엣지-디바이스 협업 환경에서 지연 시간 최소화를 위한 ResNet 동적 분할 전략 연구

정준호, 박진효, 윤주상*

동의대학교

jeong@junho.dev, jhpark3679@gmail.com, *jsyoun@deu.ac.kr

A Study on Dynamic Splitting Strategies for ResNet to Minimize Latency in Edge-Device Collaborative Environments

Junho Jeong, Jinhyo Park, Joosang Youn*

Dong-Eui Univ.

요 약

로컬 디바이스에서의 고성능 인공지능 모델 실행 수요가 증가함에 따라, AI 모델을 디바이스와 엣지 서버에서 분할하여 실행하는 분할 컴퓨팅 기술의 중요성이 대두되고 있다. 그러나 최적의 분할 지점은 디바이스의 연산 성능, 네트워크 대역폭 등 동적으로 변화하는 시스템 환경에 따라 민감하게 달라진다. 본 논문에서는 총 지연 시간 최소화를 목표로, 다양한 운영 환경에 따른 최적의 모델 분할 전략을 분석하기 위해 시뮬레이션 기반 연구를 수행한다. 이를 위해 총 지연 시간을 디바이스 연산 시간, 데이터 전송 시간, 엣지 연산 시간의 합으로 모델링하고, 대표적인 심층신경망인 ResNet-50을 대상으로 시스템 환경 변화에 따른 최적 분할 지점을 도출하였다. 실험 결과, 로컬 처리가 대부분의 저대역폭 구간에서 우위를 보였으며, 네트워크 대역폭이 높아질수록 연산 효율이 높은 SP2와 SP1으로 최적점이 전환되는 양상을 확인하였다.

Ⅰ. 서 론

최근 자율행동체 및 사물인터넷 기술의 발전으로, 제한된 컴퓨팅 자원을 가진 디바이스에서 복잡한 AI 모델을 실시간으로 실행해야 할 필요성이 급증하고 있다. 이러한 한계를 극복하기 위해 디바이스의 연산 부담을 엣지 서버로 오프로딩하는 기술이 제안되었으나, 원본 데이터 전체를 전송하는 방식은 높은 통신 지연과 프라이버시 문제를 유발할 수 있다. 이를 극복하기 위해 AI 모델을 분할하여 엣지 서버와 로컬 디바이스에서 나누어 처리하는 분할 컴퓨팅(Split Computing) 기술이 주목받고 있다[1].

분할 컴퓨팅의 핵심 과제는 모델의 최적 분할 지점을 선정하는 것이다. 최적의 분할 지점은 엣지 서버와 로컬 디바이스의 연산 성능이나 무선 네트워크의 대역폭 등 동적으로 변하는 요소에 크게 의존한다. 따라서 모든 상황에 적용할 수 있는 정적인 최적 분할 지점은 존재하기 어려우며, 환경 변화에 적응할 수 있는 동적 전략이 필수적이다. 본 논문에서는 이러한 동적 분할 전략의 기반을 마련하고자, 시뮬레이션을 통해 시스템 환경 변화에 따른 최적의 분할 지점을 분석한다.

Ⅱ. 관련 연구

엣지-디바이스 환경에서 DNN 모델을 효율적으로 실행하기 위한 분할 컴퓨팅 연구는 활발히 진행되어 왔다[1]. 분할 컴퓨팅은 DNN 모델을 디바이스에서 실행되는 Head 모델과 엣지 서버에서 실행되는 Tail 모델로 분할하여 연산량과 통신량을 절충하는 것을 목표로 한다. 분할 컴퓨팅의 접근 방법은 크게 모델의 구조를 변경하지 않는 방식과, 통신 효율을 높이기 위해 의도적으로 병목(bottleneck)을 삽입하는 방식으로 나눌 수 있다.

네트워크 수정 없는 분할 방식은 사전 학습된 모델의 구조와 가중치를 그대로 유지한 채 특정 계층을 분할 지점으로 선택하는 기법이다. 이 분야의 초기 연구인 Neurosurgeon[2]은 총 지연 시간과 에너지 소모를 최소화하는 최적의 분할 계층을 탐색했다. 해당 연구의 주요 결과 중 하나는 다수의 DNN 모델에서 최적 분할 지점이 모델의 가장 처음 또는 마지막 계층으로 나타났다는 점이다. 이는 완전한 로컬 컴퓨팅 또는 엣지 컴퓨팅이 중간에서 분할하는 것보다 더 효율적일 수 있으며, 중간 분할이 실질적인 이점을 갖는 조건에 대한 상세한 분석의 필요성을 제기한다.

본 연구는 네트워크 수정 없는 분할 방식을 기반으로 한다. 기존 연구들이 특정 조건에서 중간 분할의 효율성이 낮다고 보인 반면, 본 연구는 디바이스 성능, 네트워크 대역폭 등 다양한 시스템 환경 변수를 시뮬레이션하여 어떤 조건에서 중간 분할 지점이 최적의 전략이 되는지에 대한 정량적인 지표를 제시한다.

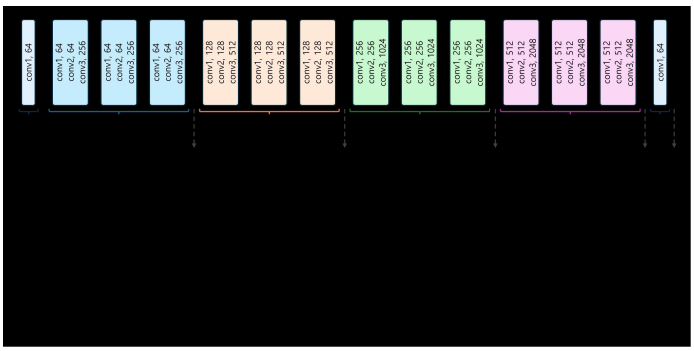

그림 1. ResNet-50 분할 후보 지점 및 후보 지점별 주요 파라미터

Ⅲ. 시뮬레이션 모델링

본 연구의 분석 대상으로는 ResNet-50[3] 모델을 선정하였다. 이는 컴퓨터 비전 분야에서 널리 사용될 뿐만 아니라, 최근 주목받는 비전-언어 모델(VLM)과 같은 더 복잡한 AI 모델의 백본으로 널리 활용되므로, 다양한 모델과 서비스에 확장하여 적용할 수 있는 대표성을 가지기 때문이다.

그림 1은 ResNet-50의 주요 블록 경계를 기준으로 선정한 분할 후보

지점과 각 지점의 핵심 파라미터를 나타낸다. 본 시뮬레이션은 표 1과 같은 조건으로 실험 환경을 구성하였다.

표 1. 시뮬레이션 환경 변수

환경 변수	값
로컬 디바이스 연산 성능(F_D)	5 - 100 GFLOPs
엣지 서버 연산 성능(F_E)	100 GFLOPs
네트워크 대역폭(B)	10 - 1,000 Mbps
분할 지점 후보 집합(K)	[SP1, SP2, SP3, SP4, LC]

분할 실행 시 발생하는 총 지연 시간(T_k)은 디바이스에서의 연산 시간(T_D^k), 데이터 전송 시간(T_{trans}^k), 엣지 서버에서의 연산 시간(T_E^k)의 합으로 정의하였다. 본 연구의 목표는 주어진 환경에서 T_k를 최소화하는 최적의 분할 지점 k^*를 찾는 것이다.

$$T_k = T_D^k + T_{trans}^k + T_E^k$$
$$= \frac{Head\ FLOPs}{F_D} + \frac{Data\ Size}{B} + \frac{Tail\ FLOPs}{F_E}$$
$$k^* = \arg\min_{k \in K} T_k$$

Ⅳ. 결과 및 분석

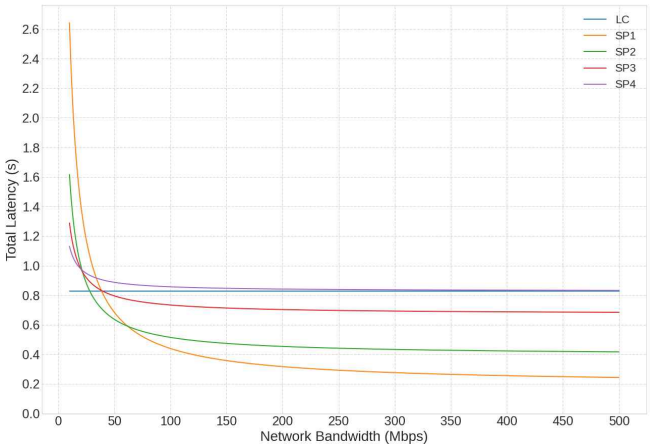

그림 2. 네트워크 대역폭에 따른 분할 지점별 총 지연 시간

본 장에서는 3장에서 정의한 시뮬레이션 결과를 제시하고 분석한다. 먼저 특정 조건에서 각 분할 지점의 성능 변화를 살펴본 뒤, 전체 환경에 대한 최적 지점 분포를 확인한다.

그림 2는 디바이스 성능을 5 GFLOPs로 고정했을 때, 네트워크 대역폭 변화에 따른 각 분할 지점(SP)의 총 지연 시간을 나타낸다. 네트워크를 사용하지 않는 로컬 컴퓨팅(LC)은 대역폭과 무관하게 일정한 지연 시간을 보여주며, 분할을 수행하는 경우 대역폭이 상승함에 따라 지연 시간이 크게 감소하는 경향을 보인다. 특히 네트워크 대역폭이 28Mbps가 넘어가는 지점에서 SP2의 성능이 LC를 넘어서며, 62Mbps가 넘어가는 구간에서는 SP1이 가장 우수한 성능을 보인다.

그림 3은 디바이스 성능과 네트워크 대역폭에 따른 최적 분할 지점을 히트맵으로 나타낸 결과이다. 네트워크 대역폭이 낮은 구간에서는 전송 지연이 시스템의 주된 병목으로 작용하므로 통신이 없는 로컬 컴퓨팅이 최적 전략으로 선택되었다. 네트워크 대역폭이 특정 임계점을 넘어서면, 최적 전략은 LC에서 SP2 및 SP1으로 전환되었다. SP3과 SP4는 모든 구간에서 최적점으로 나타나지 않은 이유는 이들이 갖는 장점이 특정 구간에서 다른 전략에 의해 항상 상쇄되기 때문이다. 네트워크가 느릴 때는 이들의 작은 전송량 이점에도 불구하고, 전송 자체가 없는 로컬 컴퓨팅에 비해 불리하며, 반면 네트워크가 충분히 빨라지는 시점에서는 전송량보다 연산량 감소가 더 중요해지기 때문에 SP1과 SP2에 비해 불리하다.

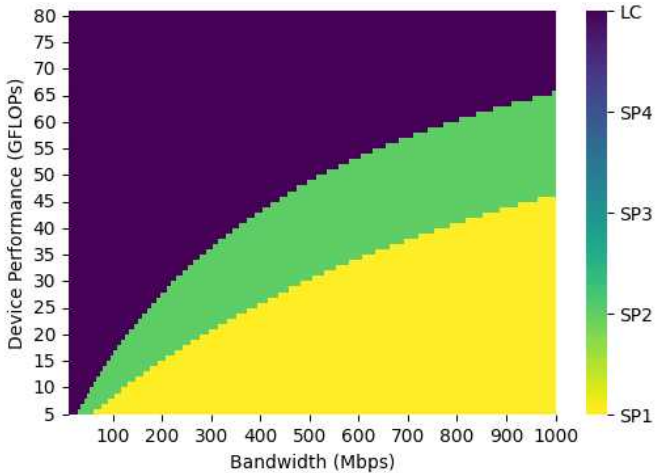

그림 3. 디바이스 성능과 네트워크 대역폭에 따른 최적 분할 지점

Ⅴ. 결 론

본 논문에서는 시뮬레이션을 통해 ResNet-50 모델의 최적 분할 지점이 디바이스 성능과 네트워크 대역폭에 따라 동적으로 변화함을 정량적으로 분석하였다. 이를 통해 실시간 분할 컴퓨팅 시스템에서 현재 시스템 상태에 맞는 최적의 분할 지점을 동적으로 선택하는 정책 수립에 필요한 정량적 근거를 제시하였다.

다만 본 연구는 ResNet-50 모델과 특정 시뮬레이션 환경에 기반하였기에, 다양한 모델 아키텍처 및 실제 네트워크 환경에서의 추가적인 검증이 필요하다. 향후 연구에서는 다양한 모델과 실제 환경에서 검증을 수행하고, 지연 시간 외에 에너지 효율성 및 프라이버시 등 다중 목표를 고려하는 심층강화학습 기반 정책 에이전트를 개발할 계획이다.

ACKNOWLEDGMENT

본 논문은 과학기술정보통신부 및 정보통신기획평가원의 복합지능자율행동체SW핵심기술개발사업(RS-2024-00346798)과 부산광역시 및 (재)부산테크노파크의 BB21plus 사업으로 지원된 연구결과임.

참 고 문 헌

[1] Matsubara, Y., Levorato, M., and Restuccia, F., "Split Computing and Early Exiting for Deep Learning Applications: Survey and Research Challenges," ACM Computing Surveys, vol. 55, no. 5, Article 90, Dec. 2022.

[2] Kang, Y., et al., "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge," in Proc. 22nd Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS '17), pp. 615-629, 2017.

[3] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer

Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.