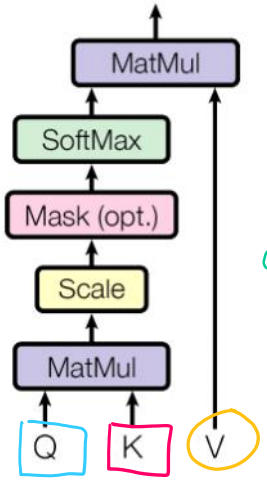
Transformers: Atención

Q: Query
K: Keys
V: Values
d_k : dimension de claves

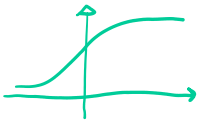


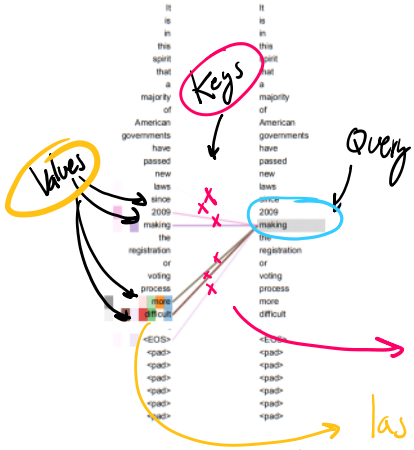
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

σ

$\sigma(z): \mathbb{R}^k \rightarrow [0,1]^k$ comprime vectores

distribución de probabilidad → Pesos W de atención





Cada fila es una consulta por un elemento

¿Que se conecta con la palabra making?

Con que elementos se quiere medir la atención

las posiciones en K

Traduce vectores a las palabras

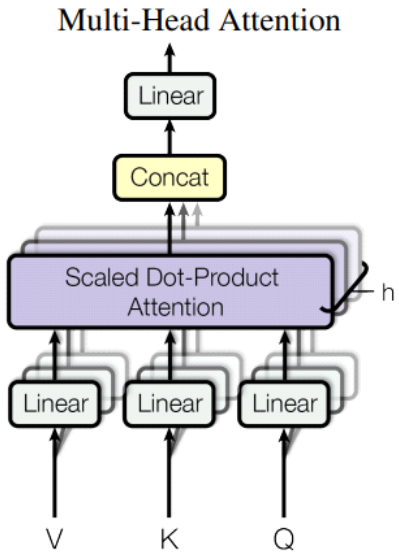
las palabras 2009, making, more, difficult

En la arquitectura Transformer, el mecanismo de atención se utiliza para capturar las relaciones entre las palabras de entrada y las palabras de salida en tareas de procesamiento del lenguaje natural, como la traducción de idiomas.

Query (consulta): Imagina que estás traduciendo una oración y te encuentras en una posición específica de la oración de salida. La consulta representa esa posición y ayuda a encontrar las palabras más relevantes en la oración de entrada que se relacionan con la palabra que estás tratando de traducir. Puedes ver la consulta como una "pregunta" que haces sobre la oración de entrada para obtener información relevante.

Key (clave): La clave es como un índice que te ayuda a buscar palabras específicas en la oración de entrada que se relacionan con la palabra en la posición de salida. Proporciona información sobre la importancia de cada palabra en relación con la consulta. Puedes considerar la clave como una "etiqueta" que te indica qué palabras son relevantes para responder la pregunta formulada por la consulta.

Value (valor): El valor es la información real asociada a las palabras de entrada. Proporciona detalles y contexto sobre las palabras en la oración de entrada. Puedes pensar en el valor como la "respuesta" a la pregunta formulada por la consulta, que contiene información útil y relevante sobre las palabras de entrada.



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Luego en el paper se incluyen 8 heads que se dividen el trabajo de atencion

Se dividen las consultas, de manera que el modelo se pueda enfocar en distintos aspectos de la informacion de entrada.

Se dividen en 8 grupos y cada grupo se encarga de realizar su propio calculo de atencion independiente para sus querys, de esa forma cada head puede aprender distintas estrategias, representandose en distintos pesos.

Cambiar la cantidad de heads permitira capturar distintas relaciones en el modelo.