AI(Claude3)가 작성한「Vision AI와 Edge AI 기술 동향과 Arm의 전략 」보고서

(2024.07.15.)

글쓴이 Claude 3(by Anthropic), 프롬프팅·편집 신동형(donghyung.shin@gmail.com)

#제가쓴거아닙니다.

#AI가작성했습니다.

Executive Summary

본 보고서는 Vision AI와 Edge AI 기술의 최신 동향과 Arm의 관련 전략을 분석한 것입니다. Vision AI는 이미지와 영상 데이터로부터 유의미한 정보를 추출하는 인공지능 기술로, 자율주행, 스마트 시티, 산업 자동화 등 다양한 분야에서 활용되고 있습니다. Edge AI는 클라우드가 아닌 엣지 단에서 AI 연산을 처리하는 새로운 컴퓨팅 패러다임으로, 데이터 폭증과 실시간성 및 보안성 요구 증대에 따라 중요성이 높아지고 있습니다.

Vision AI 분야에서는 DINOv2, SAM, Stable Diffusion 등 대규모 사전학습 모델의 등장으로 기술 혁신이 가속화되고 있습니다. Edge AI 분야에서는 모델 경량화 기술의 고도화와 함께 클라우드-엣지 협업 프레임워크에 대한 연구가 활발히 진행되고 있습니다.

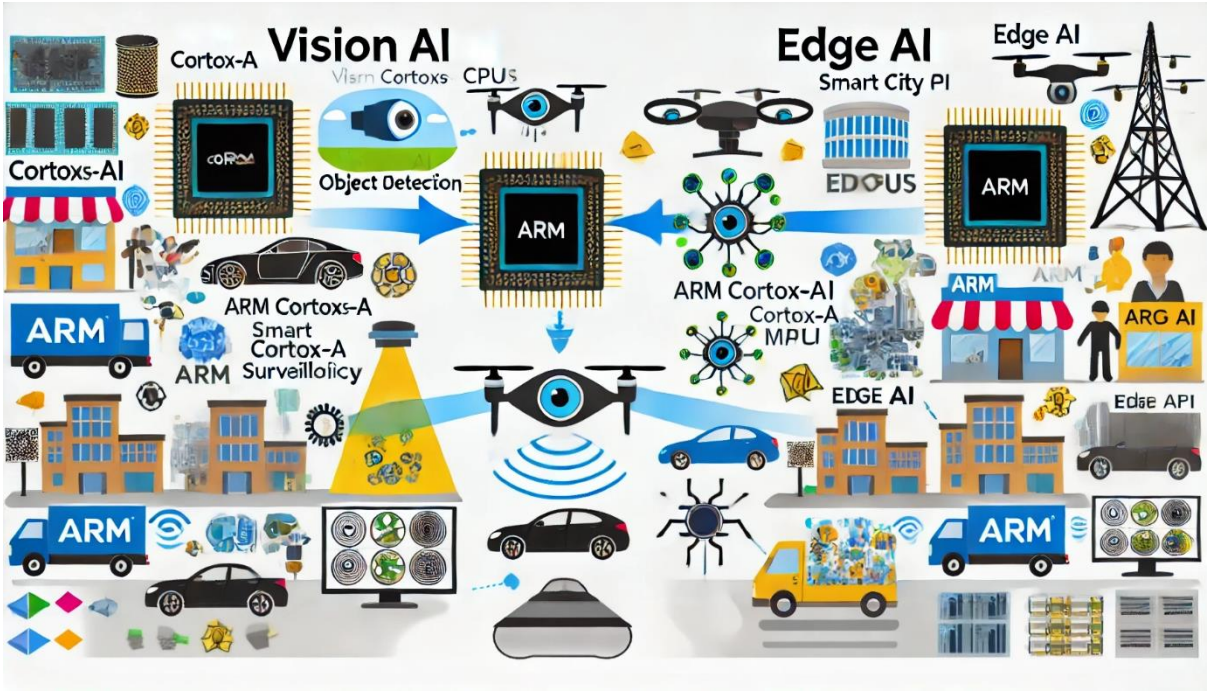
Arm은 Cortex-A, Mali, Ethos 등 고효율 IP 솔루션과 Arm NN, Arm Compute Library 등 개발자 친화적인 소프트웨어 툴을 제공함으로써, Vision AI 및 Edge AI 기술 발전을 가속화하고 있습니다. 업계 최대 규모의 개방형 생태계 역시 Arm의 강점으로, 수많은 파트너사들이 Arm 기반으로 혁신적인 제품과 서비스를 선보이고 있습니다.

그러나 Vision AI와 Edge AI의 발전 과정에서 윤리적, 법적, 기술적 도전 과제 또한 예상됩니다. 데이터 편향성 문제, 모델 설명 가능성 부족, 레이턴시와 배터리 소모 최적화, 데이터 프라이버시 및 기기 보안 등이 주요 이슈로 부각될 전망입니다. 이에 대응하기 위한 연구개발 노력과 함께, 정책적 대응 방안 마련도 요구됩니다.

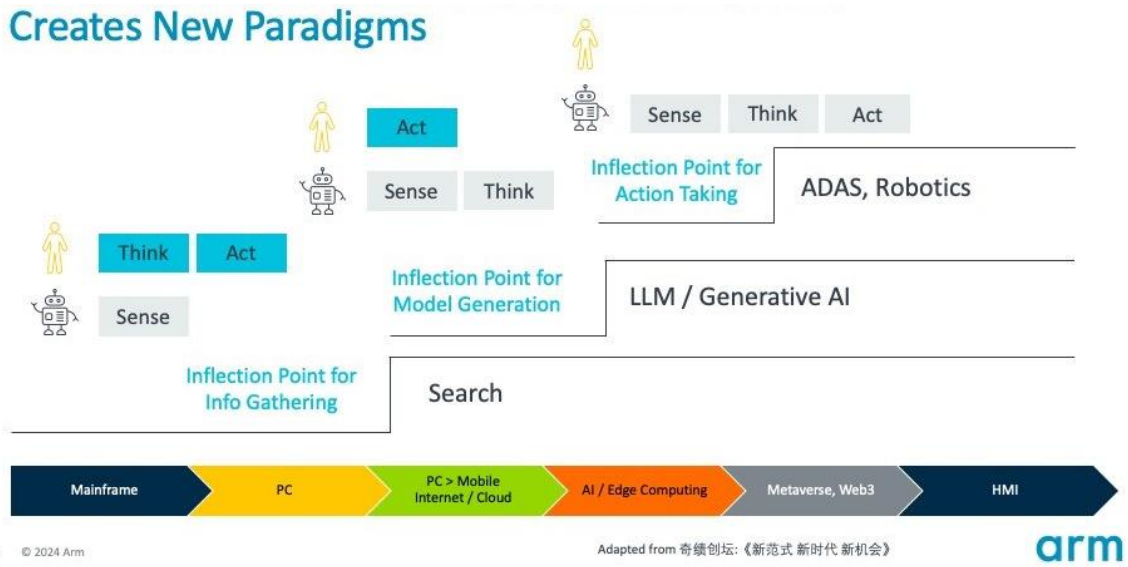
미래의 컴퓨팅은 사이버 공간과 물리적 공간의 경계가 허물어지는 방향으로 진화할 것입니다. 이러한 환경에서 Arm의 Vision AI, Edge AI 기술은 지능형 엣지 기기의 핵심 동력으로 자리매김하게 될 것이며, Arm은 다양한 도전 과제를 슬기롭게 극복하며 업계를 선도해 나갈 것으로 기대됩니다.

1. 서론

1.1. Vision AI와 Edge AI 기술의 개념과 중요성



인공지능(AI) 기술이 빠르게 발전하면서, 특히 이미지와 영상 데이터를 처리하는 Vision AI 분야가 주목받고 있습니다. Vision AI 는 컴퓨터 비전(Computer Vision) 기술과 딥러닝(Deep Learning)을 결합한 것으로, 카메라 등으로 수집한 시각 데이터에서 의미 있는 정보를 추출하고 이해하는 기술을 의미합니다.



예를 들어 자율주행차의 경우, 차량에 장착된 카메라가 도로 상황을 실시간으로 인식하고 이를 바탕으로 차량을 제어하는데, 이 때 Vision AI 기술이 핵심적인 역할을 수행합니다.

한편 Edge AI 는 중앙 집중식 클라우드 대신 데이터가 발생하는 현장(Edge)에서 AI 연산을 처리하는 기술로, 대표적으로 스마트폰, 드론, CCTV 등 엣지 디바이스에 AI 기능을 탑재하는 것을 의미합니다. Edge AI 를 활용할 경우 데이터 전송량을 줄일 수 있고, 실시간성과 보안성을 높일 수 있다는 장점이 있습니다.

Vision AI 와 Edge AI 는 밀접한 관련이 있습니다. **방대한 양의 이미지, 영상 데이터를 효과적으로 처리하기 위해서는 엣지 단에서의 데이터 전처리와 경량화된 추론(Inference)이 필수적**이기 때문입니다. 따라서 Vision AI 의 Edge 적용은 매우 중요한 기술적 과제라 할 수 있습니다.

1.2. Arm의 IP 포트폴리오와 Vision/Edge AI와의 관련성

Arm 은 모바일, 임베디드 기기용 반도체 설계 분야의 글로벌 리더로, 전 세계 스마트폰의 90% 이상이 Arm 아키텍처 기반으로 구동되고 있습니다. 스마트폰 AP 는 물론 자율주행, 드론 등 다양한 엣지 디바이스에서 Arm 아키텍처가 활용되고 있으며, Vision AI 및 Edge AI 구현에 있어서도 Arm IP 의 역할이 점차 확대되고 있습니다.

표 1 ARM IP 포트폴리오

Arm IP	**제품군**	**특징 및 활용 분야**
Cortex-A	고성능 CPU	어플리케이션 프로세서용, 스마트폰, 자율주행, 서버 등
Mali	GPU	그래픽 처리 및 AI 연산 가속, 스마트폰, 자율주행, VR/AR 등
Ethos	NPU	머신러닝 전용 프로세서, 모바일, 스마트홈, 스마트시티 등
Cortex-M	초저전력 CPU	IoT, 웨어러블 기기 등 소형 임베디드 시스템용

특히 최근 출시된 Cortex-X 시리즈 CPU, Mali-G78 GPU, Ethos-N78 NPU 등은 엣지 단에서의 고성능 AI 연산을 지원하는 IP 로, Vision AI 및 Edge AI 활용에 최적화되어 있습니다. 예를 들어 Ethos-N78 은 INT8 양자화를 통해 메모리 대역폭과 전력 소모를 최소화하면서도 높은 추론 성능을 제공합니다.

2. Vision AI 기술과 대규모 모델

2.1. Vision AI 기술의 특징과 구성 요소

Vision AI 기술은 크게 이미지 분류(Image Classification), 객체 검출(Object Detection), 의미론적 분할(Semantic Segmentation) 등으로 구분할 수 있습니다. 이미지 분류는 주어진 이미지가 어떤 범주에 속하는지 판단하는 작업이고, 객체 검출은 이미지 내의 관심 객체 위치를 찾아내는 것입니다. 의미론적 분할은 이미지의 모든 픽셀을 미리 정의된 클래스로 분류하는 기술로, 자율주행차의 주행 환경 이해 등에 활용됩니다.

이러한 Vision AI 모델을 학습시키기 위해서는 대규모의 이미지 데이터셋이 필요합니다. ImageNet, COCO, Pascal VOC 등이 대표적인 데이터셋이며, 데이터셋 확보와 레이블링(Labeling)에는 많은 비용과 시간이 소요됩니다. 최근에는 웹 크롤링을 통해 자동으로 대량의 이미지를 수집하고, 비지도 학습(Unsupervised Learning) 방식으로 레이블을 생성하는 사례도 증가하고 있습니다.

Vision AI 모델 학습에는 대규모 컴퓨팅 자원이 투입되는데, 병렬 처리에 유리한 GPU 가 주로 활용되며, Google TPU 와 같은 AI 가속기의 사용도 확산되는 추세입니다. 모델 경량화 및 엣지 단 적용을 위해 네트워크 프루닝(Network Pruning), 양자화(Quantization), 지식 증류(Knowledge Distillation) 등의 모델 최적화 기법이 연구되고 있습니다.

2.2. DINOv2, SAM, Stable Diffusion 등 대규모 Vision AI 모델 소개

최근 Vision AI 분야에서는 수억~수십억 장의 이미지로 사전학습된 대규모 모델들이 등장하고 있습니다. 2022 년 Meta AI 가 공개한 DINOv2(Vision Transformer)는 ImageNet 의 1.42 억 장 이미지로 자기지도학습된 모델로, 다양한 도메인의 이미지에 대해 우수한 성능을 보입니다. DINOv2 는 기존 CNN(Convolutional Neural Network) 기반 모델 대비 이미지 분류, 객체 검출, 실례 세분화(Instance Segmentation) 등의 태스크에서 뛰어난 성능을 나타냈습니다

메타(Meta)가 발표한 SAM(Segment Anything Model)은 무려 11 백만개의 이미지와 10 억 개 이상이 포함된 마스크를 이미지-마스크 쌍으로 학습된 Vision AI 모델입니다[1]. 이름 그대로 SAM 은 "어떤 객체든지 분할할 수 있는" 능력을 보유하고 있어, 객체 주위를 한 번만 클릭하면 쉽게 마스크를 생성할 수 있습니다. Pre-trained 모델을 새로운 분야에 바로 적용할 수 있는

[1] Input Augmentation with SAM: Boosting Medical Image Segmentation with Segmentation Foundation Model (Yizhe ZhangTao, 2023)

제로샷 전이(Zero-shot transfer)으로 다양한 도메인의 이미지에 적용 가능하다는 것이 SAM 의 가장 큰 특징입니다

스테이블 디퓨전(Stable Diffusion)은 스테이빌리티 AI(Stability AI)가 개발한 오픈소스 기반 Text-to-Image 모델입니다. Stable Diffusion 은 LAION-5B 라는 59 억 개의 텍스트-이미지 쌍 데이터셋으로 학습되었으며, 사용자의 텍스트 입력을 바탕으로 창의적인 이미지를 생성할 수 있다는 점에서 주목받았습니다. 기존의 DALL-E, Midjourney 와 같은 모델이 학습 데이터를 공개하지 않았던 반면, Stable Diffusion 은 오픈소스로 공개되어 누구나 사용 및 수정이 가능하다는 점에서 의미가 있습니다

이러한 최신 대규모 Vision AI 모델들은 아직까지는 대부분 서버 환경에서 구동되지만, 향후 On-device AI 시장 확대에 따라 모바일과 엣지 기기 환경에서의 구현도 증가할 것으로 전망됩니다.

2.3. 대규모 Vision AI 모델의 엣지 단 경량화 및 최적화 기법

대규모 Vision AI 모델을 엣지 단에서 구동하기 위해서는 모델 경량화와 최적화가 필수 과제입니다. 수백 MB 에서 수 GB 에 이르는 모델 용량과 수십 billion-flops 의 연산량을 갖는 모델을 제한된 메모리와 컴퓨팅 파워를 가진 모바일, 임베디드 기기에 탑재하기 위해서는 정교한 최적화 기술이 요구됩니다.

대표적인 모델 경량화 기법으로는 지식 증류(Knowledge Distillation), 네트워크 프루닝(Network Pruning), 양자화(Quantization) 등을 들 수 있습니다. 지식 증류는 대형 교사 모델(Teacher Model)이 작은 학생 모델(Student Model)에게 자신의 지식을 전달하여, 모델 규모를 줄이면서도 성능은 유지하는 방법입니다. BERT 의 경량화 버전인 DistilBERT 가 지식 증류의 대표적 사례입니다.

네트워크 프루닝(Network Pruning)은 학습된 신경망에서 중요도가 낮은 가중치(Weight)나 연결(Connection)을 제거하여 모델 규모를 축소하는 방식입니다. 가중치 프루닝(Weight Pruning)과 필터 프루닝(Filter Pruning) 등 다양한 프루닝 알고리즘이 제안되고 있습니다.

또한 기존 신경망에서 주로 사용되던 FP32 대신 FP16, INT8 등 낮은 비트 수로 양자화하여 모델 크기와 메모리 사용량을 대폭 감소시키는 방식도 널리 활용되고 있습니다. Arm 의 NPU IP 인 Ethos 시리즈는 INT8 양자화를 하드웨어적으로 지원함으로써 에너지 효율적인 모델 구동을 가능하게 합니다.

3. Edge AI 기술과 Arm의 IP 솔루션

3.1. Edge AI 개념과 등장 배경

Edge AI 는 클라우드가 아닌 현장(Edge)에 가까운 곳에서 AI 연산을 처리하는 새로운 컴퓨팅 패러다임입니다. 기존에는 대부분의 AI 모델이 학습과 추론 모두 클라우드 서버에서 이뤄졌으나, Edge AI 는 클라우드에서 학습된 모델을 엣지 기기로 배포하여 로컬에서 추론을 수행하는 방식입니다.

Edge AI 의 등장 배경에는 다음과 같은 요인들이 작용했습니다.

- 급증하는 엣지 데이터: IoT 기기의 확산으로 엣지 단에서 발생하는 데이터가 폭증하고 있으며, 모든 데이터를 클라우드로 전송하는 것은 현실적으로 어려워짐
- 실시간성 요구 증대: 자율주행, 로봇, AR/VR 등 실시간 대응이 필요한 애플리케이션이 확대되면서, 지연 시간을 최소화하기 위한 엣지 단 처리가 중요해짐
- 프라이버시와 보안: 카메라로 수집되는 개인 식별 정보나 민감 데이터를 클라우드로 전송하는 것은 프라이버시와 보안 문제를 야기할 수 있어, 필요한 정보만 엣지에서 추출하는 것이 바람직함
- 통신 비용 절감: 클라우드로의 데이터 전송에 따른 네트워크 비용이 만만치 않으므로, 데이터 전송량을 최소화할 수 있는 엣지 단 처리가 경제적 이점이 있음

Arm 은 이러한 Edge AI 트렌드에 선제적으로 대응하여 다양한 IP 솔루션을 제공하고 있습니다. Arm 의 저전력 및 고성능 IP 는 에지 디바이스에 최적화되어 있어 Edge AI 구현에 매력적인 옵션이 되고 있습니다.

3.2. Arm CPU, GPU, NPU IP 소개 (Cortex-A, Mali, Ethos 시리즈 등)

Arm 의 IP 포트폴리오는 크게 Cortex-A, Mali, Ethos 시리즈로 구성됩니다.

첫째, Cortex-A 시리즈는 스마트폰, 자율주행, 데이터 센터 등에 활용되는 고성능 CPU IP 제품군으로 최신 Cortex-X4 와 Coretex-A720 등이 있습니다.

- Cortex-X4: 2023 년 발표된 Cortex-X4 는 Armv9 아키텍처 기반의 초고성능 CPU IP 로, Cortex-X3 대비 15% 향상된 성능과 전력 효율성을 제공합니다. 클라우드 컴퓨팅, 고성능 모바일 기기, 자율주행 등 다양한 분야에서 활용될 것으로 기대됩니다.

- Cortex-A720: Cortex-A715의 후속 모델로, Armv9.2 아키텍처 기반의 고효율 CPU IP입니다. 전력 효율성이 20% 향상되었으며, 향상된 머신러닝 성능과 메모리 관리 기능을 제공합니다.
- 보안 강화: Arm은 모든 Cortex-A CPU에 TrustZone 기술을 탑재하여 안전한 실행 환경을 제공합니다. 또한, Confidential Computing 아키텍처를 통해 데이터 처리 중에도 데이터를 암호화하여 보안을 강화하고 있습니다.

둘째, Mali 시리즈는 모바일 기기에 가장 널리 채택된 GPU IP로 그래픽 성능과 전력 효율성이 특징입니다.

- Mali-G715: 2023년 발표된 Mali-G715는 Arm의 최신 GPU IP로, Valhall 아키텍처 기반의 향상된 성능과 전력 효율성을 제공합니다. 레이 트레이싱, 가변 레이트 쉐이딩 등 최신 그래픽 기술을 지원하며, 모바일 게임 및 XR(확장 현실) 경험을 더욱 풍부하게 만들어줍니다.
- AI 및 ML 가속: Mali GPU는 Arm NN SDK와의 통합을 통해 On-device ML 애플리케이션 개발을 지원합니다. 또한, 새로운 Mali-C81 NPU와의 연동을 통해 더욱 강력한 AI 및 ML 성능을 제공합니다.

셋째, Ethos 시리즈는 ARM의 머신러닝 전용 프로세서인 NPU용 IP입니다. 기존 CPU나 GPU에 비해서 에너지 효율이 좋습니다.

- Ethos-N79: 2023년 발표된 Ethos-N79는 Arm의 최신 NPU IP로, Ethos-N78 대비 2배 향상된 성능을 제공합니다. 8비트 부동소수점(FP8) 연산을 지원하여 더욱 정확하고 효율적인 AI 추론을 가능하게 합니다.
- 다양한 제품군: Ethos-U 시리즈는 마이크로 컨트롤러급 초저전력 디바이스를 위한 NPU 솔루션이며, Ethos-N 시리즈는 고성능 엣지 AI 애플리케이션을 위한 솔루션입니다. Arm은 다양한 Ethos 제품군을 통해 광범위한 엣지 AI 시장을 공략하고 있습니다.

표 2 ARM의 CPU, GPU, NPU 포트폴리오

IP 시리즈	주요 제품	특징
Cortex-A	Cortex-X2, Cortex-A78	고성능 애플리케이션 프로세서용 CPU

Mali	Mali-G78	최신 그래픽 기술 지원, AI 연산 가속
Ethos	Ethos-U65, Ethos-N78	초저전력 및 고성능 추론용 NPU

3.3. Arm 기반 Edge AI 구현 사례 (Axera, Hayden AI 등)

중국의 AI 칩 스타트업 Axera 는 Arm CPU, GPU, NPU 를 조합한 Edge AI 플랫폼인 AX620A 를 발표했습니다. AX620A 는 Arm Cortex-A 와 Ethos-N78 NPU 를 탑재하여 뛰어난 에너지 효율을 자랑하며, 얼굴 인식, 제스처 인식, 음성 인식 등 다양한 On-device ML 애플리케이션에 최적화된 성능을 제공합니다. Axera 는 DINOv2 Vision Transformer 모델을 AX620A 에서 에지 단 최적화하여 구동하는 데 성공했습니다.

미국의 자율주행 스타트업 Hayden AI 는 교통 감시 카메라에 Arm 기반 SoC 를 장착하고, 객체 검출 및 추적 등의 Vision AI 알고리즘을 카메라 자체에서 구동함으로써 교통 혼잡 및 사고 대응 효율을 높이는 솔루션을 제공합니다. Hayden AI 플랫폼에는 Arm Cortex-A 시리즈 프로세서가 탑재되어 있습니다.

4. Vision AI와 Edge AI 기술의 통합 및 활용 분야

4.1. 자율주행 및 스마트 모빌리티

Vision AI 와 Edge AI 기술은 자율주행 분야에서 핵심적인 역할을 수행합니다. 자율주행차는 카메라, 라이다, 레이더 등을 통해 주변 환경을 인식하고, 이를 바탕으로 실시간 주행 의사결정을 내려야 합니다. 차량에 장착된 고성능 컴퓨팅 플랫폼에서 Vision AI 알고리즘이 구동되어 도로, 차선, 표지판, 차량, 보행자 등을 인식하게 되는데, 이 때 Arm CPU, GPU, NPU 등이 활용됩니다.

자율주행 스타트업 모멘타(Momenta)는 Arm 기반 플랫폼을 활용하여 딥러닝 기반 주행 환경 인지 솔루션을 개발했습니다. 전방 카메라의 영상을 Arm CPU 와 GPU 를 활용해 분석하여 차선, 횡단보도 등을 검출하고, 이를 HD Map 과 결합하여 안전한 자율주행을 지원한다고 합니다.

4.2. 스마트 시티 및 공공 안전

CCTV 와 Edge AI 기술을 접목한 지능형 영상 감시 시스템은 스마트 시티와 공공 안전 분야의 킬러 애플리케이션으로 주목받고 있습니다. 기존 영상 감시 시스템은 대부분 사람이 CCTV 영상을 직접 모니터링하는 방식이었으나, Vision AI 기술을 활용하면 사건, 사고를 자동으로 감지하고 대응할 수 있습니다. 뿐만 아니라 영상 데이터를 분석해 혼잡도, 교통량 등 도시 현황 정보를 파악하는 데에도 활용됩니다.

Arm 의 고객사인 하이크비전(Hikvision)은 Arm CPU 및 NPU 를 탑재한 AI 카메라를 개발하여 에지 단에서 실시간 영상 분석 및 경고 알림이 가능한 솔루션을 제공합니다. 이를 활용해 투기 행위, 침입, 배회 등 의심 행위를 자동 감지하고 대응할 수 있습니다.

4.3. 산업 자동화 및 스마트 팩토리

산업 현장에서도 Vision AI 와 Edge AI 기술의 중요성이 높아지고 있습니다. 제조 공정에서의 불량품 검출, 설비 예지 보전, 작업자 안전 모니터링 등에 비전 및 센서 데이터 기반 AI 기술이 활발히 도입되고 있는데, 대부분의 경우 클라우드보다는 엣지 단에서 실시간으로 데이터를 처리하는 것이 요구됩니다.

Arm 기반 스마트 팩토리 솔루션 사례로는 화웨이의 엣지 컴퓨팅 플랫폼인 Atlas 를 들 수 있습니다. Atlas 는 Ascend AI 프로세서와 Arm CPU 를 결합한 ECS(Edge Computing Server)로, 공장에 설치된 카메라의 영상을 분석해 비정상 상황을 탐지하는 등의 기능을 제공합니다.

5. Arm의 Vision AI 및 Edge AI 생태계 전략

5.1. Arm NN, Arm Compute Library 등 소프트웨어 개발 툴

Arm 은 Vision AI 및 Edge AI 애플리케이션 개발을 위한 다양한 소프트웨어 툴과 라이브러리를 제공하고 있습니다. 대표적으로 Arm NN(Neural Network) SDK 는 Caffe, TensorFlow, TensorFlow Lite, ONNX 등 주요 딥러닝 프레임워크를 지원하는 크로스 플랫폼 추론 엔진입니다. Arm NN 은 Arm CPU, GPU, NPU 에 최적화되어 있어 에지 기기에서의 효율적인 ML 모델 배포를 가능하게 합니다.

Arm Compute Library 는 Arm CPU 와 Mali GPU 의 컴퓨팅 성능을 최대한 활용하기 위한 오픈소스 라이브러리입니다. 주로 이미지 처리, 컴퓨터 비전 관련 함수들로 구성되어 있으며, NEON SIMD 와 OpenCL 을 활용한 가속 기능을 제공합니다.

또한 Arm 은 파트너사들과 협력하여 ML 모델의 최적화 및 배포를 지원하는 Arm AI 파트너 프로그램을 운영 중입니다. X-ray 이미지 분석, 영상 기반 질병 진단 등 의료 AI 분야의 주요 기업들이 Arm AI 파트너로 참여하고 있습니다.

5.2. 파트너 협력 및 생태계 확장 전략

Arm 은 실리콘 파트너, 소프트웨어 파트너, OEM 파트너 등 폭넓은 협력 관계를 통해 Vision AI 및 Edge AI 생태계 확장을 추진하고 있습니다. 실리콘 파트너로는 삼성전자, 퀄컴, 화웨이, 엔비디아 등이 있으며, 이들은 Arm NPU 를 비롯한 Arm 의 AI 가속기를 활용한 Edge AI 칩을 개발하고 있습니다.

소프트웨어 파트너로는 아마존 웹 서비스(AWS), 마이크로소프트, 구글 클라우드 등 주요 클라우드 사업자들과 협력 관계를 맺고 있으며, Arm NN 등의 툴이 이들 클라우드 플랫폼과 호환되도록 지원하고 있습니다. 또한 Arm AI 생태계는 AI 스타트업들의 참여도 활발한데, Arm Innovator 프로그램을 통해 AI 분야 젊은 기업들에게 기술 지원을 제공하고 있습니다.

5.3. 오픈소스 커뮤니티와의 협력 (ONNX, TVM 등)

Arm 은 다양한 오픈소스 커뮤니티에도 적극적으로 참여하며 개방형 혁신을 추구하고 있습니다. 대표적으로 ONNX(Open Neural Network Exchange)는 서로 다른 딥러닝 프레임워크 간 모델 호환성을 확보하기 위한 오픈 포맷으로, Arm 은 ONNX 커뮤니티의 주요 멤버로 활동하며 Arm NN 과의 통합을 지속 추진하고 있습니다.

TVM(Tensor Virtual Machine)은 최근 많은 주목을 받고 있는 오픈소스 ML 컴파일러로, Arm Compute Library 와의 통합이 이뤄지고 있습니다. Arm 의 지원 하에 TVM 을 활용한 Arm CPU/GPU 백엔드 최적화가 이뤄지고 있습니다. 이는 개발자들이 TVM 컴파일러 스택을 통해 Arm 기반 엣지 디바이스에 손쉽게 Vision AI 모델을 배포할 수 있음을 의미합니다.

6. 결론 및 향후 전망

6.1. Vision AI와 Edge AI의 기술 발전 방향

Vision AI 기술은 향후 보다 적은 학습 데이터로 뛰어난 성능을 달성하는 방향으로 진화할 것입니다. 현재는 수천만~수억 장의 이미지를 활용한 대규모 사전학습 모델이 각광받고 있지만, 궁극적으로는 적은 수의 데이터만으로도 강건한 성능을 보장하는 few-shot learning 혹은 zero-shot learning 으로의 전환이 예상됩니다. 현실 세계의 long-tail 문제를 다루기 위해서는 대규모 데이터에 의존하는 현재의 학습 방식으로는 한계가 있기 때문입니다.

또한 Vision-Language 멀티모달 모델의 진화도 예상됩니다. 앞서 소개한 SAM 이나 Stable Diffusion 은 이미지에 대한 자연어 입력을 활용해 Vision 태스크의 정교함을 높일 수 있음을 보여주었는데, 향후에는 이미지, 텍스트를 넘어 음성, 동작 등 더욱 다양한 모달리티를 통합적으로 이해하는 모델의 등장이 예상됩니다.

Edge AI 분야에서는 클라우드-엣지 협업 학습 프레임 워크가 주목받을 것으로 보입니다. 현재의 Edge AI 는 대부분 클라우드에서 학습된 모델을 엣지 기기에 배포하는 추론 위주의 패러다임인데, 향후에는 엣지 기기가 로컬 데이터를 반영해 모델을 지속 학습하고 그 결과를 클라우드와 공유하는 방식의 연합학습(Federated Learning) 기술이 확산될 것입니다. 이는 개인정보 보호와 모델 커스터마이징 측면에서 이점이 있습니다.

기술적으로는 모델 경량화와 고속화를 위한 알고리즘 혁신이 지속될 것입니다. 지식 증류, 프루닝, 양자화 등 기존 경량화 기법의 고도화와 함께, 신경망 구조 자체를 효율화하려는 NAS(Neural Architecture Search) 연구도 활발해질 전망입니다. 아울러 차세대 메모리 기술과의 융합을 통해 메모리 병목 현상을 극복하려는 시도도 계속될 것입니다.

6.2. Arm의 역할과 비전

Arm 은 지금까지 모바일, 임베디드 시장에서의 리더십을 발판 삼아 Vision AI, Edge AI 분야에서도 선도적 역할을 수행해 왔습니다. NPU 제품군인 Ethos 시리즈를 통해 업계 최고 수준의 에너지 효율을 자랑하는 엣지 추론 가속기를 제공하고 있으며, CPU 와 GPU 역시 AI 워크로드에 빠르게 최적화되고 있습니다.

더불어 Arm NN, Arm Compute Library 등 개발자 친화적인 소프트웨어 툴을 오픈소스로 공개하고, 다양한 파트너사 및 개발 커뮤니티와의 협업을 통해 개방형 혁신의 비전을 실현하고 있습니다. 실제로 Arm 의 에코시스템 내에는 수많은 Vision/Edge AI 스타트업들이 포진해 있으며, 이들은 Arm 의 기술을 기반으로 혁신적인 제품과 서비스를 선보이고 있습니다.

미래의 컴퓨팅 환경은 클라우드와 엣지가 유기적으로 결합되고, 사이버 공간과 물리적 공간이 긴밀하게 연결되는 형태로 진화할 것입니다. 이러한 환경에서 중요해질 지능형 엣지 기기의 핵심에는 Arm 의 기술이 자리하게 될 것이며, Arm 은 Vision AI 와 Edge AI 의 혁신을 선도하는 핵심 동력으로서 그 역할을 다해갈 것입니다.

6.3. 잠재적 도전 과제 및 해결 방안

다만 Vision AI 와 Edge AI 의 발전 과정에서 몇 가지 도전 과제 또한 존재합니다. 먼저 대규모 사전학습 모델의 편향성 이슈는 지속적인 연구가 필요한 부분입니다. 학습 데이터의 편향으로 인해 특정 계층, 인종 등에 대한 차별적 결과를 초래할 수 있기 때문입니다. 이를 완화하기 위한 데이터셋 다양성 확보, 편향성 측정 기준 정립, 공정성 지표 개발 등이 요구됩니다.

또한 AI 모델의 설명 가능성(Explainability) 부족도 윤리적, 법적 이슈를 야기할 수 있습니다. 의료, 금융 등 결과에 대한 説明이 중요한 분야에서는 블랙박스 모델에 대한 규제가 강화될 가능성이 높습니다. 향후 XAI(Explainable AI) 기술 확보가 기업 경쟁력의 주요 요소로 부상할 전망입니다.

기술적으로는 Edge AI 의 레이턴시와 배터리 소모가 향후에도 지속적인 최적화 대상이 될 것입니다. 실시간성이 중요한 애플리케이션일수록 추론 지연 시간에 민감할 수밖에 없으며, IoT 기기의 배터리 수명을 고려할 때 Ultra-low Power 모델 처리 기술의 중요성은 더욱 커질 것입니다.

정책적으로는 데이터 프라이버시와 보안에 대한 규제 강화에 선제 대응할 필요가 있습니다. 특히 Vision AI 는 카메라를 통해 개인 식별 정보를 다루는 경우가 많아, GDPR, CCPA 등 각국의 개인정보보호법을 철저히 준수하는 것이 기업의 필수 과제가 될 것입니다. 나아가 Edge AI 시대에는 기기 자체에 대한 보안 강화도 요구되는데, 이 때 Arm 의 SoC 레벨 보안 기술이 핵심 솔루션으로 활용될 수 있을 것입니다.

참고자료

Advancing Smart Vision in the Large Model Era(MaChloe, 2024)

Superhuman Vision: AI Sees What You Can't | Akshat Dave | TEDxMIT (DaveAkshat, 2024)

Vision AI startup Hayden AI raises $90M to boost public transit efficiency (WHEATLEYMIKE, 2024)

#VisionAI, #비전 AI, #EdgeAI, #엣지 AI, #EdgeComputing, #엣지컴퓨팅, #Arm, #암, #ArmIP, #암 IP, #ArmCPU, #ArmGPU, #ArmNPU, # #CortexA, #Mali, #Ethos, #ArmNN, #ArmComputeLibrary, #DINOv2, #다이노 v2, #SAM, #SegmentAnythingModel, #세그먼트 애니씽 모델, #StableDiffusion, #스테이블디퓨전, #HaydenAI, #헤이든 AI, #Axera, #악세라, #ONNX, #TVM, #Federated Learning, #연합학습, #XAI, #ExplainableAI, #설명가능한 A

신동형의 AI로 작성한 보고서 시리즈

43. 20240712_AI(Gemini)가 작성한「AI for Good by ITU:지속가능한 발전을 위한 인공지능 혁신사례」보고서(LINK)

42. 20240711_AI(Claude3)가 작성한「AI for Good by ITU:지속가능한 발전을 위한 인공지능 혁신사례」보고서(LINK)

41. 20240711_AI(Claude3.5)가 작성한「갤럭시 언팩 2024」보고서(LINK)

40. 20240710_AI(Claude3)가 작성한「Arm 기술혁신; 미래를 향한 13가지돌파구」보고서(LINK)

39. 20240709_ AI(Claude3)가 작성한「Meta FAIR의 AI 연구 혁신:창의성, 효율성, 책임감의 조화로운 실현을 향하여」보고서(LINK)

38. 20240708_ AI(Claude3)가 작성한「 OpenAI 해킹 사태로 본 AI 기업의 보안 위협과 대응 전략 」보고서(LINK)

37. 20240705_AI(Claude3)가 작성한「Runway社의 Gen-3 Alpha 출시」보고서(LINK)

36. 20240704_AI(Claude3)가 작성한「Formation Bio: AI 기반 신약 개발」보고서(LINK)

35. 20240703_AI(Claude3)가 작성한「AI 평가 체계 대전환을 향한 앤트로픽의 도전」보고서(LINK)

34. 20240702_AI(Claude3)가 작성한「5G-A 시대의 개막, 화웨이의 비전과 전략」보고서(LINK)

33. 20240701_AI(Claude3)가 작성한「소셜 웹의 新패러다임, 페디버스가 열어갈 미래」보고서(LINK)

32. 20240628_ AI(Claude3)가 작성한「CriticGPT, 차세대 RLHF 위한 Human-AI 시너지」보고서(LINK)

31. 20240627_ AI(Claude3)가 작성한「Computex 2024에서 Top4 반도체 기업의 전략으로 살펴본 AI 시대의 반도체 산업 전망」보고서(LINK)

30. 20240626_ AI(Claude3)가 작성한「SLAM 기술: 공간 지능의 핵심 동력」보고서(LINK)

29. 20240625_AI(Claude3)가 작성한「EU의 AI 규제 강화와 빅테크의 대응:Meta와 Apple 중심으로」보고서(LINK)

28. 20240624_AI(Claude3)가 작성한「Intel의 AI 시대 도전과 전략」보고서(LINK)

27. 20240621_AI(Claude3)가 작성한「Claude 3.5 Sonnet: AI의 새로운 지평을 열다」보고서(LINK)

26. 20240620_ AI(Claude3)가 작성한「인공지능의 새로운 도약, 3D 공간 지능(Spatial Intelligence)의 부상) 」보고서(LINK)

25. 20240619_ AI(Claude3)가 작성한「Arm, AI 컴퓨팅의 미래를 향한 비상(飛上)」보고서(LINK)

24. 20240618_ AI(Claude3)가 작성한「AMD, AI 시대 컴퓨팅 혁신으로 지능화 가속화」보고서(LINK)

23. 20240617_ AI(Claude3)가 작성한「Apple의 차별화된 AI 전략」보고서(LINK)

22. 20240614_ 2024 컴퓨텍스 기조연설로 본 엔비디아의 미래 비전과 전략,「엔비디아, AI 시대를 이끄는 '게임 체인저'로 부상」(LINK)

21. 20240613_ AI(Claude3)가 작성한「AI PC 시대의 도래: 기술 혁신, 산업 생태계 변화」보고서(LINK)

20. 20240612_ AI(Claude3)가 작성한「대규모 언어 모델(LLM), 이렇게 생각하고 배웁니다」보고서(LINK)

19. 20240611_AI(Claude3)가 작성한「WWDC2024 애플 개인맞춤형 지능 기술로 새로운 미래 제시」 보고서(LINK)

18. 20240517_AI(Claude3)가 작성한 빅테크 기업 AI 전략 비교 분석 보고서[MS & OpenAI vs. Google vs. Meta의 AI 기술 동향과 미래 전망](LINK)

17. 20240515_AI(Claude3)가 작성한 Google I/O 2024 보고서, AI 혁신으로 만드는 더 나은 미래(LINK)

16. 20240514_AI(Claude3)가 작성한, OpenAI의 GPT-4o 공개, 멀티 모달 AI 혁명의 신호탄(LINK)

15. 20240425_AI(Claude3)가 작성한 메타의 스마트 글래스: AI Vision으로 세상을 바꿉니다(LINK)

14. 20240425_AI(Claude3)가 작성한 보고서, 온디바이스 AI 시대의 도래: Phi-3와 Llama-3이 가져올 변화와 영향(LINK)

13. 20240424_AI(Claude3)가 작성한 보고서: 경량 AI 시대의 개막, Microsoft의 Phi-3가 가져올 산업 혁신과 AI 대중화(LINK)

12. 20240423_AI(Claude3)가 작성한 메타플랫폼의 XR 생태계 新 전략(LINK)

11. 20240421_AI(Claude3)가 작성한 초등학생도 이해하는 LLAMA3과 On-Device AI 시대 도래

(LINK)

10. 20240419_AI(Claude3)이 작성한 초등학생도 이해하는 라마3(LLAMA3) 출시와 전망 보고서 (LINK)

9. 20240419_AI(Claude3)이 정리 작성한 초등학생도 이해하는 프롬프팅 프레임워크 설명(LINK)

8. 20240412_AI(Claude3)가 작성한 인텔, AI 시대를 선도하는 기술 혁신과 비전(LINK)

7. 20240408_AI(Claude3)가 작성한 2024년 중국 AI LLM 산업 발전 보고서 정리(LINK)

6. 20240408_AI(Claude3)가 작성한 Embodied AI: 현황, 전망, 그리고 미래(LINK)

5. 20240403_AI(Claude3)가 작성한 반도체 유리기판 공급망 분석 보고서 (전자신문 기획기사 참조)(LINK)

4. 20240401_AI(Claude3)가 작성한 빅테크 기업들의 AI 전략 비교 분석 보고서(LINK)

3. 20240326_AI(Claude)가 쓴 애플의 현재 AI 전략에 대한 회고: 글로벌과 개인정보보호 관점(긍정적)(LINK)

2. 20240322_AI(Claude3)가 작성한 엔비디아 파트너로서의 삼성전자: 파운드리와 HBM 사업을 중심으로(LINK)

1. 20240320_AI(Claude3)가 작성한 엔비디아 젠슨 황 CEO의 'GTC 2024' 기조연설 리뷰(LINK)