에지컴퓨팅 기반 이미지처리시스템을 위한 적응형 분산 추론

박민혁, 김동민*
순천향대학교

xlfksh9613@naver.com, *dmk@sch.ac.kr

Adaptive Distributed Inference for Edge Computing-based Image Processing Systems

Min Hyeok Park, Dong Min Kim*
Soonchunhyang University, *Soonchunhyang University

요 약

본 논문은 에지컴퓨팅 서버를 활용하여 이미지 처리 시 발생하는 서버 병목 문제 해결을 위해 적응형 분산 추론 기법을 제안한다. 서버 대기 시간에 따라 단말이 온디바이스 추론을 수행하고 중간 결과를 전송하며, 이를 통해 기존 방식 대비 총 처리 시간과 서버 부하를 유의미하게 줄이고 성능 편차도 완화한다. 실시간 환경에 효과적인 해결책이 될 수 있음을 확인하였다.

Ⅰ. 서 론

최근 로봇 비전 분야에서는 이미지 기반 키워드 추출 및 자연어 생성 모델을 활용한 다양한 응용이 활발히 연구되고 있다. 특히 로봇이 촬영한 이미지를 중앙 서버로 전송하여 Vision-Language Model(VLM)을 통해 로봇의 동작을 제어하는 시스템이 주목받고 있다[1-3]. 그러나 이러한 중앙 집중식 구조는 다수의 로봇이 동시에 이미지를 서버로 전송할 경우, 서버 측 처리 병목현상 및 시간지연 문제를 초래한다는 한계점이 존재한다. 본 연구에서는 이러한 문제를 해결하기 위해 YOLOv3 [4] 기반의 적응형 분산 추론 시스템을 제안한다. 제안된 시스템에서는 각 로봇이 서버로부터 예상 대기 시간을 수신하고, 예상 대기시간만큼 특정 레이어까지 온디바이스(On-Device) 모델 추론을 수행하여 중앙 서버로 전송한다. 이를 통해 기존의 대기시간을 온디바이스 추론하는 시간으로 활용하고 중간 결과만을 서버로 전송함으로써 전체 시스템의 처리 효율성을 향상시키는 효과를 얻을 수 있다.

Ⅱ. 본론

본 논문에서 제안하는 에지컴퓨팅 기반 이미지 처리 시스템은 다음과 같이 설계되었다. 첫째, 로컬환경(예: Single-Board Computer, SBC)에서의 실시간 객체 인식을 위해 연산 속도와 인식 정확도 간의 최적 균형을 제공하는 Darkne-53 기반의 YOLOv3 아키텍처를 채택하였다. 둘째, 중앙 서버의 처리 효율성 향상을 위해 큐잉(Queueing) 이론에 기반한 요청 관리 메커니즘을 구현하여 다중 로봇으로부터 발생하는 동시다발적 요청을 체계적으로 관리한다. 셋째, 서버의 큐 상태를 실시간으로 모니터링하고, 대기 시간이 예상될 경우 로봇 자체의 온디바이스(On-Device) 연산 자원을 활용하는 적응형 분산 추론 기법을 적용함으로써 시스템 전반의 병목 현상을 효과적으로 완화하는 방법론을 제시한다.

적응형 분산 추론 시스템은 다중 로봇 환경에서의 이미지 처리 효율성을 극대화하기 위한 계층적 구조로 설계되었다(그림 1). 시스템의 작동 메커니즘은 다음과 같다. 먼저, 현장에 배치된 각 로봇은 주변 환경을 촬영한 이미지 데이터를 중앙 서버로 전송한다. 이후 중앙 서버는 실시간 처리 대기열 상태를 모니터링하여 각 로봇 별 예상 처리 지연 시간을 산출하고, 이를 해당 로봇에게 즉시 통보한다. 로봇은 수신된 대기 시간 정보에 기반하여 온디바이스 추론 전략을 적용한다. 위 실험 환경을 기준으로 예를 들면 예상 대기 시간이 약 4.5 초 정도 측정됐을 경우 로컬환경(Raspberry Pi 3B+)에서 YOLOv3 의 13 개 레이어까지는 평균 4.4 초 이내에 처리 가능하기 때문에 YOLOv3 네트워크의 초기 13 개 레이어를 로컬에서 처리한 후 나머지 후속 레이어는 중앙 서버로 전달하여 처리한다. 최종적으로 서버는 수신된 데이터의 처리 단계에 따라 필요한 후속 연산을 수행한다.

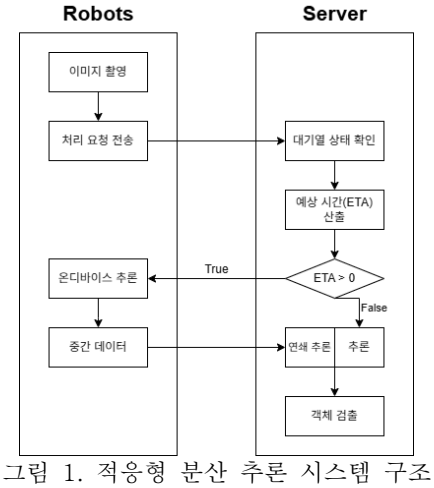

그림 1. 적응형 분산 추론 시스템 구조

본 연구에서는 제안하는 적응형 분산 추론 시스템이 기존 중앙 집중형 서버 처리 방식 대비 얼마나 효과적으로 처리 시간을 단축시키고, 서버 부하를 분산할 수 있는지를 실험적으로 분석한 결과를 제시한다. 실험에는 Raspberry Pi 3B+ 기반의 로봇 에이전트 10 대를 활용하였으며, 각 클라이언트가 동시에 이미지를 촬영하고 서버에 전송하는 상황을 가정하였다. 모든 실험은 동일한 네트워크 환경에서 100 회의 이미지 추론 요청을 기반으로 평균값을 산출하였다. 그림 2 는 각 클라이언트의 총 처리 시간을 비교한 결과이다. 중앙 집중형 방식에서는 모든 이미지가 서버로 전송되고, 서버가 순차적으로 이미지를 처리하기 때문에 후속 클라이언트로 갈수록 대기 시간이 누적되어 총 처리 시간이 선형적으로 증가하는 경향을 보인다. 반면, 제안하는 적응형 분산 추론 시스템에서는 서버로부터 받은 예상 대기 시간에 따라 각 로봇이 YOLOv3 의 일부 또는 전체 레이어를 온디바이스에서 처리한 후, 중간 출력만 서버에 전송하게 된다. 이로 인해 로봇 자체의 연산을 적극 활용할 수 있으며, 클라이언트 간 총 처리 시간의 분산을 줄이고 평균 응답 속도를 향상시킨 것을 확인할 수 있다.

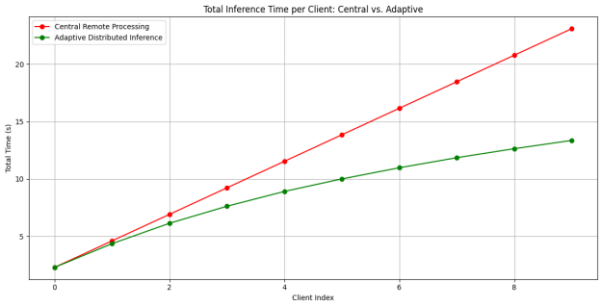

그림 2. 적응형 분산 추론 시스템과 중앙 집중형 추론 시스템의 클라이언트별 총 처리 시간 비교

그림 3 은 각 클라이언트의 처리 과정을 시간 구성 요소별로 분해하여 시각화한 것이다. 이를 통해 중앙 집중형 추론 방식과 제안하는 적응형 분산 추론 시스템 간의 서버 부하 분산 및 처리 효율성 차이를 직관적으로 확인할 수 있다. 중앙 집중형 방식에서는 모든 클라이언트가 순차적으로 서버에 접근하며, 각 클라이언트는 동일한 서버 처리 시간을 갖는다. 그러나 앞선 클라이언트의 처리 완료를 기다리는 대기 시간이 누적되어, 후속 클라이언트로 갈수록 총 소요 시간이 선형적으로 증가하는 경향을 보인다. 이 방식은 서버의 처리 능력에 전적으로 의존하며, 동시 요청이 증가할 경우 병목 현상이 발생하기 쉽다. 반면, 적응형 분산 추론 시스템에서는 서버로부터 전달받은 예상 대기 시간 정보를 기반으로 각 클라이언트가 해당 시간 동안 YOLOv3 의 일부 레이어를 온디바이스에서 처리하게 된다. 이로 인해 각 클라이언트는 대기 시간 동안 로컬 연산을 수행함으로써 서버의 부하를 줄이는 동시에 전체 추론 시간을 효과적으로 분산시킨다. 특히, 클라이언트 인덱스가 높아질수록 온디바이스 연산의 비율이 증가하고, 이에 따라 서버에서 수행되는 추론 단계가 줄어들어 중앙 서버의 처리 시간 역시 감소하는 패턴을 보인다. 결과적으로, 중앙 집중형 방식에서는 동일한 서버 처리 시간이 누적되어 총 소요 시간이 계속해서 증가하는 반면, 적응형 분산 추론 방식은 클라이언트의 내부 연산이 늘어남에도 불구하고 서버 측 부담이 지속적으로 감소함으로써 총 추론 시간이 점차 줄어드는 역전 현상을 만들어낸다. 이는 분산 추론 전략이 다중 클라이언트 환경에서 특히 효과적임을 보여준다.

그림 3. 적응형 분산 추론 시스템과 중앙 집중형 추론 시스템 세부 시간 비교

Ⅲ. 결론

본 논문에서는 에지컴퓨팅 기반 다중 로봇 환경에서의 서버 처리 지연 문제를 해결하기 위해 적응형 분산 추론 시스템을 제안하였다. 기존의 중앙 집중형 방식 대비 총 처리 시간과 서버 부하가 유의미하게 감소하였으며, 클라이언트 수 증가 시 성능 이점이 더욱 두드러졌다. 이는 실시간성이 중요한 응용 분야에서 효과적인 해결책이 될 수 있음을 시사한다. 향후 클라이언트 성능 기반 동적 분할 기법 및 다양한 모델 적용 가능성에 대한 연구가 필요하다.

ACKNOWLEDGMENT

이 논문은 2024 년 대한민국 교육부와 한국연구재단의 인문사회분야 일반공동연구지원사업(융복합연구)의 지원을 받아 수행된 연구임(NRF-2022S1A5A2A03052880). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학·석사연계 ICT 핵심인재양성사업의 연구결과로 수행되었음. (IITP-2025-RS-2024-00436500)

참 고 문 헌

[1] D. M. Kim, Y. H. Suh and I. F. Ngwalo, “Efficient Data Communication for Deep Learning Application via Latent Code Transmission in B5G Wireless Networks,” in Proc. IEEE Information and Communication Technology Convergence (IEEE ICTC 2023), Jan. 2023.

[2] C. Hu, and B. Li. “Distributed inference with deep learning models across heterogeneous edge devices.” in Proc. IEEE Conference on Computer Communications (IEEE INFOCOM 2022), May 2022.

[3] R. Firoozi, J. Tucker, S. Tian, et al., “Foundation models in robotics: Applications, challenges, and the future.” The International Journal of Robotics Research, 44(5), 701-739, 2024.

[4] J. Redmon, A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv preprint arXiv:1804.02767, 2018.