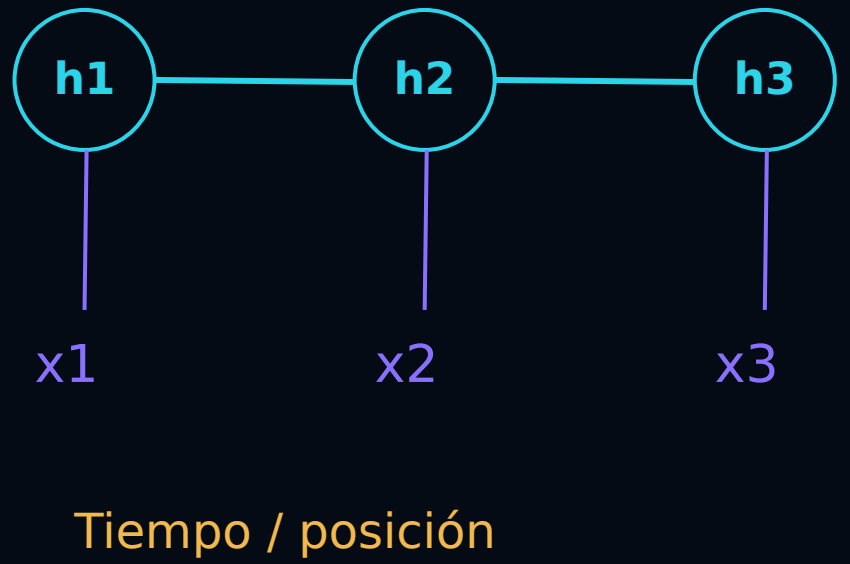
Secuencias, RNN y LSTM

Ampliación de Ciencias de Datos II · UCA

Memoria, gradientes y predicción temporal

Sergio Gevatschnaider



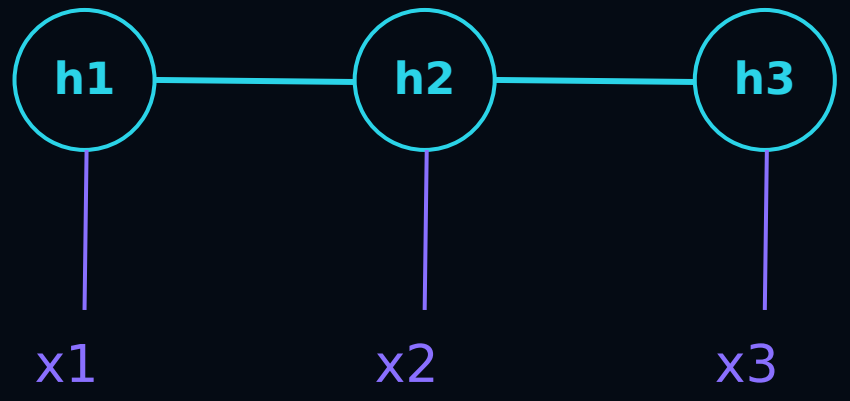
El orden puede contener información

Tabular: atributos de una unidad

Secuencial: observaciones indexadas por tiempo o posición

Ejemplo: misma venta actual, historial creciente o decreciente

Mismos pesos en cada paso



Tiempo / posición

¿Podemos modelar una serie con un árbol?

Ventanas para predecir demanda

$$X_t = [y_{t-3}, y_{t-2}, y_{t-1}, y_t]$$

Entrada t−3	Entrada t−2	Entrada t−1	Entrada t−0	Objetivo t+1
10	**12**	**11**	**14**	**15**

Ventana observada: [10, 12, 11, 14]

El objetivo 15 queda fuera de las entradas.

¿La venta de t+1 puede estar en X_t?

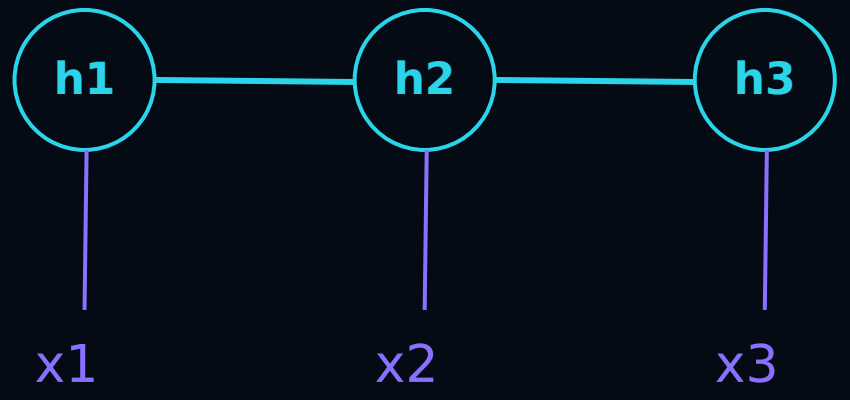
Formas de entrada y salida

Muchos a uno: pronóstico de la siguiente semana

Muchos a muchos: etiquetado por paso

Máscaras permiten longitudes diferentes en un batch

Mismos pesos en cada paso

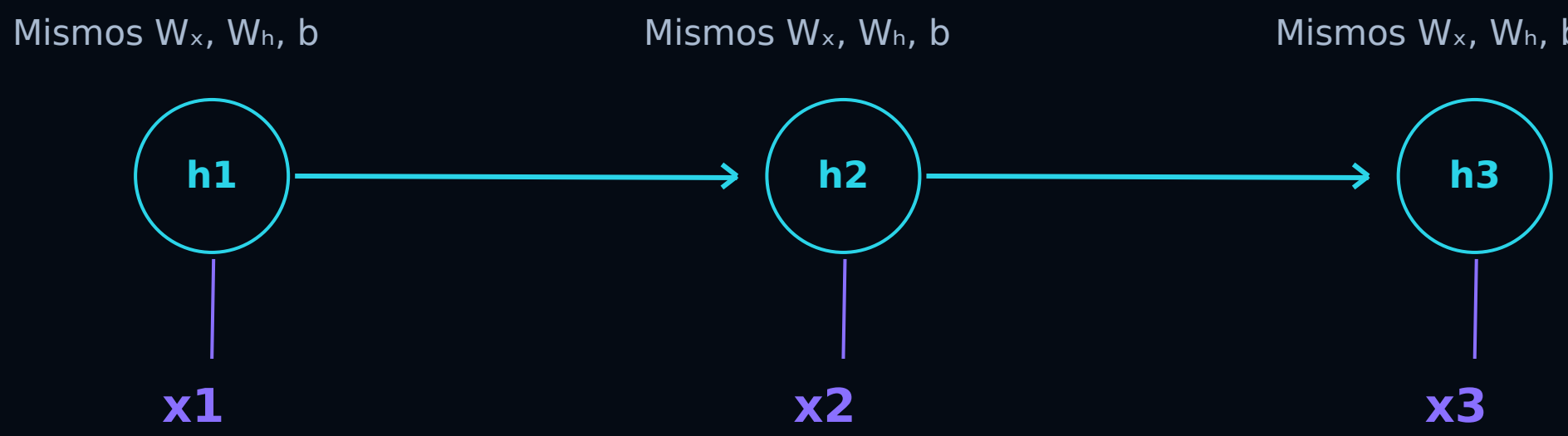


Tiempo / posición

¿Qué representa la dimensión temporal de una ventana?

La recurrencia transporta un estado

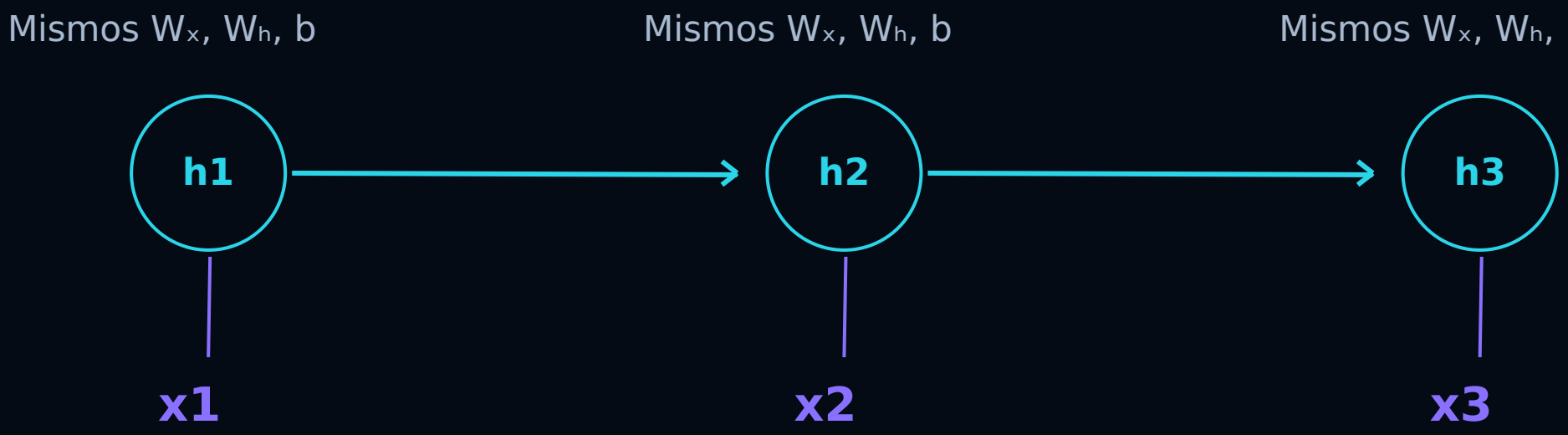
$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b)$$



$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b)$$

¿Aumentar de 10 a 20 pasos duplica parámetros?

RNN desplegada en el tiempo



$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b)$$

¿Qué pasaría si cada paso tuviera pesos distintos?

Parámetros de una RNN

$P_{oculta} = md + m^2 + m$
$d=2, m=3: 6+9+3=18$
$P_{salida}=3+1$, total=22

d atributos, m unidades ocultas y una salida

W_x es $m \times d$ y W_h es $m \times m$

Usamos un único sesgo oculto en esta convención

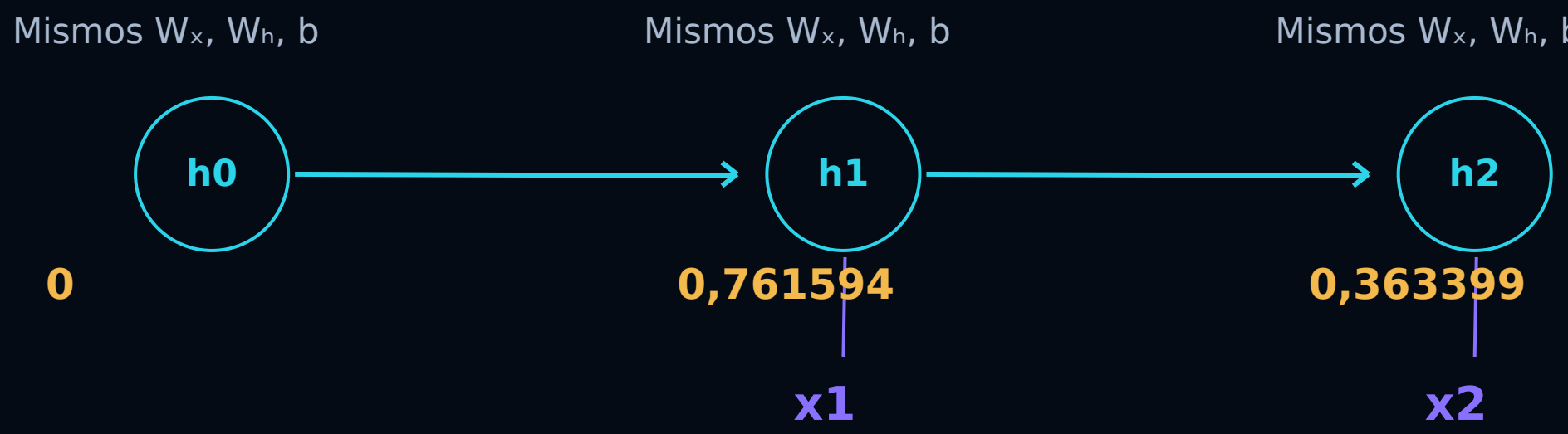
22 parámetros

Valores didácticos verificables

¿Por qué aparece m^2?

Forward de una RNN escalar

$h_1=\tanh(1)=0{,}761594$



$h_0=0 \cdot W_x=1 \cdot W_h=0{,}5 \cdot b=0 \cdot x_1=1 \cdot x_2=0 \quad \Rightarrow \quad h_2=0{,}363399$

¿Cuánto vale h_2 si $W_h=0$?

Pérdida sobre una secuencia

$$L = \Sigma_t \, \ell(\hat{y}_t, y_t)$$

Puede evaluarse en todos los pasos observados

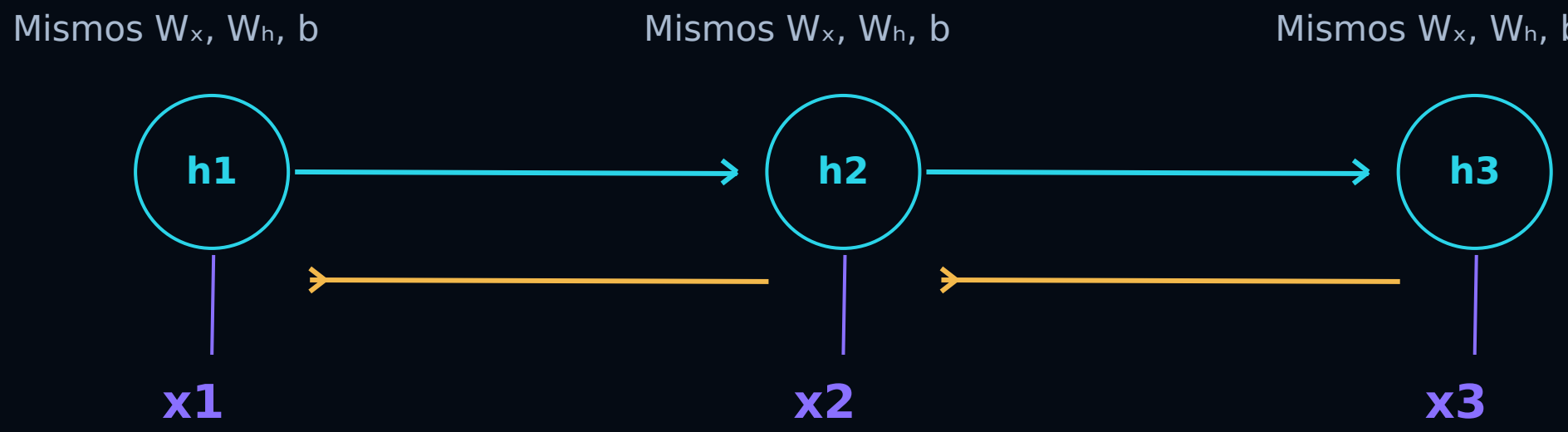
También puede usarse solo la salida final

El enmascarado evita puntuar padding

¿Dos pérdidas con igual mínimo pueden generar pasos distintos?

BPTT aplica la regla de la cadena

$\partial L/\partial W_h = \Sigma_t$ contribución del paso t



El gradiente acumula contribuciones de todos los pasos sobre los pesos compartidos.

¿BPTT necesita un optimizador especial?

Productos temporales de Jacobianos

$$J_t = \mathrm{diag}(1 - h_t^2) W_h$$
$$\partial h_t / \partial h_{t-k} = J_t J_{t-1} \dots J_{t-k+1}$$

Cada paso multiplica por una matriz local

Sus efectos pueden contraerse o amplificarse

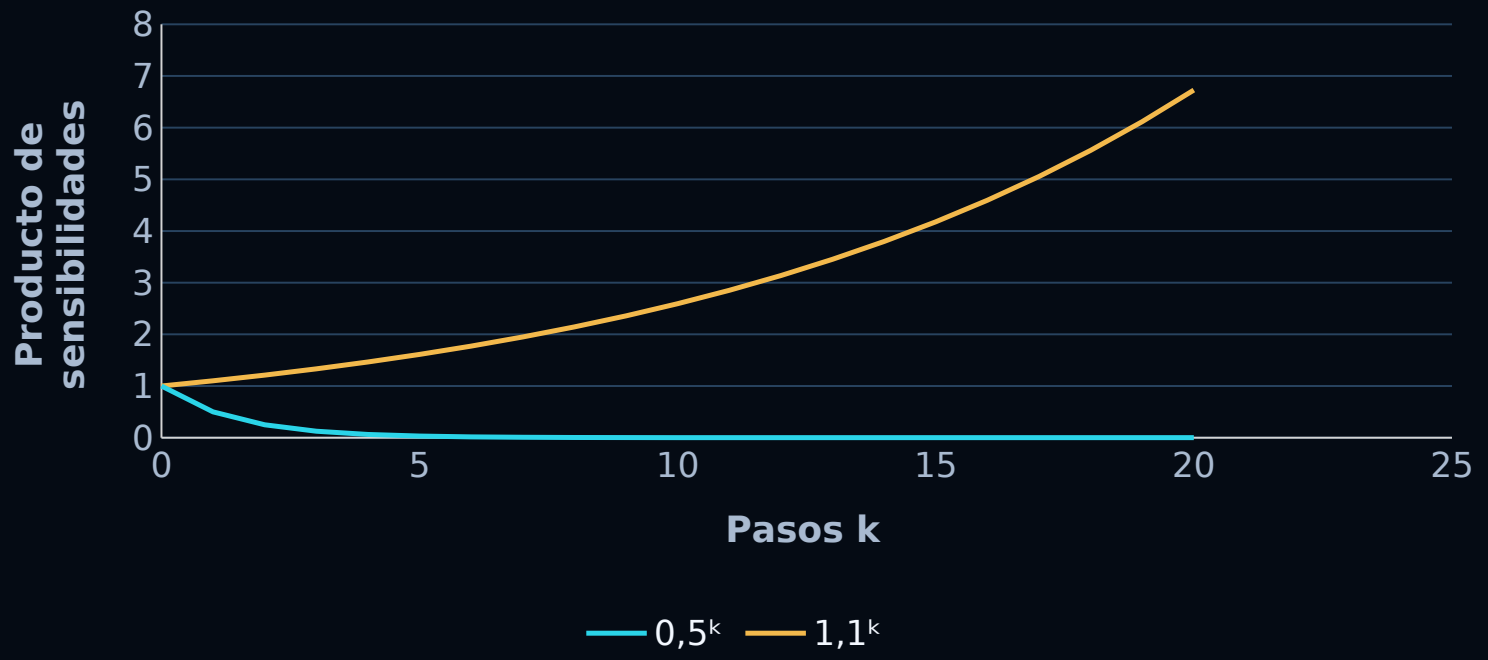
Importan activación, pesos y trayectoria



¿Por qué tanh saturada puede dificultar el aprendizaje?

Vanishing y exploding gradients

$0{,}5^{20} \approx 0{,}000000954$



El producto importa

Factores menores que 1 pueden extinguir la señal.

Factores mayores que 1 pueden amplificarla.

¿Una dependencia larga exige siempre recordar literalmente cada dato?

Clipping y BPTT truncada

$g_{clip} = g \cdot \min(1, \tau / \|g\|)$

Clipping limita la norma de gradientes grandes

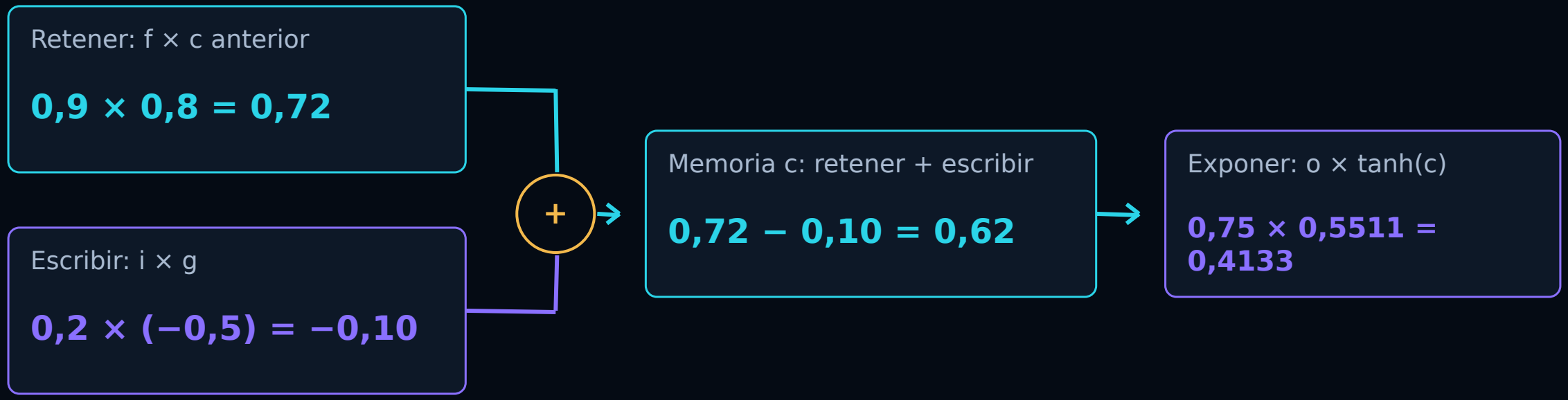
Truncar reduce el horizonte del grafo de derivadas

Mantener estado no equivale a mantener todos los gradientes

¿Truncar a 20 pasos permite atribuir directamente una pérdida a 100 pasos atrás?

LSTM agrega memoria controlada

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

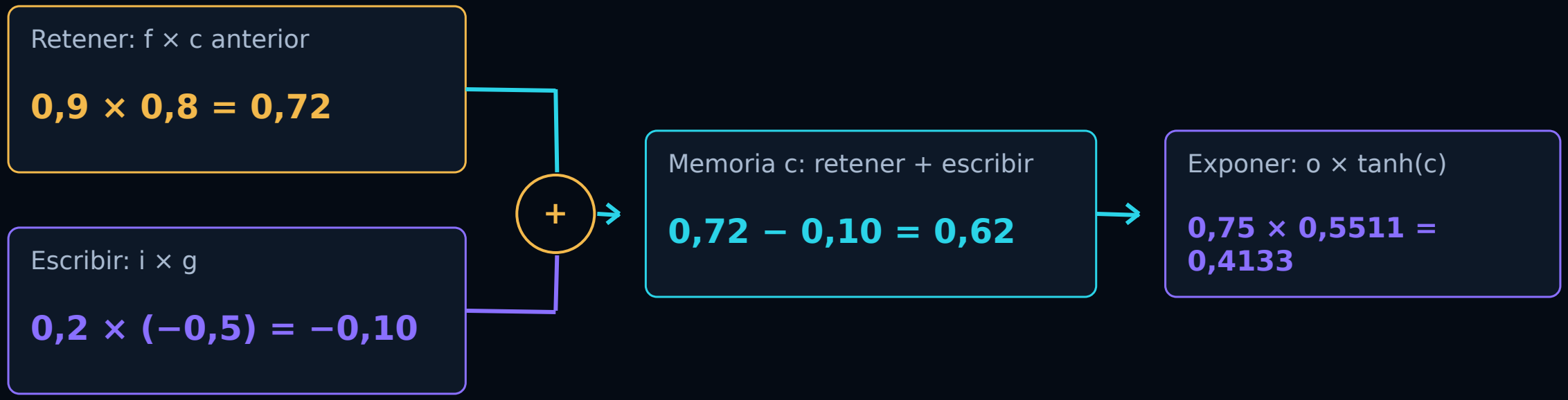


La memoria c circula entre pasos. El estado h comunica una parte transformada.

¿Por qué se necesitan c y h?

Compuerta de olvido

$u_t=[h_{t-1};x_t]$

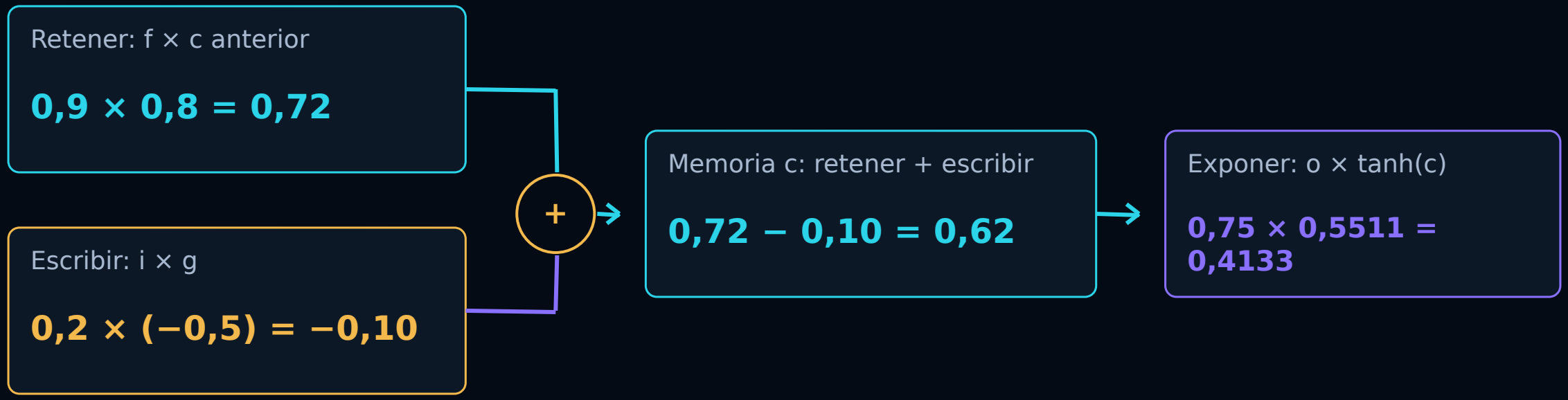


f decide qué fracción del estado anterior conserva la celda.

¿Qué implica $f_t=0$ en una componente?

Entrada y contenido candidato

$$i_t = \sigma(W_i u_t + b_i)$$

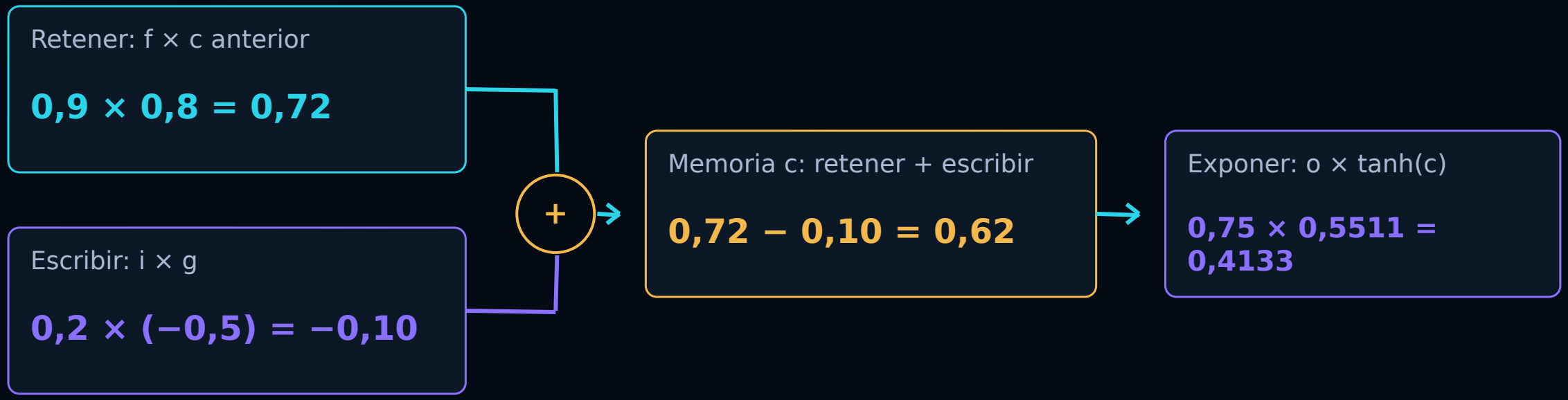


i regula la escritura. g aporta un contenido que también puede ser negativo.

¿Abrir i siempre aumenta c?

Actualización aditiva de la celda

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

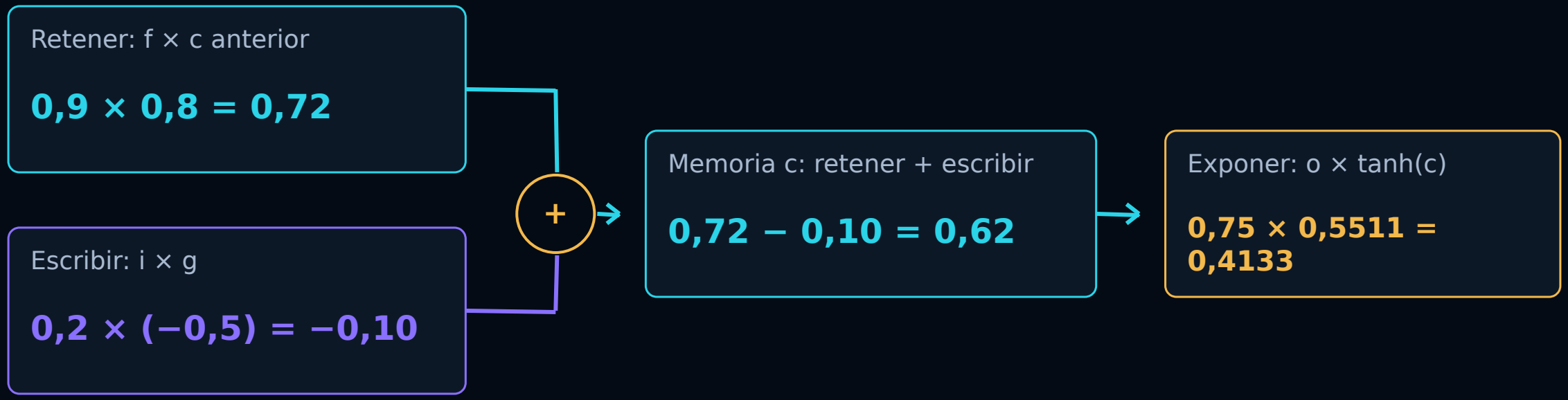


Conservar y escribir se suman en una vía aditiva.

¿Puede c=2 aunque g esté entre −1 y 1?

Compuerta de salida y estado oculto

$$o_t = \sigma(W_o u_t + b_o)$$



o regula el estado expuesto h. Cerrar o no borra c.

¿o=0 obliga a c=0?

LSTM: ecuaciones completas

$$f=\sigma(W_fu+b_f) \quad i=\sigma(W_iu+b_i)$$

$$f_t = \sigma(Wf\,[h_{t-1},x_t]+bf) \qquad i_t = \sigma(Wi\,[h_{t-1},x_t]+bi)$$

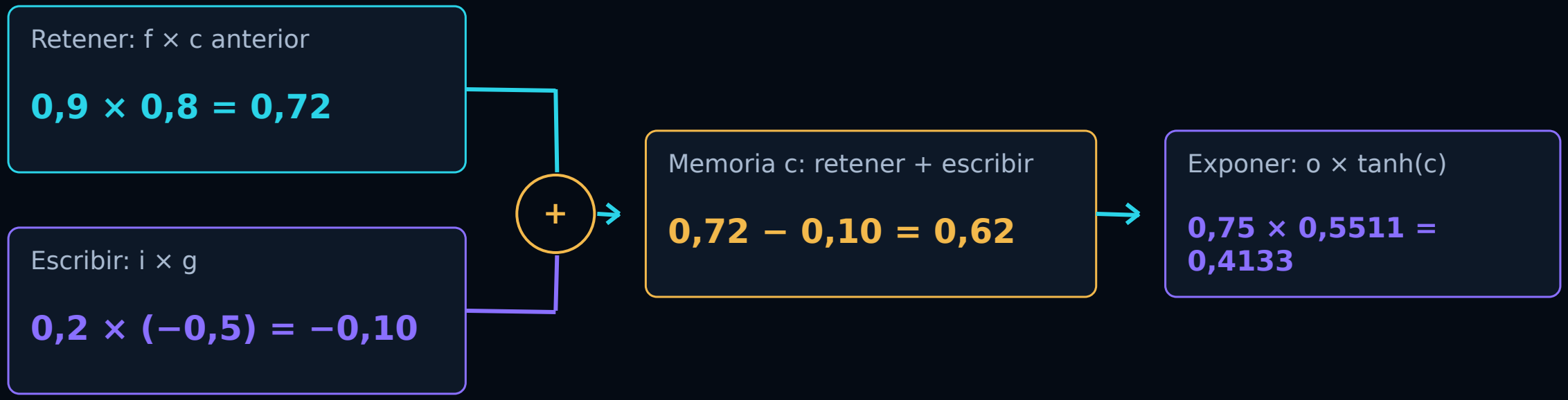
$$g_t = \tanh(Wg\,[h_{t-1},x_t]+bg) \qquad o_t = \sigma(Wo\,[h_{t-1},x_t]+bo)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \qquad h_t = o_t \odot \tanh(c_t)$$

¿Cuáles son los parámetros entrenables?

Ejemplo: conservar y escribir

$c_t = 0{,}9 \cdot 0{,}8 + 0{,}2 \cdot (-0{,}5) = 0{,}62$

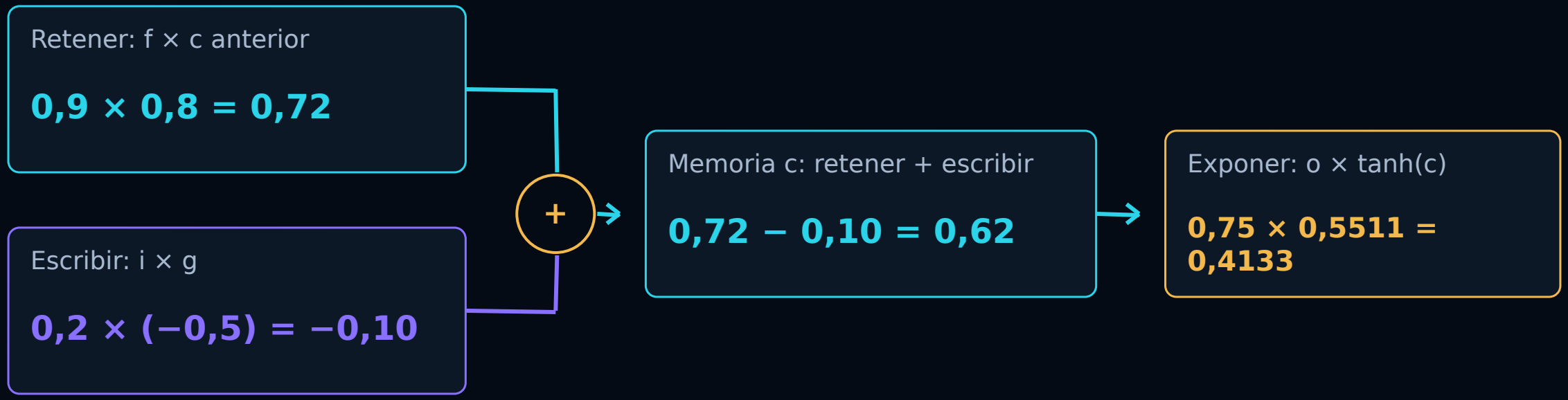


Memoria: 0,72 retenido + (−0,10) escrito = 0,62.

¿Cuánto vale c_t si i=0?

Ejemplo: exponer la memoria

$h_t = 0{,}75 \cdot \tanh(0{,}62) \approx 0{,}413346$

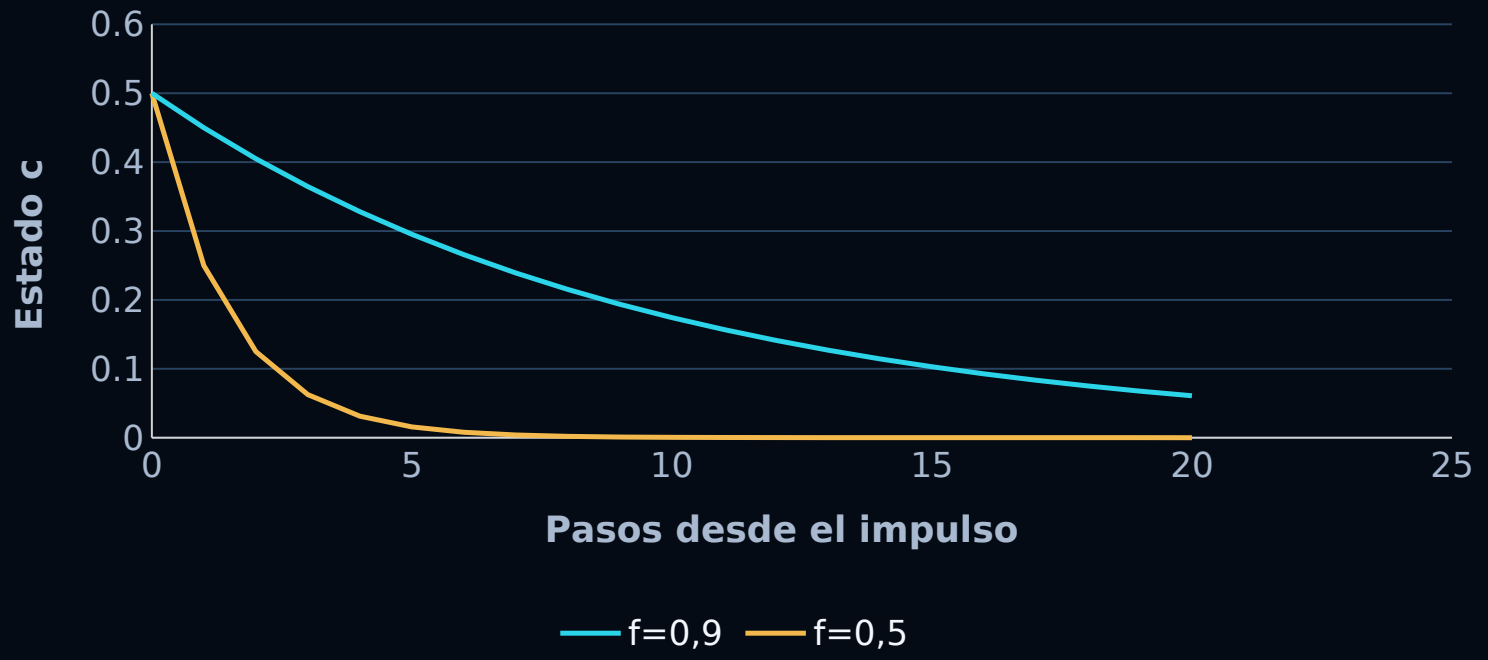


La memoria c=0,62 permanece mientras h expone solo 0,413346.

¿Qué memoria ve la capa de salida si o baja?

Persistencia con gates constantes

$c_t = 0{,}5 \cdot f^{t-1}$



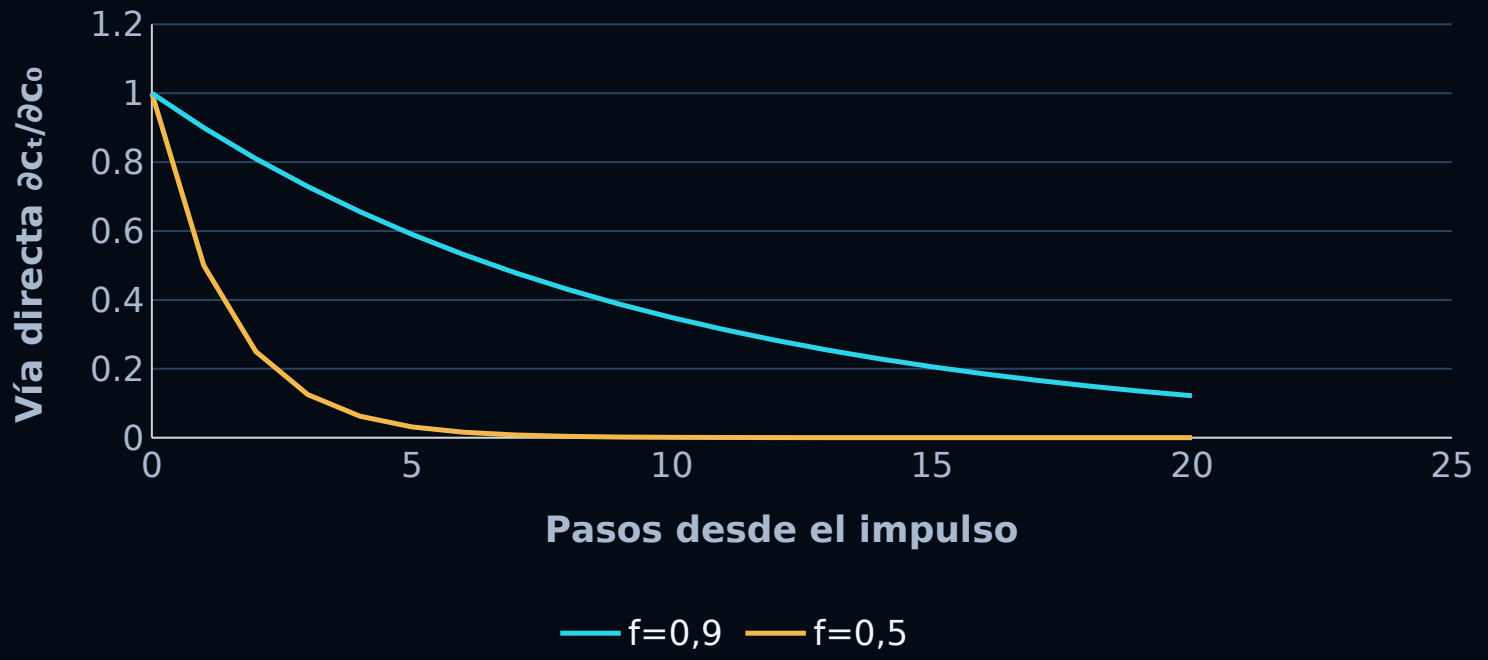
Sin nueva escritura

c_0=0,5, i=0
La retención controla la escala temporal.

¿Qué sucede si f=1 y no llega más contenido?

Qué dice el gradiente directo

$\partial c_t/\partial c_{t-1}$ (gates fijos)$=f_t$



Producto directo de f

Gates tratados constantes.
La derivada total también incluye sus dependencias.

¿LSTM elimina todo vanishing gradient?

Laboratorio: retener, alternar, sobrescribir

Impulso: observar memoria sin nueva escritura

Alternancia: contenido con cambios de signo

Sobrescritura: nueva señal tras varios pasos

Antes de simular

1. Predecir
2. Observar
3. Explicar

¿Qué diferencia observás al cerrar o en vez de cerrar f?

Conteo de parámetros LSTM

P_celda = 4m(d+m+1)
d=2,m=3: P=4·3·6=72

Cuatro transformaciones sobre d+m entradas

m salidas por transformación

Más parámetros permiten control, pero aumentan costo

72 parámetros

Valores didácticos verificables

¿La longitud de la secuencia aparece en P?

GRU: una alternativa con menos compuertas

$$z=\sigma(W_{zx}+U_{zh_{prev}}+b_z)$$
$$r=\sigma(W_{rx}+U_{rh_{prev}}+b_r)$$
$$\tilde{h}=\tanh(Wx+U(r\odot h_{prev})+b)$$
$$h=(1-z)\odot h_{prev}+z\odot\tilde{h}$$

Combina memoria y estado oculto

Usa gates de actualización y reinicio

La convención de z debe quedar explícita

¿Qué conserva el pasado en esta convención?

RNN, LSTM/GRU y Transformer

RNN: estado recurrente y transición simple

LSTM/GRU: gates y recurrencia secuencial

Transformer: atención con información de posición

RNN
Estado recurrente

LSTM / GRU
Gates y recurrencia

Transformer
Atención y posición

¿La atención recibe el orden temporal sin posición ni máscara?

Costo y paralelización dependen del uso

T: longitud de secuencia
m: dimensión del estado

RNN: $O(Tm^2)$ para recurrencia $d \approx m$

Atención densa: $O(T^2m)$ más proyecciones

Inferencia autoregresiva conserva pasos secuenciales

Comparación con evidencia

RNN: $O(Tm^2)$ para recurrencia $d \approx m$

Atención densa: $O(T^2m)$ más proyecciones

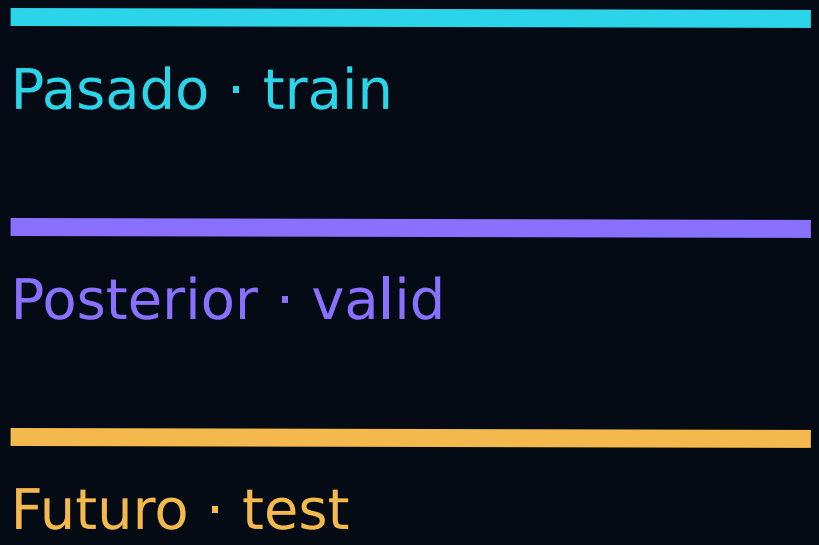
¿Una menor complejidad asintótica garantiza menos latencia para T pequeño?

Partición cronológica para demanda

Entrenar con el pasado disponible

Elegir configuración en un período posterior

Reservar el último bloque como test final



¿Una ventana test puede contener valores de train?

Walk-forward

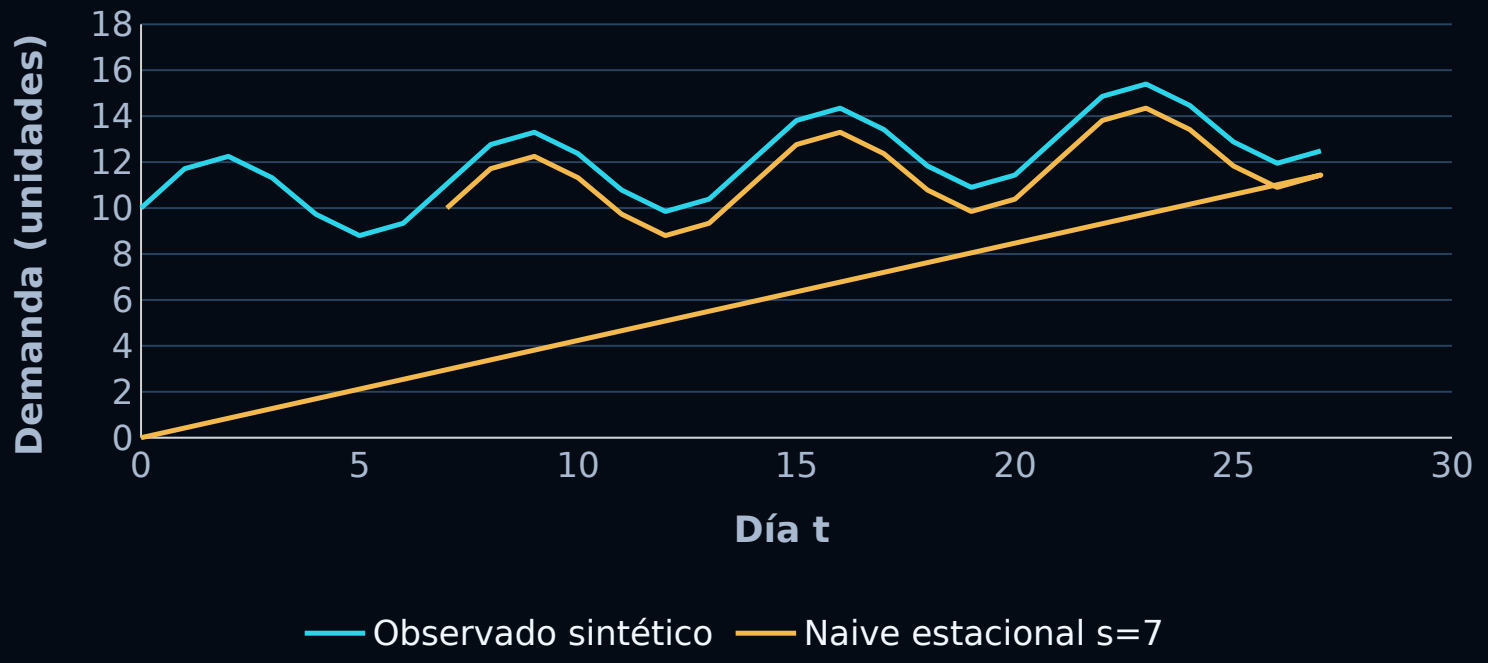


Historia creciente (cian) y horizonte de validación de 14 pasos (ámbar).

¿Por qué un split aleatorio puede sobreestimar calidad temporal?

Baseline ingenuo y estacional

Naive: $\hat{y}_{t+1} = y_t$



$$\hat{y}_t = y_{t-7}$$

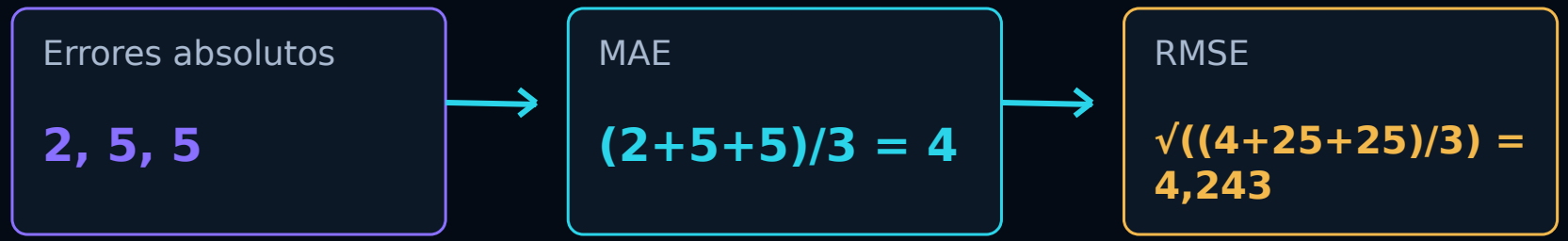
Usa el mismo día de la semana previa.

Una tendencia genera errores sistemáticos.

¿Qué baseline exigirías para demanda diaria con patrón semanal?

MAE, RMSE y límites de MAPE

MAE=mean(|e|) RMSE=√mean(e²)



MAE y RMSE mantienen la unidad del objetivo. RMSE amplifica errores grandes.

MAPE queda indefinido cuando un valor real es cero.

¿Cómo reportarías MAPE si hay y=0?

Práctica reproducible en Python

Crear ventanas y separar por fecha

Ajustar baseline estacional, modelo con rezagos y LSTM

Medir MAE/RMSE sobre idénticos orígenes

Defensa con evidencia

1. Protocolo
2. Resultado
3. Límites

¿Qué entregar además del score?

Defensa final y respuestas

Si f=1 e i=0, c se conserva

Clipping limita explosión, sin recuperar gradientes nulos

Una LSTM debe superar un baseline con evidencia temporal

Defensa con evidencia

1. Protocolo
2. Resultado
3. Límites

¿Adoptarías LSTM si la mejora media es pequeña y varía de signo por período?