주파수 자원 관리를 통한 6G 무선 통신망의 인공지능 서비스 오프로딩 최적화 기술에 관한 연구

장종규, 김민우, 양현종*
포항공과대학교

{jgjang, mwkim0210, hyunyang} @ postech.ac.kr

Optimization of AI Service Offloading in 6G Wireless Networks via Radio Resource Management

Jonggyu Jang, Minwoo Kim, Hyun Jong Yang*
Pohang University of Science and Technology (POSTECH)

요 약

본 논문에서는 최근 AI 서비스의 눈부신 발전을 기점으로 시작된 AI 서비스 오프로딩 기술에 관해 다룬다. 기존의 데이터만 주고 받는 통신 시스템과는 다르게, AI 서비스는 컴퓨팅이 주를 이루는 서비스로 통신 자원 뿐 아니라 컴퓨팅 자원의 고려 또한 추가되어야 하는 관계로 새로운 연구의 필요성이 강조되고 있다. 본 논문에서는 먼저 통신/컴퓨팅 자원을 동시에 고려하여 사용자 별 서비스 지연시간을 최적화 하는 문제를 해결하고자 한다. 결과적으로 제안하는 기술은 기존의 Pricing 기반 통신 최적화 기술에서 통용되는 구조에 맞게 디자인되었으며, 적절한 노드 간 서비스 오프로딩 기술을 설계한 결과로 기존의 통신 최적화 기술과 비교하여 짧은 지연시간을 보여준다.

Ⅰ. 서 론

근래 인공지능 기술의 전례 없는 발전에 인해 미래의 클라우드/에지 컴퓨팅 시스템은 신경망 연산과 밀접한 연관이 있는 서비스를 주로 수행할 것으로 예상되고 있다. 현재 컴퓨터 비전과 자연어 처리를 기반으로 수많은 인공지능 서비스들이 상용화 되고 있는 상황이지만, 연산 경량화 기술이 동시에 개발되고 있는 현재에도 신경망으로 인한 계산 부하는 지속해서 과중되고 있다 [1].

에지 컴퓨팅은 중앙 집중형 클라우드 서비스의 서버 과부하를 방지할 수 있을 뿐 아니라 초저지연 서비스를 제공할 수 있다는 장점을 가지고 있다. 이러한 장점들에 힘입어 에지 컴퓨팅은 차세대 네트워크의 중요한 요소기술로 자리잡고 있으며, 다양한 연구팀에서 연구주제로 선택하고 있는 추세다 [2].

하지만, 에지 컴퓨팅 역시 중앙 클라우드 컴퓨팅 시스템과 비교하자면 스케줄링 복잡도가 증가하며, 분산 배치의 특성 상 높은 설치 비용으로 인해 무한정 고성능의 서버를 설치하지 못한다는 단점이 있다. 에지 컴퓨팅과 클라우드 컴퓨팅은 서로의 장/단점을 상호 보완하고 있으며, 신경망 서비스를 위해서는 에지/클라우드 컴퓨팅이 혼용되는 상황에서 AI 서비스의 요구치에 따른 서비스를 노드에게 할당하는 기술을 개발할 필요성이 있음.

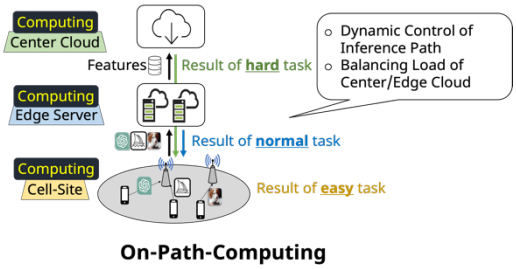

그림 1. early-exit neural network 활용 클라우드/에지 컴퓨팅 시스템의 모식도.

예를 들자면, 하나의 휴리스틱 메소드는 컴퓨팅 서비스가 고복잡도일수록 중앙 클라우드로 스케줄링하며, 저지연 서비스일 수록 에지 노드로 오프로딩하는 방식을 고려해볼 수 있다. 이점이 기존의 통신 네트워크 최적화 기술들 [3, 4]과 크게 차별되는 점으로, 두 가지 자원 (컴퓨팅 및 네트워킹) 이 서로 다른 도메인에서 동작할 뿐 아니라 소모 자원 간 큰 연관 관계가 없기 때문에 새로운 최적화 기술이 필요한 시점이다.

본 논문에서는 사용자가 요청한 서비스가 간단한 연산으로 완료될 수 있을 경우 신경망의 일부분 만을 연산하여 결과를 도출 하는 early-exit 신경망 기술을

활용하여 네트워크 전체의 연산 부하를 낮추며, 효과적인 스케줄링을 통해 사용자의 평균 지연시간을 크게 단축시킬 수 있는 방법을 제안한다.

Ⅱ. 본론

본 논문에서는 에지/클라우딩 컴퓨팅이 혼재되어 있는 네트워크를 고려하며, 요청되는 서비스는 classification 작업이며, early-exit neural network 모델은 3 개의 출력단을 가지는 것으로 가정한다. 또한, 모바일 사용자의 경우는 로컬에서 연산을 진행할 경우 배터리 소모가 커지는 trade-off 를 고려하기 위해 에너지 소모에 대한 penalty term 을 고려하였다. 그림 1 은 고려된 시스템 모델의 예시이며, 여기서 사용자/에지/클라우드는 각각 3 개의 신경망 출력 단 중 입력에 가까운 순서대로 연산하여 서비스한다. 다시 말하자면, cell-site 에서는 입력과 가장 가까운 신경망 연산을 수행하여 추론하기 쉬운 작업을 수행한다.

본 논문에서는 이 시스템 모델에서 지연시간을 최소화할 수 있는 최적화 문제를 만든 후 해결하여 아래와 같은 솔루션을 얻었다.

$$x_{ij} = \begin{cases} 1, & j = argmin_j(G-H) \ and \ (G-H) \leq 0 \\ 0, & o.w. \end{cases}$$

여기서, x_{ij} 는 하위 노드 i 와 상위 노드 j 간 연결 상태를 나타내며, G 와 H 는 각각 상위 노드의 부하와 서비스 요청량에 대한 비율과 로컬 디바이스의 배터리 소모를 나타낸다. x_{ij} 가 1 이면 연결되어 서비스를 요청하고, 0 이면 연결되지 않는다. 만일 $G-H$ 의 값이 모두 0 이하라면 상위 노드에 전달하지 않고 해당 노드에서 모든 연산을 처리한다. 변수 G 와 H 는 각각 네트워크에서 상위 노드의 부하의 정도와 로컬 디바이스의 부하 정도에 따라 업데이트 된다.

Ⅲ. 결과

이 부분에서는 제안하는 기술을 기존의 노드 연결 방식과 비교하여 제안하는 기술의 우수성을 검증한다. 비교를 위해 고려된 알고리즘은 통신 퀄리티가 좋은 노드로 우선적으로 서비스를 할당/요청하는 Max-SINR 과 랜덤 노드에게 서비스를 할당/요청하는 Random 알고리즘들이다. 에지 네트워크에 컴퓨팅 노드와 연결된 6 대의 기지국이 있다고 가정하였으며, 사용자의 수는 40 명부터 200 명까지를 고려하였다. 중앙 집중형 클라우드의 수는 1 대로 가정되었으며, 보수적으로 각 연산이 끝날 때까지의 지연시간은 일반적인 서버 컴퓨터에서 측정된 지연시간을 바탕으로 12ms 로 설계되었다.

그림 2 는 사용자 수를 변화시켜가며 서비스 별 평균 지연시간을 측정한 그래프를 나타낸다. 결과적으로, 제안하는 기술은 기존 알고리즘 대비 훨씬 낮은 지연시간을 가질 뿐 아니라, 신경망을 끝까지 연산하는 시간 대비 더 짧은 시간 (10ms 이하)를 가질 수 있다.

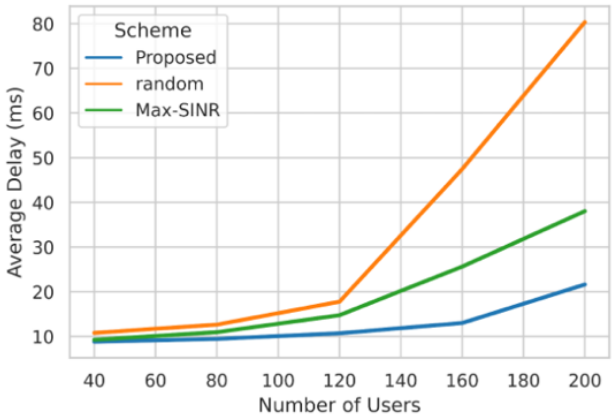

그림 2. 사용자 수 변화에 따른 서비스 당 평균 지연시간

Ⅲ. 결론

본 논문에서는 통신/컴퓨팅 자원을 동시에 고려하여 AI 서비스를 오프로딩하는 자원 관리 기술을 제안하였다. 로컬 디바이스의 배터리 소모를 고려하여 지연시간을 최소화 할 수 있는 최적화 문제를 제안하여 문제를 해결하였으며, 결과적으로 기존 알고리즘 대비 눈에 띄게 지연시간을 줄일 수 있었다. 또한, 기술에서 제안하는 솔루션은 기존의 pricing-based offloading 의 프레임 워크에서 통용되는 형태로 기존 기술들과 호환되어 운용 가능할 것으로 기대된다.

ACKNOWLEDGMENT

본 논문은 선박해양플랜트연구소 기본사업 “다개체 해양 로봇의 협력 항법 및 수중 무선 인지 네트워크 핵심 기술 개발(PES5180)" 로 수행된 연구결과입니다.

참 고 문 헌

[1] Coutinho, R. W., & Boukerche, A. Design of edge computing for 5g-enabled tactile internet-based industrial applications. *IEEE Communications Magazine*, 2022.

[2] XU, Wei, et al. Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing. *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[3] Kim, Y., Jang, J. and Yang, H.J., 2023. Distributed Resource Allocation and User Association for Max-Min Fairness in HetNets. *IEEE Transactions on Vehicular Technology*, 2023.

[4] Mawatwal, K., Roy, R., & Sen, D. A state based resource allocation game for distributed optimization in 5G small-cell networks. *IEEE Transactions on Vehicular Technology*, 2021.