LG경영연구원

2024.4.8

생성형 AI 열풍으로 성큼 다가온 온디바이스 AI

박종석 · 유지호

온디바이스 AI는 클라우드의 도움 없이 기기 자체적으로 AI 연산을 수행하는 개념으로, 수많은 기업들이 산업 성장 정체를 해소할 수 있는 기회로 여긴다. 온디바이스 AI는 클라우드 기반 AI가 가지고 있는 프라이버시, 응답 속도 등의 문제를 해결할 수 있는 대안이다. 하지만 명확한 고객 가치를 만들어내려면 ▲온디바이스 AI의 장점을 살릴 수 있는 서비스 기획 ▲작고 강력한 AI 모델 구축을 위한 경량화 기술 ▲기기 내 구동을 위한 강력한 AI 반도체를 확보하는 등의 선결 과제들을 해결해야만 한다.

오픈AI가 챗GPT를 통해 보여준 생성형 AI의 잠재력은 2023년 내내 대중의 큰 관심을 불러일으켰다. 다양한 업체들이 경쟁적으로 유사한 AI 모델을 개발하고 서비스를 출시함에 따라, 이를 뒷받침하는 클라우드 서비스까지 동시에 성장하게 되었다. 하지만, 클라우드 기반 AI가 내재하고 있는 과도한 전력 사용, 프라이버시 등의 문제점들이 대두되었고, 이를 해결하기 위한 솔루션으로서 온디바이스 AI가 부상하게 되었다.

온디바이스 AI는 전자 산업의 가장 뜨거운 트렌드로 자리매김하여 CES의 주요 테마로 선정되었다. CES 2024에서는 인텔, 퀄컴과 같은 반도체 업체부터 삼성과 같은 하드웨어 업체까지 온디바이스 AI가 적용된 제품 및 서비스를 공개하여 많은 주목을 받았다.

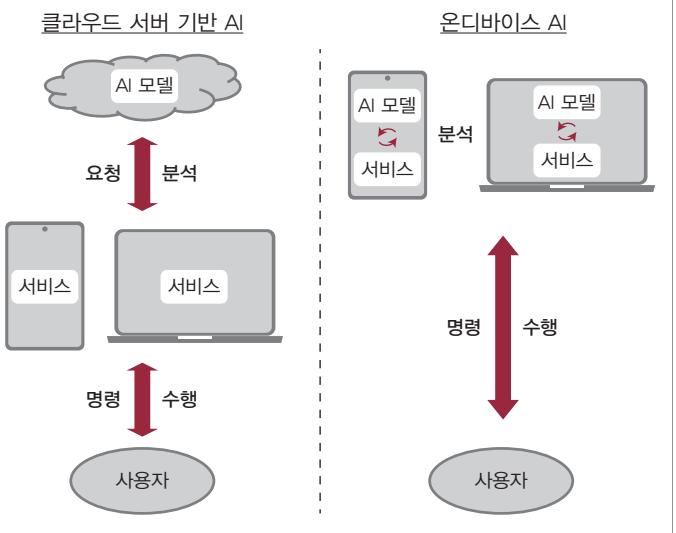
본고에서는 현재 기업 경영의 화두가 되고 있는 온디바이스 AI 트렌드를 면밀하게 이해하고자 온디바이스 AI의 개념 및 장점, 온디바이스 AI 구현을 위한 과제, 그리고 향후 전개 방향에 대해 논의해보고자 한다.

온디바이스 AI의 개념과 장점

온디바이스 AI란 인터넷 연결 없이, 기기 자체에서 AI 모델에 필요한 연산을 수행하는 것을 의미한다. 현재 클라우드 기반으로 제공되고 있는 AI와 대비되는 개념으로, 기존 방식이 갖고 있는 이슈들을 해결, 보완하기 위해 등장하였다.

온디바이스 AI는 클라우드 기반 AI에 비해 작은 용량에 따른 비용 절감과 전력 소모량 감소, 기기 자체 구동을 통한 프라이버시 확보 및 빠른 응답 속도, 그리고 최종적인 발전 방향인 맞춤형 서비스 제공을 장점으로 들 수 있다.

클라우드 기반 AI와의 비교



출처: LG경영연구원

일반적으로 모델의 크기가 커지면 학습 · 구동(추론)에 필요한 GPU 수와 함께 필요한 전력량 또한 증가하기 때문에 비용 역시 증가한다. 온디바이스 AI에서 사용되는 모델은 클라우드 기반 AI 모델 대비 크기가 훨씬 작기 때문에 들어가는 비용과 시간을 대폭 줄일 수 있

다. 심지어 특정 분야에서는 미세 조정과 고품질의 데이터학습을 통해 대형 모델에 버금가는 성능을 보여줄 수 있다.

또한 온디바이스 AI의 경우 클라우드를 거치지 않고 기기 자체적으로 정보를 수집, 연산할 수 있기 때문에 보안 문제가 해결된다. 기기 내부에서 정보를 처리함에 따라 반도체 성능이 충분하다면 클라우드 기반 AI 모델보다 더 빠른 응답 속도를 확보할 수 있다.

마지막으로 온디바이스 AI는 기기에서 개인의 데이터로 추가 학습한, 완전 개인화된 모델이 될 가능성이 높다. 이렇게 되면 기기에 내장된 AI가 이용자 사용 패턴, 개인 정보 등을 학습해 개인 맞춤형 서비스를 제공할 수 있다.

온디바이스 AI 구현을 위한 과제

다양한 가치를 지닌 온디바이스 AI를 제품에 올바르게 구현하기 위해서는 반응속도와 프라이버시를 강조한 서비스, 작고 강력한 모델, 그리고 빠른 구동을 위한 AI 반도체의 3가지 과제가 해결되어야 한다.

(1) 서비스: 기본 기능의 개선을 시작으로 개인화 서비스까지 발전

최근 온디바이스 AI 서비스를 구현한 사례를 살펴보면, 주로 스마트폰, 노트북의 기본 기능인 사진, 음성 인식 강화에 초점이 맞춰져 있다. 대화형 챗봇을 온전히 온디바이스 AI로 구현하기에는 모델 경량화와 반도체 발전이 추가적으로 필요하기 때문에, 이미 충분히 작은 모델로부터 시작하여 작지만, 아주 기본적인 기능들을 개선하는 것부터 시작하고 있다.

구글은 최신 레퍼런스폰인 픽셀8을 발표하며, 온디바이스 AI를 통해 가장 먼저 음성인식과 사진 앱의 기능을 강화했다. 경량화를 통해 데이터센터에서 사용하는 것과 동일한 텍스트 음성 변환 모델을 탑재하여, 말하는 속도, 대화 중간 멈춤 인식 기능을 추가하였고, 이미지 생성 모델을 사용하여 기본 사진앱에서 사용자가 촬영한 사진에 있는 불필요한 요소를 제거하고, 알맞은 배경을 그려 넣는 기능을 강화하였다.

기본 기능 강화부터 시작한 온디바이스 AI 서비스가 고객가치를 제대로 제공하기 위해서는 개인화된 서비스로 진화해야 한다.

윈도우 기반 노트북 키보드에 기본 탑재될 생성형 AI '코파일럿' 서비스 키
출처: 마이크로소프트

이를 위해서는 학습된 모델을 사용하는 단계를 넘어, 사용자 데이터를 안전하게 학습하여 개인화된 모델을 제공해야 한다. 마이크로소프트는 차세대 윈도우 운영체제를 통해 사용자 데이터를 통해 다음 작업을 예측, 필요한 프로그램을 제시하는 등의 사용자 씬(Scene) 구현을 예고했다. 이를테면 문서 이름이 기억나지 않는 경우 "며칠 전 피터가 왓츠앱으로 나에게 보낸 문서를 찾아주세요."하고 검색하면, 실제로 이를 이해하고 찾아줄 수 있다는 것이다. 이를 위해 마이크로소프트는 향후 모든 윈도우 노트북 키보드에 자체 AI 서비스인 '코파일럿' 전용 키를 탑재시킬 것이라고 공개하여 많은 주목을 받고 있다.

(2) 모델: 작지만 강한 모델을 위한 경량화

노트북, 스마트폰에서 AI 모델이 구동되기 위해서는 일단 모델의 크기가 작아야 한다. 클라우드 상에 구동되는 모델은 오픈AI 주도로 점차 사진, 오디오 등 다양한 입력값을 한 번에 처리하기 위해 멀티 모달 모델(GPT-4V 등)화 되고 있으나, 온디바이스 모델은 작고, 특정 기능에 한정되어 있다. 모델이 작아야 기기에 저장이 용이하며, 구동되는데 소요되는 시간이 적어 온디바이스 AI 서비스의 장점을 살릴 수 있다.

모델의 크기는 파라미터의 수, 계산에 필요한 숫자들의 자릿수 등을 포함한 모델 파일의 용량을 의미한다. GPT와 형제 격인 BERT_large 모델은 3억 4,000만 개의 파라미터를 가졌고, 모델 파일의 용량은 1.2GB 정도로 알려져 있다. 클라우드 서비스에 사용되는 대형 언어 모델은 이른바 sLLM(small Large Language Model) 모델이라 하더라도 최소 50억 개가 넘는 파라미터, 수십 GB 수준의 용량이기 때문에 계산 성능 부족, 메모리 부족으로 노트북, 스마트폰에서 활용하기 어렵다.

AI 모델의 온디바이스 AI 적용 현황

용도 구분	파라미터 수	모델 용량(GB) *FP32 기준 단순 치환	온디바이스 적용 현황	
기존 언어 AI	1억 1000만 개	0.45	서비스화 진입 중	오프라인 번역 기능, 라이브 번역 통화, 생성형 기반 사진 편집
	3억 4000만 개	1.2		
이미지 생성형 AI	약 10억 개	6.9		
언어 생성형 AI (LLM)	30억 개	14.7	추가적인 모델 경량화 후 적용 가능	빅테크 중심으로 추진 중
	70억 개	30		
	130억 개	50		
	700억 개	320		

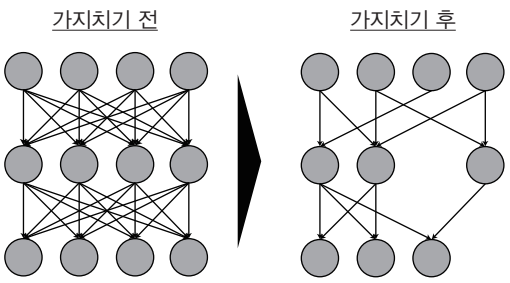
출처: Nvidia 등

모델 경량화의 대표적인 2가지 기법: 가지치기(Pruning)와 양자화(Quantization)

AI 모델의 파라미터는 각각 저장용량을 차지하고 있어, 많을수록 모델이 무거워지고 계산하는데 많은 시간을 소요하게 된다. 대형 모델은 수십억 개의 파라미터와 그를 연결하는 시냅스가 존재하지만, 모두가 중요한 것은 아니다. 가지치기는 모델의 성능에는 큰 영향이 없으나, 용량을 차지하고 있는 불필요한 파라미터와 시냅스를 제거하여 모델의 용량을 줄여주는 효과를 가져온다. 가지치기는 원본의 성능을 유지하면서도 많게는 70%에서 90% 이상까지 파라미터 수를 줄일 수 있어 온디바이스 AI를 위해 사용되는 기법으로 꼽힌다.

두 번째 기법인 양자화는 파라미터와 시냅스의 수를 변화시키지 않고 용량을 줄일 수 있는 방법이다. AI 모델의 결과를 내기 위해 컴퓨터가 계산해야 하는 숫자 형식을 변환하여 컴퓨터가 사용하는 용량을 줄이는 접근법이다. AI 모델에 사용되는 부동소수점 형식(FP32)의 숫자는 1개당 4byte의 저장용량을 차지하지만, 정수 형식(Int8)은 1개당 1byte를 차지한다. 그러므로 숫자 형식을 바꾸는 것만으로도 간단하게 모델 용량을 1/4로 줄일 수 있게 된다. 모델의 용량 감소로 메모리를 효율적으로 사용할 수 있을 뿐 아니라, 컴퓨터가 수행해야 하는 수많은 곱셈 연산에서도 복잡도는 제곱배로 줄어드는 효과까지 있어 가지치기와 함께 사용된다.

가지치기 개념도



출처: LG경영연구원

가지치기의 효과

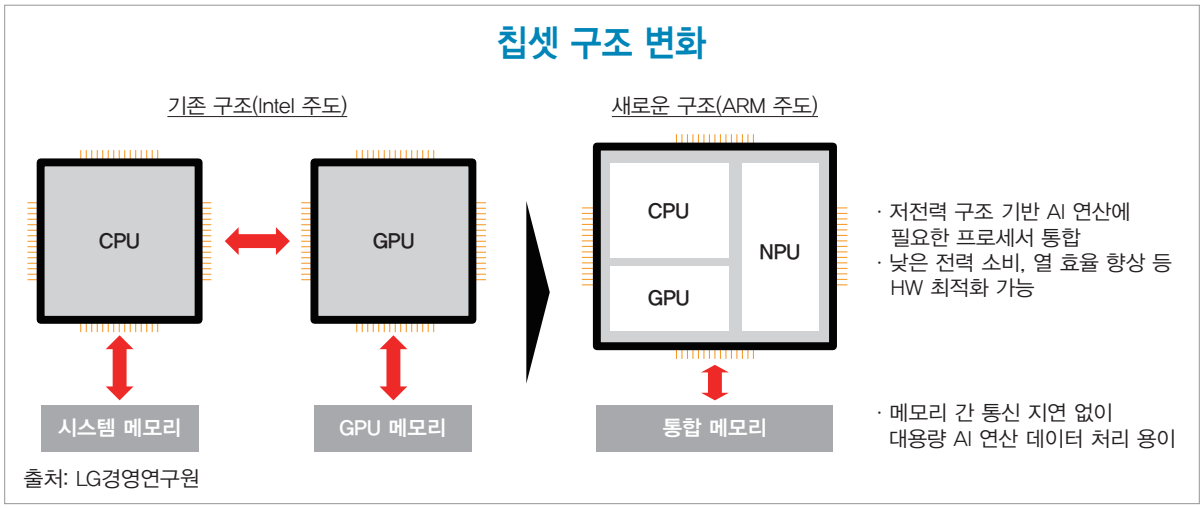
모델 명	파라미터 수		
	원본	가지치기 후	변화율
AlexNet	6100만 개	670만 개	▲89.0%
VGG-16	1억 3800만 개	1003만 개	▲92.7%
GoogleNet	700만 개	200만 개	▲71.4%
ResNet50	2600만 개	747만 개	▲71.3%

출처: MIT Han Lab

그렇기 때문에 기업들은 파라미터 수와 파일 용량을 모두 줄여, 기기에 탑재 가능한 수준까지 모델을 경량화하려 하고 있다. 모델 경량화는 크게 '처음부터 작게 만들기'와 '큰 모델을 축소'하는 두 가지 접근법이 있고, 온디바이스 AI 구현을 위해서는 후자인 '큰 모델을 축소'하는 방식을 주로 사용한다. 경쟁력 있는 오픈소스 대형 모델이 많기 때문에, 온디바이스 AI를 추진하는 퀄컴, 마이크로소프트 등 다양한 기업들은 모델 경량화 툴을 제공하고 생태계를 확보하는 데 집중하고 있다.

(3) AI 반도체: SoC와 통합 메모리 구조로 온디바이스 AI 구현

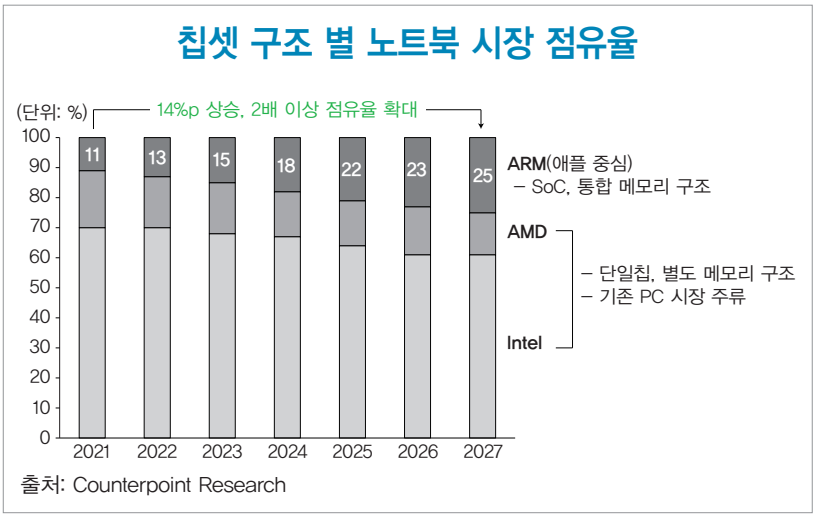
온디바이스 AI를 구현하기 위해서는 소프트웨어 관점의 모델 경량화와 더불어 하드웨어 관점에서 칩셋의 구조 변화가 필수적이다. 클라우드 기반 생성형 AI 서비스를 위한 인프라 시장에서는 엔비디아가 90% 이상의 점유율로 시장을 선점하고 있지만, 온디바이스 AI를 위한 하드웨어 시장은 상황이 약간 다르다. PC 시장을 중심으로 CPU, GPU, NPU(Neural



Process Unit) 등 별도의 칩을 하나의 칩으로 통합하는 SoC(System on Chip)화와 개별 칩이 같은 메모리를 사용하는 통합 메모리 구조[1]가 빠르게 적용되고 있다.

온디바이스 AI 트렌드 이전의 PC에는 위 두 가지 특성을 적용할 필요성이 상대적으로 적었다. PC의 경쟁력은 CPU, GPU 등 개별 칩의 성능으로 정의되었고, 메모리의 사용량이 AI 연산 대비 크지 않아 개별 칩 별 메모리 구조를 사용해도 큰 문제가 없었다.

그러나 기존 방식에서 AI 계산을 하기 위해서는 CPU와 GPU가 각각의 메모리를 통해 데이터를 주고받아야 하고, 이는 데이터 전송의 지연과 성능 저하를 유발한다. 사용되는 메모리의 양이 일반적인 소프트웨어에 비해 상당히 많은 AI 계산의 경우, 데이터 전송 시 병목 현상으로 이어질 수 있어 기존의 구조로는 온디바이스 AI 구현이 어렵다.



이에 애플을 비롯한 다양한 업체에서 AI 연산을 위한 독립적인 NPU를 탑재한 SoC와 통합 메모리 구조를 적용한 PC 제품들을 출시 중이다. 애플은 2020년부터 노트북 CPU를 자체 설계한 M1칩으로 교체하면서, 위

1 기존 별도 메모리 구조에서 AI 계산을 위해서는 데이터가 시스템 메모리–CPU–GPU–GPU 메모리 4단계를 거쳐 처리된다. 각 요소를 지날 때마다 병목 현상으로 인해 처리 속도가 느려지게 되는데, 통합 메모리 구조에서는 통합 메모리–SoC의 두 단계만 거쳐 물리적으로 병목 현상이 적어지고 처리 속도가 빨라지는 효과가 있다.

두 가지 구조를 적용하여 성능 개선에 상당한 효과를 보았다. 직전 세대(인텔 기반) 대비 15배 빠른 머신 러닝 기능을 제공함에 따라 자체적인 온디바이스 AI 기능 등을 실현할 수 있었다. 최근에는 직접 생성형 AI 모델 개발까지 착수하여, 이미 확보된 하드웨어 기반 위에 새로운 서비스를 구현하기 위해 준비 중이다.

온디바이스 AI 경쟁 본격화 전망

CES 2024뿐만 아니라 다양한 빅테크 기업 자체적으로도 온디바이스 AI 서비스 출시를 예고하고 있어, 2024년에는 온디바이스 AI 경쟁이 본격화할 전망이다. 그러나 온디바이스 AI를 무작정 적용하여 성공할 수 있는 것은 아니다. 온디바이스 AI를 통해 분명한 고객 가치를 제공하고, 이를 기기의 차별화 요소로 활용하려면 앞서 언급한 요인들을 충분히 고려해야만 한다.

서비스 측면에서 온디바이스 AI의 장점을 제대로 활용할 수 있는 사용 씬 개발이 중요하다. 최근 화두가 되고 있는 갤럭시 S24의 라이브 번역 통화는 핸드폰 본연의 기능이자, 프라이버시가 중요한 서비스에 온디바이스 AI를 적용했다는 점에서 좋은 출발점이라고 볼 수 있다. 이와 같이 온디바이스 AI 서비스는 기본 기능의 개선/발전에 적용되는 것부터 시작하여 최종적으로는 완전히 개인화된 서비스를 제공하는 방향으로 진화할 전망이다.

두 번째, 사용자가 언제 어디서나 손 안에서 필요로 하는 AI 기능을 구현하기 위한 경량화 모델 개발 흐름은 당분간 지속될 것으로 보인다. 빅테크를 중심으로 더 크고, 똑똑한 AI 모델을 개발하면, 이를 작게 만들어 온디바이스 AI화하는 개발 방향이 동시에 발생하기 때문에, 향후 온디바이스 AI 모델의 성능은 지속적으로 상승하고 더 많은 기능을 수행할 수 있을 것이다.

마지막으로 AI 반도체 관점에서 향후 온디바이스 AI 구현을 위한 칩셋 경쟁이 확대될 것으로 보인다. 기존 시장 지배자인 인텔 역시 SoC 구조를 적용한 신제품을 출시했고, 퀄컴은 스마트폰 칩셋의 노하우를 PC에 이식하기 위해 진입 중이다. 또한 서버 GPU 기반의 새로운 PC용 SoC를 개발 중인 엔비디아까지 시장에 가세했기 때문에 향후 경쟁은 더욱 치열해질 것이다. LG경영연구원