

Missing Data Completion for Telco localization by Adversarial Learning

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

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

Abstract—Telco localization, as an important complementary technique of GPS, is aimed at recovering mobility patterns of mobile devices (MDs) via the measurement report (MR) data generated in Telco networks. However, some RF signal strength (RSSI) values of MR records are commonly missing due to various reasons. As a result, MR data are often incomplete, raising the missing data completion problem for MR. To address the problem, we propose a deep learning solution, called TelcoGAN, by adapting the well-known Generative Adversarial Nets. To learn the internal correlation of MR data, we describe a serving-centric space for pre-processing. The TelcoGAN consists of three components: (i) a localizer for predicting locations of input MR data; (ii) a generator for generating complete MR data based on input missing MR data; and (iii) a discriminator to learn the data distribution between real samples and completed samples from the generator. The generator tries to generate completed MR data that can fool the discriminator. The localizer exploits available position labels to stabilize the adversarial training process and guide the optimization of TelcoGAN. Experiments on two real-world MR data sets demonstrate the improvements of TelcoGAN in RMSE and localization error, compared with state-of-the-art baselines.

Index Terms—missing data completion, Telco localization, generative adversarial nets

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. 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. 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.

State-of-the-art data completion algorithms can be classified into two categories: discriminative methods and generative methods. The discriminative models maximize the conditional probability of missing components given observed components. These methods include KNN [9], MICE [10] and matrix completion [11]. The generative methods maximize the joint probability of both observed and missing components, such as algorithms based on Expectation Maximization [12] and deep learning (e.g., denoising autoencoder (DAE)) [13], [14]. However, there are two main challenges to directly apply existing data completion methods into our scenario. 1) *Complex interactions of MR data*: Due to complicated Telco operation design [8], the internal relationship of MR data is sophisticated and nonlinear. 2) *Noisy signal strength*: The challenging RF propagation phenomena (e.g., multipath and non-line-of-sight) causes noise signal strength values. Noise caused by such fluctuation can hurt the inference of missing RSSI. Existing methods with shadow capability can not capture the distribution of MR data well. Besides, by retrieving locations from location-based services, the position labels is available in training data. However, such labels are not accessible in the testing phase. It is non-trivial to incorporate the position labels of training data into missing data completion.

With recent breakthroughs in deep learning, Generative Adversarial Nets (GANs), as a powerful technique for generative modelling, has shown impressive results in realistic data generation [15], [16]. The GANs plays a game between two networks: a generator that produces completed MR data given MR missing data and a discriminator that produces

probability distribution between the completed and real data [15]. Compared with traditional machine learning models, the competing process between two networks are better at learning sophisticated internal correlation of data.

To this end, we propose TelcoGAN that built upon GANs to address the MR missing data problem. TelcoGAN takes advantage of available GPS location labels in training data set (e.g., retrieving location coordinates from location-based services [17]), to stabilize the training process and lead to better completion performance. Our research makes two main contributions:

- We propose TelcoGAN which cooperates a *serving-centric space* and a *localizer network* to produce high quality complete MR data. The serving-centric space component helps to learn the internal correlation of MR. The localizer network utilizes available GPS location labels to guide the adversarial process towards better results. In addition, we adopt a hybrid loss trick which combines mean squared loss and adversarial loss to further improve the performance.
- Evaluations on two real-world MR data sets show that our model achieves better performance than state-of-arts and the model trained on a spatial domain can improve the performance of another one.

The rest of this paper is organized as follows. Section II gives preliminary background and some related works. In Section III, we describe the technical design of TelcoGAN in detail. After that, we conduct extensive experiments in Section IV and review some related work in Section V. Finally, we conclude in Section VI.

Table summarizes the main symbols and associated meanings used in this paper.

II. PRELIMINARIES

In this section, we first give a detailed description of MR data and the problem definition, next illustrate the solution overview, and finally give the basic of GANs [15].

A. Data Description

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, cell 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**). Normally, Telco networks set one of the nearby cells as serving cell to provide data services or communication for the connected device. The serving cell is highlighted as **Serving_eNodeBID** and **Serving_CellID**, the same as the first connected cell (**eNodeBID_1** and **CellID_1**).

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

B. Problem Definition

Problem 1: [Telco Missing Data Completion]: For a MR record with missing RSSI, Telco missing data completion problem is to recover a MR record with complete RSSI.

C. Basic of GANs

The key of Generative Adversarial Nets (GANs) [15] 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:

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

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 1.

However, the unstable training process makes GANs hard to train. It has been shown in previous work [18] 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 [19]. The proposed classification loss is described as follows:

$$L_{CLS}(x, y, l) = L_{CE}(D(x, y), \ell) \quad (2)$$

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 three parts. First, we describe the design of TelcoGAN model

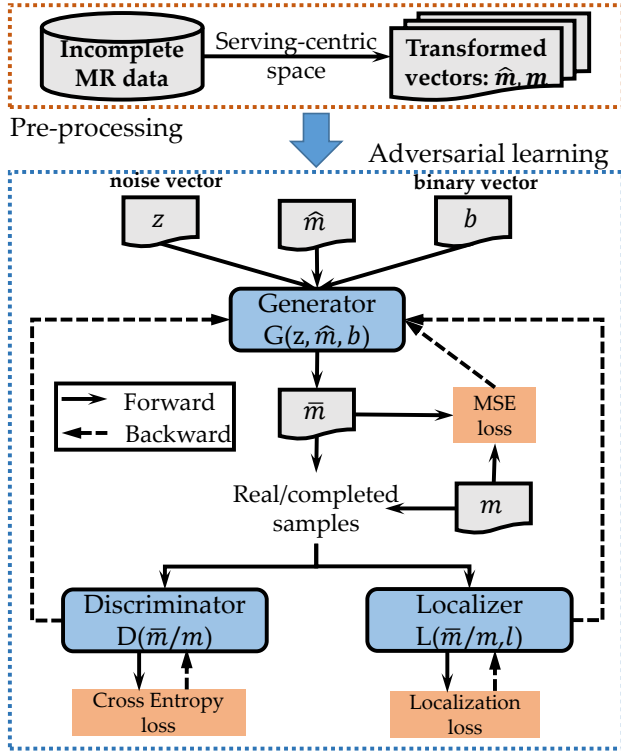


Fig. 1. The framework of TelcoGAN

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 serving-centric space. Then the sparse global coordinate-based distribution of MR data is transformed into a dense relative coordinate-based distribution. Thus TelcoGAN model can better capture the internal correlation of MR data. We assume the following notations: m denotes the transformed complete MR matrix, \hat{m} the input incomplete MR matrix, \tilde{m} the generated completed MR matrix and l the associated position label.

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

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

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

(3) **Localizer** $L(\tilde{m}/m, l)$: 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. Serving-centric Space

Here we construct a serving-centric space for MR data to better learn the internal relationship.

As seen in Table. I, each MR record has a serving cell. According to Telco operations [7], the serving cell is selected from these nearby cells with good connection, which means close distance to mobile device. Thus a serving cell indicates a specific spatial domain. Given the total MR records with extensive global spatial domain, we group them as multiple local spatial domains by the serving cell. Then the whole city-scale area is divided into multiple small spatial domains.

Based on the division above, we propose a *serving-centric space*, to transform the sparse global spatial distribution into a dense local one. In Fig. , a spatial domain is an area centered on the serving cell. Provided the cell tower database from Telco operators, we can obtain the GPS coordinate location of each cell. Based on the cell location data, we can do coordinate conversion as follows. Suppose the GPS coordinate (longitude, latitude) of the serving cell is (x_0, y_0) , those of the neighboring cells are $\{(x_i, y_i)\} (1 \leq i \leq 5)$. Under the serving-centric space, the new coordinates are $(0, 0)$ and $\{(x_i - x_0, y_i - y_0)\} (1 \leq i \leq 5)$ for the serving cell and the neighboring cells respectively.

The **rational** behind serving-centric space is not difficult to understand: suppose normal cells and similar RF environment, in principle, the received signal strengths are only decided by the relative locations of connected cells and the MD. The serving-centric space offers advantages as: 1) reveal the true internal relationship of MR data and 2) knowledge learned

from a domain can transfer to another domain and improve data quality.

C. TelcoGAN Generator

The goal here is to generate completed MR data by modelling the conditional probability distribution of missing components given the observed components. Note that to capture the uncertainty of the missing values [10], we choose to input random vectors for randomness.

We denote the input incomplete MR matrix as $\hat{m} = \{x', y', \hat{r}\}$, where x'/y' is the relative latitude/longitude vector of connected cells, and \hat{r} is the RSSI vector with missing values. We also denote the input random vector as z . To enable the randomness unified into the input MR matrix, we use the binary indicator vector b as:

$$b_i = \begin{cases} 0, & \text{if } \hat{r}_i = \text{nan} \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

so that b indicates which positions of \hat{r} are missing.

Thus, the missing values of \hat{r} is filled with $(1 - b) \odot z$, where \odot denotes element-wise multiplication. The input data matrix of generator G is computed as $\{x', y', \hat{r} + (1 - b) \odot z\}$. The generator G can be formulated with:

$$\tilde{m} = G(z, \hat{m}, b) \quad (4)$$

$$\tilde{m} = \tilde{m} \odot (1 - b) + \hat{m} \odot b \quad (5)$$

To capture the internal relationship between observed components and missing components, we construct an encoder-decoder network as generator network. The encoder network consists of convolutional layers, and the decoder network consists of deconvolutional layers. The encoder network is used to extract features that reflects context of MR record such as location, environment and etc. Because the input incomplete MR matrix and output completed MR matrix have the same size, the deconvolutional layers of decoder network are used to restore information of MR record. We use convolution kernel with 3×3 in convolutional layers to reduce the size of input data matrix.

D. TelcoGAN Discriminator

The TelcoGAN discriminator aims to differentiate completed MR records from the real ones by modelling the distribution of MR data.

The discriminator D can be formulated as a function $D(\tilde{m}/m)$, that takes a real MR record or a completed MR record from generator G as input and outputs probability over two categories (real/completed). We model D as several convolutional layers and a fully connected layer for classification.

E. TelcoGAN Localizer

The TelcoGAN localizer is designed to exploit the available position labels in training phase.

F. TelcoGAN Training

IV. EXPERIMENTS

A. Data sets and Evaluation metrics

B. Counterparts

C. Baseline Experiment

D. TelcoGAN with Different Portions of Missing RSSI

E. Ablation Study

V. RELATED WORK

Missing data completion: The pervasive missing data completion problem, also known as data imputation, has been widely studied for many years. Generally, the data completion techniques can be divided into three categories, statistical based methods,

Telco localization:

Deep learning and GANs:

VI. CONCLUSION

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] 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>
- [10] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, pp. 1–68, 2010.
- [11] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.
- [12] P. J. García-Laencina, J. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.
- [13] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, 2008, pp. 1096–1103.
- [14] 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.
- [15] 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.
- [16] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 105–114.
- [17] 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.
- [18] 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.
- [19] 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>