Modelos recurrentes

Aprendizaje profundo

Alberto Díaz Álvarez
Departamento de Sistemas Informáticos - Universidad Politécnica de Madrid

28 de marzo de 2023

En este tema hablaremos de redes neuronales recurrentes (RNN)

- Permiten tratar con información secuencial, pero no se limitan al análisis
- También pueden generar datos a partir de una o más entradas

Aprenderemos qué posibilidades nos ofrecen estos modelos y sus limitaciones

- Y técnicas para superar (casi mejor "mitigar") estas limitaciones

Las RNN tienen muchas aplicaciones prácticas en el mundo real

- Sobre todo en el procesamiento de lenguaje natural (NLP)
- De hecho ambas áreas (RNN y NLP) han evolucionado a la par en los últimos años

Series temporales

¿Qué es una serie temporal?

Es una **secuencia de observaciones** de una variable medida en el tiempo

- O dicho de otra forma, una **secuencia ordenada de datos**
- Los **datos consecutivos** tienen una **fuerte correlación** entre si
- Estos datos pueden ser tanto **numéricos** como **simbólicos**

Las series temporales se clasifican en:

- **Univariantes**: Secuencias de valores simples
  - P.ej. temperatura, precio de una acción, número de conexiones, etc.
- **Multivariantes**: Secuencias de valores de múltiples variables
  - P.ej. acelerómetros, canciones (sus múltiples pistas por instrumento), etc.
  - Se suelen representar descomponiéndolas en series univariantes

¿Por qué son importantes?

La información suele llegar como secuencia, no como muestras aisladas

- Los datos están intrínsecamente ligados a la noción de tiempo

Los humanos procesamos la información de esta manera $\rightarrow$ **Memoria secuencial**

- Mecanismo del cerebro que facilita el reconocimiento de patrones secuenciales
  i. Recita el alfabeto: ABCDEFGHIJKLMNÑOPQRSTUVWXYZ
     - Es fácil porque lo hemos aprendido así desde pequeños
  ii. Ahora al revés: ZYXWVUTSRQPOÑNMLKJIHGFEDCBA
     - Es muy difícil, a no ser que lo hayamos practicado mucho anteriormente
  iii. Ahora empieza a recitar el alfabeto desde la letra F
     - Al comienzo suele costar un poco (hay que localizar el comienzo del patrón)
     - El resto, una vez reconocido el patrón, sale de forma natural

Problemas que involucran series temporales

- **Predicción**: Predecir el valor futuro de una variable en base a su pasado
  - P.ej. pronóstico del tiempo, predicción de ventas, etc.
- **Detección de anomalías**: Detectar valores atípicos en una serie de eventos
  - P.ej. detección de fraudes, detección de intrusos, etc.
- **Reconomiento de patrones**: Extraer patrones recurrentes de una serie de eventos
  - P.ej. detección de tendencias, detección de patrones de comportamiento, etc.

Una advertencia

Cuidado con la **información futura en las características de entrada**

- Es un problema mucho más común de lo que puede parecer

En resumen, ocurre cuando entrenamos un modelo con datos pasados y datos futuros

- Especialmente cuando vamos a realizar un test del modelo
- Si separamos la serie en entrenamiento y test:
  - No basta con realizar una separación aleatoria de los ejemplos
  - Hay que separar los datos por tiempo: entrenamiento **ANTES** que test
  - Si no, estamos entrenando con información futura (**mal**)
- Insistimos: **Cuidado con añadir información futura en las entradas**

Ventanas móviles (I)

Los problemas de predicción se basan en la idea de predecir el futuro dados:

1. El valor actual de la secuencia
2. El conocimiento almacenado en su estado interno

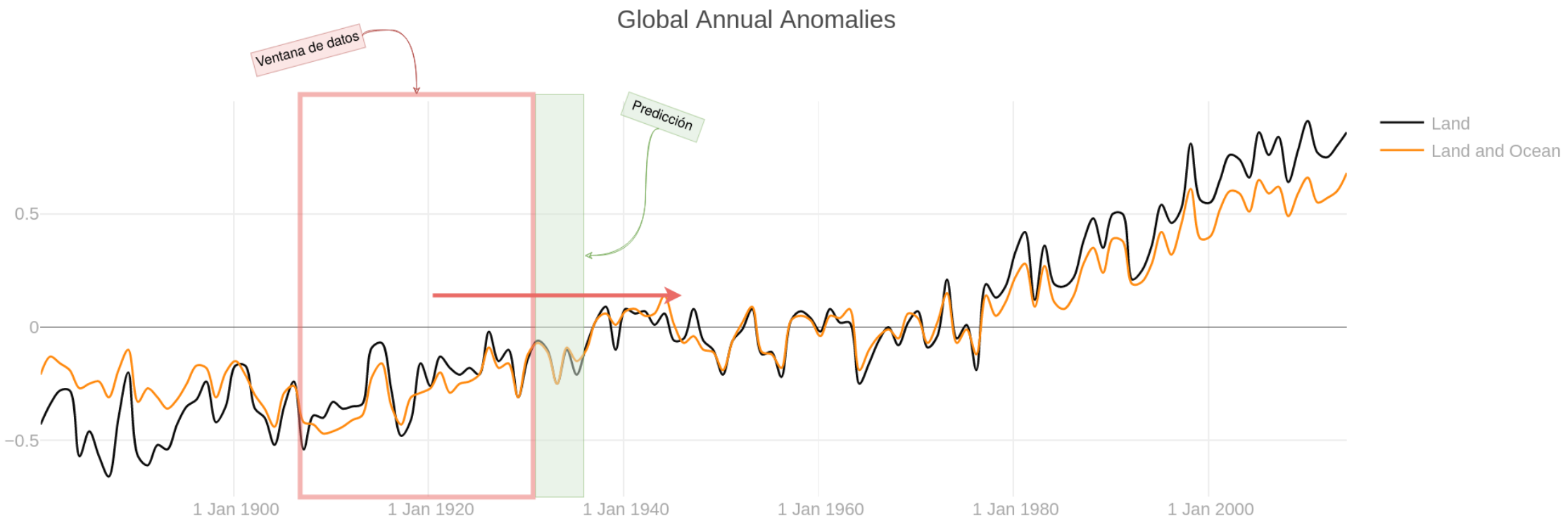
Podemos estimar este conocimiento interno como el histórico de los valores previos

- Y a esto se le conoce como **ventana móvil** (rolling windows)

El procedimiento es el siguiente:

1. Se toma una ventana de tamaño fijo en la secuencia
2. Utilizamos los datos de la ventana para predecir los valores futuros de la secuencia
3. Deslizamos la ventana hacia adelante (p.ej. la mitad del tamaño de la ventana)
4. Volvemos al paso 2.

Ventanas móviles (II)



Ventanas móviles (y III)

Este enfoque se usa en muchas aplicacinoes con éxito

- De hecho es el enfoque más simple y usado en series temporales
- En aprendizaje automático clásico se suele utilizar para mejorar las predicciones
- En el caso concreto del aprendizaje profundo hay dos aproximaciones:

  i. Alimentar la ventana entera a la red neuronal
     - Típico en perceptrones multicapa y también en redes convolucionales

  ii. Alimentar los valores uno a uno e ir manteniendo una memoria
     - ¿Cómo que memoria? Sin estrés, en unos moomentos lo vemos

Redes neuronales recurrentes

Intuición

Miles de millones de neuronas en el cerebro se interconectan sin una dirección única

- Una decisión **ahora** no se basa solo en lo que he percibido **ahora**

Razonamos usando información previa, pero las redes feed-forward:

- No pueden manejar información de entrada secuencial variable
- En la salida solo se puede usar información de la entrada actual
- No pueden memorizar entradas pasadas para predicciones

Entonces, ¿qué pasa si añadimos **retroalimentación** a una **red neuronal**?

- Pues que **tenemos una red neuronal recurrente**, eso es lo que pasa
  - Aprovechan la naturaleza secuencial de los datos para predecir
  - Cada salida depende implícitamente de todas las anteriores
  - Mantienen un conocimiento histórico de las entradas gracias a su memoria interna

¿Qué son las redes neuronales recurrentes?

Son redes neuronales que reutilizan salidas anteriores como parte de su entrada

- Fueron introducidas durante la década de los 80[1]
  - Fueron poco populares en la época por sus requisitos funcionales para entrenar

Su propia salida es parte de la entrada en la siguiente predicción

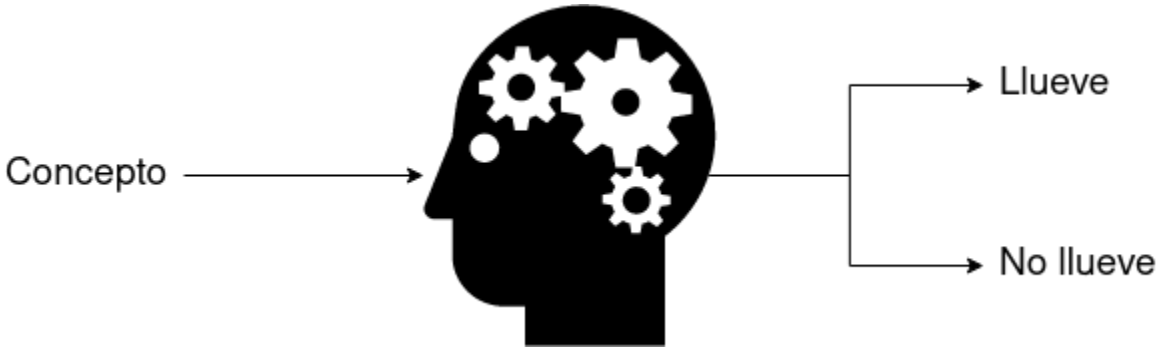
- Esto permite que la red recuerde información de entradas anteriores
- Mantienen un estado interno (memoria) que se actualiza en cada paso de tiempo
- Son muy útiles cuando el contexto es fundamental

[1] Concretamente con los trabajos de las redes de Hopfield (*Neural networks and physical systems with emergent collective computational abilities.*) y de Elman (*Finding structure in time*).

Intuición del funcionamiento de una RNN (I)

Supongamos que queremos predecir si va a llover a partir de observaciones

- Sin entrar en detalles, todo conceptual



La naturaleza secuencia de la red implica varios tipos de problemas

- Puede haber una o varias entradas y una o varias salidas
- Esto lo veremos más adelante explorando los diferentes tipos de problemas en RNN

Intuición del funcionamiento de una RNN (II)

Observación en t: `(niña, jugando, sola)`

- La observación de nuestra red está vacía
- El concepto no es indicativo de si va a llover o no
- Sin embargo, la memoria interna incluye ahora una **representación de la entrada**

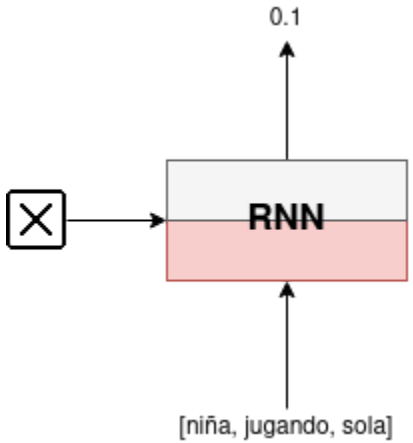


Figura 1. En el primer instante de tiempo sólo contamos con la información que acabamos de observar

Intuición del funcionamiento de una RNN (III)

Observación en $t + 1$: `(pasa, vehículo, rojo)`

- La predicción no cambia en absoluto
- Sin embargo, la memoria va **manteniendo conocimiento** de cada concepto pasados

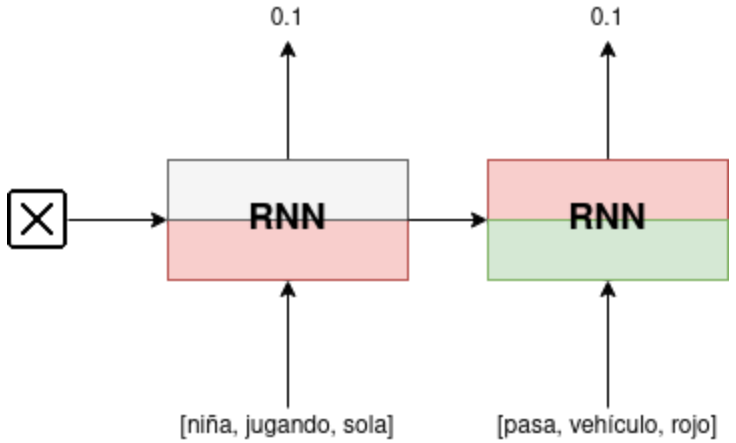


Figura 2. El segundo instante de tiempo posee conocimiento en memoria y en entrada inmediata

Intuición del funcionamiento de una RNN (IV)

Observación en $t + 2$: `(pájaros, volando, bajo)`

- La nueva observación aumenta la probabilidad a 0,7
- La memoria mantiene información acerca de las dos observaciones previas

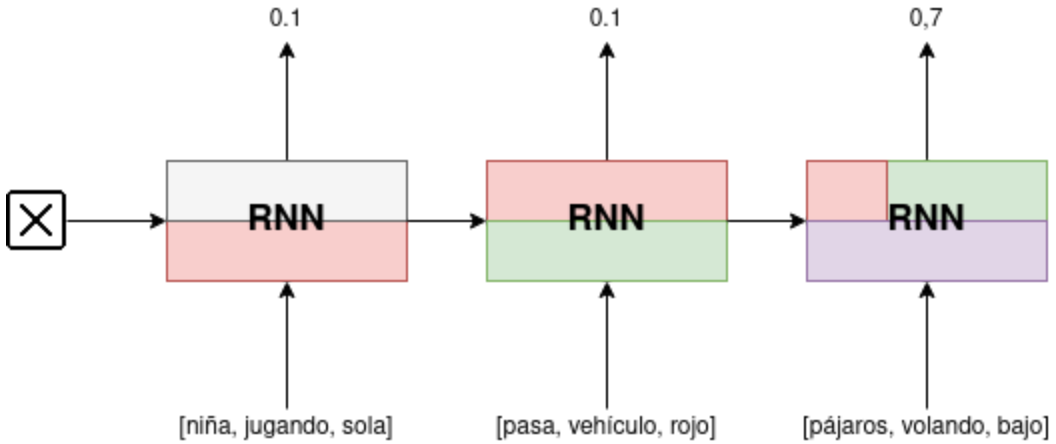


Figura 3. La memoria sigue almacenando conocimiento pasado según va llegando información nueva

Intuición del funcionamiento de una RNN (V)

Observación en $t + 3$: `(lavado, de, coche)`

- La red entiende que los últimos dos conceptos aumentan la probabilidad de lluvia
- Se sigue manteniendo el conocimiento de conceptos anteriores
  - Lo malo es que la representación de los primeros se va diluyendo rápidamente

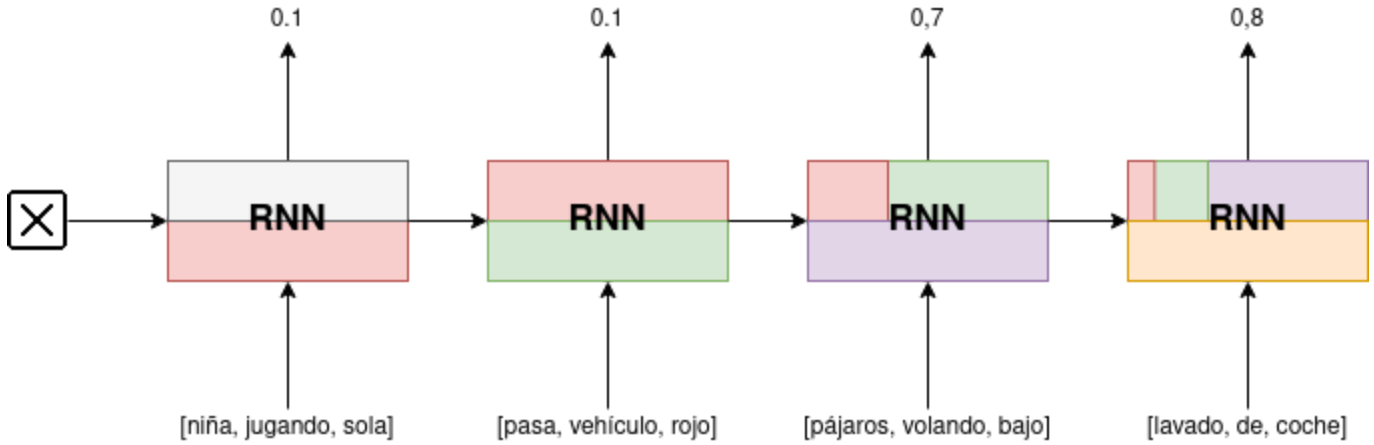


Figura 4. El conocimiento pasado se mantiene, pero el más antiguo va disminuyendo en favor del nuevo

Intuición del funcionamiento de una RNN (VI)

Observación en $t + 4$: `(cielo, con, nubarrones)`

Ha aumentado al predicción mucho

- Los conceptos anteriores "se recuerdan" y afectan a la inferencia
- Eso sí, los primeros solo la aumentan marginalmente

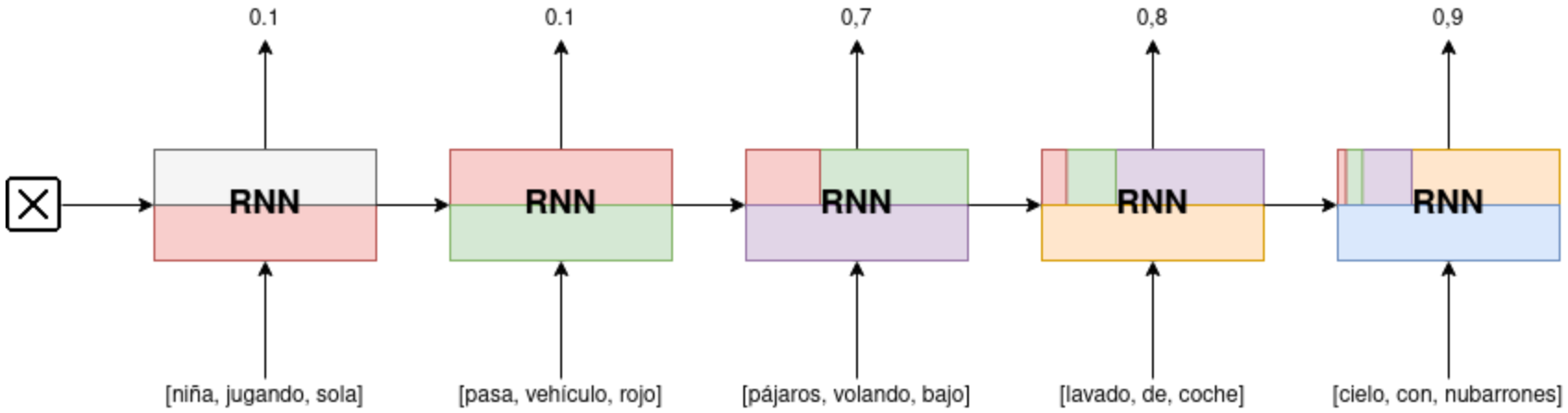


Figura 5. La predicción sigue usando conocimiento de la primera, aunque en mucha menor medida que las más recientes

Intuición del funcionamiento de una RNN (y VII)

El ejemplo ilustra cómo la red almacena la información

- Se mantiene la información de la salida anterior

Sin embargo, se intuye un problema de esta memoria

- La información anterior almacenada tiende a 0 en pocos pasos
- Ideal → Que la neurona decida qué recordar y qué olvidar
  - Bueno, lo hace (pesos de la retroalimentación) pero no es eficiente
  - Spoiler alert: Veremos una mejora más adelante

La "unidad recurrente simple" (SRU) (I)

Recurrencia $\longrightarrow$ Dependencia del valor actual con los valores anteriores

$$y_t = f(x_t, y_{t-1})$$

Las SRU son equivalentes a las neuronas en redes neuronales simples

- De hecho son muy parecidas a las neuronas de una red feed-forward
- Únicamente tienen una retroalimentación de la salida
- Una red neuronal recurrente es una combinación de una o más SRU
  - Sus arquitecturas varían con los tipos de problemas que se quieren resolver

La "unidad recurrente simple" (SRU) (II)

Existen dos formas de representar una SRU

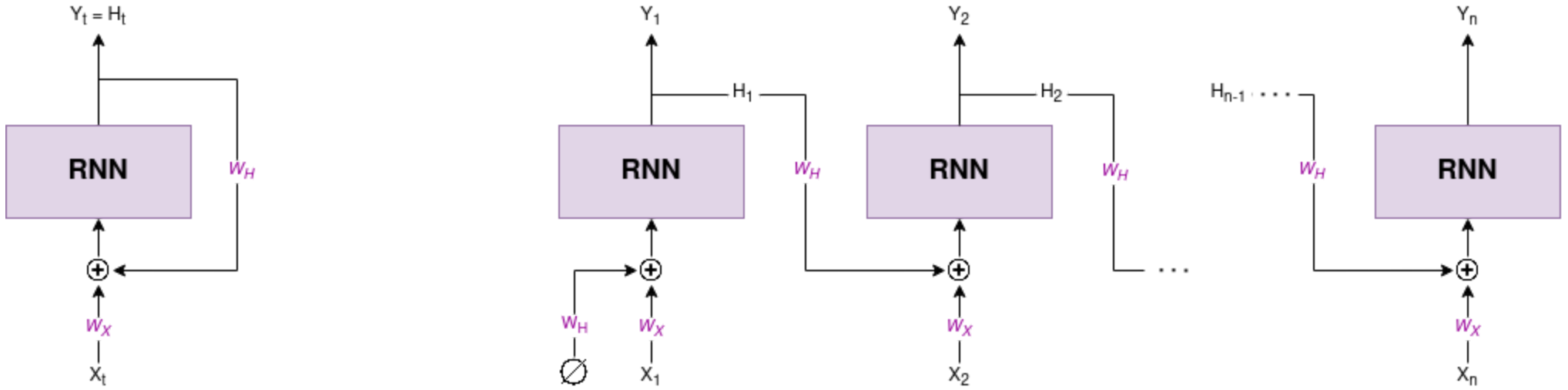


Figura 6. Las dos representaciones ípicas de una SRU, normal (izquierda) y desplegada o *unrolled* (derecha).

Ambas son equivalentes, se usan indistintamente

- Generalmente se prefiere usar la *unrolled* para facilitar la comprensión

La "unidad recurrente simple" (SRU) (y III)

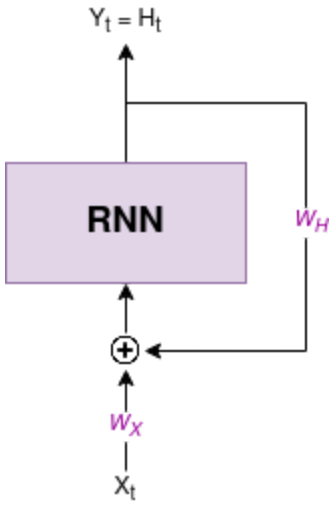


Figura 7. Representación desplegada de una SRU

Ecuación muy parecidaa a un MLP:

$$Y_t = f(W_X X_t \frown W_H H_{t-1}) = H_t$$

- X_t e Y_t: Entrada y salida de la SRU
- f: Activación (ReLU, tanh, sigm, ...)

Si pensamos en H_{t-1} como memoria

- $W_H \rightarrow$ ¿Qué recuerdos considerar?

¿Y cómo aprenden? $\rightarrow$ **Backpropagation through time (BPTT)**[2]

[2] No entramos en la implementación porque es compleja y no aporta nada al curso. Dos recursos interesantes del concepto los podemos encontrar en la entrada de la Wikipedia y el post *A Gentle Introduction to Backpropagation Through Time* de Jason Brownlee.

Un detalle al que prestar atención

Hemos visto que el valor de salida en el instante t se obtiene con la siguiente fórmula:

$$H_t = f(W_X X_t \frown W_H H_{t-1})$$

Fijémonos en que:

- Sólo intervienen dos pesos:
  - W_X que afecta a la entrada actual y W_H que afecta al valor anterior de la salida h_{t-1}
- Los pesos **no** dependen del tiempo
  - W_X: ¿A qué es interesante prestar atención ahora?
  - W_H: ¿Qué es interesante recordar de lo que pasó?

Pero las RNN sólo funcionan bien para la memoria a corto plazo

- Sobre todo si están basadas en unidades unidades recurrentes simples
- En breve veremos dos alternativas que mitigan un poco este problema

Oye, pero he leído por ahí que ...

... está demostrado que un MLP es capaz de representar cualquier función

Sí, pero terminaríamos con muchos parámetros y ninguna estructura

- Las RNN **aprovechan la estructura secuencial de los datos** en su arquitectura
- Es la misma historia que al comparar un MLP con una CNN
  - La CNN tiene una estructura interna que aprovecha la estructura espacial de los datos
- Al final utilizamos una fracción de los pesos de una forma más inteligente

También está el problema del tamaño de entrada constante

- Las redes *feed-forward* **deben tener un tamaño de entrada constante**
- El mundo real no es así; tanto imágenes como secuencias tienen distintos tamaños

Pero entonces un MLP normal vale, ¿no?

¡Por supuesto! De hecho es el mismo concepto que el de las *rolling windows*

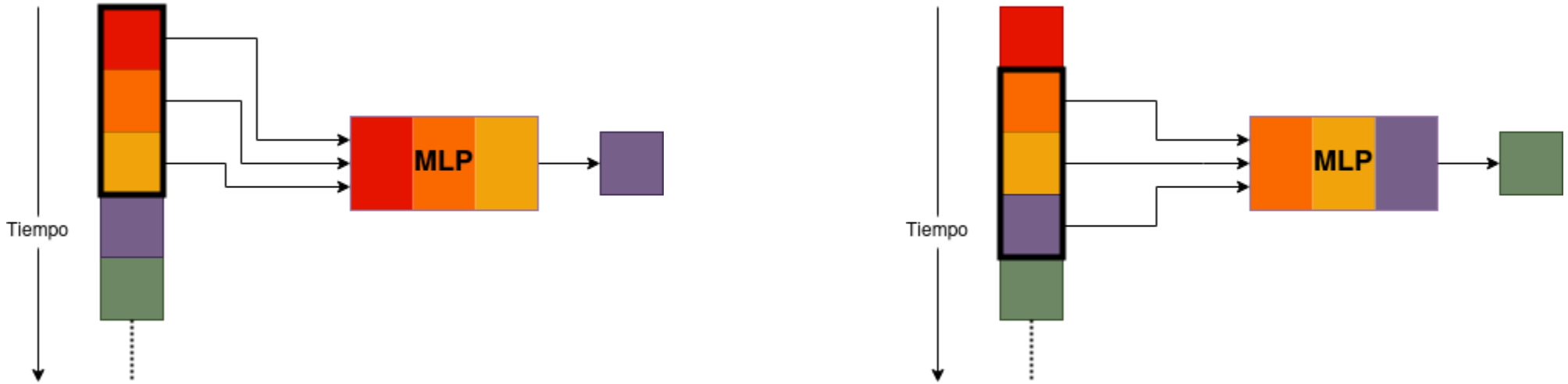


Figura 9. Al conjunto de datos se le llama *overlapping windows dataset*

Lo malo es que tienen ciertos inconvenientes

- Requieren un tamaño de entrada y salida fijos (o reentrenar)
- En realidad no hay memoria, solo lo que hay en la ventana

Además, echando cuentas ...

Supongamos un problema de clasificación de secuencias con:

- ... una longitud de secuencia de $T = 20$, ...
- ... una entrada de dimensión $D = 30$, ...
- ... una capa oculta de tamaño $M = 16$, ...
- ... y $k = 3$ clases de salida

En un MLP

$$I \to H: \quad T \times D \times M = 9600$$
$$H \to O: \quad M \times k = 48$$

Total: 9648 parámetros a ajustar

En una RNN

$$I \to H: \quad (D + M) \times M = 736$$
$$H \to O: \quad M \times k = 48$$

Total: 784 parámetros a ajustar

Implementando redes recurrentes

Conceptos generales: *Stacked* RNN

Apilar una SRU sobre otra nos lleva al concepto de RNN profundas (DRNN)

- Le damos los valores de entrada a la primera unidad ...
- ... la salida de la primera a la segunda ...
- ... y así sucesivamente ...

$$H_t^1 = f(W_X^1 X_t + W_H^1 H_{t-1}^1) \qquad (1)$$
$$H_t^2 = f(W_i^2 H_{t-1}^1 + W_H^2 H_{t-1}^2) \qquad (1)$$
$$\dots$$
$$H_t^n = f(W_i^n H_{t-(n-1)}^{n-1} + W_H^n H_{t-(n-1)}^n) \qquad (1)$$

En realidad es prácticamente lo mismo que añadir capas en un MLP

Ventajas e inconvenientes de estas arquitecturas

Ventajas

- Posibilidad de procesar entradas de longitud variable
- El tamaño del modelo no crece con el tamaño de la entrada
- La inferencia tiene en cuenta la información histórica
- Los pesos se comparten a lo largo del tiempo

Inconvenientes

- procesamiento es mucho más lento
- Es difícil acceder a información lejana

Implementación en keras (I)

Normalmente una capa produce una única salida:

```
o = Dense(128)(i)
```

En unidades recurrentes podemos devolver el estado oculto:

```
o, h = SimpleRNN(128, return_state=True)(i)
o, h = GRU(128, return_state=True)(i)
o, h, c = LSTM(128, return_state=True)(i)
```

La diferencia de `o` y `h` es... ¡ninguna!, así que ¿por qué preocuparnos en devolverlas?

- `return_sequences = True`: `o` pasa a ser una secuencia:

```
o, h = SimpleRNN(128, return_sequence=True)(i)
```

Implementación en keras (y II)

Hay tres capas RNN integradas en Keras:

- `keras.layers.SimpleRNN`: SRU donde la salida en $t-1$ se usa de entrada en t
- `keras.layers.LSTM` , Propuesta por Hochreiter y Schmidhuber[3] en 1997
- `keras.layers.GRU`: Propuesta por primera vez por Cho et al.[4] en 2014.

La primeras implementaciones de código abierto de LSTM y GRU fueron en 2015

[3] Artículo: *Long short-term memory*
[4] Artículo: *Learning phrase representations using RNN encoder-decoder for statistical machine translation*

Implementación de RNN con Modelo Secuencial

Ejemplo de implementación de un modelo simple de RNN

```
model = keras.Sequential()
model.add(layers.Embedding(input_dim=1000, output_dim=64))

# La salida de GRU será un tensor 3D de dimensión (batch_size, timesteps, 256)
model.add(layers.SimpleRNN(256, return_sequences=True))

# La salida de SimpleRNN será un tensor 2D de dimensión (batch_size, 128)
model.add(layers.SimpleRNN(128))

model.add(layers.Dense(10))

model.summary()
```

Un apunte las limitaciones de Keras

En Keras las SRU **usan secuencias de longitud fija** $N \times T \times D$

- Esto es, N ejemplos de longitud T, con dimensión de entrada D
- El código resultante es más sencillo y más rápido (operaciones en batch)
  - Se pueden almacenar en matrices de NumPy o en tensores
  - ¡Un T variable implica tener que hacer un bucle sobre todo!

Lo malo es que tenemos que elegir la longitud de secuencia T

- No caer en la trampa de escoger la más larga:
  - Las secuencias más cortas se rellenarán con ceros
  - Las más largas se truncarán (e.g. conjunto de test)
  - Son muy raras[5] (derroche de espacio y de tiempo)

[5] La larga estela.

Arquitecturas de redes recurrentes

Sobre los difentes tipos de problema

Las redes *feed-forward* tienen siempre la misma estructura

- Se presenta **un input**, se obtiene **un output**
- Prácticamente todos los problemas vistos hasta ahora son así
  - P.ej. MNIST: imagen de entrada $\longrightarrow$ un único dígito de salida
- Se conoce como ***one-to-one***

Ahora tenemos potencialmente **entradas y salidas como secuencias**

- **Entrada única, secuencia de salida**: P.ej. Subtitulado de imágenes
- **Secuencia de entrada, única salida**: P.ej. Clasificación de malware
- **Secuencia de entrada, secuencia de salida**: P.ej. Traducción automática o *name entity recognition*

Utilizaremos la notación T_x y T_y para la longitud de entrada y salida

Arquitectura *one-to-many* ($T_x = 1, T_y > 1$)

Generación de secuencias a partir de una única entrada

- Aplicaciones: generación de texto, de tráfico de red, etiquetado de imágenes, ...

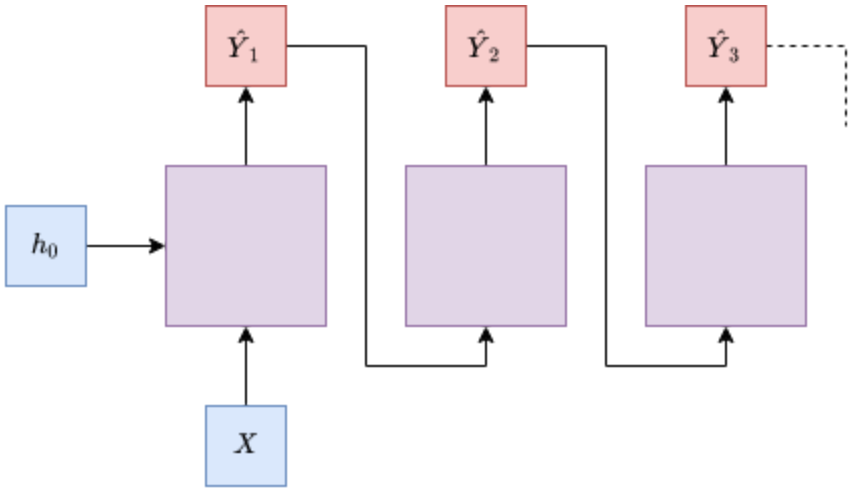


Figura 10 Ilustración de la arquitectura *one-to-many*.

Arquitectura *many-to-one* ($T_x > 1, T_y = 1$)

Una única respuesta a partir de una secuencia de entrada (típico para clasificación)

Ejemplo - Análisis de malware de $2KiB$: In: $X_1, \ldots, X_{2048}$, Out: $Y_1, \ldots, Y_{2048}$

- Qué Y_i cogemos? Tiene sentido el último, ya que así ha visto todo el ejecutable
- Keras $\rightarrow$ `return_sequences=False` para devolver la última salida

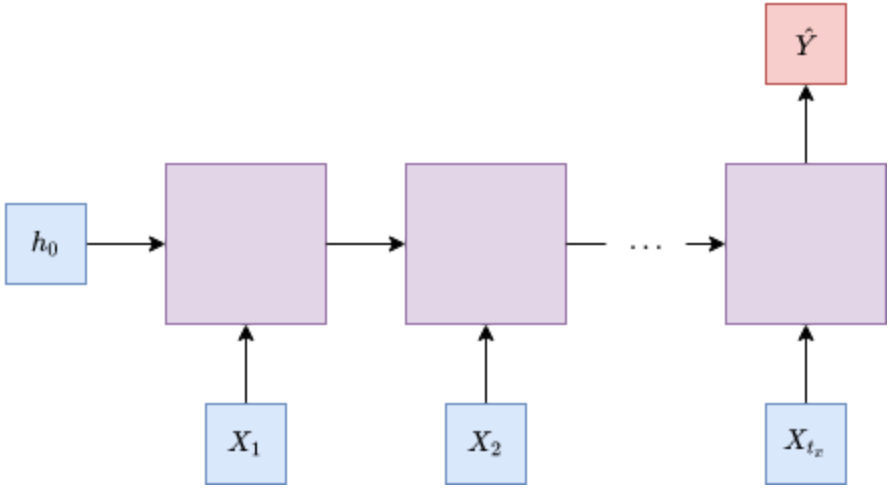


Figura 11 Ilustración de la arquitectura *many-to-one*.

Arquitectura *many-to-one* - Otra vuelta de tuerca

¿Por qué el último estado es el mejor, ¡podríamos estar perdiendo información clave!

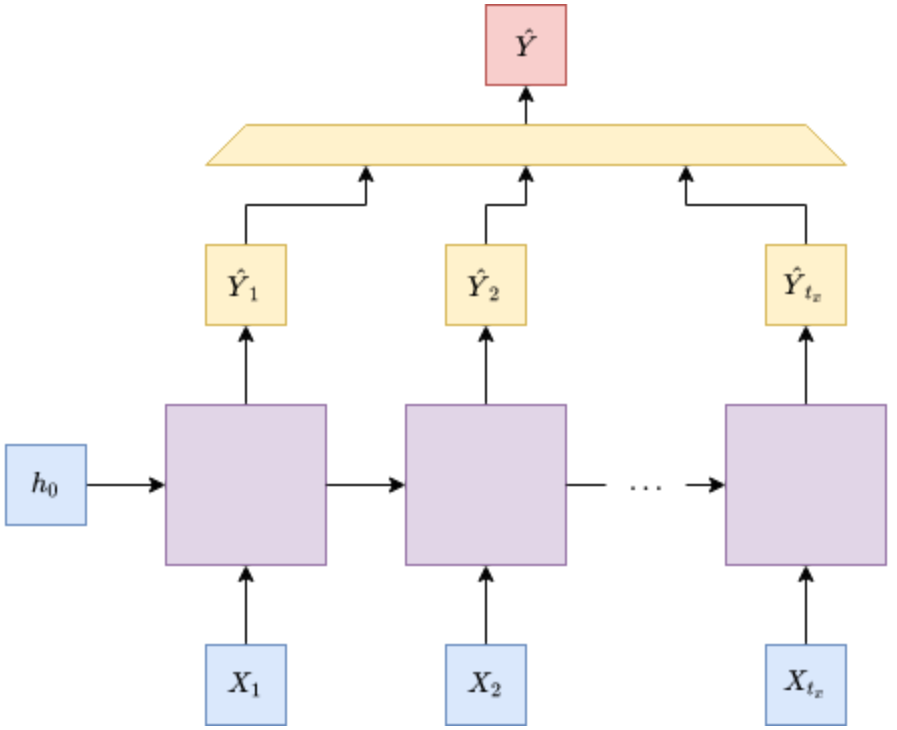


Figura 12 Ilustración de la arquitectura *many-to-one* usando *max pooling* para determinar la salida.

Arquitectura *many-to-many* ($T_x = T_y$)

Dos secuencias de entrada y salida, ambas **del mismo tamaño**

- Típicas de problemas de etiquetado gramatical y similares

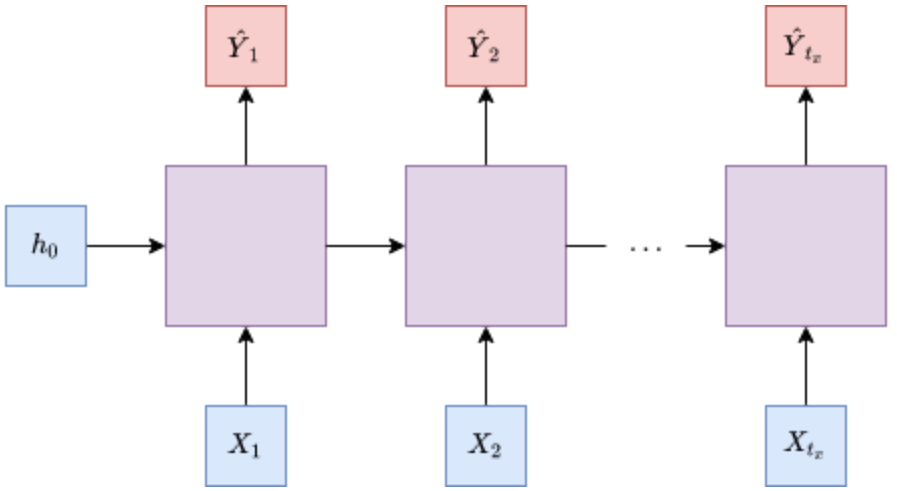


Figura 13 Ilustración de la arquitectura *many-to-many* con mismo tamaño de entrada-salida.

Arquitectura *many-to-many* ($T_x \neq T_y$)

Secuencia de entrada que genera una de **salida** de, generalmente, **distinto tamaño**

- Típicas de problemas de traducción automática

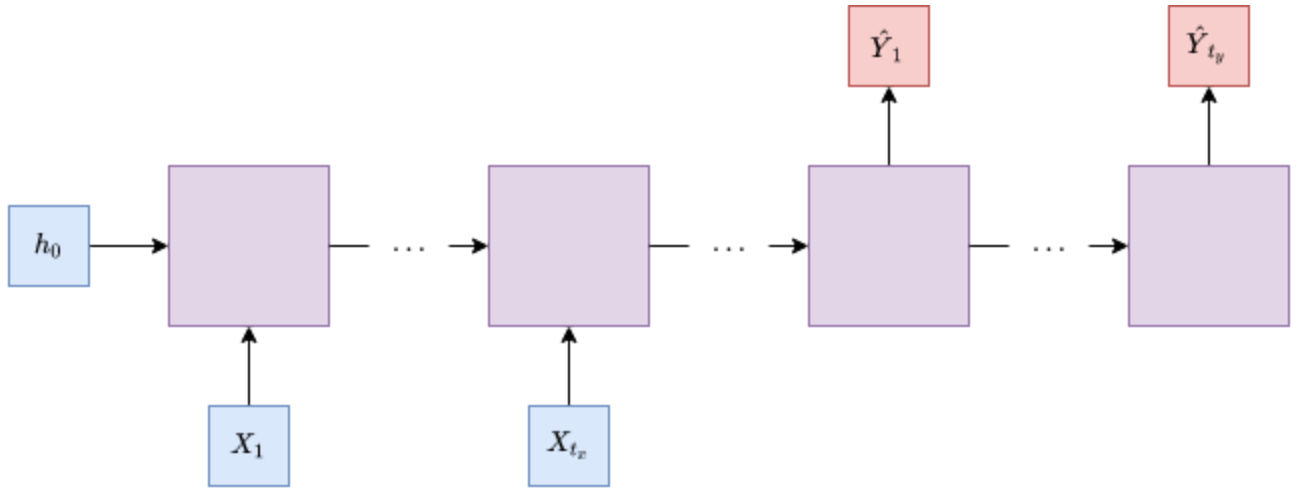


Figura 14 Ilustración de la arquitectura *many-to-many* con (potencial) diferente tamaño de entrada-salida.

Tipos de arquitecturas de redes neuronales

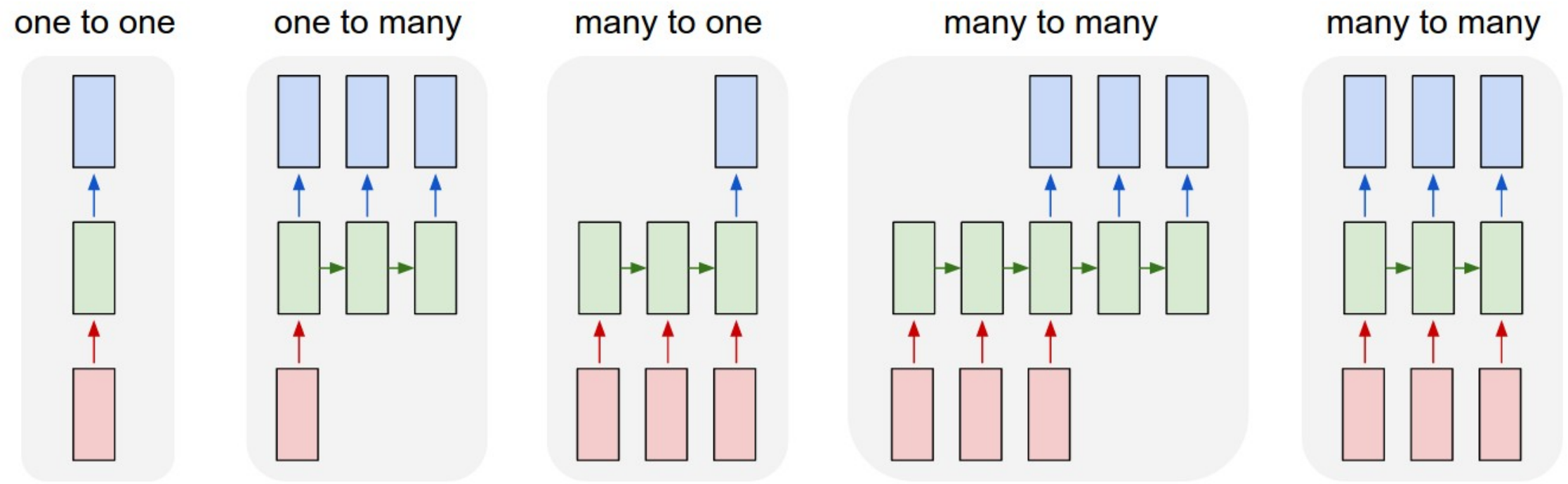


Figura 15 Diferentes problemas en redes neuronales recurrentes. Autor: Andrej Karpathy[6].

[6] Extraído de la entrada del blog del autor *The Unreasonable Effectiveness of Recurrent Neural Networks*.

¿Se diferencian las RNN de las CNN? (I)

Ambas han impulsado el rendimiento de la inteligencia artificial en los últimos años

Las redes convolucionales son redes de tipo ***feed-forward***

- Utilizan capas de filtros y de *pooling*
- Los tamañas de entrada y salida son fijos
- Se utilizan habitualmente sobre datos espaciales (p.ej. imagenes)

Por otro lado, redes recurrentes **retroalimentan** los resultados a la red

- Son adecuadas para datos secuenciales (p.ej. texto o vídeo)
- El tamaño de la entrada y la salida resultante pueden variar
- Sus casos de uso incluyen el procesamiento del lenguaje natural, la detección de anomalías o la generación de secuencias

¿Se diferencian las RNN de las CNN? (II)

¿Qué pasa si queremos entender qué ocurre en las imágenes?

Figura 16 ¿Hacia dónde se mueven la cabeza y la pelota?.

La secuencia de imágenes pasada le da contexto a la predicción

¿Se diferencian las RNN de las CNN? (III)

Figura 17 Los eventos pasados añaden información y contexto.

¿Cómo hacemos que las redes recuerden esta información previa?

- Las RNN se crearon para abordar esto, aunque ya hemos visto su principal problema
  - La información más cercana al presente tiene más influencia
  - Cuanto más nos alejamos, más se diluye la información pasada

¿Se diferencian las RNN de las CNN? (y IV)

Hemos aprendido que las redes recurrentes tienen en cuenta píxeles adyacentes

- ¿Y el mismo mecanismo para eventos adyacentes en el tiempo?
  - No es que no funcione, de hecho se usa bastante
  - De hecho es una muy buena forma de arrancar una prueba de concepto
    - Las interfaces en Keras son prácticamente iguales
  - Es un enfoque innecesariamente indirecto
    - Es usar un modelado espacial para captar un fenómeno temporal
    - Requiere mucho más esfuerzo y memoria para realizar la misma tarea

Ambas técnicas son complementarias $\rightarrow$ Redes de convolución recurrentes (CRNN):

- Etiquetado de vídeos, reconocimiento de gestos, *sentiment analysis* en lenguas logográficas, ...

Sobre las dependencias a largo plazo

Supongamos la frase "Las **nubes** están en el ..."

- La respuesta más obvia sería **cielo**
- Casi no necesitamos más contexto que la palabra **nubes**

Ahora supongamos la frase:

*"He residido en **España** los últimos 10 años, durante los cuales he visitado gran parte del país y disfrutado de su gastronomía. Puedo hablar con bastante fluidez el ...*

- La respuesta obvia es **español** pero claro, usar un cerebro humano es hacer trampa
- La información relevante está muy lejos del lugar a predecir

Las arquitecturas LSTM y GRU nos ayudan a **mitigar** este problema

- Implementan técnicas para aprender **qué** elementos **retener** y **cuáles olvidar**

Unidades Long Short-Term Memory

Long Short-Term Memory (LSTM)

Son un tipo especial de SRU capaz de aprender dependencias a largo plazo

- Diseñadas **expresamente** para el problema de las dependencias
- Tratan de entender qué datos deberían ser recordados y qué datos pueden olvidarse

Son un **recambio directo de las SRU**

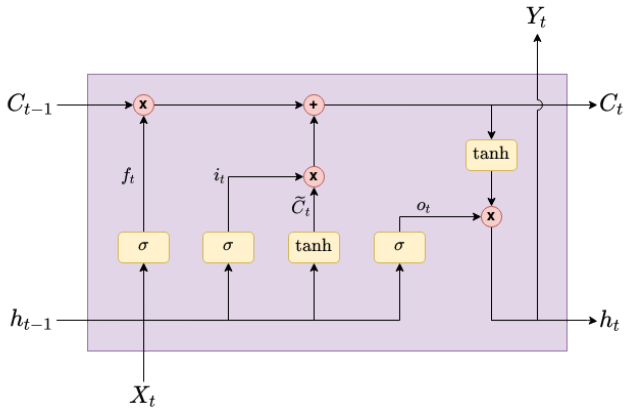


Figura 17 Los eventos pasados añaden información y contexto.

El estado de la unidad LSTM

La clave de las LSTM es la línea horizontal superior

- Es el **estado de la unidad** y la información fluye por ella, generalmente sin cambios

La unidad tiene la **capacidad de añadir o quitar información** al estado

- Esto se gestiona a través de las estructuras denominadas puertas
- Son una forma de dejar (o no) pasar información
- Están compuestas por unos pesos, una activación σ y un producto punto a punto
  - La activación determina la cantidad de cada componente que "pasa"
  - Va de 0 (no pasa nada) a 1 (pasa totalmente)

1. ¿Qué información descartamos?

Comprueba h_{t-1} y X_t y devuelve un valor entre 0 y 1 para cada componente de C_{t-1}

- De 0 (olvidar completamente) a 1 (mantener completamente)

Forget gate layer

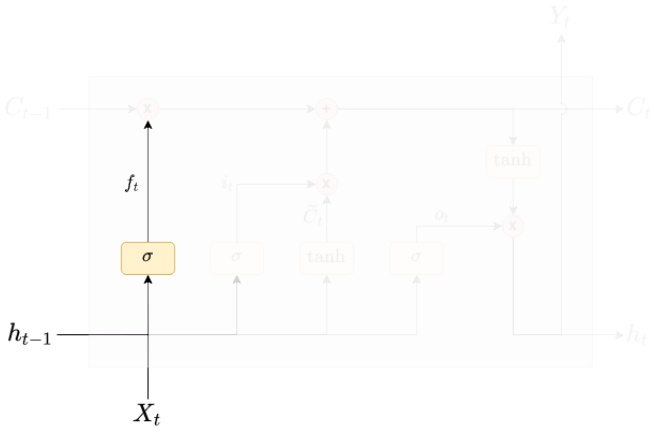


Figura 18 *Forget gate layer*.

Ejemplo

El estado mantiene el género:

- Así usa los pronombres correctos
- Si viene un nuevo sujeto, probablemente querremos olvidar el anterior

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

2. ¿A qué información prestamos atención?

Comprueba h_{t-1} y X_t y devuelve:

- Valores que vamos a actualizar (\textit{input gate layer)}
- Valores candidatos que \textit{podrían} se añadidos al estado

Input gate layer

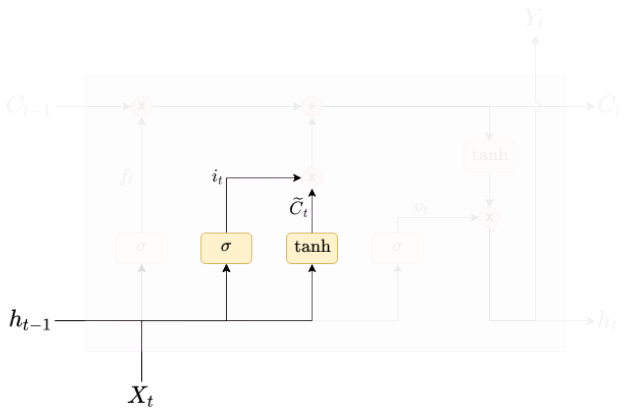


Figura 19 *Input gate layer*.

Sobre el ejemplo anterior

- Se actualizaría la información del género
- Se añadiría información si fuese necesario

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \quad (1)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, X_t] + b_C) \quad (1)$$

3. Actualización del estado

Se actualiza el estado C_{t-1} con la nueva información

- En los anteriores pasos **se decide** qué hacer con las entradas
- **Ahora** únicamente tenemos que **hacerlo**

Actualización del estado

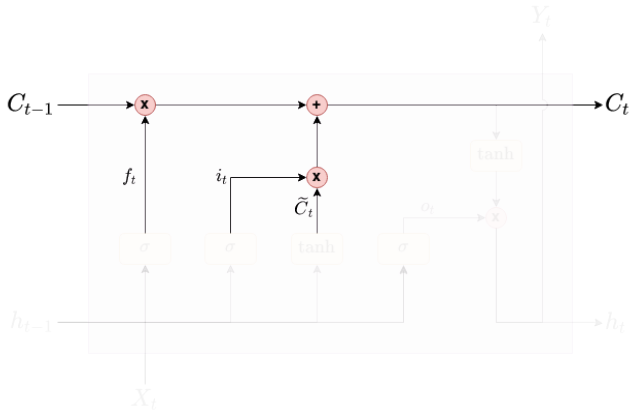


Figura 20 *Actualización del estado*.

Sobre el ejemplo anterior

- Aquí se eliminaría la información sobre el género del antiguo sujeto

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

Este nuevo estado se le pasará a la neurona en la siguiente iteración

4. ¿Qué información damos de salida?

Decidimos la salida (filtrando el estado que estamos pasando):

- Capa sigmoide que decide qué partes del estado sacar
- Tangente del estado para transformar al intervalo $(-1, 1)$
- Producto de ambas para sólo sacar las partes decididas

Actualización del estado

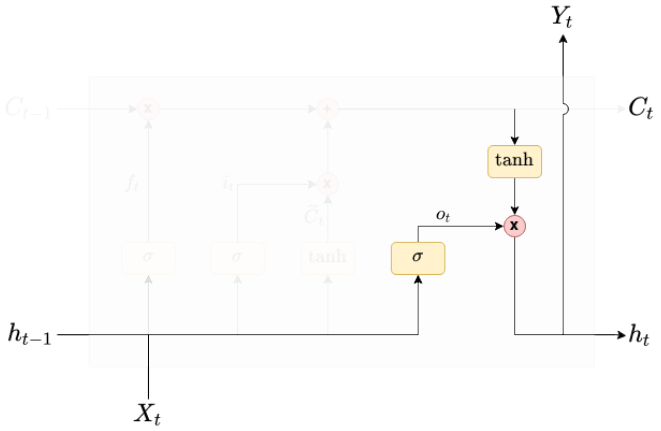


Figura 21 *Actualización del estado*.

Sobre el ejemplo anterior

- Da información relevante a:
  - ... género para construir correctamente un adjetivo
  - ... número para conjugar correctamente un verbo

$$o_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \quad (1)$$

$$h_t = o_t \odot \tanh(C_t) \quad (1)$$

Unidades Gated Recurrent Unit

Unidades *Gated Recurrent Unit* (GRU) (I)

Se pueden considerar como una **versión simplificada de las** unidades **LSTM**

- Mismo objetivo que las LSTM $\rightarrow$ mitigar el problema de la memoria a largo plazo

Su rendimiento es muy similar al de las LSTM

- Empíricamente parece que GRU se comporta mejor con \textit{datasets} pequeños
- Sin embargo, algunos autores recomiendan combinar ambas:
  - Capacidad de aprender asociaciones a largo plazo para la LSTM
  - Capacidad de aprender de patrones a corto plazo para la GRU
- ¿Cuál usar? $\rightarrow$ **Ensayo y error**

Unidades *Gated Recurrent Unit* (GRU) (y II)

Unidad recurrente GRU

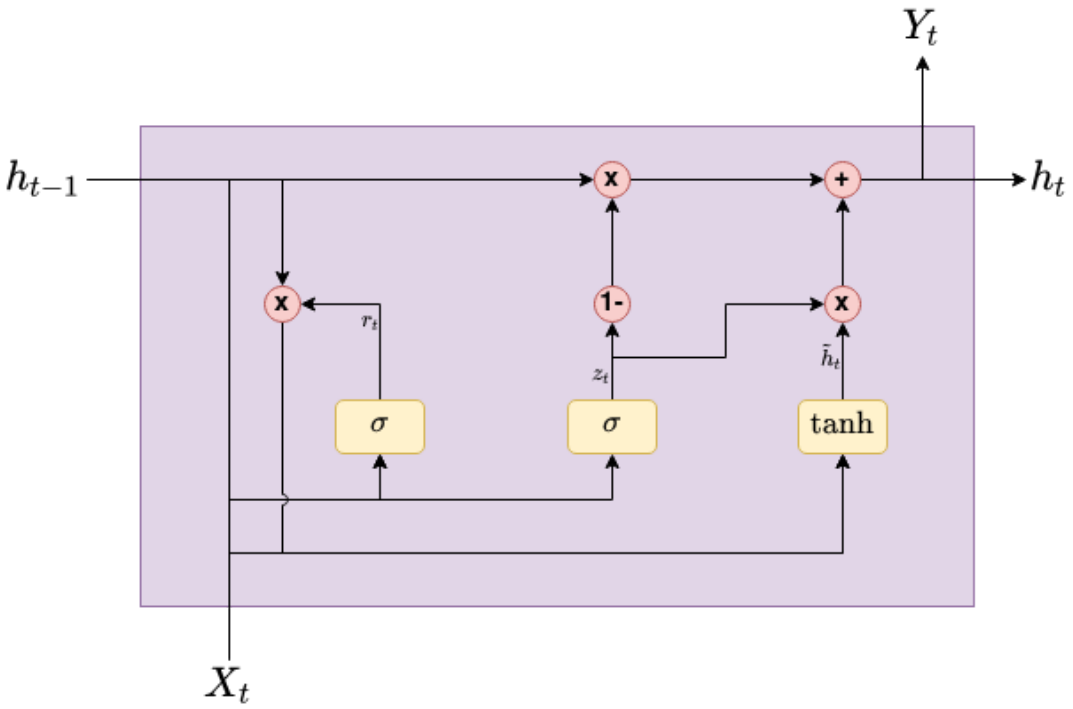


Figura 22 *Esquema de una unidad recurrente GRU*.

Características

1. No existe entrada de estado
2. Dos puertas para gestionar entradas
   i. Salida anterior y entrada actual
   ii. Salida anterior y actual

$$z_t = \sigma(W_z[h_{t-1}, X_t] + b_z) \quad (1)$$

$$r_t = \sigma(W_r[h_{t-1}, X_t] + b_r) \quad (1)$$

$$\widetilde{h_t} = \tanh(W_h[r_t \odot h_{t-1}, x_t]) \quad (1)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \quad (1)$$

Resumen de unidades recurrentes

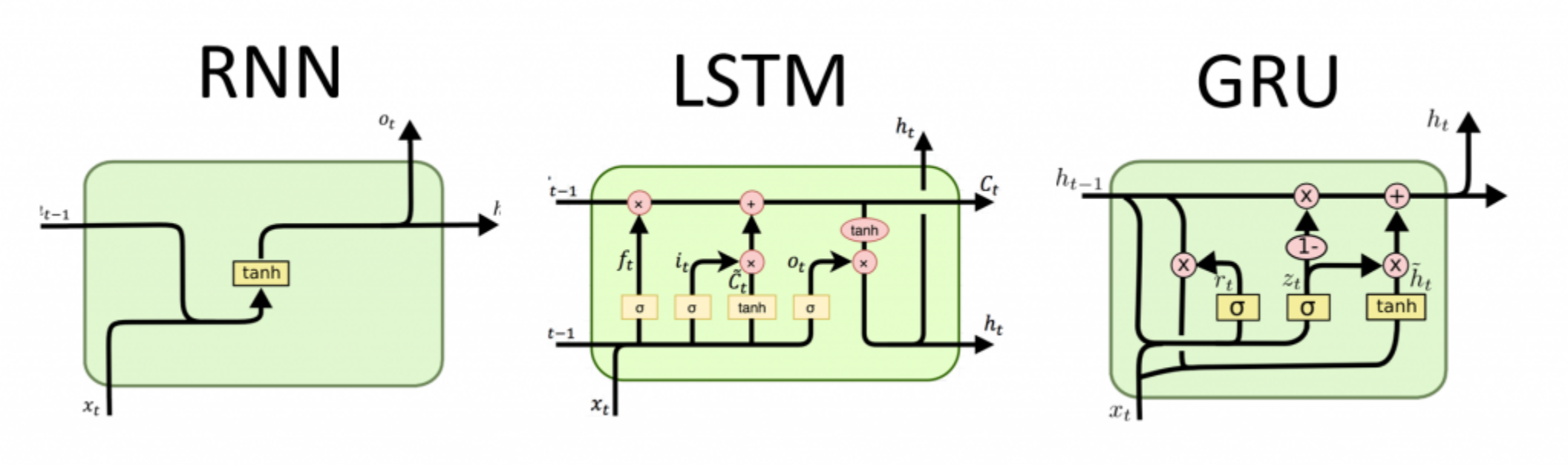


Figura 23 *Comparativa de los diferentes modelos de redes recurrentes*.

Gracias