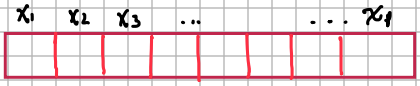
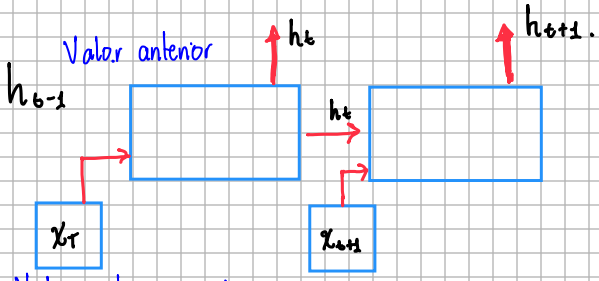
Redes recurrentes:





Definimos la función de pérdida:

$$\sum_t (y_t - h_t)^2 = \text{Loss}$$

Red recurrente de imagénes (Por ejemplo en videos) ¿ Sirve para simulaciones físicas?

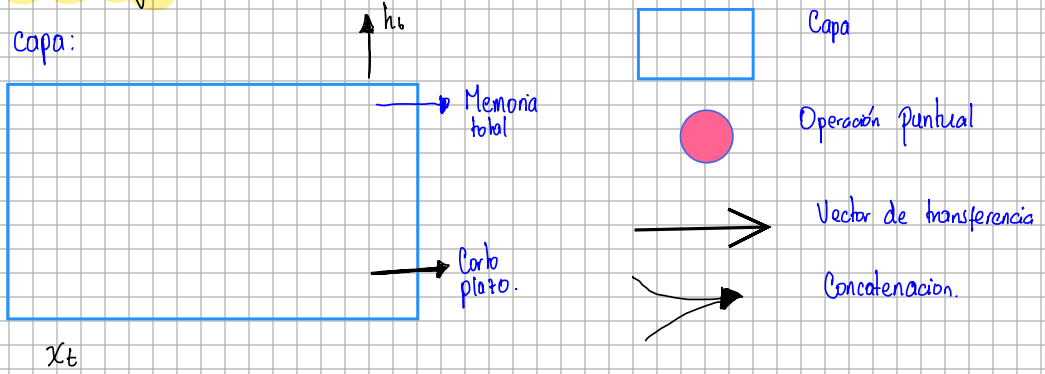
- Reconocimiento de voz
- Traducción

➡ No es algo nuevo, ideas de los años 80.

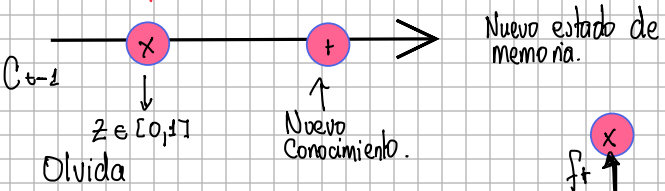
Al ir más lejos el gradiente sufrra problemas: Solución a este problema da como resultado por ejemplo redes LSTM. (1997)

Long short term memory.

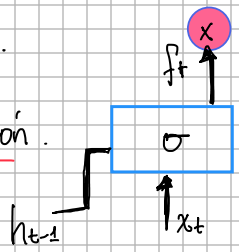
Dentro de una capa:



Banda transportadora



Compuerta: Función de activación.



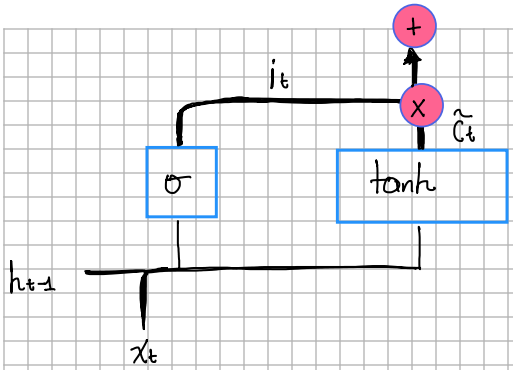
Concatenación

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Puerta de olvido

Valores entre 0 y 1 en la puerta de olvido que se multiplica con la memoria de largo plazo

Nuevo Conocimiento:



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$\tilde{C}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right)$$

Para la salida $O_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$

$h_t = O_t * \tanh(C_t)$ Se combina el preoutput y la memoria de largo plazo C_t.

Estructura Matemática de una red LSTM

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

Donde $\odot$ es el producto Hadamard.

Por otro lado, h_t se actualiza después de que C_t se haga actualizado como:

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o]$$
$$h_t = O_t \odot \tanh(C_t)$$

Importante:

Nuestra función de pérdida: $\mathcal{L} = \sum (x_t - h_t)^2 \cdot \frac{1}{N}$

Redes: Gated Recurrent Unit GRU.

Dada una secuencia $\vec{x} = (x_1, x_2, \ldots x_T)$, la RNR actualiza su estado oculto recurrente h_t a través

$$h_t = \begin{cases} 0 & \text{si } t=0 \\ \phi(h_{t-1}, x_t) & \text{otherwise.} \end{cases}$$

- Si g es una función de activación suave (Sigmoide, tanhϕ)

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

- Una RNR generativa genera una distribución de probabilidad sobre el siguiente elemento de la de la secuencia, dado su estado actual h_t,

La secuencia de probabilidad podemos descomponerla en:

$$p(x_1, \ldots, x_T) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \ldots P_T(x_T \mid x_1, x_2, \ldots, x_{T-1})$$

Modelamos cada probabilidad condicional

$$p(x_t \mid x_1, \ldots, x_{t-1}) = g(h_t)$$

→ Problemas con el gradiente.*

$$z_t = \sigma(W_z x_t + U_z h_{t+1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$
$$h'_t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b)$$

Actualización del estado recurrente:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h'_t$$