

分类号：
密 级：(不填)

单位代码：10389
学号：(S)1211193020



硕士学位论文

基于注意力机制的网络流量预测研究

学 科 门 类：工学

一级学科名称：计算机科学与技术

二级学科名称：

研 究 方 向：计算机网络与信息安全

研究生姓名：张治方

指 导 教 师：杨昇教授 舒兆港副教授

完 成 时 间：二〇二四年四月

Research on the network traffic prediction based on the attention mechanism

By

Zhifang ZHANG

Supervised by associate Prof. Zhaogang Shu and Prof. Sheng Yang

A Dissertation Submitted to

Fujian Agriculture and Forestry University

in Partial Fulfillment of the Requirements for

Master Degree of Engineering

College of Computer and Information Science

Fujian Agriculture and Forestry University

Fujian, P.R. China

Submission Date: April, 2024

独创性声明

本人声明，所提交的学位（毕业）论文，是本人在指导教师的指导下独立完成的研究成果，并且是自己撰写的。尽我所知，除了文中作了标注和致谢中已作了答谢的地方外，论文中不包含其他人发表或撰写过的研究成果。与我一同对本研究做出贡献的同志，都在论文中作了明确的说明并表示了谢意，如被查有侵犯他人知识产权的行为，由本人承担应有的责任。

学位（毕业）论文作者亲笔签名：张治方 日期：2024.5.29

论文使用授权的说明

本人完全了解福建农林大学有关保留、使用学位（毕业）论文的规定，即学校有权送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

保密，在 年后解密可适用本授权书。 ☐

不保密，本论文属于不保密。 ☒

学位（毕业）论文作者亲笔签名：张治方 日期：2024.5.29

指导教师亲笔签名： 日期：2024.5.29

目 录

专用术语对照表.....	III
摘 要.....	IV
ABSTRACT.....	VII
1 绪论.....	1
1.1 研究背景和意义.....	1
1.2 国内外研究现状.....	3
1.2.1 线性预测技术介绍.....	3
1.2.2 非线性预测技术介绍.....	3
1.3 研究目标和研究内容.....	5
1.3.1 研究目标.....	5
1.3.2 研究内容.....	6
1.4 论文结构和章节安排.....	8
2 相关理论基础.....	9
2.1 时间序列预测.....	9
2.2 网络流量相关特性.....	9
2.2.1 ARIMA 模型原理.....	10
2.2.2 CNN 模型原理.....	11
2.2.3 LSTM 模型原理.....	12
2.3 注意力机制原理和应用.....	13
2.4 本章小结.....	14
3 面向空洞卷积和相对位置表征的改进 Transformer 模型.....	15
3.1 导言.....	15
3.2 问题定义.....	15
3.3 面向空洞卷积和相对位置表征的改进 Transformer 模型组件.....	16
3.3.1 Transformer 模型.....	16
3.3.2 空洞卷积.....	19
3.3.3 注意力机制之相对位置表征.....	21
3.4 性能评估.....	22
3.4.1 数据集.....	22
3.4.2 实验环境和评价指标.....	22
3.4.3 结果与分析.....	26
3.5 本章小结.....	27
4 基于自适应注意力机制的网络流量预测模型.....	29
4.1 导言.....	29
4.2 自适应注意力机制网络流量预测模型.....	32
4.2.1 模型介绍.....	32
4.2.2 模型设计.....	33
4.2.3 自适应注意力选取算法.....	35
4.3 性能评估.....	36

4.3.1 数据集.....	36
4.3.2 仿真环境与评价指标.....	37
4.3.3 结果与分析.....	38
4.4 本章小结	44
5 结论和展望.....	46
5.1 主要工作	46
5.2 未来展望	46
参考文献.....	48
攻读学位期间的论文与研究成果.....	53
致谢.....	54

专用术语对照表

缩略词	英文全称	中文对照
5G	5th generation mobile networks	第五代移动通信网络
SDN	Software Defined Network	软件定义网络
NVF	Network Function Virtual	网络功能虚拟化
NS	Network Slice	网络切片
ARIMA	Autoregressive Integrated Moving Average Model	自回归合成移动平均模型
ML	Machine Learning	机器学习
DNN	Deep Neural Networks	深度神经网络
ANN	Artificial Neural network	人工神经网络
RNN	Recurrent Neural Network	循环神经网络
FRNN	Fully Recurrent Neural Network	递归神经网络
SNN	Stochastic Neural Network	随机神经网络
LSTM	Long short-term memory network	长短期记忆网络
SAA	Simulate Anneal Arithmetic	模拟退火
BPNN	Back Propagation Neural Network	反向传播神经网络
SWG	Sliding Window Gradient Descent	滑动窗口梯度下降
LSTF	Long Sequence Time-series Forecasting	长序列时间预测
AM	Attention Mechanism	注意力机制
DC	Dilated Convolution	空洞卷积
RP	self-attention with Relative Position representation	相对位置表征
TDCRP	modified Transformer model oriented towards Dilated Convolution and Relative Position representation	空洞卷积相对位置表征的改进 Transformer 模型
AAM	Adaptive Attention Mechanism	自适应注意力机制模型

摘 要

近年来,伴随着科技飞速发展,互联网和各种智能设备得到了广泛应用。人与互联网的关系愈发密切,全球网络流量呈现持续上涨趋势,网络的规模 and 用户数量都呈爆炸式增长,用户对互联网业务需求也不断增加。这些快速增长的流量和业务需求对网络运营商和用户来说都是一个巨大问题,因此准确地了解网络流量的变化趋势具有重要意义。

计算机网络流量是指一定时间内通过网络的数据总量,是衡量网络负载和运行状态的重要参数。在 5G、云服务和网络切片背景下,开展网络流量预测有利于网络运营商调整网络资源的同时满足用户需求。然而,实际网络流量具有实时性和突发性等特征,传统网络流量预测模型存在预测时间点少和预测精度低等问题,基于神经网络的预测模型存在梯度消失,梯度爆炸以及误差累积等问题。所以改善网络流量预测模型,避免以上问题是当前研究的重点。因此,本文提出了两种基于注意力机制的网络流量预测模型,具体的研究工作如下:

(1) 提出一种面向空洞卷积和相对位置表征的改进 Transformer 模型 (a modified Transformer model oriented towards Dilated Convolution and Relative Position representation, TDCRP), TDCRP 模型的原理是在 Transformer 模型的基础上采用空洞卷积和相对位置表征注意力重构编码器和解码器。首先,将网络流量结合一维空洞卷积提取流量序列中数据的相似性,其次,在训练模型时考虑相对位置表

征的方式替换 Transformer 模型中原本的自注意力机制，使得数据的相对位置信息得到表达，从而提升模型的性能。TDCRP 模型的优点在于能用空洞卷积扩大感受野，关注网络流量数据上下文信息（当前位置的相对应的其余时间点），与典型的 Transformer 模型计算逐点值不同，改进模型能通过数据上下文信息计算相似度，例如数据序列中蕴含的局部形状相似性。另外，TDCRP 采用相对位置表征注意力机制，其生成的查询矩阵 Q ，键值矩阵 K 能够在多头注意力中关注流量数据的相对位置信息，即相对位置信息对当前位置而言的重要性。实验表明，与原始 Transformer 和 CNN 模型相比，本文所提 TDCRP 模型在短期网络流量预测上性能较优。

（2）基于网络流量数据的时间序列理论，针对长时间序列预测问题，提出一种自适应注意力机制（Adaptive Attention Mechanism, AAM）网络流量预测模型。该模型设计一种自适应注意力选取算法，根据 KL 散度度量 q 和 k 相似性，根据排序后分数自动挑选一个数据子集，使得 AAM 能顾及网络流量中“历史久远”的流量特征，自适应注意力机制里的注意力“平均”选取避免了数据里一些重要信息的丢失，使其在长序列流量预测问题中能达到更好效果。AAM 采用动态自适应的方式处理输入数据，缩减计算量的同时又减少数据的信息损失，解码器的输入采用编码器的输出和经掩盖处理的数据序列，防止当前位置关注后续位置。另外，模型只需通过一个前向步骤就能获得长时间序列的输出，消除了推理阶段的累积误差传播，提升了长序列预测的速度。实验结果表明，AAM 模型与经典的差分自回归移动平

均模型以及长短期记忆网络方法相比，在波动明显的数据集上有更高准确率和更好预测性能。

关键词：网络流量预测，空洞卷积，相对位置表征，注意力机制，自适应注意力

ABSTRACT

In recent years, with the speedy development of science and technology, the Internet and various smart devices have been widely used. The relationship between people and the Internet has become closer, global network traffic continues to rise, the scale of the network and the number of users are growing explosively, and users' demand for Internet services is also increasing. These rapidly growing traffic and business demands are a huge problem for network governors and users, so it is significant to accurately know the changing trends of network traffic.

Computer network traffic refers to the total amount of data passing through the network within a certain period of time, and is an important parameter to measure network load and operating status. In the context of 5G, cloud services and network slicing, network traffic prediction helps network operators adjust network resources while meeting user needs. However, the actual network traffic has the characteristics of real-time and sudden, traditional models have problems such as few prediction points and little prediction accuracy. Prediction models based on neural networks have problems such as gradient disappearance, gradient explosion, and error accumulation. Therefore, improving network traffic prediction models and avoiding the above problems is the focus of current research. Therefore, this dissertation proposes two models based

on the attention mechanism for network traffic prediction. The specific research work is follows:

(1) To propose an improved Transformer model for dilated convolution and relative position representation (a modified Transformer model oriented towards Dilated Convolution and Relative Position representation, TDCRP), The principle of the TDCRP model is to use the void convolution and relative position representation of attention to constitute the encoder and decoder. First, the network traffic is combined with a one-dimensional void convolution to extract the similarity of the data in the traffic sequence. Second, consider the original self-attention mechanism in the Transformer when training the model, So that the relative positional information of the data is expressed, Thus improving the performance of the model. The advantage of the TDCRP model is that it can expand the receptive field with void convolution and capture the traffic data context information (the remaining time points of the current position). Different from the typical Transformer calculation point by point value, the similarity can be calculated by their local context information (such as local shape). In addition, TDCRP uses a relative position representation attention mechanism. The query matrix Q and key value matrix K generated by it can focus on the relative position information of traffic data in multi-head attention, that is, the importance of relative position information to the current position. Experiments show

that compared with the original Transformer and CNN models, the TDCRP model is proposed in this dissertation has better performance in network traffic prediction.

(2) Based on the time series theory of network traffic data, an Adaptive Attention Mechanism (Adaptive Attention Mechanism, AAM) network traffic prediction model is proposed for long-term series prediction problems. This model designs an adaptive attention selection algorithm that measures the similarity between q and k based on KL divergence, and a data subset is automatically selected based on the sorted scores, so that AAM can take into account the "long-standing" traffic characteristics in network traffic. The "average" selection of attention in the adaptive attention mechanism avoids some important information in the data. The loss enables it to achieve better results in long sequence traffic prediction problems. AAM uses a dynamic adaptive method to process input data, which reduces the amount of calculation and reduces the information loss of the data. The input of the decoder uses the output of the encoder and the masked data sequence to prevent the current position from focusing on subsequent positions. In addition, the model only needs one forward step to obtain the output of a long sequence, eliminating the cumulative error propagation in the inference stage and improving the speed of long sequence prediction. Experimental results show that compared with the classic differential Auto Regressive

Integrated Moving Average Model and Long Short Term Memory networks methods, the AAM model performs better on data sets with obvious fluctuations and has higher accuracy and better prediction performance.

Key Words: network traffic prediction, dilated convolution, relative position representation, attention mechanism, adaptive attention

1 绪论

1.1 研究背景和意义

近年来，随着科技的飞速发展，互联网和智能设备得到广泛应用，全球网络流量呈现出持续增长趋势。根据最新发布的《中国互联网发展现状统计报告》（第 52 期）显示，截至 2023 年的 6 月，我国的互联网用户数量已达 10.79 亿人，相比 2022 年的 12 月共增长了 1109 万人，互联网普及率达到 76.4%^[1]。该报告指出，在 2023 年上半年，我国各类互联网应用都在不断发展壮大，多种应用类型的用户数量也都有所增加。以即时通讯、在线视频和短视频为例，这三类应用的用户基数一直保持在前列。截至 23 年 6 月份，即时通讯用户数量达到了 10.47 亿，在线视频用户数量达到了 10.44 亿，短视频用户数量达到了 10.26 亿，它们的利用率分别为 97.1%、96.8%和 95.2%，如图 1-1 中清晰展示^[1]。值得注意的是，网络用户的行为具有复杂多样性，呈现出多种变化趋势。由此可见，网络在人们的日常生活中扮演着愈加重要的角色，人与互联网之间的关系愈发密切，用户对互联网业务的需求不断增加。这些快速增长的流量和业务需求和对网络运营商、服务提供商和网络规划者来说是一个巨大的挑战，因此，需要准确地了解网络流量的变化趋势。

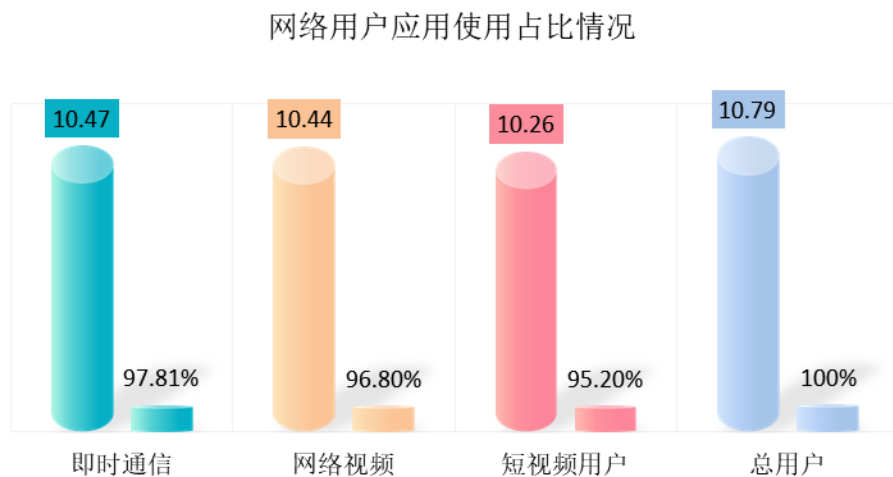


图 1-1 各网络应用人数占总用户人数

另外，伴随第五代移动通信网络（5th Generation mobile networks, 5G）、软件定义网络（Software Defined Network, SDN）^[2]，网络功能虚拟化（Network Function Virtual, NFV）^[3, 4]和网络切片（Network Slice, NS）^[5]等技术的发展，

SDN 和 NFV 可以利用虚拟化技术将特定的网络功能从专用设备分离到通用的硬件设备上,未来,网络运维和管理的新挑战在于实现更加精细化、自动化和智能化^[6]。网络流量预测可以解决这些问题,并且可以在多个级别准确感知应用程序级别的网络流量。这对于许多网络管理任务至关重要,例如流量管理、异常检测、负载均衡等。实时检测网络流量、及时发现流量变化趋势以及分析各种网络资源请求都可以帮助网络运营商调控动态智能服务并改善其资源利用率。同时,网络运营商可以根据实际业务需求合理分配资源。例如,网络基站覆盖范围和功率的动态调整对基础设施的规划和优化非常有用。因此,利用合适的模型通过历史流量信息预测未来的网络流量数据具有重要研究意义和研究价值。具体而言:

(1) 准确预测网络流量有助于网络运营商布局基础设施。当运营商配置网络资源时,如果没有考虑未来的趋势,可能导致网络过载或资源使用不合理。通过预测网络流量,可以在流量高峰到来之前扩展网络带宽并提前规划方案,以避免网络服务质量下降。

(2) 准确预测网络流量有助于网络运营商调整和扩展额外的业务服务。对于现有的网络运营商来说,服务模式相对单一,无法为用户提供不同的服务或个性化的服务。通过提前预测业务网络的流量,可以为不同的业务需求提供差异化的互联网服务。

(3) 准确预测网络流量有助于提高通信网络的性能。没有对未来状态进行预测可能会导致网络过载或资源使用不合理。准确预测网络流量可以及时应对即将到来的过载,并确保及时扩展、调整和优化网络。通过大规模部署 5G 基站,准确预测流量可以根据实际业务需求灵活分配资源。在网络需求较低时,一些负载较轻的基站可以转入节能模式,基站覆盖区域也可以动态调整。当这些基站需要被使用时,它们可以被激活,从而降低基站的能耗。

(4) 准确的网络流量预测有利于软件定义网络的性能分析和网络规划。作为一种新的尖端网络技术,SDN 将网络交换机设备的控制平面和数据平面解耦合,实现了网络测量以及管理的灵活性。基于流的转发思想的 SDN 架构为未来网络构建提供了新的思路,具有广阔的前景,网络流量预测能更好地为此技术提供支撑。

因此,准确的网络流量预测对实现各种各样的网络管理任务,如流量工程、异常检测、网络核算、网络分析、性能诊断、负载平衡等^[7-13]都具有重要意义。

1.2 国内外研究现状

随着现代社会的发展,网络诚然已经成为人们生活中不可或缺的一部分,对于网络流量预测问题,已经成为很多国内外学者研究的方向。目前,网络流量预测模型通常可以分为两种主要类型。一种是传统的线性回归模型,另一种则是基于机器学习的非线性预测模型^[14]。

1.2.1 线性预测技术介绍

经典线性流量预测模型由来已久,许多学者经常使用人工设置的多种参数化的模型。主要包括自回归合成移动平均模型(Autoregressive Integrated Moving Average Mode, ARIMA)^[15]、移动平均模型(Moving Average, MA)^[16]和自回归模型(Autoregressive Model, AR)^[17]等。随着预测数据量逐渐增加,数据变得更加复杂,时间序列中的非线性特征变得更加明显,传统线性方法的预测准确度开始难以为继。

1.2.2 非线性预测技术介绍

近年来,机器学习(Machine Learning, ML)领域迅猛发展,非线性流量预测技术和模型成为研究人员重点研究方向,一些以神经网络和深度学习模型为主的预测方法成为主流。通常来说,机器学习是一种令机器处理和分析数据并找到其中规律的方法。针对时间序列数据,计算机可以利用这些规律对其进行处理,从中提取出一些有用的信息和规律,然后利用所习得的知识对未知数据进行预测。作为数据驱动型算法,它们具有非线性预测能力以及高拟合的特性。

首先,在深度神经网络(Deep Neural Networks, DNN)还未大显身手之时,预测领域研究人员通常使用一些浅层的神经网络,如小波神经网络^[18]以及反向传播神经网络等^[19]。人工神经网络(Artificial Neural networks, ANN)是应用广泛的ML技术之一^[20],ANN可以学习众多变量之间复杂的非线性依赖关系,其一般结构是由输入层和输出层组成的有向图^[21],这些输入层和输出层通过隐藏层连接,输入值经过隐藏层处理转换到输出层。这些隐藏层的数量被称为模型的深度。深度神经网络对抽象、高维的特征提取效果比较好。相较于浅层神经网络,深度神经网络克服了浅层神经网络存在容易遗失信息的缺点,成为预测问题的解决新方案。

DNN是指由两个或两个以上的隐藏层构成的神经网络,其结构灵活,在时间序列预测中获得了惊人的广泛应用。在这种情况下,一些与神经网络相关

模型相继出现，比如循环神经网络（Recurrent Neural Network, RNN），取得了较大的进步。同时，基于 RNN 的架构，完全递归神经网络（Fully Recurrent Neural Network, FRNN）、双向神经网络（Bidirectional Neural Networks, BNN）、随机神经网络（Stochastic Neural Network, SNN）等模型应运而生。另外，众所周知的范式长短期记忆网络（Long Short Term Memory network, LSTM），根据其所采用的激活函数和神经元相互连接的方式来定义^[22]。LSTM 作为 RNN 的扩展，提出了解决 RNN 对长期依赖引起的梯度爆炸或者消失问题的方法^[23-25]。LSTM 在 RNN 架构上的创新，其具体结构是控制单元输入的三个门机制和用于数据保留的反馈回路，这使得 LSTM 可以学习长期依赖关系，并消除导致神经单元输出扰动的无效输入^[26, 27]。

其次，与神经网络和深度学习相关的组合预测模型也迅速发展。在一项研究中，任玮等人提出了一种新的网络流量预测模型^[28]，模型采用深度信念网络强大的非线性适配能力，利用贪心算法对参数开展训练，然后通过反向传播对其微调网络。张传亭及其团队提出了一种针对无线业务的深度卷积神经网络模型^[29]，他们利用深度卷积网络来捕捉数据的局部空间特征，该模型经常被用于预测具有空间特性的网络流量序列。Lazaris A 等人采用了长短期记忆（Long Short Term Memory, LSTM）以及改进的神经网络^[30, 31]。他们开发了一个基于特征的聚类框架，提出一个预处理步骤，将相似的时间序列分组在一起，并为每个组训练单个模型，从而进行预测任务的实现。Xu F、Lin Y 等人提出一种基于模拟退火（Simulate Anneal Arithmetic, SAA）优化的自回归综合移动平均模型和反向传播神经网络预测方法，该方法综合利用线性模型 ARIMA、非线性模型 BPNN 和优化算法 SAA 实现网络流量预测任务^[32]。

LAN X 和黎嘉诚用 BP 神经网络^[33]、小波神经网络^[34]等神经网络进行网络流量预测。TANG F 等人基于卷积神经网络提出了一种智能流量控制算法，名为 ST-DeLTA，它利用时空价值网络辅助深度学习^[35]，该算法针对网络流量进行处理，并在此操作的基础上进行预测。Vinchoff C 等人使用一个非线性的 GCN-GAN 模型来预测光网络中的突发流量，其通过模拟三个不同的网络状态，如稳定网络、单突发和双突发。单突发表示网络经历快速流量峰值，然后稳步下降，双突发表示网络经历快速流量峰值，然后出现意料之外的更大的流量峰值。他们通过与 LSTM 模型进行比较，验证了该模型在真实光网络中预测这些突发事件的有效性^[36]。Lohrasbinasab I 提出了一种基于统计学习和 ML 结合的方法并对现有的网络流量预测技术进行了拓展研究^[37]。Zhang L 等人提出了一个针对网络流量的端到端在线预测模型^[38]，该模型可分为小波变换和 LSTM 两部分组合而成，同时为避免突发性对模型的负面影响，适应网络流量的变化趋势，他们提出一种神经网络的权值优化算法-滑动窗口梯度下降（Sliding Window Gradient Descent, SWGD）。

韦烜等人基于 Neural-Prophet 模型对大型运营级 IP 网络的流量流向进行直接预测, 并采用改进的损失函数优化模型训练^[39], 预测结果表明 Neural Prophet 模型能够更科学、准确地预测 IP 网络流量矩阵。高志宇等人提出一种基于生成对抗网络流量预测方法^[40]。首先, 他们提出的生成网络模块, 旨在捕获网络流量的时空特征, 这些特征组合在一起, 接着输入到复合残差模块中, 以生成预测结果。而后, 生成的时间序列数值被传送到判别网络模块, 用于区分生成的流量是真实值或是预测值^[40]。最后, 经过生成和判别网络之间的博弈, 提高了产生高精度预测流量的能力。

然而, 由于短时间尺度下的网络流量一般变化更为剧烈, 波动更加严重, 数据更加复杂, 且上述多数研究者做法都是将其放在一个较长的时间尺度下, 先提取数据之间的特征, 再进行预测。线性模型对短期流量预测有较好效果, 非线性模型如 LSTM 模型难以克服梯度的影响, 因此这些模型都具有一定局限性。随着注意力机制兴起, Vaswani 等人提出的 transformer 在计算机视觉和自然语言处理以及时间序列方面的应用^[41], Zhou H 等人针对 informer 的研究在长序列时间预测 (Long sequence time-series forecasting, LSTF) 的应用等研究给网络流量预测问题提供了一些新的解决途径^[42]。Li M、Wang Y 等人提出了一种基于注意机制的无线网络流量预测深度学习方法^[43], 而 Zeng A、Chen M 等人开始利用注意力机制在时间序列相关问题进行了探索^[44]。

1.3 研究目标和研究内容

1.3.1 研究目标

网络流量是评估网络负载和运行状态的关键参数, 其准确预测对于网络管理和优化至关重要。传统的流量预测方法通常使用统计模型或时间序列模型。然而, 由于网络流量极其不确定和复杂, 受多种因素影响, 传统方法在处理这些复杂特征时存在局限, 导致预测结果不准确和不稳定, 或者对于动态变化的网络环境不适用。

近年来, 随着机器学习、神经网络和深度学习等技术的迅速发展, 网络流量预测领域取得了显著进展。这些技术具有强大的模式识别和特征提取功能, 能够从大量数据中学习复杂的非线性关系。然而, 研究人员必须专注于网络流量特性以及特殊问题限制, 设计适用于其复杂性的深度学习模型。因此尽管机器学习和深度学习在流量预测方面取得了成功, 但它们的应用仍面临一些挑战。另外, 长期时间序列中的时间依赖性在这些方法中通常难以可靠地被发现, 因为这些依赖

性可能被时间模式掩盖。同时，神经网络模型的梯度消失和梯度爆炸等也是需要解决的确定性困难问题。

在这一背景下，注意力机制作为一种强大的机制，可以捕捉序列数据中的远程依赖关系，并具有良好的并行计算性能。在自然语言处理和图像处理等领域，注意力机制已经取得了显著的成果。然而，在网络流量预测领域，针对注意力机制的应用及其在网络流量领域的预测性能研究还相对较少，需要进一步的探讨和实证研究。因此，本文的研究目标是探讨注意力机制在网络流量预测中的应用，并通过改进现有的网络流量预测方法或开发新模型，在短时间内准确预测多个时间点的网络流量。具体而言，将进行以下研究：

首先，详细研究目前流量预测方法的局限性，分析它们在处理网络流量复杂性方面的不足。通过比较分析，认识到注意力机制的重要性，然后系统介绍机器学习和深度学习在网络流量预测中的应用，包括目前广泛采用的模型和方法，为理解当前研究水平奠定基础。

其次，详细阐述了注意力机制的原理和在其他领域的成功案例，重点关注注意力机制如何从微观的角度分析序列数据中的重要特征和相关性。其中，在上述理论分析的基础上，我们设计和实现一种基于注意力机制的预测模型，适用于网络流量预测。模型的主要研究问题是怎样引入注意力机制处理流量数据的时间依赖性和复杂性。特别关注模型的结构设计、训练方法和参数调整的研究。通过实验证明和比较，评估提出模型在网络流量预测中的性能。

最后，研究了注意力机制在网络流量预测中的潜力，以促进提高预测的准确性和稳定性。同时，总结了研究的主要结果，以及模型的优点和限制，并提出未来研究的方向。这将为未来网络流量预测领域的研究方向和创新提供一定的参考。

1.3.2 研究内容

随着网络流量的随机性和非线性特征变得越来越明显，传统的线性模型在处理网络流量的预测任务时逐渐表现不出足够的准确性。深度学习算法的兴起为网络流量预测提供了新的可能性，然而，由于网络流量的动态相关性和长期预测误差的累积，现有模型仍然面临众多挑战。最近，注意力机制作为一种新技术在众多领域取得了显著的成果，研究人员已经开始探索其在预测领域的应用。因此，本文的重点在于探讨如何引入注意力机制到模型中，以实现网络流量的多阶段准确预测。

本文技术路线如图 1-2 所示，具体的内容如下：

(1) 改进 Transformer 模型：本章的首要目标是改进当前流行的预测模型——Transformer。首先，在 Transformer 模型的基础上通过添加卷积层结构来捕捉流量数据中的上下文数据信息中的依赖关系，从而提供更准确的预测结果。这个改

进旨在增强模型对数据特征的理解，特别增强是对复杂动态相关性的建模能力。其次，引入相对位置表征的注意力机制（self-attention with relative position representation）以增强模型对序列中不同位置的建模能力。在 WIDE 项目里的数据集上测试表明改进后的模型具有较好性能。

（2）从长时间序列理论出发，提出一个自适应注意力机制网络流量预测模型（Adaptive Attention Mechanism, AAM），AAM 模型包括编码器和解码器。在编码器部分，根据 KL 散度量 q 和 k 相似性，计算其运算得分而后将其进行排序，根据排序后分数自动挑选一个数据子集，这样处理使其能顾及网络流量中“历史久远”的流量特征，从而避免了计算注意力过程中数据里一些重要信息的丢失，使其在长序列流量预测问题中能达到更好效果。在解码器部分，采用编码器的输出和经掩蔽处理的数据序列作为输入，防止当前位置关注后续位置。另外，该模型只需要通过一个前向步骤就可以获得长序列输出，而不是逐步进行，模型有效避免了推理阶段的累积误差传播，提升了长序列预测的准确性。

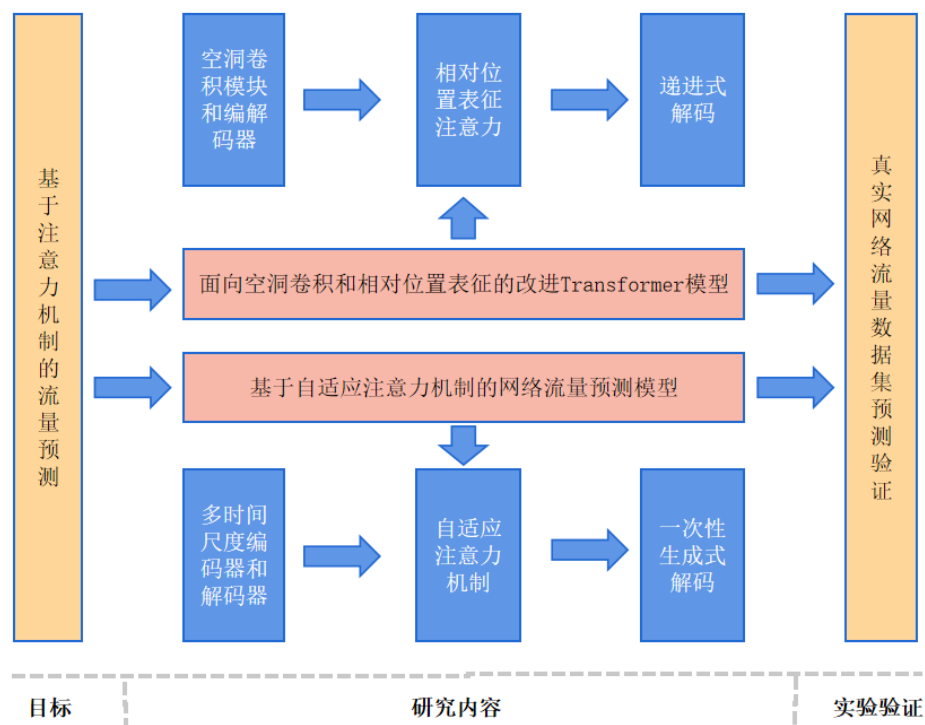


图 1-2 研究路线图

1.4 论文结构和章节安排

本章用五个部分详细介绍了论文的具体情况，下面对各个章节进行简要说明：

第 1 章，首先介绍了网络流量预测的研究背景与研究意义，指出了预测问题的重要性，总结了目前面向网络流量预测问题的国内外研究现状。接着介绍了本文的研究目标，内容以及结构安排。

第 2 章，介绍了时间序列的定义和网络流量特性，同时介绍了网络流量预测的模型 ARIMA, LSTM 以及 CNN 的原理，最后引出注意力机制的相关研究，通过对相关研究的分析，可以发现注意力机制作为一种强大的建模工具，在序列数据建模中取得了显著的成果，并具有潜力应用于流量预测任务中。为本章的研究工作提供理论支持。

第 3 章，提出了一种面向空洞卷积和相对位置表征的改进 Transformer 模型，该模型在 Transformer 的基础上采用空洞卷积和相对位置表征注意力重构编码器和解码器，最后将网络流量结合一维空洞卷积提取流量序列中数据的相似性，在训练模型时利用相对位置表征的注意力机制，使得数据直接的相对位置信息得到表达，与 Transformer 和 CNN 模型在数据集一上的实验证明了所提 TDCRP 模型的预测性能。

第 4 章，针对长序列网络流量预测问题，提出一种自适应注意力机制网络流量预测模型 AAM。该算法采用动态自适应的方式处理输入数据，缩减计算量的同时又能减少数据的信息损失。首先，详细介绍模型构建过程和实现自适应注意力机制模型流程。而后，在两个实际的网络流量数据集上验证本章方法的有效性，最后与两种经典预测问题模型 ARIMA 以及 LSTM 在作对比，以此证明模型的性能。

第 5 章，总结了本文的主要工作和论文的主要研究成果，并对模型的特点进行了分析，同时展望和分析了未来可能的研究方向。

2 相关理论基础

2.1 时间序列预测

时间序列（Time Series, TS）是一种按时间顺序排列的数据序列，代表了同一指标的数据随时间的变化情况。时间序列分析的目的在于从过去的数据中发现数据模式、趋势和周期性，以便预测未来的发展趋势。这种分析方法不仅在经济学和金融领域得到广泛应用，还在气象学、工程学、医学、社会科学等各个领域具有重要意义。通过研究时间序列数据，人们可以更好地理解数据的变化规律，从而为决策提供可靠的依据。尤其是随着 5G、SDN、NS 等技术的飞速发展，针对网络流量预测问题，需要通过创建合适数学模型或利用 ML 相关模型，来实现预测未来的数据变化趋势的目的。

长序列时间预测是时间序列预测问题中一项具有挑战性的任务，随着网络规模的增大和技术的发展，网络流量的时间序列数据变得更加庞大和复杂。解决长期预测问题需要更为复杂的方法和模型处理长时间范围内的变化和波动。因此，研究人员需将时间序列分析与先进的建模技术结合，特别在网络流量的长时间序列预测任务时，需从数据中提取更详细的特征，建立更适合长序列预测的数学模型。

2.2 网络流量相关特性

在进行网络流量预测之前，对网络流量的特征和当前问题有全面的了解可以为构建更有效的预测模型打下基础。以下 3 个方面详细讨论了网络流量的特征和该领域当前面临的问题，主要涉及实时性，突发性以及预测模型的有限性。

首先，网络流量是实时的。这表现在网络流量具有非结构化和非线性特性的特征。由于庞大的网络流量数据量，实现网络流量的预测是一项非常具有挑战性的任务，特别是对高实时性能有要求的应用场景，如在线游戏和视频通话。传统的预测模型可能面临处理大量数据和计算效率不足的问题。因此，需要更灵活、更高效的方法来满足网络流量的实时需求。

其次，网络流量是突发的。不同的应用和用户具有不同的使用特性，例如，用户行为的差异可能导致整个网络流量具有随机和突发的特征。因此，传统模型通常难以捕捉这种多样性，导致预测结果的不稳定。在动态的网络流量环境中，

需要找到更适应不同应用和用户行为挑战的适应性更强、更稳健的模型，而一些神经网络模型不具备这样的能力。

最后，网络流量预测模型的效果是有限的。许多现有的模型和方法在预测过程中只能应用于相似的数据空间，即网络流量的数据结构相似的情况。然而，总体而言，网络流量的时间和空间分布相对复杂，在不同场景中的流量特性存在明显差异。这使得在不同网络环境中应用现有模型可能会面临困难，因此，需要更多样化、更具适应性的预测模型，以便更好适应网络流量的复杂性和差异性。

综上所述，网络流量预测领域面临实时性、突发性和模型适应能力方面的许多挑战，对网络流量进行预测是热点也是难点。为了解决这些问题，研究者需要开发更灵活、更高效以及更具适应性的预测模型上，以更好地满足不同应用场景对网络流量预测的需求。

2.2.1 ARIMA 模型原理

经典模型 ARIMA 通过差分将非平稳过程转化为平稳过程解决预测问题^[45]。差分整合移动平均自回归模型，又称整合滑动平均自回归模型，经常用于具有一定趋势性的预测任务。自回归移动平均模型需要内生变量并且不需要借助其他外生变量。自回归移动平均模型中有 p , d , q 三个参数， p 代表自回归的项数， q 代表滑动平均的项数， d 代表差分次数。自回归移动平均模型的流程图如图 2-1 所示。

同时，ARIMA (p , d , q) 可定义为式 (2-1)：

$$\Delta y_t^d = c + \sum_{i=1}^p \phi_i \Delta y_{t-i}^d + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (2-1)$$

其中 Δy_t^d 为差分变量， c 为常数， ϕ_i 为自相关系数， ε_t 是白噪声序列， θ_i 是时间序列中随机项的当前和先验值的权值。本研究用两个数据集测试了 ARIMA 模型，模型的参数 p , d , q 使用网格搜索策略确定。

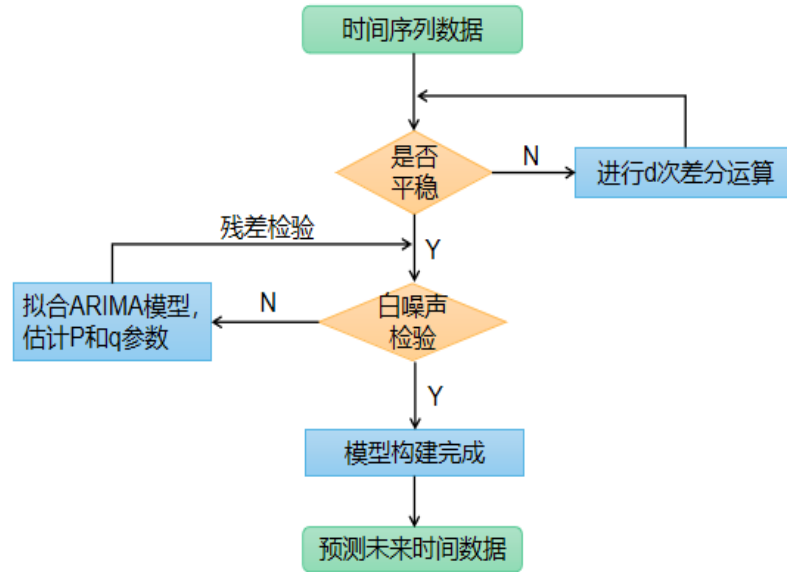


图 2-1 ARIMA 模型

2.2.2 CNN 模型原理

卷积神经网络(Convolutional Neural Network, CNN), 是一种深度学习模型。与传统神经网络相比, CNN 处理图像和序列数据的能力更好, 它能够自动学习数据的特征, 并从中提取出最有用的信息。卷积操作可以有效地减少权重数量, 卷积神经网络的关键特性之一是其卷积操作。这种操作通过在输入数据上使用滑动窗口执行计算, 利用滤波器(也称为卷积核)来提取数据的特征。同时, CNN 还利用最大池化层(Max Pooling)来进行特征下采样, 以降低数据的维度并保留最重要的特征。这种处理方式有助于网络更有效地学习和理解复杂的数据模式, 保留数据的空间结构信息。另外, CNN 还能够不改变特征图维度的前提下, 减少计算量, 提高模型的鲁棒性。Liangzhi Li 等人对输入数据进行预处理, 样本进行归一化, 将加载数据映射到[0, 1]区间, 归一化后生成的图像使用 RGB 颜色来映射负载值(将时间序列数据转换为图像数据达到利用 CNN 模型的目的)^[46]。而后用输入向量与权值向量之间的距离来寻找输出层的响应神经元, 以此实现对时间序列数据的预测。

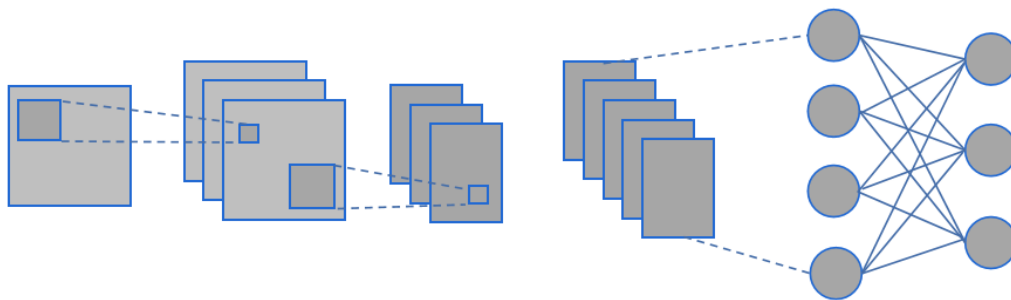


图 2-2 CNN 模型

如图 2-2 所示，这是一个简易卷积神经网络结构示意图。首先，输入数据经过第一层卷积操作，产生了深度为 3 的特征图。接着，经过池化操作得到了深度为 3 的第二层特征图。接着，这一过程重复进行，生成了第四层的深度为 5 特征图。最后，这五个特征图被展平连接成向量，并输入到全连接层。全连接层实际上是一个反向传播神经网络，其中每个特征图可以被视为排列成矩阵形式的神经元。最后经过 BP 神经网络处理输出预测结果。

2.2.3 LSTM 模型原理

LSTM 是一种常见的循环神经网络结构，用于处理具有序列结构的数据，实质上是从 RNN 进化而来，能解决 RNN 中出现的梯度消失问题^[47]。LSTM 具有较好的长期依赖性建模能力和防止梯度消失的能力，本文实验中 LSTM 的训练使用反向传播算法和梯度下降算法。它包括三个门控单元：输入门，遗忘门以及输出门。长短时记忆网络细节图如图 2-3 所示，具体构成和说明如下：

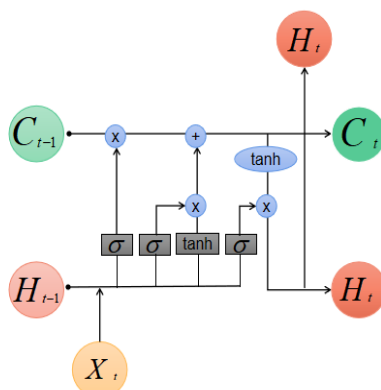


图 2-3 LSTM 模型

模型输入有三种形式：细胞状态是 C_{t-1} ，隐层状态是 H_{t-1} ，时刻输入向量为 X_t ，输出有两个：细胞状态是 C_t ，隐层状态是 H_t 。LSTM 单元是由输入门，输

出门和遗忘门组成。一般情况下，LSTM 模型可以用以下方程来表示。LSTM 模型可以用以下方程式（2-2）-（2-6）来表示：

$$i^{(t)} = \sigma(W_i x^{(t)} + U_i h^{(t-1)} + b_i) \quad (2-2)$$

$$o^{(t)} = \sigma(W_o x^{(t)} + U_o h^{(t-1)} + b_o) \quad (2-3)$$

$$f^{(t)} = \sigma(W_f x^{(t)} + U_f h^{(t-1)} + b_f) \quad (2-4)$$

$$c^{(t)} = f^{(t)} * c^{(t-1)} + i_t * \sigma(W_c x^{(t)} + U_c h^{(t-1)} + b_c) \quad (2-5)$$

$$h^{(t)} = \sigma(c^{(t)}) * o^{(t)} \quad (2-6)$$

在上述方程中， i 、 o 、 f 、 c 、 h 分别为输入、输出、遗忘门、当前状态和隐藏状态。 W 表示输入权值， U 表示门的循环权值， b 为偏差向量。在本文工作中，将LSTM 模型应用于网络流量预测，并将其转化为有监督学习问题处理。

2.3 注意力机制原理和应用

自从深度学习迅速发展以来，注意力机制（Attention Mechanism, AM）已成为语音分类、语音识别、图像检测、自然语言处理等领域中广泛采用的核心技术之一^[48]。在人类世界中，当一个情形映入眼帘，人们往往会先关注其中感兴趣的重点内容^[49]。人眼和大脑处理信息时候，就有注意力的体现，人们往往在意需要被在意的东西。例如，当人们看到一张照片时，他们会特别关注可能更倾向于人形特征区域，并忽略其他部分。在日常生活中，我们面对大量输入的信息。我们之所以能够有效处理这些信息，是因为我们的大脑能够筛选出小部分有用的信息进行重点处理。

AM 结构的明显优点是关注相关的信息并且忽略不相关的信息，不需多数神经网络中的循环结构就能直接创建输出与输入之间的依赖关系^[50,51]。以 AM 为基础的注意力机制相关模型能更好地处理时间序列数据，从而提升捕捉数据中远程依赖信息的能力，这一特性能提高模型的精度^[52,53]。下面介绍 AM 的发展和应用历程。

（1）注意力机制首先被提出和应用在视觉图像领域^[54]，它使用 AM 将传统的视觉搜索方法进行优化，降低了需处理的样本数。Google 团队将 RNN 模型与 AM 结合^[55]，利用注意力机制对特定区域进行处理。实验证实，该团队利用注意力机制简化了图像特征提取的过程，缩短了数据处理时间，并同时保留了关键信息。这使得基于注意力机制的 RNN 在处理图像信息方面比传统的循环神经网络更加高效。

(2) 随着注意力机制在视觉领域取得突破, 一些研究人员同样将这种思想应用在自然语言处理领域。最初, CHO K 等人提出了 Encoder-Decoder 结构^[56], 用于机器翻译。该模型利用两个循环神经网络结构, 其中第一个循环神经网络结构的作用是将输入序列编码成固定长度的向量, 而第二个循环神经网络结构的作用是将该向量解码成目标标记序列。这意味着编码器-解码器机制通过编码将输入序列转换为包含特定信息的向量, 然后通过解码器将该向量翻译为输出信息^[56]。然而, 这种方法只适用于固定长度的句子, 当句子长度不同时, 性能会受到影响。此外, 针对不同情况和语言, 需要选择不同的编码器和解码器^[57]。随后, Bahdanau 等人首先将 AM 引入机器翻译中, 以实现对齐和翻译的同步进行, 解决了句子长度不同的问题^[58], 他们提出的模型不再将所有输入编码为定长向量, 而是在解码过程中利用注意力机制选择输入向量的子集。

(3) 数据预测在各行各业中都举足轻重^[59]。由第一章的国内外研究现状可知, 在网络流量预测领域, 最初的预测问题由经典的线性预测发展到用与神经网络和深度学习相关的预测^[60], 现如今一些学者进行了注意力机制方面的尝试。在 2017 年, Vaswani 提出了 Transformer 模型^[41], 将注意力机制推向时间序列预测方向的高潮。

综上所述, 深入研究注意力机制在网络流量预测中的应用对于改进流量预测的准确性和效率具有重要意义。通过利用注意力机制的优势, 可以更好地捕捉复杂的非线性动态变化和时间依赖关系, 提高网络流量预测的准确性和稳定性。这将为网络规划、资源分配和服务质量保障提供有价值参考, 帮助网络服务提供商去更好管理和优化网络性能。同时, 研究结果还可以为网络领域的学术研究和实际应用提供新的思路和方法, 推动网络流量预测技术的进一步发展。

2.4 本章小结

本章首先针对时间序列问题做出介绍, 接着, 从网络流量相关特性和流量预测方面容易遇到的问题做出说明。其次, 介绍本章后续所用到非常经典的对比实验模型。最后对主要工作相关的注意力机制进行整体细致阐述, 为后续工作提供支撑。

3 面向空洞卷积和相对位置表征的改进 Transformer 模型

3.1 引言

由第一章的绪论和第二章的相关理论基础介绍可知,随着网络规模和各种网络技术的飞速发展,网络业务的复杂性和突发性日益增强,传统的线性模型适合小规模且具有一定规律相似性的网络流量序列,此类回归模型不能较好地描述网络流量特性,存在时滞性。

基于神经网络和深度学习模型在预测领域均有较好应用前景,但这种以知识为导向的用数学模型驱动方法通过对算法特定的优化或者用策略求解,从而得到模型最优解,这些方法缺少从数据中学习特征的能力。分析了网络流量预测的相关研究,发现注意力机制作为一种强大的建模工具,在序列数据建模中取得了显著的成果,有潜力应用于流量预测任务中。随着注意力机制的兴起,Vaswani 提出了 Transformer 模型是注意力成熟应用的一个阶段性体现。何迎利等人在传统时序序列预测模型的基础上设计实现了一种局部上下文信息增强的注意力机制,将提出的注意力机制分别与长短期记忆人工神经网络和门控循环单元两个时序预测模型相结合并将结合后的模型用于某运营商提供的两个不同网络流量数据集进行网络设备流量预测^[61]。

因此,本章从微观角度利用注意力机制对时间序列数据进行解释,提出一种面向空洞卷积和相对位置表征的 Transformer 改进模型进行网络流量问题预测。

3.2 问题定义

在基本时间序列模型中,网络流量预测问题本质是时间序列问题。带宽资源大小影响一个网络切片在一段时间内传输字节量。同时因为一段时间内带宽资源需求很难确定,相比较字节的时间序列数据容易收集。网络流量预测结果一定程度上反应了带宽资源需求。本文将一段时间内通过网络数据量作为一个网络切片的流量时间序列数据,可定义为公式(3-1):

$$X_i = \{x_i^0, x_i^1, x_i^2, \dots, x_i^t\} \quad (3-1)$$

其中 x_i^t 表示切片 i 在时隙 $[t, t + \tau]$ 内通过的字节量，在预测任务中， τ 为时间间隔。根据时间序列数据定义，预测任务是基于其历史的时间序列数据预测后续 n 个网络流量数据： $\{x_i^{t+1}, x_i^{t+2}, x_i^{t+3}, \dots, x_i^{t+n}\}$ ，其中 n 代表第 n 步。

单步预测和多步预测这两种预测方法在处理网络流量预测问题时具有不同的策略和特性，在本章节中，网络流量的预测被我们定义为多步预测问题，

单步预测是指每次预测未来值时只预测输入窗口中的单个值。在单步预测中，有两种不同的策略。第一种策略是将整个真实值集作为输入窗口，这意味着每个预测仅适用于未来的一个值。第二种策略是在预测未来时间点的第一个值时，使用输入窗口中的所有真实值，并且在预测后续的 $n-1$ 个值时，预测窗口包含先前预测的值。这种策略能实现单步预测实现预测未来的多个值的目的，这是一种更灵活的方法。

多步预测是指在每次预测时输入窗口预测未来的 n 个值（也称为 n 步）。多步预测策略也可以分为两种类型。第一种策略是仅预测一次，即仅输入 n 个滑动窗口，然后直接输出未来的 n 个值。在这种情况下，模型使用的输入数据完全采用历史数据的真实值。另一种策略是多次预测，即将 n 个滑动窗口作为一个输入，窗口长度为 m ，并直接输出未来的 $n * m$ 个值。在这种策略中，第一个输入是历史数据的真实值，并且在后续的预测中，滑动窗口包含先前预测的值。这种策略使模型能够在每次预测后使用已知的预测值来进一步优化未来的预测。

通过单步预测与多步预测这两种不同的策略，可以灵活应对不同的预测要求和场景。单步预测适用于需要准确预测未来单个值的情况，而多步预测更适用于需要预测未来一段时间内的多个值的情况，选择合适的预测策略取决于特定的应用场景和预测目标。

3.3 面向空洞卷积和相对位置表征的改进 Transformer 模型组件

本章受到注意力机制成功应用的 Transformer 模型引导，提出一种相对位置表征的注意力机制（attention with Relative Position representation, RP）以及空洞卷积（Dilated Convolution, DC）的改进 Transformer 模型（TDCRP），具体组件见下面详细介绍。

3.3.1 Transformer 模型

Google 团队提出一种新奇的完全基于注意机制的网络架构-Transformer 模型，模型的整体架构如图 3-1 所示^[41]。由图 3-1 可知，Transformer 架构整体是由

一组编码器-解码器组成，由红色矩形框标注的部分表示。**Transformer** 工作原理是编码器把输入数据序列进行输入嵌入，同时进行位置编码后映射为隐藏层，然后解码器通过解码处理将隐藏层与输出序列进行匹配，而后将其映射为输出序列。

编码器可拆分为 4 个部分，如图 3-1 的左侧红色标记部分所示，绿色所框第 1 部分的作用是将输入数据转换为向量形式，而后经过位置编码后输入多头注意力机制。位置编码的作用是标记数据序列之间的连续性和相关性，同时确定数据之间位置的关系。第 2 部分通过多头注意力机制获取数据之间的相关性，克服了卷积神经网络方法中缺少数据相关性的问题。在第 3 部分，模型引入了标准化和残差连接。引入了残差连接和标准化层的目的是提高模型的学习能力并加快数据收敛速度^[62]。最后，在第四部分中，通过由两个完全连接的层组成的前馈反馈层对学到的数据信息进行非线性映射，即增强具有强相关性的部分，减少具有弱相关性的部分，并最终进行统一标准化操作。因此，通过编码器学到的结果更有精准性和代表性。

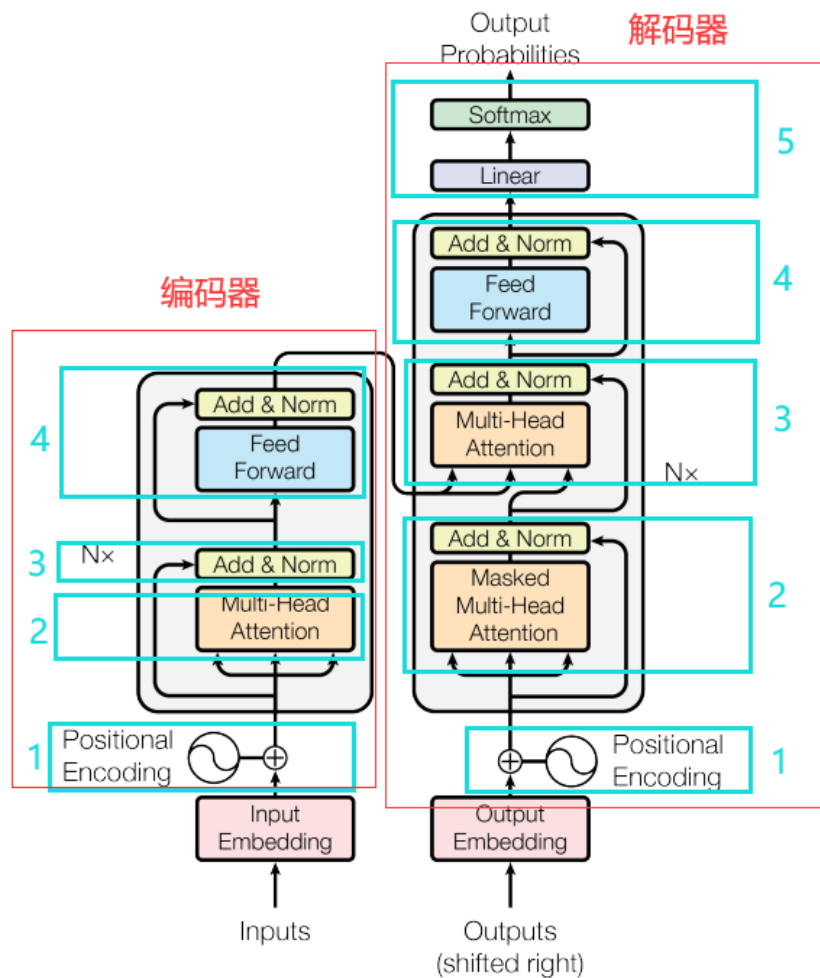


图 3-1 Transformer 模型架构图^[41]

从 Transformer 的结构可以看出，相较于左边编码器的第 2 部分，对应右边解码器具有不同的掩码多头注意力机制。由于解码器和编码器所使用的训练数据长度不相同，解码器必须使用数据中的最大长度作为训练的计算单位，并且只需要关注先前数据对当前数据的影响，而不考虑未来数据的影响。因此，未来的预测数据将被使用一个功能性的掩码填补，以防止其参与训练。解码器第 3 和第 4 部分与编码器相同。最后，线性化过程和softmax层的作用是生成输出。线性化过程和softmax层用于将向量转换为所需的输出类型。

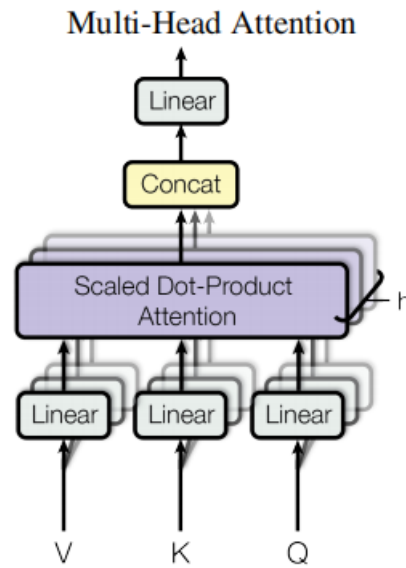


图 3-2 多头自注意力机制^[41]

如图 3-2 所示^[41]是 Transformer 模型中的核心多头注意力机制：其给定相同的查询Q，键K和值V的集合时，多头自注意力机制扩展了模型关注不同位置的能力，模型可以学习到不同的行为，从而实现关注序列内各种范围依赖关系。相较于只使用单个注意力汇集，多头注意力机制通过不同的线性投影更新查询、键和值，这些投影相对独立学习。然后，这些 h 组（ h 头）转换后的查询，键以及值同时输入到注意力机制中。最后，这些注意力输出被合并，并通过另一个可学习的线性投影进行转换，以生成最终的注意力输出。

Transformer 模型在自然语言处理和机器视觉领域已经取得很大进展，然而，在分析时间序列数据时，网络流量数值数据本身与自然语言处理问题不同，它们通常缺乏语义，也就是说秩序本身充当着关键的角色。另外，时间序列与图像序列以及文本序列还有很多不同，例如第二章所介绍的网络流量时间序列的特性。这些都给 Transformer 在时间序列预测情景中的应用带来了新的挑战，因此，单纯的把 Transformer 模型应用到时间序列预测中可能不能取得很好的效果。Zeng A 等人在 Transformer 在长时间序列预测中效果一般，但是 Transformer 模型中使

用位置编码虽然有助于保留一些排序信息，但排列不变性的自注意力机制性质不可避免地会导致时间信息的丢失^[63]。因此本章提出一种空洞卷积和相对位置表示的改进 Transformer 模型（TDCRP）运用于时间序列预测中，同时与原始的 Transformer 模型作对比。

3.3.2 空洞卷积

空洞卷积（Dilated Convolution, DC）是一种卷积操作，也被称为膨胀卷积或扩张卷积。与传统的卷积操作相比，空洞卷积引入了一个称为膨胀率（dilation rate）的参数。这个参数的设置使得在不增加参数量和计算成本消耗的情况下扩大了感受域。

传统的卷积操作，卷积核在输入特征图上滚动进行卷积运算。每一个卷积核种元素只与输入特征图上相对应位置的元素之间进行运算，而后生成输出特征图。这种操作一定程度上限制了感受域范围，即卷积核只能获取最近最相邻位置的信息。然而，在一些问题情景里，需扩大信息范围来更好地理解数据特征之间的关系。而在空洞卷积中，卷积核的元素之间通过引入膨胀率的概念来调整它们之间的间距。膨胀率定义了卷积核元素之间的空洞，用于控制卷积核在输入特征图上采样的步长。图 3-3 展示了一般卷积的过程，其卷积核大小为 3×3 ，步长 $s=1$ 。

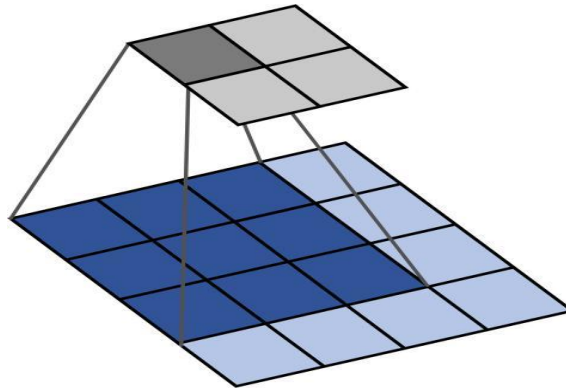


图 3-3 一般卷积示意图

具体而言，在进行空洞卷积操作时，卷积核中的每个元素都与输入特征图的不同位置开展点积运算，但它们之间的间距不再设为 1，而是根据膨胀率进行调整，膨胀率定义了卷积核元素之间的空洞或间隔，用于调整卷积核在输入特征图上采样的步长。较大的膨胀率能增加卷积核元素之间的间距，从而扩大感受域。采取此种方式可以在不增加额外参数和计算成本的情况下增加感受域的范围。例如，对于一个 3×3 的卷积核、步长为 1 以及膨胀率为 2，卷积核的中心元素将与输入特征图上间隔一个像素的位置的元素进行点积运算，如图 3-4 所示。通过这种方式，输出特征图的每个像素都可以感受到输入特征图中更广阔的上下文信息。

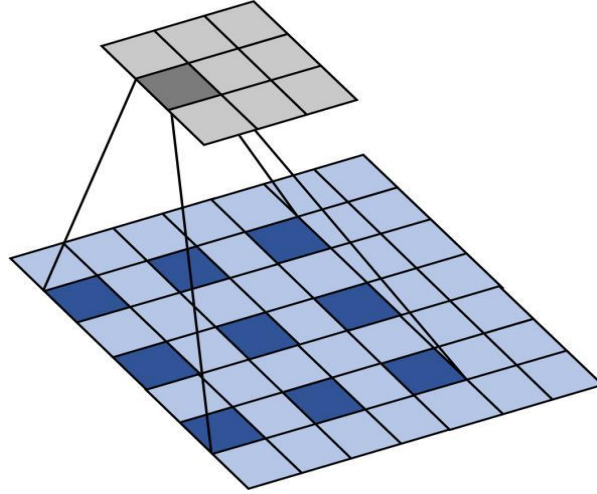


图 3-4 空洞卷积示意图

膨胀因子是衡量普通卷积和空洞卷积之间的区别所在。当膨胀因子为 1 时，表示卷积核的各个元素之间没有间隔，即表明相邻两个元素的位置之间的距离为 1，这代表了普通的卷积操作。当膨胀因子为 2 时，表示卷积核各元素之前有一个间隔，即相邻两个元素间的位置相差 2，此时就是图 3-4 所示的卷积核。空洞卷积的卷积核大小为 k ，空洞数为 d ，则其等效卷积核大小 k' 计算如式 (3-2) 所示：

$$k' = k + (k - 1) * (d - 1) \quad (3-2)$$

k' 代表等效卷积核尺寸， k 代表实际卷积核尺寸，而 d 指的是 dilation，是空洞卷积的参数大小。空洞卷积在应用有多个优点。首先，它可以增加感受野，使网络能够更好地理解输入数据的上下文信息。其次，通过调整膨胀率，可以灵活地控制感受野的大小，适应不同尺度的对象或结构。此外，空洞卷积还可以减少网络的参数量和计算成本，由于空洞卷积不需要增加额外的参数，因此可以在保持模型规模较小的同时提供更大的感受野。

在本章的研究中，我们提出了一种改进的 Transformer 模型，模型具有空洞卷积和相对位置表征两个组件。首先，时间序列数据被送入一维空洞卷积网络中，利用卷积操作来提取每个节点周围的信息。接着，我们利用多头注意力机制来学习节点之间的关系。这种方法使得注意力不仅关注每个节点的数值，还能考虑数据周围的上下文信息，从而建立起具有相似形状的区域之间的联系^[64]。空洞卷积层位置对数据处理位于图 3-1 处的左边编码器和右边解码器中的数字 1 层和 2 层之间。

3.3.3 注意力机制之相对位置表征

对 Transformer 模型而言,输入数据要添加绝对位置表示,如图 3-1 中的 input Embedding 所示。在本章工作中,考虑利用相对位置表征 (self-attention with relative position representation) 结合到 Transformer 的自注意力机制中,来达到扩展自注意力的目的,能有效地考虑相对位置间信息的重要性。普通的自注意力的输出表示为式 (3-3), (3-4), (3-5):

$$Z_i = \sum_{j=1}^n \alpha_{ij} (x_j W^V) \quad (3-3)$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^n \exp e_{ik}} \quad (3-4)$$

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}} \quad (3-5)$$

在普通自注意力的基础上,引入两表示相对位置相关关系的向量: $\alpha_{ij}^V, \alpha_{ij}^K \in \mathbb{R}^{d_\alpha}$,之所以采用 d_α 维向量的表示形式,是为了对齐 Transformer 中自注意力初始公式,方便计算。换句话说,如果我们的注意力焦点是在位置 x_i ,那么在计算 x_j 对 x_i 的注意力特征时,我们需要另外考虑与 x_i 位置有关系的向量 x_j 。因此,自注意力的计算可以修改为式 (3-6), (3-7), (3-8):

$$Z_i = \sum_{j=1}^n \alpha_{ij} (x_j W^V + \alpha_{ij}^V) \quad (3-6)$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^n \exp e_{ik}} \quad (3-7)$$

$$e_{ij} = \frac{x_i W^Q (x_j W^K + \alpha_{ij}^K)^T}{\sqrt{d_z}} \quad (3-8)$$

相对位置表征的目的是给出 $\alpha_{ij}^V, \alpha_{ij}^K$ 的计算方式, $\alpha_{ij}^V, \alpha_{ij}^K$ 只与相对位置有关。假设序列中两个元素的距离超过 n (n 设置为 6),就表明关注这两元素之间的位置信息就没有意义了,即不考虑其相对位置数据特征可能存在的关系。

注意力机制相对位置表征是针对一种自注意力机制的变种,它引入了向量里相对位置表示来捕捉序列中位置之间的关系。通过将相对位置向量与序列中的其他位置进行比较,可以调整注意力权重以更好地利用相对位置信息。相对位置表征注意力替换图 3-2 中的编码器里标记为 2 的绿色方框中的 Multi-Head Attention。

3.4 性能评估

3.4.1 数据集

本章使用的数据集是用 Wireshark 工具从 WIDE 项目获取^[65], 时间从 2020/8/1 0:00 开始至 2022/8/31 23:50 结束。每条数据以 10 分钟时长为时间单位, 共计 109585 条数据记录。数据格式是包括 6 个 IP 端口的数据量以及一个总流量数据, 同时每个记录包括日期、开始时间和结束时间、ID 号等标签。数据集的整体情况如图 3-5 所示。

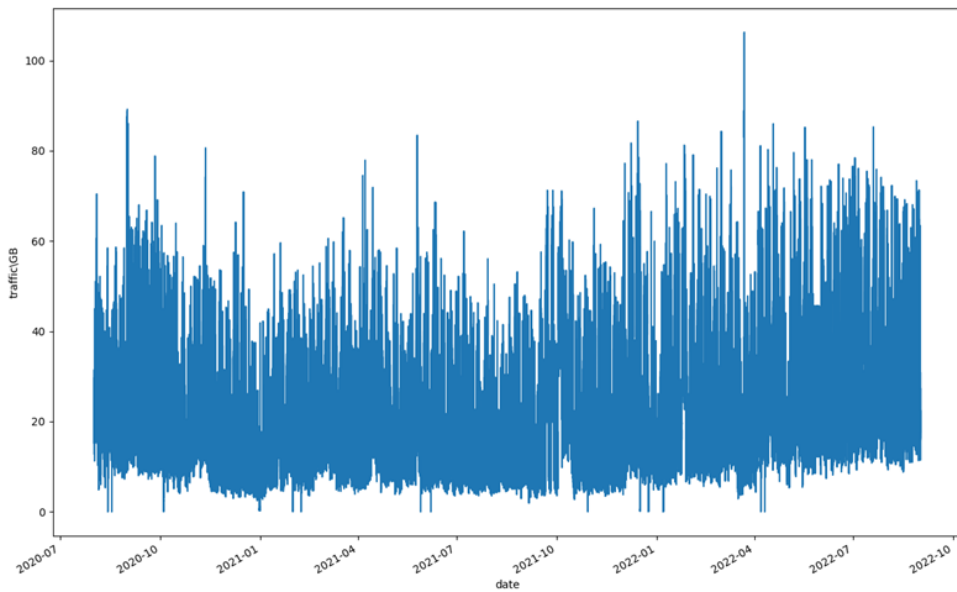


图 3-5 数据集

通常来说, 在正式进行模型训练和预测之前需要对数据集进行适当数据预处理工作, 主要包括填充缺失值和对异常值的识别与剔除。对于本文所用数据集, 缺失值和异常值采用此数据点前后两位, 共四位数据的均值进行填充。

3.4.2 实验环境和评价指标

(1) 实验环境和参数设置

表 3-1 实验环境

名称	型号
操作系统	Windows10 64bit
CPU	i7 10700k
GPU	NVIDIA GeForce RTX 1080TI
内存	1020G
编程环境	Python3.7

实验软硬件环境如表 3-1 所示，首先将处理过的网络流量数据输入本章所提 TDCRP 模型，而后对模型进行训练。

(1) 激活函数选择：激活函数（Activation Function，AF）是神经网络中的一个关键组件，用于在处理非线性问题，它的作用是对每一层的输出进行处理。通过引入激活函数，神经网络就可以引入非线性因素，从而更好地学习数据中蕴含的复杂模式。

在常见的激活函数中，Sigmoid 函数和 ReLu 函数比较常见。Sigmoid 函数图像如图 3-6 所示，其输出范围在 0 到 1 之间。Sigmoid 函数可以将输入特征值的大范围转换到 0 到 1 之间，这有助于在深层网络中确保数据幅度的稳定性。然而，当输入特别大或特别小时，Sigmoid 函数的输出接近于常数，导致梯度接近于 0，从而可能导致梯度消失问题，进而降低了收敛速度。ReLu 函数图像如图 3-7 所示。当输入为负值时，ReLu 函数的输出是 0，而在输入大于 0 时，输出值等于输入值。相较于 Sigmoid 函数，ReLu 消除了梯度消失的影响，使得深层网络更易于训练，并且收敛速度远远快于 Sigmoid 函数。因此，本章选择 ReLu 作为模型的激活函数。

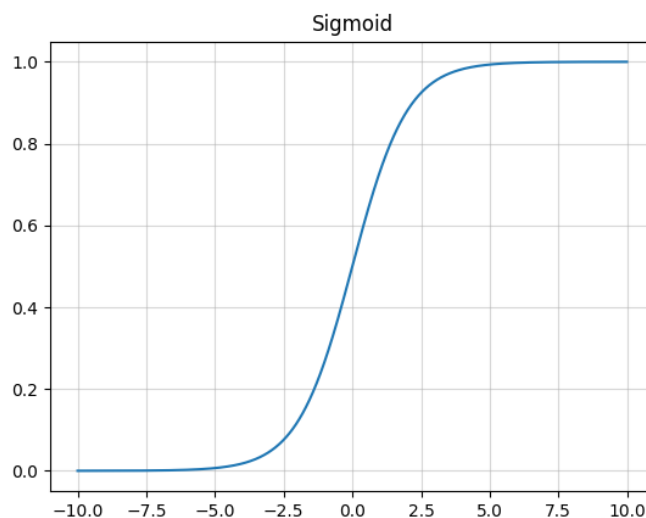


图 3-6 Sigmoid 激活函数

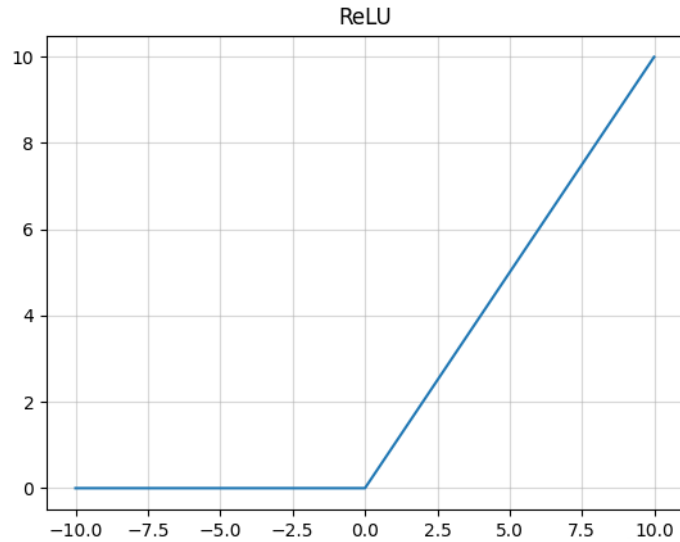


图 3-7 ReLu 激活函数

(2) 图 3-8 显示了学习率超参数 r 对训练损失的影响，其中损失采用了均方误差 (MSE)。从图中可以观察到，在训练开始阶段，所有学习率数值对应的损失都较大，随着训练轮次 (epoch) 的增加，损失逐渐减小。特别是在 $r=0.0001$ 时，验证损失的收敛速度比 $r=0.01$ 和 $r=0.001$ 更快。然而，随着训练轮次的增加，存在全局最优的最小训练损失，这可能导致模型陷入局部最优的问题。因此，鉴于 TDCRP 模型的性能和速度，将学习率大小设置为 0.0001。

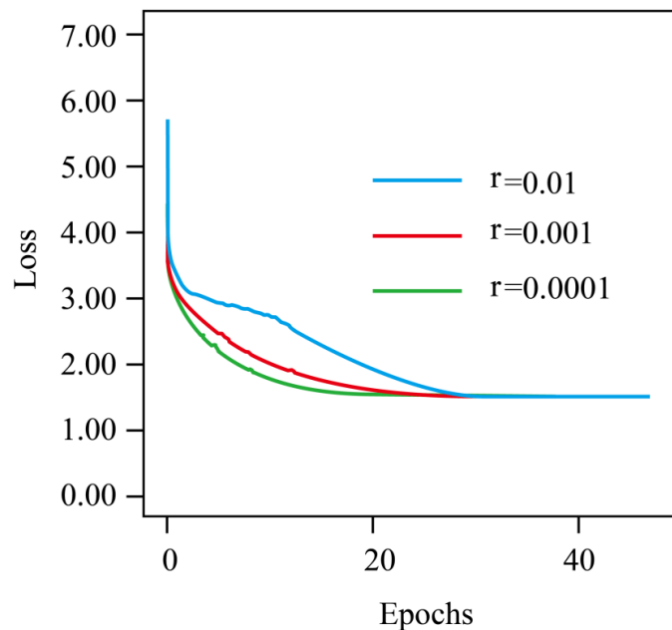


图 3-8 学习率

另外，针对模型优化器的选择，随机梯度下降 (Stochastic Gradient Descent, SGD) 和自适应优化器 (Adaptive Moment Estimation, Adam) 是两种常用优化

算法。SGD 是一种简单经典的优化算法，它通过在每个训练样本上迭代训练，逐次更新模型参数来达到最小化损失函数的目的，随机梯度下降优化器对学习率比较敏感，需要手动调整学习率的大小。此外，SGD 容易陷入到局部最优解，并在损失函数有大量局部极值点时，收敛较慢。相比之下，Adam 结合了梯度的一阶矩估计和二阶矩估计，具有较好收敛性能且更容易应用。因此，本章所提的 TDCRP 选择 Adam 作为模型优化器。

模型除了可学习的参数外，还需要手动设置指定的参数，这些参数称为超参数，会影响模型的学习效果。考虑到 TDCRP 模型的预测性能，本章选择空洞卷积核的大小为 3，选择 Adam 优化器，最小批处理为 24，学习率为 0.0001，迭代 epoch 数为 40。具体实验中设置如表 3-2 所示：

表 3-1 参数设置

名称	值
迭代 epoch 次数	40
学习率	0.0001
损失函数	MSE
优化器	Adam
批处理大小	24

(3) 评价指标

平均绝对误差(Mean Absolute Error, MAE)、均方误差(Mean Squared Error, MSE)和常用的评价指标，用于衡量预测模型在回归问题中的性能。因此本章也采用 MSE 和 MAE 来衡量模型性能。

平均绝对误差：MAE 是预测值与实际值之间差值的绝对值的平均值。在计算的时候，先对真实值与预测值求和，然后再取平均值。它衡量了预测结果与真实值之间的平均偏差程度，其计算公式如(3-9)所示：

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (3-9)$$

其中，n表示样本数量， y_i 是真实值， $\hat{y}_i$ 是预测值。当 MAE 的值较小时，表明模型的预测准确度较高。

均方误差：MSE 是预测结果与真实值之间的平均方差。其计算公式如(3-10)所示：

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (3-10)$$

其中，n表示样本数量， y_i 是真实值， $\hat{y}_i$ 是预测值。MSE 的值越小，表示模型预测精度越高。

均方误差(Root Mean Squared Error, RMSE): RMSE 是真实值和预测值之间误差的平方的平均值。也是衡量预测结果与真实值之间的偏差,其计算公式如(3-11)所示:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3-11)$$

其中, n 表示样本数量。 y_i 是真实值, $\hat{y}_i$ 是预测值。RMSE 的值越小, 代表模型的预测精度越高。

3.4.3 结果与分析

本章所提出的 TDCRP 与原来的 Transformer 模型和 CNN 做一个比较。采用 WIDE 数据集, 使用长度时间从 2020/8/1 0:00 开始至 2022/8/31 23:00 每条数据以 10 分钟时长为时间单位, 用 90% 做训练数据, 10% 做测试集。目的是预测 12, 24, 48 个步长时间点的结果, 预测时间点从 2022/6/16 20:50 开始。

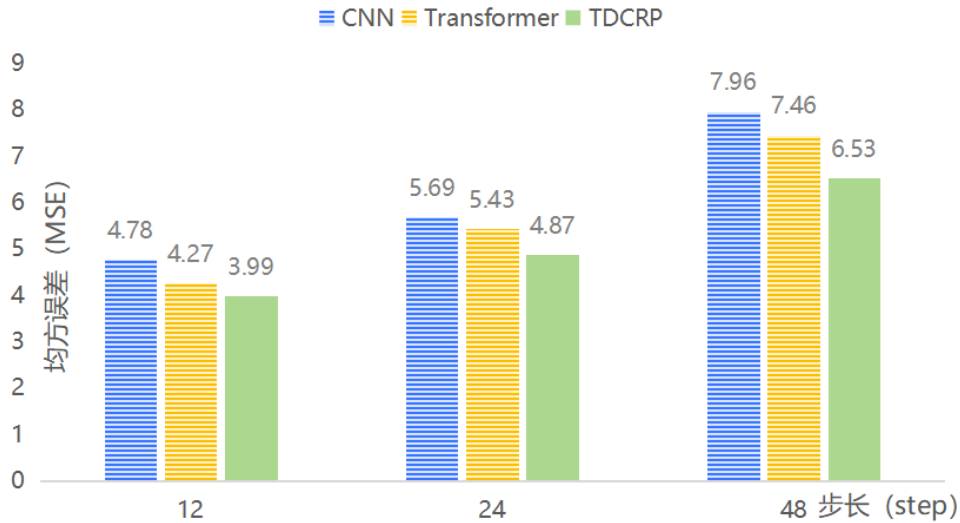


图 3-9 TDCRP、Transformer、CNN 实验 MSE 结果图

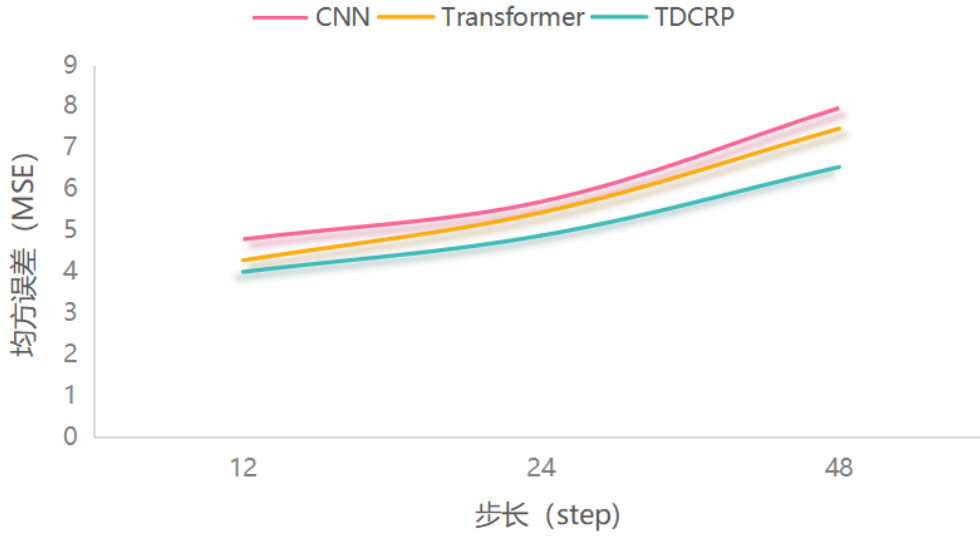


图 3-10 TDCRP、Transformer、CNN 实验 MSE

如图 3-9 和图 3-10 所示是预测 12, 24, 48 个步长时间点的 MSE 结果, 可以看出本章所提出的面向空洞卷积和相对位置表征的改进 Transformer 模型-TDCRP, 能用空洞卷积扩大感受野, 提高了模型对流量数据的特征关注能力, 与 Transformer 模型和 CNN 模型相比, 在短期网络流量预测上表现出较优性能, 但是随着预测步长的增加, 模型的均方误差指标迅速增大。

3.5 本章小结

在应对网络流量的时间序列预测问题, 本章提出一种面向空洞卷积和相对位置表征的改进 Transformer 模型, 旨在解决网络流量的时间序列预测问题。该模型在 Transformer 的基础上进行改进, 采用空洞卷积和相对位置表征注意力构成编码器和解码器, 首先将网络流量结合一维空洞卷积提取流量序列中数据的相似性, 其次在训练模型时利用相对位置表征改变了 Transformer 中原本的自注意力机制, 使得数据直接的相对位置信息得到表达, 从而提升模型的性能。本章的主要贡献如下:

(1) 本章所提出的面向空洞卷积和相对位置表征的改进 Transformer 模型-TDCRP 模型, 能用空洞卷积扩大感受野, 捕获流量数据上下文信息 (当前位置的其余时间点), 与典型的 Transformer 中注意力机制逐个计算点积不同, 这种方式能通过流量数据的上下文信息计算它们的相似度, 提高了模型对流量数据的理解能力, 从而提高预测精度。通过对网络流量中的数据相似性进行更全面和深入的分析, 改进模型能够更准确地捕捉到流量数据中的规律和变化趋势。

(2) 本章所提 TDCRP 采用相对位置表征注意力机制, 考虑到流量数据的时间间隔为 10 分钟, 故将模型中两个数据元素的相对位置 K 设为 6 (假定超过一小时内的流量数据没有意义, 不做考虑), 以便更好地捕捉流量数据中的时间依赖关系。其生成的 Q 和 K 矩阵能在关注流量数据相对位置在多头注意力中的权重, 即相对位置信息对当前位置而言的重要性。在公开数据集上 WIDE 上进行测试, 实验表明, 与 Transformer 和 CNN 模型相比, 本章所提 TDCRP 模型在短期网络流量预测上表现出较优的性能, 但是随着步长的增加, 模型的 MSE 指标快速增加, 表明模型在进行网络流量长序列预测时还有一定的局限性。

综上所述, 本章提出的面向空洞卷积和相对位置表征的改进 Transformer 模型具有一定的优势和创新点。通过引入空洞卷积和相对位置表征注意力机制, 改进模型能够更好地处理网络流量数据的特性, 提高了预测的准确性。然而, 模型在处理网络流量长序列时间预测问题时仍存在一定的挑战, 需要进一步优化和改进。

4 基于自适应注意力机制的网络流量预测模型

4.1 引言

针对第三章的工作,虽然所提 TDCRP 模型与 Transformer 和 CNN 模型相比,通过增加空洞卷积和采用相对位置表征的注意力机制具有较好的预测性能,但是随着步长的增加,模型针对网络流量长序列时间预测的局限性凸显。因此,如何利用注意力机制进行长序列网络流量预测是一个关键问题。针对此问题,本章继续对注意力机制进行探索和研究,以下是几种注意力机制的深入研究和应用。

Bryan Lim 等人引入一种时间融合的 Temporal Fusion Transformer 模型^[66],一种不同于 Transformer 的基于注意力的新架构,它结合了高性能预测以及对时间的可解释洞察力。为了处理静态协作变量和已知的优先输入,TFT 采用了一系列特殊组件,主要包括以下几个方面:(1) 门控机制,能够跳过未使用的组件,从而灵活地调整深度。(2) 控制变量选择网络,在每个时间步中选择输入变量和输入函数来实现高效的数据处理。(3) 静态变量编码器,通过对流量数据上下文信息向量编码实现对时间动态的处理。(4) 从学习到的和已知的的时间变化中学习长期和短期时间关系的处理。序列到序列和基于时间的处理组件可以关注到不同时间尺度上的时间关系,实现对复杂关系的准确处理。(5) 通过设定预测区域的输出间隔,以确定每个预测阶段可能目标值的范围。TFT 为已知的简单数据集提供了强大的预测性能,尽管在数据曲线的波动较小时,该模型会均衡地对待历史输入,并且在提取重要信息特征方面存在一定的限制。

Yang Lin 等人提出了一种状态分解神经网络(State Space Decomposition Neural Network, SSDNet),用于时间序列数据概率性和可解释性预测^[67]。其主要思想将 Transformer 结构与状态空间模型相结合,考虑其结合深度学习的优势以及注意力机制的特性。SSDNet 主要思想是基于编码器和解码器框架,使用两个关键组件:状态空间网络和 Transformer。SSDNet 的前馈过程包括两个步骤。首先,SSDNet 的 Transformer 结构处理历史时间步长,并生成潜在分量来估计状态空间模型参数以及预测的方差分布。其次,状态空间从前一个时间步长中提取状态向量,并使用它来预测分布的平均值。与经典的状态空间模型通过卡尔曼滤波分解时间序列和估计参数不同,其通过注意机制使用变压器架构学习参数,然后将分解成分(趋势、季节性和概率项)以状态空间的形式组合起来,产生最终预测结果。结果表明其提出的模型都是一种有效的方法,模型通过提取有意义的

趋势和季节性成分来提供准确和具有解释性的预测。但是该模型难以捕获结构化较差的时序数据中的时间模式。

Yan Li 等人把长期时间序列预测和有限的信息利用使得模型无法全面探索复杂的依赖关系作为出发点,提出了一种基于 Transformer 的模型变体—Conformer^[68],它与现有的长期时间序列预测方法有三个区别:(1)在滑动窗口注意和平稳即时循环网络的基础上,提出了一种包含线性复杂度的编码解码器架构;(2)利用归一化流模块,通过直接用平稳即时循环网络中的潜在变量推断输出,进一步提高信息利用;(3)时间序列数据的序列间相关和时间动态被明确建模,以促进后续的自注意力机制。其核心原理在于窗口注意和优化的循环网络的帮助下,在不牺牲预测能力的情况下,降低了自我注意力计算的复杂性,设计了一个归一化模块直接从隐藏状态推断目标序列。首先,在 Conformer 模型输入表示块预处理并相应嵌入输入的时间序列。然后,编码解码器架构从时间序列表示中探索具有窗口关注的局部时间模式,并借助循环网络和时间序列分解来学习底层分布,以生成的方式直接重建目标序列,最终产生预测结果。研究表明其在多变量和单变量预测设置下的长期时间序列预测中取得了良好的效果。但是该模型也难以捕获结构化较差的时间序列数据中的时间模式。

对于时间序列预测问题,在电子商务背景下,产品的发展趋势和季节性变化很大,促销活动等行为严重影响着销售。Xinyuan Qi 等人针对目前预测问题只从历史数据入手,弥补了 Transformer 对未来知识关注的遗漏,提出了一种基于双向 Transformer 的算法—Aliformer^[69],模型是由两个 transformer 结构并行和一个纯知识结构组成。该模型能够利用过去的经验、当前的情况以及未来的见解来预测销售业绩。他们提出了一种以知识为基础的自我关注层,通过利用已知知识的一致性来引导时间信息的传递。此外,他们还提出了一种新颖的未来关注训练策略,以使模型更加重视未来信息的利用。在训练阶段,他们引入了一种跨度掩码,用于指导模型更加关注未来信息。通过在天猫数据上进行实验,证明了该模型可以提高电子商务销售预测的准确性,为时间序列预测问题带来了新的思路和启示。然而,该模型专注于利用和应用未来知识,并不适用于那些未来信息难以预测和获取的任务。

Tian Zhou 及其团队提出了一个解决方案,即将 Transformer 与季节性趋势分解方法相结合。这个方法的提出是为了解决当前 Transformer 在计算上的昂贵以及无法捕捉时间序列数据整体趋势的问题。他们通过使用季节性趋势分解,利用大多数时间序列具有稀疏表示的基础,开发了一个频率增强模型,该模型称为频率增强分解变压器—FEDformer^[70],其结构中的傅里叶增强模块和小波增强模块,允许通过频域映射捕捉时间序列中的重要结构,能代替自注意力机制中的注意力交叉障碍。模型通过随机选择定量的傅里叶分量,在不同领域的基准数据集

上的实验表明了其选择模型的有效性。然而，对于高频率的时间序列数据，该算法总体计算成本比较高，同时可能会生成大量的冗余信息。

针对时间序列较长期预测问题，Haixu Wu 等人提出了一种 Autoformer 的模型，目的是解决 Transformer 的模型在复杂时间模式中发现可靠长期依赖关系的障碍。Autoformer 采用了稀疏版本的点自注意机制，以优化信息利用，从而实现更好的性能^[71]，Autoformer 是一种新奇的具有自相关机制的分解架构。模型利用一种高效的自相关机制，在序列层次上进行依赖性的发现和聚合，并将它作为深度模型的基本内部模块进行不断更新。Autoformer 的设计使其具备了逐步分解复杂时间序列的能力进一步提升了模型的性能。通过实验表明其所提模型可以实现更低的复杂性，具有较好的性能。然而，这个模型过于依赖于寻找时序数据的周期特性，因此不适用于那些周期性较弱的数据集。

Shizhan Liu 等人探索了时间序列的多分辨率表示，提出一种基于金字塔注意的新模型—Pyraformer^[72]，该模型可以有效地描述较低时间和空间复杂度的短时间与长时间依赖关系。模型引入了一种金字塔注意力，具体来说，他们首先利用使用粗尺度结构模块来构建一个 C-ary 树，然后为了进一步捕捉不同范围内的时间依赖性，他们引入了金字塔注意模块，利用金字塔注意机制传递信息。最后，根据下游任务，使用不同的网络结构来输出最终的预测。实验结果表明，该模型在多步预测以及单步预测任务中都表现较优，但是该模型也难以捕获结构化较差的时间序列数据中的时间模式。

Zhou H 等人一个基于高效 Transformer 的模型—Informer，该模型采用稀疏自注意力机制，其核心思想是通过计算自注意力重要程度实现注意力机制的提炼，另外生成式解码器实现较长时间序列预测，其优点是降低了计算复杂度和内存消耗量，与 Transformer 相比更适用于长序列预测。然而，尽管在预测曲线的波峰和波谷处，该模型仍存在较大的误差，而且随着序列长度的增加，对于更高精度的要求也变得更加困难。尽管它降低了计算复杂度，但这是以信息利用率的牺牲为代价的^[42]。

通过以上模型的总结和分析和优缺点对比，受上述注意力机制相关模型的启发^[42, 66-72]，本章针对自注意力机制本身，从 Informer 模型出发，提出了一种自适应注意力模型解决长序列网络流量预测问题。

4.2 自适应注意力机制网络流量预测模型

4.2.1 模型介绍

自适应注意力机制模型（Adaptive Attention Mechanism, AAM），旨在解决网络流量长序列预测问题。模型采用自适应的方式处理输入数据，分析数据之间关系，有效克服了传统模型和非传统模型局限性。尤其是，自适应注意力机制与经典的 ARIMA 模型和 LSTM 模型虽然都根植于时间序列理论，但在方法上有显著区别。传统模型难以有效捕捉网络流量特征之间的差异，这对于设计、训练和性能提升的预测模型都造成了影响。同时，一些神经网络模型例如 LSTM 参数不好确定，容易陷入梯度消失梯度爆炸等问题。自适应注意力机制模型可以有效解决上述问题。

AAM 结构如图 4-1 所示。它的特点在于通过在每个阶段动态自适应选择输入数据特征向量。下方是输入数据序列，解码器部分数据输入要预测部用 0 填充，防止提前关注到未来部分，即进行遮蔽（Mask attention）操作，具体实现表现为自适应注意力机制变为遮蔽的形式。红色虚线将整个模型分为编码器和解码器。左边编码器操作是计算多头自适应注意力，蓝色小梯形代表蒸馏操作，目的是降低计算量。数据经过编码器后经过全连接特征图处理，输入到右边解码器。右边解码器将左边编码器运算结果和解码器输入的经过遮蔽处理的数据进行运算。而后，通过一个全连接层，得到绿色格子部分的预测输出。

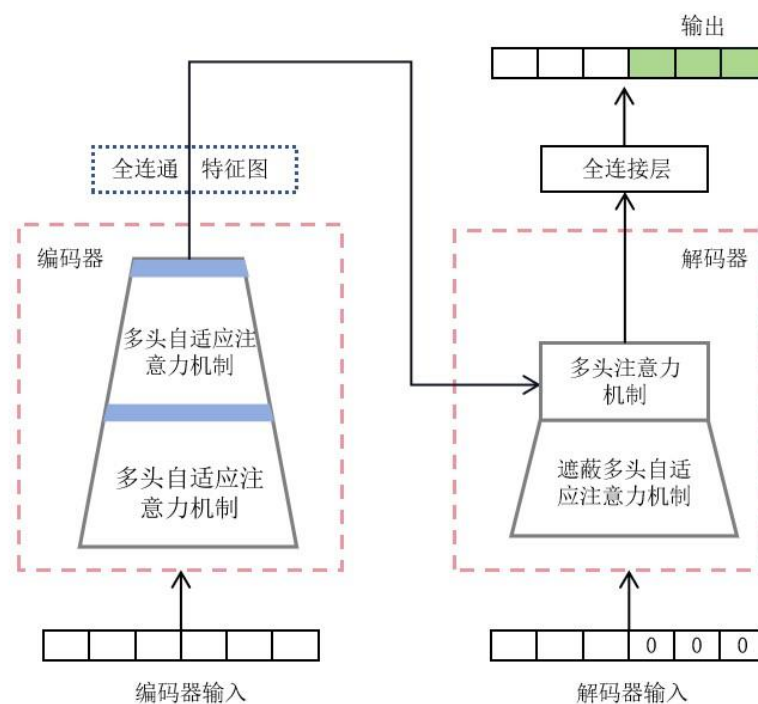


图 4-1 自适应注意力机制模型示意图

由上一章节研究可知，模型有效与否的关键在于注意力机制的设置。针对复杂的数据，不当选择方式可能造成信息丢失，而降采样率的固定一方面可能损失重要信息，另一方面又可能包含冗余特征^[42]。因此，本章提出一种采用自适应动态选取注意力矩阵的方法，在降低计算负担的同时减少信息损失，从而更好实现预测目标。其中，自适应注意力具体操作首先使用 KL 散度来度量Q和K矩阵之间相似性，根据得分进行排序。而后自适应选择一个数据子集来减少网络规模。最后，生成式解码器在推理阶段只需经过一个全连接层，避免了误差在推理过程中累积。

4.2.2 模型设计

图 4-1 所示的自适应注意力机制模型由左边编码器和右边解码器构成。其中，编码器由多个编码层构成，编码是为提取出长序列输入数据的鲁棒性远程依赖。

编码器部分的核心是计算自适应注意力。首先数据经过初始位置编码和时间编码，同时进行归一化处理，经过编码器运算得到的输出作为解码器部分注意力Q矩阵的输入。解码器由多个解码层组成，解码器数据输入采取“已知+未知”的方式，用一部分已知时间序列构成输入部分，要预测的时间序列数据用“0”填充覆盖，防止当前位置关注未来信息，避免自回归。最后通过全连接卷积实现数据逆归一化处理，得到预测结果。下面是模型各个部分具体细节介绍：

（1）编码器部分：模型的左侧编码器由多个注意力层组成，其中注意力层的数量选择为 6、4 和 2。每个注意力层包含两个子层：一是多头自适应注意力机制，另一个则是全连接前馈网络。

（2）解码器部分：与编码器类似，在每个子层使用残差连接，后进行归一化处理。输入采用编码器的输出和经遮掩处理的数据序列，防止当前位置关注后续位置。

（3）输入数据位置编码：输入数据的嵌入包含三个独立部分：一个标量投影、局部时间戳（位置）和全局时间戳嵌入（分钟、小时、日、周等）。模型不使用递归，为确保能够充分利用序列中的顺序信息，在编码器和解码器的堆栈中对输入数据序列进行嵌入时，我们引入了关于序列相对位置和绝对位置的信息，即“位置编码”。具体操作如公式（4-2），（4-3）所示。

采用一维卷积滤波器 $\text{kernel width} = 3$, $\text{stride} = 1$ 将标量上下文 f_t^t 进行投影，用一个输入矢量表示，如公式（4-1）所示，其中 $\alpha_i \in \{1, 2, 3, \dots, L_f\}$, α 是平衡标量投影与局部和全局嵌入之间大小的因素。序列输入被归一化，将其设为

1, 全局时间戳是一个可学习嵌入向量, 如公式 (4-4) 所示, 嵌入示意图如图 4-2 所示。

$$SE_{(L_f \times (t-1) + i)} \quad (4-1)$$

$$PE_{(pos, 2i+1)} = \sin(pos/10000^{2i/d_{model}}) \quad (4-2)$$

$$PE_{(pos, 2i+1)} = \sin(pos/10000^{2i/d_{model}}) \quad (4-3)$$

$$f_{feed}^t[i] = \alpha u_i^t + PE_{(L_f \times (t-1) + i)} + \sum_p [SE_{(L_f \times (t-1) + i)}]_p \quad (4-4)$$

(4) 自适应注意力机制: 最初的自注意力机制中, 注意力功能由查询矩阵Q、键矩阵K和值矩阵V组成。通过计算查询与相应键的兼容性函数^[41], 能确定每个值的权重, 然后将这些值加权求和得到输出。自适应注意力模型的注意力关注实际是“点积对关注”, 输入由维度为查询向量q, 键向量k, 以及值向量v组成。计算所有键值的查询点积, 将每个结果除以 $\sqrt{d_k}$, 应用softmax函数获得值权重, Zhou H 等人提出自注意力机制为公式 (4-5) ^[42]。

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4-5)$$

其中, $Q \in \mathbb{R}^{L_Q \times d}$, $K \in \mathbb{R}^{L_K \times d}$, $V \in \mathbb{R}^{L_V \times d}$, L示时间序列数据的长度。

与 Vaswani 和 Zhou H 等人^[41, 42]不同, 本章所提 AAM 模型将网络每个阶段输入数据特征向量的数量变为动态自适应的, 在不丢弃重要信息同时不浪费计算资源来处理冗余信息。在自适应注意力模块中, 根据 KL 散度量q和k相似性算出得分, 主要遵循公式 (4-6)。如图 4-3 所示, 多个向量q构成权重矩阵 W_q , 多个向量k构成 W_k 权重矩阵。计算 $q^1 \cdot k^1$, $q^2 \cdot k^2$, ... $q^n \cdot k^n$ 而后将得分进行排序, 根据排序后分数自动挑选一个数据子集, 具体选取流程如算法 1 所示。

$$M(q^1, K) = \ln \sum_{j=1}^{L_K} e^{\frac{q_i \cdot K_j^T}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_K} e^{\frac{q_i \cdot K_j^T}{\sqrt{d}}} \quad (4-6)$$

其中, 第i个q对所有关注定义为概率 $p(k_j|q_i)$ 是注意力概率分布。 $\frac{1}{L_K} p(k_j|q_i)$ 是均匀分布, L_K 是查询向量q长度。式子中第一项是 q_i 在矩阵K中的运算, 第二项是它们算数平均值。

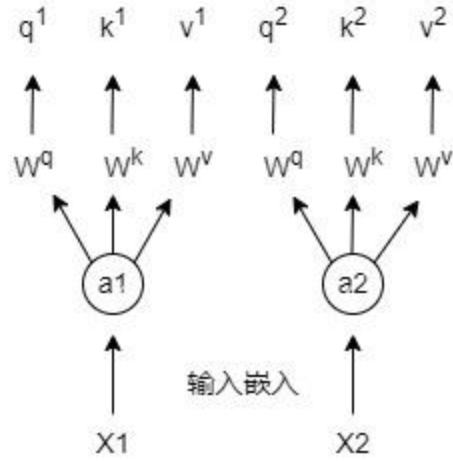


图 4-2 输入数据嵌入示意图

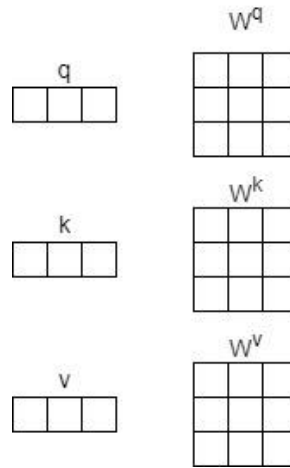


图 4-3 权重矩阵图

4.2.3 自适应注意力选取算法

自适应注意力机制选取流程如表 4-1 所示，可表示为式 (4-7)。

$$Attention(Q, K, V) = softmax(\frac{Q^v K^T}{\sqrt{d_k}})V \quad (4-7)$$

其中 Q^v 是经过自适应采样的和 q 相同大小的矩阵。

表 4-1 自适应注意力选取流程

算法 1 自适应注意力机制选取流程

前提条件和输入: $Q \in R^{m \times d}$, $K \in R^{n \times d}$, $V \in R^{n \times d}$, 流量数据

- 1 设置实验超参数, 同时设置恒定采样步长 $length$
- 2 从 K 中随机采样 m 个作为 K^y
- 3 设置标准分 $S^y = Q(K^y)^T$
- 4 计算 q 的测量值, 将测量值 $M = \max(S^y) - \text{mean}(S^y)$ 由小到大进行排序
- 5 按照步长 $length$ 从中选取 Q 个矩阵作为 Q^y
- 6 得到自适应机制注意力: $Attention(Q, K, V) = \text{softmax}(\frac{Q^y K^T}{\sqrt{d_k}})V$

输出: 自适应注意力机制

表 4-1 介绍自适应注意力机制选取流程, 同时对于自适应注意力机制模块参数, 设置前馈网络层(内层尺寸为 2048)和dropout层($p = 0.2$)。保留数 10% 的数据为验证集。因此, 所有实验是在 8 次时间顺序随机训练和验证下进行, 结果是 8 次运行均值。

4.3 性能评估

4.3.1 数据集

数据集一从 WIDE 项目获取^[64], 时间从 2020/8/1 0:00 开始至 2022/8/31 23:50 结束。每条数据以 10 分钟时长为时间单位, 共计 109585 条数据记录。数据格式是包括 6 个 IP 端口的数据量以及一个总流量数据, 同时每个记录包括日期、开始时间和结束时间、ID 号等标签。数据集二来自 ISP 互联网数据, 是英国学术网(https://github.com/rankinjl/internet-traffic_stats_project)主干网流量, 时间从 2005/6/7 07:00 开始至 2005/7/28 13:55 结束, 每隔五分钟收集一次数据, 同时数据经过处理统一保留两位小数。

在固定时间间隔内通过网络中一个观察点的一组 IP 数据包, 共享一组公共属性, 如源和目标 IP 地址或 IP 前缀。每个 IP 端口的流量可被视为一个网络切片中的原始流量数据。数据集追踪了一个月内每个 IP 端口以及其总的流量消耗。其中数据集一部分流量数据格式如表 4-2 所示, 数据集二部分流量数据如表 4-3 所示。

表 4-2 数据集一

Data	IP1	IP2	IP3	IP4	IP5	IP6	Total(GB)
2020/8/1 0:00	6.12	6.39	2.31	3.87	1.39	2.76	22.85
2020/8/1 0:10	4.61	1.91	2.92	3.14	1.74	2.57	16.90
2020/8/1 0:20	3.72	1.68	1.96	6.40	2.13	1.28	17.16
2020/8/1 0:30	5.89	1.49	5.51	1.52	1.87	3.64	19.92
2020/8/1 0:40	4.50	2.28	4.90	3.56	1.95	3.81	21.01
2020/8/1 0:50	4.64	4.21	1.93	13.30	2.31	2.76	29.13

表 4-3 数据集二

Data	Total (100MB)
2005/7/6 07:00	4.25
2005/7/6 07:05	4.42
2005/7/6 07:10	4.62
2005/7/6 07:15	4.62
2005/7/6 07:20	5.46
2005/7/6 07:25	5.98

4.3.2 仿真环境与评价指标

(1) 评价指标

为检测模型预测效果，选择评价指标：**MAE** 和 **RMSE**（第三章所介绍评价指标）用来衡量预测的误差，两者数字越小表示预测效果越好。

(2) 对比模型

对比模型有以下两种：自回归积分滑动平均模型和长短时记忆网络。**ARIMA** 和 **LSTM** 都采用单步预测第二种策略实现多个目标值的预测。

(3) 参数设置

实验所需环境如表 4-4 所示，实验参数影响模型的学习效果。经过反复测试，本章选择使用 **Adam** 优化器，激活函数选 **Relu**，最小批处理为 32，学习率为 0.0001，**dropout** 设为 0.05，迭代 **epoch** 数为 12。具体实验中设置如表 4-5 所示：

表 4-4 实验环境

名称	型号
操作系统	Windows10 64bit
CPU	i7 10700k
GPU	NVIDIA GeForce RTX 1080TI
内存	1020G
编程环境	Python3.7

表 4-5 参数设置

参数名称	值
学习率	0.0001
批处理大小	32
dropout	0.05
优化器	Adam
损失函数	MAE, RMSE
epoch	12

4.3.3 结果与分析

AAM 方法与传统时间序列预测方法 ARIMA 和深度学习 LSTM 算法进行比较。数据集一上使用长度从 2020 年 8 月 1 日 0:00 开始至 2020 年 10 月 10 日 0:00 结束,用 70%做训练数据,10%做验证数据,剩下的作为测试集。目的是预测 6, 36, 72, 144 个时间点的结果,选取测试时间点从 2020 年 9 月 26 日 02:30 开始。数据集二所使用长度从 2005 年 6 月 7 日 7:00 开始至 2005 年 7 月 28 日 13:45 结束,选取测试时间点从 2005 年 7 月 20 日 14:20 开始。

表 4-6 数据集一实验结果

预测点数 \ 方法	ARIMA		LSTM		AAM	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
6 个	3.45	4.12	3.08	3.84	2.90	3.44
36 个	3.88	4.65	2.92	3.14	3.42	4.04
72 个	5.93	7.81	5.89	7.19	4.87	6.10
144 个	8.27	10.13	5.92	7.57	3.78	4.89

表 4-7 数据集二实验结果

预测点数 \ 方法	ARIMA		LSTM		AAM	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
6 个	0.54	0.57	0.37	0.39	0.20	0.22
36 个	0.76	0.78	0.38	0.41	0.20	0.23
72 个	0.78	0.81	0.39	0.42	0.18	0.21
144 个	0.82	0.94	0.42	0.48	0.20	0.25

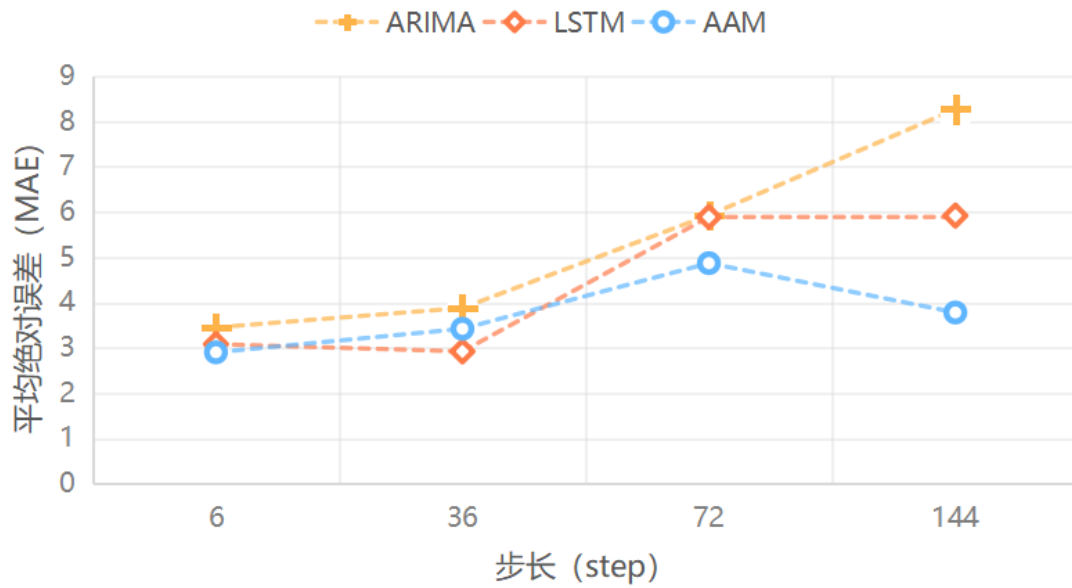


图 4-4 数据集一预测 MAE 结果图

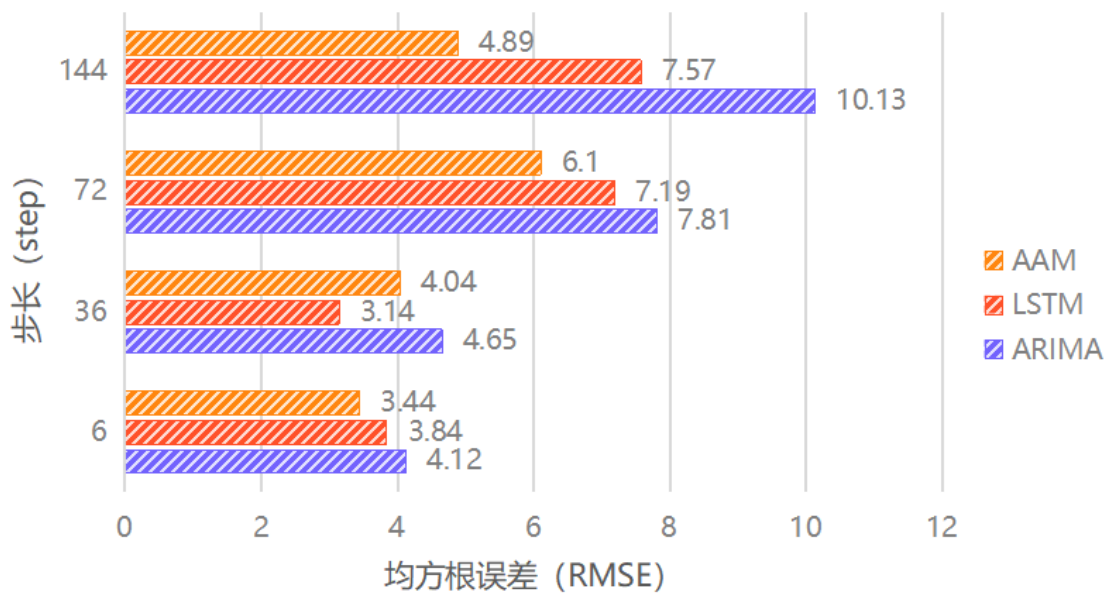


图 4-5 数据集一预测 RMSE 结果图

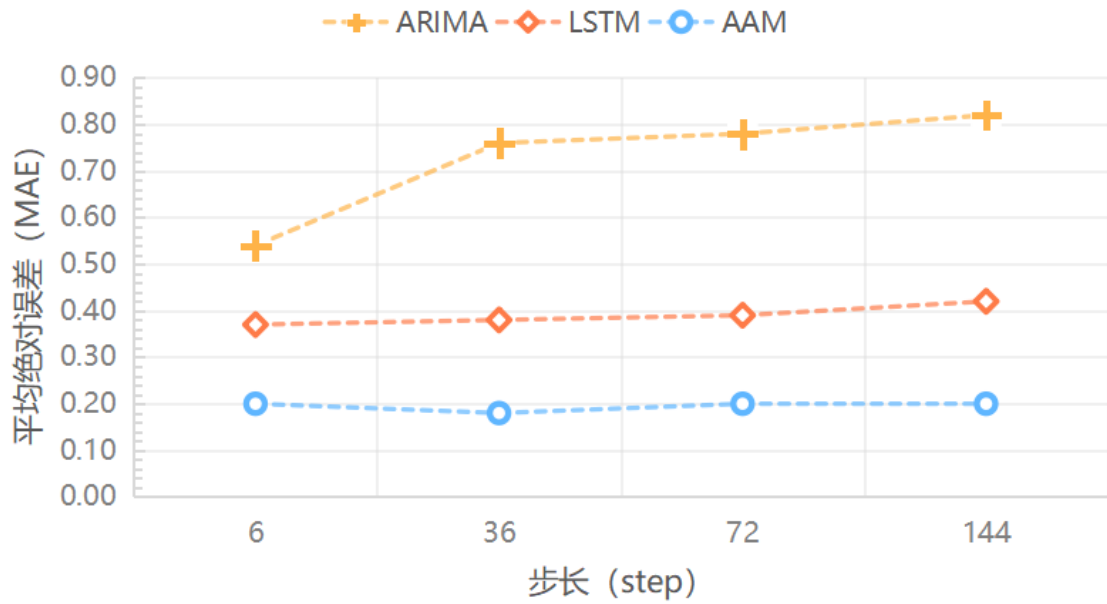


图 4-6 数据集二预测 MAE 结果图

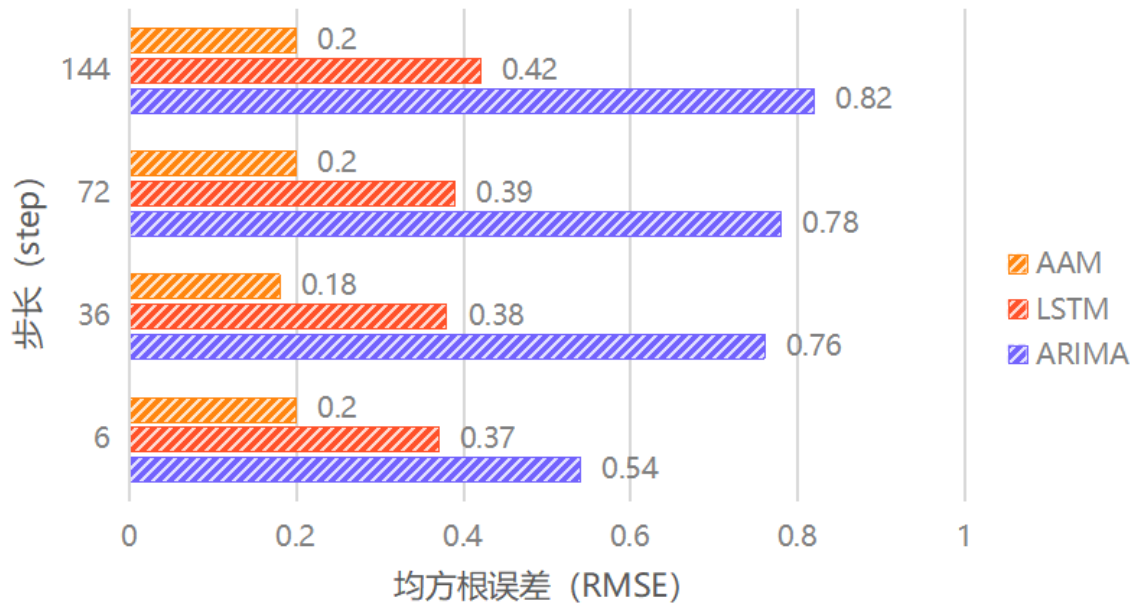


图 4-7 数据集二预测 RMSE 结果图

在本章所做实验中, ARIMA 模型利用 CPU 进行训练, LSTM, AAM 使用 GPU 进行训练。另外, ARIMA 和 LSTM 采用单步预测第二种策略以提高预测精度。AAM 采用一次性生成解码器生成预测数据。如表 4-6 和表 4-7 列出了 6, 36, 72, 144 个点的 ARIMA, LSTM 和 AAM 模型分别在两个数据集上 MAE 和 RMSE, 图 4-4 和图 4-5 是其可视化表达。由此可知, 短时间网络流量实现多步长精准预测比较困难。相较 ARIMA 模型和 LSTM 模型, 在 10 min 和 5min 时间尺度下, 本章所提自适应

注意力机制 AAM 模型的 MAPE 和 RMSE 更低。同时，由图中数据可知，数据集一上预测 144 个点时，MAE 和 RMSE 反而比预测 72 个时间点更低，准确度更高。数据集二上预测 72 个点时，MAE 和 RMSE 反而比预测 36 个时间点更低，准确度更高。这说明所提 AAM 模型可以兼顾更久远的数据信息。随着预测步长的增加，所提出的 AAM 模型和基准模型的预测误差也会随之增加。

通过对比发现，本章提出的模型始终表现优于 ARIMA 和 LSTM 模型，并且随着步长增加，预测优势会变得越来越明显，AAM 能顾及网络流量中“历史久远”的流量特征，自适应注意力机制提取数据特征时解决了注意力在计算时数据中一些重要信息丢失的问题，使其在长序列流量预测问题中能达到更好效果。

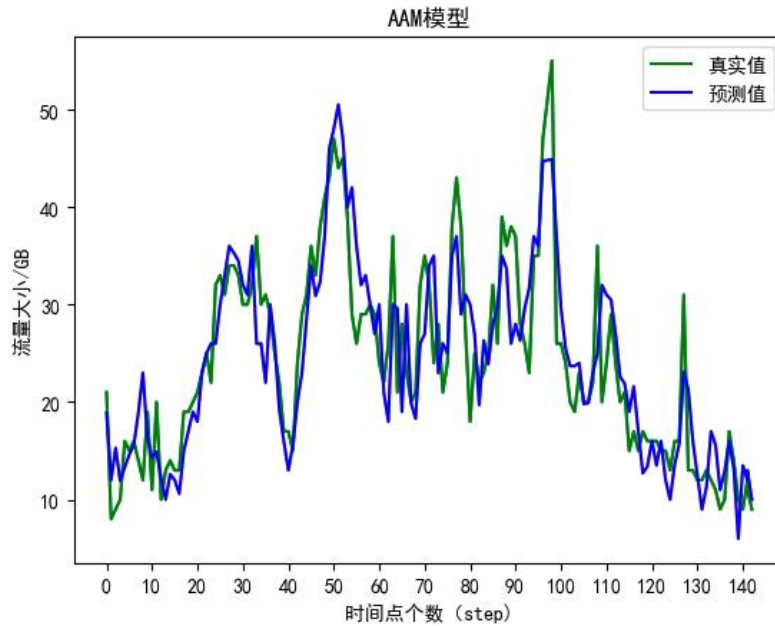


图 4-8 AAM 在数据集一的结果图

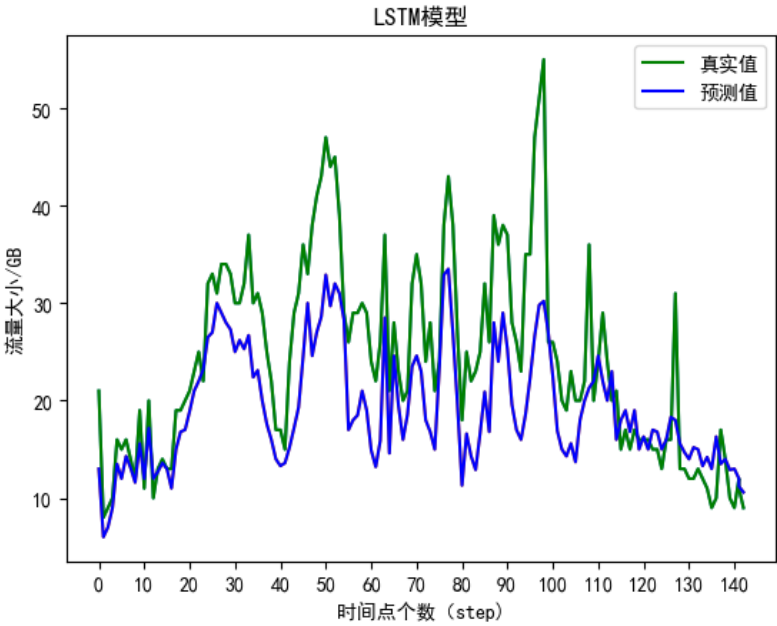


图 4-9 LSTM 在数据集一的结果图

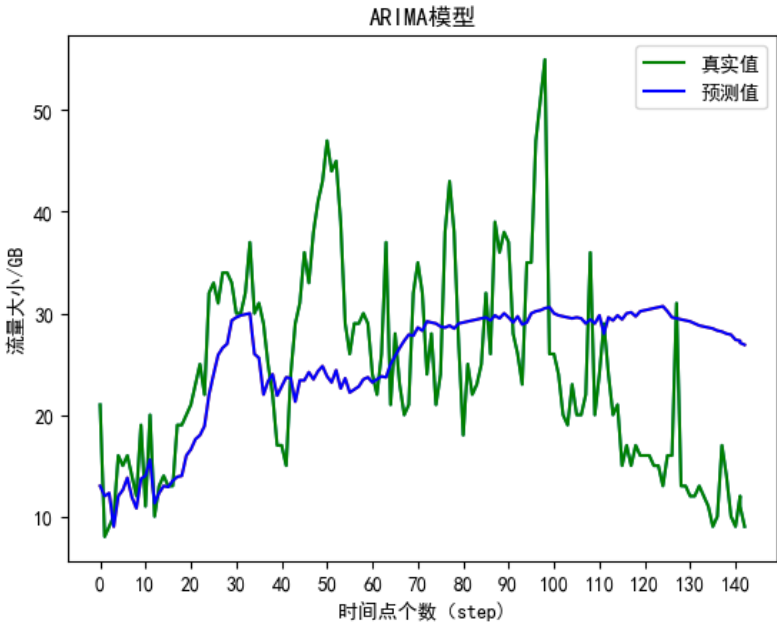


图 4-10 ARIMA 在数据集一的结果图

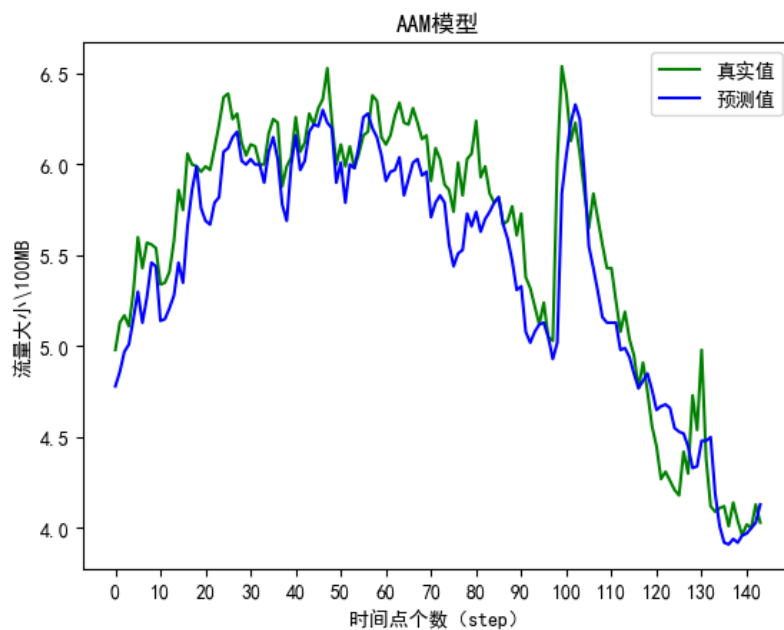


图 4-11 AAM 在数据集二的结果图

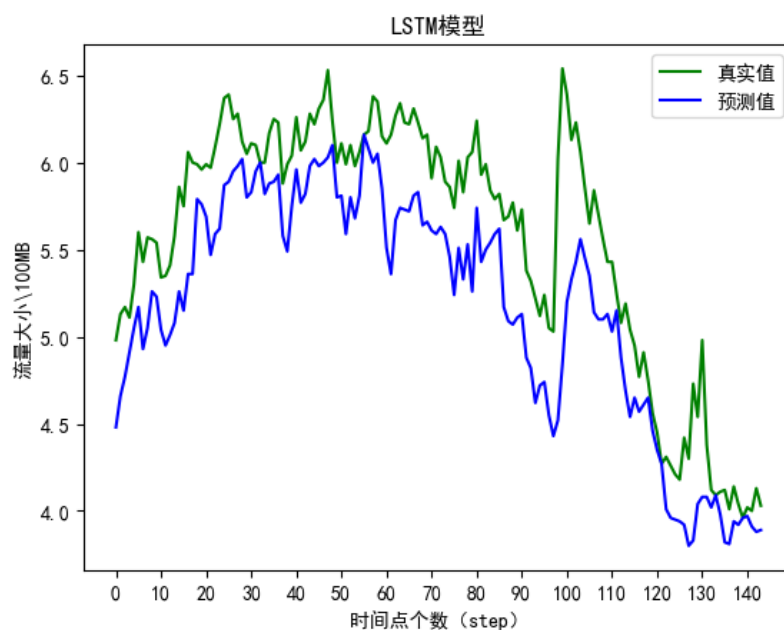


图 4-12 LSTM 在数据集二的结果图

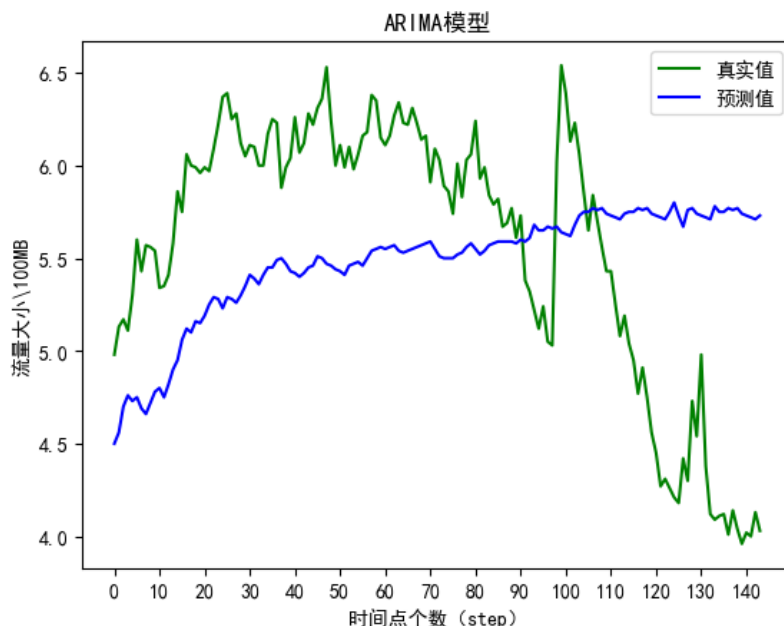


图 4-13 ARIMA 在数据集二的结果图

如图 4-8 所示，是 AAM 在数据集一上预测 144 个时间点数结果图，图 4-11 是在数据集二上预测结果。如图 4-9 所示，是 LSTM 方法在数据集一上预测 144 个时间点的结果，图 4-12 是 LSTM 方法在数据集二上预测的结果。如图 4-10 所示，是 ARIMA 方法在数据集一上预测 144 个时间点结果，图 4-13 是在数据集二上预测结果。通过两个模型在两个网络流量数据集上预测结果对比可知，AAM 模型预测效果较好，LSTM 模型由于误差反向传播而造成了性能下降，ARIMA 模型效果最差，它不适合长序列预测任务。

4.4 本章小结

针对长序列时间预测问题，本章提出一种自适应注意力机制网络流量预测算法。该算法采用动态自适应的方式处理输入数据，缩减计算量的同时又减少数据的信息损失。经实验表明，AAM 能较好地提取数据之间的特性，自适应注意力机制模型在长序列网络流量预测问题中具有较好性能。

(1) 基于自适应注意力机制的网络流量预测模型采用自适应注意力算法，根据 KL 散度量 q 和 k 相似性，计算 $q^1 \cdot k^1$, $q^2 \cdot k^2$, ... $q^n \cdot k^n$ 而后将得分进行排序，根据排序后分数自动挑选一个数据子集，这样处理使得 AAM 能顾及网络流量中“历史久远”的流量特征，自适应注意力的“平均”选取解决了注意力在计算时数据里一些重要信息丢失的问题，使其在长序列流量预测问题中能达到更好效果。

(2) AAM 解码器的输入采用编码器的输出和经掩蔽处理的数据序列,防止当前位置关注后续位置。另外,只需要通过一个生成式解码操作就可以获得长序列输出,替换了逐步生成式解码,避免了推理阶段的累积误差传播,提升了长序列预测的速度。

(3) AAM 在两个公开的数据集进行了大量训练,发现其预测误差、预测精度优于经典的 ARIMA 和 LSTM 模型,在波动明显的数据集上预测性能明显,在预测时间点数量最多为 144 个时均方根误差 (RMSE) 分别降低了约 51.77%和 73.40%,体现了模型在长序列预测问题中的有效性。

5 结论和展望

5.1 主要工作

整体而言,本论文的研究主要集中在基于注意力机制的网络流量预测模型上。为基于自注意力机制的网络流量预测研究提供了一定的参考价值。

(1) 提出一种空洞卷积和相对位置表征的注意力机制的改进 Transformer 模型-TDCRP,提高了模型在网络流量预测问题中的有效性。首先,在 Transformer 的基础上通过添加空洞卷积层结构来捕捉流量数据中的上下文数据信息中的依赖关系。其次,注意力的相对位置表征增强了 TDCRP 模型对序列中不同位置的建模能力。在 WIDE 项目里的数据集上测试表明,与 CNN 和原 Transformer 模型相比,所提改进模型在预测时具有更好性能。

(2) 为预测更长时间序列理论出发,提出一个自适应注意力机制网络流量预测模型(AAM),模型包括编码器和解码器。首先,计算注意力时根据 KL 散度衡量 q 和 k 相似性,而后将得分进行排序,根据排序后分数自动挑选一个数据“平衡注意力”子集,这样处理使得能顾及网络流量中“历史久远”的流量特征,避免了数据里一些重要信息的可能丢失问题,使其在长序列流量预测问题中能达到更好效果。解码器的输入采用编码器的输出和经掩蔽处理的数据序列,防止当前位置关注后续位置,只需要一个前向步骤就可以获得长序列输出,避免了推理阶段的累积误差传播,提升了长序列预测的速度。与 ARIMA 和 LSTM 模型相比,在预测较长时间网络流量序列时效果更好。

5.2 未来展望

本文以注意力机制作为研究出发点,致力于解决网络流量预测问题,并通过实验评估证明了其在该领域的价值。然而,通过研究发现,注意力机制在网络流量预测问题中尽管取得了一定的成效,但是仍然存在一些问题需要进一步研究和探索。

(1) 模型的可解释性是一个关键的研究方向。在当前与注意力机制相关的流量预测模型中,模型的可解释性不足。需要关注模型预测结果的解释方式,模型如何利用输入特征以及模型的决策过程,以提升对模型行为的理解和信任。通过研究诸如特征重要性分析、参数构建和超参数选择基础等可解释技术,可以提高模型的透明度和模型可解释性,以更好地指导实际应用。

(2) 可以对不同应用场景进行更深入的研究和模型探索并对模型在不同时间尺度上的预测性能进行研究,比较模型在不同数据背景下的普适性对于不同数据集是未来研究的可能方向之一。此外,随着网络技术的发展,针对网络切片背景下的预测模型,需要研究每个网络切片与整体流量预测之间的关系,以及设计实现多目标同时预测的模型。

(3) 流量预测为网络管理提供了新的思路,但在实际应用中仍存在许多挑战需要解决。例如,实现对网络流量的更准确预测。另外,实现网络流量长期准确预测后,需要进一步研究和调查以优化 SDN 路由、分配网络切片资源以及实施流量异常检测和过载警告等,这些问题还需要进一步的探索研究。

参考文献

- [1] 中国互联网络信息中心发布第 52 次《中国互联网络发展状况统计报告》[J].国家图书馆学报, 2023, 32(05):13.
- [2] McKeown N, Anderson T, Balakrishnan H, et al. OpenFlow: enabling innovation in campus networks [J].ACM SIGCOMM computer communication review, 2008,38(2):69-74.
- [3] Han B, Gopalakrishnan V, Ji L, et al. Network function virtualization: Challenges and opportunities for innovations [J].IEEE communications magazine, 2015,53(2): 90-97.
- [4] SUN J, ZHANG Y, LIU F, et al. A survey on the placement of virtual network functions [J]. Journal of Network and Computer Applications, 2022: 103361.
- [5] ALLIANCE N. Description of network slicing concept [J]. NGMN 5G P, 2016, 1 (1).
- [6] KAZMI S M A, KHAN L U, TRAN N H, et al. Network slicing for 5G and beyond networks [M].Berlin: Springer, 2019.
- [7] Augustin Soule, Kav'e Salamatian, and Nina Taft. 2005. "Combining filtering and statistical methods for anomaly detection". In Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement (IMC'05). USENIX Association, Berkeley, CA, USA, 31-31.
- [8] Matthew Roughan, Mikkel Thorup, and Yin Zhang. 2003. "Traffic engineering with estimated traffic matrices". In Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement (IMC '03). ACM, New York, NY, USA, 248-258.
- [9] Theophilus Benson, Ashok Anand, Aditya Akella, et al. 2011."MicroTE: fine grained traffic engineering for data centers". In Proceedings of the 2011 Conference on Emerging Networking Experiments and Technologies (Co-NEXT '11). ACM, Tokyo Japan.
- [10] Andrew R. Curtis, Jeffrey C. Mogul, Jean Tourrilhes, et al. 2011. "Devo Flow: scaling flow management for high-performance networks". SIGCOMM Comput. Commun. Rev. 41, 4 (August 2011), 254-265.
- [11] Cristian Estan and George Varghese. 2002. "New directions in traffic measurement and accounting". SIGCOMM Comput. Commun. Rev. 32,4 (August 2002), 323-336.
- [12] A. Yassine, H. Rahimi and S. Shirmohammadi. 2015. "Software defined network traffic measurement: Current trends and challenges," in IEEE Instrumentation & Measurement Magazine, vol. 18, no. 2 (April 2015),42-50.
- [13] Pattern Recognition. Cham: Springer International Publishing, 2020: 640-652. Tune, Paul, and Matthew Roughan. "Internet traffic matrices: A primer."Recent Advances in Networking 1 (2013).
- [14] 康梦轩, 宋俊平, 范鹏飞, 等.基于深度学习的网络流量预测研究综述[J].计算机工程与

- 应用, 2021, 第 57 卷(10): 1-9
- [15] CELENK M, CONLEY T, GRAHAM J, et al. Anomaly prediction in network traffic using adaptive Wiener filtering and ARMA modeling [C]. 2008 IEEE International Conference on Systems, Man and Cybernetics. Piscataway: IEEE, 2008.
 - [16] XU S, ZENG B. Network traffic prediction model based on auto-regressive moving average [J]. Journal of Networks, 2014, 9(3): 653
 - [17] GAN M, PENG H. Stability analysis of RBF network-based state-dependent autoregressive model for nonlinear time series[J] Applied Soft Computing, 2012, 12(1): 174-181.
 - [18] 陈皓. 基于小波神经网络的移动网络流量预测研究[J]. 电子世界, 2020.
 - [19] LAN X. Analysis and research of several network traffic prediction models[C]. 2013 Chinese Automation Congress. Piscataway: IEEE, 2013.
 - [20] Park J, Samarakoon S, Bennis M, et al. Wireless network intelligence at the edge. Proc IEEE. 2019;107(11):2204-2239. doi:10.1109/JPROC.2019.2941458.
 - [21] Jaffry S. Cellular traffic prediction with recurrent neural network; 2020. arXiv preprint arXiv:2003.02807.
 - [22] Chen M, Challita U, Saad W, et al. Artificial neural networks-based machine learning for wireless networks: a tutorial. IEEE Commun Surv Tutor. 2019;21(4):3039-3071. doi:10.1109/COMST.2019.2926625.
 - [23] Zhang X, You J. A gated dilated causal convolution based encoder-decoder for network traffic forecasting. IEEE Access. 2020;8:6087-6097. doi:10.1109/ACCESS.2019.2963449.
 - [24] Trinh HD, Giupponi L, Dini P. Mobile traffic prediction from raw data using LSTM networks. Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC; 2018:1827-1832.
 - [25] Hua Y, Zhao Z, Li R, et al. Deep learning with long short-term memory-for-time-series-prediction. IEEE-Commun-Mag. 2019;57(6):114-119. doi:10.1109/MCOM.2019.1800155.
 - [26] Mohammadi M, Al-Fuqaha A, Sorour S, et al. Deep learning for IoT big data and streaming analytics: a survey. IEEE Commun Surv Tutor. 2018;20(4):2923-2960.
 - [27] He Q, Moayyedi A, Dan G, et al. A meta-learning scheme for adaptive short-term network traffic prediction. IEEE J Select Areas Commun. 2020;38(10):2271-2283. doi:10.1109/JSAC.2020.3000408.
 - [28] 任玮. 基于深度信念网络的网络流量预测模型[J]. 山西电子技术, 2016, 2 (1): 62-64.
 - [29] 张传亭. 基于深度学习的城市尺度无线流量预测[D]. 济南: 山东大学, 2019.
 - [30] LAZARIS A, PRASANNA V K. An LSTM framework for modeling network traffic[C]//2019 IFIP/IEEE Symposium on Integrated Network and Service Management

- (IM). IEEE, 2019:19-24.
- [31] LAZARIS A, PRASANNA V K. Deep Flow: A deep learning framework for software-defined measurement [C]//Proceedings of the 2nd Workshop on Cloud-Assisted Networking.2017:43-48.
- [32] XU F, LIN Y, HUANG J, et al. Big data driven mobile traffic understanding and forecasting: A time series approach [J]. IEEE transactions on services computing, 2016, 9(5): 796-805.
- [33] LAN X. Analysis and research of several network traffic prediction models[C]/2013 Chinese Automation Congress, Changsha, 2013.
- [34] 黎嘉诚.基于小波神经网络的校园网络流量预测研究[D].南昌:南昌大学.2019.
- [35] TANG F, MAO B, FADLULLAH Z M, et al. ST- DeLTA: an novel spatial-temporal value network aided deep learning based intelligent network traffic control system [J].IEEE Transactions on Sustainable Computing, 2020 ,5 (4):568-580.
- [36] VINCHOFFV C, CHUNG N, Gordon T, et al. Traffic prediction in optical networks using graph convolutional generative adversarial networks[C]//2020 22nd International Conference on Transparent Optical Networks (ICTON). IEEE, 2020: 1-4.
- [37] LOHRASBINASAB I, SHAHRAKI A, TAHERKORDI A, et al. From statistical - to machine learning - based network traffic prediction[J]. Transactions on Emerging Telecommunications Technologies, 2022, 33(4): e4394.
- [38] ZHANG L, ZHANGH, TANG Q, et al. LNTP: An end-to-end online prediction model for network traffic [J]. IEEE Network, 2020, 35(1): 226-233.
- [39] 韦烜, 刘志华, 李青等.大型 IP 网络流量矩阵分析预测的探讨研究[J].系统工程与电子技术, 1-13[2024-03-08].
- [40] 高志宇, 王天荆, 汪悦等.基于生成对抗网络的 5G 网络流量预测方法[J].计算机科学, 2022, 49(04):321-328.
- [41] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
- [42] ZHOU H, ZHANG S, PENG J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 11106-11115.
- [43] LI M, WANG Y, WANG Z, et al. A deep learning method based on an attention mechanism for wireless network traffic prediction [J]. Ad Hoc Networks, 2020, 107: 102258.
- [44] ZENG A, CHEN M, ZHANG L, et al. Are transformers effective for time series forecasting? [J]. arXiv preprint arXiv: 2205.13504, 2022.
- [45] Xu F L, Lin Y Y, Huang J X, et al. Big data driven mobile traffic understanding and forecasting: a time series approach. IEEE Trans Serv Comput, 2016, 9: 796–805

- [46] Li, Liangzhi, Kaoru Ota, et al. (2017, June). Everything is image: CNN-based short-term electrical load forecasting for smart grid. 2017 third international symposium of creative computing (ISPAN-FCST-ISCC) (pp. 344-351).
- [47] 王红, 史金钊, 张志伟. 基于注意力机制的 LSTM 的语义关系抽取[J]. 计算机应用研究, 2018, 35(5): 1417-1420, 1440.
- [48] 申翔翔, 侯新文, 尹传环. 深度强化学习中状态注意力机制的研究[J]. 智能系统学报, 2020, 15(2): 317-322.
- [49] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational visual media, 2022, 8(3): 331-368.
- [50] 王明申, 牛斌, 马利. 一种基于词级权重的 Transformer 模型改进方法 [J] . 小型微型计算机系统, 2019, 40 (4): 744-748.
- [51] Zhao Q, Cai X, Chen C, et al. Commented content classification with deep neural network based on attention mechanism[C]//2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017: 2016-2019.
- [52] HAO S, LEE D-H, ZHAO D. Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system[J]. Transportation Research Part C: Emerging Technologies, 2019, 107: 287-300.
- [53] 周才东, 曾碧卿, 王盛玉, 等. 结合注意力与卷积神经网络的中文摘要研究 [J] . 计算机工程与应用, 2019, 55 (8): 132-137.
- [54] TSOTSOS J K, CULHANE S M, WAI W Y K, et al. Modeling visual attention via selective tuning[J]. Artificial Intelligence, 1995, 78(1/2): 507-545
- [55] MNH V, HEES N, GRAVES A. Recurrent models of visual attention[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014, 2: 2204-2212.
- [56] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arxiv preprint arxiv:1406.1078, 2014.
- [57] 贺浩, 王仕成, 杨东方, 等. 基于 Encoder-Decoder 网络的遥感影像道路提取方法 [J] . 测绘学报, 2019, 48 (3): 330-338.
- [58] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arxiv preprint arxiv:1409.0473, 2014.
- [59] 王鑫, 吴际, 刘超, 等. 基于 LSTM 循环神经网络的故障时间序列预测 [J] . 北京航空航天大学学报, 2018, 44 (4): 772-784.
- [60] HOCHREITER S. The vanishing gradient problem during learning recurrent neural nets and problem solutions[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6(2): 107-116.

- [61] 何迎利, 胡光宇, 张浩, 等. 基于局部信息增强注意力机制的网络流量预测[J]. 科学技术与工程, 2023, 23 (30): 13014-13022.
- [62] 李岚欣. 面向自然语言处理的注意力机制研究[D]. 北京: 北京邮电大学, 2019: 10-14.
- [63] Zeng A, Chen M, Zhang L, et al. Are transformers effective for time series forecasting?[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(9): 11121-11128.
- [64] Li S, Xiao Y J, Xuan Y, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting[J]. Advances in neural information processing systems, 2019, 32.
- [65] Sony C S L, Cho K. Traffic data repository at the WIDE project[C]V/Proceedings of USENIX 2000 Annual Technical Conference: FREENIX Track. 2000: 263-270.
- [66] Lim, Bryan, et al. "Temporal fusion transformers for interpretable multi-horizon time series forecasting." International Journal of Forecasting 37.4 (2021): 1748-1764.
- [67] Lin, Yang, Irena Koprinska, and Mashud Rana. "SSDNet: State space decomposition neural network for time series forecasting." In 2021 IEEE International Conference on Data Mining (ICDM), pp. 370-378. IEEE, 2021.
- [68] LI Y, LU X, XIONG H, et al. Towards long-term time-series forecasting: feature, pattern, and distribution[J]. arXiv:2301.02068,2023.
- [69] QI X, HOU K, LIU T, et al. From known to unknown: knowledge- guided transformer for time- series sales forecasting in Alibaba[J]. arXiv:2109.08381, 2021.
- [70] ZHOU T, MA Z, WEN Q, et al. FEDformer: frequency enhanced decomposed transformer for long- term series forecasting[J]. arXiv:2201.12740, 2022.
- [71] WU H, XU J, WANG J, et al. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting[C]//Advances in Neural Information Processing Systems 34, Dec 6-14, 2021: 22419-22430.
- [72] Schwöbel P, Jørgensen M, Ober S W, et al. Last layer marginal likelihood for invariance Learning[C].International Conference on Artificial Intelligence and Statistics. PMLR, 2022: 3542-3555. LIU S, YU H, LIAO C, et al. Pyraformer: low- complexity pyramidal attention for long- range time series modeling and forecasting[C]//Proceedings of the 2022 International Conference on Learning Representations, Apr 25-29, 2022: 1-20.

攻读学位期间的学术论文与研究成果

- [1] 第一作者. Zhifang Zhang, Zhaogang Shu, Sheng Yang, Shuwu Chen, Tingjiang Guo. Network traffic prediction study based on the adaptive attention mechanism. [C] International Conference on Electronic Information Engineering and Computer Communication (EIECC 2023) (EI 见刊)
- [2] 第三作者. Haoxian Feng, Zhaogang Shu*, Zhifang Zhang, Yuantao Wang, Zhiwei Liu. Bandwidth Resources Prediction for Dynamic Networking Slicing Based on Time Series Analysis[C]. 2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM). IEEE, 2022: 1-5. (EI)
- [3] 第四作者. Zhaogang Shu, Haoxian Feng, Tarik Taleb and Zhifang Zhang. A Novel Combinatorial Multi-armed Bandit Game to Identify Online the Changing Top-K Flows in Software-Defined Networks[J]. Computer Networks, 2023: 109783.(SCI, IF:5.493)

致谢

时光匆匆，岁月荏苒，我的学生时代也即将在这个炙热的夏天画上圆满的句号。不知不觉来到福建农林大学已有 3 年之久，在这 3 年时光里，无论是在学业知识上，还是校园生活中，我都收获颇丰。

首先，特别感谢我的恩师舒兆港老师。您学识渊博，治学严谨，感谢您在专业培养中对学生答疑解惑、悉心教导以及严格要求。您为人和蔼，亦师亦友，让远在他乡求学的学子倍感温暖与感动，感谢您在生活上对学生的细心关怀和排忧解难。此外，感谢杨昇老师对我学业的鼓舞与帮助。你们的支持令学生毕生难忘！

其次，感谢生我育我的父母。求学近二十载，是父母在背后一直默默无闻的支持。你们平凡的身躯在我心中无比伟岸，正因为有你们，我才能行万里路，读万卷书。愿我迅速成长，能够承担起家庭的责任，为你们分担生活的压力。

最后，感谢人生路上相遇的好友、同学、以及诸位同门。一路走到现在，得到了许多亲朋好友的支持、帮助和鼓励，无法一一提及，在这里仅致以最诚挚的谢意！