ULTRA: UNet Helps Transformer to Forecast the network states in 5G network

Anonymous Authors[1]

Abstract

Wireless communication is in strong demand for high throughput and high reliability, while dynamic scenarios bring challenges to these demands. Prediction of network states estimates the future state of the varying channel, which could provide the necessary information for the receiving terminal to adjust transmission strategies accordingly. We propose a transformer-based model called ULTRA to predict the network states. In the proposed model, self-attention is exploited to pursue features between variates, and a trend extractor is introduced to pursue local and global temporal features. We implement comprehensive experiments and an ablation study. The results with real-world data demonstrate that our model outperforms state-of-the-art models.

1. Introduction

With the development of 5G, many applications such as high definition video calls and AR/VR become possible (Andrews et al., 2014). In dynamic scenarios, due to factors such as base station handovers and fluctuations in channel quality, there is significant uncertainty in bandwidth and latency (Zhang et al., 2017). This is especially true in complex urban environments, where phenomena such as multi-path propagation and nonlinear sight (NLOS) propagation make wireless signals vulnerable to fading and interference (Rappaport et al., 2017). In addition, high-speed mobility and frequent handovers can cause delays and packet loss. Forecasting network states in dynamic scenarios helps to adapt the transmission strategy (Yin et al., 2015).

In recent years, numerous deep learning models have been developed for forecasting network states. Long Short-Term Memory (LSTM), a variant of recurrent neural networks, was employed in (Schmid et al., 2019) to enable location-independent throughput prediction. The study in (Raca et al., 2020) further investigated throughput prediction using LSTM, in conjunction with other machine learning algorithms such as Random Forest (RF) and Support Vector Regression (SVR). In (Yue et al., 2017), the performance of LSTM was evaluated against K-Nearest Neighbors (KNN), SVR, Ridge Regression, RF, and ARIMA using the dataset introduced in (Elsherbiny et al., 2020). Moreover, the spatio-temporal variability of network throughput was modeled in (Qu et al., 2020) through a hybrid architecture combining LSTM and Convolutional Neural Networks (CNN).

Following the introduction of the Transformer framework (Vaswani et al., 2017), several Transformer-based models have been proposed for time series prediction tasks. Informer (Zhou et al., 2021) employs the attention mechanism to capture dependencies across time steps and utilizes a multi-layer perceptron (MLP) to extract inter-variable features. Subsequent works, such as (Zhang & Yan, 2023; Zhou et al., 2022; Wu et al., 2021), build upon Informer by introducing enhanced attention mechanisms aimed at capturing more precise temporal relationships. However, a recent finding (Zeng et al., 2023) indicates that simple MLP-based models can outperform these complex Transformer-based architectures, as the attention mechanism may not effectively capture temporal order. iTransformer (Liu et al., 2023) addresses this limitation by inverting the input dimensions and applying the attention mechanism along the variable dimension.

Previous studies (Yin et al., 2015; Sun et al., 2016; Qiao et al., 2022) have primarily focused on forecasting throughput, treating it as a univariate network state prediction task. This work proposes a Transformer-based model designed to forecast multivariate network states. Drawing inspiration from iTransformer, we employ the attention mechanism to capture inter-variable dependencies and adopt the UNet architecture (Ronneberger et al., 2015) for temporal feature extraction. Originally developed for image segmentation, UNet leverages multi-level convolutional layers to extract semantic information across multiple scales. Given the presence of both slow and fast fading phenomena in network states, UNet is well-suited for capturing temporal features over varying time scales. The main contributions of this

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

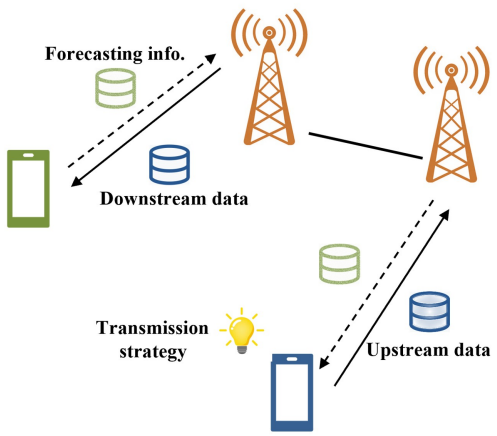


Figure 1. System overview

paper are summarized as follows:

- **Development of the Transformer-based architecture ULTRA** for multi-variant network states prediction, utilizing the attention mechanism to extract variant features of network states.
- **Design of the one-dimension UNet** as the extractor of multi-period temporal features of network states.
- **Comprehensive performance evaluation and ablation study** through experiments with real-world data, demonstrating ULTRA's capability to reduce prediction error by 5%.

2. System Model

A real-time video communication (RTVC) system using the RTCP protocol is considered as shown in Figure 1. Although RTVC is a two-orientation process, we might as well consider a one-orientation communication first. Communicating devices are labeled as device A and device B. In a time slot, device A sends a few packets to device B which records states of the channel, such as throughput, latency and jitter. The information contained in the time slot is very limited. However, the history network states stored in device B provide sufficient information to forecast states in a future time window. Then device B sends the forecasting states back to device A, which would help device 1 develop intelligent transmission strategies in the following time slots.

At each time slot, the receiver records the network states. The history network states is denoted as $\boldsymbol{s}_t \in \mathbb{R}^{N_1 \times T_1}$, where N_1 is the number of variates and T_1 is the length of the history observation slots. Forecast network states is indicated as $\hat{w}_t \in \mathbb{R}^{N_2 \times T_2}$ and ground truth is denoted as $w_t \in \mathbb{R}^{N_2 \times T_2}$, where N_2 is the number of variates and T_2 is the length of future observation slots. The forecast network states is a function of the history network states $\hat{w}_t = g_\theta(\boldsymbol{s}_t)$. The optimization target is to minimize the $\sum_{t=1}^{T_m} \|w_t - \hat{w}_t\|^2$, where T_m is the number of samples. This is a high-dimensional regression problem from one matrix to another. Consequently, the efficient design of $g_\theta(\cdot)$ should take into account the characteristics of the state matrix.

3. Method

Network state forecasting is to find a mapping from history states to future states. Extraction of temporal and variate features is the key process of the model. Given the assumption that temporal features and variate features are independent, it is reasonable to connect two feature extractors in series. The model architecture is consisted of the transformer module and the UNet module, which is shown in Figure 2.

3.1. Transformer Module

There is no order in variates of network states so the variates could be treated as a set. Experiments and analysis in (Zeng et al., 2023) demonstrate that forecasting output of Transformer is not susceptible to the order of series input.

Therefore, the extractor of variate features adopts the encoder-only architecture of Transformer (Vaswani et al., 2017), including embedding, attention, feed-forward network (FFN) and layernorm blocks. The embedding block converts the series of the same variate to the tokens $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N] \in \mathbb{R}^{N \times D}$ by MLP, where N is the the number of variate and D is the embedding length. The self-attention module works on the dimension of variate and utilizes linear projections to get queries, keys and values $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d_k}$, where d_k is the projected dimension. Each score map is formulated as $\mathbf{A} = \mathbf{Q}\mathbf{K}^\top / \sqrt{d_k} \in \mathbb{R}^{N \times N}$ which presents the multivariate correlations between paired variate tokens. Highly correlated variate will be more weighted for the next representation interaction with values $\mathbf{V}$. The FFN is leveraged on the series representation of each variate token by applying MLP and the layer normalization is applied to the series representation of individual variate in our model.

3.2. UNet Module

From the temporal perspective of network states, short-term fluctuation is always violent and long-term trend is more moderate, which is caused by fast fading and slow fading. A single-scale convolution kernel fails to extract short-term and long-term features at the same time.

The extractor of temporal features adopts multi-level archi-

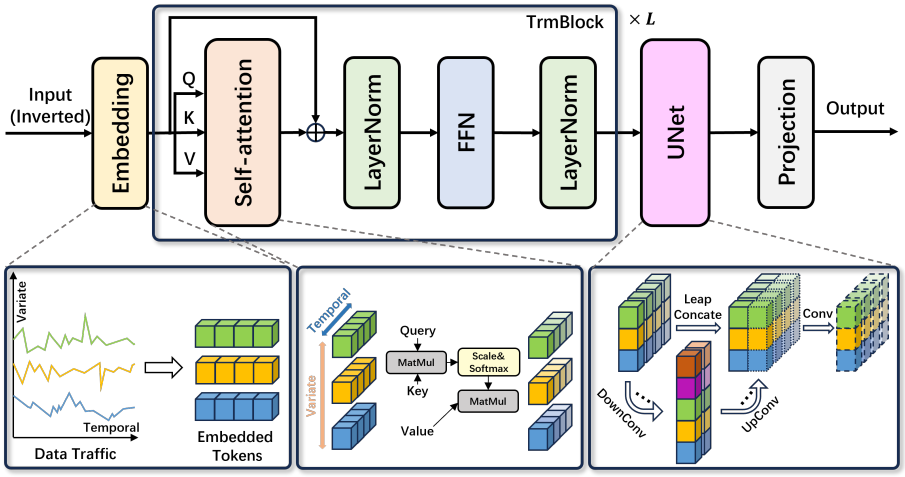


Figure 2. The architecture of ULTRA

tecture of UNet (Ronneberger et al., 2015) which contains down-convolution (DC) and up-convolution (UC). The shallow layers of DC are applied to extract locally temporal features like fluctuation and the deep layers are applied to extract globally temporal features such as periods and trends. In each layer there are multiple convolution kernels to ensure diversity of representation. UC is a restoration of the forecasting network states by levels. Due to lost of features in DC as a down-sampling process, feature fusion is implemented by concatenating and convolution. Notablly, all convolution layers in the UNet module is one-dimension. Consequently, weights are shared for series representation of different variates in the same layer.

Models		**ULTRA**		iTransformer		Informer		LSTM	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
SVF	12	**0.137**	**0.200**	0.146	0.211	0.240	0.285	0.174	0.229
	24	**0.232**	**0.262**	0.251	0.274	0.328	0.315	0.253	0.271
	36	**0.290**	**0.292**	0.291	0.293	0.408	0.358	0.303	0.298
	48	0.358	0.327	**0.351**	**0.319**	0.428	0.355	0.361	0.325
MVF	12	**0.258**	**0.225**	0.284	0.244	0.455	0.402	0.292	0.257
	24	**0.335**	**0.269**	0.357	0.283	0.487	0.406	0.354	0.300
	36	**0.378**	**0.295**	0.403	0.311	0.523	0.442	0.392	0.319
	48	**0.409**	**0.310**	0.437	0.323	0.532	0.435	0.409	0.311
Count		**14**		2		0		0	

Table 1. Full forecasting results

4. Experiments

4.1. Dataset

We select smart phone as the communication device to collect communication data in mobile scenarios. We keep one device in doors and keep the other device in mobility. These two devices are in real-time video communications and stay in 5G network. The dataset includes the traces of network state information and the information set is denoted as $\mathbb{Z}$. network states information (throughput, latency and jitter) is input of all methods. Due to the discrepancy of sampling rate between the communication module and sensors, the dataset unifies the sampling rate as 0.5 per second and consists of 120 thousand samples.

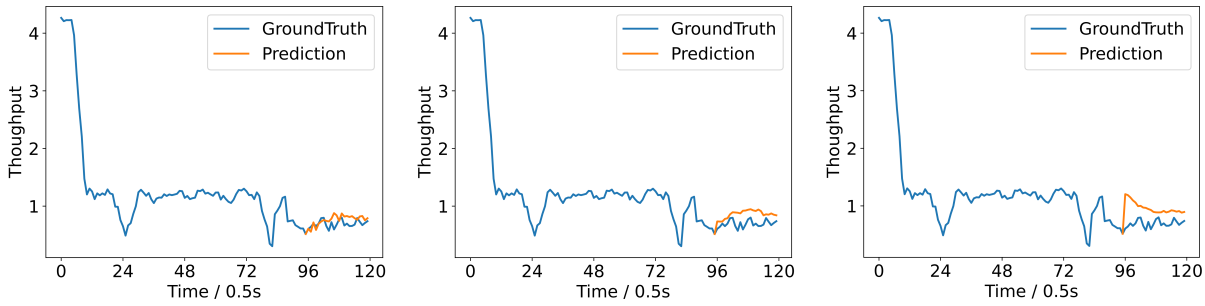
4.2. Experimental details

The proposed model has been tested on the collected dataset. It consists of a training set, a validation set and a test set with the proportions of 70%, 20% and 10% data. We have selected 3 network state forecasting methods as comparison, including iTransformer, Informer and LSTM. Then we do ablation studies to demonstrate the effectiveness of feature extractors of our model.

Inputs of the models are pre-processed as zero-mean normalized time series. We set the forecasting length T in grid, i.e., $\{12, 24, 36, 48\} \times 0.5$s. We use MSE and MAE as the evaluation metric, where MSE is the optimization objective and MAE presents the the gap between prediction and the ground truth more intuitively. All the models are trained and tested on Nvidia GeForce RTX 3090 GPUs.

4.3. Results

To verify the performance of our model comprehensively, we compare the proposed model ULTRA with three benchmarks in both single-variate forecasting (SVF) and multivariate forecasting (MVF). The benchmarks include two transformer-based models and a conventional RNN-based model. SVF only requires the extractor of temporal features and MVF requires both extractors. The results are shown in Table 1. The forecast error of Informer is much higher than that of LSTM, which shows that the self-attention mecha-



(a) ULTRA (b) iTransformer (c) LSTM

Figure 3. Presence of prediction results.

nism is not suitable to be deployed directly on the temporal dimension of network states. MSE of ULTRA is lower than iTransformer by 5% and the MAE of ULTRA is lower than iTransformer by 3.7%, which presents our model as state of the art.

Models		**ULTRA**		iTrans+conv		MLP+UNet		MLP+conv	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MVF	12	**0.258**	**0.225**	0.261	0.228	0.306	0.248	0.290	0.245
	24	**0.335**	**0.269**	0.348	0.277	0.341	0.276	0.350	0.278
	36	**0.378**	**0.295**	0.390	0.298	0.380	0.296	0.389	0.299
	48	**0.409**	**0.310**	0.418	0.312	0.416	0.311	0.419	0.316
Count		**8**		0		0		0	

Table 2. Ablation study

To validate the rational business of ULTRA components, we provide detailed ablation studies by experiments of replacing components. The transformer block is replaced by MLP and the UNet block is replaced by convolution layers with single-scale kernel. The results are listed in Table 2. ULTRA that utilizes the attention mechanism on the variate dimension and UNet structure on the temporal dimension generally achieves the best performance.

The presence of prediction results are shown in Figure 3. Future network states of 24 time slots is forecast with the history network states of past 96 time slots. The predicted receiving bitrate is selected for presence. The results show that ULTRA is able to predict not only the slow rising trend but also the short-term fluctuation. As a comparison, prediction of short-term fluctuation is weakness of iTransformer and LSTM could not predict the trend well.

Space and time consumption is always a concern of Transformer-based models. In the transformer block, the complexity of the embedding layer is $\mathcal{O}(nmd)$, where n, m, d separately denote the variate number, the look-back length and the embedding length of input. The complexity of attention is $\mathcal{O}(n^2d)$. In the UNet block, the complexity of the convolution layers is $\mathcal{O}(kn^2d)$, where k denotes the

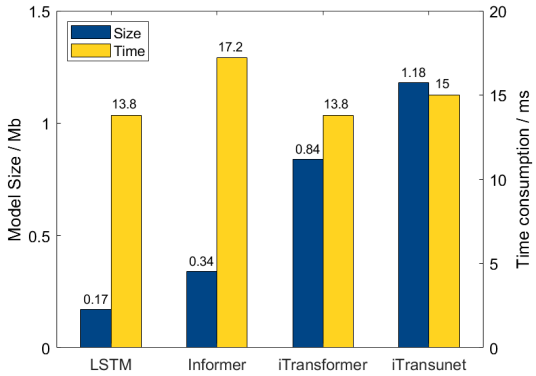


Figure 4. Consumption analysis

kernel number. Therefore, the total complexity of ULTRA is $\mathcal{O}(nd((k+1)d+m))$ which is related to the dimension of input and parameters of the neural network. We record the average sizes and reference time of each model given the same input as shown in Figure 4. ULTRA get improved performance at the cost of rise of the model size and reference time. However, space and time consumption of ULTRA is still on a low level (1.18Mb and 15ms per sample).

5. Conclusion

In this work, we have studied the characteristics of network states in mobile video calls. We propose a transformer-based model (ULTRA) for the long-term prediction of network states. Specifically, we exploit self-attention for variant feature extraction and one-dimensional UNet for temporal feature extraction. Experimentally, we demonstrate our model's ability to reduce prediction errors by 5% and achieve the SOTA level. Furthermore, we show the computation cost and inference efficiency of ULTRA.

References

Andrews, J. G., Buzzi, S., Choi, W., Hanly, S. V., Lozano, A., Soong, A. C., and Zhang, J. C. What will 5g be? *IEEE Journal on selected areas in communications*, 32 (6):1065–1082, 2014.

Elsherbiny, H., Abbas, H. M., Abou-zeid, H., Hassanein, H. S., and Noureldin, A. 4g lte network throughput modelling and prediction. In *GLOBECOM*, pp. 1–6, 2020.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. itransformer: Inverted transformers are effective for time series forecasting. In *arXiv preprint arXiv:2310.06625*, 2023.

Qiao, C., Li, G., Ma, Q., Wang, J., and Liu, Y. Trace-Driven Optimization on Bitrate Adaptation for Mobile Video Streaming. *IEEE Trans. Mobile Comput.*, 21(6): 2243–2256, June 2022. ISSN 1536-1233, 1558-0660, 2161-9875.

Qu, J., Liu, F., Ma, Y., and Fan, J. Temporal-spatial collaborative prediction for lte-r communication quality based on deep learning. *IEEE Access*, 8:94817–94832, 2020.

Raca, D., Zahran, A. H., Sreenan, C. J., Sinha, R. K., Halepovic, E., Jana, R., and Gopalakrishnan, V. On leveraging machine and deep learning for throughput prediction in cellular networks: Design, performance, and challenges. *IEEE Commun. Mag.*, 58(3):11–17, 2020.

Rappaport, T. S., Xing, Y., MacCartney, G. R., Molisch, A. F., Mellios, E., and Zhang, J. Overview of millimeter wave communications for fifth-generation (5g) wireless networks—with a focus on propagation models. *IEEE Transactions on antennas and propagation*, 65(12):6213–6230, 2017.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention (MICCAI)*, pp. 234–241. Springer, 2015.

Schmid, J., Schneider, M., HöB, A., and Schuller, B. A deep learning approach for location independent throughput prediction. In *Proc. IEEE Int. Conf. Connected Vehicles Expo*, pp. 1–5, 2019.

Sun, Y., Yin, X., Jiang, J., Sekar, V., Lin, F., Wang, N., Liu, T., and Sinopoli, B. Cs2p: Improving video bitrate selection and adaptation with data-driven throughput prediction. In *Proc. ACM SIGCOMM*, pp. 272–285, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NIPS*, volume 30, 2017.

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *NIPS*, volume 34, pp. 22419–22430, 2021.

Yin, X., Jindal, A., Sekar, V., and Sinopoli, B. A control-theoretic approach for dynamic adaptive video streaming over http. In *Proc. ACM SIGCOMM*, pp. 325–338, 2015.

Yue, C., Jin, R., Suh, K., Qin, Y., Wang, B., and Wei, W. Linkforecast: Cellular link bandwidth prediction in lte networks. *IEEE Trans. Mobile Comput.*, 17(7):1582–1594, 2017.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Zhang, H., Liu, N., Chu, X., Long, K., Aghvami, A.-H., and Leung, V. C. Network slicing based 5g and future mobile networks: Mobility, resource management, and challenges. *IEEE communications magazine*, 55(8):138–145, 2017.

Zhang, Y. and Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *ICLR*, 2023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, volume 35, pp. 11106–11115, 2021.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *ICML*, pp. 27268–27286. PMLR, 2022.