Día 07 · Tema 2 — Ejercicios: del cuello de botella RNN a la atención

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. Atención dot-product escalada, paso a paso

Consideremos una única *query* $\mathbf{q} = (2, 1)$, tres *keys* $\mathbf{k}_1 = (1, 0)$, $\mathbf{k}_2 = (0, 1)$, $\mathbf{k}_3 = (1, 1)$ y tres *values* $\mathbf{v}_1 = (10,\ 0)$, $\mathbf{v}_2 = (0,\ 20)$, $\mathbf{v}_3 = (5,\ 5)$, con $d_k = 2$.

(a) Calcula los *scores* brutos $s_j = \mathbf{q} \cdot \mathbf{k}_j$ para $j = 1, 2, 3$.
(b) Escala cada *score* dividiéndolo por $\sqrt{d_k} = \sqrt{2} \approx 1{,}4142$.
(c) Aplica softmax a los *scores* escalados para obtener los pesos de atención $\alpha_1, \alpha_2, \alpha_3$. Recuerda que

$$\text{softmax}(a, b, c) = \frac{(e^a,\ e^b,\ e^c)}{e^a + e^b + e^c}\,.$$

Valores de referencia: $e^{0{,}707} \approx 2{,}028$; $e^{1{,}414} \approx 4{,}113$; $e^{2{,}121} \approx 8{,}342$.
(d) Calcula la salida de atención $\text{Att} = \sum_{j=1}^{3} \alpha_j \mathbf{v}_j$.

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

2. Atención con máscara causal

Con los mismos datos del ejercicio anterior, supón ahora que la posición 3 queda *enmascarada* (por ejemplo, porque en un modelo autorregresivo aún no se ha generado). Antes de aplicar softmax, se asigna $s_3 \longrightarrow -\infty$ (lo que fuerza $e^{s_3} = 0$).

(a) Recalcula los pesos de atención $\alpha'_1, \alpha'_2, \alpha'_3$.
(b) Recalcula la salida Att′.
(c) Compara cualitativamente Att′ con la salida sin máscara del ejercicio 1. ¿Hacia qué *value* se desplaza la salida y por qué?

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

3. Matriz de atención completa

Ahora trabajamos con *dos* queries empaquetadas en la matriz $Q \in \mathbb{R}^{2\times 2}$, tres keys en $K \in \mathbb{R}^{3\times 2}$ y tres values en $V \in \mathbb{R}^{3\times 2}$:

$$Q = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}, \qquad K = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \qquad V = \begin{pmatrix} 10 & 0 \\ 0 & 20 \\ 5 & 5 \end{pmatrix}.$$

(a) Calcula el producto $QK^\top$ y la matriz escalada $S = QK^\top/\sqrt{d_k}$.
(b) Aplica softmax *por filas* a S para obtener la matriz de pesos $A \in \mathbb{R}^{2\times 3}$.
Valores de referencia adicionales: $e^0 = 1$; $e^{0{,}707} \approx 2{,}028$.
(c) Calcula la salida $Z = A\,V \in \mathbb{R}^{2\times 2}$.
(d) Comprueba que la primera fila de Z coincide con la respuesta del ejercicio 1.

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

4. Coste cuadrático de la atención

Un modelo Transformer procesa secuencias de T tokens con $d_k = 64$.

(a) ¿Cuántas entradas tiene la matriz $QK^\top$ cuando $T = 4096$? ¿Cuántos bytes ocupa en FP32 (4 bytes por entrada)? Expresa el resultado en megabytes.
(b) Repite el cálculo para $T = 512$.
(c) ¿Cuál es el factor de reducción al pasar de $T = 4096$ a $T = 512$? Relaciona este factor con la complejidad $O(T^2)$ de la atención.

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

5. Atención aditiva frente a dot-product

La *atención aditiva* (Bahdanau) calcula el *score* como

$$s(\mathbf{q}, \mathbf{k}_j) = \mathbf{v}^\top \tanh\big(W_1\,\mathbf{q} + W_2\,\mathbf{k}_j\big),$$

con parámetros aprendibles $W_1, W_2 \in \mathbb{R}^{3\times 2}$ y $\mathbf{v} \in \mathbb{R}^3$. Usa los mismos $\mathbf{q}, \mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3$ del ejercicio 1 y los pesos

$$W_1 = W_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0{,}5 & 0{,}5 \end{pmatrix}, \qquad \mathbf{v} = (1,\ 1,\ -1).$$

(a) Calcula $s(\mathbf{q}, \mathbf{k}_j)$ para $j = 1, 2, 3$.
Valores de referencia: $\tanh(1) \approx 0{,}762$; $\tanh(1{,}5) \approx 0{,}905$; $\tanh(2) \approx 0{,}964$; $\tanh(2{,}5) \approx 0{,}987$; $\tanh(3) \approx 0{,}995$.
(b) Compara los tres *scores* aditivos con los *scores* dot-product brutos ($s_j = \mathbf{q} \cdot \mathbf{k}_j$) del ejercicio 1. ¿Los dos mecanismos inducen el mismo ranking entre keys?
(c) ¿Cuál de los dos mecanismos es más eficiente computacionalmente y por qué?

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

6. ¿Por qué dividimos por $\sqrt{d_k}$?

Supón que cada componente de $\mathbf{q}$ y de $\mathbf{k}$ se muestrea de forma independiente según $\mathcal{N}(0, 1)$. El *score* bruto es $s = \mathbf{q} \cdot \mathbf{k} = \sum_{i=1}^{d_k} q_i\, k_i$.

(a) Calcula $\mathbb{E}[s]$ y $\mathrm{Var}(s)$ en función de d_k.
Pista: cada producto $q_i k_i$ tiene media 0 y varianza 1 (por ser producto de dos normales estándar independientes) y los sumandos son independientes entre sí.
(b) Particulariza para $d_k = 64$. ¿Cuál es la desviación típica de los *scores*?
(c) Cuando la desviación típica es grande, las entradas de softmax difieren mucho entre sí y los pesos se concentran en un solo token (softmax "saturado"). Ilustra este efecto calculando $\mathrm{softmax}(8,\ 0,\ -8)$. Valores de referencia: $e^{8} \approx 2981$; $e^{-8} \approx 0{,}000\,34$.
(d) Ahora divide por $\sqrt{d_k} = 8$ y calcula $\mathrm{softmax}(1,\ 0,\ -1)$. Valores de referencia: $e^{1} \approx 2{,}718$; $e^{-1} \approx 0{,}368$.
(e) Explica, a la luz de los apartados anteriores, por qué el factor $1/\sqrt{d_k}$ es necesario para que los gradientes de softmax no se desvanezcan durante el entrenamiento.