TinyML 기술 동향과 산업 응용

□ **박세진, 하유빈** / 감바랩스

요 약

TinyML(Tiny Machine Learning)은 초소형, 초저전력 마이크로컨트롤러(MCU)에서 머신러닝을 구현하는 기술로, 클라우드 연결 없이도 엣지 단말에서 인공지능 추론을 가능하게 하는 새로운 패러다임이다. Edge Impulse사는 TinyML 분야에서 하드웨어 개발자와 인공지능 모델 개발자를 연결하는 대표적인 플랫폼 기업으로 2025년 3월 퀄컴이 전격적으로 인수를 하였다. 이번 인수는 TinyML 생태계의 중요한 전환점으로, 단순히 글로벌 기업의 사업 영역 확장을 넘어 인공지능 산업에서 온디바이스 AI가 산업의 주류에 포함되기 시작했음을 보여준다. 이런 움직임은 글로벌 MCU 제조사들도 이미 보여주고 있었는데 대표적으로 ST마이크로일렉트로닉스의 Neural-ART 가속기 탑재 STM32N6, NXP의 eIQ Neutron NPU 통합 MCX 시리즈, TI의 AI 가속기 내장 C2000 시리즈 등 TinyML 최적화 제품을 출시하였다. 온디바이스 AI가 보편화됨으로써 스마트홈, 스마트시티, 산업 현장에서 초저전력(수 mW)으로 실시간 응답이 가능한 지능형 시스템이 구현되고 있으며, 특히 음향 센서 기반 시계열 TinyML은 음성 인식, 장비 이상 탐지, 환경 모니터링 등에서 중요한 역할을 담당하고 있다. 본고에서는 TinyML의 최신 하드웨어/소프트웨어 동향, 주요 응용 사례, 국내 기업들의 대응 현황을 살펴보고, 미래에 TinyML이 탄소배출 감축, 프라이버시 보호, 디지털 격차 해소에 기여할 수 있는 가능성을 탐색하고자 한다.

Ⅰ. 서 론

2025년 3월 10일, 퀄컴(Qualcomm)은 Edge AI 플랫폼 기업인 Edge Impulse를 인수하기로 발표하며 MCU 기반 TinyML 분야에 대한 큰 관심을 드러냈다. Edge Impulse는 소형 기기에서 음성 인식 등 머신러닝을 온디바이스로 실행하는 기술을 개척해 온 미국의 선도 기업으로, 퀄컴은 이번 인수를 통해 자사 IoT 개발 생태계를 강화하고 엣지 디바이스의 AI 성능을 한층 높일 계획을 밝혔다[1]. 한편 Science지 2025년 2월 기사에서는, 불안정한 전력 공급과 높은 비용으로 인해 이른바 글로벌 사우스와 같이 대규모 AI 활용이 어려운 지역에서 TinyML이 대안으로 부상하고 있다고 보도했다. 우표 크기(2×1.8cm)에 불과한 14달러짜리 칩으로 식물 질병 감지, 심장 리듬 이상

<그림 1> TinyML example (Science 2025.02)

탐지, 야생동물 추적, 환경 오염 모니터링까지 수행한 사례를 들어, 초저전력·저비용 AI인 TinyML 기술이 농업, 보건, 환경 보호 등 분야에서 자원 제약을 극복하는 혁신적 수단이 되고 있음을 강조했다[2]. 거대한 연산 자원이나 클라우드 연결 없이도 소형 장치 자체에서 AI 추론을 수행하는 TinyML은 민첩성과 접근성을 바탕으로 AI 활용의 폭을 넓히고 있으며, 이번 퀄컴-Edge Impulse 합병은 이러한 흐름 속에서 MCU 기반 TinyML 기술의 전략적 가치를 보여준다. 본고에서는 이 합병이 가져올 TinyML 생태계의 변화를 짚어보고, 최근 5년간 주요 MCU 제조사들의 TinyML 친화형 칩 개발 동향과 시장 응용 사례, 그리고 시계열 음향 센서 기반 TinyML 기술의 중요성을 살펴본다.

Ⅱ. TinyML을 위한 HW 및 SW 동향

1. MCU 제조사의 TinyML을 위한 칩 개발 동향

소형 MCU에서 머신러닝을 구현하는 TinyML 수요가 높아지면서, 주요 MCU 업체들은 지난 몇 년간 전용 하드웨어 가속기와 소프트웨어 도구로 TinyML 성능을 끌어올려 왔다. ST마이크로일렉트로닉스는 2018년 STM32Cube.AI같은 툴로 자사 MCU에서 신경망 추론을 지원하기 시작했고, 2023년에는 신경망 처리장치(NPU)를 내장한 MCU 제품군 STM32N6를 공개했다. STM32N6는 ST의 기존 최고급 MCU 대비 머신러닝 성능을 600배 높인 Neural-ART 가속기를 통해, 소형 임베디드 시스템에서는 어려웠던 컴퓨터 비전, 오디오 처리, 음향 분석 등의 고성능 기능을 가능케 했다[3].

NXP는 크로스오버 MCU인 i.MX RT 시리즈와 eIQ 소프트웨어로 엣지 ML을 지원해 왔으며, 2022년에 발표한 신규 MCU 플랫폼 MCX 시리즈를 통해 TinyML 성능을 대폭 향상시켰다. 특히 MCX N Advanced 계열에는 자사 설계의 NPU인 eIQ Neutron이 통합되어 Cortex-M33 코어 단독 실행 대비 30배 이상 빠른 ML 추론 성능을 발휘한다[4]. 이는 150~250MHz급 MCU에서도 DSP 코어와 NPU의 병행 동작으로 높은 연산량의 AI 알고리즘을 실시간 처리할 수 있음을 의미한다. 텍사스 인스트루먼츠(TI) 역시 2024년 자사의 32비트 MCU에 AI 가속기를 도입하며, MCU 분야 최초로 신경망 연산 전용 코어를 내장한 C2000™ TMS320F28P55x 시리즈를 선보였다. TI는 이 제품을 실시간 제어 MCU로 분류하며, 전력 변환 장치의 아크_fault 검출이나 모터 베어링 이상 탐지같이 산업 현장의 고장 징후를 99% 정확도로 실시간 판별하는 성능을 강조했다[3].

한편, MCU급 디바이스에서의 AI 성능 한계를 극복하기 위해 Arm은 코어 아키텍처와 IP 차원에서 TinyML을 지원하고 있다. 2020년에 발표된 Cortex-M55 코어(Armv8.1-M)는 Helium 벡터 연산 기술을 활용해 이전 세대 Cortex-M보다 최대 15배의 ML 성능 향상을 이루었고, 같은 시기에 등장한 마이크로 NPU Ethos-U55와 짝을 이뤄 사용할 경우, 기존 Cortex-M 단독 대비 최대 480배에 달하는 ML 워크로드 가속이 가능하다고 보고되었다. 실제 예로, 소형 MCU에서 음성비서를 실행하는 엔드투엔드 성능이 Cortex-M7만으로 구현했을 때보다 50배 이상 속도 향상되고 에너지 효율은 25배 높아진 것으로 나타났다. 2022년에는 Arm이 한층 향상된 Cortex-M85 코어와 Ethos-U65/U85 NPU를 공개하여 TinyML 성능 지평을 넓혔다. Cortex-M85는 기존 Cortex-M보다 강화된 보안과 최고 수준의 처리능력을 갖춘 MCU 코어로, 동작 주파수 내에서 DSP/ML 연산을 극대화하도록 설계되었다[5]. 여기에 연결할 수 있는 Ethos-U85는 이전 세대 대비 4배 향상된 피크 성능과 20% 향상된 전력 효율을 달성한 Arm 최신 NPU로서, 최대 2,048 MAC 연산 유닛을 구성해 INT8 기준 **초당 4조 연산(TOPS)**의 추론 성능까지 확장 가능하다[10]. 이처럼 CPU 코어와 NPU의 발전으로 MCU에서도 수 TOPS 규모의 AI 연산이 현실화되면서, 엣지 디바이스의 온디바이스 AI 구현 범위가 크게 확대되고 있다.

이러한 NPU 내장 MCU들은 추론 연산을 메인 코어와 분리된 전용 하드웨어에서 수행함으로써, MCU에도 AI 병렬처리 능력을 부여하고 전체 시스템의 지능화를 뒷받침하고 있다. 그 결과 향후에는 기존 마이크로프로세서(MPU)가 담당하던 작업도 MCU가 수행하여, 엣지 단말의 비용 및 전력 효율을 크게 높일 수 있을 것으로 기대된다.

2. TinyML 응용 시장과 요구 기술

TinyML에 대한 투자는 스마트홈, 스마트시티, 산업 등 다양한 응용 분야의 수요에 의해 가속화되고 있다. 스마트홈 분야에서는 스마트 스피커, 가전제품 등이 인터넷에 연결되지 않은 상태에서도 사용자 음성 명령을 인식하거나 손동작을 감지하는 등 지능형 동작을 수행하길 요구한다. 이러한 기능을 구현하려면 기기가 항시 대기(always-on) 상태에서 초저전력으로 센서 데이터를 모니터링하다가 즉각 응답할 수 있어야 하며, 개인 음성/영상 데이터의 프라이버시 보호를 위해 가능한 한 로컬에서 처리하는 기술이 필요하다. 스마트시티와 환경 모니터링 분야에서는 도시 곳곳의 센서 노드가 교통량, 소음, 대기 오염 등을 실시간 분석하여 이상 상황을 감지한다. 이때 수많은 센서들이 생성하는 방대한 데이터를 모두 클라우드로 전송하면 비효율적이므로, 현장에서 필요한 정보를 추출하는 엣지 분석이 필수적이다. ARM 임원도 “스마트시티의 카메라가 매일 기가바이트 단위 데이터를 쏟아내는 상황에서, IoT 단말 수십억 개의 엔드포인트 데이터를 모두 클라우드로 보내는 것은 불가능하며, AI 추론이 직접 엔드포인트에서 이루어져야 한다”라고 강조한 바 있다[5].

산업 현장에서는 장비에 부착된 MCU가 진동, 전류, 음향 등의 센서 데이터를 학습하여 설비 이상이나 품질 불량을 조기에 예측 진단하는 이상감지 기술이 중요한 이슈다. 예를 들어 태양광 발전 설비의 아크 Fault 발생이나 모터의 베어링 마모를 TinyML로 감지하면, 대형 사고를 미연에 방지하고 유지보수 비용을 절감할 수 있다.

이러한 응용 시장에서는 공통적으로 초저전력 연산, 실시간 응답성, 높은 신뢰도가 요구된다. TinyML 시스템은 수 mW 이내 전력으로 동작하면서도 센서 데이터에 대한 로컬 추론을 수 ms 내 수행해야 하며, 잡음이나 환경 변화 속에서도 일정 수준 이상의 정확도를 유지해야 한다[6].

이를 뒷받침하기 위해 MCU 설계에서는 전력 효율이 뛰어난 DSP 명령어 및 벡터 연산기(예: Arm Helium)와 하드웨어 가속기를 탑재하는 추세이고, 소프트웨어 측면에서는 임베디드 기기에 맞게 압축/최적화된 모델, AutoML을 통한 모델 경량화, OTA를 통한 모델 업데이트 등이 적극 도입되고 있다. 또한 보안 측면에서, 엣지 AI 디바이

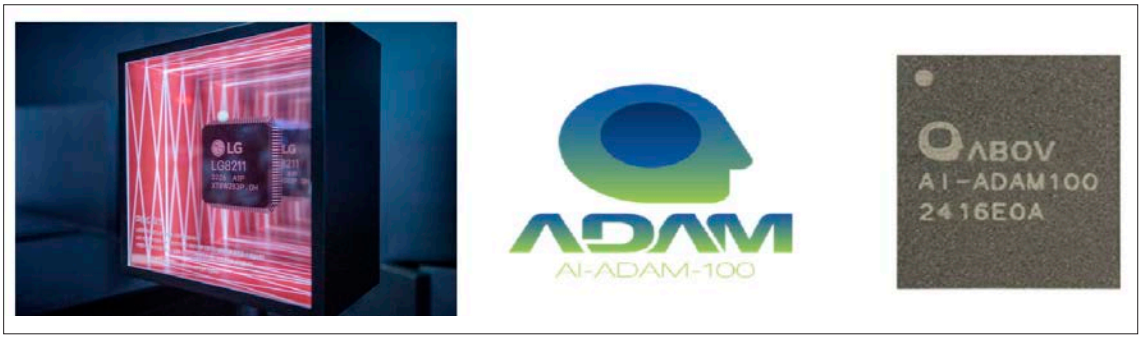
스가 사이버 공격이나 모델 탈취로부터 안전하도록 Arm TrustZone 기반 보안 영역, NXP EdgeLock 보안 서브시스템 등 온칩 보안기술도 함께 중요해지고 있다.

3. 음향 센서와 시계열 TinyML의 핵심 역할

다양한 센서 중에서도 마이크로폰을 비롯한 음향 센서는 TinyML 활용에서 특히 두각을 나타내는 분야이다. 음성은 인간-기계 인터랙션의 핵심 수단이고, 다른 형태의 데이터(예: 영상)에 비해 대역폭이 낮아 소형 MCU로도 처리하기 상대적으로 수월하다. TinyML의 대표적 성공 사례로 꼽히는 키워드 스팟팅(keyword spotting) 기반 음성 인식 기술은, 클라우드가 아닌 기기 내부의 MCU에서 “Wake Word”를 실시간 인식함으로써 스마트 디바이스의 상시대기 전력 소모와 응답 지연을 크게 줄였다. 실제로 Edge Impulse와 같은 플랫폼의 초창기 주요 적용처도 IoT 기기의 온디바이스 음성 인식 기능이었으며, 오늘날 인터넷 연결 없이도 음성 명령을 인식하여 가전제품을 비롯한 전자제품을 제어하는 데 활용하고자 한다. 나아가, 산업 및 공공 안전 분야에서는 소리를 통한 이상 탐지가 중요한 응용으로 떠올랐다. 예를 들어 공장의 장비 소음 패턴을 분석해 이상 진동이나 충돌음을 감지하거나, 스마트시티에서 총성이나 사이렌 소리를 식별하는 등의 작업은 지속적인 시계열 오디오 스트림을 실시간 처리해야 하므로 TinyML이 적합하다. 이러한 엣지 오디오 처리의 요구 조건으로는, 인터넷 업로드 없이 현장에서 빠르게 의사결정을 내릴 수 있는 저지연 처리, 배터리로 수년간 구동 가능한 지속적 초저전력, 그리고 한정된 MCU 메모리와 연산 내에서 동작하는 경량 알고리즘 등이 꼽힌다. 실제 구현 측면에서는 마이크로파워 DSP, 저전력 ADC와 함께 뉴럴 네트워크 기반 오디오 이벤트 분류기가 MCU에 탑재되어 이러한 요건을 충족시키고 있다. Science지 보도에 등장한 14달러 MCU 기반 기기는 심장 박동 신호를 분석해 부정맥을 탐지하거나 가축의 울음소리로 상태를 모니터링하는 등, 소리와 진동으로 대표되는 시계열 데이터를 TinyML로 해석해 낸 사례이다. 이처럼 음향 기반 TinyML 기술은 의료, 농축산, 보안 등 다양한 분야에서 상시 모니터링과 이상징후 감지를 가능케 하며, 온디바이스 AI의 유용성을 극대화하는 핵심 요소로 자리매김하고 있다.

4. 국내 기술 동향

한편 국내 기업들의 TinyML 대응도 시작되었다. LG전자는 CTO 부문 SoC센터 주도로 AI 가전용 SoC와 차량용 MCU 등을 자체 개발해 왔다. 가전용 MCU는 자사 가전제품에 적용이 이미 시작되었으며, 최근 자체 개발한 차량용 MCU로 국제 자동차 기능안전 표준 ISO 26262 인증을 획득함으로써 자율주행차 인포테인먼트 등에 필요한 고신



<그림 2> LG AI MCU와 어보브반도체 ADAM100

뢰성 MCU 기술력을 입증했다[7]. MCU 전문 팹리스인 어보브반도체는 2024년 온디바이스 AI MCU ‘ADAM-100’ 엔지니어링 샘플을 공개하며 초저전력 음성 인식 MCU 시장에 뛰어들었다. ADAM-100은 음성 등의 AI 추론 기능을 지원하며, 2025년 상반기 양산을 목표로 삼성전자 등 가전사에 공급을 준비 중이다[8]. TinyML 전문 스타트업 감바랩스(Gamba Labs)도 초경량 온디바이스 음성 및 화자인식 모듈을 개발하여 다양한 전자제품에 적용을 앞두고 있으며, 부산시가 추진하는 AI 허브도시 사업에서 경량형 AI 기술의 대표 사례로 선정되어 지역 지원을 받고 있다[9].

Ⅲ. 결 론

이처럼 TinyML은 학계·산업계 전반에서 차세대 임베디드 지능의 핵심 기술로 주목받고 있다. 클라우드 중심의 AI가 지닌 지연, 보안, 비용의 한계를 보완하면서 센서가 있는 곳에 직접 지능을 이식하는 TinyML은 사물인터넷 시대 온디바이스 AI의 필수 기반 기술로 확고히 자리잡을 전망이다. TinyML 기술은 이제 단순한 실험 단계를 넘어 온디바이스 AI의 필수 기반으로 자리잡을 전망이다. 마이크로컨트롤러 제조사들은 앞다투어 AI 친화적 하드웨어와 지원툴을 선보이고 있으며, Edge Impulse 인수에 나선 Qualcomm처럼 모바일/IoT 반도체 강자들의 가세도 이어지고 있다. 향후 5년 내에는 센서부터 프로세서, 통신 모듈까지 AI 최적화가 내재된 AIoT 시스템온칩(SoC)들이 등장하여, 전력 1mW 미만에서 동작하는 초소형 자율지능이 현실화될 것으로 기대된다.

TinyML은 모든 산업 분야와 융합될 것으로 전망된다. 제조업에서는 이상 탐지와 로봇 제어에, 농업에서는 스마트 팜 센싱에, 도시에서는 교통·환경 관리에 TinyML이 스며들어 현장의 데이터를 실시간 이해하고 대응하는 능력을 부여할 것이다. 클라우드 AI와의 조화도 중요해져, 중앙 AI는 복잡한 판단과 학습을 담당하고 말단 디바이스의 TinyML은 즉각적인 반응과 데이터 여과를 맡는 계층형 지능 구조가 보편화될 것이다. 이는 5G/6G 통신과 맞물려 지연 없는 서비스, 효율적인 네트워크 운용을 가능케 한다.

TinyML의 확산은 보다 인간과 환경 친화적인 기술 패러다임을 의미한다. 전력 소모를 줄여 탄소배출을 감축하고, 개인정보를 로컬에서 처리해 프라이버시를 지키며, 저가 디바이스로도 AI 혜택을 누릴 수 있어 디지털 격차 해소에도 기여할 것으로 기대된다. 한국의 산업계도 이에 대비하여 다양한 하드웨어, 소프트웨어 및 표준화 등에서 활발한 활동이 진행되기를 기대해 본다.

참 고 문 헌

[1] J. Martins, “Qualcomm acquires on-device machine learning pioneer Edge Impulse,” *audioXpress*, Mar. 2025.
[2] H. A. Phillips, “What’s tinyML? The Global South’s alternative to power-hungry, pricey AI,” *Science*, vol. 387, no. 6736, Feb. 2025.
[3] M. Ahmad, “2024: The year when MCUs became AI-enabled,” *EDN Network*, Mar. 2025.
[4] J.-L. Aufranc, “NXP unveils MCX MCU family with 30× faster machine learning performance,” *CNX Software*, Jun. 14, 2022.
[5] “Arm cores designed for TinyML devices,” *Embedded.com (EE Times)*, Jan. 2020.
[6] S.-C. Liu, “System Requirements for TinyML Edge Audio,” tinyML Summit Presentation, 2020.
[7] 서재창, “LG전자, 차량용 MCU 자체 개발로 국제 기능안전 인증 받아,” *HelloT 뉴스*, 18 Mar. 2025.
[8] 배도혁, “어보브반도체, 상반기 내 온디바이스 AI MCU 양산 착수… 삼성 LG 등 납품 기대,” *파이낸스스코프*, 8 Jan. 2025.
[9] 민건태, “부산시, 스마트시티를 ‘거대 AI 실험실’로,” *한국경제*, 13 Mar. 2025.
[10] G. Halfacree, “Arm targets the AIoT with high-performance Ethos-U85 NPU and Corstone-320 platform,” *Hackster.io News*, Apr. 2023.

저 자 소 개

박 세 진

- 1998년 : 부산대학교 공과대학 컴퓨터공학과 학사
- 2000년 : 부산대학교 공과대학 컴퓨터공학과 석사
- 2005년 : 부산대학교 공과대학 컴퓨터공학과 박사
- 2005년 ~ 2020년 : 삼성전자 DMC연구소, ST-Ericsson Korea, Sentons Korea 등
- 2021년 ~ 현재 : 감바랩스 대표이사
- 2023년 ~ 현재 : 부산대학교 정보컴퓨터공학부 겸임교수
- 주관심분야 : Embedded AI, TinyML, Recognition, Detection

하 유 빈

- 2013년 : 부산대학교 공과대학 컴퓨터공학과 공학사
- 2021년 : 부산대학교 정보융합공학과 컴퓨터공학전공 공학박사
- 2021년 ~ 2022년 : 부산대학교 BK21 박사후과정
- 2022년 ~ 2023년 : 부산대학교 연구교수
- 2023년 ~ 현재 : 감바랩스 CTO
- 주관심분야 : Embedded System, TinyML, Edge-AI