西安交通大学
XI'AN JIAOTONG UNIVERSITY

基于变分自编码器模型的轨迹用户链接问题研究

汇报人：王浩辰
指导老师：孙鹤立 教授
计算机65班

2020年06月09日

内 容 纲 要

西安交通大學
XI'AN JIAOTONG UNIVERSITY

内容一：背景及意义

背景&意义

基于地理标签的应用（GTSM）

轨迹-用户链接问题[1]（TUL）重要任务

1 个性化兴趣点推荐 POI Recommendation

2 活动识别 Activity Identification

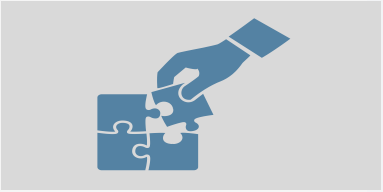
3 个人行为建模 Individual Activity

[1]Zheng Yu. Trajectory data mining: an overview. ACM Transactions on Intelligent Systems and Technology. 2015 May 12;6(3):1-41.

VAE [2]—— 一种基于神经网络的生成模型

减轻了数据稀疏性问题

挖掘隐式变量



捕捉特定语义

[2]Doersch C. Tutorial on variational autoencoders[J]. arXiv. 2016 preprint arXiv:1606.05908

西安交通大學

XI'AN JIAOTONG UNIVERSITY

内容二：TUL问题定义

TUL

轨迹数据：$t_{u_i} = \{c_{i1}, c_{i2}, \dots, c_{in}\}$

位置信息：$c_{ij} = (x_{ij}, y_{ij})$

未链接的轨迹数据：$\overline{t_i} = \{c_1, c_2, \dots, c_m\}$

$\mathcal{T} = \{\overline{t_1}, \dots, \overline{t_m}\}$　$\mathcal{U} = \{u_1, \dots, u_n\}\,(m \gg n)$

目标：$\mathcal{T} \mapsto \mathcal{U}$

未链接
轨迹数据集合

用户集合

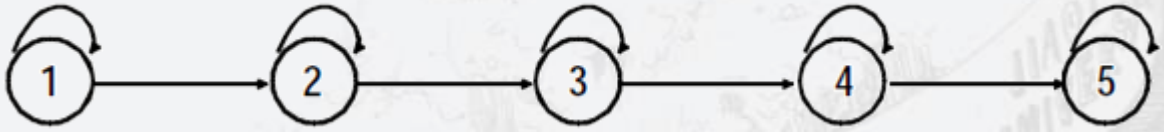
西安交通大学
XI'AN JIAOTONG UNIVERSITY

内容三：相关研究现状

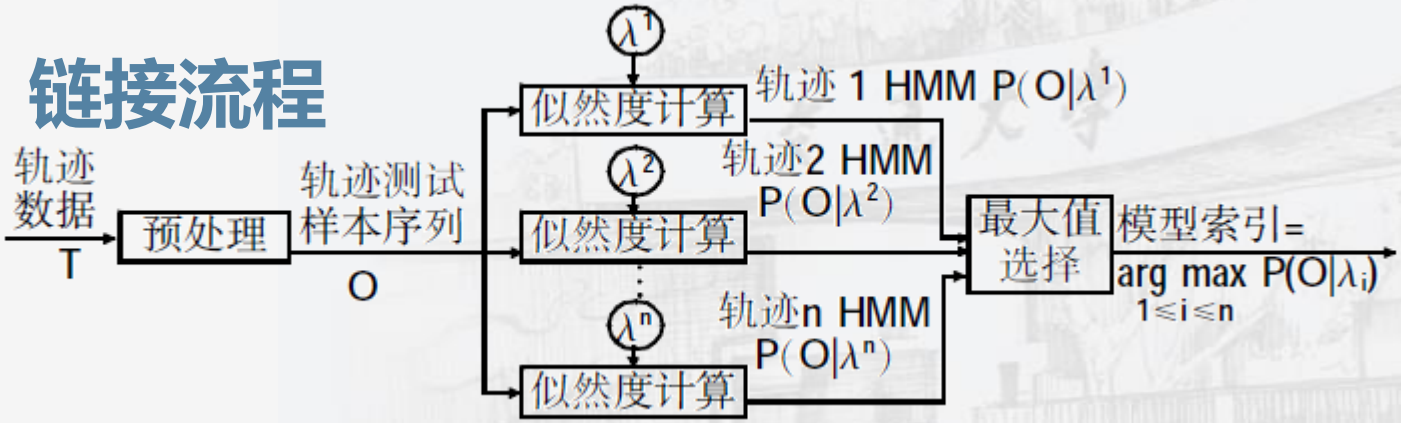
隐马尔可夫模型

- 实质上是在上述已训练好的隐马尔科夫模型中选择一个可以最佳地描述观测轨迹序列的模型[3]
- 假设在非相邻的位置间存在强独立性，失去了兴趣点对时间的依赖性，轨迹数据生硬地平均分段

状态转移图



链接流程



[3]潘奇明，程咏梅．基于隐马尔可夫模型的运动目标轨迹识别［J］．计算机应用研究．2008．25（7）：1988-1990．

循环神经网络

- 对变长序列分布建模的强大优势，例如LSTM可以很好地捕获一段时间内兴趣点序列的独立性特征[4]
- 缺乏对人类活动分层语义的理解，且无法利用未标记的数据进行深度学习

数学表示

$$h_t = \varphi\left(x_t, \varphi\left(x_{t-1}, \varphi\left(x_{t-2}, \varphi(\dots)\right)\right)\right) = f(x_{1:t})$$

[4]Wu Hao, Chen Ziyang, Sun Weiwei, et al. Modeling trajectories with recurrent neural networks[C]. IJCAI. 2017.

西安交通大學
XI'AN JIAOTONG UNIVERSITY

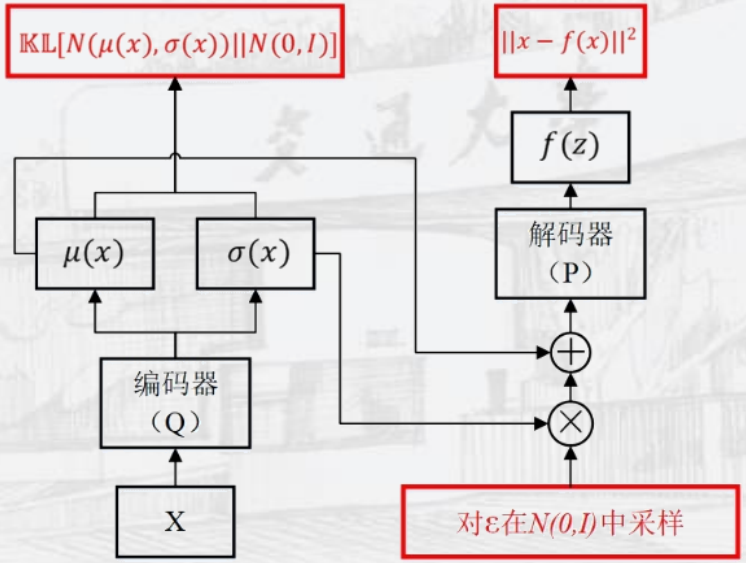
内容四：变分自编码器轨迹用户链接

VAE定义

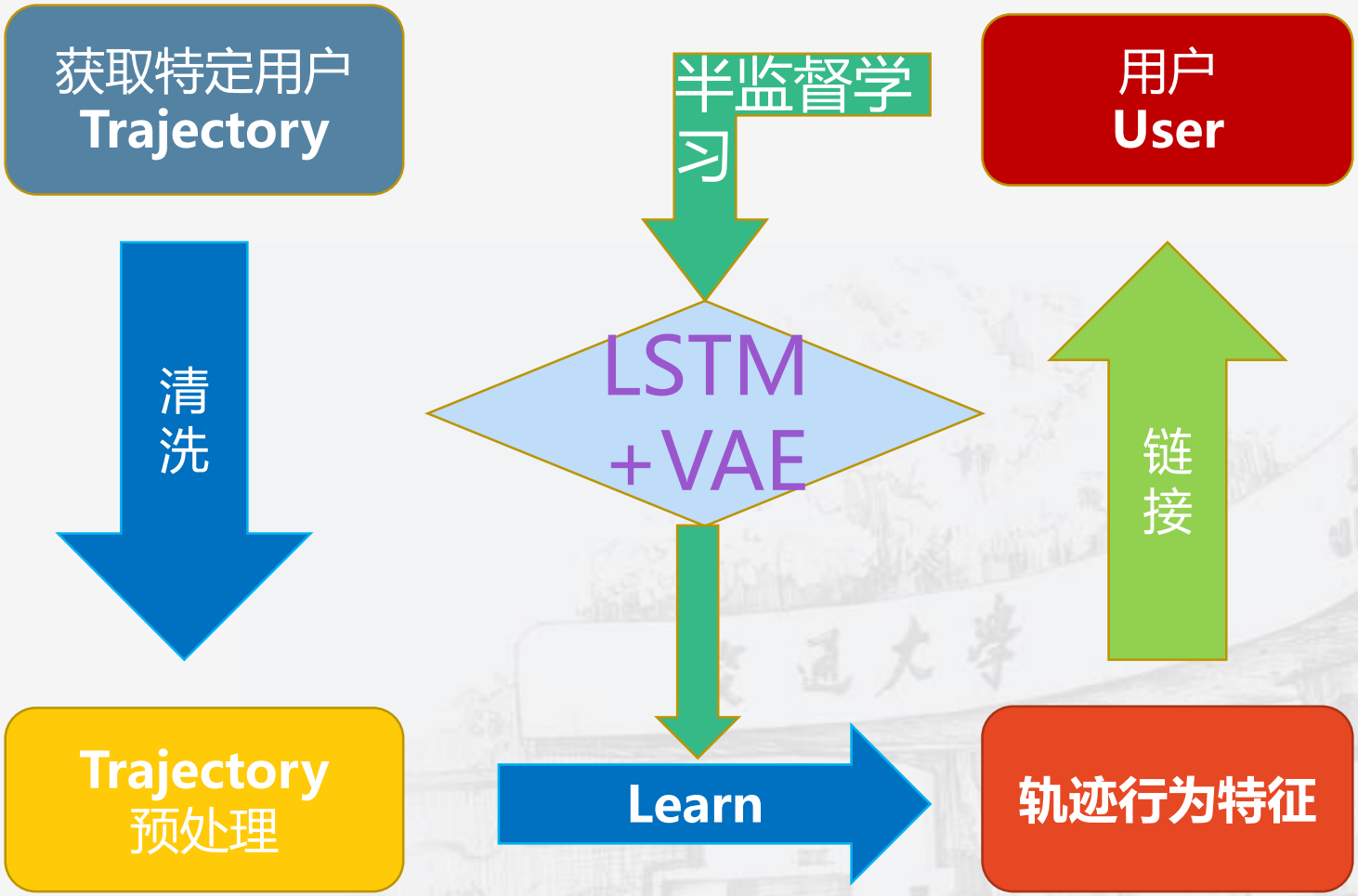
- 隐式变量：模型生成兴趣点序列的决定性因子
- 目标： $\max P(X) = \max \int P(X \mid z;\theta)P(z)dz$
- $z = sample\left(N\left(\mu(X), \Sigma(X)\right)\right)$ 重参数化"：

$$z = \mu(X) + \sigma^{\frac{1}{2}}(X) * \epsilon, \epsilon \sim N(0, I)$$

VAE模型结构



整 体 思 路


获取特定用户
Trajectory
半监督学习
用户
User
清洗
LSTM
+VAE
链接
Trajectory
预处理
Learn
轨迹行为特征

数据清洗：

无效信息：“0000000000000”---删除

错误数据：有悖于正常生活实际的平均速度

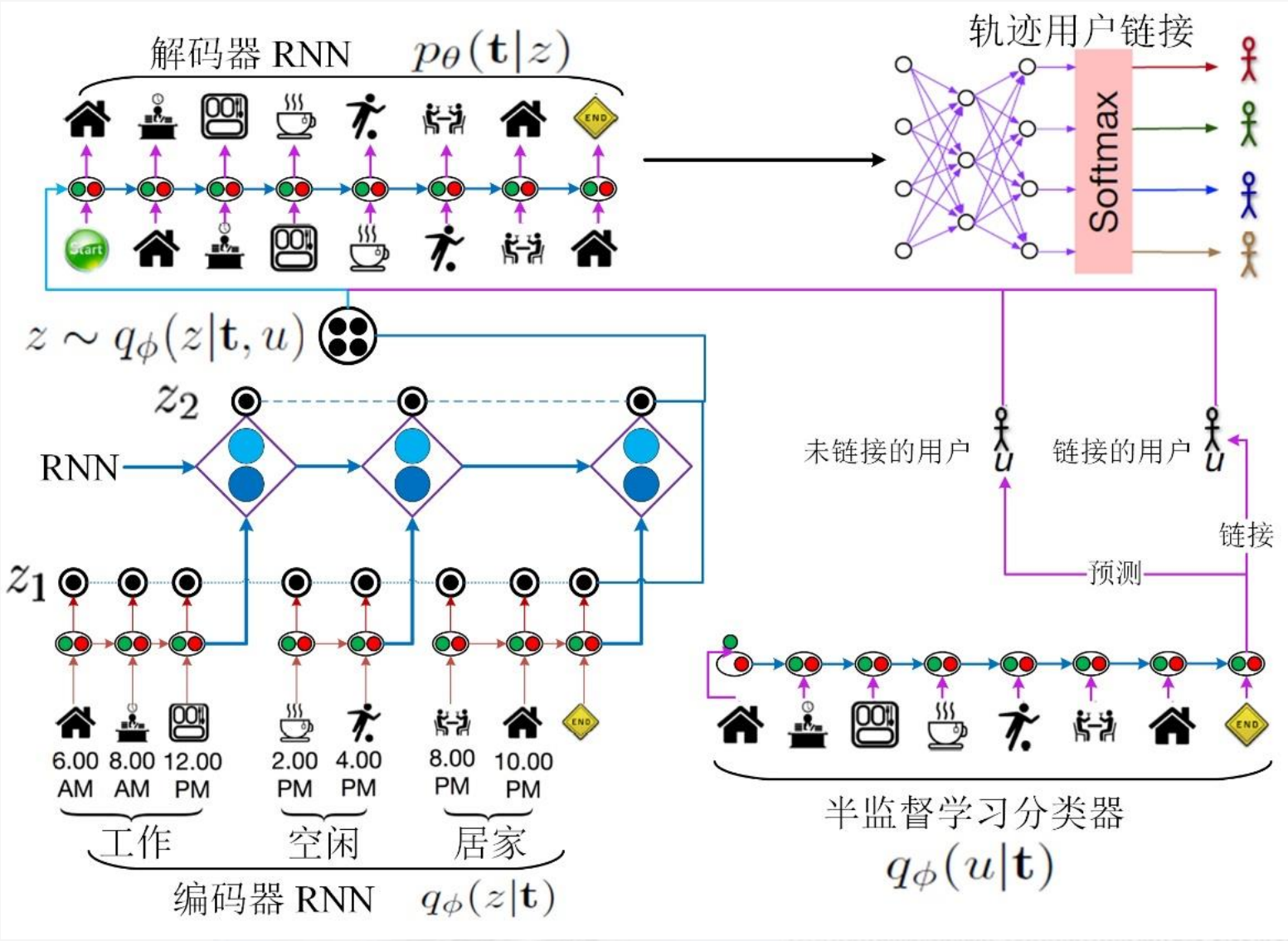
---“300探针”

轨迹数据分段：

时间阈值：4或6小时

兴趣点距离阈值：50km

实 现 思 路


解码器 RNN $p_\theta(\mathbf{t}|z)$
轨迹用户链接
Softmax
$z \sim q_\phi(z|\mathbf{t}, u)$
z_2
RNN
z_1
未链接的用户 u
链接的用户 u
链接
预测
6.00 AM
8.00 AM
12.00 PM
2.00 PM
4.00 PM
8.00 PM
10.00 PM
工作
空闲
居家
编码器 RNN $q_\phi(z|\mathbf{t})$
半监督学习分类器
$q_\phi(u|\mathbf{t})$

西安交通大學

XI'AN JIAOTONG UNIVERSITY

内容五：实验结果与分析

实验数据-实验环境

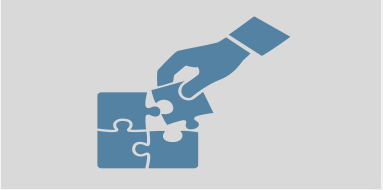
编程语言：Python 3.7、Python 2.7

机器学习框架：TensorFlow 1.14.0、Keras 2.3.1

可视化方法：Matplotlib 3.1.1、Graphviz 2.38

数据集：

Gowalla	201	Brightkite	92	Geolife	50
	112		34		36



超 参：

enbedding_dim=320
latent_dim=2
batch_size=64
epochs=300

对比方法：

（1）传统方法：LDA模型、决策树、随机森林、支撑向量机

（2）循环神经网络：LSTM、双向LSTM、双TULER[5]

评价指标：

ACC@K、Macro-P、Macro-R、Macro-F1

$$macro\text{-}F1 = 2 \times \frac{macro\text{-}P \times macro\text{-}R}{macro\text{-}P - macro\text{-}R}$$

[5] Fan Zhou, Qiang Gao, Goce Trajcevski, et al. Trajectory-User Linking via Variational AutoEncoder[C]. IJCAI. 2018.

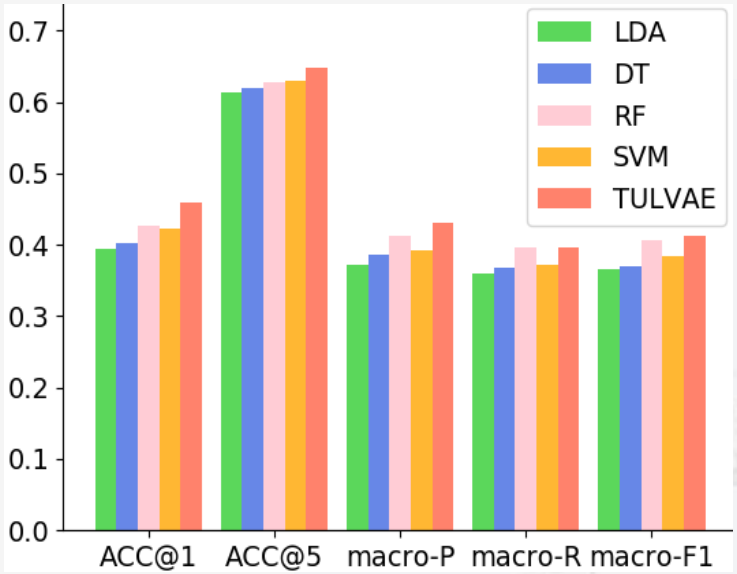
实验结果

模型方法	Brightkite数据集 $\|\mathcal{U}\| = 92$				
	ACC@1	ACC@5	macro-P	macro-R	macro-F1
TULER-LSTM	43.01%	59.84%	38.45%	35.81%	37.08%
TULER-LSTM-S	44.23%	61.00%	38.02%	36.33%	37.16%
Bi-TULER	43.54%	60.68%	38.20%	36.47%	37.31%
TULVAE	**45.98%**	**64.84%**	**43.15%**	**39.65%**	**41.32%**

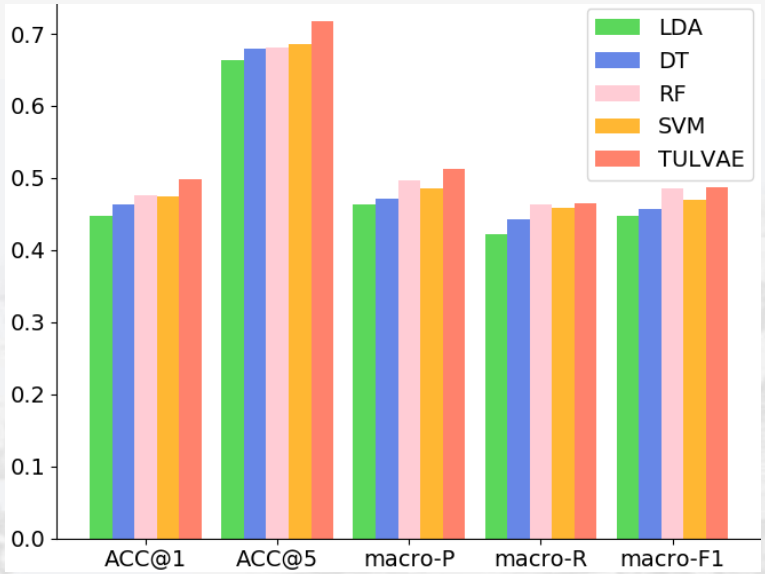
模型方法	Brightkite数据集 $\|\mathcal{U}\| = 34$				
	ACC@1	ACC@5	macro-P	macro-R	macro-F1
TULER-LSTM	48.26%	67.39%	49.90%	**47.20%**	48.51%
TULER-LSTM-S	47.88%	67.38%	48.81%	47.03%	47.62%
Bi-TULER	48.13%	68.17%	49.15%	47.06%	48.08%
TULVAE	**49.82%**	**71.71%**	**51.26%**	46.43%	**48.72%**

基于变分编码器与其他循环神经网络的轨迹用户链接性能比较

基于变分自编码器与传统方法的轨迹用户链接性能比较



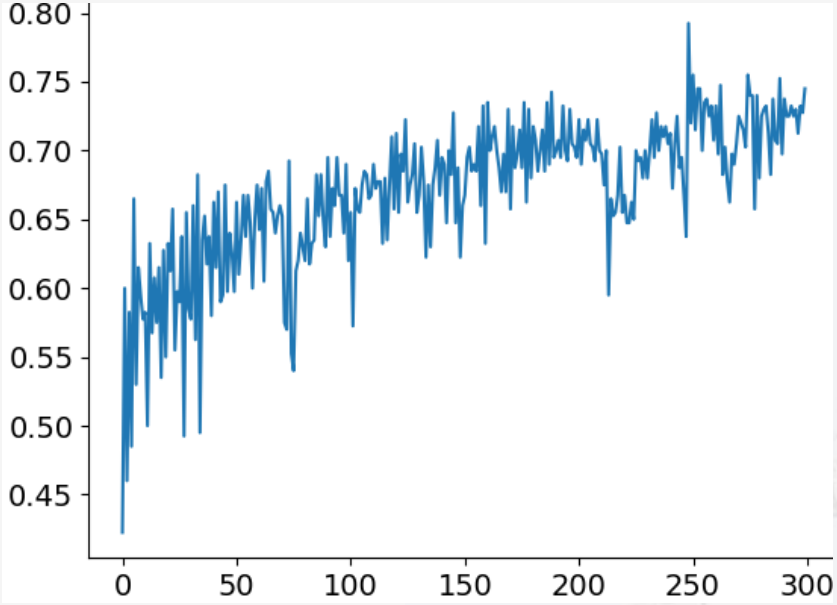
（a）$|\mathcal{U}| = 92$



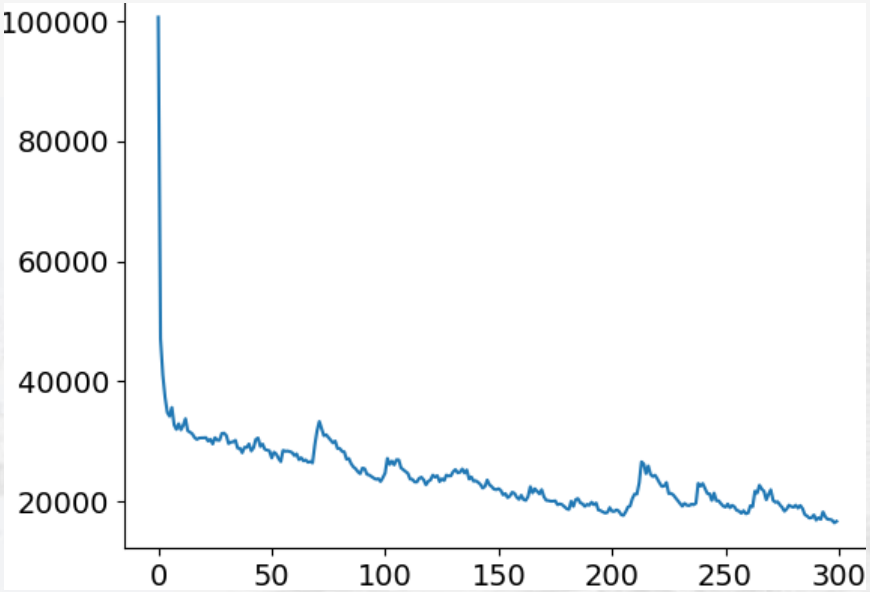
（b）$|\mathcal{U}| = 34$

Brightkite数据集

基于变分自编码器的轨迹用户链接模型
训练性能可视化表征



（a）准确率随训练轮次的变化



（b）损失值随训练轮次的变化

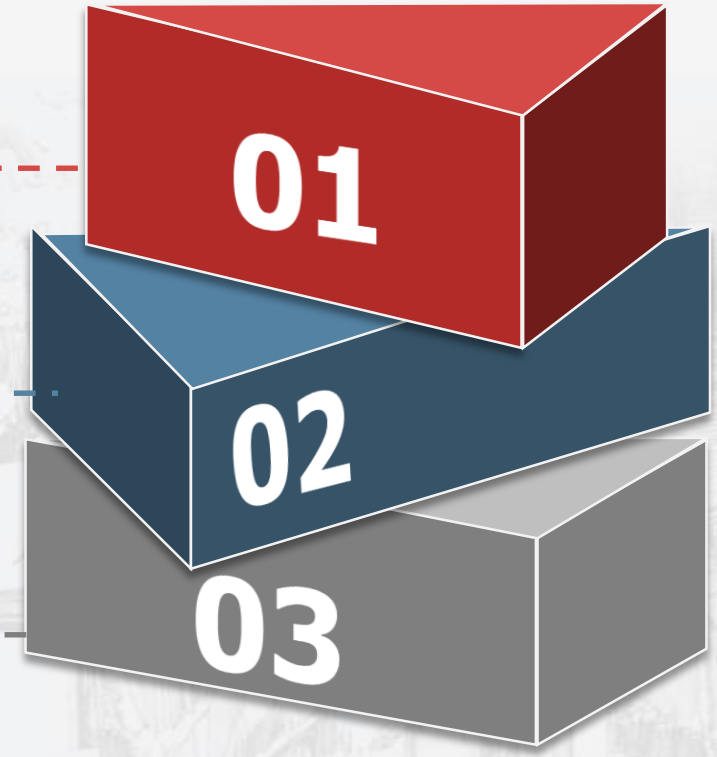
Brightkite数据集

实 验 结 论

大多数指标表现最佳

缓解了数据稀疏性问题

轨迹分层建模



西安交通大学
XI'AN JIAOTONG UNIVERSITY

内容四：总结与展望

总 结 与 展 望

- 提出基于变分自编码器的轨迹用户链接模型
- 引入半监督学习机制
- 结合循环神经网络的分层轨迹建模

编码器或解码器变形（CNN）

增强变分自编码器的推断能力

西安交通大學
XI'AN JIAOTONG UNIVERSITY

谢谢大家

欢迎专家老师批评指导

汇报人：王浩辰
指导老师：孙鹤立 教授
计算机65班

2020年06月09日