Aprendizaje profundo

Redes recurrentes

Gibran Fuentes-Pineda

Octubre 2023

Arquitecturas de redes recurrentes

- Contienen celdas recurrentes en conjunto con otras capas
- La salida de una celda alimenta otras capas u otras celdas
- Por ejemplo, para predecir el siguiente símbolo en un texto con una celda recurrente básica, a la salida podemos agregar una capa densa con función de activación *softmax*

$$\hat{\mathbf{y}}^{[t+1]} = softmax\left(\mathbf{W}_y \cdot \mathbf{h}^{[t+1]} + \mathbf{b}_y\right)$$

Arquitecturas de redes recurrentes: ejemplo

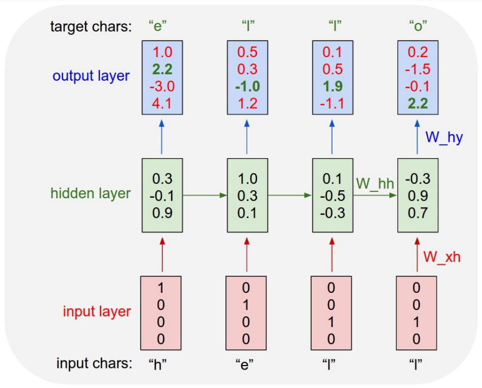


Imagen tomada de Karpathy 2015 (`http://karpathy.github.io/2015/05/21/rnn-effectiveness/`)

Arquitecturas de redes recurrentes: tareas de uno a uno

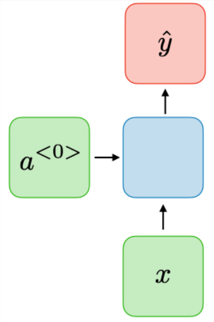


Imagen tomada de Amidi. Recurrent Neural Networks cheatsheet

Arquitecturas de redes recurrentes: tareas de uno a muchos

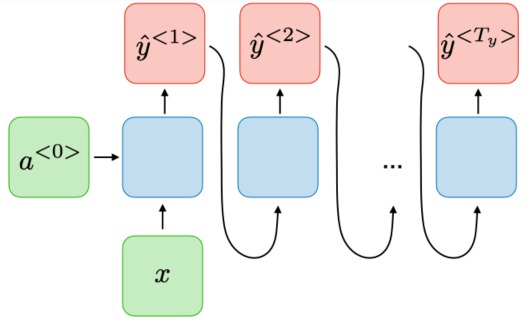


Imagen tomada de Amidi. Recurrent Neural Networks cheatsheet

Arquitecturas de redes recurrentes: tareas de muchos a uno

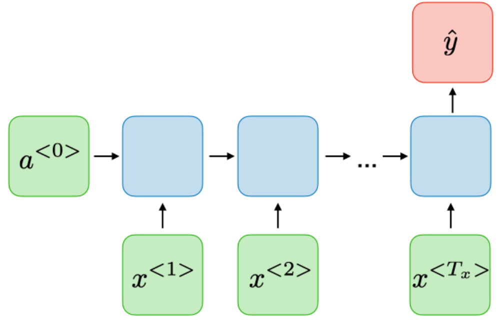


Imagen tomada de Amidi. Recurrent Neural Networks cheatsheet

Arquitecturas de redes recurrentes: tareas de muchos a muchos

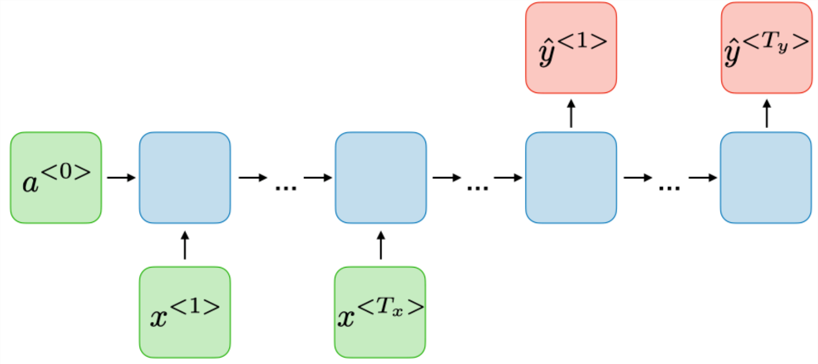


Imagen tomada de Amidi. Recurrent Neural Networks cheatsheet

Arquitecturas de redes recurrentes: LSTM/GRU bidireccional

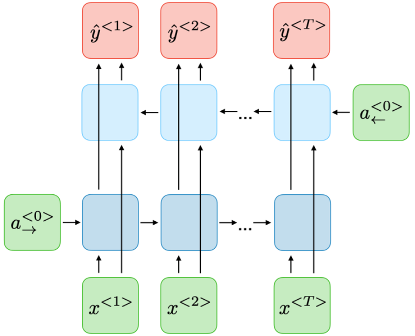


Imagen tomada de Amidi. Recurrent Neural Networks cheatsheet

Arquitecturas de redes recurrentes: celdas apiladas

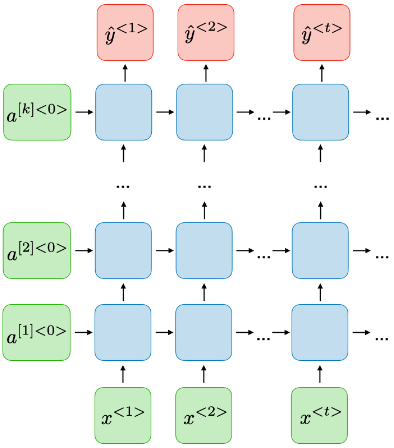


Imagen tomada de Amidi. Recurrent Neural Networks cheatsheet

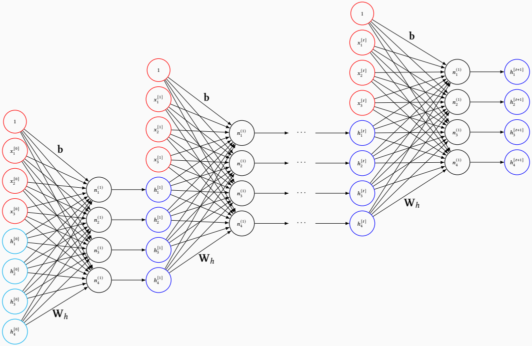
Retropropagación en el tiempo

- Pérdida en el tiempo

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{L}\left(\hat{\mathbf{y}}^{[t]}, \mathbf{y}^{[t]}\right)$$

- Retropropagación

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \mathrm{L}\left(\hat{\mathbf{y}}^{[t]}, \mathbf{y}^{[t]}\right)}{\partial \boldsymbol{\theta}}$$



Retropropagación en el tiempo para una celda básica (1)

- Para la matriz de pesos W_y y un tiempo T

$$\frac{\partial \mathcal{L}}{\partial \mathsf{W}_y} = \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}}{\partial \hat{\mathsf{y}}^{[t]}} \cdot \frac{\partial \hat{\mathsf{y}}^{[t]}}{\partial \mathsf{W}_y} \right] = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\hat{\mathsf{y}}^{[t]}} \mathsf{h}^{[t]\top}$$

$$\frac{\partial \mathrm{L}}{\partial \hat{\mathsf{y}}^{[t]}} = \frac{1}{T} \cdot \frac{\partial \mathcal{L}\left(\hat{\mathsf{y}}^{[t]}, \mathsf{y}^{[t]}\right)}{\partial \hat{\mathsf{y}}^{[t]}}$$

- Para el tiempo T

$$\frac{\partial \mathcal{L}}{\partial \mathsf{h}^{[T]}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathsf{y}}^{[T]}} \cdot \frac{\partial \hat{\mathsf{y}}^{[T]}}{\partial \mathsf{h}^{[t]}} = \mathsf{W}_y^{\top} \frac{\partial \mathcal{L}}{\partial \hat{\mathsf{y}}^{[T]}}$$

Retropropagación en el tiempo para una celda básica (2)

- Para los tiempos $t = T - 1, \ldots, 1$, la pérdida se ve afectada por $\mathsf{h}^{[t]}$ a través de $\mathsf{h}^{[t+1]}$ y $\mathsf{h}^{[t]}$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathsf{h}^{[t]}} &= \left[\frac{\partial \mathcal{L}}{\partial \mathsf{h}^{[t+1]}} \cdot \frac{\partial \mathsf{h}^{[t+1]}}{\partial \mathsf{h}^{[t]}}\right] + \left[\frac{\partial \mathcal{L}}{\partial \hat{\mathsf{y}}^{[t]}} \cdot \frac{\partial \hat{\mathsf{y}}^{[t]}}{\partial \mathsf{h}^{[t]}}\right] \\ &= \left[\mathsf{W}_{hh}^{\top} \frac{\partial \mathcal{L}}{\partial \mathsf{h}^{[t+1]}}\right] + \left[\mathsf{W}_{y}^{\top} \frac{\partial \mathcal{L}}{\partial \hat{\mathsf{y}}^{[t]}}\right]\end{aligned}$$

Retropropagación en el tiempo para una celda básica (3)

- La pérdida depende de W_{hx} y W_{hh} por $\mathrm{h}^{[1]}, \mathrm{h}^{[2]}, \ldots, \mathrm{h}^{[T]}$

- Para la matriz de pesos W_{hx}

$$\frac{\partial \mathcal{L}}{\partial \mathrm{W}_{hx}} = \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}}{\partial \mathrm{h}^{[t]}} \cdot \frac{\partial \mathrm{h}^{[t]}}{\partial \mathrm{W}_{hx}} \right] = \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}}{\mathrm{h}^{[t]}} \cdot \mathrm{x}^{[t]\top} \right]$$

- Para la matriz de pesos W_{hh}

$$\frac{\partial \mathcal{L}}{\partial \mathrm{W}_{hh}} = \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}}{\partial \mathrm{h}^{[t]}} \cdot \frac{\partial \mathrm{h}^{[t]}}{\partial \mathrm{W}_{hh}} \right] = \sum_{t=1}^{T} \left[\frac{\partial \mathcal{L}}{\mathrm{h}^{[t]}} \cdot \mathrm{h}^{[t-1]\top} \right]$$