cisco Connect

Cisco AI Inferencing Summit [아젠다]

13:00-13:30 등록 확인

환영사

추론(Inference) 시대를 선점할 인프라 혁신 전략

13:30-13:45

훈련(Training)의 시대가 가고, 추론(Inference)의 시대가 도래했습니다. 모두가 AI 모델 개발에 몰두할 때, 비즈니스의 성패를 가를 진짜 열쇠는 '어떻게 서비스할 것인가'에 있습니다. 시스코 코리아가 제시하는 AI 인프라의 새로운 게임 체인저 전략을 공개합니다.

배영국 CTO, Cisco Korea

Vijoy Pandey, GM/SVP of Outshift by Cisco

- Cisco Research, Quantum Labs, 오픈소스 오피스, DevNet 등의 핵심 조직 총괄
- 캘리포니아대학교 데이비스에서 컴퓨터공학 박사 학위 취득
- 클라우드 컴퓨팅과 AI/ML, 분산 시스템 분야 80건 이상의 특허 보유
- Google, IBM, Blade Network Technologies 등에서 대규모 인프라 개발과 글로벌 엔지니어링 팀 리딩

기조연설

Cisco Secure AI 인프라와 함께하는 AI 개발 라이프사이클 탐색

13:45-14:35

본 세션에서는 AI 개발 라이프사이클의 각 단계에서 개발자가 직면하는 주요 과제들을 살펴보고, 이러한 문제들을 해결할 수 있는 실질적인 방안들을 소개합니다. 데이터 준비, 모델 학습, 배포 및 추론에 이르는 전 과정에서 보안, 확장성, 고성능 AI 인프라의 중요성을 강조합니다. Cisco의 Feature Ready AI 인프라인 Secure AI Factory를 통해 기업이 AI 도입을 가속화하고, 모든 계층에 통합된 보안을 제공하는 방법을 알아보세요. NVIDIA, Vast Data, Red Hat 등 주요 파트너 및 ISV와의 협력을 통해 컴퓨트, 네트워킹, 스토리지, AI 소프트웨어가 결합된 검증된 아키텍처를 제공하여 AI 워크로드를 간소화하고 안전하게 지원합니다. 이 세션은 Cisco AI PODs, Secure AI Factory, 그리고 생태계 파트너들이 어떻게 복잡한 인프라 문제를 해결하고 보안을 강화하며 AI 프로젝트의 가치 실현 속도를 높이는지에 대한 종합적인 개요를 제공합니다.

Rose Skandari, AI Solution Field CTO, Cisco

- 전기공학 박사 (멜버른 대학교)
- 분석 분야 최고 리더 25인 중 한 명으로 선정
- 여성 AI 전문가상 최종 후보
- 호주 대학에서 AI 강의를 하며 차세대 전문가 양성

Session 1

GPU의 한계를 넘다: 초대규모 AI를 위한 '압도적 효율'의 차세대 추론 프레임워크

14:35-15:00

GPU 성능의 한계를 넘어 기존 인프라만으로도 수십 배의 효율을 이끌어내는 거대 언어 모델(LLM) 서비스의 최적 기법. Moreh AI 가 제안하는 '인프라 독립적' AI 추론 솔루션의 실체를 확인하십시오.

조형근, CSO/최고전략책임자, MOREH AI

- 한국인공지능·소프트웨어산업협회 AI정책협력위원
- 피지컬AI 글로벌 얼라이언스 위원
- Bain & Company consultant
- KAIST 한국과학기술원 MS, 연세대학교 BS

15:00-15:20 Break

Session 2

Agentic AI기반 휴머노이드 로봇과 Cisco UCS Edge서버의 고성능 기반 Physical AI 시스템

15:20-15:45

산업 도우미, 가정 도우미로 휴머노이드 로봇의 브레인 역할을 담당하는 Agentic AI에 있어 '최적의 성능' 및 '낮은 응답속도는(Latency)'은 중요한 부분입니다. 본 세션에서는 Cisco UCS Edge 서버의 강력한 엣지 컴퓨팅 파워와 CloAI의 경량화된 LLM, VLM 알고리즘이 결합했을 때, 로봇이 어떻게 인간 수준의 대화와 응답을 가능하게 하는지 제시합니다. 모델의 인퍼런스 성능을 높이고, 데이터 전송 병목을 제거하여 현장에서 즉각적인 추론(Inferencing)과 제어를 수행할 수 있는 가정 및 산업 도우미 로봇의 실증 사례를 소개합니다.

김성룡, CEO, CloAI

- 과기부 AI부문 과제 평가위원
- 경남 초거대AI 과제 기획 연구위원

Session 3

나보다 나를 더 잘 아는 AI: '초개인화 에이전트'가 Cisco 엣지 위에서 살아 숨 쉴 때

15:45-16:10

획일화된 챗봇은 이제 그만. 사용자의 의도를 먼저 읽고 행동하는 'AI 에이전트'가 온디바이스(On-Device) 기술과 만났습니다. 개인정보 유출 걱정 없이, 가장 가까운 곳에서 나만을 위해 작동하는 SAKAK의 프라이빗 AI 에이전트 기술을 만나보세요.

배건규, CEO & Founder, SAKAK

- 한국개발연구원(KDI) 보건의료 마이데이터 활성화 자문 워킹그룹
- DB생명 AI 자문위원

Session 4

한국어 도메인에 특화된 LLM 전체 개발 주기 : Raw 데이터부터 전문가 수준 성능 벤치마크까지

16:10-16:35

성공적인 도메인 특화 대규모 언어 모델(LLM) 개발은 일반 모델의 성능을 향상시키기 위해 독자적인 전문 지식을 활용하는 데 달려 있습니다. 이는 특히 비영어권 모델의 경우 더욱 중요하며, 사전 학습 과정에서 언어적 숙련도뿐만 아니라 해당 언어와 관련된 도메인별 지식(예: 현지 법률 체계 및 규제 문서)을 통합해야 합니다. 본 초청 강연에서는 한국어 금융 특화 LLM인 bllossom의 개발 전 과정을 상세히 소개합니다. 원시 한국어 금융 데이터의 수집 및 정제부터 전문가 지식을 통합하기 위한 맞춤형 학습 및 미세 조정 방법론에 이르기까지 전체 과정을 다룹니다. 특히 엄격한 맞춤형 벤치마크 구축 및 모델의 도메인 성능을 입증하는 검증 결과에 중점을 둘 것입니다. 발표는 전문가 수준의 다국어 LLM 구축을 위한 전략적 과제와 향후 연구 방향을 제시하며 마무리될 예정입니다.

함영균, CEO, Teddysum AI

- KAIST 컴퓨터과학 박사
- ISO/TC 37/SC 4/WG 2 간사
- K-AI 리더보드 프로젝트 PM (NIA)

16:35-16:40 맺음말