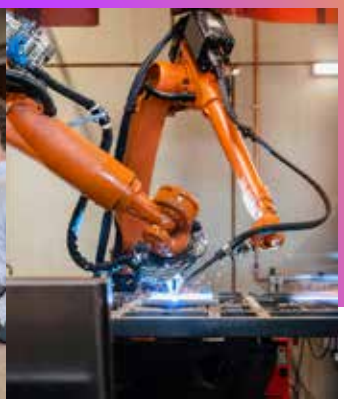
HPE GreenLake

Confía en tu competitividad en el mundo de la inteligencia artificial



Entrenar un modelo de IA puede tardar semanas o meses; aquí te enseñamos cómo evitar este problema tan común.

La inteligencia artificial (IA) cada vez se cita más como solución para cualquier problema de computación, y la gran disponibilidad que existe de herramientas de IA generativa para particulares no ha hecho más que alimentar esa reputación. Sin embargo, las empresas se están dando cuenta de que no es oro todo lo que reluce, ya que implantar un modelo de IA en producción no resulta nada fácil. Una gran parte del problema está relacionada con la limpieza de los datos, el entrenamiento de los modelos y su optimización. ¿Cómo puedes mejorar el proceso de entrenamiento de los modelos e implantarlos en menos tiempo y de forma más precisa en producción, para que la IA te permita realizar innovaciones dentro de la empresa?

Puedes empezar por identificar a los sospechosos habituales.

«El coste es una de las quejas principales», comenta Rangan Sukumar, un distinguido tecnólogo de Hewlett Packard Enterprise. «Entrenar desde cero un gran modelo de lenguaje puede suponer un coste de millones de euros solo en gastos de computación.[1] Además, una vez que el modelo haya sido entrenado, existen más costes asociados con cada consulta adicional. Cuando tienes millones de usuarios introduciendo comandos en el modelo, el coste aumenta desmesuradamente, eso sin mencionar los problemas de sostenibilidad que esto comporta: un modelo de IA como GPT-3 se estima que ha supuesto un consumo eléctrico de 1300 megavatios por hora[2], el equivalente al consumo de 1450 hogares estadounidenses durante un mes. Por lo tanto, en lo que se refiere al entrenamiento de modelos, la forma en la que hemos estado haciendo las cosas hasta ahora no resulta sostenible».

[1] «What Large Models Cost You — There Is No Free AI Lunch» («Lo que nos cuestan los grandes modelos: no hay nada gratis en el uso de la IA»), Forbes, 8 de septiembre de 2023

[2] «New tools are available to help reduce the energy that AI models devour» («Existen nuevas herramientas que ayudan a reducir la electricidad que devoran los modelos de IA»), MIT, 5 de octubre de 2023

El problema del entrenamiento

¿Por qué es tan poco rentable el entrenamiento de modelos? Sukumar afirma que el problema principal al que se enfrentan las empresas en la actualidad es la experimentación que es necesaria para conocer el efecto que tendrá en la empresa la enorme variedad de algoritmos, modelos e infraestructuras de IA. «Tienes cientos de empresas de IA intentando venderte su software, servicios, herramientas y modelos sin conocer con exactitud cuándo y qué es más útil para tu negocio», comenta.

Para ayudar a los desarrolladores a acortar al máximo los plazos para comenzar a utilizar modelos de IA, los proveedores de IA tienen que ser capaces de ofrecer diversas experiencias de producto a usuarios con distinto grado de especialización y para distintas etapas del proceso de implantación de la inteligencia artificial. Esto es algo que ha sido difícil conseguir hasta la fecha, explica Sukumar. La experiencia de entrenamiento para un usuario primerizo de la IA puede y debería ser completamente distinta a la de un usuario ya experimentado de la IA que vaya a crear un nuevo modelo, añade. Sin embargo, se trata como profesionales avanzados de la IA a casi todos los usuarios.

Para complicar aún más las cosas, existen problemas en cuanto a la escalabilidad de los modelos y la incertidumbre que rodea su entrenamiento. «Nosotros contamos con modelos sencillos lo suficientemente pequeños como para caber y poderse entrenar dentro de la memoria de una sola GPU, pero también tenemos modelos que podrían llegar a necesitar miles de GPU», señala Sukumar. Esto introduce la necesidad de trabajar con paralelismo de datos, modelos y pipelines. En estos casos, las tareas de entrenamiento deben dividir los datos de entrenamiento en fragmentos y particiones repartidos entre diversas GPU y coordinar el intercambio simultáneo de dichos fragmentos entre múltiples dispositivos, con el objetivo de que cada uno de ellos actualice el otro. Se trata de una disciplina compleja que requiere un alto nivel de conocimiento y, en ocasiones, suerte; eso sin mencionar los extraordinarios niveles de capacidad informática y de almacenamiento que son necesarios.

Este problema se agrava cuando las organizaciones no conocen bien desde el inicio el modelo de IA que están intentando entrenar. «Casi todo el mundo piensa que puede encontrar un modelo de código abierto de dominio público, descargarlo, entrenarlo con sus datos y que este va a funcionar por arte de magia», dice Sukumar. «Eso no suele ocurrir. En ocasiones, sí que resulta sencillo readaptar un modelo ya entrenado para un caso específico de uso, pero lo normal es que si estás intentando que un modelo de código abierto funcione con los datos confidenciales de tu empresa, te des cuenta de que debes cambiar por completo la arquitectura informática».

Sukumar recuerda un ejemplo publicado en 2020 sobre un modelo de visión artificial que había sido entrenado con los datos de las imágenes médicas de un hospital. El modelo funcionaba bien en el hospital que se encargó de su entrenamiento, pero cuando otros hospitales intentaron usar el sistema, el modelo proporcionaba diagnósticos erróneos. El problema estaba en las diferencias que había entre hospitales a la hora de calibrar el equipo de tomografía axial computerizada (TAC).[3]

«Solo una variación mínima del contraste provocaba que el algoritmo al completo se confundiese al recibir unos datos que no había visto nunca antes», explica Sukumar. «Hay problemas de este tipo por todos lados. La gente comienza a trabajar presumiendo ciertas cosas acerca de los datos y luego esperan que la IA simplemente comience a funcionar bien. Esto no suele ocurrir así, por lo que, por lo general, tienes que retroceder para arreglar cosas».

Otro de los problemas básicos a los que se han tenido que enfrentar las empresas en relación al entrenamiento de IA es la ya consabida falta de personal cualificado. «El nivel de conocimiento necesario para crear desde cero un modelo no suele abundar y resulta caro», dice Sukumar. «Seguramente resulta más caro contratar a alguien con ese nivel de conocimiento que entrenar el modelo mediante el uso de GPU. Crear un modelo desde cero es extremadamente difícil. Para evitar este problema, muchas organizaciones aplican una

[3] «Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans» («Dificultades habituales y recomendaciones para utilizar aprendizaje automático para detectar y pronosticar la COVID-19 mediante el uso de radiografías de tórax y tomografía axial computerizada»), Nature, 15 de marzo de 2021

estrategia de aprendizaje por transferencia y adaptación de otros modelos con buenos resultados para volver a entrenarlos usando datos nuevos».

«Con suerte, el modelo mejorará algo, pero también puede que funcione algo peor», comenta Sukumar. «Si el modelo original se había entrenado para tener una exactitud del 95 % y tú puedes vivir con una exactitud del 90 % después de haberlo entrenado con tus propios datos, esto es algo que parece suficiente para muchas personas. Gran parte del mercado se encuentra en esta situación».

En lo que respecta a los datos, su calidad es otro problema persistente que afecta bastante a todos los aspectos del entrenamiento de modelos. Sukumar comenta, «Es un problema tan grande que, cuando mencionamos otros problemas, no lo incluimos porque damos por hecho que los datos ya están limpios». Casi nunca es así, afirma, por lo que estima que hasta un 70 % del trabajo de entrenamiento de la IA implica preparar los datos para su utilización.

«Durante las primeras charlas que tengo con los clientes sobre entrenamiento de modelos, muchos afirman que su problema principal es la preparación de los datos», comenta Michael Woodacre, director de tecnología para computación de alto rendimiento (HPC) de HPE. «Requiere mucho trabajo, sobre todo en términos de la interacción humana que se necesita».

Woodacre también señala que no se puede descartar el problema del coste. Entrenar un modelo de IA resulta caro porque es necesario contar con un enorme poder de computación, no solo para hacer funcionar las GPU encargadas de dicha tarea, sino también para refrigerarlas. «Tenemos que controlar el gasto energético y las emisiones de carbono, pero, en última instancia, se trata de saber cuánta computación se puede realizar por julio», dice. «Necesitas una plataforma de hardware óptima, el conjunto adecuado de programas de software y los conocimientos necesarios para ejecutar código paralelo de la manera más eficiente posible».

Reducir la complejidad para optimizar el entrenamiento

Entonces, ¿qué podemos hacer para mejorar el proceso de entrenamiento? De acuerdo con Woodacre y Sukumar, unas cuantas cosas. «Estos problemas son parecidos a los desafíos que los especialistas en computación de alto rendimiento (HPC) llevan décadas intentando superar», afirma Woodacre. «Ahora, el mundillo de la IA trabaja para encontrar soluciones a desafíos similares en un plazo de cinco años, aprendiendo rápidamente de la experiencia de la computación de alto rendimiento».

Sukumar añade, «Que el ciclo de la IA comience como un experimento y luego se amplíe para trabajar con conjuntos más grandes de datos y modelos más grandes para después implementar esos modelos en producción, no es una cuestión trivial. La complejidad que supone llevar a cabo operaciones de datos, de desarrollo y de aprendizaje automático, en su conjunto, es mayor que cualquier otra cosa que hayan hecho antes los departamentos informáticos».

Para reducir la complejidad, Sukumar comenta que hay varios servicios en proceso de desarrollo, incluidos muchos de HPE, que permiten a las empresas abstraer muchos de esos desafíos. Contar con diversos modelos de entrega de servicios permite, por ejemplo, organizar paquetes ya preparados de soluciones de IA, tanto en entornos de nube como locales, en lugar de que lo hagan los clientes, o proporcionar acceso bajo demanda a unidades de GPU. Los proyectos de IA se pueden poner en marcha con tan solo unas cuantas líneas de código o por lotes para realizar el entrenamiento a gran escala en superordenadores.

«Podemos reducir la complejidad minimizando el número de herramientas con las que tienen que interactuar los usuarios», comenta Sukumar, lo que es cada vez más una necesidad imperiosa en el desmoralizador panorama actual donde existen cientos de herramientas de IA de código abierto. «Podemos ocuparnos de elegir por ti las mejores soluciones y,

luego, proporcionarte la infraestructura y el software que necesitas para gestionarlas todas». Un partner sólido también te puede asesorar sobre qué modelo específico usar para un proyecto de IA, buscando el equilibrio entre gasto y rendimiento, y dando pasos para preparar el sistema para el futuro.

Los cursos y talleres de formación de los empleados pueden contribuir a superar algunas carencias de conocimientos especializados del personal interno, pero optimizar las operaciones de entrenamiento puede requerir consultas mucho más especializadas. «Nosotros trabajamos directamente con profesionales de vanguardia dedicados a ampliar constantemente los límites de la tecnología», dice Sukumar, «y les ayudamos a conseguir el máximo rendimiento y la máxima eficiencia energética en sus operaciones de IA».

Por último, existe un buen número de herramientas disponibles para ayudar a las empresas a gestionar la parte de los datos, dice Woodacre. «La calidad del modelo depende directamente de la calidad de tus datos. Debes limpiar los datos que vayas a usar en el entrenamiento y registrar el seguimiento cada vez que realices una actualización, teniendo siempre en cuenta problemas como el desfase de datos. Nuestro conjuntos de programas de software HPE Machine Learning Development Environment y HPE Machine Learning Data Management (entorno de desarrollo de aprendizaje automático y gestión de datos para aprendizaje automático) se centran en el desarrollo y entrenamiento de modelos, proporcionando a sus usuarios no solo las herramientas necesarias para entrenar dichos modelos, sino también para realizar el seguimiento y anotar los datos con el fin de conocer mejor qué datos están usando sus modelos de IA. Por ejemplo, si tu modelo tiene que cumplir con ciertos requisitos regulatorios, tienes que saber bien qué se mete dentro del modelo de IA para que puedas reproducir los mismos resultados cuando así sea necesario».

Pisar el acelerador en el proceso de IA para mantener la competitividad

Recuerda que no estás solo si percibes que el mundo de la IA avanza más rápido de lo que te gustaría. Esta capacidad tecnológica está en constante evolución, por lo que ninguna persona sola puede dominar todos los aspectos de este campo.

«Resulta muy complicado para personas individuales estar al tanto de todo», comenta Woodacre. «Las personas no deberían dudar en buscar asesoramiento externo para recibir la ayuda que necesitan con respecto a su proceso de implantación de IA. En caso contrario, es posible que te veas superado por tanta complejidad».

Más información en

HPE.com/AI

Visita **HPE GreenLake**

Chat con ventas

Hewlett Packard Enterprise

a50010121ESE