Asynchronous federated learning on heterogeneous devices: A survey

Chenhao Xu a, Youyang Qu b c, Yong Xiang a, Longxiang Gao b c
Computer Science Review
Volume 50, November 2023, 100595

장치들의 다양성, 데이터의 다양성, 개인정보 및 보안을 포함한 분류 체계를 사용하여2019~2022간 125편의 연구를 기반으로 Asynchronous FL을 분석

곰탱이
wkdnffla3@gmail.com

2024-01-17

CONTENTS

1. Introduction

- ML의 발달에는 고품질의 데이터가 필수 불가결한 요소이다.
- 그중 데이터의 개인 정보 보호성이 강조되고 있으며 이로 인해 새로운 데이터를 수집할 때 어려움이 많다.
- 이러한 데이터 수집의 어려움은 isolated data islands(고립된 데이터 다양성으로 의역)문제를 야기한다.
- 이로 인해 연합 학습(FL)이라는 [각 로컬 데이터를 직접 엑세스 하는 대신 여러 디바이스간 ML 모델을 공유하는] 프레임 워크를 구글이 도입함.
- FL의 기본 목표는 데이터의 개인정보를 보호하면서 여러 참여자 또는 계산 노드를 이용해 ML 모델을 훈련하는데 있다.

1. Introduction

- FL 은 ML과 다르게 글로벌 모델의 그래디언트 기반 집계를 통해 개인 정보 보호기능을 향상시키고, 데이터를 서버로 보내는 것이 아니기 때문에 네트워크 전송 시간을 단축, 다른 장치들에서 학습한 모델들을 모아 전체 모델 성능을 향상.

- FL의 단점도 존재.

- 디바이스가 갑작스럽게 오프라인이 되어 업로드를 하지 못해 모델 업데이트에 지연이 생김(디바이스 신뢰도).

- 계산이 빠르게 끝난 디바이스는 계산이 느린 디바이스가 계산이 끝날 때 까지 대기해 효율성이 떨어짐(집계 효율성 감소).

- 디바이스 선택 알고리즘의 비효율성으로 충분한 리소스를 가진 디바이스가 학습에 참여하지 못하는 경우(낮은 자원 활용성).

- 잘못된 데이터를 집어넣어 모델의 성능을 낮추거나, 중앙 서버로 업로드한 모델의 그래디언트를 방해하는 가능성 존재(보안 위협 가능성).

1. Introduction

- 위의 문제점들을 보완하고자 비동기 연합 학습(asynchronous federated learning, AFL)이 주목을 받음.
- AFL은 로컬 모델을 수신하면 중앙 서버가 실시간으로 글로벌 모델을 업데이트
- 이로 인해 갑작스럽게 디바이스가 오프라인 되는 문제 완화 [디바이스 신뢰성 완화].
- 각각의 디바이스 들이 업로드 후 계산이 끝나지 않은 디바이스를 대기할 필요가 없어짐[운영 효율성 증가]

2. Background knowlege

- 연합 학습
- 블록체인
- 차등 개인 정보

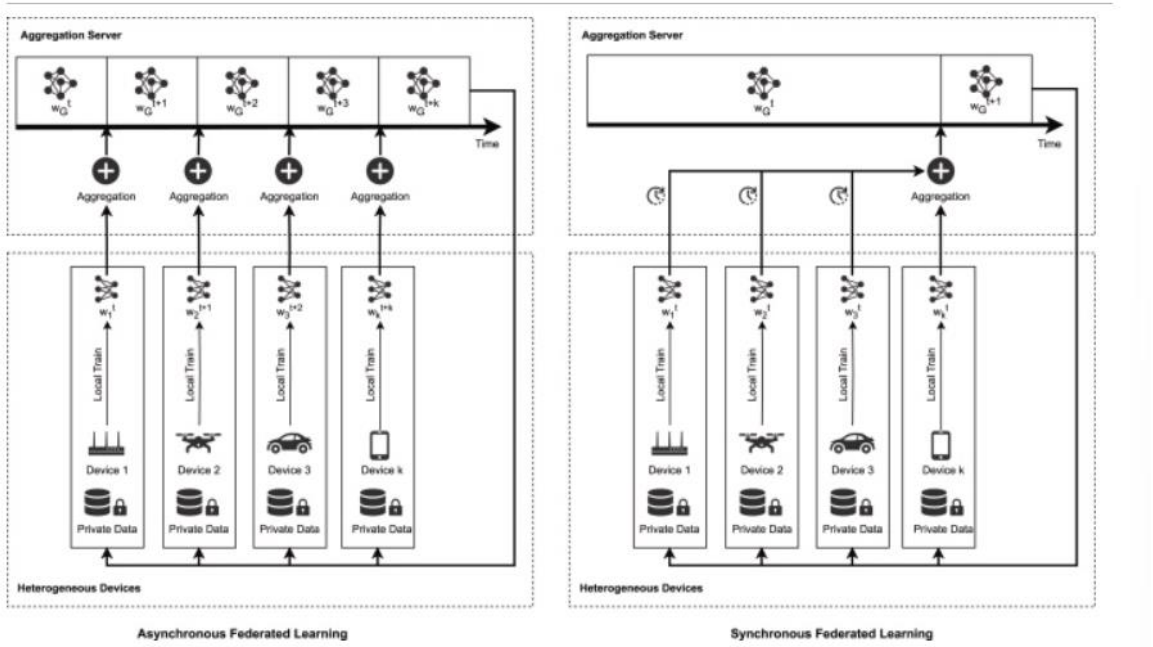
2. Background knowledge – 연합 학습

- Distribute ML(DML)은 모델을 더 잘 훈련하기 위해 다양한 컴퓨터들이 클러스터 들을 만들어 학습하는 방법

- DML은 중앙 집중식, 분산식, 완전 분산식으로 분류가 가능하며 FL은 분산 노드에서 로컬모델을 훈련하는 일종의 DML 이지만 차이점이 존재

- DML은 클라우드나 데이터 센터 환경을 가정하는 경우(데이터 공유)가 많지만 FL은 데이터가 여러 장치나 서버에 존재하고 중앙집중식이 아닌 시나리오를 위해 설계(데이터 비공유)

2. Background knowledge – 연합 학습



- FL의 프로세스
- 1. 초기화 상태 : 집계 서버에서 글로벌 모델의 초기 파라미터 값을 초기화 해서 각 노드에 전파
- 2. 로컬 모델 교육 : 로컬 디바이스에서는 받은 파라미터 값을 가지고 학습을 진행 후 집계 서버로 파라미터 값을 전송
- 3. 글로벌 모델 집계 : 전송 받은 파라미터 값을 토대로 글로벌 모델을 업데이트후 로컬 디바이스에 글로벌 모델을 배포

2. Background knowledge – 연합 학습

- 일반적인 학습은 데이터셋이 독립적이고 동일한 분포(IID)를 가진다 가정(ex : Mnist dataset 에서 모든 숫자가 있음)

- 실제 노드들 에선 IID 특성을 가지지 않는다 (non IID)(ex : 모든 노드들이 Mnist dataset에서 일부 숫자가 빠짐)

- 데이터 크기의 불일치와 노드간 연산 능력의 불일치는 학습 시간의 차이를 불러와 로컬 모델의 업데이트가 느려지는 결과가 발생해 글로벌 모델에 악영향을 끼침

2. Background knowledge – 블록체인

- 블록체인은 다양한 노드에서 거래 데이터의 통일성과 불변성을 유지하는 분산 원장 기술(DLT) 이다.
- 분산원장의 검증 방법은 스마트 컨트랙트로 검증해 보안과 신뢰성을 보장.
- 블록체인의 특성을 FL에 적용이 가능
- 분산 원장의 불변성으로 표절된 그래디언트 업로드 방지.
- 합의 알고리즘으로 작업 결과에 신뢰성을 높힌다.
- 스마트 컨트랙트로 값이 이상한 그래디언트 업로드 방지.
- 분산 집계 전략으로 집계서버가 Ddos 공격을 당하더라도 진행 가능.

2. Background knowledge – Differential privacy

- 특정 데이터 세트 내의 민감한 단일 데이터 포인트를 숨긴다.
- 핵심 목표는 데이터 분석에 필요한 특정 통계 속성을 유지하면서, 각 데이터 포인트를 비차별적으로 만드는 것.
- 여러 메커니즘들이 존재, 대표적으로 가우시안 메커니즘이 제어되는 무작위 노이즈를 데이터에 주입해서 개인정보를 보호할수 있다.
- 이러한 과정에서 보안성을 가져가는 대신 데이터의 특징이 손상될 수 도 있다.

3. Device heterogeneity

- AFL의 주요 목표 중 하나는 학습 효율성을 향상 시키기 위해 다른 기종의 장치간 자원 활용을 최적화 하는 것

- 이를 위해 여러 방법들 중 하나인 노드 선택, 가중치 집계, 기울기 압축, 반 비동기 FL, 모델 분할 방법에 대해 소개.

3. Device heterogeneity – 노드 선택

- 기존의 FL은 더 많은 훈련 데이터를 선택한 것과 달리 AFL은 계산 능력이 좋은 노드를 우선적으로 선택. -> 글로벌 모델의 성능과 오버 핏 사이에서 균형을 맞춤

- 그 방법 중 하나로 컴퓨팅 및 통신 자원을 기반으로 휴리스틱 그리디 노드 선택 전략이 있다. IID 데이터와 non-IID 데이터로 효과적임을 검증

- 많은 수의 노드가 존재할 때 동시에 훈련하는 디바이스의 수를 제한 -> 향상된 수렴 속도와 모델 정확도를 나타내지만 non-IID에 대한 부분은 좋지 않음.

- 컴퓨팅 능력과 정확도 변화에 따라 우선순위가 지정된 노드 선택 -> 더 빠른 수렴 속도와 높은 정확도 증가율을 보여준다.

- 위의 방법들은 디바이스의 신뢰성을 고려하지 않아 갑작스럽게 오프라인 되는 경우가 생김 -> 각 디바이스에 신뢰 점수를 할당해 노드 선택에서 필터링.

3. Device heterogeneity – 가중치 집계

- 기존의 FL은 더 많은 데이터로 훈련된 로컬 모델의 영향을 증폭하는 것으로 진행 했지만 AFL에서는 FL에 존재하지 않는 오래된 로컬 모델의 영향을 완화하는 방향으로 진행
- 로컬 모델에 부패도 개념의 매개변수를 도입해 시간이 지남에 따라 글로벌 모델에 끼치는 영향을 줄인다.
- 데이터셋의 크기와 로컬, 전역 모델의 기울기 간의 유사성을 고려해서 기울기 업데이트 전략을 세움.
- 로컬모델의 기울기를 자주 업로드하는 노드에 오버핏 되는 경우를 방지하기위해 모든 노드의 훈련 정확도에 따라 집계 가중치를 동적으로 조정.

3. Device heterogeneity – 기울기 압축

- 기존의 FL에서 기울기 압축은 효율성을 향상시키기 위해 사용했지만 AFL에서는 통신 비용 절감을 목표로 사용한다.
- AFL 에서 집계 및 압축 작업의 빈도가 증가하므로 기존의 FL에 비해 서버가 추가로 계산하는 경우가 발생. -> 이를 해결하기위해 AFL에 맞게 조정된 효율적인 압축 알고리즘이 제시
- Self-adaptive threshold computation – 최근 파라미터 변화를 기반으로 임계값을 계산
- Gradient communication comppression – 임계값을 기반으로 한 중복 기울기 통신을 압축
- 이외에도 통신 효율을 향상시키기 위해 모델 업로드와 다운로드를 효율적으로 스케쥴링 할수 있도록 하는 통신 프로토콜 설계가 존재.

3. Device heterogeneity – 반 비동기 FL

- AFL에서 계산이 느린 노드에서 오래된 로컬 모델을 업로드 하면 전역 모델의 정확도가 감소한다. 이러한 느린 장치의 영향을 완화하기 위해 반 비동기 FL 체계가 도입.
- 기존의 FL 방식과 AFL 방식을 결합하는 방법.
- 집계 서버는 더 일찍 도착한 로컬 모델을 캡처하고 저장한 다음 일정 시간 후 글로벌 모델 업데이트를 진행
- 부패도의 크기에 따라 로컬 모델은 다음 훈련에 참여하거나 폐기된다.

3. Device heterogeneity – 모델 분할

- 심층 신경망 모델을 분할한 후 각 노드들은 전체 모델이 아닌 분할한 계층을 학습해 통신 비용을 줄일수 있다.
- 이를 이용하면 노드가 다른 노드를 기다릴 필요성이 없어져(?) 모델의 수렴을 가속화 한다.
- 한 방법중 하나로 얕은 계층의 매개변수가 깊은 계층의 매개변수보다 더 자주 업데이트 되는 계층별 비동기 모델 업데이트 전략이 있다.
- 노드 선택 전략과 달리 모델 분할 전략은 노드에 대한 계산 요구를 줄이고 모델의 유연성을 제공하지만 다양한 모델로의 확장성은 제한적이다.

4. Data heterogeneity

- 데이터의 이질성
- 실제로 노드(디바이스) 들이 가지고 있는 데이터들은 대부분 비 IID 성을 가진다.
- 이러한 데이터 들로 학습할 경우 모델이 오버핏할 가능성이 커진다.
- 이러한 문제점들을 해결할 방안을 다음에서 소개한다.

4. Data heterogeneity

- 비 IID 데이터가 야기하는 문제점을 해결하는 연구분야로 [집계를 위한 제약 조건], [군집화된 FL], [분산 검증 전략], [수학적 최적화 매개변수] 등이 존재
- [집계를 위한 제약 조건]
- 로컬 업데이트를 전역 모델에 더 가깝게 하기위해 제시
- 유사한 업데이트 빈도를 가진 노드는 로컬 모델이 발산되는 것을 방지하기 위해 동기 및 비동기 훈련 전략을 통해 동일한 계층으로 그룹화.
- [군집화된 FL]
- 훈련 노드를 그룹화 하여 다양한 데이터 분포의 영향을 완화
- 그룹을 이룬 노드들의 데이터가 글로벌 데이터 분포에 맞게 클러스터를 이룸.

4. Data heterogeneity

- [분산 검증 전략]
- 로컬 모델을 평가하기 위해 다른 노드에 있는 데이터(5%)를 가져와 테스트.
- [수학적 최적화 매개변수]
- 전역 모델의 값이 튀는것을 완화하기위해 미리 결정된 초기 가중치 매개변수를 사용해 훈련.

5. Privacy and Security on heterogeneous devices

- FL은 로컬 훈련 데이터의 개인 정보 보호를 하기위해 도입 되었지만, 멤버 추론 공격, 속성 추론 공격, 모델 반전 공격, 및 그래디언트 공격 등의 새로운 공격 방법이 등장 했다.

- AFL 은 FL에 비해 이러한 공격에 취약해 차등 개인 정보 모델 또는 고효율 블록체인 기반 솔루션이 등장

6. Application on heterogeneous devices

- AFL은 다양한 분야에서 적용이 가능하다.
- 스마트 교통, 자율주행을 위한 스티어링 휠 각도 예측, UAV 운용, 등등에 적용되어 연구가 진행 중이다.

7. Research challenges and future directions

- 이번 절에서는 [디바이스 이질성], [데이터 이질성], [이기종 디바이스의 개인정보 보호 및 보안], [이종 기기에서의 응용]에 대한 향후 연구방향에 대해 설명한다.
- [디바이스 이질성]
- 앞에서 설명한 바와 같이 가중치 집계 및 군집화 FL 같은 방법을 사용하면 어느정도 해결이 되지만 너무 많은 방법을 사용하면 효율성이 저하된다.
- 이에 대해 다중 성능 향상 전략과 시간 소모 간의 균형에 대한 연구가 필요하다.
- [데이터 이질성]
- 군집화 훈련은 여러 AFL에서 활용이 되지만 모든 애플리케이션이나 시나리오에 대해 적용하기는 어렵다.
- 데이터 분포 유사성을 기반으로한 군집화 훈련은 적절한 유사성 평가 알고리즘의 개발이 필요하다.

7. Research challenges and future directions

- [이기종 디바이스의 개인정보 보호 및 보안]
- 차등 프라이버시를 이용한 방법은 데이터의 프라이버시를 지킬수록 모델의 성능이 저하되는 문제를 해결해야한다.
- 블록 체인을 활용한 보안 강화 최적화 방법은 훈련 모델을 주기적으로 기록하는 맞춤형 블록체인 구조로 합의 프로세스와 훈련 프로세스를 분리한다.
- [이종 기기에서의 응용]
- AFL 에 대한 시나리오가 IoV, 장애진단, IIoT 외에는 거의 없다.
- AFL은 제한된 컴퓨팅 자원을 가진 시간에 민감한 시나리오에 더 적합하므로 병원에서 환자의 상황을 예측하거나 스마트 그리드를 이용해 에너지 소비 예측을 하는데 적합하다.

8. Conclusion

- AFL을 사용할 때 기존 작업의 단점을 보완하기 위해 장치 이질성, 데이터 이질성, 이기종 장치의 보안 및 개인정보 보호 문제를 알아봤고 개선 방안도 연구한 논문들을 조사했다.