网络智能传输研究进展

Network Intelligent Transmission Technology: A Survey

廖乙鑫/LIAO Yixin[1]，王子逸/WANG Ziyi[1]，崔勇/CUI Yong[2]
（1. 北京邮电大学，中国 北京 100876；
2. 清华大学 中国 北京 100084）
(1. Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Tsinghua University, Beijing 100084, China)

DOI：10.12142/ZTETJ.202305010
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20231016.0846.002.html
网络出版日期：2023-10-17
收稿日期：2023-08-10

摘要：新兴超低时延场景的出现以及6G技术与人工智能技术的发展，促使网络智能传输成为研究热点。分析了传输层和应用层的时延组成及影响因素，对机器学习技术与传输层、应用层流媒体传输相结合的智能传输协议的发展和优缺点进行了综述。从传统网络传输协议的发展、人工智能技术的发展、网络传输和人工智能结合3个方面展望了网络智能传输面临的机遇与挑战。认为分布式机器学习训练场景的传输性能、训练数据的质量、模型的泛化能力、模型大规模部署的开销是未来网络智能传输技术的重点研究方向。

关键词：网络传输；机器学习；时延分析；拥塞控制；视频传输

Abstract: With the emergence of emerging ultra-low latency scenarios and the development of 6G technology and artificial intelligence technology, intelligent network transmission has become a research hotspot. The delay components and influencing factors of the transport layer and application layer are discussed. Then, the development, advantages, and disadvantages of intelligent transmission protocols that combine machine learning technology with transport layer and application layer streaming media transmission are reviewed. The opportunities and challenges faced by intelligent network transmission are prospected from three aspects: the development of traditional network transmission protocols, the development of artificial intelligence technology, and the combination of network transmission and artificial intelligence. It is believed that the transmission performance of distributed machine learning training, the quality of training data, the generalization ability of the model, and the cost of large-scale deployment of the model is the key research direction of future network intelligent transmission technology.

Keywords: network transmission; machine learning; analysis of delay; congestion control; video transmission

引用格式：廖乙鑫, 王子逸, 崔勇. 网络智能传输研究进展 [J]. 中兴通讯技术, 2023, 29(5): 61-67. DOI: 10.12142/ZTETJ.202305010
Citation: LIAO Y X, WANG Z Y, CUI Y. Network intelligent transmission technology: a survey [J]. ZTE technology journal, 2023, 29(5): 61-67. DOI: 10.12142/ZTETJ.202305010

随着6G技术与人工智能技术的迅速发展，我们正面临着众多机遇和挑战。6G技术提供了低时延的物理传输通道[1]，为扩展现实（XR）、远程医疗、元宇宙、车联网等新兴应用场景[2-4]开拓了广阔的发展空间。然而，新兴应用场景的发展还需要上层传输层、应用层的架构完善来支撑。新兴应用与传统应用在时延要求[5]、承载网络结构等方面存在明显差异，这对传统网络传输协议提出新的挑战。传统的传输层协议和应用层协议设计通常基于静态的规则和网络假设，难以适应现代网络中的复杂变化和多样化应用需求。借助人工智能技术[6]，传输协议可以通过学习来适应实时网络环境，调整拥塞控制和传输优化策略。比如，机器学习[7]算法能够从大量的复杂网络数据中学习并提取有用的模式和特征，动态生成网络传输协议规则。机器学习技术还具备适应的能力，能根据网络环境的变化调整模型参数和优化算法。将人工智能技术应用于传输协议的设计，可以更好地应对网络中的动态变化和多样化应用需求，提供更高效、可靠和稳定的网络传输。网络智能传输技术应运而生。

在网络传输中，时延是一个非常重要的指标，对新兴应用场景而言尤为关键。为了更准确地为网络智能传输技术设定学习的优化目标[8]，提供高质量的传输体验，我们有必要分析传输层协议和应用层协议的时延形成机制和影响因素[9]。对于传输层而言，主要的底层协议为传输控制协议（TCP）和用户数据报协议（UDP）。TCP提供面向连接的传

基金项目：自然科学基金重点项目（62132009）；创新研究群体项目（62221003）

输服务，因而会引入连接管理、可靠数据传输、拥塞控制相关的时延；UDP提供尽力而为的传输服务，因而只引入数据传输的时延。对于应用层而言，时延的组成和影响因素由应用类型决定。我们以车联网[10]和低时延直播场景[11]为例，分析两种场景的时延组成和影响因素。为了设计更符合场景需求的传输协议，减少传输时延，需要根据具体协议的时延影响因素，设置准确恰当的智能优化目标。

学界对机器学习算法应用于传输层协议和网络层协议进行了长期探索，并取得了显著成果。对于传输层，拥塞控制算法构成了传输层协议的核心组成部分，其优秀性能体现在合理调节数据端点的发送能力。利用机器学习技术理解复杂网络的流量等特性，并针对性地设计拥塞控制算法，提升其鲁棒性、泛化性、训练效率和推理效率，是传输层研究的热点。对于应用层，由于视频流量占据互联网的大部分流量，如何合理地利用机器学习技术设计高效率的视频码率自适应（ABR）算法[12]，提升视频质量和用户视频体验质量（QoE）[13]，并能恰当地在现实中进行部署，是学界的重要研究方向。

网络智能传输技术的研究取得了显著成果，但依然面临一些挑战。快速UDP网络连接（QUIC）[14–16]、截止日期感知传输协议（DTP）[17]等新兴协议的提出，Media Over QUIC[18][19]工作组的成立，进一步推动了网络传输协议研究的发展。同时，AI大模型的发展，为设计和训练更精确的网络大模型带来新的可能性。然而，很多问题的解决方案依旧亟待探索，例如：构建与现实世界分布一致的网络训练数据集，充分利用其中的带宽、时延等指标训练优秀的网络传输模型；设计网络传输协议，解决分布式机器学习的节点高效通信问题，在部署机器学习模型时，适应不同上层应用对网络传输指标的不同需求，并在网络动态变化时保持良好的泛化性。

本文对广域网点对点网络智能传输研究进展进行了综述：首先对传输层、应用层的时延成因进行详细分析；其次介绍了传输层、应用层的智能传输研究现状，包括人工智能与拥塞控制算法相结合、人工智能与流媒体传输相结合的成果进展；最后对网络传输新发展、人工智能技术的新发展、网络传输和人工智能结合这3个方面进行展望，并预测了未来的研究发展趋势和可能存在的挑战。

1 时延的形成机制

为了提高网络协议的传输质量，最大程度减少时延，有必要先讨论时延的种类和影响因素。网络层的时延主要包括分组传输时延和分组处理时延，受网络拥塞程度、路由选择、路由器缓存和处理能力影响[20]。数据链路层的时延主要包括帧传输时延和帧处理时延，受链路带宽、帧长度以及链路错误率等影响[21]。本文重点介绍传输层和应用层的时延影响因素。

1.1 UDP时延分析

UDP是无连接协议，它的时延由传输时延、传播时延、排队时延、处理时延组成。由于UDP本身的无连接性和不可靠性，它不需要处理TCP中的握手时延、挥手时延、重传时延，也不进行拥塞控制[22]。总的来说，UDP的时延通常少于TCP。

1.2 TCP时延分析

在UDP协议的基础上，TCP能够确保可靠有序的端到端传输。然而，TCP的使用会引入多种类型的时延。TCP连接的生命周期包括三次握手、数据传输和四次挥手过程。这些过程会涉及连接的建立和断开、数据的传输和传播、数据在中间端点的排队和处理，以及协议本身可能引发的队头阻塞、重传、拥塞控制等。以上过程均涉及对端通信过程中的时延开销。

传输时延是指数据包发送到链路中的时延，取决于数据包的大小以及网络的带宽。传播时延依赖于两端的物理距离和传输介质。排队时延表示数据包在网络设备的缓冲区中排队时的时延。在数据包到达目标地后，网络设备需要将数据包从接收缓冲区中取出并进行处理，这将引入处理时延。对于新建立的TCP连接，拥塞控制算法会控制数据包注入网络的速度[23]，因而会引入拥塞控制时延。这会受到拥塞控制算法的预设规则的影响（如慢启动参数）。此外，TCP保证数据有序交付的特性会引入队头阻塞时延，该时延可能由数据包分组丢失或数据包无序到达引起。当分组丢失或超时到达时，重传的数据包同样也会引入重传时延[24]。

1.3 应用层时延分析

应用层的时延影响因素由应用类型决定。以车联网场景为例，对于道路自动驾驶而言，车载计算硬件需要在毫秒级别内识别道路中的行人和障碍。在这个过程中，存在传感器数据的传输时延以及计算硬件的计算和决策时延。汽车需要实时与红绿灯等交通设施通信，并与云交换驾驶信息，以优化全局交通流量，这涉及云的决策和处理时延[25]。在交互式视频直播场景里，主播与观众之间的端到端时延可以划分为4个部分，分别是主播端视频内容上传到服务器的时延、将视频从服务器下载到客户端的时延、客户端的播放缓冲时延

和这个过程中涉及的编解码时延[26]。

2 传输层研究现状

拥塞控制算法是传输层协议的核心。良好的拥塞控制算法应该合理地控制端点发送数据的能力。拥塞控制方案的性能会受到许多因素的影响，包括流量模式、链路故障、动态延迟、数据包丢失等。针对复杂网络环境很难设计最优甚至接近最优的预定义静态规则的策略。通过人工智能技术，计算机可以生成更灵活、表现更优良的拥塞控制策略。拥塞控制算法的研究时至今日一直在进行中。本小节将介绍设计传输层协议的网络智能传输研究成果。

随着机器学习技术的发展，研究人员发现不同的机器学习训练策略效果有显著差别。2013年，WINSTEIN K.等提出Remy[27]，旨在生成用于各个通信端点的拥塞控制算法。Remy通过离线学习拥塞控制规则来执行优化，将预设定的网络条件、优化目标（如高吞吐量和低排队时延）作为目标函数输入，并通过离线学习来优化目标函数，最终生成一个规则表来指导拥塞控制操作。Remy离线优化后不再进行进一步学习，因此无法泛化到不同的网络场景。当前流量/网络条件偏离Remy的网络输入假设可能会导致性能不佳。SIVARAMAN A.[28]等对基于数据驱动的端到端拥塞控制算法的设计过程，进行了形式化讨论，量化了训练符合要求的模型的难易程度。此后，JAY N.采用了基于线性奖励函数的深度强化学习方法，提出了拥塞控制框架Aurora[29]，并证明了其训练深层神经网络以捕获复杂流量和网络条件模式的能力。该框架在拥塞控制性能上能够与最先进技术相媲美，甚至优于它们。2022年，HUANG T. C.针对互联网短视频上传的场景，提出了一种名为DuGu的拥塞控制框架[30]。该算法通过模仿最优求解器而不是使用手工设计的奖励函数，从而提高性能和学习效率。同时，这种方法降低了过度依赖奖励函数导致的学习策略不稳定行为的影响。

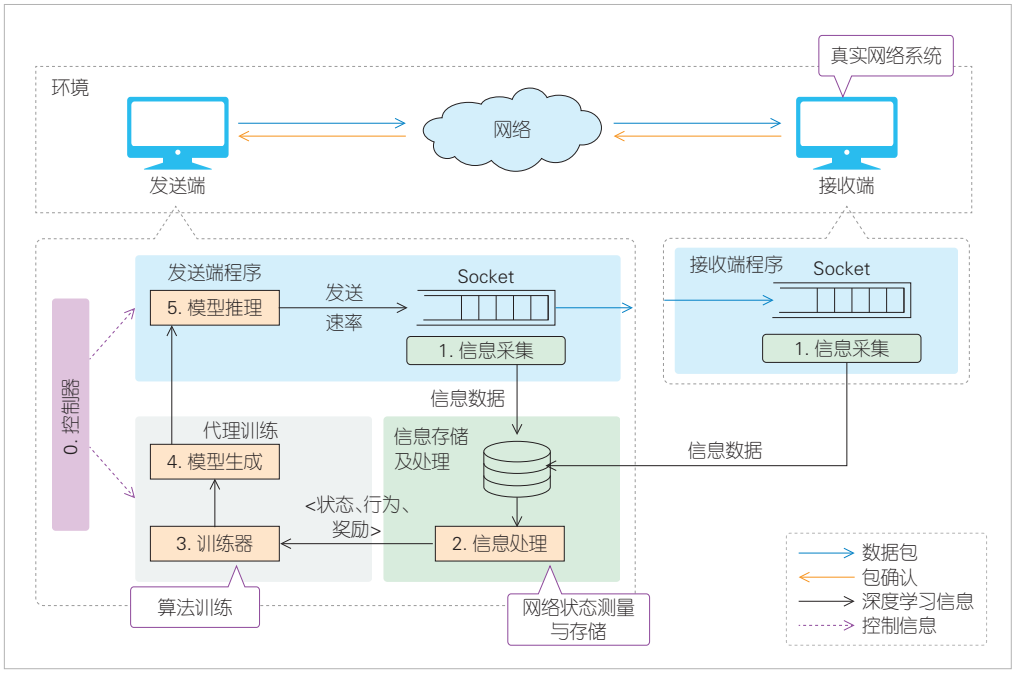
直接将AI模型与网络传输结合，可能会遇到推理效率、训练效率的问题。DONG M.提出面向性能的拥塞控制框架（PCC Vivace）[31]。Vivace将机器学习的凸优化理论应用于模型训练中，并将RTT梯度加入目标函数，来保证多个Vivace流可以收敛到一个公平和有效率的状态中，在随机丢包与收敛速度之间实现了有效均衡。此外，DONG M.使用了基于梯度的no-regret在线优化算法来调整发送速率，证明了在正确选出包含吞吐量、损耗和延迟的效用函数时，稳定的全局速率配置（纳什均衡）始终存在。ZHANG L.提出截止时间感知传输机制D3T，利用深度强化学习算法来决定发送速率和前向纠错（FEC）冗余率[32]，旨在提高用户体验质量（QoE）以满足截止时间要求。为了减小训练强化学习决策代理的难度，该研究设计了一种启发式调度算法，用于训练FEC和拥塞控制模块的强化学习代理。

由于网络流量的特征差异，训练后的机器学习策略在处理异常流量时，可能会面临严重的鲁棒性、泛化性问题。2019年，NIE X. H.[33]提出TCP-RL，根据在Web服务器端观察到的实时网络状况，为短流动态配置适合的初始窗口大小（IW），并通过演员评论家算法（A3C，一种深度学习算法）为长流动态配置合适的拥塞控制（CC）策略。将深度学习技术与传输协议的静态配置结合，可以增强协议的稳定性、鲁棒性，进一步降低智能传输协议的训练、推理成本。SOHEIL A.在2020年提出了一种名为Orcal[34]的混合拥塞控制机制，通过结合深度强化学习技术和传统拥塞控制机制，解决了应用深度学习技术可能遇到的高计算开销、低泛化性和过度收敛等问题。Orcal在洲际、洲内和蜂窝环境等复杂的网络场景中经过性能测试，并取得了良好的性能表现。因此，引入两级控制机制可以显著缓解未见场景中的性能问题。

模拟或仿真平台无法完全仿真现实世界网络流量的多变性等特性。ZHANG L.提出了拥塞控制策略学习框架ARC[35]，如图1所示。该框架在真实环境中进行异步执行，通过在发送端和接收端分别采集训练所需的网络状态，并在信息存储与处理模块中构建强化学习训练所需的数据元组，解决了训练数据的异步问题。ARC框架通过设计模型训练与真实网络系统异步运行、模型推理与数据发送异步运行，解决了模型的动态更新问题，并取得了高吞吐量和低延迟的良好性能。随后，ZHANG L.提出了一种名为Deep CC的学习训练架构[36]。Deep CC利用多目标深度强化学习算法训练深度神经网络模型。考虑到动态网络条件下应用的特定需求，Deep CC增加了在线微调功能，以提升其在应对应用程序需求变化和不同网络条件下的泛化能力。这使得在网络特征变化时，Deep CC不需重新训练就能够做出符合各种吞吐量、延迟和丢包率等均衡需求的决策。

3 应用层流媒体传输研究现状

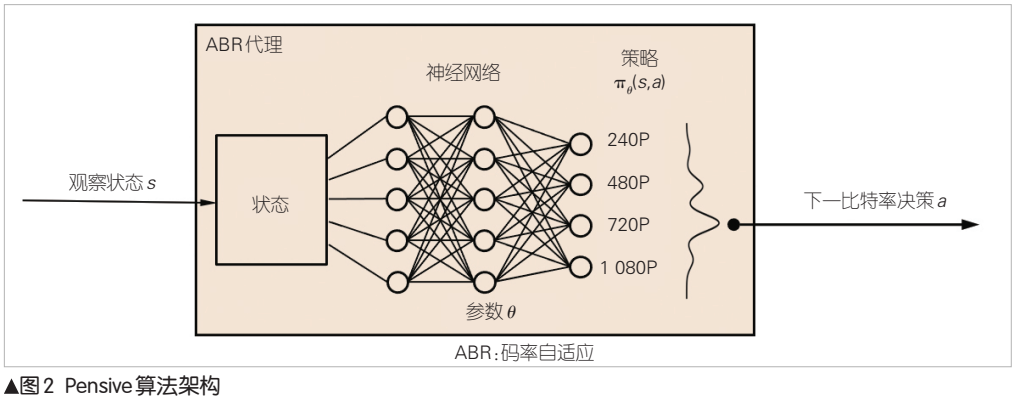
越来越多的学者开始研究将机器学习技术与传统应用层协议或算法相结合，以便更好地感知网络条件的复杂模式。近年来，视频流量快速增长，同时用户对视频质量的要求也在不断提高。ABR算法是内容提供商用来优化视频质量的主要工具，其准确性和性能会对视频流性能造成不可忽视的影响。然而，难以建模的网络变化和难以平衡的视频质量目标（如最大化比特率和最小化停顿）等因素，给设计高准确



▲图1 ARC学习框架结构

性、高效率的ABR算法带来困难。近年来，学界将人工智能技术应用于ABR算法，以增强其在底层网络变化时的适应性、效率和视频质量目标的平衡性能，并且取得了丰硕的成果。

2017年，MAO H. Z.提出Pensieve[37]，使用A3C算法生成ABR算法，以提高广泛网络条件和不同QoE目标下的性能。Pensieve基于历史传输指标选择未来视频块比特率，从而不依赖于网络环境的假设。如图2所示，观察状态s包含做出比特率决策相关的信息，如吞吐量测量值。一个由多个卷积网络（CNN）组成的2层神经网络，将状态映射到下一个比特率适应策略的概率分布。Pensive将目标设定为最大化总奖励，通过训练神经网络的参数θ，确定合适的比特率选择来提供更好的奖励。测试中，Pensieve的平均QoE提高了12%~15%。Pensieve能够深入学习特定网络场景的特征，



▲图2 Pensive算法架构

并根据经验选择适当的比特率决策进行输出，从而优化策略。然而这是一种采用离线训练并在线部署的方式，网络状况的变化可能会对其性能造成影响。为了解决这个问题，Pensieve需要支持定期更新ABR算法以应对新数据的到达。不过这会增加一定的计算开销，并需要系统从少量数据中快速收敛到优秀策略。因此，如何平衡再训练频率和性能仍然是一个待讨论的问题。

Pensieve等系统的速率控制方法侧重提供高视频比特率。但随着编码比特率的提高，视频质量的提升呈现出边际递减的效应。过分追求高比特率只会加大网络传输的压力，而视频质量提升甚微。HUANG T. C.提出了视频质量感知速率控制算法（QARC）[38]，目标是在实时视频流场景的比特率控制中平衡视频质量、延迟和带宽资源。QARC的视频质量预测网络（VQPN）模块通过历史视频帧预测未来视频质量，以解决训练过程中的“状态爆炸”问题。其视频质量强化学习（VQRL）模块使用A3C训练神经网络，根据VQPN的输出和历史时间网络状态数据，来输出满足视频质量和低延迟需求的比特率。

相似比特率和视频质量未必会为每一个用户都带来最好的QoE体验。这是因为传统的或数据驱动的ABR算法在优化QoE模型时，通常假设所有用户都具有相同的偏好[39-40]。LAI Z. Q.提出一种全景比特率适应技术。在虚拟现实（VR）场景中，当带宽不足而出现比特率降低时，该技术可优先降低用户视场外的视野质量，进而减少对用户的QoE影响[41]。ZUO Y. T.[42]提出一种视频系统Ruyi。该系统由偏好感知的QoE模型和基于神经网络的ABR算法组成。偏好感知的QoE模型输出用户偏好，并将其作为ABR算法神经网络的输入之一，从而

得出最佳比特率预测，以最大化用户特定的QoE，无须为不同用户重新训练模型。仿真实验表明，Ruyi能够提升所有用户的QoE，提升幅度高达65.22%。

先前的ABR算法研究中，预测模块都将吞吐量或交付时间视为优化目标。然而，至今还没有相关工作定量地研究这两个目标的效果差别。此外，传统的ABR算法均只考虑了网络条件波动对吞吐量造成的影响，忽略了应用程序本身行为（如On–Off周期）[43]的影响。GERUI L.等通过视频流测量平台来定量验证预测吞吐量和交付时间的所有影响因素的相关性，发现块大小和吞吐量之间存在强相关性，并深受播放器状态、播放器客户端所在平台信号强度、块索引的影响；通过分别比较多元线性回归、决策树与两种优化目标4种组合的预测误差，确定基于吞吐量的优化目标优于交付时间，并提出Lumos[43]。Lumos基于过去t个块的最大吞吐量、交付时间、客户端连接类型，以及最后块的比特率、大小、索引等训练回归决策树，通过集成到现有ABR算法中，提高吞吐量的预测精度。

4 展望与挑战

网络智能传输技术至今已经取得长足的进展，但仍面临诸多技术挑战。本文中我们从网络传输新发展、人工智能技术新发展、网络传输和人工智能结合的角度进行介绍。

4.1 网络传输新发展

传输层协议也在不断发展，基于UDP的QUIC协议被提出，并成为正式的RFC[14–16]。SHI H.提出了DTP[17]，并首次在数据块中提出截止时间、数据块优先级的概念；ZHANG J.在DTP基础之上设计了调度程序和自适应冗余机制[44]，以满足动态网络的多样化需求。

单路径的网络传输可能存在网络不稳定和吞吐量不足的问题。与单路径相比，多路径传输的优点是能够针对不同场景[45]和优化目标[46]提供网络的无缝切换和更大的聚合带宽。ZUO X. T.提出了截止日期感知的多路径调度框架[47]，根据网络状态信息对发送端缓冲区块进行排序，决定数据块的发送顺序和路径分配，以减少带宽资源的浪费。

此外，国际互联网工程任务组（IETF）成立了Media Over Quic工作组[18–19]，致力于建立利用QUIC协议发布媒体协议的机制。

4.2 人工智能技术新发展

近年来，机器学习技术在不断发展，出现了深度学习、强化学习等分支，提供了更强大的解决复杂问题的能力。然而，从网络通信的角度看，机器学习领域仍存在一些挑战。

如何构建分布比例合乎现实世界且标注正确的高质量数据集，同时提供一套方法进行验证，是训练模型场景中的挑战。例如，将机器学习应用到传输层拥塞控制策略时，由于网络流量的多变性，如何确保采集到的网络流量数据符合现实网络场景的流量分布，依旧是构建网络智能传输模型的难点。

近年来，自监督表示学习在自然语言处理领域获得了惊人的成功，成为最近一些大语言模型的基础（如LLaMA、LaMDA、Bard）[48–49]。大量参数和海量训练数据是大模型取得惊人成功的原因。然而，如何合理地对海量的网络数据（如抖动、带宽、时延）进行标注再训练，并应用于网络传输中，是一个主要挑战。

4.3 网络传输和人工智能结合

将机器学习技术应用于网络传输是人们持续关注的研究领域。通过机器学习技术对特定网络的深层规则进行建模，可能会在决策效果上超越传统协议。然而，如何更好地将网络传输与AI模型结合仍然面临一系列的挑战。

分布式机器学习通过将训练任务拆解并部署到多个节点，来加快模型的训练速度。在实际部署中，节点之间的通信方式、通信性能会对整体训练性能造成很大差异，传输性能优化是热门的研究方向之一[50–52]。如何根据分布式机器学习训练场景的数据传输周期性、数据中心流量对拥塞的敏感性的特点，及时避免拥塞，缓解交换机缓冲区的积压问题，依旧需要进一步研究。

对上层应用而言，不同的应用对网络传输的质量需求有差异，例如：实时交互性应用更关注抖动和延迟，而点播类应用更关注带宽。此外，对流媒体传输场景而言，不同用户对QoS的评判标准也有不同[42]。为每一个用户和应用都训练一个决策模型是不现实的；反之，一个统一的决策模型难以在多样的传输指标中实现平衡。对底层网络而言，由于实际网络环境是动态变化的，基于离线训练和线上部署的模型需要具有足够的泛化性。如何设计出适当的目标函数、训练框架，在底层网络的延迟、丢包率等指标变化时，做出准确的决策，并且服务好上层应用不同的需求，将对网络传输模型的设计提出更高要求。

网络智能传输模型的决策具有实时性的特点，例如需要根据动态变化的带宽、时延等网络数据做出决策。在现实网络中部署时，传统的决策结果从服务器的神经网络模型决策获得，集群计算资源的不足将会影响模型推理的时间。为了能在网络中大规模部署，模型的推理效率和开销优化都需要

进一步提高[36]。

5 结束语

人工智能技术已经应用于广域网端到端低时延网络传输中，并在传输层、应用层取得了一定的成果。本文分析网络传输时延的组成，并介绍了传输层中拥塞控制、应用层流媒体传输与机器学习技术相结合的研究。近年来，网络传输技术、机器学习模型技术的进一步发展，为网络智能传输带来更多的可能。我们认为，网络传输模型的泛化能力、大规模部署的开销、多目标学习之间的权衡、数据质量，是网络智能传输技术的新的研究主题与挑战。

参考文献

[1] LETAIEF K B, CHEN W, SHI Y M, et al. The roadmap to 6G: AI empowered wireless networks [J]. IEEE communications magazine, 2019, 57(8): 84–90. DOI: 10.1109/MCOM.2019.1900271

[2] BRAUD T, LEE L H, ALHILAL A, et al. DiOS—an extended reality operating system for the metaverse [J]. IEEE multimedia, 2023, 30(2): 70–80. DOI: 10.1109/mmul.2022.3211351

[3] ZELAYA R I, SUSSMAN W, GUMMESON J, et al. LAVA: fine-grained 3D indoor wireless coverage for small IoT devices [C]//Proceedings of the 2021 ACM SIGCOMM 2021 Conference. ACM, 2021: 123–136. DOI: 10.1145/3452296.3472890

[4] NI Y Z, ZHENG Z L, LIN X S, et al. CellFusion: multipath vehicle-to-cloud video streaming with network coding in the wild [C]//Proceedings of the ACM SIGCOMM 2023 Conference. ACM, 2023: 668–683. DOI: 10.1145/3603269.3604832

[5] ELBAMBY M S, PERFECTO C, BENNIS M, et al. Toward low-latency and ultra-reliable virtual reality [J]. IEEE network, 2018, 32(2): 78–84. DOI: 10.1109/MNET.2018.1700268

[6] WANG M W, CUI Y, WANG X, et al. Machine learning for networking: workflow, advances and opportunities [J]. IEEE network, 2018, 32(2): 92–99. DOI: 10.1109/MNET.2017.1700200

[7] JORDAN M I, MITCHELL T M. Machine learning: trends, perspectives, and prospects [J]. Science, 2015, 349(6245): 255–260. DOI: 10.1126/science.aaa8415

[8] VIOLA I, AMIRPOUR H, VEGA M T. IXR '22: 1st workshop on interactive eXtended reality [C]//Proceedings of the 30th ACM International Conference on Multimedia. ACM, 2022: 7412–7413. DOI: 10.1145/3503161.3554781

[9] ZUO X T, WANG M W, XIAO T X, et al. Low-latency networking: architecture, techniques, and opportunities [J]. IEEE Internet computing, 2018, 22(5): 56–63. DOI: 10.1109/MIC.2018.053681363

[10] AHMAD J, WARREN A. FPGA based deterministic latency image acquisition and processing system for automated driving systems [C]//Proceedings of 2018 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2018: 1–5. DOI: 10.1109/ISCAS.2018.8351472

[11] SUN L Y, ZONG T Y, WANG S Q, et al. Towards optimal low-latency live video streaming [J]. IEEE/ACM transactions on networking, 2021, 29(5): 2327–2338. DOI: 10.1109/TNET.2021.3087625

[12] PEDERSEN H A, DEY S. Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing [J]. IEEE/ACM transactions on networking, 2016, 24(2): 996–1010. DOI: 10.1109/TNET.2015.2410298

[13] XU Y D, ELAYOUBI S E, ALTMAN E, et al. Impact of flow-level dynamics on QoE of video streaming in wireless networks [C]//2013 Proceedings IEEE INFOCOM. IEEE, 2013: 2715–2723. DOI: 10.1109/INFCOM.2013.6567080

[14] LANGLEY A, RIDDOCH A, WILK A, et al. The QUIC transport protocol: design and Internet-scale deployment [C]//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. ACM, 2017: 183 – 196. DOI: 10.1145/3098822.3098842

[15] IETF. QUIC: a UDP-based multiplexed and secure transport: RFC 9000 [S]. 2021

[16] CUI Y, LI T X, LIU C, et al. Innovating transport with QUIC: design approaches and research challenges [J]. IEEE Internet computing, 2017, 21(2): 72–76. DOI: 10.1109/MIC.2017.44

[17] SHI H, CUI Y, QIAN F, et al. DTP: deadline-aware transport protocol [C]//Proceedings of the 3rd Asia-Pacific Workshop on Networking. ACM, 2019: 1 – 7. DOI: 10.1145/3343180.3343191

[18] IETF. Media over QUIC [EB/OL]. (2022-09-12) [2023-09-15]. https://datatracker.ietf.org/doc/charter-ietf-moq/

[19] IETF. MoQ relay for support of deadline-aware media transport [EB/OL]. (2023-09-11) [2023-09-15]. https://datatracker. ietf. org/doc/draft-ma-moq-relay-for-deadline/

[20] BOVY, C J, MERTODIMEDJO H T, HOOGHIEMSTRA G, et al. Analysis of end-to-end delay measurements in Internet [EB/OL]. [2023-09-15]. https://www. researchgate. net/publication/233863840_Analysis_of_end-to-end_delay_measurements_in_Internet

[21] MISHRA A, SHIN M, ARBAUGH W. An empirical analysis of the IEEE 802.11 MAC layer handoff process [J]. ACM SIGCOMM computer communication review, 2003, 33(2): 93–102. DOI: 10.1145/956981.956990

[22] PHEMIUS K, MATHIEU B. Openflow: why latency does matter [C]//2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), IEEE, 2013: 680–683

[23] CARDWELL N, CHENG Y C, GUNN C S, et al. BBR: Congestion-based congestion control: measuring bottleneck bandwidth and round-trip propagation time [EB/OL]. [2023-09-15]. https://doi. org/10.1145/3012426.3022184

[24] WOLFMAN A, VOELKER G, THEKKATH C A. Latency analysis of TCP on an ATM network [EB/OL]. [2023-09-15]. https://cseweb. ucsd. edu/~voelker/pubs/tcp-usenix94.pdf

[25] NOOR-A-RAHIM M, LIU Z L, LEE H, et al. 6G for vehicle-to-everything (V2X) communications: enabling technologies, challenges, and opportunities [J]. Proceedings of the IEEE, 2022, 110(6): 712–734. DOI: 10.1109/jproc.2022.3173031

[26] SWAMINATHAN V, WEI S. Low latency live video streaming using HTTP chunked encoding [C]//Proceedings of 2011 IEEE 13th International Workshop on Multimedia Signal Processing. IEEE, 2011: 1–6. DOI: 10.1109/MMSP.2011.6093825

[27] WINSTEIN K, BALAKRISHNAN H. TCP ex machina: computer-generated congestion control [C]//Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM. ACM, 2013: 123–134. DOI: 10.1145/2486001.2486020

[28] SIVARAMAN A, WINSTEIN K, THAKER P, et al. An experimental study of the learnability of congestion control [J]. ACM SIGCOMM computer communication review, 2014, 44(4): 479–490

[29] JAY N, NOGA R, BRIGHTEN G, et al. A deep reinforcement learning perspective on internet congestion control [EB/OL]. (2019-05-21) [2023-09-15]. https://arxiv.org/pdf/1810.03259.pdf

[30] HUANG T C, ZHOU C, JIA L C, et al. Learned Internet congestion control for short video uploading [C]//Proceedings of the 30th ACM International Conference on Multimedia. ACM, 2022: 3064 – 3075. DOI: 10.1145/3503161.3548436

[31] DONG M, MENG T, ZARCHY D, et al. PCC vivace: online-learning congestion control [C]//15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). ACM, 2018: 343–356

[32] ZHANG L, CUI Y, PAN J C, et al. Deadline-aware transmission control for real-time video streaming [C]//Proceedings of 2021 IEEE 29th International Conference on Network Protocols (ICNP). IEEE, 2021: 1–6. DOI: 10.1109/ICNP52444.2021.9651971

[33] NIE X H, ZHAO Y J, LI Z H, et al. Dynamic TCP initial windows and congestion control schemes through reinforcement learning [J]. IEEE journal on selected areas in communications, 2019, 37(6): 1231–1247. DOI: 10.1109/JSAC.2019.2904350

[34] ABBASLOO S, YEN C Y, CHAO H J. Classic meets modern: a pragmatic learning-based congestion control for the Internet [C]//Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication. ACM, 2020: 632 – 647. DOI: 10.1145/3387514.3405892

[35] ZHANG L, ZHU K W, PAN J C, et al. Reinforcement learning based congestion control in a real environment [C]//Proceedings of 2020 29th International Conference on Computer Communications and Networks (ICCCN). IEEE, 2020: 1–9. DOI: 10.1109/ICCCN49398.2020.9209750

[36] ZHANG L, CUI Y, WANG M W, et al. DeepCC: bridging the gap between congestion control and applications via multiobjective optimization [J]. IEEE/ACM transactions on networking, 2022, 30(5): 2274–2288. DOI: 10.1109/TNET.2022.3167713
[37] MAO H Z, NETRAVALI R, ALIZADEH M. Neural adaptive video streaming with pensieve [C]//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. ACM, 2017: 197 – 210. DOI: 10.1145/3098822.3098843
[38] HUANG T C, ZHANG R X, ZHOU C, et al. QARC: video quality aware rate control for real–time video streaming based on deep reinforcement learning [C]//Proceedings of the 26th ACM International Conference on Multimedia. ACM, 2018: 1208 – 1216. DOI: 10.1145/3240508.3240545
[39] ROBITZA W, GARCIA M N, RAAKE A. A modular HTTP adaptive streaming QoE model—candidate for ITU–T P.1203 ("P.NATS") [C]//Proceedings of 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017: 1–6. DOI: 10.1109/QoMEX.2017.7965689
[40] ESWARA N, ASHIQUE S, PANCHBHAI A, et al. Streaming video QoE modeling and prediction: a long short–term memory approach [J]. IEEE transactions on circuits and systems for video technology, 2020, 30(3): 661–673. DOI: 10.1109/tcsvt.2019.2895223
[41] LAI Z Q, HU Y C, CUI Y, et al. Furion: engineering high–quality immersive virtual reality on today's mobile devices [C]//Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking. ACM, 2017: 409 – 421. DOI: 10.1145/3117811.3117815
[42] ZUO X T, YANG J Y, WANG M W, et al. Adaptive bitrate with user–level QoE preference for video streaming [C]//Proceedings of IEEE INFOCOM 2022 – IEEE Conference on Computer Communications. IEEE, 2022: 1279–1288. DOI: 10.1109/INFOCOM48880.2022.9796953
[43] LV G R, WU Q H, WANG W R, et al. Lumos: towards better video streaming QoE through accurate throughput prediction [C]//Proceedings of IEEE INFOCOM 2022 – IEEE Conference on Computer Communications. IEEE, 2022: 650–659. DOI: 10.1109/INFOCOM48880.2022.9796948
[44] ZHANG J, SHI H, CUI Y, et al. To punctuality and beyond: meeting application deadlines with DTP [C]//Proceedings of 2022 IEEE 30th International Conference on Network Protocols (ICNP). IEEE, 2022: 1–11. DOI: 10.1109/ICNP55882.2022.9940391
[45] CUI Y, WANG L, WANG X, et al. FMTCP: a fountain code–based multipath transmission control protocol [J]. IEEE/ACM transactions on networking, 2015, 23(2): 465–478. DOI: 10.1109/TNET.2014.2300140
[46] ZHENG Z L, MA Y F, LIU Y M, et al. XLINK: QoE–driven multi–path QUIC transport in large–scale video services [C]//Proceedings of the 2021 ACM SIGCOMM 2021 Conference. ACM, 2021: 418–432. DOI: 10.1145/3452296.3472893
[47] ZUO X T, CUI Y, WANG X, et al. Deadline–aware multipath transmission for streaming blocks [C]//Proceedings of IEEE INFOCOM 2022 – IEEE Conference on Computer Communications. IEEE, 2022: 2178–2187. DOI: 10.1109/INFOCOM48880.2022.9796942
[48] LIU X, ZHANG F J, HOU Z Y, et al. Self–supervised learning: generative or contrastive [J]. IEEE transactions on knowledge and data engineering, 2023, 35(1): 857–876. DOI: 10.1109/TKDE.2021.3090866
[49] BROWN T, MANN B, RYDER N, et al. Language models are few–shot learners [EB/OL]. (2020–05–28) [2023–07–25]. https://arxiv. org/abs/2005.14165
[50] HU S H, ZENG G X, BAI W, et al. Aeolus: a building block for proactive transport in datacenter networks [J]. IEEE/ACM transactions on networking, 2022, 30(2): 542–556. DOI: 10.1109/TNET.2021.3119986
[51] BAI W, HU S H, CHEN K, et al. One more config is enough: saving (DC) TCP for high–speed extremely shallow–buffered datacenters [J]. IEEE/ACM transactions on networking, 2021, 29(2): 489–502. DOI: 10.1109/TNET.2020.3032999
[52] WU H T, FENG Z Q, GUO C X, et al. ICTCP: Incast Congestion Control for TCP in data center networks [C]//Proceedings of the 6th International Conference. ACM, 2010: 1 – 12. DOI: 10.1145/1921168.1921186

作 者 简 介

廖乙鑫，北京邮电大学在读硕士研究生；主要研究领域为低时延网络传输、视频传输。

王子逸，北京邮电大学副研究员；主要研究领域为低时延网络传输、流媒体传输、边缘计算等；在《IEEE/ACM Transactions on Networking》、INFOCOM等权威期刊与会议上发表论文多篇。

崔勇，清华大学长聘教授、网络技术研究所所长，教育部"长江学者"特聘教授，首届"青年长江学者"获得者，中国互联网协会学术工作委员会秘书长，中国通信学会边缘计算委员会副主任委员，《IEEE Transactions on Parallel and Distributed Systems》等4个IEEE期刊的编委，曾长期担任国际标准工作组主席；主要研究方向为低时延传输技术、视频分析、内容安全、流媒体传输、网络数字孪生、网络AI等；获国家优秀青年科学基金和教育部新世纪人才等项目持续支持，2019年成功在北京组织国际网络通信领域顶级会议Sigcomm'19并担任大会副主席，参与研制我国第一台"IPv6核心路由器"，参与建设中国下一代互联网示范工程CNGI–CERNET2；获国家科技进步二等奖1次、国家技术发明二等奖1次，多次获得国家信息产业重大技术发明；发表论文100余篇，获国家发明专利40余项，完成RFC国际标准10余项，出版学术著作4部。