DELL Technologies

미디어 및 엔터테인먼트
일본

CyberAgent

일본어 Generative AI 개발 및 디지털 광고 서비스 혁신

CyberAgent, Inc.는 8개의 NVIDIA® H100 Tensor Core GPU가 탑재된 Dell PowerEdge XE9680 서버를 사용하여 Generative AI를 가속하고 광고 효과를 높입니다.

비즈니스 요구 사항

CyberAgent, Inc.는 2016년부터 AI를 적극적으로 연구, 개발하고 이를 광고 비즈니스에 접목해 왔습니다. CyberAgent, Inc.는 직원들이 Generative AI 개발에 사용할 수 있는 최첨단 NVIDIA GPU가 탑재된, 탁월한 신뢰성을 갖춘 온프레미스 서버에 빠르고 경제적으로 액세스할 수 있도록 지원해야 했습니다.

비즈니스 성과

PowerEdge XE9680 서버를 사용하면 이전 세대에 비해 LLM(Large Language Model) 성능이 약 5.14배 가속됩니다.

NVIDIA Transformer Engine 최적화로 향후 성능이 10배 이상 향상될 것으로 예상됩니다.

최신 데이터 세트에 따라 머신 러닝 모델을 고속으로 미세 튜닝할 수 있습니다.

메인스트림 8U 대신에 6U 폼 팩터로 데이터 센터 공간을 절약하고 냉각 효율을 높입니다.

솔루션 개요

- NVIDIA® H100 GPU를 탑재한 Dell PowerEdge XE9680 서버
- Dell ProSupport

CyberAgent, Inc.는 혁신적인 TV 플랫폼인 ABEMA를 비롯한 일본 내 인터넷 광고 업계의 시장 리더이자 벤처기업으로 잘 알려진 회사입니다. 2016년에 AI 연구 조직인 AI Lab을 설립한 이후 AI 연구 개발을 적극적으로 진행해 왔습니다. 2020년, CyberAgent는 효과적인 배너 광고 문구 및 이미지 조합의 제작을 개선하여 광고 효과를 높이는 최첨단 예측 AI를 도입했습니다.

CyberAgent는 Generative AI 개발을 계속하여 130억 개의 매개변수를 사용하는 독보적인 일본어 전용 LLM(Large Language Model)을 만들었습니다. 다양한 상황에서 사용할 수 있는 범용 AI 모델로 설계된 이 LLM은 각 광고 플랫폼 사용자의 공감을 불러일으키는 광고 문구 카피를 만들도록 미세 튜닝할 수 있습니다. CyberAgent는 이미 Kiwami Prediction AI, Kiwami Prediction TD, Kiwami Prediction LP와 같은 AI 서비스에서 일본어 LLM을 사용하여 창의적인 광고 제작을 지원하고 광고 효과를 예측하고 있습니다. 앞으로 CyberAgent는 일본어 LLM뿐만 아니라 이미지까지 처리할 수 있는 멀티모달 AI를 개발할 계획입니다.

> **"사내 연구자들이 비용 걱정 없이 더 많은 리소스를 확보해 사용할 수 있습니다. 이전에는 퍼블릭 클라우드에서 GPU를 확보하지 못하거나 장기간 사용 시 더 큰 비용을 부담해야 했죠."**
>
> **Daisuke Takahashi**
> CyberAgent, Inc., CIU, Group IT 부서, Solution Architect

2023년 5월에 CyberAgent는 최대 68억 개의 매개변수를 포함하는 상용 오픈 소스 일본어 LLM인 OpenCALM(Open CyberAgent Language Model)을 출시했습니다.

ChatGPT는 채팅 용도로 튜닝되었지만, OpenCALM은 사용자의 요구에 맞게 미세 튜닝할 수 있는 범용 일본어 언어 모델에 더 가깝습니다. CyberAgent가 OpenCALM을 오픈 소스 프로젝트로 출시한 이유는 폐쇄형 환경에서 일본어 LLM을 개발하는 것보다 다른 소스로부터 피드백을 받고 다른 회사와 협업하여 일본의 AI 기술 발전에 기여하는 것이 회사에 더 유익하기 때문입니다.

CyberAgent의 AI 혁신을 지원하는 인프라스트럭처

CyberAgent가 2016년에 AI Lab을 설립했을 당시, 각 연구원은 연구용으로 GPU 기반 워크스테이션을 사용했습니다. 그러나 2020년 팬데믹 기간 동안 재택/원격 근무의 필요성이 대두되면서 각 연구원이 GPU 기반 워크스테이션을 활용하기가 어려워졌습니다. 연구원들이 필요한 컴퓨팅 리소스를 확보하도록 하기 위해, 회사는 최신 NVIDIA® A100 GPU가 출시되자 데이터 센터 또는 퍼블릭 클라우드에 GPU 기반 서버를 갖춘 중앙 집중식 ML(Machine Learning) 플랫폼을 구축하는 것을 고려하기 시작했습니다.

CyberAgent, Inc.의 CIU, Group IT 부서 Solution Architect인 Daisuke Takahashi 씨는 이렇게 말합니다. "단지 GPU를 사용하고자 했다면 퍼블릭 클라우드를 선택할 수 있었겠지만, 퍼블릭 클라우드에서는 최신 GPU를 언제 사용할 수 있게 될지 알 수가 없습니다. 또한 우리가 원할 때 GPU를 사용할 수 있을 것이라는 보장이 없으므로 사용 편의성을 고려해 GPU 리소스를 온프레미스에서 배포하기로 결정했습니다. 인프라스트럭처가 퍼블릭 클라우드와 프라이빗 클라우드 간을 유연하게 오갈 수 있도록 하기 위해 퍼블릭 클라우드 사양에 최대한 가까운 사용자 인터페이스를 고안했습니다." CyberAgent는 4개의 NVIDIA A100 GPU가 탑재된 Dell PowerEdge XE8545 서버를 사용하여 초기 온프레미스 ML 플랫폼을 구축했습니다.

CyberAgent가 NVIDIA H100 GPU가 탑재된 PowerEdge XE9680 서버를 선택한 이유

CyberAgent는 GPU 혁신, 특히 최신 NVIDIA H100 GPU를 계속 추구했습니다. "향상된 성능뿐 아니라 특정 컴퓨팅 알고리듬을 가속하는 Transformer Engine과 같은 메커니즘도 매력적이라고 생각했습니다."라고 Takahashi 씨는 설명합니다. "NVIDIA에 따르면 Transformer Engine은 이전 세대 NVIDIA A100 GPU에 비해 LLM의 AI 학습 속도를 최대 9배, AI 추론 속도를 최대 30배 높일 수 있습니다."

CyberAgent는 8개의 NVIDIA H100 GPU가 탑재된 PowerEdge XE9680 서버 모델을 선택했습니다. Takahashi 씨는 이렇게 설명합니다. "NVIDIA H100 GPU가 탑재된 Dell PowerEdge XE9680 서버가 출시될 것이라는 소식을 듣고 최대한 빨리 채택하기로 결정했습니다. 우리는 곧 출시될 PowerEdge XE9680 서버와 GPU에서 가능한 구성에 대해 Dell Technologies와 긴밀하게 의견을 나눌 수 있었습니다. 최소한의 유닛으로 가동 시간을 늘리고 싶었기 때문에 Dell Technologies가 4시간의 현장 서비스를 포함하여 높은 수준의 유지 보수를 합리적인 가격으로 제공할 수 있다는 점이 만족스러웠습니다."

DELL Technologies

130억 개의 매개변수가 포함된 LLM을 현재 5.14배, 향후 10배 이상 가속합니다.

그리고 다음과 같이 덧붙입니다. "PowerEdge XE9680 서버를 선택한 것은 이전에 설치했던 PowerEdge XE8545 서버가 안정적인 성능을 제공했고 유지 보수가 용이했기 때문이기도 했습니다. 또한 안전한 로컬 및 원격 서버 관리를 위한 Dell iDRAC 관리 툴의 사용 편의성도 중요하게 생각합니다."

Takahashi 씨는 2023년 3월에 주문한 서버가 그로부터 한 달 남짓 지난 5월 중순에 배송 완료되었다는 사실을 높이 평가합니다. "팬데믹으로 인해 공급망이 혼란스러운 가운데 Dell Technologies는 비교적 안정적인 공급망을 갖추고 있다는 점에서도 안심이 되었고 짧은 시간 안에 배송할 수 있다는 것을 알게 되어 기뻤습니다."

배송 후 구축 과정에서 다양한 혁신이 이루어졌습니다. Takahashi 씨는 당시를 이렇게 회상합니다. "수많은 매개변수가 포함된 LLM의 경우 여러 GPU를 사용해야 했기 때문에 각 서버에 8개의 400Gbps NIC(Network Interface Card)를 설치하고 RDMA(Remote Direct Memory Access) 기술을 사용하여 서버 간의 고속 상호 연결을 구축했습니다. GPU 서버는 많은 열을 발생시키므로 효율적으로 냉각되도록 설계하는 것이 중요합니다. PowerEdge XE9680 서버 6U 폼 팩터의 강력한 냉각 기능도 칭찬할 만합니다. 그 외에도 데이터 센터를 후면 도어 열 교환기를 사용할 수 있는 새로운 위치로 이전했는데, 데이터 센터가 있는 공간 전체를 냉각하는 대신에 랙 후면에 수랭식 후면 도어 열 교환기를 설치해 효과적인 냉각을 실현할 수 있게 되었습니다."

Transformer Engine 최적화로 광고 문구 정확성 향상

CyberAgent는 PowerEdge XE9680 서버를 설치하여 다양한 이점을 얻고 있습니다. Takahashi 씨는 "성능이 대폭 향상되어 일본어 LLM을 더 빠르게, 더 자주 업데이트할 수 있을 것으로 기대합니다."라고 말하면서 다음과 같이 덧붙입니다. "일본어 LLM의 발전 속도도 빨라질 것입니다. 또한 4개의 NVIDIA A100 GPU가 탑재된 PowerEdge XE8545 서버에 비해 8개의 NVIDIA H100 GPU가 탑재된 PowerEdge XE9680 서버의 성능이 약 5.14배 향상되었습니다. 향후 NVIDIA Transformer Engine을 최적화하면 성능이 10배 이상 향상될 것으로 예상됩니다. 그리고 최신 데이터 세트에 따라 ML 모델을 고속으로 미세 튜닝할 수 있어, 더 쉽게 서비스 개선 요청에 대응하고 광고 문구의 정확성을 높이며 더욱 효과적인 콘텐츠를 제공할 수 있게 될 것입니다."

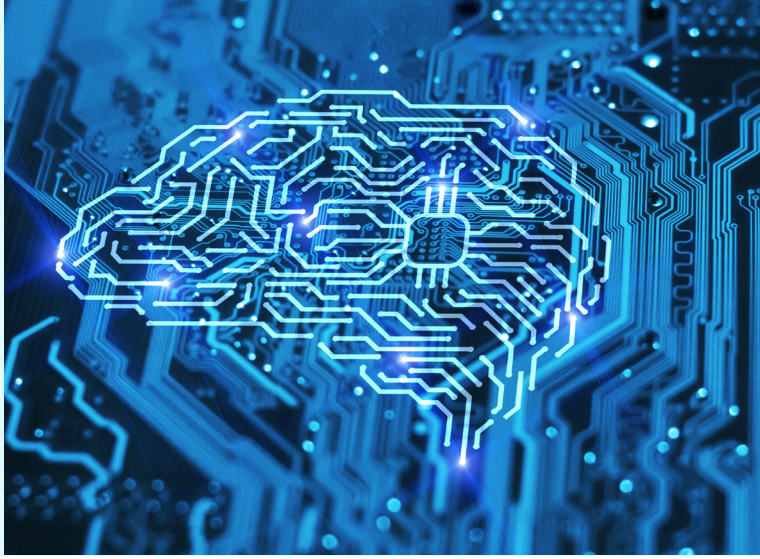
PowerEdge XE9680 서버 기반의 ML 인프라스트럭처는 사용자들로부터 높은 평가를 받았습니다. Takahashi 씨는 이렇게 말합니다. "사내 연구자들에 따르면 비용 걱정 없이 더 많은 리소스를 확보해 사용할 수 있습니다. 이전에는 퍼블릭 클라우드에서 GPU를 확보하지 못하거나 장기간 사용 시 더 큰 비용을 부담해야 했죠. 또 다른 이점은 사용자가 비즈니스에 영향을 미칠 수 있도록 상호 연결을 포함한 고사양의 인프라스트럭처를 제공할 수 있었다는 것입니다."

Takahashi 씨는 회사에서 한동안 사용해 온 Dell Technologies iDRAC 관리 툴도 관리 부담을 줄여준다는 점에서 높이 평가합니다. "우리가 항상 데이터 센터에 상주해 있는 것은 아닙니다. 따라서 OS에 액세스하지 않고도 GPU의 온도와 상태를 확인하고 펌웨어를 업데이트하는 등 원격으로 작업을 수행하는 데 iDRAC가 유용합니다."

> **"PowerEdge XE9680 서버 6U 폼 팩터의 강력한 냉각 기능도 칭찬할 만합니다."**
>
> **Daisuke Takahashi**
> CyberAgent, Inc., CIU, Group IT 부서, Solution Architect

> **일본어 LLM을 더 빠르게 업데이트할 수 있을 것으로 기대합니다. 8개의 NVIDIA H100 GPU가 탑재된 PowerEdge XE9680 서버의 성능이 약 5.14배 향상되었습니다.”**

Daisuke Takahashi
CyberAgent, Inc., CIU, Group IT 부서, Solution Architect



LLM, GPU 및 인프라스트럭처에 주력

앞으로 CyberAgent는 OpenCALM에서 수집한 피드백과 학습 내용을 활용하여 직원들이 사용하고 있는 LLM을 개선할 계획입니다. CyberAgent는 OpenCALM을 통해 광고 이외 업계의 많은 기업 및 조직과 협업할 방안도 모색하고 있습니다. 예를 들어 CyberAgent는 해당 산업 관련 데이터로부터 학습하는 산업별 LLM을 구축하기 위해 리테일 및 금융 분야의 주요 기업들과 논의를 시작했습니다.

한편 Takahashi 씨는 최신 GPU 및 관련 신기술로 계속 업데이트해 나가면서 상용화 방법을 찾을 것이라고 설명합니다. “NVIDIA가 실현한 것과 유사한 소프트웨어 생태계를 다른 공급업체들이 어떻게 구축할 수 있을지 정말 기대하고 있습니다. 그리고 개인적으로는 PCIe 버스가 GPU 성능에 병목 현상을 일으킬 수 있기 때문에 NVIDIA NVLink-C2C의 구현과 CPU 및 GPU를 연결하는 CXL(Compute eXpress Link)과 같은 새로운 표준을 구현하는 데도 관심이 있습니다. 저는 Dell Technologies가 계속해서 빠른 속도로 새로운 기술을 채택하고 성과를 실현하는 제품을 설계할 것으로 기대합니다.”

CyberAgent의 AI 연구 개발 팀은 경제적인 최신 GPU를 사용하여 사용자가 요구하는 ML 인프라스트럭처를 제공함으로써 계속 발전해 나갈 것입니다. 또한 일본어 LLM의 추가 개발을 통해 CyberAgent는 자체 광고 비즈니스뿐만 아니라 일본어 AI 시장에서도 계속해서 큰 주목을 받게 될 것입니다.

> **최소한의 유닛으로 가동 시간을 늘리고 싶었기 때문에 Dell Technologies가 4시간의 현장 서비스를 포함하여 높은 수준의 유지 보수를 합리적인 가격으로 제공할 수 있다는 점이 만족스러웠습니다.”**

Daisuke Takahashi
CyberAgent, Inc., CIU, Group IT 부서, Solution Architect

이 콘텐츠는 Dell Technologies에서 일본어 버전을 번역한 것입니다.

Dell Technologies Generative AI Solutions에 대한 **자세한 정보.**

소셜 미디어 **참여.**

DELL Technologies