# 엣지 디바이스 지능전이 환경의 AI 모델 배포 전송량 및 전송 시간 분석에 관한 연구

송순용, 배희철
한국전자통신연구원

{soony, hessed}@etri.re.kr

# A Study on the Analysis of Throughput and Payload Duration for AI Model Deployment in Edge Device Intelligence Transfer Environments

Soonyong Song, Heechul Bae
Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문은 NS3 시뮬레이터와 WIFI 통신 환경을 이용하여 지능형 엣지 디바이스 간의 지능전이를 위한 AI 모델 배포 성능을 분석하였다. 본 논문은 클라이언트 수(N=1, 4, 16, 64, 256, 1024)와 모델 파라미터 수(K=1M, 2M, 3M, 4M)에 따른 전송량과 전송 시간을 시뮬레이션을 통해 평가하였다. 실험 결과, 클라이언트 수가 1024 일 때, 4M 파라미터에서 전송량은 98.78MBps 였고, 전송 시간은 1.79 초로 측정되었다. 본 연구는 AI 모델 배포 시 고려해야 할 시간적 조건을 제시하며, 군집 드론 및 로봇의 AI 모델 배포 전략 수립을 위한 기준으로 활용할 수 있을 것으로 전망한다.

## Ⅰ. 서 론

엣지 디바이스에서 AI 모델을 사용하는 것은 데이터의 실시간 처리, 통신 대역폭 절감, 보안 유지 측면에서 뚜렷한 장점이 있다. 실시간 처리는 자율주행, 헬스케어, IoT 등에서 필수적이며, 클라우드로의 데이터 전송을 줄여 네트워크 부하를 감소시키고, 민감한 데이터의 외부 전송을 최소화하여 보안을 강화할 수 있다. 엣지 디바이스는 컴퓨팅 자원이나 전력이 부족한 제약 사항이 있는데, 이를 해결하기 위해 기기들을 군집화하고 있다. [1] 예를 들어, 분산 컴퓨팅을 활용하여 각 노드가 계산의 일부를 담당하고 계산 결과를 하나의 기기에서 통합하는 방식으로 구성할 수 있다. 또 다른 예로는, 공간적으로 분리된 디바이스들이 AI 모델의 결과를 앙상블하여 향상된 성능을 기대할 수 있다. 이러한 방식은 스마트 시티 및 스마트 팩토리와 같은 분산 환경에서 시너지를 발휘한다. 그러나 엣지 디바이스 간 협업의 이점을 극대화하려면 분산 처리를 위한 글로벌 AI 모델이 빠르게 공유되어야 한다. [2] 다른 AI 모델로 데이터를 처리하면 결과가 달라져 잘못된 의사결정을 할 가능성이 높아진다. 이러한 문제는 엣지 디바이스의 AI 모델 업데이트 시점에서 발생할 수 있다. 따라서, 적어도 AI 모델을 배포하는 동안에는 기존 모델을 유지하거나 서비스를 중단해야 한다. 군집화된 엣지 디바이스의 원활한 동작을 위해서는 AI 모델 배포에 필요한 시간을 분석하여 업데이트된 모델의 사용 시점을 스케줄링할 수 있어야 한다. 본 논문에서는 AI 모델의 배포를 통한 엣지 디바이스로의 지능 전이에 필요한 시간을 분석하였다. 본 논문에서의 배포 환경은 지능을 배포하는 서버와 지능을 사용하는 클라이언트로 구성되며, 무선 네트워크인 WIFI ac 를 사용해 토폴로지를 구성하는 조건을 고려하였다. 통신 성능 측정을 위해 WIFI 및 HTTP 프로토콜 사용이 가능한 NS3 를 사용하였다. [3] 전송할 모델의 크기에 따라 전송량과 시간의 차이가 발생할 것으로 판단되어 클라이언트 수에 따라 두 가지 성능을 시뮬레이션 하였다. 본 논문은 다수의 기기에 지능을 전이하기 위해 필요한 시간의 기준점을 제시하고 있어, 향후 군집 드론 및 군집 로봇에 필요한 AI 모델 배포를 위한 시간 스케줄링에 활용될 수 있을 것으로 전망한다.

## Ⅱ. 본론

NS3 는 네트워크 시뮬레이션을 위한 오픈 소스 소프트웨어로, 네트워크 연구 및 교육에서 널리 사용되고 있다. NS3 는 다양한 네트워크 프로토콜을 시뮬레이션 하기 위한 다양한 도구를 제공한다. 실제 네트워크 환경을 모사하여 네트워크 성능을 분석하고, 새로운 네트워크 기술과 프로토콜을 테스트를 위해 사용한다. NS3 는 특정 네트워크 프로토콜이나 시나리오를 쉽게 추가하고 수정할 수 있도록 모듈화된 아키텍처로 구성되어 있다. 또한, NS3 는 실제 네트워크 장비와의 연동이 가능하여, 시뮬레이션 결과를 실험 환경과 비교하여 검증할 수 있다. 이러한 특성 덕분에, 복잡한 네트워크 시나리오를 분석하기 위해 수많은 연구자들이 NS3 를 활용하고 있다. 본 논문은 NS3 에서 지원하는 WIFI ac 와 HTTP 프로토콜을 사용한다. 유선 네트워크를 사용한다면 통신 안정성은 확보되겠지만, 본

논문에서 무선 네트워크를 사용하는 이유는 다양한 공간에 분산 설치된 엣지 디바이스들에 연결하기에는 번거롭고 유지보수가 어려우며 설치장소 변경도 어렵기 때문이다. WIFI ac 는 IEEE 802.11ac 표준에 기반한 무선 네트워크 표준기술로, 고속 데이터 전송과 높은 네트워크 용량을 제공한다. 중심주파수로 5GHz 대역을 사용하고 최대 160MHz 에 달하는 넓은 채널 폭을 제공하여 이론상 대용량 데이터 전송이 가능하다. 또한, MU-MIMO 기술을 통해 동시에 여러 클라이언트의 고속 데이터 전송을 지원한다. 서버와 클라이언트의 통신 프로토콜은 Restful API 로 가정하였다. Restful API 는 웹을 기반으로 4 가지의 단순한 명령을 통해 구동되는 어플리케이션 레벨의 데이터 통신방법으로, HTTP 프로토콜을 기반으로 동작한다. 실제로 NVIDIA 의 Triton Inference Server 를 포함한 수많은 AI 모델 배포 서비스들이 Restful API 를 통해 운용되고 있다. [4]

본 논문에서는 NS3 시뮬레이션 실험을 위해 클라이언트 수 N={1, 4, 16, 64, 256, 1024}, 모델 파라미터 개수 K={1M, 2M, 3M, 4M}, 으로 설정하였다. 여기서 M 은 백만 개를 의미한다. 1 개의 파라미터는 4 바이트의 부동 소수점 수로 표현되는 것으로 가정하였다. 여기서 파라미터의 수는 엣지 디바이스에서 주로 사용되는 MobileNet V2 와 SqueezeNet 을 기준으로 설정하였다. 이 모델들은 각각 3.5M 및 1.2M 파라미터로, 논문에서 구축한 시나리오로 시뮬레이션이 가능하다. 서버와 클라이언트는 WIFI AP 를 중심으로 클라이언트들이 스타 토폴로지로 접속한다고 가정하였다. 시뮬레이션 시간은 3600 초로 설정하였고, 클라이언트가 1 초에서 1800 초 사이의 랜덤 시간에 GET 메시지로 모델 전송을 요청하면 서버가 해당 클라이언트로 모델 가중치를 배포하는 시나리오를 설정했다.

그림 1 은 클라이언트 수에 따른 평균 throughput 과 평균 payload duration 을 나타냈다. 그림 1 에 따르면 클라이언트 수가 64 개이고 4M 파라미터일 때 전송량이 120.18Mbps 로 최대인 것을 알 수 있으며, 클라이언트 수가 64 개이고 1M 파라미터일 때 전송 시간이 0.49 초로 최소인 것을 알 수 있다. 실험에서 설정한 시나리오에서는 해당 파라미터로 AI 모델 배포 환경을 구성한다면 통신 성능을 최적화 할 수 있을 것으로 판단할 수 있다. 그림 2 는 파라미터 수에 따른 평균 throughput 과 평균 payload duration 을 나타냈다. 그림 2 에 따르면 클라이언트 수와 모델 파라미터 수가 증가할수록 통신 성능은 저하되는 경향을 보였다. 이 결과들로, 클라이언트 수와 모델 파라미터 수가 증가할수록 통신 성능이 저하됨을 확인할 수 있었다. 클라이언트 수가 1024 일 때, 4M 파라미터에서 전송량은 98.78MBps 였고, 전송 시간은 1.79 초로 측정되었다. 이를 통해 클라이언트 수가 많아질수록, 모델 파라미터 수가 클수록 통신 성능이 저하되는 경향을 확인할 수 있었다. 1024 개의 클라이언트에 AI 모델을 배포해도 P2P 전송에 평균 2 초 미만의 시간이 소요되는 것을 확인할 수 있다.

이와 같은 결과는 AI 모델을 서버에서 클라이언트로 배포할 때, 클라이언트 수와 모델 파라미터 수를 적절히 조정하는 것이 중요하다는 것을 시사한다. 또한, AI 모델에 양자화를 적용하여 1 개 파라미터당 바이트 수를 조절하면 전송할 수 있는 파라미터의 양을 크게 증가시킬 수 있으며, 이는 AI 모델 배포의 효율성을 높일 수 있다. 예를 들어 2 바이트로 양자화하면 8M 파라미터까지 수용이 가능하며, 이는 79.8%의 acc@1 성능을 보이는 EfficientNet B1 모델의 가중치 전송이 가능하다.

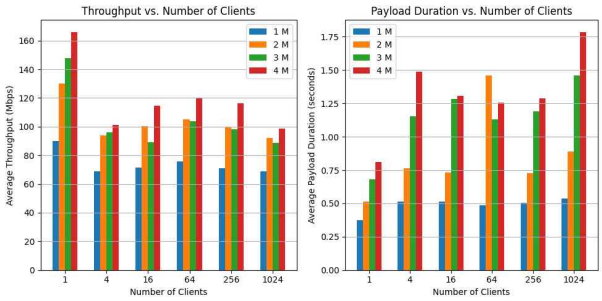


그림 1. 클라이언트 수에 따른 성능 분석 결과. (a) 평균 전송량 (b) 평균 전송시간

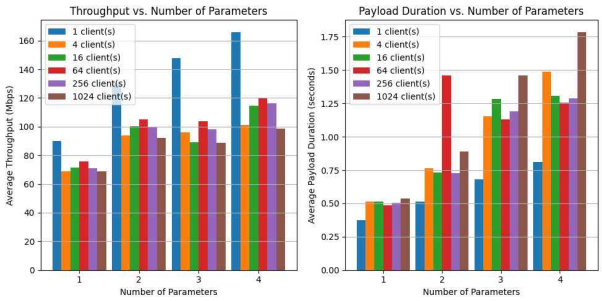


그림 2. 파라미터 수에 따른 성능 분석 결과. (a) 평균 전송량 (b) 평균 전송시간

## Ⅲ. 결론

본 논문은 NS3 시뮬레이터와 WIFI 네트워크를 이용해 엣지 디바이스 간 AI 모델 배포 시 클라이언트 수와 모델 파라미터 수에 따른 전송량과 전송 시간을 분석하였다. 실험 결과, 무선 네트워크 환경에서 Restful API 를 사용하여 4M 개 파라미터의 AI 모델 배포시 1024 개의 클라이언트에서 약 1.79 초의 시간이 필요한 것을 알 수 있었다. 이러한 결과를 바탕으로, 양자화를 적용하여 파라미터당 바이트 수를 조정하여 정보 전송 효율을 높인다면 성능이 나은 모델 배포가 가능함을 언급하였다. 이 논문을 통해 군집 드론 및 로봇의 AI 모델 배포 전략 수립에 기여할 수 있다.


## ACKNOWLEDGMENT

본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음. [24ZR1100, 자율적으로 연결·제어·진화하는 초연결 지능화 기술 연구]


## 참 고 문 헌


[1] Patel, Rahul, et al. "A Comprehensive Review on Edge Computing, Applications & Challenges." Security and Risk Analysis for Intelligent Edge Computing (2023): 1-33.

[2] Ale, Laha, et al. "Empowering generative AI through mobile edge computing." Nature Reviews Electrical Engineering (2024): 1-9.

[3] Zárate Ceballos, Henry, et al. "Network Simulating Using ns-3." Wireless Network Simulation: A Guide using Ad Hoc Networks and the ns-3 Simulator. Berkeley, CA: Apress, 2021. 65-95.

[4] https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/index.html