Chapter 01

온디바이스 AI 하드웨어 및 소프트웨어 기술개발 동향

이석준_한국전자기술연구원 선임연구원

온디바이스 AI 기술은 디바이스 자체적으로 AI 모델을 수행하기 때문에 네트워크 및 클라우드에 부하를 주지 않으면서 지능형 서비스를 제공할 수 있다. 디바이스가 수집하는 센싱 데이터의 해상도가 급격하게 높아지고, 운영하는 디바이스의 수가 폭발적으로 증가함에 따라 종래 중앙에서 데이터 처리 및 AI 분석을 수행하는 중앙집중형 방식에서는 네트워크와 클라우드의 부하로 인해 한계가 존재한다. 이에 따라 온디바이스 AI 기술에 대한 필요성이 대두되고 있다. 본 고에서는 온디바이스 AI를 실현하기 위해 AI 모델 수행에 최적화된 하드웨어 기술 개발과 AI 모델을 경량화하여 온디바이스 환경에서 AI 수행을 지원하기 위한 소프트웨어 기술 개발의 최신 동향을 분석한다.

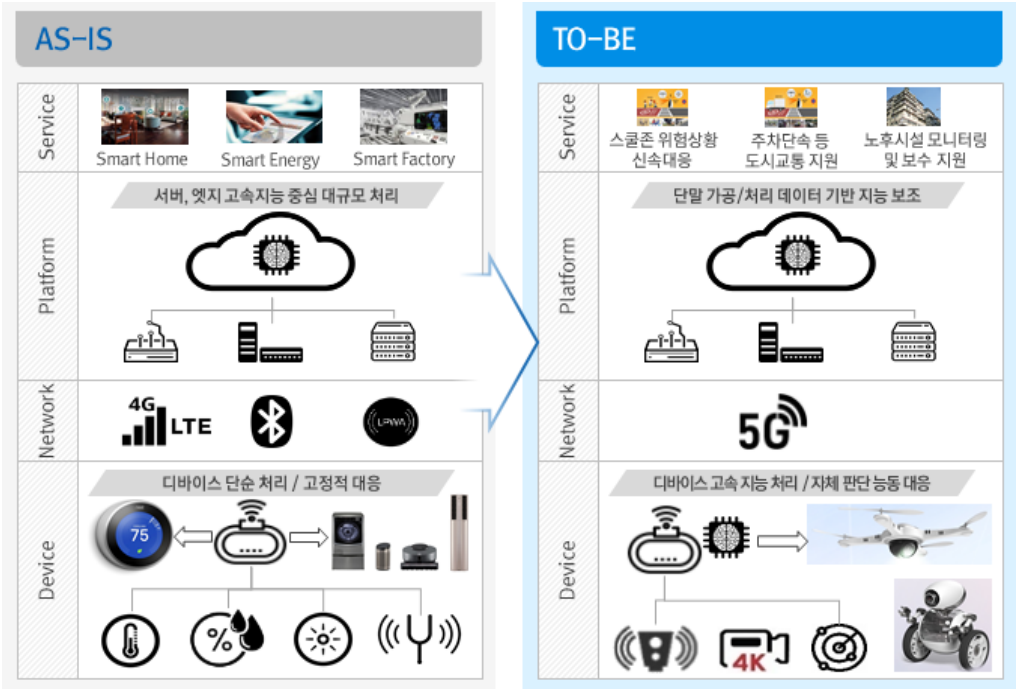
I. 서론

기존 IoT(Internet of Things) 기술은 사물과 사물 간 연결을 통해 인프라를 구축하고 이를 이용하여 데이터를 수집하거나 정해진 방식의 서비스를 제공하는 것을 목적으로 연구 및 개발되어 왔다[1]. 지능형 IoT 기술은 한발 더 나아가 인공지능 기술을 접목하여 수집되는 방대한 데이터를 인공지능 기술로 분석하여 사용자에게 최적의 맞춤 서비스를 제공하는 것을 목표로 한다[2]. 특히, 4차 산업혁명 단계로 접어듦에 따라 디바이스의 성능 향상으로 인해 고용량 데이터의 수집 및 처리가 가능해지고, 5G 기술 등 네트워크가 발달하여 대규모 데이터의 송수신이 가능해지고, 빅데이터 기술 및 인공지능 기술의 발전으로 지능형 서비스 제공이 가능해짐에 따라 빠르게 발전하고 있다.

이러한 지능형 IoT 기술은 센서, 디바이스, 네트워크, 인프라, 플랫폼 등 전방위적인 기술

* 본 내용은 이석준 선임(☎ 031-789-7577, sjlee88@keti.re.kr)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.



〈자료〉 한국전자기술연구원 자체 작성

[그림 1] 중앙집중형 AI 방식 대비 온디바이스 AI 방식

의 융합으로 이루어진다. 이 중 디바이스 자체에서 인공지능을 수행하기 위한 온디바이스 AI[3] 기술은 클라우드와 망의 부하를 최소화하면서 즉각적인 지능형 서비스 제공이 가능하다는 점에서 지능형 IoT의 핵심기술 중 하나이다.

온디바이스 AI 기술 개발에 대한 필요성은 디바이스가 수집하는 센싱 데이터의 해상도가 높아지고, 디바이스의 수가 폭발적으로 증가함에 따라 꾸준히 제기되어 왔다. 종래 디바이스에서 수집한 데이터를 클라우드나 서버 등 중앙에서 일괄적으로 분석하는 중앙집중형 AI 기술[4]은 중앙의 높은 연산능력을 기반으로 고복잡도의 인공지능 기술을 적용할 수 있다는 장점은 있으나, 수집되는 데이터 양이 증가할 경우 네트워크와 클라우드 등에 부하가 증가하여 원활한 서비스 제공이 힘든 단점이 있다. 데이터 처리를 네트워크 종단에서 수행하는 기존 엣지 컴퓨팅의 경우 단말에서 수집된 데이터를 중앙까지 보내지 않기 때문에 중앙의 부하를 감소시킨다. 그러나 최근 4K카메라, LiDAR 등 초고해상도 센서가 디바이스에서 사용됨에 따라 이러한 대용량 데이터를 엣지에서 수집하고 AI 분석까지 수행하기 위해서는

엣지와 디바이스 간 네트워크에 부하가 증가하며, 엣지의 부하 또한 커진다. 이에 따라 [그림 1]과 같이 온디바이스 AI 기술 개발에 대한 필요성이 증가하고 있으며, 기술 개발도 활발하게 진행되고 있다. 본 고에서는 온디바이스 AI 관련 하드웨어 및 소프트웨어 분야에서의 국내외 최신 동향에 대해 분석하고자 한다.

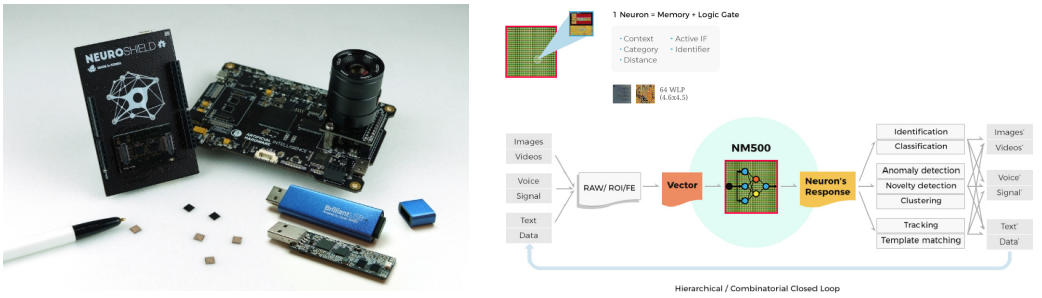
II. 온디바이스 AI 하드웨어 기술개발 동향

클라우드 및 서버 중심의 인공지능 기술이 발달함에 따라 인공지능 하드웨어는 다양한 분야에 적용할 수 있도록 범용성에 초점을 맞춰 기술이 많이 발전되어 왔다. 그러나 점차 다양한 환경에 따라 필요한 네트워크의 규모, 네트워크의 구조에 특화된 형태로 분화되고 있다. IoT 분야에서는 인공지능 기술을 리소스가 제한된 IoT 디바이스 자체에서 수행하기 위해 뉴럴 네트워크 가속 프로세서를 활용하여 인공지능 모델의 네트워크를 실행 및 처리하는 방법에 따라 고속 행렬 연산을 가능하게 하는 하드웨어 기술이 연구 및 개발되고 있다. 세부 기술에 따라 크게 인공지능 네트워크 연산에 최적화된 하드웨어와 뉴로모픽 기술을 활용한 하드웨어로 구분이 가능하다.

뉴로모픽 기술이란 인간의 뇌가 뉴런 간 상호작용을 기반으로 동작한다는 것에 착안하여 하드웨어적으로 뉴런 구조를 모사함으로써 단순 사칙연산을 넘어 인간과 유사하게 고등 사고를 수행하는 하드웨어 개발을 목표로 하는 기술이다[5]. 이는 행렬로 이루어진 인공지능 네트워크를 기반으로 학습과 추론을 수행하는 GPU(General Processing Unit) 기반 하드웨어와는 다르게 뉴로모픽 칩 내 뉴런 자체를 학습시키고 추론에 사용한다는 점에서 차이가 있다. 또한, 복잡한 부동소수점 연산을 요구하는 인공지능 네트워크와는 달리 단순히 인공뉴런만 동작시키기 때문에 소모 전력이 적고 결과 도출 속도가 빠르다. 단점으로는 기존 인공지능 네트워크를 활용하기 힘들기 때문에 확장성에 제약사항이 존재한다.

1. 국내 기술개발 동향

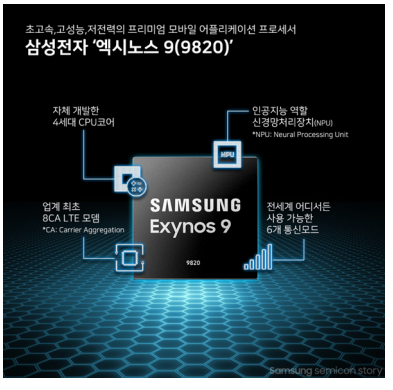
국내 인공지능 하드웨어 개발은 삼성전자와 네패스를 중심으로 활발하게 진행되고 있다. 네패스는 뉴로모픽 기술에 집중하여 엣지 및 디바이스에서 사용 가능한 인공지능 가속기



[그림 2] 네페스 뉴로모픽 칩 NM500

기술을 개발했다. 특히, 2018년 뉴로모픽 칩 양산에 세계 최초로 성공하여 상용화된 뉴로모픽 칩 기반 가속 모듈인 NM500을 출시했다([그림 2] 참조). NM500은 576개의 인공뉴런을 이용하여 고속 병렬연산을 지원한다. 또한, 에너지 효율적인 소형 폼팩터 구성요소로 NeuroMem 기술을 구현하였으며, NM500을 효율적으로 사용하기 위해 자체적으로 개발한 Knowledge Studio 소프트웨어도 제공하고 있다. 이후 NM500을 기반으로 엣지 디바이스, 스택 가능한 모듈형 디바이스, 라즈베리 파이를 위한 모듈형 디바이스 등 실제 학습 및 추론을 위한 다양한 디바이스들을 개발했다.

삼성전자는 NPU(Neural Processing Unit; 신경망처리장치) 시장 선점을 위해 지속적인 투자를 통해 리딩하고 있다. NPU란 기존에 인공지능 네트워크 연산에 사용되던 GPU와 유사한 구조이나 인공지능 기술에 특화하여 연산 효율을 높이거나 에너지 효율을 증가시키

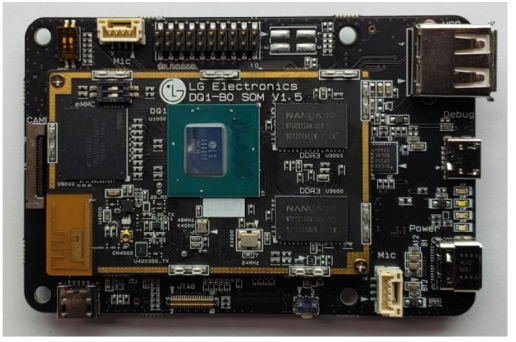


엑시노스 9820

엑시노스 2100

[그림 3] 삼성전자 NPU를 탑재한 엑시노스 AP 시리즈

LG8111 AI 칩



레퍼런스보드 Eris

[그림 4] LG의 가전 타겟 AI 칩 및 보드

기 위해 개발한 칩이다. 삼성전자는 NPU를 독자적으로 설계하고 개발하는 것에 많은 노력을 기울이고 있는데, 2018년 11월 공개한 엑시노스 9820에 처음으로 NPU를 탑재했다. 2019년 후속작인 엑시노스 9825에 이어 2020년부터 엑시노스 10시리즈(엑시노스 1080, 엑시노스 2100)에 모두 NPU가 탑재되고 있다([그림 3] 참조).

엘지는 가전 등에서 온디바이스 AI를 수행할 수 있도록 LG8111 AI SoC 솔루션을 개발했다. LG8111 칩은 엘지가 독자적으로 개발했으며, 비전 관련 작업 가속기 및 음성 관련 작업 가속기를 탑재하여 가전에서 필요한 영상 및 음성 AI 분석을 지원한다. 해당 칩을 내장한 레퍼런스보드 Eris도 제공하여 원활한 개발환경 구축을 지원한다([그림 4] 참조).

이 외에도 인포웍스, 아이원랩, 이드웨어에서 지능형 서비스를 디바이스에서 제공하기 위한 하드웨어 솔루션을 개발했다. 인포웍스는 이미지 센싱, AR 기반 안전 용품, LiDAR 등에 특화된 솔루션을 개발했으며, 아이원랩은 드론 운영 관련 온디바이스 AI를 수행하기 위한 모듈을 개발했다. 이드웨어는 음성인식을 위한 하드웨어 모듈을 개발했다.

학계에서는 카이스트 유희준 교수팀이 모바일 기기에서 심층 강화학습 수행을 지원하는 OmniDRL을 2021년 개발했다. 연구팀에 따르면 종래 기술 대비 전력 효율이 2.4배이며 성능 또한 우수하다. 한국전자통신연구원에서는 자율주행 등의 분야에 사용 가능한 고성능 가속기인 AB9와 저전력으로 비전 관련 AI 수행을 위한 VIC(Visual Intelligence Chip)를 개발했다.

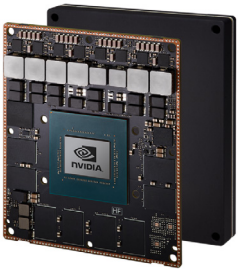
2. 국외 기술개발 동향

애플은 얼굴인식, 음성인식으로 대표되는 자사의 인공지능 서비스 품질 향상을 위해 뉴럴 엔진이라는 NPU를 독자적으로 개발하여 AP(Application Processor)와 통합된 형태로 제품을 출시하여 왔으며, XNOR를 인수하는 등 경쟁력을 지속적으로 강화하고 있다. 2017년 A11 SoC에 초당 6,000억 개의 명령을 수행하는 뉴럴 엔진 탑재를 시작으로, A12 SoC는 초당 5조 개의 명령을 수행하는 8코어 기반 뉴럴 엔진을 탑재하고, A13에서는 성능이 20% 향상된 뉴럴 엔진을 탑재했다. 특히, A13에서는 이종 멀티코어 CPU 구조에서 고성능 CPU 코어 2개를 묶어 머신 러닝 가속을 가능하게 하는 AMX 블록을 개발했다. A14에서는 16코어 기반 뉴럴 엔진을 최초로 탑재했으며, A15에서는 뉴럴 엔진 성능을 43.6% 향상시켰다.

퀄컴은 NPU 개발을 비교적 일찍부터 시작했는데, 2013년 독자적인 NPU인 제로스(Zeroth)를 개발하였다. 당초 2017년 스냅드래곤 820 AP부터 NPU를 탑재할 계획이었으나, 전용 가속기의 효용성에 대한 의문을 품은 퀄컴은 Hexagon이라는 DSP(Digital Signal Processing)를 탑재하고 CPU 및 GPU와 연계하여 AI 연산을 수행하기로 결정하였다. 이후 2021년 현재 스냅드래곤 888 AP까지 별도 NPU는 탑재하지 않고 있다. 이로 인해 순수 AI 처리 성능은 높으나 CPU 및 GPU가 많이 관여함으로써 발열 등의 문제가 발생하고 있다.

엔비디아는 AI 가속을 위한 Volta GPU 아키텍처를 기반으로 범용 GPU인 GPGPU를 개발하여 AI 하드웨어 기술을 선도하고 있으며, 최근 온디바이스 AI를 수행하기 위한 다양한 하드웨어 솔루션을 출시하고 있다. 자율주행을 위한 솔루션으로는 Drive AGX 시리즈를 출시했는데, 자율주행 레벨에 따라 각기 다른 성능의 솔루션을 제공한다. 뿐만 아니라 2022

Drive Orin



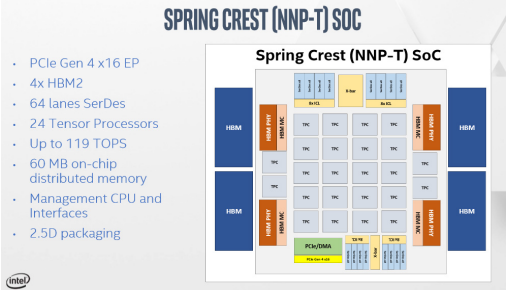
Jetson Xavier AGX

[그림 5] 엔비디아의 임베디드 환경 타겟 온디바이스 AI 솔루션

년에는 자율주행 전용의 SoC인 Drive Orin 출시를 계획하고 있다([그림 5] 참조). 임베디드 환경에서의 온디바이스 AI를 위한 Jetson 시리즈는 다양한 성능의 디바이스를 지원하기 위해 128 코어 Maxwell GPU를 탑재한 Nano부터 64개의 Tensor 코어가 장착된 512 코어 Volta GPU를 탑재한 AGX Xavier까지 4가지 성능의 솔루션을 제공하고 있다([그림 5] 참조). 2018년부터 엔비디아는 생태계 확장 및 기술개발 참여 유도를 위해 AI 추론 가속기 아키텍처를 NVDLA(NVIDIA Deep Learning Accelerator)라는 오픈소스로 공유하고 있다. NVDLA는 Xavier AGX SoC를 기반으로 설계되었으며 모듈러 구조를 통해 디바이스에서 클라우드까지 폭넓은 확장성을 제공한다.

인텔은 서버 환경에서의 AI 가속을 위한 제온 프로세스 등 고성능 CPU 기술 개발 경험을 토대로 신경망 프로세서 및 비전 처리 장치를 개발함으로써 점차적으로 온디바이스 AI 수행을 위한 하드웨어 기술로 확장하고 있다. 인텔은 NPU 대신 NNP(Neural Network Processor)라는 용어를 사용하며, 독자적으로 Nervana NNP 시리즈를 개발하여 2019년 AI 모델 훈련과 추론에 각각 특화된 NNP-T 및 NNP-I를 출시했다([그림 6] 참조). 또한, 디바이스 환경에서 영상처리 관련 AI 수행에 최적화된 Movidius Myriad X VPU(Vision Processing Unit)를 USB 기반 모듈 형태로 만든 뉴럴 컴퓨트 스틱을 출시했다. 이 외 칩 당 13만 개 뉴런과 1억 3,000만 개 시냅스로 구성된 뉴로모픽 칩 로이히 1을 2017년 개발했으며, 2021년 전력 효율과 성능을 개선한 로이히 2를 공개했다.

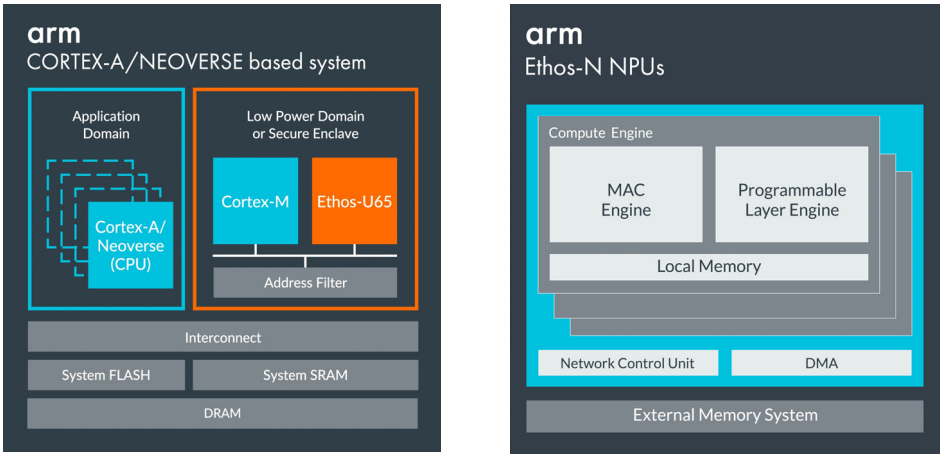
테슬라는 자사의 핵심 기능인 자율주행을 위해 NPU를 탑재한 자체 칩인 FSD(Full Self-Driving) 칩을 개발했다. 2016년 자율주행을 위한 칩의 필요성이 대두된 이후 2019년 테슬라 모델 3에 적용되기 시작했다. FSD 칩에 탑재된 NPU는 테슬라가 독자적으로 개발하였으



NNP-T

뉴럴 컴퓨트 스틱2

[그림 6] 인텔의 온디바이스 AI를 위한 신경망 프로세서



Ethos-U65 Ethos-N78

[그림 7] Arm의 온디바이스 AI를 위한 microNPU 제품군

며, 행렬 연산에 최적화된 직관적인 구조로 설계되었다.

Arm은 온디바이스 AI를 위한 microNPU 제품군을 개발했는데, 클라우드부터 엣지에서 사용 가능한 고사양의 Ethos-U 시리즈와 디바이스에서 사용 가능한 저전력의 Ethos-N 시리즈를 출시했다([그림 7] 참조). 2020년 초에 Ethos-U55를, 2020년 말에 보다 향상된 성능의 Ethos-U65를 출시했다. Ethos-N 시리즈는 Ethos-N37부터 Ethos-N78까지 4개 모델이 있으며 각각 성능이 달라 목적에 따른 선택지를 제공한다.

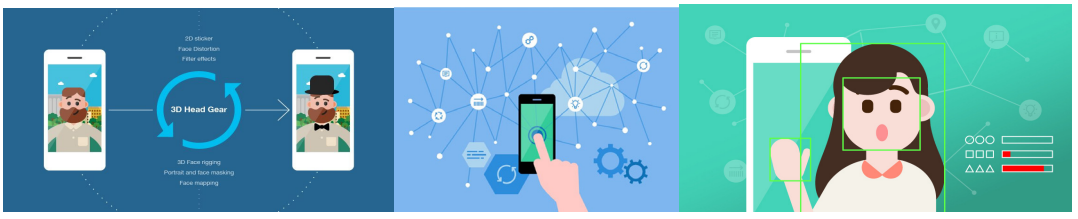
구글은 엣지 디바이스 환경 온디바이스 AI 수행을 위한 엣지 TPU를 개발했는데, 주요 특징으로 작은 사이즈, 저전력, 고성능 AI 추론 능력을 지원한다. 이를 기반으로 개발보드, USB/Mini PCIe형 엣지 TPU 등으로 이루어진 Coral 개발용 키트 제품군을 제공한다.

III. 온디바이스 AI 소프트웨어 기술개발 동향

온디바이스 AI 소프트웨어 기술은 서버와 연결 없이 단말에서 AI 연산을 수행함으로써 개인정보 보호 및 반응성 향상을 위한 기술이다. 기술개발 동향은 크게 2가지 방향으로 구분되는데, 온디바이스 환경 전용 경량화된 AI 모델 및 추론 기술 개발과 기존 AI 모델을 경량화하는 기술 개발로 나뉜다.

1. 국내 기술 동향

하이퍼커넥트는 모바일 기반 영상 채팅인 Azar를 서비스하고 있으며 이를 위해 다양한 온디바이스 AI 기술을 개발하여 실제 Azar에 적용하고 있다([그림 8] 참조). 영상 채팅 중 실시간으로 AR 기술을 도입하기 위해 영상 처리를 수행하는 모바일 증강현실 그래픽 플랫폼 HyperGraphics를 개발했다. 또 다른 기술인 HyperML은 딥러닝 기반 이미지 처리를 위한 플랫폼으로서 모바일 환경에 최적화된 데이터 처리 및 학습을 위한 파이프라인을 탑재한 추론 엔진으로, 현재 환경 감지, 부적절한 이미지 필터링 등을 모바일 디바이스 레벨에서 수행한다.



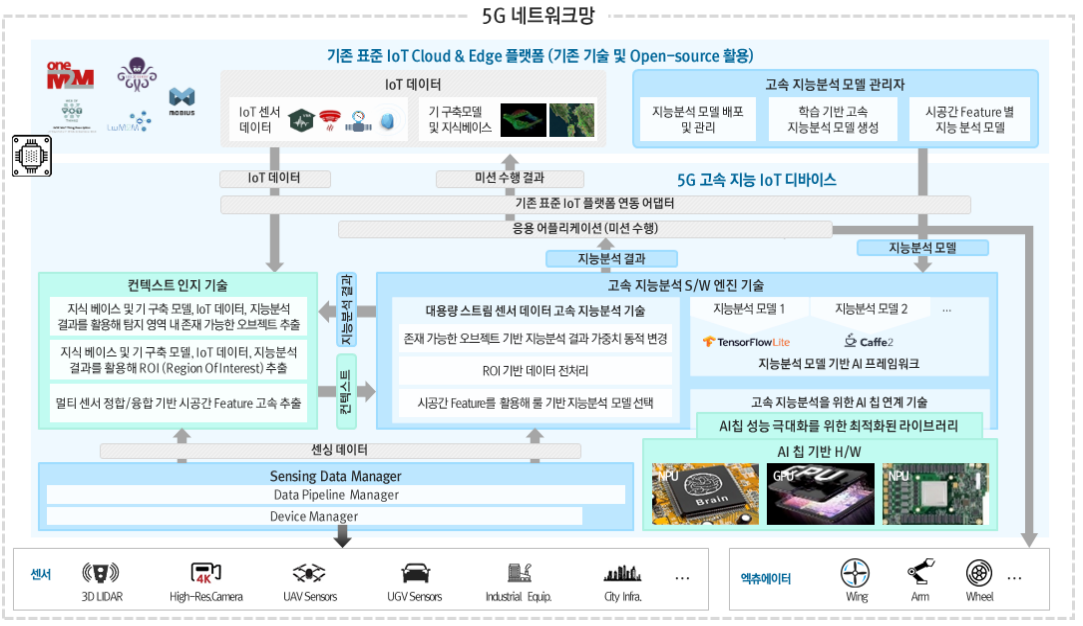
모바일용 AR　　온디바이스 AI 추론 엔진　　모바일용 이미지 인식

[그림 8] 하이퍼커넥트의 모바일 기반 영상채팅 Azar에 탑재된 온디바이스 AI 기술

삼성리서치는 리눅스의 영상 파이프라인 프레임워크인 GStreamer를 기반으로 뉴럴 네트워크 파이프라인 프레임워크인 NNStreamer를 오픈소스로 개발하여 온디바이스 환경에서 실시간 스트림 데이터에 대해 뉴럴 네트워크 파이프라인 처리를 지원한다.

노타는 AI 모델을 압축하여 연산량을 줄이면서 성능은 유지하는 NetsPresso 프레임워크를 개발했다. 가지치기 기법, 양자화 기법 등 다양한 압축방법을 사용하며, 가장 큰 특징은 실제 AI 모델이 수행될 디바이스 환경에서 반복적으로 테스트를 하여 자동으로 최적의 압축 방법을 찾는 것이다. 노타는 네이버의 투자를 받고 있으며 100억 원 이상의 투자를 유치했다.

한국전자기술연구원은 2020년부터 드론 및 로봇에서 4K 카메라, LiDAR등 대용량 스트림 센서를 사용하는 환경에서 온디바이스 AI를 수행하기 위한 고속지능분석 시스템을 개발하고 있다([그림 9] 참조). 디바이스의 주변 상황을 장면이해 방법을 기반으로 인식한 다음, 상황에 맞는 최적의 경량화된 AI 모델을 수행함으로써 범용적인 AI 모델을 수행하는 것에 비해 연산량을 감소하여 온디바이스 AI를 지원한다.



[그림 9] 한국전자기술연구원 고속지능분석시스템 구조도

2. 해외 기술 동향

온디바이스 AI 수행을 위해 [표 1과] 같이 다양한 오픈소스 경량 AI 모델들과 경량화 기술들이 연구 및 개발되고 있다[6],[7]. 모델 구조 변경, 합성곱 필터 변경, 자동 모델 탐색 기법

[표 1] 온디바이스 AI 소프트웨어 기술 동향

분류	접근방법	방향
경량 AI 모델	모델 구조 변경	잔여 블록, 병목 구조, 밀집 블록 등 다양한 신규 계층 구조를 이용하여 파라미터 축소 및 모델 성능을 개선(ResNet, DenseNet, SqueezeNet)
	합성곱 필터 변경	합성곱 신경망의 가장 큰 계산량을 요구하는 합성곱 필터의 연산을 효율적으로 감소(MobileNet, ShuffleNet)
	자동 모델 탐색	특정 요소(지연시간, 에너지 소모 등)가 주어진 경우, 강화 학습을 통해 최적 모델을 자동 탐색(NetAdapt, MNasNet)
AI 모델 경량화 기술	모델 압축	가중치 가지치기, 양자화/이진화, 가중치 공유 기법을 통해 파라미터의 불필요한 표현력을 감소(Deep Compression, XNOR-Net)
	모델 압축 자동 탐색	알고리즘 경량화 연구 중 일반적인 모델 압축 기법을 적용한 강화 학습 기반의 최적 모델 자동 탐색 연구(PocketFlow, AMC)

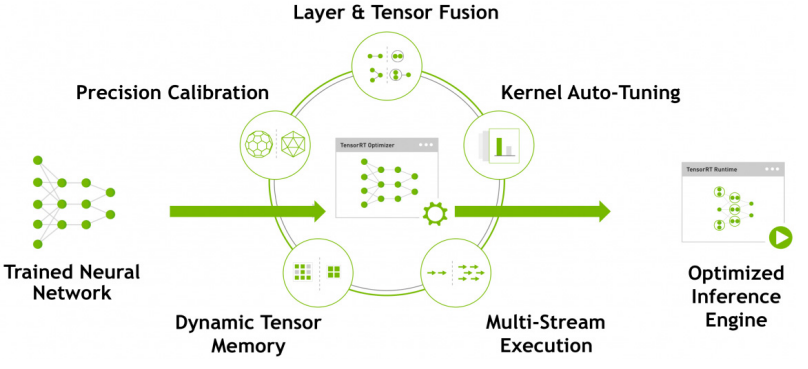
〈자료〉 이용주, 문용혁, 박준용, 민옥기, “경량 딥러닝 기술 동향”, ETRI, 전자통신동향분석 제34권 제2호, 2019, pp.40-50. 발췌

은 온디바이스 환경에 최적화된 경량 AI 모델로서, 기존 AI 모델과 유사한 성능을 유지하면서 크기가 최소화된 새로운 모델을 개발하는 것을 목표로 한다. YOLO, MobiNet, SSD 등 경량 합성곱 모델 위주로 오픈소스가 많이 존재하며, 일반적으로 입력들의 합성곱을 전부 구하는 것이 아니라 입력 값을 작은 그리드로 나눈 다음, 각 그리드의 합성곱들을 재귀적으로 구하는 방식으로 경량화하고 있다. 모델 압축, 모델 압축 자동 탐색 기법은 기존 AI 모델을 압축하여 크기를 줄이는 것을 목표로 한다.

구글은 경량 AI 플랫폼 개발을 통해 온디바이스 AI 기술을 선도하고 있다. 경량 딥러닝 네트워크 추론 엔진인 텐서플로라이트를 개발해 온디바이스 환경에서 리소스 사용을 최소화하면서 AI 추론을 수행할 수 있도록 했다. 또한, 연합학습 프레임워크를 개발하여 디바이스 레벨에서 머신러닝 학습을 수행하고 그 학습된 모델만 클라우드로 공유함으로써 개인정보를 클라우드로 보내지 않으면서 대규모 데이터 기반 머신러닝 학습을 가능하게 만들었다. 이외에 바코드 스캔, 얼굴 인식, 이미지 라벨링 등 비전 AI 관련 기능과 자연어 처리 기능을 제공하는 ML Kit을 개발했다. 구글 딥마인드는 압축 센싱(Compressed sensing) 프레임워크에서 딥 뉴럴 네트워크를 훈련하는 방법을 제시하였으며 메타-러닝 복원처리를 통해 기존 방식의 처리속도를 개선했다.

애플은 자사의 하드웨어 환경에서 애플의 뉴럴 엔진을 활용하여 온디바이스 AI를 수행할 수 있도록 Core ML 라이브러리를 제공한다. 이 라이브러리는 이미지, 비디오, 사운드 등 미디어 분석을 위해 설계된 첨단 신경망과 같은 AI 모델을 지원한다. 또한, 텐서플로나 PyTorch와 같은 라이브러리의 모델을 Core ML로 변환하는 기능도 제공한다.

엔비디아는 자사의 다양한 AI 하드웨어 기술을 기반으로 이들을 온디바이스 환경에서 활

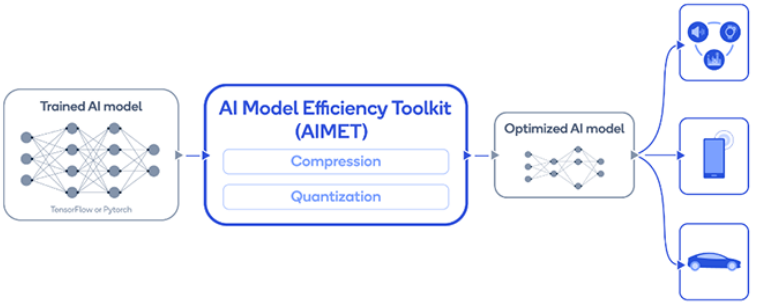


[그림 10] 엔비디아 TensorRT

용할 수 있는 플랫폼인 TensorRT를 개발했는데, 이는 고성능 딥러닝 추론을 위한 SDK로 임베디드 환경에서 학습된 모델을 기반으로 실시간 추론 및 실시간 모델 업데이트 기능을 제공한다([그림 10] 참조).

Arm은 자사의 하드웨어에 최적화하여 온디바이스 환경 비전 및 AI 분석을 지원하는 라이브러리 ACL(Arm Compute Library)를 제공한다. 사용성을 높이기 위해 핵심 기술인 코어 라이브러리와 분석이 수행될 디바이스 별 최적화가 적용된 런타임 라이브러리로 구분하여 오픈소스로 제공하고 있다.

퀄컴은 모바일에서 자주 사용되는 비전 분석 기능을 최적화한 FastCV 라이브러리를 개발했다. 해당 라이브러리는 범용 Arm 프로세서와 자사 SoC 대상의 라이브러리로 제공되는데, 자사 SoC 대상의 라이브러리는 DSP 기능을 포함한 환경도 지원한다. 또한, VeNum이라고 하는 멀티미디어 처리 엔진을 통해 고성능 실수 연산을 지원한다. 이 외에 스냅드래곤 AP에서 고효율 AI 모델 처리를 지원하는 Neural Processing SDK와 AI 모델 압축을 수행하여 온디바이스 AI를 지원하는 AIMET(AI Model Efficiency Toolkit)을 제공한다([그림 11] 참조).



[그림 11] 퀄컴 AIMET

IV. 결론

온디바이스 AI 하드웨어 및 소프트웨어 기술은 AI 분석에 요구되는 데이터 양이 폭증함에 따라 신속하고 원활한 서비스 제공을 위한 유망 기술로 주목받고 있다. 국내 대기업을 포함하여 글로벌 대기업들이 앞다투어 온디바이스 AI 수행을 지원하기 위한 하드웨어 및 소프트

웨어 기술을 선점하기 위해 노력하고 있다. 디바이스의 센서, 배터리, 연산장치 등의 관련 기술들이 나날이 발전함에 따라 온디바이스 AI의 실현도 가까워지고 있다. 실제로 스마트폰과 같은 비교적 강력한 연산능력을 갖춘 디바이스에서는 기기 자체적으로 AI 모델을 수행하고, 심지어 연합학습과 같은 기법을 통해 부분적이지만 학습까지 수행한다. 이러한 추세로 볼 때 머지않아 모든 디바이스에서 AI 모델을 수행할 정도로 기술이 성숙할 것으로 예상된다.

온디바이스 AI의 최종 목적지는 지능형 반도체 기술과 경량형 AI 기술의 발전을 통해 충분한 저전력화 · 고효율화 · 소형화를 기반으로 모든 지능형 사물들이 자체적으로 고수준의 AI 분석을 수행하는 단계로 나아가는 것이다. 자동차, 로봇, 드론, 스마트폰 등 충분한 컴퓨팅 파워를 갖춘 고가의 디바이스뿐만 아니라 CCTV, 인공지능 스피커부터 커피포트, 선풍기 등의 저가의 디바이스들도 고수준의 AI 서비스를 자체적으로 제공하게 되는 것이다. 이러한 실생활에 사용되는 주변 디바이스들이 더 이상 네트워크를 통해 수집한 데이터를 서버 또는 엣지로 전송하지 않으면서 자체적으로 AI 서비스를 수행하는 것은 개인정보 보호와 보안성 측면에서 큰 이점을 가져다줌과 동시에 보다 짧은 지연시간을 통해 다양한 실시간 서비스를 가능하게 한다. 또한, 네트워크 환경에 구애받지 않아 지능형 사물의 설치 및 운용이 훨씬 쉬워지며 재난 환경, 해양 환경 등과 같은 통신망 음영지역에서도 지능형 사물을 활용할 수 있다. 완성도가 높아진 온디바이스 AI 기술은 스마트시티, 스마트팜, 스마트팩토리 등 대규모 디바이스가 운영되는 차세대 ICT 융합기술의 실현에 핵심적인 역할을 할 것으로 기대된다.

참고문헌

[1] S. Hammoudi, Z. Aliouat, S. Harous, “Challenges and Research Directions for Internet of Things,” Telecommun. Syst., Vol.67(2), 2018, pp.367-385.

[2] IITP, “ICT R&D 중장기 기술로드맵 2022”, IITP, 2016.

[3] M. G. Sarwar Murshed, Christopher Murphy, Daqing Hou, Nazar Khan, Ganesh Ananthanarayanan, and Faraz Hussain, “Machine Learning at the Network Edge: A Survey,” ACM Comput. Surv, Vol.54(8), 2021, pp.1-37.

[4] Bisong, Ekaba, “Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners,” Apress, 2019.

[5] Indiveri et al., “Neuromorphic silicon neuron circuits,” Frontiers in Neuroscience, Vol.5(73), 2011, pp.1-23.

[6] 이용주, 문용혁, 박준용, 민옥기, “경량 딥러닝 기술 동향”, ETRI, 전자통신동향분석 제34권 제2호, 2019, pp.40-50.

[7] Sauptik Dhar, Junyao Guo, Jiayi(Jason) Liu, Samarth Tripathi, Unmesh Kurup, and Mohak Shah, “A Survey of On-Device Machine Learning: An Algorithms and Learning Theory Perspective,” ACM Trans. Internet Things, Vol.2(3), 2021, pp.1-49.