DPFL-FUSION : Fake features Utilizing SHAP-based Importance for Optimized Noise

김대현(AI학과), 김태환(AI학과), 나상현(소프트웨어학부), 박서윤(경영학부)

2024 CUAI 중앙대학교 인공지능 학회 동계 컨퍼런스
Proceeding of 2024 Chung-Ang University Artificial Intelligence Winter Conference

Abstract

본 연구에서는 연합 학습(FL) 환경에서 차등 프라이버시(DP)를 보장하면서도 모델 성능 저하를 최소화하는 프레임워크 DPFL-FUSION 을 제안합니다. FL에서는 각 클라이언트가 로컬 데이터를 기반으로 모델을 학습하고, 그래디언트만을 중앙 서버로 전송함으로써 개인정보 보호를 기대할 수 있으나, 단순 그래디언트 전송만으로도 원본 데이터를 복원할 수 있다는 보안 취약점이 제기되었습니다. 이에 원본 데이터 유추가 불가능하도록 각 피처에 노이즈를 부여하는 DPFL이 제안되었으나, 노이즈로 인한 모델 성능 하락이 나타났습니다.
이에 본 연구에서는 SHAP 기반의 피처 중요도 평가를 통해 각 피처의 기여도를 정밀하게 산출하고, 이 정보를 활용하여 중요한 피처에는 낮은 노이즈를, 비중요 피처에는 높은 노이즈를 부여하는 적응형 노이즈 주입 전략을 구현하였습니다. 또한, 연속적 매핑 함수를 적용해 노이즈 분배를 세밀하게 조절하며, 값이 0인 가짜 피처를 추가하여 주요 피처의 노이즈를 대신 흡수함으로써 전체 DP 보장을 유지하면서도 모델의 학습 안정성과 예측 정확도를 향상시키는 효과를 입증하였습니다. 이와 같은 접근법은 기존 방식에 비해 보다 정밀한 프라이버시 보호와 우수한 성능을 동시에 달성할 수 있음을 보여줍니다.

Introduction

연합 학습은 분산된 각 클라이언트가 로컬 데이터를 사용하여 개별적으로 모델을 학습한 후, 해당 클라이언트의 그래디언트나 파라미터만을 중앙 서버로 전송하여 글로벌 모델을 업데이트하는 방식입니다. 이러한 방식은 원본 데이터를 직접 공유하지 않아 개인정보 보호 측면에서 유리하다는 장점을 가지고 있습니다.

그러나 Zhu et al. (2019) 의 연구를 통해 단순 그래디언트 정보만으로도 원본 데이터를 복원할 수 있는 보안 취약점이 발견되었으며, 이에 따라 FL 환경에서도 추가적인 보안 조치가 필요하게 되었습니다. 차등 프라이버시(Differential Privacy, DP)는 이러한 문제를 해결하기 위한 핵심 기술로, 데이터에 적절한 노이즈를 추가하여 단 한 행(row)만 달라져도 결과에 큰 차이가 나타나지 않도록 합니다. 이를 통해 특정 사용자의 데이터 기여 여부를 외부에서 유추하기 어렵게 만듭니다.

한편, 최근 SHAP(Shapley Additive exPlanations) 기법은 각 피처가 모델 예측에 미치는 기여도를 정량적으로 산출할 수 있는 강력한 도구로 부상하였으며, 이를 통해 기존의 단순 민감도나 분산 기반 평가보다 세밀하고 공정한 피처 중요도 분석이 가능해졌습니다.

본 연구에서는 FL 환경에서 발생할 수 있는 보안 취약점을 극복하고, 동시에 모델의 성능 저하를 최소화할 수 있는 새로운 프레임워크를 제안함으로써, 보안성과 성능의 균형을 동시에 달성하고자 합니다. 특히, 기존에는 모든 피처에 동일하게 노이즈를 주입하는 방식이 주를 이루었으나, 본 연구에서는 SHAP 기법을 활용한 피처 중요도 평가 및 연속적 매핑 함수를 통한 정밀한 노이즈 분배와 가짜 피처를 통한 노이즈 흡수 전략을 도입함으로써, 보다 세밀한 제어가 가능하도록 하였습니다. 이러한 접근법은 분산 학습 시스템에서 프라이버시와 성능의 트레이드오프 문제를 효과적으로 해결할 수 있는 새로운 방향을 제시합니다.

Methods

각 클라이언트는 로컬 데이터를 이용해 모델을 학습한 후, SHAP(Shapley Additive exPlanations) 기법을 적용하여 모델 예측에 대한 각 피처의 기여도를 산출합니다. SHAP 기법으로 예측된 i 번째 피처의 중요도 ϕ_i 는 매핑함수 $f()$를 통해 부여할 노이즈량 $f(\phi_i)$ 으로 변환되며, 이를 정규화하여 적응형 스케일러 a_i를 구해 노이즈에 곱하여 피처에 부여합니다.

$$\mathrm{E}\left[\|n'\|_2^2\right] = \mathrm{E}\left[\sum_{i=1}^{n}(a_i n_i)^2\right] = \sum_{i=1}^{n} a_i^2 \mathrm{E}\left[n_i^2\right] = \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2$$

$$\sum_{i=1}^{n} a_i^2 = 1 \text{ 이면 노이즈 에너지 유지 } \rightarrow a_i = \frac{f(\phi_i)}{\sqrt{\sum_{j=1}^{n} f(\phi_i)^2}}$$

DP 조건 보장을 위해서는 노이즈 에너지 $\mathrm{E}\left[\|n'\|_2^2\right]$ 이 σ^2로 유지되어야 합니다. 정규화 과정은 노이즈 주입 시 전체 노이즈 에너지가 계획된 수준(σ^2)과 일치하도록 보장하여 DP 조건을 보장하는 역할을 합니다.

기존의 연구들에서는 위와 같이 수학적 바운드를 고려하지 않고 중요도가 낮은 피처의 노이즈를 증가시키는 등 나이브하게 적용하였으나, 본 연구에서는 각 피처에 할당될 노이즈 크기를 세밀하게 조절하기 위해 연속적 매핑 함수를 적용합니다. 본 연구에서는 지수 함수, 역수 함수, 시그모이드 함수 등 다양한 연속적 매핑 함수를 실험하였습니다.

또한, 본 연구에서는 입력 벡터 $x \in \mathbb{R}^d$에 모든 값이 0인 가짜 피처를 추가하여 $x' = [x, 0] \in \mathbb{R}^{d+m}$로 확장합니다. 추가된 m개의 가짜 피처는 모든 데이터셋에서 0을 가지므로 두 인접 데이터셋 간 민감도 계산에 기여하지 않습니다. 결과적으로, 가짜 피처에 우선적으로 할당된 노이즈는 DP 조건에 영향을 주지 않으면서 실제 피처에 전달되는 노이즈 양을 효과적으로 감소시켜, 모델 업데이트의 민감도를 낮추고 예측 성능을 유지하는 역할을 합니다.

Results

실험은 MNIST 데이터셋을 대상으로 진행되었으며, SHAP 기반 피처 중요도 평가를 적용한 경우 기존의 분산 기반 평가 방식보다 더욱 세밀하게 피처를 구분할 수 있음을 확인하였습니다. 또한, 연속적 매핑 함수를 활용한 노이즈 분배 방식은 단순 비율 기반의 노이즈 할당보다 모델의 학습 안정성과 예측 정확도를 크게 향상시키는 결과를 보였습니다. 가짜 피처를 추가한 경우, 실제 중요한 피처에 전달되는 노이즈 양이 효과적으로 감소하여 전체 모델의 성능이 개선되었으며, 기존의 DPFL 방식과 비교하여 보안성을 유지하면서도 우수한 학습 성능을 달성하였습니다.

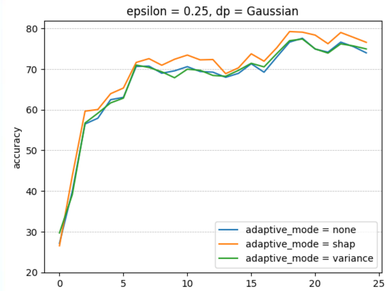

그림 1. SHAP vs Variance

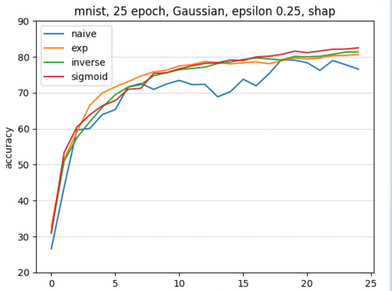

그림 2. Naïve vs Continuous Functions

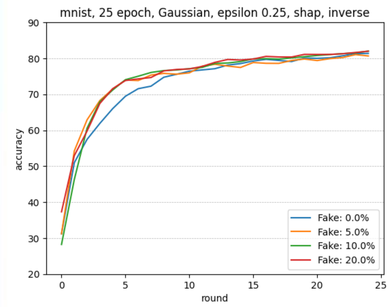

그림 3. Fake Features 0 vs 5 vs 10 vs 20

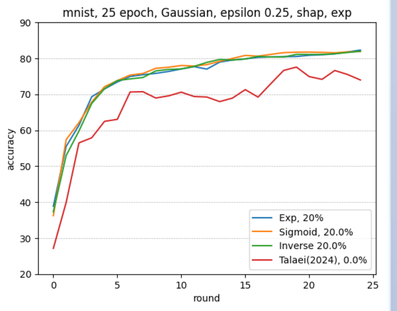

그림 4. Previous (Talaei, 2024) vs Ours

Conclusion

본 연구에서는 DPFL 환경에서 발생할 수 있는 그래디언트 기반 데이터 유출 문제를 해결하기 위해, SHAP 기반 피처 중요도 평가와 적응형 노이즈 주입, 그리고 가짜 피처를 활용한 노이즈 흡수 전략을 결합한 Adaptive DPFL-FUSION 프레임워크를 제안하였습니다. 제안된 방법은 DP 보장을 유지하면서도 모델의 학습 성능 및 일반화 능력을 향상시키는 데 효과적임을 실험적으로 입증하였으며, 이는 보안성과 성능의 트레이드오프 문제를 해결할 수 있는 새로운 접근법임을 시사합니다. 향후 다양한 데이터셋과 응용 분야에서 추가 검증 및 최적화를 통해, 보다 안전하고 효과적인 분산 학습 시스템 구축에 기여할 수 있을 것으로 기대됩니다.

Reference

[1] Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." NeurIPS. 2019.
[2] Abadi, Martin, et al. "Deep learning with differential privacy." ACM SIGSAC Proceedings. 2016.
[3] Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." Foundations and Trends® in Theoretical CS. 2014.
[4] Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." arXiv preprint. 2017.
[5] Talaei, Mahtab, and Iman Izadi. "Adaptive Differential Privacy in Federated Learning: A Priority-Based Approach." arXiv preprint. 2024.