DELL Technologies

MEDIOS DE COMUNICACIÓN Y ENTRETENIMIENTO EN JAPÓN | CyberAgent

Desarrollo de IA generativa en japonés y transformación de los servicios de publicidad digital

CyberAgent, Inc. utiliza servidores Dell PowerEdge XE9680 con ocho GPU NVIDIA® H100 Tensor Core para acelerar la IA generativa y mejorar la eficacia de la publicidad.

Necesidades empresariales

Desde 2016, CyberAgent, Inc. ha investigado y desarrollando activamente la IA y la ha incorporado en su negocio publicitario. La empresa necesitaba proporcionar a su personal un acceso rápido y asequible a servidores en las propias instalaciones con las GPU NVIDIA más avanzadas disponibles para sus tareas de desarrollo de IA generativa.

Resultados empresariales

- Aceleración del rendimiento de grandes modelos de lenguaje (LLM) en un factor 5,14 aproximadamente frente a la generación anterior con servidores PowerEdge XE9680.
- Se espera una mejora del rendimiento en más de 10 veces en el futuro con las optimizaciones de NVIDIA Transformer Engine.
- Permite el ajuste a alta velocidad de los modelos de aprendizaje automático conforme a los conjuntos de datos más recientes.
- Ahorra espacio en el centro de datos y proporciona refrigeración eficiente con un factor de forma de 6U frente al estándar de 8U.

Resumen de soluciones

- Servidores Dell PowerEdge XE9680 con GPU NVIDIA® H100
- Dell ProSupport

CyberAgent, Inc. es una empresa conocida por ser líder en el mercado del sector de la publicidad en Internet para usuarios domésticos y distintas colaboraciones que incluyen ABEMA, una innovadora plataforma de TV. En 2016, la empresa estableció una organización de investigación de IA llamada AI Lab y, desde entonces, ha investigado y desarrollado IA activamente. En 2020, CyberAgent presentó una IA predictiva de vanguardia que mejora la producción de banners, eslóganes y combinaciones de imágenes de alto impacto para impulsar la eficacia de la publicidad.

CyberAgent continuó su desarrollo de IA generativa para crear un modelo de lenguaje de grandes dimensiones (LLM) específico del idioma japonés con 1300 millones de parámetros. Este LLM está diseñado como modelo de IA de uso general aplicable a distintas situaciones y se puede refinar para crear textos de eslóganes atractivos para los usuarios de cada plataforma publicitaria. CyberAgent ya está utilizando su LLM en japonés en servicios de IA como Kiwami Prediction AI, Kiwami Prediction TD y Kiwami Prediction LP con el fin de asistir en la producción publicitaria creativa y predecir la eficacia de la publicidad. En el futuro, CyberAgent pretende desarrollar una IA multimodal que no solo pueda gestionar LLM en japonés, sino también imágenes.

> **"Nuestros propios investigadores pueden asegurarse una cantidad mayor de recursos y utilizarlos sin tener que preocuparse por el coste. que antes no podían reservar GPU en la cloud pública o se les cobraba más por su utilización a largo plazo".**
>
> **Daisuke Takahashi**
> Arquitecto de soluciones, CIU, departamento de TI del grupo, CyberAgent, Inc.

En mayo de 2023, CyberAgent lanzó una LLM en japonés de código abierto de carácter comercial llamada OpenCALM (Open CyberAgent Language Models), que incluye hasta 6800 millones de parámetros.

Mientras que ChatGPT está optimizado para una conversación, OpenCALM es más bien un modelo de lenguaje japonés de uso general que se puede ajustar según las necesidades de los usuarios. CyberAgent lanzó OpenCALM como proyecto de código abierto porque resulta más beneficioso para la empresa poder recibir comentarios de otras fuentes y colaborar con otras empresas para contribuir al desarrollo de la tecnología de IA en Japón en lugar de desarrollar una LLM en japonés en un entorno cerrado.

La infraestructura que impulsa la innovación de CyberAgent en IA

Cuando CyberAgent estableció su AI Lab en 2016, cada investigador disponía de una estación de trabajo con tecnología de GPU para su investigación. Sin embargo, la necesidad de trabajar a distancia durante la pandemia de 2020 dificultó a los investigadores poder trabajar con sus estaciones de trabajo con GPU. Para garantizar que los investigadores tuviesen los recursos de computación que necesitaban, la empresa empezó a pensar en construir plataformas de aprendizaje automático (ML) centralizadas con servidores basados en GPU en sus centros de dato o en la cloud pública cuando se lanzaron GPU NVIDIA® A100 más recientes.

Daisuke Takahashi, arquitecto de soluciones, CIU, departamento de TI del grupo en CyberAgent, Inc. explica que: "Podríamos haber seleccionado una cloud pública si hubiésemos querido utilizar GPU, pero en una cloud pública no es posible saber cuándo habrá GPU de las más recientes disponibles. Además, no hay ninguna garantía de que las GPU vayan a estar disponibles cuando las necesitamos, así que decidimos implementar recursos de GPU en nuestras propias instalaciones para facilitar el uso. Para hacer realidad la flexibilidad de las infraestructuras para cambiar de ida y vuelta entre la cloud privada y la cloud pública, diseñamos una interfaz de uso lo más cercana posible a las especificaciones de la cloud pública". CyberAgent desarrolló su plataforma de ML inicial en las propias instalaciones utilizando servidores Dell PowerEdge XE8545 con cuatro GPU NVIDIA A100.

Por qué CyberAgent seleccionó los servidores PowerEdge XE9680 con GPU NVIDIA H100

CyberAgent continuó siguiendo la línea de innovación con GPU, especialmente las GPU NVIDIA H100 más recientes. "Pensamos que resultaba atractivo, no solo por su rendimiento mejorado, sino también por mecanismos como su Transformer Engine, que aceleran algoritmos de computación específicos", explica el Sr. Takahashi. "Según NVIDIA, Transformer puede acelerar el entrenamiento de IA de LLM en hasta nueve veces y la inferencia de IA en hasta 30 veces que con las GPU NVIDIA A100 de la generación anterior".

CyberAgent eligió el modelo de servidor PowerEdge XE9680 con ocho GPU NVIDIA H100. El Sr. Takahashi explica que, "Cuando vimos que se iban a lanzar los servidores Dell PowerEdge XE9680 con GPU NVIDIA H100, decidimos adoptarlos lo más rápido posible. Pudimos comunicarnos de cerca con Dell Technologies sobre las configuraciones que iban a ser posibles con los servidores PowerEdge XE9680 y las GPU siguientes. Queríamos aumentar el tiempo de funcionamiento con el mínimo de unidades posible, así que valoramos mucho que Dell Technologies pudiese proporcionar un nivel elevado de mantenimiento, incluido el servicio de cuatro horas in situ, por un precio razonable".

DELL Technologies

Acelera un LLM con 1300 millones de parámetros en un factor 5,14 hoy y un factor mayor que 10 en el futuro.

El Sr. Takahashi continúa, "También seleccionamos los servidores PowerEdge XE9680 porque las instalaciones previas de servidores PowerEdge XE8545 nos habían proporcionado un rendimiento estable y facilidad de mantenimiento. También valoramos el uso de la herramienta de gestión Dell iDRAC para la gestión local y remota segura de los servidores".

El Sr. Takahashi valora el hecho de que después de confirmar el pedido en marzo de 2023, la entrega se completó al cabo de poco más de un mes, a medidos de mayo. "Con las cadenas de suministro afectadas por la pandemia, me dio tranquilidad pensar que Dell Technologies tiene una cadena de suministro relativamente estable, y saber que podían entregar los equipos en tan poco tiempo".

En el período de desarrollo después de la entrega se introdujeron varias innovaciones. El Sr. Takahashi recuerda que, "para un LLM con un gran número de parámetros, necesitábamos utilizar varias GPU, de modo que instalamos una red con ocho tarjetas de interfaz de red (NIC) de 400 Gbit/s en cada servidor y utilizamos la tecnología RDMA (Acceso remoto a memoria directa) para crear una interconexión de alta velocidad entre los servidores. Los servidores con GPU generan mucho calor, así que es importante que estén bien diseñados para poder refrigerarlos de forma eficiente. El factor de forma 6U de los servidores PowerEdge XE9680 para refrigeración sólida también es muy notable. Además, el centro de datos se trasladó a una nueva ubicación, donde hay intercambiadores de calor disponibles en las puertas traseras de los racks, de modo que es posible aplicar una refrigeración eficiente por agua instalando intercambiadores refrigerados por agua en la parte trasera de los racks, en lugar de tener que refrigerar todo el recinto que alberga el centro de datos".

Mejora de la precisión de los eslóganes con las optimizaciones de Transformer Engine

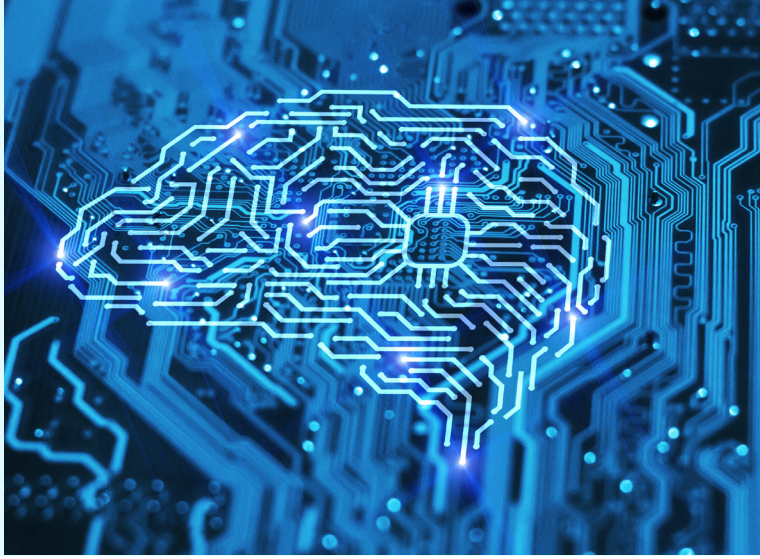
Con la instalación de servidores PowerEdge XE9680, CyberAgent se beneficia de varios aspectos distintos. "Esperamos poder actualizar nuestro LLM más rápido y con más frecuencia gracias a la notable mejora en el rendimiento", explica el Sr. Takahashi. "La velocidad de evolución de los LLM en japonés también mejorará. Además, en comparación con los servidores PowerEdge XE8545 equipados con cuatro GPU NVIDIA A100, los servidores PowerEdge XE9680 con ocho GPU NVIDIA H100 alcanzaron una mejora en el rendimiento en un factor 5,14 aproximadamente. También prevemos un aumento en un factor mayor que 10 del rendimiento con la optimización de NVIDIA Transformer Engine en el futuro. Además, podemos realizar el refinado de modelos de ML a alta velocidad conforme a los conjuntos de datos más recientes, lo que facilita la labor de responder a solicitudes para hacer evolucionar nuestros servicios, mejorar la precisión de los eslóganes y proporcionar un contenido más eficaz".

La infraestructura de ML con tecnología de servidores PowerEdge XE9680 ha recibido muchas alabanzas de los usuarios. "Nuestros propios investigadores han dicho que ahora pueden asegurarse una cantidad mayor de recursos y utilizarlos sin tener que preocuparse por el coste, mientras que antes no podían reservar GPU en la cloud pública o se les cobraba más por su utilización a largo plazo", explica el Sr. Takahashi. "Otro beneficio es que podemos proporcionar una infraestructura con especificaciones exigentes, incluida la interconexión, de modo que los usuarios pueden tener más impacto en el negocio".

El Sr. Takahashi también valora la herramienta de gestión iDRAC de Dell Technologies, que la empresa ha estado utilizando durante un tiempo, porque reduce la carga de gestión. "No siempre estamos trabajando en el centro de datos, así que iDRAC resulta útil para hacer cosas a distancia, como comprobar la temperatura y el estado de las GPU o actualizar el firmware sin tener que acceder al sistema operativo".

> **"El factor de forma 6U de los servidores PowerEdge XE9680 para refrigeración sólida también es muy notable".**
>
> **Daisuke Takahashi**
> Arquitecto de soluciones, CIU, departamento de TI del grupo, CyberAgent, Inc.

> **Esperamos poder actualizar los LLM en japonés más rápido. Los servidores PowerEdge XE9680 con ocho GPU NVIDIA H100 alcanzaron una mejora del rendimiento en un factor 5,14 aproximadamente".**
>
> **Daisuke Takahashi**
> Arquitecto de soluciones, CIU, departamento de TI del grupo, CyberAgent, Inc.



Centrados en LLM, GPU e infraestructura

En el futuro, CyberAgent planea utilizar los comentarios y lo aprendido con OpenCALM para mejorar el LLM que sus empleados utilizan. Mediante OpenCALM, CyberAgent también analiza las colaboraciones con empresas y organizaciones de sectores distintos del de la publicidad. Por ejemplo, CyberAgent ha iniciado conversaciones con actores de los sectores del comercio minorista y las finanzas para desarrollar LLM específicos de sectores que aprendan de sus datos propios de cada sector.

Mientras tanto, el Sr. Takahashi explica que se seguirán actualizando con las GPU más recientes y otras nuevas tecnologías relacionadas para ver cómo se comercializan. "También esperamos ver cómo otros proveedores pueden crear un ecosistema parecido al que NVIDIA ha logrado. Asimismo, me interesa la implementación de NVIDIA NVLink-C2C y de nuevos estándares como CXL (Compute eXpress Link) para la conexión entre CPU y GPU, ya que el bus PCIe se puede convertir en un cuello de botella para el rendimiento de la GPU. Espero que Dell Technologies siga adoptando nuevas tecnologías con rapidez y diseñe productos pensados para el rendimiento".

Con el uso de las GPU más recientes y rentables, el equipo de investigación y desarrollo en IA de CyberAgent continuará evolucionando proporcionando la infraestructura para ML que los usuarios exigen. Además, con la continuación del desarrollo del LLM en japonés, CyberAgent seguirá captando la atención de forma notable, no solo en su propio negocio de publicidad, sino también en el mercado japonés de la IA.

> **Queríamos aumentar el tiempo de funcionamiento con el mínimo de unidades posible, así que valoramos mucho que Dell Technologies pudiese proporcionar un nivel elevado de mantenimiento, incluido el servicio de cuatro horas in situ, por un precio razonable".**
>
> **Daisuke Takahashi**
> Arquitecto de soluciones, CIU, departamento de TI del grupo, CyberAgent, Inc.

Dell Technologies ha traducido este contenido a partir de su versión en Japonés.

Más información sobre las soluciones de IA generativa de Dell Technologies.

Conecte con nosotros en redes sociales.

DELL Technologies

Copyright © 2023 Dell Inc. o sus filiales. Todos los derechos reservados. Dell Technologies, Dell y otras marcas comerciales pertenecen a Dell Inc. o sus filiales. Otras marcas comerciales pueden pertenecer a sus respectivos propietarios. Este caso práctico se ofrece exclusivamente con fines informativos. Dell considera que la información de este caso práctico es precisa en el momento de su publicación, en septiembre de 2023. La información está sujeta a cambios sin aviso previo. Dell no ofrece ninguna garantía, ni expresa ni implícita, sobre este caso práctico.