쿠버네티스 기반 연합학습 및 스플릿 컴퓨팅의 최신 연구 동향

임동건[§], 정택준[§], 김주령[§], 방인규[†■], 김태훈[§■]

[§]국립한밭대학교 컴퓨터공학과, [†]국립한밭대학교 지능미디어공학과

{20211893, 20211929, 20222021}@edu.hanbat.ac.kr, {ikbang, thkim}@hanbat.ac.kr

Recent Research Trends in Kubernetes-Enabled Federated Learning and Split Computing

Donggeon Im[§], Taekjun Jeong[§], Juryeong Kim[§], Inkyu Bang[†■], Taehoon Kim[§■]

[§]Department of Computer Engineering, Hanbat National University

[†]Department of Intelligence Media Engineering, Hanbat National University

요 약

연합 학습(Federated Learning, FL)과 스플릿 컴퓨팅(Split Computing)은 데이터 프라이버시를 보존하고 분산된 엣지 환경에서 인공지능(AI) 모델을 효율적으로 운영하기 위한 핵심 패러다임이다. 특히, 쿠버네티스(Kubernetes, K8s)는 컨테이너화된 워크로드의 오케스트레이션 및 관리를 위한 표준 플랫폼으로 부상하면서, 이 두 분야의 확장성과 효율성을 높이는 데 중추적인 역할을 하고 있다. 이 논문은 쿠버네티스 기반의 FL 플랫폼 및 프레임워크, 그리고 대규모 기반 모델(Large Foundation Model, LFM) 추론을 위한 동적 스플릿 컴퓨팅 오케스트레이션에 대한 최신 연구 동향을 조사한다. K8s의 컨테이너 오케스트레이션 기능이 데이터 이질성 및 시스템 이질성을 포함하는 엣지 환경의 동적 자원 관리, 프라이버시 보존, 그리고 서비스 품질(Quality of Service, QoS) 보장에 어떻게 기여하는지 분석한다.

Ⅰ. 서 론

분산된 엣지 디바이스에서 수집되는 데이터의 증가와 함께 데이터 프라이버시 및 보안에 대한 요구가 높아지면서, 중앙 집중식 모델 학습의 대안으로 연합 학습(FL)이 주목받고 있다 [2, 5]. FL은 데이터를 로컬에 유지한 채 모델 업데이트만 공유하여 중앙 서버에서 통합하는 협업적 모델 학습 방식을 가능하게 한다 [3]. 나아가, 대규모 AI 모델(LFM)의 등장으로 인해, 자원 제약적인 엣지 환경에서 LFM 추론을 효율적으로 수행하기 위한 분산 스플릿 추론(Distributed Split Inference, DSI), 즉 스플릿 컴퓨팅이 중요해졌다 [4]. 이 방식은 모델의 레이어를 여러 노드에 분할하여 순차적으로 실행함으로써 클라이언트 장치의 계산 부담을 줄인다.

쿠버네티스(K8s)는 이러한 분산 AI 워크로드를 관리하고 확장하는 데 이상적인 플랫폼을 제공한다 [1, 3]. K8s는 애플리케이션의 배포, 확장, 관리를 용이하게 하며, 특히 엣지 환경에서의 계산 탄력성(computational elasticity)과 효율성을 향상시키는 데 기여한다 [5]. 그러나 K8s의 기본 네트워크 모델(flat networking)은 잠재적인 프라이버시 위험을 초래할 수 있으며 [2], 이질적인 엣지 노드와 동적인 워크로드에 대한 효율적인 자원 할당 및 QoS 보장이라는 과제가 남아있다 [5, 4].

이 논문은 K8s를 활용하여 FL의 프라이버시 및 관리 용이성을 개선하는 연구와 [2, 5], 스플릿 컴퓨팅에서 동적 오케스트레이션 및 QoS를 달성하는 최신 방법론을 탐구한다 [4].

Ⅱ. 쿠버네티스 기반 연합 학습 환경의 발전 동향

FL 시스템을 K8s 기반으로 구축하는 최신 연구는 프라이버시 강화 및 동적 자원 관리라는 두 가지 핵심 영역에서 발전하고 있다.

1. 시스템 격리 및 데이터 프라이버시 강화 기법

FL의 핵심 원칙은 데이터 프라이버시 보존이지만, K8s의 기본 네트워킹 모델은 모든 파드(Pod) 간의 통신을 허용하여 악의적인 FL 클라이언트가 다른 리소스에 접근할 수 있는 잠재적인 위험을 야기한다 [2]. kubeFlower 프레임워크는 K8s 기반 FL 환경에서 프라이버시를 보존하기 위해 Isolation-by-design 원칙을 구현한다 [2]. 이 원칙은 소프트웨어 정의 네트워크(SDN) 기법을 적용하여 논리적으로 격리된 가상 네트워크 파티션(Virtual Network Partitions)을 생성함으로써, FL 클라이언트 파드가 오직 FL 서버 파드와만 통신하도록 제한하여 클라이언트 간의 교차 접근을 근본적으로 차단한다 [2]. 또한, kubeFlower는 Privacy-Preserving Persistent Volume Claims (P3-VC) 개념을 도입하여 FL 클라이언트의 데이터셋에 차분 프라이버시(DP)를 적용하며, K8s 환경에서 데이터 마운트 시 프라이버시 메커니즘을 통합한다.

2. 동적 자원 할당 및 탄력적 확장성 확보 방안

엣지 컴퓨팅 환경의 시스템 이질성과 데이터 이질성은 FL의 효율성에 주요한 제약이 되며, K8s의 자동 확장 기술이 이 문제를 해결하는 핵심 도구로 사용된다 [5]. Elastic FL (EFL) 프레임워크는 Kubernetes Vertical Pod Autoscaler (VPA)를 활용하여 FL 애플리케이션의 성능을 향상시킨다 [5]. VPA는 파드의 실시간 및 이력 자원 사용량을 기반으로 CPU 및 메모리 요청(request)과 제한

■ Corresponding Authors: Inkyu Bang (ikbang@hanbat.ac.kr), Taehoon Kim (thkim@hanbat.ac.kr)

(limit)을 동적으로 조정하여 자원 할당을 최적화한다 [5]. 이 방식은 자원 부족으로 인한 메모리 부족(OOM) 오류 및 파드 퇴거(Eviction)를 방지하여 학습 안정성을 높이고, 이질적인 데이터 샘플 크기에 비례하여 자원을 효율적으로 할당함으로써 FL의 계산 탄력성을 확보한다 [5]. 한편, FedEdge [3] 및 Flautim [1]과 같은 플랫폼들은 K8s와 Flower, TensorFlow FL을 통합하여 FL 실험의 이식성, 확장성, 사용자 친화성을 높이는 데 중점을 둔다.

Ⅲ. 대규모 모델 추론을 위한 동적 스플릿 컴퓨팅 오케스트레이션

FL이 모델 학습에 중점을 둔다면, 스플릿 컴퓨팅은 주로 대규모 기반 모델(LFM) 추론의 효율성과 QoS 보장에 초점을 맞춘다 [4]. 기존의 분산 스플릿 추론(DSI) 전략은 엣지 환경의 동적인 변화에 적응하지 못하는 정적 분할(static split) 방식을 채택하여 성능 저하 문제를 야기한다 [4].

1. 적응형 배치 및 동적 재분할 프레임워크 설계

최신 연구는 LFM 레이어의 배치(placement)와 분할(partitioning)을 런타임에 동적으로 조정하는 적응형 스플릿 추론 오케스트레이션 프레임워크를 제안한다 [4]. 이는 K8s의 컨테이너 수준 관리를 넘어 LFM 그래프 구조를 인식하고 제어하는 세분화된 접근 방식이다. 이 프레임워크는 LFM 추론 워크로드를 관리하기 위해 세 가지 핵심 기능을 통해 네트워크 변동 및 노드 부하와 같은 동적 환경 변화에 실시간으로 대응한다. 첫째, 용량 인식 워크로드 분배는 지속적인 노드 자원 프로파일링을 기반으로 최적의 엣지 노드 부분 집합을 선택하는 역할을 한다. 둘째, 동적 파티션 마이그레이션은 활용률이 높아지거나 네트워크 조건이 변화할 때, 미리 분할된 LFM 세그먼트를 투명하게 재배치하여 병목 현상을 방지한다. 셋째, 실시간 재구성(동적 재분할)은 레이어를 재분할하여 대기 시간(L), 처리량, 그리고 프라이버시를 균형 있게 맞추도록 보장한다. 오케스트레이션 결정은 대기 시간(L), 자원 사용 불균형 또는 노드 과부하(U), 그리고 프라이버시 위반 페널티(P)를 최소화하는 다목적 함수 Φ를 기반으로 설계된 최적화 문제를 통해 이루어진다. 특히, 민감한 데이터를 처리하는 파티션이 신뢰할 수 있는 노드에 배치되도록 보장하는 프라이버시-인식 계층 배치를 통해 규정 준수 및 보안을 확보한다 [4].

2. QoS 기반 성능 개선 및 정량적 평가

적응형 스플릿 추론은 동적인 환경 변화에 대응함으로써 정적 방식 대비 종단 간 대기 시간을 50% 이상 감소시키며, 동시에 처리량을 5배 이상 향상시키는 정량적 성능 이점을 보인다 [4]. 이 방식은 엣지 AI 서비스의 QoS 목표(SLA) 준수율을 95% 이상으로 유지하는 등 월등한 안정성과 효율성을 입증하며, 동적 오케스트레이션이 LFM 추론의 성능을 확보하는 데 결정적인 역할을 함을 시사한다.

Ⅳ.결론

쿠버네티스를 활용한 연합 학습 및 스플릿 컴퓨팅 연구는 분산 AI 시스템의 효율성, 확장성, 그리고 프라이버시 보존이라는 핵심 과제를 해결하는 데 있어 중요한 진전을 이루었다.

FL 분야에서 kubeFlower와 EFL 프레임워크는 K8s의 오케스트레이션 능력에 프라이버시-강화 메커니즘(isolation-by-design, DP)과 동적 자원 관리(VPA)를 성공적으로 통합함으로써 이질적인 엣지 환경에 대한 실질적인 솔루션을 제시했다 [2, 5].

스플릿 컴퓨팅 분야에서는 LFM 추론의 QoS 보장을 위해 모델의 배치와 분할을 런타임에 동적으로 조정하는 적응형 오케스트레이션이 필수적인 방향으로 자리 잡고 있다 [4].

향후 연구는 AI 기반의 의사 결정 메커니즘(예: 강화 학습)을 오케스트레이션에 더 깊이 통합하여 추론 최적화를 강화하고, 보안 멀티파티 컴퓨테이션(Secure Multi-Party Computation, MPC) 및 동형 암호화(Homomorphic Encryption)와 같은 고급 프라이버시 보존 기술을 분산 추론 환경에 적용하는 데 중점을 둘 것이다 [4]. 또한, 자원 인식 클라이언트 배치(resource-aware client placement)를 통해 FL 성능을 더욱 최적화하고, 6G 네트워크의 핵심 목표에 부합하는 적응형 네트워크 인식 분할 전략을 개발하는 것이 중요한 연구 과제이다 [4].

ACKNOWLEDGMENT

본 과제(결과물)는 2025년도 교육부 및 대전광역시의 재원으로 대전RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.(2025-RISE-06-002)

참 고 문 헌

[1] P. H. S. S. Barros, M. Q. A. Oliveira, O. Orang, F. A. R. da Silva, F. J. Erazo-Costa, A. T. Bastos, P. C. L. Silva, G. S. dos Santos, A. A. F. Loureiro, M. Gómez Ravetti, M. A. Costa, F. G. Guimarães, H. S. Ramos, “Flautim: A Federated Learning Platform using K8S and Flower,” XXII Workshop de Ferramentas e Aplicações (WFA 2024), Brazilian Computer Society, 2024.

[2] J. M. Parra-Ullauri, H. Madhukumar, A.-C. Nicolaescu, X. Zhang, A. Bravalheri, R. Hussain, X. Vasilakos, R. Nejabati, D. Simeonidou, “kubeFlower: A privacy-preserving framework for Kubernetes-based federated learning in cloud–edge environments,” Future Generation Computer Systems, vol. 157, pp. 558–572, 2024.

[3] M. Hassan, L. L. Custode, K. S. Yildirim, G. Iacca, “FedEdge: Federated Learning with Docker and Kubernetes for Scalable and Efficient Edge Computing,” Dept. of Information Engineering and Computer Science, University of Trento, Italy, 2023.

[4] F. Koch, A. Djuhera, A. Binotto, “Intelligent Orchestration of Distributed Large Foundation Model Inference at the Edge,” arXiv preprint arXiv:2504.03668, 2025.

[5] K. Q. Pham, T. Kim, “Elastic Federated Learning with Kubernetes Vertical Pod Autoscaler for edge computing,” Future Generation Computer Systems, vol. 158, pp. 501–515, 2024.