Atención: el puente al Transformer

Cuando una RNN ya no llega: del cuello de botella al acceso directo.

Conceptos clave

- **Cuello de botella RNN**: toda la historia se comprime en un único vector h_T. Con 100 pasos en $\mathbb{R}^{64}$, hay pérdida garantizada.
- **Atención**: cada query atiende a todos los keys con un peso, sin pasar por la cadena temporal.
- Fórmula canónica: $\text{Att}(Q,K,V) = \text{softmax}(QK^\top/\sqrt{d_k})\,V$.
- Coste: $O(T^2)$ en memoria y tiempo. Para $T = 4\,000$ son 16 millones de elementos.

Ejemplo a mano

Atención con números. Query $q = (1,0)$, keys $k_1 = (1,0)$, $k_2 = (0{,}5, 0{,}5)$, $k_3 = (-1, 0{,}5)$, values $v_1 = 10$, $v_2 = 20$, $v_3 = 30$.

$$\begin{aligned} \text{scores} &= q \cdot K^\top = (1{,}0,\ 0{,}5,\ -1{,}0), \\ \alpha &= \text{softmax} = (0{,}56,\ 0{,}34,\ 0{,}10), \\ \hat{y} &= \alpha \cdot V = 0{,}56 \cdot 10 + 0{,}34 \cdot 20 + 0{,}10 \cdot 30 = \mathbf{15{,}40}. \end{aligned}$$

La query "se parece"más a k_1, así que la salida tira hacia v_1. Si cambia q, cambian los pesos.

Concepto → aplicación: RNN vs. atencional

Caso	Mejor opción	Por qué
Forecast 12 sem. con 156 hist.	LSTM/GRU.	Memoria suficiente.
Traducción de un párrafo	Transformer.	Dependencias largas.
Asistente 16K tokens	Transformer + KV cache.	Acceso paralelo.
Audio en streaming (latencia)	RNN/GRU o SSM (Día 10).	Estado fijo online.

Idea clave del Día 07. **Self-attention compra calidad pagando memoria cuadrática.** La RNN no es obsoleta: sigue ganando cuando importa la latencia y el procesamiento es online.

Para llevarte: lectura Alammar, *The Illustrated Transformer* (sección self-attention); ejercicio: implementar la atención en numpy y comparar con `F.scaled_dot_product_attention`.