노이즈가 가미된 연합학습 환경에 대한 클라이언트 기여도 측정 방법의 적합성 평가

신성국, 김동희*, 김광수**

성균관대학교

davidshyn@skku.edu *ym.dhkim@skku.edu **kim.kwangsu@skku.edu

Suitability Assessment on Client Contribution Estimation Methods for Federated Learning in Noisy Environments

Shyn Sung Kuk, Kim Dong Hee*, Kim Kwangsu**

Sungkyunkwan University

요 약

연합학습은 분산된 환경에서 직접 데이터를 접근하지 않고 각 클라이언트에서 학습한 모델 파라미터를 통합하여 연합 모델을 생성시키는 분산 머신러닝 기술이다. 연합 모델 성능 향상을 위해 연합학습 통합 알고리즘에 대한 연구가 활발하게 진행되고 있는 반면, 클라이언트 기여도 측정 방법 및 클라이언트 제거 기술에 대한 연구도 하나의 연합학습 연구 분야로 급부상하고 있다. 특히, 노이즈가 투입되어 데이터가 훼손될 수 있는 환경에서 '훼손된 클라이언트'(corrupted clients)를 클라이언트 기여도로 선별하는 기술이 필요하다. 본 논문에서는 연합학습에서 클라이언트 기여도 측정으로 기존에 연구된 대표적인 두 가지 방법, Federated LOO와 Federated SV를 소개한다. 이후 이 두 방법이 노이즈가 가미된 연합학습 환경에서 적절하게 작동되는지 노이즈가 투입된 환경에서 실험을 통해 적합성을 평가한다.

Ⅰ. 서 론

개인 정보의 유출에 민감한 현대 사회에서 연합학습은 데이터의 접근 없이 분산된 각 클라이언트가 각각의 데이터로 직접 학습을 진행한 후 학습된 모델 가중치들만을 중앙 서버에 전송하여, 중앙 서버가 이를 통합(FedAvg)[1]하여 하나의 연합 모델을 생성하는 프로세스로 진행된다.

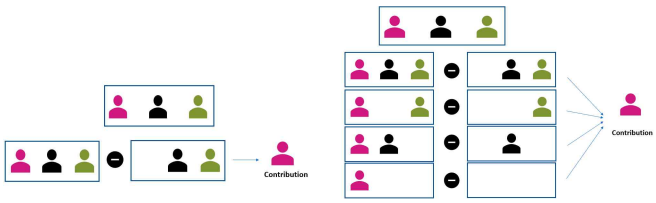
이 때, 연합 모델의 성능 향상을 위하여 각 클라이언트의 성능 기여도를 평가하여 중요하지 않은 클라이언트를 선별하는 클라이언트 제거 기술이 중요한 연구 분야로 대두되고 있다. 특히, 데이터가 불균등하게 분포된 non-IID 환경이나 노이즈가 포함된 데이터를 다수 보유한 상황에서 연합학습을 방해하는 클라이언트를 사전에 미리 선별하여 제거하는 것 [2]이 효과적이다.

하지만, 데이터를 직접 열람할 수 없는 연합학습 환경에서 데이터의 특징을 파악할 수 없기에 클라이언트의 성능 기여도 측정에 많은 제약이 따른다 [3]. 연합학습 환경에서 중앙 서버가 파악할 수 있는 정보는 클라이언트 모델 통합(FedAvg) 시 활용되는 각 클라이언트의 모델 가중치(local weights)와 활용하는 데이터의 크기(data size) 뿐이다 [1].

본 논문에서는 기존 연구에서 등장한 클라이언트 기여도 측정 방법으로, 클라이언트가 보유한 데이터를 직접 열람하지 않고 각 클라이언트의 성능 기여도를 측정할 수 있는 Federated LOO와 Federated SV를 소개하고 연합학습 환경에서의 적절성을 평가하고자 한다. 두 방법의 적절성을 판단하기 위해, 특정 클라이언트에 노이즈를 추가한 상황에서 기여도를 측정하고자 한다.

Ⅱ. 연합학습 클라이언트 성능 기여도 측정 방법

훼손된 데이터를 보유한 클라이언트가 존재하는 분산 환경에서 학습을 방해하는 '훼손된 클라이언트'(corrupted client)를 선별하기 위해서는 정확한 클라이언트 성능 기여도 측정이 필요하다.



(a) Federated LOO (b) Federated SV

그림 1. 클라이언트 기여도 측정 방법 설명

(1) Federated Leave-One-Out(Federated LOO)

Federated LOO[2]는 특정 클라이언트의 성능 기여도를 파악하기 위해 통합 라운드(r)마다 전체 클라이언트에서 그 클라이언트를 연합학습에 포함할 때와 포함하지 않을 때의 성능 변화를 비교하여 합한 것이다. 클라이언트의 가중치를 포함/제거를 반복하여 Federated LOO를 구할 수 있기 때문에 연합학습에서 활용이 가능하다. 클라이언트 i의 Federated LOO 수식은 다음과 같다.

$$loo_r(i) = acc(I_r^{all}) - acc(I_r^{all/\{i\}}), \qquad loo(i) = \sum_{r=1}^{R} loo_r(i)$$

(2) Federated Shapley Value(Federated SV)

Federated SV[2, 4]는 클라이언트들이 모두 참여하는 상황에서만 비교하는 Federated LOO와 달리, 모든 클라이언트들의 조합에서 특정 클라이언트를 포함할 때와 포함하지 않을 때의 성능 변화를 평균하는 방법이다. 그림 1(b)와 같이 세 명의 클라이언트를 가정하였을 때 첫 번째 클라이언트(빨강)의 기여도를 측정하기 위해 각 통합 라운드(r)마다 그 클라이언트를 포함한 모든 조합의 결과에서 포함하지 않은 조합의 결과를 뺀 값을 평균화한다. 클라이언트 i의 Federated SV는 다음 수식으로 계산된다.

$$SV_r(i) = \frac{1}{|I_r|} \sum_{S \subseteq I_r/\{i\}} \frac{1}{\binom{|I_r|-1}{|S|}} [acc(I_r^{S\cup\{i\}}) - acc(I_r^S)],$$

$$SV(i) = \sum_{r=1}^{R} SV_r(i)$$

Ⅲ. 실험

(1) 실험 구성

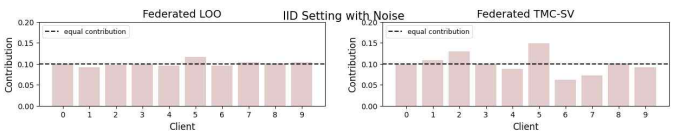
두 클라이언트 기여도 측정 방법의 적절성을 평가하기 위해 이미지 데이터셋인 MNIST[6]을 활용한다. 데이터 분포는 IID 환경과 non-IID 환경으로 구성한다. MNIST은 10개의 클래스로 이루어져 있어, 확실한 non-IID 환경으로 설정하기 위해 10개의 클라이언트로 구성하여 각 클라이언트별 하나의 클래스의 이미지를 보유하도록 설정한다. 또한, 10개의 클라이언트 중 **{클라이언트1,클라이언트6}**에게 보유한 데이터의 40%를 label noise를 주어 데이터를 훼손시키는 환경을 구성한다 [2]. 기여도 측정 방법으로 Federated LOO와 Federated SV을 사용하되, Federated SV의 경우 계산 시간을 단축하고자 TMC-SV[5]라는 근사 방법을 적용한다.

(2) Client Contribution Index(CCI) 측정

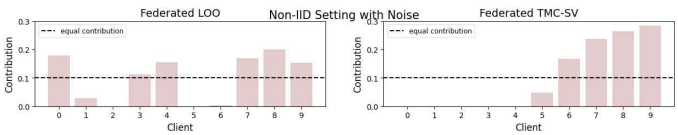
두 방법으로 측정된 값의 범위가 상이하기 때문에, 측정된 값들을 클라이언트 중요도 비율로 표준화한다. 주의할 점은 특정 클라이언트가 음수의 값을 가질 때 표준화하기 전에 0으로 미리 설정한다. 음수의 값을 가지지 않은 기여도(v_i)에서 클라이언트 i의 CCI는 다음과 같이 계산된다.

$$CCI_i = \frac{v_i}{\sum_{j=0}^{9} v_j}, \qquad where\, v_i \geq 0$$

(3) 실험결과



(a) IID 환경과 노이즈 설정



(b) non-IID 환경과 노이즈 설정

그림 2. 클라이언트별 기여도 비율(CCI) 결과.

일반적으로 IID 환경에서는 클라이언트들은 비교적 고른 분포의 CCI 값을 가지게 된다. 하지만 **{클라이언트1,클라이언트6}**에 노이즈를 추가하게 되면, 두 클라이언트의 기여도는 낮게 측정되어야 한다. 그림 2.(a)에서 확인할 수 있듯이, Federated LOO는 0.1의 CCI에서 소폭 하락한 결과를 볼 수 있지만, Federated SV는 클라이언트 6은 크게 하락한 반면, 클라이언트 1은 0.1보다 높게 측정되었다. 훼손된 데이터가 학습에 방해하는 현상을 반영하지 못한 결과라고 볼 수 있다.

Non-IID 환경(그림 2.(b))에서는 두 '훼손된 클라이언트'를 제외하고도 각 클라이언트의 분포가 크게 차이나는 결과를 볼 수 있다. 그럼에도 Federated LOO는 두 '훼손된 데이터'의 기여도(0.1보다 작은 CCI)를 잘 측정한 반면, Federated SV는 클라이언트 6의 기여도(0.1675)를 다른 클라이언트에 비해 비정상적으로 높게 측정하였다.

Data Setting	Evaluation Method	Contribution (equal = 0.1)					
		Client 1			Client 6		
		No Noise	Noise	Diff	No Noise	Noise	Diff
IID	Federated LOO	0.1025	0.0916	-0.0109	0.1008	0.0952	-0.0056
	Federated SV	0.1287	0.1085	-0.0202	0.0790	0.0626	-0.0164
non-IID	Federated LOO	0.1604	0.0289	-0.1315	0.1540	0.0024	-0.1516
	Federated SV	0	0	+0	0.0027	0.1675	+0.1648

표 1. '훼손된 클라이언트'의 기여도 변화

이는 노이즈를 추가하지 않은 환경과 비교하였을 때 비정상적인 측정과정임을 확인할 수 있다. 표 1에서 볼 수 있듯이, IID 환경과 non-IID 환경의 Federated LOO는 훼손된 클라이언트의 기여도가 정상적으로 하락하였지만, non-IID 환경의 Federated SV는 오히려 크게 상승하였다. 이러한 결과는 특정 환경에서 Federated SV가 적절하지 않음을 보여준다.

Ⅲ. 결론

본 논문에서는 연합학습 환경에서 학습에 방해하는 노이즈 데이터가 가미된 '훼손된 클라이언트'를 미리 발견하여 제거하기 위해 데이터를 직접 접근하지 않고 각 클라이언트의 성능 기여도를 측정하는 Federated LOO와 Federated SV를 소개한다. 이 두 방법을 훼손된 데이터를 보유하는 상황을 가정하여 실험하였다. 그 결과, Federated LOO가 '훼손된 클라이언트' 측정을 Federated SV에 비해 효과적으로 수행한다. 후에 진행할 연구로, non-IID 환경에서 특정 클라이언트의 기여도가 0인 것에 대한 적절성을 판단하고자 한다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신산업진흥원의 지원을 받아 수행된 헬스케어 AI 융합 연구개발 사업임 (No.S0316-21-1006)

참 고 문 헌

[1] B. McMahan, E. Moore, D. Ramage, S.Hampton, and B.A. y Arcas." Communtion-efficient Learning of Deep Netowrks from Decentralized Data. In Artificial Intelligence and Statstics, p. 1273-1282, PMLR, 2017.

[2] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. Song. "A Principled Approach to Data Vauation for Federated Learning." In Federated Learning, p. 153-167, Springer, 2020.

[3] P. Vepakomma, T. Swedish, R. Raskar, O. Gupta, and A. Dubey. "No peek: A Survey of Private Distributed Deep Learning." arXiv preprint, arXiv:1812.03288, 2018.

[4] L. S. Shapley. "A value of n-person games." In Theory of Games II, 1953..

[5] Ghorbani and J. Zou. "Data Shapley: Equitable valuation of data for machine learning." In International Conference on Machine Learning, p.2242-2251, PMLR, 2019

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278 - 2324, 1998