Generación de contenidos con Inteligencia Artificial

Introducción a los LLM

valgrai.eu

Valencian Graduate School and Research Network of Artificial Intelligence

Índice

1. Introducción a los LLM

Primero, PNL

El **Procesamiento del Lenguaje Natural (PLN)** es un subcampo de la inteligencia artificial (IA) que se centra en la interacción entre ordenadores y humanos a través del lenguaje natural. El objetivo final de la PNL es leer, descifrar, comprender y dar sentido al lenguaje humano de una manera valiosa.

El NLP consta de dos componentes principales: Comprensión del Lenguaje Natural (NLU) y Generación del Lenguaje Natural (NLG). La comprensión implica tareas como la **traducción** de idiomas, el **análisis** de sentimientos y la **respuesta** a preguntas, mientras que la generación consiste en **crear frases** y oraciones significativas desde cero, como en chatbots o resúmenes de texto.

1. Introducción a los LLM

Deep Learning: un punto de inflexión en PNL

Antes del aprendizaje profundo, las técnicas tradicionales de PNL tenían dificultades para comprender el contexto y los matices del lenguaje. El aprendizaje profundo, con su capacidad para aprender representaciones jerárquicas, ha mejorado significativamente la capacidad de las máquinas para comprender las sutilezas del lenguaje humano.

Con la introducción de la arquitectura **Transformer**, que utiliza mecanismos de atención para sopesar la importancia de las palabras en una oración, el aprendizaje profundo dio otro gran salto en la PNL. Esto llevó a modelos como GPT y BERT que han demostrado un rendimiento notable en la comprensión y generación de texto similar al humano.

A pesar de estos avances, el aprendizaje profundo en PNL aún enfrenta desafíos, como el requerimiento de grandes cantidades de datos y la dificultad para interpretar los modelos.

2. Definición de LLM

Qué es un modelo de lenguaje (grande)

Un modelo de lenguaje es un tipo de modelo de inteligencia artificial que entiende, genera y trabaja con lenguajes humanos. Se entrena con grandes cantidades de datos de texto y aprende a predecir la probabilidad de la siguiente palabra en una oración.

Imagina que estás tratando de adivinar la siguiente palabra en una oración, "El gato es muy ___". Basándote en tus conocimientos de lenguaje, puedes adivinar "lindo" o "pequeño". Los modelos de lenguaje hacen lo mismo, pero su "conocimiento" proviene de los datos con los que han sido entrenados.

La implementación común actual de los LLM se basa en la arquitectura de red neuronal llamada **transformer**.

2. Definición de LLM

Más información sobre los LLM

Los modelos de lenguaje se utilizan en una amplia gama de aplicaciones, desde la corrección ortográfica y gramatical, las funciones de autocompletar, la clasificación de textos, la traducción, hasta los chatbots y los asistentes virtuales. Cada vez que un sistema necesita comprender o generar lenguaje humano, es probable que un modelo de lenguaje esté funcionando.

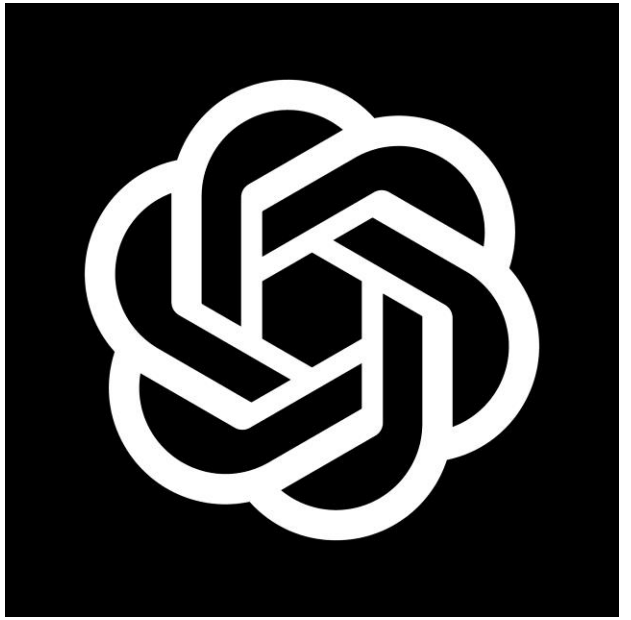
Pueden escribir ensayos, responder preguntas e incluso crear poesía. Sin embargo, **no "entienden" realmente el lenguaje de la forma en que lo hacen los humanos**, simplemente son muy buenos para detectar patrones en los datos con los que han sido entrenados.

3. Entrenando LLMs

Diferentes enfoques en función de los objetivos

El proceso de entrenamiento de un LLM depende de cuál sea su objetivo. Todos ellos tienen en común principalmente la primera etapa de entrenamiento, pero luego podemos encontrar diferencias en función de los objetivos finales y los diferentes planteamientos.

Nos centraremos principalmente en el caso de la familia ChatGPT.



1. **Aprendizaje autosupervisado:** La primera fase del entrenamiento de ChatGPT implica el aprendizaje autosupervisado. En este paso, el modelo se entrena para **predecir la siguiente palabra de una oración**. Esto se hace proporcionándole grandes cantidades de datos de texto, y aprende tratando de predecir cada palabra en función del contexto de las palabras anteriores. Esta etapa permite a ChatGPT aprender gramática, datos sobre el mundo, así como algunos de los sesgos en los datos de entrenamiento.

3. Entrenando LLMs

Diferentes enfoques en función de los objetivos

2. **Ajuste fino supervisado basado en instrucciones:** Después de la fase de aprendizaje autosupervisado, ChatGPT se somete a un proceso de puesta a punto. En esta etapa, **revisores humanos** proporcionan una guía instructiva para el modelo, siguiendo pautas específicas. En el caso de ChatGPT, se trata de un proceso supervisado llevado a cabo por humanos reales que proporcionan las respuestas esperadas a la entrada proporcionada, por lo que se utiliza para cambiar el modelo de forma que, en lugar de limitarse a proporcionar la siguiente colección de palabras más probable, las palabras se alinean con lo que realmente esperamos en una interacción basada en el diálogo.

Ejemplo:

- Un LLM base delante de una entrada como '¿Cuál es la capital de Francia?', la respuesta podría ser '¿Cuál es la ciudad más grande de Francia?' como la forma más probable de seguir la oración.
- Un LLM afinado por instrucciones, después de un proceso de ajuste fino supervisado, podría proporcionar como respuesta 'París'

3. Entrenando LLMs

Diferentes enfoques en función de los objetivos

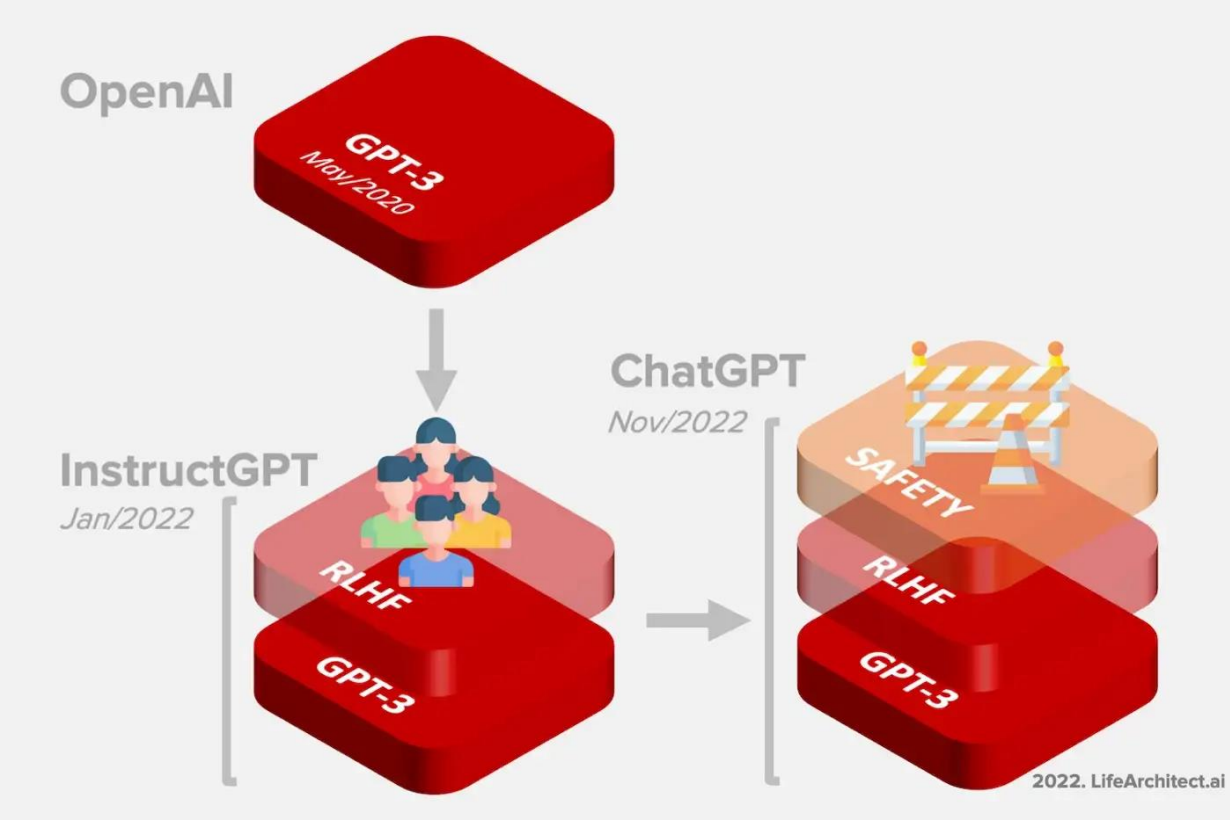
3. **Reinforcement Learning from Human Feedback (RLHF):** La etapa final del entrenamiento de ChatGPT utiliza un método llamado Reinforcement Learning from Human Feedback (RLHF). Aquí, los revisores clasifican las respuestas de los diferentes modelos según su **calidad**. Esta retroalimentación se usa para crear un modelo de recompensas, y el modelo se ajusta para optimizar estas recompensas mediante la optimización de políticas próximas. Este enfoque permite que el modelo generalice a partir de los comentarios de los revisores y mejore su capacidad para responder a una amplia gama de entradas de los usuarios.

El proceso de ajuste y RLHF es altamente iterativo, lo que implica interacciones y aclaraciones continuas con los revisores. Este bucle de retroalimentación continua permite que el modelo mejore con el tiempo.

Esta etapa también permite maximizar **la alineación de la IA**: Cerrar la brecha entre los valores y objetivos humanos y la IA. La alineación de la IA se refiere al proceso de garantizar que los sistemas de inteligencia artificial se comporten de una manera que sea beneficiosa para los humanos y en línea con nuestros valores y estándares éticos. Implica diseñar y entrenar modelos de IA para comprender y respetar las intenciones y los principios humanos.

3. Entrenando LLMs

ChatGPT



4. Memoria y LLMs

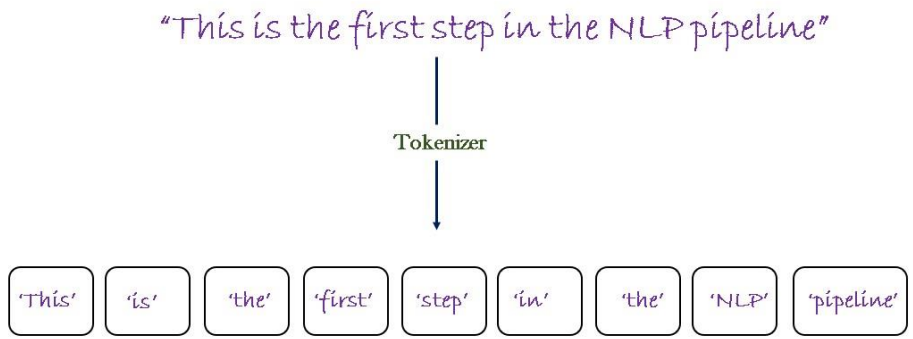
Memoria, límites de tokens e información extra de interés

Hablemos de tokens.

En el contexto de los modelos de lenguaje grandes (LLM), un "token" es una unidad de datos que el modelo lee al generar texto.

Esto podría ser tan pequeño como un carácter, o tan grande como una palabra, o incluso más grande. El tamaño exacto de un token puede variar según el idioma y el modelo específico.

Por ejemplo, en inglés, un token suele tener alrededor de 4 caracteres de media, por lo que un límite de 4.000 tokens correspondería aproximadamente a unos 16.000 caracteres, o unos 3.000-4.000 palabras.



4. Memoria y LLMs

Memoria, límites de tokens e información extra de interés

El **límite de tokens de un modelo** afecta tanto a las peticiones de datos que se le pueden dar como a las respuestas que se pueden generar.

En términos de ingeniería de prompts, el límite de tokens restringe la longitud del prompt puede proporcionar. Si un prompt es demasiado largo y supera el límite de tokens del modelo, deberá acortarlo antes de que el modelo pueda procesarlo. Esto podría implicar la eliminación de detalles menos importantes, el resumen de puntos complejos o la división de un mensaje grande en varios mensajes más pequeños.

El límite de tokens afecta a la cantidad de **historial de conversaciones** recientes que el modelo puede tener en cuenta al generar una respuesta. Si una conversación supera el límite de tokens del modelo, el modelo solo tendrá en cuenta los tokens más recientes hasta su límite. Esto significa que el modelo podría olvidarse de las partes anteriores de la conversación. Por lo tanto, para conversaciones prolongadas o tareas complejas que requieren mantener mucho contexto, los modelos con límites de token más altos pueden ser ventajosos.

4. Memoria y LLMs

Memoria, límites de tokens e información extra de interés

En el momento de escribir esta información, tenemos ahora mismo estos límites de tokens con los LLM comerciales más populares (los rangos de soluciones de código abierto difieren bastante):

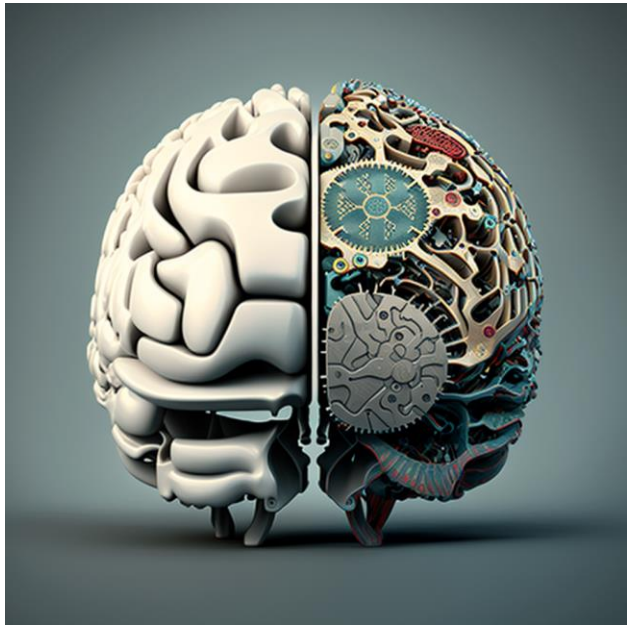
- OpenAI GPT-4: 8192 tokens (8K), 4K GPT-3, pronto 32K tokens (esto podría ser unas 25.000 palabras, o unas 50 páginas de un libro)
- Anthropic Claude: muy pronto 200.000 tokens en su última versión.
- Bard de Google: 1000 tokens

Tenga en cuenta que estos límites de token se aplican a cada interacción con el modelo, abarcando tanto el mensaje de entrada como la salida del modelo. Por ejemplo, si utiliza el modelo GPT-4 estándar y proporciona un mensaje de 4000 tokens, el modelo podría generar hasta 4192 tokens en respuesta.

4. Memoria y LLMs

Memoria, límites de tokens e información extra de interés

Vamos a tratar de explicar las diferentes estrategias y conceptos relacionados con la memoria y la información en los LLM y hacer una analogía con los humanos:



1. Memoria de entrenamiento de redes neuronales: Esto es como la memoria a largo plazo de un ser humano.

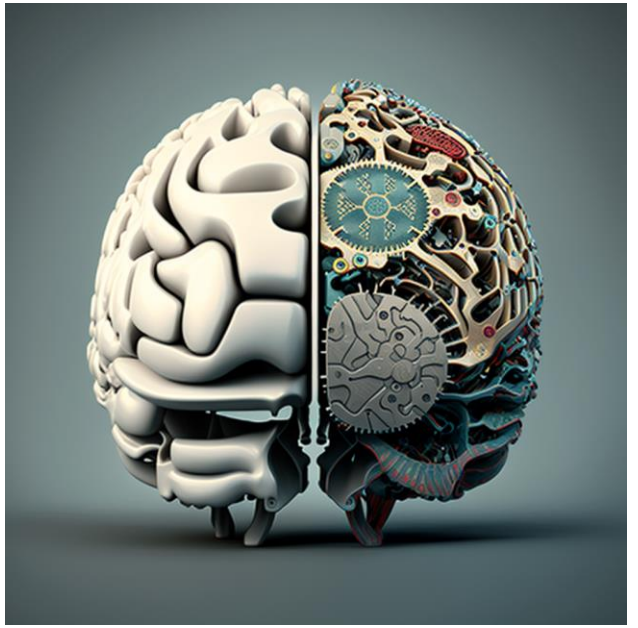
Al igual que aprendemos y recordamos hechos, conceptos y experiencias a lo largo de nuestras vidas, la red neuronal de un LLM aprende patrones y estructuras en los datos durante el proceso de entrenamiento. Puede recordar esta información al generar respuestas. Sin embargo, al igual que la memoria humana a largo plazo, no es perfecta: el modelo puede recordar cosas de forma incorrecta o incompleta, y no recuerda puntos de datos o fuentes específicas, solo los patrones generales.

4. Memoria y LLMs

Memoria, límites de tokens e información extra de interés

2. Memoria de tokens y aprendizaje en contexto: Esto se parece más a la memoria de trabajo de un ser humano o a la memoria a corto plazo.

Al igual que tenemos en cuenta la información reciente mientras mantenemos una conversación o trabajamos en una tarea, un LLM realiza un seguimiento de los tokens recientes que ha procesado (el "contexto") al generar una respuesta. El límite simbólico es análogo a la capacidad de la memoria de trabajo humana, que también es limitada. Si el contexto supera el límite de tokens del modelo, es como intentar hacer malabarismos con más elementos de la memoria de trabajo de los que puede manejar: algunos de ellos se descartarán inevitablemente.



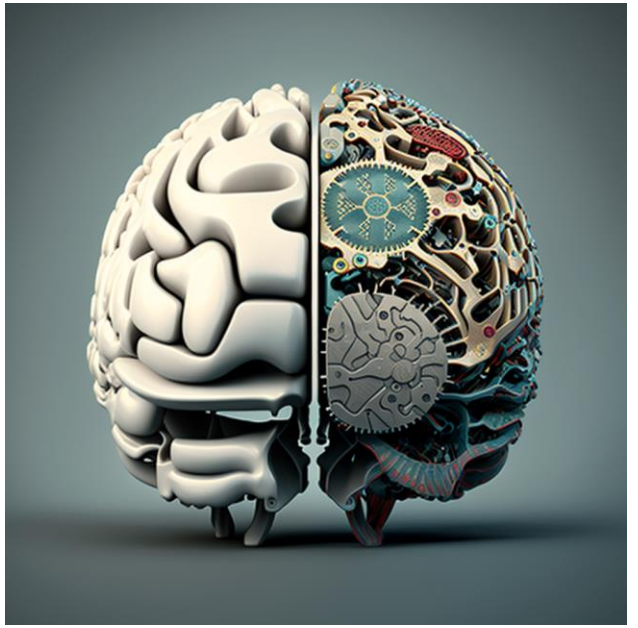
4. Memoria y LLMs

Memoria, límites de tokens e información extra de interés

3. External Knowledge Bases y Document Embeddings: Estas estrategias pueden considerarse similares al uso de materiales de referencia o a la búsqueda de información en Internet.

Al igual que no recordamos todo y a veces necesitamos consultar un libro, un conjunto de notas o un motor de búsqueda para encontrar la información que necesitamos, un LLM puede utilizar bases de conocimiento externas e incrustaciones de documentos para acceder a información que no está en sus datos de entrenamiento o contexto reciente. Por ejemplo, algunos LLM más nuevos pueden usar incrustaciones de documentos para hacer referencia a documentos específicos en una base de datos externa, que es como sacar un libro o documento específico para consultarlo cuando lo necesite.

En cada uno de estos casos, el LLM utiliza un tipo diferente de "memoria" para generar sus respuestas, de forma muy parecida a como los humanos utilizamos diferentes tipos de memoria y recursos externos en nuestra vida cotidiana. Sin embargo, es importante recordar que esta es una analogía y que hay muchas diferencias entre la memoria humana y las formas en que los LLM procesan y generan información.



5. Habilidades emergentes y limitaciones

Sorpresas y límites de los LLM

Emergentes, en el contexto de la teoría de sistemas y la inteligencia artificial, se refiere al fenómeno en el que el sistema en su conjunto demuestra propiedades, comportamientos o habilidades que no son obvias a partir de los componentes individuales del sistema. Estas habilidades no están programadas explícitamente en el sistema, sino que surgen como resultado de las interacciones entre los componentes del sistema.

Las habilidades emergentes en los grandes modelos de lenguaje (LLM) se refieren a habilidades o competencias que surgen de las complejas interacciones de las muchas capas y parámetros del modelo. Estas habilidades no se programaron directamente en el modelo, sino que se hicieron evidentes como resultado del proceso de entrenamiento del modelo.

Por ejemplo, un modelo como GPT-4 podría desarrollar la capacidad de escribir poesía o generar fragmentos de código. Si bien los desarrolladores no programaron específicamente estas habilidades en el modelo, surgen como resultado del proceso de aprendizaje del modelo, donde se ha expuesto y aprendido de una variedad de datos de texto, incluida la poesía y el código de programación. En muchos casos, el aprendizaje de "pocos ejemplos" también se considera una habilidad emergente.

5. Habilidades emergentes y limitaciones

Sorpresas y límites de los LLM

Sin embargo, también existen límites bien conocidos de los LLM actuales

Razonamiento

Si confías en ChatGPT para tareas complejas, probablemente hayas notado que carece de un "modelo del mundo" y tiene dificultades con varias formas de razonamiento. Es decir, se enfrenta a dificultades para:

- Razonamiento espacial: comprender y manipular las relaciones entre objetos en el espacio físico.
- Razonamiento temporal: razonar y hacer predicciones sobre eventos y su ordenación en el tiempo.
- Razonamiento físico: comprensión y manipulación de objetos físicos y sus interacciones en el mundo real.
- Razonamiento psicológico: comprender y hacer predicciones sobre el comportamiento humano y los procesos mentales.

5. Habilidades emergentes y limitaciones

Surprises and limits of LLMs

Planificación y lógica.

Esta es una de las tareas más difíciles para un LLM. Si el proceso de planificación o lógica significa un proceso exploratorio en grandes espacios de búsqueda o etapas muy complejas, aún se requieren enfoques híbridos.

Matemáticas y aritmética

Del mismo modo, sigue siendo necesario el uso de sistemas externos para apoyar estas tareas.

Alucinación

La alucinación se refiere a los casos en los que la IA genera información que no se basa en sus datos de entrenamiento o en la realidad fáctica. Es esencialmente cuando la IA inventa cosas, a menudo debido a la falta de información precisa o explícita en el mensaje dado o debido a limitaciones inherentes en la comprensión del mundo del modelo

Sesgo y discriminación

Independientemente de las barreras de seguridad que agreguemos, los LLM han sido capacitados en Internet, por lo que es difícil evitar el sesgo.

5. Habilidades emergentes y limitaciones

Surprises and limits of LLMs

Sentido común y modelos del mundo.

Los LLM actuales tienen capacidades limitadas de sentido común, principalmente solo aquellas que pueden obtenerse de la información textual. Carece de modelos del mundo, por lo que, por ejemplo, no entienden realmente la gravedad y las consecuencias de las cuestiones relativas a las acciones en las que se requieren implicaciones gravitatorias.

Los modelos multimodales, entrenados no solo con texto, y los sistemas híbridos, pueden proporcionar alternativas. Todavía hay un debate abierto sobre si LLM será capaz o no, aunque sea el único uso de redes neuronales, de proporcionar soluciones generales.

Valencian Graduate School and Research Network of Artificial Intelligence

valgrai.eu

valgrai.eu

Valencian Graduate School and Research Network of Artificial Intelligence