Día 10 · Tema 3 — Ejercicios: Mamba, MoE y alternativas al Transformer

Cómo se usa esta hoja. Hay dos tipos de ejercicio, y sólo dos.

El ejercicio del día. Uno por hoja. Es el obligatorio y es el que entra en el examen, con el cálculo incluido. Basta con hacer ese para seguir el curso con solvencia.

Los ejercicios de ampliación. El resto. Están para enseñar de qué es capaz el método, no para exigir: **no se piden calculados en el examen**. De ellos sólo puede preguntarse la idea, a nivel conceptual.

EL EJERCICIO DEL DÍA *obligatorio · entra en el examen*

1. Paso forward de un SSM (caso escalar, $N = 1$)

Consideremos un modelo de espacio de estados discretizado en el caso escalar ($N = 1$):

$$h_t = \bar{a}\, h_{t-1} + \bar{b}\, x_t, \qquad y_t = c\, h_t,$$

con parámetros $\bar{a} = 0{,}9$, $\bar{b} = 0{,}5$, $c = 1{,}0$ y estado inicial $h_0 = 0$.

Dada la secuencia de entrada $x_1 = 2$, $x_2 = 1$, $x_3 = 3$:

a) Calcula h_1, h_2 y h_3.
b) Calcula y_1, y_2 e y_3.
c) Expresa h_3 como combinación lineal de x_1, x_2 y x_3 e identifica los coeficientes. ¿Qué papel juega $\bar{a}$ en la memoria del sistema?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

2. SSM frente a RNN clásica: memoria efectiva

Ahora comparamos el SSM del ejercicio anterior con una RNN clásica de un solo estado:

$$h_t^{\text{RNN}} = \tanh\!\left(w_h\, h_{t-1}^{\text{RNN}} + w_x\, x_t\right),$$

con $w_h = 0{,}9$, $w_x = 0{,}5$ y $h_0^{\text{RNN}} = 0$. Usamos la misma entrada: $x_1 = 2$, $x_2 = 1$, $x_3 = 3$.

a) Calcula h_1^{RNN}, h_2^{RNN} y h_3^{RNN} paso a paso (escribe la pre-activación antes de aplicar tanh). Valores útiles: $\tanh(1{,}0) \approx 0{,}7616$, $\tanh(1{,}1854) \approx 0{,}8292$, $\tanh(2{,}2462) \approx 0{,}9779$.
b) Para medir la «memoria efectiva», calcula la sensibilidad $\partial h_3 / \partial x_1$ en cada modelo.
   - En el SSM lineal, usa la expresión analítica directa.
   - En la RNN, aplica la regla de la cadena a través de tanh: $\frac{\mathrm{d}}{\mathrm{d}z}\tanh(z) = 1 - \tanh^2(z)$.
c) ¿Cuál de los dos modelos retiene más información de x_1 en el instante $t = 3$? ¿Por qué?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

3. Memoria de la matriz de atención

En la atención estándar de un Transformer, la matriz $QK^\top$ tiene tamaño $T \times T$.

a) Para $T = 4\,096$ y con almacenamiento en FP32 (4 bytes por elemento), ¿cuánta memoria ocupa la matriz $QK^\top$? Expresa el resultado en MB.
b) Repite el cálculo para $T = 65\,536$ y $T = 1\,048\,576$.
c) Si la GPU disponible es una A6000 con 48 GB de VRAM, ¿a partir de qué longitud de secuencia T la matriz de atención por sí sola supera la memoria disponible? Plantea la inecuación y resuélvela.
d) Reflexiona: la estimación anterior solo cuenta *una* cabeza de *una* capa. ¿Cómo cambia la conclusión con $H = 32$ cabezas y $L = 32$ capas? (Se puede compartir memoria entre cabezas; explica cómo.)

AMPLIACIÓN *opcional · no se pide calculado en el examen*

4. Enrutamiento en una capa MoE con top-r

Una capa Mixture-of-Experts tiene $K = 4$ expertos y usa enrutamiento top-r con $r = 2$. El router produce logits $\mathbf{z} = W_g\,\mathbf{x} = (2{,}0;\ 0{,}5;\ 1{,}0;\ -0{,}5)$.

a) Calcula las probabilidades completas $\mathrm{softmax}(\mathbf{z})$. Valores útiles: $e^{2,0} \approx 7{,}389$, $e^{0,5} \approx 1{,}649$, $e^{1,0} \approx 2{,}718$, $e^{-0,5} \approx 0{,}607$.
b) Identifica los dos expertos con mayor probabilidad.
c) Renormaliza las probabilidades de los dos expertos seleccionados para que sumen 1. *Pista*: observa que la ratio de las dos probabilidades seleccionadas es $e^2/e^1 = e$. Expresa los pesos en función de $\sigma(1)$ (la función sigmoide evaluada en 1).
d) Si las salidas de los expertos son $E_1(\mathbf{x}) = (1, 0)$, $E_2(\mathbf{x}) = (0, 1)$, $E_3(\mathbf{x}) = (1, 1)$, $E_4(\mathbf{x}) = (-1, 0)$, calcula $\mathrm{MoE}(\mathbf{x})$.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

5. Aritmética de parámetros en una capa MoE

Consideramos un modelo Transformer con dimensión $d = 4\,096$. Cada bloque MLP expande la dimensión a $4d = 16\,384$, por lo que un MLP denso tiene dos matrices: $W_{\text{up}} \in \mathbb{R}^{d\times 4d}$ y $W_{\text{down}} \in \mathbb{R}^{4d\times d}$ (ignoramos sesgos).

Reemplazamos cada capa MLP por una capa MoE con $K = 8$ expertos y $r = 2$.

a) ¿Cuántos parámetros tiene un solo experto MLP? ¿Y la capa MoE completa (incluyendo el router $W_g \in \mathbb{R}^{d\times K}$)?
b) ¿Cuántos parámetros se activan *por token* (los $r = 2$ expertos elegidos más el router)?
c) ¿Cuál es la ratio parámetros totales / parámetros activos?
d) Si almacenamos todos los expertos en FP16 (2 bytes por parámetro), ¿cuánta VRAM consume esta capa MoE? ¿Y una capa MLP densa con el mismo número de parámetros activos? Comenta la implicación para el despliegue en producción.
e) Aplica la misma lógica a Mixtral 8×7B ($K = 8$, $r = 2$, 56B totales, 14B activos) y DeepSeek-V3 ($K = 256$, $r = 8$, 671B totales, 37B activos). ¿Cuántas GPUs A100 de 80 GB necesita cada uno en FP16 solo para alojar los pesos?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

6. Atención estándar frente a atención lineal

Sean las matrices:

$$Q = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad K = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix},$$

donde $Q \in \mathbb{R}^{2\times 2}$, $K, V \in \mathbb{R}^{3\times 2}$ y $d_k = 2$.

a) **Atención estándar.** Calcula $\mathrm{Att}(Q, K, V) = \mathrm{softmax}\big(QK^\top/\sqrt{d_k}\big)\,V$ paso a paso: primero $QK^\top$, después el escalado por $1/\sqrt{2} \approx 0{,}7071$, después el softmax por filas, y finalmente la multiplicación por V. Valores útiles: $e^{0,707} \approx 2{,}028$, $e^{1,414} \approx 4{,}113$, $e^{2,121} \approx 8{,}342$.
b) **Atención lineal.** Con φ = identidad (sin función de características), calcula $\mathrm{Att}_{\text{lin}}(Q, K, V) = Q\,(K^\top V)$. Nótese el cambio en el orden de multiplicación.
c) Compara numéricamente las dos salidas. ¿Por qué difieren? ¿Qué propiedad de softmax se pierde al eliminarla?
d) **Coste computacional.** Si las dimensiones generales son $Q \in \mathbb{R}^{T\times d}$, $K, V \in \mathbb{R}^{T\times d}$: ¿cuántas operaciones requiere cada método en función de T y d? ¿Para qué relación entre T y d es ventajosa la atención lineal?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

7. Tabla comparativa de costes: atención, SSM y atención lineal

Para una secuencia de longitud $T = 10\,000$, dimensión del modelo $d = 512$ y tamaño de estado SSM $N = 16$, rellena la siguiente tabla calculando los FLOPs (operaciones de punto flotante) de cada mecanismo:

Mecanismo	Complejidad	FLOPs	FLOPs (not. científica)
Atención densa	$O(T^2 d)$		
SSM (Mamba)	$O(TNd)$		
Atención lineal	$O(Td^2)$		

a) Completa la columna de FLOPs sustituyendo los valores numéricos.
b) ¿Cuántas veces más rápido es el SSM respecto a la atención densa? ¿Y la atención lineal?
c) ¿Para qué valor de T (con $d = 512$ y $N = 16$ fijos) la atención densa y la atención lineal cuestan lo mismo?
d) Si $d = 512$ y $N = 16$, ¿existe algún T para el que la atención densa sea más eficiente que el SSM? Justifica.