프라이버시 보호를 위한 연합학습 기반 핑거프린팅 측위 시스템

박준하, 김효원, 김선우
한양대학교 전자컴퓨터통신공학과

{eric0725, khw870511, remero}@hanyang.ac.kr

Federated Learning-based Privacy-preserving Fingerprinting Localization Systems

Junha Park, Hyowon Kim, and Sunwoo Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 Wi-Fi의 수신신호세기(RSSI) 데이터 학습을 통한 연합학습(federated learning) 기반 핑거프린팅 측위 시스템을 제안한다. 제안하는 측위 시스템은 분산형의 학습(decentralized learning)을 이용하여 사용자의 프라이버시 보호가 가능하다. 시뮬레이션 결과를 통해 프라이버시 보호가 불가능한 중앙집중형의 학습(centralized learning)을 이용한 측위 시스템과의 측위 정확도 비교하였다.

Ⅰ. 서 론

전세계적으로 모바일 단말을 사용하면서 생성되는 사용자의 센서(GPS, 카메라, 마이크 등) 및 5G 이동통신 데이터를 이용한 다양한 머신러닝 연구가 활발히 진행되고 있다[1]. 다양하고 막대한 양의 데이터에는 개인정보에 민감한 데이터들이기 때문에 보안성과 데이터의 익명성이 보장되어야 한다. 이를 해결하기 위한 머신러닝 기법으로는 대표적으로 연합학습(federated learning)이 있다. 연합학습은 스마트폰 같은 모바일 단말에 저장되어 있는 분산 데이터를 통해 개개인이 서버의 머신러닝 모델을 훈련하는 분산 머신러닝의 한 종류이다[2, 3].

이와 같은 연학학습을 이용하여 본 논문에서는 수신신호세기(RSSI) 데이터 학습을 통한 연합학습(federated learning)기반 측위 시스템을 제안한다. 또한 기존 방식의 머신러닝 기반 측위 시스템[4]과 ㅁ 논문에서 제시하는 알고리즘의 측위 성능을 비교하고 적용 가능성을 평가한다.

Ⅱ. Federated Learning 기반 실내 측위

서버에서 모든 데이터를 학습하는 기존의 중앙 집중형 학습과 달리, 연합학습은 사용자의 단말에서 각 단말들이 얻은 데이터를 이용하여 학습을 진행한다. 각 단말은 학습된 모델의 가중치를 서버로 보내고, 서버는 수신한 모든 가중치를 통합한다. 연합학습은 데이터가 아닌 학습된 가중치만 서버로 보내, 사용자의 프라이버시를 보호할 수 있다.

학습 데이터 수집을 위한 네트워크에는 총 J개의 Wi-Fi AP(access point)가 분포되어 있으며, N명의 사용자 단말은 각 AP로부터 신호세기를 수집한다. 학습을 위한 각 단말의 MLP(multi-layer perceptron) 모델은 입력층, 다중의 은닉층(hidden layer), 그리고, 출력층으로 구성된다. 단말의 i번째 샘플의 입력벡터는 다음과 같다.

$$x_i = [r_{i1}, r_{i2} \dots, r_{ij}, \dots, r_{ij}], \quad (1)$$

r_{ij}는 j번째 Wi-Fi AP(access point)가 수신하는 단말의 신호세기이다. 서버는 최초 광역모델(global model)을 생성하며, 모든 단말은 다음과 같이 손실함수(loss function)를 최소화한다.

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i) \quad (2)$$

각 단말 u는 가중치 $\boldsymbol{w}$를 추정하기 위해 각자의 데이터를 사용하여 학습을 진행한다. 이때, 광역모델은 학습에 사용되는 데이터의 형태에 따라 매번 다르게 디자인된다. 학습을 마친 각 단말들은 자신들의 가중치들을 서버로 보내게 된다. 서버는 전송 받은 가중치들을 각 단말 마다 학습시킨 데이터의 개수의 비율을 곱하여 평균을 취하게 된다. 이때, 광역모델의 가중치는 다음 두 식을 이용한 것과 같다.

$$w^{t+1} = \frac{1}{H^t} \sum_{u=1}^{N} m_u^t \, w_u^t, \quad (3)$$

$$H^t = \sum_{u=1}^{N} m_u^t, \quad (4)$$

m는 학습 샘플들의 개수, $\boldsymbol{x}_i$, $\boldsymbol{y}_i$는 각각 입력 과 출력 레이블들을 나타낸다. 또한, H는 전체 학습 샘플의 개수, t는 각 단말들의 학습을 거치고 서버로 돌아오는 광역모델 epoch이다. 이와 같이 가중치들을 평균내는 1 번의 epoch를 거친 후 서버는 첫 번째 학습을 마치게 된다. 업데이트된 광역모델은 각 단말들이 이전에 진행한 한 epoch의 일련의 과정들을 다시 거치는데 사용된다. 이 과정들은 광역모델의 가중치가 수렴할 때 까지 학습을 진행하게 된다. 연합학습의 일련의 과정들은 그림 1 과 같다.

Ⅲ. 시뮬레이션 성능평가

서버의 모델은 중앙집중형의 학습과 동일한 MLP 로 디자인하고, 각 사용자들의 단말들이 가지고 있는 수신신호세기 값들이 갖는 특징을 얻기 위한 학습을 통해 실내에서 단말들의 좌표를 출력 레이블로 도출한다. 시뮬레이션에 사용되는 데이터베이스로는 UJIIndoorLoc 데이터베이스를 이용하였다[5]. 본 논문의 시뮬레이션에 사용된 매개변수들은 표 1 과 같다. 비교를 위한 중앙집중형의 학습 모델 구조 또한 동일한 모델 구조와 매개변수들로 구성하여 학습에 사용한다. 학습에 사용되는 데이터베이스는 520 개의 AP 가 각 단말에게 획득한 수신신호세기값들을 입력 데이터로 이용한다. 측정되지 않은 수신신호세기는 일정한 상수 $C = 100$으로 설정하였고 출력 레이블은 측정된 단말의 위치로 UTM 좌표계(미터 단위)로 표현된 위도와 경도로 나타낸다.

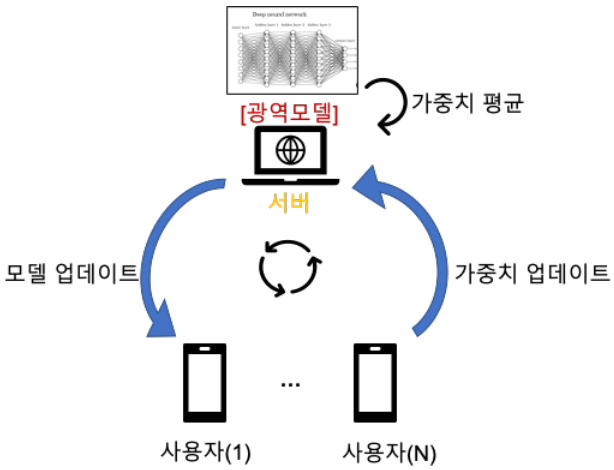

그림 1. 연합학습 모델 구조

연합학습의 마지막 레이어의 활성함수(activation function)는 연속된 좌표값을 출력 레이블로 가지기 위해 회기모델(regression model)에 적절한 linear 함수를 이용한다. 각 사용자들은 자신의 단말로 수집한 데이터로 20 번의 epoch를 수행하고 15 명의 사용자들이 각자 학습시킨 데이터를 자신의 데이터 개수에 비례하여 도출되는 가중치들에 가중치를 주어 평균을 취해 서버에서 새로운 광역 모델의 가중치에 적용한다.

본 논문의 시뮬레이션에서는 위와 같은 과정들을 25 번의 round를 수행하여 연합학습의 과정을 마친다. 그림 2 는 연합학습 알고리즘이 매 round마다 도출되는 오차 평균을 중앙집중형의 학습을 통한 결과와의 비교를 보여준다. 중앙집중형의 학습을 통한 결과는 약 7m의 오차범위를 가지고 있음을 보여준다. 이에 연합학습을 거친 결과는 round가 진행됨에 따라 약 10m정도로 중앙집중형의 학습 결과에 2~3m의 거리 차를 보여준다. 이를 통해 서버에서 모든 데이터를 가지고 학습을 진행하지 않고도 각 단말들이 연합학습을 통해 도출한 결과값은 데이터 프라이버시를 보존함에 동시에 측위 오차 또한 중앙집중형 학습으로 도출된 결과에 근접함을 보여준다.

표 1. 연합학습 모델 매개변수

Parameters	Value
optimizer	Adam
learning rate	0.0001
β_1, β_2	0.1, 0.99
hidden layer format	512x256x128x64
drop out	0.2
activation function	sigmoid x 4, linear
batch size	100
local model epoch	20
global model epoch	25
train data set	19937
the number of users	15
loss function	mean absolute error

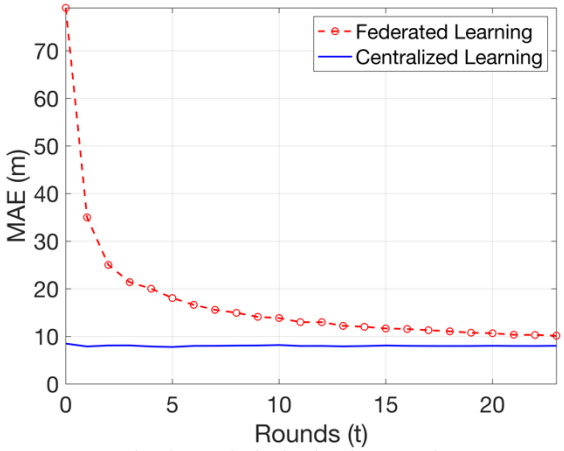

그림 2. round 에 따른 실내 측위 평균 오차

IV. 결 론

본 논문에서는 비중앙집중형의 학습을 통해 개인 정보를 보호하면서 실내 측위가 가능한 연합학습 기반 핑거프린팅 기법 알고리즘을 제안하였다. 서버에서 사용자들의 데이터를 받지 않고 10m 이내의 오차범위를 가지고 실내 측위가 가능함을 얻을 수 있었다. 데이터의 양이 방대해짐에 따라 연합학습을 이용하여 더 빠른 학습을 진행할 수도 있음을 기대할 수 있다. 추후에는 연합학습 기반으로 무선 신호의 다른 측정값을 이용한 실내 측위에 대한 연구도 진행될 예정이다.

ACKNOWLEDGMENT

본 연구는 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터(No.2017-0-00316, 차세대 공공안전통신 원천기술 연구), 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터 육성 지원사업(IITP-2020-2017-0-01637)의 연구결과로 수행되었음.

참 고 문 헌

[1] 김효원, 강규식, 서현덕, 정민수, 최정애, 강정완, 김선우, “5G 및 무인이동체 기술 동향 및 미래 전망,” 한국통신학회지 (정보와통신) 34 권, 7 호, 34-60, 2017.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1-19, Jan. 2019.

[3] T. Li, A.K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50-60, May 2020.

[4] L. Xiao, A. Behboodi, and R. Mathar, “A deep learning approach to fingerprinting indoor localization solutions,” in *Proc. 2017 Int. Telec. Netw. Appl. Conf. (ITNAC)*, Nov. 2017, pp. 22-24.

[5] J. Torres-Sospedra, R. Montoliu, A. Martnez-Us, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, “UJIIndoorLoc: A new multi-buliding and multi-floor database for WLAN fingerprint-based indoor localization problems,” in *Proc. 2014 Int. Conf. Indoor Positioning Indoor Navi. (IPIN)*, Oct. 2014, pp. 261-270.