CES 2025: 온오프 AI, Vertical AI, AI 에이전트의 시대가 왔다

(이경전 경희대 교수, klee@khu.ac.kr)

1. 온디바이스 AI, 온프레미스 AI, 클라우드 AI가 결합되는 온오프 AI

이번에도 인공지능 기술의 최첨단을 달리는 회사들은 전시장에 부스를 설치하지 않고 미디어에 최신 기술을 알렸다. Nvidia는 CEO가 CES 2025 개막 연설을 했고, 2024년 12월 OpenAI는 12 Days of OpenAI를 통해 Sora, O3 등 놀랄만한 신상품을 12일간 매일 하나씩 발표했다. OpenAI, Meta AI, Anthropic, X.AI, Perplexity.AI 등 현 AI를 주도하는 기업들은 CES에 부스를 설치하지 않았다. 흥미롭게도 SK관은 제휴 협력하는 미국회사 퍼플렉시티를 자사의 AI 포트폴리오에 포함시켰다. 구글은 CES 2024에는 참가했으나, 첨단 AI 기술보다는 기존 서비스에 AI를 입히는 상당히 소극적인 전시 전략으로 실망시킨바 있는데, CES 2025에는 아예 참가하지 않았다. 애플은 1992년을 마지막으로 CES에 참여하지 않고 있으며, 마이크로소프트도 2012년 이후 참가하지 않고 있다.

엔비디아는 Project DIGITS라는 개인용 AI 슈퍼컴퓨터를 발표했다. GB10 Grace Blackwell Superchip을 탑재한 데스크톱 크기의 장치로, 최대 2000억 개의 파라미터를 가진 AI 모델을 실행할 수 있는 성능을 제공한다. DIGITS는 연구자, 데이터 과학자, 학생 등을 대상으로 설계되었고, AI 모델 개발, 프로토타이핑, 미세 조정 및 추론 작업을 로컬 환경에서 수행할 수 있도록 지원한다. GB10 Grace Blackwell Superchip은 NVIDIA의 최신 Grace Blackwell 아키텍처를 기반으로 설계된 고성능 시스템 온 칩(System-on-a-Chip, SoC)이다. 최대 1 페타플롭(PFLOPS, Floating-Point Operations Per Second, 초당 부동소수점 연산 횟수)의 AI 성능(FP4 기준: 4비트로 숫자를 표현하여 계산)을 보이는데, 메가가 10의 6승, 기가가 10의 9승, 테라가 10의 12승, 페타가 10의 15승으로, 초당 최대 1,000조(10^{15})번의 부동소수점 연산을 수행할 수 있음을 의미하는데, 인간이 계산기를 사용해 초당 1번 계산한다고 가정하면, 1 PFLOPS 성능의 컴퓨터는 인간이 3200만 년 동안 할 계산을 단 1초 만에 처리할 수 있다. NVIDIA의 A100 GPU가 약 624 테라플롭(FP16 기준: 16비트로 숫자를 표현하여 계산)으로, 고성능 AI 훈련 및 데이터 분석에 사용되어 왔는데, 이것의 약 1.6배 높은 성능을 가진다고 할 수 있는데, 이 비교는 서로 다른 정밀도(FP4와 FP16)를 기준으로 하므로 실제 응용에 따라 성능 차이는 달라질 수 있다.

두 대의 DIGITS를 NVIDIA ConnectX 네트워크 어댑터로 연결하면, 최대 4050억 파라미터 모델도 실행이 가능하다. 2025년 5월에 약 3천불 가격으로 협력사들과 함께 출시하겠다고 발표했다.

이번 CES 2025에서 나온 최고의 뉴스라고 생각한다.

ChatGPT 3.5버전이 1750억개의 파라미터이므로, 2025년부터는 개인이 자신의 컴퓨터에서 ChatGPT 3.5 수준의 AI 모델을 자유 자재로 파인튜닝해가면서 나만의 강력한 AI를 만들 수 있다는 뜻이된다. 헨리 포드가 나만의 자동차 T-모델을 출시하고, 스티브잡스가 나만의 컴퓨터 Apple II와 나만의 스마트폰 아이폰을 출시한 것에 비견할만한 사건이기 때문이다. DIGITS는 인류 최초의 나만의 AI 컴퓨터 출시 사건이다. OpenAI와 Meta AI, Anthropic, x.AI 등은 DIGITS에 탑재할 기본 AI 모델을 서로 팔기 위해 경쟁할 것이다. 가격은 떨어질 것이다. 소비자들은 즐거운 비명을 지르게 될 것이다.

LG전자 역시 AI모델을 구동할 수 있는 LG 그램 모델을 1월 7일을 기준으로 국내외 시장에 동시에 출시했다. 16Z90TP와 17Z90TP 모델인데, 16인치와 17인치가 있다는 뜻이고, Z는 초경량 노트북 그램 라인업이라는 뜻이며, 90은 프리미엄 등급이고, T는 2025년 출시를 의미하며, P는 그래픽 프로세서를 내장했다는 뜻이다. '그램 챗 온디바이스(On-Device)'와 '그램 챗 클라우드'를 모두 활용할 수 있는 점이 특징이다. 실수로 삭제된 데이터를 복구하거나, 오래된 파일이나 문서 속 텍스트를 검색하거나, 캘린더와 메일 연동을 통해 일정 관리를 지원하는 등과 같은 개인 맞춤형 AI 기능인 그램 챗 온디바이스는, LG AI연구원이 개발한 최대 78억 개의 파라미터를 가진 소형 언어 모델 엑사원(EXAONE) 3.0을 활용한다. 2000억 파라미터를 가진 AI를 운영할 수 있는 엔비디아의 DIGITS보다는 25분의 1 수준의 규모지만, 첫술밥에 배부르랴. LG가 자체 개발한 AI 모델을 온디바이스 형태로 제공한다는 것은 LG전자의 정공법이라 할 수 있으며, 기대해볼만 하다. 오픈에이아이의 GPT-4o를 기반으로 한 그램 챗 클라우드도 제공하는데, GPT-4o를 1년간 무료로 제공하는 형태다. 결국 고객은 LG의 엑사원 3.0과 오픈에이아이의 GPT-4o를 기본적으로 장착한 PC를 사용하게 되는 것으로 매우 현실적인 가치제안이다.

CES 2024에서는 Samsung S24, Pixel 8 Pro 등 삼성전자와 구글이 온디바

이스 AI 폰을 출시하였으나, CES 2025에는 온디바이스 스마트폰은 이슈가 되지 않는 상황이고, 엔비디아와 LG가 온디바이스 AI 랩탑컴퓨터를 발표하여, AI 내재 컴퓨터의 크기가 커지고, 내재되는 AI 모델의 크기가 커지는 추세임을 보여준다. 향후 계속적으로 개인용 AI 탑재장치의 용량과 성능이 커질 것이라는 것을 자연스럽게 유추할 수 있다.

2024년 9월에 출간된 <AI 에이전트와 사회변화>라는 책에서 다음과 같이 언급한 바 있다. “초거대 AI도 잘 따져 보면, 개인용 컴퓨터에 설치 가능한 규모다. 1조 개의 숫자는 4조 바이트, 즉 4테라 바이트이므로 시중에 나온 4테라 바이트 하드디스크를 가진 PC에 충분히 설치할 수 있는 크기인 것이다 (GPT3.5의 경우는 약 700기가바이트). 물론 이것을 운용할 수 있는 GPU를 개인이 갖추는 것이 아직 어렵고 비싸다. 그러나 현재 세계에서 가장 큰 AI 모델이 지금 일반인이 가질 수 있는 컴퓨터에 설치 정도는 가능하다는 것은 시사점이 크다. AI는 현재 그 규모만 가지고도 어떤 큰 국가나 기업, 조직이 독점할 수 있는 성질의 것이 아니라는 것이다. 즉, 인공지능이 전 세계 사람 누구나 소유할 수 있고 관리할 수 있는 AI 에이전트의 방향으로 갈 수밖에 없다는 것을 시사한다”

엔비디아의 Digits 프로젝트의 5월 출시는 위 언급이 8개월만에 벌써 실현되어가고 있다는 것을 의미한다.

CES 2025에서는 XanderGlasses, Xreal 등의 스마트 글래스들이 자체로 구동되는 온디바이스 AI 탑재를 내세웠다. 지연(latency)이 없고, 프라이버시를 보호한다는 점에서 온디바이스 AI는 필수적인 기능이다. 다만 기술적 구조와 사용자 제공가치는 상당히 다르다. 중국 북경의 XReal은 독자적 프로세서 X1을 통해 몰입형 증강 현실 경험을 제공한다. 멀티모달 AI 카메라와 공간 컴퓨팅 기술을 통해 실시간 이미지 추적, 평면 탐지 등을 지원한다. 미국 보스톤의 XanderGlasses는 미국 뉴욕의 뷰직스(Vuzix)의 하드웨어를 가지고, 청각 보조를 위한 실시간 자막 생성에 특화된 AI 기술을 탑재한 스마트 글래스를 발표했는데, 이는 스마트 글래스에서 온디바이스 AI 기능을 제공하기 위해 설계된 증강현실(AR) 플랫폼인 퀄컴(Qualcomm) 스냅드래곤 AR1 Gen 1을 사용한다. 이 플랫폼은 실시간 번역, 비주얼 검색, 오디오 개선 등 스마트 글래스에 필요한 핵심 기능을 효율적으로 처리할 수 있도록 최적화되어 있다. XReal은 번역 기능은 클라우드 기반 AI 모델과 스마트폰 앱을 활용한다.

XanderGlasses는 고급 번역, 대화형 AI는 클라우드 기반 AI를 사용한다. 중국 Sharge사의 서브브랜드인 Loomos.AI는 ChatGPT 기반 AI를 활용하여 실시간 텍스트 생성, 4K 사진 촬영, 1080p 영상 녹화를 지원하는 스마트 글래스를 공개했는데, 이는 온디바이스AI가 아니라 클라우드AI이다. 결국 스마트글래스라는 이름으로 불리지만, 이를 구현하는 기술구조는 다르며, 결국 온디바이스AI와 클라우드AI가 혼합되고 있다는 것을 알 수있다.

그런데, 보청과 같은 특수 목적 온디바이스 AI 스마트 글래스는 과거 삐삐나 시티폰과 같이, 일종의 중간 제품으로 그칠 가능성이 있다. 즉, 보청기가 필요한 사람에게 보청 스마트 글래스를 사주어야 할지, 좀더 기다렸다가 범용 스마트 글래스를 사주어야 할지 등을 고민하게 된다는 것이다. 물론 XanderGlasses는 처음엔 특수 목적 스마트 클래스로 시작해서 앱을 보강하는 형태로, 범용 스마트 글래스로 진화하는 전략을 가지고 있는 것으로 판단한다. 즉, 처음에 어떤 큰 이유로 스마트 글래스를 사게 만들것인가가 관건이다.

퀄컴은 플랫폼을 만들어서, 뷰직스와 같은 하드웨어 회사와 협업하고, 뷰직스는 Xander와 협업하여 Xander가 완성품을 만들어 판매한다. 그런데, 과연 스마트 글래스가 스마트 폰처럼 보편적 기기가 될지는 여전히 미지수이다. 사람들이 보청이 필요한 노인이 자신의 앞에서 스마트 글래스를 착용하는 것은 용인할지 모르나, 정상인이 프라이버시가 요구되는 여러 공간(까페, 식당, 술집 등)에서 스마트 글래스를 착용하는 것을 용인할지는 아직 속단하기 어렵다. 스마트글래스는 디스플레이 기기인 동시에, 카메라가 달린 녹화기기로 여겨지기 때문이다.

XanderGlasses는 이러한 시장의 우려를 세심하게 해결하는 모습을 보여준다. 음성을 텍스트로 변환하는 작업을 온디바이스 AI로 처리하며, 클라우드에 데이터를 전송하지 않는다. 이를 통해 대화 내용이 외부 서버로 유출될 가능성을 차단하고, 사용자의 개인정보를 보호한다. 그리고, XanderGlasses는 카메라 없이 음성을 텍스트로 변환하는 기능만 제공하며, 이를 통해 대중의 신뢰를 확보하고 있다. 어쩌면 스마트글래스는 카메라 기능없이 디스플레이만 존재하는 제품으로 시작해야 하는지도 모른다. XanderGlasses는 이러한 면에서 매우 사려깊은 접근 방법을 취하고 있다.

한국 기업 deepX는 2024년에 이어서 온디바이스 AI에 집중한 기술로 빠른

영상 처리를 시연하여 기업 파트너 고객들의 많은 방문을 유도하였다. AI가 현장에서 활용될 때는 속도 경쟁이며, 기업 고객들이 보안을 중시하므로, 여기서 반도체 기술이 기여하고 있다. 쿠퍼티노에 소재한 AIZip도 해양에서 사용할 Vision AI, 자동차 등에서 활용될 작은 언어 모델이 내재된 하드웨어를 선보였다. 한국과 룩셈부르크의 협력 사례인 Data Design Engineering도 우주 탐사와 같은 극한 환경에서 사용되는 rover와 같은 기계에서 작동하는 온디바이스 AI를 전시하였는데, 오프라인 AI라고 명명하고 있는 것이 특징이다.

한국의 수프리마AI도 ATM에 설치하여 기승하고 있는 보이스 피싱 범죄를 예방하는 AI 내재 제품을 출시하여 CES 2025 최고 혁신상을 수상했다. 수프리마는 출입통제 부문 글로벌 Top 5 기업으로, AI를 통해 고객을 유지하고, 시장을 확대하는 AX전략의 모범적 사례다.

역시 CES 2025 최고혁신상을 받은 웅진씽크빅의 독서 지원 제품인 북스토리도, AI가 내재된 제품이다. 어떤 책이든지 올려놓으면, 책을 읽어주고, 다른 언어로 번역도 해준다. 소장하고 있는 모든 책의 오디오북화가 가능하며, 번역판을 구하지 않아도 자신이 모르는 언어의 책을 읽을수 있는 기회를 준다. 필수 기능은 온디바이스 AI로 구현하고, 부가적인 기능은 ChatGPT와 같은 Cloud AI로 해결하는 방식을 채택한다. 이렇게 현실에서는, 오프라인 AI와 온라인 AI가 결합되는 경우가 더 많다. 북스토리는 아티젠스페이스(ArtiGenSpace)와의 협력을 통해 개발되었는데, 아티젠스페이스는 모델 파라미터를 40억 개 이하로 줄인 경량화된 생성형 AI 모델을 개발하여, 텍스트 변환(TTS, STT), 실시간 번역, 음성 합성 등 다양한 작업을 수행한다. Bookstory 디바이스에는 이 모델을 구현한 신경망 처리 장치(NPU, Neural Processing Unit)가 내장되어, AI 작업을 빠르고, 저전력으로, 인터넷연결없이도 수행한다. 모든 데이터 처리는 디바이스 내부에서 이루어지며, 클라우드 서버로 전송되지 않기 때문에 개인정보 보호와 보안성이 강화된다.

한국 판교의 Maum.AI도 기업 환경에서 프라이버시와 보안을 지켜주면서 생산성과 성과를 달성하는 온프레미스 AI, 온디바이스 AI로 발빠르게 전환하고 있는 모습을 보여주었다. 마음AI는 메타AI의 라마(Llama) 3.1 모델에 기반한 자체 온프레미스(On-Premise) LLM(대규모 언어 모델)인 MAAL(Multilingual Adaptive Augmentation Language-model, 다국어 적응형 증강 언어 모델)을 Apple의 Mac Mini M4에 탑재, 구현하여, 클라우드 AI에의 의존도를 줄이

고 데이터 보안을 강화하면서도, 기업과 개인 사용자가 AI 기술을 효율적으로 활용할 수 있도록 설계했다.

Mac Mini M4는 10코어 CPU, 10코어 GPU, 그리고 16코어 Neural Engine을 탑재하고 있다. 각각이 10개, 10개, 16개의 작업을 동시에 병렬적으로 할 수 있고, 메모리가 이들 연산 장치에 초당 120GB의 데이터를 제공할 수 있어서, 온프레미스 AI 모델 실행에 필요한 고성능 연산 능력을 제공한다. MAAL은 기업 내부 데이터와 인프라에 통합되어, 특정 도메인(법률, 금융 등)에 특화된 지식을 학습하여 사내 문서 분석, 요약, 질의응답 등 반복 작업을 자동화하여 생산성을 높이면서, 고객 데이터를 외부로 전송하지 않고 로컬 환경에서 처리하여 보안을 유지하면서 실시간 상담을 제공한다. Mac Mini M4의 용량을 고려할 때, 80억 파라미터 또는 700억 파라미터짜리 라마 3.1을 사용할 것으로 추정된다.

MAAL이 온프레미스 AI라면 SUDA는 온디바이스 AI이다. SUDA(Seamless Uninterrupted Dialogue Assistant)는 라마 v3.2 1B 모델을 기반으로 한 음성 인식, 텍스트 변환, 자연스러운 실시간 응답 등을 수행하는 음성 대화 솔루션이다. 산업 및 상업용 IoT 애플리케이션에서의 AI 연산을 위해 퀄컴이 설계한 고성능 System-on-Chip(SoC)인 QCS6490 프로세서에 기반한 하드웨어 및 소프트웨어 개발 키트인 RB3 Gen 2 플랫폼을 활용하여 구현되었다. 지연 제로를 추구하므로, 음성을 실시간으로 처리하며, 사용자가 말을 하는 동안에도 동시에 데이터를 분석하고 응답을 생성한다. 사용자가 질문을 중단하거나 새로운 명령을 삽입하더라도 즉각적으로 반응할 수 있다. 스마트 홈 IoT나, 키오스크 및 POS 시스템, 물류 및 헬스케어 장치에서 고객 응대 및 주문 처리, 실시간 명령 처리와 데이터 관리를 할 수 있다. 클라우드 기반 AI 대비 해킹 위험이 낮아 보안성이 높고, 응답 시간이 1.5초 이하로 단축되었다.

마음AI의 이번 CES2025 전시는 이 회사가 고객의 까다로운 요구사항인 지연 제로 문제와, 프라이버시와 보안 문제에 얼마나 민감하게 반응하고 있는지 보여주었고, 고객 요구에 대한 재빠른 대응이 경쟁력있는 신제품을 만들어낸다는 것을 보여주었다. 이러한 고객 요구에 대한 재빠른 대응은 AI 에이전트, 그리고 Vertical AI와 연결된다.

2. AI 에이전트와 Vertical AI

2024년 10월 블룸버그는 AI Agent Economy의 규모를 1조달러로 산정하고, MS는 'Autonomous Agent'로, Anthropic은 'Computer Use'로, Salesforce.com은 'Agentforce', Tencent는 'AppAgent'라는 이름의 제품을 내는 등, 빅테크들이 경쟁적으로 AI Agent 프레임워크를 발표하는 상황이다. 구글은 Astra가, 오픈에이아이는 가칭 Operator가 곧 발표된다는 소식이다. 오픈에이아이의 경우 2025년 3월경에 발표할 o3가 결국 AI 에이전트 서비스가 될 가능성이 크다. 이미 OpenAI의 o1 Pro(월 200달러의 구독료)가 Agent 서비스라는 이름을 붙이지는 않지만, 일종의 에이전트 서비스를 제공하고 있다. 오픈에이아이는 o1이나 o3의 본질을 추론이라는 키워드로 설명하지만, 현재 AI 모델 규모에 대한 Scaling Law가 정체되고 있는 상태에서, AI 모델을 여러번 호출하는 형태(Scaffolding, 스캐폴딩)로 AI 시스템의 성능을 향상시키는 것은, 결국 멀티 Agent 구조를 활용하는 것이다.

즉, 추론을 한다고 표현되지만, 내부적으로는 LLM을 여러번 호출하는 형태의 Agent를 구현하는 상황이고, 그 AI 시스템이 외부의 다른 AI시스템이나 정보시스템과 연결되어 기술적으로는 RAG와 결합하고 마지막에는 인간과 소통하기 위하여 LLM을 활용하는 것이다. 이러한 모델을 가장 모범적이면서, 선구적으로 구현하고 있는 기업이 Perplexity인데, 퍼플렉시티가 이 구조를 계속 정형화해 나가는 과정에서, 다른 기업들이 이에 영향을 받아, 각 분야에 재현해 나가는 과정이 AI Agent 또는 Vertical AI라는 관점에서 볼 수 있는 현상이다.

AI Agent이면서 Vertical AI의 모습을 보여준 CES 2025의 전형적인 사례가 SourcingGPT.ai였다. 홍콩에 소재한 소싱지피티AI는 BuyHive의 자회사로, 고객사의 각종 구매 의사결정을 돕는 서비스를 제공한다. Perplexity의 B2B상거래 버전이라고 할 수 있으며, BuyHive가 가진 방대한 정보와 구성원이 가진 조달(Procurement) 관련 지식을 기반으로, 고객의 질문을 RAG, 프롬프트엔지니어링, 그리고 LLM을 결합하여 에이전트형태로 구현한 것이다. 무역 또는 B2B 거래라는 특정 도메인에 AI를 개발하는 Vertical AI의 전형적인 사례이며, 이러한 서비스의 성능을 올리기 위해서는 다른 여러 AI Agent와도 협력해야 하므로, 자연스럽게 Multi-Agent구조로 진화 발전할 수 있다.

아이티원(IT-One)이 포스코이앤씨와 공동 개발한 '콘크리트 시공이음부 요철

생성 로봇'은 CES 2025에서 로봇 분야 혁신상을 수상했다. 건설 현장에서 콘크리트층 사이의 결합력을 높이고, 작업 효율성과 안전성을 크게 향상시키고 철근 사용량도 줄이는 데 중점을 둔 솔루션으로, 콘크리트를 여러 층으로 타설해야 하는 공사(예: 댐, 교량, 초고층 건축물 기초)에서, 덜 굳은 콘크리트 표면에 요철을 생성하여 층간 결합력을 강화하여, 구조물의 안정성과 내구성을 향상시키는데, 기존 수작업 대비 작업 시간을 최대 85% 단축하며, 자동화된 로봇 시스템으로 균일하고 정밀한 요철 생성이 가능하다. 작업자가 노출된 철근에 의해 다칠 위험을 줄인다. AI모델이 내재되진 않았지만, 라이다 기술로, 어느 정도의 자동주행을 구현하고 있는 적정 기술 적용 사례이며, 건축분야의 골칫거리중의 하나를 해결하는 진통제를 개발한 모범사례로 최고혁신상을 수상할만하다고 평가된다.

웅진씽크빅이나, 소싱GPT.ai, 아이티원 사례 모두 기존의 고객 관계와 전문 지식이 풍부한 기업이 AI를 통해서 성공적으로 전환하면서, 새로운 제품과 서비스를 출시하여 매출을 올리는 동시에, 기존의 고객 관계도 강화하는 결과를 얻어내는 모범사례이다.

2025년 1, 2월에 출간될 <비즈니스 모델과 인공지능>이라는 책에서 다음과 같이 언급한 바 있다. “AI를 BM 혁신에 활용하는 첫걸음은 매출과 연결짓는 것이다. 고객의 새로운 요구를 해결하고, 수주활동의 애로 사항을 없애고, 영업활동을 혁신할 수 있는 방법을 찾는 과정에서 AI를 활용하는 것이다”

기업들은 아이티원, 소싱지피티AI, 웅진씽크빅의 사례를 잘 참고하여 자신들의 기업에 맞는 AX전략을 채택하고 그 결과로 새로운 제품, 서비스를 개발하여 사업화해야 할 것이다. 이런 것들이 AI에이전트이고, Vertical AI이다. 고객이 빠른, 그리고 보안과 프라이버시보호가 강한 AI를 원하면, 그것을 온디바이스, 온프레미스 AI로 구현하는 것이 하나의 옵션이 되면 된다. 그리고, 결국 그것은 온라인 AI와 오프라인 AI가 결합되는 온오프 AI가 될 것이다.

이경전: 경희대학교 교수로 재직 중이며, 카이스트(KAIST)에서 경영과학 학사·석사·박사 학위를, 서울대학교에서 행정학 석박사를 수료했으며, 카네기멜론대학(Carnegie Mellon University), UC 버클리(Berkeley), 매사추세츠공과대학(MIT)에서 초빙과학자·교수로 활동했다. 미국인공지능학회(AAAI)로부터 네 차례 (1995, 1997, 2020, 2024년) 혁신적 인공지능 응용상(IAAI Award)을 수상했다. 2024년에 저서 <AI 에이전트와 사회변화>, <Life with Intelligence: AI는 어떻게 인생의 무기가 되는가>를 출간했고, 2025년 1월에는 저서 <비즈니스 모델과 인공지능>을 출간한다.