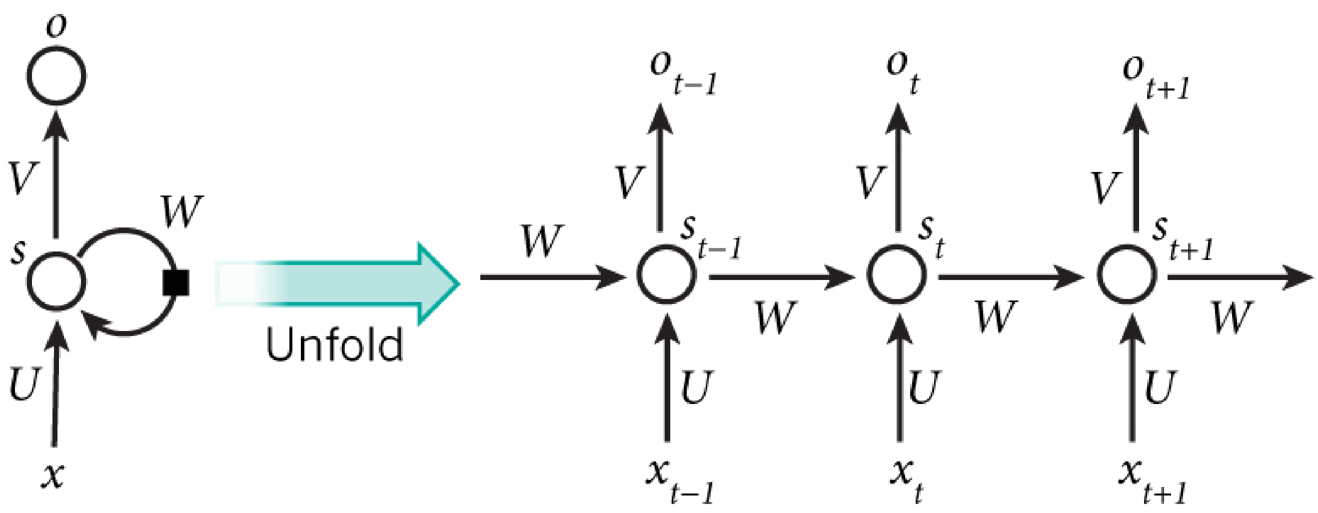
Machine Learning

Recurrent Neural Network



1. Basics

sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\sigma'(x) = \sigma(x) \cdot [1 - \sigma(x)]$$

hyperbolic function:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\tanh'(x) = 1 - \tanh^2(x)$$

rectified linear unit(ReLU):

$$f(x) = \max(0, x)$$

softmax function:

$$\mathbf{y} = \mathrm{softmax}(\mathbf{x})$$

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

$$\frac{\partial y_i}{\partial x_j} = \begin{cases} -y_i \cdot y_j, & i \neq j \\ y_i \cdot (1 - y_i), & i = j \end{cases}$$

2. Model

input:

$$x = (x_1, x_2, \ldots, x_T) \quad x_t \in \mathbb{R}^n$$

initialize hidden state:

$$s_0 \in \mathbb{R}^k$$

forward propagation:

$$\begin{aligned} s_t &= \tanh(Ux_t + Ws_{t-1}) \quad (t = 1, 2, \ldots, T) \\ \hat{y}_t &= \mathrm{softmax}(Vs_t) \quad (t = 1, 2, \ldots, T) \end{aligned}$$

output:

$$\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T) \quad \hat{y}_t \in \mathbb{R}^m$$

3. Backpropagation Through Time

cost function:

$$E(\hat{y}) = \sum_{t=1}^{T} E_t(\hat{y}_t)$$

definition:

$$\begin{aligned} h_t &= Ux_t + Ws_{t-1} \quad (t = 1, 2, \ldots, T) \\ z_t &= Vs_t \quad (t = 1, 2, \ldots, T) \end{aligned}$$

gradient for V:

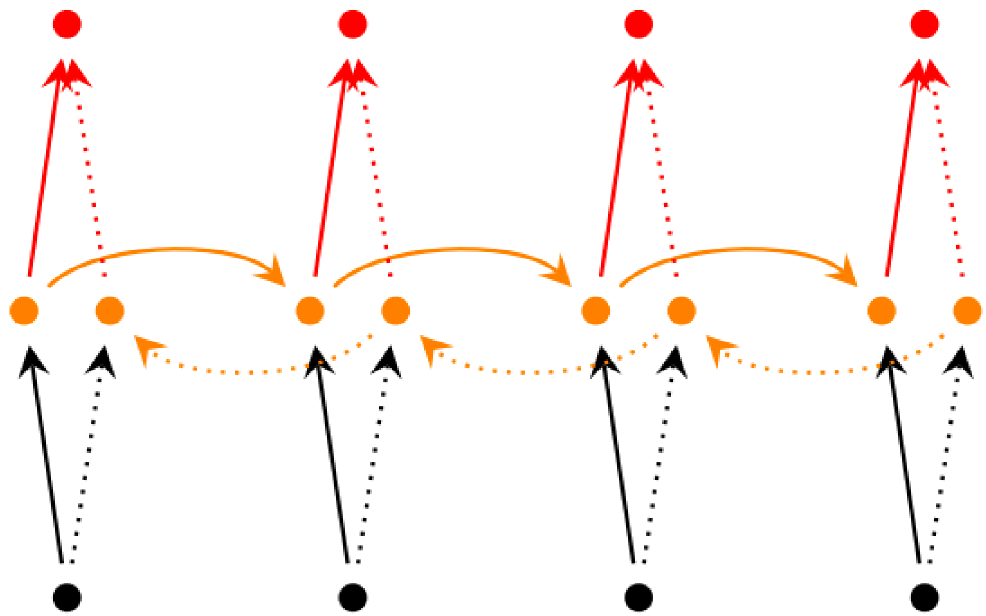
$$\begin{aligned}\frac{\partial E_t}{\partial V} = \frac{\partial E_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial V} &= \frac{\partial E_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial V} \\ &= \left(\frac{\partial E_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t}\right) \cdot s_t^T \quad (\text{need } \hat{y}_t, s_t; t = 1, 2, \ldots, T)\end{aligned}$$

gradient for W:

$$\begin{aligned}\frac{\partial s_1}{\partial W} &= \frac{\partial s_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial W} \quad (\text{need } s_1, s_0) \\ \frac{\partial s_t}{\partial W} &= \frac{\partial s_t}{\partial h_t} \cdot \left(\frac{\partial h_t}{\partial W} + W \cdot \frac{\partial s_{t-1}}{\partial W}\right) \quad (\text{need } s_t, s_{t-1}; t = 2, 3, \ldots, T) \\ \frac{\partial E_t}{\partial W} &= \frac{\partial E_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial s_t} \cdot \frac{\partial s_t}{\partial W} \\ &= \left(\frac{\partial E_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t}\right)^T \cdot V \cdot \frac{\partial s_t}{\partial W} \quad (\text{need } \hat{y}_t; t = 1, 2, \ldots, T)\end{aligned}$$

4. RNN Extensions

Bidirectional RNNs:



Deep (Bidirectional) RNNs: