**用于连续POI推荐的内容感知层次兴趣点嵌入模型**

**摘要**

基于时间和空间上下文信息的用户下一步访问感兴趣点(POI)是移动应用中的一项重要任务。近年来，已有几种基于传统顺序数据建模方法的POI推荐模型被提出。然而，这种模型只关注用户的签入序列信息和POI之间的物理距离。此外，它们没有利用POI的特性或POI之间的关系。为了解决这个问题，我们提出了CAPE，这是第一个基于内容的POI嵌入模型，它利用文本内容提供关于POI特性的信息。CAPE由签入上下文层和文本内容层组成。签入上下文层从用户的签入序列中获取POI的地理影响，而文本内容层则从文本内容中捕获POI的特征。实验评价表明，简单地将CAPE应用于POI推荐模型，可以显著提高POI推荐模型的性能。

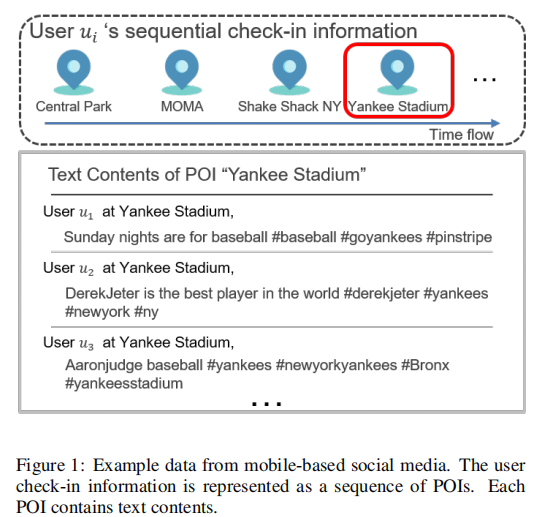
**1 介绍**

近年来，许多基于移动的社交媒体平台，如twitter、facebook、instagram和Foursquare都越来越受欢迎。这些平台上的用户产生大量的数据，其中包括包含时间和空间信息的文本内容。这些信息对于理解用户对感兴趣点(POI)的行为和偏好特别有用，而POI是某个人感兴趣的特定位置。在基于移动的应用程序中，根据时间和空间信息推荐用户下一步访问的位置是必要的，但也是具有挑战性的。

最近，已经提出了几种基于传统的顺序数据建模方法并且被设计用于推荐下一个POI的模型。基于马尔可夫链的模型[1,2]在考虑时间和空间约束的同时，利用用户的签入序列信息。基于张量分解，STELLAR [13]模型POI对相继访问。此外，ST-RNN [5]采用了递归神经网络(RNN)，并在顺序数据建模方面取得了最先进的性能。

虽然ST-RNN取得了最先进的性能，但仍有改进的余地。由于ST-RNN采用随机初始化向量作为POI表示，因此它不利用POI的特性或POI之间的关系。ST-RNN只使用两个签入之间的时间间隔和两个POI之间的距离作为推荐下一个POI的因素。

尽管最近提出POI嵌入模型[6,3,14]来捕获POI之间的关系，但它们尚未用于连续的POI推荐任务中。此外，现有的POI嵌入模型仅使用用户登记上下文来捕获POI的地理影响，但是它们不利用提供关于POI的特征的信息的文本内容。例如，如果POI的文本内容包括单词“latte”、“咖啡”和“饮料”，则POI具有与咖啡馆相关联的特性。图1显示了用户编写的POI“扬基球场”的文本内容示例。



在本文中，我们首先描述了我们新构建的包含Instagram文本内容的数据集，并分析了用于POI嵌入研究的数据集的特点。其次，我们提出了CAPE，它是一种用于POI推荐的内容感知分层POI嵌入模型。从文本内容上看，CAPE不仅抓住了POI的地理影响，而且还抓住了POI的特点。CAPE由签到上下文层和文本内容层组成。 在登记上下文层中，使用Skip-gram [7]训练用户登记序列中的POI的嵌入向量在向量空间中彼此接近。同时，在文本内容层中，对POI嵌入向量进行训练，以便从POI文本内容中使用的单词捕获POI的特征。在实验部分，我们实现了连续的POI推荐模型，并将我们的CAPE模型应用到POI推荐模型中。在我们的数据集上的实验评估表明，CAPE有助于提高连续POI推荐模型的性能。评价结果还表明，CAPE捕捉到了POI的地理影响和POI的特征。CAPE的性能优于以前提出的POI嵌入模型。

本文的贡献总结如下：

* 我们构建了一个新的大规模POI数据集，并对该数据集进行了分析，以探讨其在POI推荐研究中的适用性。说明了文本内容在POI嵌入中的重要性。
* 提出了一种新的基于内容感知的层次POI嵌入模型-CAPE，该模型利用了用户的签入序列和有关POI的文本内容。据我们所知，CAPE是第一个具有内容感知的POI嵌入模型。
* CAPE捕捉了POI的地理影响和特性，极大地提高了后续POI推荐模型的性能。
* 我们发布了新构建的大规模数据集，其中包括基于移动的社交网络用户编写的文本内容。我们在开放库中实现的源代码也是公开的。

本文的其余部分按以下方式组织。在第二节中，我们总结了以前提出的连续POI推荐模型和POI嵌入模型。在第三节中，我们介绍了我们新构建的包含文本内容的数据集，并研究了基于内容感知的POI嵌入的必要性。在第四节中，我们描述了我们提议的CAPE模型。在第五节中，我们将我们提出的模型与现有的POI嵌入模型进行了比较，并分析了实验结果。最后，在第六节中，我们给出了本文的结束语。

**2 相关工作**

**2.1 连续POI推荐**

目前已有几种基于顺序数据建模方法的POI推荐模型。赵等人[13]提出了一种时空潜在排序（STELLAR）方法，该方法采用基于排序的成对张量分解框架来建立连续的POI。为了捕获和利用登记时间对POI推荐的时间影响，STELLAR不仅模型POI-POI对，还模拟时间POI对。但是，STELLAR无法对用户的整个签到序列进行建模，因为它侧重于成对的连续POI。Cheng等[2]和Cheng等 [1]提出了基于因子化个性化马尔可夫链（FPMC）的推荐模型[8]，它通过分解概率转换来扩展马尔可夫链模型。考虑到用户的运动限制，Cheng等人的模型[2]嵌入了个性化的马尔可夫链。陈等人的模型 [1]将个性化马尔可夫链与全局马尔可夫链相结合，捕捉全球行为模式。然而，由于基于FPMC的模型线性地结合了所有嵌入的马尔可夫链和约束，因此它们需要在不同因素之间具有很强的独立性假设[Wang等人，2015]。随着RNN在顺序数据建模中的成功，RNN也被用来为后续POI推荐建模用户的签入序列。刘等人[Liu等人，2016a]提出的ST-RNN将RNN扩展为具有特定时间和距离的转移矩阵，以分别建模局部时间和空间上下文。

虽然上述模型在某些应用中取得了令人满意的结果，但模型无法利用POI的特征和POI之间的关系，因为它们采用随机初始化的向量作为POI表示。在本文中，我们使用我们预先训练的POI嵌入模型角来改进现有的POI推荐模型的性能，该模型抓住了POI的特点。

**2.2 POI嵌入**

由于单词嵌入已被证明在NLP领域取得了成功，研究已经提出了POI嵌入模型，该模型能够依次从用户的检查中获取POI的地理影响。刘等人[Liu等人，2016 b]假设POI在用户签入顺序中受其他POI的影响。他们将每个POI视为“单词”，将每个用户的顺序检入信息视为“句子”，并使用Skip-gram[Mikolov等人，2013年]训练POI的潜在表示向量。此外，他们还提出了一种基于时间感知的POI推荐模型，该模型使用了时间隐式表示向量。冯等人[von等人，2017]提出了POI2Vec，这是一个包含地理影响的POI潜在代表模型。赵等人[赵等人，2017年]考虑了POI的时间特征。由于在不同的日子访问的POI之间的相关性很低，他们将用户一天的签入作为一个“句子”来对待。然后，他们提出了一个时态POI嵌入模型，利用特定的时间状态区分工作日和周末。

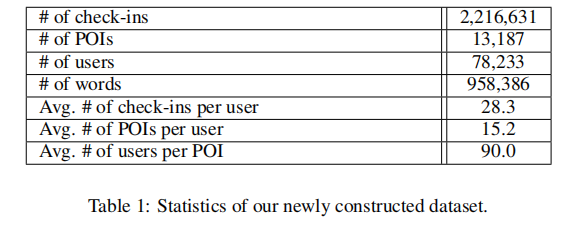
很少有将文本内容用于POI嵌入的研究。然而，我们的模型利用用户的签入序列信息和文本内容来嵌入POI。

**3 数据描述与分析**

在本节中，我们首先描述了包含文本内容的大规模数据集，然后分析了数据集的属性。

**3.1数据说明**

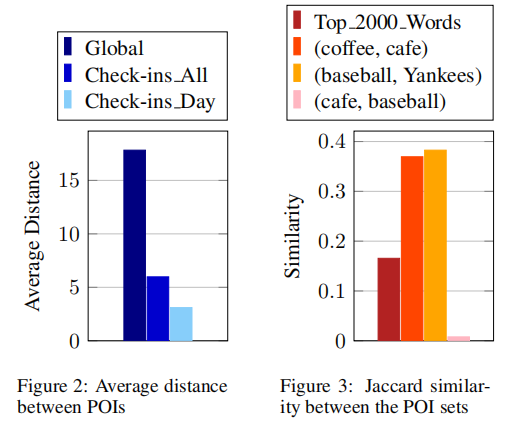
由于现有的POI推荐研究大多不使用文本内容，因此没有合适的数据集包含文本内容。尽管Yang等人。[Yang等人，2013]使用文本内容来理解用户情绪并提高POI推荐性能，他们的数据集不适合学习POI表示。首先，数据集中的许多POI没有用户编写的文本内容。第二，数据集不足以训练POI表示。因此，我们构建了一个新的数据集，其中包含引用POI的文本内容。我们从Instagram收集数据，Instagram是最受欢迎的移动社交网络之一。Instagram数据不仅包括用户POI签入信息，还包括用户编写的文本内容。我们收集了在纽约市创建的Instagram数据，并使用相同的方法对收集到的数据进行预处理。[赵等人，2017年]。我们删除了不到5次登录的POI和不到10条帖子的用户。经过预处理后，我们的新数据集包括2216631个登录点，共有78233名用户中的13187个POI。我们的数据集的统计数据汇总在表1中。



**3.2实证分析**

Instagram每天生成大量的位置感知数据。然而，Instagram数据集仅用于一项POI推荐研究[Wang等人，2017]，使用图片了解用户对POI的偏好。因此，我们进行了数据分析，以确定我们新构建的数据集是否适合连续的POI推荐任务。我们还分析了单词和POI之间的关系，以捕捉文本内容在POI嵌入中的重要性。

**地理影响** 根据POI推荐研究[ye等人，2011年；Liu等人，2013年]，基于位置的社交网络用户倾向于访问用户已经访问过的POI附近的POI。我们将考察数据集中的用户是否有类似的倾向。我们测量用户登记信息（签到）中POI之间的平均距离，并将该距离与数据集中所有POI（全局）之间的平均距离进行比较。我们使用Liu等人[刘等人，2016b]和赵等人[Zhao，2017]的方法在信息集中进行两次用户检查。第一个集合，我们称之为(签入所有)集，包含每个用户的所有签入。第二组称为(签入日)集，包含按日组织的所有顺序POI签入。如图2所示，我们观察到（Check-ins All）集和（Check-ins Day）集中的POI之间的平均距离分别比所有POI（全局）集合中的对之间的平均距离小约2.98和5.74倍。这些结果表明，每个用户访问的POI都非常接近，并且在地理上相互影响。他们还显示，Instagram用户倾向于访问用户已经访问过的POI附近的POI。



**文本对POI的影响** 定性地分析了数据集中词汇与POI之间的关系。我们通过计算POI在文本内容中的Jaccard相似度来度量单词之间的相关性。我们首先删除停止词，并为每个单词构造一组POI。只有文本内容提到一个单词超过五次的POI才会包含在这个单词的POI集中。例如，当在POIs中 l1，l2和l3的文本内容中使用单词w1时，单词w1的POI集合是{l1，l2，l3}。我们为按频率排序的前2000个单词制作POI集，并计算所有单词对（Top 2000 Words）的平均Jaccard相似度值。我们还计算以下单词对的相似度值:(咖啡，咖啡馆），（棒球，洋基队）和（咖啡馆，棒球）。结果显示在图3中。正如预期的那样，（咖啡，咖啡馆）和（棒球，洋基队）具有比（咖啡馆，棒球）高得多的Jaccard相似性，因为前者对共享类似的语义上下文，因此具有更大的POI重叠。同样有趣的是，（咖啡馆，棒球）的Jaccard相似度远低于Top 2000 Words集中所有对的Jaccard相似度。这个结果证明了文本内容与POI相关，这促使我们在POI嵌入中使用文本内容。

**4 方法**

在本节中，我们描述了内容感知的分层POI嵌入模型CAPE，它不仅捕获POI的地理影响，还捕获来自文本内容的POI的特征。CAPE由签入上下文层和文本内容层组成.

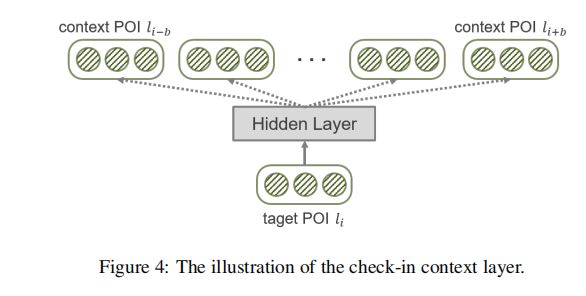
**4.1签入上下文层**

正如数据分析部分所解释的那样，每个用户访问的POI，特别是在用户签入上下文中连续访问的POI，在地理上是相互影响的。为了捕捉POI在地理上的影响，检入上下文层采用了Liu等人在研究中提出的基于跳格的POI嵌入模型[Liu等人,2016年b]。为了更好的理解签入上下文层，我们定义了一些数据概念如下。

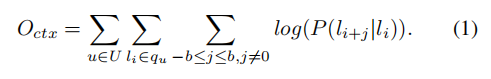
**定义1（签入数据）**每个签入数据d由用户u、POI l、文本内容x和签入时间t组成。我们将数据集D和登记数据d分别表示为D =（d1，d2，...）和d =（u，l，x，t）。

**定义2（用户签入信息）**用户签入信息是按签入时间排序的POI列表。用户u∈U的签入信息定义为Qu=(l1，l2，…)其中U是D中的用户集合。

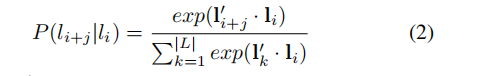
**定义3（目标和背景POI）**在用户登记序列中，在一个目标POI li之前和之后访问的b POI被定义为目标POI li的上下文POI。 b是预定义的上下文窗口大小。



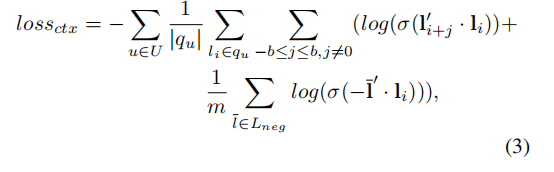
上下文层如图4所示。上下文层将用户的签入信息视为“句子”，将顺序中的每个POI作为“单词”使用word2vec[Mikolov等人，2013年]进行POI嵌入。据赵等人说。[赵等人，2017]，用户在同一天访问的POI之间的关系更密切。基于此，我们将用户在同一天访问的POI序列视为“句子”。 登记上下文层通过以下方式最大化目标函数Octx来训练POI嵌入向量



给定目标POI li，通过使用soft-max函数如下计算上下文POI li + j的概率P（li + j | li），



其中l和l’分别是POI l的输入向量和输出向量，而| L|数据集中的POI数。为了加快学习过程，我们应用了负抽样技术[Mikolov等人，2013年]，然后将检入上下文层lossctx的损失函数定义为：



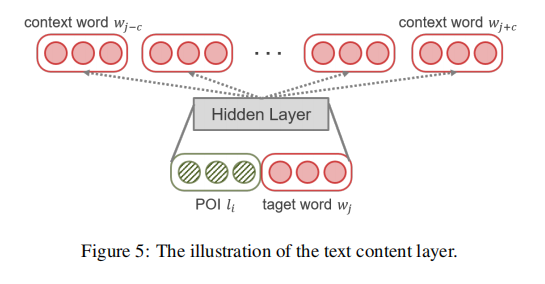
其中σ(·)是Sigmoid函数，Leng是m个负POI样本的集合，l’是负样本l的输出表示向量。

4.2文本内容层

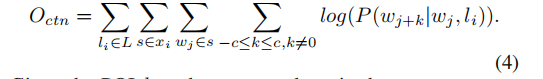
现有的POI嵌入模型只利用用户的签入信息捕获POI的地理影响，因而无法捕捉到POI的特性。Instagram帖子的文本内容提供了人们在每个POI上所做的活动的信息。为了从文本内容中捕捉POI的特征，我们提出了一个文本内容层。在描述该层之前，我们定义了另外两个概念，如下所示。

**定义4（文本内容）**每个文本内容si由单词组成，定义为si={w1，w2，.}，其中w是一个单词。每个POI li包含多个关于POI li的文本内容。关于POI li的多内容xi集定义为xi={s1，s2，.}

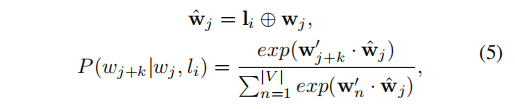
**定义5（目标和上下文词）**在文本内容si中，位于目标词wj之前和之后的每个c个词被定义为wj的上下文。c是文本内容的预定义上下文窗口大小。



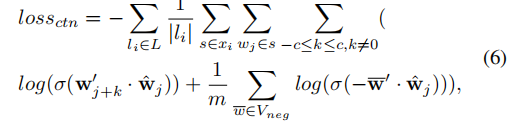
文本内容层如图5所示。我们首先将文本内容视为“句子”，并使用word2vec训练单词嵌入向量w[Mikolov等人，2013年]。然后，内容层将POI向量和字嵌入向量一起训练，最大化目标函数Octn。



给定文本内容中的POI li和目标词WJ，利用上下文词WJ+k的概率和计算目标函数。我们使用soft-max函数表示概率P（wj + k | wj，li），如下所示：



在⊕是级联函数的情况下，{V}是数据集中的字数。w和w’分别是soft-max层中的输入和输出字表示矢量。与公式3类似，我们使用负抽样技术制定内容层损失的损失函数，如下所示：

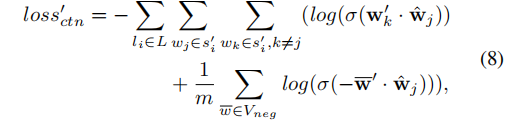


其中，Vneg是一组负采样m字，W’是负样本词w的输出向量。

在本文中，我们将与POI li高度相关的单词作为文本内容s’i进行样本，以过滤许多无意义的单词，如“自拍”和“跟随”。 我们计算了一个单词和一个POI之间的关联权重，它类似于TF-IDF，如下所示：



其中，TF(WJ，li)是LI文本内容中使用WJ的频率，DF(WJ)是文本内容中包含单词WJ的POI的数量。然后，我们对每个POI按关联权重排序的顶部k个单词进行采样。损失函数修改为：



其中si’是POI li的抽样词集。

最后，我们结合了签入上下文和文本内容层。我们的感知内容的POI嵌入模型CAPE的最终目标函数用超参数α表示如下：



CAPE同时学习POIS的地理影响和语义特征，使目标函数Oemb最大化。

**5实验评价**

为了定性地评估CAPE对后续POI推荐任务的影响，我们进行了实验评估。

**5.1 实验装置**

**连续POI推荐**

给定一个时间顺序的用户签入顺序，推荐模型使用第一个(n-1)POI作为输入，并预测第n个POI作为推荐POI。我们实现了以下POI推荐模型，利用用户签入序列。(1) **STELLAR** [Zhao et al., 2016], 基于张量因式分解的模型，该模型训练一对连续的POI来预测下一个POI；（2）**LSTM**，提出了一种基于LSTM信元的简单RNN模型；（3）**GRU**， 基于GRU单元的简单RNN模型；（4）**ST-RNN**[Liu et al., 2016a]，最先进的POI推荐模型采用RNN，并将两个检查点之间的时间间隔和两个POI之间的距离作为因子。

**基线**

我们将CAPE与以下四个基线POI嵌入模型进行了比较。（1）**Random** 一种简单的基线POI表示模型，它使用随机初始化向量，并被大多数使用用户签入序列数据的POI推荐研究所采用；（2）**SG** [Liu et al., 2016b], 提出了一种基于跳格的POI嵌入模型，对用户签入序列进行建模；（3）**POI2Vec** [Feng et al., 2017]，一种基于层次二叉树的利用空间位置信息的模型；（4）**Geo-Teaser** [Zhao et al., 2017], 利用各POI的地理影响和时间特性，提出了一种地时序列嵌入秩模型。当我们将每个POI嵌入模型应用于推荐模型时，我们测量性能的变化。

**数据集准备**

我们在**数据描述和分析**部分描述的数据集上评估我们的POI嵌入模型CAPE。为了将CAPE与现有的POI嵌入基线模型进行比较，我们将数据集划分为训练、验证和测试集，采用与刘氏等人[Liu等人， 2016a]研究方法相同的方法。最近的签入(占每个用户总签入的20%)用于测试集。较近期的签入(10%)用于验证集，其余的签入(70%)用于训练集。

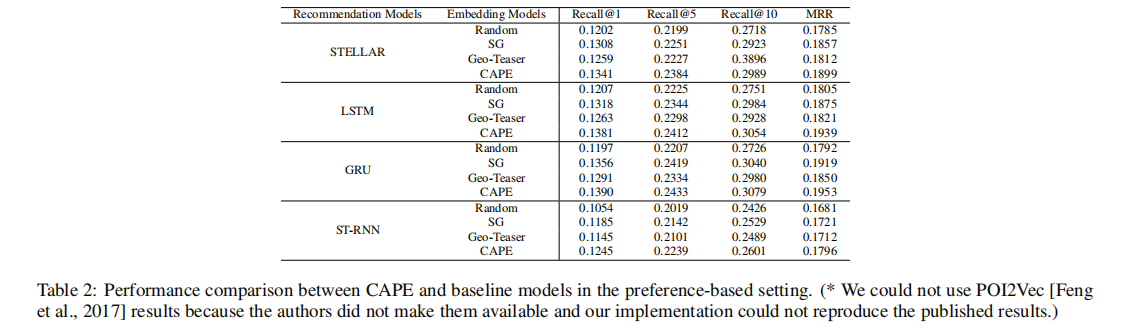
**指标**

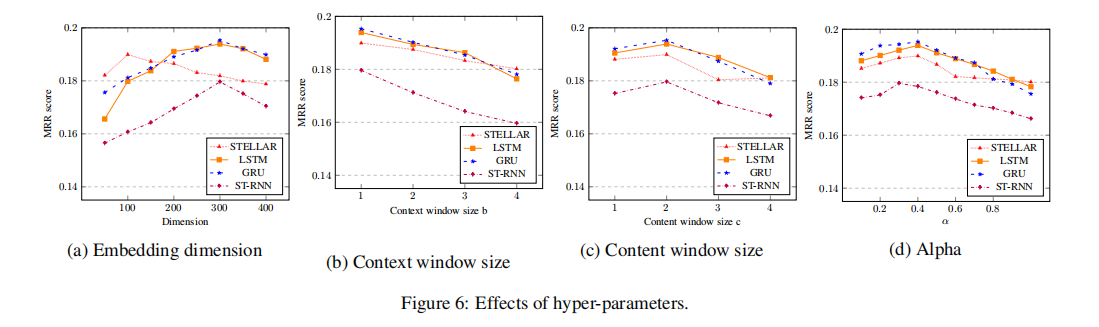
为了比较推荐模型的性能，我们采用了两种评价指标：Recoke@k和MeanEveryRank(MRR)。**Recoke@k**是一种常用的任务排序标准，当正确的后续POI位于顶部-k推荐的POI列表时，就会计算它。由于在我们的连续POI推荐任务中只有一个正确的答案，我们不使用Liu等人[Liu et al., 2016a]使用的F1-score @ k（或Precision @ k）。我们为k=1、5和10报告召回@k。**MRR**用于衡量排名列表的质量。当下一个正确的POI在推荐列表中被高分时，评估分数也很高。

**5.2评价结果**

**性能比较**

表2总结了使用指标Recovery@N和MRR评估的实验结果。结果表明，预训练的POI嵌入模型提高了推荐性能。当应用于推荐模型时，CAPE，SG和Geo-Teaser在MRR方面的表现优于随机的7.4％，4.3％和1.8％。此外，应用于推荐模型时，CAPE在召回@k和MRR度量方面优于POI嵌入模型。特别是，尽管CAPE不使用Geo-teaser使用的位置信息，但它的性能优于GeoTeaser。这些结果表明，文本内容为理解POI的特点提供了有用的信息。与Liu等人[Liu et al。，2016a]的研究报告的结果不同。利用连续登记时间间隔作为时间因素的ST-RNN的性能低于LSTM。在我们的Instagram数据集中，两个连续签入之间的平均时间间隔大约是Gowalla数据集的10倍。因此，无法正确构造作为ST-RNN输入的目标POI的上下文。当输入上下文不适当时，ST-RNN的状态仅由先前的状态计算。我们报告说，CAPE，SG-CWARP和Geo-Teaser分别在单台GTX Titan X GPU机器上进行了大约3小时40分钟，3小时和7小时的优化。



**超参数优化**

当我们改变CAPE的超参数时，我们观察到性能的变化。图6总结了结果。我们首先将POI嵌入维数的范围设置为50-450。然后，我们观察到从增加维度到增加50的性能变化。如图6a所示，当维度设置为300时，性能最好，但当维度设置为100时，STELLAR的性能最好。图6b和图6c分别显示了更改签入上下文和文本内容窗口大小时的性能变化。我们将范围设置为1-4，并在增加窗口大小1之后观察性能变化。当我们将上下文和内容窗口大小分别设置为1和2时，得到最佳性能。CAPE使用超参数α对线性组合的lossctx和lossctn进行优化。我们还研究了超参数α的影响。当超参数α为0.4时，除ST-RNN外，POI推荐模型的性能最好. 当超参数为0.3时，ST-RNN的性能最好。

**6 结论**

在本文中，我们介绍了我们新构建的大型数据集，并研究了在POI嵌入中使用文本内容的必要性。我们还提出了一种新的POI推荐层次POI嵌入模型-CAPE。CAPE从用户的文本内容和签入上下文中捕捉到POI的特点和地理影响。实验结果表明，CAPE对提高POI推荐性能有一定的促进作用。