엣지 디바이스에서의 AI 이용에 관한 산업분석

윤종혁 정보통신기획평가원 동향분석팀 책임

1. 머리말

2022년 오픈에이아이(OpenAI)가 대화형 챗봇 인공지능(ChatGPT)을 발표하였다. 과거 알파고와 이세돌의 바둑 대국과 같은 놀라움을 선사하였으나, 이전과는 확실한 차이점이 있었다. 우리 모두가 인공지능 기술을 직접 체감할 수 있었다는 것이다. 챗GPT는 이전에 공개된 AI 서비스들보다 접근성과 효용성 차원에서 대폭 발전하여 대중에게 상용화되기 시작하였다. 대형언어모델(LLM)로 데이터센터에서 엄청난 수의 반도체로 학습된 초거대 인공지능 서비스이며, 대량의 데이터 처리와 다량의 컴퓨팅 리소스 소비 등을 특징으로 하는, 공급자 중심의 클라우드 기반 AI(Cloud-based AI)로 발전하였다. 방대한 양의 파라미터를 통한 자연어 학습으로 고성능 대화형 인공지능 서비스를 다수에게 제공할 수 있는 장점이 있으나, 개인정보 유출과 보안 위협, 높은 유지비용, 서비스 속도 지연 등의 단점 역시 존재했다. 대중에 인공지능 서비스의 공급이 확대되며, 기존 방식에 대한 이용자들의 개선 의견이 증가하였고, 공급자 역시 서비스를 운영하는 데 많은 유지비용이 수반됨에 따라, 새로운 형태의 인공지능 서비스 공급 방안이 필요해졌다. 결과적으로 '에지(Edge)에서의 인공지능 운영' 방식이 대안으로 제시되었다.

에지(Edge) AI는 클라우드 기반 AI와 대조적 개념의 운영 방식으로, 클라우드나 중앙 데이터센터가 아닌 네트워크 종단에서 AI 알고리즘 컴퓨팅이 이뤄지는 것을 의미한다. 에지 AI는 중앙 서버에 의존하지 않고 IoT 기기 자체 또는 물리적으로 근거리에 위치한 에지 서버를 주 매개로 하여 데이터 분석과 기기 동작이 이뤄지는 방식이다[1].

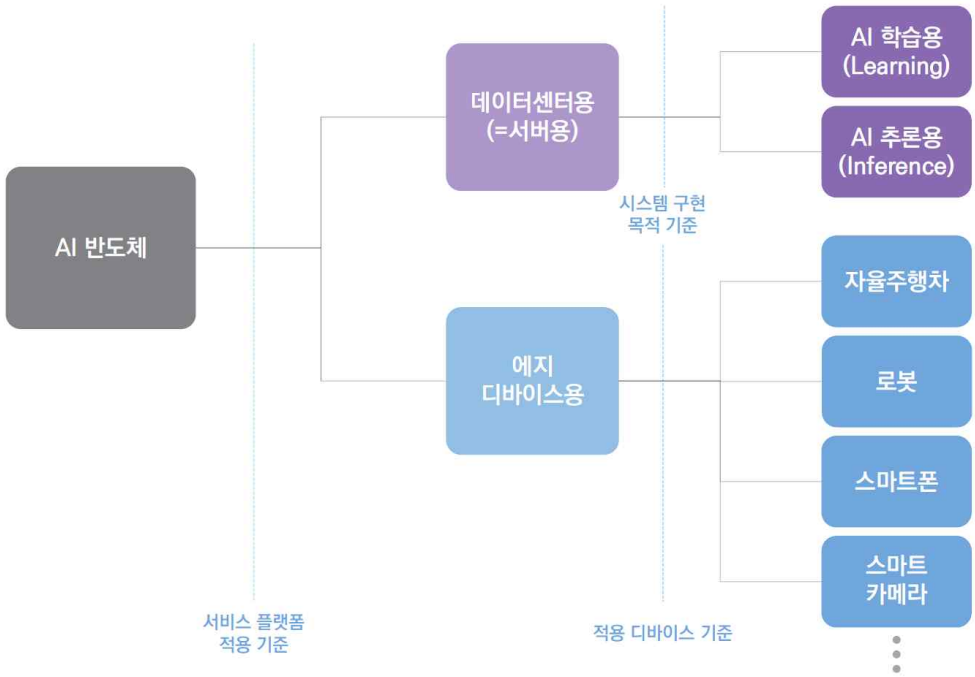
클라우드 AI와 에지 AI의 구현은 기술적으로 반도체에 의해 구분된다. 중심부에서 모든 주변부를 제어하는 방향으로 발전해 왔던 초기 반도체 시장은 데이터센터용 반도체인 CPU, GPU가 주류를 이루고 있었다. 에지로 인공지능의 구동 위치가 이동하게 되면서, 기존 반도체보다 고효율·저전력·소형화 특징을 갖춘 고성능 반도체가 필요해졌다. 이는 에지 디바이스용 반도체로 구분되며, 인공지능 전용 반도체의 플랫폼 기준을 확장하였다([그림 1] 참조).

본고에서는 클라우드 기반 AI 서비스가 에지로 확장되는 이유를 살펴보면서, 에지 AI 구현을 위한 연구현황과 새로운 반도체 시장을 형성하게 된 인공지능 반도체 기업 동향을 살펴보고자 한다.

2. 에지(Edge)로의 인공지능 이동

2.1 클라우드 기반 초거대 AI의 한계

챗GPT 등 생성형 AI의 상용화 성공과 AI 서비스이용자 및 기업 활용이 증가함에 따라 이전엔



※ 출처: 이동수, '지능형 반도체의 새로운 패러다임(2023 ICT 산업전망컨퍼런스 발표)', 네이버, 2022

[그림 1] 인공지능 반도체 응용 분야[2]

개념적인 고민이었던 AI 서비스 사용에서의 개인정보유출, 보안 위협에 대한 이슈가 당장 풀어야 할 핵심 문제로 떠오르게 되었다. 이 외에도 클라우드 기반 서비스 이용에 따른 서버 유지비용, 네트워크 병목 등은 클라우드 기반의 AI 서비스 제공 방식과는 다른 새로운 형태의 서비스 요구로 이어졌다.[3]

2.1.1 정보보안과 데이터조작

CPU·GPU 반도체 기반 대규모 데이터센터의 AI서비스는 이용자들이 입력하는 모든 데이터를 처리·활용·공유하므로, 의도하지 않은 쉐도우데이터(shadow data)[1]가 남아 유출될 가능성이 상존한다. 지난 2023년 3월 삼성전자 디바이스솔루션 부문 사업장 내에서 반도체 '설비계측, 수율·불량' 등 총 3건의 보안정보가 챗GPT를 통해 유출되어 AI 학습 데이터로 활용되는 사고가 있었고, 한번 유출된 데이터는 회수할 수 없었다. 이런 이유로 JP모건체이스, 씨티그룹, 골드만삭스 등 월가 주요 은행은 사내 데이터로 챗GPT 등 AI 챗봇을 사용하는 것을 제한하였다. 또한, 챗GPT 플러그인 서비스가 해킹되어 연동된 앱과 고객 데이터가 노출되는 사고가 발생함에 따라 생성형 AI에 앱을 연결하는 것에 대한 우려를 표하는 의견도 제시되고 있다.

생성형 AI에 조작된 데이터를 입력하거나 AI가 학습한 기존 데이터에서 개인정보만 뽑아내는 방식의 해킹이 대표적인 위변조 공격 방법이며, 인공지능 서비스가 확대되는 만큼 사이버범죄 역시 증가하는 추세이다. 글로벌 보안업체 체크포인트에 따르면 프롬프트 엔지니어링을 통해 해킹에 악용될 수 있는 질문을 차단하고 있으나, 방어자의 보호 조치를 우회하는 방법은 챗GPT에도 열려있어 언제든 사고가 생길 수있는 상황이다.([그림 2] 참조)

1) 쉐도우데이터(shadow data): 인지 혹은 통제 없이 생성·저장·유통되는 데이터, 모니터링 시스템 밖에 존재하며, 승인되지 않거나 알려지지 않은 애플리케이션에 저장됨

※ 출처: "'챗GPT악용' 범죄 늘었는데…양날의 검, 국내업체엔 기회?" 중앙일보, https://www.joongang.co.kr/article/25171421#home

[그림2] 생성형 AI 모델을 대상으로 한 데이터조작 위협

2.1.2 응답지연과 유지비용

클라우드 기반의 생성형 AI 서비스는 중앙 데이터센터로의 데이터 수·송신 과정이 필요하며, 이용자가 과도하게 집중되거나 데이터의 양이 커질 경우 데이터 병목 현상이 발생하게 된다. 챗GPT 서비스의 시작 초기, 하루 약 1,500만 명 이상의 이용자가 접속하여 메모리 부하와 함께 간단한 질문에도 응답속도에 일부 장애가 발생하는 현상이 발견되었다.

데이터센터는 대용량 데이터의 동시 처리를 위해 수많은 반도체의 묶음으로 구성되며, 이를 운용하기 위해서는 높은 전력이 필요하고 과열을 방지하기 위한 냉각시스템 비용이 발생하는 등 유지비용이 과다하게 발생한다. 이 비용은 서비스 이용자에게 부과되며, 챗GPT 운용을 위해 하루에 약 70만 달러(약 9억원) 이상의 비용이 발생한다.

2.1.3 실시간 현장 대응

데이터를 수집하거나 입력하는 디바이스와 데이터를 처리하는 중앙 서버가 물리적으로 떨어져 있는 경우, 디바이스 주변 실시간 환경을 인지하기 어려워 사고 대응에 어려움이 있을 수 있다. 가령 수천분의 1초라도 데이터 분석에 지연이 발생할 경우, 큰 사고를 초래할 수 있는 자율주행차, 산업·보안시설, 항공 등 분야에서는 클라우드 기반 AI 외에도 추가적인 보완 프로그램을 필요로 한다.

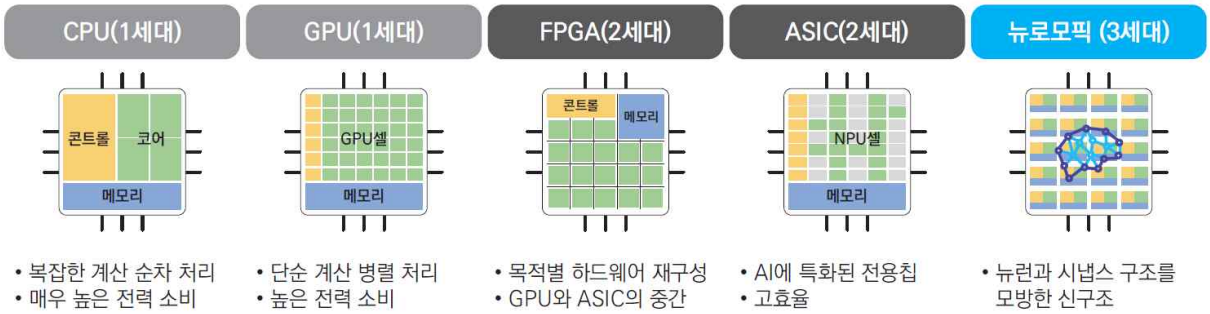
2.2 사용자 맞춤형 AI 서비스 확산

개인 디바이스 보급이 늘어나고, 디바이스 의존도가 높아짐에 따라 서비스 공급 기업들은 경쟁적으로 개별 디바이스에 AI 애플리케이션을 탑재한 서비스 공급을 확대하고 있다. 가장 오랜 시간 소지하는 스마트폰은 사용자와 밀접한 환경에서 수집·검토된 데이터를 '개인화'하여 사용자 맞춤형 서비스를 제공할 수 있다.

2024년 1월 삼성전자는 갤럭시 S24에 온디바이스 AI를 탑재한 최초의 인공지능 스마트폰을 출시하였다. 실시간 통화 통역으로 언어 장벽을 허물고, 카메라 촬영 후 결과물의 편집까지 스마트폰 자체 AI가 분석하여 맞춤 도구로 제안을 해준다. 퀄컴과 협력하여 개발한 갤럭시용 스냅드래곤8 3세대 AP를 탑재하였으며, 비행기모드에서도 스마트폰 자체적으로 AI 서비스를 제공한다.

3. 저전력·고성능·소형 반도체와 알고리즘 경량화

개별 디바이스에서 AI 알고리즘 작동이 가능케 하는 인공지능 반도체는 소형화·저전력·고성능 설계 및 호환의 유연성이 필요하여 ASIC[2], PIM[3] 중심으로 발전하게 되었다.([그림 3] 참조) 한정된 전력과 소규모 알고리즘으로 외부 명령에 따른 AI 처리를 수행하고 결과를 도출해야 하므로 학습된 정보를 기반으로 추론 작업을 수행할 추론용 AI 반도체가 필요하며, 추론용 반도체는 연산 가속·초저전력·경량화의 특징을 갖는다. 2세대 반도체보다 진보한 3세대 반도체는 기존 폰노이만 방식을 탈피한 PIM이나 뉴로모픽 반도체가 대표적이며, 현재 상용화를 위해 글로벌 기업과 대학들이 연구에 참여하고 있다.



※ 출처: 이선재, 'AI와 AI반도체 생태계 특징 및 시사점-팹리스 스타트업을 중심으로', 2022

[그림 3] 생성형 AI 모델을 대상으로 한 데이터조작 위협[4]

물리적인 반도체의 소형화와 함께 소프트웨어인 알고리즘의 경량화도 함께 구현되어야 에지에서의 AI연산 성능 확보가 가능하다. 경량 딥러닝 연구는 기존 클라우드 기반의 학습된 인공지능 모델을 에지 디바이스에 내장하기 위한 필수 기술이다. 경량 딥러닝 기술은 알고리즘 자체를 효율적 구조로 재설계하여, 기존 모델 대비 효율을 극대화하는 경량 딥러닝 알고리즘 기술과 설계된 알고리즘의 파라미터를 이후에 축소하는 알고리즘 경량화 기술로 구분된다.

4. 주요 업체 동향

4.1 1세대 AI 반도체

엔비디아는 에지 AI 확산의 선두로 나서고 있으며, 자체 에지 제품군을 판매할 뿐 아니라 다양한 모바일 칩 아키텍처에 자체 GPU 기술을 '쿠다(CUDA)'로 통합하여 제공하고 있다. 자율주행을 비롯한 에지 컴퓨팅용 젯슨(Jetson) 시리즈를 출시하였으며, 중국 3대 전기차 기업인 리오토(Li Auto) 신차의 자율주행기술에 젯슨 AGX 오린(Jetson AGX Orin)을 적용하였다.

인텔은 CPU와 함께 비전프로세서(VPU)를 에지 AI 구동을 위한 하드웨어 라인업으로 추가하였다. 각 라인업마다 소프트웨어 툴을 동시 제공하고 있어 편의성을 더한다. 하바나 랩을 인수하여 기존 CPU 및 GPU 이외 ASIC 기반의 추론용 가속기인 가우디2와 고야(GOYA)의 후속 제품 그레코(GRECO)를 출시하여 인공지능 반도체 라인업을 강화하고 있다.

2) ASIC(Application specific integrated circuit): 명확한 애플리케이션과 목적을 가진 시스템을 저전력으로 구동하기 위해 활용하는 주문형 칩(SoC)
3) PIM(Processor-in-memory): 메모리 반도체와 시스템 반도체 기능을 결합한 인공지능 반도체

4.2 2세대 AI 반도체

애플은 2018년 자체 설계한 A12 Bionic부터 영상처리, 음성/얼굴인식, 게임 등의 다양한 응용을 위해 NPU[4]를 탑재, 자사 모바일 및 컴퓨터 제품군에 적용하여 경쟁력을 확보하고 있다. 애플의 NPU는 다층 인공신경망의 추론이 단말기에서도 정상 작동할 수 있도록 CPU에 비해 맥(MAC) 연산 작업의 효율을 50배나 높인 가속기이며, 2020년 이후에는 랩탑용 프로세서에도 탑재하고 있다. 또한 온디바이스 AI 스타트업 실크랩스와 엑스노AI를 인수하는 등 인공지능 반도체 개발에 투자를 확대하고 있다.

테슬라는 자율주행 차량의 인공지능 연산을 처리하는 FSD(Full Self Driving) 칩을 자체 개발하였다. 2024년 1월 베타 버전12를 출시하였으며, 종단간(end-to-end) 신경망이라 불리는 기능이 포함된다. 테슬라가 개발한 FSD 칩은 1개의 GPU 코어, CPU 4개를 가지고 있는 ARM의 Quad A72 칩을 3개(총 12개의 CPU) 사용한다.

퀄컴 AI리서치는 15억 개 파라미터를 포함하는 이미지 기반 이미지생성(Image-to-Image) 모델 컨트롤넷(ControlNet)으로 스마트폰에서 실행가능한 데모 모델을 공개하였다. 또한 최신 칩셋인 스냅드래곤8 2세대를 통해 전 세대 대비 AI 성능이 4.35배 좋아졌고, 추론 연산에서 와트당 60%의 성능 향상을 보였다.

구글은 2019년 소형장치에서 추론이 가능한 에지 TPU 칩셋을 발표하고, 칩셋이 장착된 코랄개발보드(Coral Dev. board)를 출시하여 음성 및 영상 인지작업에 활용하였다. 구글 픽셀폰 내 온디바이스 AI 작동을 위한 픽셀 비주얼 코어(Pixel Visual Core)를 탑재하기도 하였다.

4.3 국내 AI반도체

메모리 반도체 분야의 초격차 기술을 보유한 삼성전자와 SK하이닉스는 메모리에 연산 유닛을 집적한 PIM 구조의 차세대 메모리 개발에 매진하고 있다. 삼성전자는 '핫칩스 2023'에서 고대역폭메모리(HBM) PIM과 저전력 더블데이터레이트(LPDDR)[5] 등 두 연구성과를 공개하였으며, AMD의 GPU와 대비했을 때 성능과 전력효율 측면에서 2배 이상임을 확인하였다.

SK하이닉스는 PIM 기술이 적용된 사피온 AI반도체와의 결합 결과물인 X330을 공개하였다. X330은 사피온의 AI 반도체 가운데 처음으로 SK하이닉스 D램을 탑재한 제품이다. 기존에는 엔비디아, AMD, 인텔 등 글로벌 반도체 기업의 최신 GPU에만 공급되었으나 SK그룹 차원에서 AI 반도체 사업을 전략적으로 육성하기 위해 X330에도 D램을 탑재하기 시작하였다.

국내 중소·중견 팹리스 기업의 인공지능 전용 반도체 개발 현황은 <표 2>와 같다.

5. 맺음말

최근 클라우드를 기반으로 한 인공지능 서비스의 확장성 한계를 확인할 수 있었으며, 이에 따라 이를 극복할 수 있는 새로운 개념의 '에지 AI' 시장이 빠르게 성장하고 있다. 소비자 개인 맞춤형 서비스를 제공할 수 있다는 장점으로 시장 선점을 위한 경쟁이 치열하며, 특히 고성능·고효율·저

4) NPU(Neural Processing Unit): 수많은 신경세포와 시냅스로 연결되어 신호를 주고받으며 여러 작업을 병렬(동시)로 처리하는 인간의 뇌 묘사

5) LPDDR(Low Power Double Data Rate): 극단적인 저전력을 목표로 만든, '모바일용 DDR'의미

<표 2> 국내 팹리스 업체 인공지능 반도체 개발현황(2023 인공지능 반도체[6] 재구성)

	개발 내용
사피온	• 자율주행 전용 AI반도체 X340 생산을 위해 텔레칩스와 협력, 창사 이래 처음 설계자산(IP) 타사 공급 사례이며, 2026~2027년 양산을 목표로 하고 삼성전자 파운드리에서 칩 생산 예정 • 스마트폰 등 에지 디바이스용 AI반도체 X350 2024년 상반기 공개 예정
퓨리오사AI	• 엔비디아 T4 대비 4배 성능의 영상인식에 특화된 NPU 'Warboy' 개발 • 추론성능에서 엔비디아 A100과 경쟁 가능한 기술로, GPT-3와 같은 자연어 처리를 위한 거대 인공신경망용 반도체 출시 계획 발표 • 자율주행차·클라우드·의료분야 영상진단 등 최첨단 기술에 활용과 카카오, 네이버 등 컴퓨터비전-메타버스-하이퍼스케일 분야의사업화 추진 중
리벨리온	• 인텔 Goya보다 성능 우위 30%의 금융에 특화된 NPU 아이온 칩 발표(2021.12) • 실시간 트레이딩과 같이 빠른 처리속도가 중요한 금융 분야 AI응용에서 엔비디아 A100보다 연산속도는 10배 빠르고, 전력소모는 10W로 경쟁제품의 10% 수준으로 발표
딥엑스	• 에지 디바이스, 자율주행차 등 각 애플리케이션에 특화된 NPU인 '제네시스' 개발(2022) • CES 2024에서 스몰 센서부터 AI서버까지 적용 가능한 '올인포 AI토탈 솔루션' 공개
디퍼아이	• 팹리스 기업으로서 NPU를 내장한 CCTV 및 로봇용의 AI 반도체 SoC 양산 예정 • 다수의 AI 반도체 개발 특허를 보유하고 있으며 에지용 AI 반도체를 위한 NPU 기술 독자 확보

전력 인공지능 전용 반도체의 개발과 알고리즘 경량화 연구가 활발하다. 에지 AI는 클라우드 기반 서비스가 효율적으로 작동하지 못하는 영역을 담당하는 개념으로, 클라우드 시장을 '대체'하는 것이 아니라 클라우드와 '동시성장'하는 별도 서비스 시장을 형성할 것이다. 데이터센터용 반도체와 더불어 에지 디바이스용 반도체로의 확장을 통해 인공지능 전용 반도체 시장은 계속해서 성장할 전망이다.

일반적으로 우리나라는 반도체 강국이라는 인식이 있으나, 이는 메모리 반도체에 국한된 이야기이다. 시스템 반도체를 기반으로 하는 AI 전용 반도체 분야에서 한국 업체 글로벌 시장 점유율은 3% 정도의 미미한 수준이다. 대기업 위주의 반도체 산업 환경이 형성되어 있는 우리나라로서는 새롭게 형성되는 에지 AI 반도체 시장의 경쟁력 확보를 위해서 팹리스(Fabless)-파운드리(Foundry) 연계 지원 등을 통해 국내 시스템 반도체 스타트업 육성에 집중해야 할 것이다.

[참고문헌]

[1] 신성식 외 3인, '엣지 컴퓨팅 시장 동향 및 산업별 적용 사례', 한국전자통신연구원, 2019
[2] 이동수, '지능형 반도체의 새로운 패러다임(2023 ICT 산업전망컨퍼런스)', 네이버, 2022
[3] 홍정하 외 2인, '엣지 컴퓨팅 기술 동향', 한국전자통신연구원, 2020
[4] 이선재, 'AI와 AI반도체 생태계 특징 및 시사점-팹리스 스타트업을 중심으로', 2022
[5] 이용주 외 3인, '경량 딥러닝 기술 동향', 한국전자통신연구원, 2019
[6] 채명식, 이호윤, '2023 인공지능 반도체', 한국과학기술기획평가원, 브리프 65호, 2023

※ 출처: TTA 저널 제211호