온디바이스(On-device) AI 기술 동향: 초경량 모델에서 LLM 적용까지

□ 조용훈, 김태구, 백윤주 / 부산대학교

요 약

온디바이스 AI(On-device AI)는 클라우드 중심 AI가 지닌 프라이버시, 지연 시간, 비용 등 본질적인 제약을 극복하기 위한 핵심적인 기술로써, 외부 네트워크로 민감한 정보가 유출되지 않아 프라이버시 보호 측면에서 뛰어나며, 데이터 전송으로 인한 네트워크 지연 없이 빠른 추론이 가능하다는 장점을 가진다. 온디바이스 AI 기술 구현을 위해선 MCU(Microcontroller Unit)부터 고성능 모바일 AP(Application Processor), NPU(Neural Processing Unit)에 이르는 하드웨어 플랫폼의 발전과, Tensorflow Lite와 같은 경량 추론 프레임워크와 최적화 라이브러리가 필수적이다. 또한, 자원 제약적인 임베디드 디바이스에 AI 모델을 탑재하기 위해 모델의 성능 저하를 최소화하며 크기와 연산량을 줄이는 모델 경량화 기법과 경량 딥러닝 모델 구조가 핵심 기술로 부각되고 있다. 본고는 이러한 기반 기술들에 대해 간략히 살펴본다. 이후 소개된 기술들을 바탕으로 구현된 경량 음성 인터페이스와 경량 영상 인식 시스템, 시계열 예측과 이상치 진단, 나아가 경량화된 대규모 언어 모델(Small Large Language Model, sLLM)의 적용 가능성에 대해 살펴보고, 구체적인 응용 사례들을 통해 온디바이스 AI의 가능성을 확인하고자 한다.

I. 서 론

인공지능(Artificial Intelligence, AI)은 1950년에 앨런 튜링이 튜링 테스트를 제안하면서 처음 그 개념이 등장하였다. 초창기 AI 연구는 주로 인간의 논리적 사고 과정을 기호 연산으로 모방하려는 시도에 집중되었으나, 방대한 복잡성과 이를 처리하기 위한 연산 능력의 한계, 그리고 학습에 필요한 데이터의 절대적인 부족으로 인한 시대상의 한계를 맞이했다. 2000년대 이후 인터넷 보급 확산으로 인한 빅데이터 시대의 개막과 GPU의 발전으로 인해 가능해진 고성능 병렬 컴퓨팅 환경과 딥러닝 알고리즘의 등장은 기존 AI의 시대적 한계를 극복하고 이미지 분류, 자연어 처리 등 다양한 분야에서 기존 접근법을 상회하는 성능을 달성하면서 기술의 급격한 발전을 이끌었다. 이러한 발전은 막대한 컴퓨팅 자원을 요구하기에, 자연스럽게 클라우드 인프라를 중심으로 AI 기술과 서비스가 구현되는 패러다임이 형성되었다.

하지만 ChatGPT, Gemini 등의 딥러닝 서비스 제공에 필수적이라고 여겨지는 클라우드 기반 AI 기술은 본질적인 한계를 가진다. 모든 데이터를 중앙 서버로 전송하고 처

리하는 과정에서 발생하는 데이터 프라이버시 유출 문제와 서버 과부하, 물리적인 통신 거리와 네트워크 상태에 따라 발생하는 응답 지연, 지속적인 데이터 전송에서 발생되는 통신 비용이 대표적이다. 최근 소형 MCU(Microcontroller Unit)의 발전으로 경량 AI 모델의 충분한 성능 구현이 가능해짐에 따라, 네트워크 연결 없이 기기 내부에서 자체적으로 AI 연산을 수행하는 온디바이스 AI 기술이 그 필요성을 인정받으며 대안으로 부상하고 있다[1]. 본고에서는 온디바이스 AI 구현을 위한 핵심 기술과 플랫폼, 경량화 기법과 주요 활용 사례를 중심으로 논하고자 한다.

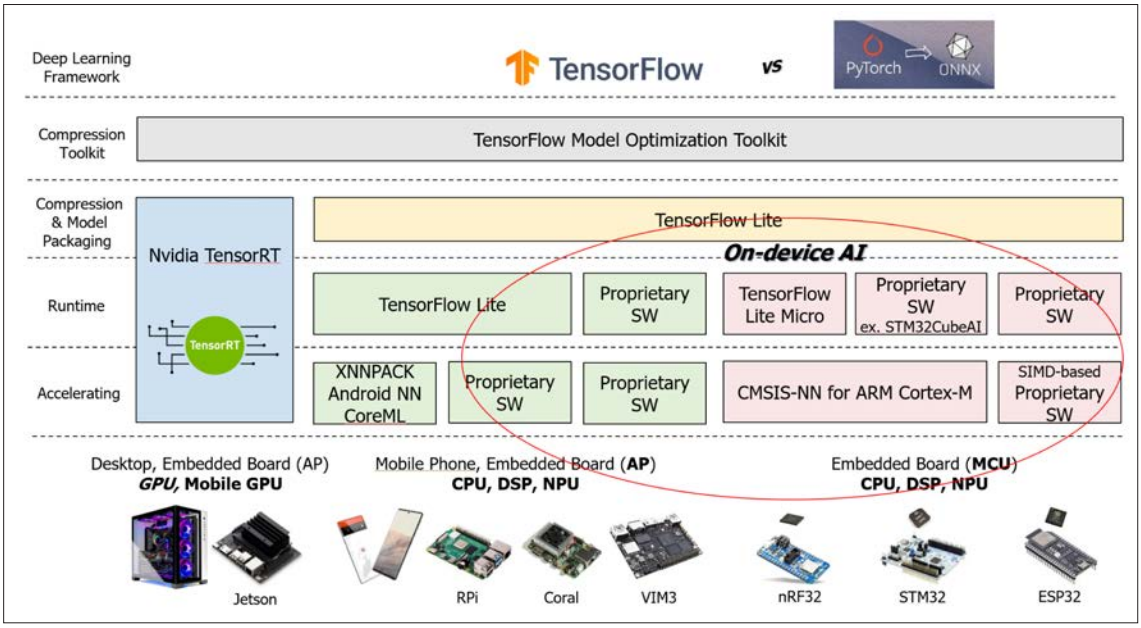
Ⅱ. 온디바이스 AI 기술과 플랫폼

온디바이스 AI는 딥러닝 모델의 추론 과정에 클라우드 서버가 아닌, 스마트폰이나 IoT 기기 등의 단말(Edge Device)에서 수행되는 기술을 의미한다. 사용자 데이터가 외부 네트워크로 전송되지 않아 민감한 정보 유출을 원천적으로 차단해 프라이버시 보호 측면에서 뛰어나며, 기기 자체에서 추론하므로 네트워크 지연이 없어 빠른 응답 속도를 가진다. 온디바이스 AI는 딥러닝 모델이 단말에서 수행되므로 하드웨어와 소프트웨어 양단에서의 기술 발전으로 인해 구현된다.

1. 하드웨어 플랫폼

온디바이스 AI를 구현하기 위한 하드웨어 플랫폼은 요구되는 성능, 전력 소모, 비용 등에 따라 다양하게 분류될 수 있다. 큰 분류로는 MCU, 저사양 AP 수준, 모바일 AP 수준의 세 단계로 구분할 수 있다.

MCU는 CPU 코어, 메모리, I/O 포트가 하나의 칩에 통합된 형태로, 운영체제(OS) 없이 펌웨어나 실시간 운영체제(RTOS) 기반으로 동작하는 자원 제약적인 임베디드 시스템에 최적화된 프로세서다. 저전력 동작과 저렴한 비용, 초소형 크기로 인해 웨어러블 디바이스나 스마트 센서, IoT 기기에 널리 사용되고 있다. MCU는 많게는



<그림 1> 온디바이스 AI 구현을 위한 소프트웨어 및 하드웨어 플랫폼 아키텍쳐

수 MB에서 KB 단위의 메모리를 보유하여 복잡한 모델 보단 초경량 AI 모델을 탑재한 온디바이스 AI 시스템을 개발하는데 적합하다. 온디바이스 AI를 위한 하드웨어로는 ARM Cortex-M 기반 MCU인 STMicroelectronics의 STM32 제품군이나 Espressif Systems의 ESP32 제품군이 대표적이며, 각 제조사에서 자체적인 모델 변환 도구를 제공하는 등 개발 용이성이 뛰어나 다양한 연구에 적극적으로 사용된다.

저사양 AP는 리눅스와 같은 범용 운영체제가 동작할 수 있을 만한 성능을 가진다. 수 GB 수준의 메모리를 통해 MCU보다 복잡한 AI 모델을 실행할 수 있으며, 표준 AI 프레임워크를 활용해서 학습과 추론이 가능하다. 또한, 다양한 인터페이스를 갖춰 높은 범용성을 가진다. 대표적으로 ARM Cortex-A 기반의 AP인 라즈베리파이(Raspberry Pi)가 있으며, Python을 포함한 다양한 언어와 Tensorflow, PyTorch와 같은 프레임워크를 실행할 수 있어, 온디바이스 AI 기반의 단순한 이미지 분류나 자연어 처리 시 우선적으로 고려되는 하드웨어이다.

모바일 AP는 스마트폰이나 태블릿 등 모바일 기기에 주로 탑재되는 고성능 프로세서로, 수십 CPU, GPU와 함께 전용 NPU를 탑재하여 AI 연산 효율성과 추론 속도를 극대화한 것이 특징이다. 수십 GB 수준의 메모리가 필요한 생성형 AI와 같이 복잡하고 무거운 AI 모델을 실시간으로 처리할 수 있으며, 고해상도 영상 데이터 처리나 자율주행에도 활용된다. 퀄컴 스냅드래곤(Qualcomm Snapdragon) 뿐만 아니라 엔비디아 젯슨(NVIDIA Jetson) 시리즈도 모바일 AP 수준의 하드웨어 플랫폼으로 분류할 수 있다.

2. 소프트웨어 플랫폼과 라이브러리

개발한 AI 모델을 실제 기기에 배포하고 효율적으로 실행하기 위한 소프트웨어 플랫폼도 중요하다. 소프트웨어 플랫폼을 통해서 학습된 모델을 특정 하드웨어에 맞게 변환하고 최적화하는 과정을 거치며, 다양한 하드웨어 가속기를 활용해 추론하는 기능을 제공한다.

TFLite(Tensorflow Lite)와 TFLite Micro(Tensorflow Lite for Microcontrollers)가 대표적인 프레임워크이며, 추론 성능을 극대화하기 위해 ARM CMSIS-NN과 같은 최적화 라이브러리도 중요하다.

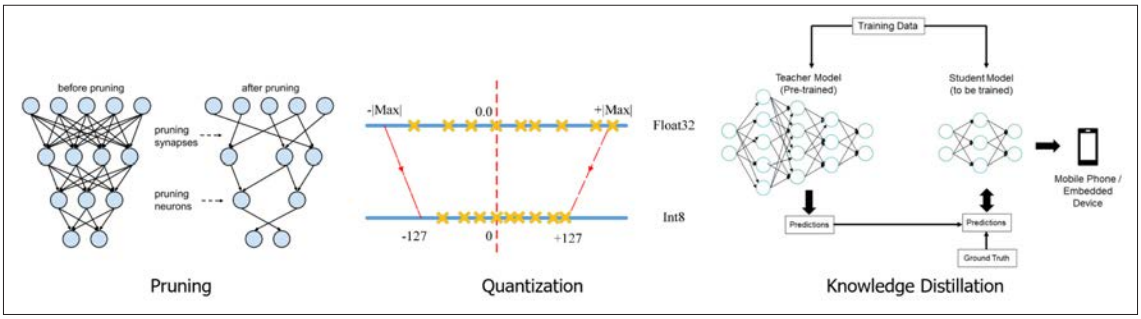
TFLite는 Google에서 제공하는 모바일과 임베디드 환경을 위한 추론 프레임워크다. Tensorflow 모델을 FlatBuffers라는 특수 형식으로 변환한다. 해당 변환을 통해 추가적인 파싱이나 압축해제 단계 없이 직접 데이터를 액세스할 수 있어 자원 제약적인 환경에서도 효율적인 TFLite의 실행이 가능하다. Android나 iOS, Linux와 MCU 등 다양한 플랫폼을 지원하며, CPU 외에도 GPU, DSP, NPU 등의 하드웨어 가속 및 모델 최적화를 사용한다. 일반적으로 스마트폰이나 라즈베리파이와 같은 AP 기반의 임베디드 보드에서 사용된다[2].

TFLite Micro는 TFLite를 더욱 경량화하여 수십 KB 수준의 메모리를 지닌 MCU 환경에 적합하도록 설계된 C++ 기반의 프레임워크다. 운영체제나 동적 메모리 할당을 사용하지 않고, 필수적인 연산자만을 포함하여 STM32나 ESP32 등 저전력 MCU 기반의 온디바이스 AI 구현에 핵심적인 역할을 수행한다[3].

ARM CMSIS-NN은 ARM Cortex-M 시리즈 MCU를 위해 개발된 커널 라이브러리로, 컨볼루션(Convolution) 연산과 같이 딥러닝 네트워크 추론에 필요한 함수들을 최적화한다. INT8, INT16 등 정수 연산에 특화되었으며, TFLite Micro가 CMSIS-NN을 활용해 MCU 환경에서 연산 효율성을 극대화한다.

Ⅲ. 온디바이스 AI 구현을 위한 경량화 기법

수 GB 수준의 AP나 수 MB, KB 단위의 메모리를 가지는 MCU에 딥러닝 모델을 탑재하는 온디바이스 AI를 구현하기 위해선 학습된 모델의 정확도를 유지하면서 경량화하는 기술이 필수적이다. 경량 딥러닝 모델



<그림 2> 모델 경량화 기술

의 개발을 위해 경량화에 일반적으로 사용되는 기술인 가지치기(Pruning), 양자화(Quantization), 지식 증류(Knowledge Distillation)를 적용하거나, CNN과 같은 일반적인 딥러닝 모델 구조가 아닌, 컨볼루션 연산을 효율적으로 개선한 DS-CNN(Depthwise Separable CNN)을 활용하는 등 경량 딥러닝 모델 구조를 사용한다.

가지치기는 모델을 구성하는 가중치가 모델 성능에 얼마나 기여하는지를 평가하고, 설정된 임계값보다 낮은 중요도를 가지는 가중치를 제거하는 방식으로 모델을 경량화한다. 가중치의 중요도를 평가하는 기법으로는 일반적으로 크기 기반(Magnitude-based)의 방식을 사용하는데, 절대값이 0에 가까운 가중치일수록 중요도가 낮다고 가정하여 제거하는 방식이다. 드롭아웃(Dropout) 방식과 유사할 수 있으나 가지치기는 제거한 가중치를 보관하지 않는데 반해 드롭아웃은 학습 시 뉴런들을 일부 비활성화해서 학습하고, 추론 시에는 모든 뉴런들을 활용한다는 차이가 있다.

양자화는 모델의 가중치나 활성화 값을 표현하는데 사용되는 데이터의 정밀도를 낮춰 모델을 경량화한다. 일반적으로 딥러닝 모델은 32-bit 부동소수점(FP32)으로 학습되는데, 양자화는 이를 8-bit 정수형(INT8)이나 16-bit 부동소수점(FP16)으로 변환하여 파라미터를 표현하는 비트 수를 줄인다. 온디바이스 AI에서 양자화는 특히 중요한 과정이다. 이는 일부 MCU에서 부동소수점 연산을 지원하지 않거나 정수 연산에 최적화되어, 정수형으로의 변환을 통해 추론 속도를 크게 향상시킬 수 있고 하드웨어 성능을 최대한 활용할 수 있기 때문이다.

지식 증류는 사전에 잘 학습된 대용량 모델인 교사 모델의 지식을 작은 학생 모델에게 전달한다. 해당 경량화 방식은 학생 모델이 교사 모델의 출력을 모방하도록 학습하는데, 단순히 최종 결과만 학습하는 것이 아니라 교사 모델이 예측하는 클래스별 상세한 확률 분포 전체를 모방하도록 학습한다. 지식 증류는 최근 온디바이스 AI 구현 시 높은 정확도를 달성하는데 필수적으로 사용되는 기법이다. 성능 저하를 최소화하면서 모델을 효과적으로 경량화할 수 있으며, 다른 경량화 기법들을 함께 사용하여 MCU에 탑재 가능한 KB 수준의 경량화된 모델을 구현하기에 용이하다.

Ⅳ. 온디바이스 AI 주요 활용 사례

삼성은 2024년 1월 31일, 세계 최초로 온디바이스 AI를 지원하는 스마트폰인 갤럭시 S24 시리즈를 출시하여 많은 관심을 받았다. 전화 통화 시 별도의 앱 없이 실시간으로 양방향 음성 및 텍스트 번역을 제공하는 실시간 통역 기능을 선보였으며, 통화 내용이 기기 외부 서버로 전송되지 않고 스마트폰 자체 NPU를 활용해 번역하여 프라이버시를 보호한다. 삼성 노트 앱에서 작성한 글을 자동으로 요약하거나, 미리 설정된 서식에 맞춰 정리하고 표지를 생성

<그림 3> 갤럭시 S24의 온디바이스 AI를 통한 실시간 번역 주요 처리 기술 (이미지=삼정KPMG)[4]

하거나 번역하는 노트 어시스트 기능과 음성 녹음 앱에서 녹음된 내용을 텍스트로 변환하고, 여러 화자를 구분하여 변환된 텍스트를 요약하거나 번역하는 기능인 텍스트 변환 어시스트 기능도 선보였다.

삼성 갤럭시 S24는 온디바이스 AI가 일상적인 스마트폰 사용 경험을 어떻게 혁신할 수 있는지 보여주는 대표적인 사례이다. 특히 실시간 통역이나 노트 어시스트와 같은 기능들은 빠른 응답 속도, 강력한 프라이버시 보호, 오프라인 구동 등 온디바이스 AI가 가지는 핵심적인 가치를 분명히 드러낸다. 갤럭시 S24의 사례가 사용자 편의 기능에 집중했다면, 온디바이스 AI는 스마트폰 AP에서의 시도를 넘어 수 자원 제약적인 MCU에서의 구현을 통해 보다 넓은 영역으로 확장되고 있다.

1. 초저지연-초경량 음성 인터페이스

핵심어 검출(Keyword Spotting, KWS)은 온디바이스 AI의 대표적인 응용 분야이다. 음성 인식은 발화된 문장 전체를 텍스트로 변환하는 복잡한 시퀀스 변환 문제이므로 음향 모델링과 언어 모델링을 위해 LSTM, Transformer 등 복잡한 신경망으로 학습된 대규모 언어 모델이 필요하다. 하지만 핵심어 검출은 특정 단어가 발화되었는지에 대한 유무만을 판단하는 단순한 탐지나 분류 기술이므로 작고 효율적인 신경망 구조로 구성되어 적은 메모리와 연산력을 가지는 자원 제약적인 임베디드 디바이스에 탑재되기에 적합하다.

현재까지의 핵심어 검출 시스템은 음성 비서를 호출하는 Wake-up Word를 인식하는 기술로써 활용됐다. 사용자가 특정 키워드를 발화했을 때, 기기 자체에서 저전력으로 단어를 인식하고 후속 명령 처리를 위해 오디오 데이터가 서버로 전송되거나 간단한 명령의 경우 기기 내에서 추가 처리되는 방식이다. 이는 시스템의 전력 소모를 최소화할 수 있으며, 불필요한 오디오 데이터가 외부 네트워크로 유출되는 것을 최소화해 개인 프라이버시 보장에 효과적이다. 이외에도 핵심어 검출 시스템을 보다 다양한 분야에 활용하고자 하는 시도 또한 활발하다. MCU 기반 임베디드 시스템에 탑재한 핵심어 검출 시스템은 도입 비용이 저렴하며, 초소형·저전력 설계로 배터리 기반 동작이 가능해 화자와 인접한 곳에 설치가 가능하다. 이를 통해 소음이 극심한 산업 현장에서 긴급 상황 발생 시 사전 학습된 작업자의 키워드를 인식해 설비를 즉각 정지시키는 긴급 정지 서비스로 안전한 작업 환경의 제공[5]이 가능하다. 또한, 화자 인식 기술과의 결합을 통해 특정 작업자의 음성에만 반응하여 시스템을 제어하는 등 고도화된 온디바이스 핵심어 검출 서비스를 위한 경량화 기술도 지속적으로 연구되고 있다.

2. 경량 영상 인식 시스템

온디바이스 AI 기반 영상 인식 시스템은 주로 이미지

처리에 요구되는 높은 연산력과 메모리 사용량으로 인해 GPU가 탑재된 고성능 임베디드 디바이스에서 구현된다. 이는 1차원 시간 축에 따른 소리의 진폭 변화를 나타내는 오디오 데이터나 매우 낮은 차원을 가지는 센서 데이터 대비 이미지 데이터가 공간적으로 표현해야 할 정보의 양이 많기 때문이다. 그럼에도 임베디드 디바이스에 영상 인식 모델을 탑재하고자 하는 수요는 꾸준히 증가하고 있다. 내부 설비 유출에 민감한 공장 등 산업 환경에서 온디바이스 AI 기반의 불량 탐지 기술이 대표적이다.

AI 기반 불량 탐지 시스템은 제품 표면의 손상이나 불순물을 감지하기 위해 고해상도 이미지를 활용한다. 이는 제품의 미세한 부분을 검사하기 위함이나, 해상도가 높아질수록 실시간 인식에 필요한 연산력과 메모리 사용량이 증가하므로 경량 네트워크 구조를 기반으로 모델을 학습하거나 모델 경량화 기법 적용, 입력 데이터 전처리를 통해 연산 부하를 낮추는 방향으로 온디바이스 AI를 구현한다. 저해상도 이미지를 활용해 단순하고 특정 목적에 최적화된 모델을 개발하는 연구도 활발하다. 사람의 재실 여부나 시각적 이벤트(Visual Wake Words) 감지, 간단한 손동작을 인식하는 등의 다양한 활용이 가능하며, AP가 아닌 MCU 기반의 임베디드 디바이스로도 온디바이스 AI의 구현이 가능하다. 혹은 저해상도 이미지 대신 다른 열화상 이미지 등 다른 도메인으로부터 유용한 특징을 추출하여 객체를 인식하는 연구[6]가 수행되고 있다.

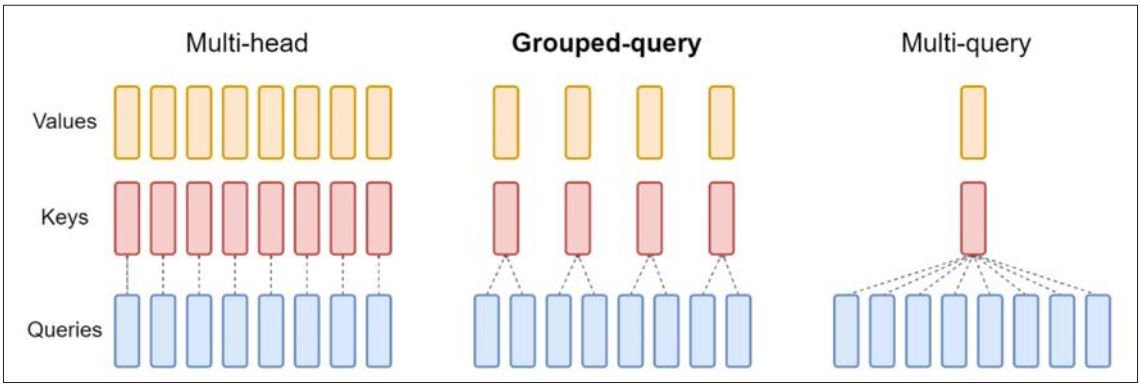
3. 시계열 예측 및 이상치 진단

이상치 진단 시스템은 1D 센서 데이터나 다변량 시계열 데이터를 입력으로 추론하기에 학습된 모델이 상대적으로 작아 MCU에 탑재하기에 용이하다. 이상치 진단의 경우 두 가지 형태로 구현되는데, 첫 번째는 정상 데이터와 라벨링(Labeling)된 이상 데이터를 함께 학습하는 방식이다. 이는 분류(Classification) 기반의 시스템으로 이상 종류를 분류할 수 있다는 점에서 유용하나 학습에 충분한 이상 데이터를 수집하기가 어렵다. 두 번째는 정상 데이터로만 모델을 학습시킨 후, 재구성 오차(Reconstruction Error, RE)를 기반으로 이상을 판단하는 방식이다. 이는 정상 데이터만으로 학습된 모델은 정상 데이터를 잘 재구성할 수 있으나, 이상 데이터는 재구성하지 못할 것이라는 가정을 기반으로 한다. 이때 발생하는 원본 데이터와 복원된 데이터 간의 차이, 재구성 차이를 측정해 이상을 판별하며, 수집이 쉬운 정상 데이터만으로 이상을 감지할 수 있다는 장점을 가진다.

온디바이스 AI 기반 이상치 진단은 스마트 워치에서 사용자의 심박수나 활동량 등의 생체 데이터를 분석하여 불규칙한 심장 박동 등 건강 이상 징후를 조기에 판별하는데 활용된다[7]. 혹은 워치 내부의 IMU 센서를 기반으로 낙상을 감지하는 등의 위험행동을 인식하는데도 효과적이다[8]. 산업 현장에서도 기술 도입에 대한 수요가 높은데, 기존 설비에 이상치 모델을 탑재한 온디바이스 AI 시스템을 부착하는 것으로 기기에 발생하는 이상을 감지할 수 있다. 수중에 설치된 펌프에 온디바이스 AI 시스템을 부착하여 이상을 판별하는 것으로, 펌프의 고장을 즉각적으로 인식하여 시간이 많이 소요되던 기존 점검 방식을 개선하는 연구가 수행된 바 있다[9].

Ⅴ. 경량 sLLM 기반 온디바이스 AI

대규모 언어 모델(Large Language Model, LLM)은 인간의 언어를 이해하고 생성하는 능력을 갖춘 AI 모델로, OpenAI의 ChatGPT나 Google의 Gemini가 대표적이다. LLM은 뛰어난 성능에 비례해 막대한 파라미터 수로 인해 모델 크기가 매우 크고 연산량이 많아 클라우드 환경에서만 구동되었다. 고성능의 LLM을 훈련시키기 위해서 방대한 양의 데이터와 고성능 컴퓨팅 자원, 엄청난 전력이 필수적이며, 훈련 이후 LLM 서비스를 제공하는데 발생하는 추론 비용도 막대하다. sLLM은 LLM보다 파라미터 수와 모델 크기를 줄여 보다 가볍고 효율적으로 만든 경량화된 언어 모델을 뜻한다. 모델 경량화를 위해 거대 모델이



<그림 4> MHA와 GQA, MQA의 QKV 헤드 구성

가지는 범용성 대신 특정 도메인에 특화되도록 개발하는 경우가 많으며, 이 과정에서 모델 구조 최적화나 앞서 소개된 경량화 기법들이 적용된다.

sLLM을 위해 LLM 모델을 구성하는 기본 블록을 보다 효율적으로 개선하고자 하는 시도가 이어지고 있다. LLM 모델의 기본이 되는 Transformer 모델의 핵심인 어텐션(Attention) 매커니즘에서 사용하는 표준적인 MHA(Multi-Head Attention) 대신 MQA(Multi-Query Attention)나 GQA(Grouped-Query Attention)[10]를 사용하여 추론 시 메모리 사용량과 연산량을 줄이는 연구가 수행되었다. MHA는 입력된 쿼리(Query), 키(Key), 값(Value) 벡터를 여러 개의 헤드로 나누어 각각 어텐션 연산을 수행한 후 이를 병합해 최종적인 출력을 얻는 방식이다. MQA는 이를 모든 쿼리 헤드가 단일 키와 값 헤드를 공유하는 방식으로 단순화하였으며, 이를 통해 모델의 품질이 저하될 수 있다는 문제를 개선하고자 두 방법의 절충안인 GQA가 제안되었다.

LLM 경량화를 위한 양자화 방법으로는 학습 후 양자화하는 PTQ(Post-Training Quantization)와 모델 학습 과정에서 가중치를 조절하는 방식인 QAT(Quantization-Aware Training)가 있다. 각 양자화 기법은 LLM 모델에 따라 적합한 방식을 사용하며, 일부 LLM 기반 모델에는 가중치를 업데이트 하지 않는 경우가 존재한다[11]. 이 경우 PTQ 기법을 통해 LLM 모델을 양자화하여 저사양 AP 수준에서 sLLM을 구현할 수 있다.

Ⅵ. 결 론

온디바이스 AI는 하드웨어 플랫폼과 경량화 기술의 발전으로 인해 현실의 다양한 영역에서 지능형 AI 서비스를 구현하는 핵심 동력으로 자리 잡고 있다. 이는 저전력 고효율 하드웨어의 발전, 모델 배포 및 실행을 위한 소프트웨어, 그리고 AI 모델 자체를 경량화하는 지속적인 알고리즘 개발이 복합적으로 작용한 결과로 보인다. IoT 환경에서 구현되는 단순 센서 데이터 처리 수준을 넘어, 복잡한 자연어 처리와 고해상도 영상 분석까지 기기 자체에서 수행하려는 시도가 활발히 이루어지고 있으며, 특히 sLLM의 등장은 온디바이스 AI의 지능 수준을 한 단계 끌어올릴 중요한 변곡점이 될 것으로 기대된다. 프라이버시에 민감한 개인화 서비스부터 즉각적인 반응성이 요구되는 산업 자동화 및 안전 시스템, 네트워크 환경이 열악한 환경에서의 AI 활용에 이르기까지, 온디바이스 AI는 응용 분야를 지속적으로 확장하며 사용자 경험을 혁신하고 산업의 지능화를 가속하고, 클라우드 너머 우리 일상과 산업 현장 곳곳에 안전하고 효율적으로 통합시키는 중추적인 역할을 수행할 수 있을 것으로 기대한다.

참 고 문 헌

[1] 최철준. (2024). 온디바이스 AI 동향 및 활용 분야. 전자공학회지, 18-25.
[2] https://www.tensorflow.org/lite/guide?hl=ko
[3] https://www.tensorflow.org/lite/microcontrollers?hl=ko
[4] 삼정PKMG 경제연구원, Issue Monitor 제165호, 생성형 AI에게 펼쳐진 새로운 무대, 온디바이스 AI
[5] 신채림, 임재봉, 백윤주. (2024). 극심한 소음 환경에서 위험 감지 시스템을 위한 지속적 학습 기반 음성 인식 기술 구현. 한국정보통신학회논문지, 28(11), 1311-1320. 10.6109/jkiice.2024.28.11.1311
[6] 윤현석, 김응태. (2024). 엣지 디바이스용 실시간 열화상 객체 검출을 위한 YOLOv5기반 경량화 방법론. 방송공학회논문지, 29(5), 703-712. 10.5909/JBE.2024.29.5.703
[7] Šabić, Edin, et al. "Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data." *Ai & Society* 36.1 (2021): 149-158.
[8] Mauldin, Taylor R., et al. "SmartFall: A smartwatch-based fall detection system using deep learning." *Sensors* 18.10 (2018): 3363.
[9] Antonini, Mattia, et al. "An adaptable and unsupervised TinyML anomaly detection system for extreme industrial environments." *Sensors* 23.4 (2023): 2344.
[10] Ainslie, Joshua, et al. "Gqa: Training generalized multi-query transformer models from multi-head checkpoints." *arXiv preprint arXiv:2305.13245* (2023).
[11] Liu, Yong, et al. "Autotimes: Autoregressive time series forecasters via large language models." *Advances in Neural Information Processing Systems 37 (2024): 122154-122184.*

저 자 소 개

조용훈

- 2020년 : 동의대학교 컴퓨터공학과 공학사
- 2020년 ~ 현재 : 부산대학교 정보융합공학과 석박사통합과정
- 주관심분야 : Embedded System, On-device AI, Healthcare, Anomaly Detection

김태구

- 2017년 : 인제대학교 전기전자컴퓨터공학과 공학사
- 2020년 : 부산대학교 컴퓨터공학과 공학석사
- 2020년 ~ 현재 : 부산대학교 컴퓨터공학과 박사과정
- 주관심분야 : Embedded System, TinyML, Human Activity Recognition, Driver Behavior Analysis

저 자 소 개

백 윤 주

- 1990년 : 한국과학기술원 전산학과 공학사
- 1992년 : 한국과학기술원 전산학과 공학석사
- 1997년 : 한국과학기술원 전산학과 공학박사
- 1999년 ~ 2002년 : 네이버 CTO
- 2003년 ~ 현재 : 부산대학교 정보컴퓨터공학과 교수
- 주관심분야 : Embedded System, TinyML, RTLS System, Wireless Sensor Networks