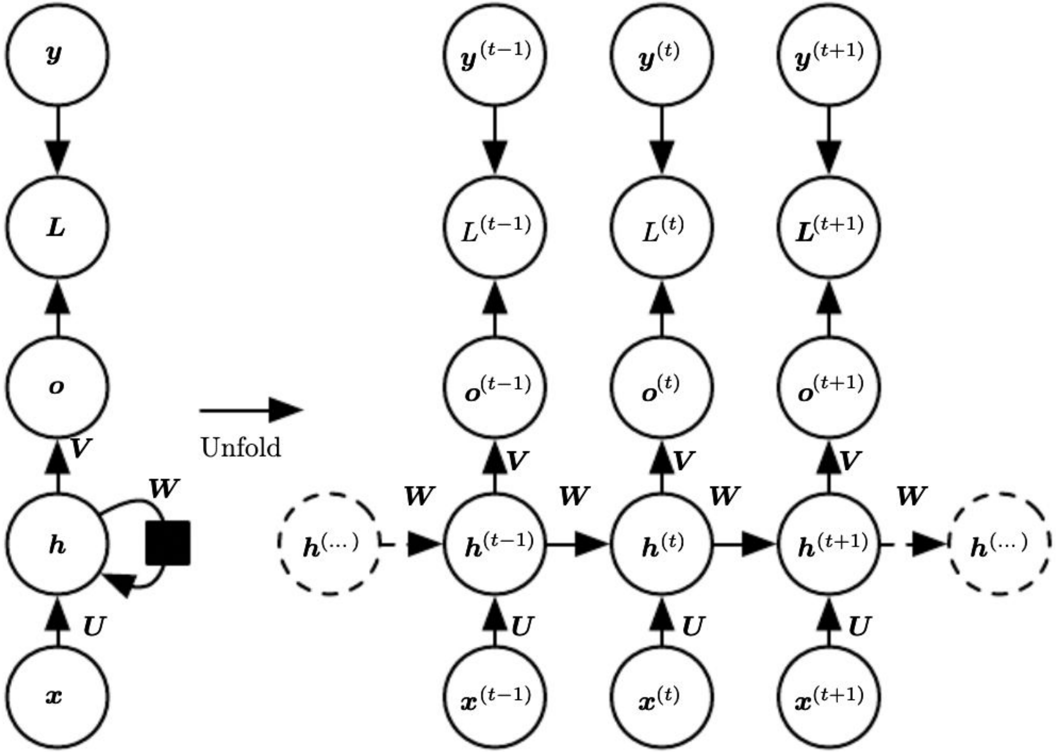
Advanced Machine Learning

Likhit Nayak

Sequence modelling



Backpropagation through time (BPTT)

$$\begin{cases} h_t = Ux_t + Wh_{t-1} \\ o_t = Vh_t \end{cases}$$

where :

h_t : is the hidden state

o_t : is the output at time step t

x_t : is the inputat time step t

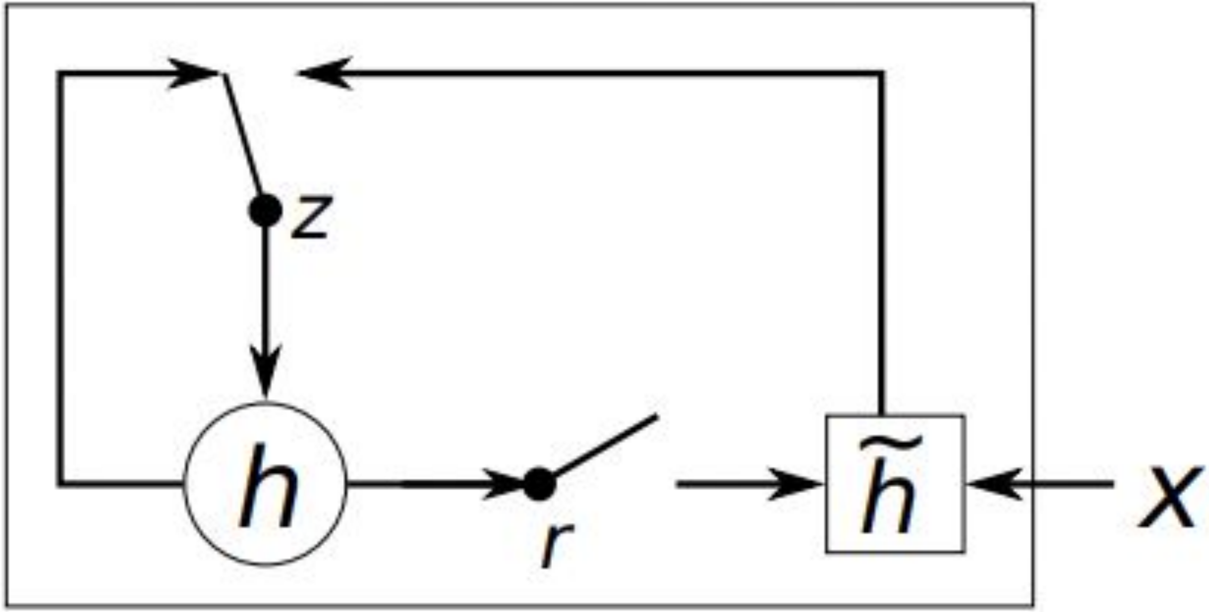
Backpropagation through time (BPTT)

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{T} \frac{\partial l}{\partial o_t} . h_t^{\intercal}$$

$$\frac{\partial L}{\partial W} = \sum_{t}^{T} \sum_{k=1}^{t+1} \frac{\partial L_{t+1}}{\partial o_{t+1}} . \frac{\partial o_{t+1}}{\partial h_{t+1}} . \frac{\partial h_{t+1}}{\partial h_k} . \frac{\partial h_k}{\partial W}$$

$$\frac{\partial L}{\partial U} = \sum_{t}^{T} \sum_{k=1}^{t+1} \frac{\partial L_{t+1}}{\partial o_{t+1}} . \frac{\partial o_{t+1}}{\partial h_{t+1}} . \frac{\partial h_{t+1}}{\partial h_k} . \frac{\partial h_k}{\partial U}$$

Gated Recurrent Unit (GRU)



Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).

Gated Recurrent Unit (GRU)

$$r_j = \sigma\left(\left[\mathbf{W}_r\mathbf{x}\right]_j + \left[\mathbf{U}_r\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

$$z_j = \sigma\left(\left[\mathbf{W}_z\mathbf{x}\right]_j + \left[\mathbf{U}_z\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

$$h_j^{\langle t\rangle} = z_j h_j^{\langle t-1\rangle} + (1 - z_j)\tilde{h}_j^{\langle t\rangle}$$

$$\tilde{h}_j^{\langle t\rangle} = \phi\left(\left[\mathbf{W}\mathbf{x}\right]_j + \left[\mathbf{U}\left(\mathbf{r}\odot\mathbf{h}_{\langle t-1\rangle}\right)\right]_j\right)$$

Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).