DBpia

탈중앙화된 환경에서의 연합 학습을 위한 평판 모델

김윤재○ 박상현 김재윤 문수묵

서울대학교 전기·정보공학부

gabriel1120@naver.com, {lukepark, jaeykim}@altair.snu.ac.kr, smoon@snu.ac.kr

Reputation Model for Decentralized Federated Learning

Yunjae Kim○ Sanghyeon Park Jae-Yun Kim Soo-Mook Moon

Department of Electrical and Computer Engineering, Seoul National University

요 약

연합 학습은 클라이언트들이 각자 학습을 수행하여 중앙화된 모델을 학습시키는 기계 학습의 방법이다. 탈중앙화된 분산 네트워크 환경에서도 연합 학습을 수행할 수 있는데, 이 경우 고의적으로 학습을 방해하려는 악의적인 노드가 존재할 수 있다. 악의적인 노드는 잘못된 학습 정보를 배포함으로써 정상적인 노드의 학습을 방해한다. 본 연구에서는 악의적인 노드가 포함된 분산 네트워크에서 연합 학습을 수행할 때, 평판 설정을 통해 정상적인 노드들이 받는 영향을 최소화하는 방법을 제안한다. 실험 결과 본 방법을 통해 얻어진 모델은 악의적인 노드가 없는 환경에서와 유사한 수준의 정확도를 보였다.

1. 서 론

연합 학습(Federated Learning)은 많은 수의 클라이언트가 각자의 데이터를 가지고 학습을 수행하여 중앙화된 모델을 학습시키는 기계 학습(machine learning) 방법이다[1]. 연합 학습은 개인 모바일 기기들의 정보를 취합하는 상황에서 사용될 수 있는데, 기기별로 학습을 수행한 뒤 모델의 변경 사항을 서버에서 취합한다[2]. 가령 인공신경망의 경우 클라이언트의 가중치 갱신 정보를 서버가 취합하여 통합하는 것으로 중앙화된 모델의 성능을 향상시킨다.

기존의 연합 학습은 클라이언트-서버 구조를 상정한다. 그러나 이러한 구조는 서버의 성능에 해당하는 상한을 가지며, 대규모 학습을 수행하는 것에 한계가 있다[3]. 이를 해결하고자 분산 네트워크에서 연합 학습을 수행하는 방법이 있다. 노드들은 중앙화된 서버와 통신할 필요가 없고, 이를 통해 보안의 강화와 학습 효율 향상의 효과를 얻을 수 있다[4].

그러나 이러한 탈중앙화 환경에서 연합 학습을 적용하면 학습이 어떻게 일어나고 있는지 실시간으로 판단하는 중앙 서버가 없기 때문에 학습의 방향성을 바꾸기가 어렵다. 특히 네트워크 참여자 중 악의적으로 학습을 방해하는 노드가 존재하는 상황이 문제시된다. 기존의 연합 학습에서는 중앙 서버에서 해당 노드를 배제해 학습할 수 있지만, 탈중앙화 연합 학습에서는 중재할 서버가 없다[5].

따라서 연합 학습 과정에서 악의적인 노드를 배제하고 정상적인 노드가 제시한 가중치를 포함해 긍정적인 방향으로 학습을 하기 위한 방안이 필요하다.

본 연구에서는 탈중앙화된 연합 학습 환경에서 정상적인 노드가 악의적인 노드를 학습에서 배제하기 위해 다른 노드들에 대한 평판(reputation)을 설정하는 방법을 제시한다. 이를 위해 악의적인 노드가 존재하는 연합 학습 환경을 만들고, 평판 설정을 위한 정책(policy)을 제시하고 구현했다.

2. 실험 환경

본 연구에서는 실험 및 평가에 Fashion-Mnist 데이터셋을 사용하였다. Fashion-Mnist 데이터셋은 10개의 범주를 가진 28 x 28 픽셀(pixel)로 이루어진 이미지 데이터셋이며, 총 60,000개의 학습을 위한 데이터셋과 10,000개의 테스트셋으로 구성되어 있다[6]. 가정하는 환경은 다음과 같다.

- 네트워크에 참여하는 노드의 수는 총 100개이다. 각 노드는 600개의 학습 데이터와 99개의 테스트 데이터를 가진다.
- 연합 학습 정도를 평가하기 위해 별도의 마스터 테스트셋을 둔다. 마스터 테스트셋은 100개의

* 이 논문은 BK21플러스 사업 및 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2020R1A2B5B02001845)

www.dbpia.co.kr

데이터로 구성된다. 노드들은 마스터 테스트셋에 접근할 수 없다.

- 연합 학습 과정은 100번의 반복(Rounds)을 통해 진행한다. 학습에 참여하는 노드들은 각 라운드 시작 시에 자신의 학습 데이터를 통해 1회 학습을 진행하고, 라운드 종료 전에 자신의 가중치를 다른 노드들과 공유한다.
- 각 라운드가 끝날 때 각 노드의 가중치는 다음과 같이 갱신된다.

$$W_i \leftarrow \frac{1}{len(N)}\left(W_i + \sum_{j \in N-i} W_j\right)$$

W_i : i번째 노드의 가중치
N : 모든 노드의 집합
$len(N)$: 노드의 총 개수

- 악의적인 노드가 포함된 네트워크는 전체 참여자 수의 70%를 악의적인 것으로 가정한다.

2.1 악의적인 노드 설정

악의적인 노드는 Fashion-Mnist에 대한 혼동행렬(Confusion Matrix)에서 주로 혼동되는 범주를 반대로 학습하도록 설정했다[7]. 본 실험에서는 Shirt를 T-shirt로, Pullover를 Coat로 학습한다. 고의적으로 특정 범주를 다른 범주로 인식하게 학습하면 가중치가 잘못된 방향으로 갱신되고, 가중치 공유 시 다른 정상적인 노드들의 가중치에 영향을 줄 수 있다.

3. 평판 시스템

본 연구의 궁극적인 목표는 악의적인 노드들이 더 많은 환경에서도 정상적인 노드들이 정상적인 노드들만 선택해 연합 학습을 하여, 올바른 학습 방향을 가지게 하는 것이다. 어떠한 노드가 다른 노드들에 대한 평판을 설정하는 정책은 다음과 같다.

① 노드는 학습에 앞서 가상 노드(virtual node)를 생성한다.
② 가상 노드에 현재 가중치 정보를 업데이트한다.
③ 노드가 가진 테스트셋으로 가상 노드의 모델 정확도 $acc_{current}$를 구한다.
④ 수신한 다른 노드의 가중치를 가상 노드에 대입한다.
⑤ 노드가 가진 테스트셋으로 가상 노드의 모델 정확도 $acc_{neighbor}$를 구한다.
⑥ 다음 조건을 만족하면 다른 노드에 대한 평판을 1, 그렇지 않으면 0으로 설정한다.

$$acc_{neighbor} \geq acc_{current} \times 0.9$$

본 실험에서는 평판을 설정할 때 10%의 허용 범위를 두는 것이 가장 좋은 판단력을 보였다. 판단력이란 정상적인 노드들에게 평판을 1로, 악의적인 노드들에게 평판을 0으로 설정할 확률을 의미한다. 적절한 허용 범위를 두면 보다 다양한 노드들의 가중치를 받아와서 편향되지 않게 학습하는 효과가 있다.

각 노드는 평판 설정을 완료한 후 다음의 공식에 따라 가중치를 갱신한다.

$$W_i \leftarrow \frac{1}{len\left(N_{rep=1}\right)+1}\left(W_i + \sum_{j \in N-i}\left(rep_j \times W_j\right)\right)$$

W_i : i번째 노드의 가중치
$N_{rep=1}$: 평판이 1인 노드의 집합
$len\left(N_{rep=1}\right)$: 평판이 1인 노드의 개수

4. 평가 및 분석

시스템의 우수성을 평가하기 위해 다음의 세 가지 환경에 대한 실험을 수행했다.

1) 악의적인 노드 없이 정상적인 노드만 존재하는 환경
2) 악의적인 노드 70%와 정상적인 노드 30%로 구성된, 평판 설정을 하지 않는 환경
3) 악의적인 노드 70%와 정상적인 노드 30%로 구성된, 평판 설정을 하는 환경

그림 1의 좌측 그래프는 위의 세 실험에 대한 정상적인 노드들의 정확도(accuracy)를 비교한 것이다. 정상적인 노드만 존재하는 환경과 평판 설정을 하는 환경에서의 정확도가 유사하게 나타났다. 반면, 평판 설정을 하지 않는 경우에는 다른 두 경우보다 낮은 정확도를 보였다. 그림 1의 우측 그래프는 손실(loss)을 나타낸 것이다. 평판 설정을 하는 정상적인 노드들은 악의적인 노드가 없는 환경에서의 손실과 유사한 면모를 보였다. 반면, 평판 설정 과정을 거치지 않은 경우 높은 손실을 보였다. 평판 설정 과정을 거친 경우 50번의 학습 이후 손실이 0.5 이하로 줄어든 반면, 평판 설정을 거치지 않은 경우 손실이 1.5 이하로 줄어들지 않았음을 확인할 수 있다.

평판 설정을 하지 않으면 악의적인 이웃 노드와 정상적인 이웃 노드의 가중치를 모두 고려한 평균으로 가중치를 갱신한다. 따라서 정상적인 노드도 라운드가 거듭될수록 부정적 영향을 받는다. 반면, 평판 설정을 하는 경우에는 악의적인 노드들의 가중치를 무시할 수 있다. 비록 평판 설정을 하지 않는 환경에서 악의적인 노드들의 수가 70%로 다수를 차지하고 있으나, 30%의

www.dbpia.co.kr

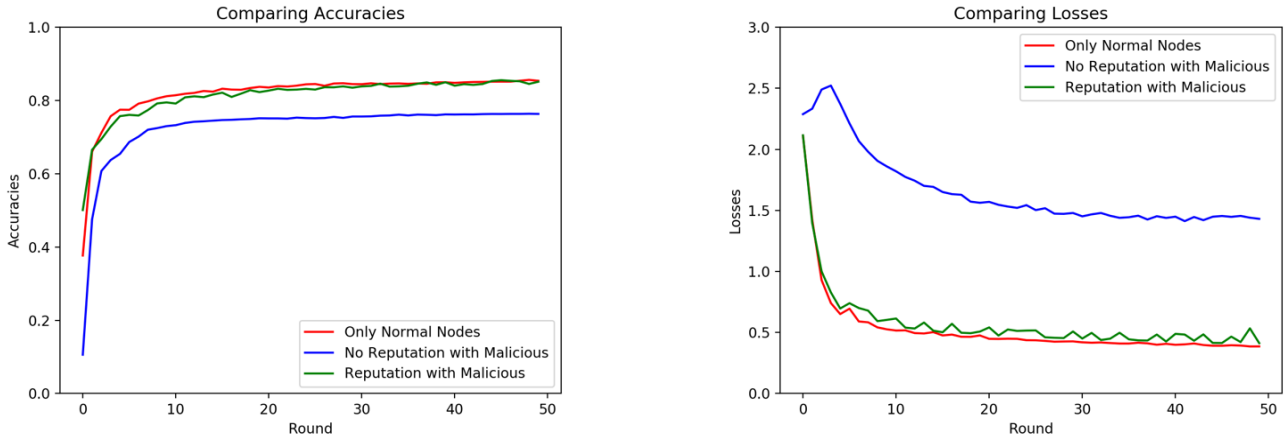

그림 1 정상적인 노드들의 평균 정확도(左) 및 평균 손실(右)

정상적인 노드들의 학습이 전체 노드들의 학습에 긍정적인 영향을 미친다는 것을 알 수 있다. 실제로 정확도의 경우, 50번의 학습 후 0.7 이상으로 증가한다는 것을 확인할 수 있었다. 그러나 라운드가 거듭될 수록 정상적인 노드가 다수의 악의적인 노드들로부터 영향을 받게 되고, 학습에 긍정적인 영향을 줄 수 없게 된다. 그 결과 그림 1의 우측 그래프에서처럼 손실이 더 이상 줄어들지 않게 된다.

위의 평가 결과로부터 다음과 같은 두 사실을 추론할 수 있다. 첫 번째는 범주를 바꿔서 학습한 악의적인 노드가 정상적인 노드들의 가중치에 악영향을 미칠 수 있다는 것이다. 중앙 서버가 없는 환경에서 서버의 중재 없이 연합 학습을 할 때, 악의적인 노드가 존재한다면 정상적인 노드의 학습 효율이 떨어지고 잘못된 방향으로 학습될 수 있음을 시사한다. 두 번째는 평판 설정을 통해 악의적인 노드들을 학습에서 배제할 수 있다는 것이다. 탈중앙화된 연합 학습 환경에서 평판을 공유할 수 없는 환경이라면, 노드 스스로 다른 노드들에 대한 평판을 매길 수 있어야 한다. 본 연구에서 제시한 평판 설정 정책을 이용하면 다른 노드의 평판 정보를 받지 않고 연합 학습에서 주고받는 가중치만으로 다른 노드들에 대한 평판 설정을 할 수 있다. 이는 모바일 기기 등에서 연합 학습을 활용할 때 사용자가 자신의 평판 설정 정책을 개인별로 설정하여 보다 목적에 맞는 연합 학습을 할 수 있음을 의미한다.

5. 결론 및 향후 연구

본 연구에서는 탈중앙화된 연합 학습에서 생길 수 있는 문제점인 악의적인 노드의 존재에 대한 대처 방안인 평판 시스템을 제안하고 실험하였다. 악의적인 노드가 존재하는 연합 학습 환경을 만들어서 해당 문제가 실제로 존재함을 보였고, 평판 시스템을 통해 문제를 해결했다.

비록 본 연구에서 제시한 평판 시스템이 정상적인 노드들의 학습 정확성을 향상시키지만, 별도의 평판 설정 과정을 요구하기 때문에 학습에 소요되는 총 시간이 증가하게 된다. 보다 효율적인 평판 설정 방법을 찾아내는 것이 후속 연구이다.

참 고 문 헌

[1] Konečný Jakub et al, "Federated learning: Strategies for improving communication efficiency", 2016.

[2] LIM, Wei Yang Bryan et al., Federated learning in mobile edge networks: A comprehensive survey. arXiv preprint arXiv:1909.11875, 2019.

[3] LI, Tian et al., Federated learning: Challenges, methods, and future directions, 2019.

[4] J. Kang, Z. Xiong, D. Niyato, S. Xie and J. Zhang, "Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to Combining Reputation and Contract Theory," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 10700-10714, Dec. 2019.

[5] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang and M. Guizani, "Reliable Federated Learning for Mobile Networks," IEEE Wireless Communications, 2020.

[6] Xiao, Han, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.", 2017.

[7] Dong, Manqing & Yao, Lina & Wang, Xianzhi & Benatallah, Boualem & Zhang, Shuai, GrCAN: Gradient Boost Convolutional Autoencoder with Neural Decision Forest, 2018

www.dbpia.co.kr