Redes Neuronales Recurrentes

Modelado de secuencias profundas



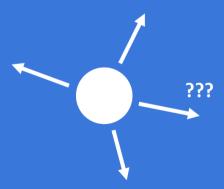
Marco Teran



Octubre 2020- Bogotá

Dada la imagen de una pelota, ¿puedes predecir dónde irá a continuación?

Dada la imagen de una pelota, ¿puedes predecir dónde irá a continuación?



Dada la imagen de una pelota, ¿puedes predecir dónde irá a continuación?



Secuencias en la naturaleza



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 6 / 69

Secuencias en la naturaleza

character:

Introducción al aprendizaje profundo

word:

Texto

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 7 / 69

Un problema de modelado de secuencias:

Predecir la siguiente palabra

Un problema de modelado de secuencias: predecir la siguiente palabra

"Esta mañana yo saqué a mi gato para un paseo."

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 9 / 69

Un problema de modelado de secuencias: predecir la siguiente palabra

"Esta mañana yo saqué a mi gato para un paseo." dada estas palabras

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 10 / 69

Un problema de modelado de secuencias: predecir la siguiente palabra

"Esta mañana yo saqué a mi gato para un paseo."

dada estas palabras predecir la

siguiente palabra

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 11/69

Idea 1: usar una ventana fija

"Esta mañana yo saqué a mi gato para un paseo."

dada estas palabras predecir la
siguiente palabra

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 12 / 69

Idea 1: usar una ventana fija

```
"Esta mañana yo saqué a mi gato para un paseo."

dada estas palabras predecir la
siguiente palabra
```

La codificación de características one-hot: nos dice qué es cada palabra



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 13 / 69

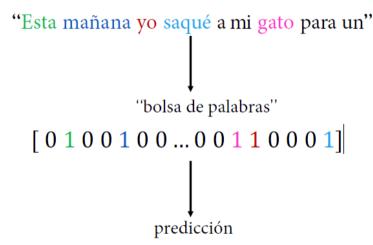
Problema 1: no se pueden modelar las dependencias a largo plazo

"Colombia es donde crecí, pero ahora vivo en Chicago. Yo hablo con fluidez ____."

Necesitamos información del **pasado distante** (contexto) para poder predecir la palabra correcta.

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 14 / 69

Idea 2: usar la secuencia completa como un conjunto de conteos



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 15 / 69

Problema 2: los recuentos no preservan el orden

La comida estaba buena, nada mal.

VS.

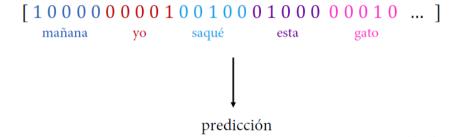
La comida estaba mala, nada buena.

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 16 / 69

Idea 3: usar una ventana fija realmente grande

"Esta mañana yo saqué a mi gato para un paseo."

dada estas palabras predecir la
siguiente palabra



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 17 / 69

Problema 3: no se comparten los parámetros

Cada una de estas entradas tiene un parámetro separado:

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 18 / 69

Problema 3: no se comparten los parámetros

Cada una de estas entradas tiene un parámetro separado:

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 19 / 69

Problema 3: no se comparten los parámetros

```
[ 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 ... ]

esta mañana tomé el gato
```

Cada una de estas entradas tiene un parámetro separado:

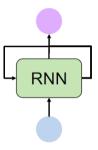
Las cosas que aprendemos sobre la secuencia **no se transfieren** si aparecen en **cualquier parte** de la secuencia

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 20 / 69

Modelado de secuencias: criterios de diseño

Para modelar las secuencias, tenemos que:

- Manejar secuencias de longitud variable
- Seguimiento de las dependencias a largo plazo
- Mantener la información sobre el orden
- Compartir los parámetros a través de la secuencia



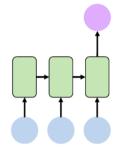
Hoy: Las redes neuronales recurrentes (RNN) como un enfoque para los problemas de modelado de secuencias

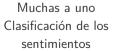
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 21 / 69

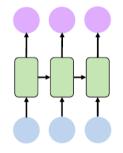
Recurrent Neural Networks (RNNs)

Red neuronal de alimentación estándar







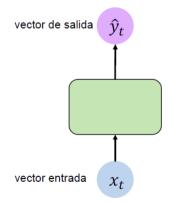


De muchas a muchas Generación de la música

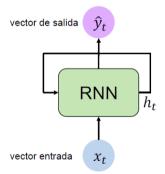
... y muchas otras arquitecturas y aplicaciones

Una a una

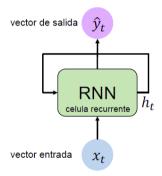
Red neuronal estándar



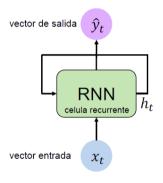
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 24 / 69



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 25 / 69

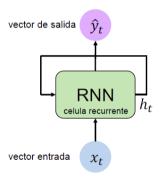


Marco Teran Octubre 2020 Redes Neuronales Recurrentes 26 / 69

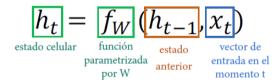


Aplicar una relación de recurrencia en cada paso de tiempo para procesar una secuencia:

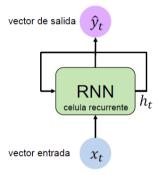
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 27 / 69



Aplicar una relación de recurrencia en cada paso de tiempo para procesar una secuencia:



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 28 / 69



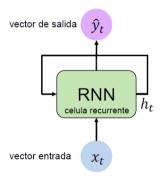
Aplicar una relación de recurrencia en cada paso de tiempo para procesar una secuencia:



Nota: se utilizan la misma función y el mismo conjunto de parámetros en cada paso de tiempo

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 29 / 69

Actualización y salida del estado de una RNN



Vector de salida

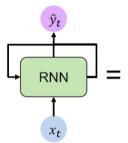
$$\hat{y}_t = \mathbf{W}_{hy} h_t$$

Actualización del estado oculto

$$h_t = \tanh(\mathbf{W}_{hh}h_{t-1} + \mathbf{W}_{xh}x_t)$$

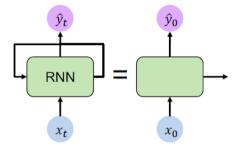
Vector de entrada

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 30 / 69

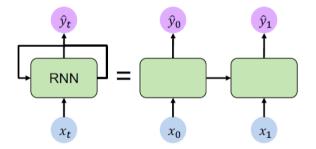


Representar como grafo computacional desenrollado a través del tiempo

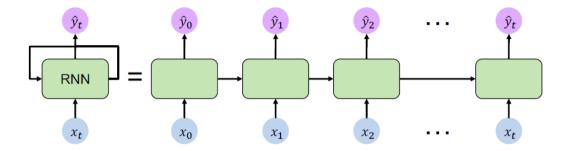
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 31 / 69



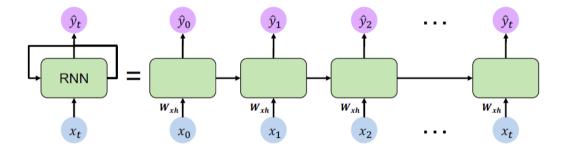
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 32 / 69



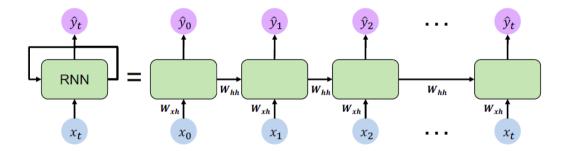
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 33 / 69



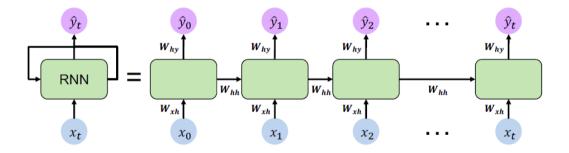
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 34 / 69



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 35 / 69



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 36 / 69,

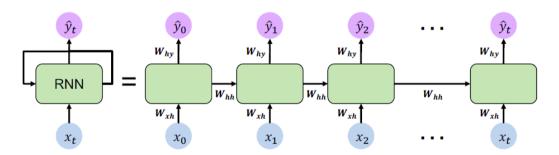


Marco Teran Octubre 2020 Redes Neuronales Recurrentes 37 / 69



RNNs: grafo computacional a través del tiempo

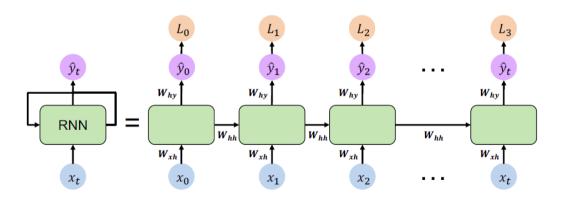
Reutilizar las mismas matrices de peso en cada paso de tiempo



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 38 / 69

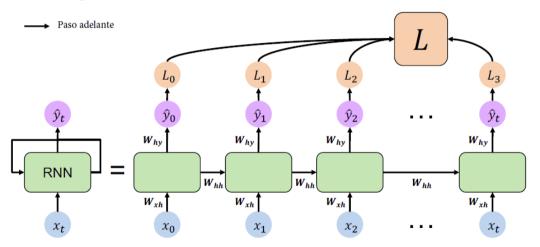
RNNs: grafo computacional a través del tiempo

→ Forward pass



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 39 / 69

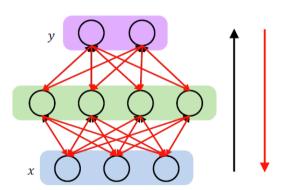
RNNs: grafo computacional a través del tiempo



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 40 / 69

La retropropagación a través del tiempo (BPTT)

Recordatorio: la retropagación en los modelos de avance

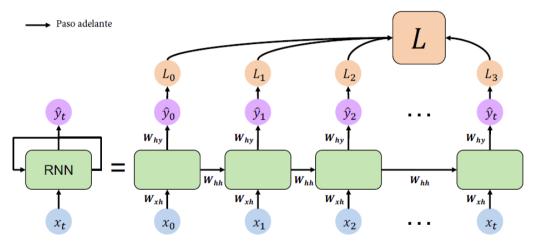


Algoritmo de retropropagación:

- 1 Tomar la derivada (gradiente) de la pérdida con respecto a cada parámetro
- Cambiar los parámetros para minimizar la pérdida

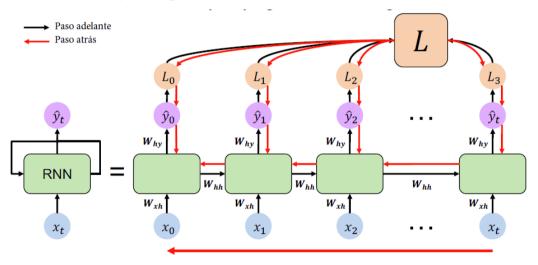
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 42 / 69

RNNs: retropropagación a través del tiempo



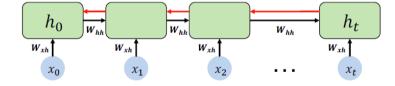
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 43 / 69

RNNs: retropropagación a través del tiempo



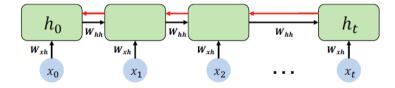
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 44 / 69

Flujo estándar del gradiente en RNN



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 45 / 69

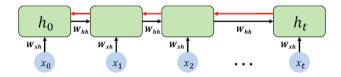
Flujo estándar del gradiente en RNN



Computar el gradiente respecto a h_0 implica **muchos factores** de W_{hh} (y f^\prime repetida muchas veces)

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 46 / 69

Flujo estándar del gradiente en RNN: gradientes explosivos

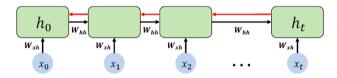


Computar el gradiente respecto a h_0 implica **muchos factores** de W_{hh} (y f^\prime repetida muchas veces)

Muchos valores > 1:
 gradientes explosivos
Recorte de gradientes para escalar
 grandes gradientes

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 47 / 69

Flujo estándar del gradiente en RNN: gradientes desvanecidos



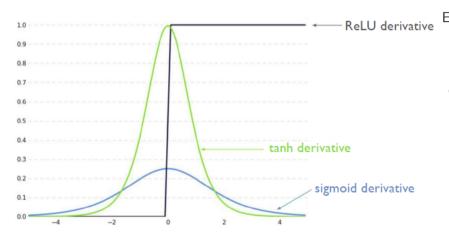
Computar el gradiente respecto a h_0 implica **muchos factores** de W_{hh} (y f^\prime repetida muchas veces)

Muchos valores > 1:
 gradientes explosivos
Recorte de gradientes para escalar
 grandes gradientes

Muchos valores > I: gradientes desvanecidos

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 48 / 69

Truco 1: funciones de activación



El uso de ReLU evita que f' reduzca los gradientes cuando x > 0

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 49 / 69

Truco 2: Inicialización de parámetros

Inicializar los pesos a la matriz de identidad

Iniciar los sesgos a cero

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Esto ayuda a evitar que los pesos se reduzcan a cero.

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 50 / 69

Solución 3: células cerradas (gated cells)

Idea: usar una unidad **recurrente más compleja con puertas** para controlar la información que pasa por ella

gated cell
LSTM, GRU, etc.

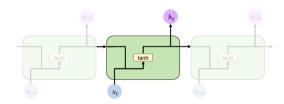
Las redes de memoria de largo y corto plazo (Long Short Term Memory, LSTM) se basan en una célula cerrada para rastrear la información a través de muchos pasos de tiempo.

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 51 / 69

Redes de memoria a largo y corto plazo (LSTM)

RNN estándar

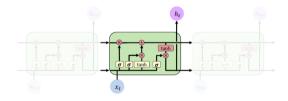
En una RNN estándar, los módulos de repetición contienen un simple nodo de cálculo



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 53 / 69



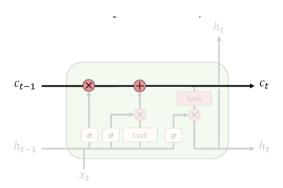
Los módulos de repetición del LSTM contienen capas interactivas que controlan el flujo de información



Las células LSTM son capaces de rastrear la información a través de muchos pasos de tiempo

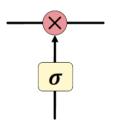
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 54 / 69

Los LSTM mantienen un estado celular c_t donde es fácil que la información fluya



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 55 / 69

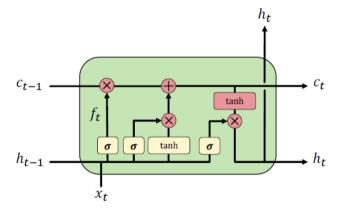
La información se **añade** o se **elimina** al estado celular a través de estructuras llamadas **gates** (puertas)



Las puertas permiten opcionalmente el paso de la información, a través de una capa de red neural **sigmoide** y la multiplicación punto a punto

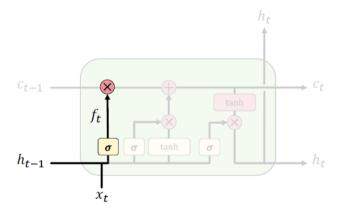
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 56 / 69

¿Cómo funcionan las LSTM?



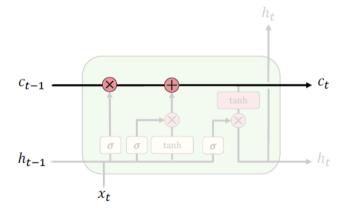
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 57 / 69

Las LSTMs olvidan información irrelevante del estado previo



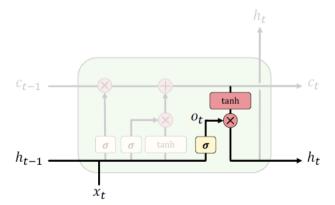
Marco Teran Octubre 2020 Redes Neuronales Recurrentes 58 / 69

Las LSTMs actualizan selectivamente los valores del estado de las células



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 59 / 69

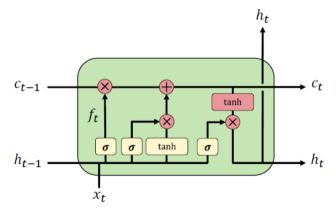
Las LSTMs usan una **puerta de salida** para dar salida a cierta información del estado de la células



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 60 / 69

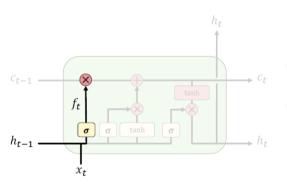
¿Cómo funcionan las LSTM?

1) Olvidar 2) Actualizar 3) Salida



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 61 / 69

LSTMs: olvida la información irrelevante



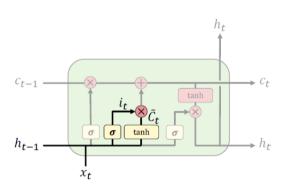
$$f_t = \sigma\left(\mathbf{W}_i[h_{t-1}, x_t] + b_f\right)$$

- Usar la salida v entrada de la célula anterior
- Sigmoide: valor 0 y 1 "olvidar completamente" vs. "mantener completamente"

ej: Olvida el pronombre de género del sujeto anterior en la oración.

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 62 / 69

LSTMs: identificar la nueva información que se va a almacenar



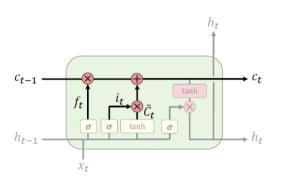
$$\begin{split} i_t &= \sigma\left(\mathbf{W}_i[h_{t-1}, x_t] + b_f\right) \\ \tilde{C}_t &= \tanh\left(\mathbf{W}_c[h_{t-1}, x_t] + b_c\right) \end{split}$$

- Capa sigmoide: decidir qué valores actualizar
- Capa de Tanh: generar un nuevo vector de "valores candidatos" que podría añadirse al estado

ej: Agregar el género del nuevo sujeto para reemplazar el del antiguo sujeto

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 63 / 69

LSTMs: actualizar el estado de las células



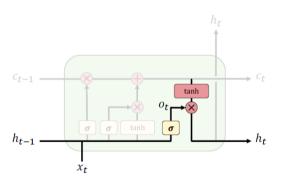
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

- Aplicar la operación "forget" al estado previo de la célula interna: $f_t \cdot C_{t-1}$
- Agregar nuevos valores candidatos, escalados según lo que decidimos actualizar: $i_t \cdot \tilde{C}_t$

ej: En realidad, dejar la información antigua y añadir nueva información sobre el género del sujeto.

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 64 / 69

LSTMs: versión filtrada de salida del estado de las células



$$\begin{aligned} o_t &= \sigma\left(\mathbf{W}_o[h_{t-1}, x_t] + b_o\right) \\ h_t &= o_t \cdot \tanh(C_t) \end{aligned}$$

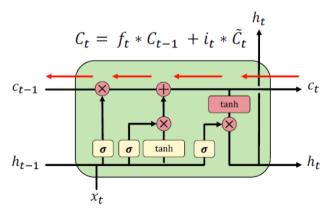
- Capa sigmoidea: decidir qué partes del estado se deben producir
- Capa de Tanh: valores de calabaza entre -1 y 1
- $o_t \cdot anh(C_t)$: versión filtrada de salida del estado de la célula

ej: Habiendo visto un sujeto, puede producir información relativa a un verbo.

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 65 / 69

Flujo de gradientes de las LSTM

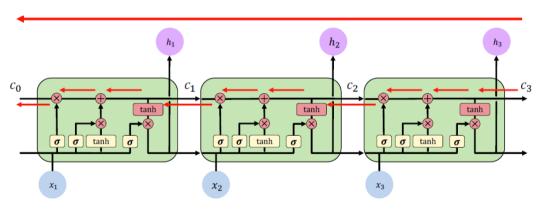
La retropropagación de C_t a C_{t-1} sólo requiere una multiplicación entre elementos No hay multiplicación de la matriz \to evitar el problema de los gradientes desvanecidos



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 66 / 69

Flujo de gradientes de las LSTM

¡Flujo de gradiente ininterrumpido!



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 67 / 69

LSTMs: conceptos clave

- Mantener un estado celular separado de lo que se emite
- Usar puertas para controlar el flujo de información
 - Olvida que la puerta se deshace de la información irrelevante
 - Actualizar selectivamente el estado de las células
 - La puerta de salida devuelve una versión filtrada del estado de la célula
- lacksquare La retropropagación de C_t a C_{t-1} no requiere multiplicación de matrices:
 - flujo de gradiente ininterrumpido

Marco Teran Octubre 2020 Redes Neuronales Recurrentes 68 / 69

Muchas gracias por su atención

; Preguntas?

Contact: Marco Tulio Teran De La Hoz e-mail: marco.teran@usa.edu.co



Marco Teran Octubre 2020 Redes Neuronales Recurrentes 69 / 69

