When Wireless Video Streaming Meets AI: A Deep Learning Approach

Lu Liu, Han Hu, Yong Luo, and Yonggang Wen

Abstract

Wireless multimedia big data contains valuable information on users' behavior, content characteristics and network dynamics, which can drive system design and optimization. The fundamental issue is how to mine data intelligence and further incorporate them into wireless multimedia systems. Motivated by the success of deep learning, in this work we propose and present an integration of wireless multimedia systems and deep learning. We start with decomposing a wireless multimedia system into three components, including end-users, network environment, and servers, and present several potential topics to embrace deep learning techniques. After that, we present deep learning based QoS/QoE prediction and bitrate adjustment as two case-studies. In the former case, we present an end-to-end and unified framework that consists of three phases, including data preprocessing, representation learning, and prediction. It achieves significant performance improvement in comparison to the best baseline algorithm (88 percent vs. 80 percent). In the latter case, we present a deep reinforcement learning based framework for bitrate adjustment. Evaluating the performance with a real wireless dataset, we show that the perceived video QoE average bitrate, rebuffering time and bitrate variation can be improved significantly.

Introduction

Motivated by the success of various artificial intelligence applications, their milestone technique, that is, deep learning, has been considered to be an effective means to alleviate the ever-increasing pressure of wireless video traffic. Globally, IP video traffic will grow four-fold from 2017 to 2022, and account for 82 percent of all IP traffic by 2022, up from 75 percent in 2017. The explosive trend is caused by several changes, including the growth of video viewing on mobile devices, the video definition shift from standard-definition to ultra-high-definition (UHD) [1], and abundant short-form video and long-form video applications [2]. This massive data contains valuable information of human-beings and wireless systems, and provides various potentials for data-driven system design and optimization [3]. Due to the great success of deep learning in other fields, the design of wireless multimedia systems based on deep learning has attracted attention from both industry and academia. In particular, we can mine the information propagation pattern based on social network analysis and further utilize the insights for video prefetching or dissemination [4]; QoS/QoE modeling based on context and user interest mining can guide the encoding strategy design.

Applying deep learning into the design and optimization of wireless multimedia systems faces two fundamental challenges. First, the unique characteristics of wireless multimedia data embarrass data analysis and intelligence mining. In particular, wireless multimedia data is usually unstructured, multi-modal and heterogeneous. The status information of multimedia systems and network environment is organized into structured time-sequential format. Video contents are unstructured, including meta-data, audio and video information. Multimedia consumers related information, such as emotions, interests and social relationships, is a mixture of structured and un-structured data, which is difficult to represent. Existing deep learning methods perform well for single-domain data, for example, computer vision and natural language processing, but not for wireless multimedia communication. Second, wireless multimedia systems, consisting of end-users, outlets, wireless network environments, and server systems, are complex and evolving rapidly. Multimedia communication techniques include 4G/5G, software defined networking (SDN), information centric networking (ICN) [5], network function virtualization (NFV), and so on. Multimedia applications cover interactive video streaming, ultra-definition videos, cloud gaming, augmented reality and virtual reality (AR/VR), and so on. Multimedia outlets shift from PC to tablet PCs or smart phones. Distinctive techniques request customized learning and optimization methods, and different applications/outlets correspond to the fast changing access pattern. Incorporating deep learning into multimedia systems should be in an online-manner.

There exist some efforts to tackle the aforementioned two challenges, which can be categorized into machine learning based methods and model based methods. For the former, the traditional machine learning methods often rely on hand-crafted features based on prior knowledge and domain expertise, which are designed for specific datasets, and are difficult to extend to other scenarios directly. To address the drawback, deep learning based methods utilize multi-layer neural networks for feature representation and classification (or regression). However, some multimedia data share

This research is supported in part by National Natural Science Foundation of China (NSFC) under No. 61971457, and Youth Program of the National Social Science Fund of China under No.16CXW008.

Lu Liu is with Sichuan Normal University; she is also with Fudan University; Han Hu (corresponding author) is with the Beijing Institute of Technology, China; Yong Luo and Yonggang Wen are with Nanyang Technological University, Singapore.

Digital Object Identifier: 10.1109/MWC.001.1900220

1536-1284/19/$25.00 © 2019 IEEE

Deep neural networks, inspired by the working principle of the human brain, are a set of algorithms (e.g., convolutional neural network, long short-term memory, and deep reinforcement learning) that are used to recognize patterns, including clustering, classification, and regression. They can model complex trends and detect non-linear relationships among input data.

significant similarities in features and processing methods. Various models develop separate frameworks for feature engineering and training, neglecting these similarities. They often lead to inefficiency in the model training and developing process. For the latter, existing methods also utilize a given dataset to model the running mechanism of wireless multimedia systems or build a simulator, and then devise various optimization algorithms. Due to the fast evolving characteristic of multimedia systems, the well-characterized simulators or models in the experimental environment inevitably differ from target systems. This bias between simulator and real systems presents an obstacle to any model based method.

To further bridge the performance gap, we jointly study deep learning based data analysis and system optimization for wireless multimedia systems. First, we decompose a wireless multimedia system into three components, including end-users, network environment, and servers, and present several potential topics to embrace deep learning techniques. After that, we present deep learning based QoS/QoE prediction and bitrate adjustment as two case-studies. In the first case, we present an end-to-end and unified framework that consists of three phases, including data pre-processing, representation learning, and prediction. Our framework supplies a complete pipeline and is applicable for a variety of datasets. In the second case, we present a deep reinforcement learning (DRL) based framework for video bitrate adjustment. Our proposed methods in two examples can be extended to other topics. For example, the QoE prediction framework can be utilized to analyze user access patterns and network status monitoring, and the DRL framework can be applicable for content prefetching or caching.

The rest of the article is organized as follows. The following section introduces related work for deep learning and its applications in wireless video communications. Then we provide an end-to-end framework and discuss potential topics to embrace the deep learning technique. Following that we show a novel framework to predict video QoS/QoE. Then we develop a DRL framework for bitrate adjustment in wireless video streaming. The final section concludes this article and discusses future work.

Overview of Deep Neural Network Based Video Streaming

This section presents a brief introduction to adaptive video streaming and deep neural network based video streaming.

Adaptive Video Streaming

HTTP-based adaptive streaming (HAS) is the de-facto standard for video streaming, and has been widely supported by many solutions, including MPEG-DASH, Adobe HTTP Dynamic Streaming, and Apple HTTP Live Streaming. Under such a framework, each video is divided into a set of chunks with equal length (e.g., 2–5 seconds). Each of these chunks is then encoded into multiple quality levels corresponding to different bitrate and resolution. Clients can dynamically download chunks with the appropriate bitrate and resolution according to the network conditions and outlet specifics to guarantee the desired quality requirement.

Many previous efforts focus on bitrate adaptation to improve the QoS/QoE. In particular, the most widely adopted objective QoS/QoE metric is in the weighted summation form of startup delay, rebuffering time, average bitrate, and video quality variation. To achieve this goal, Huang *et al.* [6] designed a buffer-fullness based approach which directly chose the video bitrate by considering the current buffer occupancy. Yin *et al.* [7] proposed a model predictive control approach, which combined throughput and buffer occupancy information, for bitrate adaptation. In this line of work, other control theoretic approaches, such as Proportional-Integral-Derivative (PID) controller [8], are also investigated for bitrate adjustment.

Neural Network Based Video Streaming

Deep neural networks, inspired by the working principle of the human brain, are a set of algorithms (e.g., convolutional neural network, long short-term memory, and deep reinforcement learning) that are used to recognize patterns, including clustering, classification, and regression. They can model complex trends and detect non-linear relationships among input data. A typical deep neural network model makes up of one input layer, many hidden layers and one output layer. The input layer refers to the input data variables. Each hidden layer consists of a number of neurons that process its inputs from the previous layer using an activation function, which transforms the input data variables to an output data. The neurons between each layer are connected by connections that have numeric weights. The output layer generates results (e.g., class) for the given inputs according to the interconnection weights defined through the hidden layer. A convolutional neural network is comprised of a number of convolutional and subsampling layers and followed by several fully connected layers. The major benefits are that CNNs have quite fewer parameters than fully connected networks with the same number of hidden neurons, and can grasp the 2D structure of images. Long short-term memory (LSTM) is a variation of recurrent neural networks, which takes not only the current input example, but also those it has previously perceived (i.e., feedback connections). LSTMs help preserve the error that can be backpropagated through time and layers, and widely used in sequential data analysis. DRL, considering long-term accumulative rewards for sequential and far-sighted decision making, combines deep neural networks and reinforcement learning. It utilizes the neural networks to approximate the relationship between states and actions, and thus address the curse of dimensionality.

Currently, deep neural networks outperform traditional machine learning algorithms in many files, including computer vision, speech recognition, and natural language processing. Traditional machine learning algorithms consist of two major steps, including feature extraction, and (un)supervised feature learning. The performance highly relies on feature extraction, which needs prior knowledge and domain expertise. By contrast, deep neural networks replace these two steps with a unified neural architecture, and learn more latent fea-

tures via a huge amount of connected neurons to gain better performance. Although deep neural networks show excellent performance in many machine learning tasks, their applications to wireless systems have not yet been widely explored.

Motivated by the success of deep learning techniques, our previous work, called DeepQoE [9], presented an end-to-end framework by combining different neural network models to predict QoE, and outperformed the traditional QoE metric. Traditional video adaptation algorithms rely on model building and parameter estimation, such as network bandwidth. Deep learning based approaches [10, 11] directly learn how the past bitrate adjustment decisions impact the perceived video quality and make better choices. Gao *et al.* [12] proposed an user-interest-aware rate adaptation approach by inferring viewer interest based on video semantics.

Framework of Deep Neural Network Based Wireless Video Streaming

A generic system framework for wireless multimedia systems, as illustrated in Fig. 1, consists of three participatory segments, that is, end-users, network environment and media servers.

End-Users: They use different outlets for video consumption. Typical outlets include TV, PC, and smart phones.

Network Environment: Refers to various underlying networking, such as WiFi, 4G, and device-to-device. In general, the wireless channel condition is changing fast and the available bandwidth for video streaming is unstable.

Media Servers: Commonly deployed over the cloud infrastructure. In this service model, there is a resource pool powered by virtualization technology in a set of data centers. Those resources can be utilized on demand to provide elastic computation and storage capability to the upper layers [13]. This framework accepts all possible content sources, including live video streaming and video on-demand, and diverse video applications.

According to the characteristics of different segments, we can utilize deep learning to mine data intelligence and optimize system performance from different aspects, as explained below:

- For end-users, users may move around and access different mobile stations or WiFi for video consumption. We can utilize user trajectory information to analyze user behavior. In addition, we can utilize users' interest from the historical watching list to infer users' preference and recommend videos accordingly.
- For network environments, as the wireless channel is unstable and shared among multiple users, we can utilize historical bandwidth information to estimate the available bandwidth for each user. Furthermore, the spectrum resource is shared among multiple users and its utilization is coupled with many factors [14]. We can learn the optimal spectrum allocation in an online manner.
- For media servers, content caching affects the delay of each request. We can learn content popularity, predict which content will have more requests, and prefetch it to reduce delay. In addition, we can adjust the bitrate to reduce the jitter according to the available bandwidth.

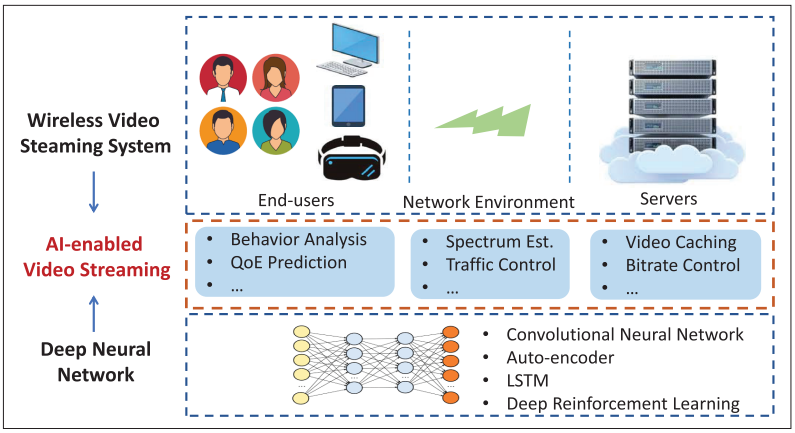

FIGURE 1. Integration of deep learning and wireless multimedia system. A wireless multimedia system consists of three components, including end-users, network environment, and servers. The deep learning technique can be incorporated into all the three components, such as behavior analysis for end-users, spectrum estimation for network environment, and video caching for multimedia servers.

QoE Prediction

In this section, we introduce our DNN based architecture for QoS/QoE prediction, and compare its performance with the traditional machine learning algorithms.

Problem Statement: QoS/QoE Prediction

Video QoS/QoE metrics, categorized into objective measurements and subjective measurements, guide the video compression, storage, transmission, deployment and operation of video related services and applications. QoS/QoE metrics assess the service quality from the perspective of end users or video systems. Objective measurements identify objective parameters that contribute to service quality. For instance, rate distortion and structural similarity indexes are used for video compression; the startup delay time and average bitrate are critical for video streaming. Subjective measurements solicit participants, give them a series of tested video sequences, and require them to provide scores on the video quality under laboratory environment.

Video QoS/QoE prediction, depending on various inter-related factors, remains to be a challenging task. In particular, video QoE relies on system factors, context factors and human factors. For example, system factors include frame rate and resolution; context factors include geo-location and video semantics; human factors include gender and age. Using these factors, traditional machine learning based methods utilize hand-crafted features and data representations for feature extraction, and further select an appropriate classification or regression model for QoS/QoE prediction. However, these methods develop separate frameworks for feature engineering and model training, resulting in inefficiency in the model development and training process.

Motivated by the success of deep neural networks, we aim to develop a deep learning based prediction framework to solve the following challenges. First, the framework should be in the end-to-end manner to combine feature extraction and representation, and thus eliminate the dataset spe-

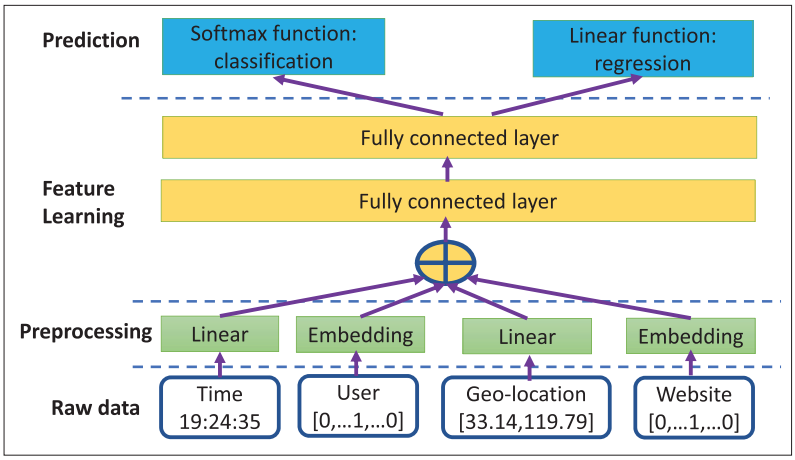

FIGURE 2. DNN architecture for mobile access pattern prediction. It consists of four layers, including one input layer, two hidden layers and one output layer. For the two hidden layer, we first utilize one entity embedding layer to transform the spare one-hot encoding into a vector with lower dimension for each type of input, and a dense layer to fully connected the neurons in the entity embedding layer.

cific representation and feature engineering. Second, the framework should be flexible to handle different types of datasets, coming from different sources, as well as data heterogeneity. In particular, datasets are heterogeneous in data modality and type. Some datasets contain numerical values and categorical information, while others include text descriptions, image information, and sequential information (e.g., video sequences). Furthermore, datasets are heterogeneous in representation approaches. It is straightforward to encode gender information with 0 and 1. The categorical information can be represented using one-hot vector (vector of zeros and one). Image information can be represented using CNN methods. It is not obvious which representation will lead to the best performance.

A DNN-Based Unified Framework for QoE Prediction

To tackle this problem, we propose a novel deep neural network architecture [9] as illustrated in Fig. 2. It consists of three phases for predicting video QoE: feature preprocessing, representation learning, and QoE prediction. Different types of raw data are transformed into initial feature vectors in the feature preprocessing phase. These initial feature vectors are then fused to learn a better representation, which is used for classification or regression in the QoE prediction phase.

Pre-Processing: The input data can be categorized into five types: text, video, categorial information, continuous information, and sequence data (e.g., bandwidth information). For each data type, we utilize a specific approach to extract features. In particular, we use GloVe, C3D, embedding layer, and dense layer and LSTM to extract the features for text, video, categorical information, continuous values and sequence data, as shown in Fig. 2. Let $\mathbf{x_i}$ denote the i-th input data and $\mathbf{v_i}$ denote the extracted feature vector. Data prepossessing can be denoted as $\mathbf{v_i} = f_i(\mathbf{x_i}; \delta) \in \mathbb{R}^U$, where f_i is the feature extraction method for the i-th data type and δ is the learned parameters.

Learning Representation: In this phase, we combine the feature vectors from the preprocessing phase to generate a single feature vector. In particular, the simple concatenation operation is employed and has the following form: $\mathbf{s} = C(g_1(\mathbf{v_1}), g_2(\mathbf{v_2}), ..., g_i(\mathbf{v_i}))$, where $\mathbf{v_i}$ is the i-th feature vector and $g_i(\mathbf{v_i})$ assigns different "weights" for distinctive features, which can be tuned as hyper-parameters in the training process. C is the concatenation operation and $\mathbf{s}$ represents the fused feature vector. Other fusion methods, such as 1D CNN, cannot provide noticeable performance improvement.

The fused feature vector is fed into several fully connected layers to generate a high-level representation. It should be noted that the number of fully connected layers can be adjusted according to the size of the dataset. Interested readers are referred to the hyper-parameter tuning. The representation output of layer j, that is, $\mathbf{r}^j$, is given as $\mathbf{r}^{(j)} = f(\mathbf{W}^{(j)}\ \mathbf{s}^{(j-1)} + \mathbf{b}^{(j)})$, where $\mathbf{s}^{(j-1)}$ is the input of layer j, $\mathbf{W}^j$ is the weight matrix between layer j–1 and layer j, $\mathbf{b}^j$ is the bias, and f represents the activation function of layer j. To prevent overfitting, we utilize the dropout technique in these hidden layers while training.

Predicting Video QoE: In this phase, we use a single-layer NN to process the output or the learned representation from the learning representation phase for either classification or regression. The learned representation is denoted as $\mathbf{s}^{(l)}$, and the associated ground truth or label of the training sample is denoted as $\mathbf{Y}$. Let $\mathbf{W}^{l+1}$ denote the weight matrix of the single layer NN. For classification, the loss function is defined as the cross-entropy function; for regression, the loss function is defined as the mean square error.

Performance Evaluation

To evaluate the performance of our proposed method, we utilize a dataset from [15] to drive the experiments. The dataset includes about 13,000 video sessions from four popular video service platforms (e.g., Youku and iQiyi), the overall volume of which is more than 10 TB. 320 students are recruited to conduct the standard subjective measurement, in which different operations (e.g., device ID, watching time, forward and suspend) and scores are recorded.

In the feature pre-processing phase, the pre-trained GloVe model is employed to transform video types to 50-dimension vectors, which are reduced to 5-dimension. The resolution information is mapped to an 8-dimension vector using an embedding layer. The bitrate and age information is normalized to range [0,1] and a dense layer is used to get two vectors of one dimension. For user gender information, an embedding layer is used to get a vector of one dimension. The bandwidth information is transformed to a 20-dimension vector using the pre-trained LSTM model. The video feature is extracted using the 3D CNN. All these vectors are concatenated into a single vector, which is fed to two fully connected layers with dropout technique applied to prevent overfitting. The output layer uses cross entropy loss function and softmax activation function for training and prediction, respectively. For baseline algorithms, we use SVM, decision tree, random forest Adaboost, and Naive Bayes.

The performance comparison is shown in Fig. 3. The accuracy of our algorithm is 88.74 percent, which is higher than the results generated by state-of-the-art machine learning algorithms.

Content Aware Video Streaming

In this section, we incorporate our QoE prediction model into the bitrate adaptation problem for video streaming, under the framework of deep reinforcement learning.

Problem Statement: Content-Aware Video Streaming

The prevailing bitrate adaptation methods for video streaming optimize the objective QoS/QoE metrics, such as the average bitrate and rebuffering time, and ignore viewers' subjective feelings on different video chunks when they watch video content. A hidden hypothesis for these methods is that different chunks of a video are of the same importance to viewers. As a result, these methods only need to increase the average bitrate of video chunks and reduce the initial startup delay and rebuffering time. However, the fact is that the human visual system is selective for distinctive video content and viewers will pay different attention to different content. For instance, pet lovers may be more interested in video clips that contain pets. If these video clips are in a higher resolution, they will enjoy them more.

This work aims to design a content-aware bitrate adaptation policy with the objective to prefetch a higher resolution version for video clips that is in line with viewers' interests. To achieve this goal, we need to address the following challenges.

User Interest Analysis: The interest of viewers is subjective and subtle, it is difficult to describe quantitatively by certain models. In addition, video content is quite complex, and it is challenging to analyze video content in real time.

Content-Aware Bitrate Control: Traditional methods adjust bitrate from two dimensions, including network conditions and subjective QoS/QoE metrics. By contrast, one extra dimension, that is, content intentness, is added to the optimization problem.

Deep Reinforcement Learning Based Approach

To tackle the aforementioned challenges, we propose a content-aware video streaming architecture (as shown in Fig. 4) based on deep learning techniques. It consists of three components, including a streaming environment, an agent, and the deepQoE. The streaming environment refers to the typical video streaming procedure. Given a video content, the streaming server divides it into a set of video chunks with an equal duration, and encodes each of them into different bitrate versions (e.g., 240p, 360p). The meta information of each video content, including the available bitrate, video description and video semantic information, is included in the Media Presentation Description (MPD) manifest file. While starting a video playback session, a client or video player requests the MPD of a video file and analyzes the available bitrates and semantic information of the video content. Based on measurement of network parameters, the video player can dynamically select the appropriate bitrate version of different video chunks to guarantee smooth playback.

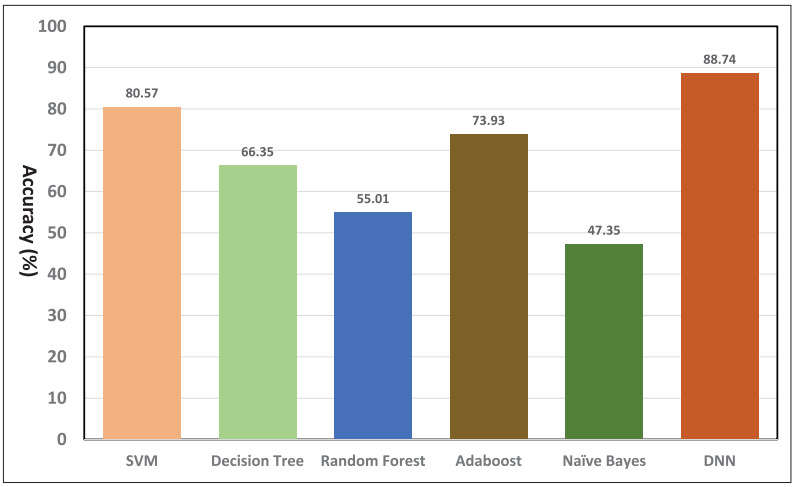

FIGURE 3. Prediction accuracy comparison of different algorithms. Our proposed method shows the best performance comparing to the other state-of-the-art machine learning algorithms.

The deepQoE component is responsible for interest analysis and QoS/QoE modeling, which are implemented using the DNN architecture as shown in Fig. 2. For interest analysis, 3D ConvNets is used to learn spatiotemporal features. In particular, we extract 16 images for each video chunk and use 3D ConvNets to generate video features. These features are fed into two fully-connected layers with the Rectifier activation function. The output layer has one node with the Softmax activation function. The output value is real-valued, and a larger value represents a higher level of video interestingness. To enable real-time bitrate control, the interestingness information is analyzed offline and included in the MPD file sent to a video player for decision making.

The agent interacts with the streaming environment and serves as the core for bitrate adjustment. We formulate the interest-aware bitrate adaptation as a reinforcement learning problem and learn the optimal policy in an online manner. In particular, after downloading the video chunk $t-1$, the agent observes the system state s_t, takes an action to select the best bitrate version for video chunk t, and gets rewards r_t after downloading video chunk t. This process continues until the end of the session. The physical meaning of different notations is listed as below:

- The state describes the bandwidth of the streaming service, the buffer occupancy, and the interestingness of the following video chunks, and so on.
- The control action is to select the appropriate bitrate version for the next requested video chunk.
- We adopt the QoE metrics above for measuring the reward during a time slot.
- Our objective is to derive the optimal rate adaptation policy for maximizing the overall discounted rewards over a video session.

We adopt DQN for learning the rate adaptation policy.

Performance Evaluation

We use real-world traces to conduct our experiments. The FCC broadband dataset and the 3G/HSDPA mobile dataset are used to simulate different network conditions. The available bitrate ver-

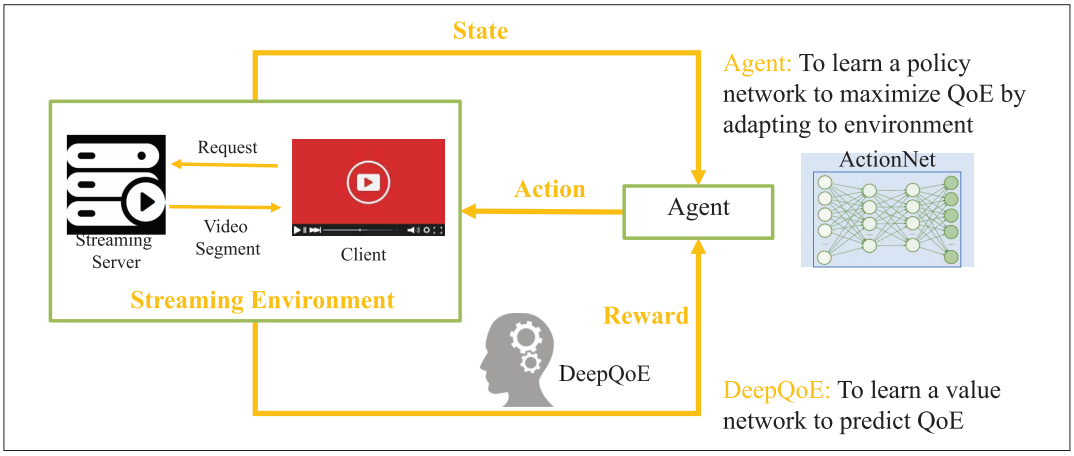

FIGURE 4. Learning based video adaptation system. It consists of three components, including streaming environment, agent and deepQoE. The streaming environment refers to the typical video streaming procedure. The deepQoE component is responsible for interest analysis and QoS/QoE modeling. The agent interacts with the streaming environment and serves as the core for bitrate adjustment.

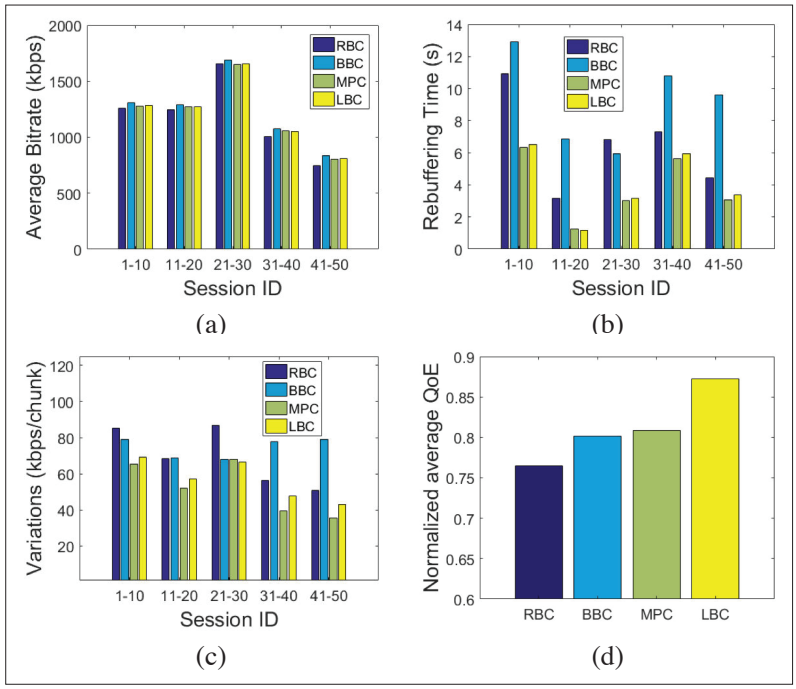

FIGURE 5. Performance comparison with other bitrate adjustment algorithms.

sions are 350kb/s, 600kb/s, 1000kb/s, 2000kb/s, and 3000kb/s. For the DQN agent, we utilize a fully-connected neural network with two hidden layers. The dimensions of the two hidden layers are 256 and 512. The activation function is ReLu, and the output layer uses a linear activation function to generate an approximated Q value for the state-action pair.

Our proposed method (called LBC) is compared with three state-of-the-art algorithms:

- Buffer-based control (BBC) chooses the bitrate version for the next video chunk according to the buffer occupancy.
- Rate-based control (RBC) selects the largest available bitrate version which is less than the estimated bandwidth.
- Model prediction control (MPC) uses the MPC method to select the bitrate version that maximizes the overall QoE function. The prediction horizon is three time slots.

The performance comparison results are shown in Fig. 5. Our proposed method outperforms the baseline algorithms in terms of average bitrate, rebuffering time, bitrate variation, and weighted QoE metric.

Summary and Future Direction

In this work, we discussed how to integrate wireless multimedia systems with the fast growing deep learning techniques, and presented several potential topics for different components in a wireless multimedia system. Furthermore, we introduced two cases, including deep learning based QoS/QoE prediction and bitrate adjustment. In the former case, an end-to-end and unified DNN architecture was devised to fuse different types of multimedia data and predict the QoS/QoE value. In the latter case, a deep reinforcement learning based framework was designed for bitrate adjustment according to the viewers' interestingness.

References

[1] B. Tan *et al.*, "Toward a Network Slice Design for Ultra High Definition Video Broadcasting in 5G," *IEEE Wireless Commun.*, vol. 25, no. 4, 2018, pp. 88–94.

[2] "Cisco Visual Networking Index: Forecast and Trends, 2017c2022 white paper," https://www.cisco.com/c/en/us/solutions/collateral/ service-provider/visual-networking-index-vni/white-paper-c11-741490. html, accessed: 2019-05-08.

[3] J. Huang *et al.*, "Big Data Routing in D2D Communications with Cognitive Radio Capability," *IEEE Wireless Commun.*, vol. 23, no. 4, 2016, pp. 45–51.

[4] H. Hu, Y. Li, and Y. Wen, "Toward Rendering-Latency Reduction for Composable Web Services via Priority-Based Object Caching," *IEEE Trans. Multimedia*, vol. 20, no. 7, 2018, pp. 1864–75.

[5] H. Hu *et al.*, "Orchestrating Caching, Transcoding and Request Routing for Adaptive Video Streaming over ICN," *ACM Trans. Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, 2019, p. 24.

[6] T.-Y. Huang *et al.*, "A Buffer-Based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service," *ACM SIGCOMM Computer Commun. Review*, vol. 44, no. 4, 2015, pp. 187–98.

[7] X. Yin *et al.*, "A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP," *ACM SIGCOMM Computer Commun. Review*, vol. 45, no. 4. ACM, 2015, pp. 325–38.

[8] Y. Qin *et al.*, "A Control Theoretic Approach to ABR Video Streaming: A Fresh Look at PID-Based Rate Adaptation," *Proc. IEEE INFOCOM 2017-IEEE Conf. Computer Commun. IEEE*, 2017, pp. 1–9.

[9] H. Zhang *et al.*, "DeepQoE: A Unified Framework for Learning to Predict Video QoE," *Proc. 2018 IEEE Int'l. Conf. Multimedia and Expo (ICME)*, IEEE, 2018, pp. 1–6.

[10] H. Mao, R. Netravali, and M. Alizadeh, "Neural Adaptive Video Streaming with Pensieve," *Proc. Conf. ACM Special Interest Group on Data Communication*, ACM, 2017, pp. 197–210.
[11] G. Gao *et al.*, "Content-Aware Personalised Rate Adaptation for Adaptive Streaming via Deep Video Analysis," arXiv preprint arXiv:1811.06663, 2018.
[12] G. Gao *et al.*, "Optimizing Quality of Experience for Adaptive Bitrate Streaming via Viewer Interest Inference," *IEEE Trans. Multimedia*, vol. 20, no. 12, 2018, pp. 3399–3413.
[13] J. Wu *et al.*, "Cloud Radio Access Network (C-RAN): A Primer," *IEEE Network*, vol. 29, no. 1, 2015, pp. 35–41.
[14] H. Hu, Y. Wen, and D. Niyato, "Spectrum Allocation and Bitrate Adjustment for Mobile Social Video Sharing: Potential Game with Online QoS Learning Approach," *IEEE JSAC*, vol. 35, no. 4, 2017, pp. 935–48.
[15] L. Zhou *et al.*, "Seeing Isn't Believing: QoE Evaluation for Privacy-Aware Users," *IEEE JSAC*, vol. 37, no. 7, 2019, pp. 1656–65.

Biograpahies

Lu Liu is an associate professor with the College of Movie and Media at Sichuan Normal University, China. Concurrently she is a postdoctoral fellow with the School of Journalism and Communication at Fudan University, China. She received her Ph.D. degree in Journalism from Sichuan University in 2011. She was a visiting scholar at the Communication University of China, and the School of Computer Science and Engineering (SCSE) at Nanyang Technological University (NTU), Singapore in 2018. She has published more than 20 papers and two monographs. Her research focuses on applying quantatitive techniques to emerging communication topics, including new media communication, political communication, and cultural product.

Han Hu is currently a professor with the School of Information and Electronics, Beijing Institute of Technology, China. His research interests include multimedia networking, edge intelligence and data analytics. He received several academic awards, including the Best Paper Award at IEEE TCSVT 2019, the Best Paper Award from *IEEE Multimedia Magazine* in 2015, the Best Paper Award at IEEE Globecom 2013, among others. He served as an associate editor of IEEE TMM, and a TPC Member of Infocom, ACM MM, AAAI, IJCAI, among others.

Yong Luo received the B.E. degree in computer science from the Northwestern Polytechnical University, Xi'an, China, in 2009, and the D.Sc. degree from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2014. He is currently a research fellow with the School of Computer Science and Engineering, Nanyang Technological University. His research interests are primarily in machine learning and data mining with applications to visual information understanding and analysis. He has authored several scientific articles at top venues including IEEE T-PAMI, T-NNLS, IEEE T-IP, IEEE T-KDE, IEEE T-MM, IJCAI and AAAI. He received the IEEE Globecom 2016 Best Paper Award, and was nominated as the IJCAI 2017 Distinguished Best Paper Award. He is the corresponding author of this work.

Yonggang Wen [S'99, M'08, SM'14] is a full professor and the Associate Dean (research) of College of Engineering (CoE) at Nanyang Technological University (NTU), Singapore. He also served as the Acting Director of the Nanyang Technopreneurship Centre at NTU. He received his Ph.D. degree in electrical engineering and computer science (minor in Western literature) from Massachusetts Institute of Technology (MIT), Cambridge, USA, in 2007. Previously he has worked at Cisco to lead product development in content delivery networks, which had a revenue impact of 3 Billion US dollars globally. He has worked extensively in learning-based system prototyping and performance optimization for large-scale networked computer systems. His work in Multi-Screen Cloud Social TV has been featured by global media (more than 1600 news articles from over 29 countries) and received 2013 ASEAN ICT Awards (Gold Medal). His work on Cloud3DView, as the only academia entry, has won 2016 ASEAN ICT Awards (Gold Medal) and 2015 Datacentre Dynamics Awards 2015 C APAC (Oscar award of data centre industry). He is a co-recipient of the 2015 IEEE Multimedia Best Paper Award, and a co-recipient of Best Paper Awards at 2016 IEEE Globecom, 2016 IEEE Infocom MuSIC Workshop, 2015 EAI/ICST Chinacom, 2014 IEEE WCSP, 2013 IEEE Globecom and 2012 IEEE EUC. He received the 2016 IEEE ComSoc MMTC Distinguished Leadership Award. He serves on editorial boards for *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Wireless Communications Magazine*, *IEEE Communications Survey & Tutorials*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Signal and Information Processing over Networks*, *IEEE Access*, and *Elsevier Ad Hoc Networks*. He was elected Chair for the IEEE ComSoc Multimedia Communication Technical Committee (2014–2016). His research interests include cloud computing, green data center, distributed machine learning, big data analytics, multimedia networks and mobile computing.