Trabajo Práctico de Procesamiento del Lenguaje Natural – Ejercicio 2

Alumno: Demarré, Lucas Federico – D.N.I.: 44.523.677

Año: 2024

Estado del Arte de las Aplicaciones Actuales de Agentes Inteligentes Usando Modelos LLM Libres

Introducción

En la actualidad, los Modelos de Lenguaje de Gran Tamaño (LLMs) están redefiniendo el panorama de la inteligencia artificial (IA), impulsando innovaciones y remodelando nuestra interacción con la tecnología. La disponibilidad de LLMs de código abierto está democratizando el acceso a estas tecnologías avanzadas, permitiendo a investigadores y desarrolladores explorar sus complejidades y adaptarlas para tareas específicas. En este informe hablaremos de tres LLM Libres que más me interesaron:

- **LLaMA 2**: es un modelo de lenguaje grande y generativo desarrollado por *Meta AI*, representando un avance significativo respecto a su predecesor *LLaMA*. Este modelo se basa en la arquitectura original de Transformers y ha sido actualizado con una serie de mejoras que amplían su capacidad y eficacia en una amplia gama de tareas de procesamiento de lenguaje natural.
- **BLOOM:** desarrollado por *BigScience*, es un modelo de lenguaje grande (LLM) multilingüe y de código abierto con 176 mil millones de parámetros, diseñado para generar texto coherente en 46 idiomas naturales y 13 lenguajes de programación.
- **Falcon 180B:** desarrollado por el *Instituto de Innovación Tecnológica* (TII) de Abu Dhabi, se destaca como el modelo de lenguaje grande (LLM) de código abierto más grande y de alto rendimiento disponible actualmente, con 180 mil millones de parámetros.

LLaMA 2

Características principales:

- **Arquitectura**: *LLaMA 2* mantiene la estructura fundamental de los modelos *LLaMA*, incluyendo características como la pre-normalización al estilo de *GPT-3*, el uso de la función de activación *SwiGLU*, y los embeddings posicionales rotatorios (RoPE) para mejorar el rendimiento en entrenamiento.
- **Conjunto de Datos de Entrenamiento Ampliado**: El modelo ha sido entrenado con un conjunto de datos un 40% más grande que el utilizado para el modelo *LLaMA* original, alcanzando un total de 2 billones de tokens. Este aumento en la cantidad de datos mejora la capacidad del modelo, incluso en sus versiones más pequeñas, y se ha hecho un esfuerzo consciente para excluir datos de sitios que contienen información personal y privada.
- **Variantes de Chat y RHLF**: *LLaMA 2* introduce variantes de chat, ajustadas finamente en base a preferencias humanas mediante el aprendizaje por refuerzo con feedback humano (RHLF). Estas variantes representan un avance significativo en la interactividad humana y se han afinado mediante técnicas de supervisión, RHLF, y afinamiento iterativo.
- **Escalabilidad hasta 70 Mil Millones de Parámetros**: El modelo más grande de *LLaMA 2* cuenta con 70 mil millones de parámetros, lo que le permite competir favorablemente incluso con modelos de código cerrado como *ChatGPT (GPT3.5)*, aunque aún se encuentra detrás de modelos como *GPT-4*. Sin embargo, se espera que la brecha se reduzca a medida que la comunidad de código abierto continúe ajustando el modelo.

Responsabilidad y Comunidad: *Meta A*I ha enfatizado su compromiso con el desarrollo responsable de IA, proporcionando guías de uso responsable, realizando pruebas de *red-teaming* para identificar y mitigar respuestas problemáticas, y estableciendo un foro comunitario para fomentar la deliberación y la toma de decisiones informadas en torno a la IA generativa. Además, han lanzado un programa de becas Llama *Impact* para incentivar la utilización de *LLaMA 2* en la solución de problemas importantes en áreas como el medio ambiente y la educación.

Aplicaciones y Uso: *LLaMA 2* es accesible para investigadores, desarrolladores y empresas, permitiendo una amplia gama de aplicaciones en generación de texto, comprensión del lenguaje, y más. La disponibilidad de este modelo en plataformas como *Hugging Face* y su integración en entornos de desarrollo como *Gradient* facilita su uso y experimentación por parte de la comunidad de IA.

Conclusión: Este modelo representa un paso significativo hacia adelante en la democratización del acceso a modelos de lenguaje avanzados, ofreciendo una alternativa poderosa y flexible a modelos de código cerrado. Con su lanzamiento, *Meta AI* no solo busca impulsar el avance tecnológico en el campo de la IA, sino también promover un ecosistema de innovación colaborativa y responsable.

BLOOM

Características y uso:

- **Preprocesamiento de Datos y Afinamiento con Datasets Provocados**: *BLOOM* se benefició de un proceso detallado de preprocesamiento de datos, incluyendo la deduplicación y la redacción de privacidad, y empleó el afinamiento multitarea provocado para mejorar la generalización de tareas en un escenario de cero disparos.
- **Cómo Usar BLOOM**: Se puede acceder a BLOOM fácilmente a través del ecosistema de *Hugging Face*, utilizando bibliotecas como *Transformers* y *accelerate* para su implementación. El modelo está diseñado tanto para la generación de texto como para servir como base preentrenada para tareas específicas más finas.

Aplicaciones e Impacto: *BLOOM* ha sido diseñado para impulsar la investigación pública sobre LLMs, con aplicaciones que van desde la generación de contenido multilingüe y desarrollo de software hasta la investigación académica. Sus capacidades multilingües lo hacen particularmente valioso para la generación de contenido inclusivo y la asistencia en la codificación y desarrollo de software.

Consideraciones Éticas y Limitaciones: A pesar de sus avances, *BLOOM* no está exento de limitaciones y consideraciones éticas. El manejo de datos sensibles, la toma de decisiones de alto riesgo, y la sustitución de la interacción humana son usos fuera de su alcance previsto. Además, como con cualquier LLM, hay preocupaciones sobre el sesgo de datos, la privacidad, y el riesgo de generación de desinformación.

Usuarios Directos e Indirectos: Los usuarios directos de *BLOOM* incluyen a desarrolladores, científicos de datos, investigadores y creadores de contenido, mientras que los usuarios indirectos abarcan desde negocios y organizaciones hasta el público general, beneficiándose de las mejoras tecnológicas y la accesibilidad de contenido multilingüe que *BLOOM* facilita.

Conclusión*: BLOOM* se destaca por su enfoque colaborativo y abierto, buscando democratizar el acceso a tecnologías de LLM y fomentar la investigación y aplicaciones responsables. Su lanzamiento subraya la importancia de la colaboración abierta en el avance de la IA, ofreciendo una herramienta potente y versátil para diversas aplicaciones lingüísticas y de programación.

Falcon 180B

Características principales:

- **Arquitectura Innovadora y Rendimiento**: *Falcon 180B*, con sus 180 mil millones de parámetros, marca un hito como el modelo de lenguaje grande (LLM) de acceso abierto más grande y avanzado disponible. Incorpora la atención multiconsulta para reducir los requerimientos de ancho de banda de memoria en inferencias, destacándose por un rendimiento comparable a modelos de vanguardia como *PaLM 2* de Google y cercano a *GPT-4*.
- **Entrenamiento Extensivo con Datos Diversificados**: Fue entrenado con 3.5 trillones de tokens, lo que contribuye significativamente a su capacidad para entender y generar texto en una amplia gama de temas y contextos. Este modelo también incluye una versión específica para chat, ajustada en conjuntos de datos de instrucciones, ampliando sus aplicaciones potenciales.

Accesibilidad y Aplicaciones: Disponible para uso comercial bajo condiciones específicas*, Falcon 180B* se presenta en el ecosistema de *Hugging Face*, facilitando su acceso y aplicación en una variedad de contextos, desde investigación hasta desarrollo de productos, permitiendo a los usuarios beneficiarse de su avanzada comprensión del lenguaje y capacidades de generación de texto.

Multilingüismo y Aplicación Global: A pesar de su enfoque en inglés, *Falcon 180B* exhibe comprensión en alemán, español, francés, y capacidades básicas en múltiples otros idiomas europeos, lo que amplía su aplicabilidad a contextos lingüísticos diversos y fomenta la inclusión de comunidades no anglófonas en la investigación y aplicaciones de IA.

Requisitos de Hardware y Consideraciones de Costo: Aunque el acceso al modelo es abierto, su implementación demanda recursos significativos, con requisitos de memoria y GPU que pueden representar barreras para individuos y organizaciones con limitaciones de hardware. Este aspecto subraya la necesidad de considerar el equilibrio entre avance tecnológico y accesibilidad.

Responsabilidad y Comunidad:

- **Licencia y Uso Comercial**: *Falcon 180B* se ofrece bajo una licencia que permite el uso comercial, aunque con restricciones específicas, promoviendo un marco de uso responsable y ético en aplicaciones comerciales y de investigación.
- **Contribución a la Innovación y el Conocimiento Abierto**: Al ser el LLM de código abierto más grande disponible*, Falcon 180B* empodera a la comunidad científica y de desarrollo, ofreciendo nuevas posibilidades para la exploración, innovación, y desarrollo de soluciones basadas en IA sobre una base tecnológica avanzada y accesible.

Aplicaciones y Uso:

- **Plataforma para Investigación y Desarrollo**: La incorporación de *Falcon 180B* en el ecosistema de *Hugging Face* y su disponibilidad para ajuste fino y aplicaciones específicas lo convierten en una herramienta valiosa para una amplia gama de tareas de procesamiento de lenguaje natural (NLP), desde la generación de texto hasta la comprensión del lenguaje.
- **Potencial para Innovación Multilingüe**: La capacidad de *Falcon 180B* para trabajar con múltiples idiomas abre el camino para su uso en aplicaciones globales, incluyendo traducción automática, generación de contenido multilingüe, y asistencia lingüística, contribuyendo a la reducción de barreras lingüísticas y culturales en la tecnología y la comunicación.

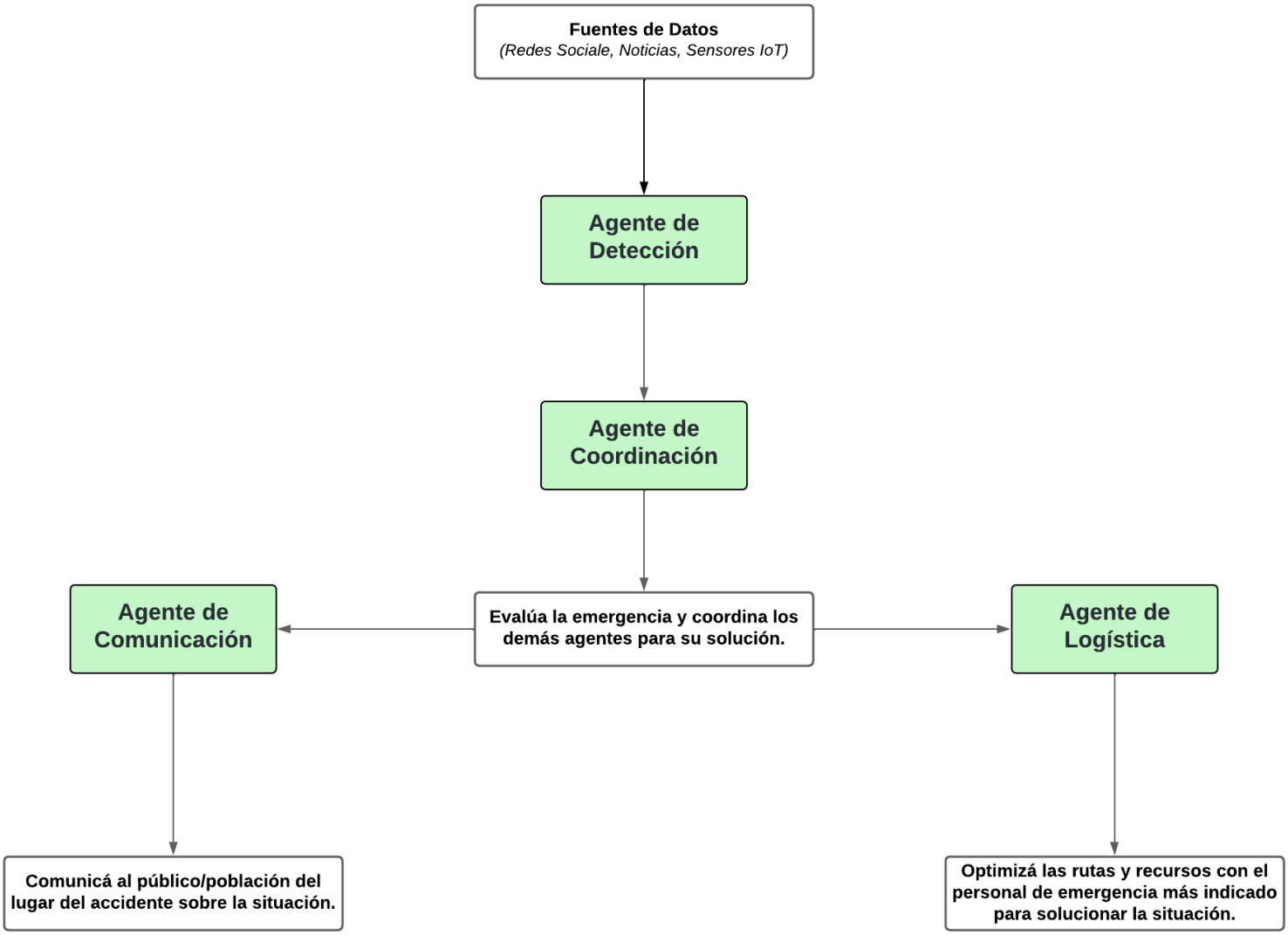
Conclusión: *Falcon 180B* se destaca no solo por su escala y rendimiento sino también por su modelo de acceso abierto, que fomenta una mayor colaboración y experimentación dentro de la comunidad de IA. A pesar de los desafíos relacionados con los requisitos de hardware y costos de implementación, ofrece un avance significativo en la democratización del acceso a tecnologías de LLM avanzadas, promoviendo un ecosistema de innovación inclusivo y colaborativo.

Problemática a Solucionar con un Sistema Multiagente

Una problemática relevante en la sociedad actual es la gestión eficiente de las respuestas ante emergencias, como desastres naturales o accidentes urbanos. La coordinación rápida y efectiva entre los diferentes servicios de emergencia puede ser la diferencia entre salvar vidas y mitigar daños.

El sistema multiagente propuesto incluiría los siguientes agentes, cada uno especializado en una tarea crítica de la gestión de emergencias:

- **Agente de Detección**: Utilizaría tecnologías de procesamiento de lenguaje natural y análisis de datos para monitorear redes sociales, noticias y sensores IoT (sensores inteligentes) para detectar posibles emergencias en tiempo real.
- **Agente de Coordinación**: Este agente, basado en modelos de toma de decisiones y optimización, sería el encargado de evaluar la información recopilada y coordinar la respuesta entre los diferentes servicios de emergencia
- **Agente de Logística**: Especializado en la optimización de rutas y recursos, este agente garantizaría la asignación eficiente de recursos y personal de emergencia al lugar del incidente.
- **Agente de Comunicación**: Utilizaría LLM para interactuar con el público y los servicios de emergencia, proporcionando información actualizada, recomendaciones y recopilando datos relevantes del terreno.



Esquema del Sistema Multiagente

Ejemplo de Conversación:

Agente de Detección: *"He detectado un aumento significativo de mensajes en redes sociales sobre un posible incendio forestal en la región de la Sierra Norte. Los sensores de temperatura en la zona también muestran lecturas anómalas."*

Agente de Coordinación: *"Recibido, Agente de Detección. Solicitando al Agente de Logística que evalúe la disponibilidad y la ubicación de los equipos de bomberos más cercanos y la mejor ruta para llegar al lugar del incendio."*

Agente de Logística: *"El equipo de bomberos más cercano está en la estación de San Miguel, a 10 km del incendio. La ruta más rápida está actualmente despejada y tomará aproximadamente 15 minutos. Estoy coordinando la movilización inmediata."*

Agente de Comunicación: *"Informando a la población local sobre el incendio. Mensaje enviado: 'Alerta de incendio en la Sierra Norte. Por favor, eviten el área y sigan las instrucciones de evacuación si se encuentran cerca. Manténganse informados para más actualizaciones.'"*

Bibliografía:

"8 Top Open-Source LLMs for 2024 and Their Uses."

"Llama 2: open source, free for research and commercial use."

"LLaMA 2: a model overview and demo tutorial with Paperspace Gradient."

"Exploring BLOOM: A Comprehensive Guide to the Multilingual Large Language Model."

"HuggingFace: BLOOM."

"Abu Dhabi Releases Largest Openly-Available Language Model Falcon 180B."

"Spread Your Wings: Falcon 180B is here."

"Falcon 180B: Model Overview."

"Introducing Falcon 180b: A Comprehensive Guide with a Hands-On Demo of the Falcon 40B."

"¿Para qué sirven los agentes inteligentes? Ejemplos de aplicación."