电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

学士学位论文

BACHELOR THESIS

论文题目　基于图神经网络预训练的在线信息传播流行度预测

学　　院　公共管理学院

专　　业　信息管理与信息系统

学　　号　2017120101002

作者姓名　佟计呈

指导教师　冯小东

摘　要

在互联网社交计数日渐发达的现代社会，在线社交平台吸引了数以亿计的用户参与其中，可以说人们的日常生活已经离不开在线社交平台了。人们为了从分享、浏览中收获精神物质上的满足而不断发布、获取信息。其中，微博短文本以其精炼、易读等特点快速收获大量用户，为了向更具有目标性的把特定的信息推送给特定的用户，越来越多的研究学者开始向在线社交平台信息流行度预测的领域靠拢。与此同时，此领域的研究对于舆情监控、个性推送等方面的应用具有重大意义，即个人、企业或政府能够根据预测的结果对相关热点话题进行及时回应。综上所示，对于在线信息流行度预测的研究正处于上升期。本文首先构建了基于图神经网络预训练模型，接着提取了数据中转发有向图自身图特征，最后在流行度预测中加入应用 Dropout 的深度神经网络模型，并进行模型对比：

(1)　建立图神经网络迁移学习模型并转发有向图特征。首先，本文的实验使用现有数据集，由微博转发链构建微博转发网络，并划分区间并提取转发图自身特征。然后通过自编码门控循环网络进行预训练，来预测自身转发网络的整体特征，之后将自编码门控循环网络的解码层删去，仅留下其前半的编码层。随后利用训练好的编码层更好的提取图度分布特征数据的深度信息，以更好地完成在线信息传播流行度预测。

(2)　利用深度学习预测。现有预测的研究多是采用人为特征选取，然而人为选择的特征难以代表原始数据，数据特征挖掘不够深入。本文在在线信息流行度预测的研究中应用深度神经网络技术，构建了相关深度神经网络模型，同时利用 Dropout 减少拟合。

关键词：图神经网络，预训练，迁移学习，深度神经网络，流行度预测

ABSTRACT

In the modern society where Internet social counting is increasingly developed, online social platforms have attracted hundreds of millions of users to participate in it. People keep posting and acquiring information in order to reap spiritual and material satisfaction from sharing and browsing. In order to target specific information to specific users, more and more research scholars are approaching the field of online social platform information popularity prediction. As shown above, the research on online information popularity prediction is on the rise. In this paper, we first builds a pre-training model based on a graph neural network, then extracts the characteristics of the forwarded directed graph itself from the data, and finally adds a deep neural network model using Dropout to the popularity prediction, and compares the models:

(1) Constructing a graph neural network migration learning model and forwarding directed graph features. First, the experiments in this paper use the existing dataset to construct a microblog forwarding network from microblog forwarding chains, and divide the intervals and extract the forwarding graph own features. Then the self-coding gated recurrent network is pre-trained to predict the overall features of its own retweet network, after which the decoding layer of the self-coding gated recurrent network is deleted, leaving only its first half of the coding layer. The trained coding layer is then used to better extract the depth information of the graph degree distribution feature data to better complete the online information dissemination popularity prediction.

(2) Building deep neural network models. Most of the existing studies on microblog popularity prediction use human feature selection, however, the human selected features are not representative of the original data and the data feature mining is not deep enough. In this paper, we applied deep neural network technology to construct a deep neural network model for online information popularity prediction, and used Dropout to improve the model in order to reduce the model overfitting problem. The experimental results show that the deep neural network model has good performance in predicting popularity.

Keywords: Graph Neural Network, Pre-training, Migration Learning, Deep Neural Network, Prevalence Prediction

目 录

第一章 绪论

1.1 研究现状及发展态势

社交网络不仅吸引了众多的使用者参与到聊天讨论当中，而且其中相关推送的流行度预测也带来了许多问题值得有研究。在线内容流行性研究往往从是否流行、流行等级、具体流行度预测这三个角度出发。

流行等级预测则是利用信息自身特征进行，的基于二分类的多分类的过程。Hong 等用预 Twitter 帖子的转发次数的间隔作为目标，通过逻辑回归作为分类器进行预测，最终以 99%的正确率预测了哪些信息将被转发一万次以上[1]。

本文主要是进行具体的流行度预测，根据研究进行的角度，在线信息流行度预测的研究可以分为：在流行度宏观预测和流行度微观预测[2]。

在宏观层面，研究多通过广泛收集流行度信息，获得统计层面的规则与规律，而非着眼于某一特定用户。在宏观层面的信息流行度预测也取得可观的成果，但这种研究在跨平台的应用中表现出较多不足之处，是的最终的预测出现偏差。

在微观层面，可分为发布前预测和发布预测，主要是通过分析影响其流行度的因素进行预测。

对于根据发布前影响因素预测，因为影响信息流行度的因素过多，故其预测值经常存在大量偏差，最终导致准确数值难以预测[3]。所以此方向的大部分研究是进行流行度的分类，即按不同流行程度分类，预测某一条信息处于哪一层级。虽然这种预测研究可以获得较为准确的预测，满足一定范围内研究的需要，但还存在着一定的局限性——一定程度上忽视了整体信息传播的大趋势，仅仅过分着眼于流行度属性的预测。

至于内容发布后流行度预测的研究，大范围使用目标发布初期点击、转发等属性，以更长的数据获取时段换取更好的预测性能，进而实现准确率的提升。在信息发布后开展的预测工作，多是基于信息属性的预测，按照人为经验，提取其相关发布者、转发者、时间的等特征，利用回归或分类问题的预测模型去拟合流行度预测问题[4]。然而，上述人工定义选取特征的方法往往需要较强的领域经验，且具有不确定性，耗时耗力。

随着硬件计算能力的提升，流行度预测领域渐渐应用了众多深度学习的方法，模型从大量信息中自动提取特征，进而预测在线信息流行度。但深度学习想要获得较好的预测结果，就需要大量的数据，而现实中这些大量数据往往难以获得。

1.2 选题依据及意义

本文选取在中国具有广泛受众群体的社交网络平台——微博，作为研究数据的来源。微博信息可以通过信息发布者的粉丝来进行信息的广泛、快速传播，因此根据信息主题的关注程度，成千上万的微博浏览者可以同时查看其他用户分享的信息。真是由于微博的粉丝、浏览转发功能，被发布的信息才能如此迅速的得到广泛的传播，用户为了分享浏览到的推送，而使用此功能转发到自己的账户上。综上，如果微博发布者拥有广泛的粉丝基础，他发布的消息则会在一一次被转发与转发中被快速的传播开来，甚至呈几何式增持。

但是，上述方法都存在一定的不足之处。是否流行和流行等级预测均致力于探究话题最终的流行程度，而难以描述其整体流行趋势；具体流行度预测是从具体数值预测角度出发的预测方法，可以做短期流行度预测，但在内容长期流行趋势预测上表现欠佳[5]。所以，在这一领域预测方法的研究还并不完善。

1.3 课题研究内容

首先进行流行度预测领域文章的越多并进行综述，研究并提出用于在线信息传播流行度预测的图神经网络预训练模型，随后进一步开展在线信息流行度预测实证研究。本文实验数据来自于微博，以微博转发网络预测转发量作为标准来衡量微博流行度。使用 Python 中 PyTorch 框架，利用 networkx、keras 等 Python 工具包，通过对微博数据处理、构建转发网络、提取转发网络自身特征并构建图神经网络预训练模型、基于预训练的图神经网络进行微博流行度预测等问题的研究，实现在线信息传播流行度的准确预测。

本文将基于真实的社交网络，通过构建微博转发网络、提取并利用转发网络自身特征预训练图神经网络，对信息流行度预测开展实证研究，旨在准确高效地预测信息流行度。

(1) 微博数据的处理：本文以大量的微博随机用户公开样本中获取数据，数据以某一用户的某条微博为单位，包括该微博编号（一条微博无论如何转载仅有一个编号）、作为根节点的该用户的编号、根节点用户发布或转发该微博的时间、其他用户从根节点用户下转发该微博的数目、从根节点用户下转发该微博的其他用户编号及转发时间、24 小时内根节点用户下该微博转发数目的增长量。

(2) 构建转发网络：通过第一步获得的微博数据，构建微博转发网络，即以某条微博为单位，由最后的转发用户追溯其被转发用户，在追溯此用户转发微博来源的上一层被转发用户，循环往复，直至微博的最初发布用户。

(3) 构建图神经网络预训练模型：通过 Python 中 PyTorch 框架，利用

networkx、keras 等 Python 工具包，首先提取转发网络自身特征（如各个节点的度和图的中心势），之后利用图神经网络拟合特征，进行图神经网络预训练，最后将完成预训练的网络作为流行度预测网络的前 k 层，然后重新训练后面 n-k 层。

(4) 基于预训练的图神经网络进行微博流行度预测：利用真实微博转发数据进行实验分析，将提出的预测模型与现有方法的应用效果进行比较，最终确定一种对在线信息传播流行度预测准确率最高的模型。

1.4 论文特色或创新点

(1) 构建转发网络，并应用图神经网络预测在线信息传播流行度：因为图数据包含大量显性及隐性信息、特征，可以模拟多种任务场景，能更好地胜任社交网络中的预测任务

(2) 利用转发网络自身特征预训练图神经网络：仅仅利用收集到的真实微博转发数据，我们很难训练出一个性能较好的模型，容易产生过拟合。本文利用转发网络自身特征预训练图神经网络，这样更容易获得理想的结果，并减少数据过拟合的可能性。

1.5 论文章节安排

本文分五章对相关工作进行了阐述，各章节的内容安排如下：

第一章：对流行度预测研究领域的背景进行描述，并说明了此研究的必要性。随后通过这项研究目前的研究现状和所面临的挑战，进而引出本文的研究方法总述。

第二章：分析了文章所应用到的模型及理论，包括：层感知器（MLP）、门控循环网络（GRU）、前向后向门控循环网络（BI-GRU）、自编码门控循环网络（GRU AutoEncoder）、自注意前向后向门控循环网络（GRU SelfAttention）、网络预训练和在线信息传播流行度，最后说明了本文实验中分类效果评估指标。

第三章：本章主要分为两个部分。本章的第一部分介绍了实验的整体流程，随后描述了微博转发链的数据集及预处理，然后进行数据的标准化并进行预训练。第二部分描述为了减少过拟合而使用的 Dropout 方法，最后，进行实验并确定最佳网络结构。

第四章：本章主要为实验结果分析。将不同模型之间的预测效果，和相同模型预训练前和预训练后的预测效果进行对比，得出实验结论。

第五章：总结与展望。本章对课题所完成的研究进行整理概述，针对本研究进行中发现的不足之处进行讨论，并对今后类似领域的研究进行展望。

第二章 相关模型及理论

2.1 神经网络模型

2.1.1 多层感知器（MLP）

MLP 的前身，单层感知器是神经网络中最为简单的一种，它主要由输入、输出两个部分，而且这两层之间为直接连接。但是单个感知器存在许多缺陷，面对分线性模型往往束手无策，虽然可以将多个单层感知器进行前后连接，来解决一部分的非线性问题，但显而易见，这样的组合模型仍然为线性分类器。其实只要在单层感知器中稍作改变——增加中间的隐藏层，就能解决许多问题，也就是说 MLP 是单层感知器的推广[6]。

MLP 神经网络再连接上同单层感知机类似，为全连接形式，即输出层众多每一个节点都连接隐藏层的所有节点，隐藏层的每一个节点又连接输出层的每一个节点，如图 2-1 所示：

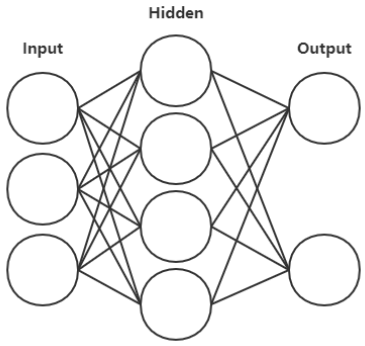

图 2-1 多层感知器结构图

为了避免单层感知机的线性模型缺陷，以增加非线性激活函数的方式给 MLP 增加了对非线性模型的拟合能力。激活函数一般需要有这几种特点：首先保证可微性，以便于神经网络使用梯度下降方法进行优化，其次要保重单调性，需要保证单层网络为凸函数以便于优化，所以需要激活函数保持单调性[7]。

激活函数常使用 Sigmoid 、Tanh 、Relu 等函数，本文主要采用 Sigmoid 函数。Sigmoid 非线性激活函数可以将正负无穷内的数值“压缩”至 0，1 之间，其计算公式为：

$$f(x) = \frac{1}{1+e^{-x}} \tag{2-1}$$

其中，当进行数据输入时，会为每一个输入数据添加一个截距项b。但是根据

每层网络的定义的不同，他们学习的方法也可能大相径庭。我们多通过 BP 传播算法来训练多层感知机，使其收敛。即多层感知机模型会从每次预测偏差中学习“知识”，微调模型参数。MLP 中输入层、隐藏层和输出层之间都存在大量可学习的参数，学习的目的就是为了寻找能以某一类型输入准确预测输出的参数。 MLP 通常采用随机梯度下降的优化策略，求解各个层之间参数的最优解的问题：首先将全部的参数进行随机初始化，之后通过循环进行参数的更新，不断通过目前参数的梯度方将来更新参数，以朝更小的损失值前进，最终达到准确预测的训练目的。

2.1.2 门控循环网络（GRU）

门限循环单元（Gated Recurrent Unit， GRU）神经网络改进于长短型记忆神经网络（LSTM）的模型，由其改进而来，而 LSTM 神经网络又是循环神经网络的一种，是其变形模型，GRU 神经网络和 LSTM 神经网络因为能够减少循环神经网络中所面临的长期依赖的问题[8]。虽然有这样的优势，但 LSTM 神经网络模型十分复杂，计算复杂度较高，也就是说面对同等规模的数据，LSTM 需要更长的时间已完成训练，更多的算力以支持其优化。GRU 神经网络对 LSTM 神经网络的这些缺陷进行了改善。如 LSTM 神经网络中为了遗忘或加强某些信息的各种“门”模块——输入门、输出门和遗忘门，LSTM 神经网络通过这些模块对输入进行调整，时模型中最为重要的部分。

GRU 神经网络对 LSTM 的改进就是从这些最重要的模块入手，它简化了 LSTM 神经网络中这些“门”的设定，即利用更新门来取缔遗忘门和输入门，进而将循环模块从三个变为了两个，考虑到神经单元的循环次数，GRU 神经网络极大的减少了模型的复制程度，提高了模型的泛化性能[9]。整体的 GRU 神经网络结构如图 2-2 所示。

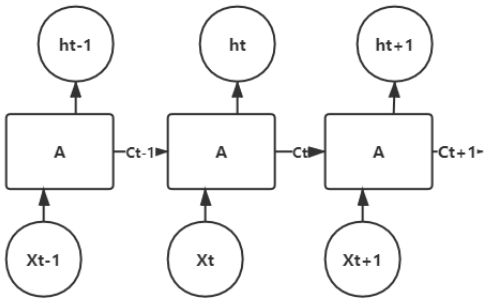

图 2-2 门控循环网络结构图

在模型的组成模式上，GRU 神经网络和其他循环神经网络类似，都是由重复的循环神经单元前后相连而成。但是不同于传统循环神经网络只能使用简单的激

活函数，GRU 神经网络可以使用门限模块作为其激活神经元，能拟合更加复杂的模型。GRU 神经网络中神经元的结构如图 2-3。

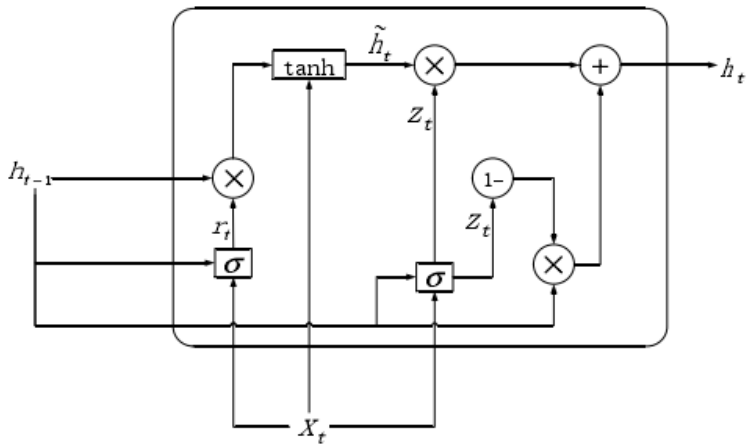

图 2-3 GRU 神经网络神经元结构

GRU 神经网络的神经元使用数学公式表达为：

$$z_t = \sigma\left(W_z \cdot \left[h_{t-1},\ x_t\right]\right) \tag{2-2}$$

$$r_t = sigmoid\left(W_r \cdot \left[h_{t-1},\ x_t\right]\right) \tag{2-3}$$

$$\widetilde{h_t} = \tanh\left(W \cdot \left[r_t * h_{t-1},\ x_t\right]\right) \tag{2-4}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \tag{2-5}$$

z_t表示更新门（Update Gate），r_t表示重置门（Reset Gate），h_{t-1}表示上一个神经元的输出，x_t表示本次神经元的输入，W_z表示更新门的权重，W_r表示重置门的权重，sigmoid表示 sigmoid 函数，经过实验发现 GRU 不仅能取得 同 LSTM 相差无几的成绩，还更容易收敛，训练所需的时间和算力都更少[10]。

2.1.3 前向后向门控循环网络（Bi-GRU）

Bi-GRU 神经网络的设计为能够同时进行前向传播和后向传播，其网络结构如图 2-4，其中 h 意为的是当前时刻输入的信息在与之相对应的正向 GRU 单元的迭代优化后的结果，那么，与之相对应的 h 则意为当前时刻输入在与之相对应的反向 GRU 单元的迭代优化后的结果，t 时刻的输出最终所对应的就是，h 该种情况下，时刻所对应的输出就不仅仅有与之相对应的正向的历史信息，还有该时刻与之相对应的反向的历史信息，因此，就可以很好的去表示文本的上下文信息[11]。

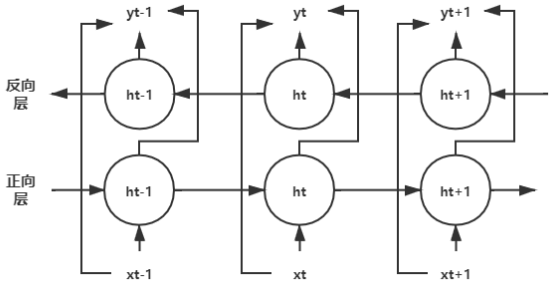

图 2-4 前向后向门控循环网络结构图

2.1.4 自编码门控循环网络（GRU AutoEncoder）

自编码器时模拟输入等于输出映射的多层神经网络，如图 2-5。从左至右第一、二层即使模型编码（encode）的过程，第三层为编码所提取出来的输入深度特征，第四、五层则相当于是一个解码的过程。我们首先向自编码器中输入非监督数据，模型就会以自身为目标进行迭代学习，我们最终希望解码层能得到一个和输入类似数据。如果得到的输出和输入相同，我们就可以认为中间层得到的隐含特征能够代表模型的输入。模型利用梯度下降来调整其中的各种权重，最终达到最小结构误差的目的，此时能够得到编码，从第三层（隐含层）到第四、五层（输出层）是正向运算，所以说自编码的意思就是由自编码器提取出来的中间特征可以最大程度的代表输入的各种特征，能以最小的损失找到输入数据的另一种表达方式[12]。

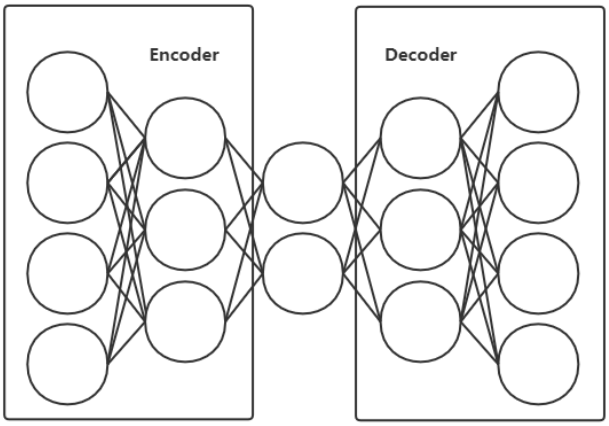

图 2-5 自编码门控循环网络

2.1.5 自注意前向后向门控循环网络（GRU SelfAttention）

自注意力（self-attention）机制是一种优化改经的注意力模型，加强了模型对内部数据特征的提取，而减少了对输入信息的依靠，在内部关系等研究层面给，自

注意力机制被广泛的应用，得到了令人满意的成绩[10]，如图 2-6。如果采用 LSTM 或 RNN 利用依赖特征 Target，则需要按照输入的顺序依次进行预测，对于处于模型较远端的特征，模型需要更长的时间、更多的算力完成时间步的累计才能学习二者的关系，距离越远，特征捕获越困难。但是利用自注意力模型在计算时可以无视两目标之间的距离，减少间隔造成的影响，更加等效的利用全部数据。

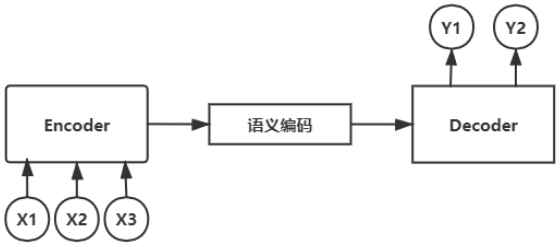

图 2-6 自注意前向后向门控循环网络结构图

2.2 迁移学习

在某些领域中，训练模型所需要的大量数据也许难以获得，也就是如果仅用便于得到的数据作为输入，很有可能无法得到预期的、理想的结果。而且，在某一任务中已经训练好的模型，在目标任务进行细微调整后，整个模型就难以被简单的套用。为了解决以上两种问题，迁移学习应运而生（Transfer Learning）[13]。

迁移学习之后，我们可以训练好的模型采用微调、冻结部分层等方法进行模型迁移，甚至可以直接使用训练好的模型到目标任务中。预训练模型可以花费较小的代价通过众多易获得的数据学习一部分的通用经验。即使这部分数据对于本任务来说有一定的偏差，但偏差主要体现在输出种类的差异上，在输入上相差无几。这种关系不太复杂的网络，节点和边的规模相对较小，因此可以将预训练学习中得到的广范通用的知识应用到这样结构的网络当中，而且往往会得到较好的结果。利用迁移学习，可以将训练好的某一任务的模型进行微调而应用到零一类似任务中。

对模型进行完善的“泛用化”也是这种技术预训练技术的另外一个优点，过拟合问题往往是深度学习网络无法避免的问题，也就是说对于未曾训练过的真实情况下的真实数据有更好的预测能力。原因在于不同的数据类型，不同的噪声让迁移学习模型鲁棒性更强，可以更好的训练其基础规则 [13]。

神经网络模型的规模和所需数据量的规模之间呈现一种正相关的关系。对于特定的问题，模型必须足够大，以便得到数据和预测结果之间的关系。问题越是复杂，需要的层数也就越深，模型需要训练的参数和计算量也会水涨船高。借用迁移学习技术，能够更好的解决类似的问题。

但同时，迁移学习也存在一些缺点：(1)迁移学习模型大、参数多、设定完成的结构难以变更、灵活性差，计算量大，实际应用受限较大。(2)不同任务，如分类和回归之间目标函数和中间层差距比较大。(3)虽然可以在小范围的差异可以通过预训练微调而进行泛化，但是当模型间的差异超过了这个限度，迁移的效果就不明显了[13]

2.3 在线信息传播流行度

流行度描了某一话题在用户范围内地讨论激烈程度，与其自身内容和发布人的特殊属性相关。某一话题的浏览、点赞、评论、转发等等形式都可以算作是本话题流行度程度的描述。因此，评价一个话题的流行度，常常采用该话题浏览量、转发次数、评论数量等作为评判指标。多种多样的因素都会对话题流行度造成影响：

(1) 内部因素：包括研究话题本身的内容、属性等影响其被论激烈程度的因素。如：①内容本身所包括的事件、情绪等信息，都会在用户阅读兴趣等方面吸引浏览、点赞和转发；②话题长度、文采、包含图片等直接影响浏览者第一感官的因素。包括以上两部分的内部因素会直接对浏览用户的观看、分享意愿产生影响，从而直接影响此话题的流行度。

(2) 外部因素：指包括话题本身内容以外其他因素，比如在线社交平台以外的舆论、社会环境、观念等对于这类话题流行度的影响。还比如话题发布者自身的影响力，最直接的体现就是话题发布者的粉丝数量。当一个具有庞大粉丝的用户发布某一话题时，几乎没有例外的都会产生大量的讨论和转发，从而带来极高的话题流行度，甚至可以一定程度上忽略话题本身的内容。

2.4 预测效果评价指标

均方误差（Mean-Square Error、MSE）在统计中，样本量的均方误差（用于估计未观测量的过程）测量误差平方的平均值，即估计值之间的平均平方差。MSE是评估预测值与实际值的质量的度量，它总是非负的，接近零的值更好。对于无偏估计，MSE 是估计量的方差。与方差一样，MSE 拥有与估计数量的平方相同的度量单位。均方误差有下式定义：

$$\mathrm{MSE}=\frac{1}{n}\sum_{I=1}^{n}(y_i-x_i)^2 \tag{2-6}$$

其中，n代表样本总个数，只代表第i个样本标签，y_i代表第i个样本通过模型后的对应输出。

2.6 本章小结

本章阐述介绍了本流行度预测方法所需要的模型及其基础理论，便于后面对算法进行展开分析。首先介绍了了多层感知器、门控循环网络、前向后向门控循环网络、自编码门控循环网络、自注意前向后向门控循环网络的定义和计算。接着，介绍了迁移学习的相关概念，说明了迁移学习使用的方法和其优缺点。然后，介绍了在新信息传播流行度。最后，对本文使用的预测效果评价指标做简单描述。

第三章 基于图神经网络预训练的在线信息传播流行度预测

3.1 整体流程

本文主要研究的是基于图神经网络预训练的在线信息传播流行度预测。本文提出的图神经网络预训练模型将采用利用自编码门控循环网络强大的特征提取能力来提取特征，将在线信息传播的各项图特征数据的深度信息提取出来，然后放到神经网络中进行预测。

基于图神经网络预训练的在线信息传播流行度预测的整体流程如图 3-1 所示。该方法划分为 3 部分：数据预处理；自编码门控循环网络预训练，提取多模在线信息传播图特征；流行度预测，实验结果分析。

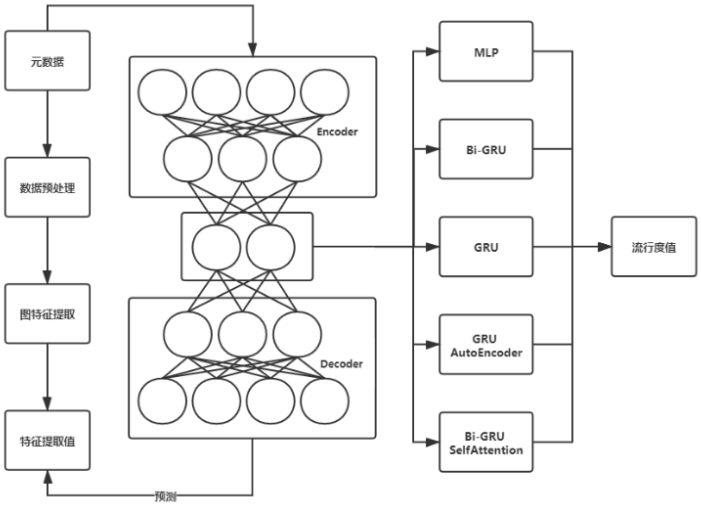

图 3-1 基于图神经网络预训练的在线信息传播流行度预测结构图

首先，本文的实验使用现有数据集，由微博转发链构建微博转发网络，并按时间划分为多个区间，在同一区间内提取转发图的度分布和整体特征。我们对预处理后的微博转发网络数据集进行预训练，将图的以时间划分区间的度分布特征输入自编码门控循环网络，来预测自身转发网络的整体特征，之后将自编码门控循环网络的解码层删去，仅留下其前半的编码层。然后，将用于训练、测试、验证的以时间划分区间的度分布特征输入此前得到的编码层，就能更好的提取图度分布特征数据的深度信息，以更好地完成在线信息传播流行度预测。

之后分别将未经特征提取的原始时间划分度分布和从其中提取出的特征输入多层感知器（MLP）、门控循环网络（GRU）、前向后向门控循环网络（Bi-GRU）、

自编码门控循环网络（GRU AutoEncoder）、自注意前向后向门控循环网络（Bi-GRU SelfAttention）模型中，预测在线信息传播流行度。比较各个神经网络的预测准确率，确定更适合在线信息传播流行度预测的神经网络，并得出图神经网络预训练是否会增加预测准确度的结论。

3.2 微博转发链的数据及预处理

3.2.1 实验数据集

本节将详细介绍实验所使用的数据集。我们使用现有的新浪微博数据集，数据集包括中国社交网络的代表之一——新浪微博 2016 年 6 月 1 日一天内发布的全部微博，以及此条微博从发布时起 24h 内的转发信息，同时将发布后 24h 的转发数作为最终流行度。

该数据集包含 42，183 条带标签的数据，每一条数据是一条微博信息在新浪微博网站上被转发的相关信息，表 3-1 展示了新浪微博数据集的 6 个属性及其描述说明。

捕获转发数为 10 到 1000 条之间的微博，并剔除那些未来 24 小时转发量没有增加的微博。为了减少微博发布时间对于转发的影响，本研究剔除 18：00 至次日 8：00 发布的的微博。最终将剩余的新浪微博数据中的 70%作为训练集，15%作为验证集，15%作为测试集，最终训练集 29，529 条，验证集和测试集各 6，327 条。

表 3-1 微博数据及属性及其描述

属性	描述
message_id	收集到的每条微博信息的 ID，如 65877
user_root_id	发布该条微博的用户 ID，即为本转发链或转发网络的根用户节点，如 2117059
publish_time	根用户节点发布该条微博的时间，如 1464766217
retweet_number	给微博在 2016 年 6 月 1 日一天内的转发量
user_a，userb:time	在某一时间上，由用户 b 转发用户 a 的该条微博，如 2117059，1167998:873
increment retweets in 24hours	该条微博在未来 24 小时内的转发增量，即本实验的预测目标

3.2.2 微博转发链的预处理

根据新浪微博数据集的数据格式，将全部微博转发链添加到一有向网络图中，并按照预定的时间区间和预定的度分布区间划分转发链，最终构成用于接下来预训练模型的 n * time_interval * degree_interval 的度分布列表。本文的微博转发链的预处理算法流程如下：

算法 3-1 微博转发链的预处理

输入：graphs：微博 ID 及转发链。

labels：一条微博 24 小时增加转发量。

time_interval：按时间划分区间的间隔。

degree_interval：按度分布划分区间的间隔。

算法过程：

for id， graph in graphs:

将该条微博在当天的转发量的加入转发量列表。

获取 graph 中的全部节点。

将全部节点构建为列表。

建立有向图，并添加全部节点列表，建立图节点。

将节点列表按转发时间降序排列。

for walk in graph:

将转发链中的每次转发按时间离散化到 time_interval 各的区间中。

按转发链建立有向图的边。

将按时间划分后的转发链，其度分布再次离散化到 degree_interval 各区间中，得到 time_interval * degree_interval 的数据。

输出：微博 ID，度分布列表（n * time_interval * degree_interval），微博 24 小时增加转发量列表，微博转发有向图。

利用得到微博转发有向图，获取图的边总数、节点总数、图密度、图的点度中心势、中介中心势和接近中心势。其中点度中心势用来刻画一个图在多大程度上表现出向某个节点集中的趋势，可用公式表示为：

$$C = \frac{\sum_{i=1}^{N}[C_{max} - C(n_i)]}{max\sum_{i=1}^{N}[C_{max} - C(n_i)]} \tag{3-1}$$

拥有最高中介中心度的节点与其他节点的中介中心度的差距即为中介中心势，可用公式表示为：

$$C_B = \frac{\sum_{i=1}^{N}[C_{Bmax} - C_B(n_i)]}{(N-1)^2(N-2)} \tag{3-2}$$

接近中心是反映在网络中某一节点与其他节点之间的接近程度，可用公式表示为：

$$C_C = \frac{\sum_{i=1}^{N}[C_{Cmax} - C_C(n_i)]}{(N-1)(N-2)}(2N-3) \tag{3-3}$$

3.2.3 数据标准化处理

为了将数据保持在某一区间内，而进行数据标准化，实际上就是按照某一方法的数学共识对数据实行缩放。对于量纲不同的数据，标准化的目的就是为了将它们转化为单纯的数值，忽视其原本携带的量纲，进而可以进行不同数据之间的比较和运算。本文通过 Z 得分法对以上获得的六项微博转发图特征进行标准化处理。则 Z 得分法标准化处理公式如下：

$$S_X = \frac{X - \bar{X}}{S} \tag{3-4}$$

3.2.4 自编码门控循环网络预训练

表 3-2 自编码器预训练结构图

Encoder	
INPUT	10
GRU1	40
GRU2	20
Dense	10
Decoder	
GRU1	20
GRU2	40
Dense	6

再利用之前的得到的度分布预测图特征，进而训练自编码器。最后再通过训练后的自编码器提取度分布深度特征，并作为随后对于微博 24 小时转发增量的预测自变量。本文的自编码器预训练算法流程如下：

算法 2 自编码器预训练

输入：nx_G：微博转发有向图。

degree_x：由微博转发链预处理得到的度分布列表。

time_interval：按时间划分区间的间隔。

自编码器神经网络的初始参数。

算法过程：

```
for graph in nx_G:
   for i in range(0,  time_interval):
```

获取有向图在相应时间间隔区间内的边总数。

获取节点总数。

获取图的密度。

获取图的点度中心势。

获取图的中介中心势。

获取图的接近中心势。

构建预训练图特征列表 degree_pre_x，并按时间间隔填入以上六项图特征，得到 time_interval * 6 列表。

确定自编码器编码层及解码层结构，并对网络参数进行初始化。

确定优化器的主要参数。

利用 degree_x 预测 degree_pre_x 训练自编码器

自编码器训练完成

利用训练后的自编码器提取度分布 degree_x 深度特征，并赋值为 x

输出：度分布深度特征 x

3.3 神经网络结构

门控循环网络（GRU）

本文的门控循环网络框架包括两个门控循环单元层、一个 Flatten 层和两个全连接层。门控循环单元由 Tanh 激活，全连接层由 Relu 激活，使用 MSE 作为损失函数，如表 3-3 所示：

表 3-3 门控循环网络结构图

GRU	
GRU1	6， 20
GRU2	6， 20
Flatten	120
Dense	16
Dense	1

自注意前向后向门控循环网络（Bi-GRU SelfAttention）

本文的自注意前向后向门控循环网络框架包括两个前向后向门控循环单元层、一个自注意层，一个 Flatten 层和三个全连接层。自注意层由 Sigmoid 激活，全连接层由 Relu 激活，使用 MSE 作为损失函数，如表 3-4 所示：

表 3-4 自注意前向后向门控循环网络结构图

Bi-GRU SelfAttention	
GRU1	6， 20
GRU2	6， 20
SeqSelfAttention	6，20
Flatten	120
Dense	16
Dense	4
Dense	1

前向后向门控循环网络（Bi-GRU）

本文的前向后向门控循环网络框架包括两个门控循环单元层、一个 Flatten 层和两个全连接层，全连接层由 Relu 激活，使用 MSE 作为损失函数，如表 3-5 所示：

表 3-5 前向后向门控循环网络结构图

Bi-GRU	
GRU1	6， 20
GRU2	6， 20
Flatten	120
Dense	16
Dense	1

多层感知器 MLP

本文的门控循环网络框架包括一个 Flatten 层和三个全连接层，全连接层由 Relu 激活，使用 MSE 作为损失函数，如表 3-6 所示：

表 3-6 前向后向门控循环网络结构图

MLP	
Flatten	120
Dense	32
Dense	8
Dense	1

自注意门控循环网络（GRU AutoEncoder)）

本文的门控循环网络框架中编码层和解码层各包括两个门控循环单元层和一个全连接层。门控循环单元由 Tanh 激活，全连接层由 Relu 激活，使用 MSE 作为损失函数。前面的门控循环单元提取相对特定的特征，而全连接层提取相对抽象的特征，并将所有的特征融合起来。度分布深度特征。

经训练好的编码层后，再将输出结果输入表示层。表示层有两个全连接层 构成，由 Relu 激活，如表 3-7 所示：

表 3-7 自注意门控循环网络结构图

Encoder	
GRU1	40
GRU2	20
DENSE	20
Decoder	
GRU1	20
GRU2	40
DENSE	10
Representation	
Dense	16
Dense	1

3.4 基于深度神经网络模型的流行度研究

深度神经网络通多大量的隐藏层，从原始数据中提取了更加复杂、难以理解的各种参数。这样的作法虽然使深度神经网络的泛化性和鲁棒性更强，更能做出准确

的预测，但同时也仍然存在着一些问题：

(1) 深度神经网络为了能够提取信息更加深层的特征，而相比传统神经网络增加大量的参数，最终导致其不容易优化收敛。显然，这样大量的参数也就造成了其学习速度的缓慢，并且需要花费大量的计算能力，同时模型过拟合的现象也更容易出现。

(2) 模型训练时长增长快。深度神经网络中怎加一个节点，就会增加大量节点之间的连接，参数也随之增加，因此学习所花费的时间也急剧增加。

在本研究中，针对深度神经网络所面临的以上问题，通过 Dropout 随机消除全连接层节点的方式预防模型过拟合。如图 3-2 所示。

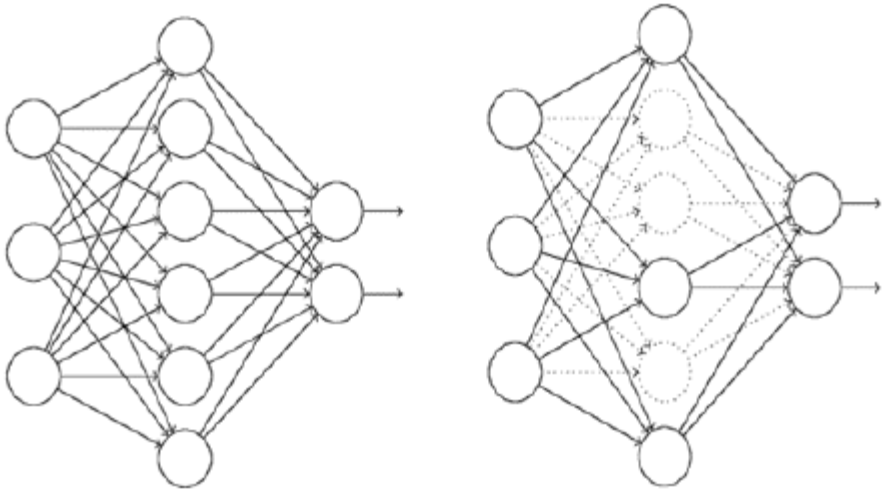

图 3-2 深度神经网络 Dropout 模型

Dropout 通过在深度神经网络的前向传播中，按照一定概率随机的舍弃一定数量的全连接层节点以及和这些节点相连的全部的边，如图 3-2 中由实线变为虚线的节点和边，也就是在 Dropout 过程中被随机删去的部分。在结束了前向传播过程之后，便进入到误差的后向传播之中，参数的更新自然就只发生在违背 Dropout 过程删去的节点之中。这样的作法让整个深度神经网络在训练的过程中接收了更多种类的噪声，减少了各层或本层之间各节点之间的相互依赖，进而提高了模型的泛化性和鲁棒性。又因为 Dropout 每一次仅仅抛弃一小部分的节点权重，而这些网络共用模型的所有参数[15]。因此 Dropout 不仅使模型不容易过拟合，还能提高准确率。

3.5 本章小结

本章首先主要说明了基于图神经网络预训练的在线信息传播流行度预测算法

设计的整体流程，本框架由自注意门控循环网络特征提取和多种神经网络流行度预测两部分组成。首先手动提取图自身特征，其次为机器学习图特征提取，这两部分两部分组成自注意门控循环网络的特征提取过程。转发图自身特征提取包括转发有向图的边总数、节点总数、图密度、图的点度中心势、中介中心势和接近中心势，并阐述其实现细节和算法流程。

本章随后详细说明了微博转发链的数据及预处理，介绍了各个神经网络的结构。然而，大量的节点数既是深度学习的优势又是其劣势所在，不足之处就是更容易发生数据过拟合，故本文采用 Dropout 方法减少过拟合现象，以提高最终的预测成绩 。

第四章 实验结果与分析

4.1 实验数据集

在本章节的结果分析部分，使用的数据集与上一章介绍的数据集以及其划分方法保持一致，即使用全部微博数据集中各 15%的数据作为测试集和验证集，70%的数据作为训练集，利用训练集中 29529 条数据对各个模型进行 20 次迭代。

4.2 实验结果对比与分析

根据上述算法过程，本节将详细介绍算法的结果，并且对这些结果进行详细的分析。

本文算法实施结果如表 4-1 所示：

表 4-1 算法对比

	MAX	MIN	GOAL
MLP	4.27644	2.307554	2.1297654
Pre_MLP	4.18663	2.187848	2.061957
BiLSTM	2.487351	2.065499	2.0055389
Pre_BiLSTM	2.201748	2.086187	1.966474

表格中展示的是各个神经网络模型利用原始数据和预训练数据所得到训练结果的对比，其中 MAX 表示模型在 20 轮训练中，每一次迭代得到的参数对于测试集预测结果均方误差的最大值；其中 MIN 表示模型在 20 轮训练中，每一次迭代得到的参数对于测试集预测结果均方误差的最小值；而 GOAL 表示模型参数在取得预测结果均方误差最小值时，对于训练集的预测结果均方误差，本实验以此作为模型预测能力的最终成绩。各模型 GOAL 如图 4-1 所示。

在图 4-1 中， Pretraining 代表经过预训练后的模型，Normal 代表未经预训练的初始模型。我们分别用了 5 种模型作为本文提出的基于图神经网络预训练的在线信息传播流行度预测算法的回归器，并在其中的 2 中模型上取得了较好的成绩。由图、表可以看出，本文采用的预训练算法获得的数据应用各模型中，在相同的测试集上，MLP 和 BiLSTM 模型的得分分别提高了 3.18% 和 1.94% ，对比利用未经预训练数据进行训练的这两种模型，预训练后的模型取得了更好的成绩，由此可知于图神经网络预训练的模型对于在线信息传播流行度预测的确能够取得预测准确率的提高，也就是说明在线信息流行度预测中加入深度神经网络和迁移学习

的方法是可行的。

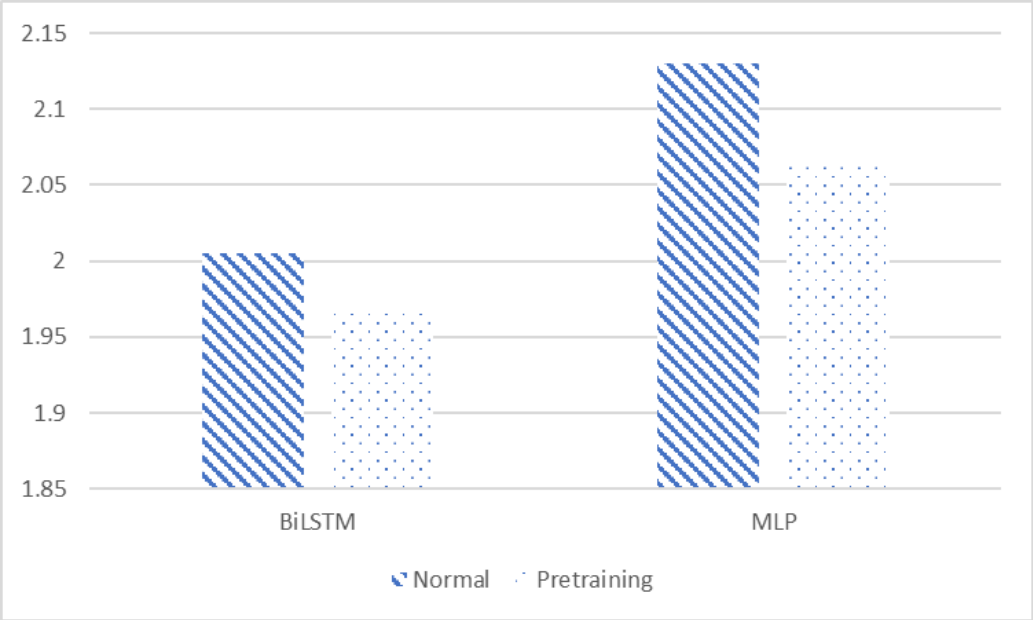

图 4-1 各模型预测得分对比

第五章 总结与展望

5.1 论文总结

在互联网社交计数日渐发达的现代社会，在线社交平台吸引了数以亿计的用户参与其中，可以说人们的日常生活已经离不开在线社交平台了。人们为了从分享、浏览中收获精神物质上的满足而不断发布、获取信息。其中，微博短文本以其精炼、易读等特点快速收获大量用户，为了向更具有目标性的把特定的信息推送给特定的用户，越来越多的研究学者开始向在线社交平台信息流行度预测的领域靠拢。与此同时，此领域的研究对于舆情监控、个性推送等方面的应用具有重大意义，即个人、企业或政府能够根据预测的结果对相关热点话题进行及时回应。综上所示，对于在线信息流行度预测的研究正处于上升期。本文首先构建了基于图神经网络预训练模型，接着提取了数据中转发有向图自身图特征，最后在流行度预测中加入应用 Dropout 的深度神经网络模型，然后进行不同算法之间的结果比较。内容如下：

(1) 设定图神经网络迁移学习模型并转发有向图特征。首先，本文的实验使用现有数据集，由微博转发链构建微博转发网络，并划分区间并提取转发图自身特征。然后通过自编码门控循环网络进行预训练，来预测自身转发网络的整体特征，之后将自编码门控循环网络的解码层删去，仅留下其前半的编码层。随后利用训练好的编码层更好的提取图度分布特征数据的深度信息，以更好地完成在线信息传播流行度预测。

(2) 构建深度神经网络模型。现有的对于微博流行度预测的研究多是采用人为特征选取，然而人为选择的特征难以代表原始数据，数据特征挖掘不够深入。本文在在线信息流行度预测的研究中应用深度神经网络技术，构建了相关深度神经网络模型，同时添加 Dropout 方法，以实现减少模型过拟合的问题，并由此确实取得了预测准确率的提高。所以，在流行度预测的问题上，本文所使用的基于图神经网络预训练的在线信息传播流行度预测模型能够获得比一般预测方法更加准确的预测结果。

5.2 工作展望

本研究通过使用预训练模型和深度神经网络进行在线信息流行度预测，取得了一定的准确率，但此模型还仍然存在着许多缺陷和不足之处。特别值得注意的是进化后的研究可以对深度神经网络方法进行算法、结构层面的改进。今后的相关研

究可以从以下几个本文的缺陷和不足之处进行改进：

(1) 可以向流行度预测模型中加入更多能够影响其预测的细节因素，而且可以实验将预训练和神经网络与其他方法相结合，以进一步提升特征提取的深度和流行度预测的准确度。

(2) 为了节约训练模型所需要的算力和时间，今后的研究中应更优先考虑利用 GPU 的并行计算方法来进行实验以提高效率，与此同时，还可以通过多次实验确定更为合适的学习率、网络结构等超参数以改进算法，提高模型预测能力。

(3) 随着硬件算力的不断提升，目前深度神经网络算法越来越复杂，而在本研究之中，仅仅使用了几种通用效果较好、结构相对简单的深度神经网络模型。所以，关于具体模型的选取，值得之后的研究对此进行更加深入的探索。

致 谢

我首先要感谢我的论文指导老师、电子科技大学公共管理学院的冯小东老师对我论文的研究方向做出了指导性的推荐和帮助，在论文撰写过程中及时对我遇到的困难和疑惑给予悉心指点，投入了大量的心血和精力。冯小东老师对我的帮忙和关怀表示诚挚的谢意！同时，还要感谢电子科技大学公共管理学院信息管理与信息系统的授课老师们和所有同学们，大家在大学四年的学习中互相学习，互相帮忙，共同度过了一段完美难忘的时光。

最后，谢谢论文评阅老师们的辛苦工作。衷心感谢我的家人、朋友，以及同学们，真是在他们的鼓励和支持下我才得以顺利完成此论文。

参考文献

[1] Liangjie Hong,Ovidiu Dan,Brian D. Davison. Predicting popular messages in Twitter[P]. World wide web,2011.

[2] 吴越,陈晓亮,蒋忠远.微博信息流行度预测研究综述[J].西华大学学报(自然科学版),2017,36(01):1-6.

[3] 韩凤娟,肖春静,王欢.基于多任务学习的微博流行度预测[J].河南大学学报(自然科学版),2017,47(05):544-551.

[4] 鲍鹏,徐昊.基于图注意力时空神经网络的在线内容流行度预测[J].模式识别与人工智能,2019,32(11):1014-1021.

[5] 刘洋. 基于 GRU 神经网络的时间序列预测研究[D].成都:成都理工大学,2017,23-25.

[6] 张荣,李伟平,莫同.深度学习研究综述[J].信息与控制,2018,47(04):385-397+410.

[7] 王治权. 基于注意力机制和改进型 RNN 的 Web 文本情感分析研究[D]. 兰州:兰州大学,2018,5-16.

[8] 司新红,王勇.CNN 结合 BLSTM 的短文本情感倾向性分析[J].软件导刊,2019,18(11):15-20.

[9] 陈红松,陈京九.基于循环神经网络的无线网络入侵检测分类模型构建与优化研究[J].电子与信息学报,2019,41(06):1427-1433.

[10] 李卫疆,李涛,漆芳.基于多特征自注意力 BLSTM 的中文实体关系抽取[J].中文信息学报,2019,33(10):47-56+72.

[11] 李天旭. 基于深度强化学习的多智能体协同算法研究[D]. 江苏:中国矿业大学,2020,7-15.

[12] 朱志鹏,杜建强,余日跃,聂斌,喻芳.融入深度学习的偏最小二乘优化方法[J].计算机应用研究,2017,34(01):87-90.

[13] 庄福振,罗平,何清,史忠植.迁移学习研究进展[J].软件学报,2015,26(01):26-39.

[14] 贺智超. 基于深度学习和迁移学习的多任务图像分类[D]. 广州:华南理工大学,2017,10-26.

[15] 顾淑琴. 基于集成学习的去相关正则化深度神经网络[D]. 天津:天津大学,2018,12-25.

外文资料原文

Pre-Training Graph Neural Networks for Generic Structural Feature Extraction

Ziniu Hu, Changjun Fan, Ting Chen, Kai-Wei Chang, Yizhou Sun
University of California, Los Angeles
{bull, cjfan2017, tingchen, kwchang, yzsun}@cs.ucla.edu

Abstract

Graph neural networks (GNNs) are shown to be successful in modeling applications with graph structures. However, training an accurate GNN model requires a large collection of labeled data and expressive features, which might be inaccessible for some applications. To tackle this problem, we propose a pre-training framework that captures generic graph structural information that is transferable across tasks. Our framework can leverage the following three tasks: 1) denoising link reconstruction, 2) centrality score ranking, and 3) cluster preserving. The pre-training procedure can be conducted purely on the synthetic graphs, and the pre-trained GNN is then adapted for downstream applications. With the proposed pre-training procedure, the generic structural information is learned and preserved, thus the pre-trained GNN requires less amount of labeled data and fewer domain-specific features to achieve high performance on different downstream tasks. Comprehensive experiments demonstrate that our proposed framework can significantly enhance the performance of various tasks at the level of node, link, and graph.

1 Introduction

Graphs are a fundamental abstraction for modeling relational data in physics, biology, neuroscience and social science. Although there are numerous types of graph structures, some graphs are known to exhibit rich and generic connectivity patterns that appear general in graphs associated with different applications. Taking network motifs, which are some small sub-graph structures, as an example, they are considered as the building blocks for many graph-related tasks [3], e.g., triangular motifs are crucial for social networks, two-hop paths are essential to understand air traffic patterns, etc. Despite its importance, previous researchers are required to design various specific rules or patterns to extract motif structures, in order to serve as features for different applications manually. This process is tedious and ad-hoc.

Recently, researchers have adopted deep representation learning into graph domain and proposed various Graph Neural Network architectures [22, 18, 41] to alleviate this issues by automatically capturing complex information of graph structures from data. In general, GNNs take a graph with attributes as input and apply convolution filters to generate node embeddings with different granularity levels layer by layer. The GNN framework is often trained in an end-to-end manner towards some specific tasks and has shown to achieve competitive performance in various graph-based applications, such as semi-supervised node classification [22], recommendation systems [47] and knowledge graphs [38].

Despite the success, most of the GNN applications heavily rely on the domain-specific input features. For example, in PPI dataset [49], which is widely used as a node classification benchmark task, positional gene sets, motif gene sets, and immunological signatures are used as features. However, these domain-specific features can be hard to obtain, and they cannot generalize to other tasks. Without access to these domain-specific features, the performance of GNNs is often suboptimal.

Preprint. Under review.

外文资料译文

用于通用结构特征提取的图谱神经网络的预训练

作者：Ziniu Hu, Changjun Fan, Ting Chen, Kai-Wei Chang, Yizhou Sun

研究单位：加州大学洛杉矶分校

主要内容：

图形神经网络（GNNs）被证明在建模应用中是成功的。训练一个准确的 GNN 模型需要大量的标记数据和表达式特征，但这些标记数据对某些应用来说可能是难以获得的。为了解决这个问题，我们提出了一个预训练框架捕获可跨任务转移的通用图结构信息。我们的框架可以利用以下三个任务。(1)去噪链路重建，(2)中心性得分排名(3)集群保存。预训练过程可以纯粹在合成图上进行，而训练后的 GNN 可用于下游的应用。通过本文提出的预训练程序，通用的结构信息被学习和保留，因此，预训练的 GNN 需要较少的标记数据和较少的特定领域特征，能够在不同的下游应用中获得更高性能。综合实验表明，我们提出的框架可以显著地在节点、链接和图的层面上提高各种任务的性能。