LITORALPRESS

Fecha: 03-03-2025
Medio: El Mercurio
Supl. : El Mercurio - Cuerpo B
Tipo: Noticia general
Título: Por qué el gasto en inteligencia artificial no está disminuyendo

Pág. : 4
Cm2: 813,5

Tiraje: 126.654
Lectoría: 320.543
Favorabilidad: No Definida

Desembolso proviene de las grandes tecnológicas, gobiernos nacionales y capitalistas de riesgo:

Por qué el gasto en inteligencia artificial no está disminuyendo

CHRISTOPHER MIMS
The Wall Street Journal

La enorme demanda de modelos de razonamiento consumirá electricidad, microchips e inmuebles de centros de datos en un futuro previsible.

DOMINIC BUGATTO

Los modelos de razonamiento de IA pueden utilizar fácilmente más de 100 veces más recursos computacionales que los modelos grandes de lenguaje convencionales.

A pesar de un breve período de dudas de los inversionistas, el dinero está fluyendo hacia la inteligencia artificial (IA) proveniente de las grandes compañías tecnológicas, los gobiernos nacionales y los capitalistas de riesgo a niveles sin precedentes. Para entender el porqué, es útil apreciar la forma en que la misma IA está cambiando.

La tecnología se está alejando de los modelos grandes de lenguaje convencionales y se está acercando a los modelos de razonamiento y agentes de IA. Capacitar los modelos grandes de lenguaje convencionales —el tipo que se encuentra en las versiones gratuitas de una mayoría de chatbots de IA— requiere enormes cantidades de energía y tiempo computacional. Pero estamos descubriendo rápidamente formas de reducir la cantidad de recursos que necesitan para funcionar cuando un ser humano recurre a ellos. Los modelos de razonamiento, los que se basan en modelos grandes de lenguaje, son diferentes en que su operación real consume muchas veces más recursos, en términos tanto de microchips como de electricidad.

Desde que OpenAI presentó un avance de su primer modelo de razonamiento, llamado o1, en septiembre, las compañías de IA se han movido de prisa para lanzar sistemas que puedan competir. Esto incluye R1 de DeepSeek, que sacudió al mundo de la IA y las valoraciones de muchas compañías tecnológicas y de energía a principios de este año, y xAI de Elon Musk, que acaba de lanzar su modelo de razonamiento Grok 3.

DeepSeek provocó una especie de pánico porque demostraba que un modelo de IA se podía capacitar por una fracción del costo de otros modelos, algo que podría reducir la demanda de centros de datos y chips de avanzada caros. Pero lo que hizo realmente DeepSeek fue impulsar la industria de la IA incluso en forma más intensa hacia modelos de razonamiento que requieren un uso intensivo de recursos, lo que significa que la infraestructura computacional sigue siendo muy necesaria.

Debido a sus mejores capacidades, es probable que estos sistemas de razonamiento pronto sean la forma predeterminada en que las personas utilicen la IA para muchas tareas. El director ejecutivo de OpenAI, Sam Altman, señaló que la próxima gran actualización del modelo de IA de su compañía incluirá capacidades de razonamiento avanzadas.

¿Por qué los modelos de razonamiento —y los productos de los que forman parte, como los instrumentos de "investigación profunda" y los agentes de IA— necesitan mucha más energía? La respuesta está en cómo funcionan.

Los modelos de razonamiento de IA pueden utilizar fácilmente más de 100 veces más recursos computacionales que los modelos grandes de lenguaje convencionales, escribió en un blog post reciente la vicepresidenta de administración de producto para IA de Nvidia, Kari Briski. Ese multiplicador proviene de modelos de razonamiento que pasan minutos o incluso horas conversando con ellos mismos —no todo lo cual ve el usuario— en una larga "cadena de pensamiento". La cantidad de recursos computacionales que utiliza un modelo es proporcional a la cantidad de palabras generadas, por lo tanto un modelo de razonamiento que genera 100 veces más palabras para responder una pregunta va a utilizar mucha más electricidad y otros recursos.

Las cosas pueden necesitar un uso incluso más intensivo de recursos cuando los modelos de razonamiento tienen acceso a internet, como lo hacen los modelos de "investigación profunda" de Google, OpenAI y Perplexity.

Estas demandas de energía computacional son solo el comienzo. Como un reflejo de eso, Google, Microsoft y Meta Platforms están planeando destinar colectivamente al menos US$ 215 mil millones a gastos de capital —mucho de eso para centros de datos de IA— en 2025. Eso representaría un aumento de un 45% en su gasto de capital en relación al año pasado.

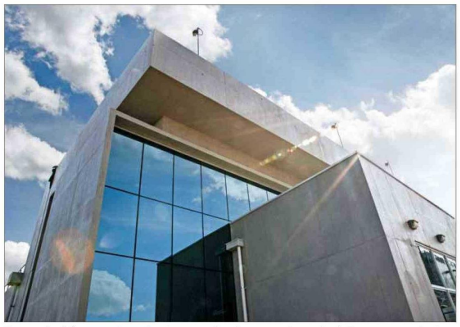
AVANCE
La tecnología se está alejando de los modelos grandes de lenguaje convencionales y se está acercando a los modelos de razonamiento.

Para demostrar las proyecciones de una futura demanda de IA, podemos desplegar una simple ecuación.

El primer valor en nuestra ecuación es la cantidad de recursos computacionales que se necesitan para procesar un solo token de información en una IA como la que alimenta ChatGPT.

En enero, parecía que el costo por token —tanto en energía computacional como en dólares— se vendría abajo inmediatamente después del lanzamiento de DeepSeek R1, el modelo de IA chino. DeepSeek, junto con el artículo que venía, demostraba que era posible capacitar y ofrecer IA en una forma que era radicalmente más eficiente que los enfoques que los laboratorios de IA estadounidenses habían revelado anteriormente.

A primera vista, esto parecería indicar que la demanda futura de energía computacional de la IA sería una fracción de su cantidad actual; es decir, una décima parte, o incluso menos. Pero el aumento en la demanda de los modelos de razonamiento cuando están respondiendo preguntas podría compensarlo con creces. Para verlo en una forma más simplista, si los modelos de IA nuevos, más eficientes basados en la información que se utilizó en DeepSeek disminuyen radicalmente la demanda de energía computacional para la IA en una décima parte, pero los modelos de razonamiento llegan a ser el estándar y aumentan la demanda de esos modelos en un factor de 100, aun así es un incremento de 10 veces en la demanda futura de energía para la IA.



Es probable que el gasto siga creciendo en centros de datos como esta instalación de Google en Santiago de Chile.

Esto es solo el punto de partida. A medida que las empresas están descubriendo que los nuevos modelos de IA son más capaces, los están solicitando cada vez más a menudo. Esto está cambiando la demanda de potencial computacional de capacitar modelos a utilizarlos; o lo que se llama "inferencia" en la industria de IA.

Tuhin Srivastava, director ejecutivo de Baseten, que proporciona recursos computacionales de IA a otras empresas, asegura que este giro hacia la inferencia ya está en marcha. Entre sus clientes hay compañías tecnológicas que utilizan IA en sus aplicaciones y servicios, como Descript, que permite que los creadores de contenido editen audio y video directamente de una transcripción de una grabación, y PicnicHealth, un emprendimiento que procesa registros médicos. Los clientes de Baseten están viendo que necesitan más potencia de procesamiento de IA a medida que la demanda de sus propios productos crece rápidamente, dice Srivastava.

"Para un cliente, bajamos sus costos probablemente un 60% hace seis meses, y dentro de tres meses ya estaban consumiendo a un nivel más alto que en un comienzo", agrega.

Todos los grandes laboratorios de IA en compañías como OpenAI, Google y Meta siguen tratando de ser mejor que el otro mediante la capacitación de modelos de IA cada vez más capaces. Sea cual fuere el costo, el premio es capturar tanto mercado aún incipiente de IA como sea posible.

"Creo que es totalmente posible que los laboratorios importantes tengan que seguir invirtiendo cantidades impresionantes de dinero con el fin de ampliar las fronteras", observa Chris Taylor, director ejecutivo de Fractional AI, una empresa emergente con sede en San Francisco que ayuda a otras compañías de *software* a crear e integrar IA personalizadas. Su compañía, al igual que Baseten y muchas otras en el floreciente ecosistema de IA, depende de aquellos modelos de avanzada para ofrecer resultados a sus propios clientes.

Durante el próximo par de años, las nuevas innovaciones y más microchips específicos de IA podrían significar que los sistemas que entregan IA a clientes finales sean mil veces más eficientes que lo que son hoy, indica Tomasz Tunguz, capitalista de riesgo y fundador de Theory Ventures. La apuesta que los inversionistas y las grandes compañías tecnológicas están haciendo, agrega, es que durante el transcurso de la próxima década, la cantidad de demanda de modelos de IA podría aumentar en un factor de un billón o más, gracias a los modelos de razonamiento y a la rápida adopción.

"Cada pulsación de una tecla en su teclado, o cada fonema que exprese en un micrófono, va a ser transcrito o manipulado al menos por una IA", afirma Tunguz. Y si ese es el caso, agrega, el mercado de IA pronto podría ser mil veces más grande de lo que es hoy.

Artículo traducido del inglés por "El Mercurio".