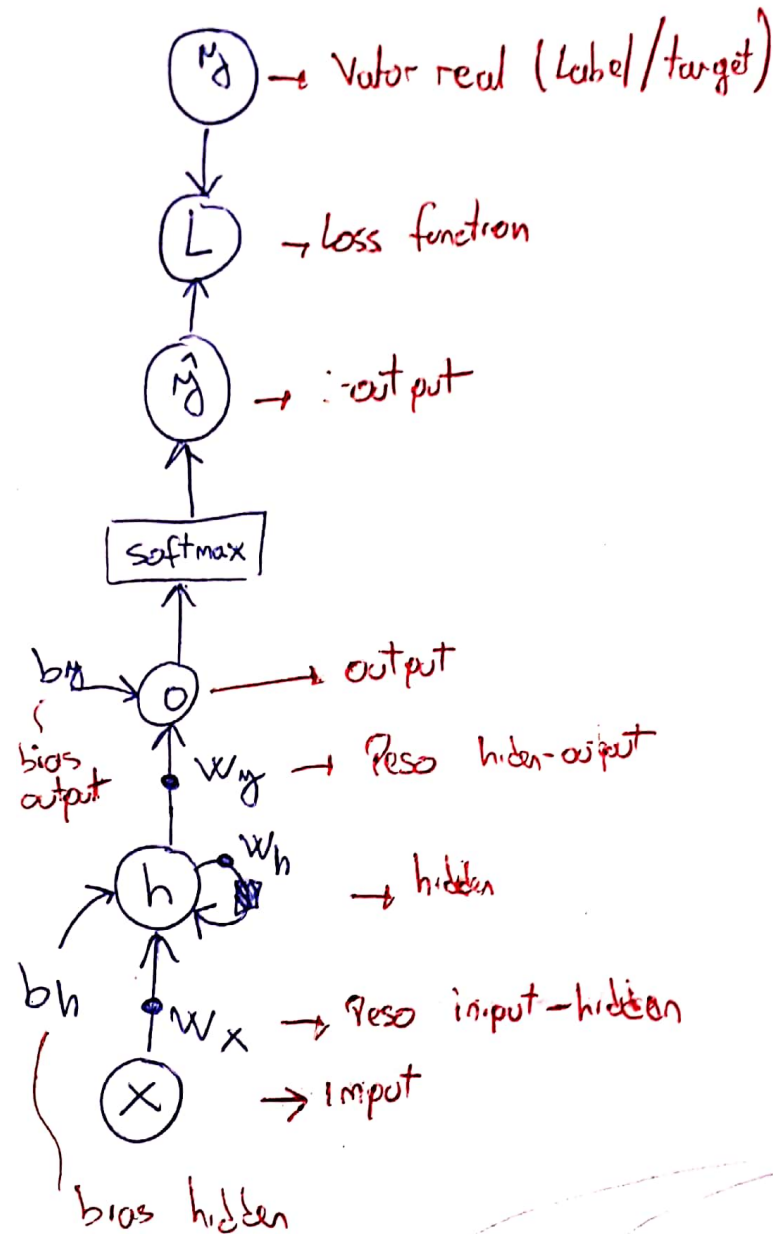


Back propagation Through Time (BPTT)

①



Ecuaciones

$$L_T = \sum_{t=0}^T L_t \quad (1)$$

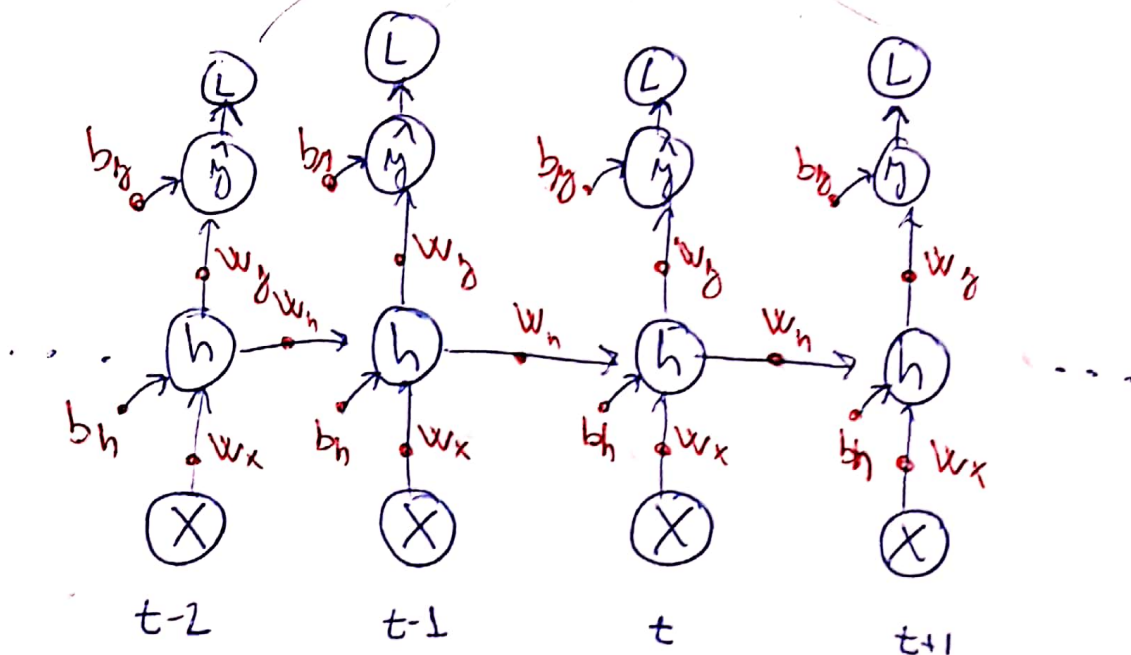
$$L_t = -y_t \log \hat{y}_t \quad (2)$$

$$\hat{y}_t = \text{softmax}(o_t) = \quad (3)$$

$$o_t = h_t \cdot w_y + b_y \quad (4)$$

$$h_t = \text{Tanh} [X_t w_x + h_{t-1} w_h + b_h] \quad (5)$$

$$L_T = \sum_{t=0}^T L_t$$



Parámetros a optimizar.

(II)

- 1º $W_y \rightarrow$ Peso hidden - output
- 2º $b_y \rightarrow$ bias output
- 3º $W_h \rightarrow$ Peso hidden - hidden
- 4º $W_x \rightarrow$ Peso input - hidden
- 5º $b_h \rightarrow$ bias hidden

¿Como varia mi L_T respecto de ellos?

$$\frac{\partial L_T}{\partial W_y} = \frac{\partial}{\partial W_y} \sum_{t=0}^T L_t = \sum_{t=0}^T \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial W_y}$$

Los pesos no varían con los Δt

sumo todos los t para tener una L global

derivada de Loss function respecto de una softmax

Ya fue demostrado $\left(\hat{y}_t - y_t \right)$

de ecuacion (4) en hoja (I)

$$\frac{\partial o_t}{\partial W_y} = h_t$$

entonces llegamos a que

$$\frac{\partial L_t}{\partial W_y} = \left(\hat{y}_t - y_t \right) h_t \Rightarrow$$

$$\boxed{\frac{\partial L_T}{\partial W_y} = \sum_{t=0}^T \left(\hat{y}_t - y_t \right) h_t} \quad (6)$$

2º $\frac{\partial L}{\partial b_y}$ se calcula igual q 1º reemplazando la última $\frac{\partial o_t}{\partial w_y}$ por $\frac{\partial o_t}{\partial b_y}$

$$\boxed{\frac{\partial L_T}{\partial b_y} = \sum_{t=0}^T (\hat{y}_t - y_t) \cdot 1} \quad \text{⊕}$$

$\rightarrow \frac{\partial o_t}{\partial b_y}$

3º $\frac{\partial L_T}{\partial w_h}$ ¿Cómo varia L cuando vario w_h ?

ojo! $h_t = \text{Tanh} \begin{pmatrix} x w_x \\ h w_h \\ b_h \end{pmatrix}$

$$\frac{\partial L_{t+1}}{\partial w_h} = \underbrace{\frac{\partial L_{t+1}}{\partial \hat{y}_{t+1}} \cdot \frac{\partial \hat{y}_{t+1}}{\partial o_{t+1}} \cdot \frac{\partial o_{t+1}}{\partial h_{t+1}}}_{\text{ya lo vieron}} \cdot \underbrace{\frac{\partial h_{t+1}}{\partial w_h}}_{\text{a resolver...}} \quad \text{⓪}$$

ya lo vieron $\leftarrow (\hat{y}_{t+1} - y_{t+1})$

de ecuación ④
en hoja ①
= w_y

$\frac{\partial h_{t+1}}{\partial w_h}$

$$h_{t+1} = \text{Tanh}(X_L w_x + h_t w_h + b_h)$$

h_{t+1} es función de w_h y h_t
q es función de w_h y h_{t-1}
...

$$\frac{\partial h_{t+1}}{\partial w_h} = \frac{\partial h_{t+1}}{\partial h_t} \frac{\partial h_t}{\partial w_h} \Rightarrow \text{Podemos escribir:}$$

$$\frac{\partial h_{t+1}}{\partial w_h} = \sum_{k=0}^{t+1} \frac{\partial h_{t+1}}{\partial h_k} \frac{\partial h_k}{\partial w_h} \rightarrow \text{veamos que da eso}$$

$$\frac{\partial h_3}{\partial w_h} = \sum_{k=1}^3 \frac{\partial h_3}{\partial h_k} \frac{\partial h_k}{\partial w_h}$$

$$= \underbrace{\frac{\partial h_3}{\partial h_1} \frac{\partial h_1}{\partial w_h}} + \underbrace{\frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial w_h}} + \underbrace{\frac{\partial h_3}{\partial h_3} \frac{\partial h_3}{\partial w_h}}$$

h_3 varia porque
 h_1 Varia por que
 w_h Varia

h_3 Varia xq
 h_2 Varia xq
 w_h Varia

h_3 Varia xq
 w_h Varia

Rescapitulamos ecuacion (8) de hoja (III)

$$\frac{\partial L_{t+1}}{\partial w_h} = \frac{\partial L_{t+1}}{\partial \hat{y}_{t+1}} \cdot \frac{\partial \hat{y}_{t+1}}{\partial o_{t+1}} \cdot \frac{\partial o_{t+1}}{\partial h_{t+1}} \cdot \underbrace{\sum_{k=0}^{t+1} \frac{\partial h_{t+1}}{\partial h_k} \cdot \frac{\partial h_k}{\partial w_h}}_{\frac{\partial h_{t+1}}{\partial w_h}}$$

Esto es otra regla de la cadena mas!!

$$\frac{\partial h_{t+1}}{\partial h_k} = \prod_{j=k}^t \frac{\partial h_{j+1}}{\partial h_j} \rightarrow \underbrace{\frac{\partial h_{k+1}}{\partial h_k}}_{\text{algunos términos}} \cdot \dots \cdot \underbrace{\frac{\partial h_{t+1}}{\partial h_t}}_{\text{el último}}$$

Ahora si unamos todo

$$\frac{\partial L_{t+1}}{\partial w_h} = \frac{\partial L_{t+1}}{\partial \hat{y}_{t+1}} \cdot \frac{\partial \hat{y}_{t+1}}{\partial o_{t+1}} \cdot \frac{\partial o_{t+1}}{\partial h_{t+1}} \cdot \sum_{k=1}^{t+1} \left[\prod_{j=k}^t \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \cdot \frac{\partial h_k}{\partial w_h} \right]$$

obl. es para 1 solo tiempo.

$$\frac{\partial L_T}{\partial W_h} = \sum_t^T \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial h_t} \cdot \sum_{k=1}^t \left[\prod_{i=k}^t \left(\frac{\partial h_i}{\partial h_{i-1}} \right) \cdot \frac{\partial h_k}{\partial W_h} \right] \quad \textcircled{V}$$

9

4º Para $\frac{\partial L_T}{\partial W_x}$ se computa igual que 3º ($\frac{\partial L_T}{\partial W_h}$)

Ya que cada hidden state h_t es función de W_x y h_{t-1}

$$\frac{\partial L_T}{\partial W_x} = \sum_t^T \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial h_t} \cdot \sum_{k=1}^t \left[\prod_{i=k}^t \left(\frac{\partial h_i}{\partial h_{i-1}} \right) \cdot \frac{\partial h_k}{\partial W_x} \right]$$

↑
de acá cambia

5º Igual procedimiento

$$\frac{\partial L_T}{\partial b_h} = \sum_t^T \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial h_t} \cdot \sum_{k=1}^t \left[\prod_{i=k}^t \left(\frac{\partial h_i}{\partial h_{i-1}} \right) \cdot \frac{\partial h_k}{\partial b_h} \right]$$

↓
= 1 de ecuación
5º hacia I