SISTEMAS INTELIGENTES

Lecture 10:Large language Models - Introducción

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024

Introducción

En el ámbito de la inteligencia artificial (IA), los modelos de lenguaje de gran tamaño (LLMs, por sus siglas en inglés) se han convertido en una fuerza impulsora para el avance en la comprensión y la generación del lenguaje.

Estos modelos, entrenados en conjuntos masivos de datos, han demostrado capacidades impresionantes en tareas como traducción, resumen de textos, generación creativa y respuesta a preguntas, revolucionando la forma en que interactuamos con la información y la tecnología.

Que son los LLMs

- ✓Los LLMs son redes neuronales profundas entrenadas con una gran cantidad de datos de texto.
- ✓Estos modelos aprenden las relaciones entre las palabras y las frases, lo que les permite generar texto, traducir idiomas, responder preguntas y realizar otras tareas lingüísticas de forma similar a los humanos.
- ✓La clave de su poder radica en su capacidad para capturar patrones complejos y relaciones semánticas en el lenguaje natural.
- ✓Un LLM es un modelo estadístico que determina la probabilidad de ocurrencia de una secuencia de palabras en una oración.

Que son los LLMs

1 Computación de alto rendimiento

Los LLMs requieren una gran cantidad de poder de computación para su entrenamiento y operación. Los avances en la arquitectura de las GPUs y los chips de IA han hecho posible el entrenamiento de modelos de lenguaje con miles de millones de parámetros.

2 Grandes conjuntos de datos

Para que los LLMs aprendan patrones complejos en el complejos en el lenguaje, necesitan ser entrenados en entrenados en conjuntos masivos de datos de texto. texto. Estos conjuntos pueden incluir libros, artículos, artículos, código fuente, conversaciones y mucho más. mucho más.

3 Arquitecturas avanzadas

Los LLMs utilizan arquitecturas de redes neuronales neuronales avanzadas como Transformers, que les les permiten procesar secuencias de texto de manera manera eficiente y aprender relaciones complejas complejas entre palabras.

4 Aplicaciones diversas

Los LLMs tienen aplicaciones en una amplia gama de gama de campos, incluyendo traducción automática, automática, generación de texto, respuesta a preguntas, chatbot, análisis de sentimientos y mucho mucho más.

Arquitectura de los LLMs

✓ **Las arquitecturas de los LLMs se basan en diferentes enfoques para modelar el lenguaje. Algunos de los más populares incluyen los modelos de lenguaje autoregresivos, los modelos de lenguaje codificador-decodificador y los modelos de lenguaje Transformers.**

Modelos autoregresivos

Los modelos autoregresivos predicen la predicen la siguiente palabra en una una secuencia basada en las palabras palabras anteriores. Ejemplos de modelos autoregresivos incluyen GPT-3 GPT-3 y Jurassic-1 Jumbo.

Modelos codificador-decodificador

Los modelos codificador-decodificador primero codifican una entrada de texto en una representación vectorial y luego la decodifican para generar una salida de texto. Ejemplos incluyen T5 y BART.

Modelos Transformers

Los modelos Transformers utilizan mecanismos de atención para aprender las relaciones entre las palabras en una secuencia de texto. BERT y RoBERTa son ejemplos de modelos Transformers.

Modelos autoregresivos

- ✓ **Los modelos de lenguaje autoregresivos funcionan generando texto de forma secuencial, palabra por palabra. Estos modelos usan el contexto de las palabras anteriores para predecir la siguiente palabra en una secuencia.**
- ✓ **Los modelos autoregresivos se entrenan mediante una función de probabilidad que maximiza la probabilidad de una secuencia de palabras dada la secuencia anterior.**

1

Entrada de texto

El modelo recibe una entrada de texto como una secuencia de palabras.

2

Predicción de la siguiente palabra

El modelo utiliza las palabras anteriores en la secuencia para predecir la siguiente palabra con mayor probabilidad.

3

Generación de texto

La predicción de la siguiente palabra se agrega a la secuencia de texto, y el proceso se repite hasta que se genera una secuencia completa de texto.

Modelos de lenguaje codificador-decodificador

- ✓ **Los modelos de lenguaje codificador-decodificador están diseñados para tareas que requieren la traducción entre dos representaciones diferentes de texto. Estos modelos constan de dos partes principales: un codificador y un decodificador. El codificador convierte la entrada de texto en una representación vectorial, mientras que el decodificador usa esa representación para generar una salida de texto.**

Codificador

El codificador procesa la entrada de texto y la convierte en una representación vectorial.

Decodificador

El decodificador recibe la representación vectorial y la usa para generar la salida de texto.

Salida de texto

El modelo genera una salida de texto que puede ser una traducción, un resumen o una respuesta a una pregunta.

Modelos de lenguaje transformers

- ✓ **Los modelos de lenguaje Transformers usan mecanismos de atención para aprender las relaciones entre las palabras en una secuencia de texto. A diferencia de los modelos autoregresivos, que procesan las palabras en orden secuencial, los Transformers pueden procesar todas las palabras en paralelo, lo que permite una mayor eficiencia y precisión.**

Mecanismo de atención	Descripción
Auto-atención	El modelo aprende la importancia relativa de las palabras en una en una secuencia de texto para predecir la salida.
Atención multi-cabeza	El modelo utiliza múltiples cabezales de atención para capturar capturar diferentes aspectos de las relaciones entre las palabras. palabras.

Ventajas y desafíos de las arquitecturas de LLMs

✓Las arquitecturas de LLMs presentan una serie de ventajas y desafíos. Las ventajas incluyen:

Alto rendimiento en tareas lingüísticas

Los LLMs han demostrado un alto rendimiento en una variedad de tareas lingüísticas, incluyendo traducción, resumen de textos, generación creativa y respuesta a preguntas.

Capacidad de aprendizaje de patrones complejos

Los LLMs pueden aprender patrones complejos en el lenguaje natural, lo que les permite generalizar y realizar tareas que no se les han enseñado explícitamente.

Generación de texto de alta calidad

Los LLMs pueden generar texto de alta calidad, que es que es coherente y gramáticamente correcto. La generación de texto de los LLMs a menudo es indistinguible del texto humano.

✓Sin embargo, también hay desafíos:

Requisitos computacionales intensivos

El entrenamiento y la ejecución de los LLMs requieren una gran cantidad de poder de computación, lo que puede ser costoso y difícil de escalar.

Sesgos y errores

Los LLMs pueden reflejar los sesgos y errores presentes en los datos de entrenamiento, lo que puede llevar a resultados sesgados o impredecibles.

Dificultad para explicar las predicciones

A veces es difícil comprender por qué un LLM produce una predicción en particular, lo que puede ser un problema en aplicaciones donde la transparencia es esencial.

Aplicaciones de los LLMs

✓Los LLMs tienen aplicaciones en una amplia gama de campos, incluyendo:

Chatbots y asistentes virtuales

Los LLMs pueden usarse para crear chatbots y asistentes virtuales más inteligentes y capaces de interactuar con los usuarios de forma natural y fluida.

Traducción automática

Los LLMs han revolucionado la traducción automática, permitiendo traducciones más precisas y naturales entre diferentes idiomas.

Generación creativa de contenido

Los LLMs pueden usarse para generar contenido creativo, como poemas, historias, guiones y código fuente, abriendo nuevas posibilidades para la creación artística y técnica.

Búsqueda de información y respuesta a preguntas

Los LLMs pueden ayudar a los usuarios a encontrar información de forma más eficiente y a obtener respuestas a sus preguntas con mayor precisión.

Consideraciones éticas y de privacidad

✓**El desarrollo y la implementación de LLMs presenta importantes consideraciones éticas y de privacidad. Es fundamental abordar cuestiones como:**

1 Sesgos en los datos de entrenamiento

Los LLMs pueden reflejar los sesgos sesgos presentes en los datos de entrenamiento, lo que puede llevar a la a la generación de texto discriminatorio discriminatorio o inapropiado.

2 Protección de la privacidad

Los LLMs a menudo se entrenan con datos personales, lo que plantea preocupaciones sobre la privacidad de la información. Es importante desarrollar mecanismos para proteger la privacidad de los datos.

3 Uso responsable de la tecnología

Es fundamental promover el uso responsable de los LLMs para evitar la evitar la propagación de información información falsa, el abuso o la manipulación.

GPT (Generative Pretrained Transformer)

Arquitectura: Transformer unidireccional. La arquitectura Transformer se basa en mecanismos de atención, donde GPT utiliza una variante causal de la atención. Esta arquitectura permite que el modelo prediga la siguiente palabra en una secuencia utilizando únicamente el contexto previo.

Modelo: GPT sigue un enfoque de preentrenamiento y ajuste fino. Primero, el modelo se entrena con grandes cantidades de datos textuales sin etiquetas (preentrenamiento) y luego se ajusta para tareas específicas mediante conjuntos de datos etiquetados (fine-tuning). El modelo es autoregresivo, lo que significa que genera texto secuencialmente, palabra por palabra.

Soporte: Se ejecuta en infraestructuras de alto rendimiento, como GPU y TPU. Frameworks como PyTorch y TensorFlow son comúnmente utilizados para implementar y entrenar versiones de GPT. OpenAI también utiliza infraestructura distribuida en múltiples servidores para ejecutar las versiones más grandes.

Aplicaciones: Generación de texto, asistencia en redacción, traducción automática, chatbots, resumen de texto, escritura creativa, e incluso en la investigación de código (Codex para generación de código).

Parámetros: GPT-4: Aunque OpenAI no ha revelado la cantidad exacta de parámetros, se estima que GPT-3 tiene 175 mil millones de parámetros, y GPT-4 es aún mayor, con modelos que probablemente superen los 200 mil millones de parámetros.

Gemini (Google DeepMind)

Arquitectura:Multimodal Transformer. Gemini se basa en una versión mejorada de la arquitectura Transformer, que no solo maneja texto, sino también imágenes, y posiblemente otras modalidades como audio y video. Esto lo convierte en un modelo multimodal, capaz de integrar y procesar información de diferentes tipos de datos.

Modelo:Gemini combina capacidades avanzadas de lenguaje y visión, lo que le permite resolver tareas más complejas que van más allá del texto. Está diseñado para ser altamente escalable y eficiente, beneficiándose de mejoras en el manejo de memoria y la eficiencia computacional de los Transformers. Se preentrena en vastas cantidades de datos y puede ajustarse para tareas específicas multimodales, como la integración de texto e imagen.

Soporte:Principalmente soportado en la infraestructura de TPUs de Google, especialmente optimizado para funcionar en grandes entornos distribuidos en centros de datos de alto rendimiento. Se basa en frameworks internos desarrollados por Google, como JAX y TensorFlow.

Aplicaciones:Procesamiento multimodal que abarca visión y lenguaje, como interpretación de imágenes con texto, generación de contenido visual y textual integrado, y aplicaciones avanzadas en áreas como el análisis de sentimientos y comprensión profunda de imágenes junto con texto.

Parámetros:Aunque Google no ha revelado el número exacto de parámetros de Gemini, se espera que sea comparable o superior a los modelos más grandes de GPT, lo que implicaría un tamaño de alrededor de 500 mil millones de parámetros o más, considerando su enfoque multimodal.

LLaMA (Large Language Model Meta AI)

Arquitectura:Transformer con optimizaciones de eficiencia. LLaMA se basa en la misma arquitectura Transformer, pero con modificaciones que permiten una mayor eficiencia en términos de uso de recursos y memoria. Esto le permite competir con modelos más grandes utilizando una cantidad significativamente menor de parámetros.

Modelo:El enfoque de LLaMA es proporcionar un rendimiento robusto con modelos de menor tamaño. Aunque sigue siendo un modelo preentrenado y ajustado para tareas específicas, su enfoque es reducir el uso de recursos manteniendo una precisión comparable a modelos más grandes. Es altamente eficiente para aplicaciones con limitaciones de hardware.

Soporte:LLaMA está diseñado para ser más flexible en términos de soporte. Se puede ejecutar en GPUs comunes en configuraciones de escritorio, aunque también escala bien en infraestructuras de servidores distribuidos. Es compatible con frameworks como PyTorch y TensorFlow, lo que facilita su implementación en entornos de investigación y producción.

Aplicaciones:Procesamiento de lenguaje natural en tareas comunes como generación de texto, análisis de sentimientos, resumen de texto, y otras aplicaciones de NLP. Es particularmente útil en entornos que necesitan reducir costos computacionales sin perder precisión.

Parámetros:La versión más reciente, LLaMA 2, tiene variantes de diferentes tamaños:

- **LLaMA 2-7B: 7 mil millones de parámetros.**
- **LLaMA 2-13B: 13 mil millones de parámetros.**
- **LLaMA 2-70B: 70 mil millones de parámetros.**

Conclusiones y perspectivas futuras

- ✓Los LLMs representan un avance significativo en el campo de la inteligencia artificial, con aplicaciones en una amplia gama de campos. A medida que la tecnología continúa desarrollándose, podemos esperar ver aplicaciones aún más innovadoras y transformadoras.
- ✓Sin embargo, también es importante abordar las consideraciones éticas y de privacidad relacionadas con la tecnología. A través del uso responsable y la investigación continua, podemos aprovechar el potencial de los LLMs para mejorar nuestra comprensión del lenguaje y la comunicación, impulsando avances en campos como la educación, la investigación y la atención médica.

PREGUNTAS??

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024