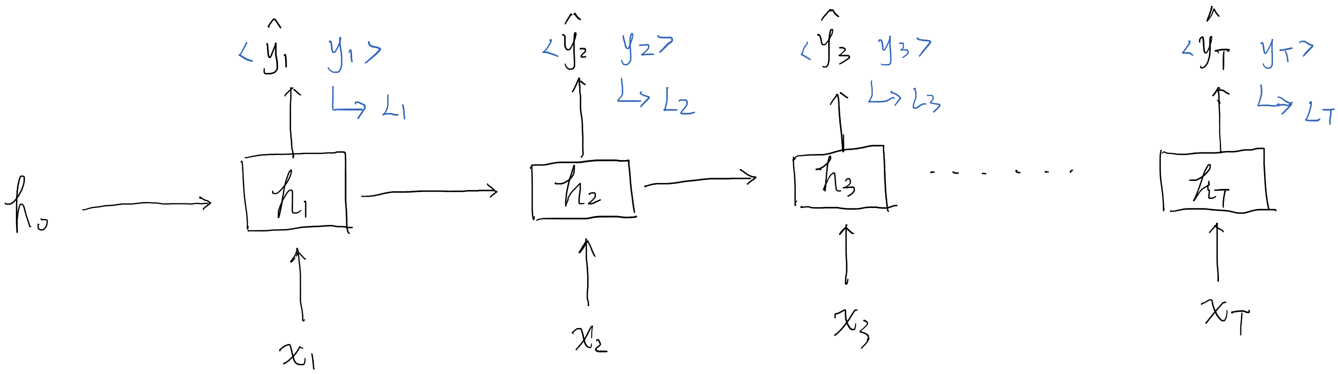
BPTT

2021年8月16日 10:45

Backprop through time (BPTT)



$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t$$ (summation of losses)

cross entropy $-y \log \hat{y}$

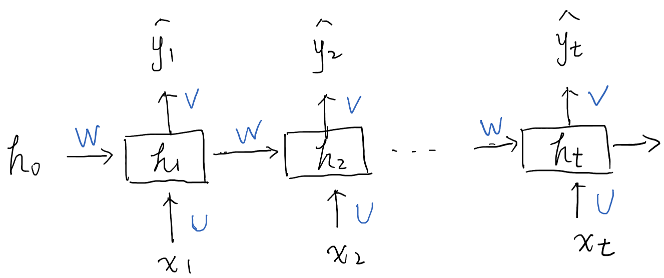
MSE $(y-\hat{y})^2$

non-linear function

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t + b) \qquad \hat{y}_t = g(W_{hy} h_t)$$

$W_{hh} \to W$, $W_{xh} \to U$, $W_{hy} \to V$

$$h_t = \tanh(W h_{t-1} + U x_t + b) \qquad \hat{y}_t = g(V h_t)$$



W, V, U are shared layers

To update weights :

Compute $\frac{\partial \mathcal{L}}{\partial W}$, $\frac{\partial \mathcal{L}}{\partial V}$, $\frac{\partial \mathcal{L}}{\partial U}$

Assumption :

① loss function is CE

② g = softmax function

① $\frac{\partial \mathcal{L}}{\partial V}$

$$\begin{cases} h_t = \tanh(W h_{t-1} + U x_t + b) \\ \hat{y}_t = g(V h_t) \\ \mathcal{L}_t = -y_t \log \hat{y}_t \\ \mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t \end{cases}$$

Let $z_t = V h_t$

Chain rule :

$$\frac{\partial \mathcal{L}}{\partial V} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial V}$$

$$= \sum_{t=1}^{T} \underbrace{\frac{\partial \mathcal{L}_t}{\partial \hat{y}_t}}_{A} \cdot \underbrace{\frac{\partial \hat{y}_t}{\partial z_t}}_{B} \cdot \underbrace{\frac{\partial z_t}{\partial V}}_{C}$$

$$A: \frac{\partial \mathcal{L}_t}{\partial \hat{y}_t} = \frac{\partial(-y_t \log \hat{y}_t)}{\partial \hat{y}_t} = -y_t \frac{\partial \log \hat{y}_t}{\partial \hat{y}_t} = -\frac{y_t}{\hat{y}_t}$$

$$B: \frac{\partial \hat{y}_t}{\partial z_t} = \frac{\partial (g(V h_t))}{\partial z_t} = g' \cdot V$$

$$\begin{cases} z_t = V h_t \\ g(z_t) = \mathrm{softmax}(z_t) = \frac{e^{z_t}}{\sum_{k=1}^{K} e^{z_k}} \end{cases}$$

Compute g' :

① case 1 : $t = k \Rightarrow e^{z_t} = e^{z_k}$

$$g' = \frac{\partial \frac{e^{z_t}}{\sum_{k=1}^{K} e^{z_k}}}{\partial z_t} = \frac{e^{z_t}}{\sum_{k=1}^{K} e^{z_k}} - e^{z_t}\left[\frac{e^{z_t}}{(\sum_{k=1}^{K} e^{z_k})^2}\right] = \hat{y}_t(1-\hat{y}_t)$$

② case 2 : $t \neq k \Rightarrow e^{z_t} \neq e^{z_k}$ (treat e^{z_k} as constant)

$$g' = \frac{-e^{z_t} \cdot e^{z_k}}{(\sum_{k=1}^{K} e^{z_k})^2} = -\hat{y}_t \hat{y}_k$$

① and ②,
$$\frac{\partial \hat{y}_t}{\partial z_t} = \begin{cases} \hat{y}_t(1-\hat{y}_t), & t = k \\ -\hat{y}_t \hat{y}_k, & t \neq k \end{cases}$$

$$A \times B = \frac{\partial \mathcal{L}_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t} = -\frac{y_t}{\hat{y}_t} \begin{cases} \hat{y}_t(1-\hat{y}_t), & t = k \\ -\hat{y}_t \hat{y}_k, & t \neq k \end{cases}$$

$$A \times B \sim \frac{\quad}{\partial \hat{y}_t} \cdot \frac{\quad}{\partial z_t} = -\frac{\partial L}{\hat{y}_t} \begin{cases} \\ -\hat{y}_t \hat{y}_k, & t \neq k \end{cases}$$

$$= \begin{cases} \hat{y}_t y_t - y_t, & t = k \\ y_t \hat{y}_k, & t \neq k \end{cases}$$

Summation of all k,

$t=k \Rightarrow \hat{y}_k y_k - y_k$

$$A \times B = \{ \hat{y}_t y_t - y_t \} + \sum_{t \neq k} y_t \hat{y}_k$$

$$= -y_k + \hat{y}_k \left[y_k + \sum_{t \neq k} y_t \right]$$

$$= -y_k + \hat{y}_k \sum_{t=1}^{K} y_k \qquad y_k = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ one-hot vector}$$

$$= \hat{y}_k - y_k$$

$$C: \frac{\partial z_t}{\partial V} = h_t$$

$$\frac{\partial L}{\partial V} = \underbrace{\sum_{t=1}^{T} (\hat{y}_t - y_t)}_{\text{vector}} \otimes \underbrace{h_t}_{\text{vector}} \longrightarrow \text{matrix}$$

outer product

② $\frac{\partial L}{\partial W}$

$$\begin{aligned} h_t &= \tanh(W h_{t-1} + U x_t + b) \\ \hat{y}_t &= g(V h_t) \\ L_t &= -y_t \log \hat{y}_t \\ L &= \sum_{t=1}^{T} L_t \end{aligned}$$

Chain rule:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W}$$

$$= \sum_{t=1}^{T} \underbrace{\frac{\partial L_t}{\partial \hat{y}_t}}_{A} \cdot \underbrace{\frac{\partial \hat{y}_t}{\partial h_t}}_{B} \cdot \underbrace{\frac{\partial h_t}{\partial W}}_{C}$$

$$A: \frac{\partial L_t}{\partial \hat{y}_t} = -\frac{y_t}{\hat{y}_t}$$

$\partial \hat{y}_t \qquad \hat{y}_t$

$$A \times B: \quad \frac{\partial \hat{y}_t}{\partial h_t} = (\hat{y}_t - y_t) V$$

$$C: \quad \frac{\partial h_t}{\partial w} = \frac{\partial \tanh(\overbrace{W h_{t-1} + U x_t + b}^{z_t})}{\partial w}$$

$$= (1 - \tanh^2(z_t)) \cdot \left(h_{t-1} + W \cdot \frac{\partial h_{t-1}}{\partial w}\right)$$

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

$$\frac{\partial \tanh(x)}{\partial x} = 1 - \tanh^2(x)$$

$$h_{t-1} = \tanh(W h_{t-2} + U x_{t-1} + b)$$

(h_{t-1} is a function of W)

recursion

$$\frac{\partial h_{t-1}}{\partial w} = (1 - \tanh^2(z_{t-1})) \cdot \left(h_{t-2} + w \cdot \frac{\partial h_{t-2}}{\partial w}\right)$$

③ $\frac{\partial L}{\partial U}$

$$h_t = \tanh(W h_{t-1} + U x_t + b)$$

Chain rule:

$$\frac{\partial L}{\partial U} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial U} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \frac{\partial h_t}{\partial U}$$

$$= \sum_{t=1}^{T} (\hat{y}_t - y_t) V \cdot \frac{\partial h_t}{\partial U}$$

$$\frac{\partial h_t}{\partial U} = (1 - \tanh^2(z_t)) \left(x_t + \frac{\partial (W h_{t-1})}{\partial U}\right)$$

$$\frac{\partial (W h_{t-1})}{\partial U} = W \cdot \frac{\partial h_{t-1}}{\partial U} + \boxed{h_{t-1} \cdot \frac{\partial W}{\partial U}} = 0$$

$$= W \cdot \frac{\partial (\tanh(W h_{t-2} + U \cdot x_{t-1} + b))}{\partial U}$$

(h_{t-1} is a function of U)

$$\frac{\partial h_t}{\partial U} = (1 - \tanh^2(z_t)) \left(x_t + W \cdot \frac{\partial h_{t-1}}{\partial U}\right)$$

recursion