엣지 디바이스에서 인공지능 응용 서비스 실행에 관한 연구

홍태철, 전승협
한국전자통신연구원

taechori@etri.re.kr, shjeon00@etri.re.kr

A Study on the Running Artificial Intelligence Application Service Execution on Edge Devices

Hong Tae Chul, Jeon Seung Hyup
ETRI

요 약

인공지능 특히 딥러닝의 경우 좋은 성능을 위해서는 많은 연산 능력을 필요로 하여 대부분 클라우드 기반으로 활용되었으나, 지연시간, 사생활 이슈, 전용 하드웨어 등장 등의 다양한 요인으로 인해 점차 엣지 디바이스에서 직접 활용하는 방향으로 발전하고 있다. 본 논문에서는 엣지 디바이스에서 인공지능 응용을 적용하기 위한 다양한 접근 방법에 대한 분석하였다.

Ⅰ. 서론

엣지 디바이스의 경우 기존 인공지능(또는 기계학습) 모델을 개발하고 실행하던 클라우드 환경에 비해 전력소모, 연산능력, 메모리 등 다양한 제약사항으로 인하여 기존 활용하던 모델을 그대로 활용하기에 어려움이 있다. 따라서 그림 1 과 같이 크게 3 가지 접근 방식을 통해 엣지 디바이스에서 인공지능 응용을 적용하려고 한다. 첫번째 접근 방법(그림 1 - Lightweight ML Model)은 기존 활용되던 인공지능 모델을 최대한 성능저하 없이 경량화하고, 최적화하는 방법이다. 두번째 접근 방법(그림 1 - Hardware Performance Enhancement)은 저전력으로 인공지능 연산의 성능을 높일 수 있는 전용 하드웨어(인공지능 가속기)의 개발 및 활용이다. 마지막 세번째 접근방법(그림 1- Load Balancing)은 엣지 디바이스의 부족한 인공지능 연산 능력을 클라우드의 도움을 통해 해결하는 방법이다. 본 논문에서는 이러한 3 가지 접근 방법에 대해 분석하였다.

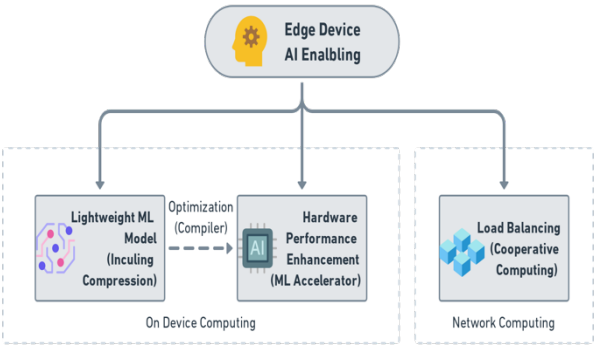

그림 1. 엣지 인공지능 응용을 위한 접근 방법

Ⅱ. 본론

기계학습 기반, 특히 딥러닝 기반 인공지능 방식의 경우 하드웨어의 성능 개선과 빅데이터로 인해 크게 성능이 개선되어 현재 주도적인 방식으로 활용되고 있다. 초기에는 CPU(그림 2 - ML-CPU)를 이용하여 활용되다가, GPU(그림 2 - ML-GPU)를 사용하게 되면서 인공지능의 성능이 크게 개선되었다. 그런데, 인공지능 모델 개발 및 활용에 컴퓨팅 자원 소모가 극심해지면서, 점차 에너지 효율적으로 많은 연산량을 지원하는 전용 하드웨어(그림 2 - ML-NPU(cloud))들이 개발되고 있다.

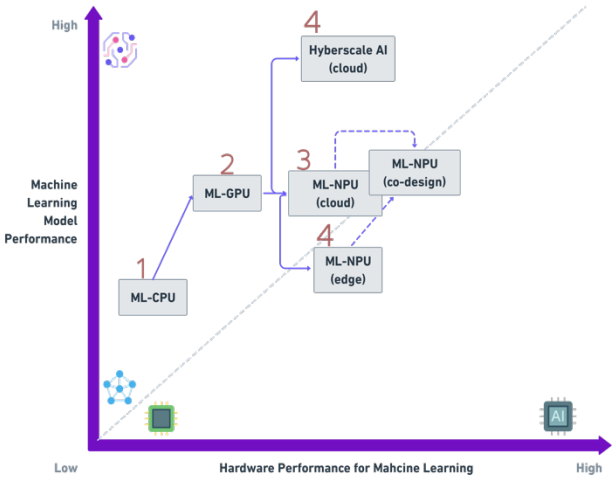

그림 2. ML 모델 성능과 HW 성능 연관 관계

이렇게 클라우드 기반에서도 많은 연산자원을 필요로 했던 인공지능 응용을 엣지 디바이스에서 실행하기

위해서, 기존의 모델의 성능을 일부 희생한 경량화를 수행하고 또한 엣지 디바이스 환경에서 저전력으로 높은 연산량을 제공할 수 있는 전용 가속기(NPU: Neural Processing Unit)를 활용(그림 2 - ML-NPU (edge))하고 있다. 엣지 디바이스에서 최적의 성능을 제공하기 위해서는 기존 모델의 단순 경량화뿐만 아니라 전용 가속기에 대한 최적화가 필요하다. 최적화의 경우 인공지능 모델 컴파일러와 또한 모델 생성 및 경량화 과정에 모두 반영할 수 있다 [1][2].

최종적으로는 엣지 디바이스에서 제공하려는 서비스에 최적화할 수 있도록 인공지능 모델과 하드웨어를 함께 설계하는 형태(그림 2 - ML-NPU(codesign))로 발전하고 있다[3]. 그림 2 는 여기서 설명한 인공지능 모델 성능과 하드웨어 성능이 연관되어 발전하는 것을 그래프 형태로 보여준다.

이렇게 인공지능 모델과 하드웨어 성능을 개선하더라도 현재 수준에서 엣지 디바이스에서 인공지능 응용을 활용하는데 연산능력이 부족할 수 있다. 이러한 경우에는 인공지능 모델을 분할하여 실행하는 부하분산 방법을 고려한다. 5G 를 이용하는 엣지 디바이스의 경우, 5G 네트워크에서 제공하는 Mobile Edge Computing (MEC)와 응용 서비스를 제공하는 클라우드를 활용하여 인공지능 연산을 분할하여 수행할 수 있다. 3GPP 에서도 이렇게 인공지능 모델을 분할하여 실행하는 모델 형태를 지원하기 위한 분석작업을 완료하였다[4]. 그림 3 은 5G 네트워크를 이용하여 딥러닝 모델 분할 실행의 예를 보여준다. 엣지 디바이스의 연산 능력이 부족한 경우 그림 3 에서와 같이 인공지능 모델을 분할하여, 일부분을 엣지 디바이스에서 수행하고 다른 부분들은 MEC 나 클라우드에서 실행하는 형태로 인공지능 응용 실행할 수 있다. 인공지능 모델 분할을 통해 인공지능 서비스를 제공하는 경우에는 5G 네트워크에서 전송속도와 지연시간 등을 보장할 수 있어야 한다.

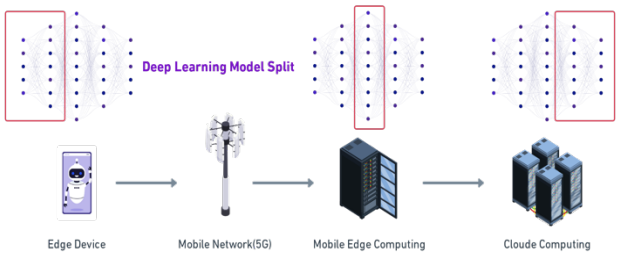

그림 3. 5G 네트워크를 이용한 딥러닝 모델 분할 실행

Ⅲ. 결론

본 논문에서는 엣지 디바이스에서 인공지능 서비스를 활용하기 위한 다양한 방법을 소개하였다. 향후 엣지 디바이스에서 다양한 인공지능 서비스를 제공하기 위해서는, 인공지능 모델 경량화와 하드웨어 최적화, 그리고 엣지 디바이스 환경에 적합한 저전력, 고성능 전용 하드웨어(인공지능 가속기) 개발이 필요하다. 또한 단순히 엣지 디바이스의 자원만으로 인공지능 서비스를 제공할 수 없는 경우도 발생할 수 있으므로, 5G 네트워크를 이용하여 MEC 와 클라우드의 연산 자원을 활용하는 방안도 함께 고려되어야 한다. 엣지 디바이스를 이용하여 다양한 인공지능 서비스를 제공하기 위해서는 본 논문에서 소개한 방법들을 상황에 맞게 결합하는 통합적 접근방법이 필요하기 때문에, 이러한 것을 쉽게 개발할 수 있도록 지원하는 플랫폼이 향후 중요해질 것으로 판단된다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00454, 스마트 엣지 디바이스 SW 개발 플랫폼 개발)

참 고 문 헌

[1] Mingzhen Li, et al, "The Deep Learning Compiler: A Comprehensive Survey", IEEE Transactions on Parallel & Distributed Systems, vol. 32, no. 03, pp. 708-727, 2021

[2] Hadjer, B., et al, "A Comprehensive Survey on Hardware-Aware Neural Architecture Search", International Joint Conference on Artificial Intelligence, pp. 4322-4329, 2021.

[3] Lukas, S., " Neural Architecture Search and Hardware Accelerator Co-Search: A Survey" IEEE Access, vol. 9, pp.151337-151362, Nov. 2021

[4] 3GPP TR22.874 v18.2.0, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS(Release 18), Dec. 2021