JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 12, pp. 19-27, December 2021
https://doi.org/10.9708/jksci.2021.26.12.019

Privacy-Preserving Traffic Volume Estimation by Leveraging Local Differential Privacy

Yang-Taek Oh*, Jong Wook Kim*

*Student, Dept. of Computer Science, Sangmyung University, Seoul, Korea
*Professor, Dept. of Computer Science, Sangmyung University, Seoul, Korea

[Abstract]

In this paper, we present a method for effectively predicting traffic volume based on vehicle location data that are collected by using LDP (Local Differential Privacy). The proposed solution in this paper consists of two phases: the process of collecting vehicle location data in a privacy-presering manner and the process of predicting traffic volume using the collected location data. In the first phase, the vehicle's location data is collected by using LDP to prevent privacy issues that may arise during the data collection process. LDP adds random noise to the original data when collecting data to prevent the data owner's sensitive information from being exposed to the outside. This allows the collection of vehicle location data, while preserving the driver's privacy. In the second phase, the traffic volume is predicted by applying deep learning techniques to the data collected in the first stage. Experimental results with real data sets demonstrate that the method proposed in this paper can effectively predict the traffic volume using the location data that are collected in a privacy-preserving manner.

▸ **Key words**: Local difference privacy, Traffic volume estimation, Deep learning, Data Privacy

[요 약]

본 논문에서는 지역 차분 프라이버시(Local Differential Privacy, LDP) 기법을 이용하여 프라이버시를 보호하면서 수집한 차량 위치 데이터와 딥러닝 기법을 이용하여 교통량을 예측하기 위한 기법을 제시한다. 제시한 기법은 데이터를 수집하는 과정과 수집한 데이터를 이용하여 교통량을 예측하는 과정으로 구성된다. 첫 번째 단계에서는 데이터 수집 과정 중에 발생할 수 있는 프라이버시 침해 문제를 해결하기 위해 LDP 기법을 적용하여 차량의 위치 데이터를 수집한다. LDP 기법은 데이터 수집 시 원본 데이터에 노이즈를 추가해 사용자의 민감한 데이터가 외부에 노출되는 것을 방지한다. 이를 통해 운전자의 프라이버시를 보존하면서 차량의 위치 데이터를 수집할 수 있다. 두 번째 단계에서는 첫 번째 단계에서 수집한 데이터에 딥러닝 기법을 적용하여, 교통량을 예측한다. 또한, 본 논문에서 제안한 기법의 우수성을 입증하기 위해, 실데이터를 이용한 성능 평가를 진행한다. 성능 평가 결과는 본 논문에서 제안한 기법이 사용자의 프라이버시를 보호하면서 수집된 데이터를 이용하여 효과적으로 교통량을 예측할 수 있음을 입증한다.

▸ **주제어**: 지역 차분 프라이버시, 교통량 예측, 딥러닝, 데이터 프라이버시

• First Author: Yang-Taek Oh, Corresponding Author: Jong Wook Kim
*Yang-Taek Oh (oyt950418@gmail.com), Dept. of Computer Science, Sangmyung University
*Jong Wook Kim (jkim@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2021. 10. 12, Revised: 2021. 12. 07, Accepted: 2021. 12. 08.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

IoT 기술의 발전으로 인하여 일상생활에서 사용하는 여러 제품들이 인터넷을 통하여 서로 데이터를 주고받는 것이 가능해졌다. 이런 제품들은 컴퓨팅자원을 이용하여 사용자에게 다양한 서비스를 제공한다. '스마트폰'을 시작으로 '스마트 워치', '스마트 티비' 등의 기기들이 등장하고 이러한 개념을 도입해 차량에도 IoT 기술이 적용되는데 이를 IoV(Internet of Vehicles)라고 한다 [1,2]. IoV는 인터넷을 통해 데이터를 교환할 수 있는 차량 네트워크이다. 각각의 차량은 이동하면서 무수히 많은 수의 데이터들을 생성한다. 무선연결을 통해서 네트워크에 연결되어있는 차량들은 실시간으로 교통에 대한 정보를 전송할 수 있으며, 이러한 데이터들을 활용하여 여러 가지 교통과 관련된 문제들을 해결을 기대할 수 있다.

지난 10년 동안 국내 차량 등록 수는 꾸준히 상승하여, 2011년에 약 1800만 대에서 2020년에 약 2400만 대까지 증가하였다 [3]. 차량 수가 증가함에 따라 교통 혼잡 문제도 증가하였고, 이를 해결하기 위한 많은 연구들이 진행되어 왔다. 특히 최근 들어 빅데이터를 활용하여 교통 혼잡도를 예측하여 교통 문제를 해결하기 위한 방식이 많은 관심을 받고 있다. 기존에는 도로에 설치된 인프라를 통하여 차량 위치 데이터를 수집하였다. 그러나 IoV의 등장으로 인하여 각각의 차량으로부터 직접적으로 위치 데이터를 수집하는 것이 가능하게 되었다. 이러한 방식을 통해 실시간으로 수집된 차량 위치 데이터를 이용하여, 교통량을 예측할 수 있으며, 이를 통해 교통 혼잡도를 줄이는 것이 가능하게 되었다.

IoV 발달로 인하여 도로 위의 차량으로부터 대용량의 위치 데이터 수집이 가능하게 되었다. 하지만 차량으로부터 수집된 위치 데이터들은 민감한 정보를 포함하고 있기 때문에, 무분별하게 위치 데이터를 수집하면 프라이버시 침해 문제가 발생할 수 있다. 즉, 차량으로부터 수집된 위치 데이터를 이용하여 운전자의 집 혹은 직장의 주소가 외부에 노출될 수 있다. 또한, 병원 방문과 같은 민감한 정보도 외부에 노출될 수 있다. 이로 인하여, 차량 운전자들은 자신의 차량 위치 데이터를 외부에 제공하는데 반감을 가지고 있다. 그러므로 프라이버시를 보호하면서, 차량으로부터 위치 데이터를 수집하기 위해서는 데이터 수집 과정 중에 차량의 실제 위치 정보가 외부에 노출되지 않게 해야 한다. 최근 들어 사용자로부터 프라이버시를 보호하면서 민감한 정보를 수집하기 위한 방법이 연구되어 왔으며, 그 중 지역 차분 프라이버시(Local Differential Privacy, LDP) 기술이 크게 주목 받고 있다. LDP 기법은 사용자의 민감한 데이터 수집 시, 데이터를 수집하는 과정에서 원본 데이터에 노이즈를 추가하여, 원본 데이터가 외부에 노출되는 것을 방지할 수 있는 방식이다 [4,5,6,7]. 본 논문의 기여는 다음과 같다.

- LDP를 이용하여 운전자의 프라이버시를 보존하면서 차량으로부터 위치 데이터를 수집하기 위한 방식을 제안한다.
- 또한, 프라이버시를 보존하면서 수집된 차량의 위치 데이터를 이용하여, 교통 혼잡도를 예측하기 위한 방법을 개발한다.
- 마지막으로, 실데이터를 이용하여 제안 기법의 우수성을 평가한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 본 논문과 연관된 관련 연구에 대해서 설명하고, 3장에서는 배경지식에 대하여 설명한다. 4장에서 제안 기법을 제시한다. 5장에서는 제안 기법의 성능을 평가한 후, 6장에서 결론을 맺는다.

II. Related Work

2.1 Traffic Volume Estimate

최근 들어 교통량이 증가함에 따라 발생하는 사회 문제가 더욱 심각해지고 있다. 교통량 문제는 동일한 시간에 급격히 한곳으로 몰리는 차량들로 인하여 발생하는 경우가 대부분이다. 이를 미리 교통량 데이터를 활용하여 해결하기 위한 다양한 연구들이 진행되고 있다. *N. K. Jain et al.* [8]는 교통량 문제를 해결하기 위한 모니터링 시스템을 제안하였다. 제안 기법은 이미지 및 비디오 처리 기술과 미래 예측 기법을 이용하여 현재 및 미래 교통량을 모니터링 및 예측한다. 다양한 데이터 예측 기법들이 발달함에 따라 모니터링 시스템을 기반으로 수집한 빅데이터를 이용하여 교통량을 예측하기 위한 다양한 연구들이 진행되었다 [9,10,11].

머신러닝, 딥러닝 기술이 비약적으로 발달함에 따라, 이를 활용하여 교통량을 예측하기 위한 다양한 연구가 진행되었다 [12,13]. 또한, 지능형 교통 시스템(Intelligent Transportation System, ITS)에서 필수적으로 교통량 예측이 요구되므로, 이러한 요구 사항을 해결하기 위해 더욱 정확한 교통량 예측에 대한 연구가 이루어지고 있다 [14].

2.2 Privacy-Preserving Data Collection

최근 들어 우리 일상생활 중에 IoT 기기를 통해 다양한 데이터가 생성되고 있다. 이러한 일상생활 중에 생성되는 데이터들을 수집하여 다양한 분야에서 활용하고 있다. 그러나 무분별한 데이터 수집으로 인하여 개인정보가 침해되는 문제가 발생할 수 있다. 이에 따라, 데이터 수집 시 민감한 개인 정보를 보호하기 위한 많은 연구들이 진행되어 왔다. 가장 대표적인 방법으로 차분 프라이버시 기법(Differential privacy, DP)이 있다. DP는 민감한 개인 정보를 보호하기 위하여 데이터 소유자로부터 수집한 데이터에 노이즈를 추가하여 데이터 사용자에게 배포하는 방식이다 [15].

최근 들어, LDP 기법을 이용하여 민감한 데이터를 수집하기 위한 다양한 연구가 진행되었다 [4,5,6,7,16]. *Moon et al.* [17,18]은 스마트와치 사용자로부터 프라이버시를 보존하면서, LDP를 이용하여 사용자의 건강 데이터(예, 심박수, 누적 걸음수)를 수집하기 위한 방법을 제안하였다. *Lim et al.* [19]는 포그-클라우드 환경에서 사용자로부터 민감한 데이터를 프라이버시를 보호하면서 효과적으로 수집하기 위한 방법을 제안하였다. 또한, LDP 기법보다 데이터 보안성을 더욱 높인 분산 차분 프라이버시(Distributed Differential Privacy, DDP)기법도 프라이버시 보존 데이터 수집을 위해 사용되고 있다. *Lim et al.* [20]는 DDP를 이용하여 민감한 데이터를 안전하게 수집하기 위한 기법을 제안하였다. 최근 연구에서 *Arachchige et al.* [21]는 딥러닝에 사용되는 데이터에 대한 개인정보 유출을 막기 위해 LATENT이라는 새로운 LDP 기반 알고리즘을 제시하였다.

III. Background

3.1 Local Differential Privacy

DP는 프라이버시를 보호하면서 데이터 소유자로부터 민감한 데이터를 수집하기 위한 기법이다. DP는 신뢰할 수 있는 데이터 수집가가 존재한다고 가정한다. 신뢰할 수 있는 데이터 수집가가 데이터 소유자로부터 원본 데이터를 수집한 후, 수집한 데이터에 노이즈를 추가하여 배포하는 방식이다 (그림 1(a)).

그러나 DP는 신뢰할 수 있는 데이터 수집가가 필수적으로 존재해야 한다는 단점이 있다. 이러한 단점을 보완하기 위해 제시된 방식으로 LDP가 있다. LDP는 신뢰할 수 있는 데이터 수집가가 없는 환경을 가정한다. (그림 1(b)). 데이터 소유자가 자신의 원본 데이터에 대해 차분 프라이버시를 만족하도록 직접 변조를 수행하고, 변조된 데이터를 데이터 수집가에게 전송한다. 데이터 수집가는 변조된 데이터를 수집하여 데이터 사용자에게 제공한다.

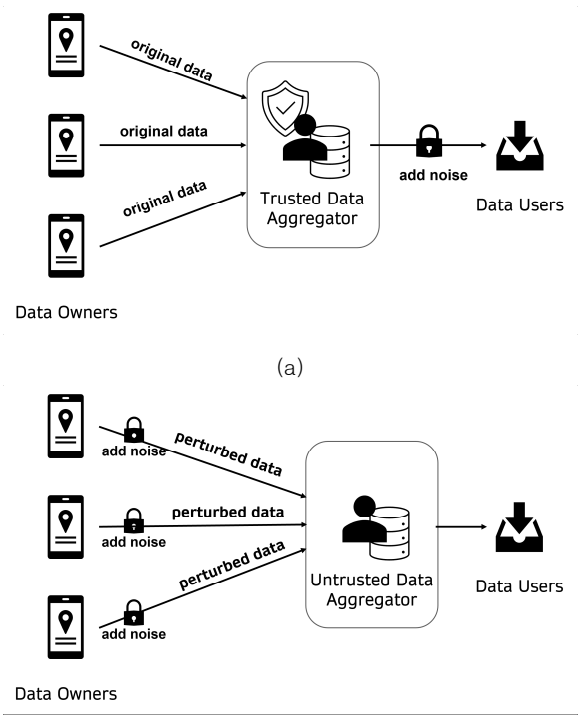


Fig. 1. Data collection using (a) DP and (b) LDP

3.2 Deep Learning

RNN (Recurrent Neural Network)은 입출력을 시퀀스 단위로 처리하는 모델이다 [22]. RNN은 은닉층의 노드에서 활성화 함수를 통해 나온 값을 출력층과 은닉층의 노드의 다음 계산의 입력으로 보내는 특징을 가지고 있다 (그림 2). 이런 특징이 의미하는 바는 현재 시점의 활성화 함수의 결과값이 이전 시점의 결과 값의 영향을 받는다는 것이다.

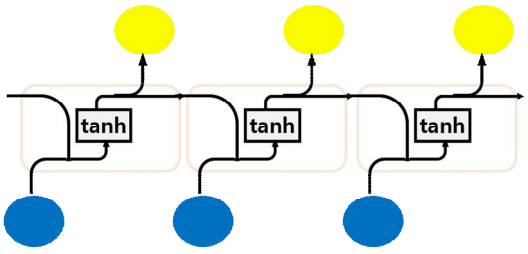


Fig. 2. RNN Structure

앞에서 설명한 RNN은 시계열 데이터를 처리하기에 효과적인 방법이다. 하지만 시계열 데이터의 시점이 길어질수록 앞선 데이터의 값이 잘 전달되지 않는 현상이 발생한

다. 이러한 현상을 “긴 기간의 의존성(long-term dependencies)”라고 하는데 이를 해결하기 위하여 RNN의 한 종류인 LSTM이 제시되었다 [23][24]. LSTM은 기존의 RNN의 셀을 변형시켜 입력 게이트, 망각 게이트, 출력 게이트를 포함하는 셀 상태를 추가했다 (그림 3). 이로 인해 비교적 긴 시퀀스의 입력을 처리하는데 더 효과적인 성능을 보인다.

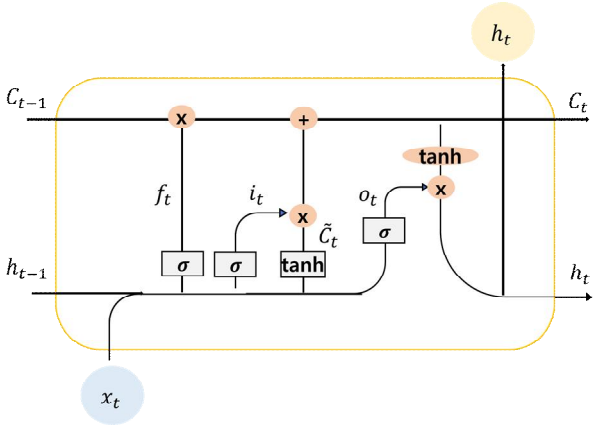


Fig. 3. LSTM Structure

기본 LSTM 네트워크 아키텍처에서 (x_1,x_2, …, x_t)로 표시되는 입력 시퀀스가 주어지면 출력 시퀀스 y_t는 식 1을 반복적으로 계산하여 얻을 수 있다.

$$h_t = LSTM(h_{t-1}, x_t; W)$$
$$y_t = W_{hy} h_t + b_y \quad \text{(식)1}$$

이때, W 항은 다른 가중치 행렬을 나타내며 b_y는 출력 y_t에 대한 편향 벡터를 나타내고 h는 은닉 상태를 나타낸다. 셀 함수 LSTM (·)에서 은닉 상태는 식 2를 통해 입력 게이트 i, 망각 게이트 f, 출력 게이트 o 및 셀 상태 c에 의해 결정된다.

$$f_t = \sigma(W_{xh_f} x_t + W_{hh_f} h_{t-1} + b_{h_f})$$
$$i_t = \sigma(W_{xh_i} x_t + W_{hh_i} h_{t-1} + b_{h_i})$$
$$o_t = \sigma(W_{xh_o} x_t + W_{hh_o} h_{t-1} + b_{h_o})$$
$$\widetilde{C_t} = \tanh(W_{xh_C} x_t + W_{hh_C} h_{t-1} + b_{h_C}) \quad \text{(식)2}$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C_t}$$
$$h_t = o_t \circ \tanh(C_t)$$

이를 통해 예측된 결과 값을 실제 결과 값과 유사도를 측정하기 위해 두 값을 손실함수를 사용해서 비교한다. 이 과정을 통해 손실 함수 값을 줄이는 방향으로 가중치를 갱신하는 방식으로 학습을 진행한다.

IV. The Proposed Scheme

본 장에서는 본 논문의 제안 기법을 설명한다. 그림 4는 제안 기법의 구성도에 해당한다.

- LDP 기법을 이용하여 프라이버시를 보존하면서 차량의 위치 데이터를 수집한다.
- 프라이버시를 보존하면서 수집한 차량 위치 데이터를 사용하여 교통량을 예측할 수 있는 딥러닝 모델을 학습한다.
- 학습한 딥러닝 모델을 이용하여 실시간으로 교통량을 예측한다.

4.1 Data collection

차량의 위치 데이터를 수집하고자 하는 지역이 m개의 구역으로 나누어 있다고 가정하자. 또한, 차량의 현재 위치가 k번째 구역에 해당한다고 가정하자. 이때, 차량의 위치는 k번째 요소가 1이고, 나머지 요소는 0인 m-차원 벡터 v로 표현할 수 있다 (즉, v = [d_1, … ,d_k, …, d_m] = [0, …, 1, …, 0]).

m-차원 벡터 v를 데이터 수집가에게 전송하면, 차량 운전자의 위치가 외부에 노출되어 프라이버시 침해 문제가 발생한다. 그러므로 차량 운전자의 프라이버시를 보호하기 위해 LDP의 데이터 변조 기법을 m-차원 벡터 v에 적용하여, 변조된 m-차원 벡터 z를 식 3을 이용하여 구한다 [5,6].

$$\Pr[z[i]=1] \begin{cases} p = \frac{1}{2} & \text{if } v[i]=1 \\ q = \frac{1}{e^{\epsilon}+1} & \text{if } v[i]=1 \end{cases} \quad \text{(식)3}$$

식 3에 의해서, 원본 벡터 v의 i번째 요소가 1인 경우, 변조된 벡터 z의 i번째 요소가 1일 확률은 0.5에 해당한다, 반면, 원본 벡터 v의 i번째 요소가 0인 경우, 변조된 벡터 z의 i번째 요소가 1일 확률은 $q = \frac{1}{e^{\epsilon}+1}$ 에 해당한다.

벡터 v의 각각의 요소들이 확률적으로 0 혹은 1로 변환하도록 데이터 변조가 발생하므로, 민감한 원본 위치 정보가 외부에 노출되는 것을 방지할 수 있다. 데이터 변조 후, 각각의 차량 운전자는 변조된 m-차원 벡터 z를 데이터 수집 서버에 전송한다.

4.2 Data aggregation

데이터 수집 단계에서는 차량 운전자로부터 전송 받은 변조된 벡터들을 이용하여, 특정 시간 동안 (예, 30분 간격) 각 지역별 차량 수를 구한다. 특정 시간 동안 n개의 변

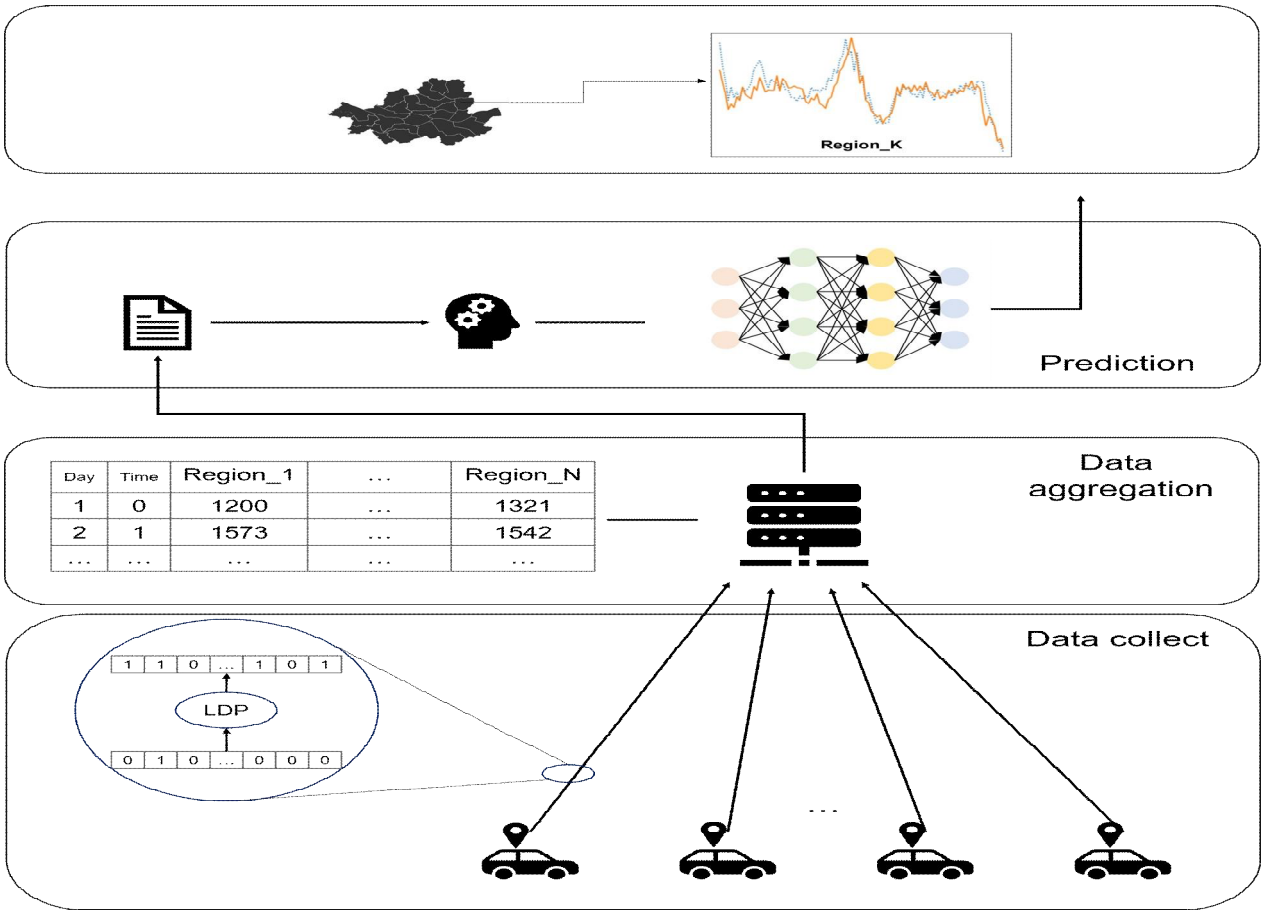


Fig. 4. The architecture of the proposed approach

조된 벡터를 수집했다고 가정하자. 이때, 각각의 변조된 m-차원 벡터를 z_1, z_2, ..., z_n으로 표현하자. 또한, $\bar{z}$를 n개 벡터들의 합이라 가정하자 (식 4).

$$\bar{z} = \sum_{i=1}^{n} z_i \quad \text{(식)4}$$

식 3을 참고해 LDP의 데이터 변조가 적용되는 과정을 역으로 적용하여, k번째 구역에 위치한 차량 수 x_k는 다음 식 5와 같이 예측할 수 있다 [5,6].

$$x_k = \frac{\bar{z}[k] - nq}{p - q} \quad \text{(식)5}$$

4.3 Prediction

이 절에서는 LDP 기법을 이용하여 프라이버시를 보존하면서 수집한 데이터들을 이용하여 각 지역별 차량 수를 예측하기 위한 딥러닝 모델을 만든다.

LDP 기법을 이용하여 수집한 데이터를 이용하기 위해서는 먼저 데이터에 대한 전처리 과정이 필요하다. 수집한 데이터들의 값의 범위는 다양하다. 그러므로 원본 값을 그대로 사용하여 딥러닝 모델을 훈련하게 되면, 예측 모델의 성능 저하가 발생할 수 있다. 그러므로 데이터들의 값의 범위가 0에서 1사이의 수로 표현되도록, 정규화 과정을 적용한다. 정규화하고자 하는 값들의 집합을 X라고 가정하고, 이중 최솟값과 최댓값을 각각 x_{max}, x_{min}이라고 가정하자. 이때, 0에서 1사이의 값들로 정규화된 값 x'는 다음 식 6을 이용하여 구할 수 있다.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad \text{(식)6}$$

정규화 과정을 거친 후 예측모델을 생성하기 위해서 딥러닝 기법중 하나인 LSTM을 사용한다. LSTM을 적용하기 위해서는 입력 데이터가 시퀀스 형태로 표현되어야 한다. x_k^t를 특정 현재 시점 t에서 k번째 구역에 위치한 차량의 수라 가정하자. x_k^t는 4.1절과 4.2절에서 설명하였듯이, LDP 기법을 이용하여 프라이버시를 보존하면서 차량 위치 데이터를 수집하여 예측한 값이다. 본 연구에서는 그림 5와 같이 현재 시점 t에서 과거 t개의 데이터 x_k^1, x_k^2, ..., x_k^{t-1}, x_k^t를 이용하여 미래 시점 (t+r)까지의 차량 수(즉, x_k^{t+1}, x_k^{t+2}, ..., x_k^{t+r-1}, x_k^{t+r})를 예측한다. 즉, 본 논문에서

는 t개의 입력으로 r개의 출력을 예측하는 LSTM 모델을 훈련 시킨다.

그림 5에서 예측되는 r개의 출력 x_k^{t+1}, x_k^{t+2}, ..., x_k^{t+r}를 y_1, y_2,...,y_r라 가정하자(즉 $x_k^{t+1}=y_1$, $x_k^{t+2}=y_2$, ..., $x_k^{t+r}=y_r$). 이때, LSTM 모델을 훈련시키기 위하여, 손실함수는 평균 제곱 오차 (Mean Squared Error, MSE)를 사용한다. MSE는 다음 식 7과 같이 실제값 $Y=[y_1, y_2, ..., y_r]$와 예측값 $\overline{Y}=[\overline{y}_1, \overline{y}_2, ..., \overline{y}_r]$의 차를 제곱하여 합한 값이다. 이 값을 줄이는 방향으로 최적화 함수가 적용된다.

$$MSE = \frac{1}{r}\sum_{i=1}^{r}(y_i - \overline{y}_i)^2 \qquad \text{(식)7}$$

위와 같은 방식을 사용하여 m개의 구역에 대하여 특정 현재 시점 t에서 미래 시점 (t+r)까지의 차량 수를 예측함으로써 전체 구역에 대한 교통량을 예측할 수 있다.

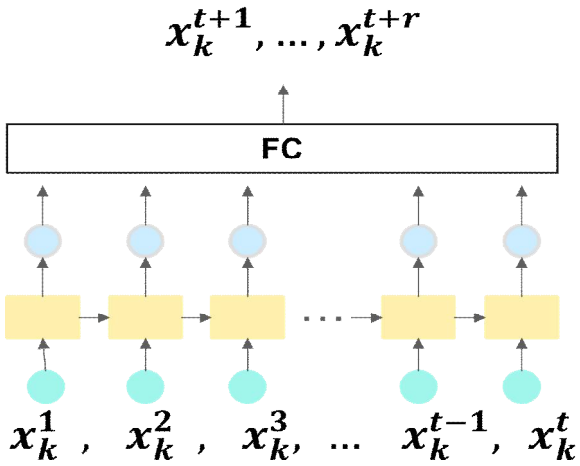


Fig. 5. LSTM process

V. Experiment

5.1 Experiment Setup

본 논문의 실험은 구글에서 제공하는 Colaboratory를 활용하여 구글 클라우드 환경에서 진행하였다. CPU는 인텔 제온 2.30GHz, GPU는 Nvidia K80, Ubuntu 18.04 환경에서 실험을 진행하였다. 실험에 사용된 데이터는 서울특별시 빅데이터 캠퍼스에서 배포한 택시 데이터(서울시 주요도로 택시운행 분석데이터 정보)[25]를 사용하여 실험을 진행하였다. 택시 데이터는 7개의 열(링크ID, 요일, 시간, 날씨, 목적지, 승차건수, 하차건수, 공차건수)로 구성되어 있다. 7개의 열로 구성된 데이터에서 요일(월~금요일을 1~7로 표현), 시간(24시간을 30분 단위로 나눈 값들로 0~47로 표현), 목적지(각 행정구역 코드)에 해당하는 데이터를 추출하였다. 목적지의 값이 의미하는 서울시의 25개 행정구역을 요일, 시간에 따라 구분하여 해당 요일 해당 시간에 각 행정구역의 택시 수를 표현하도록 데이터를 구성하였다. 실험에 사용한 데이터는 2015년 1월부터 12월까지의 총 1년간 수집한 데이터를 사용하였다. 전체 데이터 중 70%는 훈련 데이터로, 20%는 검증 데이터로, 10%는 테스트 데이터로 나누어 사용하였다.

실험에서 사용한 LDP의 프라이버시 예산 ε 값은 0.5, 1.0, 2.0이다. 예측모델은 Tensorflow 2.6 라이브러리를 사용하여 구현하였다 [26]. LSTM 모델 학습에 사용한 배치 크기는 64이며, 에포크는 200에 해당한다. 또한, 4.3절에서 설명하였듯이 MSE를 사용하였고, 최적화 함수로는 ADAM을 사용하였다 [27].

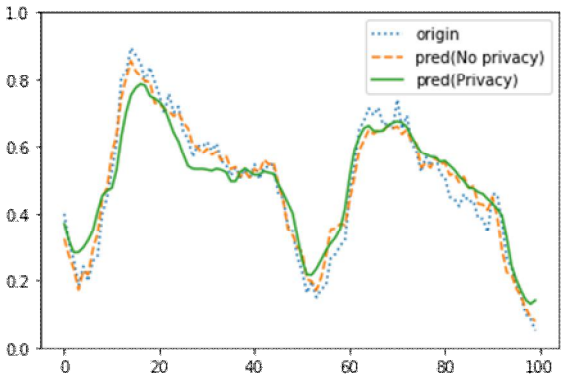
5.2 Results

그림 6은 4.3에서 언급한 t와 r 값을 변화시켰을 때, Loss, MAE, RMSE 값을 나타낸다 [28][29]. 이 실험에서 프라이버시 예산 ε값은 1로 설정하였으며, 실험에서 사용한 t와 r의 값은 각각 t=12/24/48, r = 1/2/4/8이다. 가령, t=12, r=8일 경우, 4.1절과 4.2절에서 설명하였듯이, LDP를 이용하여 프라이버시를 보존하면서 수집한 과거 6시간의 교통량 데이터를 이용하여(30분 단위로 수집), 미래 4시간의 교통량을 30분 단위로 예측하는 것이다.

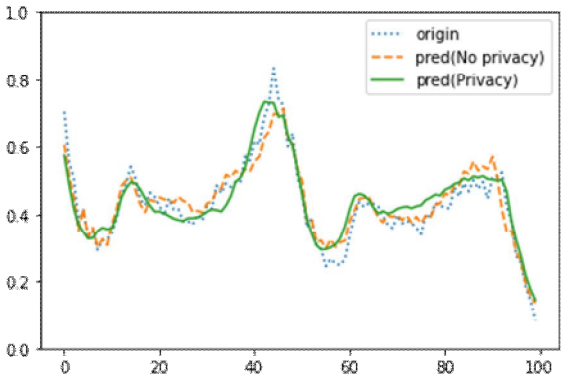
그림 6의 결과 값들을 보면 r의 값이 감소할수록, Loss, MAE, RMSE 값이 감소하여, 예측 모델의 성능이 증가함을 보인다. 이는 r의 값이 감소할수록, 현재 시점과 상대적으로 가까운 값들만을 예측하므로 예측 모델의 성능이 증가하기 때문이다.

t/r	Loss	MAE	RMSE
12/1	0.0075	0.0683	0.0864
24/1	0.0077	0.0692	0.0877
48/1	0.0078	0.0696	0.0881
12/2	0.0078	0.0696	0.0883
24/2	0.0075	0.0685	0.0867
48/2	0.0079	0.0701	0.0891
12/4	0.0077	0.0694	0.0879
24/4	0.0076	0.0687	0.0890
48/4	0.0085	0.0724	0.0921
12/8	0.0081	0.0712	0.0902
24/8	0.0079	0.0703	0.0892
48/8	0.0079	0.0701	0.0890

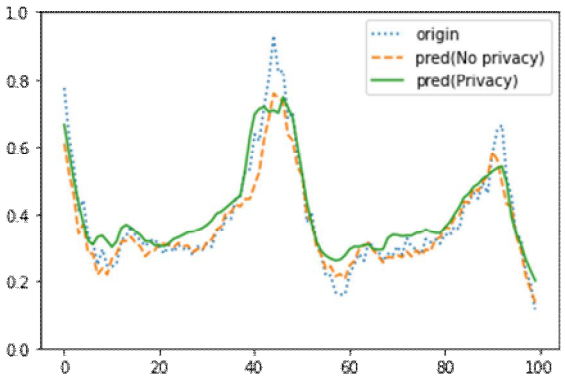
Fig. 6. Loss, MAE, RMSE for varying t and r



(a) region code= 0



(b) region code= 10



(c) region code= 20

Fig. 7. Actual vs. Estimated normalized number of vehicles

그림 7은 25개의 행정 구역 중에서 행정 코드가 0, 10, 20에 해당하는 구역의 시간대별 교통량 예측 결과를 보여준다. 이 실험에서 t=24, r=2로 설정하였으며, 프라이버시 예산 ε값은 1로 설정하였다. 각 그래프에서 성능 비교 목적으로 다음과 같은 3가지 경우의 값을 표현하였다.

- Origin은 실제 교통량을 나타낸다.
- pred(No privacy)는 프라이버시가 보존되지 않은 데이터를 사용하여 예측한 교통량을 나타낸다. 즉. 이 경우 차량의 원본 위치 데이터가 데이터 수집 서버에 전송된다.
- pred(Privacy)는 본 논문에서 제안한 기법에 해당한다. 즉, 프라이버시를 보존하면서 수집한 데이터를 이용하여 교통량을 예측한 방법이다.

그림 7의 결과에서 알 수 있듯이, 프라이버시를 보존하지 않고 수집한 데이터를 이용하여 예측한 값이 실제 값과 더 유사함을 보인다. 그러나 본 연구에서 제안한 기법으로 예측한 값도 이와 매우 유사한 예측 성능을 보이는 것을 알 수 있다. 따라서 실험의 결과에서 알 수 있듯이, 본 연구에서 제안하는 기법이 차량 운전자의 프라이버시를 보존하면서도 실제 교통량과 매우 유사하게 교통량을 예측할 수 있다.

그림 8은 프라이버시 예산 ε값이 예측 결과에 미치는 영향을 보여준다. 실험에서 t=24, r=2로 설정하였으며, ε값을 0.5, 1, 2로 변경하였다. 그림에서 알 수 있듯이, ε값이 감소할수록 예측 성능이 저하된다. 이는 ε값이 감소할수록 사용자의 프라이버시 보호 수준이 증가하여, 원본 데이터에 대한 심한 변조가 발생하기 때문이다. 반면에 ε값이 증가할수록 예측 값이 실제 값과 유사해짐을 알 수 있다. 이는 ε값이 증가할수록 사용자의 프라이버시 보호 수준이 감소하여, 원본 데이터에 대한 변조가 작게 발생하기 때문이다. 그림 8의 실험 결과를 통하여 본 논문에서 제안하는 기법은 사용자의 프라이버시 보호 수준에 맞추어 예측 결과의 정확도를 조절할 수 있음을 보인다. 본 장의 실험 결과는 본 논문에서 제안한 기법이 사용자의 프라이버시를 보호하면서 수집된 데이터를 이용하여 효과적으로 교통량을 예측할 수 있음을 입증한다.

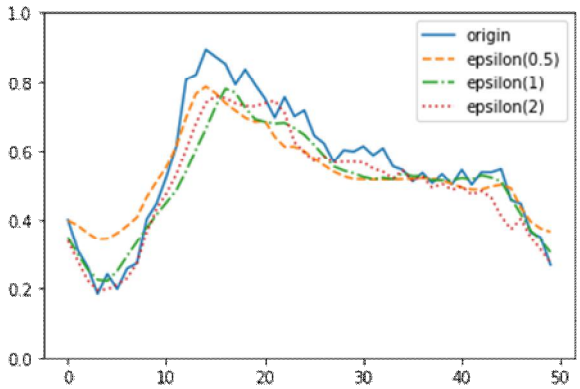


Fig. 8. Actual vs. Estimated normalized number of vehicles for varying privacy budget ε

VI. Conclusions

본 논문에서는 LDP 기법을 이용하여 수집한 교통량 데이터와 LSTM을 활용하여 교통량을 예측하기 위한 기법을 제안하였다. 특히, 본 연구에서 사용한 데이터 수집 기법을 통하여, 차량 위치 데이터에 대한 프라이버시를 보존할 수 있다. 또한 실데이터를 이용한 실험 결과를 통하여, 본 논문에서 제안한 기법이 사용자의 프라이버시를 보호하면서 수집된 데이터를 이용하여 효과적으로 교통량을 예측할 수 있음을 입증하였다. 본 연구를 통해 개인 프라이버시를 보호하면서 수집한 차량데이터로 교통량을 예측하여 교통량 문제를 해결할 수 있을 것으로 기대된다.

REFERENCES

[1] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," IEEE internet of things journal, Vol. 1, No. 4, pp. 289-299, Aug. 2014.

[2] F. Yang, S. Wang, J. Li, Z. Liu, and Q. Sun, "An overview of internet of vehicles," China communications, Vol. 11, No. 10, pp.1-15, Oct. 2014.

[3] Ministry of Land, Infrastructure and Transport, Automobile Operation Insurance Division, https://index.go.kr

[4] J. W. Kim, K Edemacu, J. S. Kim, Y. D. Chung, B. Jang "A Survey of Differential Privacy-based Techniques and their Applicability to Location-Based Services." Computers and Security Sep. 2021.

[5] J. W. Kim, S. M. Moon, S. Kang, B. Jang "Effective privacy-preserving collection of health data from a user's wearable device." Applied Sciences Oct. 2020.

[6] J. W. Kim, and B. Jang. "Workload-aware indoor positioning data collection via local differential privacy," IEEE Communications Letters, Vol. 23, No. 8, pp. 1352-1356, Aug. 2019.

[7] J. W. Kim, J. H. Lim, S. M. Moon, and B. Jang, "Collecting health lifelog data from smartwatch users in a privacy-preserving manner," IEEE Transactions on Consumer Electronics, Vol. 65, No. 3, pp. 369-378, Aug. 2019.

[8] N. K. Jain, R. K. Saini, and P. Mittal, "A review on traffic monitoring system techniques," Soft Computing: Theories and Applications. Springer, Singapore, pp.569-577. Aug. 2019.

[9] H. Feng, and Y. Shu, "Study on network traffic prediction techniques," Proceedings. 2005 International Conference on Wireless Communications, Networking and Mobile Computing, 2005., IEEE, Vol. 2, pp. 1041-1044, Dec. 2005.

[10] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," Proceedings of the AAAI conference on artificial intelligence, Vol. 33, No. 01, pp. 5668-5675, Jul. 2019.

[11] A. M. Nagy, and V. Simon, "Survey on traffic prediction in smart cities," Pervasive and Mobile Computing, Vol. 50, pp. 148-163, Oct. 2018

[12] J. Z. Zhu, J. Z. Cao, and Y. Zhu, "Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections," Transportation Research Part C: Emerging Technologies, Vol. 47, pp. 139-154, Oct. 2014.

[13] M. Moniruzzaman, H. Maoh, and W. Anderson, "Short-term prediction of border crossing time and traffic volume for commercial trucks: A case study for the Ambassador Bridge," Transportation Research Part C: Emerging Technologies, Vol.63, pp. 182-194, Feb. 2016.

[14] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "A comprehensive survey on traffic prediction,." arXiv preprint arXiv:2004.08555 Apr. 2020.

[15] C. Dwork, "Differential privacy," International Colloquium on Automata, Languages, and Programming, Springer, Berlin, Heidelberg, pp. 1-12 Jul. 2006.

[16] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin, "Collecting and analyzing data from smart device users with local differential privacy," arXiv preprint arXiv:1606.05053 Jul. 2016.

[17] S. M. Moon, and J. W. Kim, "Privacy-Preserving Method to Collect Health Data from Smartband," Journal of the Korea Society of Computer and Information, Vol. 25, No. 4, pp. 113-121, Apr. 2020.

[18] S. M. Moon, and J. W. Kim, "Collecting Health Data from Wearable Devices by Leveraging Salient Features in a Privacy-Preserving Manner," Journal of the Korea Society of Computer and Information, Vol. 25, No. 10, pp. 59-67, Oct. 2020.

[19] J. H. Lim, and J. W. Kim, "Privacy-Preserving IoT Data Collection in Fog-Cloud Computing Environment," Journal of the Korea Society of Computer and Information Vol. 24, No. 9, pp. 43-49, Sep. 2019.

[20] J. H. Lim, and J. W. Kim. "Privacy-Preserving Aggregation of IoT Data with Distributed Differential Privacy," Journal of the Korea Society of Computer and Information Vol. 25, No. 6, pp. 65-72, June. 2020.

[21] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "Local differential privacy for deep learning," IEEE Internet of Things Journal, Vol. 7, No.7, pp. 5827-5842, Nov. 2019.

[22] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," Interspeech. Vol. 2, No. 3, pp. 1045-1048, Jul. 2010.

[23] S. Hochreiter, and J. Schmidhuber, "Long short-term memory,"

Neural computation, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997.
[24] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." Sep. 2014.
[25] Seoul Metropolitan Government.Big Data Campus, https://bigdata.seoul.go.kr/
[26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, et al. "Tensorflow: A system for large-scale machine learning." 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16). Nov. 2016.
[27] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, Dec. 2014.
[28] C. J. Willmott, and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.", Climate research Vol. 30, No. 1, pp. 79-82, Dec. 2005.
[29] T. Chai, and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," Geoscientific model development, Vol. 7, No. 3, pp. 1247-1250, Jun. 2014.

Authors

Yang-Taek Oh received the B.S. degree from Sangmyung University in 2020, where he is currently pursuing the master's degree with the Department of Computer Science. His research mainly focuses on data privacy and Artificial Intelligence.

Jong Wook Kim received the Ph.D. degree in Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013.

Dr. Kim is currently an Associate Professor with the Department of Computer Science at Sangmyung University. His primary research interests include the area of data privacy, distributed databases, and query optimization.