METODOS NUMERICOS

[Actividad extracurricular 07]

Redes neuronales Transformers

Nombre: Luis Enrique Pérez Señalin.

1. **¿Qué es, para qué sirve?**

   Es una arquitectura de red neuronal diseñada para tratar secuencias de datos de manera eficiente y utilizan mecanismos de atención para captar contextos y dependencias a larga distancia sin necesidad de recurrencia, como se observa en las RNN y LSTM, esto significa que es capaz de usar el texto tanto antes como después de una palabra, lo que permite una mejor comprensión de la entrada de texto.

   Los Transformers revolucionaron la generación y análisis de lenguaje natural (NLP) gracias al contexto, permitiendo un mejor análisis en comparación con otros modelos.

2. **Qué es un embedding, cuál es el tamaño del embedding en los principales modelos de lenguaje (ChatGPT 3.5, 4, Claude, Mistral, etc)**

   Los embeddings son representaciones vectoriales de palabras que permiten a los modelos procesar texto como datos numéricos. El tamaño de estos embeddings varía entre los modelos:

   - **GPT-3.5 y GPT-4**: Usan embeddings de 768 y 1280 dimensiones respectivamente.
   - **Claude y Mistral**: No se especifica claramente en la información disponible, pero típicamente estos modelos operan en rangos similares a GPT para mantener la compatibilidad con aplicaciones de procesamiento de lenguaje.

3. **Ventajas con respecto a otro tipo de redes neuronales (i.e. CNN, LSTM)**

   Los Transformers gestionan dependencias a larga distancia sin degradar el rendimiento, gracias a su mecanismo de atención que no se limita por la secuencia de procesamiento de los datos.

   Permite un entrenamiento en paralelo, reduciendo significativamente el tiempo requerido para entrenar grandes volúmenes de datos y la mejora en el análisis de lenguaje natural (NLP).

4. **En qué parte de la arquitectura transformer existe factorización de matrices?**

   Se usa a través de la función de atención, donde las matrices de 'query', 'key' y 'value' son manipuladas para generar una distribución de atención sobre los inputs. Esta operación se basa en la multiplicación de matrices seguida de una normalización softmax