一种基于深度强化学习的动态路由算法

肖扬　吴家威　李鉴学　刘军

(北京邮电大学人工智能学院,北京 100876)

摘要: 路由是网络基础架构稳定运行的保障,是支撑下一代网络持续发展的关键功能。如今,网络流量的快速增长和服务需求的不断变化使传统路由算法面临严峻的挑战。近年来,深度强化学习在解决复杂连续控制问题上表现出良好的效果。为了解决传统路由算法的一系列弊端,将深度确定性策略梯度(Deep Deterministic Policy Gradient,DDPG)与路由场景相结合,提出一种基于深度强化学习的新型动态路由算法(DDPG4Net);随后,在自行开发的网络模拟器 RL4Net 上对该算法的效果进行了验证。

关键词: 深度强化学习; 路由算法; 网络流量工程

1　引言

路由是为网络中的数据包选择传输路径的过程,是通信网的核心功能之一。路由优化的标准取决于网络服务提供商或网络管理人员,通常根据网络质量指标(如时延、丢包率、链路利用率、QoS 等)确定。路由算法通常分为静态路由算法和动态路由算法两类。静态路由算法由网络管理员根据经验手工配置并储存在路由表内,在实际运行阶段,每一个到达的数据包根据路由表检索需要转发到的下一跳。静态路由算法较简单且易于设置,在其用于小规模网络时具有良好的表现。但静态路由算法在没有人工干预的条件下无法根据网络变化做出相应改变,因此不适用于大型或易变的网络。从 20 世纪 90 年代起,动态路由算法成为了网络路由算法的主流。动态路由算法可以通过分析监测到的网络状态信息,即时地或周期性地对路由策略进行相应的调整。动态路由算法在实际中应用广泛,较为常见的有基于距离矢量的路由算法 RIP、IGRP,基于链路状态的路由算法 OSPF、IS-IS 和结合距离矢量和链路状态的路由算法 EIGRP。尽管这些算法在大量环境下被部署且使用成熟,但随着网络流量的快速增长和服务需求的不断变化,对不同特征的网络数据流进行区分化的路由策略使传统路由算法面临新的难题。虽然有的学者根据分析优化或基于局部搜索提出了一些启发式路由算法,但仍因过于依赖人类经验,其收敛性、鲁棒性和泛化性难以得到保证。近年来,机器学习技术的引入给路由算法设计带来了更多的可能性。机器学习技术强大的数据表征能力适用于日益复杂化的网络拓扑结构和动态化的网络流量特性,不仅如此,机器学习技术还可以从更深层次挖掘路由选路与反馈的网络指标之间的关系,预测路由决策所带来的全局影响。在所有的机器学习分支中,强化学习在路由领域最具发展前景。

强化学习的本质在于智能体和环境的不断交互,这一交互过程常被建模为马尔可夫决策过程。其中,智能体循环不断地观测状态,选择动作并从环境反馈中获得奖励,在试错过程中学习最大化累积奖励的行动策略。然而,对于大型网络而言,虽然传统的强化学习算法能够收敛,但因为必须探索并获得整个网络系统的知识,需要耗费大量的算力和时间。因此,传统的强化学习方法在实际中的应用十分有限。深度学习作为一种新的突破性技术,在与传统强化学习结合后可以弥补传统强化学习的局限性,即形成深度强化学习。深度强化学习利用深度神经网络强大的提取高维数据特征的能力,大幅提高了传统强化学习算法的收敛速度和性能表现,在游戏对抗、自动驾驶、机器人控制等领域得到了长足的发展。随着通信技术的不断发展,路由器算力的不断提高为强化学习智能体的部署提供

了硬件支撑，SDN 等新型网络架构的提出[1]为强化学习算法的设计提供了开放平台，而深度确定性策略梯度(Deep Deterministic Policy Gradient，DDPG)[2]等一系列深度强化学习算法的完善为深度强化学习应用到路由领域提供了技术基础。根据现有的研究提出更为有效的动态路由算法，对于构建新一代智能化通信网络至关重要。

2 相关研究工作

2017 年，西班牙加泰罗尼亚大学的 G. Stampa 团队首次尝试将深度强化学习技术应用于路由优化问题[3]，提出将 DDPG 算法作为主体，状态定义为流量矩阵(包含每对源—目的地节点对的带宽需求)，动作定义为链路度量(Metric)向量，奖励定义为网络平均时延，以最大化累积奖励(即最小化网络平均时延)为优化目标持续迭代以得到最优策略。2018 年，解放军信息工程大学的 C. Yu 等提出了一种基于 SDN 的深度强化学习路由算法[4]，在奖励机制部分除了考虑网络时延等单一变量外，还拓展到吞吐量、带宽等，可用一个综合多种指标的函数作为奖励。来自北京邮电大学的 X. Huang 等则针对多媒体通信传输过程中对于体验质量(Quality of Experience，QoE)的特殊要求，基于 DDPG 设计了针对多媒体路由控制的算法[5]，其状态包含了网络带宽、时延、抖动以及丢包率，奖励则使用衡量多媒体服务质量的平均意见分数(Mean Opinion Score，MOS)，该算法相比于其他路由策略显著提升了 QoE。

许多学者也通过将深度强化学习与其他深度学习技术相结合提升路由策略的可靠性。T. A. Q. Pham 等提出将深度强化学习与卷积神经网络(Convolutional Neural Network，CNN)结合[6]，在 Actor 网络和 Critic 网络的全连接层之前增加多个卷积层，将 CNN 在计算机视觉领域出色的提取空间邻域信息特征的能力用于探索不同信息流之间的相互影响。P. Sun 等则提出网络流量具有显著的周期特性[7]，并设计将深度强化学习与循环神经网络(Recurrent Neural Network，RNN)结合，从而提取网络流量的时域相关性。以上两种算法仍以 DDPG 算法作为深度强化学习部分的实现。根据以上一系列基于深度强化学习的路由算法，可将学术界常用的解决方案总结为：状态通常代表某一时刻网络中所有节点和链路的路由状态，动作代表所有节点到其可选的下一跳节点的度量，随后通过最短路径优先(Shortest Path First，SPF)算法得到每一对源——目的地节点之间的最短路径；对于每一状态——动作对都根据下一时间步的一个或多个网络指标(包括时延、吞吐量、带宽、拥塞水平、链路利用率、重传次数等)反馈奖励。由于状态空间和动作空间都会在大规模网络中具有较高的维度，以 DDPG 为代表的一系列演员——评论家算法(Actor-Critic，AC)往往被采用。由上可知，以往的算法设计中动作空间仍参考传统动态路由算法的度量，即对于分配了度量的链路集合，得到的最优路径只有一条。然而在实际转发时，可能出现最短路径出现拥塞，而非最短路径带宽占用量低的情况——在这种情况下，如果能将部分数据包转发至其他非最短路径，既能降低最短路径的拥塞程度，又能提高网络整体的链路利用率。因此，可考虑将 SPF 输出单一最短路径替换为基于概率路由，同时将链路指标替换为链路权重(Weight)，路由器不再根据指标对源节点和目的节点相同的数据包指定唯一一条路径，而是根据所有可达路径权重的反比得到转发到任意下一跳节点的概率。根据这一思路，笔者考虑在动作空间的设计中直接对转发概率建模，设计基于深度强化学习的新型路由算法 DDPG4Net。

3 基于深度强化学习的路由算法设计

强化学习模型包含状态、动作、奖励和策略 4 种要素(见图 1)，可对 DDPG4Net 的模型要素做如下定义。

(1)状态：在强化学习模型中，状态需要反映智能体所处环境的实时特征。对于路由场景中智能体的状态，通常使用业务流的传输状况。定义所有数据包的集合为 P，对于每个经过网络传输的数据包 p_i，将从源

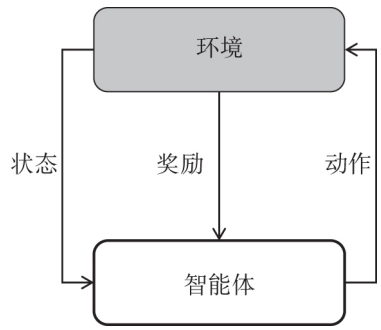

图 1 强化学习流程

节点 d_i 转发进入网络并从目的节点 d_j 转发离开网络。假设网络中节点的总数为 N,且所有节点都可被经由传入和传出网络,则所有源——目的地节点对总数为 $N \times N$。若将状态 s 定义为一维度 $N \times N$ 的流量矩阵(Traffic Matrix,TM),其中的每个元素 d_{ij} 为单位时间从源节点 d_i 传输到目的节点 d_j 的总流量,则:

$$s = \begin{bmatrix} d_{11} & \cdots & d_{1N} \\ \vdots & \ddots & \vdots \\ d_{N1} & \cdots & d_{NN} \end{bmatrix} \tag{1}$$

(2) 动作: 动作需要智能体根据当前状态和策略做出,是每一时间步智能体对网络下发的具体路由规则。与传统路由策略不同,基于概率的路由不再赋予每条边指标,而是为每条边指定权重。定义连接节点 i 和任意一个邻接节点 j 的边所对应的链路权重为 w_{ij},则对于有 m 个邻接节点的某节点 i 都有一权重向量 $W_i = <w_{i1}, w_{i2}, \cdots, w_{im}>$。若将动作 a 定义为全局链路权重的集合$<W_1, W_2, \cdots, W_N>$,用于在当前时间步下发至网络中的每个路由器,则 $a = <W_1, W_2, \cdots, W_N>$。

(3) 奖励: 根据当前时刻网络的状态和智能体做出的行为,在下一时刻智能体将收到网络的指标反馈作为奖励,奖励可根据不同的网络优化需求设置为综合不同指标的函数。若将奖励定义为整个网络的平均端到端时延,假设每个数据包的时延为 $delay_i$,则:

$$r = \sum_{i=1}^{P} delay_i / P \tag{2}$$

(4) 策略: 智能体的策略指从当前状态到对应当前动作的映射,指导着智能体在特定环境下的行为方式,是强化学习的核心。由于 DDPG 应用确定性策略,智能体根据当前状态和动作输出的是叠加随机噪声的确定性动作,而非条件概率分布,因此能够有效减少算法的计算量,提高收敛速度。

以上状态空间、动作空间和奖励的设计与网络状态紧密相关,囊括了路由场景的关键组成部分,且不包含多余的信息内容。如果对于整个网络设置一个全局智能体以运行 DRL 算法,在每个时间步能够观测网络全局状态,由最佳策略生成对应的控制动作,将新的路由规则(权重)通过网络控制器或 SDN 下发到网络,并在下一时间步收获奖励和记录训练经验。由于状态空间和动作空间都包含连续域,故可将成熟的连续域 AC 算法 DDPG 作为深度强化学习路由算法 DDPG4Net 的框架。

DDPG 是一种基于确定性策略梯度(Deterministic Policy Gradient,DPG)[8] 的离线 AC 算法,其中的 Actor 负责对策略建模,即对于智能体观察到的状态输出动作并和环境交互; Critic 负责对价值函数建模,即实时地通过价值量化评估 Actor 的表现,并通过误差更新修正 Actor 的策略。相较于其他 AC 算法,DDPG 的 Actor 不再通过随机策略输出动作的概率分布,而是选择最大概率的动作作为输出。DDPG 的创新点除了输出确定性动作而非概率分布,还使用了经验回放和目标网络两项技术。经验回放技术将智能体和环境交互过程中的状态转移储存为元组(s_t, a_t, r_t, s_{t+1})并放入经验回放池,在更新策略时从经验回放池中随机取样一批用于计算误差。经验回放有效地减少了采样数据的时间相关性,防止神经网络因使用高度时间相关的数据用于训练从而导致陷入局部最优的问题; 目标网络技术则将神经网络中计算目标值的部分解耦,使用另一结构相同的目标网络计算目标值,其网络参数则周期性从当前网络复制。因此,DDPG4Net 拥有当前 Actor 网络、目标 Actor 网络、当前 Critic 网络和目标 Critic 网络 4 个神经网络,其各自所负责的功能可总结为: 当前 Actor 网络根据当前状态选择动作并和环境交互,观察智能体的下一状态,计算奖励; 目标 Actor 网络根据从经验回放池中取样的下一状态选择下一动作; 当前 Critic 网络根据当前状态和动作计算当前 Q 值; 目标 Critic 网络根据下一状态和下一动作计算目标 Q 值。

当前 Actor 网络的参数通过反向传播均方误差 loss 来更新,$loss = \frac{1}{N}\sum_i [y_i - Q(s_i, a_i)]^2$,其中 N 为随机取样批容量,y_i 为目标 Q 值; 而当前 Critic 网络的参数则通过反向传播策略梯度来更新。训练收敛后,DDPG4Net 输出的动作将给每条链路下发对应的权重; 而智能体在数据包路径上的每一跳时,将根据权重确定转发到每一可达路径的概率,从而实现基于概率的路由。具体而言,DDPG4Net 将根据权重的反比进行归一化后得到转发概率向量。假设节点 i 具有 m 个邻接节点,且到达节点 i 的数据包都可经由这 m 个邻接节点到达目的节点,节点 i 的权重向量为 $W_i = <w_{i1}, w_{i2}, \cdots, w_{im}>$,则当某一数据包到达节点 i 时,将其转发到第 k 条边的概率为:

$$P_{ik}=\frac{\frac{1}{w_{ik}}}{\sum_{j=1}^{m}\frac{1}{w_{ij}}} \tag{3}$$

同理，如果某到达节点 i 的数据包只能经由这 m 个邻接节点中的 n 个到达其目的节点，则当此数据包到达节点 i 时，将其转发到第 q 条边（第 q 条边可达目的节点的 n 条边之中）的概率为：

$$P_{iq}=\frac{\frac{1}{w_{iq}}}{\sum_{j=1}^{n}\frac{1}{w_{ij}}} \tag{4}$$

4 试验与评估

4.1 RL4Net

学术界和产业界尚未有成熟的支持基于深度强化学习的路由算法试验框架。为了更好地开展针对 DDPG4Net 的后续试验，笔者基于常用的网络仿真软件 ns3[9] 自行开发了一套支持深度强化学习的网络仿真工具 RL4Net[10]。参照强化学习的通用框架，RL4Net 主要由智能体和环境两部分组成（见图 2）。智能体部分包含多个 DRL 算法的容器，可根据试验内容选择一个或多个 DRL 算法进行对比试验，算法实现支持 TensorFlow 和 pyTorch；环境部分主要基于 ns3 网络模拟器开发，并在其基础上拓展了如下 6 个针对路由场景的组件。

（1）指标提取组件：用于从 ns3 中提取必要数据计算时延、吞吐量、丢包率等性能指标。

（2）指标转化组件：用于转化性能指标，从而可供智能体直接用于奖励等。

（3）动作下发组件：用于在网络环境端获取智能体输出的动作。

（4）动作执行组件：根据智能体下发的动作执行对应的 ns3 实际操作。

（5）ns3Env：将 ns3 对象转化为强化学习环境。

（6）envInterface：用于 ns3 数据和强化学习参数之间的转换。

4.2 试验设计

为了验证 DDPG4Net 算法的有效性，笔者基于路由场景设计了对应的测试网络拓扑和测试流量，并在试验中将 DDPG4Net 算法的效果与基于链路状态的路由算法 OSPF 和随机概率路由进行了对比。本试验使用了 6 节点网络拓扑（见图 3），每个节点分别用 n0～n5 标号，所有邻接的节点都由箭头相连。设计的测试业务流量为 8 Mbit/s，从源节点 n0 发往目的节点 n5，网络中每条链路的带宽都为 5 Mbit/s。由 SPF 算法容易得出从 n0 到 n5 的最短路径为 n0→n3→n5，然而业

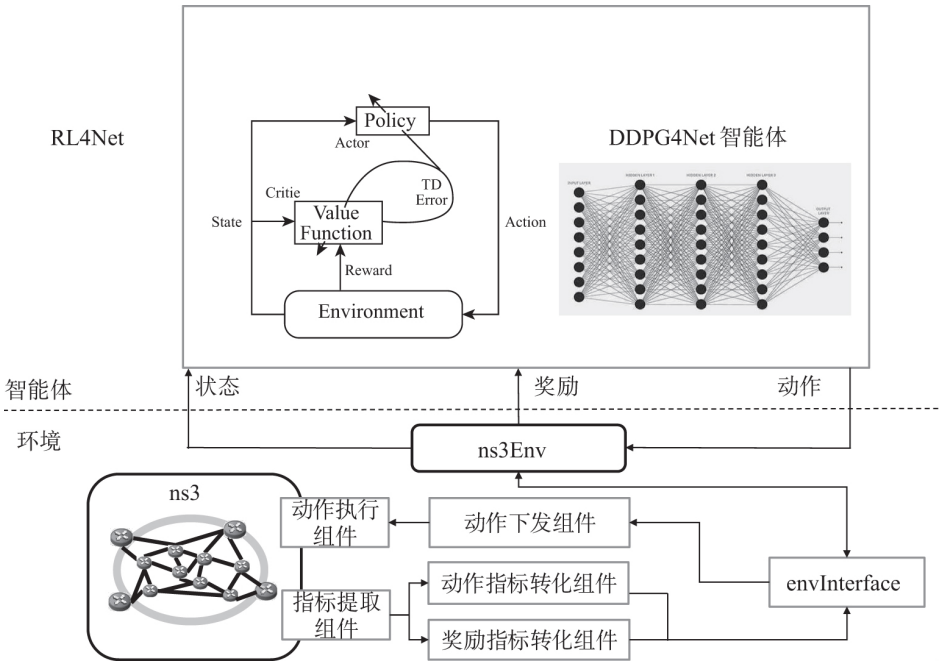

图 2 RL4Net 框架

务流量超过了最短路径的带宽,因此必然发生网络拥塞。可想而知,如果网络能够智能地将部分流量经由 n0→n1→n3→n5 和 n0→n2→n4→n5 两条路径传输,则一方面可以缓解最短路径的拥塞(n0→n2→n4→n5,而分往 n0→n1→n3→n5 仍将导致 n3→n5 发生拥塞);另一方面也可以增加网络整体的链路利用率。笔者设计了两组对比试验,分别将 DDPG4Net 算法和 OSPF、随机概率路由进行对比。OSPF 算法不基于概率路由,而包含人为预设的经验 SPF; 随机路由算法基于概率路由,但不经过算法根据优化目标的迭代学习。通过两组对比试验,可以分别验证概率路由在部分场景中相对于传统路由算法的优越性,以及深度强化学习技术对于基于概率路由算法的性能提升。

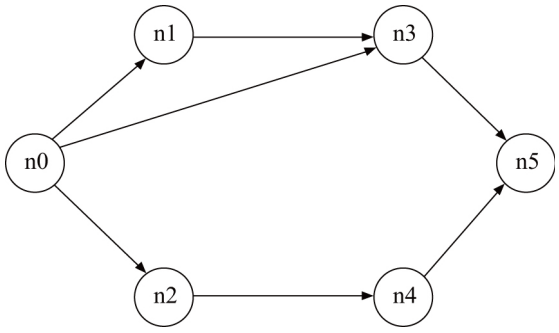

图 3　网络拓扑图

4.3　试验结果评估

图 4 和图 5 分别为智能体在训练 75 个 Episode 之后 Actor 和 Critic 的损失函数曲线。从两图中可以看出,Actor 和 Critic 的损失函数分别在约 50 个 Episode 和 65 个 Episode 之后达到收敛,即 Actor 可以在仿真网络中输出对应价值函数值较高的动作; Critic 能够对动作价值函数做出与实际奖励相对应的合理估计。图 6 所示为在上一节所述网络拓扑和业务流需求的条件下,DDPG4Net、OSPF 和随机路由三者的网络平均时延对比情况。如图 6 所示,OSPF 算法(上部曲线) 的平均时延最高,而随机概率路由(中间曲线) 相对于 OSPF 的平均时延有了大幅降低,这与试验的预期相符合: 虽然随机概率路由没有任何针对路径最优化的转发策略,但是在本试验的特定条件下,随机策略也会将部分经过 n3→n5 的数据包转移到 n0→n2→n4→n5 路径,而 n0→n2→n4→n5 路径原本没有占用任何带宽,因此出现了相对于 OSPF 网络平均时延显著下降的情况。DDPG4Net(下部曲线) 则相对于 OSPF 算法和随机概率路由都有了更为显著的性能提升,不仅平均网络时延明显下降,且收敛性较好。DDPG4Net 从约 60 个 Episode 之后时延收敛到极小值,这也与图 4、图 5 中 Actor 和 Critic 共同收敛的阶段相近。由于 DDPG4Net 能够以平均时延为优化目标,在试错过程中不断调整路由器转发到不同下一跳的概率,而从源节点 n0 到目的节点 n5 的总带宽高于业务流量 8 Mbit/s,所以能够完全避免网络的拥塞。试验结果证明,DDPG4Net 在较短的时间内达到了符合理论预期的良好路由控制。

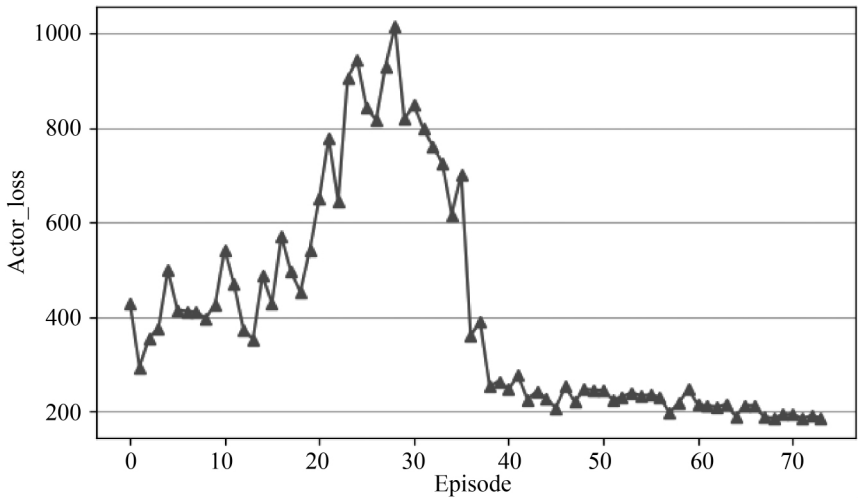

图 4　Actor Loss 曲线

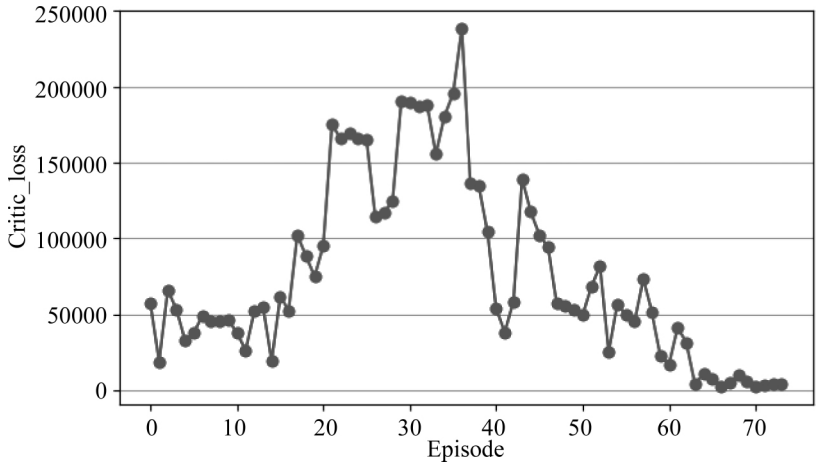

图 5 Critic Loss 曲线

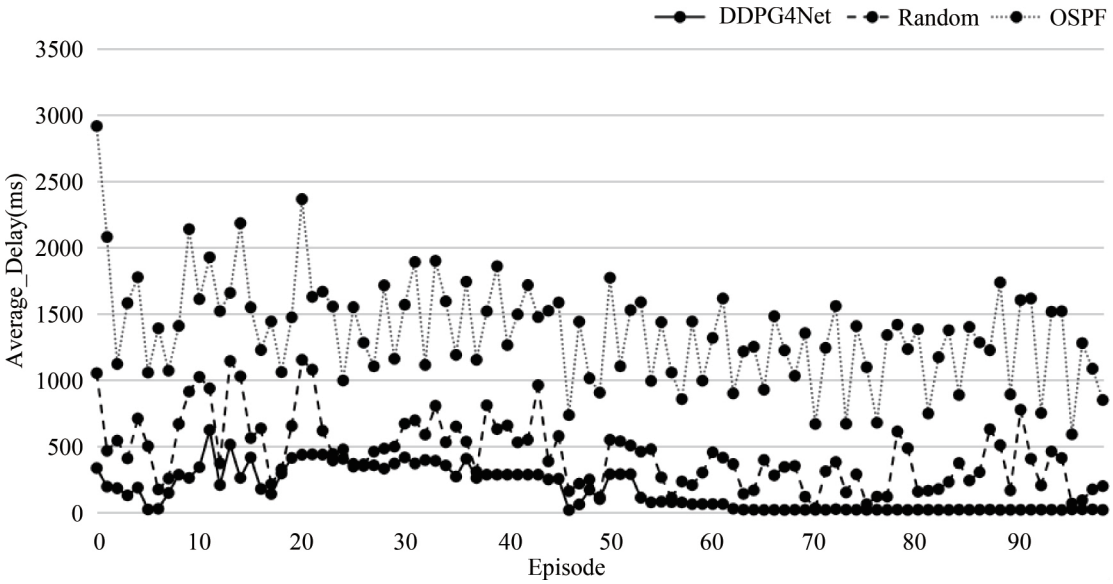

图 6 网络平均时延对比

5 结束语

路由是通信网的核心功能，然而随着网络流量的剧增和网络拓扑的复杂化，传统路由策略在一些路由场景下不再适用。随着路由器等网络硬件设备的性能提升和深度强化学习技术的不断发展，基于深度强化学习技术设计智能的动态路由算法成为了可能。笔者基于当前最先进的深度强化学习 AC 算法之 DDPG 提出了一种基于概率的动态路由算法 DDPG4Net，并在自行开发的网络仿真工具 RL4Net 上对算法进行了性能测试。试验表明，DDPG4Net 算法能够有效地减少特定场景下的网络拥塞程度，降低网络时延。在测试网络中，DDPG4Net 相对于传统路由算法 OSPF 和随机概率路由具有显著的性能提升。

参考文献

[1] N. McKeown, T. Anderson, H. Balakrishnan, et al. OpenFlow: enabling innovation in campus networks [J]. Acm Sigcomm Computer Communication, 2008, 38(2): 69-74.

[2] T. P. Lillicrap, J. J. Hunt, A. Pritzel, et al. Continuous control with deep reinforcement learning [R]. The 4th International Conference on Learning Representations, 2016.

[3] G. Stampa, M. Arias, D. Sanchez-Charles, et al. A deep-reinforcement learning approach for software-defined networking routing optimization [J]. arXiv: 1709.07080, 2017.

[4] C. Yu, J. Lan, Z. Guo, et al. DROM: optimizing the routing in software-defined networks with deep reinforcement learning [J]. IEEE Access, 2018, 6: 64533-64539.

[5] X. Huang, T. Yuan, G. Qiao, et al. Deep reinforcement learning for multimedia traffic control in software defined networking [J]. IEEE Network, 2018, 32(6): 35-41.

[6] T. A. Q. Pham, Y. Hadjadj-Aoul, A. Outtagarts. Deep reinforcement learning based QoS-aware routing in knowledge-defined networking [J] Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, 2019, 272: 14-26.

[7] P. Sun, J. Li, J. Lan, et al. RNN deep reinforcement learning for routing optimization [R]. 2018 IEEE 4th Int. Conf, 2018.

[8] D. Silver, G. Lever, N. Heess, et al. Deterministic policy gradient algorithms [R]. The 31st International Conference on Machine Learning, 2014.

[9] Klaus Wehrle, Mesut Gnes, James Gross. Modeling and tools for network simulation [M]. Springer Publishing Company, Incorporated, 2010.

[10] GitHub. RL4Net. [2020-08-05]. https://github.com/bupt-ipcr/RL4Net.

作者简介:

肖扬 北京邮电大学人工智能学院智能感知与计算教研中心博士研究生,主要研究基于强化学习的自主网络

吴家威 北京邮电大学人工智能学院智能感知与计算教研中心硕士研究生,主要研究基于强化学习的网络路由算法

李鉴学 北京邮电大学人工智能学院智能感知与计算教研中心硕士研究生,主要研究基于强化学习的网络路由算法

刘军 北京邮电大学人工智能学院智能感知与计算教研中心副教授,博士生导师,北京邮电大学数据科学中心主任,北京大数据协会常务理事,主要研究基于强化学习的自主路由

A dynamic routing algorithm based on deep reinforcement learning

XIAO Yang, WU Jiawei, LI Jianxue, LIU Jun

(School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Routing maintains the stable operation of network infrastructure and supports the sustainable development of next-generation networks. Nowadays, the rapid growth of network traffic and the continuous changes in network services make traditional routing algorithms face severe challenges. In recent years, deep reinforcement learning has shown good results in solving complex continuous control problems. In order to solve a series of shortcomings of traditional routing algorithms, we combined the Deep Deterministic Policy Gradient (DDPG) algorithm with routing scenarios, and proposed a new dynamic routing algorithm based on deep reinforcement learning—DDPG4Net.

Key words: deep reinforcement learning; routing algorithm; network traffic engineering

(收稿日期: 2020-08-05)