El nuevo juguete favorito de Silicon Valley nos va a costar caro

Tiempo de lectura: 15 min.

Nitish Pahwa

Sáb, 02/09/2023 - 12:47

Hay una gran razón por la cual cada empresa que espera tener algún tipo de relación con la inteligencia artificial está gastando o recaudando miles de millones de dólares en este momento, y no es solo el entusiasmo desbordado de los inversionistas. Estas enormes sumas de dinero son necesarias para cubrir los costos de construir, entrenar y mantener generadores de contenido que consumen mucha energía y recursos, como ChatGPT, así como los conjuntos de datos, redes neuronales y grandes modelos de lenguaje, o LLM, usados para entrenarlos, y que también consumen mucha energía y recursos. Un ejemplo es el GPT-4 de OpenAI, cuya interfaz de programación de aplicaciones (API) fue recientemente puesta a disposición de clientes dispuestos a pagar con experiencia de programación.

Alguien que comprende muy bien el problema de la energía es el propio presidente ejecutivo de OpenAI, Sam Altman. En mayo, mientras testificaba ante el Congreso de Estados Unidos sobre los desafíos causados por la carrera armamentista de la inteligencia artificial que su empresa había iniciado en el mundo, Altman admitió algo curioso: que preferiría que su popularísimo bot ChatGPT, en ese momento la aplicación de crecimiento más rápido en la historia, tuviera menos usuarios. “No estamos tratando de lograr que lo usen más“, afirmó. “De hecho, nos encantaría que lo usen menos, porque no tenemos suficientes GPU”.

Por “GPU”, Altman se refería a las unidades de procesamiento de gráficos, que son los procesadores especializados utilizados para renderizar imágenes en videojuegos, minar Bitcoins y potenciar varios tipos de inteligencia artificial. Debido a la gran popularidad de estos tres sectores, es difícil encontrar GPUs asequibles. Ejecutivos interesados en la inteligencia artificial como Mark Zuckerberg y Elon Musk están acumulando grandes cantidades de GPU en sus empresas, y los inversionistas están buscando fabricantes de chips que puedan producir suficientes unidades para satisfacer la demanda.

La demanda colectiva de GPU ha aumentado tanto que Nvidia ha agotado valiosas unidades como la H100 por el resto del año. Mientras tanto, algunos entusiastas de las criptomonedas están readaptando sus máquinas de minería que consumen mucha energía para usarlas en el entrenamiento de inteligencia artificial, y Google está apostando por sus TPU (unidades de procesamiento tensorial, inventadas por Google específicamente para manejar los requisitos de cálculo para la tecnología de aprendizaje automático).

Ya antes de que la demanda de GPU se disparara, la tecnología no era barata. A principios de este año, Altman admitió a un colega ejecutivo de inteligencia artificial que un “gran margen” de los gastos de OpenAI estaban relacionados con “cómputo”, definido como los recursos técnicos necesarios para entrenar, ajustar y desplegar LLM. En 2018, OpenAI publicó un informe ahora citado con frecuencia titulado ““AI and compute”, que señala que “desde 2012, la cantidad de cómputo utilizada en las ejecuciones de entrenamiento de IA más grandes ha estado aumentando de manera exponencial” y señala que “más cómputo parece conducir de manera predecible a un mejor rendimiento“. El artículo también menciona que “creemos que las ejecuciones de entrenamiento más grandes en la actualidad emplean hardware que cuesta millones de dólares”, incluyendo GPU y TPU. Como es lógico, los modelos avanzados de IA no solo utilizaban cientos de esas unidades, sino que también empleaban versiones de estos modelos con un rendimiento más alto.

En otras palabras: la tecnología que permite a ChatGPT redactar escritos legales inadmisibles y publicaciones de blog llenas de errores en cuestión de segundos utiliza mucho hardware que consume mucha electricidad. Y si estas herramientas son efectivas en este momento, es porque los conjuntos de datos en los que se entrenan no hacen más que aumentar y aumentar –y la infraestructura física en la que funcionan también debe crecer y escalar en consecuencia.

Como es de esperarse, entonces, “los costos de cómputo son exorbitantes” cuando se trata del desarrollo de la inteligencia artificial, como tuiteó Altman en diciembre, explicándole a un usuario entusiasta por qué el ChatGPT, en su mayoría gratuito para usar, tendría que ser “monetizado”. Altman ha estado muy consciente de este hecho durante un tiempo y ha sido notablemente sincero al respecto. “Los costos de cómputo se vuelven significativos para nosotros“, le dijo a un usuario de Twitter en agosto pasado, explicando por qué el generador de imágenes DALL-E 2 de OpenAI aún no tenía un plan de precios más “generoso”.

Esto es clave para entender por qué el sector de la inteligencia artificial se presenta de la manera en que lo hace: está principalmente controlado por corporaciones tecnológicas gigantes que poseen recursos diversos y abundantes, dependen de grandes y constantes flujos de efectivo, tienen esperanzas en proyectos ambiciosos desde hace mucho tiempo en campos como la computación cuántica y la fusión nuclear, menosprecian a competidores más pequeños que no pueden esperar alcanzar los asombrosos avances de las empresas más grandes y son discretos acerca de los factores técnicos detrás de sus insumos energéticos.

Incluso Andreessen Horowitz, la firma de capital de riesgo cuyos fundadores son extremadamente optimistas sobre el futuro de la inteligencia artificial, ha admitido que “el acceso a recursos de cómputo, al costo total más bajo, se ha convertido en un factor determinante para el éxito de las empresas de IA. … De hecho, hemos visto a muchas empresas gastar más de 80% de su capital total recaudado en recursos de cómputo”. Aquí, OpenAI tiene una gran ventaja sobre cualquier competidor recién llegado gracias a miles de millones de dólares de inversión por parte de Microsoft, además de la disposición de esa empresa de invertir sumas considerables en supercomputadoras exclusivas hechas a la medida.

Con el mayor poder ha llegado una menor transparencia. La API de GPT-4 es visible para más partes del mundo, pero el conocimiento público sobre su funcionamiento sigue siendo limitado: cuando el informe de OpenAI sobre el modelo salió en marzo, controversialmente excluyó “detalles adicionales sobre la arquitectura (incluido el tamaño del modelo), hardware, cómputo de entrenamiento, construcción del conjunto de datos, método de entrenamiento”.

El temor constante hacia una conciencia robótica similar a la singularidad tecnológica a menudo no tiene en cuenta los límites físicos muy reales de la inteligencia artificial actual, y como resultado, su impacto muy real en el planeta. Sabemos mucho menos de lo que deberíamos acerca de eso, al tiempo que soportamos temperaturas récord causadas por el cambio climático. No es que no se haya estudiado ni advertido sobre la huella de carbono de la inteligencia artificial: en 2019, mi antigua colega April Glaser entrevistó a un investigador que había copublicado un destacado artículo académico ese año sobre los efectos climáticos de la inteligencia artificial. Pero ese mismo artículo, titulado “Green A.I.”, sigue siendo la principal fuente en la que se basan los reporteros tecnológicos hasta el día de hoy para cuantificar el problema de la inteligencia artificial y el clima. No hace falta decir que mucho ha cambiado en los cuatro años transcurridos desde entonces,

en términos de capacidades tecnológicas, inversión y eficiencia energética (o la falta de esta).

Entonces, si OpenAI y otros actores importantes como Google se niegan a compartir detalles que podrían inspirar un escrutinio sobre su uso de energía en la inteligencia artificial y sus repercusiones ambientales, ¿cómo debemos percibir las capacidades en constante avance de la tecnología y sus contribuciones al cambio climático? Para responder a esta pregunta, desglosemos los componentes exactos de lo que sabemos sobre cómo funciona ChatGPT.

Primero, veamos el fundamento que representa el acrónimo “GPT”: un Generative Pre-trained Transformer o Transformador Generativo Preentrenado. El “transformador” que se señala aquí es “una novedosa arquitectura de red neuronal basada en un mecanismo de autoatención” que fue inventada por Google en 2017. Una red neuronal es, en términos muy simples, un modelo técnico formado por la interconexión de un conjunto de “nodos”, que básicamente son funciones matemáticas individuales, en un arreglo destinado a parecerse al del cerebro humano. (No te preocupes por esto.)

Las redes neuronales han existido por un tiempo, pero lo que hace único al Transformador es que, según Google, cuando se trata de detectar patrones y contextos en el lenguaje, “requiere menos cómputo para entrenarse” que los tipos anteriores de redes neuronales. Podrías alimentar a un Transformador con mucha más información que los modelos neuronales anteriores mediante la introducción de unidades de datos conocidas como “tokens”, que la red puede procesar, comprender y memorizar de manera económica, utilizando mucho menos energía, tiempo y dinero de lo que podrían requerir redes neuronales menos eficientes. Por eso los modelos de inteligencia artificial actuales tienen mejores capacidades predictivas y generativas: muchos de ellos están entrenados ahora en cientos de miles de millones de estos tokens, lo que establece así miles de millones de “parámetros“, también conocidos como las “sinapsis” de las redes neuronales (más sobre eso más adelante).

Eso es lo del “T”, pero ¿qué pasa con el “GP”? La innovación “Generativo Preentrenado” es lo que OpenAI añadió a la invención de Google para el año 2018. “Preentrenado” se refiere a que el Transformador de OpenAI ha sido alimentado con un conjunto de datos específico –en el caso de los modelos GPT, fragmentos de texto extraídos de libros y páginas web–, que el sistema procesa para establecerse

como “aprendido” en varios patrones y contextos de lenguaje, expresados en parámetros. “Generativo” se refiere a la capacidad de estos modelos para, de manera natural, generar texto que es (a menudo) legible y (a veces) coherente, basado en lo que han sido preentrenados a través del Transformador.

Cada parte de este proceso requiere una cantidad considerable de energía. Algunos académicos, al discutir la huella de carbono de la inteligencia artificial, se centran en todas las etapas del desarrollo de la tecnología, desde la obtención de los materiales necesarios hasta su envío a través de cadenas de suministro, pasando por los vuelos que los investigadores individuales de IA realizan para colaborar entre sí o asistir a conferencias. Sin embargo, para nuestros propósitos, mantengamos las cosas simples y concentremos nuestra atención en el proceso que va desde el entrenamiento del sistema de texto hasta la salida final, probada y desplegada en un laboratorio con todas las piezas ensambladas y listas. (Para abordar la generación de imágenes, videos y audio se requeriría un análisis más detallado).

Primero, los datos. En inteligencia artificial, gran parte de los datos de texto se obtienen en línea de varios sitios web utilizando un método de recopilación masiva que a menudo aumenta bruscamente el número de solicitudes enviadas a un sitio específico y puede sobrecargar sus servidores, externalizando así el consumo de energía a los millones de sitios que están siendo rastreados. Los datos recopilados deben ser almacenados en algún lugar. Microsoft y otras empresas que incursionan en la inteligencia artificial están construyendo campus de centros de datos a “hiperescala“, a menudo en ciudades grandes o en regiones europeas con climas más fríos, lo que proporciona la ventaja de moderar naturalmente las temperaturas operativas de estos centros de datos.

La necesidad de tener en funcionamiento constante, mantener y estabilizar estos centros de datos libera cientos de toneladas métricas de emisiones de carbono. En climas cálidos, enfriar los centros de datos no relacionados con la inteligencia artificial requiere miles de millones de galones de agua. La firma de análisis tecnológicos Tirias Research estima que el consumo de energía global de los centros de datos podría aumentar en un 21,200 por ciento en cinco años, generando costos operativos que superen los $76 mil millones (en dólares actuales). Para satisfacer esta creciente demanda de energía de manera sostenible, necesitaremos mucha más energía renovable.

Está el asunto de mantener los datos que has recopilado a mano y listos en todo momento. Y luego está el proceso de entrenar realmente tu red neuronal, que te gustaría que fuera lo más grande posible, quizás incluyendo billones de nodos y parámetros y capas interconectadas. ¿Por qué tan grande? Porque, como señaló OpenAI en el informe mencionado anteriormente en 2018, cuanto más grande sea el modelo, más rápido y preciso será su resultado, o al menos eso es lo que parece demostrar el historial de OpenAI, desde su primer modelo GPT hasta su iteración actual GPT-4.

Como señalaron los investigadores de Meta y de la academia en un artículo de mayo, “los modelos de lenguaje grandes se entrenan en dos etapas: (1) preentrenamiento no supervisado a partir de texto sin procesar, para aprender representaciones de propósito general, y (2) ajuste de instrucciones a gran escala y aprendizaje por refuerzo, para alinearse mejor con las tareas finales y las preferencias del usuario”. En otras palabras: está el primer paso de incorporar montones de datos a partir de los cuales el modelo crece y aprende, y luego está la cuestión de afinar más el modelo después de que termina su primer “preentrenamiento”.

Esto incluye refinar y ampliar el modelo posteriormente, a través de procesos como el ajuste fino y el aprendizaje por refuerzo a partir de comentarios humanos, o RLHF. Lo primero se refiere a la práctica técnica de agregar más datos de ejemplos del mundo real para beneficiar al LLM, de modo que establezca un conocimiento más amplio sin comenzar el entrenamiento desde cero. El RLHF es la forma en que un entrenador humano asiste al entrenamiento, ya sea calificando ciertas partes de la salida o proporcionando datos refinados que (con suerte) ayudarán a producir un resultado deseado. Por ejemplo: ¿ves cuando le haces tus preguntas tontas a ChatGPT y luego haces clic en el ícono de pulgar hacia arriba o hacia abajo según lo que recibas, o le dices explícitamente a ChatGPT que hizo algo bien o mal y le ofreces una manera de corregirse? Eso es RLHF en acción, externalizado hasta tu escritorio o teléfono.

El ajuste fino se lleva a cabo en el extremo de la investigación y desarrollo, pero el RLHF tiene un alcance mayor: son los enorme grupos de trabajadores mal remunerados etiquetando fragmentos de datos para facilitar que la computadora aprenda cosas fácticas, y somos nosotros, los humanos, diciéndole a ChatGPT por qué su resumen de la historia de la energía estaba mal, mal, mal. De hecho, gran parte de la razón de existir de ChatGPT era para que OpenAI pudiera acelerar la

mejora del modelo en el que estaba trabajando, en el caso del chatbot, GPT-3, y llevarlo al siguiente nivel.

Pero cuando se trata de hacer que ChatGPT sea más competente, contar con entrenadores voluntarios dispuestos no significa automáticamente un ahorro de costos. A diferencia del ajuste fino, que modifica directamente la mecánica de una red neutral, tener 100 millones de usuarios realizando RLHF significa que el modelo también se está desplegando simultáneamente para su uso, se está aplicando al mundo real a través de una acción conocida como “inferencia”.

Los GPT pueden tener su preentrenamiento, pero aún requieren cómputo y energía para producir respuestas y párrafos cuando se les solicita. Según el informe de la firma de investigación y consultoría en semiconductores SemiAnalysis, “los costos de inferencia superan con creces los costos de entrenamiento al implementar un modelo a cualquier escala razonable. De hecho, los costos de inferencia de ChatGPT superan los costos de entrenamiento semanal”. Según los cálculos propios de SemiAnalysis, “los costos de operación de ChatGPT son de $694,444 dólares por día en costos de hardware de cómputo”, lo que equivale a aproximadamente 36 centavos por interacción.

Todo eso se suma al costo que llevó simplemente preparar ChatGPT tal como lo conoces. Según el analista de inteligencia artificial Elliot Turner, el costo de cómputo para la ejecución inicial de entrenamiento probablemente sumó $12 millones de dólares, 200 veces el costo de entrenamiento de GPT-2, que solo tenía 1.5 mil millones de parámetros. A principios de 2021, investigadores de Google y la Universidad de California-Berkeley estimaron que solo el entrenamiento de GPT-3 consumió hasta 1,287 megavatios-hora de electricidad, suficiente para alimentar aproximadamente 360 hogares durante un año, y todo eso antes de entrar en la inferencia. Y todo esto es solo para la generación de texto, hay que tener en cuenta que los costos de energía y emisiones aumentan significativamente cuando se trata de generación de imágenes y videos.

Mapear todo esto nos ayuda a comprender la cantidad abrumadora de recursos monetarios y físicos que serán necesarios si se supone que la inteligencia artificial controlará el futuro.

Para muchos desarrolladores, el objetivo actual es asegurarse de que la inteligencia artificial generativa no necesite depender de una infraestructura tan masiva.

Investigadores en la Universidad de Stanford en California están trabajando en el desarrollo de modelos neuronales avanzados que podrían ser aún más eficientes en términos de consumo de energía que los Transformadores, tanto en su entrenamiento como en su implementación. Google y Meta están esperando que un preentrenamiento lo suficientemente avanzado para los LLM pueda reducir la necesidad de un ajuste fino intensivo, lo que haría que la implementación fuera mucho más económica y accesible en formas más pequeñas de hardware. Diferentes partes del proceso de potencia de la inteligencia artificial, como la ubicación y eficiencia de los centros de datos, mejoras en la arquitectura de redes neuronales, atajos en el entrenamiento, obtención de electricidad de cómputo a partir de energía solar, eólica y conexiones nucleares, o de redes alimentadas por energías renovables, pueden ser ajustadas en el camino para reducir el impacto.

Sin embargo, lo que resulta alarmante es que la emoción, la competencia, la energía y el dinero que se están destinando a la inteligencia artificial en este momento amenazan con abrumar y socavar las inversiones que finalmente estamos realizando para mitigar las amenazas del cambio climático. Necesitamos esas fuentes de energía, limpias y sucias, para nuestras necesidades cotidianas mientras hacemos la transición de los combustibles fósiles a energías más verdes; necesitamos esos mismos semiconductores y chips utilizados en los centros de datos y la computación de inteligencia artificial para configuraciones de energía limpia y vehículos eléctricos; necesitamos esas extensiones de tierra que se dedican a los centros de datos de inteligencia artificial para la agricultura, el refugio y el mantenimiento ambiental; necesitamos el agua utilizada para enfriar esos centros de datos para el consumo, el riego y la protección de la vida silvestre; necesitamos aliviar la presión y la demanda en nuestras redes eléctricas, que ya están abrumadas en gran parte debido a eventos climáticos extremos provocados por el cambio climático.

En una línea de tiempo en la que la humanidad hubiera tomado medidas más tempranas y decisivas para prevenir y reducir los daños del calentamiento global, una versión más sostenible de esta carrera de desarrollo de inteligencia artificial podría haber sido posible. Pero en un momento en el que los costos de la inacción ya han contribuido a temperaturas récord, desastres climáticos frecuentes y crisis de biodiversidad que amenazan con trastornar los ecosistemas de la Tierra, la rápida manifestación de esta visión estrecha de la inteligencia artificial parece más difícil de justificar. ~

Este artículo es publicado gracias a la colaboración de Letras Libres con Future Tense, un proyecto de Slate, New America, y Arizona State University.

No.297 / septiembre 2023

Letras Libres

https://letraslibres.com/ciencia-tecnologia/future-tense-inteligencia-ar...