Unidad 4.2. Ampliando el chat. Hablar con tus datos y generación aumentada

> Con el avance de la tecnología de la realidad virtual, pronto llegaremos a un punto donde no podremos distinguir entre lo que es real y lo que es un juego
>
> **Elon Musk**, CEO de SpaceX y Tesla y cofundador de OpenAI

Introducción

La inteligencia artificial y, en particular, los modelos de procesamiento del lenguaje natural (PLN), han experimentado avances significativos en estos dos últimos años, avances que se traducen en un crecimiento exponencial de aplicaciones en todos los ámbitos y sectores empresariales, gubernamentales y de cualquier entidad u organización.

Modelos de lenguaje, como GPT, BERT, Llama y otros, han demostrado ser extremadamente potentes para comprender y generar texto en lenguaje natural proporcionando mecanismos para facilitar y automatizar la gestión de la información y del conocimiento. Sin embargo, para aprovechar al máximo su potencial, a menudo es necesario personalizarlos y adaptarlos a conjuntos de datos específicos o a dominios particulares. No hay que olvidar que estos modelos se entrenan con datos de fuentes diversas como Wikipedia pero que no están actualizados por lo que en muchas ocasiones deben tener la posibilidad navegar en internet para accder a contenidos más específicos o actuales. No solo eso sino que hay información poco o nada accesible que los modelos desconocen.

En la siguiente tabla podemos apreciar el coste, tanto en tiempo como en dinero del entrenamiento de los modelos de lenguaje más utilizados

Modelo de Lenguaje	Empresa	Año de Creación	Estimación de Tiempo de Entrenamiento	Estimación de Costo de Entrenamiento	Código Abierto
GPT-3	OpenAI	2020	Varios meses	Millones de dólares	No
BERT	Google	2018	Semanas a meses	Cientos de miles a millones de dólares	Sí
T5	Google	2020	Meses	Millones de dólares	Sí
GPT-4	OpenAI	2023	Meses	Decenas de millones de dólares	No
GPT-2	OpenAI	2019	Semanas a meses	Cientos de miles a millones de dólares	Sí
Transformer	Google	2017	Semanas	Decenas a cientos de miles de dólares	Sí
XLNet	Google/CMU	2019	Semanas a meses	Cientos de miles a millones de dólares	Sí
AlphaFold	DeepMind	2020	Meses	Millones de dólares	Sí
MuZero	DeepMind	2020	Meses	Millones de dólares	No
LLaMA	Meta	2023	No disponible	No disponible	Sí

> Es importante tener en cuenta que la disponibilidad de los modelos como código abierto varía significativamente. Algunos modelos, especialmente los más avanzados como GPT-3 y GPT-4 de OpenAI, no son de código abierto, aunque OpenAI ofrece acceso a través de su API. Por otro lado, muchos modelos desarrollados por Google y otros investigadores académicos suelen ser de código abierto para fomentar la investigación y el desarrollo en la comunidad científica. La información sobre el tiempo y el costo de entrenamiento del modelo LLaMA de Meta no está claramente disponible, ya que la compañía no ha divulgado estos detalles.

Dados los costes inasumibles se requieren métodos de actualización de dichos modelos, métodos que no deben pasar por el re-entrenamiento que es absolutamente inasumible por pequeñas empresas o usuarios individuales.

Para ello existen diversas estrategias o como suele decirse 'workarounds' que están implantándose con rapidez en todos los chatbots actuales.

¿Por qué Personalizar?

La personalización de un modelo de lenguaje es crucial cuando trabajamos con datos específicos de un dominio particular o cuando queremos que el modelo realice tareas muy concretas. Los modelos de lenguaje preentrenados son generalistas; han sido entrenados en grandes cantidades de texto de internet, lo que los hace versátiles, pero no necesariamente expertos en áreas específicas. Personalizar estos modelos con nuestros propios datos nos permite ajustarlos para que se alineen mejor con nuestras necesidades particulares, mejorando así su rendimiento y relevancia.

Transfer-Learning: Hablando con tus Datos

Una de las técnicas más comunes para personalizar modelos de lenguaje es el llamado Transfer-Learning o transferencia de conocimiento.

Este proceso implica tomar un modelo preentrenado y continuar su entrenamiento en un conjunto de datos de un dominio específico, evitando del coste de entranar el modelo de nuevo.

El "transfer learning" o aprendizaje por transferencia, es una técnica en el campo de la inteligencia artificial y el aprendizaje automático. Para entenderlo mejor, podemos usar el símil de un chef aprendiendo a cocinar un nuevo tipo de cocina.

Imagina que un chef ya es experto en cocina italiana. Sabe cómo preparar una variedad de platos italianos y entiende los principios básicos de esta cocina. Ahora, si quiere aprender a cocinar comida japonesa, no necesita empezar desde cero. Puede aprovechar muchas de las habilidades y conocimientos que ya posee, como técnicas de corte, manejo de ingredientes frescos y presentación de platos. Este chef solo necesita aprender las diferencias específicas de la cocina japonesa, como trabajar con ingredientes típicos de Japón o técnicas de cocción únicas para esa cocina.

De manera similar, en el aprendizaje por transferencia, un modelo de IA que ha sido entrenado en una tarea (como reconocer objetos en imágenes) puede reutilizar su conocimiento previo para aprender una nueva tarea relacionada más rápidamente y con menos datos. Por ejemplo, si un modelo se ha entrenado para reconocer automóviles en imágenes, y luego se desea entrenar para reconocer motocicletas, no tiene que aprender desde cero. Puede adaptar lo que ya sabe sobre vehículos y características visuales para aprender la nueva tarea con más eficacia.

Esta técnica, que coloquialmente se suele denominar 'habla con tus datos' ha sido una de las principales derviaciones de la IA tectual al permitir a las organizaciones hablar y procesar información propia de manera muchos más intelegiente y específica.

Hasta hace poco este proceso se hacía mediante programación. Hoy en día, herramientas como chatGPT ya permiten la generación de modelos personalizados para campos específicos.

Por ejempo puedo crear un chatBot personalizado y especialista en el campo de la historia medieval y compartirlo con mis alumnos, o centrar mi chatBot en la programación de videojuegos.

En el proceso de creación de estos chats puedo agregar prompts específicos, urls, bases de datos propias e incluso documentos en pdf, vídeos y audios.

Este proceso hace sólo unos meses era muy complejo y requería conocimientos de programación, sin embargo actualmente ya hay herramientas que lo facilitan enormemente, por supuesto también en chatGPT como podemos apreciar en este vídeo dónde el propio SAm Altman (cofundador de chatGPT) desarrolla un sencillo ejemplo de uso de creación de un chatGPT presonalizado:

https://www.youtube.com/embed/q1dcs0biFWU

Vídeo en el que Sam Altman demuestra la nueva funcionalidad de chatGPT para contruir chatBots personalizados

En la siguiente tabla indicamos algunas de las principales herramientas para ello

Herramienta	Tipo de Datos	Descripción	Características Clave
ChatGPT	Texto	Interfaz de chat para interactuar con grandes cantidades de texto, generando respuestas y análisis.	Procesamiento de lenguaje natural, generación de texto.
ChatDoc	Documentos de texto	Herramienta diseñada para analizar y extraer información relevante de documentos de texto.	Extracción de texto, análisis de contenido de documentos.
ChatPDF	Documentos PDF	Especializada en extraer y analizar información de documentos PDF.	Extracción de texto, análisis de contenido de PDF.
PageChat	Páginas web	Permite interactuar con el contenido de páginas web para extraer y analizar información relevante.	Extracción y análisis de contenido web, fácil de usar.
Chatbase	Bases de datos	Herramienta de análisis y consulta de bases de datos mediante una interfaz de chat.	Interfaz de chat para SQL, análisis de datos.
Dante AI	Análisis de texto avanzado	Herramienta para analizar y obtener insights de grandes volúmenes de texto.	Análisis de texto profundo, aprendizaje automático.

Tableau	Datos visuales	Visualización de datos para crear y compartir cuadros de mando y gráficos interactivos.	Visualizaciones interactivas, integración de datos.
Power BI	Datos de negocios	Herramienta de Microsoft para visualizar datos y compartir insights a través de la organización.	Análisis de datos, informes interactivos.
Google Data Studio	Datos web y marketing	Convierte datos en informes y cuadros de mando personalizables e informativos.	Integración con Google Analytics, fácil de usar.
Domo	Datos empresariales	Combina herramientas para la integración, visualización y colaboración en datos.	Visualización de datos, colaboración en tiempo real.

Aumento de Datos o RAG

El aumento de datos es otra estrategia clave para mejorar el rendimiento de los modelos de lenguaje en conjuntos de datos específicos. Consiste en generar variaciones de los datos de entrenamiento para crear un conjunto de datos más amplio y diverso. Esto puede incluir técnicas como la paráfrasis, la traducción a otros idiomas y la vuelta al idioma original, y la manipulación sintáctica.

"Retrieval Augmented Generation" (RAG), que traducido sería "Generación Aumentada por Recuperación", es una técnica en el procesamiento del lenguaje natural que combina la recuperación de información relevante con la generación de texto. Es una metodología avanzada usada en modelos de inteligencia artificial para mejorar la generación de respuestas más informadas y precisas. Aquí te explico con más detalle:

Componentes de RAG

1. **Recuperación de Información**:
   - En la fase de recuperación, el modelo busca en una gran base de datos o repositorio de documentos para encontrar fragmentos de texto que sean relevantes para la pregunta o el prompt dado.
   - Este repositorio puede incluir una amplia gama de documentos, como artículos de Wikipedia, publicaciones de blogs, libros, etc.
2. **Generación de Respuestas**:
   - Utilizando los fragmentos de texto recuperados, el modelo de lenguaje luego genera una respuesta.
   - Esta generación no es una simple repetición de los fragmentos recuperados, sino que el modelo los utiliza como contexto para construir una respuesta coherente y contextualizada.

Funcionamiento de RAG

- **Integración de Recuperación y Generación**:
  - RAG efectivamente integra dos componentes principales de la inteligencia artificial: un sistema de recuperación de documentos (como un motor de búsqueda) y un modelo de generación de texto (como GPT-3).
  - Cuando se formula una pregunta, primero activa su componente de recuperación para encontrar la información relevante. Luego, el modelo de generación utiliza esta información para formular una respuesta informada.
- **Mejora de la Calidad de las Respuestas**:
  - Al basar sus respuestas en información específica y relevante recuperada, RAG puede proporcionar respuestas más precisas, detalladas y contextualizadas.
  - Esto es particularmente útil para preguntas que requieren conocimiento especializado o actualizado.

Aplicaciones de RAG

- **Asistentes Virtuales y Chatbots**: Mejorando la precisión y relevancia de las respuestas en aplicaciones de conversación.
- **Herramientas de Búsqueda y Análisis de Datos**: Proporcionando respuestas más detalladas y contextualizadas a consultas de búsqueda.
- **Educación y Aprendizaje Automático**: Como una herramienta para generar explicaciones educativas o para responder preguntas de estudio.

Ventajas de RAG

- **Respuestas Basadas en Evidencia**: Al usar documentos y datos reales como base para las respuestas, RAG ofrece una forma de generar respuestas que están respaldadas por evidencia concreta.
- **Adaptabilidad**: Puede adaptarse a una amplia gama de temas y preguntas, gracias a su capacidad para buscar y utilizar información de numerosas fuentes.

Desafíos de RAG

- **Dependencia de la Calidad de los Datos**: La efectividad de RAG depende en gran medida de la calidad y actualidad de la base de datos que utiliza para la recuperación de información.
- **Complejidad y Recursos**: Implementar un sistema RAG efectivo puede ser complejo y requerir recursos computacionales significativos.
- La técnica de "Retrieval Augmented Generation" (RAG) se centra principalmente en el procesamiento del lenguaje natural y la generación de texto. Sin embargo, el concepto subyacente de combinar la recuperación de información con la generación o transformación de contenido puede, en teoría, ser aplicado en el campo de la imagen y el video, aunque con diferentes técnicas y tecnologías. En el contexto de imágenes y videos, el proceso sería diferente y se basaría en técnicas de visión por computadora y aprendizaje profundo. Aquí hay un par de aplicaciones hipotéticas en estos campos:

Aplicaciones en Imágenes

1. **Recuperación y Mejora de Imágenes**:
   - Un sistema podría buscar en una base de datos imágenes similares a una dada y utilizar esa información para mejorar o editar la imagen original (por ejemplo, mejorar la resolución, corregir colores, etc.).
   - Por ejemplo, si se tiene una imagen borrosa, el sistema podría buscar imágenes claras y nítidas con características similares y utilizarlas como referencia para mejorar la calidad de la imagen original.
2. **Generación de Contenido Basado en Imágenes Existentes**:
   - Un modelo podría generar nuevas imágenes o modificar las existentes basándose en características y estilos de imágenes recuperadas de una base de datos amplia. Esto sería útil en diseño gráfico, publicidad, y arte digital.

Aplicaciones en Videos

1. **Mejora y Restauración de Videos**:
   - Similar a las imágenes, un sistema podría mejorar la calidad de un video (por ejemplo, resolución, claridad, estabilización) basándose en datos recuperados de videos de alta calidad.
2. **Generación de Secuencias de Video**:
   - Crear nuevas secuencias de video o editar videos existentes basándose en características y estilos de otros videos. Esto podría aplicarse en la producción de películas, publicidad y realidad virtual.

Consideraciones Técnicas

- **Complejidad de Datos**: Los datos de imagen y video son significativamente más complejos que el texto, lo que requiere modelos más sofisticados y más recursos computacionales.
- **Técnicas de Visión por Computadora**: La implementación de una técnica similar a RAG en imágenes y videos requeriría el uso de avanzadas técnicas de visión por computadora y redes neuronales convolucionales.
- **Desafíos en la Recuperación**: La recuperación de información relevante y útil a partir de imágenes y videos es un desafío significativo debido a la variabilidad y riqueza de los datos visuales.

Aunque el concepto de RAG como tal es específico del procesamiento del lenguaje, sus principios fundamentales de combinar la recuperación con la generación o transformación pueden inspirar enfoques similares en otros campos como el de las imágenes y los videos. Sin embargo, estas aplicaciones requerirían un desarrollo tecnológico considerable y enfrentarían desafíos únicos inherentes a estos medios.

RAG representa un paso adelante significativo en la creación de sistemas de IA más sofisticados y útiles, capaces de manejar preguntas complejas y proporcionar respuestas bien informadas y precisas.

https://www.youtube.com/embed/T-D1OfcDW1M

Vídeo introductorio del concepto de RAG

Consideraciones Éticas y de Sesgo

Al personalizar modelos de lenguaje, es importante tener en cuenta las consideraciones éticas y el potencial de sesgo en los datos. Los modelos aprenden de los datos en los que son entrenados, y si esos datos contienen sesgos, el modelo los replicará. Es crucial ser consciente de esto y tomar medidas para mitigar los sesgos tanto como sea posible.

Vectores de datos (embeddings)

Aunque ya hemos comentado este tipo de objetos en el módulo 2 sobre fundamentos, lo retomamos d enuevo ya que además de ser esencales en el entrenamiento de modelos también se usan para tareas típicas de NLP.

Los almacenes de datos que utilizan datos vectorizados están diseñados para mejorar el rendimiento de las consultas y operaciones analíticas en grandes conjuntos de datos. La vectorización es un método de procesamiento de datos en el que se operan vectores enteros de datos, en lugar de procesar un único elemento de datos a la vez. Esto se alinea con las capacidades de las CPU modernas que pueden realizar operaciones en vectores de datos simultáneamente, resultando en un rendimiento significativamente mejorado. A continuación, se describen algunos de los usos y beneficios de los almacenes de datos con datos vectorizados:

¿Qué son los Word Embeddings?

Los word embeddings son, en esencia, una forma de convertir palabras en vectores numéricos. Imagina que cada palabra es una persona y cada persona tiene una lista de características que la describen. En el caso de los word embeddings, estas características son números. Este proceso permite que las computadoras trabajen con palabras y textos, realizando operaciones matemáticas sobre ellos.

¿Cómo Funcionan?

Para entender cómo funcionan, podemos usar un símil: Imagina un mapa de una ciudad donde cada punto en el mapa representa una tienda. Las tiendas que venden productos similares están más cerca unas de otras. De manera similar, en el espacio de word embeddings, palabras con significados similares están "más cerca" unas de otras en términos numéricos. Por ejemplo, "gato" y "perro" estarían más cerca que "gato" y "avión".

Aplicaciones

1. **Búsqueda y Recomendación de Textos**: Ayudan a encontrar textos similares o relacionados.
2. **Análisis de Sentimientos**: Identifican la emoción o el sentimiento detrás de un texto.
3. **Traducción Automática**: Facilitan la traducción de un idioma a otro.
4. **Asistentes Virtuales y Chatbots**: Mejoran la comprensión del lenguaje humano.

Ventajas

- **Mejor Comprensión del Lenguaje**: Permiten a las máquinas entender mejor las sutilezas del lenguaje humano.
- **Versatilidad**: Son útiles en una amplia gama de aplicaciones de NLP.
- **Eficiencia**: Mejoran la eficiencia en el procesamiento de grandes volúmenes de texto.

Desafíos

- **Contexto Limitado**: Pueden no capturar completamente el contexto en el que se usa una palabra.
- **Sesgo en los Datos**: Pueden heredar y amplificar sesgos presentes en los datos con los que fueron entrenados.

Los word embeddings son una herramienta poderosa en el campo del NLP, proporcionando una manera para que las computadoras "entiendan" y trabajen con el lenguaje humano. Al convertir palabras en vectores numéricos, abren un mundo de posibilidades para el procesamiento y análisis de texto, aunque no están exentos de desafíos y limitaciones. Su uso continuará siendo fundamental en el desarrollo de tecnologías relacionadas con el lenguaje.

Conclusión final

En muchos casos, estas técnicas se utilizan juntas en aplicaciones de NLP. Por ejemplo, un modelo de lenguaje podría ser afinado para una tarea específica, y luego las representaciones vectoriales generadas por este modelo podrían ser almacenadas y consultadas utilizando un almacén de vectores de datos como Pinecone. Esto permite tanto la personalización del modelo (a través del fine-tuning) como la búsqueda eficiente y la similitud semántica (a través del almacén de vectores de datos).

El afinamiento (fine-tuning) y el uso de almacenes de vectores de datos son técnicas complementarias más que excluyentes, y cada una tiene su lugar en el procesamiento del lenguaje natural (NLP).

La personalización de modelos de lenguaje para adaptarlos a nuestros propios datos es un paso crucial para aprovechar al máximo el potencial de la inteligencia artificial en el procesamiento del lenguaje natural. Mediante técnicas como el "fine-tuning", la transferencia de conocimientos, el aumento de datos y la inyección de conocimiento, podemos ajustar los modelos para que se alineen mejor con nuestras necesidades específicas, mejorando así su rendimiento y relevancia en tareas concretas. Sin embargo, es importante abordar este proceso con un enfoque reflexivo y crítico, teniendo en cuenta las consideraciones éticas y los potenciales sesgos en los datos. Con un enfoque cuidadoso y metódico, podemos personalizar los modelos de lenguaje para desbloquear

nuevas posibilidades y obtener insights valiosos de nuestros datos.

Revision #18
Created 7 October 2023 11:14:16 by Pedro López
Updated 7 December 2023 09:08:53 by Luis Hueso