경로 탐색을 위한 FPGA 기반 MLP 가속 온디바이스 시스템 설계

백승원, 권은혜, 백돈규

충북대학교

seungwon@sdalab.cbnu.ac.kr, eunhye@sdalab.cbnu.ac.kr, donkyu@cbnu.ac.kr

On-Device System Design With FPGA-Based Acceleration of MLP for Path Finding

Seungwon Baek, Eunhye Kwon, Donkyu Baek

Chungbuk National Univ.

요 약

자율이동로봇(AMR)은 사람과 공유되는 환경에서 동적 장애물에 신속하고 신뢰성 있게 대응해야 하며, 이를 위해 저지연성과 결정성이 보장된 의사결정이 요구된다. 그러나 많은 기존 시스템은 여전히 외부 서버나 클라우드 기반 추론에 의존하고 있어, 무선 통신과 원격 처리의 변동성으로 인해 예측 불가능한 지연이 발생한다. 이러한 문제를 해결하기 위해, 본 연구에서는 실시간 AMR 내비게이션을 위한 Zynq-7020 SoC 기반의 온디바이스 신경망 가속기를 제안한다. 제안한 가속기는 고정소수점 정책 신경망을 FPGA 패브릭 상에서 전적으로 실행하며, LiDAR 기반의 압축된 로컬 맵으로부터 AMR의 행동을 직접 추론한다. 성능 평가 결과, 하드웨어 가속기는 단일 추론에서 CPU 소프트웨어 대비 1.68배의 지연 감소를 달성하였으며, 반복 추론이 요구되는 실시간 AMR 내비게이션 환경에서는 최대 12배의 성능 향상을 보였다. 또한 실제 환경에서 주행한 결과 제안한 하드웨어 시스템이 AMR을 목표 위치까지 안정적으로 유도하는 올바른 내비게이션 동작을 지속적으로 생성함을 보여주며, 임베디드 AMR 시스템에 요구되는 신뢰성 있는 실시간 의사결정을 제공함을 입증한다.

Ⅰ. 서 론

기존의 자동 유도 차량(AGV)은 규칙 기반 제어와 사전에 정의된 경로에 의존하여, 구조화된 공장 레이아웃과 같이 제한된 환경에서만 운용이 가능하다. 이러한 방식은 동적인 장애물이나 환경 변화에 유연하게 대응하기 어렵다는 한계를 가지며, 이에 따라 보다 높은 자율성을 갖춘 자율이동로봇(AMR)으로의 전환이 촉진되고 있다. AMR은 센서 기반 인식과 학습 기반 내비게이션을 통합함으로써, 비구조적이고 지속적으로 변화하는 환경에서도 자율적인 이동을 수행할 수 있다.

학습 기반 제어 방식은 대규모 데이터와 반복적인 연산을 요구하기 때문에, 다수의 선행 연구에서는 신경망 모델의 학습 과정을 클라우드 서버에서 수행하는 방식을 제안하였다. 이러한 접근에서는 서버에서 학습된 모델을 로봇 시스템에 배포하여 매핑, 위치 추정 및 경로 계획과 같은 내비게이션 기능에 활용하고자 하였다[1], [2]. 이러한 방식은 실제 운용 환경에서 추론 단계까지 서버에 의존할 경우, 무선 통신 지연과 네트워크 상태 변화로 인해 실시간성이 저하될 수 있다.

서버 기반 추론의 한계를 보완하기 위해, A* 알고리즘을 FPGA로 가속한 연구[3]와 RRT 알고리즘을 GPU 병렬 구조로 구현한 연구가 제안되었다[4]. 이러한 로컬 가속 기반 접근은 서버 의존성을 감소시키고 처리 지연을 완화하는 데 기여하였으나, 환경 복잡도 증가에 따른 탐색 공간 확장, 전력 효율 및 엄격한 실시간성 보장 측면에서 여전히 한계를 가진다. 본 연구에서는 이러한 문제를 보완하기 위해 DQN 기반 경로 추론을 적용하고, 시퀀셜 연산 구조에 부분 병렬성을 결합한 FPGA 기반 온디바이스 추론 가속기를 설계한다. 이를 통해 복잡한 환경에서도 빈번한 재탐색 없이 안정적인 경로 결정을 수행하면서, 전력 및 자원 효율을 고려한 실시간 AMR 내비게이션을 구현하였다.

Ⅱ. 전체 시스템 구조

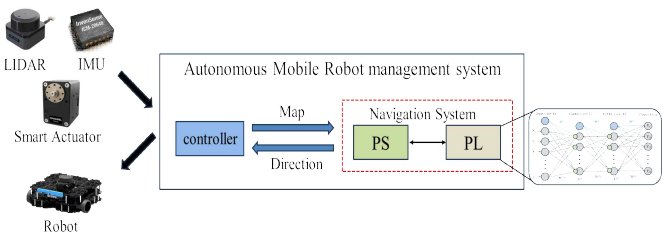

그림 1 제안하는 시스템 구조

그림 1은 본 연구에서 구현한 온디바이스 자율주행 로봇 제어 시스템의 전체 구성을 나타낸 것이다. 이 프레임워크에서는 센싱, 전처리및 추론이 로봇의 임베디드 플랫폼에서 완전히 실행된다. 외부 서버에 대한 의존성을 제거함으로써, 이 시스템은 동적인 환경에서 실시간 AMR 내비게이션에 적합한 결정론적이고 저지연의 의사결정을 보장한다.

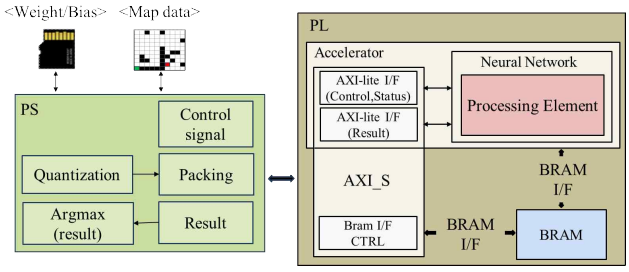

그림 2 내비게이션 시스템 구조

그림 2는 본 연구에서 제안한 내비게이션 시스템의 전체 구조를 나타낸 것이다. 시스템은 PS와 PL로 구성되며, 맵 데이터의 전처리와 신경망 기반 경로 추론 기능이 각각 분리된 형태로 동작한다.

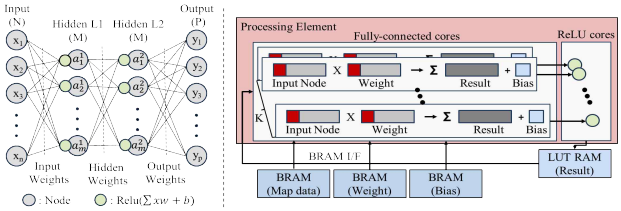

그림 3 가속기 구조

제안하는 신경망 가속기는 Processing Element(PE)를 기본 연산 단위로 사용한다. 각 PE는 완전연결층 연산을 수행하는 Fully-connected core와 활성화 함수를 처리하는 ReLU core로 구성되며, 하나의 뉴런에 대한 MAC 연산과 활성화 연산을 담당한다.

처리량 향상을 위해 K개의 PE를 병렬로 배치하여 하나의 run cycle 동안 K개의 뉴런 출력을 동시에 계산하도록 설계하였다. 입력, 가중치, 바이어스는 BRAM을 통해 공급되며, 연산 결과는 내부 메모리에 저장된다. 은닉층 또는 출력층의 뉴런 수가 K보다 큰 경우에는 동일한 PE 집합을 반복적으로 재사용하여 필요한 뉴런 수를 모두 처리한다. 이러한 구조는 제한된 하드웨어 자원 내에서 병렬성과 자원 효율을 동시에 확보할 수 있다.

Ⅲ. 실험 결과

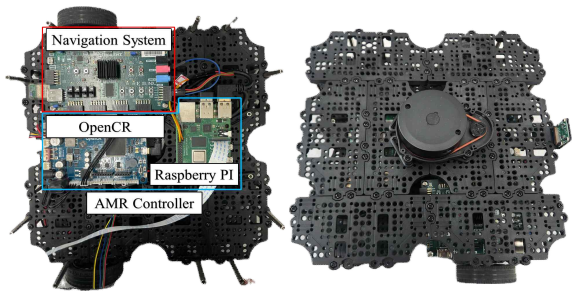

그림 4 AMR 플랫폼의 전체 하드웨어 구성

그림 4는 AMR 플랫폼의 실제 하드웨어 구성을 나타낸다. OpenCR 보드는 IMU 및 오도메트리 데이터를 수집하고 저수준 모터 제어를 수행하는 동시에, 수집된 센서 데이터를 라즈베리 파이로 전송한다. 라즈베리 파이는 LiDAR 스캔 데이터를 수신하고 OpenCR로부터 전달된 센서 정보와 융합하여 서버로 전송한다. 서버에서 생성된 로컬 지도는 내비게이션 시스템의 입력으로 사용되며, 이를 기반으로 다음 이동 방향이 결정된다. 이후 선택된 동작 명령은 다시 OpenCR로 전달되고, OpenCR은 해당 명령을 해석하여 모터 제어 신호를 생성함으로써 AMR의 실제 주행을 수행한다.

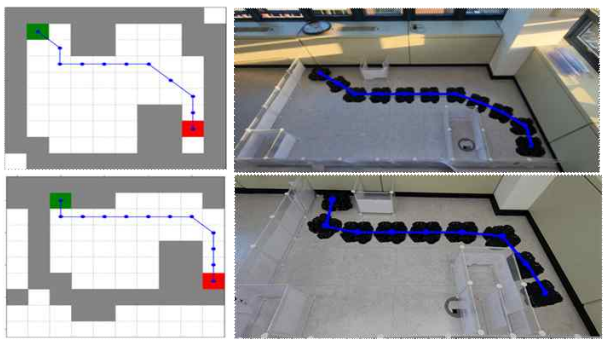

그림 5 서버/온디바이스 경로 추론 결과

그림 5는 서버 기반 추론과 온디바이스 기반 추론 결과의 비교를 나타낸다. 그림 5(a)는 학습된 DQN 정책 신경망을 이용하여 서버 환경에서 추론한 기준(reference) 경로를 보여주며, 그림 5(b)는 동일한 입력에 대해 FPGA 기반 온디바이스 추론 결과를 바탕으로 실제 AMR이 주행한 경로를 나타낸다. 비교 결과, 온디바이스 기반 추론을 통해 생성된 실제 주행 경로는 서버 기반 추론 결과와 동일한 경로추론을 보였다.

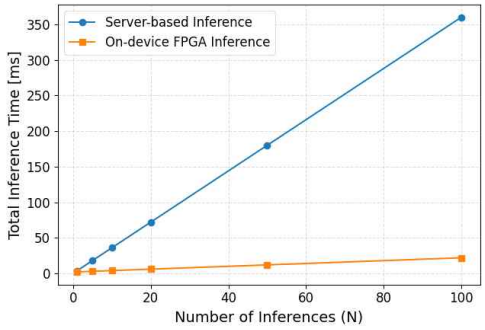

그림 6 추론 지연 시간 비교

그림 6은 반복 추론 환경에서 누적 지연 시간을 비교한 결과를 나타낸다. 단일 추론에서는 제안한 가속기가 서버 기반 추론 대비 1.68배의 지연 감소를 달성하였으며, 반복 추론 환경에서는 누적 지연 측면에서 최대 12배의 성능 향상을 보였다.

Ⅳ. 결론

본 연구에서는 Zynq-7020 플랫폼 상에서 센싱, 전처리, 고정소수점 정책 추론을 통합한 완전 온디바이스 신경망 가속기를 제안하였다. 병렬 Processing Element 구조, 완전연결 연산의 folding 기법, 그리고 온칩 고정소수점 정렬을 적용함으로써, 외부 서버에 의존하지 않는 결정론적이고 저지연의 추론을 구현하였다. 서버 기반 추론 경로와 온디바이스 기반 추론 결과를 비교한 결과, 두 방식 모두 동일한 경로를 도출함을 확인하였으며, 이를 통해 제안한 온디바이스 추론 시스템이 서버 기반 추론과 동등한 경로 결정 성능을 제공함을 검증하였다, 이를 통해 기능적 신뢰성을 입증하였다. 성능 평가 결과, 단일 추론에서는 서버 대비 1.68배의 지연 감소를 달성하였고, 실시간 AMR 내비게이션에 요구되는 반복 추론 환경에서는 최대 12배의 성능 향상을 보였다. 또한 전체 아키텍처 및 추론 파이프라인을 시스템 수준에서 검증함으로써, 제안한 설계가 임베디드 AMR 플랫폼에 적용 가능함을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2026년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. RS-2020-NR049604).

참 고 문 헌

[1] J. Abbenseth et al., “Cloud-based cooperative navigation for mobile service robots in dynamic industrial environments,” in Proc. ACM SAC, 2017, pp. 460 - 467.

[2] N. Zagradjanin et al., “Cloud-based multi-robot path planning in complex and crowded environment using fuzzy logic and online learning,” Inf. Technol. Control, vol. 50, no. 2, pp. 357 - 374, 2021.

[3] A. Kosugeand T. Oshima, “A 1200×1200 8-Edges/Vertex FPGA-Based Motion-Planning Accelerator for Dual-Arm-Robot Manipulation Systems,” *Proc. IEEE*, 2020.

[4] J. Bialkowski, S. Karaman, and E. Frazzoli, “Massively parallelizing the RRT and the RRT*,” Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3513 - 3518, 2011.