Self-Attention-based Uplink Radio Resource Prediction in 5G Dual Connectivity

Jewon Jung, Sugi Lee, Jaemin Shin and Yusung Kim

***Abstract*—Mobile communication technology is evolving rapidly and becoming increasingly ubiquitous, thereby increasing the demand for uplink data-intensive applications (e.g., personal broadcasting and live augmented/virtual reality videos). Recently, to facilitate a cost-effective and smooth transition from 4G to 5G networks, most carriers leverage existing 4G infrastructures using a dual connectivity (DC) feature. DC increases uplink throughput and mobility robustness; however, it also causes unprecedented dynamic fluctuations in radio channels due to the coverage discrepancy between 4G and 5G networks. Thus, in this paper, we propose a self-attention-based deep learning model to predict uplink radio resources in 5G DC. We trained the proposed model on commercial 5G DC traffic data from three major carriers in South Korea and obtained an average prediction accuracy of 95.08% under various mobility and cell-load conditions. The proposed model explains the rationale for the obtained predictions by highlighting the parts of the input time-series data that are important to realize accurate prediction. We also demonstrate the usability of the proposed model using a network emulator based on real-world 5G trace data. Extensive evaluations demonstrate that the existing congestion control algorithms can achieve excellent performance when used with the proposed model.**

***Index Terms*—5G Uplink Prediction, Dual Connectivity, Deep Learning, Transformer**

I. Introduction

THE emergence of 5G New Radio (NR) technology has enabled the development of various multimedia and IoT services that leverage its high bandwidth, ultra-low latency, and massive connectivity. In contrast to 4G, 5G multimedia services, such as high-resolution live streaming (e.g. up to 16K) and virtual/augmented reality, require huge resources for uplink transmission [1]–[4]. 5G will also enable new capabilities in vehicle-to-everything applications where vehicles equipped with cameras and sensors generate increasing amounts of multi-modal data to improve services such as autonomous driving, advanced driver assistance system, and infotainment [5], [6]. Uplink performance has therefore become a crucial factor in 5G, especially given the impact that any transmission delays could have on human life. However, in 5G networks, a user equipment (UE) or an IoT device freely moves within or between cells, and a base station (BS) must allocate an uplink radio resource to the UE (or an IoT device) dynamically according to their radio channel conditions. Unpredictable channel fluctuations degrade throughput significantly and increase end-to-end latency, and such frequent fluctuations severely reduce the performance of mobile applications.

To address this problem, several studies [7]–[12] have investigated forecasting uplink radio resources using rule-based or machine learning (ML) methods. The available bandwidth can be calculated according to the predicted radio resources by applying bandwidth prediction for their congestion control algorithms or application-level rate controls. However, the prediction models were specifically designed for traditional 4G networks. Thus, additional considerations are required to design a prediction model for the unique characteristics of 5G network systems.

5G provides high bandwidth and low latency due to its physical-layer innovations (e.g., massive multiple-input and multiple-output and advanced channel coding). However, for a smooth migration to 5G, most carriers employ the non-standalone (NSA) mode to reuse the existing 4G infrastructure. In the NSA mode, a UE can be connected to both 5G and 4G radio cells simultaneously. Although this dual connectivity (DC) increases the aggregate throughput and mobility robustness, the UE will experience frequent fluctuations in radio channels due to the coverage gap between these two different technology cells.

Thus, in this study, we measured large amounts of 5G DC traffic from three major carriers in South Korea. For approximately 6,000 minutes, various scenario data were collected, e.g., different mobility cases (highway driving, downtown driving, downtown walking, and stationary cases), several time zones, and different cell loads. By conducting in-depth data analysis, we found that uplink throughput (i.e., the sum of 4G and 5G throughput) fluctuates more frequently than when using only conventional 4G networks, which is consistent with previous measurement results [13], [14]). These frequent fluctuations occur in both moving and stationary scenarios due to the poor channel conditions of 5G radio.

In this paper, we propose the learning-based Self-attention-based Uplink Radio resource Estimation (SURE) model to predict uplink radio resources in 5G DC. The proposed SURE model is designed with a lightweight Transformer [15] architecture. The Transformer introduces a self-attention mechanism to efficiently process sequential input data, e.g., natural

Jewon Jung is with the College of Computing and Informatics, Sungkyunkwan University, Suwon 03063, Republic of Korea (e-mail: doubele112@skku.edu). (Co-first author)

Sugi Lee is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 03063, Republic of Korea (e-mail: sglee0323@gmail.com). (Co-first author)

Jaemin Shin is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 03063, Republic of Korea (e-mail: alex93950@gmail.com).

Yusung Kim is with the College of Computing and Informatics, Sungkyunkwan University, Suwon 03063, Republic of Korea (e-mail: yskim525@skku.edu). (Corresponding author)

Manuscript received , 2022; revised , 2023.

language. The attention mechanism is effective at differentially weighing the importance of each part of the input sequence. We demonstrate that this process helps SURE understand the dynamics of 5G DC channels affected by various situations. The proposed model uses channel information that can be collected by a UE without modification of commercial 5G modems. With a given sequence of channel information, the proposed SURE model predicts radio resources to be allocated within a short period (e.g., 100 ms) for both 4G and 5G technologies.

Our primary contributions are summarized as follows.

- We thoroughly examined the input features, the input sequence length, and the model structures to efficiently learn the 5G DC dynamics. We found that training to predict 4G and 5G radio resources together using a single shared model can achieve higher accuracy than training each process separately.
- The proposed SURE model can learn which parts of the time-series input data have major impacts on future predictions. This attention mechanism improves the representation learning of input data and provides an explanation of how accurate predictions could be determined under various conditions.
- In a real trace-driven network emulator [16], we evaluated the usability of the proposed SURE model by integrating it into existing congestion control algorithms, e.g., CUBIC [17], BBR [18], and Indigo (a recurrent neural network-based approach) [19]. By simply providing an upper bound of the available bandwidth, all of these algorithms exhibit better throughput while maintaining a low queuing delay.

II. Background and Motivation

A. Uplink Resource Allocation in Cellular Networks

In cellular networking systems, a BS monitors the states of randomly moving a UE in each cell and conducts uplink channel scheduling by assigning the UE resource blocks (RB) dynamically according to the uplink traffic load of each cell. An RB is a minimum unit of radio resource allocation, and a BS initiates the uplink RB allocation when receiving a scheduling request message generated by a UE that has data to transfer [20]. In addition, the BS adopts the proper modulation coding scheme (MCS) in consideration of the current channel conditions and determines the transport block size (TBS) based on the number of allocated RBs and MCS, i.e., the amount of data that can be transferred during a single transmission time interval (TTI) [1]. This mechanism effectively helps a BS to assign optimal radio resources to a UE with high mobility but also causes it to experience fluctuating throughput and latency, thereby resulting in severe performance degradation in mobile applications.

B. Deployment of 5G Networking System

The 3rd Generation Partnership Project (3GPP) specifies various options for efficient 5G deployment [2]. The standalone (SA) architecture (option 2), operating on an infrastructure dedicated to 5G, fully exploits the advantages of 5G networks. However, it is challenging and time-consuming to completely replace existing 4G infrastructures with 5G infrastructures. To save time and reduce costs, most carriers have switched to an NSA 5G architecture (option 3), which provides 5G networking on existing 4G infrastructures; thus, NSA 5G architectures have been widely deployed worldwide. Therefore, in this study, we primarily focused on predicting uplink radio resources in NSA 5G networks, although we expect that the proposed SURE model can work effectively in both SA and NSA 5G networks.

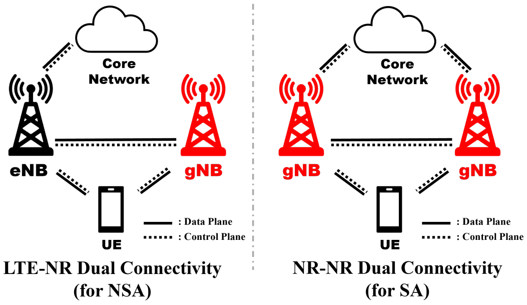


Fig. 1. **Dual Connectivity overview for NSA/SA architecture:** The LTE-NR DC utilizes both eNodeB and gNodeB for data transmission. In NR-NR DC of the SA architecture, a UE connects to two gNodeBs for data transmission.

C. Dual Connectivity in 5G Networks

DC is an important feature of 5G networks that effectively improves both throughput and mobility robustness. Fig. 1 shows two types of DC in 5G networks, i.e., LTE-NR DC and NR-NR DC [14]. The LTE-NR DC supported in NSA 5G enables a UE to perform parallel data transmission through 4G and 5G channels based on simultaneous connections with a 4G BS (eNodeB) and a 5G BS (gNodeB). In the LTE-NR DC, an eNodeB with greater coverage is a master node (MN) that manages the control plane, and a gNodeB, as a secondary node (SN), is connected to a UE via an MN. The SA 5G architecture provides NR-NR DC, which utilizes mm-wave and mid/sub-6 GHz channels using two gNodeBs. In the NR-NR DC, both gNodeBs manage the control plane; however, a single gNodeB with mid/sub-6 GHz frequencies functions as an MN based on its greater coverage than the SN. These mechanisms help a UE boost throughput and experience seamless networking services because an MN effectively covers the high-performance but unstable data plane of an SN. However, new types of handovers caused by the different coverages between an MN and an SN cause unprecedented fluctuations of the radio resources to be assigned to a UE, which can seriously degrade mobile application performance. Fig. 2 shows three types of handovers in 5G DC when a UE moves between SNs that are close to each other (Fig. 2a) or distant from each other (Fig. 2b) within the coverage of an MN and moves between MNs (Fig. 2c).

D. Uplink Prediction Challenges in 5G Networks

Data-intensive mobile applications are becoming increasingly ubiquitous; thus, the demand for efficient uplink trans-

[1] In 4G, the TTI is fixed to 1ms, while scalable TTI (from 62.5μs to 1ms) is used in 5G networks.

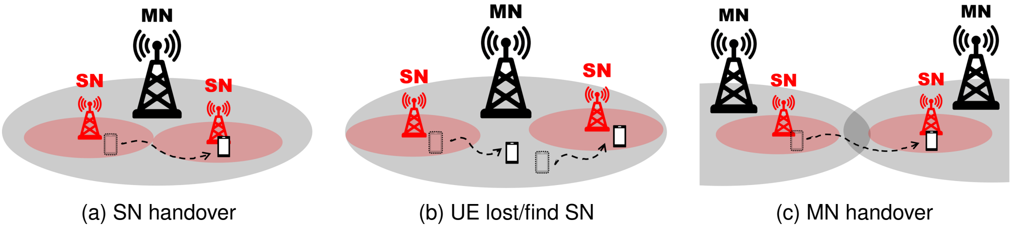


Fig. 2. **Various handover scenarios in 5G:** (a) A UE moves to the adjacent SN causing a handover between SNs while maintaining DC. (b) As the UE gets out of the coverage of the connected SN, DC is deactivated. When the UE enters the coverage of another SN, DC is reactivated. (c) If the UE approaches the edge of the connected MN, handover between MNs occurs with temporary deactivation of DC.

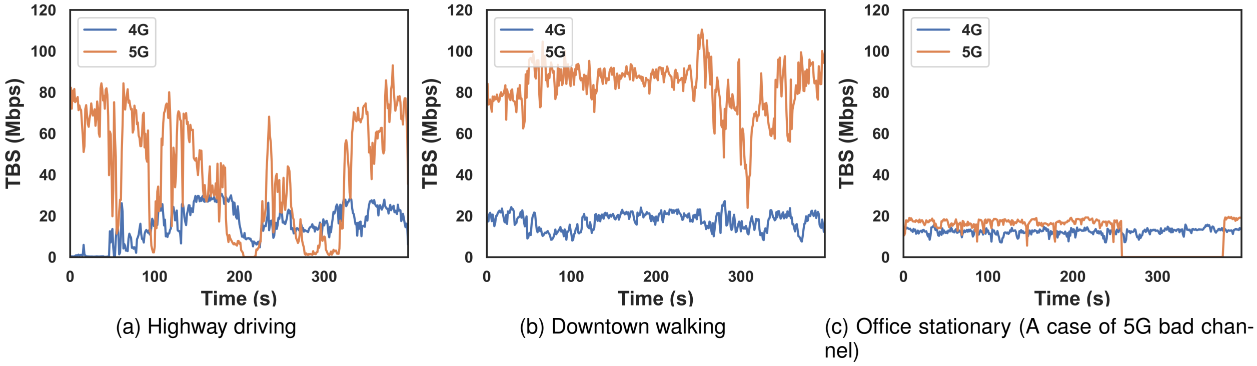


Fig. 3. **5G TBS fluctuation for each scenario:** (a) In the highway driving scenario, 5G TBS fluctuates drastically due to the rapidly changing channel quality, frequent cell switching, and handover caused by the small cell coverage of a gNodeB, whereas 4G TBS fluctuates much less frequently. (b) In the downtown walking scenario, 5G TBS still fluctuates due to the high cell load and small cell range. (c) An unstable 5G signal makes a UE disconnect from gNodeB even in the office stationary scenario.

mission is also increasing in 5G networks. However, such applications frequently have degraded performance because a 5G UE experiences more frequent handovers than 4G networks due to the small communication range (e.g., micro/pico/femto-cells) of the high-band signals [21] and DC operations [14]. To demonstrate this, we measured uplink resource fluctuations caused by handovers in real-world NSA 5G networks. In Fig. 3, 4G, and 5G TBSs are shown for three scenarios: (1) highway driving, (2) downtown walking, and (3) office stationary. The TBSs in the office stationary scenario are lower than those in the other scenarios due to poor channel conditions. We observe that the 5G TBS fluctuates severely due to the rapidly changing channel quality and frequent handover in the highway driving and downtown walking scenarios, whereas the 4G TBS oscillates less. In particular, we note that the 5G TBS fluctuates even in the office stationary scenario, and we can reasonably assume that this is because the UE was located in the coverage boundary of a gNodeB, and the signal from the gNodeB suffered attenuation or path loss frequently. These analyses and observations well explain the difficulty in forecasting uplink radio resources in 5G networks accurately. Several studies [9]–[12] have proposed uplink prediction models for cellular networks to improve the performance of uplink-centric mobile applications. These models predict the near-future available bandwidth based on the history of various information related to uplink resource allocation. Specifically, in recent studies about uplink resource prediction [10], [12], it was shown that the length (or time window) of history strongly affects prediction accuracy, and PERCEIVE [10], which forecasts uplink resources based on long short-term memory (LSTM), successfully improved accuracy using an additional LSTM model that selects an appropriate time window dynamically in response to channel condition changes. However, in our systematic analysis (Section IV-A), the time window selection, which requires a single one-time window value from limited options (100 ms, 300 ms, and 1000 ms), was insufficient to maximize prediction accuracy under extremely fluctuating 5G channel conditions. In addition, most existing prediction models, including PERCEIVE, are specifically designed for 4G networks without considering the aforementioned challenges associated with 5G networks. Thus, in this study, we propose a Transformer-based prediction model that performs precise uplink prediction in 5G networks by automatically focusing on the important parts of the input time-series data.

E. Transformer

The Transformer [15] model, which is a contemporary deep learning model, processes sequential data (e.g., language or vision data) based on a self-attention mechanism that provides context to each part of an input sequence by estimating the independent weights of those parts. Here, the context represents how closely one part of the input sequences is relevant to other parts, and the Transformer model effectively predicts missing or succeeding parts of the input sequences based on the available context information. The Transformer model utilizes an encoder and decoder that each perform self-attention. The encoder comprises multiple encoding layers, each of which processes the input sequences and passes the results to the following layer. The decoder with multiple decoding layers generates output sequences from the encoding results based on comprehensive context information. Thus, compared to recurrent neural networks (RNN) traditionally used to process sequential input data, the Transformer model works efficiently regardless of the length of the input sequences and prevents the vanishing gradient problem [15], which significantly reduces the accuracy of RNNs.

III. Design of SURE

In this section, we describe the architecture and implementation of the proposed SURE model, which learns an uplink scheduling pattern and performs millisecond-level uplink prediction in 5G DC.

A. System Overview

Fig. 4 shows an overview of the SURE-based uplink radio resource prediction system. In this system, the UE collects uplink scheduling information every 20 ms and creates a feature group based on this information. We denote this feature group by $F_t = \{f_1^t, f_2^t, ..., f_k^t\}$ where f_k and t refer to the selected input feature and specific time point t, respectively. The Transformer-based prediction model utilized in the proposed SURE model processes an input sequence with 50 feature groups gathered over the past 1000 ms, which can be denoted by the input sequence $S_t = \{F_{t-49\phi}, F_{t-48\phi}, ..., F_{t-\phi}, F_t\}$ where ϕ refers to the time period of 20 ms. With S_t consisting of 50 feature groups, our prediction model imposes different weights on every feature group by computing self-attention distribution and passes them across normalization and feed-forward layers, just like the original encoder of the Transformer model. After predicting the 4G and 5G TBSs at the same time, we inform the UE of 4G or 5G TBSs to be allocated by the BSs (eNodeB or gNodeB) for the next 100 ms. Finally, the UE estimates the available bandwidth based on TBSs and applies the predictive bandwidth to various congestion control schemes to improve application performance against severely fluctuating uplink radio resources. Fig. 7a shows the performance of the proposed SURE model with different input sequence lengths. As can be seen, SURE achieves the lowest root mean square error (RMSE) when the length of the input sequence is equal to or greater than 1000 ms because the Transformer model can focus on important parts of long input sequences dynamically based on the self-attention mechanism.

B. Data Collection

To maximize the accuracy of the proposed SURE model, we collected a large amount of training data and extensively measured 5G DC traffic over the commercial NSA 5G networks of three major Korean carriers (SKT, KT, and LG U+). In these measurements, we used a rooted phone (Samsung Galaxy S20) and XCAL-Solo [22], which is a COTS monitoring tool, to acquire a cellular signal and uplink scheduling information directly from a 5G modem chipset. We executed iPerf [23] to generate massive UDP traffic to fully utilize the available uplink bandwidth allocated by BSs (eNodeB or gNodeB). As shown in TABLE I, we collected measurement data in various locations, i.e., five highways, four downtown areas, and three stationary locations (residences, offices, and department stores). Here, equal amounts of data were collected from each of the three carriers. The entire dataset corresponds to a duration of over 100 hours and covers different time periods: morning (7:00–12:00), afternoon (12:00–18:00), evening (18:00–00:00), and dawn (00:00–7:00). Our test dataset and brief learning code are available publicly at GitHub repository [53].

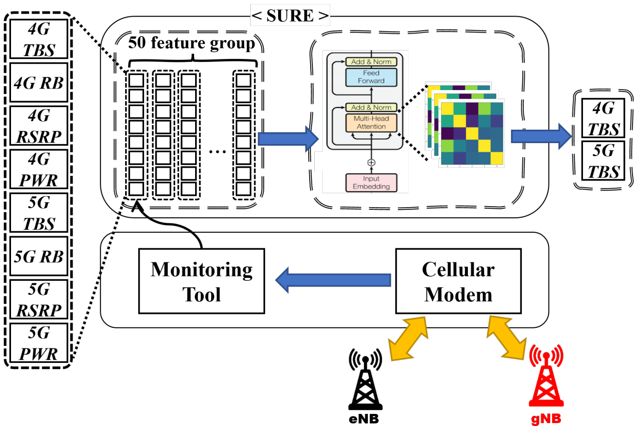


Fig. 4. **SURE Overview:** 50 feature groups consisting of the past 1000 ms of data gathered from the cellular monitoring tool form the input sequence for the Transformer-based prediction model that estimates the average 4G/5G TBSs for the next 100 ms.

TABLE I

Data collection statistics in minutes: Extensive 5G DC traffic traces collected in various time zones and locations.

	Highway	Downtown	Stationary	Total (min)
Morning	455	222	68	**745**
Afternoon	1124	372	713	**2209**
Evening	888	593	1296	**2784**
Dawn	29	28	461	**518**
Total (min)	**2496**	**1215**	**2545**	**6256**

C. Input Feature Selection

We select input features by performing an in-depth analysis of the uplink scheduling and DC operations in 5G networks. Here the TBS, RB, and reference signals received power (RSRP) are selected because they tend to have a close relation with radio resource allocation in cellular networks [10]–[12], [24]. In addition, we discover an unrevealed correlation between the TBS to be allocated to the UE and the transmission power (Tx-Power) required when the UE transfers uplink data to a BS. This may be because the Tx-Power is selected by considering the channel quality between a BS and a UE; however, Tx-Power also impacts the TBS for the following reasons. As in the literature [25], for 5G DC, a UE is assigned independent Tx-Powers by two BSs. However, the UE cannot use both BSs simultaneously when each Tx-Power or the sum of the Tx-Powers exceeds the output power capability of the UE. Several power-sharing schemes have been proposed previously to fully exploit 5G DC on power-constrained UEs [26], [27], and we find that Tx-Power is a key element in these schemes. Thus, we select 4G/5G Tx-Powers as our final learning feature. Fig. 5 shows the nonlinear correlation coefficients between the learning features and the 4G and 5G TBSs. The coefficients

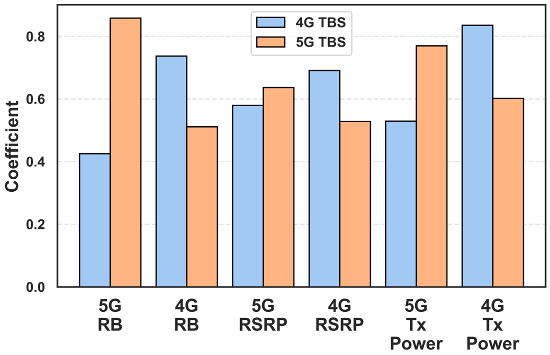


Fig. 5. **Nonlinear correlation coefficient of each feature with 4G/5G TBS:** 4G features tend to have a relationship with both 4G TBS and 5G TBS (and vice versa).

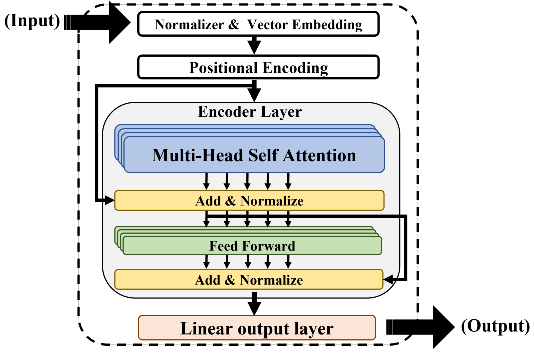


Fig. 6. **Transformer Model Structure in SURE:** A normalized input sequence is an input to the encoder through an embedding and positional encoding layer. Rather than a decoder layer, here, a linear output layer is connected directly to the encoder.

are calculated in reference to the literature [28], and all features exhibit a high correlation with both TBSs. Thus, the proposed SURE model achieves 10% higher accuracy than the combination of two Transformer models that learn either the 4G TBS or 5G TBS separately, as shown in Fig 7b. We also evaluated the performance of the proposed SURE model with different feature sets, and it achieved the highest accuracy when using all selected features (RSRP, RB, TBS, and Tx-Power), as shown in Fig.. 7c. Our novel feature, i.e., Tx-Power, significantly affects the accuracy, whereas an additional feature, i.e., MCS, reduces the accuracy of the proposed model.

D. Model Architecture

Decoder-free Transformer Model. Fig. 6 illustrates the structure of our Transformer model. Here, we replaced the decoder with a linear output layer to obtain the output results. Since our goal is to predict 4G and 5G TBSs at the same time, we set the size of the output dimension to 2. This decoder-free Transformer model has the following advantages compared to the original Transformer. (1) According to the literature [29], our proposed Transformer model is less likely to suffer overfitting, and it effectively solves regression problems using context information from an encoder directly when deriving the final results. (2) Our Transformer model requires fewer parameters, which reduces the computational resources required for training and inference by less than 1%. (3) Its lightweight design ensures that the proposed SURE model can be implemented easily on a UE. As shown in Fig. 7d, our model achieves approximately 30% lower RMSE than the Transformer model with the decoder.

TABLE II
TRAINING HYPERPARAMETER OF SURE

Batch size	128
Dropout rate	0.03
Model dimension	128
Feed-forward dimension	256
Number of heads	8
Number of encoder block	1
Epoch	150
Initial learning rate	0.00025
Optimizer	RAdam optimizer
Normalization layer	LN
Activation function	GELU

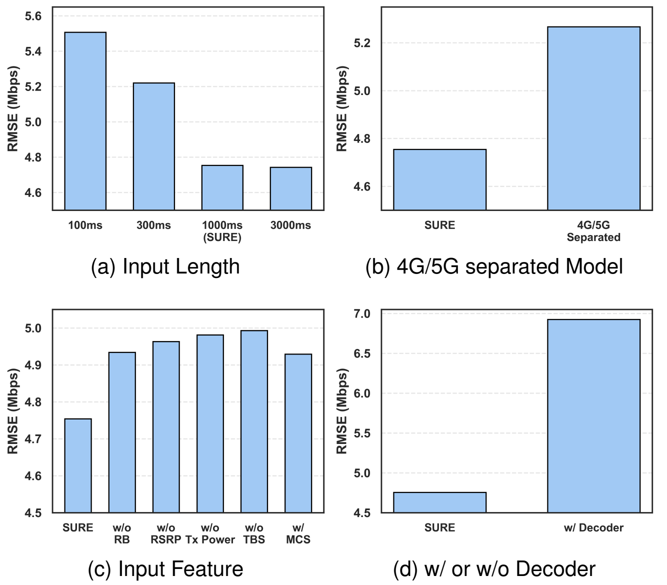
Learning Process. In the proposed model, training and test datasets are normalized using the standardization method for each feature. A total of 8 selected features (TBS, RB, RSRP, and Tx-Power of 4G and 5G) form a feature group F_t for every time point and an input sequence S_t consisting of 50 F_ts gathered over the past 1000 ms. After that, S_t is linearly projected through a vector embedding module to be transformed into a group of single vectors that can be processed by the encoder. Specifically, every feature group F_t consisting of 8 features (k=8) passes a single linear layer to be compressed to a single vector. These vectors are used as queries, keys, and values at the self-attention layer of the Transformer model. We also utilize positional encoding [15] to assign a sequential meaning of the time-series data to the input vectors. And then, the encoder with eight self-attention heads processes the input vectors and the final sequence of continuous representation vectors $Z_t = \{z_{t-49\phi}, z_{t-48\phi}, ..., z_{t-\phi}, z_t\}$ is generated and concatenated into a single vector $\bar{z}_t$ which is used as an input to the linear output layer. Finally, the output results of average 4G and 5G TBSs for the next 100 ms are estimated through the linear output layer:

$$\hat{y} = W\bar{z}_t + b \tag{1}$$

where $\hat{y}$ refers to the prediction result and W, b are learnable weight and bias value of the linear layer (note that we set the output dimension to 2, the 4G and 5G TBS predictions are estimated). We also applied the RMSE loss function:

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y_i} - y_i)^2} \tag{2}$$

where N and y_i refer to the number of samples and ground truth, respectively. We empirically selected hyperparameters through several experiments. Our model performed best when composed of one encoder block with an internal embedding dimension size of 128. A set of hyperparameters that showed the best performance during the model training is described in TABLE II.



(a) Input Length

(b) 4G/5G separated Model

(c) Input Feature

(d) w/ or w/o Decoder

Fig. 7. **Ablation Study of SURE:** (a) Comparison of prediction accuracy with different input sequence lengths. (b) The proposed SURE model outperforms the separated prediction model due to the strong relation between 4G and 5G features. (c) Prediction accuracy which achieved by SURE with different feature combinations. (d) For our regression task, a model without a decoder layer exhibits optimal performance.

E. Implementation

Model Complexity. Developing an efficient model for a UE with limited resources, e.g., RAM and CPU, is challenging, and certain deep learning models (e.g., Transformer models) require extensive resources, which may exceed a UE's capabilities. The proposed model is designed as a lightweight version of the Transformer model. Here, the number of parameters in our model is 146,434, which is less than 1% of the number of parameters in the original Transformer model [15]. In addition, we reduce the memory size of the model by approximately 30% using quantization without compromising prediction accuracy (accuracy is reduced by only 0.035%). We also implemented the proposed SURE model on several COTS 5G smartphones (Galaxy S20, Galaxy A90 5G, and LG Q92 5G) and evaluated its practicality. We observed that the inference of the proposed SURE model is completed within 12.23–16.88 ms using minimal resources (5.25–8.657% of the CPU resources and 1.223–2.957% of the RAM resources).

Integration with congestion control protocols. Congestion control is a core component that efficiently utilizes network resources while avoiding congestive collapse. Many congestion control algorithms have been studied, and some have been implemented in operating system kernels. Each algorithm is designed to realize a specific purpose, e.g., high throughput [17], low tail latency [18], and seamless live streaming [30]. Thus, we attempt to generalize the proposed model such that it can be incorporated into various congestion control algorithms rather than being tied to a specific algorithm. Here, we introduce a simple concept to integrate the proposed SURE model with several congestion control algorithms. We limit the number of in-flight packets without affecting the original congestion control algorithms. For this purpose, we calculate the average congestion window (or bit rate) for the next 100 ms based on 4G/5G TBSs predicted by the proposed SURE model every 100 ms and determine whether to send a packet based on two strategies, i.e., the *min-strategy* and *avg-strategy*. The *min-strategy* limits the generation of packets when the number of in-flight packets is greater than the smaller value of the original congestion window (*cwnd*) and the predicted average congestion window ($cwnd_{pred}$). In contrast, the *avg-strategy* regulates the average number of in-flight packets never exceeding $cwnd_{pred}$. Here, we utilize the *min-strategy* due to its simplicity, even though its expected throughput is less than that of the *avg-strategy*. Note that other smart strategies that maximize network performance may be available; however, we observed that the simple *min-strategy* achieves high throughput and low tail latency in various environments (Section IV-B).

IV. Evaluation

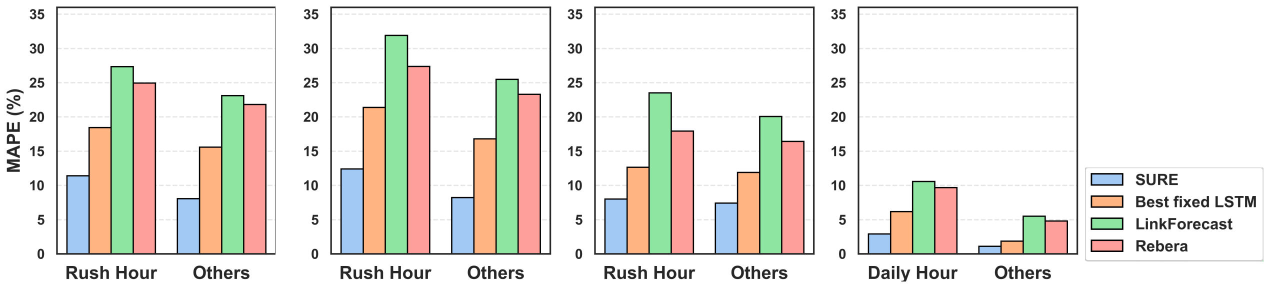
We evaluated the performance of the proposed SURE model based on real-world DC traffic traces collected from NSA 5G networks. In this evaluation, we validated the prediction accuracy and usability of the proposed SURE model based on extensive trace-driven emulations.

A. Prediction Accuracy

Training and Test Datasets. We collected a large number of datasets involving a total of approximately 20 million samples. Based on the date and location of the data collected, we split the collected datasets into training and test datasets at a ratio of 80:20, where 20% of the training datasets were used as validation sets in the learning process. In addition, 50% of the test datasets comprised data samples collected at different locations from the location where the training datasets were collected, which we refer to as untrained-location test datasets. Unless otherwise noted, we utilized the comprehensive test datasets, including the untrained-location test datasets, in most evaluations.

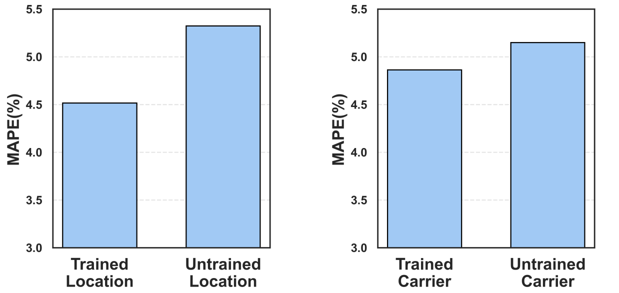
Baseline model. We implemented several uplink resource prediction models for comparison, i.e., LinkForecast [12], Rebera [11], and Best Fixed LSTM (BF-LSTM). The LinkForecast model trains a random forest model with input sequences including TBS, RSRP, and reference signal received quality (RSRQ) gathered over the previous 1000 ms. For Rebera, we used the history-based prediction model, where the history of TBSs is managed based on an exponentially weighted moving average. The BF-LSTM model is implemented based on PERCEIVE [10]. Here, we trained three LSTM models processing the input time-series data with different lengths or time windows of 100 ms, 300 ms, and 1000 ms, and BF-LSTM select the most accurate value among the TBSs inferred by the three LSTM models (LSTM$_{100ms}$, LSTM$_{300ms}$, and LSTM$_{1000ms}$) every 100 ms. We also modified all models to predict both 4G and 5G TBSs because they were designed specifically for 4G networks only, and we set other details (selected features, hyperparameters, and model structures) to be the same as in the referenced studies.

Accuracy in various scenarios. We evaluated the prediction accuracy of the proposed SURE model compared to the baseline models. Here, we considered four unique scenarios classified based on a combination of locations and mobility



(a) Highway driving (b) Downtown driving (c) Downtown walking (d) Stationary

Fig. 8. **Aggregated TBS MAPE comparison between SURE, BF-LSTM, LinkForecast [12], and Rebera [11] for each scenario:** The proposed SURE model achieves the lowest prediction error for both the rush hour and daily hour periods in all scenarios, which indicates that the SURE model can handle a high cell-load environment properly. The proposed SURE model outperforms other models in (a)–(c) mobility scenarios and (d) a stationary scenario, which demonstrates that the proposed model can effectively handle cell switching and handover.



(a) MAPE of trained and untrained location

(b) MAPE of trained carrier(A) and untrained carriers(B and C)

Fig. 9. **Generalizability on test sets with different locations or carriers:** (a) shows that SURE can predict well on test datasets collected in different locations that were not used for training. In (b), we first trained SURE only on datasets for carrier A, and then evaluated SURE on the test sets for carriers B and C.

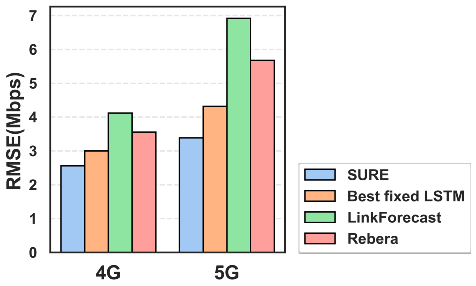


Fig. 10. **Average 4G/5G prediction RMSE:** SURE achieves the lowest prediction error for both 4G and 5G networks. The compared methods exhibit larger errors in 5G prediction.

states, i.e., highway driving, downtown driving/walking, and stationary scenarios. Fig. 8 shows the mean absolute percentage error (MAPE) values between the measured and expected throughput in all scenarios. The expected throughput was estimated based on the sum of the 4G and 5G TBSs predicted by the proposed SURE model or the baseline models. To highlight the impact of cell load on uplink radio resource prediction, we show the MAPE values of specific time periods with high cell loads: rush hours (morning and evening) in the highway/downtown scenarios and daily hours (morning, afternoon, and evening) in the stationary scenario. The proposed SURE model achieved the highest accuracy in high mobility scenarios (highway driving and downtown driving), and it demonstrated excellent performance in high cell-load environments (rush hours and daily hours). We found that the proposed SURE model achieved approximately 8.96%, 19.48%, and 15.06% lower MAPE values than the BF-LSTM, LinkForecast, and Rebera models, respectively.

Generalization to different locations or carriers. We also validated the generalizability of the proposed SURE model by evaluating accuracy based on the trained and untrained test datasets. Fig. 9a shows the prediction performance of SURE for the trained-location and untrained-location test datasets. As can be seen, the proposed SURE predicted the uplink throughput at untrained locations precisely while maintaining a low MAPE value of 5.32%. In addition, to further evaluate generalizability, we trained the proposed SURE model using training datasets for a single carrier (carrier A) and measured the MAPE when it predicted the uplink throughput based on the test datasets corresponding to the trained carrier (carrier A) or untrained carriers (carriers B and C). As shown in Fig. 9b, the SURE model exhibits excellent generalizability by achieving a MAPE value of 5.14%, even on the test datasets for untrained carriers.

Detailed Analysis of SURE. To demonstrate the performance of the proposed SURE model extensively, we examined the performance for both 4G and 5G uplink prediction tasks. As shown in Fig. 10, SURE achieves the highest accuracy for both 4G and 5G uplink prediction because it effectively considers 5G DC operations based on our novel features (4G and 5G Tx-Powers). Here, we consider RMSE rather than MAPE because it is difficult to calculate MAPE values for 5G throughput that frequently becomes zero (Section II-D). In addition, Fig. 11 shows an example of severely fluctuating throughput measured in the highway driving scenario and the instantaneous predictive throughput of the SURE model and baseline models against the fluctuation. As can be seen, the proposed SURE model forecasts the ground truth throughput much more accurately and promptly than the baseline models.

Effect of Self-Attention. The impact of each part of the input time-series data on uplink prediction depends on the characteristics of the dynamically changing cellular environment, e.g., cell-load dynamics, channel conditions, and handovers. To validate this based on real-world traces, we analyzed the prediction accuracy of the three LSTM models used in the BF-LSTM method in various scenarios. Here, we found that the best-performing model depends on the dynamicity of the uplink resources in each scenario. Fortunately,

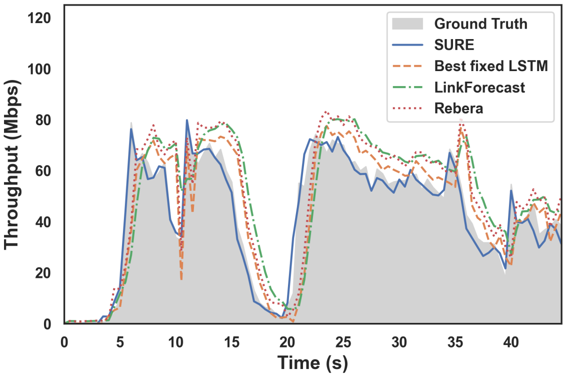


Fig. 11. **Predicted uplink throughput traces of SURE and other prediction models on 5G DC traffic:** SURE effectively predicts uplink throughput even in the presence of drastic fluctuations.

the Transformer-based SURE model obtains highly accurate predictions against this dynamicity because the self-attention mechanism in SURE properly focuses on different parts of the input sequences in response to changes in the cellular environment. Fig. 12 shows the accuracy of the three LSTM models and SURE for the three scenarios. In addition, attention maps are shown, which describe the average attention for each part of the input sequences estimated while SURE performs the prediction process. For the stationary scenario, where LSTM$_{100ms}$ achieves the highest accuracy, SURE gives high attention to only recent parts (0–100 ms) of the input time-series data, and other parts (0–300 ms or 0–1000 ms) receive more attention in the downtown walking and highway driving scenarios, where LSTM$_{300ms}$ and LSTM$_{1000ms}$ work well, respectively. From these results, we conclude that the high performance of SURE is due to the self-attention mechanism. In addition, long-term information and the latest information are both important to realize precise uplink prediction against severe fluctuations in the uplink radio resources. We also expect that the attention-map-based explainability of SURE will be helpful for future traffic analysis in both 5G and 6G networks.

B. Usability

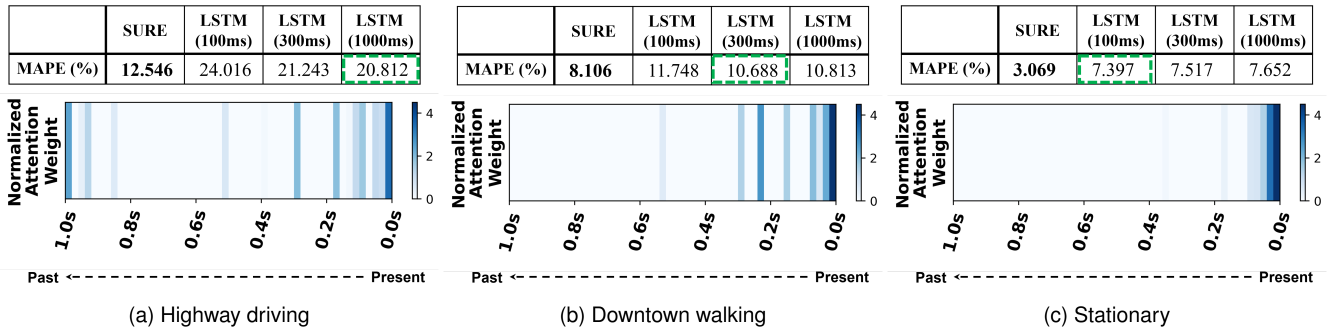
Emulation Setup. We performed extensive emulations to evaluate the performance of the congestion control protocols integrated with the proposed SURE model. Here, we used the mahimahi trace-driven network emulator [16] and generated uplink traffic over the emulated 5G networks based on real-world DC traces [31]. We integrated SURE into well-known congestion control protocols (CUBIC [17] and BBR [18] in QUIC [32]) and Indigo [19], which control network congestion based on two ML models, i.e., LSTM and DAgger [33]. To compare the SURE-based congestion control protocols with various other protocols [34]–[37], we also utilized Pantheon [19], which is a widely used congestion control evaluation platform. In these emulations, the one-way propagation delay and link-queue size were set to 30 ms and 1.5 × bandwidth-delay product (BDP), respectively (unless otherwise noted).

Performance in various scenarios. Fig. 13 shows the performance of various congestion control schemes for highway driving, downtown walking, and stationery scenarios. To increase the reliability of this evaluation, for each scenario and congestion control scheme, we repeated the emulation 10 times with different 50-s traces and show the average throughput and 95^{th} percentile one-way delay. We observed different degrees of uplink channel fluctuations in each scenario. The uplink radio resources allocated to the UE fluctuate most severely in the highway driving scenario, whereas it is the most consistent in the residence-stationary scenario. In all scenarios, the SURE-based protocols exhibit excellent throughput with low latency compared to various congestion control protocols. We found that SURE-QUBIC achieves approximately 49.74% lower tail latency than QUBIC, and SURE-Indigo increases throughput by up to 4.66% compared to Indigo. In particular, SURE-BBR significantly improves the throughput and latency of BBR by up to 10.2% and 52.8%, respectively. It is remarkable that SURE significantly improves the performance of existing congestion control schemes while preserving their original objectives, including the high throughput of QUBIC and low latency of Indigo.

Performance with various network configurations. We evaluated the SURE-based congestion control protocols with various one-way propagation delays and link-queue sizes. In this evaluation, the one-way propagation delay varied from 10 to 100 ms, and the queue size varied from 0.5 x BDP to 2 x BDP. Here, we only considered two scenarios, i.e., the highway driving and stationary scenarios. We repeated the emulation 10 times with the same trace for each scenario to clearly demonstrate the impact of one-way propagation delay and link-queue size on congestion control [2]. As shown in Fig. 14, the SURE-based protocols improve the performance of the original protocols in the stationary scenario. As the one-way propagation delay or queue size increases, SURE significantly reduces the tail latency of QUBIC and BBR by 22.67–69.13% and 25.7–59.23%, respectively, without compromising throughput. In addition, SURE-based QUBIC and SURE-based BBR achieve 6.5% and 6.7% higher throughput than CUBIC and BBR, respectively, when the one-way propagation delay is 100 ms. As shown in Fig. 15, CUBIC and BBR cannot work efficiently due to the severely fluctuating uplink radio resources in the highway driving scenario, and the tail latency increases drastically as the queue size increases. However, the proposed SURE model effectively prevents the tail latency of the original protocols from increasing. In addition, the throughput of BBR decreases greatly as the one-way propagation delay grows; however, SURE effectively improves the throughput of BBR by up to 34.5%. In all scenarios, we found that Indigo efficiently controls network congestion against uplink radio resource fluctuations; however, SURE-based Indigo achieves approximately 3.7% higher throughput than Indigo while preserving its excellent latency.

Performance on video streaming. We conducted the evaluation for the effectiveness of SURE in adaptive video streaming which is a real uplink data-intensive application. We utilized tyStream [49], a performance test tool that emulates

[2] Despite using the same trace, we could obtain slightly different results from each emulation due to the randomness of the traffic generation process in the mahimahi emulator.



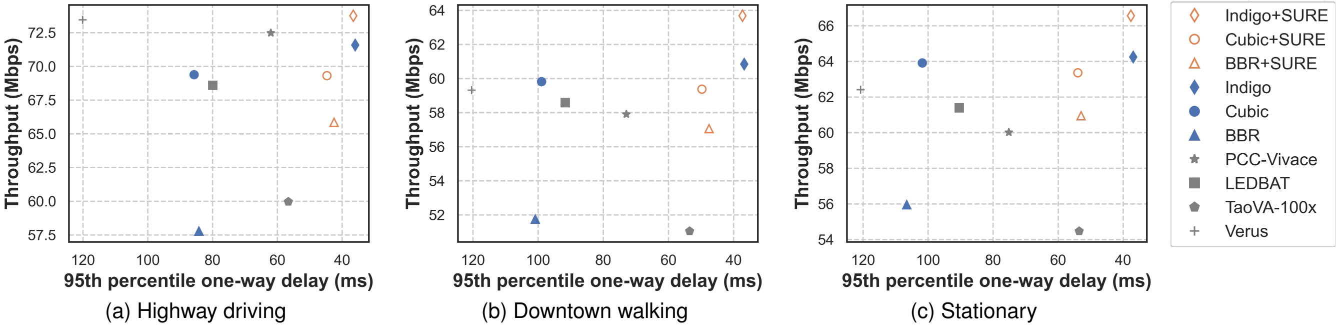
	SURE	LSTM (100ms)	LSTM (300ms)	LSTM (1000ms)
MAPE (%)	**12.546**	24.016	21.243	20.812

	SURE	LSTM (100ms)	LSTM (300ms)	LSTM (1000ms)
MAPE (%)	**8.106**	11.748	10.688	10.813

	SURE	LSTM (100ms)	LSTM (300ms)	LSTM (1000ms)
MAPE (%)	**3.069**	7.397	7.517	7.652

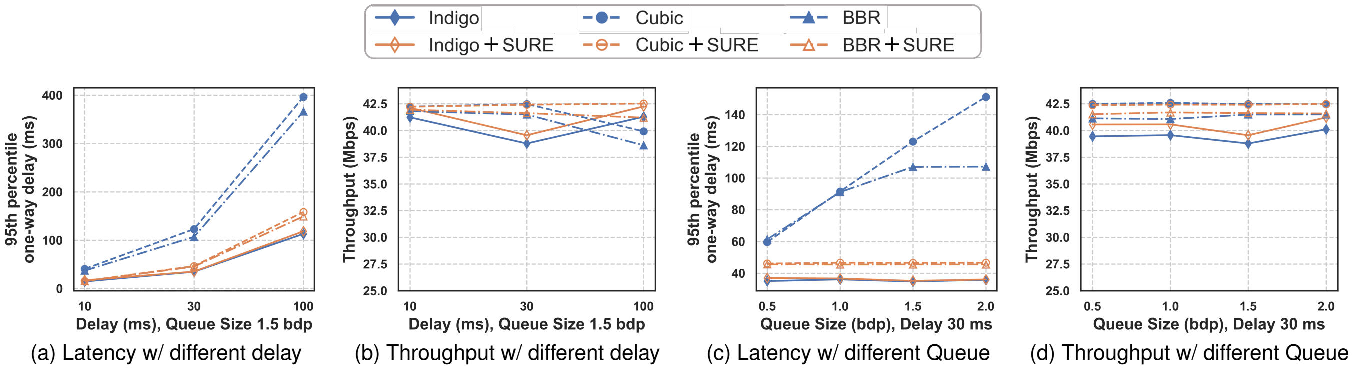
(a) Highway driving (b) Downtown walking (c) Stationary

Fig. 12. **MAPE comparison between LSTM model of PERCEIVE with three different ITW and SURE, and normalized attention weight map estimated by SURE for each scenario:** By utilizing the self-attention mechanism of the Transformer model, high attention weights are imposed on proper parts of the input sequence.



(a) Highway driving (b) Downtown walking (c) Stationary

Fig. 13. **Comparison of various congestion control schemes using Pantheon for each scenario:** We calculated the average throughput and latency from 10 different traces for each scenario. Here, the delay was 30 ms, and the queue size was 1.5 × BDP.



(a) Latency w/ different delay (b) Throughput w/ different delay (c) Latency w/ different Queue (d) Throughput w/ different Queue

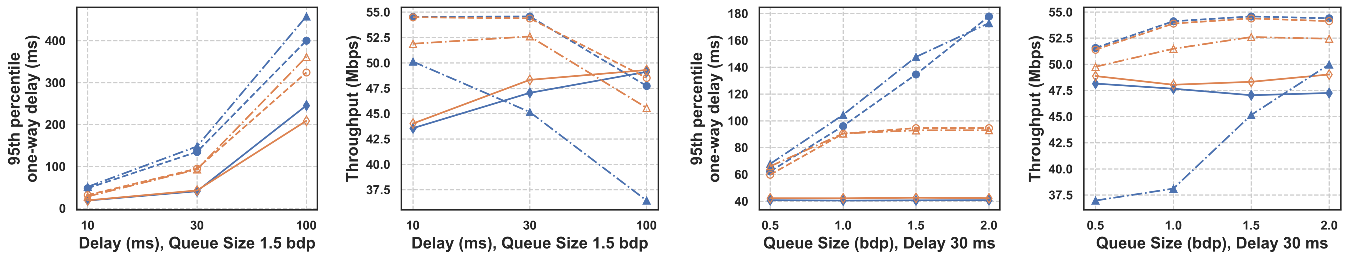
Fig. 14. **Comparison of throughput and latency of original and SURE-based protocols in stationary scenario:** We calculated the average throughput and latency of 10 emulations using the same trace by varying the delay and queue size.

a video streaming environment between an ABR server and a DASH [50] client over wireless networks emulated by mahimahi. The DASH client was set up to fetch bitrate selection decisions from the ABR server using fastMPC [51] as its adaptive streaming algorithm. To conduct the evaluation, we used an 8K video [52] encoded at bit rates of 17, 22, 25, and 30 Mbps, which resolutions are 3840x2160, 5120x2880, 7680x4320, and 7680x4320, respectively. For the 5G network emulation, our highway driving traces are used to represent unstable and fluctuating mobile link conditions. We compared the performance of adaptive streaming when using BBR versus SURE-based BBR in the 5G emulation.

Fig.16a displays the throughput achieved by BBR and SURE-based BBR during video uploading. SURE-based BBR achieves 32.26% higher throughput than BBR because SURE effectively improves the performance of BBR. With this advantage, higher bitrate video chunks (4-second video block) could be selected more, as illustrated in Fig.16b. As a result, the percentage of video chunks transmitted at the highest bitrate (30 Mbps) was 70% when using SURE-based BBR, but only 17.8% when using BBR only.

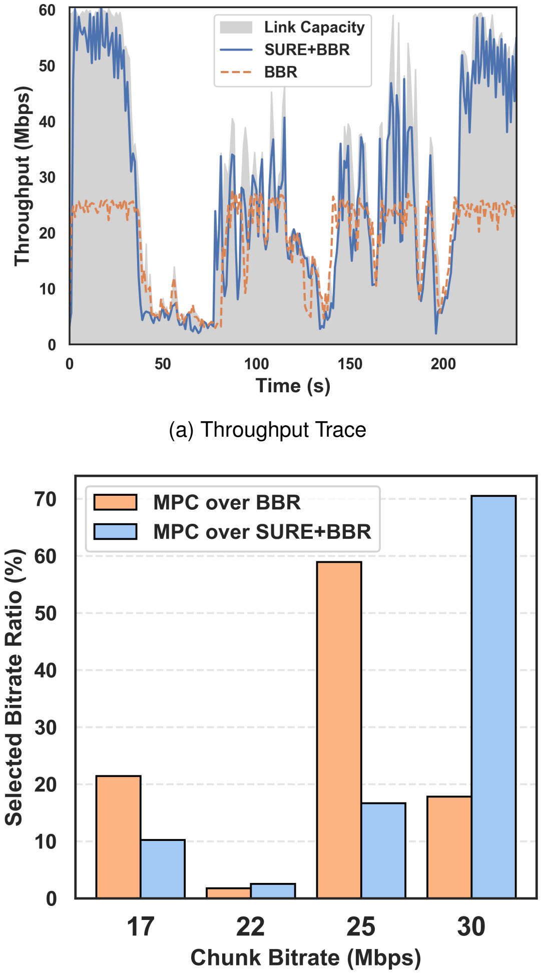
V. Related Work

Cellular Network Prediction. Numerous studies have proposed methods to predict the next network state in cellular networks, e.g., PROTEUS [9] for 3G networks and Rebera [11] and PERCEIVE [10] for 4G networks. QCut [38] demonstrated that packet queuing delay can be reduced significantly by estimating the throughput of the 4G network accurately. PCC Vivace [34] considers 4G networks in its congestion control algorithm. A previous study [7] attempted to improve a video codec by collaborating with the transport layer using deep imitation learning. While these schemes only utilize



(a) Latency w/ different delay (b) Throughput w/ different delay (c) Latency w/ different Queue (d) Throughput w/ different Queue

Fig. 15. **Comparison of throughput and latency of original and SURE-based protocols in highway driving scenario:** We calculated the average throughput and latency of 10 emulations using the same trace by varying the delay and queue size.



(a) Throughput Trace

(b) Selected Chunk Bitrates Ratio

Fig. 16. **Video Streaming Comparison between BBR and SURE based BBR:** Both scheme used same ABR algorithm, fastMPC.

upper-layer information, e.g., throughput and latency, LinkForecast [12] demonstrated the possibility of predicting 4G networks using physical channel information, e.g., RSRP and RSRQ. In addition, PBE-CC [39] optimized TCP congestion control for a 4G downlink channel by providing the physical-layer information to the server. [45] predicted 4G channel quality using LSTM and DNN (Deep Neural Network). Reinforcement learning (RL) has become increasingly promising, and OWL [8] provides a congestion control algorithm based on predicting network conditions, including cellular networks, using an RL agent that utilizes upper-layer information. In a previous study [24], RL was employed to predict networks using physical channel information. However, the methods proposed in these studies were not designed to consider the specific characteristics of 5G networks. In [40] and [41], ML-based prediction models were proposed to forecast cellular radio resources in vehicular networks based on NSA 5G or 6G. [47] used ML for predicting SINR of 5G, and HYPER [46] used ARMA(AutoRegressive Moving Average) model to predict 5G intra-cell bandwidth. However, these studies did not sufficiently consider severe uplink resource fluctuations caused by DC operations in 5G networks. In contrast, the proposed SURE model accurately predicts uplink radio resources in DC-enabled 5G networks based on the lightweight Transformer model, which effectively learns resource allocation patterns with novel features that are closely related to 5G DC.

Cell Load Estimation. Previous studies have attempted to evaluate the cell load, which is also an important aspect of cellular network performance. For example, the piStream method [42] measures the subcarrier-wise energy level to evaluate the cell resource element occupancy and predict the bandwidth that will be given. In addition, CLAW [43] estimates the cellular downlink load using RSRQ, which can be obtained easily by a UE. CASTLE [44] improved the accuracy of estimating the cell load by considering inter-cell interference using a nonlinear support vector machine model. A2T-Boost [48] utilized the ML model for 5G cell selection to minimize handovers and improve network performance in vehicular networks.

VI. Conclusion

In this paper, we have proposed a self-attention-based learning model to predict uplink radio resources in 5G DC. The proposed SURE model realizes accurate predictions with excellent explainability based on extensive measurements of COTS NSA 5G traffic. In addition, we integrated our prediction information into existing congestion control algorithms to prove its usability. The trace-driven evaluation results demonstrate that this integration can improve throughput significantly while maintaining a sufficiently low queuing delay.

Acknowledgments

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00952, Development of 5G Edge Security Technology for Ensuring

5G+ Service Stability and Availability)(No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems)(No. 2019-0-00421, Artifcial Intelligence Graduate School Program(Sungkyunkwan University)).

Furthermore, we would like to express our deep gratitude to Innowireless Co., Ltd. for providing us with the 5G commercial network measurement analysis tool.

References

[1] IMT Traffic Estimates for the Years 2020 to 2030, Standard ITU-RM.2370-0, 2015.

[2] A. Narayanan, X. Zhang, R. Zhu, A. Hassan, S. Jin, X. Zhu, X. Zhang, D. Rybkin, Z. Yang, Z. M. Mao, and F. Qian, "A variegated look at 5G in the wild: performance, power, and QoE implications," in *Proc. 2021 Annual conference of the ACM Special Interest Group on Data communication on the applications, technologies, architectures, and protocols for Computer Communication*, 2021, pp. 610–625.

[3] M. Uitto and A. Heikkinen, "Evaluation of Live Video Streaming Performance for Low Latency Use Cases in 5G," *2021 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, Porto, Portugal, 2021, pp. 431-436, doi: 10.1109/EuCNC/6GSummit51104.2021.9482605.

[4] M. Ghoshal, I. Khan, Q. Xu, Z. J. Kong, Y. C. Hu, and D. Koutsonikolas, "NextG-up: A Tool for Measuring Uplink Performance of 5G Networks," in Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, in MobiSys '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 638–639. doi: 10.1145/3498361.3539694

[5] A. Aliyu et al., "Towards video streaming in IoT Environments: Vehicular communication perspective," Computer Communications, vol. 118, pp. 93–119, 2018, doi: https://doi.org/10.1016/j.comcom.2017.10.003.

[6] M. Boban, C. Jiao and M. Gharba, "Measurement-based Evaluation of Uplink Throughput Prediction," 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), Helsinki, Finland, 2022, pp. 1-6, doi: 10.1109/VTC2022-Spring54318.2022.9860971.

[7] A. Zhou, H. Zhang, G. Su, L. Wu, R. Ma, Z. Meng, X. Zhang, X. Xie, H. Ma, and X. Chen, "Learning to Coordinate Video Codec with Transport Protocol for Mobile Video Telephony," in *Proc. ACM 25th Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2019, pp. 1–16.

[8] A. Sacco, M. Flocco, F. Esposito, and G. Marchetto, "Owl: Congestion Control with Partially Invisible Networks via Reinforcement Learning," in *Proc. IEEE Conference on Computer Communications*, 2021, pp. 1–10.

[9] Q. Xu, S. Mehrotra, Z. Mao, and J. Li, "PROTEUS: Network Performance Forecast for Real-Time, Interactive Mobile Applications," in *Proc. ACM 11th annual international conference on mobile systems, applications, and services*, 2013, pp. 347–360.

[10] J. Lee, S. Lee, J. Lee, S.D. Sathyanarayana, H. Lim, J. Lee, X. Zhu, S. Ramakrishna, D. Grunwald, K. Lee, and S. Ha, "PERCEIVE: Deep Learning-Based Cellular Uplink Prediction Using Real-Time Scheduling Patterns," in *Proc. ACM 18th international conference on mobile systems, applications, and services*, 2020, pp. 377–390.

[11] E. Kurdoglu, Y. liu, Y. Wang, Y. Shi, C. Gu, and J. Lyu, "Real-Time Bandwidth Prediction and Rate Adaptation for Video Calls over Cellular Networks," in *Proc. ACM 7th International Conference on Multimedia Systems*, 2016, pp. 1–11.

[12] C. Yue, R. Jin, K. Suh, Y. Qin, B. Wang, and W. Wei., "LinkForecast: Cellular Link Bandwidth Prediction in LTE Networks," *IEEE Trans. Mobile Computing*, vol. 17, no. 7, pp. 1582–1594, 2018.

[13] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, F. Qian, and Z. L. Zhang, "A First Look at Commercial 5G Performance on Smartphones," in *Proc. ACM Web Conference 2020*, New York, NY, USA: Association for Computing Machinery, 2020, pp. 894–905.

[14] M. Agiwal, H. Kwon, S. Park and H. Jin, "A Survey on 4G-5G Dual Connectivity: Road to 5G Implementation," in *IEEE Access*, vol. 9, pp. 16193–16210, 2021.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *30th Advances in Neural Information Processing Systems*, 2017.

[16] R. Netravali, A. Sivaraman, S. Das, A. Goyal, K. Winstein, J. Mickens, and H. Balakrishnan, "Mahimahi: Accurate Record-and-Replay for HTTP," in *2015 USENIX Annual Technical Conference (USENIX ATC 15)*, 2015, pp. 417–429.

[17] S. Ha, I. Rhee, and L. Xu, "CUBIC: A New TCP-Friendly High-Speed TCP Variant," *ACM SIGOPS Operating Systems Review*, vol. 42, no. 5, pp. 64–74, 2008.

[18] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, "BBR: Congestion-Based Congestion Control," *Communications of the ACM*, vol. 60, no. 2, pp. 58–66, 2017.

[19] F. Y. Yan et al., "Pantheon: the training ground for Internet congestion-control research," in *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, 2018, pp. 731–743.

[20] NR; Medium Access Control (MAC) protocol specification, 3GPP TS 38.321 v16.9.0, July. 2022

[21] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[22] Innowireless XCAL Solo, [Online]. Available: https://www.accuver.com/sub/products/view.php?idx=6

[23] "iPerf - The ultimate speed test tool for TCP, UDP and SCTP," June 2016

[24] M. Chen, R. Li, J. Crowcroft, J. Wu, Z. Zhao and H. Zhang, "RAN Information-Assisted TCP Congestion Control Using Deep Reinforcement Learning With Reward Redistribution," *IEEE Trans. Communications*, vol. 70, no. 1, pp. 215–230, 2022.

[25] NR; Radio Resource Control (RRC); Protocol specification, 3GPP TS 38.331 v16.9.0, July 2022.

[26] MediaTek, "5G NR Uplink Enhancements Better Cell Coverage & User Experience," 2018, [Online]. Available: https://newsletter.mediatek.com/hubfs/mwc/download/ul-enhancements.pdf

[27] ZTE, "5G Uplink Enhancement Technology White Paper," 2020, [Online]. Available: https://www.zte.com.cn/content/dam/zte-site/res-www-zte-com-cn/mediares/zte/files/newsolution/wireless/ran/white_paper/5G_Uplink_Enhancement_Technology_White_Paper.pdf

[28] P. Laarne, M. A. Zaidan, and T. Nieminen, "Ennemi: Non-linear correlation Detection with Mutual Information," SoftwareX, vol. 14, 100686, 2021

[29] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based Framework for Multivariate Time Series Representation Learning," in *Proc. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2114–2124.

[30] S. Holmer, H. Lundin, G. Carlucci, L. De Cicco, and S. Mascolo, "Google Congestion Control Algorithm for Real-Time Communication on the World Wide Web," *IETF Draft*, 2015.

[31] H. Zhang, A. Zhou, R. Ma, J. Lu, and H. Ma. "Arsenal: Understanding Learning-based Wireless Video Transport via In-depth Evaluation," *IEEE Trans. Vehicular Technology*, vol. 70, no. 10, pp. 10832–10844, 2021.

[32] QUIC Working Group [Online]. Available: https://quicwg.org.

[33] S. Ross, G. J. Gordon, and J. A. Bagnell, "No-Regret Reductions for Imitation Learning and Structured Prediction," in *Proc. 14th Intl. Conference on Artificial Intelligence and Statistics*, 2011.

[34] M. Dong, T. Meng, D. Zarchy, E. Arslan, Y. Gilad, P. B. Godfrey, and M. Schapira, "PCC Vivace: Online-Learning Congestion Control," in *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, 2018, pp. 343–356.

[35] Shalunov, Sea, Greg Hazel, Janardhan Iyengar, and Mirja Kuehlewind, "Low Extra Delay Background Transport (LEDBAT)", RFC6817, 2012.

[36] A. Sivaraman, K. Winstein, P. Thaker, and H. Balakrishnan, "An Experimental Study of the Learnability of Congestion Control," *ACM SIGCOMM Computer Communication Review 44(4)*, 2014, pp. 479–490.

[37] Zaki, Yasir, Thomas Po¨tsch, Jay Chen, Lakshminarayanan Subramanian, and Carmelita Gorg, "Adaptive Congestion Control for Unpredictable Cellular Networks," in *Proc. 2015 ACM Conference on Special Interest Group on Data Communication*, 2015, pp. 509–522.

[38] Y. Guo, F. Qian, Q. A. Chen, Z. M. Mao, and S. Sen, "Understanding On-Device Bufferbloat for Cellular Upload," in *Proc. 2016 Internet Measurement Conference*, 2016, pp. 303–317.

[39] Y. Xie, F. Yi, and K. Jamieson, "PBE-CC: Congestion Control via Endpoint-Centric, Physical-Layer Bandwidth Measurements," in *Proc. 2020 Annual conference of the ACM Special Interest Group on Data communication on the applications, technologies, architectures, and protocols for Computer Communication*, 2020, pp. 451–464.

[40] B. Sliwa, H. Schippers and C. Wietfeld, "Machine Learning-Enabled Data Rate Prediction for 5G NSA Vehicle-to-Cloud Communications," in *2021 IEEE 4th 5G World Forum (5GWF)*, 2021, pp. 299–304.

[41] B. Sliwa, R. Adam and C. Wietfeld, "Client-based Intelligence for Resource Efficient Vehicular Big Data Transfer in Future 6G Networks,"

in *IEEE Trans. Vehicular Technology*, vol. 70, no. 6, pp. 5332–5346, 2021.

[42] X. Xie, X. Zhang, S. Kumar, and L. E. Li, “piStream: Physical Layer Informed Adaptive Video Streaming over LTE,” in *Proc. 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 413–425.

[43] X. Xie, X. Zhang, and S. Zhu, “Accelerating Mobile Web Loading using Cellular Link Information,” in *Proc. 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 427–439.

[44] J. Lee, J. Lee, Y. Im, D. Sathyanaryana, P. Rahimzadeh, X. Zhang, M. Hollingsworth, C. Joe-Wong, D. Grunwald, and S. Ha, “CASTLE over the air: Distributed Scheduling for Cellular Data Transmissions,” in *Proc. 17th Annual International conference on Mobile Systems, Applications, and Services*, 2019, pp. 417–429.

[45] N. Diouf, M. Ndong, D. Diop, K. Talla, M. Sarr and A. C. Beye, ”Channel Quality Prediction in 5G LTE Small Cell Mobile Network Using Deep Learning,” 2022 9th International Conference on Soft Computing Machine Intelligence (ISCMI), Toronto, ON, Canada, 2022, pp. 15-20, doi: 10.1109/ISCMI56532.2022.10068487.

[46] Y. Lin, Y. Gao and W. Dong, ”Bandwidth Prediction for 5G Cellular Networks,” 2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS), Oslo, Norway, 2022, pp. 1-10, doi: 10.1109/IWQoS54832.2022.9812912.

[47] A. and M. E. and G. Yu. V. and S. S. Bobrikova Ekaterina and Platonova, “Using Neural Networks for Channel Quality Prediction in Wireless 5G Networks,” Distributed Computer and Communication Networks: Control, Computation, Communications, pp. 132–143, 2022.

[48] A. AlAblani and M. A. Arafah, ”A2T-Boost: An Adaptive Cell Selection Approach for 5G/SDN-Based Vehicular Networks,” in IEEE Access, vol. 11, pp. 7085-7108, 2023, doi: 10.1109/ACCESS.2023.3237851.

[49] tystream, https://github.com/KevinRSX/tystream

[50] DASH.js, https://github.com/Dash-Industry-Forum/dash.js

[51] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, “A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP,” in Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, in SIGCOMM ’15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 325–338. doi: 10.1145/2785956.2787486.

[52] B. Taraghi, H. Amirpour, and C. Timmerer, “Multi-Codec Ultra High Definition 8K MPEG-DASH Dataset,” in Proceedings of the 13th ACM Multimedia Systems Conference, in MMSys ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 216–220. doi: 10.1145/3524273.3532889.

[53] Test dataset and learning code, https://github.com/Doubb/Self-Attention-based-5G-Uplink-Resource-Prediction