새로운 AI 패러다임의 시작, 온디바이스 AI

글. 한영주 한국방송통신전파진흥원 연구위원 / 언론정보학 박사

#온디바이스 AI #온디바이스 AI의 구현 방식 #온디바이스 AI를 주목하는 이유

AI 패러다임은 계속 진화하며 확장한다. 지난해 AI 논점은 기술이 구현될 수 있는 환경 인프라를 만드는 것으로, AI 기술의 개발이나 고도화와 같은 후방 산업에 주력했다. 반면 올해부터 AI 논점은 영역별 특성에 맞게 AI 기술을 잘 수용하고 활용할 수 있을지에 초점을 둔다. 수요자의 접점이 형성하는 서비스 부분에서 AI 가치를 찾으며 전방 산업에 집중하고 있다. 올해 초 세계적인 가전박람회 CES 2024에서도 AI 기술이 본격적으로 응용과 확장을 통해 기술 적용 분야가 더 다양해질 것으로 전망한 바 있다. 이와 유사하게 지난 4월 미국 라스베이거스에서 열린 2024 NAB Show(전자미디어 쇼)에서는 AI 기술을 어떻게 수용하고, 또 어떻게 공존할 것인지에 대한 전시와 컨퍼런스가 주를 이루기도 했다.

이러한 기술 전망처럼 매번 놀라운 결과들로 화제의 중심이었던 생성형 AI는 어느새 전방 산업을 통해 우리 일상으로 스며들고 있으며, 그중 하나는 온디바이스 AI(on-device AI)를 통해 개인 디바이스로 생활 밀착형의 밀도 높은 침투가 시작되었다는 것이다.

온디바이스 AI의 구현 방식

온디바이스 AI는 각각의 디바이스에서 자체적인 AI 모델을 구현해서 데이터를 수집하고 처리하는 방식을 말한다. 그간 AI 기반 서비스는 대부분이 클라우드 시스템을 활용했는데, 이는 방대한 양의 데이터를 클라우드로 수집하고 분석한 후, 그 결과를 각 디바이스로 전송하는 형태이다. 클라우드 AI 시스템은 오픈AI ChatGPT, 구글 Gemini, 마이크로소프트 Copilot 등 많은 생성형 AI 기반의 서비스에 대규모 언어 모델(Large Language Model, LLM)을 구현하기 위해 채택하는 방식이다. 클라우드 AI 시스템은 이용자가 특정 데이터, 혹은 태스크를 요구하면 외부에 있는 대형 데이터센터에서 데이터 분석, AI 모델 학습, 추론 등 주요 프로세스를 처리하기 때문에 LLM 같은 큰 규모의 모델을 활용할 수 있는 장점이 있다.

반면 기술적, 기능적 측면에서 온디바이스 AI는 개별 디바이스 단위로 AI 모델이 구현되므로 반응 속도를 빠르게 끌어낼 수 있고, 이용자에게 최적화된 맞춤형 기능을 강화할 수 있다. 기존 클라우드 AI 시스템은 데이터 요청과 전송을 외부 서버에 존재하는 클라우드를 거치며 많은 인프라 자원을 활용했는데, 온디바이스 AI는 개별 디바이스에서 AI가 구동되므로 개별 데이터를 외부 서버로 이동할 필요가 없어서 이전보다 적은 인프라 자원을 사용하며 안정적이고 효과적인 결과를 제공할 수 있다. 즉 이용자의 디바이스 안에서 AI 모델을 통해 데이터의 처리, 분석, 추론 등 프로세스를 비교적 빠른 속도로 처리해서 결과를 생성할 수 있고, 디바이스 내부에서 자체적으로 AI 모델을 구현하므로 이용자 데이터를 외부로 전송할 필요가 없어서 높은 보안 효과까지 기대할 수 있다.

아마 지금쯤이면, 온디바이스 AI가 전혀 생소한 새로운 개념이 아니라는 것은 어느 정도 눈치챘을 것이다. 온디바이스 AI는 AI 기술을 처리하는 방식이 디바이스 내부에서 이뤄진다는 것인데, 이미 스마트폰에서 날씨, 시간, 전화 걸기 등으로 활용해 온 삼성의 빅스비나 애플의 시리와 같은 AI 음성 비서로 경험해봤기 때문이다. AI 음성 비서는 이용자가 요구한 정보를 검색해서 그 결과를 음성이나 디바이스로 제공하는데, VUI(Voice User Interface) 시나리오가 고정된 형태라서 비교적 간단한

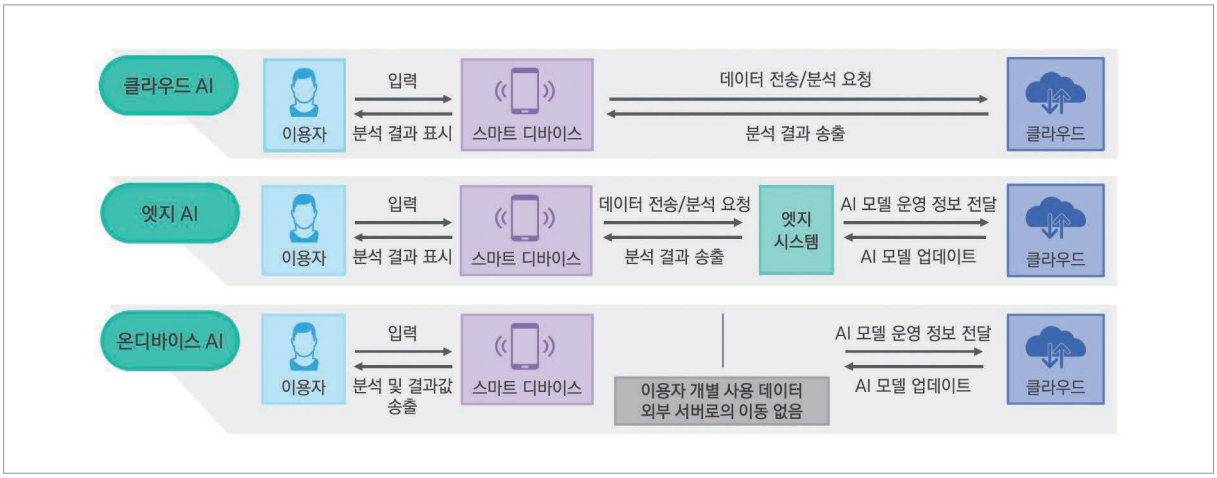

그림 1. AI 운영 방식의 변화 / 출처 : 삼정KPMG 경제연구원(2024. 6).

정보와 기능만을 제한적으로 제공할 뿐이었다. 스마트폰에 탑재한 AI 음성 비서는 혁신적이었지만 이용자가 별도로 이용 방법을 숙지하고 정해진 질문과 답변 순서를 통해 대화가 가능한 형태였기 때문에 예상했던 기대 수준에는 미치지 못했다.

이와 달리, 온디바이스 AI는 디바이스 내 AI 모델로 얼굴 인식, 음성 인식, 사진 보정 등 인물, 사물, 동물 등을 인식하고 식별하며 데이터를 자동으로 처리해준다. 즉 온디바이스 AI는 기존 AI 음성 비서의 개념을 혁신한 것이자, AI 컴퓨팅 시스템을 축소하여 개개인의 손안으로 넣어준 셈이다. 개별 디바이스에서 AI 모델이 구현된다고 해도 AI 모델의 운영 정보나 업데이트는 클라우드를 거쳐 주고받아야 한다. 이는 엣지 AI 시스템에서 이용자와 물리적으로 가까운 곳에 엣지 서버를 두고 데이터를 처리하는 방식과도 유사하다. 이런 면에서 온디바이스 AI는 엣지 AI 시스템의 장점을 극대화하고 클라우드 AI 시스템의 단점을 보완하여 개개인의 디바이스 소유자에게 최적화된 서비스를 구현할 것으로 전망한다.

스마트폰의 온디바이스 AI 혁신

온디바이스 AI는 개별 디바이스에서 이용자 데이터를 수집, 분석, 처리한다는 점에서 이용자의 접점과 경험을 위한 효용 높은 가치를 창출해 낼 수 있다. 이러한 가치를 먼저 간파하고 시장에서는 올해를 기점으로 온디바

그림 2. 삼성 갤럭시 AI / 출처 : 삼성 갤럭시 AI 공식 페이지(2024. 6. 20 기준)

이스 AI 구현을 위해 스마트폰 제조사와 생성형 AI 모델을 지닌 빅테크 기업이 서로 협업하여 AI 기능을 탑재한 스마트폰을 앞다투어 출시하기 시작했다.

올해 1월 삼성전자는 구글과 협력하여 구글의 Vertex AI, Gemini Pro, Imagen 2를 통합한 최초의 AI 스마트폰 '갤럭시 S24'를 출시했다. 삼성은 갤럭시 S24를 출시한 1월부터 3월까지 3개월간 글로벌 시장 점유율 20.8%를 기록했다. 당시 애플의 글로벌 시장 점유율이 17.3%였던 점을 감안할 때, 삼성이 '갤럭시 S24'를 통해 생성형 AI 기능을 먼저 탑재한 것은 신의 한 수였다. 덕분에 올해 1분기 실적에서 애플은 출하량이 5,010만 대로 집계되며 지난해 1분기와 비교할 때 10%가 감소했다.

Smartphone Shipments Worldwide, by Brand, 2022 & 2023

millions, % of total, and % change

	2022	% of total	2023	% of total	% change
Apple	232.2	19%	229.1	20%	-1%
Samsung	257.9	22%	225.5	20%	-13%
Xiaomi	152.7	13%	146.1	13%	-4%
Oppo	113.4	10%	100.7	9%	-11%
Transsion	73.1	6%	92.6	8%	27%
Other	364.1	31%	347.9	30%	-4%
Total	1,193.4	100%	1,141.9	100%	-4%

Note: numbers may not add up to 100% due to rounding
Source: Canalys as cited in press release, Jan 31, 2024

284964 EMARKETER

그림 3. 전 세계 스마트폰 출하량 / 출처 : Emarketer(2024. 4. 15).

그림 4. WWDC 2024 애플 CEO 팀 쿡 / 출처 : WWDC 2024

지난달 6월, 애플의 CEO 팀 쿡은 세계개발자회의 2024(WWDC 2024)에서 'AI는 우리가 계속 시간과 노력을 투자하는 분야'라고 언급하며, 오는 9월에 출시할 신제품 '아이폰16'은 기존 AI 음성 비서인 시리를 혁신해서 새로운 AI 기능을 탑재할 것으로 밝히며 본격적으로 온디바이스 AI에 합류할 것을 예고했다. 이를 팀 쿡은 애플 인텔리전스(Apple Intelligence)라고 소개했다. 이름에서 느껴지듯, 삼성보다 늦은 만큼 애플은 자사 디바이스에 생성형 AI 기능을 탑재하기 위해 심혈을 기울이는 것처럼 보인다.

블룸버그(Bloomberg)에 의하면, 애플은 OpenAI ChatGPT를 iOS 18에 통합할 준비를 하고 있으며, 시리는 생성형 AI를 구현할 핵심 기능으로 재탄생할 것으로 소개했다. 애플의 시리는 14년 전 가장 먼저 스마트폰의 AI 음성 비서로 탑재되며 많은 관심을 받았지만, 구글의 어시스턴트나 아마존의 알렉사에 비해 다소 뒤처진 모습을 보여왔다. 이번 애플의 온디바이스 AI를 계기로 시리가 어떤 형태로 어떤 기능을 구현할지 벌써 많은 궁금증을 자아낸다. 온디바이스 AI로 시리의 혁신은 아이폰의 변화만을 의미하는 것이 아니라, 아이패드, 맥북, 애플워치에 이르기까지 애플의 약 22억 대 이상의 기기에서 AI의 혁신을 의미하기 때문이다. 최근 애플은 다양한 AI 스타트업을 인수하며 자체적인 AI 데이터센터 운용을 위해 하드웨어 개발에도 매진하는 것으로 알려졌다.

온디바이스 AI, 왜 주목할까?

손안으로 생성형 AI를 가져올 수 있는 온디바이스 AI는 개별 이용자에게 집중한 소형 sLLM(small Large

구분	대형 AI 모델	소형 AI 모델
주요 형태	대규모 언어 모델(Large Language Model)	소형 언어 모델[sLM(small Language Model), sLLM(small Large Language Model)]
Parameter 수	1천억 개 이상	10억 ~ 수백억 개 수준
사용 목적	다양한 태스크를 수행하는 범용 AI	특정 태스크에 특화된 목적 기반형 AI
주요 특징	· 대용량의 데이터를 학습하고 다수의 Parameter를 기반으로 운영됨 · 모델의 규모가 크기 때문에 태스트 수행을 위해 필요한 인프라의 수준이 높음 - AI 서비스 운영 과정에서 발생하는 비용이 크며, 결과물 도출 속도가 비교적 오래 걸림	· 대형 AI 모델이 학습한 데이터를 기반으로 특정 태스크에 특화된 형태로 조정함 · 작은 규모의 모델로 운영되어 개별 태스크 수행을 위해 필요한 인프라의 수준이 낮음 - AI 서비스 운영 비용이 비교적 저렴하며, 결과물 도출 시간이 빠른 편임

표 1. 대형 AI 모델과 소형 AI 모델 비교 / 출처 : 삼정KPMG(2024. 6).

Language Model)을 구현한다. sLLM은 10억~수백억 개 수준으로 LLM에 비해 적은 파라미터 수를 보이지만, 목표 지향적인 특화 AI 구현에 매우 탁월하다. 온디바이스 AI가 본격화되며 스마트폰에 공격적으로 탑재되는 상황을 볼 때, 이제 개인마다 생성형 AI를 항시 휴대하고 다니는 셈이 된 것이다. 스마트폰 하나로 생성형 AI를 이용해 사진과 동영상을 생성하거나 편집할 수 있고 정보 검색과 제공에서 또 다른 편의성으로 시간을 크게 절약할 수도 있다. 이러한 가상 시나리오는 스마트폰이 등장했던 초기 시절과 유사해서 묘한 기시감마저 든다.

궁극적으로 온디바이스 AI를 주목하는 가장 큰 이유는 AI 모델을 통한 속도 향상, 외부 시스템 이용에 대한 비용감소, 데이터 보안의 우려 감소라고 할 수 있다. AI 모델을 통한 속도 측면에서 외부 통신 없이 내부 디바이스에서 분석하고 처리하기 때문에 분석 속도가 향상될 것으로 기대한다. 또한 외부 시스템 이용에 대한 비용 측면에서는 외부 클라우드나 데이터센터와 연동한 모델의 업데이트 용도로만 제한적으로 사용해서 클라우드나 데이터센터를 이용했던 비용을 절약할 수 있을 것으로 기대한다. 마지막으로 데이터 보안에서는 이용자의 개별 디바이스라는 폐쇄적인 형태로 데이터 처리가 실행되므로 외부로 데이터를 요청·전송하는 과정에서 발생할 수 있는 보안 문제를 원천적으로 차단할 수 있다. 더 나아가 AI 기반 서비스에서 데이터 수집, 활용, 처리가 중요하고 이를 바탕으로 차별화된 서비스나 마케팅을 실현할 수 있지만 개인정보로 인해 민감 데이터 접근에서 한계가 존재했었다. 그러나 온디바이스 AI는 기존 AI 모델이나 시스템과 달리 폐쇄적으로 개인 디바이스 안에서 처리되는 특성 때문에 민감 데이터 접근의 한계를 어느 정도 해소하게 될 것으로 기대한다.

최근 삼성과 애플처럼 스마트폰 제조사와 여러 빅테크 기업이 협력한 AI는 효율적인 AI 구현을 위해 온디바이스 AI에 주목한다고 볼 수 있다. 기술 수준별로 차이는 있지만, AI 에브리웨어(everywhere)라고 해도 무방할 정도로 AI는 비즈니스를 위한 필수요건이 되었다. 하지만 AI 적용에서 끝이 아니라 기술 고도화 및 업데이트, 인프라와 같은 제반 사항에 필요한 비용들이 계속해서 발생한다. 이런 상황이 계속되면 사업적으로 부담이 될 수밖에 없는데, 이는 작은 기업뿐만 아니라 대형 기업도 마찬가지이다. 사업 측면에서 AI가 동반한 허와 실은 결국, 투자 대비 고효율을 창출할 수 있는 sLLM 소형 AI 모델처럼 안정적으로 AI를 구동하며 비용을 절약할 방법을 고민한 끝에 온디바이스 AI를 주목하기 시작했다. 스마트폰을 중심으로 촉발된 온디바이스 AI가 어떤 AI 패러다임을 만들어갈지 주목해보자.

참고문헌

· 삼정KPMG 경제연구원(2024. 6). 생성형 AI에게 펼쳐진 새로운 무대, 온디바이스 AI. Issue Monitor, 제165호.
· Emarketer(2024. 4. 15). Samsung leads global smartphone-shipment rebound.
· Emarketer(2024. 5. 13). Apple could revolutionize Siri with ChatGPT integration in iOS 18
· WWDC 2024. official page.