Missing Data Completion for Telco Localization via Adversarial Learning

Jinhua Lv
Dept. of Software Engineering
Tongji University
Shanghai, China
jhlv@tongji.edu.cn

Yige Zhang
Dept. of Software Engineering
Tongji University
Shanghai, China
yigezhang@tongji.edu.cn

Weixiong Rao
Dept. of Software Engineering
Tongji University
Shanghai, China
wxrao@tongji.edu.cn

***Abstract*—This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX. This document is a model and instructions for LATEX.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Recent years witnessed the rapid spreading of cellular networks and pervasive mobile devices (MDs). The telecommunication (Telco) data, as trace of MDs in cellular networks, has many important applications for Telco operators, e.g., city-scale Telco localization [1], churn prediction of subscribers [2] and user experience assessment [3]. In particular, as an important complementary technique of GPS, Telco localization is aimed at recovering mobility patterns of MDs at fine grained level (e.g., 20 meters). Unlike the call detail records (CDRs), Telco localization techniques mainly focus on measurement records (MRs), which measures radio signal strengths (RSSIs) between MD and its connected cells in Telco networks. In the past years, a plenty of Telco localization methods have been proposed to improve the performance under challenging city environment [4]–[8].

However, these localization models are suffering from missing signal strength (RSSI) values with varying degrees. Zhu [1] found that nearly 50% of real world MR records have RSSI with only 1 or 2 cells. Ray [7] proposed a localization model based on that RSSI values from neighboring cells are all missing. In the worst case, there is no RSSI at all, left only the serving cells in MR records [9]. The missing data problem significantly deteriorates the performance of Telco localization. There are two main reasons of missing values in MR records. One is that the mobile phones do not provide API to access neighboring cells. The other is that the RSSI is lost, due to communication failure or data corruption. Consequently, it is desired to develop a methodology with high completion performance to estimate the missing data.

The most frequently used methods for data completion are interpolation, statistics and nearest neighbors [10]. These methods simply fill missing values by part of the data set, do not generate high quality complete data. . The recent proposed algorithms built on deep learning (e.g. denoising autoencoders (DAE)) [11] learn the conditional probability from complete parts to missing parts, based on the whole data set. Given existing data completion methods, there are two major challenges in applying them directly for MR data.

(1) The existing methods cannot capture the correlation of cell locations and associated RSSI, which is most important in completing missing RSSI.

(2) Due to complex RF propagation phenomena (e.g., multi-path and non-line-of-sight), the signal strengths often fluctuate in a wide range. Noise caused by such fluctuation can hurt the inference of missing RSSI.

To this end, we propose TelcoGAN that generalizes Generative Adversarial Nets [12] with to generate complete data set.

Besides, there is a major difference in our problem. The missing MR data to complete is collected without location labels (e.g., GPS coordinates) due to GPS issues [13]. However, by retrieving locations from location-based services, it is available in historical collected MR data [14]. How to exploit these labels to help the generation of missing RSSI remains challenging.

II. Preliminaries

In this section, we first give a detailed description of MR data and some basic notations, then followed by the problem definition and overview of GANs [12] with its variants.

TABLE I
AN EXAMPLE OF 4G LTE MR RECORD

Field	Value	Field	Value
MRTime	**2017/5/31 14:12:06**	IMSI	*****012**
Serving_eNodeBID	**99129**	Serving_CellID	**1**
eNodeBID_1	**99129**	CellID_1	**1**
RSRP_1	**-93.26**	RSSI_1	**-67.18**
...	...	...	...
eNodeBID_6	**99145**	CellID_6	**5**
RSRP_6	**-90.02**	RSSI_6	**-50.92**

A. Data Description and Notations

Telco localization techniques mostly focus on MR data, which are generated when MDs connect to nearby cell towers in Telco networks. Generally, the MR data can be categorized into two aspects: (1) connected cells data including cell ids, GPS locations; (2) continuous signal strength data, such as RSRP, RSSI. Table I shows an example of MR record in 4G LTE network. A piece of MR record contains up to 6 nearby cells (eNodeBID and CellID) and radio signal strength indicators (RSSI) for each. Besides, it also marks a user ID (IMSI: International Mobile Subscriber identification Number) and connection time stamp (MRTime).

For the rest of this paper, a MR record and associated GPS label are denoted by m and l respectively. For a MR record vector m, it consists of cell id vector c, cell coordinate vector d and RSSI vector r with equal length as 6. Note that to protect privacy of involved users, we delete IMSI to reduce risk.

B. Problem Definition

The MR missing data problem refers to missing of RSSI from neighboring cells. We assume that RSSI from any neighboring cell can be missing randomly (i.e., from r_2 to r_6). In addition, the binary vector $b = \{b_i\}_{i=1}^{6}$ taken values in $\{0, 1\}$ indicates which part of RSSI vector r could be lost. Thus, an incomplete RSSI vector $\hat{r}$ can be defined as follow:

$$\hat{r}_i = \begin{cases} r_i, & \text{if } b_i = 1 \\ nan, & \text{otherwise} \end{cases} \quad (1)$$

Note that we can recover binary vector b from $\hat{r}$.

The detail of the problem studied in this paper is described in Problem 1.

Problem 1: **[Telco Missing Data Completion]**: For an incomplete MR record m with RSSI vector $\hat{r}$, we aim at filling the missingness of $\hat{r}$ and generating r, given the cell id vector c and cell coordinate vector d.

C. Basic of GANs

The key of Generative Adversarial Nets (GANs) [12] is to play a competing game between two networks. The generator network G takes noise vector z as input and generates fake data. The discriminator network D takes a data sample (real/generated) as input and try to classify the sample accurately. In contrast, the generator G try to generate realistic data to fool discriminator D. Hence, the two networks G and D play a minimax game which can be formulated as:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \quad (2)$$

where $p(\mathbf{x})$ denotes real data distribution, $p(\mathbf{z})$ is noise distribution such as the uniform distribution or the normal distribution. The two networks are trained iteratively towards the optimization of objective function 2.

However, the unstable training process makes GANs hard to train. It has been shown in previous work [15] that the label information can help stabilize training, leading to improved quality of generated data. The auxiliary classification procedure is employed to help discriminator distinguish input samples from different label categories. A most recent work on multi-modality data completion employs the auxiliary label information and verifies the improvement [16]. The proposed classification loss is described as follows:

$$L_{CLS}(x, y, l) = L_{CE}(D(x, y), \ell) \quad (3)$$

where x, y and l denote observed variables, missing variables and associated label respectively. The discriminator is trained to minimize the classification loss of data samples, and meanwhile distinguish data samples (real/generated).

Inspired by the above methodologies, we propose our TelcoGAN model by adopting a deep localization model in a unified generative adversarial network, utilizing the available GPS label information in training data.

III. TELCOGAN MODEL

In this section, we introduce our TelcoGAN model for Telco missing data completion. The descriptions are separated into four parts. First, we describe the design of TelcoGAN model as well as the interaction of different components. Second, we give the details of relative coordinate space and explore the real-world relationship between connected cells and RSSI. Next, the designs of generator, discriminator and localizer is introduced respectively. Finally, we give TelcoGAN's training process.

A. Overview of TelcoGAN

The TelcoGAN model consists of pre-processing step and three basic components. The framework of TelcoGAN is illustrated as Fig. 1.

In the pre-processing step, due to sparse extensive cell locations in MR data, we propose to apply a relative coordinate space. The sparse global coordinate-based distribution of MR data is transformed into a dense relative coordinate-based distribution. Thus the TelcoGAN model can better capture the internal correlation of cell locations and corresponding RSSI.

The adversarial learning step consists of three interacting components described as follows:

(1) **Generator $\bar{\mathbf{m}} \sim$ G($\mathbf{z}$,$\hat{\mathbf{m}}$, $\mathbf{b}$)**: The generator takes a random vector, an incomplete MR matrix and corresponding binary vector as input, and generates a completed MR matrix $\bar{\mathbf{m}}$ that fools the discriminator as well as makes localizer produce accurate predictions.

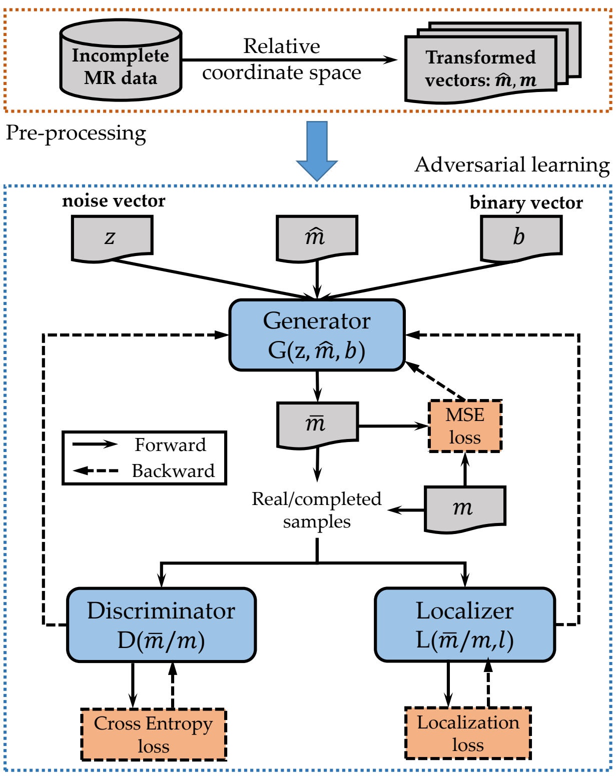

Fig. 1. The framework of TelcoGAN

(2) **Discriminator** $D(\bar{\mathbf{m}}/\mathbf{m})$: The discriminator takes either a generated MR sample or a real MR sample as input, and gives each sample the probability over two categories (real/generated).

(3) **Localizer** $L(\bar{\mathbf{m}}/\mathbf{m}, \mathbf{l})$: The localizer takes a pair of MR sample and corresponding location label as input. It tries to predict the position of MR sample and minimize the localization loss.

The intuition of how our TelcoGAN model can generate high quality MR data for Telco localization is as follows. The generator tries to recover complete MR data samples based on observed variables to fool the discriminator; The discriminator distinguishes input data samples and computes probability distribution that the samples comes from real data or generated data; The localizer predicts locations of MR samples and produces a score for each sample that reflects its quality. During the adversarial game between the generator and the discriminator, the localizer can guide the optimization towards better data quality by utilizing available location labels. When the training reaches the optimality, the generator will have learnt the mapping from incomplete MR data to complete MR data.

We choose to implement these three components as neural networks. We will discuss their detailed structures for generating complete MR data in the following subsections.

B. Relative Coordinate Space

C. TelcoGAN Generator

D. TelcoGAN Discriminator

E. TelcoGAN Localizer

F. TelcoGAN Training

REFERENCES

[1] F. Zhu, C. Luo, M. Yuan, Y. Zhu, Z. Zhang, T. Gu, K. Deng, W. Rao, and J. Zeng, "City-scale localization with telco big data," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, 2016, pp. 439–448.

[2] Y. Huang, F. Zhu, M. Yuan, K. Deng, Y. Li, B. Ni, W. Dai, Q. Yang, and J. Zeng, "Telco churn prediction with big data," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, 2015, pp. 607–618.

[3] C. Luo, J. Zeng, M. Yuan, W. Dai, and Q. Yang, "Telco user activity level prediction with massive mobile broadband data," *ACM TIST*, vol. 7, no. 4, pp. 63:1–63:30, 2016.

[4] M. Ibrahim and M. Youssef, "A hidden markov model for localization using low-end GSM cell phones," in *Proceedings of IEEE International Conference on Communications, ICC 2011, Kyoto, Japan, 5-9 June, 2011*, 2011, pp. 1–5.

[5] S. Hara, D. Anzai, T. Yabu, T. Derham, and R. Zemek, "Analysis on TOA and TDOA location estimation performances in a cellular system," in *Proceedings of IEEE International Conference on Communications, ICC 2011, Kyoto, Japan, 5-9 June, 2011*, 2011, pp. 1–5.

[6] M. Ibrahim and M. Youssef, "Cellsense: An accurate energy-efficient GSM positioning system," *IEEE Trans. Vehicular Technology*, vol. 61, no. 1, pp. 286–296, 2012.

[7] A. Ray, S. Deb, and P. Monogioudis, "Localization of LTE measurement records with missing information," in *35th Annual IEEE International Conference on Computer Communications, INFOCOM 2016, San Francisco, CA, USA, April 10-14, 2016*, 2016, pp. 1–9.

[8] R. Margolies, R. A. Becker, S. D. Byers, S. Deb, R. Jana, S. Urbanek, and C. Volinsky, "Can you find me now? evaluation of network-based localization in a 4g LTE network," in *2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1-4, 2017*, 2017, pp. 1–9.

[9] K. Perera, T. Bhattacharya, L. Kulik, and J. Bailey, "Trajectory inference for mobile devices using connected cell towers," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, November 3-6, 2015*, 2015, pp. 23:1–23:10.

[10] O. G. Troyanskaya, M. N. Cantor, G. Sherlock, P. O. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001. [Online]. Available: https://doi.org/10.1093/bioinformatics/17.6.520

[11] L. Gondara and K. Wang, "MIDA: multiple imputation using denoising autoencoders," in *Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III*, 2018, pp. 260–272.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 2672–2680.

[13] Y. Zhang, W. Rao, M. Yuan, J. Zeng, and H. Yang, "Confidence model-based data repair for telco localization," in *18th IEEE International Conference on Mobile Data Management, MDM 2017, Daejeon, South Korea, May 29 - June 1, 2017*, 2017, pp. 186–195.

[14] X. Huang, Y. Li, Y. Wang, X. Chen, Y. Xiao, and L. Zhang, "CTS: A cellular-based trajectory tracking system with gps-level accuracy," *IMWUT*, vol. 1, no. 4, pp. 140:1–140:29, 2017.

[15] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 2642–2651.

[16] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 1158–1166. [Online]. Available: https://doi.org/10.1145/3219819.3219963