上海交通大學

# 学生实习报告

**在校外做的需要在封面盖单位章**

实习单位: 丁家昕老师

实习时间：2021.7.1 至 2021.8.31

学院(系): 电子信息与电器工程学院

专　　业: 信息工程

学生姓名: 王 XX 学号: 5180XXXXXXXX

2021 年 9 月 16 日

实习报告主要内容包括实习目的与任务、实习单位、实习内容、实习收获等

实验单位是丁家昕老师的实验室，实验的任务是尝试利用不同方式完成轨迹数据的降维并计算命中率，我决定最主要是利用图神经网络完成任务，并使用 pca 降维的结果作为 baseline。

在此实验中，我们将 10000 个轨迹作为节点，并以轨迹之间的相似度建构出边（这也是一个实验点，要测试不同构建原则下的效果），因此本实习的重点在于学习图数据的处理，如何对每个节点的信息进行降维，对每个节点分别得到一个尽可能包含较多信息的向量（称为嵌入向量 embedding vector）。

图是一种由若干个结点(Node)及连接两个结点的边(Edge)所构成的图形，用于刻画不同结点之间的关系。图是一种非欧空间，我们常用的图像是欧式空间。传统的卷积神经网络在文本和图像领域有很好的效果，但是它仅能处理欧氏空间数据，所以针对图数据，要发展不同的理论和模型。

在做这个任务之前我先尝试阅读图神经网络的文献，但是由于当时我只有学习过最基本的机器学习和深度学习的概念，很多文献中提到的概念不是很理解，因此我尝试用最短的时间学习了深度学习，包括 deep learning 的原理、backpropagation 的推导、Convolutional Neural Network、Recurrent Neural Network，到深度学习在 NLP 上的发展，深度学习在 NLP 领域比较重点的三大突破分

别是：Word Embedding、NN/LSTM/GRU+Seq2Seq+Attention+Self-Attention 机制和 Contextual Word Embedding(Universal Sentence Embedding)，Word Embedding 解决了传统机器学习方法的特征稀疏问题，它通过把一个词对映到一个低维稠密的语义空间，从而使得相似的词可以共享上下文资讯，从而提升泛化能力。而且通过无监督的训练可以获得高质量的词向量(比如 Word2vec 和 Glove 等方法)，从而把这些语义知识迁移到资料较少的具体任务上。但是 Word Embedding 学到的是一个词的所有语义，比如 bank 可以是”银行”也可以是”水边。如果一定要用一个固定的向量来编码其语义，那么我们只能把这两个词的语义都编码进去，但是实际一个句子中只有一个语义是合理的，这显然是有问题的。

虽然 Word Embedding(比如 Word2vec 和 Glove 等方法)有它的缺点，但是我们可以将 word2vec 的思想用在图数据的处理中，也就是 node2vec。word2vec 任务主要是将人类语言符号转化为可输入到模型的数学符号（向量）。与之类似类似，拥有网络结构数据的图，通常也无法直接输入到模型中进行计算，这就需要我们用相类似的方法，将一个图所包含的信息尽可能的用向量（embedding vector）表示。

图数据其实非常常见，例如社交网络关系、分子结构、论文相互引用的关系网络等等，所以如何表达网络节点的特征就十分重要（嵌入向量）。表达好了节点的特征，我们就可以用它来做下游的分类、预测、聚类、可视化等等的任务，而本实习研究的重点就是

如何产生一个嵌入向量（embedding vector），以便能运用在下游的任务中。

在 node2vec 之后，由于深度学习的发展，图神经网络又有很大的进展，《Deep Learning on Graphs: A Survey》将现有应用于图的不同深度学习方法分为三个大类：半监督方法、无监督方法和近期进展。具体来说，半监督方法包括图神经网络（GNN）和图卷积网络（GCN 和 GAT），无监督方法主要包括图自编码器（GAE），近期进展包括图循环神经网络和图强化学习。

图卷积网络分为两种：基于谱域（基于图卷积定理的图神经网络）和基于空域的（基于聚合函数的图卷积网络）。在基于图卷积定理的图神经网络中，利用卷积定理，我们可以对谱空间的信号做乘法，再利用傅里叶逆变换将信号转换到原空间来实现图卷积，从而避免了图数据不满足平移不变性而造成的卷积定义困难问题。但最早提出的谱方法计算量太大且不具局部性。切比雪夫网络（ChebyNet）通过参数化卷积核实现局部性，同时降低参数复杂度和计算复杂度。GCN 是基于谱域的模型中代表性的一个，这个方法是在 ChebyNet 的基础上又进行了参数的近似所提出的，主要是来解决使用一阶近似简化计算的方法，提出了一种简单有效的层式传播方法。

用于图的 AE 来源于稀疏自编码器（Sparse Autoencoder，SAE）其基本思路是，将邻接矩阵或其变体作为节点的原始特征，从而将 AE 作为降维方法来学习低维节点表征。与上述自编码器

AE 不同，变分自编码器（VAE）是另一种将降维与生成模型结合的深度学习方法，VAE 首次在《Variational graph auto-encoders》中提出用于建模图数据，简称为 VGAE（此实验主要使用的图神经网络模型），其解码器是一个简单的线性乘积，至于均值和方差矩阵的编码器，作者采用 GCN。

了解整个深度学习和图神经网络的发展脉络和各个主要模型的思路之后，我开始阅读这些主要模型的提出论文，并尝试推导公式、代码復現，最后并着手构建针对本任务的模型，一开始先用 pca 跑出一个结果，作为 baseline，这一步很简单，花的时间少。接下来要进行图神经网络模型的选择和构建。最初尝试的是 LINE 模型，直接使用论文的代码来做修改，但是由于任务的目标不同，因此效果非常差，这是可预期的，但是总归是跑通了一个图网络模型。接着使用基于 GCN 的 VGAE 模型，好几次跑出来的结果都不合理，在经历很多天的尝试的检查，终于发现是一个代码的细节错误（在 topk.py 的错误，与模型无关），修改完这个错误之后，模型顺利的跑出结果，接着就是在此模型基础上进行参数和边建构方式的尝试，尝试不同的参数和边建构的方法（我们采用的是设定一个阈值，相似度大于该阈值的两个轨迹之间才会建立一条边，因此这个阈值也是实验的一个参数之一），在多天尝试之后（由于数据量大，因此是用租的服务器跑的，并且每个尝试都要跑 3 个多小时），模型的效果有些微进步，我画出参数和命中率的散点图，在论文中尝试分析此结果。最后就是学习怎么写论文，整理自己这两

个月的学习，写成综述作为 related work 和 algorithm 部分，并在 result 和 analysis 的部分写上结果的分析。

由于一开始的知识和经验都不足，因此前期的学习花了蛮长的时间，感谢丁老师给我很大的自由度安排学习和工作的进度，而且会在我遇到困难的时候给予协助和指导，感谢老师给我这个实习机会，过程中学习了很多，包含：理论知识、检查模型代码、文献阅读、论文写作，收获很大。

| | |
|---|---|
| 指导教师对学生实习情况的评价意见 | 王 XX 同学在实习期间完成了对轨迹表征学习的探索，完成了基础文献的阅读任务并进行细致总结，其后动手实践了图神经网络对于轨迹的表征学习，具有一定的工作量，较好地完成了暑期实习的预期目标。<br><br>**在校外做的 需要在此盖单位章**<br><br>指导教师（签名）：<br><br>2021 年 9 月 17 日 |