Survey on Machine Learning for IntelligentEnd-to-End Communication Toward 6G: From Network Access, Routing to Traffic Control and Streaming Adaption

1.M.Mounika,

Assistant Professor in Computer Science(School of Engineering), Mallareddy University, Hyderabad.

2.Shravani Amar,

Assistant Professor in Computer Science(School of Engineering), Mallareddy University, Hyderabad.

***Abstract*—The end-to-end quality of service (QoS) and quality of experience (QoE) guarantee is quite important for network optimization. The current 5G and conceived 6G network in the future with ultra high density, bandwidth, mobility and large scale brings urgent requirement of high efficient end-to-end optimization methods. The conventional network optimizationmethods without learning and intelligent decision ability are hardto handle the high complexity and dynamic scenarios of 6G. Recently, machine learning based QoS and QoE aware network optimization algorithms emerge as a hot research area and attractmuch attention, which is widely acknowledged as the poten- tial solution for end-to-end optimization in 6G. However, there are still many critical issues of employing machine learning in networks, especially in 6G. In this paper, we give a compre- hensive survey on the recent machine learning based network optimization methods to guarantee the end-to-end QoS and QoE. To easy to follow, we introduce the investigated works following the end-to-end transmission flow from network access, routing to network congestion control and adaptive steaming control.Then we discuss some open issues and potential future research directions.**

***Key Words*—End-to-end, quality of service (QoS), quality of experience (QoE), machine learning (ML), deep learning (DL), network access, resource allocation, channel assignment, routing, congestion control, adaptive streaming control, adaptive bitrate streaming (ABR).**

I. INTRODUCTION

5G IS widely deployed in many countries and brings high quality of service (QoS) and experience (QoE) of enhanced mobile broadband (eMBB), ultra-reliable and low-latency communications (uRLLC), and massive machine type communications (mMTC) [1], [2] to users. Meanwhile, the next 6G network is conceived by researchers with new features of ultra high density of connections, ultra high frequency bandwidth (Terahertz), ultra high mobility (>500km/H) and extremely large network scale (> 10^6 devices) [3], [4], which also brings higher QoS and QoE requirement such as Ultra low latency, ultra big throughput and high energy efficiency for end-to-end communications.

The guarantee of end-to-end QoS and QoE is critical in the network to provide high-quality information and accelerate the communication efficiency of users. To ensure the end-to-end QoS and QoE, various protocols are proposed to optimize network function from the data-link layer to the application layer [5]–[8]. However, the most widely used media access control (MAC) layer control, routing protocol, transmission control protocol (TCP) are designed decades ago, which is hard to adapt to the highly complex and dynamic environmentof next generation network. For example, the conventional TCPin the complex environment is more than 10 times from optimalvalue in terms of packets loss rate [9]. Notably, as shown in Fig. 1, the conceived 6G network coverage from underwater area to air and space and constructed with various devices and heterogeneous structures, which makes the QoS and QoE much harder to be guaranteed with conventional methods.

In the past years, the researches mainly focus on the network function optimization based on fixed policy in the certain layer in the Open System Interconnection (OSI) modelto optimize the end-to-end communication. However, some researches recently propose that the QoS guarantee in the6G should be scheduled intelligently with overall cross-layer end-to-end optimization instead of the single layer and single link [10], [11]. For example, the conventional MAC protocol manages the radio access most focuses on the performance of direct link within one-hop communication. This is not optimalin end-to-end communication, as there exists a multi-player competition and the performance might be affected during the multi-hop transmission [12]. Thus, the end-to-end QoS aware access control should be considered instead of the direct link state based access control. After the media access, the network routing, congestion control in the network/transportation layer, and steaming adaption in the application layer mostly designed based on fixed rules or policy. Those fix rule based algorithm can not aware of the dynamic network changing and causes QoS and QoE drop when the assumed environment varying. To conquer those problems, many researchers propose thatthe future end-to-end QoS guarantee algorithm should also

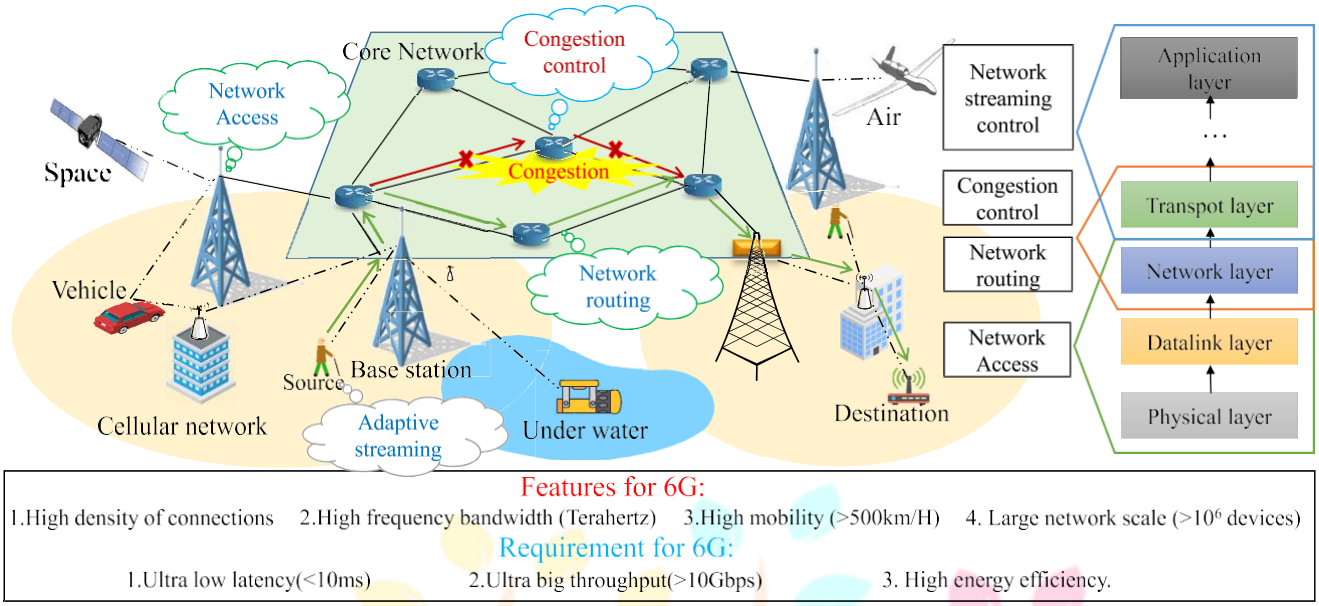


Fig. 1. The features and requirements for the end-to-end communication in 6G.

be designed to be aware of the end-to-end QoS/QoE and intelligent make decision by adjusting to the varying network state [3], [4], [13].

Nearly 30 years ago, the AI technology of machine learn- ing was initially employed for network function optimization of network state prediction and routing to ensure end-to- end QoS [14]–[16]. However, due to the limitation of boththe computation ability of computer and complexity of early machine learning algorithms, the initial machine learning based networking do not show significant improvement on end-to-end QoS and QoE. Fortunately, with the rapid devel- opment of computer technology and breakthrough of AI algorithm, powerful AI technology especially deep learning is widely proved to be efficient to endure intelligence for things. Naturally, deep learning is once again considered as an potential tool to improve end-to-end communication, and proved to be efficient on various network optimization areas such as network traffic prediction, traffic classification and congestion control [17], [18]. One emerging concern of employing deep learning to the network is the computation ability of the network infrastructures. Recently, the new generation network structure of Software Defined Network (SDN) and network hardware of SDN-switch, Graphics Processing Unit (GPU)-accelerated router and Network Functions Virtualization (NFV) aided base-station have been widely emerged to ensure the high computation and networking ability for next generation network [19], [20]. The 6G network is conceived to support ubiquitous Artificial Intelligence (AI) services from the core to the end devices of the network [3], which makesthe wide deployment of the machine learning system become possible. With both computation and communication evolution, the machine learning based network may pave the way to the future intelligent 6G [4].

In this paper, we give a comprehensive survey on machine learning for intelligent end-to-end communication service guarantee from the data-link layer to the application layer. The structure of the paper is constructed as shown in Fig. 2. We introduce the machine learning based network optimization approaches in terms of the objectives following the layers structure of OSI. At the beginning of end-to-end communication, machine learning is employed for dynamic network access control including power, channel and joint resource allocation. We introduce the machine learning based network access in terms of the three aspects in Section II. Then, the machine learning based routing approaches are investigated in terms of the different learning types (e.g., supervised, reinforcement, and imitation learning) in Section III. In the third part of Section IV, we survey the machine learning based adaptive streaming control for end-to-end QoE optimization depending on different transmission types (e.g., single-path and multi-path transmission). Beyond the transportation layer, the machine learning based adaptive streaming control for end-to-end QoE optimization in the application layer is introduced in Section V. In the last part, we discuss some open research issues and summarize potential future directions in Section VI.Finally, in Section VII, the article is concluded.

II. Machine Learning for Network Access in MAC Layer

The MAC is widely researched for decades, the main task of the MAC protocol is to avoid collisions from interfering nodes [21]. A large amount of untapped spectrum resources at terahertz in 6G creates new opportunities for access acceleration but also brings new challenges for the MAC design. In conventional MAC design, most of the access control algorithms mainly focus on the optimization of direct connection in the MAC layer. However, the good performance of direct connection did not mean a high level of end-to-end quality, as the direct access control is selfish and is unfair to other nodes. Especially in the ultra-high dense 6G network, the unfairness always causes high congestion not inner cell but

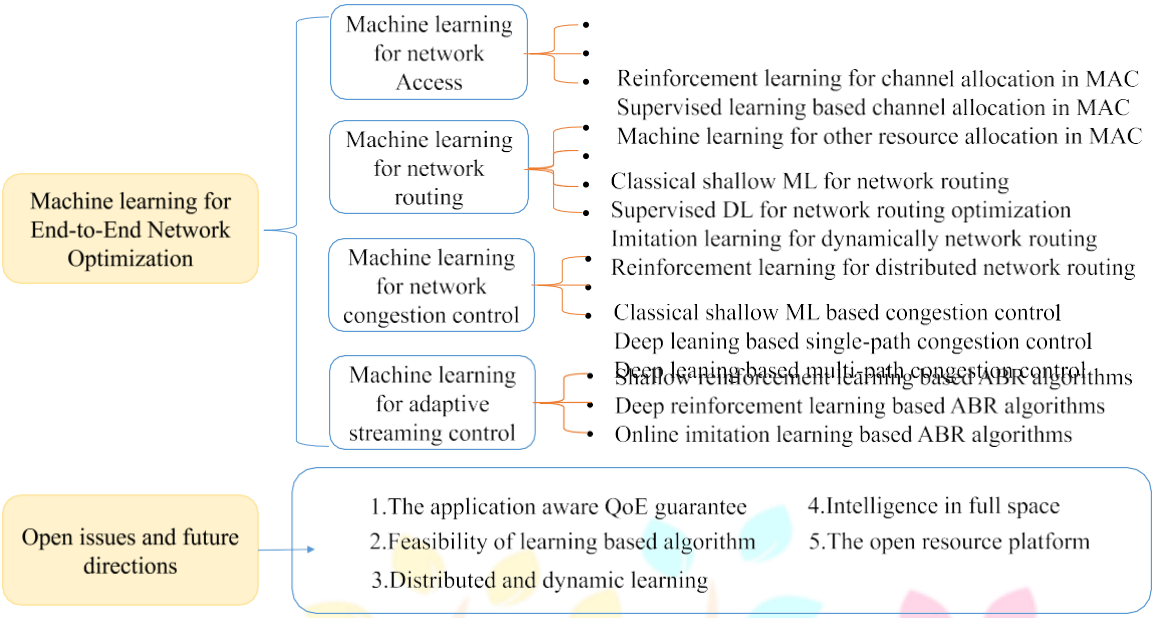


Fig. 2. The structure of the survey.

inter cells and significantly reduces the end-to-end QoS [22]. Furthermore, the only MAC layer information not reflecting the end-to-end QoS is insufficient for MAC optimal deci-sion making. Especially in the wireless network of 6G, the high dynamic change of both network topology and channel state significantly affects the performance of the MAC proto- col. Tackling this issue, the cross-layer end-to-end QoS aware MAC algorithms are widely proposed [12], [23]. However, the conventional end-to-end QoS aware MAC algorithms based on fixed rules or predefined policy are hard to handle the complex and dynamic environment of the next-generation network. In this section, we in detail introduce the novel machine learn- ing (ML) based approaches in MAC design and illustrate how the enabled intelligence surmounts the dynamic state-changing problem in wireless network.

A. Reinforcement Learning for Channel Allocation in MAC

Due to the high dynamic of the wireless network (e.g., nodes moving, channel changing and network traffic waving), the fixed rule based channel allocation algorithms without intelligence are always stuck in the performance drop when the state changes. Benefiting from the development of the powerful intelligent tool, the machine learning based MAC algorithms are widely proposed to be the potential solution to adding intelligence to the MAC protocol. The earliest machine learning based MAC algorithm can trace back to a decade ago, the initial Q-learning based dynamic multichannel Access control is proposed to schedule the channel resources in mobile cognitive radio system [24]. The authors model the multiple channel access as a game process and evaluate the game utility based on the usage of the sub-channels. With the simple Q-learning, the agent can learn and select the best channel access strategy by maximizing the utility function (i.e., accumulated reward).

Recently, the more powerful deep learning architecture brings more efficient learning ability and inspires a lot of deep learning based channel allocation algorithms. The researchers in [25] considers the network state is time-varying and hard to be simply modeled with fixed rules. A deep q-learning based channel allocation algorithm is proposed in their proposal. Comparing with the conventional method, the proposal can extract the network feature and predict the future network state without knowing the system statistics. Both simulation and real data trace based test show that the proposal achieves near-optimal performance in more complex situations (e.g., large scale and time-varying), comparing with conventional Myopic policy [5] and Whittle Index-based heuristic algorithms [26].

With the explosive increase of connections of mobile devices, power consumption becomes a critical criterion of the measurement of QoS. By jointly considering the power consumption and network performance, the work in [27] proposes a deep reinforcement learning based multi-access control and dynamic energy harvesting algorithm. Instead of the DNN used in the previous work, a time-varying sensitive deep learning structure called Long short-term memory (LSTM) is employed in this work for features extracting, and the channel allocation policy is dynamically decided by the updated reward of sum rate of all the uplink transmissions. The work in [28] also uses LSTM based deep Q-learning for dynamic channel allocation. In this work, the unlicensed spectrum in Long Term Evolution (LTE) is considered with limited resources, the proposal proactively detects carrier state and dynamically select channels based on the reward measured with end-to-end throughput.

The previous reinforcement learning mainly focuses on self-learning with local information and lacks consideration of multi-agent competition. Li and Guo propose a distributed multi-agent deep reinforcement learning based channel allocation framework for D2D underlay communications [29]. With the multi-agent actor-critic (state-of-art policy gradient in reinforcement learning), the proposed channel allocation algorithm is divided into off-line learning and on-line execution phases.

In the off-line learning period, the agents share global historical states, actions and policies and do centralized training. Then, in the online phase, the agents can cooperate with each other to optimize the end-to-end QoS by allocating channels following trained actions.

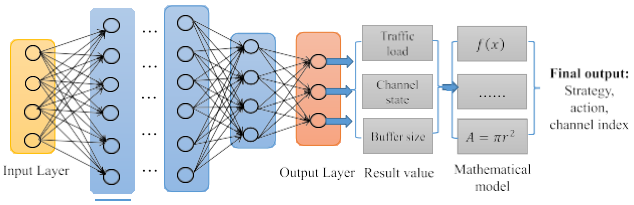
In order to accurately measure the radio resource competition between nodes, the following two works describe the multi-agent channel allocation process with an auction model [30], [31]. To find the best strategy to the auction gameand speed up the convergence speed, the work in [30] pro- poses a deep reinforcement learning based channel allocation algorithm for the time-vary channel for carrier sensing multiple access (CSMA). In the proposed learning algorithm, users justobserve and collect the channel state of single channel without information exchange with other nodes. Without a redundant game process, the proposal converges to the optimal strategy achieving an order optimal regret of O (log T), where T is the length of time horizon length.

Besides the channel allocation in the ground network, the work in [32] proposes a deep reinforcement learning based channel allocation algorithm for satellite network. Instead of the reward calculated with channel utilization or end-to-end throughput used in previous works, the work calculated the reward depending on the sum number of the satisfied services. Although the dynamic access is handled in this work, the dynamic topology change of the satellite itself is not fully considered. Therefore, the researchers in [33] further consider the vary topology issue and proposes a Graph Convolutional Network (GCN) based channel allocation algorithm for the wireless network. In the proposal, the GCN is employed to extract the features of the channel vectors with topology information, and reinforcement learning is online processed to fine-tune the value matrix. The simulation demonstrates the proposal achieves significant performance improvement in the dynamic network with varying topology changes.

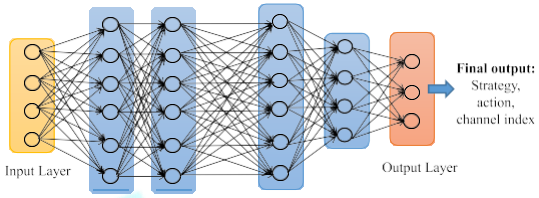
B. Supervised Learning Based Channel Allocation in MAC

Besides reinforcement learning which directly explores actions, supervised learning is also considered as a solution for the dynamic channel allocation. The supervised learning trains the agent with labeled data which is collected from conventional methods, thus hard to directly used for allocation (the approximated function can not exceed the benchmark without additional optimization). Supervised learning is widely considered for part or combined use with other optimization methods as shown in Fig. 3(a). In such approach, the deep learning structure is used for prediction the state value such as traffic load, remaining buffer size and channel state. Then, a mathematical model is leveraged to the final actions such as chosen resources allocation strategy.

For example, the algorithm in [34] proposes a liquid state machine (LSM) algorithm to predict the content requirement from users for the high dynamic unmanned aerial vehicle (UAV) network. Then, a joint spectrum allocation and con- tent caching strategy is processed based on the predicted user requirement. To accurately predict the requirement of users, the vector of the context information including age, gender,



(a) The deep learning combined with other optimization methods



(b) The one step deep learning based resource allocation

Fig. 3. Two kinds of deep learning based resources allocation.

occupation, and device type is used as input, and the output is characterized with a vector of discrete probabilities of content requirement density. With simulation, the supervised LSM based algorithm shows even better performance than the Q-learning based algorithms.

Besides LSM, the DNN is used for traffic load prediction and joint channel and power allocation in [35]. The workshows that the deep learning based algorithm achieves 85% accuracy of the joint channel and power allocation, and gets more than 10 times convergence speed than the conventional algorithm.

Considering the spatial features is hard to be extracted by DNN, a deep CNN structure is used for traffic load prediction and channel allocation is proposed in [36] for the dynamic Internet of Things (IoT). In this work, the deep CNN is employed to extract the features of the network and predict the future traffic load. Then, a DNN is followed to quickly allocate the channel based on the predicted traffic load. Three types of prediction and allocation mechanisms based on different kinds of control methods (e.g., centralized, semi-centralized and distributed) are analyzed and compared for different scenarios. To simplify the learning structure, the improved work in [37] integrates the prediction and channel allocation parts into a deeper CNN, which shows same-level performance and less memory overhead. Those supervised channel allocations show good performance in the complex networks.

Besides CNN, the time sequence sensitive deep recurrent neural network (RNN) is also exploited in supervised channel allocation. The authors in [35] propose a gated recurrent unit (GRU) based RNN structure to model the long-term temporal dependencies and predict the traffic load. With deep RNN, the predicted loads achieve almost half accuracy improvement than Auto-Regressive Integrated Moving Average (ARIMA) or multi-layer perceptron (MLP) based predictor.

C. Machine Learning for Other Resource Allocation in MAC

The next-generation network constructed with a heterogeneous network structure causes frequent interference and

TABLE I
THE LIST OF MACHINE LEARNING BASED NETWORK ACCESS ALGORITHMS

Learning type	System model	Task/Action	Value network	Approximation	Reward/Output	Literature
Reinforcement learning	NONE	multichannel access control	Q-table	Q-learning	Binary vector of the usage of the sub-channel	[24]
Reinforcement learning	MDP	multichannel access control	DBN	Q-learning	NONE	[25]
Reinforcement learning	MDP	multi-access contrd and Battery Prediction	LSTM	Q-learning	The sum rate of all the uplink transmissions	[27]
Reinforcement learning	Game theory	Dynamic channel selection	LSTM	Q-learning	Transmission throughput	[28]
Reinforcement learning	MDP	Channel allocation	CNN	Q-learning	Sum number of the satisfied services	[32]
Supervised learning	Liquid model	Spectrum allocation	LSM	LSM-algorithm	Probabilities of content requested by user	[38]
Supervised learning	Game theory	Channel allocation	CNN	SGD	taffic load	[37]
Supervised learning	Game theory	Channel allocation	CNNDBN	SGD	taffic load	[36]
Supervised learning	NONE	radio resource allocation	RNN	SGD	taffic load	[35]
Reinforcement learning	Markov game	Channel allocation	DBN	Actor-critic	the D2D link rate and the SINR	[29]
Reinforcement learning	Auction theory	Channel allocation	Q-table	Auction	estimated QoS of the link	[31], [30]
Reinforcement learning	MDP	Channel allocation	GCN	Q-learning	total end-to-end throughput	[33]

access competition. Such interference and competition conventionally are controlled via distributed resource allocation methods. However, with the increase of network complexity, the fixed rule based distributed resource allocation algorithms can not predict the environment changing and response in real-time. Employing machine learning for dynamic network resource allocation has been proposed for heterogeneous networks recently. Besides the deep learning based prediction aided resource allocation mentioned in Fig. 3(a). Another approach is the one step deep learning based resource allocation as shown in Fig. 3(b). Instead of the predicted value, the output of the neural network is simply characterized as the final actions, which is also widely used in the reinforcement learning and imitation learning based network optimization system [39].

To ensure the QoS for users and fairness inter heterogeneous cells, the work in [40] proposes a Q-learning based power resource allocation algorithm. In this work, the users in different cells are considered as competitive agents, with the construction of inter-cell interference, the optimal multi-access and power resource allocation is modeled as a non-convex optimization problem. With one step deep Q-learning, the proposal intelligently allocate power resources to users to reduce interference and maximize the QoS of users. Besides, the work in [41] proposes a supervised learning based power allocation algorithm for links based on the offline trained deep belief networks (DBNs).

Besides power resource allocation, the time resource is also considered via machine learning for intelligent alloca- tion. Due to the inter-cell and D2D-cellular interference, the conventional Time-division multiple access (TDMA) faces the problem of distributed allocation of time slots for users without knowing real-time information of other cells. Thepaper of [22] proposes a deep Q-learning based time slotallocation algorithm based on the past links state and traffic patterns. Without collecting the global information, the pro- posal shows the superiority in high dynamic heterogeneous networks.

Instead of individual resource allocation, to achieve the optimal end-to-end QoS and fairness to users, the practi- cal access control is always optimized with joint resource allocation including channel, power and time resources. For example, the works of [42]–[44] jointly allocate channel and power resources with distributed Q-learning. The worksof [45], [46] propose reinforcement learning based joint computation, channel and power allocation algorithms for ensuring end-to-end QoS.

D. Summary and Discussion

From the above surveys, we can see that machine learning is widely applied in intelligent resource allocation and network access control, and the different types of machine learning are efficient for various scenarios depending on the definite problems. Compared with classical shallow learning, deep learning shows more power in the highly dynamic and complex network. However, it does not mean the classical machine learning with shallow structure is useless for thefuture network access problem. Shallow machine learning is low computation cost and power consumption which is more efficient in some scenarios where the power and computation are limited. Furthermore, the classical machine learning

algorithm with stronger interpretability is more suitable for the case where the ultra-low fault tolerance rate is required.

III. Machine Learning in Network Layer for End-to-End Network Routing

Beyond link reliability, the end-to-end reliability is mainly depending on the network routing mechanism and congestion control algorithm. Network routing is an end-to-end path selection process for network traffic from transmitter to receiver. The network routing can be the single path or multi-paths and is processed in a single network or across heterogeneous networks. In the conventional network, the routing protocol mainly depends on the IP protocol and has been developed for decades. The classical routing protocol includes distance vector based routing (BGP [47], Babel [48]) and link state based routing (OSPF [6], IS-IS [49]) mainly depends on the maximum or minimum value or metric for instance. However, these methods have different shortcomings such as week adaptability for dynamic environment, slow convergence speed for large scale network and poor maintainability. For example, the traditional Shortest Path (SP) based routing algorithm has the problem of slow convergence, which is not suitable for dynamic networks as the slow response to the network changes can lead to severe congestion [6], [17]. The end-to-end QoS suffered from those weaknesses and can not meet the needs of the future high dynamic and large scale network in 6G. To conquer the shortage of the conventional routing protocol, machine learning based routing algorithms are proposed.

A. Classical Shallow ML for Network Routing

In the early age when the Internet emerged, the classical machine learning algorithms such as the Bayesian classi- fier, Logistic regression, Perceptron model, decision tree and K-Nearest Neighbor (KNN) algorithms are already widelyused for many areas including pattern recognition and natural language processing (NLP). However, the initial implementation of machine learning for network routing was proposed until the machine learning algorithm of neural network (NN) was proposed. In 1989, Zhang and Thomopoulos, proposed a neural network aided shortest path selection algorithm in network [14]. The proposal uses a neural network to approximate the weights of links and chooses multi-link paths by minimizing the cost function which is measured as network delay in this work. The work is an initial attempt at exploiting machine learning for network routing problems. However, the early neural network without proper back-propagation and active function is not capable of high dimension input and is with slow convergence speed. Besides, this initial machine learning based routing algorithm did not consider the traffic patterns, changing network state and signaling overhead which is not suitable for the real case. After the initial machine learning based routing, some other shallow machine learning based routing algorithms are proposed. For example, [15] treats the wireless sensors as modified neural nodes to build a self-organized integrating network wireless sensor network (WSN), the [16] solved the shortest path selection problem in the computer network, and the [50] expend the path selection problem from a single path to multi-path and solve it with Q-learning algorithm. However, none of them considered the dynamic network traffic patterns and complex network state, and it is hard to handle the real network. Due to those issues, the initial machine learning based routing algorithms have not been able to enter the spotlight for network researchers.

This situation continues till the great breakthrough of machine learning technology. In 1986, Rumelhart *et al.* first proposes the classical back-propagation method to solve both the convergence and scalability problem in conven- tional machine learning algorithm [51]. In 2006, Hinton and Salakhutdinov first propose the concept of Deep Learning(DL), which is a generative deep architecture designed to characterize the high-order correlation properties of the input data for synthesis purposes [52], the machine learning especially deep learning becomes powerful to handle complex input data with sophisticated state space and can efficiently avoid the curse of dimensions. Naturally, the powerful deep learning quickly becomes the most exploited machine learning tools for network routing recently. In the next parts, we focus on the introduction of the recent works of using different types of deep learning tools for network routing.

B. Supervised DL for Network Routing Optimization

To exploit the possibility of deep learning in network traffic control, our previous works [20], [53] at first give a proof-of-concept on how the deep learning based routing intelligently control network traffic and improves the network performance in terms of packets delay, network throughput and packets loss rate.

In the initial works, the supervised learning with deep belief network (DBN) is employed as the routing controller in edge nodes in the network to dynamic schedule network traffic flow to optimal paths based on the historical traffic patterns. The work flow is shown as in Fig. 4.

As shown in the figure, the edge nodes collect local information including historical traffic patterns and characterize them as input data. With the deep neural network, the learning system continues to fine-tune the neural network and predict the possible information of the unknown part of the core network. Based on the predicted information, the learn- ing system can intelligently make routing decisions almost as accurate as of the one in which information of the core network is known. Comparing with conventional routing algorithms, the proposed deep learning can efficiently decide optimal rout-ing path just based on the information of edge nodes, which significantly decreases the signaling overhead.

However, the deep belief network (DBN) used in the proposal is highly dependent on the known and fixed network topology. One big challenge of the proposal is the adapt-ability of the varying network topology. The Graph Neural Network (GNN) is widely researched to deal with the varying topology problem. The authors in [54] employ the GNN to distributed network routing problem. Compared with previousworks, this work considers the varying network topology caused by moving node and join/break links. By adding the edge information to the input data, the proposal can achieve

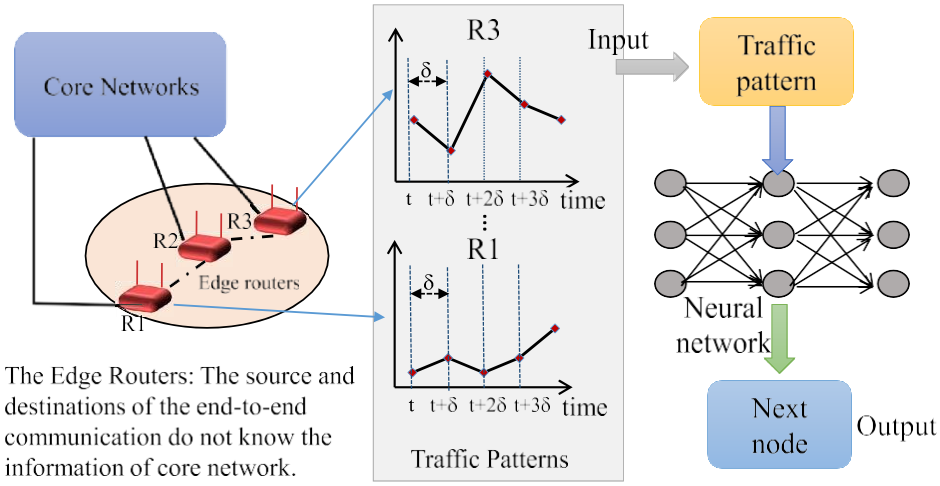


Fig. 4. The idea of supervised deep learning based routing.

above 95% decision accuracy even in the scenario of frequent changing network topology.

The main idea of the above research is utilizing the powerful relation extraction ability of deep learning to recover the correction just based on incomplete information. Another usage of supervised deep learning in network routing is to classify traffic flow and dynamic schedule the routing path based on the classified traffic flows. In [55], the authors consider using deep learning to jointly optimize wavelength assignment and routing problem in the optical network. In the work, the authors model the joint wavelength assignment and routing problem as a machine learning task classification problem and employ the Logistic Regression and DNN to train the Software-Defined Optical Network with collected training data set.

The previous researches show the outperformance of deep learning based routing algorithm comparing with conventional routing method. However, a big question is why deep learning is suitable for routing collection extraction? In the research of [56], the authors proved the hypothesis proposed by the previous researches that there is a mapping relationship between the optimization problem instance expressed in graphs and the optimum decisions. Furthermore, in addition to the traffic measurement, this work uses enhanced supervised learning with a deep feed-forward network (DFN) for accurate link-state evaluation, and the links state based routing scheduling algorithm is proposed.

However, network performance improvement is not evaluated in the above work. Thus, in [57] a supervised graph-aware deep conventional learning (GCN) approach is proposed. Comparing with previous works, one contribution is that the authors further characterized the traffic patterns including both historical traffic tracks and links state information such as link connectivity, link latency and packets queuing tracks. Another contribution is that the authors proposed that thegraph information such as connections of neighbors are neces-sary to be trained in the learning system. With simulation, the network performance of the proposal is proved to be improved than conventional methods.

The supervised learning based routing algorithms are considered as the potential solution for intelligent routing, which may efficiently reduce the computation and signaling overhead compared with the conventional method. However, supervised learning faces one big problem in the routing and resource allocation problem in the network: the accuracy of the routing decision based on supervised learning is hard to outperform the benchmark method that the training data collected from. Furthermore, even the supervised learning based algorithms show improved performance in a specific scenario, which might be overfitting and less of universality for more general scenarios. Thus, the imitation based learning and reinforcement learning based routing algorithms are proposed to get rid of the benchmark training data and improve both the adaptability and universality.

C. Imitation Learning for Dynamically Network Routing

Imitation learning is a prediction and imitation based intelligent algorithm, which is an online algorithm and not depends on labeled data. The main idea of imitation learning is to learn from the expert and imitate the optimal behaviors in an online manner.

Thus, with the concept of imitation learning, some researchers consider the network routing with the best routing decision as the expert behaviors, and the agent of the network learn the state information and imitate the optimal decision continuously. Fig. 5 shows the main difference between imitation learning and supervised learning. The supervised learning shows the adequate fitting ability to “copy” the behaviors of the benchmark. However, such kind of off-policy method mayfall into a bad situation when the benchmark makes mistakes such as compounding errors. Thus, imitation learning contin- uously explores the possible decisions and imitate the optimal decisions (expert behaviors), which learned to efficiently avoid possible bad actions and recover from failures.

Following the idea of imitation learning, a CNN based network traffic control algorithm was proposed in [58], [59]. In this work, the network continuously explores different routing

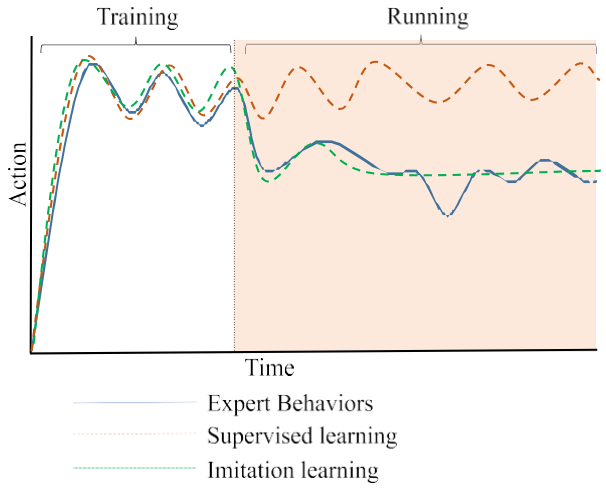


Fig. 5. The difference between the imitation learning and supervised learning.

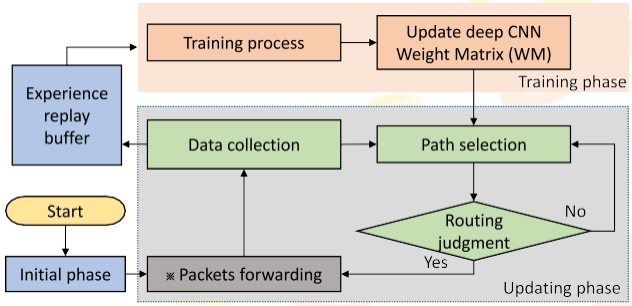


Fig. 6. The work flow of imitation learning based network traffic control.

paths (potential policies) and selects the valid path combinations (expert behaviors) based on the future congestion state. Then, a CNN based training system is employed to imitate the valid paths combinations. The workflow is simply introduced as in Fig. 6.

As the figure demonstrates, the proposed imitation based routing algorithm is parallelly running with training and updating phases. In the initial phase, the actions space is initialized with potential path combinations, then the updating phase continuously explores the actions and measure the expert behaviors based on the future congestion state. The explored triples of state, action and future state are recorded in the experience replay buffer. Meanwhile, the training phase exploits the actions and learn the explored expert behaviors with the training data in the experience replay buffer. Such training and running process are in an online manner, which is the first work considering using imitation learning to dynamically adjust the routing strategy to the high dynamic traffic. The simulation result shows that the proposed imitation learning based traffic control mechanism can efficiently avoid potential congestion and significantly reduced the packet loss rate and improve end-to-end throughput. However, one shortage of this work is that the high computation overhead and unsuitable for changing topology.

One solution to the varying topology adaption is to include the geometric and connection information of nodes in theinput of the learning system. In [60], another routing approachwith imitation learning called ST-DeLTA is proposed to deal with large scale network with varying topology. To adapt to the changing network topology, the proposed Spatial-Temporal model employs the 3D convolutional neural network to extract not only the temporal features of traffic patterns but also consider the location information of nodes. Besides, to decrease the computation overhead, the ST-DeLTA changes the action of routing strategy from binary routing feasibility judgmentof each path combination to the next-hop router selection. In this algorithm, the value matrix is proposed to judge the pos- sible reward of each action (chosen next hop) and records the expert behaviors when the topology is dynamically changing. The mathematical analysis shows that the proposal can significantly reduce both the time and space complexity of the decision-making process, and the end-to-end throughput and delay are proved to be improved in the simulation result.

Also considering the imitation for routing, the workof [61] proposed a network partition based intelligent routing. Compared with the next node selection strategy of previous works, this proposal chooses the whole path vector from the source and destination as the action space. Such improve-ment solved the local optimum caused by the only next node prediction of action space. Any route prediction error of the next route node will inevitably lead to the failure of the entire routing path prediction. Also, take care of this issue, the work in [62] proposes an online extreme learning machine (ELM) based transmission path selection algorithm. In addition to the normal traffic congestion caused by busy connections,the authors in this work employ similar imitation learningfor secure path selection to avoid potential attack and abnor- mal congestion in the edge computation system. In this work, instead of binary state prediction in [58], the authors evaluate the transmission probability of all possible paths and choose the best paths correspondingly. Such best path selection strat- egy increases the convergence speed of the learning algorithmand improves end-to-end QoS.

Another solution to adapting to the varying topology isthe flexible learning structure. The Graph Neural Networks (GNN) which directly operates on the graph structure is provedto be efficient for the flexible situation in fields including image classification and video processing to speech recogni- tion and natural language understanding [63]. Naturally, the authors in [64] considered Leveraging Graph Neural Networks for network modeling and control. In this work, the GNN is employed to predict the Key Performance Indicators (KPI) of the network such as end-to-end delay, packet loss rate and jitter ratios. Based on the predicted KPI, intelligent routing and network traffic control can be implemented. Besides, the work in [65] proposes a joint route prediction and forwarding algorithm based on the global visibility of SDN central controller. This work employs imitation learning to train SDN controller to predict future state, based on the predicted network state, a stream access control strategy of stream re-transmission is further proposed to dynamic schedule potential congestion data stream.

The main difference between imitation learning and reinforcement learning is that imitation learning learns from expert behaviors without the redundant exploration process. Imitation learning can have a faster convergence speed and save both

computation and communication costs. However, imitation learning may fall into local optimal state and it is hard to catch the varying state in competitive tasks such as multi-agent competition game.

D. Reinforcement Learning for Distributed Network Routing

Recently, reinforcement learning is widely considered for the state-of-the-art solution of the distributed game without known global information. The multi-agent routing process can be modeled as a competitive game. Comparing with imi- tation learning, reinforcement learning with online explorationshows a better possibility to approximate the global optimal. To deal with the heavy traffic in the wireless network, a deep reinforcement learning based routing strategy is proposed in [66]. Instead of the online imitation learning based approach predicting the congestion and treat the judgment of conges- tion as a classification task, the work in this paper uses deep reinforcement learning to explore the potential next-hop selection. The authors also considered a value matrix in this work to measure the utility of actions. Different from theaction space of paths combination selection used in imita-tion based learning in [58], the authors in the work use the options of the next hop router to construct the action space. With reinforcement learning, the central controller continu- ously explores the actions (i.e., next-hop routers) and updates the Q-value in the value matrix of all actions. With online learning, the corresponding action is chosen based on the updated value matrix with ϵ *gready* algorithm. Comparing with the imitation based algorithm, The next hop selection based routing algorithm with reinforcement learning achieves less computation overhead and is suitable for larger-scale network.

However, the central controller based routing strategy is high communication cost and hard to adapt to the heteroge- neous network where the central control mechanism is infea- sible. For such a heterogeneous scenario, the authors in [67] propose a distributed Multi-agent Reinforcement Learning based routing strategy. In this work, instead of central control, the authors consider the network as a multi-agent competition game, the Multi-Agent Reinforcement Learning (MARL) [73] is leveraged to distribute the action decision task to each router. Without waiting for global information, the proposal achieves less communication overhead and fast response speed. A simi- lar approach is proposed in [74], this work models the routing problem of satellite network with a Stackelberg game, and then proposes a deep reinforcement Learning based routing algorithm for solving the multi-agent competitive game with near-optimal performance. In this work, the next node selec- tion is also considered as the action space, however, instead of the CNN used in the previous works, the time sequence sensitive LSTM is used to extract the features and build a relation between traffic patterns and actions. However, such an approach is high topology dependency which is not suitable for the topology varying environment such as IoT, Internet of Vehicles (IoV) and Space-air-ground integrated network (SAGIN).

Wang *et al.* try to conquer the dynamic moving problem of satellite communication and propose a deep reinforcement learning based state-aware routing algorithm for high mobility LEO Satellite network [68]. In this work, the link states of two-hops neighbors are continuously collected as state space in the input of the reinforcement learning system. Then, the next hop is decided with the double Q-learning algorithm. The training process is off-line processed in this work to save communica- tion and computing cost. However, there is an assumption that the dynamic moving satellites can always keep static virtual topology. Furthermore, it is hard for offline learning to catch the emergency and changing situation of complex networks.

All the previous reinforcement learning based routing algo- rithms all use the next-hop selection as the action space for learning. However, as the same concern to the imitation learn- ing based routing, the end-to-end performance is satisfying only when all hops are correctly chosen but is dropped when any hop is wrongly selected. That is to say, the wrongly selected one-hop in the whole end-to-end path of next-hop based routing chained influences the end-to-end QoS of the entire path. Some researchers think the whole path selection is the better choice for machine learning based routing in some high QoS required scenarios. The papers [69], [70] all con- sider the flow based whole end-to-end path routing as the action space for the reinforcement learning algorithm. TheNN and RNN are respectively used in the two works for fea- tures extraction. Both the end-to-end throughput and delay areproved to be improved for SDN enabled network.

The previous works all consider single-path transmission, the packets of a given flow take the same routing path which may cause path resource waste and high end-to-end delay. Therefore, the researchers propose the deep reinforce- ment learning can be employed for multi-path routing and achieves substantially lower latency than single-path routing. Xu *et al.* first consider the multi-path flow routing problemand propose a deep reinforcement learning based routing algorithm [72]. For the multi-path transmission problem, this work proposes a two-step model-free learning to explore flow splitting and choose the best flow routing paths. In the pro- posal, the reinforcement learning first does traffic engineering-aware exploration and records rewards correspondingly. The proportions allocated to all paths are calculated as a vector of continuous values. Then, actor-critic based prioritized experi- ence replay is off-line proposed to train the neural network to choose the best proportion vector in an online manner. another contribution of this work is that the authors consider continuous action control for learning, however, whether the continuous control is high computation cost and may not help the network performance improvement in some cases. The authors in [71] propose a SDN based direct flow routing algo- rithm for the multi-path transmission. The simple Q-learningis employed in this work to achieve a significant performance improvement than conventional SPF routing and single-path routing. An open source code is also provided in this work. However, the two proposals only consider the throughput and delay of seasons as state space do not well describe the traffic state and link condition of networks. Which might be inca- pable of the high dynamic network especially when frequent

TABLE II
THE FEATURES OF MACHINE LEARNING BASED NETWORK ROUTING ALGORITHMS

Learning type	Depth	Control type	Vary topology	Routing type	Value network	Action	Literature
Supervised learning	Shallow	Centralized	No	Single-path	NN	Path selection	[14]
Supervised learning	Shallow	Centralized	No	Single-path	NN	Path selection	[15]
Supervised learning	Shallow	Centralized	No	Single-path	NN	Path selection	[16]
Supervised learning	Deep	Distributed	No	Single-path	DNN	Next-hop node selection	[20], [53]
Supervised learning	Deep	Distributed	No	Single-path	DNN	Path selection	[55]
Supervised learning	Deep	Distributed	Yes	Single-path	GNN	Path selection	[54]
Supervised learning	Deep	Centralized	Yes	Single-path	DFN	Link prediction and path selection	[56]
Supervised learning	Deep	Centralized	Yes	Single-path	GCN	Path selection	[57]
Imitation learning	Deep	Centralized	No	Single-path	CNN	Path selection	[58], [59]
Imitation learning	Deep	Centralized	Yes	Single-path	3D-CNN	Next-hop node selection	[60]
Imitation learning	Deep	Centralized	No	Single-path	CNN	Path selection	[61]
Imitation learning	Deep	Centralized	No	Single-path	ELM	Path selection	[62]
Imitation learning	Deep	Centralized	Yes	Single-path	GNN	Path selection	[64]
Imitation learning	Deep	Centralized	Yes	Single-path	DNN	Next-hop node selection	[65]
Reinforcement learning	Deep	Centralized	No	Single-path	CNN	Next-hop node selection	[66]
Reinforcement learning	Deep	Distributed	No	Single-path	LSTM	Next-hop node selection	[67]
Reinforcement learning	Deep	Distributed	Yes	Single-path	DNN	Next-hop node selection	[68]
Reinforcement learning	Deep	Centralized	No	Single-path	NN	Path selection	[69]
Reinforcement learning	Deep	Centralized	No	Single-path	DNN	Path selection	[70]
Reinforcement learning	Shallow	Centralized	No	Multi-path	Q-table	Best multi-path selection	[50]
Reinforcement learning	Shallow	Centralized	No	Multi-path	Q-table	Division ratio for multi-path	[71]
Reinforcement learning	Deep	Centralized	No	Multi-path	DNN	Division ratio for multi-path	[72]

bursty traffic exists. Furthermore, the state-action space grows exponentially with the number of network nodes which may cause high computing overhead.

As we introduced above, imitation learning is quick but lacks full exploration, the reinforcement learning is the oppo- site. An interesting idea is to employ imitation learning integrated with reinforcement learning to balance the tradeoff between exploration and exploitation. The work in [75] pro- poses an integrated reinforcement learning algorithm for MECresource allocation, which shows accelerated convergence speed and superiority of learning accuracy.

The machine learning based routing algorithms from single network layer optimization to cross-layer cooperative optimization, from the single path to multiple-paths, shows step-to-step evolution. One potential research direction is to combine the machine learning based routing with congestion control in the transportation layer. In the next section, we detailed introduce machine learning based intelligent end-to-end network congestion control.

E. Summary and Discussion

In recent years, machine learning based routing algorithms are widely proposed and show significant improvement compared with conventional routing algorithms. The routing strategy is developed from single-path to multi-path, from single layer to cross-layer design, and shows great potential for future deployment in the real world. At the same time that related technologies have been rapidly developed, one big issue for the machine learning based routing algorithm appears in front of everyone. What is the unified performance measurement standard for the machine learning based routing algorithm? For example, the performance of supervised learning based routing can be measured with the training accuracy and prediction accuracy, and the performance of reinforcement learning based routing can be measured with accumulated reward and convergence speed. However, the training accuracy, reward, delay, and throughput can be different when the deployed network environment such as topology and traffic generation model is changed. To build a unified performance measurement standard recognized by most researchers is necessary.

IV. MACHINE LEARNING FOR END-TO-END NETWORK CONGESTION CONTROL

The retransmission mechanism is critical for reliable end- to-end data transmission. However, frequent retransmission causes high packet overhead and congestion. Smaller transmission window size reduces retransmission frequency and leads to less congestion however affects the end-to-end throughput and delay. The automatic repeat request (ARQ) one of the most reliable insurance mechanisms in transmission control protocol (TCP) to adaptively change the transmit window size and improve the end-to-end transmission QoS. The main idea of the ARQ is that all transmitters reduce window size when congestion happens, and increase the window size when the transmission is smooth. However, how to judge the congestion state and when to change the window size is always a critical challenge for ARQ designers.

To address the challenge, there are many classical congestion control algorithms for dynamic ARQ. Reno is the classical and most used ARQ algorithm in TCP [76], which guesses a reasonable initial window size and quickly increase the window size when packets loss not happen. When Reno received any packet loss notification, it slowly decreases window size to avoid congestion and ensure reasonable end-to-end

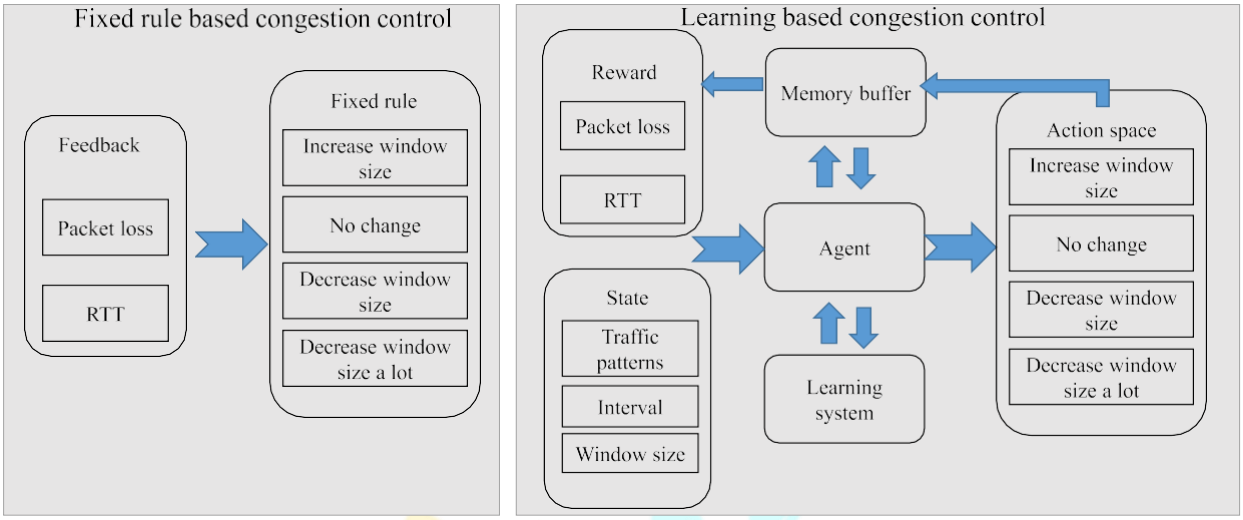


Fig. 7. The difference between the machine learning based congestion control and conventional methods.

throughput. However, packet loss may not be the best feature of congestion. Another classical ARQ algorithm called TCP-Vegas is proposed to dynamically change window size based on the packets delay instead of packets loss [7]. The TCP-Vegas is shown to be more efficient in some scenar-ios. Considering both delay and packet loss as the feedback, the work called TCP-Illinois proposes a joint loss and delay based congestion control algorithm [77]. And some new algo- rithm considering more features such as packets loss reason judgment [78], transmission flow classification, wireless link state [79] and direction of target rate [80]. An enhanced ver- sion of Reno referred to as new-Reno [81] is widely consideredas the state-of-the-art algorithm within conventional TCP algo- rithms. However, the reason behinds the feedback including packet loss and long RTT is complex in the practical com-plex network. For example, the packet loss can be caused by self-errors, congestion with other flows, insufficient buffer andother random loss. All those conventional algorithms usingthe fixed rules to address the adaptive TCP problem can not capture the reasons in the complex network. Besides, noneof them considers the possible traffic changing trend and intelligent reflection mechanism in more generic scenarios.

A. Classical Shallow ML Based Congestion Control

As shown in Fig. 7, the conventional methods make a decision based on the feedback and fixed rules, but the learning based congestion control will intelligently choose actions based on the updated pair of network state and reward. Thus, the congestion control agent can adapt to the changing network state based on learned experience. The initial attempt of using machine learning for transmission congestion control is from 20 years ago [82]. The main idea of this work is to adaptive change transmit window size based on the reasonable judgment of packets loss (similar to the work in [78]). Based on the different reasons for packet loss (e.g., queuing packets drop or link error broke), the TCP chooses a different window size changing strategy. The various machine learning algorithms including decision trees, Decision tree ensembles, Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN) are employed for the packet loss classification. With simulation, the decision tree boosting based TCP is turned out as the optimal solution and shows more almost 300% performance comparing with conventional TCP-Veno. However, machine learning is only used for state classification, the transmission windows size scheduling is still based on conventional strategy. Furthermore, this work only considers the performance of the transmitter which is selfish and not fair for other nodes in the network.

The work in [83] models the packets sending process as a cooperative game. Considering the cooperation of nodes, machine learning is employed to map the memory to an optimal action (i.e., window size change). This work further improves fairness for all nodes in the network and intelligent make decision just based on the learning algorithm. Nevertheless, this proposed learning algorithm is a simple mapping strategy,no specific machine learning technology is used.

The previous work modeled the congestion control with game theory, which is naturally considered to be solved with reinforcement learning. The initial reinforcement learn- ing based TCP is proposed in 2016 [84]. The work proposesa Q-learning based congestion control algorithm. Four state variables including inter-arrival time of sending packets, inter-arrival time of received acknowledgment (ACK), the ratioof current round trip time (RTT) and best RTT, slow start threshold, and the values of state variables are partitioned by discretization. The reward function is calculated based on the weighted sum of end-to-end delay and throughput. With simulation, the work shows better end-to-end network performance than the classical new-Reno algorithm. Furthermore, it demonstrates how memory plays a critical role in building the exploration space of reinforcement learning, and proposes a way with the FUZZY KANERVA approximation algorithm to accelerate the convergence process and reduce this memory overhead.

Kong *et al.* proposes both supervised learning and reinforcement learning based TCP congestion control algorithms

to conquer the optimal feedback-action mapping problem in complex network [85]. In this work, random forests based supervised learning is used to predict the packets loss based on the feedback state. The state characterization of feedback is also including four variables as the same as the work in [84]. The window size is scheduled based on the predicted packet loss. Then, an on-policy SARSA based reinforcement learn- ing is proposed to map the relation between the feedback and action (i.e., window size change). From the simulation for bothsingle sender and multi-senders, the proposed two algorithms both outperform the conventional new-Reno.

Besides static network, the work in [86] further considering the high mobility of nodes and the varying network topology, proposes a centralized and localized unsupervised learning based congestion control algorithm for dynamic vehicular ad-hoc networks. In this work, the packets are clustered with an unsupervised K-means algorithm based on the message size, the validity of messages, and type of messages. Then, each cluster header correspondingly changes the value of transmis-sion range and rate, contention window size, and arbitration inter-frame spacing in terms of different cluster types.

The reinforcement learning based congestion control algo-rithm is mostly based on the global vision with a central controller. However, which is not practical for heterogeneous networks and causes high signaling overhead for the large-scale network. The work in [9] proposed a totally distributed online learning based congestion control algorithm called PCC-Vivace. The proposal is an imitation learning based con-gestion control algorithm, based on the powerful optimization ability of machine learning, the learning system continuously exploits action and calculates the state-action based utility of the next time slot. The utility function is mainly based on the feedback including RTT and end-to-end loss rate, which decides whether the exploited action is expert behavior or not. The utility reflects network state and is collected in sender-side. Thus, the proposed algorithm is totally distributed and is friend with current TCP. Then, an improved algorithm called PCC Proteus is further proposed to consider the different primary of traffic flows [87]. However, those imitation based congestion control algorithm modeled the network as a non-cooperative game and selfish to maximize its utility, which brings unfairness to other nodes and causes unexpected QoS decrease for users in the network. Furthermore, it is hard for the shallow learning structure to handle high dimension data. Thus the shallow learning structure is not applicable for the large-scale network.

B. Deep Learning Based Single-Path Congestion Control

A recently attractive imitation learning based congestion control algorithm is called Indigo [88]. Comparing with previous works, this work employs the deep learning struc- ture of LSTM to predict the state-action utility. The proposal with a higher dimension capacity can handle large scale data. However, this training process is offline and the mapping is fixed after the offline training is finished. Thus, this proposal may be not suitable for the dynamic network especially the varying topology network.Another direction is to deepen the value network of rein- forcement learning. The authors in [89] leverage deep rein- forcement learning to capture intricate patterns in data traffic and network conditions and train the policies of network con- gestion control actions. Compared with the previous PCC family protocols [9], [87]. The proposal further expends the state of the input to include three main features of latency gradient, latency ratio and sending ratio, and the reward func- tion is constructed with a linear combination of delay and packet loss. With simulation, the proposal with deep structure shows better performance in terms of both throughput and delay compared with the previous shallow one.

Another advantage of reinforcement learning compared with supervised and imitation learning is fairness. Inspired from the previous work [83]. Li *et al.* propose a deep Q-learning based TCP congestion control to adapt to flexible scenarios [90]. In the same year, another group proposed a deep reinforcement learning with Asynchronous Advantage Actor Critic (A3C) for dynamic congestion control [91]. Comparing with previous works that only process the congestion window adaption, this work separately schedules the suitable (initial window) IW for short flows through group-based RL, and dynamically con-figures a suitable Congestion Control (CC) scheme for long flows. However, this work mainly focuses on Web service optimization which is not tested for applicability for generic networks. The work in [92] proposes a centralized deep learn-ing based IW scheduling algorithm for MEC scenarios. This work focuses on short flow congestion control and improves the flow completion time (FCT).

C. Deep Learning Based Multi-Path Congestion Control

The Multipath TCP (MPTCP) is a testing transmission con-trol standard allowing a TCP connection to use multiple paths to transmit packets. The MPTCP is processed based on the multi-path routing introduced in the previous section, which carries new challenges to the congestion control algorithm. However, conventional MPTCP is mostly based on fixed rules or optimized with the current state. Those conventional meth-ods are hard to be considered to handle the congestion control in the high complex network with complex rules and state change.

To adapt the MPTCP to modern and future complex network, the work in [93] proposes an experience-driven deep reinforcement learning based MPTCP congestion con- trol algorithm. In addition to the single congestion window changing of previous reinforcement learning based algorithms, the proposal synchronously change all congestion window of all flows for multi-paths. This work shows better network performance than conventional methods and is flexible to highly-dynamic networks. Another work in [94] proposes reinforcement Learning based MPTCP congestion control in heterogeneous networks. In addition to the states of flow,the proposal also includes the states of routers as input. Furthermore, the author in this work proposes a novel func- tion estimation algorithm to improve Q-learning efficiency and improve the control convergence speed. This proposal out-performs conventional MPTCP mainly in terms of aggregate throughput. However, one concern is that the collection of

the information of routers may cause unexpected signaling overhead. A centralized reinforcement learning based MPTCP congestion control algorithm is proposed depending on the efficient SDN central controller [95].

Besides on the short-range ground network, the work of [96] claims that the reinforcement learning can also be implemented to dynamically control multi-path congestion for satellite network. However, this work is very initial for considering the distinctive features of the satellite network. Compared with ground networks, the satellite network becomes more sensitive to packet loss. Besides, the satellite links are highly flexible which should be carefully considered for the congestion control algorithm design. An improvement of deep learning based congestion control can be further considered by the researchers in the future.

D. Summary and Discussion

In this section, we introduce the applications of machine learning for network congestion control dealing with definite problems. Compared with conventional algorithm, the machine learning based congestion control algorithm shows improvement for both single and multi-path communication. Thus, machine learning based congestion control algorithm has been attempted to be deployed in both academic and industrial environments recently. The most popular congestion algorithms are mainly based on reinforcement learning. However, as far as we knows, the imitation learning based algorithm is more efficient for some cases with small data set, clear requirements and lightweight application.

V. Machine Learning in Application Layer for End-to-End Adaptive Streaming Control

QoS measures the service quality mainly related to the network itself does not satisfy the new requirement of users in the next-generation network. Instead, the QoE is emerged to measure the service from users' perspective with parameters including efficiency, fairness, stability and so on.

The end-to-end streaming control to ensure the QoE ofusers emerges as the new challenge for both network and service provider. The QoE optimization with streaming con- trol is deployed in both the server and the client side, which reaches a high-level OSI to the application layer. The adaptive bitrate streaming (ABR) is a typical traffic control technology for end-to-end streaming multimedia QoE guarantee. The mainidea of ABR algorithm is to dynamically adjust the quality (i.e., sending rate of the streaming) of media to the real-time changing network resources.

The Conventional ABR algorithms are based on fixed designed control rules. The rules can be categorized into three types. One is the estimated network throughput based ABR, for example, the work in [8] designs a bandwidth estimation based ABR algorithm considering players sharing the same battle link. The work in [107] proposes a throughput predictionbased ABR algorithm. The throughput based ABR algorithms improve the QoE of users including efficiency, fairness, and stability. Meanwhile, another type of buffer based ABR algorithms is proposed. Different from the throughput based ABR, the buffer based ABRs such as in [108] and [109] design the ABR for the scenarios where the throughput is highly dynamic and varies frequently, it makes the network state hard to be estimated accurately. In those buffer-based proposals, the ABR is scheduled based only on the playback buffer occupancy. The simulations show that the buffer-based ABRs outperform throughput based ABRs in the frequent throughput varying environment. The third type of conventional ABR algorithm is the hybrid ABRs such as the work in [110] schedule the ABR based on the combination of parameters including buffer and throughput. The most consensus state-of-the-art ABR algorithm is called MPC [111], which employs a combination of various features of future chunks as the feedback and shows superiority in terms of both delay and fairness.

Although the state-of-the-art conventional ABR shows good performance for some scenarios, two shortcomings make the conventional ABR algorithms hard to handling the high QoE requirement in the next-generation network in 6G. Firstly, the fixed rule based algorithms need significant tuning over a horizon of feedback. The tuning makes the proposals over-fitting for the specific scenario and hard to be adapted to the generic scenarios. Secondly, the conventional ABRs using fixed heuristics depending on the modeling of the network, the estimation of the future state of the network based on the assumed model significantly affects the ABR performance. Therefore, when the network environment changes, the simplified or inaccurate model no longer adapt to the new environment and leads to inevitably performance loss of the ABR algorithm.

A. Shallow Reinforcement Learning Based ABR Algorithms

In order to overcome the shortcomings of conventional ABR algorithms and further improve the QoE of users, the machine learning based ABR algorithms is proposed. Petrangeli *et al.* first propose a multi-agent Q-Learning based ABR algorithm to achieve fairness for users [97]. In this proposal, the multi- agent ABR is modeled as a stochastic game, and the q-learning is employed to update the reward continuously. Comparing with conventional ABR algorithms, the proposal schedule ABR without the requirement of global information, thus achieve fairness with less signaling overhead. Furthermore, the proposed deep learning based ABR learn and adapt the ABR policy from varying network condition which can adjust to generic dynamic networks.

Besides the game theory based model, the work in [98] models the state changing process of the network as a Markov decision process (MDP). Considering the high mobil- ity vehicular network, this work explores more advanced decision-making tools to improve the QoE balance ability and achieves better network performance comparing with conven- tional algorithms. Another reinforcement learning based ABR is proposed in [100]. In this proposal, a concept of Post-Decision States (PDSs) is proposed as an intermediate stateto divide the transition from the current state to the next step into two steps. Then, the ABR is scheduled based on the estimated value of PDS, which can be evaluated by the employed Temporal Difference (TD) learning.

Previous works design the reward function mainly focus on the network state of the next time slot, which may not

TABLE III
THE LIST OF ML BASED CONGESTION CONTROL ALGORITHMS

Year	Learning type	Depth	Name	Learning Method	Input	Output	Literature
2006	Supervised	Shallow	None	Decision trees, ANN, KNN	Traffic patterns	Congestion window changing	[82]
2013	Supervised	Shallow	TCP-Remy	Simple mapping strategy	Interval between sending packets, interval between receiving ACKs, average RTT	Congestion window changing	[83]
2016	Reinforcement learning	Shallow	TCP-Learning	Q-learning	Delay of received ACKs, delay of sending pakcets, ratio of RTT and best RTT, slow start threshold	Congestion window changing	[84]
2016	unSupervised	Shallow	ML-CC	K-means	Message size, validity of messages, and type of messages	Transmission range and rate, contention window size, and arbitration interframe spacing	[86]
2018	Supervised/ Reinforcement learning	Shallow	LP-TCP	Random forests, Q-learning	Delay of received ACKs, delay of sending pakcets, ratio of RTT and best RTT, slow start threshold, window size	Congestion window changing	[85]
2018	Imitation learning	Shallow	PCC-Vivace	Utility Gradients	Sending rate, observed loss rate, RTT gradient	Congestion window changing	[9]
2020	Imitation learning	Shallow	PCC-Proteus	Utility Gradients	×	Congestion window changing	[87]
2018	Imitation learning	Deep	Indigo	Q-learning	×	Congestion window changing	[88]
2019	Reinforcement learning	Deep	Aurora	Q-learning	Latency gradient, latency ratio, sending ratio	Sending rate	[89]
2019	Reinforcement learning	Deep	QTCP	Q-learning	Interval between sending packets, interval between receiving ACKs, average RTT	Congestion window changing	[90]
2019	Reinforcement learning	Deep	TCP-RL	A3C with LSTM	Throughput, RTT, packets loss	Initial window, congestion window changing	[91]
2019	Reinforcement learning	Deep	DRL-CC	Actor-Critic with LSTM	Sending rate, goodput, average RTT, mean deviation of RTTs and congestion window size of all flows	Congestion window changing of all flows	[93]
2020	Supervised learning	Deep	NeuroIW	CNN	Flow completion time, inter-arrival time, inter-departure time, and response size	Initial window changing	[92]
2019	Reinforcement learning	Deep	SmartCC	Q-learning	Data rate, delay, and available bandwidth of sub-routes, interval between receiving ACKs and sending rate of flows	Congestion window changing of all flows	[94]
2019	Reinforcement learning	Deep	NeuroIW	A3C with CNN	Flow completion time, inter-arrival time, inter-departure time, and response size.	Initial window changing	[92]
2020	Reinforcement learning	Deep	FNDRL-CC	Actor-Critic with LSTM	Congestion window size, mean deviation of RTT, average RTT, goodput and sending rate	Congestion window changing of all flows	[95]
2019	Reinforcement learning	Deep	×	A3C with CNN	Sending packets ratio; RTT; packet loss rate; cumulative rate number of retransmissions	Congestion window changing	[96]

accurately measure the QoE level of users. The work in [99] split the resulting stream into small segments and design the reward function by introducing more features including favoring the selection of higher qualities, the amplitude of the quality variation and penalization factor of video freezes to measure the satisfying level of QoE of users.

The use of shallow reinforcement learning for ABR achieves reasonable performance improvement comparing with conventional algorithms. It paves the way for using more powerful machine learning tools such as deep learning for further improving the QoE of the ABR algorithm.

B. Deep Reinforcement Learning Based ABR Algorithms

In this part, we introduce the most recently proposed deep reinforcement learning based ABR algorithm. Comparedwith conventional algorithms, all of those deep learning based proposals show better performance than the state-of-art MPC [111].

The first proposal of using deep reinforcement learning to ABR called Pensieve is proposed by Mao *et al.* in [101]. Comparing with conventional ABR algorithms, the Pensieve utilizes DRL to select bitrate for next video chunks, does not rely on any pre-programmed models or assumptions about the environment. That means the proposal is generic to various networks. And compared with previous shallow reinforcement learning, in this work, a deep CNN is employed for extracting the features of the input, and the actor-critic algorithm is proposed to polish the value network and select the policy correspondingly, which is considered as the state-of-the-art learning-based ABR scheme due to the excellent performance.

However, one question arises that the reinforcement learning based ABR mostly based on the measurement of QoE. The level of QoE is always characterized as a simple metric to reflect the optimization objective. It is hard to guide the optimization in different scenarios only with the simple metric of QoE, as the QoE can be changed due to different requirement of users. To solve this, the work called Tiyuntsong is proposed in [104] of using a generative adversarial network

TABLE IV
THE LIST OF ML BASED ADAPTIVE STREAMING CONTROL METHODS

Learning type	Value network	Depth	Learning method	Input	Reward	Action	Literature
RL	Q-table	Shallow	Q-learning	Local perceived bandwidth, the buffer filling level	Estimated total future quality level	Changing quality level	2014-[97]
RL	Q-table	Shallow	Q-learning	Buffer size of agent	Expiration probability of next chunk	Changing quality level	2015-[98]
RL	Q-table	Shallow	Q-learning	Buffer size of agent, current quality level, estimated bandwidth	Quality variation amplitude and video freezes penalization	Changing quality level	2016-[99]
RL	PDSs-table	Shallow	TD-learning	Estimated bandwidth, the complexity of next content, current buffer state	Weighted sum of quality reward and buffer reward	Choose best representation to download	2016-[100]
RL	CNN	Deep	A3C	Network throughput, past download time, available next chunk sizes, current buffer level, last chunks number and bitrate	According to QoS requirement	Changing bitrate	2017-Pensieve [101]
IL	Decision tree	Shallow	Gradient descent	Rebuffer, past download time	Square loss of max/min bitrates	Changing bitrate	2019-PiTree [102]
IL	CNN	Deep	Gradient descent	Network throughput, chunk size, past download time, perceptual video quality-buffer occupancy	Future RTT, video render time, buffer size occupancy	Probability of the bitrate being selected	2019-Comyco [103]
RL	CNN, LSTM	Deep	GAN	Network throughput, past download time, previous bitrate, remaining playback time, buffer length, next video sizes	Win rate of two agents for each epoch	Probability distribution of bitrate candidate	2019-Tiyuntsong [104]
RL	LSTM	Deep	Q-learning	Network throughputnext time-slot file sizes, download time, buffer occupancy, last bitrate, remaining chunks number	Weighted mean quality	Choice of quality level	2019-DeepVR [105]
RL	NONE	Deep	Combination	Quality of previously video, estimated bandwidth, complexity of next video, buffer length	User QoE	Changing bitrate	2019-IAMS [106]

(GAN) to dynamically measure the QoE with two agents competition with each other. This work first considers the multi-objective optimization of ABR and uses the actor-critic method to update the value network for the agent. Compared with the previous works getting training samples with the communicated signal, the proposal only trains the samples generated by the agent itself. Thus, in addition to the better measurement of QoE, the proposal can explore and train efficiently and saves signaling overhead.

Considering the fine-tune speed of the deep reinforcement learning is slow and high computation cost, the proposal in [105] improves the traditional one-step deep Q-learningto multi-step learning to improve the learning efficiency. In this work, the agents make decisions for several steps and multiple parameters, which speeds up the training process and saves computation resources. To adapt the multi-step learning,a time-sequence sensitive LSTM is used as the value network to predict the field of view (FoV) of the user in the nextfew seconds. In addition to the simulation, this work imple- ments a prototype of the proposed algorithm to evaluate its effectiveness.

Although the learning based ABR algorithms outperform the conventional methods in many scenarios, however, the simpleness and the interpretability make the conventional algorithms more efficient in some specific cases. In order to allocate the best efficiency of both conventional and learning based ABR, the authors in paper [106] proposes an ensemble framework to dynamic switch ABRs adapting to different environments. The proposed InstAnt Method Switching (IAMS) and InterMittent Method Switching (IMMS) based framework is shown to be simple yet very effective, which exploits method pool including many candidate ABR methods such as throughput based conventional ABR algorithm [112], control theory based conventional ABR algorithm [113] and previously introduced online learning based ABR algorithm [100]. Based on the dynamic choosing process of the best ABR algorithm from the method pool, the proposal shows a better QoE level in various scenarios.

C. Online Imitation Learning Based ABR Algorithms

As illustrated in the routing part, efficient reinforcement learning relying on the large-scale exploring process is high computation and time cost. Imitation learning which sim- ply explores several policies and follows the evaluated expert behaviors with faster convergence speed is considered to be another learning approach for ABR. Meng *et al.* at first propose a decision tree based ABR algorithm [102].The authors in the work consider the scenario where the complex deep learning based ABR is too heavyweight for client devices with less power. The proposal employs lightweight machine learning structure of decision trees to deploy the ABR into practical devices for saving power. However, with the save of expense, the proposal shows slight performance degradation.

Compared with the above reinforcement learning based algorithms, more efficient learning based ABR called Comyco is proposed by Huang *et al.* in [103]. The proposed video quality-aware ABR algorithm enormously improves the learning efficiency by training the policy via imitating expert trajectories given by a designed instant solver. With imitation learning, the proposal avoids redundant exploration and makes better use of the collected samples. The simulation shows

that the proposal can achieve equal performance but better convergence speed to reinforcement learning based ABRs.

D. Summary and Discussion

The machine learning based ABR mainly includes two directions. One uses reinforcement learning and the other use imitation learning. Both two types of learning can effi- ciently learn the past experiment and alleviate the failure and congestion ratio during end-to-end transmission. For ABR,the machine learning system enables more knowledge about future available transmission window size and connection sta- tus, which enables necessary configurations towards improved transmission efficiency. However, the existing machine learning base ABR is mostly designed for a specific application, which is trained with distinctive data and is not efficient to transfer to other applications. A convenient transfer system or federated learning platform might be a potential solution to covering various applications.

VI. Open Research Issues and Future Directions

Machine learning brings new research directions to the network. Although there are many superior works emerges recently, there are still many challenges that lead to future research directions. In this section, we summarise some potential open research issues and future directions.

A. The Application Aware QoE Guarantee

The current machine learning based network functions including access, routing and congestion control mainly focus on the local layer optimization. However, different users rely on various application requires totally different QoE. For example, video streaming based applications require high throughput and low delay but less security. Meanwhile, the payment software requires high security but relatively lower throughput. Therefore, to design a cross-layer protocol and do an action based on the requirement of different applications is an interesting direction to satisfy the various requirement while balancing network resources. To ensure such a cross-layer design, one big challenge is the reasonable feedback control from the application layer, as the feedback from the application layer is always slower than the lower layers [114]. One solution to the quick cross-layer design is to employ machine learning to predict the possible application of nodes based on the historical patterns and intelligently allocate access, routing and congestion strategy by cross-layer cooperation.

B. Feasibility of Learning Based Algorithm

One big challenge of the network functions is the feasibility of machine learning based algorithm. To build a unique learning system that can handle all tasks in different scenarios is quite hard under the current computation limitation. One solution to the feasible deployment for dealing with different tasks of various scenarios is to transfer the trained learning struc- ture to the new scenario by storing some knowledge. Suchkind of learning transfer process is called transfer learning. The transfer learning of reusing the learned knowledge can significantly improve the learning efficiency and deal with the situation lacks training samples [115]. For example, the learning system trained for channel allocation can be transferred to the tasks of routing, congestion control as the features of links state and traffic patterns is overlapping and related. However, the obvious ties between the two scenarios might be limited, how to choose the reasonable features and design a proper transfer learning structure to improve the transfer efficiency isstill an open research issue.

Another direction to enable the feasibility for machine learning based network functions is the meta-learning, which is also referred to as learning to learn [116]. One handicap for the feasible deployment of learning based network system is the distinctive inductive bias of training data of different scenarios [117]. With meta-learning which learned to change the learning structure and parameters by itself, the learning based system can intelligently adjust the changing environment. How to efficiently deploy the meta-learning to the network system and embed it to existing learning algorithms should be an interesting research direction.

C. Distributed and Dynamic Learning

As we introduced in the previous section, both the centralized and distributed control based learning systems are widely researched for network end-to-end service guarantee. Due to the highly complex and heterogeneous structure of future generation network, the distributed control based learning system attract more attention recently. However, the total distributed learning without the gradients sharing is less of universality and sometimes is trucked in an over-fitting state. On the other hand, the information sharing based distributed learning faces challenges of high communication overhead, security and privacy issue. Federated learning is widely researched for efficient distributed learning considering information sharing delay, learning security and user privacy [118]. Both the horizontal transfer learning and vertical federated learning canbe deployed [119]. For example, for the congestion control problem, each router can train the routing policy based on local information and share the updated gradient. Such kind of horizontal transfer learning can significantly reduce communication overhead of training data sharing and protect the privacy of users (no unencrypted content sharing) [120]. Besides, for the scenario two training data-sets sharing the same sample ID space but differ in feature space, such as the distributed routing problem, the router can employ the vertical federated learning to federated learn one piece of training sample including features of link state from the router and other pieces of samples including features of traffic flow features from end-clients. As the knowledge as we know, there are still no vertical federated learning related works are published for network end-to-end optimization. Furthermore, federated learning can be combined with transfer learning for more complex tasks [118]. For example, the transfer federated learning can train both the routing and congestion control policy by collecting samples from both routers and clients in distributed manner.

D. Intelligence in Full Space

The high-quality end-to-end services are guaranteed with the increase of network resources and intelligent resource allocation. However, both the network resource and cov- erage are limited to terrestrial communication. To better utilize the space resources, the full space network such as Space-air-ground (SAGIN) and SAGIN-underwater are widelyproposed as the potential solution to better resource utiliza- tion in next-generation network [121]. Naturally, to enablethe intelligence for the full space network should be a mean- ingful future research direction. The machine learning based resource allocation and traffic control are also been researchedrecently [122], [123]. Due to the heterogeneous structure ofthe full space network, it is hard for the end-to-end QoS to be ensured in a centralized manner. However, as is men- tioned above, the distributed learning system faces many issues including information sharing and security. Furthermore, there are still two distinctive issues that should be carefully consid-ered in the full space network. The first one is the scarce energy resources of the space and underwater nodes (e.g., satellite with polar power and underwater vehicles with lim- ited battery). The second one is the super high mobility of space nodes (e.g., satellite, planes and UAVs are continuously moving), which leads to frequent changing of network topol- ogy and varying link states. To ensure the stable quality of end-to-end service, the researchers should take those serious issues into consideration for future machine learning based network optimization algorithms.

E. The Open Resource Platform

Unlike the famous Computer Vision (CV) and Natural Language Processing (NLP) research areas, where the open source codes are widely publicized and several common data sets such as MNIST, are recognized as the benchmark [124], it is hard to find the open resource code in the machine learning based network optimization area due to many issues such as business conflict of interest and deployment difficulty. Recently, a common simulation platform of machine learn-ing based networking is proposed in [125] called as ns3-gym, which is combined with network simulator ns3 [126] and machine learning tools of OpenAI Gym [127]. Meanwhile, some open source code in network routing, congestion control and APR are released recently [71], [88], [104]. However, one problem of the NS3 simulator based platform is the simula- tion speed is hard to support a large scale network (a complete simulation may cause days or weeks). Another problem is that the real traffic trace is hard to be built with simulator. Without a commonly recognized date set, the training performance has always been doubted.

VII. Conclusion

The next-generation network with ultra-reliable and low-latency communications (uRLLCs) brings a high requirement of the end-to-end QoS and QoE. However, it is hard for the conventional end-to-end QoS and QoE aware network ptimization methods to handle the complex and high dynamic scenarios of the next-generation network. In this paper,we indicated the machine learning based end-to-end aware network optimization approach and investigated the related works including machine learning based network access, routing, congestion control and streaming adaption. In each section, we have listed the features of the surveyed works and analyzed the advantages and disadvantages. In the last part, the open research issues and potential future directions are summarized.