

UNIVERSIDAD DEL VALLE DE GUATEMALA

DEEP, sección 20



Hoja de Trabajo # 2

Joaquin Campos – 22155

Sofía Velásquez - 22049

José Marchena - 22398

Guatemala, Agosto 2025

Repaso Teoría

1. ¿Cuál es la principal innovación de la arquitectura Transformer?

La atención que específicamente el mecanismo de autoatención (self-attention) entonces, elimina por completo el uso de convoluciones y recurrencias con el objetivo que le permita modelar dependencias a lo largo del alcance entre palabras sin importar la posición en la secuencia, con mayor paralelización y eficiencia en el entrenamiento.

2. ¿Cómo funciona el mecanismo de atención del scaled dot-product?

Se basa en las 3 entradas, query, keys y values de forma que se calcula el producto punto entre cada query y todos los keys. Luego el resultado se divide por la raíz cuadrada de d_k , para evitar valores muy grandes que saturan el softmax. Luego, se aplica softmax para obtener un conjunto de pesos de atención. Y por último se usan esos pesos para obtener una combinación lineal de los valores.

3. ¿Por qué se utiliza la atención de múltiples cabezales en Transformer?

Para que pueda poner “atención” en diferentes partes de la secuencia al mismo tiempo, entonces cada cabeza usa diferentes entradas, para que puedan obtener diferentes datos y luego conectarlos como una especie de paralelización tipo pragma openMP que se juntan al final de la cláusula.

4. ¿Cómo se incorporan los positional encodings en el modelo Transformer

Al no tener información sobre el orden de las palabras, se añaden los positional encodings a los embeddings de entrada, las cuales son vectores calculados con funciones de seno y coseno a diferentes frecuencias. Este diseño, permite que aprenda de relaciones relativas entre tokens.

5. ¿Cuáles son algunas aplicaciones de la arquitectura Transformer más allá de la machine translation?

- Biología computacional. (jumper, 2021)
- Audio y voz.
- Modelos con imágenes, texto audio y demás.

Jumper, J., (et al). (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>