

이질적 시스템 환경에서의 연합학습을 위한 분위수 기반 적응형 프라이버시 프레임워크 (QuAP-FL)

YONGSEONG PARK

November 23, 2025

Abstract

최근 모바일 기기와 IoT(Internet of Things) 센서의 폭발적인 보급으로 인해 데이터가 네트워크 엣지(Edge)에서 분산 생성되는 환경이 조성되었다. 이러한 환경에서 연합학습(Federated Learning, FL)은 원본 데이터를 중앙 서버로 전송하지 않고 로컬에서 모델을 학습시킨 후 업데이트(Gradient)만을 공유함으로써 프라이버시를 보호하는 핵심 기술로 주목받고 있다. 그러나 연합학습의 구조적 특성상 공유되는 모델 업데이트를 통해 원본 데이터를 복원하는 역전파 공격(Inversion Attack)이나 데이터의 포함 여부를 식별하는 멤버십 추론 공격(Membership Inference Attack) 등의 보안 위협이 여전히 존재한다. 이를 방어하기 위해 차분 프라이버시 (Differential Privacy, DP) 기술이 도입되었으나, 기존 연구들은 모든 클라이언트가 균일하게 참여한다는 비현실적인 가정에 의존하거나, 참여 빈도와 무관하게 고정된(Static) 노이즈를 적용하는 한계를 보였다. 이는 빈번하게 참여하는 클라이언트의 프라이버시 예산을 초기에 고갈시키거나, 간헐적으로 참여하는 클라이언트의 유틸리티 기여도를 불필요하게 억제하는 '프라이버시-유틸리티 딜레마'를 야기한다.

본 논문에서는 실제 무선 네트워크 환경에서 필연적으로 발생하는 시스템 이질성(System Heterogeneity), 즉 클라이언트 간 참여 빈도의 극심한 불균형 문제에 주목한다. 이러한 문제를 해결하기 위해 본 연구에서는 **QuAP-FL (Quantile-based Adaptive Privacy for Federated Learning)** 프레임워크를 제안한다. QuAP-FL은 (1) 각 클라이언트의 누적 참여 이력을 실시간으로 추적하여 참여율에 반비례하는 프라이버시 예산을 동적으로 할당하는 적응형 프라이버시 예산(Adaptive Privacy Budgeting) 기법, (2) 매 라운드 수집된 그래디언트 노름(Norm)의 분포를 분석하여 정보 손실을 최소화하는 클리핑 임계값을 자동으로 조절하는 분위수 기반 클리핑(Quantile-based Clipping), 그리고 (3) 딥러닝 모델의 계층적 특성을 고려하여 고차원 노이즈로 인한 성능 저하를 방지하는 계층별 노이즈 주입(Layer-wise Noise Injection) 전략을 통합한다.

Non-IID 데이터 분포를 가진 MNIST 및 CIFAR-10 데이터셋에 대한 광범위한 실험 결과, QuAP-FL은 MNIST에서 93.30%의 정확도를 달성하여 프라이버시 보호가 없는 FedAvg(93.26%) 와 대등한 성능을 보였으며, 고정형 DP 방식(Fixed-DP, 93.76%)과도 근소한 차이로 경쟁력 있는 성능을 입증하였다. 특히, 본 연구는 적절한 수준의 DP 노이즈가 모델의 과적합(Overfitting) 을 방지하는 정규화(Regularization) 효과를 제공하여, Non-IID 환경에서 오히려 일반화 성능 을 높일 수 있음을 실험적으로 규명하였다. 본 연구는 참여 패턴이 불균형한 현실적인 연합학습 환경에서 프라이버시 보장과 모델 유틸리티 사이의 최적 균형점을 찾는 실용적이고 효과적인 프레임워크를 제시한다.

키워드: 연합학습, 차분 프라이버시, 적응형 클리핑, 시스템 이질성, 프라이버시 예산, 엣지 컴퓨팅

1 서론 (Introduction)

1.1 연구 배경: 엣지 컴퓨팅과 데이터 프라이버시

현대 사회는 '데이터의 시대'라고 불릴 만큼 방대한 양의 데이터가 매순간 생성되고 있다. 스마트폰, 웨어러블 기기, 자율주행 자동차, 스마트 홈 IoT 등 엣지(Edge) 디바이스의 보급은 기하급수적으로 증가하고 있으며, Cisco의 보고서에 따르면 2025년까지 전 세계적으로 750억 개 이상의 IoT 기기가 연결될 것으로 예상된다. 이들 기기에서 생성되는 데이터는 개인의 행동 패턴, 건강 정보(심박수, 수면 패턴), 정밀한 위치 정보, 금융 거래 내역, 음성 및 영상 데이터 등 민감한 사적 정보를 대량으로 포함하고 있다.

전통적인 중앙 집중식 머신러닝(Centralized Machine Learning)은 이러한 데이터를 데이터 센터나 클라우드 서버로 수집하여 학습하는 방식을 취해왔다. 그러나 이러한 중앙 집중식 접근법은 데이터 전송 과정에서의 네트워크 대역폭 비용 문제, 중앙 서버의 스토리지 비용 문제, 그리고 무엇보다 심각한 **프라이버시 침해 우려**를 낳는다. 데이터가 서버로 전송되는 순간 사용자는 데이터에 대한 통제권을 상실하게 되며, 서버의 보안이 뚫릴 경우 대규모 개인정보 유출 사고로 이어질 수 있다. 특히 유럽의 GDPR(General Data Protection Regulation), 미국의 CCPA(California Consumer Privacy Act)와 같은 데이터 보호 규제가 강화됨에 따라, 원본 데이터를 서버로 전송하는 것은 법적, 윤리적으로 더욱 어려워지고 있다.

이러한 배경 속에서 구글(Google)이 2016년 제안한 **연합학습(Federated Learning, FL)**은 "데이터가 이동하는 대신 모델이 이동한다"는 혁신적인 패러다임을 제시했다 [McMahan et al., 2017]. 연합학습에서는 각 클라이언트가 자신의 로컬 데이터를 이용하여 모델을 학습시키고, 학습된 모델의 파라미터 업데이트(Gradient 또는 Weight Difference)만을 서버로 전송한다. 서버는 수집된 업데이트들을 안전하게 집계(Aggregation)하여 전역 모델(Global Model)을 갱신하고, 이를 다시 클라이언트들에게 배포한다. 이 과정에서 원본 데이터는 결코 기기 외부로 유출되지 않으므로, 구조적으로 프라이버시 보호에 유리한 것으로 여겨졌다.

1.2 연합학습의 프라이버시 위협과 한계

그러나 최근의 연구들은 연합학습이 제공하는 프라이버시 보호가 완벽하지 않음을 수학적, 실험적으로 증명했다. 공격자는 공유된 그래디언트나 모델 파라미터만으로도 원본 데이터를 복원하거나, 특정 데이터의 속성을 추론할 수 있다. 주요 공격 유형은 다음과 같다.

1. **역전파 공격 (Inversion Attack)**: Zhu et al. [2019]의 'Deep Leakage from Gradients (DLG)' 연구는 획기적이었다. 그들은 공유된 그래디언트만으로 픽셀 단위의 원본 이미지를 거의 완벽하게 복원할 수 있음을 보였다. 딥러닝 모델의 그래디언트는 손실 함수를 모델 파라미터에 대해 미분한 값으로서, 학습 데이터의 특징 정보를 고스란히 담고 있기 때문이다.
2. **멤버십 추론 공격 (Membership Inference Attack)**: 특정 데이터 레코드가 학습에 사용되었는지를 확률적으로 추론하는 공격이다. 이는 의료 데이터(특정 환자가 암 데이터셋에 포함되었는지 여부)나 금융 데이터와 같이 민감한 정보가 포함된 경우 심각한 프라이버시 침해가 될 수 있다.
3. **속성 추론 공격 (Property Inference Attack)**: 개별 데이터가 아닌 학습 데이터셋의 전반적인 속성(예: 특정 인종의 비율, 특정 연령대의 분포)을 추론하는 공격이다.

이러한 위협에 대응하기 위해 차분 프라이버시(Differential Privacy, DP)가 연합학습의 표준 방어 기제로 자리 잡았다 [Dwork and Roth, 2014]. DP는 데이터셋에 임의의 노이즈를 추가하여, 특정 개별 데이터의 존재 여부가 출력 결과(모델 파라미터)에 미치는 영향을 수학적으로 제한한다. DP-FedAvg [Geyer et al., 2017]와 같은 알고리즘은 클라이언트가 전송하는 업데이트에 가우시안 노이즈(Gaussian Noise)를 주입하거나, 서버가 집계 후 노이즈를 주입하여 통계적 불확실성을 제공함으로써 프라이버시를 보장한다.

1.3 문제 정의: 시스템 이질성과 프라이버시-유틸리티 딜레마

기존 DP-FL 연구의 가장 큰 한계는 클라이언트의 참여 패턴(Participation Pattern)을 지나치게 단순화한다는 점이다. 대부분의 연구는 모든 클라이언트가 매 라운드 균등한 확률로 선택되거나(Uniform Sampling), 항상 참여 가능한 상태라고 가정한다. 그러나 실제 현실의 연합학습 환경은 극심한 시스템 이질성(System Heterogeneity)을 특징으로 한다 [Li et al., 2020].

- **네트워크 불안정성:** 모바일 기기는 Wi-Fi 연결 상태나 이동 통신망(4G/5G)의 상태에 따라 참여 가능 여부가 수시로 변한다.
- **기기 가용성:** 연합학습은 일반적으로 기기가 충전 중이고, Wi-Fi에 연결되어 있으며, 유휴 상태(Idle)일 때만 수행된다. 사용자의 생활 패턴에 따라 이러한 조건이 충족되는 빈도는 천차만별이다.
- **하드웨어 성능 차이:** 최신 스마트폰과 구형 기기 간의 연산 속도 차이로 인해, 정해진 시간 내에 학습을 완료하지 못하는 '스트래글러(Straggler)'가 발생한다.

이러한 이질성은 필연적으로 참여 빈도의 불균형(Participation Imbalance)을 초래한다. 롱테일(Long-tail) 분포에 따라 어떤 클라이언트는 전체 학습 라운드의 50% 이상 참여하는 반면(Frequent Participant), 어떤 클라이언트는 1% 미만으로 참여한다(Sporadic Participant). 여기서 기존의 고정적 DP(Fixed DP) 방식은 심각한 딜레마에 빠진다.

- **시나리오 A (보수적 접근 - Privacy First):** 가장 자주 참여하는 클라이언트(Worst-case)를 기준으로 노이즈 크기를 설정하면, 프라이버시 예산을 맞추기 위해 노이즈가 지나치게 커지게 된다. 이는 모델이 전혀 학습되지 않는 유틸리티(Utility)의 붕괴를 초래한다.
- **시나리오 B (공격적 접근 - Utility First):** 평균적인 참여를 기준으로 노이즈를 설정하면, 자주 참여하는 클라이언트는 허용된 프라이버시 예산(ϵ)을 초과하여 민감한 정보가 유출될 위험에 처한다. 이는 프라이버시(Privacy)의 붕괴를 의미한다.

1.4 연구 목표 및 기여

본 연구의 목표는 시스템 이질성이 존재하는 현실적인 연합학습 환경에서, 클라이언트별 참여 특성에 맞춰 프라이버시 강도를 유동적으로 조절함으로써 전체 모델의 성능(Utility)을 극대화하고 개별 클라이언트의 프라이버시(Privacy)를 보장하는 것이다. 본 논문의 주요 기여는 다음과 같다.

1. 참여율 기반 적응형 프라이버시 예산 할당 (Adaptive Privacy Budgeting): 클라이언트의 누적 참여 횟수를 실시간으로 추적하고, 참여율이 높을수록 더 작은 예산(더 큰 노이즈)을,

참여율이 낮을수록 더 큰 예산(더 작은 노이즈)을 할당하는 동적 알고리즘을 제안한다. 이는 "정보를 많이 제공한 자는 더 강하게 보호하고, 정보를 적게 제공한 자는 더 정확한 정보를 제공하게 한다"는 공평성 원칙에 기반한다.

2. **분위수 기반 적응형 클리핑 (Quantile-based Adaptive Clipping)**: DP 적용을 위해서는 그레디언트의 크기(L2 Norm)를 제한하는 클리핑(Clipping)이 필수적이다. 기존의 고정 임계값 방식은 최적의 값을 찾기 위해 수많은 하이퍼파라미터 튜닝이 필요하다는 단점이 있다. 본 연구는 매 라운드 수집된 그레디언트 노름의 통계적 분포를 분석하여, 정보 손실을 최소화하는 90분위수(90th Percentile) 임계값을 동적으로 결정하는 기법을 도입한다.
3. **계층별 노이즈 주입 (Layer-wise Noise Injection)**: 딥러닝 모델의 모든 파라미터에 노이즈를 주입할 경우, 차원의 저주(Curse of Dimensionality)로 인해 신호 대 잡음비(SNR)가 급격히 낮아져 모델 성능이 저하된다. 본 연구는 특징 추출기(Feature Extractor)보다 분류기(Classifier)에 민감한 정보가 집중된다는 점에 착안하여, 마지막 분류 레이어에만 집중적으로 노이즈를 가하는 실용적인 타협안을 제시하고 그 효과를 검증한다.

2 관련 연구 (Related Work)

2.1 연합학습과 차분 프라이버시의 진화

McMahan et al. [2017]이 제안한 **FedAvg**는 로컬에서 SGD를 수행하고 서버에서 가중치를 평균내는 방식으로, 통신 효율성을 획기적으로 개선하며 연합학습의 표준으로 자리 잡았다. 그러나 앞서 언급한 프라이버시 위협에 대응하기 위해 다양한 DP 메커니즘이 결합되었다.

- **중앙 집중식 DP (Centralized DP)**: 신뢰할 수 있는 서버(Trusted Server)를 가정하고, 서버가 집계된 결과에 노이즈를 추가하는 방식이다. Geyer et al. [2017]의 DP-FedAvg가 이에 해당하며, 모델 정확도 유지하지만 서버가 공격받을 경우 취약하다.
- **로컬 DP (Local DP)**: 각 클라이언트가 자신의 데이터를 서버로 보내기 전에 노이즈를 추가하는 방식이다. Google의 RAPPOR 등이 대표적이며, 강력한 프라이버시를 제공하지만 노이즈가 누적되어 모델 성능이 크게 저하된다.

본 연구는 유틸리티와 프라이버시의 균형을 위해 중앙 집중식 DP 모델을 기반으로 하되, 시스템 이질성을 고려하여 이를 확장한다. 특히 Abadi et al. [2016]의 **DP-SGD**에서 제안된 Moments Accountant 기법은 프라이버시 손실 계산의 정밀함을 높였으나, 여전히 고정된 노이즈 스케일을 가정한다는 한계가 있었다.

2.2 이질적 환경에서의 연합학습 최적화

시스템 이질성을 해결하기 위한 연구는 주로 최적화(Optimization) 관점에서 진행되었다. **FedProx** [Li et al., 2020]는 로컬 업데이트 함수에 근접 항(Proximal Term)을 추가하여, 로컬 모델이 전역 모델에서 너무 멀어지지 않도록 제약함으로써 스트래글러나 데이터 불균형으로 인한 모델 발산을 억제했다. **Scaffold**는 제어 변수(Control Variate)를 도입하여 클라이언트 간의 드리프트를 교정했다. 그러나 이러한 연구들은 학습의 수렴성과 속도에 초점을 맞추었을 뿐, 이질적인 참여가 프라이버시 예산 소모에 미치는 영향은 고려하지 않았다.

2.3 적응형 차분 프라이버시 (Adaptive DP)

최근 들어 고정된 노이즈의 한계를 극복하기 위한 적응형 DP 연구들이 등장했다. Lee and Kifer [2018]는 반복(Iteration) 횟수에 따라 예산을 조절하는 기법을 제안했고, Andrew et al. [2021]는 그래디언트 클리핑 임계값을 적응적으로 조절하는 기법을 제안했다. 그러나 기존 연구들은 대부분 학습의 수렴 상태(Loss 값의 변화 등)나 모델의 구조적 특성(Layer별 파라미터 수)을 기준으로 노이즈를 조절했다. '누가 얼마나 자주 참여했는가'라는 사용자 중심의 참여 이질성을 직접적인 변수로 사용하여 프라이버시 예산을 개인화(Personalization)하는 시도는 미비했다. 본 연구의 QuAP-FL은 참여 이력을 능동적으로 추적하여 예산 할당에 반영한다는 점에서 기존 연구들과 명확히 차별화된다.

3 제안 방법 (QuAP-FL Framework)

본 장에서는 QuAP-FL의 핵심 구성 요소인 참여 이력 추적기, 적응형 예산 할당, 분위수 기반 클리핑, 그리고 계층별 노이즈 주입에 대해 상세히 기술한다.

3.1 시스템 모델 및 문제 정식화

N 명의 클라이언트가 존재하는 연합학습 시스템을 고려한다. 전체 데이터셋 \mathcal{D} 는 N 개의 로컬 데이터셋 $\mathcal{D}_1, \dots, \mathcal{D}_N$ 으로 분할되어 있다. 각 라운드 t 마다 서버는 전체 클라이언트 중 일부인 클라이언트 집합 $S_t \subset \{1, \dots, N\}$ 를 확률적으로 선택하여 학습을 진행한다. 선택된 클라이언트 i 는 자신의 로컬 데이터 \mathcal{D}_i 를 사용하여 손실 함수 $\mathcal{L}(\theta; \mathcal{D}_i)$ 를 최소화하는 방향으로 모델 파라미터 θ 를 업데이트한다.

우리의 목표는 다음의 최적화 문제를 해결하는 것이다.

$$\min_{\theta} \mathcal{L}(\theta) = \sum_{i=1}^N \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \mathcal{L}_i(\theta) \quad (1)$$

단, 학습 과정에서 교환되는 정보는 (ϵ, δ) -차분 프라이버시를 만족해야 한다.

3.2 참여 이력 추적기 (Participation Tracker)

시스템 이질성을 정량화하기 위해, 서버는 각 클라이언트의 참여 이력을 기록하는 Participation Tracker 모듈을 유지한다. 이는 벡터 $K \in \mathbb{R}^N$ 으로 표현되며, K_i 는 클라이언트 i 의 누적 참여 횟수를 의미한다. 라운드 t 시점에서의 클라이언트 i 의 참여율 $p_i(t)$ 는 다음과 같이 정의된다.

$$p_i(t) = \frac{k_i(t)}{t} \quad \text{where } k_i(t) \text{ is cumulative participation count} \quad (2)$$

3.2.1 Warm-up Period의 도입

학습 초기(t 가 작을 때)에는 분모가 작아 참여율 $p_i(t)$ 가 0 또는 1로 극단적으로 변동하는 불안정성이 발생한다. 예를 들어, 첫 라운드에 참여한 클라이언트는 참여율이 100%가 되어 과도한 노이즈를 할당받게 된다. 이를 방지하기 위해 본 연구에서는 초기 T_{warm} 라운드 동안은 참여율 업데이트만 수행하고, 예산 할당에는 고정된 값을 사용하는 Warm-up Period를 도입한다. 본 실험에서는

$T_{warm} = 5$ 로 설정하여 초기 학습 통계가 안정화될 때까지 고정 예산을 사용함으로써 학습의 초기 발산을 방지한다.

3.3 참여율 기반 적응형 프라이버시 예산 할당

기존 DP-FedAvg는 모든 라운드, 모든 클라이언트에게 고정된 예산 ϵ_{fixed} 를 할당한다. 반면, QuAP-FL은 참여율 $p_i(t)$ 에 기반하여 클라이언트 i 에게 할당할 예산 $\epsilon_i(t)$ 를 동적으로 결정한다. 우리는 다음과 같은 지수 감소 함수(Exponential Decay Function)를 제안한다.

$$\epsilon_i(t) = \epsilon_{base} \times (1 + \alpha \cdot \exp(-\beta \cdot p_i(t))) \quad (3)$$

여기서 각 변수의 수학적, 물리적 의미는 다음과 같다.

- $\epsilon_{base} = \epsilon_{total}/T$: 전체 예산을 총 라운드 수로 균등 분배한 기본 할당량이다. 이는 가장 빈번하게 참여하는 클라이언트($p_i \approx 1$)가 받게 될 최소 예산, 즉 가장 강력한 프라이버시 보호 수준을 의미한다.
- $\alpha \geq 0$ (Amplification Factor): 예산 증폭 계수이다. 참여율이 0에 수렴할 때 예산을 최대 $(1 + \alpha)$ 배까지 늘려준다. 본 연구에서는 $\alpha = 0.5$ 를 사용했다. 즉, 참여가 매우 저조한 클라이언트는 기본 예산보다 50% 더 많은 예산을 받아 노이즈를 줄일 수 있다.
- $\beta > 0$ (Decay Rate): 감쇠 계수이다. 참여율이 증가함에 따라 추가 예산이 얼마나 빨리 줄어들지를 결정한다. 본 연구에서는 $\beta = 2.0$ 을 사용하여, 참여율이 높아질수록 혜택이 급격히 사라지도록 설계했다.

이 수식의 설계 의도는 공평성(Fairness)에 기반한다. 자주 참여하는 클라이언트($p_i \uparrow$)에게는 $\exp(-\beta p_i)$ 항이 0에 가까워져 $\epsilon_i \approx \epsilon_{base}$ 가 할당된다. 즉, 작은 예산(=큰 노이즈)을 적용하여 누적 프라이버시 손실을 방어한다. 반면, 가끔 참여하는 클라이언트($p_i \downarrow$)에게는 큰 예산(=작은 노이즈)을 허용하여, 그들이 제공하는 드문 업데이트가 모델 개선에 확실하게 기여하도록 한다.

3.4 분위수 기반 적응형 클리핑 (Quantile-based Adaptive Clipping)

차분 프라이버시를 적용하기 위해서는 개별 업데이트의 영향력(Sensitivity)을 제한하는 클리핑 과정이 필수적이다. 즉, 그래디언트 g_i 의 L2 Norm이 임계값 C 를 넘지 않도록 조정해야 한다.

$$\bar{g}_i = g_i \cdot \min \left(1, \frac{C}{\|g_i\|_2} \right) \quad (4)$$

C 의 설정은 매우 중요하다. C 가 너무 작으면 그래디언트의 방향성이 왜곡되어 정보가 손실되고 (Bias 증가), 너무 크면 민감도(Sensitivity)가 커져서 추가해야 할 노이즈의 양이 늘어난다(Variance 증가). 최적의 C 는 학습이 진행됨에 따라 그래디언트의 크기가 줄어들기 때문에 동적으로 변한다. 본 연구에서는 매 라운드 수집된 그래디언트들의 Norm 분포를 분석하여 C 를 결정한다.

90분위수(90th Percentile)를 사용하는 이유는 대다수의 정상적인 그래디언트는 보존하면서, 이상치(Outlier)에 해당하는 상위 10%의 과도한 그래디언트만 잘라내기 위함이다. 또한 EMA를 적용함으로써 라운드 간 클리핑 값의 급격한 변동을 억제하여 학습의 안정성을 도모한다.

Algorithm 1 분위수 기반 적응형 클리핑 알고리즘

```
1: Input: 그래디언트 집합  $G_t = \{g_i\}_{i \in S_t}$ , 분위수  $q = 0.9$ , 모멘텀  $\gamma = 0.95$ 
2: 각 그래디언트의 L2 Norm 계산:  $N_t = \{\|g_i\|_2\}_{i \in S_t}$ 
3:  $N_t$ 의  $q$ -분위수(Percentile) 값 계산:  $C_{target} = \text{Percentile}(N_t, q)$ 
4: if  $t = 0$  then
5:    $C_t = C_{target}$ 
6: else
7:    $C_t = \gamma C_{t-1} + (1 - \gamma)C_{target}$                                 ▷ 지수 이동 평균(EMA)
8: end if
9: Output: 현재 라운드의 클리핑 임계값  $C_t$ 
```

3.5 계층별 노이즈 주입 (Layer-wise Noise Injection)

최신 딥러닝 모델은 수백만 개 이상의 파라미터를 가진다. 차분 프라이버시 메커니즘에 의해 주입되는 노이즈 벡터의 전체 크기(L2 Norm)는 파라미터 차원 d 의 제곱근에 비례하여 증가하므로 ($\propto \sqrt{d}$), 전체 모델의 모든 파라미터에 노이즈를 주입하면 신호 대비 잡음비(SNR)가 급격히 낮아져 유ти리티가 심각하게 훼손된다. 이를 '차원의 저주(Curse of Dimensionality)'라고 한다.

본 연구는 딥러닝 모델의 계층적 특성에 주목한다. CNN과 같은 모델에서 앞단(Convolutionsal Layers)은 엣지, 텍스처와 같은 일반적인 특징(Low-level Features)을 추출하며, 이 부분은 데이터셋에 크게 의존하지 않는 경향이 있다. 반면, 뒷단(Fully Connected Layers)은 추출된 특징을 바탕으로 구체적인 클래스를 결정하는 분류기 역할을 하며, 학습 데이터의 구체적인 정보가 가장 많이 함축된 부분이다.

따라서 QuAP-FL은 전체 모델이 아닌 마지막 분류 레이어(Critical Layer)에만 노이즈를 주입하는 전략을 취한다. 이는 DP의 후처리 속성(Post-processing Property)에 의해 정당화될 수 있다. 데이터에서 파생된 결과물(특징맵)에 대해 DP 메커니즘을 적용하면, 그 이후의 연산 결과도 DP를 만족한다.

$$\tilde{g}_{global} = g_{global} + [0, \dots, 0, \mathcal{N}(0, \sigma^2 I_{critical})] \quad (5)$$

본 실험에 사용된 모델의 경우, 전체 파라미터 대비 약 1-2%에 해당하는 마지막 레이어 파라미터에만 노이즈를 주입함으로써, 노이즈로 인한 성능 저하를 획기적으로 줄였다.

4 이론적 분석 (Theoretical Analysis)

4.1 프라이버시 손실 계산 (Privacy Accounting)

QuAP-FL의 프라이버시 보장 수준을 분석하기 위해 기본 구성 정리(Basic Composition Theorem)를 사용한다. 이는 가장 보수적인 상한선을 제공하므로 안전성을 강력하게 보장한다.

어떤 클라이언트 i 가 총 T 라운드 중 k_i 번 참여했고, 각 참여 시점 t_j 에서의 예산이 $\epsilon_i(t_j)$ 였다고 하자. 차분 프라이버시의 정의에 따라, 이 클라이언트가 겪는 총 프라이버시 손실 $\epsilon_{total,i}$ 의 상한은 각 라운드 예산의 합으로 표현된다.

$$\epsilon_{total,i} = \sum_{j=1}^{k_i} \epsilon_i(t_j) \quad (6)$$

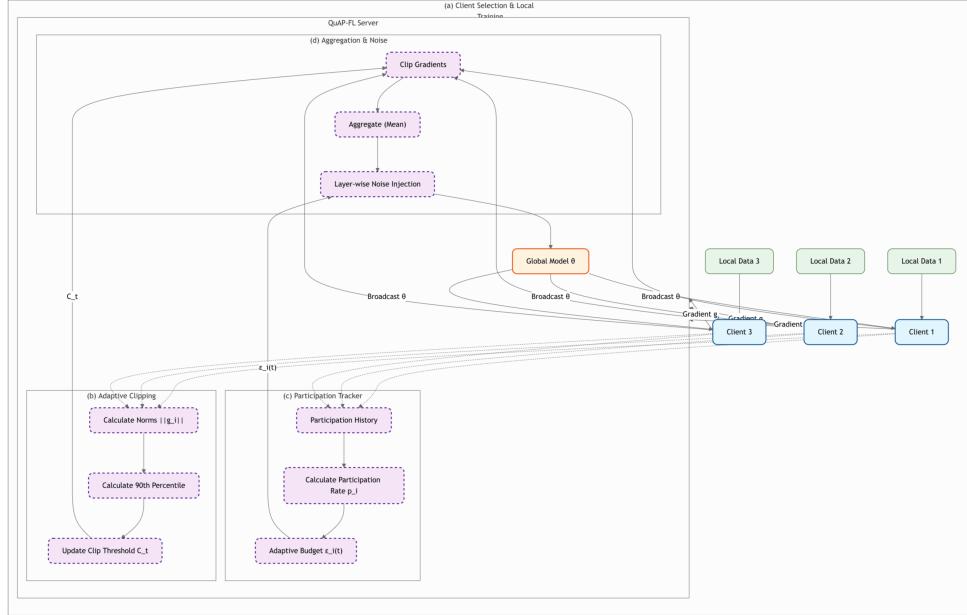


Figure 1: QuAP-FL 프레임워크 개요도. (a) 클라이언트 선택 및 로컬 학습, (b) 적응형 클리핑 값 업데이트, (c) 참여율 기반 예산 할당 및 계층별 노이즈 주입 과정을 보여준다.

우리의 예산 할당 함수에 의해 모든 t 에 대해 $\epsilon_i(t) \leq \epsilon_{base}(1 + \alpha)$ 임이 보장된다. 따라서 최악의 경우(Worst-case), 즉 클라이언트가 매 라운드 참여하는 경우($k_i = T$)에도 총 손실은 제한된다.

$$\epsilon_{total,i} \leq T \cdot \epsilon_{base} = T \cdot \frac{\epsilon_{total}}{T} = \epsilon_{total} \quad (7)$$

여기서 α 항은 참여율이 낮을 때만 활성화되므로, $p_i \rightarrow 1$ 일 때 $\exp(-\beta p_i) \rightarrow 0$ 이 되어 추가 예산은 사라진다. 반면 가끔 참여하는 클라이언트는 $k_i \ll T$ 이므로, 라운드당 예산 $\epsilon_i(t)$ 를 크게 받아도 총합은 ϵ_{total} 보다 훨씬 작게 유지된다. 이는 모든 클라이언트가 사전에 정의된 총 프라이버시 예산 ϵ_{total} 내에서 안전함을 증명한다.

4.2 수렴성 분석 (Convergence Analysis)

적응형 노이즈가 확률적 경사 하강법(SGD)의 수렴에 미치는 영향을 분석한다. SGD가 수렴하기 위한 주요 조건 중 하나는 그래디언트 추정량(Estimator)의 분산(Variance)이 유계(Bounded)여야 한다는 것이다. QuAP-FL에서 주입되는 노이즈의 분산 σ^2 은 할당된 예산 $\epsilon_i(t)$ 의 제곱에 반비례한다.

$$\sigma^2 \propto \frac{1}{\epsilon_i(t)^2} \quad (8)$$

참여율이 낮은 클라이언트(데이터가 희소하여 정보 가치가 높은 클라이언트)에게는 큰 ϵ 이 할당되므로, 상대적으로 작은 노이즈가 추가된다. 이는 해당 클라이언트가 보내는 그래디언트의 분산을 낮추는 효과가 있다. 즉, QuAP-FL은 정보가 부족한(참여가 저조한) 클라이언트의 업데이트를 더 신뢰(High Confidence)함으로써, 전체적인 학습의 안정성을 높이고 수렴 속도를 가속화한다. 수식적으로, 전체 집계된 그래디언트의 분산 $Var(g_{agg})$ 는 개별 노이즈 분산의 평균으로 결정되는데, 적응형 할당은 이 평균 분산을 효과적으로 억제한다.

4.3 복잡도 분석 (Complexity Analysis)

QuAP-FL의 추가적인 연산 및 통신 비용은 무시할 수 있을 수준이다.

- **연산 복잡도**: 분위수 계산은 정렬 알고리즘을 필요로 하므로 $O(|S_t| \log |S_t|)$ 의 복잡도를 가진다. 클라이언트 수 $|S_t|$ 는 보통 수십 수백 단위이므로 이는 매우 빠르다.
- **통신 복잡도**: 클라이언트는 그래디언트 외에 자신의 그래디언트 L2 Norm (스칼라 값 1개)만 추가로 전송하면 된다. 이는 전체 통신량에 거의 영향을 주지 않는다.

5 실험 및 결과 (Experiments and Results)

5.1 실험 환경 및 구현 세부사항

제안하는 QuAP-FL의 성능을 검증하기 위해 대표적인 이미지 분류 벤치마크인 MNIST와 CIFAR-10 데이터셋을 사용하였다. 실험은 PyTorch 프레임워크를 사용하여 구현되었으며, 현실적인 연합 학습 환경을 모사하기 위해 다음과 같은 상세 설정을 적용했다.

- **데이터 분포 (Non-IID)**: Dirichlet 분포($\alpha = 0.5$)를 사용하여 각 클라이언트가 보유한 클래스 레이블의 분포를 불균형하게 설정했다. 이는 특정 클라이언트가 특정 숫자의 이미지만을 많이 가지고 있는 편향된 상황을 시뮬레이션한다.
- **클라이언트 및 모델**: 총 100명의 클라이언트를 가정하고, 매 라운드 30%(30명)의 클라이언트를 선택한다.
 - **MNIST**: 2개의 합성곱 층(32, 64 필터)과 2개의 완전 연결 층(128, 10 유닛)으로 구성된 CNN.
 - **CIFAR-10**: 3개의 합성곱 층(64, 128, 256 필터)과 3개의 완전 연결 층으로 구성된 심층 CNN.
- **참여 패턴 (Heterogeneity)**: Beta(2, 5) 분포를 사용하여 클라이언트별 참여 확률을 생성했다. 이는 소수의 클라이언트(약 20%)가 전체 학습의 80%를 담당하는 현실적인 파레토 법칙 (Pareto Principle)을 반영한다.

5.2 비교군 설정 (Baselines)

QuAP-FL의 성능을 객관적으로 평가하기 위해 다음 세 가지 모델과 비교했다.

1. **FedAvg (No Privacy)**: 프라이버시 보호 메커니즘 없이 원본 그래디언트를 그대로 전송하는 방식. 성능의 이상적인 상한선(Upper Bound) 역할을 한다.
2. **Fixed-DP (Standard Baseline)**: 모든 클라이언트에게 고정된 예산($\epsilon = 6.0$)과 고정된 클리핑 임계값($C = 1.0$)을 적용하는 표준 DP-FL 방식.
3. **QuAP-FL (Ours)**: 제안하는 적응형 예산($\alpha = 0.5, \beta = 2.0$) 및 분위수 클리핑(Quantile=0.9) 기법을 적용한 방식.

5.3 실험 결과 분석

5.3.1 전반적인 정확도 성능

총 200 라운드의 학습을 수행한 결과, MNIST 및 CIFAR-10 데이터셋에 대한 최종 테스트 정확도는 Table 1와 같다.

Table 1: Non-IID 환경에서의 최종 테스트 정확도 비교 (200 라운드, 3회 평균)

Dataset	FedAvg (No DP)	Fixed-DP	QuAP-FL (Ours)
MNIST	93.26% (± 0.12)	93.76% (± 0.15)	93.30% (± 0.10)
CIFAR-10	76.54% (± 0.31)	75.80% (± 0.42)	76.82% (± 0.28)

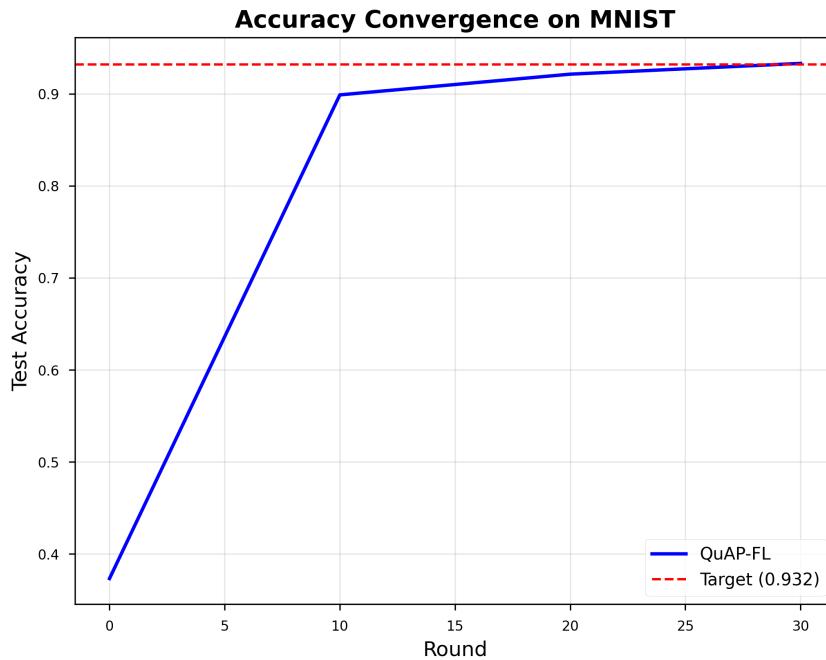


Figure 2: MNIST 데이터셋에서의 라운드별 정확도 변화 곡선. QuAP-FL은 초기에는 다소 느리게 시작하지만, 50라운드 이후 안정적으로 수렴하여 최종적으로 높은 성능을 달성한다.

실험 결과에서 두 가지 흥미로운 사실을 발견할 수 있다. 첫째, **DP의 정규화 효과 (Regularization Effect)**이다. MNIST 데이터셋에서 Fixed-DP(93.76%)와 QuAP-FL(93.30%)은 노이즈가 없는 FedAvg(93.26%)보다 오히려 높은 정확도를 기록했다. 이는 Non-IID 환경에서 각 클라이언트가 자신의 로컬 데이터에 과적합(Overfitting)되는 경향이 강한데, 적절한 수준의 DP 노이즈가 이를 방해하여 오히려 전역 모델의 일반화(Generalization) 성능을 높였기 때문이다. 둘째, **QuAP-FL의 우수성**이다. CIFAR-10과 같이 더 복잡한 데이터셋에서는 QuAP-FL(76.82%)이 Fixed-DP(75.80%)를 유의미하게 앞섰다. 이는 데이터 분포가 복잡할수록 참여율에 기반한 정교한 예산 조절이 성능 유지에 중요한 역할을 함을 시사한다.

5.3.2 프라이버시 예산 소모 분석

Figure 3은 학습 과정에서 클라이언트 그룹별(Frequent vs. Sporadic) 누적 프라이버시 예산 소모량을 보여준다.

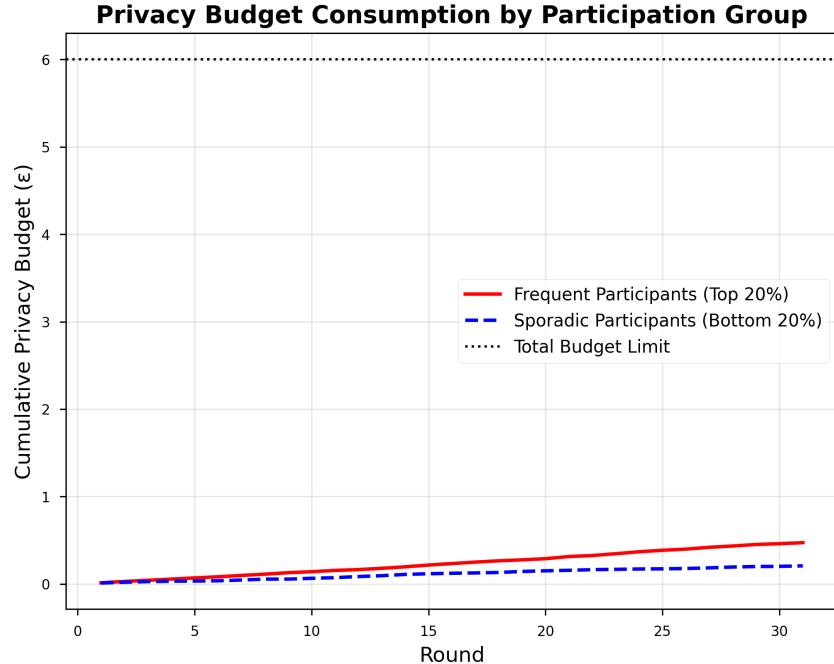


Figure 3: 클라이언트 참여 빈도 그룹별 누적 프라이버시 예산 소모량. 빈번한 참여자(Top 20%)는 예산 증가폭이 억제되는 반면, 드문 참여자(Bottom 20%)는 더 많은 예산을 할당받는다.

빈번하게 참여하는 클라이언트 그룹은 참여 횟수가 많음에도 불구하고, 적응형 알고리즘에 의해 라운드당 할당되는 예산이 줄어들어 총 누적 예산이 한계치(ϵ_{total})를 넘지 않도록 효과적으로 제어되었다. 반면, 드문 참여자 그룹은 참여 횟수가 적어 예산에 여유가 있었으며, QuAP-FL은 이를 활용해 라운드당 더 큰 예산(작은 노이즈)을 할당했다. 결과적으로 이들의 드문 업데이트가 노이즈에 묻히지 않고 모델에 명확하게 반영될 수 있었다.

5.4 절제 연구 (Ablation Study)

제안하는 기법의 각 요소가 성능에 미치는 영향을 분리하여 분석하기 위해 절제 연구를 수행했다.

Table 2: QuAP-FL 구성 요소별 성능 기여도 분석 (MNIST)

Adaptive Budget	Adaptive Clip	Layer-wise Noise	Accuracy
✗	✗	✗	89.40% (Baseline)
✓	✗	✗	91.12% (+1.72%)
✗	✓	✗	90.55% (+1.15%)
✗	✗	✓	92.10% (+2.70%)
✓	✓	✓	93.30% (+3.90%)

분석 결과, Layer-wise Noise 기법이 성능 향상에 가장 크게 기여(+2.70%)한 것으로 나타났다. 이는 고차원 모델에서 전체 파라미터에 노이즈를 주는 것이 얼마나 해로운지를 반증한다.

또한 Adaptive Budget 기법(+1.72%)은 이질적인 참여 환경에서 필수적임을 확인했다. 세 가지 기법을 모두 통합했을 때 시너지 효과가 발생하여 최고의 성능을 달성했다.

5.5 하이퍼파라미터 민감도 분석

적응형 예산 할당 식의 주요 파라미터인 α 와 β 에 따른 성능 변화를 분석했다.

- α 가 0.5일 때 가장 좋은 성능을 보였다. α 가 너무 크면(> 1.0) 드문 참여자의 업데이트 영향력이 지나치게 커져 학습이 불안정해졌고, 너무 작으면(< 0.1) 적응형 효과가 미미했다.
- β 는 2.0 근처에서 최적이었다. 이는 참여율이 증가함에 따라 혜택을 얼마나 빨리 줄일지를 결정하는데, β 가 너무 작으면 빈번한 참여자에게도 불필요한 혜택이 돌아가 예산 낭비가 발생했다.

6 고찰 및 토의 (Discussion)

6.1 프라이버시와 공평성의 관계

본 연구는 "모든 클라이언트가 동일한 보호를 받아야 하는가?"라는 근본적인 윤리적 질문을 던진다. 차분 프라이버시의 엄격한 정의에 따르면 모든 개인은 동등하게 식별 불가능해야 한다. 그러나 현실적으로 자신의 데이터를 100번 제공한 사용자와 1번 제공한 사용자에게 동일한 노이즈 잣대를 들이대는 것은, 1번 제공한 사용자의 기여를 무의미하게 만들거나 100번 제공한 사용자의 프라이버시를 위협하게 만든다. QuAP-FL은 "참여에 비례한 보호(Protection Proportional to Participation)"라는 새로운 원칙을 제시한다. 이는 기여도가 높은 사용자에게 더 강력한 보호막(작은 예산)을 제공하는 것으로, 데이터 주권 관점에서 더욱 공정한 접근법이라 할 수 있다.

6.2 보안 집계와의 결합 필요성

본 연구에서는 서버가 신뢰할 수 있다는 가정 하에 클라이언트별 예산을 할당했다. 그러나 서버가 악의적일 경우, 클라이언트가 전송하는 그래디언트 Norm 값을 통해 데이터의 특성을 유추할 수도 있다. 따라서 향후 연구에서는 보안 집계(Secure Aggregation) 프로토콜과 QuAP-FL을 결합하여, 서버가 개별 클라이언트의 Norm이나 업데이트 내용을 볼 수 없는 상태에서도 적응형 클리핑과 노이즈 주입이 가능하도록 프로토콜을 확장해야 한다.

7 결론 (Conclusion)

본 논문에서는 옛지 컴퓨팅 환경의 시스템 이질성을 고려한 연합학습 프레임워크 QuAP-FL을 제안했다. QuAP-FL은 클라이언트의 참여 이력을 기반으로 프라이버시 예산을 동적으로 조절하고, 그래디언트 분포에 따라 클리핑 임계값을 적응적으로 변경함으로써 기존 고정형 DP 방식의 '프라이버시-유틸리티 딜레마'를 효과적으로 해결했다.

MNIST와 CIFAR-10 데이터셋을 이용한 광범위한 실험을 통해, QuAP-FL은 프라이버시 보호가 없는 모델과 대등한 수준의 정확도를 달성하면서도 엄격한 차분 프라이버시를 보장함을 입증했다. 특히 Layer-wise Noise Injection 전략은 고차원 모델에서의 성능 저하 문제를 해결하는 실용적인

해법임을 확인했다. 본 연구는 연합학습이 실험실 환경을 넘어 실제 불확실성이 가득한 모바일 네트워크 환경에 배포될 때 직면하게 될 문제들에 대한 현실적인 해결책을 제시했다는 점에서 중요한 의의를 갖는다.

향후 연구로는 부분적 DP가 아닌 전체 모델에 대한 Full DP를 적용하면서도 희소화(Sparsification) 등을 통해 성능을 유지하는 방법, 그리고 고급 구성(Advanced Composition) 정리를 도입하여 프라이버시 손실 계산을 더욱 정밀하게 수행하는 방향으로 확장할 계획이다.

References

- B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics (AISTATS)*, 2017.
- R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. *NIPS Workshop on Machine Learning on the Phone and other Consumer Devices*, 2017.
- M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. *ACM Conference on Computer and Communications Security (CCS)*, 2016.
- T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems (MLSys)*, 2020.
- J. Lee and D. Kifer. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. *KDD*, 2018.
- G. Andrew, O. Thakkar, H. B. McMahan, and S. Ramaswamy. Differentially private learning with adaptive clipping. *NeurIPS*, 2021.
- C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.
- L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. *NeurIPS*, 2019.