연합학습에서 이상치 데이터가 미치는 영향 평가

정성우, 유준혁*
대구대학교, *대구대학교
learningstaedy0j0@gmail.com, *joonhyuk@daegu.ac.kr

Evaluating the Impact of Out of Distribution Data on Federated Learning

Seong-Woo Jeong, Joonhyuk Yoo*
Daegu University, *Daegu University

요 약

중앙 집중화된 전통적인 기계학습 방식은 개인정보 보호 문제와 컴퓨팅 자원이 제한된 에지 디바이스에서의 학습이 어려운 문제가 존재한다. 이를 해결하기 위해 새로운 분산 학습 패러다임인 연합학습에 대한 연구가 최근 활발하게 이루어지고 있다. 본 논문은 연합학습 프레임워크에서 사용자의 실수로 발생할 수 있는 이상치 데이터에 대한 영향력을 분석하기 위해 Non-IID CIFAR-10 데이터세트의 이미지 분류 작업을 위한 연합학습에서 SVHN 데이터 셋을 이상치 데이터로 하여 이상치 데이터가 포함되어 있는 클라이언트 비율과 이상치 데이터 비율에 따른 정확도를 비교한다. 실험 결과, 모든 클라이언트에서 10%의 이상치 데이터 비중을 가지고 있을 때 약 3% 정도의 정확도 감소가 나타나고, 이상치 데이터 비중이 증가하면서 정확도 감소폭도 함께 증가함을 보여준다,

Ⅰ. 서 론

최근 자율주행, 의료, 스마트 시티 등과 같이 개인정보 보호가 중요한 다양한 분야에서 연합학습이 적용되고 있다. 기존의 기계학습 방법은 에지 디바이스에서 생성된 데이터를 서버로 전송하여 중앙 집중화 방식으로 학습하기 때문에 데이터의 개인정보 보호에 대한 우려가 높다. 이와 더불어 에지 디바이스가 생성하는 방대한 양의 학습 데이터를 서버에서 처리하는데 필요한 컴퓨팅 자원이 기하급수적으로 증가하여 시간적/금전적 부담이 급증하였다. 이를 해결하기 위해 McMahan et al.이 연합학습(Federated Learning)이란 새로운 분산 학습 패러다임을 제시한 바 있다[1].

연합학습은 개인정보보호를 위해 클라이언트라 불리는 다수의 사용자 에지 디바이스가 저장하고 있는 원시 데이터를 중앙서버에 전송하지 않으면서 딥러닝 모델을 최적화하는 분산 학습 방법 중 하나이다. 연합학습의 최적화 방식은 그림 1과 같다. 먼저 각각의 클라이언트가 중앙 서버의 초기 모델을 수신하고 로컬 데이터를 통한 확률적 경사하강법(Stochastic Gradient Descent)을 수행한 뒤 생성된 매개변수를 서버로 전송한다. 그런 다음 서버는 모든 클라이언트의 매개변수를 수신한 후에 평균하여 초기 학습 모델을 업데이트 한다. 이러한 일련의 과정을 여러 번 반복 하여 최적화를 진행한다. 훈련에 사용되는 원시 데이터가 클라이언트에서만 관리되는 학습 방식 덕분에 개인정보 보호가 가능해지며 클라이언트의 로컬 데이터 전체를 보내는 것이 아닌 매개변수 값만 전송하기 때문에 상대적으로 서버에서 처리해야 하는 컴퓨팅 자원의 부담이 줄어든다.

하지만, 연합학습은 각 클라이언트에서 사용자 상호 작용에 의해 학습 데이터가 생성되기 때문에 사용자의 실수나 악의로 인한 이상치(OOD, Out of Distribution) 데이터가 발생하기 쉽다. 이상치 데이터란 일반적인 데이터 패턴과 다르게 매우 다른 패턴을 가지고 있는 데이터를 뜻한다. 예를 들면 '개' 이미지 데이터 집합에 '고양이' 이미지가 섞여 있는 경우 '고양이' 이미지를 이상치 데이터라고 볼 수 있다. 이러한 이상치 데이터는 다양한 실험을 통해 딥러닝 모델의 성능을 크게 좌우하는 것이 입증되었고, 이를 해결하기 위해 다양한 연구가 진행되고 있다[2,3]. 따라서 본 논문에서는

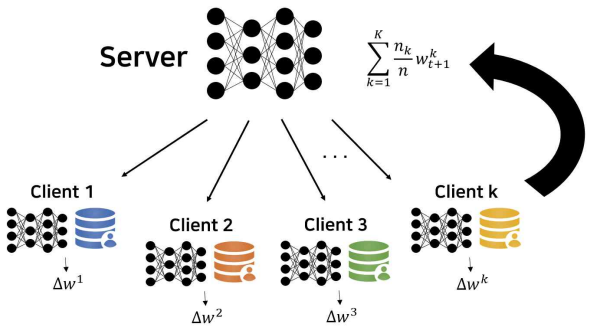


그림 1. 연합학습 최적화 방식

연합학습 프레임워크에서 이상치 데이터가 미치는 영향을 평가하여 문제점을 분석하고 초기 결과를 제시하고, 향후 연합학습에서 이상치 데이터의 적절한 처리방법의 필요성을 제기하고자 한다. 이와 더불어 이미지 분류 작업을 위한 연합학습시 이상치 데이터가 포함되어 있는 클라이언트의 비율과 이상치 데이터의 비율에 따른 영향력을 정확도 차이로 측정하였으며, 연합학습의 실제 적용 시 허용 가능한 목표 이상치 비율을 실험을 통해 제시한다.

본 논문은 다음과 같이 구성된다. Ⅱ장에서는 이미지 분류 작업에서 이상치 데이터가 미치는 영향에 대해 평가하기 위한 실험 방법을 설명한다. Ⅲ장에서는 실험 결과를 분석하고 Ⅳ장에서는 결론을 제공한다.

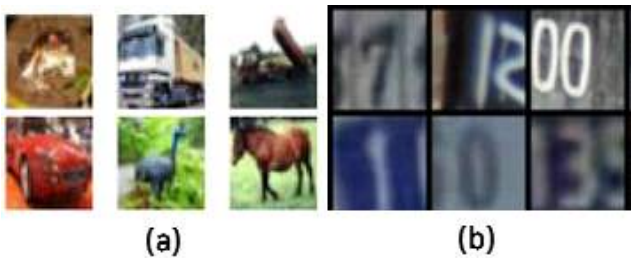
Ⅱ. 실험 평가 방법

이미지 분류를 위한 학습 데이터세트는 Non-IID CIFAR-10 데이터세트를 사용한다. 그림 2(a)에서 볼 수 있는 CIFAR-10 데이터세트는 컴퓨터 비전 연구에서 널리 사용되는 데이터세트 중 하나로 비행기, 자동차, 새, 고양이, 사슴, 개, 개구리, 말, 배 그리고 트럭 클래스에 대한 총 60,000개(학습 데이터 50,000개, 평가 데이터 10,000개)의 32x32 RGB 이미지 데이터이다[4]. Non-IID란 데이터가 독립적이지 않고 동일하지 않은 확률 분

포를 가지고 있다는 의미이다. 연합학습은 물리적으로 떨어진 곳에서 생성된 각 클라이언트의 로컬 데이터를 사용하기 때문에 학습 데이터를 Non-IID 데이터로 상정하고 성능 평가를 진행한다[5]. 따라서 CIFAR-10 학습 데이터세트 50,000개에서 10개의 클래스 중 5개의 클래스를 무작위로 선택하고, 이미지 데이터 수 또한 무작위로 하여 하나의 클라이언트에서 로컬 데이터세트를 생성한다. 이어서 모든 데이터가 포함되도록 총 10개의 로컬 데이터세트를 생성하였고, 이를 Non-IID CIFAR-10이라 정의하였다.

학습 데이터에 포함되지 않은 패턴을 가지는 이상치 데이터로는 SVHN(Street View House Numbers) 데이터세트를 사용한다. 그림 2(b)에서 볼 수 있는 SVHN 데이터세트는 Google이 Google 지도를 만드는 과정에서 촬영한 집들의 번호판 이미지이다. 이는 0부터 9까지 숫자 10개로 이루어져 있고 총 73,257개의 32x32 크기의 RGB 이미지 데이터이다[6].

이상치 데이터의 영향력 평가를 위해 본 논문에서는 SVHN을 Non-IID CIFAR-10 학습 데이터세트에 다양한 비율로 추가하여 실험을 진행하였고, 테스트 데이터세트로는 CIFAR-10의 테스트 데이터세트 10,000개를 그대로 사용하였다.



(a) (b)

그림 2. (a) CIFAR-10 데이터세트의 일부 이미지, (b) SVHN 데이터세트의 일부 이미지

실험 환경으로는, AMD RYZEN™ 7 1700 8-Core 3.0 GHz CPU와 GeForce RTX 2080 TI GPU 2개가 장착되어 있는 서버 컴퓨터를 사용하였고, Ubuntu 20.04 LTS 환경에서 PyTorch ver.1.7.1을 사용하여 실험을 진행하였다. 모델은 ResNet56을 사용하였고, 10개의 클라이언트를 가정하여 시뮬레이션을 진행하였다. 각 클라이언트에서는 10의 배치크기와 0.03의 학습률과 함께 확률적 경사하강법을 사용하여 5 epoch만큼 매 라운드 학습을 진행하였고, 학습된 각 클라이언트의 매개변수를 평균하여 총 100 라운드 동안 서버 모델의 업데이트를 진행하였다.

Ⅲ. 실험결과 분석

표 1은 클라이언트 및 이상치 데이터 비율에 따른 서버 모델의 정확도 값이다. 표에 나타나 있는 클라이언트 비율이란 이상치 데이터를 포함하는 클라이언트의 비율을 나타낸다. 다음으로 이상치 데이터 비율은 각 로컬 데이터세트에 추가된 이상치 데이터의 비율을 뜻한다.

Baseline은 이상치 데이터 없이 Non-IID CIFAR-10 데이터만으로 학습했을 때의 정확도 값이다. 클라이언트와 이상치 비율 모두 10%로 아주 작은 비율을 가지고 있을 때에는 0.1%의 아주 미미한 정확도 감소를 나타냈으며, 모든 클라이언트가 10%의 이상치 비율을 가지고 있을 때 약 3% 정도의 정확도 감소가 나타났다. 이 후 모든 클라이언트에서 이상치 비율의 증가에 따른 비교를 하였을 때 상당히 큰 폭으로 정확도 감소가 일어나는 것을 관측할 수 있었다.

이러한 결과를 바탕으로 연합학습의 실제 적용시 모든 클라이언트에서 이상치 데이터가 생성된다고 가정하였을 때 로컬 데이터세트에서 이상치 데이터가 10%의 비율까지는 허용할 수 있는 오차범위로 간주될 수 있다고 판단된다.

표 1. 클라이언트 및 이상치 데이터 비율에 따른 정확도 비교

Index	Client Ratio (%)	OOD Ration (%)	Acc (%)
Baseline	**0**	**0**	**69.5**
1	10	10	69.4
2	30	10	67.7
3	50	10	69.2
4	70	10	67.2
5	100	10	66.2
6	100	25	61.6
7	100	50	55.4
8	100	75	50.1
9	100	100	48.2

Ⅳ. 결론

본 논문에서는 연합학습시 이상치 데이터의 영향력을 평가하기 위해 클라이언트와 이상치 데이터 비중에 따른 비교실험을 진행하였다. 실험을 통해 모든 클라이언트에서 10%의 이상치 데이터 비중을 가지고 있을 때 약 3%의 허용 가능한 정확도 감소가 나타나는 것을 관측할 수 있었고, 연합학습의 실제 적용시 이를 허용 가능한 목표 비율로 제안하였다. 향후 본 연구를 확장하여 분산학습 프레임워크에서 이상치 데이터를 적절하게 보상할 수 있는 연합학습 알고리즘에 대한 연구를 진행할 계획이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020R1A2C1014768).

참 고 문 헌

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” In Artificial intelligence and statistics, vol. 54, pp. 1273-1282, Apr. 2017.

[2] D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep anomaly detection with outlier exposure.” arXiv preprint arXiv:1812.04606. Dec. 2018.

[3] D. Hendrycks, and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” arXiv preprint arXiv:1610.02136. Oct. 2016.

[4] A. Krizhevsky, and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[5] P. Kairouz, B. McMahan, ... & S. Zhao, “Advances and Open Problems in Federated Learning,” arXiv preprint arXiv:1912.04977. Dec. 2019.

[6] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning”, NIPS Workshop on Deep Learning and Unsupervised Feature Learning, Dec. 2011.