저빈도 통신 활용 Split Federated Learning 기법을 통한 다중 디바이스 학습 효율 최적화

배준우, 김윤지, 안진현*

명지대학교

ag660340@mju.ac.kr, yoonji1004@mju.ac.kr,

*wlsgus3396@mju.ac.kr

A Study on the Optimization of Learning Efficiency in Multi-Device Environments via Split Federated Learning with Low-Frequency Transmission

JunWoo Bae, Yoonji Kim, Jin-Hyun Ahn*

Myongji Univ.

요 약

본 연구에서는 smashed data 및 gradient 교환을 위한 통신을 저빈도로 수행하여 서버-디바이스 간 통신 효율을 향상시키고 학습 성능 저하를 최소화하는 Split Federated Learning(SFL) 기법을 제안한다. 해당 기법은 초기 구간에 일반적인 분할 학습 (Split learning) 방식을 통해 학습을 진행한 뒤, 특정 라운드 이후부터 데이터 교환 빈도를 낮추는 방식을 적용한다. 이를 통하여 종단 디바이스에서 부담하는 통신량을 감소시키면서도 모델 성능 유지를 최대화하는 것을 목표로 한다. 결과적으로 자원 제약이 큰 분산 디바이스 환경에서도 모델 학습을 효율적으로 수행할 수 있도록 하는 알고리즘을 제시한다.

Ⅰ. 서 론

Split Federated Learning(SFL)[1]은 다중 디바이스 환경에서의 모델 학습을 위한 방법론으로써 연합 학습(Federated Learning)[2]의 병렬성과 분할 학습(Split Learning)[3]의 모델 분할 방식을 결합한 알고리즘이다. 이에 따라 학습에 참여하는 각 디바이스는 신경망을 특정 지점을 기준으로 분할하여 입력단을 포함하는 얕은 서브 모델(이하 전단부 모델)을, 서버는 출력단을 포함하는 깊은 서브 모델(이하 후단부 모델)을 보유한다. 이러한 모델 구조를 갖는 각 디바이스는 개별적으로 학습을 진행한 파라미터를 업로드하며 서버에서는 이를 FedAvg 알고리즘을 통해 전역 모델을 업데이트한다. 앞서 언급한 특징들로 인해 개별 디바이스는 모델 학습에 참여하기 위하여 요구 받게 되는 연산 및 메모리 자원에 대한 부담이 감소한다. 일반적으로 제한된 조건을 갖는 엣지 디바이스 환경에서도 신경망 모델 학습 진행이 보다 수월하게 되는 것이며 또한 개별 서버-디바이스 상호작용 과정이 순차적으로 진행되는 분할 학습과 달리 SFL에서는 해당 과정이 독립적으로 진행 된 후 동기화가 진행되기 때문에 기존의 학습 지연 및 효율 문제를 개선할 수 있다.

그러나 SFL은 매 미니배치에 대한 디바이스 측 모델의 중간 출력값인 smashed data를 서버로 전송하고 서버가 계산한 gradient를 다시 역으로 반환하는 직접적인 데이터 교환 구조 기반 업데이트 방식을 유지한다. 이러한 고빈도의 서버-디바이스 상호작용은 필연적으로 매우 큰 통신 오버헤드를 발생시키게 되며, 통신 자원이 충분하지 않은 네트워크 환경에서 전체 학습 시스템의 지연과 학습 효율을 저해하는 주된 병목 요인으로 작용한다. 특히 참여 클라이언트의 수가 증가할수록 통신 비용의 선형적 증가로 인해 실시간 학습 환경에서의 효율이 심각하게 저하되는 한계를 가진다. 따라서 본 논문은 SFL 상호작용 기반 업데이트 방식의 장점을 유지하면서도 통신에 따른 부하를 감소시키기 위해 smashed data와 gradient 교환을 특정 시점 이후 저빈도로 수행하는 기법을 제안한다. 이는 학습 초기 단계에서 기존과 같이 매 배치에 대한 데이터 통신을 통해 모델의 수렴 기반을 확보하고, 안정화 단계 이후에는 통신 주기를 제어하여 학습 안정성과 자원 효율성 사이 최적점을 찾는 것을 목표로 한다.

Ⅱ. 본론

Ⅱ-a) 제안 기법

제안하는 저빈도 통신 기반 SFL 기법은 서버 측에서 FedAvg 알고리즘으로 모델 집계를 수행하는 SFLv1 체계를 기본 프레임워크로 활용하며, 전체 학습 과정을 고빈도 통신 구간과 저빈도 통신 구간으로 분리하여 수행한다. 학습 초기인 고빈도 구간에서는 분할 학습(이하 SL) 방식을 적용하여 매 미니배치 단위로 서버와 디바이스 간의 즉각적인 데이터 교환 및 이를 통한 업데이트가 이루어지도록 한다. 디바이스는 전단부 모델의 순전파 연산으로 생성된 smashed data를 서버로 전송하며, 서버는 이에 대해 후단부 모델에서 역전파를 수행하여 생성된 gradient를 전송한다. 디바이스는 전송받은 gradient를 이어서 역전파한 후 전단부 모델을 업데이트하는 과정을 반복한다. 이 과정은 학습 초반에 서버로부터 충분한 피드백 정보를 확보함으로써 모델이 디바이스가 갖는 로컬 데이터에 대한 경향성을 충분히 학습하는 방향으로 학습되어 이후 저빈도 통신 단계를 거치더라도 안정적으로 수렴할 수 있도록 하는 역할을 한다. SL의 단일 글로벌 라운드 학습에 참여하는 디바이스들은 각자의 로컬 데이터를 사용하여 일반화 성능을 개선하는 방향으로 순차적 서버-디바이스 상호 작용 기반 전역 모델 업데이트를 진행하게 된다. 이때 전단부 모델과 후단부 모델에 대한 전역 업데이트는 각 서버-디바이스 상호 작용마다 수행되며, 이러한 과정에서 FedAvg 와 같은 알고리즘은 사용되지 않는다.

설정된 특정 라운드에 도달하면 통신 비용 최적화를 위한 저빈도 통신 단

계로 전환된다. 후반 구간에서는 모든 배치에 대한 데이터 교환을 수행하는 대신, 사전에 정의된 방식을 통하여 데이터 교환 빈도를 의도적으로 낮추는 방식으로 업데이트를 진행한다. 디바이스는 모든 로컬 데이터의 미니배치에 대해 전단부 모델의 순전파 연산만을 독립적으로 수행한 후, 한 번의 통신을 통하여 서버측으로 smashed data를 일괄 전송한다. 서버에서는 이에 대응하는 gradient들을 후단부 모델 역전파로 생성해 동일하게 디바이스로 일괄 전송한다. 이때 서버와 디바이스는 통신이 이뤄지는 동안 즉시 모델 업데이트를 수행하지 않고 smashed data와 gradient의 교환이 완료된 이후 교환된 데이터를 기반으로 전단부 및 후단부 모델을 업데이트한 후 개별 글로벌 라운드 종료 시점에 집계를 수행한다. 이러한 저빈도 통신 기반 업데이트 방식은 네트워크 지연이 빈번한 실제 환경에서 서버와 디바이스 간의 데이터 교환에 따른 전송 횟수를 줄이고 사용가능한 네트워크 자원을 효율적으로 사용할 수 있도록 한다. 이러한 실시간으로 매 배치에 대하여 업데이트되는 전단 및 후단부 모델을 통해 생성된 smashed data와 gradient를 사용하는 것이 아닌 특정 통신 주기에서 생성된 smashed data와 gradient 기반 업데이트 방식을 채택함으로 인해 실시간성이 저하되고 이로 인해 모델 수렴이 불안정해질 수 있게 된다. 이를 제어하기 위해, 통신 방식이 전환되는 시점부터 학습률을 초기 설정값의 10% 수준으로 대폭 감쇠시킴으로써 초기 단계에서 큰 학습률을 통해 로컬 데이터의 경향에 대한 충분한 학습을 수행하고, 저빈도 통신 단계에서는 낮은 학습률로 학습된 경향성에 대해 정교한 미세 조정을 수행하여 성능 하락을 최소화하는 데에 도움을 주도록 설계하였다. 이러한 주장에 대한 근거는 실험 단계의 그림 1.에서 학습률 조정 여부를 다르게 수행한 비교 실험을 통해서 제시한다.

I-b) 실험

알고리즘의 실증적 검증을 위해 제안 기법과 분할 학습에 대해 ResNet-18 모델과 CIFAR-10 데이터셋을 이용하여 이미지 분류 성능 실험을 수행하였다. CIFAR-10의 50,000개 학습 데이터를 50명의 클라이언트에게 균등하게 분배하여 각 디바이스가 1,000개의 이미지 데이터 및 레이블을 갖도록 구성하였다. 매 글로벌 라운드에서는 전체 디바이스 중 5개의 디바이스를 무작위로 선택하여 학습 라운드에 참여시켰으며, 총 글로벌 라운드 수는 200회, 로컬 에폭은 1회로 설정하였다. 미니배치 크기는 10으로 설정하였고, 초기 학습률 0.03에서 시작하여 저빈도 통신이 개시되는 시점에 초기 학습률의 10%로 조정되도록 하였고, 저빈도 전환 시점은 전체 글로벌 라운드의 각각 25%와 50%를 완료한 시점을 기준으로 최종 수렴 성능을 측정하였다. 본 기법은 SFL의 직접 피드백 이점을 유지하면서도 전체 학습 과정에서 발생하는 통신 횟수를 감소시키고 일정 수준 이상의 성능을 보임을 그림 2.의 실험 결과를 통해 확인하였다.

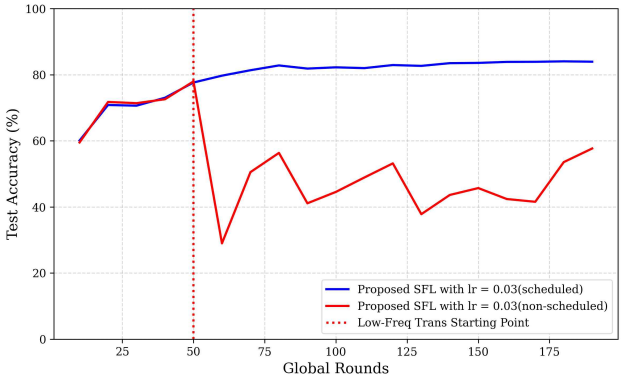

그림 1 학습률 조정 여부에 따른 학습 안정성 비교

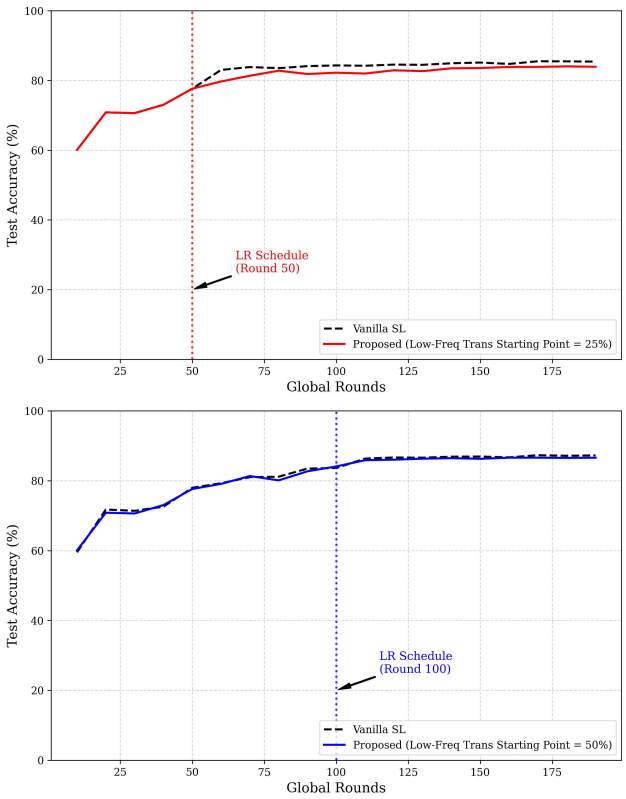

그림 2 저빈도 통신 시작 지점에 따른 성능 비교

Ⅲ. 결론

본 논문은 다중 디바이스 SFL 환경에서 발생하는 매우 큰 통신 오버헤드 문제를 해결하기 위해 저빈도 데이터 교환 기반 업데이트 방식을 제안하고 그 효용성을 검증하였다. 제안된 알고리즘은 학습 초기의 고빈도 통신을 통한 성능 기반 구축과 후기 단계의 저빈도 전송 주기를 결합함으로써 통신 자원 효율적인 분산 학습 모델을 제시한다. 특히 통신 빈도 조정을 고려한 학습률 조정 방식은 데이터 교환 횟수 감소로 인해 우려되는 모델 정확도 손실을 완화하고 최종 수렴의 안정성을 유지하는데에 기여하였다. 향후 연구에서는 고정된 라운드 기반 전환 방식에서 나아가, 실제 네트워크 상황에 따른 변수를 고려하여 통신 빈도 조정 시점과 학습률 조정 방식을 동적으로 결정하는 적응형 제어 알고리즘으로 발전시키려 한다. 최종적으로는 통신 오버헤드 절감과 모델 성능 유지 사이의 트레이드-오프를 최적화하는 연구를 지속할 계획에 있다.

ACKNOWLEDGMENT

본 과제(결과물)는 2025년도 교육부 및 경기도의 재원으로 경기RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2025-RISE-09-A15)

참 고 문 헌

[1] C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, “SplitFed:When federated learning meets split learning,” in Proc. AAAIConf. Artif. Intell. (AAAI), Vancouver, BC, Canada, Feb. 2022,pp. 8485-8493.

[2] McMahan, B. Moore, E. Ramage, D. Hampson, S. andy Arcas, B. A. 2017. Communication-Efficient Learning ofDeepNetworks fromDecentralized Data. In Proc. AISTATS,1273-1282.

[3] Gupta, O. and Raskar, R. 2018. Distributed learning of deepneural network over multiple agents. J. Network and Computer Applications, 116: 1-8.