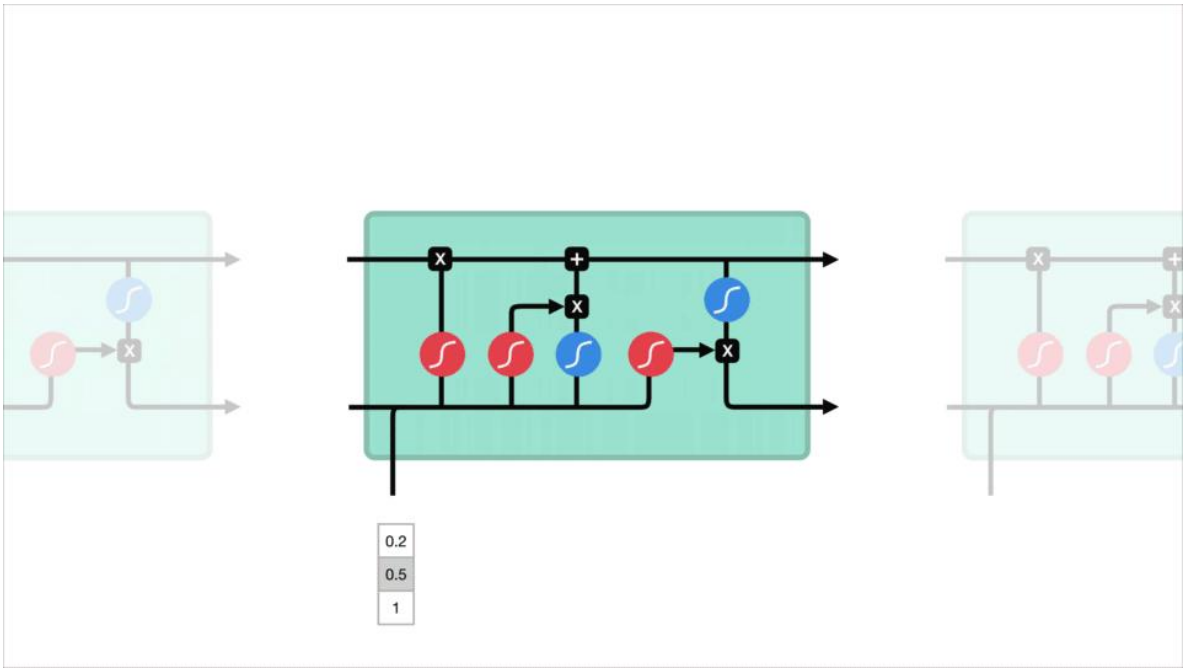
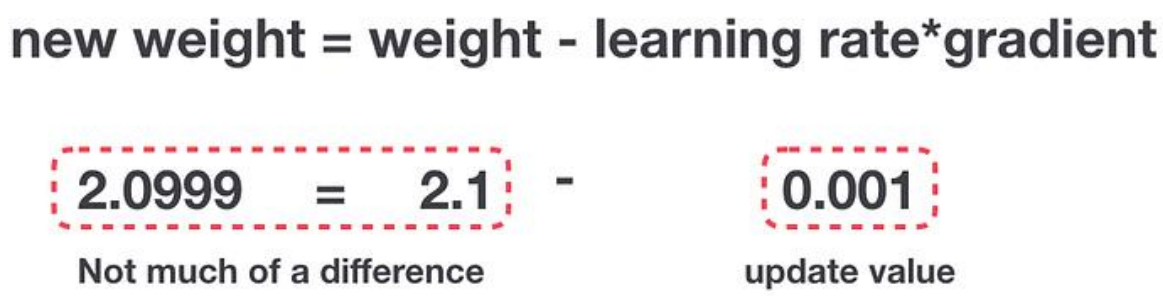
LSTM y GRU



El problema, la memoria a corto plazo

Las redes neuronales recurrentes sufren de memoria a corto plazo. Si una secuencia es lo suficientemente larga, tendrán dificultades para llevar información de pasos de tiempo anteriores a los posteriores. Entonces, si está tratando de procesar un párrafo de texto para hacer predicciones, las RNN pueden omitir información importante desde el principio.

Durante la propagación hacia atrás, las redes neuronales recurrentes sufren el problema del gradiente de desaparición. Los gradientes son valores que se utilizan para actualizar los pesos de una red neuronal. El problema del gradiente de desaparición es cuando el gradiente se encoge a medida que se propaga a través del tiempo. Si un valor de gradiente se vuelve extremadamente pequeño, no aporta demasiado aprendizaje.

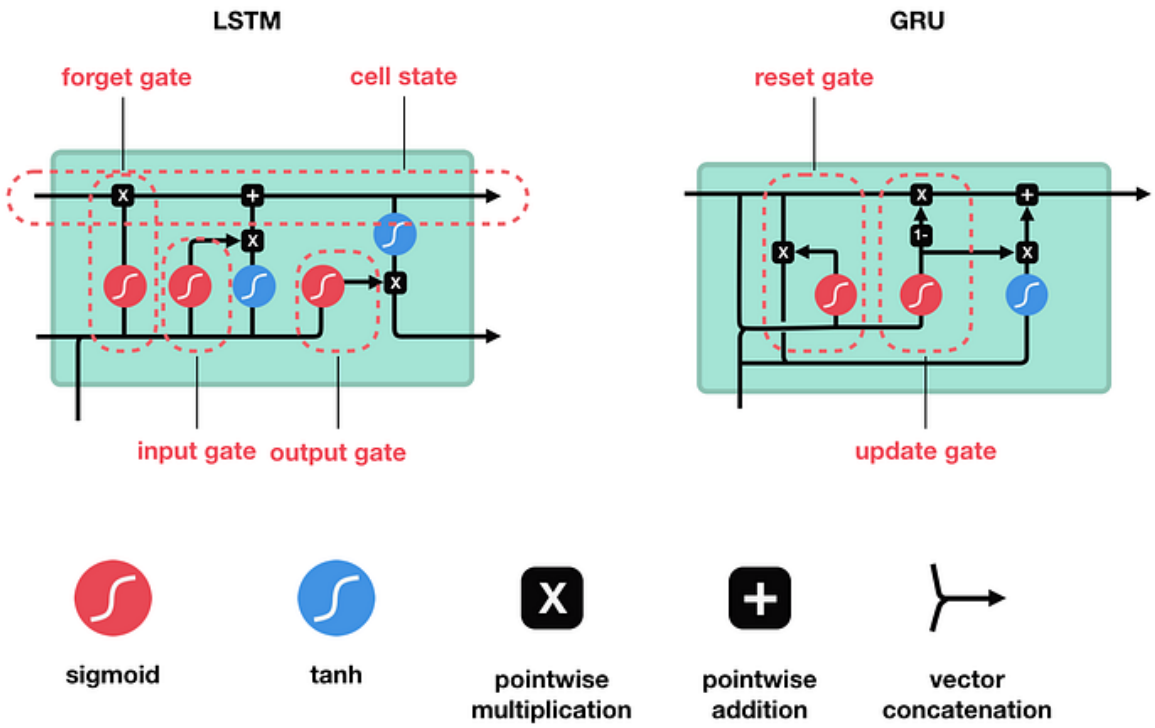


Regla de actualización de degradado

Entonces, en las redes neuronales recurrentes, las capas que reciben una pequeña actualización de gradiente dejan de aprender. Esas suelen ser las primeras capas. Entonces, debido a que estas capas no aprenden, las RNN pueden olvidar lo que vieron en secuencias más largas, por lo que tienen una memoria a corto plazo. Si quieres saber más sobre la mecánica de las redes neuronales recurrentes en general, puedes leer mi post anterior aquí.

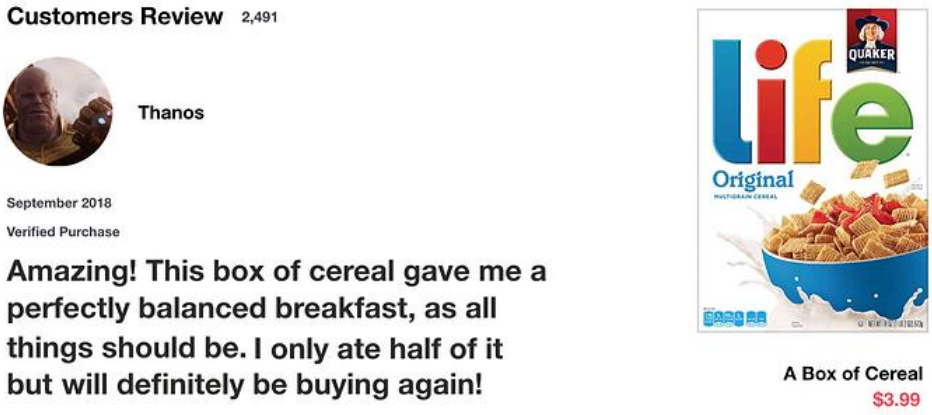
LSTM y GRU como solución

LSTM y GRU se crearon como la solución a la memoria a corto plazo. Tienen mecanismos internos llamados puertas que pueden regular el flujo de información.

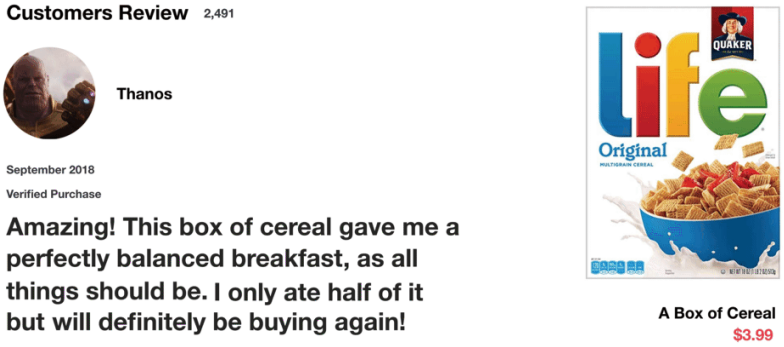


Estas puertas pueden aprender qué datos de una secuencia es importante conservar o desechar. Al hacer eso, puede pasar información relevante a lo largo de la larga cadena de secuencias para hacer predicciones. Casi todos los resultados de vanguardia basados en redes neuronales recurrentes se logran con estas dos redes. LSTM y GRU se pueden encontrar en reconocimiento de voz, síntesis de voz y generación de texto. Incluso puede usarlos para generar subtítulos para videos.

Digamos que está mirando reseñas en línea para determinar si desea comprar cereales Life (no me pregunte por qué). Primero leerá la reseña y luego determinará si alguien pensó que era bueno o si era malo.



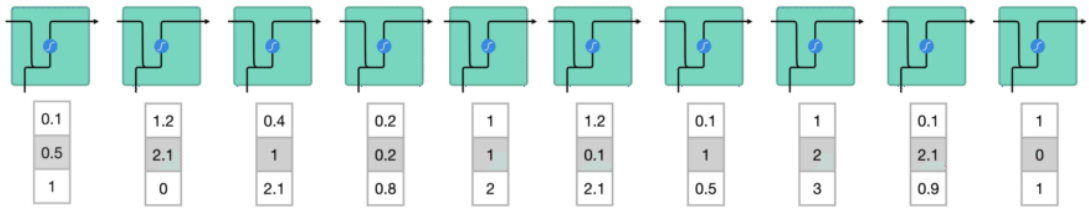
Cuando lees la reseña, tu cerebro inconscientemente solo recuerda palabras clave importantes. Aprendes palabras como "increíble" y "desayuno perfectamente equilibrado". No te importan mucho palabras como "esto", "dio", "todos", "debería", etc. Si un amigo te pregunta al día siguiente qué dice la reseña, probablemente no lo recordarías palabra por palabra. Sin embargo, es posible que recuerde los puntos principales como "definitivamente volveré a comprar". Si eres muy parecido a mí, las otras palabras se desvanecerán de la memoria.



Y eso es esencialmente lo que hace un LSTM o GRU. Puede aprender a guardar solo información relevante para hacer predicciones y olvidar datos no relevantes. En este caso, las palabras que recordabas te hacían juzgar que era bueno.

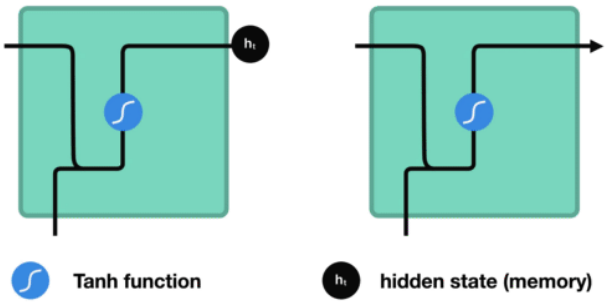
Revisión de redes neuronales recurrentes

Para comprender cómo LSTM o GRU logra esto, revisemos la red neuronal recurrente. Una RNN funciona así; Las primeras palabras se transforman en vectores legibles por máquina. Luego, la RNN procesa la secuencia de vectores uno por uno.



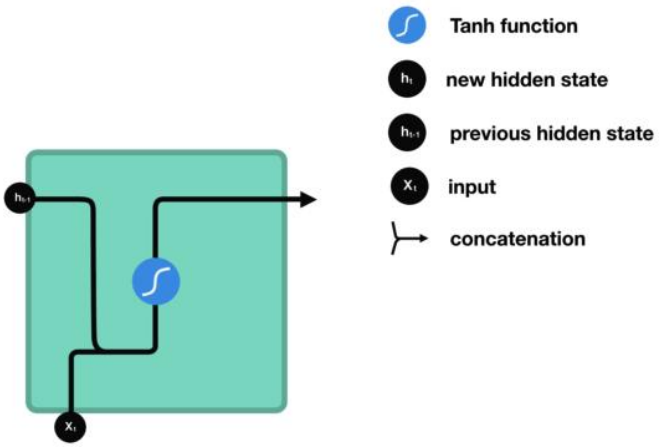
Secuencia de procesamiento uno por uno

Durante el procesamiento, pasa el estado oculto anterior al siguiente paso de la secuencia. El estado oculto actúa como la memoria de las redes neuronales. Contiene información sobre datos anteriores que la red ha visto antes.



Pasar el estado oculto al siguiente paso de tiempo

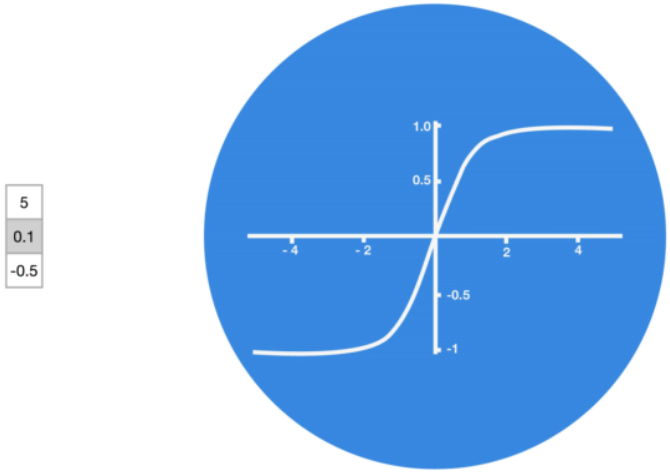
Veamos una celda de la RNN para ver cómo calcularía el estado oculto. Primero, la entrada y el estado oculto anterior se combinan para formar un vector. Ese vector ahora tiene información sobre la entrada actual y las entradas anteriores. El vector pasa por la activación tanh y la salida es el nuevo estado oculto o la memoria de la red.



Célula RNN

Activación de Tanh

La activación tanh se utiliza para ayudar a regular los valores que fluyen a través de la red. La función tanh aplasta los valores para que siempre estén entre -1 y 1.

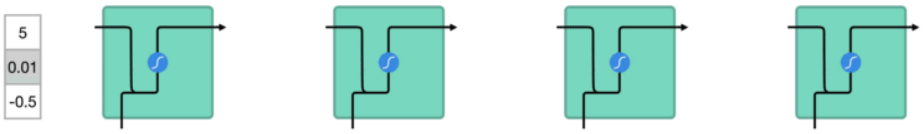


Tanh aplasta los valores para que estén entre -1 y 1

Cuando los vectores fluyen a través de una red neuronal, sufre muchas transformaciones debido a varias operaciones matemáticas. Así que imagina un valor que continúa multiplicándose por, digamos, ***3***. Puede ver cómo algunos valores pueden explotar y volverse astronómicos, lo que hace que otros valores parezcan insignificantes.

Transformaciones vectoriales sin tanh

Una función tanh asegura que los valores permanezcan entre -1 y 1, regulando así la salida de la red neuronal. Puede ver cómo los mismos valores de arriba permanecen entre los límites permitidos por la función tanh.

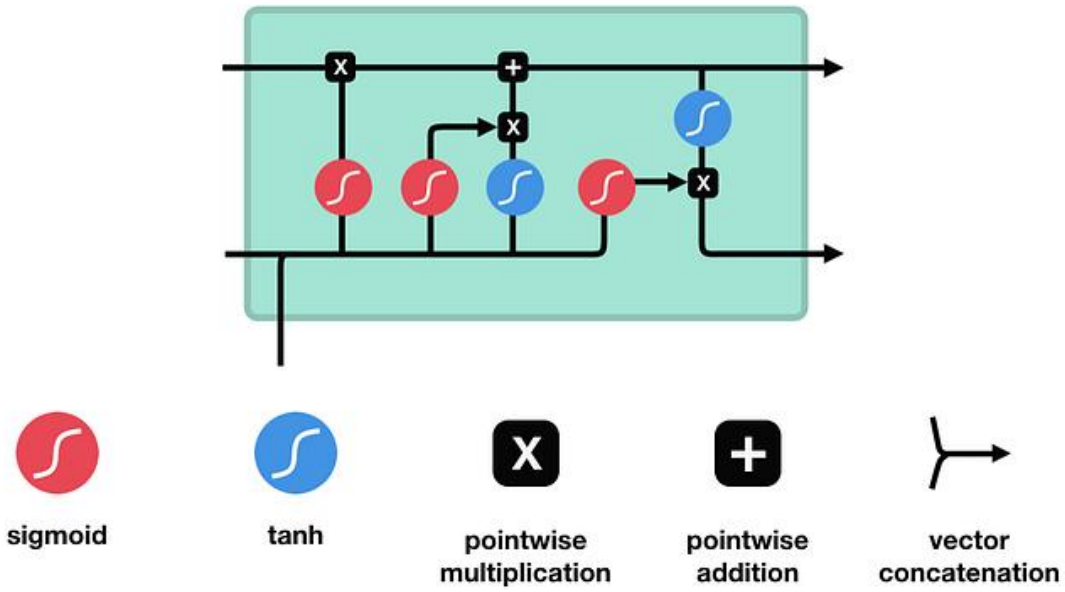


Transformaciones vectoriales con tanh

Así que eso es una RNN. Tiene muy pocas operaciones internamente, pero funciona bastante bien dadas las circunstancias adecuadas (como secuencias cortas). RNN utiliza muchos menos recursos computacionales que sus variantes evolucionadas, LSTM y GRU.

LSTM

Un LSTM tiene un flujo de control similar al de una red neuronal recurrente. Procesa los datos que transmiten información a medida que se propagan hacia adelante. Las diferencias son las operaciones dentro de las celdas del LSTM.



La celda LSTM y sus operaciones

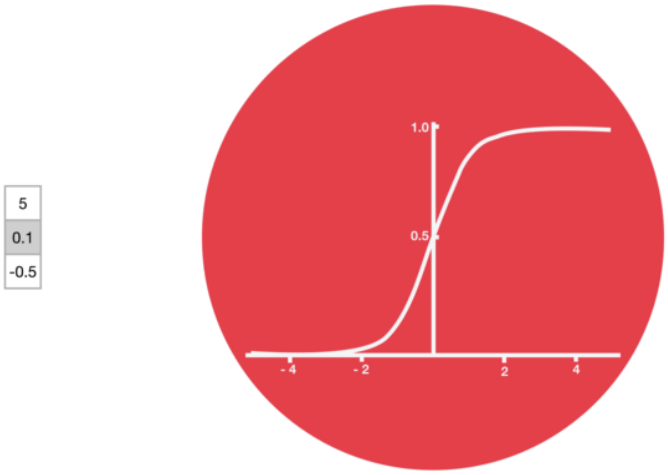
Estas operaciones se utilizan para permitir que el LSTM conserve u olvide información. Ahora, mirar estas operaciones puede ser un poco abrumador, así que repasaremos esto paso a paso.

Concepto central

El concepto central de LSTM es el estado de la celda y sus varias puertas. El estado de la celda actúa como una autopista de transporte que transfiere información relativa a lo largo de la cadena de secuencia. Puedes pensar en ello como la "memoria" de la red. El estado celular, en teoría, puede transportar información relevante a lo largo del procesamiento de la secuencia. Por lo tanto, incluso la información de los pasos de tiempo anteriores puede llegar a pasos de tiempo posteriores, reduciendo los efectos de la memoria a corto plazo. A medida que el estado de la celda continúa su viaje, la información se agrega o elimina al estado de la celda a través de puertas. Las puertas son diferentes redes neuronales que deciden qué información se permite sobre el estado de la célula. Las puertas pueden aprender qué información es relevante conservar u olvidar durante la capacitación.

Sigmoide

Gates contiene activaciones sigmoides. Una activación sigmoidea es similar a la activación tanh. En lugar de aplastar valores entre -1 y 1, aplasta valores entre 0 y 1. Esto es útil para actualizar u olvidar datos porque cualquier número que se multiplique por 0 es 0, lo que hace que los valores desaparezcan o se "olviden". Cualquier número multiplicado por 1 es el mismo valor, por lo tanto, ese valor permanece igual o se "mantiene". La red puede aprender qué datos no son importantes, por lo tanto, pueden olvidarse o qué datos son importantes conservar.

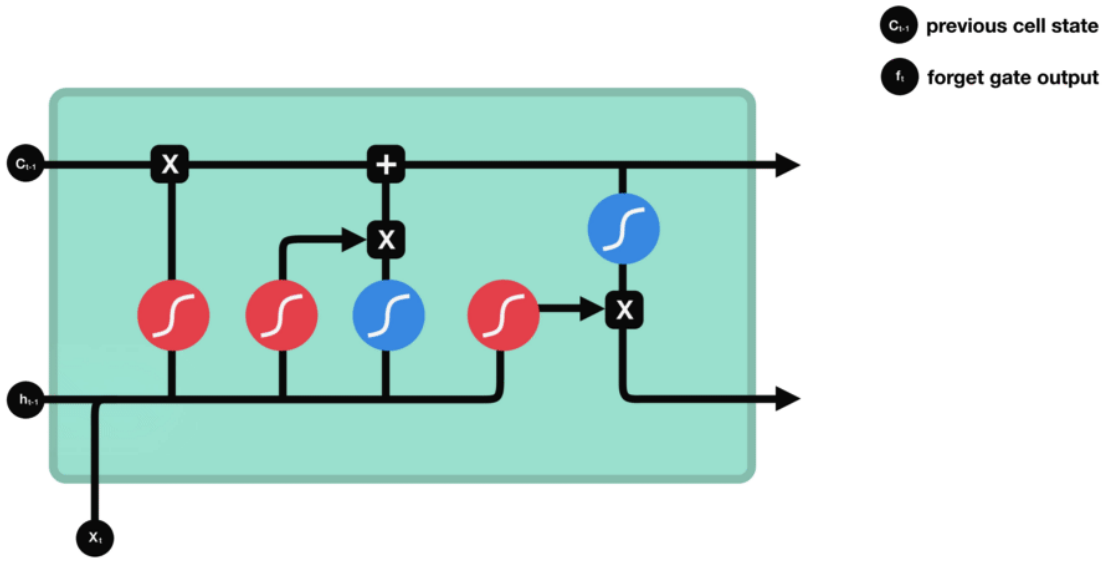


Sigmoid aplasta los valores para que estén entre 0 y 1

Profundicemos un poco más en lo que están haciendo las distintas puertas, ¿de acuerdo? Así que tenemos tres puertas diferentes que regulan el flujo de información en una celda LSTM. Una puerta de olvido, una puerta de entrada y una puerta de salida.

Puerta de olvido

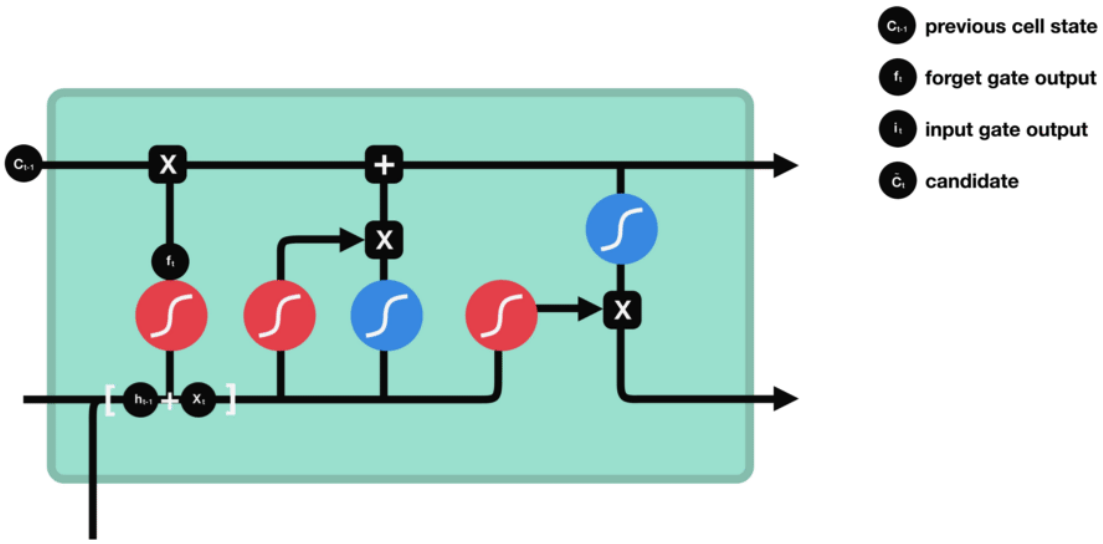
Primero, tenemos la puerta del olvido. Esta puerta decide qué información debe desecharse o conservarse. La información del estado oculto anterior y la información de la entrada actual se pasan a través de la función sigmoide. Los valores salen entre 0 y 1. Cuanto más cerca de 0 significa olvidar, y más cerca de 1 significa mantener.



Operaciones de la puerta de olvido

Puerta de entrada

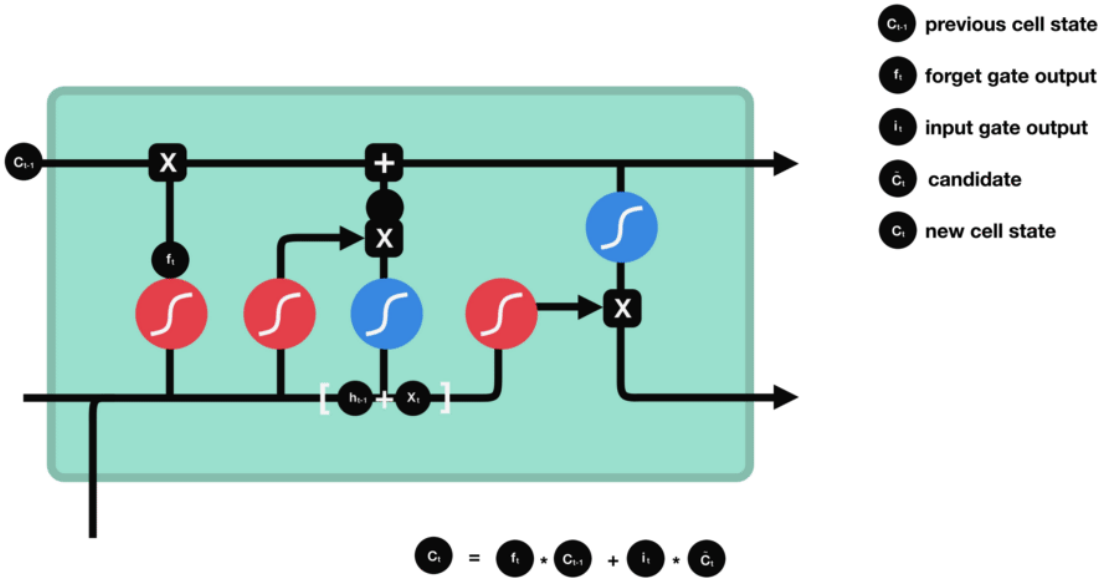
Para actualizar el estado de la celda, tenemos la puerta de entrada. Primero, pasamos el estado oculto anterior y la entrada actual a una función sigmoide. Eso decide qué valores se actualizarán transformando los valores para que estén entre 0 y 1. 0 significa no importante y 1 significa importante. También pasa el estado oculto y la entrada de corriente a la función tanh para aplastar valores entre -1 y 1 para ayudar a regular la red. Luego multiplicas la salida tanh por la salida sigmoidea. La salida sigmoide decidirá qué información es importante mantener de la salida tanh.



Operaciones de puerta de entrada

Estado de la celda

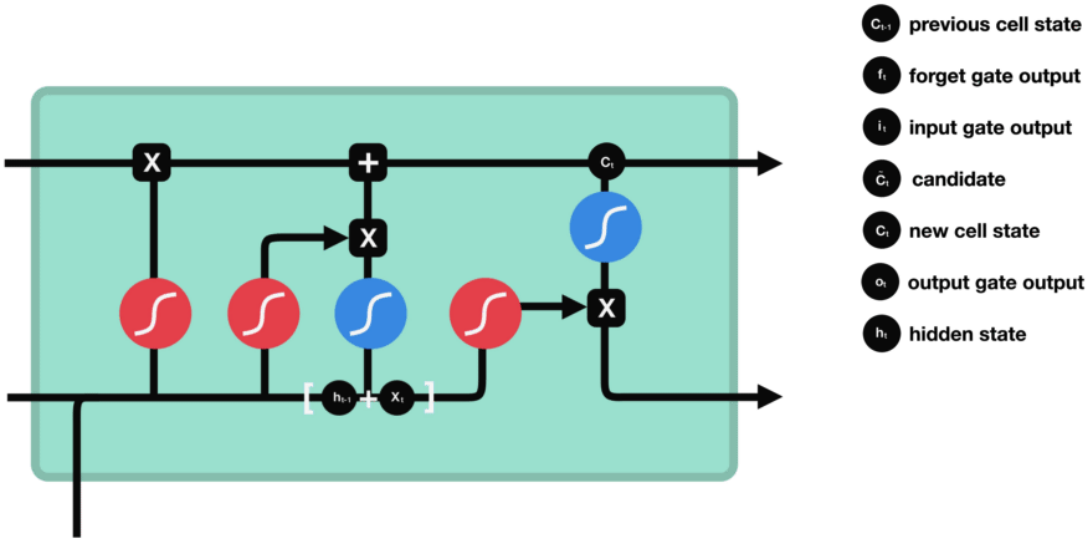
Ahora deberíamos tener suficiente información para calcular el estado de la celda. Primero, el estado de la celda se multiplica puntualmente por el vector de olvido. Esto tiene la posibilidad de eliminar valores en el estado de la celda si se multiplica por valores cercanos a 0. Luego tomamos la salida de la puerta de entrada y hacemos una suma puntual que actualiza el estado de la celda a nuevos valores que la red neuronal considera relevantes. Eso nos da nuestro nuevo estado celular.



Cálculo del estado de la celda

Puerta de salida

Por último, tenemos la puerta de salida. La puerta de salida decide cuál debe ser el siguiente estado oculto. Recuerde que el estado oculto contiene información sobre entradas anteriores. El estado oculto también se utiliza para predicciones. Primero, pasamos el estado oculto anterior y la entrada actual a una función sigmoide. Luego pasamos el estado celular recién modificado a la función tanh. Multiplicamos la salida tanh con la salida sigmoide para decidir qué información debe llevar el estado oculto. La salida es el estado oculto. El nuevo estado de celda y el nuevo oculto se transfieren al siguiente paso de tiempo.

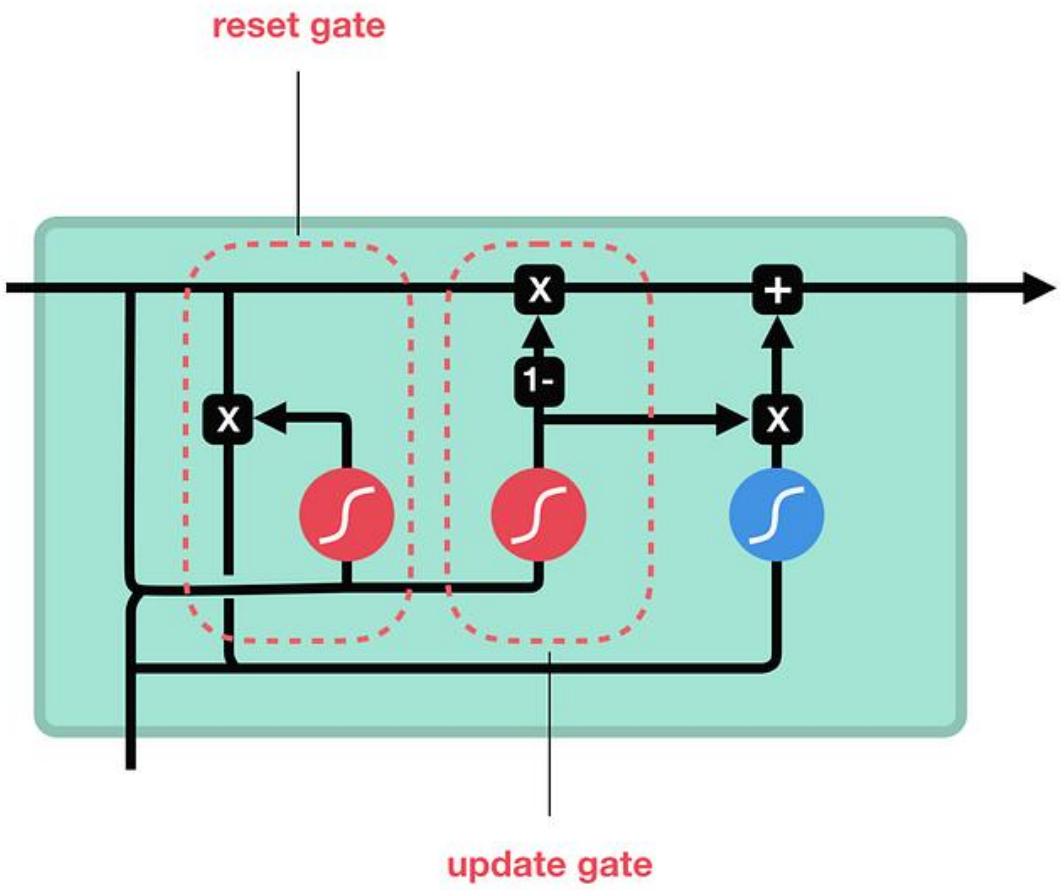


Operaciones de puerta de salida

Para revisar, la puerta Olvidar decide qué es relevante mantener en los pasos anteriores. La puerta de entrada decide qué información es relevante agregar desde el paso actual. La puerta de salida determina cuál debe ser el siguiente estado oculto.

GRU

Así que ahora que sabemos cómo funciona un LSTM, veamos brevemente el GRU. El GRU es la nueva generación de redes neuronales recurrentes y es bastante similar a un LSTM. Los GRU se deshicieron del estado de la celda y usaron el estado oculto para transferir información. También solo tiene dos puertas, una puerta de reinicio y una puerta de actualización.



Célula GRU y sus puertas

Puerta de actualización

La puerta de actualización actúa de manera similar a la puerta de entrada y olvido de un LSTM. Decide qué información desechar y qué nueva información agregar.

Puerta de reinicio

La puerta de reinicio es otra puerta que se utiliza para decidir cuánta información pasada olvidar.

Y eso es un GRU. El GRU tiene menos operaciones tensoriales; por lo tanto, son un poco más rápidos de entrenar que los LSTM. No hay un ganador claro sobre cuál es mejor. Los investigadores e ingenieros generalmente intentan ambos para determinar cuál funciona mejor para su caso de uso.

En resumen, las RNN son buenas para procesar datos de secuencia para predicciones, pero sufren de memoria a corto plazo. LSTM y GRU se crearon como un método para mitigar la memoria a corto plazo utilizando mecanismos llamados puertas. Las puertas son solo redes neuronales que regulan el flujo de información que fluye a través de la cadena de secuencias. LSTM y GRU se utilizan en aplicaciones de aprendizaje profundo de última generación como reconocimiento de voz, síntesis de voz, comprensión del lenguaje natural, etc.