온디바이스 AI 산업·기술 경쟁력 강화 방안에 관한 연구

양현

정보통신기획평가원

myway@iitp.kr

A Study on Strengthening the Competitiveness of On-Device AI Industry and Technology

Hyun Yang

Institute of Information & Communications Technology Planning & Evaluation

요 약

본 연구에서는 온디바이스 AI 글로벌 시장을 선점하기 위한 온디바이스 AI 경쟁력 강화 방안을 모색하였다 먼저는 다가올 무한경쟁의 파고를 넘을 수 있는 튼튼한 산업경쟁력이 요구되며, 누구에게도 종속되지 않는 우리 주도의 산업생태계 구축도 필요하다. 또한, 좁은 국내를 넘어 넓은 세계를 향해 과감히 도전할 필요가 있다. 전 산업 파급력이 큰 온디바이스 AI 기술로 미래 디지털 시대를 주도할 수 있도록 발빠르게 대응해 나갈 수 있길 기대한다.

Ⅰ. 서 론

AI 일상화 시대가 개화하며 클라우드를 통한 대규모 데이터 처리를 넘어, 기기 자체에서도 AI가 구동하는 '온디바이스 AI'가 크게 부각되고 있다. AI 내재화를 통해 전 산업의 경쟁 구도가 새롭게 재편되는 상황에서 기기에 최적화된 AI가 경쟁력의 핵심으로 부상 중이다. 스마트폰·로봇·헬스케어 등 다양한 기기에 AI가 융합된 제품이 등장하며 온디바이스 AI 경쟁이 가시화되고 있다. 온디바이스 AI는 지속가능성·경제성 등에 유리하여 AI반도체의 새로운 수요를 창출할 수 있다. 또한, 온디바이스 AI는 기기 자체에서 AI 추론을 수행함에 따라 데이터센터의 AI 연산 부담을 줄이는 데도 크게 기여할 수 있다. 본 연구에서는 이러한 온디바이스 AI의 글로벌 시장 선점을 위한 산업·기술 경쟁력 강화 방안에 대해 모색하고자 한다.

Ⅱ. 본론

가. 온디바이스 AI의 정의

온디바이스 AI는 해외에서 조금씩 다르게 정의하고 있으나 본질적으로는 일맥상통한다. 퀄컴은 온디바이스 AI를 '데이터를 로컬에서 처리하여 클라우드 기반 솔루션에 대한 의존도를 줄여 개인정보보호 및 보안성을 강화한 것'이라고 정의한다.[1] 가트너의 경우는 온디바이스 AI의 대표적인 분야인 AI PC에 대해 '디바이스에서 AI 작업을 최적화하도록 설계된 전용 AI 가속기가 장착된 PC'라고 정의하고 있다.[2] 이렇듯 온디바이스 AI는 디바이스 자체에서 인공지능 학습·추론을 수행하고 데이터 수집·연산·처리가 가능한 '디지털 기기에 탑재된 AI'를 의미한다고 정의할 수 있다. 한편 온다바이스 AI 범위는 AI HW(AI반도체), AI SW(언어모델 등), 디바이스(자율차·스마트폰·CCTV 등) 등 온디바이스 AI의 전·후방 산업 전반을 포괄한다. 또한 핵심기술로는 적은 학습량과 저비용으로 기기 단에서 AI가 구동되도록 지원하는 HW 고성능화, SW 경량화 및 시스템SW 최적화 기술을 둘 수 있다.

<표1> 온디바이스 AI 핵심기술

구분	특성	주요내용
HW기술	고성능화	• 기존 범용 GPU와 달리, 특정 알고리즘에 특화하여 연산과 에너지 효율을 높일 수 있는 'AI 반도체' * 신경망처리장치(NPU), 주문형 반도체(ASIC), PIM 반도체 등
SW기술	경량화	• 한정된 연산 능력과 메모리에서, 최적화된 인공지능 학습과 추론성능을 낼 수 있는 'AI 알고리즘' * 경량 알고리즘(모델 구조 변경), 알고리즘 최적화(불필요한 파라미터 제거) 등
최적화	최적화	• HW 성능을 향상시키거나 효율적으로 작동시키기 위해 HW-SW 간의 상호작용을 최적화하는 '하드웨어 인지 SW' * 그래픽카드 드라이버, IoT 임베디드 시스템 SW 등

나. 온디바이스 AI의 중요성

온디바이스 AI는 AI 일상화 시대를 이끄는 새로운 혁신의 추동력이 될 수 있다. 대규모 데이터 처리가 가능한 기존의 방식(클라우드 AI)과 함께, 최근 소규모 연산에 적합한 효용성 높은 온디바이스 AI가 부각되고 있다. 클라우드 AI의 보완재로 저전력·저비용·보안성·맞춤 서비스 등에 유리한 온디바이스 AI의 발전과 확산이 전망된다.[3]

<표2> 클라우드 AI vs. 온디바이스 AI 특징 비교

구분	클라우드 AI	온디바이스 AI
파라미타수	수천억 ~ 2조 개	수십억 무 100억 개
응답속도	수 sec ~ 수십 sec	~ msec
소비전력	수십 ~ 수백 W (서버)	10mW ~ 수 W(스마트폰)
장점	학습 및 추론성능 극대화	저전력·저비용·보안성·맞춤서비스·빠른 응답 등
단점	고전력·고비용, 네트워크 부하 상승 등	추론 정확도 감소, 주기적 최신화 요구 등

또한, 온디바이스 AI는 반도체·SW·디바이스 등 디지털 전 산업 재편의 축이 될 것으로 예상된다. AI 추론과 저전력·고효율에 유리한 NPU가 대두되며, 온디바이스 AI가 AI반도체 산업의 새로운 수요를 창출하는 성장 동력으로 부상하고 있다. 온디바이스 AI가 탑재되는 기기·서비스 등 전방 산업까지 파급되며 디지털 산업 전반의 성장을 견인할 것으로 예측된다. 특히, 스마트폰·자동차 등 기존 산업의 재도약을 이끄는 동시에, 웨어러블·로봇·CCTV 등 신산업 확산의 핵심동인이 될 전망이다.[4]

<표3> 온디바이스 AI 품목별 시장 전망(마켓엔마켓, 백만달러)

구분	2023년	2027년	CAGR
스마트폰	22,870.78	32,072.76	8.8%
자동차	458.5	3,120.74	61.5%
웨어러블	15.93	209.71	90.5%
로봇	13,253.16	82,939.67	58.2%
CCTV	711.95	2,005.98	29.6%

이러한 온디바이스 AI는 글로벌 디지털 산업 생태계를 주도할 새로운 기회를 형성할 수 있다. 과거 정보통신 산업의 전환기(아날로그→디지털TV, 피처폰→스마트폰 등) 때마다, 시장·기술 대응 여부가 해당 산업의 성공과 실패를 결정하였다. 노키아가 휴대폰 세계 시장 1위('98~'11) 기록 후, 시장 대응 실패로 MS에 휴대폰 부문을 매각('13)한 사례가 대표적이다. AI 대전환기, 온디바이스 AI에 대한 대응 여부가 미래 디지털 산업에서의 낙오와 도약을 판가름하는 시금석이 될 예정이다.

다. 국내외 동향

먼저 해외동향을 살펴보면, 글로벌 팹리스 중심으로 NPU 등 AI반도체 개발에 박차를 가하고 있다. 퀄컴·미디어텍·엔비디아·인텔 등은 기기 내 생성형 AI 실행을 위해 빠른 데이터 처리·전력 효율성에 강점을 갖는 NPU 기능을 강화하고 있다. 퀄컴은 생성형 AI용 AP '스냅드래곤8 3세대'를 출시하였고, 미디어텍은 자체 개발 NPU 탑재 AP '디멘시티 9300'를 출시하였다.[5] 또한, 엔비디아는 AI PC용 신형 GPU 'RTX 4080 슈퍼'를, 인텔은 CPU 최초 NPU 탑재 '메테오레이크'를 개발하였다. 한편, 언어모델 부분에 있어서는 글로벌 빅테크 주도로 대규모언어모델(LLM)의 경량화에 집중하고 있다. 구글·MS·메타·애플 등은 컴퓨팅이 한정된 기기 내에서 원활히 구동될 수 있도록 자체 대규모 모델의 경량화를 적극 추진 중이다. 구글은 경량언어모델 '제미나이 나노'를 픽셀 8 프로에 적용하였고, MS는 소규모 언어모델 '파이2'를 개발하였다. 메타는 대규모 언어모델을 경량화한 '라마2'를, 애플은 iOS 17에 '온디바이스 AI 관련 기능 추가'하였다. CES 2024와 NWC 2024에서도 디바이스 기업의 온디바이스 AI 적용 가시화되고 있음이 확인된다. 스마트폰·CCTV·로봇·헬스케어 등 다양한 분야에서 온디바이스 AI가 탑재된 기기·서비스가 비로소 선보이기 시작했다. 보쉬의 영상분석 AI 탑재 '총기 감지 CCTV', 도이치텔레콤의 앱 설치가 필요 없는 '앱프리 AI폰', 니콘의 협동 로봇용 '비전 AI 카메라', 바라코다의 생성형 AI 기반 감정케어 '스마트 거울' 등이 대표적이다.

국내 동향도 살펴보면, 온디바이스 AI 시대를 대비하는 AI HW 기업의 성장 잠재력이 충분해 보인다. 종합반도체기업, 중소 팹리스 등 국내 기업은 연산 효율성이 높고 소비전력이 낮은 NPU를 개발 또는 준비하는 초기 단계 진입하고 있다. 특히, 기술경쟁력을 갖춘 클라우드용 NPU를 기반으로 한 온디바이스용 NPU 분야 진출이 가시화되며 향후 높은 성장이 기대된다. 또한, 민첩한 온디바이스 전용 생성형 AI모델 개발로 AI SW 시장 주도도 기대된다. AI SW 기업은 자사 기기 탑재를 목표로 생성형 AI를 직접 개발하거나, AI 알고리즘 고도화를 통한 소형언어모델(sLLM) 개발을 추진 중이다. 특히, 대기업뿐만 아닌, 벤처·스타트업에서도 온디바이스 AI에 적합한 언어모델을 개발·준비하며 AI SW 시장 선점에 주력하고 있다.

라. 진단 및 방향

먼저는 다가올 무한경쟁의 파고를 넘을 수 있는 튼튼한 산업경쟁력이 요구된다. 각국이 HW고성능화·SW경량화 등 온디바이스 AI시대를 준비하는 세계적인 추세 속, 우리 기업 또한 성장 잠재력이 충분하다. 우리의 잠재력을 실제 고성장으로 이끌 수 있도록, 온디바이스 AI 전략품목 육성과 HW·SW 핵심기술 선제 확보 추진해야 할 것이다.

누구에게도 종속되지 않는 우리 주도의 산업생태계 구축도 요구된다. 해외기업은 글로벌 AI 생태계에서 이미 확보한 막대한 영향력을 바탕으로, 온디바이스 AI 분야 진출을 적극 모색 중이다. 특히, 엔비디아는 GPU와 AI개발 플랫폼(CUDA) 등을 통해 클라우드 AI 생태계 주도하고 있다. 따라서, 글로벌 기업에 종속(lock-in)되지 않도록, 연구환경·인재·법제도 등 산업 기반을 고도화하여 경쟁력 있는 온디바이스 AI 생태계 구축이 중요하다.

마지막으로 좁은 국내를 넘어 넓은 세계를 향해 과감히 도전할 필요가 있다. 국내 시장만으로는 우리 기업의 성장에 한계가 있으며, 우리 기업 간 상생보다는 경쟁이 우선시 될 우려가 있다. 모든 가능성이 열려 있는 온디바이스 AI 시장에서 적극적인 수출지원과 글로벌 협력체계 구축으로 글로벌 초기시장을 주도해야 할 것이다.

Ⅲ. 결론

본 연구에서는 우리의 디지털 역량을 결집, 온디바이스 AI 글로벌 시장을 선점하기 위한 온디바이스 AI 경쟁력 강화 방안을 모색하였다. 온디바이스 AI 전·후방 산업의 핵심 경쟁력과 생태계 기반을 조기 확보하고, 세계 초기시장을 적극적으로 공략하며 성장의 기틀을 다질 필요가 있다. 그간 정부는 AI반도체를 미래 신산업 성장을 위한 필수 인프라로 주목하고 정책적 지원과 투자를 강화하였다. 이를 통해, 논문·특허 등 기술적 성과가 향상되었으며, 특히 AI반도체 기업의 매출 성장과 고용 창출을 견인하였다. 축적된 AI반도체 토대 위에, 전 산업 파급력이 큰 온디바이스 AI 기술로 미래 디지털 시대를 주도할 수 있도록 발 빠르게 대응해 나갈 수 있길 기대한다.

ACKNOWLEDGMENT

참 고 문 헌

[1] Jilei Hou, “The future of AI is hybrid”, Qualcomm Technologies, June, 2023

[2] "Gartner Predicts 295 Million AI-Enabled PCs and AI-Enabled Smartphones Will Ship in 2024", Gartner, February, 2024

[3] 김민진,"클라우드 기반 AI에 대한 엣지 AI의 도전과 영향“, 정보통신정책연구원, 2020.8

[4] “Artificial Intelligence Market, Global forcast”, MARKETS AND MARKETS, July 2023

[5] Joseph Soriaga, “Generative AI at the edge”, QCOMResearch.. November 8, 2023