计算机研究与发展 DOI:10.7544/issn1000-1239.2020.20190866
Journal of Computer Research and Development 57(4): 671-687, 2020

基于机器学习的智能路由算法综述

刘辰屹 徐明伟 耿 男 张 翔
(清华大学计算机科学与技术系 北京 100084)
(liucheny19@mails.tsinghua.edu.cn)

A Survey on Machine Learning Based Routing Algorithms

Liu Chenyi, Xu Mingwei, Geng Nan, and Zhang Xiang
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract The rapid development of the Internet accesses many new applications including real time multi-media service, remote cloud service, etc. These applications require various types of service quality, which is a significant challenge towards current best effort routing algorithms. Since the recent huge success in applying machine learning in game, computer vision and natural language processing, many people tries to design "smart" routing algorithms based on machine learning methods. In contrary with traditional model-based, decentralized routing algorithms (e.g. OSPF), machine learning based routing algorithms are usually data-driven, which can adapt to dynamically changing network environments and accommodate different service quality requirements. Data-driven routing algorithms based on machine learning approach have shown great potential in becoming an important part of the next generation network. However, researches on artificial intelligent routing are still on a very beginning stage. In this paper we firstly introduce current researches on data-driven routing algorithms based on machine learning approach, showing the main ideas, application scenarios and pros and cons of these different works. Our analysis shows that current researches are mainly for the principle of machine learning based routing algorithms but still far from deployment in real scenarios. So we then analyze different training and deploying methods for machine learning based routing algorithms in real scenarios and propose two reasonable approaches to train and deploy such routing algorithms with low overhead and high reliability. Finally, we discuss the opportunities and challenges and show several potential research directions for machine learning based routing algorithms in the future.

Key words machine learning; data driven routing algorithm; deep learning; reinforcement learning; quality of service (QoS)

摘 要 互联网的飞速发展催生了很多新型网络应用,其中包括实时多媒体流服务、远程云服务等.现有尽力而为的路由转发算法难以满足这些应用所带来的多样化的网络服务质量需求.随着近些年将机器学习方法应用于游戏、计算机视觉、自然语言处理获得了巨大的成功,很多人尝试基于机器学习方法去设计智能路由算法.相比于传统数学模型驱动的分布式路由算法而言,基于机器学习的路由算法通常是

收稿日期:2019-12-17;**修回日期**:2020-02-23
基金项目:国家自然科学基金项目(61625203,61832013);国家重点研发计划项目(2017YFB0801701)
This work was supported by the National Natural Science Foundation of China (61625203, 61832013) and the National Key Research and Development Plan of China (2017YFB0801701).
通信作者:徐明伟(xmw@cernet.edu.cn)

数据驱动的,这使得其能够适应动态变化的网络环境以及多样的性能评价指标优化需求.基于机器学习的数据驱动智能路由算法目前已经展示出了巨大的潜力,未来很有希望成为下一代互联网的重要组成部分.然而现有对于智能路由的研究仍然处于初步阶段.首先介绍了现有数据驱动智能路由算法的相关研究,展现了这些方法的核心思想和应用场景并分析了这些工作的优势与不足.分析表明,现有基于机器学习的智能路由算法研究主要针对算法原理,这些路由算法距离真实环境下部署仍然很遥远.因此接下来分析了不同的真实场景智能路由算法训练和部署方案并提出了2种合理的训练部署框架以使得智能路由算法能够低成本、高可靠性地在真实场景被部署.最后分析了基于机器学习的智能路由算法未来发展中所面临的机遇与挑战并给出了未来的研究方向.

关键词 机器学习;数据驱动路由算法;深度学习;强化学习;服务质量

中图法分类号 TP393

近年来随着互联网的高速发展,包括工业互联网、4K+视频及全息通信、网络游戏、远程云服务在内的很多新兴应用大量涌现.这些新兴的网络应用带来了高度差异化的服务质量需求.然而以往单纯通过对设备提速扩容来提升网络服务质量的方式已经逐渐触及天花板,进一步提升性能需要很高的成本,与此同时,研究表明:现有网络仍然存在巨大的优化空间[1].因此,对现有网络资源进行更好地优化利用成为提升用户服务体验的重要途径.

在传统的计算机网络体系结构中,网络层通常采用尽力而为的数据包分组转发模式,路由算法所关注的重点是数据包的可达性、算法的性能和可扩展性.近几年随着计算机网络的飞速发展,网络规模变得越来越大的同时网络上层的应用服务类型数量也在飞速增长.日益增长的服务类型数量带来了多样化的服务性能优化目标,这些优化目标涉及时延、带宽、吞吐、丢包率和网络稳定性等.尽力而为的传统路由算法使得现有计算机网络体系结构对于这些性能评价指标进行优化时存在一定的局限性.图1给出了传统路由算法局限性的示例,在本示例中网络流负载需求500 Mbps的带宽,传统基于最短路径的路由算法将所有流量导入瓶颈链路中,所选择的路径可用带宽(100 Mbps)远小于服务需求带宽.这不仅会大幅降低用户体验,同时还可能带来严重的网络拥塞问题并造成网络资源的巨大浪费.对上述流量进行恰当的路由分流能够很好地避免此示例中的问题,然而由于真实网络环境中路径可用带宽随时间动态变化,传统路由算法很难实现精确感知当前网络状态并据此进行恰当的动态路由调度.

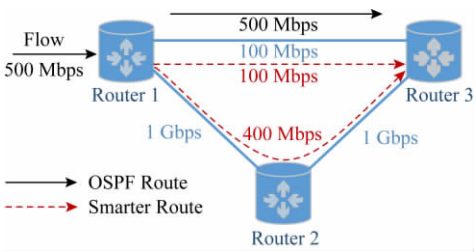


Fig. 1 A suboptimal routing decision made by OSPF

图1 OSPF所生成的非最优路由决策示意图

此外,数据中心网络等新兴网络应用场景的出现为路由优化与流量工程领域提出了新的挑战.相比于传统网络,数据中心网络带宽更大,同时存在的大流、长流更多,对于流量调度的需求与难度也更高.虽然现在已经有一些路由与流量工程的方法尝试解决各种数据中心场景下的网络优化问题,然而在数据中心网络场景中,现有路由与流量调度优化方法仍然很难满足高效利用链路以及负载均衡的需求[2].

为了满足复杂的网络应用场景以及多样化的服务质量需求,很多基于数学模型的网络层优化方案被提出[2-7].这些路由优化或流量工程方案在建模时通常会针对应用场景进行一些假设来简化问题,以使得优化问题能够利用现有数学方法高效求解.然而真实网络应用场景往往难以完全符合这些理想化假设,这使得基于数学模型的路由优化算法无法保证其在真实场景下部署的效果.实际上,即使是在经过假设简化过的场景下,很多路由优化问题的求解仍是十分复杂的,目前尚未存在一个通用的模型能够同时求解不同类型的路由优化问题[3].由于传统的路由优化任务需要针对每一种特定的场景以及特定的优化目标单独建模,将这些方法部署在真实网络环境下可能会对网络设施的可扩展性带来影响,因此传统基于数学模型的路由优化方案目前仍难以大规模部署在实际场景中.

近几年,基于深度学习的人工智能技术飞速发展并被广泛应用于自然语言处理[8]、图像识别[9]、游戏策略计算[10]等领域中.对深度学习模型的研究和

CPU,GPU 等计算机硬件的发展使得人工智能模型能够学习到的策略越来越复杂,训练和执行效率越来越高.设备算力以及模型表达能力的提升使得的人工智能模型具备了强大的学习能力和良好的泛化性,利用人工智能模型去解决路由优化问题、为网络层赋予智能正逐渐变得可能.相比于传统模型驱动的路由优化算法,数据驱动的智能路由优化算法具有 3 方面优势:1)准确性.利用真实数据对机器学习算法模型进行训练,不需要对网络环境进行复杂的假设和建模.2)高效性.多项式时间内可根据输入数据快速推理得到优化后的路由决策.3)通用性.相同的机器学习模型根据训练数据不同可以用来求解不同网络优化问题.上述 3 个优势使得数据驱动的智能路由方法相比传统路由方法能够更好地适应不同网络应用场景和路由优化目标,并使得智能路由方法在部署的过程中存在较好的可扩展性.

除了人工智能技术的飞速发展,近些年兴起的软件定义网络(software defined networking, SDN)[11]与可编程路由设备[12-13]的相关研究同样为智能路由算法提供了部署的可能.这些工作使得路由层可以完成更多、更复杂的任务.SDN 架构的出现使得基于机器学习的智能路由算法能够作为一个应用运行在具有强大运算能力的 SDN 服务器中并且有效地对路由和流量进行控制[14].不过现有基于机器学习的智能路由方案研究仍然处于比较初步的阶段,研究主要针对智能路由算法的正确性以及收敛性,智能路由算法在真实场景下的训练与部署方案仍不够完善.此外,当前路由设备的计算能力对于智能路由算法的大规模部署而言仍然远远不够[15].

本文从方法与应用场景等角度介绍了现有基于机器学习的数据驱动智能路由算法的相关工作并分析了不同智能路由方法的优劣.之后本文进一步对现有智能路由算法的训练与部署方法进行了分析总结并提出了 2 种适用于不同应用场景的智能路由算法训练部署框架.最后本文分析了基于机器学习的智能路由算法未来发展中所面临的机遇与挑战并给出了智能路由算法未来的研究方向.

1 数据驱动的智能路由算法概览

早在 1994 年 Boyan 等人[16]就提出了基于 Q-Learning 的、应用在通信网络中的智能路由算法 Q-routing.实验表明:相比于传统的最短路径路由,Q-routing 方案能够有效避免网络拥塞并降低数据包传输时延.然而虽然后续有很多相关工作对该方法进行了完善和优化[17-18],受限于路由器的计算能力以及网络层结构设计,智能路由算法难以被真正部署到真实网络场景中.

2010 年 Hu 等人[19]提出了 QELAR 方法,将 Q-Learning 的思想应用于无线传感器网络(wireless sensor network, WSN),用来优化无线传感器网络的能耗和寿命.相比于传统网络,无线传感器网络所处环境复杂多变,路由服务质量需求多样,传统路由算法在该应用场景下往往难以取得令人满意的效果.此外 WSN 与传统网络相比结构较为独立,因此基于 Q-Learning 的智能路由方法的部署难度更小.后续 Basagni 等人[20-21]进一步将 Q-Learning 方法用于无线传感器网络的可靠传输和加速转发上,取得了良好的效果.

近几年,随着深度学习技术的飞速发展,深度学习正越来越多地被应用于网络领域,并已经在包括传输层拥塞控制[22]、网络安全检测[23]、视频流传输优化[24]等领域取得了显著进展.利用深度学习解决路由优化问题也得到了更多的关注,一些基于深度学习和深度强化学习的路由算法被提出[25].这些智能路由算法既有利用深度学习对传统路由算法进行改进[26],也有针对数据中心网络流量调度、骨干网流量工程[14]等近些年新出现的网络应用场景进行全局性能优化.

随着越来越多的智能路由算法的提出,如何将数据驱动的智能路由算法部署在真实环境中同样成为了一个备受关注的问题.Mao 等人[15]的工作对基于深度学习的智能路由算法在真实场景下部署的前景进行了探讨并提出了一种利用配备了 GPU 的软件定义路由器(SDR)来部署基于深度学习的智能路由算法的框架设想.然而根据我们的研究,现有研究工作仍然没有给出一套切实可行的将智能路由算法部署在现有计算机网络体系架构中的方案.

经过调研,近年来数据驱动的智能路由算法依照其所应用的机器学习方法类型主要分为基于监督学习的智能路由算法以及基于强化学习的智能路由算法.

2 基于监督学习的智能路由算法

2.1 应用于智能路由中的监督学习方法概述

监督学习是指利用已知的输入输出样本训练模型,使得模型能够准确地完成从输入到输出映射的

一类机器学习任务[27].近年来所提出的基于监督学习的智能路由方法主要基于深度学习模型.相比于传统监督学习方法,深度学习模型能够通过带标签的数据学习得到更加复杂的策略,为实现应用于复杂网络环境下的智能路由方法提供了可能.在本节中我们将对现有智能路由方法中常用的深度学习方法进行简单介绍.

最常见的深度学习模型是深度神经网络(deep neural network, DNN),其模型设计模拟了生物神经元的工作原理,工作过程包括前馈过程和反馈过程.图 2 中给出了其模型结构与工作过程.在 DNN 的前馈过程中,模型将输入向量利用线性加权与激活函数相结合的方式逐层向前传递,最终实现输入到输出的映射.在 DNN 的反馈过程中,模型将实际输出结果与期望结果的偏差逐层反向传递完成模型参数的调整过程,达到自动学习的效果.作为对 DNN 模型的改进,Hinton 等人[29]于 2006 年提出了深度置信网络(deep belief network, DBN).DBN 模型将传统 DNN 模型与受限玻尔兹曼机(restricted Boltzmann machine, RBM)相结合,训练过程可以被视作利用 RBM 对 DBN 模型的参数进行初始化和利用梯度反向传递过程对 DBN 模型参数根据任务进行微调 2 部分.作为一个基础深度学习模型,DBN 模型可以被用于包括路由优化在内的多种任务中.

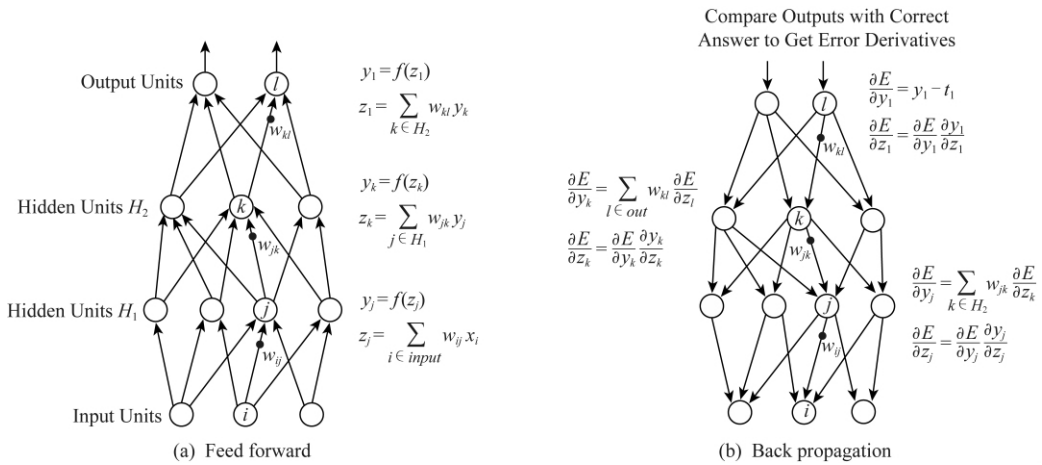


Fig. 2 Feedforward and backpropagation of deep neural networks[28]

图 2 深度神经网络的前馈与反馈原理示意图[28]

在智能路由方案中很多时候需要处理维度不定的序列化信息,例如路径信息提取[30],基于过往流量信息预测下一时刻流量[31].在这些任务中仅仅通过 DNN 模型就很难达到期望的效果,这时往往需要用到循环神经网络(recurrent neural network, RNN).RNN 能够很好地处理不定长度的序列化输入[28],对于网络流量信息的时序性、路径特征的有序性具有良好的保证.图 3 中给出了 RNN 网络的模型结构.作为 RNN 模型的改进,长短期记忆单元(long short-term memory, LSTM)[32]以及门控循环单元(gated recurrent unit, GRU)[33]在现有工作中具有更好的效果并被广泛使用.

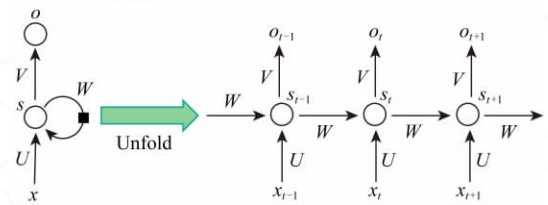


Fig. 3 Recurrent neural network (RNN) and unfolding[28]

图 3 循环神经网络及其模型展开示意图[28]

在智能路由方案中,当前网络的局部或全局拓扑信息是完成智能路由决策的重要依据,然而由于网络拓扑的动态变化性,传统深度学习模型往往难以很好地处理这部分信息.图神经网络(graph neural network, GNN)是近年来被提出的,被认为能够有效处理拓扑信息提取问题的新型神经网络结构[34].GNN 模型将网络节点与边的特性进行向量化表示,并进行若干轮迭代.每一轮迭代过程中,这些节点和边信息的向量化表示会根据拓扑依赖关系利用基于深度学习模型的更新函数进行更新.最终这些节点与边的向量化表示将收敛到确定值,代表着 GNN 模型已经将拓扑信息转化为了可被深度学习模型利用的向量化表示信息.研究表明,GNN 模型具有

良好的可扩展性与泛化性,并已经被广泛应用于网络拓扑信息提取任务中[35].

2.2 基于深度学习的智能路由算法

深度学习在路由优化问题中最直接的应用就是利用深度学习模型去代替原本基于数学模型的路由求解算法.一个普遍的路由求解模型如图 4 所示,即将网络拓扑以及网络状态信息作为输入,深度学习模型根据输入信息做出符合当前网络环境状态的恰当路由决策.

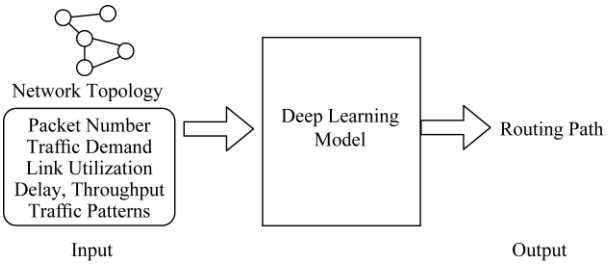


Fig. 4 Scheme of deep learning based routing model

图 4 基于深度学习的智能路由算法框架

Mao 等人[15]在 2017 年提出了一种基于深度置信网络(DBN)的路由决策方案.图 5 给出了该方案的整体模型示意图,Mao 等人的智能路由方案应用场景为骨干网络,该方案将路由器分为域内路由器与边界路由器.数据包在经由边界路由器进入主干网时部署于边界路由器上的 DBN 模型会根据当前网络各节点流量状态为每个数据包计算其在主干网内的转发路径,其后数据包经由域内路由器转发到目的边界路由器并最终离开改主干网.在上述模型中,域间路由器只负责路由转发和网络状态信息收集,从而避免了传统分布式路由算法中频繁的网络拓扑信息交换.该方案的路由决策模型为每个路由节点到每个目的边界路由器路单独训练一个 DBN 模型用来根据网络状态信息输出恰当的下一跳节点,路由路径计算过程采用逐跳的方式依次通过对应的 DBN 模型生成.Mao 等人的工作表明基于深度学习模型的路由策略能够达到 95%准确率,与此同时,深度学习模型所具有的基于部分网络状态特征进行路由决策的特点也使得基于深度学习的智能路由方法相比传统路由方法具有更低的信息交换成本以及当网络环境发生变化时更快的路由收敛速度.然而,上述方案的部署不仅需要骨干网路由器具备极强的模型计算能力,同时还需要对现有路由协议进行修改,因此在现有计算机网络体系结构下部署上述方案需要极高的成本并且会严重影响网络的可扩展性.

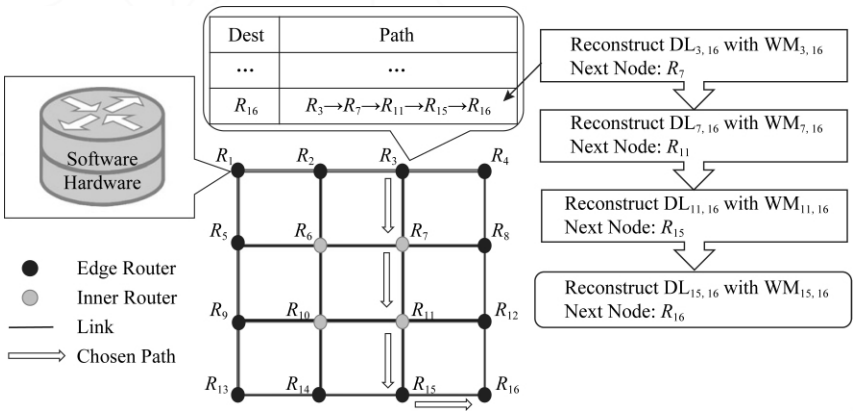


Fig. 5 Considered system model of the DBN-based routing protocol[15]

图 5 基于 DBN 的智能路由算法系统模型[15]

除了 DBN 模型,其他深度学习模型同样被尝试应用于智能路由任务中.Zhuang 等人[36]的工作对于应用不同深度学习模型学习路由决策的效果进行了对比,该工作中将逐跳智能路由决策过程形式化表示为:

$$\langle src, dst\rangle_{n+1} = F(\langle src, dst\rangle_n, dst, G),$$

其中,src,dst 分别表示源、目的节点,$\langle src, dst\rangle_n$ 是从 src 到 dst 的路由中的第 n 个路由节点编号;$F()$ 是路由决策函数;G 代表拓扑结构信息.通过实验发现将基于拓扑结构的特征提取方式与深度学习模型相结合的方案(graph-aware deep learning, GADL)相比单纯采用 DBN,CNN 等现有深度学习模型能够有效提升模型测试准确率并降低模型训练时间.

更进一步地利用拓扑结构信息,Geyer 等人[26]基于 GRU 和 GNN 设计了分布式智能路由算法.为了使得 GNN 模型能够更好地表现路由网络结构特点并使得 GNN 建模的网络特征信息能更方便地用于

路由决策过程,该方案将路由器接口作为额外节点加入图模型中.图 6 中给出了将路由器接口作为额外节点加入后的图模型示意图.当 GNN 完成了拓扑结构建模之后,每个路由器接口对应的节点信息向量化表示 $\boldsymbol{h}_v$ 不仅包含了自身信息,同时由于 GNN 的信息传递特性使得该节点同时会包含路由决策所需的全网结构和状态信息.利用路由接口信息 $\boldsymbol{h}_v$,每个路由器能够在本地计算出到对应目的节点所应该通过的路由器接口.由于 GNN 的模型特性,上述 GNN 拓扑结构建模的迭代过程可以通过将 GNN 参数更新函数部署在每个路由器上的方式分布式地完成,因此该方法天然具有良好的可扩展性与分布式路由决策的能力.该工作的仿真实验表明,基于 GNN 的分布式智能路由算法在路由收敛速度、准确性、鲁棒性、故障适应性方面表现良好,其中对于最短路径路由,经过训练的 GNN 模型能够在 15 轮迭代之内达到 98%的准确率,而对于最大最小公平路由[37]算法能够在 15 轮迭代之内达到 95%的准确率.

结合图 7 中的内容能够发现,现有基于深度学习模型的智能路由方案主要通过逐跳的方式生成路由路径.与逐跳路由生成方式相对应的另一种路由模式是预先计算所有可能路径,通过深度学习模型根据网络状态选择恰当的路径.这种基于路径选择的方式能够避免路径生成模型所带来的路由环路等问题,具有更好的效果保障.然而网络中的可选路径数会随着网络规模的增大指数级增长,其巨大的输出维度使得基于路径选择的深度学习模型的学习难度以及模型参数数量处于难以承受的数量级[38].此外由于网络路径特征与拓扑结构具有很强的相关性,基于路径选择的深度学习模型很难具有足够的通用性和泛化性.相比于路径选择的方式,采用逐跳生成路径的方式能够显著降低输出维度以及模型决策难度,使得路由决策的准确率明显提升[38-39].

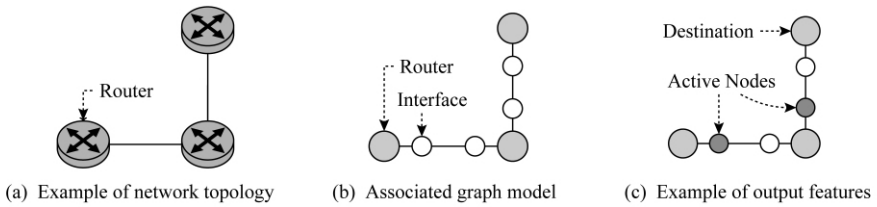


Fig. 6 Graph model with extra nodes for router interfaces[26]

图 6 将路由器接口作为额外节点的网络图模型[26]

Learning Mode	Training Mode	Controlling Mode	Deployment Mode	Routing Policy	ML Algorithm	Reference
Supervised Learning	Offline	Packet-controlled	Decentralized	Path Generation	DBN	Ref[15,25,39]
				Path Generation	GNN	[26]
				(A)Congestion Prediction	DNN	Ref[40]
			Centralized	(A)Delay and Jitter Prediction	GNN	Ref[30]
		Flow-controlled	Centralized	Path Generation	GADL	Ref[36]
Reinforcement Learning	Online	Packet-controlled	Decentralized	Path Generation	Q-Learning	Ref[16-21]
		Epoch-controlled	Centralized	Setting Splitting Ratio	DDPG	Ref[14]
	Offline	Epoch-controlled	Centralized	Setting Link Weights	TRPO	Ref[31]
					MADDPG	Ref[38]

"(A)" denotes that the machine learning algorithm only acts as an auxiliary part of the routing model

Fig. 7 Summary of machine learning based routing model

图 7 基于机器学习的路由方法概述

现有工作表明,基于深度学习的智能路由算法能够基于部分网络状态信息快速、准确地计算出对应的路由决策,并且在信息传递成本、路由收敛速度等方面相比传统分布式路由展现出了一定的优势.

基于GNN的分布式路由决策在拓扑信息建模、鲁棒性以及故障适应性等基于传统深度学习模型的智能路由方案难以解决的问题上面已经取得了一定的进展.然而现有基于深度学习模型的智能路由算法主要学习的是基于最短路径的路由算法,其能否很好地学会更多复杂的动态路由算法是值得更进一步探讨的.此外,现有基于深度学习的智能路由算法无法保证其在复杂多变的网络环境下的安全性和鲁棒性,并且需要高昂的部署成本,因此基于深度学习的路由算法想要替代传统路由算法仍有很长的一段路要走.

2.3 利用智能模块辅助路由计算

现有的深度学习方法在网络建模、流量预测、拥塞检测方面已经取得了一定的成果[31,41-42],利用深度学习方法在这些领域的成果来辅助路由计算是使得路由算法变得更加智能的另一种途径.在路由优化问题中,有很多时候传统基于模型优化或者启发式的方法都需要涉及网络环境建模、流量预测、拥塞检测等模块,用深度学习方法来替代这些模块有时会取得比较好的效果.

Barabas等人[40]的工作利用基于多任务学习的深度神经网络预测器根据链路历史状态数据为每条链路进行链路拥塞预测,并将预测得到的结果与基于规则的拥塞避免和重路由方案相结合,使得路由方法能够在拥塞发生前主动调整路由而不是发生后被动地亡羊补牢.

Rusek等人[30]的工作将GNN与LSTM模型相结合,用基于图神经网络的深度学习模型对路由路径时延和时延抖动与网络拓扑结构、流量矩阵以及路由路径之间的关系进行建模,并利用所建立的模型辅助启发式路由优化算法进行路由策略计算.研究结果表明基于GNN的网络建模能够根据输入信息准确预测路由路径时延和时延抖动,并且对于没有在训练中出现的拓扑以及动态变化的路由路径展现了良好的泛化性.数据驱动的网络建模方法为基于探索的启发式路由优化算法提供了一个准确、高效的路由策略试验环境,使得启发式的路由优化算法能够以低成本完成路由优化求解过程,同时避免了因为网络建模与真实环境不符所带来的路由策略效果损失.

利用深度学习模型辅助传统路由算法的方案能够有效提升传统路由优化算法性能,与此同时传统路由优化算法保证了智能路由方案具有更强的可靠性与可解释性.因此未来将传统路由优化算法与深度学习模型相结合可能是智能路由算法发展的一个途径.

3 基于强化学习的智能路由算法

3.1 应用于智能路由中的强化学习方法概述

一个标准的强化学习过程可以被看做一个强化学习单元在离散时间步数内与环境交互的过程.在每个时间点 t,强化学习单元根据状态 s_t,采取行动 a_t,并且受到一个反馈奖励 r_t.强化学习的目标是寻找到一个策略 $\pi(s)$,该策略函数是一个从状态到行动的映射并且能够最大化递减奖励和 $R_0=\sum_{t=0}^{T}\gamma^t r_t$,$\gamma\in[0,1]$是奖励折扣因子.

Q-Learning方法采用一个Q函数来预测时刻 t 观测到的状态 s_t 和动作 a_t 对应的最大递减奖励和,Q函数的定义为:

$$Q(s_t,a_t)=\max_{\pi}\{E[R_t|s_t,a_t,\pi]\}.$$

对于Q函数的计算有基于模型和模型无关2种方法.其中基于模型的方法通过Markov决策过程中各状态间的关联模型对Q函数进行直接求解,形式化表示为:

$$Q(s_t,a_t)=r_t+\gamma\sum_{s_{t+1}\in S}P^{a_t}_{s_ts_{t+1}}V(s_{t+1}),$$

$$V(s)=\max_{a}\{Q(s,a)\},$$

其中,V 函数为状态价值函数,表示对应状态下所能够获取的最大递减奖励和,$P^{a_t}_{s_ts_{t+1}}$ 代表了强化学习任务对应Markov决策过程的状态转移概率.在强化学习任务中,状态转移概率并不总是易于获取的,此时可以采用状态无关的方法估计Q函数:

$$Q(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha[r_t+\gamma V(s_{t+1})],$$

其中,α 是模型学习速率,相比于基于模型的Q函数计算方法,模型无关的Q函数计算方法通常需要更长的收敛时间.

在传统的Q-Learning方法中,Q函数是一个从有限状态决策空间 $S\times A$ 到实数空间$\mathbb{R}$的映射,为了处理连续高维状态决策空间上的强化学习问题,研究者们将深度学习模型引入强化学习框架,设计出了多种深度强化学习(deep reinforcement learning, DRL)模型.

Google Deep Mind机构提出了深度Q值学习(deep Q-Learning, DQN)[43].DQN采用一个深度神经网络(DNN)来代替原本的Q值表来近似估计Q函数,并通过平方误差进行训练:

$$L(\theta^Q)=E[(y_t-Q(s_t,a_t|\theta^Q))^2],$$

这里 θ^Q 是 DQN 的参数，y_t 是目标值，可计算为：

$$y_t=r_t+\gamma Q(s_{t+1},\pi(s_{t+1})|\theta^Q),$$

其中，$\pi(\cdot)$是一个能够最大化预期总收益的策略函数，一个常用的异步策略是采用贪心的方式选择动作：

$$\pi(s_t)=\arg\max_{a_t} Q(s_t,a_t).$$

与基于 Q 函数估计的 DQN 方法相对应的是策略梯度方法[44]，策略梯度法利用深度学习模型作为策略函数 $\pi_\theta(s,a)$，通过计算策略梯度的方式直接优化策略函数.

为了进一步提升策略梯度方法的性能，加速强化学习模型的收敛速度，可以将 Q 值学习与策略梯度方法结合起来，通过价值估计函数来预测当前状态下采用行动后续会得到的价值，并利用预测结果对策略模型进行训练，这就是强化学习的演员-评价者(actor-critic, AC)框架.

一种目前常用的基于在线策略(on-policy)的 AC 框架利用一个动作优势函数 $A(s,a)$来对策略优劣进行估计，引入优势函数后的策略梯度为

$$\nabla J(\theta)=E_{\tau\sim p_{\theta(\tau)}}[\nabla_\theta \log \pi_\theta(s,a)A(s,a)],$$

其中，τ 代表状态-动作元组(s_t,a_t).

基于在线策略的强化学习方法需要将训练过程与数据收集同步进行，经过多轮数据收集-参数更新的迭代过程达到参数收敛，为了将数据收集和模型训练过程解耦合，可以采用基于离线策略(off-policy)的强化学习方法，一个常用的基于离线策略的 AC 框架深度强化学习模型是确定性策略梯度算法(deterministic policy gradient, DPG)[45].该方法直接利用价值网络梯度回传的方式计算策略梯度，在连续动作空间强化学习问题上取得了良好的效果.该方法的改进版深度确定性策略梯度算法(deep deterministic policy gradient, DDPG)[46]在解决连续动作空间的路由优化问题上有比较广泛的应用.

近些年来的最新工作中，为了解决传统基于随机梯度下降算法的策略优化方法所存在的策略更新过度问题，Schulman 等人[47]提出了二阶强化学习方法——置信域策略优化方法(trust region policy optimization, TRPO).虽然二阶方法具有比一阶方法更好的收敛性保证，其过高的计算复杂度限制了它的应用场景.基于 TRPO 的思想，OpenAI 与 DeepMind 提出了近端策略优化方法(proximal policy optimization, PPO)[48]，该方法兼具了传统一阶方法的高效和易于实现的特性以及置信域算法的数据效率和可靠表现，成为了当前的主流强化学习算法之一.

3.2 基于 Q-Learning 的智能路由算法

1994 年 Boyan 等人[16]的工作 Q-routing 第一次将 Q-Learning 用在了路由算法上面.Q-routing 将路由转发过程用 Markov 决策过程(Markov decision process, MDP)进行建模，将每个路由节点视作 MDP 中的状态，路由下一跳所选择的邻居节点作为 MDP 中的动作，路由每一跳所花费的时延作为强化学习一次动作所获得的反馈值.Q-routing 中用 Q 值函数 $Q_x(d,y)$来预测从当前节点 x 到目标节点 d 采用下一跳节点 y 所需花费的时间.每当节点 x 向邻居节点 y 发送一次数据包，节点 y 立刻会返回预估的剩余路程时延 t 给 x：

$$t=\min_{z\in \text{neighbors of } y} Q_y(d,z),$$

此时利用基于模型的 Q-Learning 方法，节点 x 可以动态更新自身对应的 Q 值函数信息，形式化地：

$$\Delta Q_x(d,y)=\eta(q+s+t-Q_x(d,y)),$$

其中，η 是算法学习速率，q 和 s 分别是从 x 到 y 的队列时延和传输时延.根据动态更新的 Q 值函数，Q-routing 能够自适应动态变化的网络状态并为每个数据包选择时延最短的路由路径.相比于传统最短路径路由算法，Q-routing 将时延而不仅仅是路由跳数作为衡量路径长短的指标，因此能够有效避免网络拥塞的发生.

虽然 Q-routing 能够很快地感知网络拥塞的发生并调整路由路径来实现拥塞避免，该方法很难快速地感知到拥塞消除情况.由于 Q-Learning 模型所限，对于因对应路径发生拥塞而导致短时间内不被采用的邻居节点，Q-routing 方法中路由器只能通过向邻居节点发送额外的请求数据包的方式来更新其对应的 Q 值表，这不仅带来了额外的数据传输成本，而且受限于额外请求数据包的发送频率，在全网范围内完成拥塞消除情况的传递需要一个较长的时间，这使得 Q-routing 实际上难以达到最优的路由调度效果.为了做到快速感知拥塞恢复，Choi 等人[17]对于 Q-routing 中的拥塞恢复过程与时间的关系进行了建模，提出采用 R 函数来对 Q 函数随时间的变化速率进行估计，并将 R 函数用于路由决策时对当前各邻居节点对应 Q 值的计算.实验表明基于 Q 值变化预测的 Q-routing 方案在网络拥塞频繁出现的情况下相比于原本的 Q-routing 方案具有更好的收敛速度和稳定性.此外，Kumar 等人[18]利用对偶强化学习对于 Q-routing 进行了改进并获得了更好的性能.

2010 年 Hu 等人[19]的工作将 Q-Learning 的方法应用在了无线传感器网络(WSN)中,提出了 QELAR 方案.由于 WSN 的工作环境复杂,网络拓扑结构经常变动,所以传统路由方法应用在 WSN 环境下往往无法取得很好的效果.QELAR 主要解决 WSN 的寿命问题,类似于 Q-routing,QELAR 同样将数据包在网络中传输的过程用 Markov 过程进行建模,不同的是 QELAR 将当前节点及其邻居节点的剩余能量状态与路径跳数相结合作为强化学习的反馈,使得路由算法能够根据当前系统剩余能量状态进行智能路由决策,以保证 WSN 网络正常工作的时间尽可能长.

在 QELAR 之后,Basagni 等人[20-21]又提出了 MARLIN 和 MARLIN-Q 模型,将 WSN 网络的数据包发送与重传过程用 MDP 进行建模.图 8 中展示了 MARLIN-Q 方案中每个路由节点控制数据包进行转发的状态转移模型示意图.在 MARLIN 与 MARLIN-Q 工作中,数据包 p 在每个路由节点的状态空间 S 根据当前数据包重传次数进行定义

$$S=\{0,1,\cdots,K-1\}\bigcup\{rcv,drop\}.$$

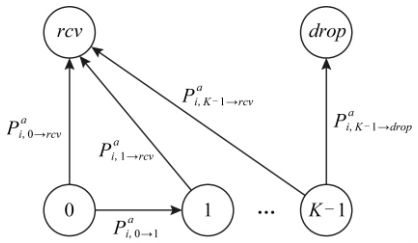


Fig. 8 States and transitions of node i handling packet p as shown in MARLIN-Q[21]

图 8 MARLIN-Q 中节点 i 处理数据包 p 的状态转移模型[21]

每个路由节点 i 在状态 s 可进行的动作空间包括选择的调制解调器类型以及对应的调制解调器所能到达的下一跳路由节点:

$$A_i^M(s)=\{a=\langle j,m\rangle \mid m\in M, j\in Neighbor_i^m\},$$

其中,M 为该节点所具有的调制解调器类型集合,$Neighbor_i^m$ 表示该节点采用调制解调器类型 m 所能到达的邻居节点集合.MARLIN 系列算法通过巧妙地设计反馈函数使得每个节点的强化学习模型所获的反馈值与数据包传输时延正相关,与此同时对于一个丢包($drop$)行为施加以很大的惩罚,这使得系统能够用于水下传感器网络的可靠低时延数据传输.在真实场景下,通过不停的尝试和学习,MARLIN 系列模型可以通过历史数据自适应地计算出状态转移概率 $P_{i,s\to rcv}^{\langle j,m\rangle}$,继而通过基于模型的 Q-Learning 方法保证 WSN 网络的路由传输质量(quality of service, QoS).此外通过改变 MDP 过程的最大重传次数 K,MARLIN-Q 能够支持不同类型的 QoS 需求,例如需求低时延的加速转发服务以及需求保障可靠性的可靠传输服务.MARLIN-Q 在仿真环境中对不同网络参数和负载下的算法性能进行了测试,结果表明相比于现有最先进的水下传感器网络路由传输算法 CARP[49]以及针对网络寿命进行优化的 QELAR 算法,MARLIN-Q 算法能够有效避免数据包传输过程中的失败重传,在有效吞吐、时延和能耗方面具备更好的性能.

经过调研,现有基于 Q-Learning 的智能路由算法大都将数据包在网络中的转发过程用 MDP 进行建模,之后将路由优化问题转化为基于模型的 Q-Learning 问题,并在此基础上构建智能路由算法.由于 MDP 建模以及基于模型的 Q-Learning 本身的特点,其优化目标主要为时延、吞吐、能耗等可逐跳累加的性能评价指标.利用基于模型的 Q-Learning 方法设计的智能路由算法本身能够自适应动态变化的网络环境,且由于其 MDP 模型已知,其决策过程相比于其他基于深度学习的方法具有更好的可解释性,因此在网络状态波动性很大的应用场景中,例如 WSN 网络,具有比较广泛的应用.然而对于输入输出维度更高、优化目标更复杂的路由优化问题显式地建立 MDP 模型十分困难,此外现有基于 Q-Learning 的路由优化方法普遍采用的包级别的路由控制方式难以满足主干网的高性能需求,因此现有基于 Q-Learning 的智能路由算法的应用场景仍然具有很大的局限性.

3.3 基于深度强化学习的智能路由算法

随着近几年深度学习技术的发展,研究者们开始尝试将深度强化学习技术(DRL)应用到智能路由与流量工程方案设计中.相比于 Q-Learning,DRL 方法能够学习到更复杂的策略,以解决状态、决策空间更大以及优化目标更复杂的路由优化问题.

Xu 等人[14]将深度强化学习用于域内流量工程问题中提出了基于深度强化学习的流量工程方案 DRL-TE.类似于 2018 年 Kumar 等人[7]提出的经典的半状态无关流量工程方案 SMORE,DRL-TE 将流量工程问题划分为静态多路径求解以及在线动态调整路径分流比 2 部分.DRL-TE 采用传统方法生成路径,并利用一个深度强化学习单元来完成在线动态调整路径分流比过程.DRL-TE 方案中深度强化

学习模型将当前每个会话对应的时延和吞吐作为强化学习的状态,将路径分流比作为强化学习的动作,将每个会话的性能评价函数作为强化学习的反馈,从而动态感知网络状态信息,控制各条路径的分流比,并根据各会话反馈结果自适应地学习最优分流策略.为了处理分流比所带来的连续动作空间问题,DRL-TE 采用深度确定性策略梯度算法(DDPG)作为强化学习模型,并采用了专为流量工程设计的经验回放方式来保证强化学习模型的收敛性和稳定性.相比于 SMORE 需要准确预测下一时刻的流量矩阵才能利用线性规划模型解出最优的分流比并且只能优化有限的目标(例如最大链路利用率),DRL-TE 只需根据各会话当前流量特征信息即可自动预测未来的流量变化情况,并做出能最大化各会话总效益函数值的决策.因此,DRL-TE 相比于 SMORE 方法对应用场景需求更少的假设,具有更好的通用性和鲁棒性.DRL-TE 在 ns-3 环境下进行了仿真实验,实验结果表明:相比于传统路由以及流量工程算法,DRL-TE 不论在时延、吞吐还是文中定义的效用函数指标上都具有明显优势.此外直接采用原始 DDPG 算法的对比实验表明利用机器学习模型解决流量工程问题时对原有机器学习算法进行针对性地改进是十分必要的,直接将现有机器学习模型应用在路由优化与流量工程问题中可能难以达到十分理想的效果.

除了流量工程领域,深度强化学习同样被应用于智能路由配置优化任务中.Valadarsky 等人[31]尝试利用深度强化学习单元根据历史流量数据对未来的网络流量进行预测,并基于强化学习模型的流量预测能力计算出恰当的路由配置.在这篇工作中,Valadarsky 等人将历史流量矩阵作为强化学习模型的输入,每条链路的权值作为强化学习模型的输出,强化学习模型(TRPO)根据学习到的经验和知识通过历史流量矩阵对未来流量进行预测并通过调整链路权值来进行路由配置,以达到优化全网最大链路利用率并完成负载均衡的目标.Valadarsky 等人的工作中指出,路由规则的表现形式与强化学习模型的收敛性有很强的相关性.对于一个网络拓扑 $G(V,E)$,如果直接采用一个输出维度为 $|V|\cdot|E|$ 的基于目的节点的路由规则形式作为上述强化学习模型的输出动作,即为每个节点 v 针对每个目的节点 d 设置一个对其所有邻居节点的分流比,那么由于输出维度过高上述强化学习模型将难以收敛.因此该工作中强化学习模型的动作为每条链路设置一个实数权值,链路权值通过一个传统基于规则的方式映射成为路由规则.这使得强化学习模型的输出维度降为 $|E|$,以降低强化学习模型的动作空间大小,减轻探索和学习难度,达到加速收敛的效果.该工作采用了稀疏和非稀疏的重力/双峰模型生成了不用类型的流量矩阵序列用于检测算法性能.仿真实验结果表明对于具有明显规律特征的流量矩阵,强化学习模型能够通过流量预测来实现良好的路由配置,达到优于流量无关最优路由[50]并且接近最优的路由配置效果.然而当流量矩阵不再具有明显规律特征时,该方法的性能就会显著下降.实际上,真实场景下的流量变化可能是无规律的,包含许多突发流量的,因此对于上述模型在真实流量数据下的流量预测和路由配置能力仍然是一个值得探索的问题.

虽然 DRL 模型理论上能够根据网络状态数据或历史信息对未来的流量进行预测并作出最优的路由决策,在目前实验中 DRL 模型的结果还远远没有达到最优.Xu 等人[38]的工作对比了若干种强化学习模型在路由任务上的效果,提出了将强化学习模型用来解决路由问题的指导性建议.首先作者通过一个 Q-routing 模型[16]简单场景部署实验表明包级别路由控制的强化学习智能路由模型对于吞吐较高的应用场景难以适用,采用时间段级别路由控制模型将会是比较推荐的方式.其次,将显式的路径选择方式作为强化学习单元动作的智能路由方案难以收敛到理想结果.正如 2.2 节所提到的,路径数目随网络规模的增长而指数增长,基于路径选择的方案无疑会大幅增加强化学习模型的学习和探索能力.基于上述 2 点,本文最终同样选择了通过强化学习模型来控制链路权值继而间接实现路由控制的方案.与 Valadarsky 等人直接生成实链路权值相比,Xu 等人的方案将链路权值离散化处理,进一步将动作空间大小从无限降为了有限,并对每一条链路对应的权值选择过程单独采用一个强化学习模型进行处理,进一步减小了每个强化学习模型的决策难度和探索空间.生成的链路权值作为最短路径算法的边权来进行路由计算.为了保证这个多智能体的合作路由模型的策略一致性,Xu 等人利用最新的多智能体深度确定性策略梯度算法[51](multi-agent deep deterministic policy gradient, MADDPG)来对模型进行训练.最终的实验结果表明基于离线链路权值的强化学习智能路由算法相比于最短路径路由具有更好的负载均衡特性,即更短的路由器平均等待队长.

现有基于深度强化学习的智能路由方案在域内流量工程和智能路由优化任务上已经取得了一定的成果.深度强化学习模型具有良好的通用性与泛化性,其既可以优化网络全局性能评价指标,例如全网最大链路利用率、路由器平均等待队长等,也可以优化每个会话对应的私有效益值函数.此外相比于传统基于规则或数学模型的路由优化算法,基于深度强化学习的智能路由算法无需对环境做出假设,并且能够自适应动态变化的网络环境.然而,不难发现,深度强化学习模型的收敛性与其生成路由规则的形式间具有很强的关联性,过高的输出维度往往使得深度强化学习模型无法收敛.因此现有研究工作中,深度强化学习模型普遍通过控制路径分流比或链路权值的方式间接完成流量控制,而非通过路径选择或路径生成的方式直接生成路由路径.实际上即使现有工作已经尽量降低深度强化学习单元的路由决策难度,并取得了显著进展,现有方案在复杂应用场景下的表现仍然有很大的提升空间.另外受限于深度强化学习的模型性能,现有方案大部分都采取时间段级别的路由控制方式,包级别的路由控制方式则不太适合于此类智能路由方案.对于路由算法而言,鲁棒性和可靠性是十分重要的性质,然而现有基于深度强化学习的智能路由算法在这方面的研究还远远不够.

4 智能路由算法的训练与部署

虽然近些年已经有很多基于机器学习的智能路由算法相关工作,但是这些工作主要针对智能路由算法的原理设计和算法准确性、收敛性等问题进行研究,而对于智能路由算法在真实场景下的训练与部署还尚未有一个成熟且完整的框架.本文对智能路由算法不同的训练方式与部署方式的优势与不足进行了讨论,并提出了2类较为合理的智能路由训练与部署框架以使得智能路由算法能够低成本、高可靠性地在真实场景被部署.

4.1 训练方式:在线与离线

智能路由算法模型的训练方式主要分为在线和离线2种.图7中给出了现有智能路由方案的训练方式.其中基于监督学习的智能路由模型全部采用离线训练的方式;而基于强化学习的模型则既可以在真实环境下在线训练也可以在仿真环境下进行离线训练.

通常来说,模型的离线训练过程首先需要从真实环境中收集数据,这些数据可能是流量矩阵、网络各节点状态信息以及对应的路由决策标签等.数据经过处理后被用于机器学习模型在服务器上的离线训练过程.训练完成后模型被部署到真实环境中进行在线路由决策.离线训练和在线测试、部署是深度学习领域常见的训练部署方式,然而对于智能路由算法,离线训练往往面临着3个挑战:1)训练数据的收集可能需要比较高的成本;2)真实场景下的网络状态可能与训练数据集不同,导致路由算法无法达到预期效果甚至出现错误;3)对于强化学习来说,搭建与真实环境近似的仿真训练环境可能很困难.

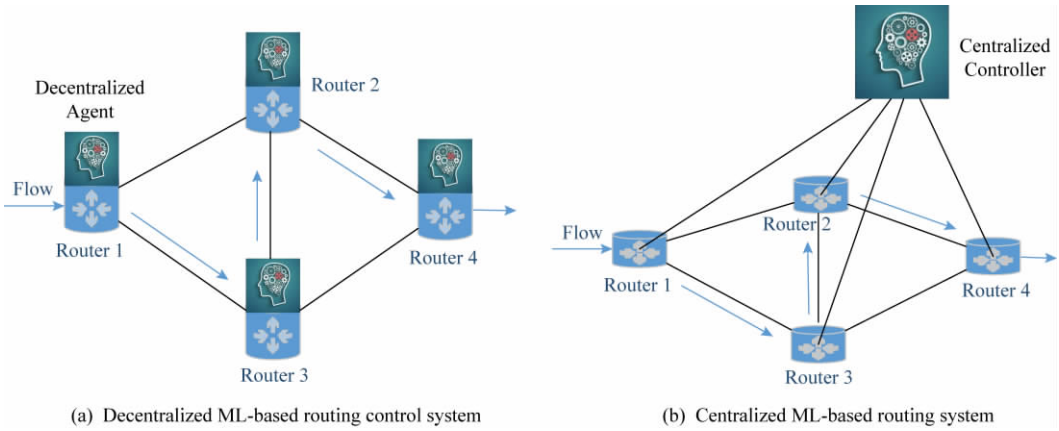
对于强化学习方法,在线训练可以保证模型自适应网络环境的变化,并且避免离线仿真环境搭建所带来的困难与额外成本.然而在线训练所带来的路由安全性和可靠性问题使得实际部署中往往难以部署需要在线训练的智能路由方法.实际上,在在线强化学习中,安全问题是一个已经被广泛研究的问题[52-53],强化学习模型在训练的初始阶段以及训练过程中的探索阶段都可能会产生难以预测的行为,当强化学习方法应用在路由任务中时,这些难以预测的行为可能造成包括路由环路、链路拥塞等严重后果.因此,保证在线强化学习路由算法训练过程的安全性与可靠性将是其在真实场景下部署的重要前提.

4.2 部署方式:集中式与分布式

随着越来越多的智能路由算法的提出,如何在现有计算机网络体系结构中部署这些算法正受到越来越多的关注.智能路由算法的部署方式主要分为分布式与集中式2种.

图9中给出2种部署方案的框架结构示意图.智能路由算法部署于集中式控制器中,根据控制器所收集到的网络状态信息动态进行路由决策,路由决策通过集中式控制器下发至各路由节点中.SDN网络结构的提出为智能路由算法的集中式部署在理论上提供了可能,通过将智能路由控制单元作为SDN控制器上的一个应用可以完成上述集中式控制过程.在数据中心网络流量工程这样相对独立的应用场景下,采用集中式方法部署智能路由调度方案是一种现阶段较为可行的方案.

集中式方案部署需要在网络中部署一个集中式的路由控制器,并设计一个集中式的路由控制协议,然而当前计算机网络体系结构中路由协议依然以分布式路由协议为主.相比于集中式路由协议,分布式路由协议具有更好的可扩展性.从图7中可以看出,现有智能路由算法中有很多能够支持分布式路由



(a) Decentralized ML-based routing control system (b) Centralized ML-based routing system

Fig. 9 Comparison between decentralized and centralized machine learning based routing control system

图 9 分布式智能路由控制系统与集中式智能路由控制系统结构对比

决策,这些分布式智能路由算法在收敛性、鲁棒性等方面已经取得了进展,然而想要真正部署,还需要对应的路由器硬件的进一步发展和完善[15].随着可编程路由设备的发展,未来在真实网络中部署分布式智能路由算法将会成为可能.然而现有分布式智能路由算法主要关注路由方法的准确性以及收敛性,并没有考虑对现有网络层结构与协议的兼容.对于分布式智能路由算法而言,如何在兼容现有网络层结构的基础上进行增量式部署将是一个未来值得思考的问题.

4.3 智能路由训练与部署模型设计

本节基于上述讨论总结并提出了 2 类未来具备可行性的智能路由训练与部署框架:1)集中式离线训练与在线决策相结合的智能路由框架;2)保证安全的在线强化学习路由框架.

图 10 中给出了集中式离线训练与在线路由决策相结合的智能路由部署框架的工作流程图.在这种智能路由部署方案下,路由器数据平面需要收集网络流量特征信息并向上传递给控制层用来完成智能路由模型的训练以及在线路由决策过程.智能路由决策模型在一个单独的具有足够计算能力的节点利用历史网络状态信息以及网络仿真环境完成离线训练,并将训练好的模型参数发布到在线路由决策单元中.对应的路由决策单元既可以采用分布式部署的方式将在线智能路由单元部署到每个路由器的控制平面,也可以采用集中式部署的方式将智能路由单元放在一个集中式的路由控制器中,例如 SDN 控制器.为了适应随时间动态变化的网络拓扑结构以及流量特征,上述模型采用闭环学习的方式定期根据最新的网络流量特征对智能路由模型进行增量式训练.基于机器学习的智能路由模型的训练过程需要消耗大量计算和存储资源,采用集中式的离线训练使得网络各路由节点不需要额外部署这些资源,能够有效降低智能路由算法的部署成本.

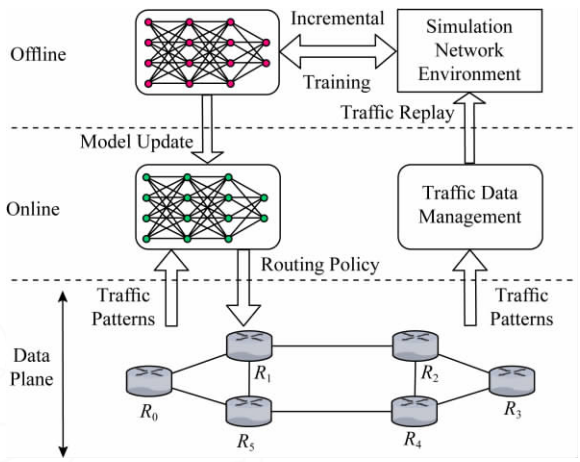


Fig. 10 Centralized offline training and online deployment model for machine learning based routing protocol

图 10 集中式离线训练与在线部署相结合的机器学习智能路由部署方案

集中式离线训练加在线路由决策的智能路由部署方案适用于大多数现有智能路由算法,并且与机器学习离线训练、在线决策的思想相吻合.然而对于强化学习模型而言,无论是在线策略(on-policy)模型还是离线策略(off-policy)模型,与环境的交互是其学习过程必不可少的部分.不同于游戏任务,在路由优化问题中搭建一个与真实网络环境相一致的仿真环境往往依赖于对网络场景的精确建模,是一件十分困难的事情[30].与之相对应,深度强化学习模型开始阶段糟糕的策略以及其学习过程中的探索

行为,使得直接将基于深度强化学习的智能路由模型放在真实网络环境中进行训练很可能会为网络带来严重的安全性和可靠性问题.为了解决基于深度强化学习的智能路由策略在训练过程中所面临的挑战,本文参考安全在线强化学习的思想[53],提出了具有可靠性保证的深度强化学习智能路由模型在线训练方案,图 11 给出了该方案的工作流程图.相比于传统的强化学习方法,该方案引入安全监测模块对强化学习单元所做出的路由决策是否安全进行了基于规则的判断,当强化学习单元所做出的路由决策可能存在安全隐患时,例如包含路由回路、引发网络拥塞等,强化学习单元采用一个简单可靠的路由决策(例如最短路径路由)对原本的路由决策进行替换,并同时给强化学习单元施加一个惩罚因子 p,以避免强化学习单元之后再次生成类似的路由策略.在线安全学习在其他网络应用场景下的相关工作表明,基于在线安全学习的深度强化学习智能路由方案有能力在不影响原本路由优化目标的同时保证路由学习过程的可靠性[53].它不仅能解决由于模型尚未收敛以及探索过程所带来安全性问题,而且可以在不保证模型可解释性的前提下保证模型的可靠性,一定程度上解决了深度学习智能路由模型不可解释性以及网络突发状况下路由行为不可预测性所带来的担忧.

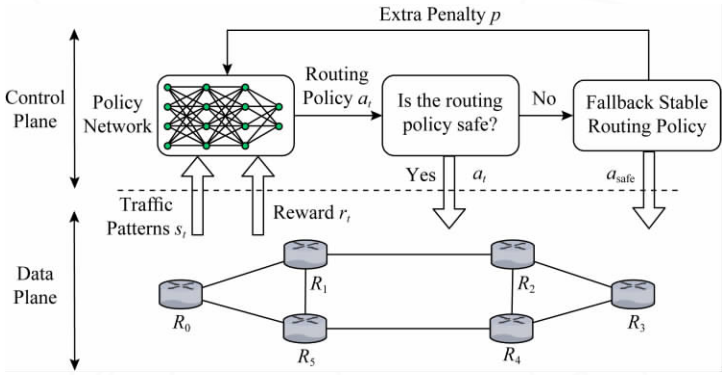


Fig. 11 Safe online learning model for deep reinforcement learning based routing algorithm

图 11 具有可靠性保证的基于深度强化学习的智能路由算法在线学习方案

对于智能路由的训练与部署框架,现有研究工作还比较少,但是本文认为由智能路由方案所带来的模型不可解释性、路由行为的不可预测性将是其训练部署框架设计的重要挑战.而利用基于规则的方案来对智能路由控制单元进行约束可能是保证智能路由的可靠性的一个有效手段.

5 智能路由算法所面临的机遇与挑战

近年来,智能路由算法受到越来越多的关注,本节针对智能路由算法在解决路由优化问题上所具有的优势以及其未来发展过程所面临的的挑战进行了探讨.

5.1 智能路由算法的优势

数据驱动的智能路由算法通常基于深度学习或强化学习,其主要具有 5 个优点:

1) 网络状态敏感.相比于传统基于模型的路由算法,智能路由算法能够处理更高维度的网络状态特征信息,这使得智能路由算法对网络状态的变化更加敏感,当网络状态发生变化时能快速收敛,做出更适合当前网络状态的路由决策.

2) 数据驱动.相比于传统路由算法基于固定的模型求解路由策略,智能路由算法由数据驱动,基于更少的环境假设,利用历史数据信息以及对环境的自发探索来自动对应用场景进行建模并完成路由优化,因此能够自适应不同应用场景与网络环境变化.

3) 面向服务质量.智能路由能够更好地支持区分服务质量的路由请求.相比于传统服务质量路由优化方案基于大量对应用场景的假设为每种 QoS 需求单独设计复杂的优化模型,数据驱动的智能路由算法能够根据 QoS 需求自动学习得到恰当的路由决策.

4) 经验驱动与记忆特性.相比于传统基于模型和规则的路由算法,基于机器学习的智能路由算法能够通过学习历史数据来把过往经验记忆下来,使得模型能像人类一样“吃一堑长一智”,随着经验的增长逐步提升路由优化效果.

5) 路由决策考虑过去、现在和未来.循环神经

网络结构(RNN)及其相应扩展(GRU,LSTM)能够很好地将过往历史信息进行建模,而强化学习模型则赋予了智能路由算法不仅着眼于当前路由效果,更可预测未来网络状态变化,提前避免未来可能发生的网络拥塞的能力.

5.2 智能路由算法面临的挑战

与智能路由算法的优势相对应的,智能路由方法的未来发展过程同样面临着很多挑战:

1) 网络特征信息提取.智能路由方法中,网络状态信息可能是按照拓扑结构的形式进行组织的,并且由于网络场景的动态变化,使得网络状态信息的维度可能发生改变.传统的机器学习方法对于这种类型的网络状态信息的处理上存在困难.现有智能路由算法尝试利用图神经网络模型(GNN)对网络状态信息进行建模和提取[26,30].GNN 方法对于不同拓扑结构具有良好的泛化性,然而现有 GNN 方法是否能够对于路由优化问题真实场景中动态变化的大规模拓扑结构完成建模还缺乏足够的实验支撑.

2) 算法收敛性.相比于游戏、图像识别、自然语言处理等已经广泛应用机器学习的场景,路由优化问题的输入输出维度更高,目标策略更复杂,现有的研究表明对于输入输出维度很高的复杂路由优化问题,现有机器学习方案往往难以收敛到最优解.为了解决模型难以收敛的问题,往往需要通过降低输入输出维度,将决策空间离散化,或者采用间接控制路由决策以简化策略复杂度的方式来降低模型的收敛难度,然而即使采用了这些方案,很多模型最终的收敛结果依然距离理论最优值存在很大差距.

3) 算法可扩展性.可扩展性是路由算法所需要满足的重要特性.现有基于机器学习的智能路由算法主要基于不超过 20 个节点的小拓扑进行设计和实验.更大的拓扑意味着指数增长的网络状态数以及更高的路由决策难度,如何保证智能路由算法在大拓扑上依然能取得良好的效果将是未来智能路由算法设计面临的一个挑战.此外当拓扑规模很大时,集中式的路由控制算法可能带来很高的数据交换成本以及网络状态传输延时,影响可扩展性;而分布式的智能路由算法如何在大拓扑下保证各节点路由策略的一致性将是未来需要解决的问题.

4) 算法可解释性.智能路由方法所面临的另一个问题是路由策略的不可预测性以及不可解释性,相比于传统路由基于数学模型的传统路由算法,基于深度学习的方法其行为往往具有不可预测性,当出现一个糟糕的路由决策时,操作员很难去定位错误原因,至于针对错误去更正模型更是一件几乎不可能的事情.因此,如何提升智能路由算法的可解释性将是未来智能路由方法发展过程中面临的一个挑战.

5) 模型训练成本.对于基于监督学习的智能路由算法而言,收集足够多、足够准确的带标签数据有时是一个成本很高昂的事情.不同于人脸识别等一次训练一劳永逸的应用场景,随着网络环境的变化,现有智能路由可能需要重复收集训练数据并重新进行训练.因此如何提升智能路由训练过程的数据效率是智能路由方案部署过程中所面临的重要挑战.面对类似的问题时,通过元学习来降低训练成本是一个可行的解决方案[54],然而路由领域对这方面尚未有很完善的研究.此外对于基于深度强化学习的智能路由方法,无论是在线训练还是离线训练,其高昂的训练成本以及训练过程中对于系统所带来的可靠性隐患都是亟待解决的挑战.

6) 网络突发情况处理.对于智能路由方法来说,如何处理网络突发状况是另一个智能路由算法未来发展过程将面临的挑战.流量突发、网络设备故障带来的网络状态变化是现实中非常常见的情况,然而这些突发情况种类多样,很多突发情况在训练数据中从未出现过,现有数据驱动的智能路由算法很难保证当面对这些突发情况时能够处理得当.实际上,即使是 Q-Learning 这类能够动态适应环境变化的方法也无法很好地应对网络突发且剧烈的波动,利用"安全在线强化学习"[53]的思想来应对网络突发状况变化也许是未来一个可能的解决方案,但如何精确感知网络突发状况同样是一个挑战.

7) 真实场景部署.对于智能路由方法来说,如何在真实场景部署是一个巨大的挑战.相比于传统路由算法来说,智能路由需要更多的计算资源、更高的路由性能,与此同时训练数据收集与路由感知过程需要对于原有的路由协议重新设计以使得智能路由算法所需要的数据能够被智能单元所获得.SDN 网络以及可编程路由设备的提出使得路由器控制层的计算能力变得更强,然而即便如此智能路由算法也很难在现有网络体系结构下进行大规模部署.在优化智能路由算法性能并增强其对传统路由算法兼容性以及可扩展性的同时,设计与智能路由方案相匹配的路由设备也许会是未来智能路由算法发展的趋势.

6 总 结

本文经过调研发现,现有智能路由算法主要分为基于监督学习与基于强化学习2类:1)基于监督学习的智能路由方法主要通过用深度学习模型替代现有路由算法或辅助传统路由算法完成路由求解.深度学习方法使得智能路由算法对环境感知更敏感、收敛速度更快,数据驱动的辅助模块也能够使得传统路由算法所做出的路由决策更准确,并在拥塞发生之前提前避免.2)基于强化学习的路由算法能够自适应不同的路由应用场景,并优化多种网络性能指标.其中基于模型的Q-Learning方法被广泛用于无线传感器网络的路由优化过程,而深度强化学习方法则被应用于域内流量工程、基于流量预测的智能路由算法等多样化的复杂路由优化问题.

本文分析了在线与离线的智能路由训练方案以及集中式和分布式2种智能路由部署方案的优缺点,并进一步提出了离线集中式训练加在线部署的闭环学习框架以及自适应在线训练与安全学习相结合的有可靠性保证的智能路由部署框架.这2种框架为基于机器学习的智能路由算法在真实场景下低成本、高可靠性地部署提供了可能.

本文讨论了智能路由算法在未来发展过程中的机遇与挑战,并针对这些挑战提出了基于机器学习的智能路由算法未来可能的研究方向.

参 考 文 献

[1] CAIDA. The cooperative associate for Internet data analysis (CAIDA) [EB/OL]. [2019-11-01]. http://www.caida.org/data

[2] Ghorbani S, Yang Zibin, Godfrey P, et al. DRILL: Micro load balancing for low-latency data center networks [C] //Proc of the Conf of the ACM Special Interest Group on Data Communication. New York: ACM, 2017: 225-238

[3] Cui Yong, Wu Jianping, Xu Ke, et al. A survey on quality-of-service routing algorithms for the Internet [J]. Journal of Software, 2002, 13(11): 2065-2075 (in Chinese)
(崔勇, 吴建平, 徐恪, 等. 互联网络服务质量路由算法研究综述[J]. 软件学报, 2002, 13(11): 2065-2075)

[4] Chen Shigang, Nahrstedt K. Distributed quality-of-service routing in high-speed networks based on selective probing [C] //Proc of the 23rd Annual Conf on Local Computer Networks. Piscataway, NJ: IEEE, 1998: 80-89

[5] Korkmaz T, Krunz M. Multi-constrained optimal path selection [C] //Proc of IEEE INFOCOM 2001 Conf on Computer Communications, the 20th Annual Joint Conf of the IEEE Computer and Communications Society. Piscataway, NJ: IEEE, 2001: 834-843

[6] Ma Qingming, Steenkiste P. Supporting dynamic inter-class resource sharing: A multi-class QoS routing algorithm [C] //Proc of the 18th Annual Joint Conf of the IEEE Computer and Communications Societies. Piscataway, NJ: IEEE, 1999: 649-660

[7] Kumar P, Yang Yuan, Yu C, et al. Semi-oblivious traffic engineering: The road not taken [C] //Proc of the 15th Symp on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2018: 157-170

[8] Xu Jinghang, Zuo Wanli, Liang Shining, et al. Causal relation extraction based on graph attention networks [J]. Journal of Computer Research and Development, 2020, 57(1): 159-174 (in Chinese)
(许晶航, 左万利, 梁世宁, 等. 基于图注意力网络的因果关系抽取[J]. 计算机研究与发展, 2020, 57(1): 159-174)

[9] He Kaiming, Zhang Xiangyu, Ren Shaoping, et al. Deep residual learning for image recognition [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[10] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge [J]. Nature, 2017, 550(7676): 354-359

[11] Kreutz D, Ramos F, Verissimo P, et al. Software-defined networking: A comprehensive survey [J]. arXiv preprint arXiv: 1406.0440, 2014

[12] Bosshart P, Gibb G, Kim H S, et al. Forwarding metamorphosis: Fast programmable match-action processing in hardware for SDN [J]. ACM SIGCOMM Computer Communication Review, 2013, 43(4): 99-110

[13] Bosshart P, Daly D, Gibb G, et al. P4: Programming protocol-independent packet processors [J]. ACM SIGCOMM Computer Communication Review, 2014, 44(3): 87-95

[14] Xu Zhiyuan, Tang Jian, Meng Jingsong, et al. Experience-driven networking: A deep reinforcement learning based approach [C] //Proc of 2018 IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2018: 1871-1879

[15] Mao Bomin, Fadlullah Z M, Tang Fengxiao, et al. Routing or computing? The paradigm shift towards intelligent computer network packet transmission based on deep learning [J]. IEEE Transactions on Computers, 2017, 66(11): 1946-1960

[16] Boyan J A, Littman M L. Packet routing in dynamically changing networks: A reinforcement learning approach [C] //Proc of Advances in Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 1994: 671-678

[17] Choi S P M, Yeung D Y. Predictive Q-routing: A memory-based reinforcement learning approach to adaptive traffic control [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1996: 945-951

[18] Kumar S, Miikkulainen R. Dual reinforcement Q-routing: An on-line adaptive routing algorithm [C/OL] //Proc of the Artificial Neural Networks in Engineering Conf. 1997: 231-238. [2019-11-01]. http://www.cs.utexas.edu/~ai-lab/pubs/kumar.drqrouting.pdf

[19] Hu Tiansi, Fei Yunsi. QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks [J]. IEEE Transactions on Mobile Computing, 2010, 9(6): 796-809

[20] Basagni S, Di Valerio V, Gjanci P, et al. Finding MARLIN: Exploiting multi-modal communications for reliable and low-latency underwater networking [C] //Proc of 2017 IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2017: 1-9

[21] Basagni S, Di Valerio V, Gjanci P, et al. MARLIN-Q: Multi-modal communications for reliable and low-latency underwater data delivery [J]. Ad Hoc Networks, 2019, 82: 134-145

[22] Jay N, Rotman N, Godfrey B, et al. A deep reinforcement learning perspective on Internet congestion control [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2019: 3050-3059

[23] Sirinam P, Imani M, Juarez M, et al. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 1928-1943

[24] Mao Hongzi, Netravali R, Alizadeh M. Neural adaptive video streaming with pensieve [C] //Proc of the Conf of the ACM Special Interest Group on Data Communication. New York: ACM, 2017: 197-210

[25] Fadlullah Z M, Tang Fengxiao, Mao Bomin, et al. State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems [J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2432-2455

[26] Geyer F, Carle G. Learning and generating distributed routing protocols using graph-based deep learning [C] //Proc of the 2018 Workshop on Big Data Analytics and Machine Learning for Data Communication Networks. New York: ACM, 2018: 40-45

[27] Russell S J, Norvig P. Artificial Intelligence: A Modern Approach [M]. Kuala Lumpur, Malaysia: Pearson Education Limited, 2016

[28] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444

[29] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554

[30] Rusek K, Suárez-Varela J, Mestres A, et al. Unveiling the potential of graph neural networks for network modeling and optimization in SDN [C] //Proc of the 2019 ACM Symp on SDN Research. New York: ACM, 2019: 140-151

[31] Valadarsky A, Schapira M, Shahaf D, et al. Learning to route [C] //Proc of the 16th ACM Workshop on Hot Topics in Networks. New York: ACM, 2017: 185-191

[32] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780

[33] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv: 1412.3555, 2014

[34] Zhou Jie, Cui Ganqu, Zhang Zhengyan, et al. Graph neural networks: A review of methods and applications [J]. arXiv preprint arXiv: 1812.08434, 2018

[35] Mao Hongzi, Schwarzkopf M, Venkatakrishnan S B, et al. Learning scheduling algorithms for data processing clusters [C] //Proc of the ACM Special Interest Group on Data Communication. New York: ACM, 2019: 270-288

[36] Zhuang Zirui, Wang Jingyu, Qi Qi, et al. Graph-aware deep learning based intelligent routing strategy [C] //Proc of 2018 IEEE 43rd Conf on Local Computer Networks. Piscataway, NJ: IEEE, 2018: 441-444

[37] Nace D, Pióro M. Max-min fairness and its applications to routing and load-balancing in communication networks: A tutorial [J]. IEEE Communications Surveys & Tutorials, 2008, 10(4): 5-17

[38] Xu Qian, Zhang Yifan, Wu Kui, et al. Evaluating and boosting reinforcement learning for intra-domain routing [C] //Proc of the 16th IEEE Int Conf on Mobile Ad Hoc and Sensor Systems. Piscataway, NJ: IEEE, 2019

[39] Kato N, Fadlullah Z M, Mao Bomin, et al. The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective [J]. IEEE Wireless Communications, 2016, 24(3): 146-153

[40] Barabas M, Boanea G, Dobrota V. Multipath routing management using neural networks-based traffic prediction [C/OL] //Proc of the 3rd Int Conf on Emerging Network Intelligence. 2011: 118-124. [2019-11-01]. https://www.thinkmind.org/download.php?articleid=emerging_2011_6_30_40129

[41] Hua Yuxiu, Zhao Zhifeng, Liu Zhiming, et al. Traffic prediction based on random connectivity in deep learning with long short-term memory [C] //Proc of the 88th Vehicular Technology Conf. Piscataway, NJ: IEEE, 2018: 1-6

[42] Nie Laisen, Jiang Dingde, Guo Lei, et al. Traffic matrix prediction and estimation based on deep learning in large-scale IP backbone networks [J]. Journal of Network and Computer Applications, 2016, 76: 16-22

[43] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning [J]. arXiv preprint arXiv: 1312.5602, 2013

[44] Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 1057-1063

[45] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms [C] //Proc of the 31st Int Conf on Machine Learning. New York: ACM, 2014: 387-395

[46] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning [J]. arXiv preprint arXiv: 1509.02971, 2015

[47] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 1889-1897

[48] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms [J]. arXiv preprint arXiv: 1707.06347, 2017

[49] Basagni S, Petrioli C, Petroccia R, et al. CARP: A channel-aware routing protocol for underwater acoustic wireless networks [J]. Ad Hoc Networks, 2015, 34: 92-104

[50] Azar Y, Cohen E, Fiat A, et al. Optimal oblivious routing in polynomial time [J]. Journal of Computer and System Sciences, 2004, 69(3): 383-394

[51] Lowe R, Wu Yi, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 6379-6390

[52] Garcıa J, Fernández F. A comprehensive survey on safe reinforcement learning [J]. Journal of Machine Learning Research, 2015, 16(1): 1437-1480

[53] Mao Hongzi, Schwarzkopf M, He Hao, et al. Towards safe online reinforcement learning in computer systems [C/OL] // Proc of NeurIPS Machine Learning for Systems Workshop. 2019. [2020-02-16]. http://mlforsystems.org/assets/papers/neurips2019/towards_mao_2019.pdf

[54] Mao Hongzi, Venkatakrishnan S B, Schwarzkopf M, et al. Variance reduction for reinforcement learning in input-driven environments [J]. arXiv preprint arXiv: 1807.02264, 2018

Liu Chenyi, born in 1996. PhD candidate at the Department of Computer Science and Technology of Tsinghua University. Received his B. Eng degree in computer science and technology from Tsinghua University in 2019. His main research interests include traffic engineering and machine learning.

Xu Mingwei, born in 1971. Full professor at the Department of Computer Science and Technology of Tsinghua University. Received his BSc and PhD degrees from Tsinghua University. Senior member of CCF. His main research interests include computer network architecture, Internet routing, network protocol design and network security.

Geng Nan, born in 1993. PhD candidate at the Department of Computer Science and Technology of Tsinghua University. Received his B.Eng degree in communication engineering from Beijing University of Posts and Telecommunications in 2016. His main research interests include traffic engineering, destination and source IP routing, and machine learning.

Zhang Xiang, born in 1995. PhD candidate at the Department of Computer Science and Technology of Tsinghua University. Student member of CCF. Received his B. Eng degree in information security from Tongji University in 2018. His main research interests include satellite network and machine learning.