都市における人の移動性に対して VAE に基づく生成モデルの検討

A Variational Autoencoder Based Generative Model of Urban Human Mobility

学籍番号　47-166810
氏　　名　黄 豆（Huang, Dou）
指導教員　柴崎 亮介 教授

Abstract

Recently, big and heterogeneous human mobility data inspires many revolutionary ideas of implementing machine learning algorithms for solving some traditional social issues. However, incomplete datasets were provided owing to both the concerns of invasion of privacy and some technique issues in many practical applications, which leads to some limitations of the utility of collected data.

Inspired by the generative model used for reconstructing images in image processing domain, we want to build a generative model which can tackle the human mobility data. Variational Autoencoder (VAE), which uses a well-constructed latent space to capture salient features of the training data. By combining VAE and sequence-to-sequence (seq2seq) model, a Sequential Variational Autoencoder (SVAE) is built for the task of human mobility reconstruction. It is the first time that this kind of SVAE model is implemented for solving the issues about human mobility reconstruction. We use navigation GPS data of selected greater Tokyo area to evaluate the performance of the SVAE model. Experimental results demonstrate that the SVAE model can efficiently capture the salient features of human mobility data and generate more reasonable trajectories.

Keywords: big data, urban computing, GPS trajectory, generative model

Introduction

Privacy is a very complex topic. To prevent some risks of invasion of privacy of mobile device users, although many locational data is collected, the owner of such kind of data is not willing to provide the data for researchers or other research institutes, which leads to a limitation of the usage of this kind of locational data. There is a trade-off between the implement of collected locational data and the concerns about invasion of privacy of users, which is often that even the owner of the data is willing to provide the data for some research purpose, they are not going to provide the whole data, instead, a very small part of the data will be provided, such as 1% of the entire data set.

Despite the privacy concerns, there are some other problems which can also lead to the low sampling rate of the collected data. An example of such situation is that fishery data in the world. This kind of fishery data is an open data, but we still cannot get the data reflect the real trajectory patterns of all fishing boats since some of small fishing boats lack efficient device to record their trajectories and thus cannot be obtained.

Both privacy concerns of mobile device users and the lack of techniques of collection method in some cases will lead to the difficulty to obtain the human mobility data to reflect the real trajectory patterns in real situations. There are many research and implement based on the human mobility data, for instance, human mobility prediction. These applications usually need to use previous steps of trajectories to predict the human mobility in the future. We are not talking about the accuracy or performance of such methods, what we concern is that if we cannot get the data which can reflect the real situation of human mobility in a target area, it is difficult to predict the future human mobility in a proper way.

A generative model is not designed for the transportation planning and applications directly, but we can use this kind of model to improve the existing datasets to match the requirement of implementation of other applications.

Related Work

To simulate human behavior and moving patterns, various generative models have been developed in recent years. However, HMMs cannot completely model the temporal dependency of states. To improve the HMMs, Baratchi et al. proposed Hidden Semi-Markov Model(HSMM), which including the duration of the state into the hidden variables. In general, their works are all based on Hidden Markov Model, and focus on reconstruct the trajectories of human mobility following a specific probability distribution.

Very recently, a non-Parametric generative model for human trajectories has been proposed. They use Generative Adversarial Network (GAN) to produce data points after a simple and intuitive yet effective embedding for locations traces designed. It is the first time that deep learning methods implemented in building a generative model for human mobility in our knowledge.

The Sequntial Variational Autoencoder we build in this research has significant differences comparing with their model. Their work is a GAN based model which aims to generate fake data that can be recognized as true data by the trained discriminator. While the SVAE model in this research aims to learn the approximated latent distribution of training data first, then resample the fake data from this learned latent space. Besides, there is no need of trajectory transformation for trajectories when using SVAE model.

Methodology

Le $\boldsymbol{x} = (x_1, x_2, \dots, x_t)$ denote a high dimensional sequence, such as a trajectory of human mobility wit t steps. We use a LSTM as recurrent encoder to capture the information of the input trajectory x. Then we will obtain a series of hidden state s_t, and a series of output o_t. In actual case, what we really care about is the final output o rather a sequence of output value o_t. Since we only keep the final output, we can obatin an intermediate non-sequential vector o to represent the information captured from the input sequence using this recurrent encoder. After intermediate vector o is obtained, we treat this vector as the input of the Variational Autoencoder part. Then we can write the joint probability of the model as $p(o, z) =$

$p(o|z)p(z)$. $p(z)$ is a prior latent distribution, and $p(o|z)$ is the likelihood. Then we need to calcalute the posterior latent distribution $p(z|o)$ given observed data:

$$p(z|o) = \frac{p(o|z)p(z)}{p(o)}$$

by marginalizing out the latent distribution:

$$p(z|o) = \frac{p(o|z)p(z)}{\int p(o|z)p(z)dz}$$

This is an exponential time-consuming process.

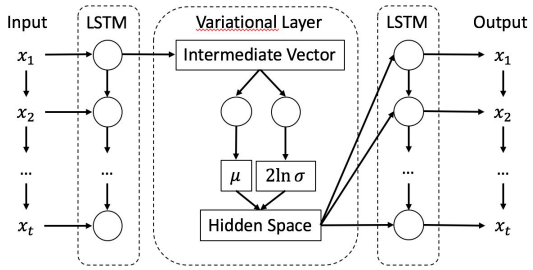

Figure 1 Framework of Sequential Variational Autoencoder

Therefore, variational inference approximates the posterior with a family of distribution $q_\lambda(z|o)$. Usually, we choose q to follow a Gaussian distribution, then λ would be the mean and variance of the latent distribution $\lambda = (\mu, \sigma)$. Kullback-Leibler divergence is used for measuring the information lost when using q to approximate p, the optimal approximate posterior is thus:

$$q_\lambda^*(z|o) = argmin_\lambda\, KL(q_\lambda(z|o)||p(z|o))$$

In the SVAE model, we parametrize approximate posterior $q_\theta(z|o)$ using an inference network, approximate likelihood $p_\phi(o|z)$ using a generative network. Then the loss of the model will be:

$$loss = -E_{q_\theta(z|o)}[logp_\phi(o|z)] + KL(q_\theta(z|o)||p(z))$$

Finally, we use another LSTM neural network as recurrent decoder to reconstruct the trajectories of human mobility, x from parameters in learned latent distribution. Training process of the model is shown as figure 1.

Experiment

We conduct the experiment using navigation GPS data after data preprocessing. The visualization of results is given by figure 2. It consists of training data, reconstructed data, resampled data, and 10 times resampled data, which give an intuition that the results look reasonable.

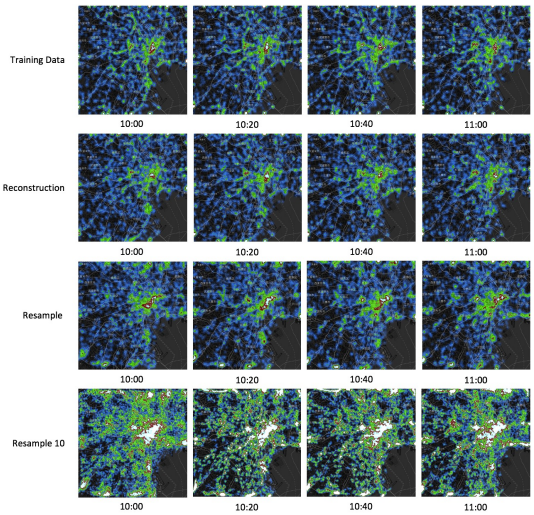

Figure 2 Visualization of results

Mean Distance Error (MDE) is used to quantitatively evaluate the results of SVAE model.

$$E_j = \frac{\sum_{i=1}^{N} dis(l_{i,j}, \hat{l}_{i,j})}{N}$$

Where $dis(l_{i,j}, \hat{l}_{i,j})$ calculate the distance of ground truth and reconstructed points. Comparison of MDE between different parameter settings is shown as figure 3.

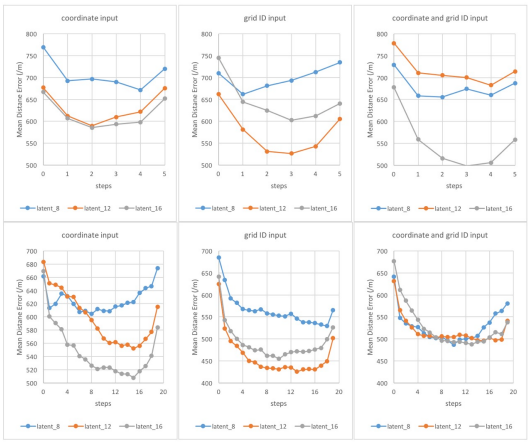

Figure 3 Mean Distance Errors

Conclusions and Future Work

In this research, we make a brief introduction about VAE and seq2seq model, then combine these two frameworks to build a Sequential Variational Autoencoder (SVAE). It is believed that this SVAE is first time implemented in modeling trajectories of human mobility. We use navigation GPS data of cars in Tokyo to evaluate the performance of SVAE model. The performance of SVAE with different parameter settings and its explanation have been discussed. In general, the SVAE model can capture the salient features of input trajectories using a latent space constructed by following Gaussian distribution, then reconstruct the input trajectories. As a generative model, the ability of generating fake resampled trajectories of SVAE is also proved. Using this SVAE model, we can generate more trajectories of human mobility which have similar pattern with training data to solve the low sampling rate problem. Besides, it is a good choice for preventing the risk of privacy invasion by implementing SVAE model to learn the salient features of confidential data. Then we can reconstruct the dataset which has similar patterns but has no privacy information. In addition, this model can improve the performance of practical applications by improve the dataset on which they based.

We also note the limitation of SVAE model when implemented in trajectories of human mobility, which is that many points of reconstructed trajectories is not located in road network. A solution for that problem is A road network based generative model should be built, which makes sure the reconstructed trajectories can all located in the road network.

Reference

[1] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[2] Ouyang, Kun, et al. "A Non-Parametric Generative Model for Human Trajectories."

[3] Kullback, Solomon, and Richard A. Leibler. "On information and sufficiency." The annals of mathematical statistics 22.1 (1951): 79-86.

[4] Baratchi, Mitra, et al. "A hierarchical hidden semi-Markov model for modeling mobility data." Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2014.