# Día 09 · Tema 3 — Ejercicios: self-attention y 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.

Estos ejercicios son de cálculo analítico (lápiz y papel). Todos los resultados numéricos deben obtenerse paso a paso; se proporcionan los valores de referencia de las exponenciales necesarias para que el cálculo sea factible sin calculadora científica.

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

### 1. Self-attention desde cero (3 tokens, $d$=2)

Considera una secuencia de $T$=3 tokens con embeddings de dimensión $d$=2, recogidos en la matriz

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix},$$

donde cada fila es un token. Las matrices de proyección son

$$W^Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad W^K = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad W^V = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(a) Calcula las proyecciones $Q = XW^Q$, $K = XW^K$, $V = XW^V$.
(b) Calcula la matriz de puntuaciones $S = QK^\top/\sqrt{d_k}$ con $d_k = 2$.
(c) Aplica softmax por filas para obtener la matriz de atención $A$. Utiliza los siguientes valores de referencia:

$$e^0 = 1, \qquad \alpha = e^{1/\sqrt{2}} \approx 2{,}0281, \qquad \alpha^2 = e^{\sqrt{2}} \approx 4{,}1133.$$

(d) Calcula la salida $Z' = AV$.
(e) Interpreta los pesos de atención: ¿a qué token presta más atención cada token? ¿Por qué?

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

### 2. Máscara causal

Parte de la matriz de puntuaciones $S$ obtenida en el ejercicio anterior (antes de softmax).

(a) Aplica la máscara causal: sustituye $S_{ij}$ por $-\infty$ para todo $j > i$, de modo que cada token solo pueda atender a sí mismo y a los anteriores. Escribe la matriz $S^{\text{causal}}$.
(b) Calcula la nueva softmax $A^{\text{causal}}$. Recuerda que $e^{-\infty} = 0$.
(c) Calcula $Z'_{\text{causal}} = A^{\text{causal}}V$.
(d) Compara $Z'$ con $Z'_{\text{causal}}$. ¿Qué fila no cambia? ¿Por qué?

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

### 3. Conteo de parámetros: Multi-Head Attention

Considera un modelo con dimensión $d = 512$ y $H = 8$ cabezas de atención. Las proyecciones por cabeza son $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times d_k}$ y la proyección de salida es $W^O \in \mathbb{R}^{d \times d}$. No consideres sesgos (biases).

(a) ¿Cuánto vale $d_k = d/H$?
(b) ¿Cuántos parámetros tiene cada proyección $W_h^Q$ de una cabeza?
(c) ¿Cuántos parámetros suman las tres proyecciones ($W_h^Q$, $W_h^K$, $W_h^V$) de una cabeza?
(d) ¿Cuántos parámetros suman las proyecciones de las $H$ cabezas?
(e) ¿Cuántos parámetros tiene $W^O$?
(f) ¿Cuál es el total de parámetros del bloque MHA? Exprésalo también en función de $d$.

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

**4. Conteo de parámetros: bloque Transformer completo (BERT-base)**

Considera la arquitectura BERT-base: $d = 768$, $H = 12$ cabezas, MLP con expansión $\times 4$ (es decir, $d_{\text{ff}} = 4d = 3072$), y $N = 12$ bloques. Cada bloque consta de:

$$X' = X + \text{MHA}(\text{LN}(X)), \qquad Z = X' + \text{MLP}(\text{LN}(X')),$$

con $\text{MLP} = \text{Linear}(d, 4d) \to \text{GELU} \to \text{Linear}(4d, d)$.

(a) Calcula los parámetros del bloque MHA (sin sesgos): $4d^2$ (tres proyecciones de entrada más $W^O$).
(b) Calcula los parámetros del MLP (sin sesgos): $\text{Linear}(d, 4d)$ y $\text{Linear}(4d, d)$.
(c) Calcula los parámetros de las dos capas LayerNorm. Cada LayerNorm tiene un vector de escala $\gamma$ y uno de desplazamiento $\beta$, ambos de dimensión $d$.
(d) Suma para obtener el total por bloque. Exprésalo como $12d^2 + 4d$.
(e) Multiplica por $N = 12$ bloques. ¿Cuántos millones de parámetros resultan solo en los bloques Transformer (sin contar embeddings)?

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

**5. Positional encoding sinusoidal**

La codificación posicional del Transformer original se define como

$$P_{t,2i} = \sin\!\left(t/10000^{2i/d}\right), \qquad P_{t,2i+1} = \cos\!\left(t/10000^{2i/d}\right).$$

Toma $d = 4$, de modo que las cuatro dimensiones de cada posición son:

$$\underbrace{\sin(t)}_{\text{dim } 0}, \quad \underbrace{\cos(t)}_{\text{dim } 1}, \quad \underbrace{\sin(t/100)}_{\text{dim } 2}, \quad \underbrace{\cos(t/100)}_{\text{dim } 3}.$$

(a) Completa la tabla de codificación posicional para $t = 0, 1, 2, 3$, usando los valores de referencia:

| $t$ | $\sin(t)$ | $\cos(t)$ | $\sin(t/100)$ | $\cos(t/100)$ |
|---|---|---|---|---|
| 0 | 0,0000 | 1,0000 | 0,0000 | 1,0000 |
| 1 | 0,8415 | 0,5403 | 0,0100 | 0,9999 |
| 2 | 0,9093 | −0,4161 | 0,0200 | 0,9998 |
| 3 | 0,1411 | −0,9900 | 0,0300 | 0,9996 |

(b) Dada la matriz de embeddings

$$X = \begin{pmatrix} 0{,}5 & 0{,}3 & -0{,}1 & 0{,}2 \\ 0{,}1 & -0{,}4 & 0{,}6 & 0{,}0 \\ -0{,}2 & 0{,}7 & 0{,}3 & -0{,}5 \\ 0{,}8 & 0{,}1 & -0{,}3 & 0{,}4 \end{pmatrix},$$

calcula $X + P$ (la entrada al primer bloque Transformer).
(c) Observa las columnas 2 y 3 (las dimensiones de baja frecuencia). ¿Qué diferencia hay en su comportamiento respecto de las columnas 0 y 1? ¿Qué implicación tiene esto para posiciones lejanas ($t = 500$ frente a $t = 501$)?

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

**6. Coste computacional de la atención**

Considera una capa de atención con $T$ tokens, dimensión del modelo $d = 512$, $d_k = 64$ y $H = 8$ cabezas.

(a) Para una cabeza, calcula el número de operaciones en coma flotante (FLOPs) del producto $QK^\top$, donde $Q, K \in \mathbb{R}^{T \times d_k}$. Usa la aproximación $\text{FLOPs} \approx 2\,T^2\,d_k$. Evalúa para $T = 1024$.
(b) Calcula la memoria necesaria para almacenar la matriz de atención $A \in \mathbb{R}^{T \times T}$ en FP32 (4 bytes por elemento). ¿Y para las $H = 8$ cabezas simultáneamente? Evalúa para $T = 1024$.
(c) Repite los cálculos para $T = 8192$. ¿Cuál es el factor de incremento en FLOPs y en memoria respecto de $T = 1024$? Justifica el resultado en función de la complejidad $O(T^2)$.

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

**7. Conexiones residuales: preservación de la información**

Considera un token con representación $\mathbf{x} = (1,\ 2)$ a la entrada de un bloque Transformer. Supón que los sub-bloques producen:

$$\mathrm{MHA}(\mathrm{LN}(\mathbf{x})) = (0{,}3,\ -0{,}1), \qquad \mathrm{MLP}(\mathrm{LN}(\mathbf{x}')) = (0{,}5,\ 0{,}2).$$

(a) Calcula $\mathbf{x}' = \mathbf{x} + \mathrm{MHA}(\mathrm{LN}(\mathbf{x}))$.
(b) Calcula $\mathbf{z} = \mathbf{x}' + \mathrm{MLP}(\mathrm{LN}(\mathbf{x}'))$.
(c) Expresa $\mathbf{z}$ como la suma de tres términos: la representación original $\mathbf{x}$, la corrección de atención y la corrección del MLP.
(d) ¿Qué ocurriría si no existieran las conexiones residuales y el bloque simplemente aplicara $\mathbf{z} = \mathrm{MLP}(\mathrm{MHA}(\mathbf{x}))$? Reflexiona sobre la preservación de la información original y el problema del desvanecimiento del gradiente.