

LoRA 파인튜닝과 양자화를 활용한 경량 Vision-Language Model의 객체 선택 성능 평가

이준혁¹⁾, 송기훈²⁾, 송현욱¹⁾, 안양진¹⁾, 박형준²⁾, 강창묵^{*1,2)}
한양대학교 전기공학과¹⁾, 한양대학교 인공지능학과²⁾

Performance Evaluation of a Lightweight Vision-Language Model with LoRA Fine-Tuning and Quantization for Object Selection

Junhyuk Lee¹⁾, Gihoon Song²⁾, Hyeonwook Song¹⁾, Yangjin An¹⁾, Hyeongjun Park²⁾, Changmook Kang^{*1,2)}
Department of Electrical Engineering¹⁾, Department of Artificial Intelligence Hanyang University²⁾

Abstract - 본 연구는 물류 현장에서 요구되는 톨 선택 문제를 반영하기 위해, 현업 적용 가능성을 고려한 튜토리얼형 시각-언어 추론 작업을 설계하고 이를 기반으로 경량 모델의 성능을 검증하였다. 실제 산업 환경에서는 작업 대상과 작업 공간의 형태를 관찰한 뒤 어떤 톨이 가장 적합한지를 선택해야 하며, 이 과정에서 추론 정확도뿐 아니라 비용, 지연시간, 네트워크 의존성, 하드웨어 제약까지 함께 고려되어야 한다. 이러한 이유로 클라우드 기반 대형 모델보다 온디바이스에서 동작 가능한 경량 모델의 활용 가능성이 중요하다. 이에 해당 연구는 복잡한 현업 문제를 직접 다루기보다, 핵심 의사결정 구조를 반영한 튜토리얼형 테트리스 객체 선택 문제로 단순화하여 모델의 성능을 평가하였다. 입력은 단일 Top-view 이미지이며, 모델은 작업대와 후보 객체를 바탕으로 가장 적합한 객체를 선택해야 한다. 이를 위해 Qwen3-VL-4B 기반 경량 비전언어모델에 LoRA(Low-Rank Adaptation) 파인튜닝을 적용하였고 양자화 모델 및 GPT-5.1 기반 API 추론과 성능을 비교하였다. 실험 결과, LoRA 파인튜닝 모델은 51.49%의 정확도로 GPT-5.1 대비 우수한 성능을 보였으며, 4-bit 양자화를 적용한 경우 정확도는 일부 감소하였으나 GPU 메모리 사용량은 약 8.9GB에서 3.24GB로 크게 감소하였다. 또한 추론 지연시간은 약 524ms 수준으로 유지되어, 제한된 자원 환경에서도 효율적인 온디바이스 추론이 가능함을 확인하였다.

1. 서 론

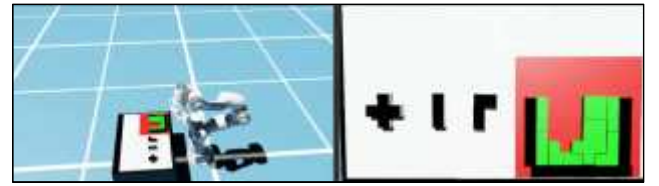
최근 물류 자동화 현장에서는 단순 이송이나 반복 작업을 넘어 작업 대상의 형상과 작업 공간의 구조를 함께 고려하여 적절한 톨이나 조작 방식을 선택하는 지능형 의사결정의 중요성이 점차 커지고 있다. 이러한 방식은 단순한 객체 인식이나 분류 문제와는 다르며, 형상 이해, 제약 조건 반영, 후보 간 비교, 최종 선택이 결합된 멀티모달 추론 문제에 해당한다. 그러나 실제 물류 현장에 대형 상용 API 모델을 그대로 적용하는 데에는 현실적인 제약이 존재한다. 첫째, API 호출 기반 방식은 지속적인 사용 비용이 발생하므로 반복적인 추론이 필요한 산업 환경에서는 운영 부담이 커질 수 있다. 둘째, 외부 네트워크 연결에 의존하는 구조는 지연시간 증가나 응답 불안정으로 이어질 수 있어 실시간성이 요구되는 작업에 불리하다. 셋째, 산업 현장에서는 로봇 컨트롤러 또는 엣지 디바이스 인근에서 직접 추론을 수행할 수 있는 구조가 운영 안정성과 보안 측면에서 더 적합할 수 있다. 따라서 현업 적용 관점에서는 단순히 최고 정확도를 보이는 모델을 선택하는 것보다 제한된 하드웨어 자원 내에서 충분한 성능을 유지하면서도 온디바이스 구동이 가능한 경량 모델을 확보하는 것이 중요하다. 이러한 문제의식을 바탕으로 실제 물류 현장의 톨 선택 문제를 곧바로 다루기보다 그 핵심 의사결정 구조를 반영한 튜토리얼형 문제를 먼저 설계하였다. 본 연구에서 사용한 테트리스형 객체 선택 작업은 빈 공간과 후보 객체를 함께 제시하고 그중 가장 적합한 객체를 선택하도록 구성된다. 이는 실제 산업 현장의 복잡한 환경을 단순화한 형태이지만, 본질적으로는 주어진 작업대와 후보 톨을 보았을 때 어떤 선택이 가장 적합한가를 판단하는 구조와 유사하다. 따라서 해당 연구는 실제 물류 현장 문제로 확장하기에 앞서 모델이 형상 기반의 적합성 판단과 제약 조건 반영을 얼마나 안정적으로 수행할 수 있는지를 검증하는 실험적 기반으로 활용될 수 있다. Qwen3-VL-4B-Instruct 기반 경량 비전언어모델에 LoRA 미세조정을 적용하여 작업 특화 성능 향상을 유도하였다. 또한 양자화를 통해 모델의 경량화를 추가로 수행하고, 이를 범용 상용 모델인 GPT-5.1 기반 API 추론 결과와 비교하였다. 그 결과, 제한한 미세조정 모델은 특정 작업에서 GPT-5.1 기중선보다 우수한 성능을 보였으며, 전반적으로 유료 API 사용 수준에 근접하는 정확도를 달성하였다. 이는 작업 특화 미세조정을 통해 경량 모델의 성능을 실용 수준까지 끌어올릴 수 있음을 보여주며, 나아가 물류 로봇 분야에서 범용 상용 모델에 대한 의존도를 낮추고 엣지 디바이스 기반의 독립적 추론 시스템을 구현할 가능성을 시사한다.

2. 본 론

2.1 시뮬레이션 환경 구성 및 작업 정의

로봇 실험 환경은 NVIDIA Isaac Sim 기반의 시뮬레이션 환경에서 실험을 수행하였다[1]. 작업 공간은 그림 1과 같이 단일 평면 테이블로 구성되며 테이블 위에는 테트리스 형태의 목표 구조와 여러 개의 후보 객체가 배치된다. 시스템의 전체 시나리오는 먼저 초기에 로봇은 initial pose로 들어간다. 다음 Vision-Language Model(VLM)으로 작업 공간의 Top-view 이미

지와 프롬프트를 입력으로 받아 후보 객체 중 하나를 선택한다. 이후 선택된 객체에 대해 로봇 매니퓰레이터가 접근하여 해당 객체를 테트리스 구조의 내부 목표 영역으로 밀어 넣는 pushing 조작 동작을 수행한다. 이러한 구조를 통해 시각 기반 객체 선택과 로봇 조작 동작을 연계한 통합 작업 시나리오를 구성하였다.



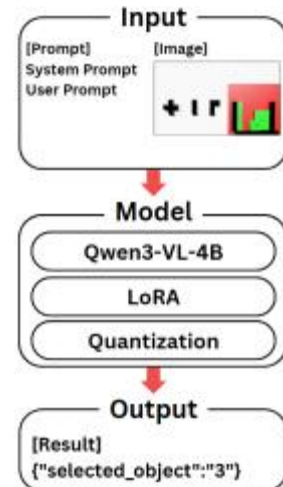
〈그림 1〉 Isaacsim environment

2.1.1 로봇 및 센서 구성: Franka Panda

시뮬레이션 환경에서 로봇 플랫폼으로는 Franka Emika Panda 매니퓰레이터를 사용하였다. Panda는 7자유도(7-DoF) 관절 구조를 갖는 대표적인 연구용 로봇팔로, 시뮬레이션 환경에서 파지 및 접촉 기반 조작 시나리오를 반복 검증하기에 적합하다. 로봇의 동작 계획과 실행은 MoveIt2 프레임워크를 기반으로 수행하였다. 구체적으로, 도구 파지 전 접근 동작, 파지 자세 정렬, 그리고 파지 이후 푸싱 동작을 시작하기 위한 초기 자세로의 이동 등 일련의 동작을 MoveIt2의 모션 플래닝 기능을 통해 생성하였다. 또한 경로 생성 과정에서 로봇과 환경 간의 충돌 가능성을 고려한 안전한 경로를 계획하도록 구성하였다.

2.2 Qwen3-VL-4B-Instruct 기반 파인튜닝 및 경량화

본 연구에서는 물류 현장형 톨 선택 문제를 온디바이스 환경에서 수행하기 위해, 비교적 작은 규모의 비전언어모델인 Qwen3-VL-4B-Instruct를 기반 모델로 채택하였다[2]. 해당 모델은 시각 정보와 텍스트 지시를 함께 처리할 수 있어 작업 테이블의 상태와 후보 객체를 동시에 이해하고 적합한 대상을 선택해야 하는 행동에 적합하다. 특히 4B 규모의 모델은 초대형 모델에 비해 모델 크기와 연산 부담이 과도하지 않아 제한된 자원 환경에서도 파인튜닝 및 추론 실험을 반복적으로 수행하기에 유리하다. 또한 해당 연구에서 유료 API에 의존하기보다 작업 특화 데이터셋을 활용해 경량 모델을 직접 적용시키는 방향을 택하였다. 이는 좁고 구조화된 문제에서 대규모 일반 지식보다 규칙에 대한 일관된 적용이 더 중요할 수 있다는 가정으로 선택하였다. 따라서 본 연구에서는 Qwen3-VL-4B-Instruct를 대상으로 LoRA 기반 파인튜닝과 양자화를 함께 적용하여 실제 로봇 시스템에 탑재 가능한 VLM을 그림 2와 같이 파이프라인으로 구성하였다.



〈그림 2〉 Proposed VLM Object Selection Pipeline

[6] A. Yadav and R. C. Bhargavi, "Optimizing LLMs Using Quantization for Mobile Execution," arXiv preprint, arXiv:2512.06490, 2025.