基于图注意力模型的交通网络流量预测

1 引言

随着社会的进步和经济的发展，人民生活水平不断提高，城市中车辆的数量也与日俱增。随之而来的是城市交通拥堵问题，以及由此产生的出行时间浪费、城市环境污染、交通事故频发等等现象，不仅不利于城市治理，也降低了市民生活满意度。实现对城市交通的流量预测，能够有效帮助人们判断交通拥堵情况，做出合理的出行安排，缓解交通压力。

过去人们只能依靠经验判断交通状况，如今信息技术高速的发展使得城市交通系统的逐步智能化成为可能，信息技术、传感技术、计算机处理技术等都可以作为交通治理的有力辅助工具。在智能交通系统中，交通流量的预测是交通管理与调控的基础性环节，而流量的预测数据也正是其他工作开展的重要依据。因此，如何选择一个科学有效的方法进行预测模型构建就成为了一个关键问题。

目前，交通流量预测问题的主要研究方法有三大类：经典时间序列预测方法、传统机器学习方法和深度学习方法。早期以经典序列预测方法为主，如历史均值法、向量自回归模型、自回归积分滑动平均模型及其变体模型等等。之后，以传统机器学习方法为主，主要通过构建集成模型，如将经验模式分解与 BP 神经网络相结合、小波支持向量机等来实现。随着深度学习在计算机视觉、语音文本处理等任务中不断取得突破，深度学习技术开始被应用于交通流量预测任务，出现了将 RNN 类模型与传统机器学习方法相结合进行流量预测等预测方法。后来，为解决传统卷积神经网络在图数据上的局限性，图卷积技术应运而生，ChebNet 模型、GCN 模型、GAT 模型等被用于交通预测。本文将要探索的就是基于图注意力网络模型（GAT）的交通流量预测方法。

2 相关工作

本文的目标为学习图注意力模型，探索如何实现基于图注意力网络模型的交通流量预测方法，从而对社会网络分析的实际应用有更深刻的理解。所做的主要工作有：

- 理论调研。查询了有关图卷积神经网络的资料，对其产生与发展历程，各种经典模型的架构与优缺点，以及实际应用有了基本的了解。
- 问题定义。对“交通流量预测”这一任务进行分析，经过抽象，得到问题的数学表示方法，用于指导后面的实验。
- 数据集选取与处理。选取美国加州高速路网交通流量数据集的子集 PeMS-04 作为实验数据集，并对其进行合适的处理，便于模型训练和测试。
- 算法设计与实验。根据数据情况设计 GAT 模型，并进行模型训练和测试，得到结果。

3 实验方法

在这部分将对实验进行设计，包括问题定义、数据集选取和算法介绍三部分内容。

3.1 问题定义

定义 1（交通网络）用G表示交通节点形成的交通网络，$G=(V,E,\boldsymbol{A})$，其中，V表示所有节点，v表示某个节点，$v\in V$，$|V|=N$为节点个数，E表示边集，边代表节点之间存在直连的道路，$\boldsymbol{A}$表示邻接矩阵。

定义 2（邻居图）用G_N表示根据节点位置相邻关系形成的邻居图，$G_N=(V,E_N,\boldsymbol{A}_N)$，其中，$E_N$表示边集，边代表节点之间的相邻关系，$\boldsymbol{A}_N$表示邻接矩阵。

定义 3（流通流量图）用G_F表示根据节点间流通流量关系形成的流通流量图，$G_F=(V,E_F,\boldsymbol{A}_F)$，其中，$E_F$表示边集，边代表流量强弱关系，$\boldsymbol{A}_F$表示邻接矩阵。

定义 4（出入网络记录）四元组$T_k(p,v,\tau,\kappa)$ 表示个体p从站点v进出网络产生的一条记录，其中，p表示个体标识，v表示站点标识，τ表示时间标识，κ表示进出标识，$\kappa\in\{in,out\},k\in\mathbf{N}$表示记录序号，若某个体多次进出网络，则有记录链$T_0\to T_1\to\cdots\to T_k\to\cdots$。

定义 5（连续出入网络记录组）二元组$T_{oi}(T_k,T_{k+1})$ 表示某个体连续进出一次网络产生的记录组，其中，$T_k\cdot\kappa=in,T_{k+1}\cdot\kappa=out,T_k\cdot p=T_{k+1}\cdot p,T_k\cdot\tau<T_{k+1}\cdot\tau$。

定义 6（站点出入流量）$t=[t_{start},t_{end})$ 表示某个时间区间，$\Delta t=t_{start}-t_{end}$表示区间长度，则在时间区间$t$内，站点$v$上产生的进出流量分别为：

$$x_t^{in,v}=|\{T_k(p,v,\tau,\kappa)|T_k\cdot\kappa=in\wedge T_k\cdot v=v\wedge T_k\cdot\tau\in t\}|$$

$$x_t^{out,v}=|\{T_k(p,v,\tau,\kappa)|T_k\cdot\kappa=out\wedge T_k\cdot v=v\wedge T_k\cdot\tau\in t\}|$$

将所有站点在时间区间t内的出入流量表示为张量$X_t\in\boldsymbol{R}^{N\times 2}$，其中，$(X_t)_{v,0}=x_t^{in,v}$，$(X_t)_{v,1}=x_t^{out,v}$。

定义 7 本文预测任务为根据各站点若干个历史时间区间的出入流量观测值，预测各站点在未来若干个时间区间内的进出流量值。$\boldsymbol{X}=(X_{t-T_h+1},X_{t-T_h+2},\dots,X_t)\in\boldsymbol{R}^{T_h\times N\times 2}$ 表示 T_h 个历史区间观测值，$\boldsymbol{Y}=(X_{t+1},X_{t+2},\dots,X_{t+T_p})\in\boldsymbol{R}^{T_p\times N\times 2}$表示$T_p$个待预测的区间值，本文目标是学习如下式所示的映射$f(\cdot)$：

$$(X_{t-T_h+1},X_{t-T_h+2},\dots,X_t)\xrightarrow{f(\cdot)}(X_{t+1},X_{t+2},\dots,X_{t+T_p})$$

3.2 数据集选取

鉴于大范围跨时空交通数据采集的难度，高质量的数据集对于交通预测问题的研究至关重要。交通数据类别丰富，包括交通运输数据、交通管理数据以及用于辅助的气象数据和事件数据等。本文选取的实验数据集为美国加州高速路网交通流量数据集的子集 PeMS-04。

PeMS 为美国加利福尼亚运输局的性能测量系统(Performance Measurement System)的缩写。该数据集的内容由 39000 余个独立传感器以 5 分钟为时间间隔实时收集，其范围覆盖了加利福尼亚州所有主要城市区域的高速公路系统。由于 PeMS 全部数据体量较大，目前衍生出多个常用的子数据集，如 PeMS-03、PeMS-

04、PeMS-07、PeMS-08、PeMS-SF 以及 PeMS-BAY 等。子数据集覆盖了不同大小的区域，包含不同数量传感器采集的不同时间跨度以及不同时间粒度的交通流信息。其中，PeMS-04 子数据集覆盖 307 个节点，时间范围是 2018 年 1 月 1 日至 2018 年 2 月 28 日，采样时间间隔为 5 分钟。

3.3 算法介绍——GAT 模型

GAT 模型，即图注意力网络模型（Graph Attention Networks），将注意力机制引入到基于空间域的图神经网络，不需要进行复杂的计算，仅是通过一介邻居节点的表征来更新节点特征。算法流程如下：

①计算注意力系数。对所有节点训练一个共享的权重矩阵$\boldsymbol{W} \in R^{F\times F'}$，得到每个邻居节点的权重，这个权重矩阵就是输入的$F$个特征与输出的$F'$个特征之间的关系，起到映射的作用。计算注意力值时，将节点i和节点j的表示分别使用$\boldsymbol{W}$做映射，并将其结果向量拼接起来。之后使用前馈神经网络$\vec{a}^T$将拼接向量映射到实数上，并通过 LeakyReLU 函数激活，经过归一化后得到最终的注意力系数，计算公式为：

$$e_{ij} = LeakyReLU\left(\vec{a}^T[\boldsymbol{W}\vec{h}_i||\boldsymbol{W}\vec{h}_j]\right)$$

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k\in N_i} exp(e_{ik})}$$

其中，||符号表示向量拼接。

②加权求和。得到注意力系数后，对邻居节点特征进行加权求和，得到节点i的输出特征：

$$\vec{h}_i' = \sigma(\sum\nolimits_{j\in N_i} \alpha_{ij}\boldsymbol{W}\vec{h}_j)$$

③引入多头注意力机制。为了使自注意力能够稳定地表示节点，引入了多头注意力机制来提高模型的表征能力。对于中间层输出特征，使用K个$\boldsymbol{W}$计算自注意力，然后将注意力头得到的结果拼接得到输出向量，对于最终的结果，则是对各个注意力头的输出向量采用取平均的策略：

$$\vec{h}_i' = ||_{k=1}^{K}\sigma(\sum\nolimits_{j\in N_i} \alpha_{ij}^k\boldsymbol{W}^k\vec{h}_j)$$

GAT 模型为邻接节点分配不同的权重，考虑到了节点特征之间的相关性。对于交通流量预测任务而言，某道路的交通状态在一定程度上必然受到直接相邻或间接相邻道路的影响，这种影响的范围以及程度随位置、距离以及时间等因素不断变化。因此，使用 GAT 模型来预测交通网络的流量是切实可行的。

4 实验及结果

在上一部分中，我们对问题进行了定义，确定了实验数据集与算法，接下来进行实验，并对得到的结果进行分析。

4.1 数据分析与处理

使用数据进行模型训练与测试前，必须对数据内容有所了解，并对数据进行

合适的处理，以嵌合模型需要。

第一步，加载数据，查看数据形状并使数据可视化。在这一步中，我们得知数据由三个维度组成，分别是时间、节点数和节点特征，对于 PeMS-04 数据集，时间点有 16992 个，即 59 天中的每 24 小时里，每小时采样 12 次（即采样时间间隔 5 分钟）；节点数为 307，即网络中共有 307 个节点；节点特征数量为 12。选取节点 25、106 和 286 进行特征查看，得到结果分别如下图 4.1.1、图 4.1.2 和图 4.1.3 所示。

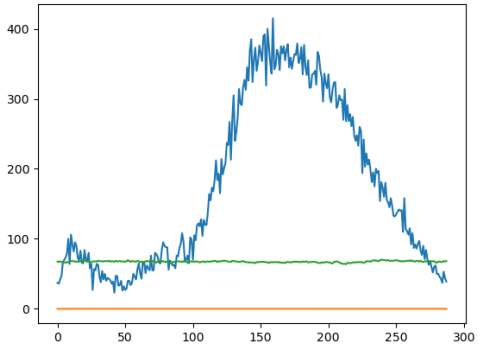

图 4.1.1 节点 25 的节点特征

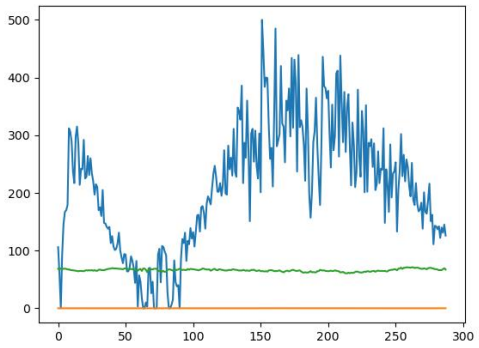

图 4.1.2 节点 106 的节点特征

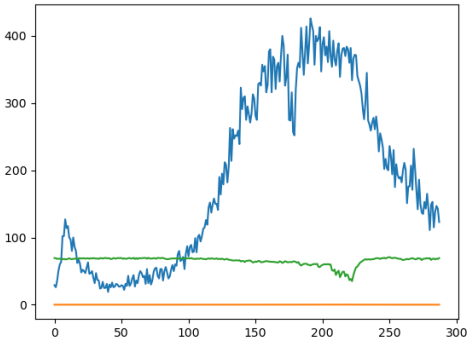

图 4.1.3 节点 286 的节点特征

其中，蓝色线代表第一维特征，橘色线代表第二维特征，绿色线代表第三维特征。从图中可以看出，每个节点有三个特征，但第二、第三维特征基本平稳不变，因此，我们只需要取第一维特征进行分析处理即可。

第二步，对数据集进行处理，使之成为 PyTorch 需要的数据集的类。首先，将节点关系处理成邻接矩阵，相邻节点权重设为 1。接着，读入流量数据，即上一步中每个节点的特征数据。然后，通过滑动窗口，将读入的数据处理成模型所需要的训练集和测试集，训练集取 45 天的数据，测试集取 14 天的数据。

至此，数据预处理过程完成。

4.2 GAT 模型构建

得到处理好的数据后，接下来开始 GAT 模型构建工作。

第一步，定义模型。GAT 模型的实现可简单归结为计算节点之间的关联度、得到每个节点的注意力系数并进行归一化、利用注意力系数对邻域节点进行有区别的信息聚合三个步骤。整个模型基于 PyTorch 实现，模型中，设置了一个线性层，用于关联度计算，设置了循环，用于增加多头注意力，设置模型输入层层数

为 6，隐藏层层数为 6，输出层层数为 1，注意力头数为 2。

第二步，定义损失函数和优化器。损失函数使用均方损失函数，学习率为默认的 1e-3。

4.3 模型训练与测试

模型训练过程的主要步骤为取出数据、梯度清零、计算损失、反向传播和参数更新，训练次数设为 10。模型测试过程的主要步骤为将测试集放入训练好的模型中，得到预测值，并与标签值进行比较，得到模型的评价指标值。这里选择的评价指标为 MAE（平均绝对误差，Mean Absolute Error）、MAPE（平均绝对百分比误差，Mean Absolute Percentage Error）和 RMSE（均方根误差，Root Mean Square Error）。

取节点 68、130 和 256 的结果进行可视化，查看预测值与实际值的重合情况，如图 4.3.1、图 4.3.2 和图 4.3.3 所示。从图中可知，预测值和真实值的变化趋势基本一致，在值的精确度上仍有差距，这可能是训练次数较少所致。

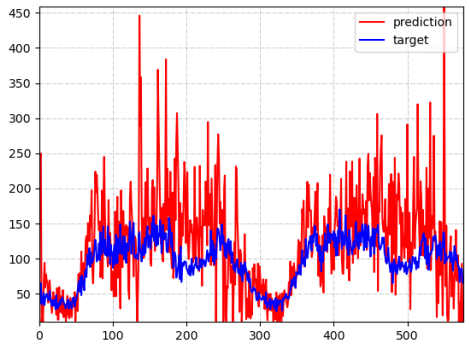

图 4.3.1 节点 68 的预测结果

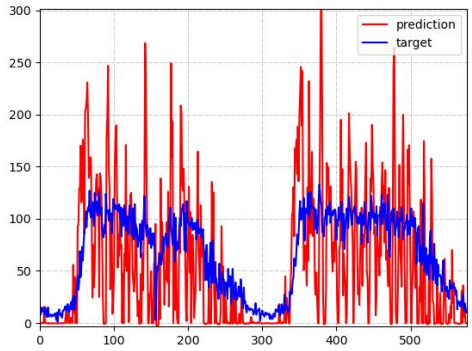

图 4.3.2 节点 130 的预测结果

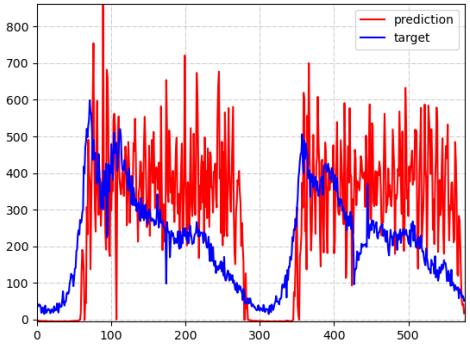

图 4.3.3 节点 256 的预测结果

下表展现了节点 68、130 和 256 的预测评价指标值的情况。

表 4.3.1 模型评价指标值

节点编号	Loss	MAE	MAPE	RMSE
68	0.9456	85.50	1.80%	123.17
130	1.3548	110.14	1.59%	149.95
256	1.0205	94.69	1.86%	131.19

5 总结

本文针对基于图注意力模型的交通网络流量预测问题进行了探究和实现，得到了一个基于 PyTorch 实现的可用于预测交通流量的 GAT 模型，预测结果在流量变化趋势上较为准确，在流量值的精确度上仍有改进空间。作者通过这一实践，增加了对社会网络分析的实际应用的了解，同时看到了图神经网络在城市交通网络流量预测乃至其他网络的分析中有着巨大的潜力。