전자통신동향분석, 41권 3호 2026, 66-74

온디바이스 환경에서의 경량형 AI 프레임워크 기술 동향과 표준화 이슈

Technical Trends and Standardization Issues of Lightweight AI Frameworks in On-device Environments

최영환 (Y.H. Choi, yhc@etri.re.kr) 전략표준연구실 표준전문위원/책임연구원

ABSTRACT

As the amount of AI data generated from various devices such as videos and sensors rapidly increases, the paradigm of AI processing is shifting from centralized clouds to on-device AI. On-device AI ensures real-time responsiveness, personalization, and data security by performing direct inference directly where the data are generated. However, independent developments by manufacturers have led to significant technical fragmentation, hindering interoperability and reuse. This study defines a lightweight AI framework as an integrated execution structure covering the entire lifecycle, from model development and optimization to deployment and management. We analyze the latest global hardware and software trends, identify key standardization issues arising from fragmentation, and explore domestic and international activities the ITU-T, ISO/IEC, and TTA. Finally, we propose a strategic roadmap for standardization centered on infrastructure, interworking, and performance evaluation to enhance national AI competitiveness.

KEYWORDS Lightweight Framework, On-device AI, Standardization

I. 서론

인공지능(AI: Artificial Intelligence) 기술의 모든 출발점은 데이터라고 할 수 있다. 오늘날 사물인터넷(IoT) 등 ICT 기술과 융합되면서 영상, 센서, 사용자 상호작용 등과 같은 방대한 AI 데이터는 우리 주변의 다양한 스마트 디바이스에서 실시간으로 끊임없이 생성되고 있다. 이에 따라 고성능 처리 시스템이 필요한 기존의 인공지능 서비스는 데이터를 중앙집중형 클라우드 인프라로 전송하여 학습과 추론을 수행하는 구조를 주로 취해 왔으나, 데이터 생성량의 폭발적 증가에 따른 네트워크 연결성 의존도 심화와 인프라 유지 비용의 확대, 그리고 데이터 전송 과정에서의 지연시간 발생 등 물리적인 한계가 점

* DOI: https://doi.org/10.22648/ETRI.2026.J.410306
* This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) [No. 2018-0-00278, Development of Big Data Edge Analytics SW Technology for Load Balancing and Active Timely Response].

차 드러나고 있다.

이러한 한계를 극복하기 위해 인공지능 처리 영역은 클라우드를 넘어 데이터가 생성되는 위치인 디바이스 내부에서 직접 추론을 수행하는 '온디바이스 AI(On-device AI)'로 연구[1]가 확장되고 있다. 온디바이스 AI는 데이터 처리를 디바이스 내부로 이전함으로써 네트워크 지연 없이 즉각적인 응답을 제공하는 실시간성을 확보하며, 사용자의 이용 패턴을 기반으로 한 개인화된 서비스를 제공하고, 무엇보다 민감한 데이터를 외부로 전송하지 않아 데이터 보안 및 프라이버시 보호를 동시에 실현할 수 있는 차세대 AI 핵심 기술이다.

보이지 않던 인공지능 서비스 기술이 사용자의 눈앞에서 직접 동작하게 됨에 따라 온디바이스 AI는 서비스 경쟁력을 좌우하는 필수적인 ICT 기반 기술로 부상하고 있으나, 온디바이스 환경은 연산 자원과 메모리, 전력 소모 측면에서 엄격한 제약이 존재한다. 따라서 자원 제약 환경에 최적화된 모델 경량화와 실행 최적화가 필수이며, 이를 모델 개발부터 관리까지 전주기적으로 지원하는 '경량형 온디바이스 AI 프레임워크'의 중요성이 더 강조되고 있다.

II. 경량형 온디바이스 AI 기술 동향

1. 핵심 개념 및 기술 요소

그림 1과 같이, 경량형 온디바이스 AI 프레임워크는 단순히 디바이스 내에서 AI 모델을 실행하는 추론 엔진의 역할을 넘어, 자원 제약 환경에서 모델의 변환, 최적화, 배포, 실행 및 관리에 이르는 전 과정을 지원하는 통합 실행 구조로 정의된다[1]. 특히 온디바이스 환경은 연산 자원, 메모리, 전력 측면에서 제약이 크기 때문에, 이러한 제약을 고려한 경량화 및 실행 최적화 기술이 필수로 요구된다.

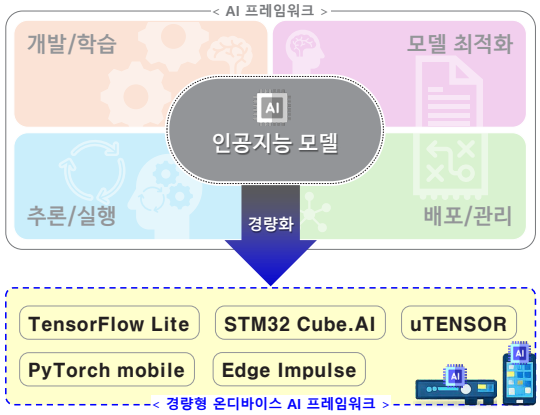

그림 1 "온디바이스 AI" 기술 개념

대표적인 핵심 기술로는 모델의 가중치 정밀도를 낮춰 연산량을 줄이는 양자화(Quantization), 중요도가 낮은 연결을 제거하여 모델 크기를 축소하는 가지치기(Pruning), 그리고 대규모 모델의 지식을 소형 모델로 이전하는 지식 증류(Knowledge Distillation) 기술이 있다. 이러한 기술들은 제한된 하드웨어 환경에서도 실시간 추론 성능을 확보하면서 모델 정확도를 유지하는 데 중요한 역할을 한다.

그러나 이러한 경량화 기술은 프레임워크마다 상이한 방식으로 구현되고 있으며, 동일한 모델이라 하더라도 적용 방식과 최적화 결과가 달라지는 문제가 발생한다. 이는 프레임워크 간 상호운용성을 저해하고, 모델 재사용성을 제한하는 구조적 한계로 작용하고 있다.

2. 국외 기술 동향

표 1과 같이, 국외에서는 온디바이스 AI 확산에 대응하여 글로벌 빅테크 기업을 중심으로 다양한 경량형 프레임워크가 개발 · 확산되고 있다.

Google은 TensorFlow Lite[2]를 기반으로 모바일 및 임베디드 환경에서 동작하는 경량 추론 환경을

제공하고 있으며, Android NNAPI와 연계하여 디바이스 내 하드웨어 가속기(NPU, GPU)를 활용한 최적화를 지원하고 있다. 또한, 최근에는 생성형 AI 모델의 온디바이스 실행을 위해 Gemma 기반 경량 모델과 Edge TPU를 결합한 형태로 확장하고 있다.

Meta는 PyTorch Mobile[3] 및 ExecuTorch를 통해 온디바이스 환경에서의 추론 최적화를 추진하고 있으며, 특히 다양한 디바이스에서 일관된 실행을 지원하기 위한 런타임 경량화와 모델 최적화 기술에 집중하고 있다.

Microsoft는 ONNX Runtime[4]을 중심으로 다양한 프레임워크 간 상호운용성을 지원하는 실행 환경을 제공하고 있으며, ONNX 모델 포맷을 통해 프레임워크 독립적인 배포 구조를 구축하고 있다. 이는 이기종 디바이스 환경에서의 모델 호환성 확보 측면에서 중요한 역할을 하고 있다.

Apple은 Core ML 프레임워크를 통해 iOS 및 macOS 환경에서 온디바이스 AI 실행을 지원하며, Neural Engine 기반의 하드웨어 가속과 긴밀하게 통합된 구조를 제공하고 있다. Qualcomm 또한 Snapdragon 플랫폼 내 AI Engine을 통해 모바일 디바이스에서의 온디바이스 AI 실행을 최적화하고 있다.

이러한 국외 기술은 플랫폼 기업 중심으로 프레임워크와 하드웨어가 통합된 형태로 발전하고 있으나, 기업별로 독자적인 구조와 실행 환경을 채택하고 있어 프레임워크 간 상호운용성 확보에는 한계가 존재한다.

3. 국내 기술 동향

국내에서도 온디바이스 AI 기술 확산에 대응하여 주요 ICT 기업과 연구기관을 중심으로 관련 기술 개발이 활발히 진행되고 있다.

삼성전자는 스마트폰 및 가전제품에 온디바이스 AI 기능을 적용하기 위해 자체 NPU 기반의 AI 처리 구조를 개발하고 있으며, Exynos 플랫폼과 연계된 AI 실행 환경을 구축하고 있다. 또한, 삼성은 생성형 AI 모델을 온디바이스 환경에 적용하기 위한 경량화 기술 개발을 병행하고 있다.

LG전자는 스마트가전 및 로봇 분야를 중심으로 온디바이스 AI 기술을 적용하고 있으며, 디바이스 내부에서의 실시간 인식 및 제어 기능을 강화하는

표 1 주요 온디바이스 AI 프레임워크 비교

항목	TensorFlow Lite(Google)	PyTorch Mobile / ExecuTorch(Meta)	ONNX Runtime(Microsoft)
구조	경량 런타임 기반 추론 구조, NNAPI 및 Edge TPU 등 하드웨어 가속 연계	PyTorch 기반 모델을 모바일 환경에 최적화, ExecuTorch로 경량 런타임 확장	중간 표현(IR) 기반 실행 구조, 다양한 프레임워크 모델 지원
모델 포맷	TensorFlow Lite FlatBuffer(.tflite)	TorchScript(.pt)	ONNX(.onnx)
특징	모바일 및 임베디드 환경 최적화, 다양한 디바이스 지원	개발 유연성 및 빠른 모델 배포, 연구-서비스 연계 용이	프레임워크 간 호환성 지원, 범용 실행 환경 제공
장점	하드웨어 가속 연계 용이, 안정적 생태계	개발 편의성 높음, 동적 모델 지원	높은 이식성, 다양한 플랫폼 지원
한계	TensorFlow 생태계 의존성, 타 프레임워크와 호환 제한	실행 최적화 및 경량화 측면에서 제한 존재	완전한 호환성 미흡, 변환 과정에서 성능 저하 가능
상호운용성	제한적(자체 생태계 중심)	제한적(PyTorch 중심)	상대적으로 우수(중립 포맷 기반)
적용 분야	모바일, IoT 디바이스, Edge AI	모바일, 연구/서비스 연계	다양한 디바이스 및 클라우드-엣지 연계

방향으로 경량 AI 기술을 활용하고 있다.

네이버는 HyperCLOVA[5] 기반의 소형화 모델을 개발하여 온디바이스 및 엣지 환경에서의 활용 가능성을 확대하고 있으며, 클라우드-엣지-디바이스를 연계하는 구조를 통해 AI 서비스 확장성을 확보하고 있다.

ETRI를 비롯한 국내 연구기관에서는 온디바이스 AI 프레임워크 구조, 경량화 기술, AI 반도체 연계 기술 등에 대한 연구를 수행하고 있으며, 특히 표준 기반 구조 설계를 위한 기술 개발이 진행 중이다.

그러나 국내 기술 역시 개별 기업 및 플랫폼 중심으로 발전하고 있어, 프레임워크 구조와 실행 환경이 상이하게 구성되는 경향이 있으며, 글로벌 프레임워크 의존도가 높은 상황에서 독자적인 기술 생태계 구축과 국제 표준 연계 전략이 중요한 과제로 남아 있다.

III. 기술 파편화에 따른 표준화 이슈 및 한계

Ⅱ장에서 소개한 기술적 차이는 단순한 구현상의 다양성을 넘어 프레임워크 수준에서의 구조적 파편화로 심화하고 있으며, 이는 온디바이스 AI 확산의 핵심 제약 요인으로 작용하고 있다.

1. 기술 파편화 현상

경량형 온디바이스 AI 프레임워크는 글로벌 ICT 기업을 중심으로 빠르게 발전하고 있으며, 다양한 디바이스 환경에서의 적용이 확대되고 있다. 그러나 이러한 발전은 기업별 플랫폼과 생태계를 중심으로 이루어지고 있어, 프레임워크 구조, 모델 표현 방식, 실행 환경 등이 상이하게 구현되는 기술 파편화 현상이 심화하고 있다.

특히 Google, Apple, Qualcomm 등 주요 기업은 자사 하드웨어 및 운영체제에 최적화된 프레임워크를 중심으로 온디바이스 AI 생태계를 구축하고 있으며, 이러한 구조는 특정 플랫폼에 종속되는 경향을 보인다. Meta와 Microsoft 역시 각각 PyTorch Mobile 및 ONNX Runtime을 통해 온디바이스 실행 환경을 제공하고 있으나, 프레임워크 간 완전한 호환성은 확보되지 않은 상태이다.

이와 같은 기술 파편화는 단순한 기술 다양성의 문제가 아니라, 온디바이스 AI 서비스의 확산과 생태계 형성에 직접적인 영향을 미치는 구조적 문제로 작용하고 있다. 특히 동일한 AI 모델이라 하더라도 프레임워크 및 디바이스 환경에 따라 재구현이 요구되는 경우가 많아 개발 비용이 증가하고, 서비스 확장성이 제한되는 문제가 발생한다.

2. 3대 핵심 표준화 이슈

기술 파편화로 인해 나타나는 표준화 이슈는 크게 구조, 인터페이스, 실행 환경의 세 가지 측면에서 도출될 수 있다.

2.1 프레임워크 구조의 비표준화

경량형 온디바이스 AI 프레임워크는 모델 변환, 최적화, 배포, 실행, 관리 등 전주기 기능을 포함하는 통합 구조를 가지나, 각 프레임워크는 이러한 기능을 상이한 구조로 구현하고 있다. 예를 들어, 모델 최적화 과정이나 런타임 구성 방식, 하드웨어 연동 방식이 프레임워크마다 다르게 설계되어 있어, 동일한 기능을 수행하더라도 구조적 차이가 존재한다.

이러한 구조적 비표준화는 프레임워크 간 호환성을 저해하며, 새로운 디바이스나 플랫폼에 적용하기 위해 추가적인 변환 과정이 요구되는 문제를 발생시킨다. 결과적으로 프레임워크 구조에 대한 공

통 참조 모델 또는 아키텍처 표준의 부재가 기술 확산의 제약 요인으로 작용하고 있다.

2.2 모델 표현 및 연동 인터페이스의 비표준화

온디바이스 AI 프레임워크는 모델 실행을 위한 다양한 모델 표현 방식과 연동 인터페이스(API/SDK)를 포함하고 있으나, 프레임워크별로 상이한 형식을 채택하고 있다. 대표적으로 TensorFlow Lite, PyTorch, Core ML 등은 각각 독자적인 모델 포맷을 사용하고 있으며, ONNX와 같은 중립적 모델 표현 방식이 존재함에도 불구하고 모든 환경에서 일관되게 적용되지는 못하고 있다.

또한, 프레임워크 간 연동을 위한 API 및 SDK 역시 플랫폼별로 상이하게 제공되고 있어, 모델의 이식성과 재사용성이 제한되는 문제가 발생한다. 특히 모델을 다른 프레임워크나 디바이스로 이전하는 과정에서 별도의 변환 도구가 필요하며, 이 과정에서 성능 저하나 기능 제한이 발생할 수 있다.

결과적으로 모델 표현 방식과 연동 인터페이스의 비표준화는 온디바이스 AI 서비스의 상호운용성을 저해하는 핵심 요인으로 작용하고 있으며, 이를 해결하기 위한 공통 모델 포맷 및 인터페이스 표준 정의가 요구된다.

2.3 실행 환경 및 하드웨어 의존성

온디바이스 AI는 디바이스 내부에서 직접 실행되기 때문에 CPU, GPU, NPU 등 다양한 하드웨어 자원에 대한 의존성이 높다. 그러나 디바이스마다 하드웨어 구성과 성능 특성이 상이하며, 이를 지원하기 위한 프레임워크의 실행 방식 역시 서로 다르게 구현되고 있다.

예를 들어, 특정 프레임워크는 특정 NPU에 최적화된 실행 환경을 제공하는 반면, 다른 프레임워크는 GPU 중심의 실행 구조를 채택하는 등 하드웨어 종속성이 강하게 나타난다. 이러한 차이는 동일한 모델이라 하더라도 디바이스에 따라 성능 편차가 발생하거나, 일부 기능이 제한되는 문제로 이어질 수 있다.

그뿐만 아니라 하드웨어 가속기와 프레임워크 간 연동 방식이 표준화되어 있지 않기 때문에, 새로운 디바이스 환경에 적용하기 위해 추가적인 최적화 작업이 요구된다. 이는 개발 복잡도를 증가시키고, 온디바이스 AI 기술 확산을 저해하는 요인으로 작용한다.

마지막으로 디바이스별 하드웨어 특성 차이로 인해 동일 모델이라도 성능 편차가 발생하며, 이를 객관적으로 비교할 수 있는 공통 성능평가 기준의 부재 역시 주요 문제로 지적된다.

IV. 국내외 표준화 추진 현황 및 분석

1. ITU-T SG13 & SG20

ITU-T는 네트워크 및 서비스 관점에서 인공지능 표준화를 추진하고 있으며, 특히 SG13[6]과 SG20[7]을 중심으로 온디바이스 AI와 연계된 기술이 논의되고 있다.

SG13은 미래 네트워크 및 관련 신기술(Future Networks and Emerging Network Technologies)을 주요 대상으로 하며, 전통적인 네트워크 기능을 넘어 클라우드, 엣지 등 분산된 인프라 환경에서 AI 기능이 내재화되는 구조를 다루고 있다. 특히 AI가 인프라 내부의 핵심 기능으로 동작하는 AI-native 인프라 개념, 즉 인프라 내 AI 기능의 배치, 실행 구조, 자원 관리 및 서비스 제공을 위한 아키텍처 모델을 정의하는 데 중점을 두고 있다.

이는 단순히 네트워크에 AI를 적용하는 수준을 넘어, 인프라 전반에서 AI 기능이 통합적으로 동작하는 구조를 표준화하는 접근으로 볼 수 있다. 또한

AI 기반 자원 최적화, 서비스 품질 향상, 자율적 운영을 위한 핵심 기술 요소들이 SG13 내에서 함께 논의되고 있다.

SG20은 지능형 사물인터넷을 기반으로 지능화된 서비스 제공을 위한 AI 핵심 기능 및 응용 기술에 대한 표준화를 추진하고 있다. 특히 IoT 디바이스 및 서비스가 인간의 개입 없이 상황을 인지하고 판단하며, 자율적으로 동작하는 구조를 구현하기 위한 기능적 요구사항과 서비스 모델 정의에 중점을 두고 있다.

이러한 관점에서 SG20은 AI 기반의 상황 인지, 의사결정, 지능형 제어 등 자율형 IoT 구현을 위한 핵심 기능을 중심으로 표준화를 진행하고 있으며, 다양한 산업 및 도시 환경에서 적용할 수 있는 서비스 프레임워크를 정의하고 있다. 또한, 디바이스 수준에서 생성되는 데이터를 기반으로 지능형 서비스를 수행하기 위한 구조와 기능 요구사항을 포함하고 있어, 온디바이스 AI 기술과의 연계성이 높다.

특히 스마트시티, 스마트홈, 산업 IoT 등 다양한 응용 분야에서 자율형 서비스 구현을 위한 AI 적용 구조가 논의되고 있으며, 서비스 관점에서의 상호운용성과 기능 연계를 주요 고려사항으로 다루고 있다.

2. ISO/IEC JTC 1/SC 42

ISO/IEC JTC 1/SC 42[8]는 인공지능 분야의 국제 표준화를 담당하는 핵심 기구로, AI 기술 전반에 대한 기반 체계를 정의하고 있다.

대표적으로 AI 시스템의 참조 모델, 용어 정의, 데이터 품질, 신뢰성 및 위험 관리 등에 관한 표준이 개발되고 있으며, AI 기술의 전반적인 구조를 이해하기 위한 공통 기준을 제공한다. 예를 들어, AI 시스템 라이프사이클, 데이터 처리 흐름, 신뢰성 확보를 위한 요구사항 등이 주요 표준화 대상이다.

이러한 표준은 온디바이스 AI 프레임워크를 포함한 AI 시스템 전반의 설계 기준으로 활용될 수 있으나, 실제 디바이스 환경에서 요구되는 경량화 구조, 모델 최적화, 실행 인터페이스와 같은 구체적인 기술 요소까지는 충분히 반영하지 못하고 있다.

따라서 SC 42 표준은 기반 개념과 참조 모델을 제공하는 역할은 수행하고 있으나, 온디바이스 AI 프레임워크 수준의 구체적 구현을 직접적으로 지원하기에는 한계가 있다.

3. 국제 사실표준화

IEEE[9]는 AI 시스템의 신뢰성, 안전성, 윤리성, 그리고 시스템 설계 및 운영 측면에서의 표준화를 추진하고 있다. 특히 AI 시스템의 투명성, 설명 가능성, 데이터 거버넌스 등과 관련된 표준이 활발히 논의되고 있으며, AI 기술의 사회적 수용성과 신뢰성을 확보하는 데 중요한 역할을 하고 있다. 그러나 이러한 표준은 시스템 수준의 요구사항과 가이드라인 중심으로 구성되어 있어, 온디바이스 AI 프레임워크의 구조나 실행 방식과 같은 기술적 세부 요소까지는 직접적으로 다루지 않는다.

한편, IETF[10], oneM2M[11] 등과 같은 사실표준화 기구에서는 데이터 교환 프로토콜, 디바이스 간 통신, 서비스 플랫폼 연동 등 구현 중심의 기술이 정의되고 있다. 예를 들어, IETF는 인터넷 기반 데이터 전송 및 프로토콜 표준을 통해 디바이스 간 연결성과 상호운용성을 지원하며, oneM2M은 IoT 서비스 플랫폼 간 연동 구조를 정의한다.

이러한 사실표준은 실제 시스템 구현과 밀접하게 연관되어 있어 온디바이스 AI 서비스의 연동 기반을 제공하는 형태로 특정 기술 또는 플랫폼 중심으로 개발되고 있다.

4. 국내표준화

국내에서는 한국정보통신기술협회(TTA)[12] 및 표준화 포럼을 중심으로 인공지능 및 응용기술 표준화가 추진되고 있다.

TTA PG1005(인공지능)를 중심으로 AI 시스템 구조, 데이터 처리, 신뢰성 및 응용 서비스 관련 표준화가 진행되고 있다. 해당 그룹에서는 AI 서비스 아키텍처, 데이터 품질, AI 신뢰성 확보 등 기반 기술에 대한 표준을 다루고 있으며, 온디바이스 AI와 관련된 기술 요소 역시 일부 포함되고 있다.

사물인터넷 분야에서는 PG1001 및 P1002를 중심으로 인공지능 기반 IoT 응용/서비스, 디바이스 및 플랫폼 기술에 대한 표준화가 추진되고 있다. 또한, 네트워크 및 서비스 인프라 관점에서는 PG1003(클라우드) 등을 통해 AI 기반 네트워크 및 지능형 서비스 구조와 관련된 기술이 다루어지고 있으며, 이는 ITU-T SG13과 연계되는 방향으로 발전하고 있다.

이와 함께 ETRI와 KETI 등 연구기관들은 온디바이스 AI 프레임워크 구조, 경량화 기술, AI 반도체 연계 기술 등을 기반으로 국내 및 국제 표준화 기구와의 연계를 추진하고 있으며, 국내 기술의 국제 표준 반영을 위한 활동을 수행하고 있다.

V. 온디바이스 AI 기술 확산을 위한 표준화 전략 및 정책 효과

1. 3대 핵심 추진 전략

온디바이스 AI 프레임워크의 기술 파편화 문제를 해소하고 상호운용 가능한 생태계를 구축하기 위해서는 프레임워크 구조, 모델 표현 및 연동 인터페이스, 실행 환경을 포괄하는 통합적 표준화 전략이 필요하다. 이를 위해 국제표준, 사실표준, 국내표준 간의 역할 분담과 협력 메커니즘을 기반으로 한 체계적인 접근이 요구된다.

우선, 프레임워크 구조 표준화는 AI 시스템 참조 모델과 인프라 기반 구조를 연계하여 디바이스까지 확장할 수 있는 공통 아키텍처를 정의하는 방향으로 추진되어야 한다. ISO/IEC JTC 1/SC 42에서 정의된 AI 참조 모델 및 라이프사이클 구조를 기반으로, ITU-T SG13에서 논의되는 AI-native 인프라 구조를 결합함으로써 상위 구조를 정립하고, 이를 온디바이스 환경까지 확장하는 접근이 필요하다. 특히 SC 42와 SG13 간에는 연락문서(Liaison Statement)를 통한 기술 요구사항 공유 및 표준 개발 방향의 정합성 확보를 기반으로 협력 관계를 유지함으로써, 참조 모델과 인프라 구조 간의 일관성을 확보해야 한다. 또한, 이러한 구조를 국내 TTA 표준화 활동을 통해 구체화하고, 국제표준과의 정합성을 확보하는 단계적 전략이 요구된다.

다음으로, 모델 표현 및 연동 인터페이스의 표준화는 프레임워크 간 상호운용성 확보를 위한 핵심 요소로, 국제표준과 사실표준 간의 연계를 중심으로 추진되어야 한다. SC 42의 데이터 및 AI 모델 관련 기본 체계를 기반으로 IEEE의 데이터 거버넌스 및 신뢰성 기준을 반영하고, IETF 및 oneM2M과 같은 사실표준화 기구에서 정의된 데이터 교환 및 서비스 연동 기술을 통합적으로 활용하는 접근이 필요하다. 특히 ONNX와 같은 중립적 모델 표현 방식을 중심으로 프레임워크 간 모델 호환성을 확보하고, 이를 국제표준 체계 내에서 수용함으로써 다양한 플랫폼 간 연계가 가능한 구조를 마련해야 한다.

마지막으로, 실행 환경 및 응용 서비스 측면에서는 자율형 IoT 기반의 AI 기능과 디바이스 중심 실행 구조를 반영한 표준화가 필요하다. ITU-T SG20에서 정의되는 자율형 사물인터넷 기반의 상황 인지, 의사결정, 지능형 제어 기능을 중심으로 서비스

구조를 정립하고, 사실표준화 기구에서 제공하는 디바이스 연동 및 서비스 구현 기술과의 정합성을 확보해야 한다. 또한, 이러한 구조를 국내 TTA 표준화를 통해 산업 적용 수준으로 구체화하는 것이 중요하다. 아울러 다양한 디바이스 환경에서 발생하는 이기종 온디바이스 AI 프레임워크 간 성능 편차 문제를 해소하기 위해, 연산 성능, 지연시간, 에너지 효율 등을 포함하는 공통 성능평가 기준 및 벤치마크 체계를 표준화할 필요가 있다. 이를 통해 디바이스 간 비교 가능성과 서비스 품질의 일관성을 확보할 수 있다.

2. 정책 및 사회적 기대효과

온디바이스 AI 프레임워크 중심의 표준화 전략은 기술적 문제 해결을 넘어 산업 및 사회 전반에 걸쳐 다양한 파급효과를 기대할 수 있다.

첫째, 기술 생태계 측면에서는 프레임워크 간 상호운용성 확보를 통해 개발 비용을 절감하고, 다양한 디바이스 환경에서의 서비스 확산을 촉진할 수 있다. 특히 표준 기반 기술을 활용함으로써 중소기업 및 스타트업의 시장 진입 장벽을 낮추고, 개방형 기술 생태계를 조성할 수 있다.

둘째, 산업 경쟁력 측면에서는 온디바이스 AI 관련 기술의 국제표준 반영을 통해 글로벌 시장에서의 주도권 확보가 가능하다. 온디바이스 AI는 스마트폰, 가전, 자동차, 로봇 등 다양한 산업과 밀접하게 연관되어 있어, 프레임워크 수준의 표준 선점은 국가 기술 경쟁력 확보에 중요한 요소로 작용한다.

셋째, 서비스 측면에서는 실시간성, 개인화, 보안성이 강화된 지능형 서비스 제공이 가능해진다. 온디바이스 AI는 데이터가 생성되는 위치에서 직접 처리함으로써 지연시간을 최소화하고, 민감한 데이터의 외부 전송을 줄여 사용자 신뢰 기반의 서비스 확산에 이바지할 수 있다.

마지막으로, 정책적 측면에서는 AI 기술의 신뢰성과 책임성을 확보하기 위한 기반이 마련된다. 표준화된 프레임워크 구조와 인터페이스, 그리고 성능평가 기준은 AI 시스템의 투명성과 검증 가능성을 향상시키며, 향후 AI 규제 및 정책 수립을 위한 기술적 기준으로 활용될 수 있다.

VI. 결론

온디바이스 AI는 데이터가 생성되는 디바이스 내부에서 직접 인공지능 처리를 수행함으로써, 기존 클라우드 및 엣지 중심 구조에서 발생하는 지연시간, 네트워크 의존성, 데이터 프라이버시 문제를 근본적으로 개선할 수 있는 차세대 핵심 기술로 부상하고 있다. 특히 실시간성, 개인화, 보안성을 동시에 요구하는 다양한 서비스 환경에서 온디바이스 AI의 역할은 더 확대될 것으로 예상된다.

그러나 본고에서 분석한 바와 같이, 현재 온디바이스 AI 기술은 프레임워크 구조, 모델 표현 방식 및 연동 인터페이스, 실행 환경 전반에서의 비표준화로 인해 기술 파편화가 심화되고 있으며, 이는 단순한 기술적 차이를 넘어 상호운용성 저하, 개발 복잡도 증가, 서비스 확장성 제한 등 산업 전반의 구조적 제약으로 작용하고 있다. 특히, 동일한 AI 모델이라 하더라도 프레임워크 및 디바이스 환경에 따라 재구현이 요구되는 현상은 온디바이스 AI 생태계 확산의 핵심 저해 요인으로 볼 수 있다.

현재 ISO/IEC JTC 1/SC 42, ITU-T SG13 및 SG20, IEEE, IETF, oneM2M 등 다양한 표준화 기구에서 관련 기술을 다루고 있으나, 대부분이 특정 계층 또는 기능 중심으로 분산되어 있어 온디바이스 AI 프레임워크의 전주기 구조를 통합적으로 포괄하기에는 한계가 존재한다. 이는 기술 파편화를 근본

적으로 해소하기 위한 표준화 체계가 아직 충분히 정립되지 않았음을 의미한다.

이러한 분석을 바탕으로 본고는 프레임워크 구조, 모델 표현 및 연동 인터페이스, 실행 환경 및 성능평가를 중심으로 한 3대 표준화 전략을 제시하였다. 특히 국제표준과 사실표준, 국내표준 간 역할 분담과 연계를 기반으로 한 협력적 표준화 접근이 필요하며, 이를 통해 상호운용 가능한 온디바이스 AI 생태계를 구축할 수 있다.

시사점 측면에서, 온디바이스 AI 표준화는 단순한 기술 정합성 확보를 넘어 산업 경쟁력과 직결되는 전략적 요소로 작용한다. 표준 기반의 프레임워크 구조는 다양한 디바이스 및 플랫폼 간 호환성을 확보하고, 개발 비용 절감과 서비스 확산을 촉진하며, 중소기업 및 신규 사업자의 시장 진입 장벽을 낮추는 효과를 가져올 수 있다. 또한, 온디바이스 AI 기술의 국제표준 반영은 글로벌 시장에서의 기술 주도권 확보와 직결되며, 국가 차원의 디지털 전환 및 AI 산업 경쟁력 강화에 중요한 기반이 될 것이다.

향후 온디바이스 AI는 생성형 AI와의 결합을 통해 디바이스 중심의 지능형 서비스로 빠르게 확산할 것으로 예상되며, 이에 따라 경량화 기술, 실행 최적화, 성능평가 체계 등 표준화 범위 역시 지속적으로 확대될 필요가 있다. 따라서 온디바이스 AI 프레임워크에 대한 통합적 표준화는 기술 발전과 산업 생태계 형성을 동시에 견인하는 핵심 수단으로 지속적인 연구와 정책적 지원이 요구된다.

용어해설

NPU(Neural Processing Unit) 딥러닝 연산을 가속하기 위해 설계된 전용 프로세서로, 신경망 추론에 최적화된 행렬 연산을 효율적으로 수행하는 하드웨어

ONNX(Open Neural Network Exchange) 서로 다른 인공지능 프레임워크 간 모델의 상호운용성을 지원하기 위한 중립적 모델 표현 형식으로, 다양한 플랫폼 간 모델 변환 및 실행을 가능하게 함

NNAPI(Android Neural Networks API) Android 환경에서 AI 연산을 가속하기 위한 인터페이스로, CPU, GPU, NPU 등 다양한 하드웨어 가속기를 활용할 수 있도록 지원하는 API

IR(Intermediate Representation) 서로 다른 프레임워크 간 모델 변환 및 실행을 위해 사용되는 중간 표현 형식으로, 모델의 구조와 연산을 공통된 형태로 표현하는 방식

SDK(Software Development Kit) 특정 플랫폼이나 환경에서 애플리케이션을 개발하기 위해 제공되는 도구 모음으로, 라이브러리, API, 문서 등을 포함

API(Application Programming Interface) 소프트웨어 간 상호작용을 가능하게 하는 인터페이스로, 기능 호출 및 데이터 교환을 위한 규격과 방법

TPU(Tensor Processing Unit) Google에서 개발한 AI 전용 가속기로, 대규모 행렬 연산과 딥러닝 모델 추론 및 학습을 효율적으로 처리하도록 설계된 하드웨어

참고문헌

[1] X. Wang et al., “Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models,” ACM Comput. Surv., vol. 57, no. 9, 2025, pp. 1-39.

[2] Google, “TensorFlow Lite,” https://www.tensorflow.org/lite

[3] Meta AI, “PyTorch Mobile/ExecuTorch,” https://pytorch.org/mobile/

[4] Microsoft, “ONNX Runtime,” https://onnxruntime.ai

[5] 네이버, “HyperCLOVA,” https://clova.ai/hyperclova

[6] ITU-T SG13, “Future networks and emerging network technologies,” https://www.itu.int/en/ITU-T/studygroups/2025-2028/13/Pages/default.aspx

[7] ITU-T SG20, “Internet of Things, digital twins and smart sustainable cities and communities,” https://www.itu.int/en/ITU-T/studygroups/2025-2028/20/Pages/default.aspx

[8] ISO/IEC JTC 1/SC 42, “Artificial Intelligence,” https://www.iso.org/committee/6794475.html

[9] IEEE Standard Association. https://standards.ieee.org/

[10] IETF. https://www.ietf.org/

[11] oneM2M. https://www.onem2m.org/

[12] TTA. https://www.tta.or.kr/