¿Qué son los modelos de lenguaje de gran tamaño?

Los modelos de lenguaje de gran tamaño, también conocidos como LLM, son modelos de aprendizaje profundo muy grandes que se preentrenan con grandes cantidades de datos. El transformador subyacente es un conjunto de redes neuronales que consta de un codificador y un decodificador con capacidades de autoatención. El codificador y el decodificador extraen significados de una secuencia de texto y comprenden las relaciones entre las palabras y las frases que contiene.

Los transformadores LLM son capaces de entrenarse sin supervisión, aunque una explicación más precisa es que los transformadores llevan a cabo un autoaprendizaje. Es a través de este proceso que los transformadores aprenden a entender la gramática, los idiomas y los conocimientos básicos.

A diferencia de las redes neuronales recurrentes (RNN) anteriores que procesaban las entradas de forma secuencial, los transformadores procesan secuencias enteras en paralelo. Esto permite a los científicos de datos utilizar las GPU para entrenar LLM basados en transformadores, lo que reduce significativamente el tiempo de entrenamiento.

La arquitectura de las redes neuronales del transformador permite el uso de modelos muy grandes, a menudo con cientos de miles de millones de parámetros. Estos modelos a gran escala pueden incorporar cantidades masivas de datos, a menudo de Internet, pero también de fuentes como Common Crawl, que comprende más de 50 000 millones de páginas web, y Wikipedia, que tiene aproximadamente 57 millones de páginas.

¿Por qué son importantes los modelos de lenguaje de gran tamaño?

Los modelos de lenguaje de gran tamaño son increíblemente flexibles. Un modelo puede realizar tareas completamente diferentes, como responder preguntas, resumir documentos, traducir idiomas y completar oraciones. Los LLM tienen el potencial de alterar la creación de contenido y la forma en que las personas utilizan los motores de búsqueda y los asistentes virtuales.

Si bien no son perfectos, los LLM están demostrando una notable capacidad para hacer predicciones basadas en un número relativamente pequeño de indicaciones o entradas. Los LLM se pueden utilizar en la IA (inteligencia artificial) generativa para producir contenido basado en indicaciones de entrada en lenguaje humano.

Los LLM son grandes, muy grandes. Pueden considerar miles de millones de parámetros y tienen muchos usos posibles. A continuación, se indican varios ejemplos:

- El modelo GPT-3 de OpenAI tiene 175 000 millones de parámetros. Su primo, ChatGPT, puede identificar patrones a partir de datos y generar resultados naturales y legibles. Si bien no sabemos el tamaño de Claude 2, puede aceptar entradas con hasta 100 000 tokens en cada indicación, lo que significa que puede funcionar en cientos de páginas de documentación técnica o, incluso, en un libro completo.

- El modelo Jurassic-1 de AI21 Labs tiene 178 000 millones de parámetros y un vocabulario simbólico de partes de 250 000 palabras y capacidades de conversación similares.
- El modelo Command de Cohere tiene capacidades similares y puede funcionar en más de 100 idiomas diferentes.
- El Paradigm de LightOn ofrece modelos básicos con capacidades declaradas que superan las del GPT-3. Todos estos LLM vienen con las API que permiten a los desarrolladores crear aplicaciones únicas de IA generativa.

¿Cómo funcionan los modelos de lenguaje de gran tamaño?

Un factor clave en el funcionamiento de los LLM es la forma en que representan las palabras. Las formas previas de machine learning utilizaban una tabla numérica para representar cada palabra. Sin embargo, esta forma de representación no podía reconocer las relaciones entre las palabras, como las palabras con significados similares. Esta limitación se superó mediante el uso de vectores multidimensionales, también denominados incrustaciones de palabras, para representar palabras de modo que las palabras con significados contextuales similares u otras relaciones estén cerca unas de otras en el espacio vectorial.

Al utilizar incrustaciones de palabras, los transformadores pueden preprocesar el texto como representaciones numéricas a través del codificador y comprender el contexto de palabras y frases con significados similares, así como otras relaciones entre palabras, como las partes del discurso. Entonces es posible que los LLM apliquen este conocimiento del idioma a través del decodificador para producir un resultado único.

¿Qué son las aplicaciones de modelos de lenguaje de gran tamaño?

Hay muchas aplicaciones prácticas para los LLM.

Redacción de textos publicitarios

Además de GPT-3 y ChatGPT, Claude, Llama 2, Cohere Command y Jurassic pueden escribir copias originales. AI21 Wordspice sugiere cambios en las oraciones originales para mejorar el estilo y la voz.

Respuesta a la base de conocimientos

La técnica, que a menudo se denomina procesamiento del lenguaje natural intensivo en conocimiento (KI-NLP), se refiere a los LLM que pueden responder a preguntas específicas a partir de la información en los archivos digitales. Un ejemplo es la capacidad de AI21 Studio Playground para responder a preguntas de conocimiento general.

Clasificación de textos

Mediante la agrupación en clústeres, los LLM pueden clasificar textos con significados o sentimientos similares. Los usos incluyen medir la opinión de los clientes, determinar la relación entre los textos y buscar documentos.

Generación de código

Los LLM dominan la generación de código a partir de indicaciones en lenguaje natural. Algunos ejemplos incluyen Amazon CodeWhisperer y el Codex de Open AI utilizado en GitHub Copilot, que puede codificar en Python, JavaScript, Ruby y varios otros lenguajes de programación. Otras aplicaciones de codificación incluyen la creación de consultas SQL, la escritura de comandos shell y el diseño de sitios web.

Generación de texto

Al igual que la generación de código, la generación de texto puede completar oraciones incompletas, escribir la documentación del producto o, como Alexa Create, escribir un cuento infantil corto.

¿Cómo se entrenan los modelos de lenguaje de gran tamaño?

Las redes neuronales basadas en transformadores son muy grandes. Estas redes contienen varios nodos y capas. Cada nodo de una capa tiene conexiones con todos los nodos de la capa subsiguiente, cada uno de los cuales tiene un peso y un sesgo. Los pesos y los sesgos, junto con las incrustaciones, se conocen como parámetros del modelo. Las grandes redes neuronales basadas en transformadores pueden tener miles y miles de millones de parámetros. El tamaño del modelo generalmente se determina mediante una relación empírica entre el tamaño del modelo, la cantidad de parámetros y el tamaño de los datos de entrenamiento.

El entrenamiento se lleva a cabo mediante un gran corpus de datos de alta calidad. Durante el entrenamiento, el modelo ajusta, de forma iterativa, los valores de los parámetros hasta que predice correctamente el siguiente token a partir de la secuencia anterior de tokens de entrada. Lo hace mediante técnicas de aprendizaje autónomo que enseñan al modelo a ajustar los parámetros para maximizar la probabilidad de los siguientes tokens en los ejemplos de entrenamiento.

Una vez entrenados, los LLM se pueden adaptar fácilmente para realizar múltiples tareas mediante conjuntos relativamente pequeños de datos supervisados, un proceso que se conoce como ajuste fino.

Existen tres modelos de aprendizaje comunes:

- Aprendizaje de disparo cero: los LLM básicos pueden responder a una amplia gama de solicitudes sin entrenamiento explícito, a menudo a través de indicaciones, aunque la precisión de las respuestas varía.
- Aprendizaje de pocos disparos: al proporcionar algunos ejemplos de entrenamiento relevantes, el rendimiento del modelo fundacional mejora de manera significativa en esa área específica.

- Ajuste fino: se trata de una extensión del aprendizaje de pocos disparos en la que los científicos de datos entrenan un modelo fundacional para ajustar sus parámetros con datos adicionales relevantes para la aplicación específica.

¿Cuál es el futuro de los LLM?

La introducción de modelos de lenguaje de gran tamaño, como ChatGPT, Claude 2 y Llama 2, que pueden responder preguntas y generar texto, apunta a interesantes posibilidades en el futuro. De forma lenta pero segura, los LLM están logrando un rendimiento similar al humano. El éxito inmediato de estos LLM demuestra un gran interés en los LLM de tipo robótico que emulan y, en algunos contextos, superan al cerebro humano. A continuación, se mencionan algunas reflexiones sobre el futuro de los LLM:

Mayores capacidades

Por impresionantes que sean, el nivel tecnológico actual no es perfecto y los LLM no son infalibles. Sin embargo, las versiones más recientes mejorarán la precisión y las capacidades a medida que los desarrolladores aprendan a mejorar su rendimiento y, al mismo tiempo, reducir los sesgos y eliminar las respuestas incorrectas.

Entrenamiento audiovisual

Si bien los desarrolladores entrenan a la mayoría de los LLM con texto, algunos han empezado a entrenar modelos con entrada de video y audio. Este tipo de entrenamiento debería conducir a un desarrollo de modelos más rápido y abrir nuevas posibilidades en términos de uso de LLM para vehículos autónomos.

Transformación del lugar de trabajo

Los LLM son un factor disruptivo que cambiará el lugar de trabajo. Es probable que los LLM reduzcan las tareas monótonas y repetitivas de la misma manera que lo hicieron los robots con las tareas de fabricación repetitivas. Entre las posibilidades se incluyen tareas administrativas repetitivas, chatbots de servicio al cliente, y redacción automatizada y simple de textos publicitarios.

IA conversacional

Sin duda, los LLM mejorarán el rendimiento de los asistentes virtuales automatizados como Alexa, Google Assistant y Siri. Podrán interpretar mejor la intención del usuario y responder a comandos sofisticados.

¿Cómo puede ayudar AWS con los LLM?

AWS ofrece varias posibilidades para los desarrolladores de modelos de lenguaje de gran tamaño. Amazon Bedrock es la forma más fácil de crear y escalar aplicaciones de IA generativa con modelos de lenguaje de gran tamaño (LLM). Amazon Bedrock es un servicio totalmente administrado que permite que los LLM de Amazon y de las principales *startups* de

IA estén disponibles a través de una API, de modo que pueda elegir entre varios LLM para encontrar el que mejor se adapte a su caso de uso.

Amazon SageMaker JumpStart es un centro de machine learning con modelos fundacionales, algoritmos integrados y soluciones de ML preintegradas que puede implementar con unos pocos clics. Con SageMaker JumpStart puede acceder a modelos previamente entrenados, incluidos los modelos fundacionales, para realizar tareas como el resumen de artículos y la generación de imágenes. Los modelos preentrenados se pueden personalizar completamente para su caso de uso con sus datos, y puede implementarlos fácilmente en producción con la interfaz de usuario o el SDK.