \cdot [X_t, r_t * h_{t-1}] + b_g)$$

$$z_t = \sigma(W_z \cdot [X_t, h_{t-1}] + b_z)$$

$$h_t = z_t * \tilde{h}_t + (1 - z_t) * h_{t-1}$$