ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

METODOS NUMERICOS ICCD412

JUAN FRANCISCO PINTO ANDRANGO

Actividad Extracurricular 06B Factorio en transformers

GR1CC

FECHA DE ENTREGA 28 DE ENERO DEL 2026

Factoreo en transformers

Investigue sobre el uso de factoreo de matrices en la arquitectura de redes neuronales Transformers. Indique las razones y las ventajas de realizar esta operación.

Redes Neuronales en modelos de lenguaje

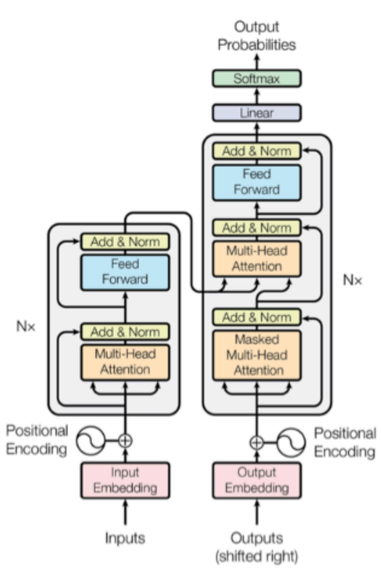
Una red neuronal es el método de la inteligencia artificial que se encarga de enseñar a las computadoras cómo procesar datos de manera similar a la del cerebro humano. Es un proceso de Machine Learning que se denomina aprendizaje profundo, que utiliza los nodos o las neuronas interconectadas en una estructura de estilo capas. Se crea un sistema adaptable que las computadoras utilizan para aprender de errores y realizar una mejora continua. Así, las redes neuronales artificiales intentan resolver problemas complejos como resúmenes de documentos o reconocimiento de rostros, todo esto con mayor precisión

Redes Transformer

Las Redes Transformer son un nuevo tipo de red neuronal que planea transformarse en un gran avance para el Machine Learning. La red Transformer nació como una alternativa al problema de la traducción de texto de un idioma a otro. En estas redes, la totalidad de la secuencia de entrada se procesa en paralelo por la red. La nueva secuencia es convertida en representación numérica usando un embedding.

Una red de tipo Transformer es solo una red neuronal para secuencias que se basa en autoatención, que se adapta al texto. En la actualidad se impulsan avances en el procesamiento de lenguaje natural y esta es su arquitectura.

Out[1]=



Attention is all you need

La Red Transformer descrita en el artículo de 2017 , desarrollado por investigadores de Google, y nacieron inicialmente como una alternativa al problema de la traducción de texto de un idioma a otro.

Los Transformers son una arquitectura de red neuronal desarrollada para tareas de procesamiento de lenguaje, que se ha consolidado como herramienta principal para abordar problemas como el procesamiento de texto, audio, imágenes, aprendizaje y otros dominios de datos. Su aspecto destacado es el mecanismo de autoatención, que hace que el modelo pondere dinámicamente diferentes partes de una secuencia de entrada al procesarla.

Factoreo de matrices en la arquitectura de redes neuronales Transformers

El factoreo de matrices en las redes neuronales Transformers no forma parte de un proceso que hace una red de manera natural. Es una nueva técnica de optimización inteligente usada principalmente en los modelos de lenguaje gigantes, como GPT o Llama, para que trabajen de manera más rápida o para poder reentrenarlos (fine-tuning) sin la necesidad de una supercomputadora.

Una red Transformer tiene parámetros ("pesos") que son matrices gigantescas de números. El factoreo de matrices trata de representar una tabla gigante como la multiplicación de dos matrices más pequeñas. El factoreo de matrices es una técnica de compresión y eficiencia que permite a los modelos de lenguaje poder ser más manejables, rápidos y entrenables.

Link del Repositorio de Git-Hub

https://github.com/JuanfranPinto/Metodos-Numericos-/blob/main/Actividad-extracurricular-06b-Factoreo-en-transformers-PJF.ipynb