TRABAJO PRÁCTICO FINAL

PROCESAMIENTO DE LENGUAJE NATURAL

EJERCICIO 2

(Parte 1)

Introducción

Los modelos de lenguaje de gran tamaño (LLM) han revolucionado el campo del procesamiento del lenguaje natural, permitiendo a los agentes inteligentes comprender y generar texto de manera más sofisticada que nunca, traducir idiomas, escribir diferentes tipos de contenido creativo y responder preguntas de manera informativa.

Esto ha sido impulsado por la disponibilidad de grandes conjuntos de datos y avances en la computación en la nube. Los LLM son muy buenos para comprender el lenguaje debido a la amplia capacitación previa que se ha realizado para los modelos básicos en billones de líneas de texto de dominio público, incluido el código.

A medida que los LLM se vuelven cada vez más sofisticados, se pone énfasis en democratizar el acceso a ellos, ofreciendo a los desarrolladores la oportunidad de profundizar en sus complejidades y ajustarlos para tareas específicas.

En este informe, exploramos el estado actual de las aplicaciones de agentes inteligentes que hacen uso de LLMs de código abierto, analizando sus capacidades, limitaciones y las implicaciones éticas asociadas.

Estado general

Analizando el estado actual se puede decir que si bien los LLMs de código abierto han demostrado ser herramientas poderosas para una variedad de aplicaciones, también presentan desafíos significativos en términos de la originalidad y precisión del texto generado. Por ejemplo, si bien algunos modelos exhiben un discurso de alta calidad, también muestran una propensión a la memorización de contenido, lo que plantea interrogantes sobre la naturaleza misma del aprendizaje en estos sistemas.

Además, se observan diferencias significativas entre los distintos modelos en cuanto a su capacidad para generar contenido libre de errores factuales, lógicos y de consistencia discursiva. Estos hallazgos sugieren la necesidad de un enfoque integral y reflexivo para evaluar y utilizar estos modelos de manera ética y efectiva en aplicaciones de agentes inteligentes.

Aplicaciones actuales

Las aplicaciones de agentes inteligentes basados en LLM se encuentran en una etapa temprana de desarrollo, pero ya se han visto casos de éxito en diversos campos:

Chatbots

Los chatbots son agentes conversacionales que pueden interactuar con los usuarios a través de texto o voz. Los LLM permiten crear chatbots más sofisticados que pueden comprender mejor las preguntas y solicitudes de los usuarios, y proporcionar respuestas más relevantes y personalizadas. Algunos ejemplos:

- LaMDA de Google se ha utilizado para crear chatbots que pueden conversar sobre diversos temas, desde la actualidad hasta la vida personal.
- GPT-3 se ha utilizado para crear chatbots para atención al cliente que pueden responder preguntas sobre productos y servicios.

Asistentes virtuales

Los asistentes virtuales son agentes inteligentes que pueden ayudar a los usuarios con tareas cotidianas como programar citas, realizar reservas o buscar información. Los LLM permiten crear asistentes virtuales más conversacionales y adaptables a las necesidades de cada usuario. Algunos ejemplos:

- Google Assistant y Siri utilizan LLM para mejorar la comprensión del lenguaje natural y la capacidad de respuesta a las solicitudes de los usuarios.
- Amazon Alexa se ha integrado con LLM para ofrecer una experiencia más personalizada e interactiva.

Generación de contenido

Los LLM se pueden utilizar para generar contenido textual de forma automática, como artículos de noticias, publicaciones en redes sociales o incluso guiones de películas. Algunos ejemplos:

- GPT-3 se ha utilizado para generar artículos de noticias que son indistinguibles de los escritos por periodistas humanos.
- LaMDA se ha utilizado para crear guiones para cortometrajes.

Educación

Los LLM se pueden utilizar para crear plataformas de aprendizaje personalizadas que se adapten a las necesidades de cada estudiante. Algunos ejemplos:

- Carnegie Learning utiliza LLM para crear experiencias de aprendizaje personalizadas para estudiantes de matemáticas.
- Pearson utiliza LLM para crear herramientas de evaluación que pueden identificar las fortalezas y debilidades de los estudiantes.

Resumen de textos

Los LLM se pueden utilizar para resumir textos largos de forma automática, lo que puede ser útil para estudiantes, investigadores o cualquier persona que necesite procesar grandes cantidades de información. Algunos ejemplos:

- Google Summarize utiliza LLM para resumir artículos de noticias, páginas web y otros documentos.
- SMMRY utiliza LLM para resumir correos electrónicos, documentos y libros.

Modelos de código abierto destacados

Llama 2

Llama 2 ha sido la consecuencia de una colaboración entre Meta y Microsoft, con lo cual se buscó garantizar la confiabilidad. El modelo se ha sometido a un riguroso entrenamiento para minimizar las "alucinaciones", la desinformación y los sesgos.

Está optimizado para plataformas como AWS, Azure y la plataforma de alojamiento de modelos de IA de Hugging Face. Está diseñado para alimentar una gama de aplicaciones de última generación.

Características principales de LLaMa 2:

- Diversos datos de entrenamiento: Los datos de entrenamiento de Llama 2 son extensos y variados, lo que garantiza una comprensión y un rendimiento completos.
- Disponibilidad abierta: A diferencia de su predecesor, Llama 2 está disponible para un público más amplio, listo para ajustarse en múltiples plataformas.
- Versiones optimizadas: Llama 2 viene en dos versiones principales: Llama 2 y Llama 2-Chat, y esta última está especialmente diseñada para conversaciones bidireccionales. Estas versiones varían en complejidad de 7 mil millones a 70 mil millones de parámetros.
- Entrenamiento mejorado: Llama 2 se entrenó con dos millones de tokens, un aumento significativo de los 1.4 billones de tokens originales de Llama.

BLOOM

Un proyecto en el que participaron voluntarios de más de 70 países y expertos de Hugging Face. Tiene la capacidad de generar texto coherente y preciso en 46 idiomas y 13 lenguajes de programación.

Características principales de Bloom:

- Acceso de código abierto: El código fuente del modelo y los datos de capacitación están disponibles públicamente, lo que promueve la transparencia y la mejora colaborativa.
- Generación de texto autorregresivo: Diseñado para continuar el texto a partir de un mensaje determinado, BLOOM se destaca en extender y completar secuencias de texto.
- Recuento masivo de parámetros: Con 176 mil millones de parámetros, BLOOM se erige como uno de los LLM de código abierto más poderosos que existen.
- Accesibilidad gratuita: Los usuarios pueden acceder y utilizar BLOOM de forma gratuita a través del ecosistema Hugging Face, mejorando su democratización en el campo de la IA.
- Capacitación a escala industrial: El modelo se entrenó con grandes cantidades de datos de texto utilizando importantes recursos computacionales, lo que garantiza un rendimiento sólido.

Falcon LLM

Opera como un modelo decodificador autoregresivo, lo que significa que predice el token subsiguiente en una secuencia basándose en los tokens precedentes, algo que recuerda al modelo GPT. Falcon ha demostrado un rendimiento superior a GPT-3, logrando este hito con solo el 75% del presupuesto de cómputo de entrenamiento.

Se programó con un sistema de canalización de datos que se escaló a decenas de miles de núcleos de CPU. Esto permitió un procesamiento rápido y la extracción de contenido de alta calidad de la web, logrado mediante procesos exhaustivos de filtrado y deduplicación.

Características principales de Falcon LLM:

- Parámetros extensos: Falcon-40B está equipado con 40 mil millones de parámetros, lo que garantiza un aprendizaje y un rendimiento integrales.

- Canalización de datos de alta calidad: La canalización de datos de TII garantiza la extracción de contenido de alta calidad de la web, crucial para el entrenamiento del modelo.
- Variedad de modelos: Además de Falcon-40B, TII ofrece Falcon-7B y modelos especializados como Falcon-40B-Instruct y Falcon-7B-Instruct.
- Disponibilidad de código abierto: Falcon LLM ha sido de código abierto, lo que promueve la accesibilidad y la inclusión en el dominio de la IA.

BERT

Lanzado en 2018 por Google como un LLM de código abierto, BERT (siglas de Bidirectional Encoder Representations from Transformers), alcanzó rápidamente un rendimiento de vanguardia en muchas tareas de procesamiento del lenguaje natural.

Gracias a sus características innovadoras en los inicios de los LLM y a su naturaleza de código abierto, Bert es uno de los LLM más populares y utilizados. Por ejemplo, en 2020, Google anunció que había adoptado Bert a través de Google Search en más de 70 idiomas.

Actualmente hay miles de modelos de Bert de código abierto, gratuitos y preentrenados disponibles para casos de uso específicos, como el análisis de sentimientos, el análisis de notas clínicas y la detección de comentarios tóxicos.

Desafíos y oportunidades

A pesar de los avances, aún existen algunos desafíos que deben abordarse para que las aplicaciones de agentes inteligentes basados en LLM alcancen su máximo potencial:

- **Sesgo**: Los LLM pueden ser sesgados, lo que puede afectar negativamente al rendimiento y la precisión de las aplicaciones que los utilizan. Esto mismo puede propiciar la discriminación de todo tipo y aumentar las desigualdades.
- **Seguridad**: Los LLM pueden ser utilizados para crear contenido malicioso, como deepfakes o noticias falsas.
- **Interpretabilidad**: Es difícil entender cómo funcionan los LLM, lo que dificulta la depuración de errores y la evaluación de su confiabilidad.

Sin embargo, las oportunidades que ofrecen los LLM para el desarrollo de agentes inteligentes son enormes. A medida que se superen los desafíos mencionados, podemos esperar ver una proliferación de aplicaciones innovadoras en diversos sectores.

Conclusión

El ámbito de los LLM es vasto y está en constante expansión, y cada nuevo modelo supera los límites de lo que es posible. A medida que continuamos siendo testigos de rápidos avances en este campo, está claro que los modelos de código abierto desempeñarán un papel crucial en la configuración del futuro de la IA.

Por eso para mantener un avance de manera responsable, es crucial continuar investigando y desarrollando técnicas para mejorar la calidad del discurso generado por los agentes inteligentes.

Al hacerlo, podemos aprovechar al máximo el potencial de los LLMs de código abierto para crear aplicaciones de agentes inteligentes que sean verdaderamente útiles, confiables y éticas.

Estos avances están impulsando el desarrollo de aplicaciones de agentes inteligentes en una variedad de campos como la atención médica, la educación, la atención al cliente, el comercio electrónico y muchos más. Por esa razón, debemos actuar con responsabilidad y ética y procurar que estas implementaciones tengan un acceso libre y sean beneficiosas para toda la sociedad.

Fuentes

- Los 5 mejores LLM de código abierto (febrero de 2024) - https://www.unite.ai/es/mejores-pel%C3%ADculas-de-c%C3%B3digo-abierto/
- Los 5 mejores modelos de IA (LLM) de Open Source en 2024 - https://openexpoeurope.com/es/los-5-mejores-modelos-de-ia-llm-de-open-source-en-2024-quieres-conocerlos/
- 8 Top Open-Source LLMs for 2024 and Their Uses - https://www.datacamp.com/blog/top-open-source-llms
- OpenAI API - https://beta.openai.com/docs/api-reference/
- Google Cloud Natural Language API - https://cloud.google.com/natural-language/
- Hugging Face Transformers - https://huggingface.co/transformers/
- BigScience Bloom - https://bigscience.huggingface.co/blog/bloom
- Open LLMs - https://github.com/eugeneyan/open-llms
- Criterios para elegir tu modelo de LLM - https://www.youtube.com/watch?v=GwxuFoxRjUE
- Should You Use Open Source Large Language Models? - https://www.youtube.com/watch?v=y9k-U9AuDeM