***Redes Neuronales-Transformers***

***Leandro Bravo***

***Julio 19, 2024***

1. **¿Qué es, para qué sirve?**

Es un tipo de modelo de inteligencia artificial diseñado principalmente para procesar secuencias de datos, como texto. Se introdujo en el artículo "Attention is All You Need" en 2017 y ha revolucionado el campo del procesamiento del lenguaje natural (NLP) y más allá.

En el contexto de las redes neuronales, un Transformer es una arquitectura específica diseñada para procesar datos secuenciales de manera eficiente y efectiva, utilizando principalmente el mecanismo de atención.

*¿Para qué sirve?*

* ***Traducción automática:*** Los Transformers pueden traducir textos de un idioma a otro.
* ***Resumen de textos:*** Pueden generar resúmenes coherentes de textos largos.
* ***Generación de texto:*** Capaces de generar textos creativos y coherentes basados en un contexto dado.
* ***Respuesta a preguntas:*** Pueden responder preguntas basadas en un corpus de texto dado.
* ***Análisis de sentimientos:*** Pueden determinar el sentimiento detrás de un texto.
* ***Tareas de visión por computadora:*** Con adaptaciones, también se usan en clasificación de imágenes y otros problemas de visión.

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

Es un vector de números que representa una palabra, frase o cualquier otra unidad de texto en un espacio continuo y de alta dimensión. La idea es que palabras con significados similares tendrán representaciones vectoriales similares.

*Tamaños de embedding en los principales modelos de lenguaje:*

1. ***ChatGPT 3.5:*** Utiliza embeddings generados por el modelo base. Aunque no se especifica un tamaño fijo.
2. ***ChatGPT-4 y Claude:*** Los detalles específicos sobre sus tamaños no están disponibles en los resultados proporcionados.
3. ***Mistral Large (LLM):*** Este modelo utiliza 12,288 dimensiones para codificar su vocabulario.
4. ***Text-embedding-3-small:*** Un modelo más pequeño y eficiente con dimensiones no especificadas.
5. ***Text-embedding-3-large:*** Un modelo más grande y potente con hasta 3072 dimensiones.
6. **Ventajas con respecto a otro tipo de redes neuronales (i.e. CNN, LSTM).**

***Ventajas de los Transformers sobre LSTM***

1. ***Paralelización****:*

* *Transformers:* Pueden procesar secuencias en paralelo, lo que los hace mucho más rápidos en el entrenamiento y la inferencia.
* *LSTM:* Procesan secuencias de manera secuencial, lo que limita la paralelización y puede llevar a tiempos de entrenamiento más largos.

1. ***Manejo de Dependencias Largas****:*

* *Transformers:* Utilizan un mecanismo de atención que les permite capturar dependencias a largo plazo de manera eficiente, incluso en secuencias muy largas.
* *LSTM:* Aunque pueden manejar dependencias a largo plazo mejor que las RNN tradicionales, todavía tienen limitaciones y pueden olvidar información pasada en secuencias muy largas.

1. ***Escalabilidad:***

* *Transformers:* Son altamente escalables y han demostrado ser efectivos en modelos de gran escala, como GPT-3 y BERT.
* *LSTM:* Su escalabilidad es limitada debido a la naturaleza secuencial del procesamiento.

1. ***Versatilidad:***

* *Transformers:* Pueden ser adaptados a una variedad de tareas, tanto en procesamiento de lenguaje natural (NLP) como en otras áreas como visión por computadora.
* *LSTM:* Principalmente utilizados en tareas de secuencias y series temporales.

***Ventajas de los Transformers sobre CNN***

1. ***Contexto Global:***

* *Transformers:* Capturan el contexto global de la secuencia mediante el mecanismo de atención, lo que les permite entender relaciones a larga distancia.
* *CNN:* Capturan relaciones locales a través de convoluciones, lo que puede limitar su capacidad para entender el contexto global sin capas adicionales como convoluciones dilatadas o pooling.

1. ***Flexibilidad:***

* *Transformers:* Pueden manejar secuencias de longitud variable sin necesidad de operaciones adicionales.
* *CNN:* Están diseñadas para trabajar con entradas de tamaño fijo, aunque se pueden adaptar mediante técnicas como el padding.

1. ***Adaptabilidad a Diferentes Tipos de Datos:***

* *Transformers:* Originalmente diseñados para NLP, se han adaptado para tareas de visión por computadora, series temporales, y más.
* *CNN:* Son extremadamente efectivas para tareas de visión por computadora, pero no son tan versátiles en otras áreas sin modificaciones significativas.

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

La factorización de matrices en la arquitectura Transformer ocurre principalmente en:

* Las proyecciones lineales de *Q, K*, y *V* en el mecanismo de atención.
* El cálculo de los scores de atención mediante el producto punto escalado *QKT*.
* La proyección final después de concatenar las cabezas de atención.
* Las capas feed-forward que siguen a las capas de atención.

***Bibliografía***

[1] “¿Qué son los ‘embeddings’?”, Codificando Bits. [En línea]. Disponible en: <https://www.codificandobits.com/blog/embeddings-y-llms/>. [Consultado: 14-jul-2024].

[2] R. Fernandez, “El modelo Embeddings (Incrustaciones) de Palabras”, ▷ Cursos de Programación de 0 a Experto © Garantizados, 04-sep-2018. [En línea]. Disponible en: <https://unipython.com/el-modelo-embeddings-incrustaciones-de-palabras/>. [Consultado: 14-jul-2024].

[3] DimensionIA, “Descifrando el Enigma de los Modelos de Lenguaje: Text Embeddings”, DimensionIA, 27-may-2023. [En línea]. Disponible en: <https://www.dimensionia.com/text-embeddings-en-los-modelos-de-lenguaje/>. [Consultado: 14-jul-2024].

[4] “ChatGPT”, *Chatgpt.com*. [En línea]. Disponible en: <http://chatgpt.com>. [Consultado: 14-jul-2024].