Machine Learning and Deep Learning for Throughput Prediction

Dongwon Lee and Joohyun Lee
Department of Electrical and Electronic Engineering
Hanyang University
Ansan, South Korea
{dw2689, joohyunlee}@hanyang.ac.kr

Abstract— Wireless communication contains many fluctuations than wired networks. In this paper, we present several machine learning and deep learning models to predict future network throughput, which is crucial for reducing latency in online streaming services. This paper explains the main components of the throughput prediction system. The throughput prediction model includes data input, data training, and prediction computation parts. This model accepts network throughput for the training data of the model and forecasts future data. We also present the advantages and limitations of utilizing AI models for throughput prediction. Finally, we believe that this study highlights the impact of deep learning techniques for throughput prediction.

Keywords—machine learning; deep learning; throughput prediction;

I. Introduction

Throughput prediction plays an important role in real-time video and audio streaming services that require low-latency network protocols. For example, the throughput prediction can be a solution that can improve quality in video streaming applications such as YouTube or NetFlix [1]. Current video streaming service determines video quality depending on the network quality. Therefore, the throughput prediction is required to guarantee stable Quality of Service (QoS) and Quality of Experience (QoE). In addition, the throughput prediction can assist downloading large files by efficiently handling downloads without adding excessive loads in the network [2]. Previous throughput estimation methods use exponential moving average, arithmetic mean, and the last sample. However, these methods cannot lead to accurate prediction and thus novel methods are required [3].

The throughput in wireless networks quickly fluctuates by numerous factors. Wireless networks have many random characteristics. Besides, radiowave interference and the number of connected users influence the throughput. Therefore, the throughput has various factors that are too complicated to find relationships using traditional models [4]. In this paper, we investigate how machine and deep learning techniques can assist precise throughput prediction. Then, we classify some literature with the components of a basic throughput prediction system.

II. Background

A. Machine and Deep Learning

Machine learning has a broad range of applications, including search recommendation, text recognition, stock forecasting, and so on. Machine learning enables one to learn from data by finding patterns and to classify objects or predict a value. It also attempts to generate algorithms and models that can directly compute data without certain formulas. Current machine learning algorithms can be classified into supervised learning, unsupervised learning, and reinforcement learning. Throughput prediction corresponds to supervised learning because it conducts regression tasks from labeled data, which correspond to the previous throughput.

A machine learning algorithm has resolved many problems by processing raw data and extracting key features that are beneficial to the accurate prediction. For example, in an image classification task, edges and corners are mainly extracted as features from image pixels. Then, machine learning algorithm processes these features to identify the object. However, processing these features is complex for high-resolution images, and takes considerable time to draw results. Therefore, a deep learning algorithm is introduced to solve these complex problems. A deep learning model has several layers to extract features by learning from the input data. As a result, a deep learning algorithm significantly reduces computing time and cost and is often used for classification and regression tasks.

III. Throughput Prediction

The throughput prediction system consists of three components: data collection and preprocessing, data training, and data prediction. The data collection obtains cellular data through network protocols and elaborates collected data to use in the prediction model. The data training creates and updates the model from the training process. The data prediction estimates values that have not been used for training. This system is illustrated in Figure 1.

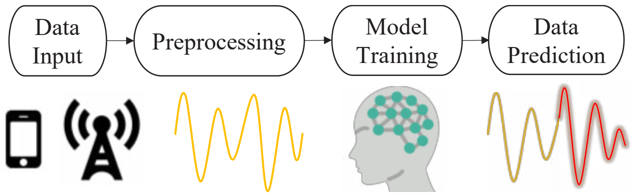

Fig. 1. Components of the throughput prediction system

A. Data Collection and Preprocessing

The data collection obtains various data that is divided into channel-specific (e.g., channel quality indicator (CQI) [5], network-specific (e.g., cell load), application-specific (e.g., application throughput), and context-specific (e.g., mobility mode). These data are collected from the mobile device using its operating system application programming interface (API) (e.g., Telephony). To compute the throughput, congestion control algorithms are typically used: TCP Reno, TCP Vegas, or TCP Cubic [6]. These algorithms compute congestion window (CWND), and the throughput can be obtained by dividing CWND by round trip time (RTT).

In the machine or deep learning algorithm, the input data is preprocessed before training. Preprocessing includes normalization, eliminating outliers, and so on. Normalization performs a transformation that subtracts the mean and divides by the standard deviation, i.e., zero mean and unity standard deviation. This process is also called z-score normalization. Another type of normalization is min-max normalization. It subtracts the minimum value of the dataset and divides it by the difference in the maximum and minimum value of the dataset. Additionally, removing outliers, which are distinct from the remaining data points, is important because they can degrade the accurate training.

B. Data Training and Prediction

The machine or deep learning algorithm is used for data training and prediction. We describe several algorithms that are mainly used for throughput prediction in the literature.

- Random Forest (RF): RF is one of the ensemble learning methods [7]. RF constructs a series of decision trees in the training process. In [8], Schmid et al. use RF with important features such as Reference Signal Receiving Power (RSRP), Reference Signal Received Quality (RSRQ), Reference Signal Strength Indicator (RSSI), and RTT. Khan et al. propose an RF model for real-time throughput prediction for two distinct datasets with different number of features [10].
- Support Vector Regression (SVR): SVR is a version of a support vector machine (SVM) for regression [11]. It follows linear regression, where the output value has a linear relationship with the input value. Wei et al. [12] apply SVR to model the relationship between the future throughput and CQI. They test the model in different moving scenarios, such as static place, walking, and riding on a bus or train. They conclude that the SVR model with parameters of throughput and RSSI performs better than the single parameter SVR model, harmonic mean, stochastic model, and last sample. They also conclude that the SVR model decreases prediction error by a maximum of 26.47% in a static user scenario. Mirza et al. [15] develop and test an SVR-based model investigating the relationship between TCP throughput and path properties including available bandwidth (AB), queuing delay, and packet loss. They conclude that queuing delay and packet loss are both significant for accurate prediction, but AB rarely affects the model performance.
- Long Short-Term Memory (LSTM): LSTM is an improved model of recurrent neural network (RNN). RNN is more suitable for handling sequential or time-series data than other neural network models. B. Wei et al. [13] propose a method for throughput prediction with an RNN model from cellular data. They also evolve the model by replacing RNN with LSTM and thus enhance the performance [14]. They compare the prediction results with the arithmetic mean, harmonic mean, last sample, moving average, hidden Markov model (HMM), a hybrid model of autoregressive model and HMM, and stochastic model. They also add RSSI, Cell ID, time, and location as the input features of the LSTM model to identify the user movement pattern and vary the prediction length with 5, 20, 100, and 200 seconds. The result says that LSTM can precisely estimate throughput than the baseline models. Schmid et al. [9] apply a two-layer bidirectional LSTM model with the input features of throughput, signal-to-interference-plus-noise ratio (SINR), RTT, RSSI, and RSRQ. They compare the performance with SVR, RF, and feedforward neural network models. They discover that the LSTM model is robust to dynamic fluctuation and thus results in the best performance.

We summarize which model is used in each literature in TABLE I. Please remind that other models may be used for comparison rather than RF, SVR, and LSTM.

TABLE I. A SUMMARY OF THE PREDICTION MODEL IN THE LITERATURE

Ref.	Prediction Model		
	RF	**SVR**	**LSTM / RNN**
[8]	○	○	
[9]	○	○	○
[10]	○	○	
[12]		○	
[13]			○
[14]			○
[15]		○	

C. Performance Metrics

There are various metrics to evaluate the prediction accuracy of the model. However, these metrics are generally based on the size of the error, the difference between the estimated and the true value. In general, there exist a lot of estimated values in one prediction set, so it is better to summarize each error into a single value. One common solution is averaging these errors: the mean absolute error (MAE), the root mean squared error (RMSE) is evaluated. For output value y_i and estimated value $\hat{y}_i$, they are defined as the following,

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|^2} \tag{2}$$

The reason for using squared error is that it is easy to detect large errors. The other metrics denote error as a ratio or percentage, the mean absolute percentage error (MAPE), R2 score. They are defined as the following,

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

$$\text{R}^2 = 1 - \frac{\text{MSE}}{\sigma_y^{\ 2}} \tag{4}$$

, where MSE is the mean squared error, the square of RMSE, and $\sigma_y^{\ 2}$ is the variance of y.

Since MAE, RMSE, and MAPE compute the amount of error, these metrics should be small for higher accuracy. On the other hand, R2 score computes the accuracy, and a higher score implies the better performance of the model.

In addition to these metrics, relative error is used for performance metrics. In [9], the mean relative error (MRE) and mean squared relative error (MSRE) are used in addition to MAE and MSE. MRE and MSRE are suitable for finding severe outliers in the prediction data.

We summarize which metric is used in each literature in TABLE II. Please remind that other metrics such as CDF of error or normalized error may be used in some literature.

TABLE II. A SUMMARY OF PERFORMANCE METRICS IN THE LITERATURE

Ref.	Performance Metrics				
	MAE	**MSE (RMSE)**	**MRE**	**MSRE (RMSRE)**	**R2**
[8]	○		○	○	
[9]	○	○	○	○	
[10]	○	○			○
[12]			○	○	
[13]		○	○		
[14]		○	○		
[15]			○		

Acknowledgment

This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C1005126). This work was supported by Institute of Information & communications Technology Planning & evaluation(IITP) grant funded by the Korea government(MSIT) (No.2021-0-01673, Establishment of Medium and Long Term Network Strategy for New Digital Growth).

References

[1] T. Mangla et al., "Video through a crystal ball: Effect of bandwidth prediction quality on adaptive streaming in mobile environments," Proceedings of the 8th International Workshop on Mobile Video. 2016.

[2] C. E. Andrade et al., "Managing massive firmware-over-the-air updates for connected cars in cellular networks," Proceedings of the 2nd ACM International Workshop on Smart, Autonomous, and Connected Vehicular Systems and Services. 2017.

[3] D. Raca et al., "Empowering video players in cellular: Throughput prediction from radio network measurements," Proceedings of the 10th ACM Multimedia Systems Conference. 2019.

[4] Q. He, C. Dovrolis, and M. Ammar, "On the predictability of large transfer TCP throughput," ACM SIGCOMM Computer Communication Review 35.4 (2005): 145-156.

[5] S. N. Donthi, and N. B. Mehta, "Joint performance analysis of channel quality indicator feedback schemes and frequency-domain scheduling for LTE," IEEE Transactions on Vehicular Technology 60.7 (2011): 3096-3109.

[6] H. Jamal, and K. Sultan, "Performance analysis of tcp congestion control algorithms," International journal of computers and communications 2.1 (2008): 18-24.

[7] A. Liaw, and M. Wiener, "Classification and regression by randomForest," R news 2.3 (2002): 18-22.

[8] J. Schmid, M. Schneider, A. Höß and B. Schuller, "A Comparison of AI-Based Throughput Prediction for Cellular Vehicle-To-Server Communication," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019, pp. 471-476, doi: 10.1109/IWCMC.2019.8766567.

[9] J. Schmid, M. Schneider, A. HöB and B. Schuller, "A Deep Learning Approach for Location Independent Throughput Prediction," 2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE), 2019, pp. 1-5, doi: 10.1109/ICCVE45908.2019.8965216.

[10] M. A. Khan, et al., "Real-time throughput prediction for cognitive Wi-Fi networks," Journal of Network and Computer Applications 150 (2020): 102499.

[11] H. Drucker et al., "Support vector regression machines," Advances in neural information processing systems 9 (1997): 155-161.

[12] B. Wei et al., "A History-Based TCP Throughput Prediction Incorporating Communication Quality Features by Support Vector Regression for Mobile Network," 2017 IEEE International Symposium on Multimedia (ISM). IEEE, 2017.

[13] B. Wei, M. Okano, K. Kanai, W. Kawakami and J. Katto, "Throughput Prediction Using Recurrent Neural Network Model," 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), 2018, pp. 107-108, doi: 10.1109/GCCE.2018.8574877.

[14] B. Wei, W. Kawakami, K. Kanai, J. Katto and S. Wang, "TRUST: A TCP Throughput Prediction Method in Mobile Networks," 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 1-6, doi: 10.1109/GLOCOM.2018.8647390.

[15] M. Mirza et al., "A machine learning approach to TCP throughput prediction," ACM SIGMETRICS Performance Evaluation Review 35.1 (2007): 97-108.