“¿Queremos quemar el planeta para producir ilustraciones baratas con IA?”

Por: Elena de Sus. 17/04/2025

Wim Vanderbauwhede es profesor de Informática en la Universidad de Glasgow, donde dirige el grupo de investigación de Informática Sostenible y de Bajas Emisiones.

Ha escrito sobre el elevado consumo de energía de los grandes modelos de inteligencia artificial generativa como ChatGPT, cuya expansión tal y como se está planteando considera que “no nos podemos permitir”. Recientemente, se ha mostrado escéptico con la idea de que los avances en eficiencia puedan producir un descenso de las emisiones de esta industria.

Atiende a CTXT por videollamada.

Investiga sobre la informática de bajo consumo. ¿Cómo empezó a interesarse por esta cuestión?

Estoy concienciado sobre el cambio climático desde hace mucho. Al fin y al cabo, esto no es nada nuevo. Nací en Bélgica y cuando vivía allí, hacía trabajo voluntario en una organización ecologista.

> Nuestro modelo social está pensado para animarnos a usar más recursos y más energía, pero ese no es un modelo sostenible

En mi carrera académica, me he centrado en mejorar la eficiencia de los ordenadores. Pero se sabe desde hace mucho que si mejoras la eficiencia de algo, normalmente pasa a ser más barato, así que la demanda aumenta y como hay más demanda, las emisiones de carbono aumentan, no disminuyen.

Toda la historia de la Revolución Industrial ha sido una historia de mejora de la eficiencia. El aumento en la eficiencia de la máquina de vapor nos llevó a quemar muchísimo carbón.

Los ordenadores son, literalmente, millones de veces más eficientes que en los años treinta o cuarenta del siglo XX. Pero eso ha hecho que su uso sea ubicuo. Así que las emisiones totales de la computación han crecido a pesar de todas las mejoras en eficiencia.

Tenía este conflicto con el trabajo sobre la eficiencia y quería contribuir a la sostenibilidad de una forma más amplia. Hace unos años tuve la oportunidad de iniciar una nueva actividad investigadora en el departamento donde trabajo, con el apoyo del jefe de departamento, y así se creó el grupo de Informática Sostenible y de Bajas Emisiones.

El término que uso cuando doy charlas es informática frugal. El mensaje de la informática frugal es que debemos usar menos recursos informáticos, igual que debemos usar menos de cualquier recurso si no queremos un cambio climático catastrófico.

No deberíamos perseguir el crecimiento entendido como crecimiento del consumo de recursos y energía porque eso es destructivo. Nuestro modelo social está pensado para animarnos a usar más recursos y más energía, pero ese no es un modelo sostenible.

El *bitcoin* no es una moneda viable para un Estado nación

Sin embargo, desarrollos recientes como la IA generativa y el *bitcoin* requieren mucha energía.

En algún momento, antes de que empezara la fiebre de la IA, tuvimos la burbuja del *bitcoin* y parecía que el *bitcoin* iba a consumir una gran cantidad de recursos. Pero el *bitcoin* no es una moneda viable para un Estado nación. El exministro de Finanzas griego, Yanis Varoufakis, ha escrito mucho sobre ello. Si hace falta alguna prueba, El Salvador ha abandonado el *bitcoin* como moneda nacional. Esto significa que el *bitcoin* y sus derivados seguirán siendo populares en algunos círculos, pero no crecerán demasiado. Por lo tanto, sus emisiones tampoco aumentarán mucho. Además, otras criptomonedas como Ethereum, basadas en el protocolo *proof of stake,* en lugar de *proof of work*, han crecido en popularidad. Su huella de carbono es 100 veces menor. Así que las emisiones de las criptomonedas no han experimentado un crecimiento espectacular, y el actual nivel de emisiones no es

terrible. Si sigue así no será un gran problema.

La IA generativa es diferente porque cuenta con el apoyo de muchos gobiernos. Todo el mundo parece creer que es mágica y que creará un crecimiento ilimitado. O igual no lo creen, pero actúan como si lo creyeran. Por lo tanto, es un impulso importante para producir más chips, más centros de datos y generar más electricidad. En este momento, el 70% de la electricidad sigue procediendo de combustibles fósiles. Así que vamos a quemar más carbón.

¿Entonces el problema es el apoyo estatal a la IA?

El apoyo estatal retrasa el proceso. Las burbujas normalmente explotan solas porque la gente empieza a darse cuenta de que no hay nada ahí. Pero si los gobiernos piensan que esto es una buena idea, invertirán en ello y estas inversiones se llevarán a cabo incluso si la gente ya se está dando cuenta de que no vale la pena, porque las instituciones son lentas. Así que todo se retrasa. Y con ese retraso, por supuesto, se generan más emisiones.

Hoy en día es muy difícil lograr crecimiento económico. Y si crees que tienes que conseguirlo, cualquier cosa que prometa ese crecimiento te va a interesar. El gobierno del Reino Unido, por ejemplo, es así. También el de los Estados Unidos. Piensan que la IA les va a dar crecimiento, así que están dedicando inversiones a esa área y esas inversiones se producirán aunque la burbuja reviente este año. Y por supuesto, si el gobierno está diciendo que la IA es buena, es mucho más difícil para una persona corriente decir que la IA es mala.

Con el lanzamiento de los modelos de IA generativa de la empresa china DeepSeek, más eficientes, parece que la burbuja ha reventado. Al menos, las acciones de Nvidia han caído y ha habido mucho debate sobre el hecho de que quizás no tenga sentido una inversión tan grande en centros de datos.

He estado analizando la información que DeepSeek ha querido facilitar.

Para empezar, la narrativa de que han tenido que usar GPUs de menor rendimiento por las restricciones a las exportaciones del gobierno estadounidense es falsa. Voy a explicar por qué es falsa.

Para cumplir con las restricciones a las exportaciones de 2022, Nvidia creó una serie especial de GPUs para el mercado chino que rinden peor en un apartado específico.

Se llama rendimiento de punto flotante de doble dirección. Pero la IA no necesita rendimiento de punto flotante de doble dirección.

Lo necesitan los superordenadores que hacen cálculos científicos. Pero los chinos han producido sus propios superordenadores para cálculos científicos, no están comprando GPUs de Nvidia para eso. Las están comprando para la IA, y para la IA esto es irrelevante.

Los modelos de IA de OpenAI, Google y las demás empresas estadounidenses funcionan con unas GPUs de Nvidia llamadas A100. Para entrenar estos modelos, usan unas superiores llamadas H100.

Para el mercado chino, Nvidia vendía las equivalentes A800 y H800 [Nota: Estados Unidos también prohibió la exportación de estas al año siguiente, en 2023]. En su *paper*, DeepSeek dice que su modelo funciona con la H800. La H800 es superior a la A100 en casi todos los aspectos, solo es un poco peor en conectividad. Así que si combinas varias de estas GPUs en una red, el ancho de banda de la red es menor, y en el *paper* DeepSeek explica cómo resolvieron esto. Es una muestra de buena ingeniería, pero no te da un gran beneficio.

Así que no estamos hablando de una capacidad de cómputo restringida. Esto es tope de gama. Es mejor que lo que usan la mayoría de las empresas en sus centros de datos ahora mismo.

> Si el precio de DeepSeek es competitivo, más gente lo usará. Así que es probable que el resultado no sea una reducción del consumo energético

DeepSeek ha sido muy listo con dos cosas. Han lanzado una *app* que le ha gustado a la gente. Su precio es competitivo. Y tienen un montón de modelos más pequeños de código abierto con los que la gente puede jugar. Y creo que esto es en lo que se han fijado los medios, pero tampoco es nuevo. Meta ya había lanzado modelos más pequeños de código abierto con Llama. No son realmente de código abierto, ni unos ni otros, porque los datos que han utilizado no son públicos, pero este es otro tema.

El caso es que el modelo que hace las inferencias principales no es tan pequeño. En comparación con GPT4, por ejemplo, DeepSeek ha conseguido utilizar menos parámetros al mismo tiempo en un momento determinado, así que su modelo será un poco más eficiente energéticamente.

La idea es ingeniosa, han demostrado que funciona y eso es bueno. Pero volvemos al mismo problema. Si su precio es competitivo, más gente lo usará. Así que es probable que el resultado no sea una reducción del consumo energético. Puede ser un aumento si la empresa se hace muy grande.

Se ha puesto mucho el foco en el coste de entrenar estos modelos de IA generativa, pero usted ha escrito que el coste de usarlos es mucho mayor.

Sí. Eso es cierto tanto si hablamos de costes ambientales como de costes financieros. No soy el único que ha escrito sobre esto. Mucha gente está observando que el coste económico del entrenamiento está pasando a ser anecdótico. Yo lo he calculado.

Los costes de inferencia [uso] escalan con el número de usuarios. Los costes de entrenamiento solo escalan si haces un modelo más grande. Aquí es probablemente donde DeepSeek ha sido más inteligente porque su clúster de GPUs no es muy grande, se las arreglaron para entrenar el modelo en un clúster más pequeño, ya que es una empresa pequeña. Eso les permite ahorrar en el coste inicial. Pero si se convierten en una gran empresa, van a necesitar muchos centros de datos para responder a todas las consultas de los usuarios. Ese será el coste dominante.

Hace unos años, los costes de entrenamiento eran mucho más altos porque los modelos se entrenaban de forma poco eficiente. No sabían cómo hacerlo bien. Así que necesitaban muchos recursos para obtener un modelo no muy bueno, y seguramente tenían que repetir los procesos. Pero ahora los costes de la inferencia son los dominantes, definitivamente. Y también las emisiones derivadas de la inferencia.

¿Piensa que los mercados han reaccionado al lanzamiento de DeepSeek de forma exagerada?

Absolutamente, sí. Sobre todo los mercados estadounidenses porque esto viene de China y están asustados. Pero pienso que Nvidia no debería preocuparse.

Quiero decir, por las razones que he explicado, sus ventas dependen más del hecho de que la gente proyecta un crecimiento enorme de la IA.

> No es posible multiplicar la producción de semiconductores por 100 y es probable que toda esta gente lo sepa

Los CEOs de las grandes empresas han estado diciendo que necesitan multiplicar por 100 la fabricación de chips en los siguientes diez años o así. Esas cosas han hecho que las acciones suban. El problema es que los centros de datos ya se están construyendo. Y también las centrales eléctricas para darles suministro porque un centro de datos necesita electricidad en cuanto esté construido.

Así que incluso si nada de esto llega a suceder con la IA, habrán empezado a construir y después querrán usar esas infraestructuras porque si no, habrán hecho un muy mal negocio. Ese es el daño que creo que se está haciendo.

No es posible multiplicar la producción de semiconductores por 100 porque, como mucho, podemos multiplicar la capacidad de minado de los materiales necesarios por dos. Así que esto no va a pasar. Y es probable que toda esta gente lo sepa.

¿Puede que todo el mundo sepa que es una burbuja?

Sí. Pero hace mucho daño porque da a la industria de los combustibles fósiles la excusa perfecta para producir más, por toda la energía que dicen que hará falta para algo que probablemente nunca va a ocurrir.

> Antes de que existiera la IA generativa, la gente no deseaba tenerla. Ha sido un empuje tecnológico, no un tirón del mercado

Usted considera que estos grandes modelos de lenguaje no valen la pena, ¿verdad? Incluso si son útiles para algunas cosas.

Sí, personalmente pienso que la IA generativa que está siendo impulsada por OpenAI y el resto de compañías que compiten con ellos no es muy útil. Quiero decir,

es útil para escenarios específicos, pero cuando tienes un escenario específico puedes usar un modelo mucho más pequeño para hacer lo mismo.

Tenemos estos grandes modelos que pueden hacer de todo para todo el mundo desde 2020 o así, y la productividad global definitivamente no ha crecido.

Las empresas que usan Copilot y otros grandes modelos de lenguaje para programar ven que es problemático, porque es mucho más difícil *debuguear* [corregir] código que no ha sido escrito por tus propios desarrolladores, sino por una máquina. Puedes pensar que el código se escribirá más rápido porque lo hace la máquina, pero la máquina no garantiza que sea correcto. No puede. Un modelo de lenguaje de IA generativa no tiene noción de lo que significan las cosas.

Y hay muchas cosas así. Si te fijas en la IA que genera imágenes, puede parecer brillante, pero en realidad es mediocre. No puede sustituir a los buenos ilustradores porque quien quiera una ilustración de calidad no puede usar eso. ¿Queremos quemar el planeta para producir ilustraciones baratas?

Antes de que existiera la IA generativa, la gente no deseaba tenerla. Ha sido un empuje tecnológico, no un tirón del mercado. El problema es que al crear esta tecnología creamos una gran cantidad extra de emisiones en un momento en el que no podemos permitirnos eso. Las emisiones deben bajar. Si la IA generativa es útil o no es irrelevante. Podría ser extremadamente útil, pero si aun así hace que el planeta arda, no es buena.

Y según mis cálculos, si las proyecciones de estos hombres de negocios se hicieran realidad, la IA por sí sola sería suficiente para saltarnos todos los objetivos climáticos. Como he dicho, es muy improbable que esto pase. Pero están diciendo que no les importaría que pasara. Y no nos podemos permitir ese aumento de las emisiones.

Simple y llanamente.

Podemos permitirnos modelos más pequeños. En informática, hacemos una gran distinción entre lo que preferimos llamar *machine learning* y lo que se está llamando IA, que normalmente es IA generativa.

> El modelo puntero en detección de cáncer de colon, con un 99% de

acierto, tiene 7,6 millones de parámetros, mientras GPT4 tiene más de un billón

De acuerdo. Creo que hay mucha confusión con esto. ¿Podría explicar cuál es la diferencia?

El gobierno del Reino Unido también comete este error. Hablan de que la IA puede hacer grandes cosas como detectar un cáncer en una imagen de una resonancia magnética o en una radiografía y por lo tanto, debemos construir más centros de datos para la IA generativa. Pero SegNet, el modelo puntero en detección de cáncer de colon, con un 99% de acierto, tiene 7,6 millones de parámetros, mientras GPT4 tiene más de un billón.

Esto supone que SegNet utiliza 100.000 veces menos energía que GPT4. Puede funcionar en un PC en el hospital. No necesitas construir ningún centro de datos para obtener mejores diagnósticos. Solo unos pocos servidores en hospitales.

¿Y qué es lo que tienen en común entre esas diferentes cosas que llamamos IA?

La mayoría de los modelos hoy en día utilizan redes neuronales. Una red neuronal es una abstracción inspirada en el cerebro en la que, esencialmente, cada neurona recibe unas señales o inputs, que son números, los multiplica por pesos y luego los suma y luego normaliza el resultado y lo envía a otra neurona. Y si haces eso el suficiente número de veces, obtienes algo que puede hacer extrapolaciones en un espacio de parámetros muy amplio. Así que se le da muy bien… vamos a decir adivinar cosas, pero es hacer aproximaciones estadísticas.

El modelo que se usa para detectar cánceres es una red neuronal convolucional. Esas son las que se usan para imágenes. Las que se usan para textos se llaman redes neuronales recurrentes. En una imagen, los píxeles están uno junto a otro. En un texto, las palabras van una detrás de otra. Los grandes modelos de inteligencia artificial generativa son versiones mucho más avanzadas de estos dos tipos de redes neuronales.

No es lo mismo un modelo que detecta un patrón en una imagen que un modelo generativo que tiene que producir un texto o una imagen nueva. Eso es más trabajo. Por eso los modelos generativos son más costosos en términos energéticos, porque

tienen que hacer más cálculos.

> Si le das a un modelo de IA contenido generado por IA, tiende a empeorar su rendimiento muy rápido. Se llama envenenamiento

He leído que podríamos estar alcanzando un límite en los datos disponibles para entrenar estos grandes modelos, que ya no se pueden encontrar muchos más. No sé si eso es cierto.

Es peor que eso. Ahora hay mucho contenido generado con IA en internet. Esta no es mi especialidad, pero se ha demostrado que si le das a un modelo de IA contenido generado por IA, tiende a empeorar su rendimiento muy rápido. Se llama envenenamiento. No es fácil de evitar porque los *bots* que *scrapean* internet [programas que recogen datos de las páginas web, en este caso para entrenar la IA] no pueden distinguir si una página ha sido generada por IA o no. Eso significa que los mejores datos para modelos generalistas serán los previos a 2022.

Además, no puedes seguir haciendo modelos más grandes, tienes que empezar a hacer cosas como lo que ha hecho DeepSeek. De hecho, OpenAI ya estaba haciendo cosas similares. Tiene un modelo con 1,76 billones de parámetros, pero solo usa 200.000 millones al mismo tiempo. Simplemente porque no es posible acceder a todos a la vez. DeepSeek ha demostrado que se puede hacer bien con menos aún.

En cualquier caso, no puedes seguir haciéndolos más grandes y esperando que su funcionamiento mejore porque hay límites tanto en la calidad de los datos como en la ingeniería necesaria. Así que sí, los resultados probablemente empezarán a estancarse, no llegarán a ser mucho mejores.

> No hay ninguna posibilidad de que un generador de patrones estadísticos se vuelva inteligente

Entonces la idea de que podemos alcanzar una inteligencia artificial general a partir de estos modelos…

Eso es absurdo. Quienes promueven esa idea saben que es una distracción. Se dedican a decir “oh, la inteligencia artificial general será muy peligrosa y tenemos

que adoptar todo tipo de salvaguardas para asegurarnos de que si tenemos una, se comporte correctamente". Esa es la distracción perfecta para no tener que preocuparnos de las verdaderas consecuencias negativas de estos productos.

No hay ninguna posibilidad de que un generador de patrones estadísticos se vuelva inteligente. No hay nada en esos modelos que de verdad imite a la inteligencia.

Llevamos más de 50 años pensando en la inteligencia artificial. Muy profundamente. Y creo que cualquiera que se haya dedicado a ello estaría de acuerdo en que los modelos de IA generativa o cualquier cosa que ahora llamemos IA no son del tipo que nos llevaría a un *software* autoconsciente.

Parecen inteligentes porque todo lo que sabemos está ahí. Se ha introducido un resumen de todo el conocimiento que el ser humano ha puesto online en esos modelos. Así que tienen una aproximación de todo.

LEER EL ARTÍCULO ORIGINAL PULSANDO AQUÍ

Fotografía: CTXT. Wim Vanderbauwhede

Fecha de creación
2025/04/17