로컬 셀프 어텐션 업데이트 및 앙상블 증류를 통한 트랜스포머용 개인화 연합 학습에 관한 연구

황지영, 이주형*
가천대학교

hjy0714@gachon.ac.kr, j17lee@gachon.ac.kr*

Personalized Federated Learning for Vision Transformers via Local Self-Attention Updates and Historical Ensemble Distillation

Jiyoung Hwang, Joohyung Lee*
Gachon Univ.

요 약

본 논문은 비전 트랜스포머(ViT)의 구조적 특성을 활용하여 개인화와 일반화 성능을 동시에 향상시킨 새로운 개인화 연합 학습(PFL) 프레임워크인 **FedSAH (Federated Self-Attention with Historical Buffer in Federated Learning)**를 제안한다. 클라이언트 간 데이터 분포가 상이한 Non-IID 환경에서 발생하는 글로벌 모델의 성능 저하 문제를 해결하기 위해, 본 연구는 과거 글로벌 모델들의 이력 앙상블(Historical Ensemble)과 지식 증류(Knowledge Distillation) 기반의 로컬 셀프 어텐션 계층 최적화 기법을 도입한다. 실험 결과, FedSAH 는 기존 FedAvg 방식 대비 약 15 배 빠른 수렴 속도를 기록하였으며, 데이터 이질성이 심한 조건에서도 높은 정확도와 견고한 일반화 성능을 달성하였다.

Ⅰ. 서 론

스마트 기기의 폭발적인 보급과 함께 발생하는 방대한 양의 데이터는 중앙 집중식 학습 방식에 통신 비용 및 프라이버시 침해라는 문제를 야기한다. 이러한 배경에서 데이터를 로컬 장치에 유지하며 협력적으로 모델을 학습시키는 연합 학습(Federated Learning, FL)이 유망한 대안으로 주목받고 있다. 그러나 클라이언트 간 데이터 분포가 다른 Non-IID 환경에서 글로벌 모델은 모든 사용자에게 균일하게 일반화되지 못하며, 결과적으로 많은 사용자에게 최적화되지 않은 상태에 머물게 된다. 이를 해결하기 위해 각 클라이언트가 로컬 데이터에 특화된 표현을 학습하고 유지하는 개인화 연합 학습(PFL)이 제안되었다 [1]. 그러나 기존 PFL 연구는 대부분 합성곱 신경망(CNN)을 중심으로 발전해왔기 때문에 비전 트랜스포머(Vision Transformer) 고유의 아키텍처 특성을 반영한 개인화 기법은 초기 단계에 머물러 있다 [2]. 최근 연구를 통해 비전 트랜스포머의 셀프 어텐션(Self-Attention) 계층이 도메인 특화 정보를 학습하는 개인화의 핵심 요소임이 확인되었으나 [3], 이를 단순히 로컬에서만 갱신하는 방식은 글로벌 공유 지식과의 일관성을 저해하여 결과적으로 글로벌 표현의 붕괴와 일반화 성능 저하라는 한계를 야기한다. 따라서 본 연구에서는 이를 극복하기 위해 이력 앙상블과 지식 증류를 결합하여 ViT의 개인화와 일반화를 동시에 달성하는 FedSAH 프레임워크를 제안한다.

Ⅱ. 본론

1. FedSAH 의 상세 설계 및 방법론

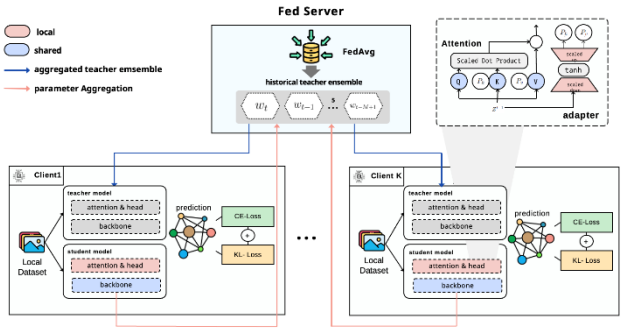

그림 1. FedSAH 아키텍처: 공유 백본(Patch Embedding, MLP)과 로컬 개인화 모듈(Self-Attention, Classifier)의 분리 구조.

본 논문에서 제안하는 FedSAH 프레임워크는 그림 1 과 같이 모델 파라미터를 공유 백본과 로컬 개인화 파라미터로 분리하여 관리하며, 다음의 3 단계 과정을 거친다. 첫 번째 단계인, **선택적 집합(Selective Aggregation)**에서 클라이언트는 로컬 학습 후 백본 파라미터 만을 서버로 전송하며, 수식 (1)과 같이 서버는 이를 가중 평균하여 글로벌 모델을 갱신한다.

$$u^{(t+1)} = \sum_{i=1}^{K} p_i u_i^{(t)} \qquad (1)$$

이어서 서버는 **이력 교사 앙상블 (Historical Teacher Ensemble)**을 수행하며, 최신 글로벌 모델과 버퍼에 저장된 과거 M개의 모델을 가중 앙상블 하여 수식 (2)와 같이 '이력 교사' 모델을 구성한다.

$$u_{teacher}^{(t)} = \sum_{j=0}^{M} \beta_j \cdot u^{(t-j)} \qquad (2)$$

이는 Non-IID 환경에서 발생할 수 있는 라운드 간 업데이트 변동성을 줄이고 최적화 궤적을 안정화하여 글로벌 표현의 붕괴를 방지한다. 마지막으로 각 클라이언트는 **지식 증류 기반 업데이트(Knowledge Distillation)**를 통해 이력 교사 모델로부터 지식을 흡수하며 로컬 데이터에 최적화된 학습을 수행한다. 전체 손실 함수는 수식 (3)과 같다.

$$\min_{h_i} L_{total} = L_{CE}(f(x; h_i), y) + \lambda_{KD} D_{KL}\left(q\left(x; u_{teacher}^{(t)} \cup v_i\right) \,||\, q(x; h_i)\right) \qquad (3)$$

2. 실험 환경 및 성능 분석 결과

본 연구의 성능을 검증하기 위해 CIFAR-100 데이터셋을 바탕으로 Dirichlet (alpha=0.25) Non-IID 환경에서 실험을 수행하였다. 제안된 FedSAH 의 성능을 평가하고자 표준 연합 학습 알고리즘인 FedAvg, 최신 개인화 기법인 FedPerfix, 단순 부분 집합 방식인 Vanilla Attention 과 비교를 진행하였다.

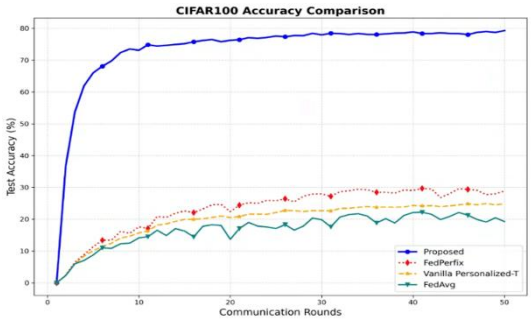

그림 2. 알고리즘별 정확도 비교

실험 결과, 알고리즘별 정확도 FedSAH 는 이력 앙상블을 통해 라운드 간 업데이트 변동성을 줄여 안정적인 글로벌 특징 학습을 보장하며, 50 라운드에 79%라는 높은 정확도에 도달하였다.

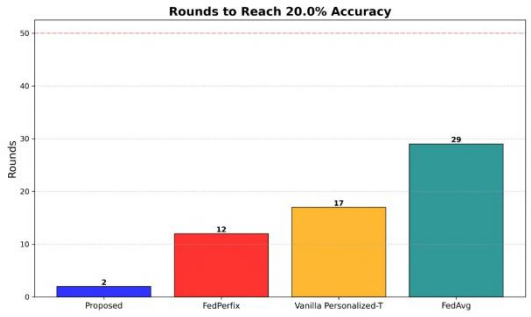

그림 3. 알고리즘별 수렴 성능 비교

수렴 속도 비교에서도 제안된 FedSAH 는 기존의 대표적 알고리즘인 FedAvg 대비 약 15 배 빠른 수렴 속도를 기록하였다. 구체적으로 목표 정확도인 20%에 도달하는 데 있어 FedAvg 는 29 라운드가 소요되었으나, FedSAH 는 단 2 라운드 만에 도달하여 압도적인 통신 효율성을 보여주었다.

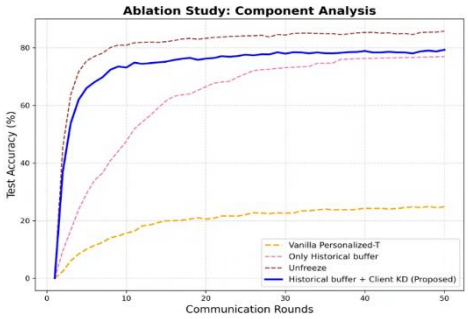

그림 4. 주요 구성 요소별 성능 기여도 분석 결과

소거 연구(Ablation Study) 결과, 이력 버퍼가 성능 향상의 가장 결정적인 요인인 것으로 확인되었으며, 과거 모 수렴을 유지할 수 있었다. 마지막으로 로컬 업데이트 시 백본 전체를 해제(Unfreeze)하여 학습하는 방법이 로컬 데이터에 대한 적응력을 높여 추가적인 성능 향상을 이끌어냈다.

Ⅲ. 결론

본 논문은 ViT 의 구조적 특성을 활용해 개인화와 일반화를 동시에 달성하는 새로운 연합 학습 프레임워크 FedSAH 를 제안하였다. 과거 모델의 이력 앙상블을 교사 모델로 활용하는 지식 증류 기법을 도입하여 Non-IID 환경에서도 학습의 안정성을 확보했다. 실험 결과 기존 FedAvg 대비 15 배 빠른 수렴 속도를 기록하며, 통신 자원이 제한된 실제 엣지 환경에서의 높은 실용성을 입증하였다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2026 년도 SW 중심대학사업의 결과로 수행되었음" (2021-0-01389)

참 고 문 헌

[1] Michael Zhang, Karan Sapra, S. Fidler, Serena Yeung, and J.` Alvarez. Personalized federated learning with first order model optimization. ArXiv, 2020.

[2] Hongxia Li, Zhongyi Cai, Jingya Wang, Jiangnan Tang, Weiping Ding, Chin-Teng Lin, and Ye Shi. Fedtp:Federated learning by transformer personalization. IEEE Transactions on Neural Networks and Learning Systems, 35(10):13426-13440, 2024

[3] Sun, G., Mendieta, M., Luo, J., Wu, S., & Chen, C. (2023). FedPerfix: Towards partial model personalization of vision transformers in federated learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 4988-4998)