MÍDIA E ENTRETENIMENTO NO JAPÃO

Desenvolvimento de IA generativa em japonês e transformação dos serviços de publicidade digital

A CyberAgent, Inc. usa servidores Dell PowerEdge XE9680 com oito GPUs NVIDIA® H100 Tensor Core para acelerar a IA generativa e melhorar a eficácia da publicidade.

Necessidades dos negócios

Desde 2016, a CyberAgent, Inc. pesquisa e desenvolve ativamente a IA, incorporando-a em sua empresa de publicidade. Eles precisavam fornecer à equipe acesso rápido e acessível a servidores no local que fossem altamente confiáveis, com as GPUs NVIDIA mais avançadas disponíveis para seus esforços de desenvolvimento de IA generativa.

Resultados de negócios

- Acelera o desempenho de grandes modelos de linguagem (LLM) em aproximadamente 5,14 vezes em comparação à geração anterior com servidores PowerEdge XE9680.
- Expectativa de melhoria no desempenho em mais de dez vezes no futuro com as otimizações do NVIDIA Transformer Engine.
- Permite o ajuste fino em alta velocidade de modelos de aprendizado de máquina de acordo com os conjuntos de dados mais recentes.
- Economiza espaço no data center e oferece resfriamento eficiente com um formato de 6U em comparação com o convencional de 8U.

Resumo geral das soluções

- Servidores Dell PowerEdge XE9680 com GPUs NVIDIA® H100
- Dell ProSupport

A CyberAgent, Inc. é uma empresa conhecida por ser líder de mercado no setor doméstico de publicidade na Internet e em empreendimentos que incluem a ABEMA, uma plataforma de TV inovadora. Em 2016, a empresa estabeleceu uma organização de pesquisa de IA chamada AI Lab. Desde então, ela realiza pesquisas e desenvolve ativamente a IA. Em 2020, a CyberAgent introduziu uma IA preditiva de ponta que melhora a produção de frases de efeito em banners de alto impacto e combinações de imagens, aumentando a eficácia da publicidade.

A empresa continuou seu desenvolvimento da IA generativa, criando um grande modelo de linguagem (LLM) específico para o idioma japonês, exclusivo com 13 bilhões de parâmetros. Projetado como um modelo de IA de uso geral que pode ser usado em uma variedade de situações, o LLM pode ser ajustado para criar cópias de frases de efeito que repercutam nos usuários de cada plataforma de publicidade. A CyberAgent já usa seu LLM japonês em serviços de IA, como Kiwami Prediction AI, Kiwami Prediction TD e Kiwami Prediction LP para auxiliar na produção de publicidade criativa e prever sua eficácia. No futuro, a CyberAgent pretende desenvolver uma IA multimodal que possa lidar não apenas com LLMs japoneses, mas também com imagens.

> **"Nossos pesquisadores internos podem proteger uma quantidade maior de recursos e usá-los sem se preocupar com o custo. Anteriormente, não era possível proteger as GPUs na nuvem pública ou havia cobrança adicional pelo uso em longo prazo."**
>
> **Daisuke Takahashi**
> Arquiteto de soluções, CIU, Departamento de TI do grupo da CyberAgent Inc.

Em maio de 2023, a CyberAgent lançou um LLM japonês de código aberto disponível comercialmente chamado OpenCALM (Open CyberAgent Language Models), que inclui até 6,8 bilhões de parâmetros.

Enquanto o ChatGPT esteja ajustado para um estilo mais conversacional, o OpenCALM é um modelo de idioma japonês de uso geral, que pode ser ajustado conforme as necessidades dos usuários. A CyberAgent lançou o OpenCALM como um projeto de código aberto porque é melhor para a empresa receber feedback de outras fontes, colaborando com outras empresas a fim de contribuir para o desenvolvimento da tecnologia de IA no Japão, em vez de desenvolver um LLM japonês em um ambiente fechado.

A infraestrutura que promove a inovação na IA da CyberAgent

Quando a CyberAgent estabeleceu seu laboratório de IA em 2016, cada pesquisador tinha uma workstation habilitada por GPU para pesquisa. No entanto, com a necessidade de trabalho remoto durante a pandemia em 2020, os pesquisadores tiveram dificuldade para utilizar workstations habilitadas por GPU. Para garantir que os pesquisadores tivessem os recursos computacionais necessários, a empresa começou a pensar em construir plataformas centralizadas de aprendizado de máquina (ML) com servidores habilitados por GPU em seus data centers ou em nuvem pública quando as mais recentes GPUs NVIDIA® A100 foram lançadas.

Daisuke Takahashi, arquiteto de soluções, CIU, do Departamento de TI do grupo da CyberAgent, Inc. diz: "Poderíamos escolher uma nuvem pública se quiséssemos apenas usar GPUs, mas com uma nuvem pública, você nunca sabe quando as GPUs mais recentes estarão disponíveis. Além disso, não há garantia de que as GPUs estariam disponíveis quando necessário. Por isso, decidimos implementar recursos de GPU no local com facilidade de uso. Para perceber a flexibilidade da infraestrutura para alternar entre a nuvem pública e a nuvem privada, desenvolvemos uma interface de usuário que está mais próxima possível das especificações da nuvem pública." A CyberAgent desenvolveu sua plataforma inicial de ML no local usando servidores Dell PowerEdge XE8545 com quatro GPUs NVIDIA A100.

Por que a CyberAgent selecionou os servidores PowerEdge XE9680 com GPUs NVIDIA H100

A CyberAgent continuou a acompanhar a inovação da GPU, principalmente a mais recente GPU NVIDIA H100. "O que nos atraiu foi o desempenho aprimorado, assim como os mecanismos como o Transformer Engine, que acelera algoritmos computacionais específicos" explica Takahashi. "De acordo com a NVIDIA, o Transformer Engine pode acelerar o treinamento de IA de LLMs em até nove vezes e a inferência de IA em até 30 vezes em comparação com as GPUs NVIDIA A100 da geração anterior."

A CyberAgent escolheu o modelo de servidor PowerEdge XE9680 com oito GPUs NVIDIA H100. Takahashi explica: "Quando descobrimos que os servidores Dell PowerEdge XE9680 com GPUs NVIDIA H100 seriam lançados, decidimos adotá-los o mais rápido possível. Conseguimos trabalhar em conjunto com a Dell Technologies abordando as configurações que seriam possíveis com os futuros servidores PowerEdge XE9680 e as GPUs. A empresa gostaria de aumentar o tempo de atividade com o mínimo de unidades possível. Por isso, ficamos satisfeitos quando a Dell Technologies conseguiu fornecer um alto nível de manutenção, incluindo quatro horas de serviço no local, a um preço razoável."

DELL Technologies

Acelera um LLM com 13 bilhões de parâmetros em 5,14 vezes hoje e mais de 10 vezes no futuro.

Takahashi explica: “Também escolhemos os servidores PowerEdge XE9680 porque as instalações anteriores dos servidores PowerEdge XE8545 proporcionaram desempenho estável e facilidade de manutenção. Além disso, valorizamos a facilidade de uso da ferramenta Dell iDRAC para gerenciamento seguro de servidores locais e remotos."

Para Takahashi, foi importante fazer o pedido em março de 2023 e recebê-lo pouco mais de um mês depois, em meados de maio. “Com as cadeias de suprimentos desordenadas devido à pandemia, também tive a certeza de que a Dell Technologies tem uma cadeia de suprimentos relativamente estável. Também foi bom saber que eles poderiam fazer as entregas em um período de tempo tão curto."

Uma variedade de inovações foram realizadas no processo de construção após a entrega. Takahashi lembra: “Para um LLM com um grande número de parâmetros, precisávamos usar várias GPUs. Desse modo, instalamos oito placas de interface de rede (NICs) de 400 Gbps em cada servidor e usamos a tecnologia RDMA (Remote Direct Memory Access) para criar um interconexão de alta velocidade entre os servidores. Os servidores com GPUs geram muito calor, por isso é importante que sejam projetados para serem resfriados de modo eficiente. O formato de 6U dos servidores PowerEdge XE9680 para resfriamento consistente também é recomendável. Além disso, o data center também foi realocado para um novo local em que os trocadores de calor da porta traseira estão disponíveis. Assim, é possível ter resfriamento eficaz por meio da instalação de trocadores de calor pela porta traseira, resfriados a água, em vez de resfriar a sala inteira que abriga o data center."

Melhorando a precisão das frases de efeito com otimizações do Transformer Engine

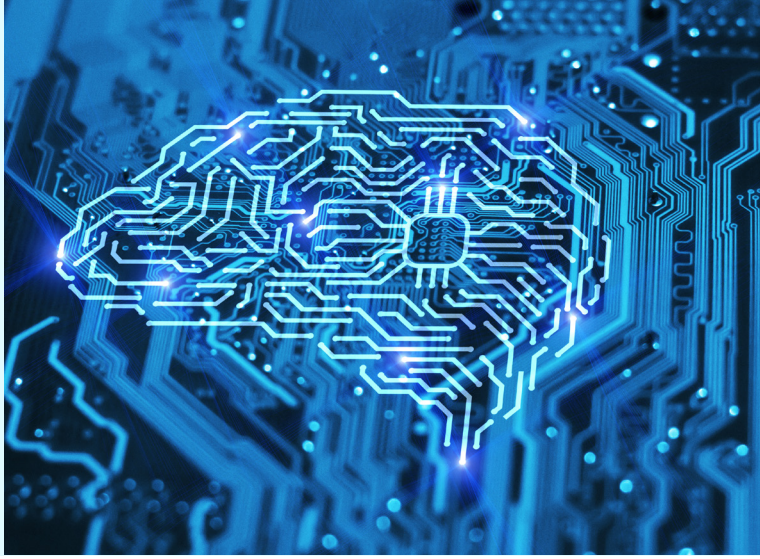
Ao instalar servidores PowerEdge XE9680, a CyberAgent obtém vários benefícios. “Esperamos atualizar nossos LLMs japoneses com mais rapidez e frequência devido à melhoria significativa no desempenho”, afirma Takahashi. “A velocidade de evolução dos LLMs japoneses também vai melhorar. Além disso, em comparação com os servidores PowerEdge XE8545 equipados com quatro GPUs NVIDIA A100, os servidores PowerEdge XE9680 com oito GPUs NVIDIA H100 tiveram uma melhoria de desempenho de aproximadamente 5,14 vezes. Também prevemos um aumento de desempenho de mais de 10 vezes com a otimização para o NVIDIA Transformer Engine no futuro. Além disso, podemos fazer ajustes em alta velocidade nos modelos de ML de acordo com os conjuntos de dados mais recentes. Desse modo, podemos responder mais facilmente às solicitações para evoluir os serviços, melhorar a precisão das frases e fornecer conteúdo mais eficaz."

A infraestrutura de ML habilitada por servidores PowerEdge XE9680 recebeu muitos elogios dos usuários. “Segundo os nossos pesquisadores internos, eles podem conseguir uma quantidade maior de recursos e usá-los sem se preocupar com o custo. Anteriormente, não era possível conseguir as GPUs na nuvem pública ou havia cobrança adicional pelo uso em longo prazo", diz Takahashi. “Outro benefício é que conseguimos fornecer uma infraestrutura de alta especificação, incluindo interconexão, para que os usuários possam causar um impacto nos negócios."

Takahashi também valoriza a ferramenta de gerenciamento iDRAC da Dell Technologies, que a empresa usa há algum tempo, porque reduz a carga de gerenciamento. “Nem sempre usamos o data center. Por isso o iDRAC é útil para realizar tarefas remotamente, como verificar a temperatura e o status das GPUs, assim como atualizar o firmware sem precisar acessar o sistema operacional."

> **“O formato de 6U dos servidores PowerEdge XE9680 para resfriamento consistente também é recomendável."**
>
> **Daisuke Takahashi**
> Arquiteto de soluções, CIU, Departamento de TI do grupo da CyberAgent Inc.

> **"Esperamos atualizar nossos LLMs em japonês mais rapidamente. Os servidores PowerEdge XE9680 com oito GPUs NVIDIA H100 tiveram uma melhoria de desempenho de aproximadamente 5,14 vezes."**
>
> **Daisuke Takahashi**
> Arquiteto de soluções, CIU, Departamento de TI do grupo da CyberAgent, Inc.



Foco em LLMs, GPUs e infraestrutura

Olhando para o futuro, a CyberAgent planeja utilizar o feedback e o aprendizado coletados no OpenCALM para melhorar o LLM usado pelos funcionários. Por meio do OpenCALM, a CyberAgent também explora colaborações com empresas e organizações em outros setores além da publicidade. Por exemplo, a CyberAgent iniciou discussões com participantes do setor de varejo e financeiro para construir LLMs específicos que aprendam com seus dados.

Enquanto isso, Takahashi explica que continuará a acompanhar as GPUs mais recentes e as novas tecnologias relacionadas para conferir como elas são comercializadas. "Estamos ansiosos para ver como outros fornecedores podem criar um ecossistema de software semelhante ao da NVIDIA. Também estou interessado na implementação do NVIDIA NVLink-C2C e de novos padrões como CXL (Compute eXpress Link) que conectam CPU e GPU, já que o barramento PCIe pode ser um gargalo no desempenho da GPU. Espero que a Dell Technologies continue a adotar novas tecnologias em um ritmo rápido e a projetar produtos que ofereçam desempenho."

Ao usar as GPUs mais recentes e econômicas, a equipe de pesquisa e desenvolvimento de IA da CyberAgent continuará a evoluir, fornecendo a infraestrutura de ML que os usuários exigem. Além disso, com o desenvolvimento do LLM japonês, a CyberAgent continuará a atrair atenção significativa, não apenas para seu próprio negócio de publicidade, mas também para o mercado japonês de IA.

> **"A empresa gostaria de aumentar o tempo de atividade com o mínimo de unidades possível. Por isso, ficamos satisfeitos quando a Dell Technologies conseguiu fornecer um alto nível de manutenção, incluindo quatro horas de serviço no local, a um preço razoável."**
>
> **Daisuke Takahashi**
> Arquiteto de soluções, CIU, Departamento de TI do grupo da CyberAgent Inc.

Este conteúdo foi traduzido da versão em japonês pela Dell Technologies.

Saiba mais sobre as soluções de IA generativa da Dell Technologies.

Conecte-se pelas redes sociais.

DELL Technologies

Copyright © 2023 Dell Inc. ou suas subsidiárias. Todos os direitos reservados. Dell Technologies, Dell e outras marcas comerciais pertencem à Dell Inc. ou suas subsidiárias. Outras marcas comerciais podem pertencer aos respectivos proprietários. Este estudo de caso serve apenas a fins informativos. A Dell acredita que as informações neste estudo de caso estão corretas na data de publicação, em setembro de 2023. As informações estão sujeitas a alterações sem aviso. A Dell não oferece garantias expressas nem implícitas neste estudo de caso.