A Transformer-based Modelling Approach for Robust QoE Estimation in Video Streaming

MohammadAli Hamidi, Simone Porcu, Alessandro Floris, and Luigi Atzori
DIEE, University of Cagliari, 09123 Cagliari, Italy
CNIT, University of Cagliari, 09123 Cagliari, Italy
{mohammadali.hamidi, simone.porcu, alessandro.floris84, l.atzori}@unica.it

***Abstract*—Network management is crucial to ensuring adequate Quality of Experience (QoE) for all connected users. In recent years, artificial intelligence (AI) has been leveraged to support the development of accurate QoE prediction models. In this paper, we leverage the learning characteristics of transformer architectures to implement a novel transformer-based model for estimating the QoE of video streaming services, the most consumed media services on the Internet. Dedicated steps have been defined to collect, encode, and sequentialize the data concerning streaming session-related key performance indicators (KPIs). These steps are needed to prepare the data in a sequential form appropriate for the transformer's learning processes. The model has been trained using two open datasets from the ITU-T P.1203 standardization procedure, including 82 videos (watched and rated on mobile and PC devices) impaired by video quality switching, delay, and stalling events. The proposed model outperforms the P.1203 model in predicting the QoE rated on both mobile and PC devices in terms of root mean square error (RMSE), Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC). Moreover, our model achieved robust performance in the cross-device scenario, i.e., it accurately estimates the QoE of a video watched on a device different from that used for training the model.**

***Index Terms*—Transformer model, Quality of Experience, Video streaming, Objective QoE model, Machine Learning.**

I. Introduction

In today's rapidly evolving digital landscape, the seamless delivery of multimedia content has become integral to our daily lives. From streaming services and online gaming to virtual communication platforms, the demand for high-quality multimedia experiences has surged exponentially. However, ensuring user satisfaction in these dynamic environments presents a multifaceted challenge. Quality of Experience (QoE) encapsulates the holistic perception of users concerning the quality of multimedia services, encompassing various factors, such as audiovisual quality, interactivity, responsiveness, and overall satisfaction [1].

In recent years, there has been increasing interest in measuring the QoE perceived by the user in both objective and subjective ways. Objective QoE models are computational models designed to estimate the perceived quality of multimedia content without relying on direct subjective feedback from users. Instead, these models are built using ground-truth subjective feedback collected during well-controlled subjective assessments. Nevertheless, users have diverse preferences, expectations, and contexts (e.g., cultural background, and past experiences) that may influence their perceptions of multimedia quality [2].

Capturing and modeling these individual differences remains a challenge. Therefore, the goal is to develop a QoE estimation model that effectively accounts for the variability of the considered scenario, which may include, for instance, diverse network conditions and device characteristics. Although today's methods utilize machine learning (ML) techniques, such as auto-regressive model [3], deep neural networks [4], and ensemble learning [5], we still see that these QoE estimation models have the limitation of adapting to diverse and dynamic contexts. Hence, they are prone to a lack of generalization and over-fitting.

One promising approach to tackle the aforementioned challenges is the transformer deep learning architecture [6]. Transformers have shown great success in natural language processing (NLP) tasks, including sentiment analysis [7] and language modeling [8]. Transformers excel at capturing long-range dependencies between different data points by relying on a "multi-head attention mechanism" that allows them to focus on the most relevant parts of the input data for each specific prediction. This helps them discover complex relationships between different variables, even if they are far apart in the sequence. These characteristics make transformers well-suited for analyzing the impact of variability in short and long multimedia streaming sessions on the user's QoE, which may be due, for example, to diverse occurrences of video quality switches and stalling events. Moreover, the attention mechanism could learn differences in QoE perception of the same streaming sessions on different devices or contexts. These considerations lead us to believe that a transformer-based approach can produce more robust and accurate QoE estimators by incorporating well-suited sequential data.

In this paper, we investigate a transformer-based approach to QoE modeling. In particular, we designed a novel transformer-based QoE estimation model for video streaming services. Accordingly, we propose a specific workflow to collect, encode, and sequentialize the data concerning video and streaming

This work has been partially supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP C29J24000300004, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"), by the European Union under the NRRP of NextGenerationEU, "Sustainable Mobility Center" Centro Nazionale per la Mobilità Sostenibile, CNMS, CN_00000023, and by the NRRP - M4C1 - Inv. 3.4 and M4C1 - Inv. 4.1, Ministerial Decree no. 351/2022.

session-related key performance indicators (KPIs). These steps are needed to prepare the data in a sequential form appropriate for the learning processes carried out by the transformer. For training the model, we have considered two open datasets from the ITU-T P.1203 standardization procedure, including a total of 82 video sequences impaired by video quality switching, delay, and stalling events [9]. These videos are accompanied by subjective quality feedback provided for PC and mobile devices. We have computed the QoE estimation performance of the proposed transformer-based model in terms of root mean square error (RMSE), Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC). For comparison, we have reported the performance achieved by the official ITU-T P.1203 standard model. The proposed model outperforms the P.1203 model in predicting the QoE rated on both Mobile and PC devices. Furthermore, our model proved to be robust even in the cross-device scenario, i.e., it can accurately estimate the QoE of a video watched on a device different from that used for training the model.

The paper is structured as follows. Section II discusses the related work in this area. Section III presents the motivations and the methodology of the proposed transformer-based QoE model. In Section IV, we describe the implementation details of the proposed model, while Section V discusses the achieved results. Finally, Section VI concludes the paper.

II. Related Work

Assessing the QoE for telecommunications services is of paramount importance [10]–[12]. Objective QoE models aim to predict the quality of multimedia content and applications by relying on mathematical models [13]. Several models have been developed for video streaming services in the last decade [14]. Among these models, the ITU-T Rec. P.1203 is the most recently approved recommendation for estimating the QoE of adaptive audiovisual streaming services [15]. The P.1203 model adopts a scalable approach using the same components for spatial and temporal scaling video degradations across 4 modes of operations, reflecting different types of media stream encryption. The 4 modes range from *Mode0*, corresponding to the highest considered level of encryption and lowest computational requirements (only access to codec, target bitrate, resolution, frame rate, and segment durations and sizes) to *Mode3*, having access to the full video bitstream. From the available video-related information, the P.1203 model estimates the final media session quality score in terms of the Mean Opinion Score (MOS) scale, ranging from 1 to 5.

In recent years, significant research studies have adopted ML-based methods to improve the accuracy of QoE models [16]. In [3], the nonlinear autoregressive with exogenous variables (NARX) model was proposed to process a nonlinear aggregate of subjective QoE measurements as inputs, including video quality during intervals of normal playback, rebuffering traces, and memory of prior events affecting QoE. Numerous ML models were trained in [17] for estimating QoE, startup delay, video resolution, video bitrate, and rebuffering occurrence. Moreover, the authors investigated the potential solutions for improving the cross-applicability of the ML models on different datasets. The DeepQoE framework in [18] combines different deep learning techniques, such as word embedding and 3D convolutional neural network (CNN), to extract generalized features. These features are then combined and fed into a neural network to create a learned representation that will serve as input for classification or regression tasks. The deep learning approach proposed in [4], called DeSVQ, combines multiple feature processing stages to capture the complex dependencies underlying the QoE prediction process. To this, DeSVQ utilizes a framework consisting of CNN and Long Short Term Memory (LSTM) networks. This integrated model improves the linear correlation coefficient performance compared to their counterparts.

To address the limitation of these deep learning models of adapting to diverse and dynamic contexts, transformer architectures have been considered. Transformers become a state-of-the-art solution for sequence modeling and transduction problems, such as language modelling, speech recognition, and machine translation, where they surpassed their counterparts based on a shallow stack of LSTM recurrent neural network layers [8]. However, transformers have also become widely utilized in the multimedia and network management fields. The FlowFormers algorithm, proposed in [19], is a transformer-based model for real-time network flow classification, which can be utilized for network security and traffic management applications. The use of transformer-encoders allowed this model to outperform existing deep learning approaches based on CNN and LSTM. A transformer-based model aimed at predicting user satisfaction for proactive interaction mechanisms in spoken dialogue systems is proposed in [20]. This transformer-based model extracts information from both the structured and text data and grasps the temporal dependency between the current turn and the previous turns for user satisfaction prediction. The results show that the transformer-based model allows the proactive interaction mechanism to achieve a 19% improvement in the accuracy of user satisfaction prediction and a 2.3% increase in user experience. A hierarchical combination of transformer models is proposed in [21] as a novel method for video quality assessment. Two types of transformers are used to extract clip-level quality embeddings (QE) of an input video and subsequent frame-level QE and video-level QE. Finally, a linear regressor predicts the quality value of the input video. The proposed hierarchical transformer model outperformed most other state-of-the-art methods. However, this model is only focused on video quality, and it does not consider streaming impairments, such as delay and stalling events.

These studies showed that the utilization of transformers provides enhanced performance in several contexts when analysis of sequential data is required. To the best of the authors' knowledge, this is the first study investigating the utilization of the transformer architecture to implement a model for estimating the QoE of video streaming services. The proposed model considers both video quality and session-related data and outputs the estimated QoE on the MOS scale.

III. Methodology

The variability and dynamic nature of QoE are inherent in human personal subjectivity and influenced by several aspects, which include system, human, and context-related factors [1]. For this reason, QoE objective models typically design the relationship between a few QoE influence factors and the QoE perceived for a specific multimedia service. As an example, QoE models for video streaming services observe session parameters (e.g., video resolution changes, occurrence and duration of stalling events) to predict the user's perceived QoE. However, the prediction accuracy of these models strongly relies on the specific session parameters (and related ranges of variation for these parameters) they are trained on. The P.1203 model overcame these issues because it is scalable to the type of input information available to it and can also be fed with per-second video and streaming session-related data [22].

Inspired by this modelling approach, we propose a transformer-based QoE estimation model that leverages sequential data encoding per-second video and session-related information. In the following, we refer to this information as the KPIs. The rationale behind encoding the observed KPIs into sequential data is the transformer's capability to process this kind of data and understand long-range dependencies between different data points by relying on the multi-head attention mechanism that allows it to focus on the most relevant parts of the input data. By capturing per-second level KPIs, we believe the transformer would be able to identify patterns and connections that might be missed with coarser data aggregation. Moreover, by feeding sequential data, we directly exploit the transformer architecture's strength in capturing complex relationships within the data sequence.

The following are the research questions we aim to address in this study:

- Q1:*"What performance can the transformer-based model achieve in predicting the QoE of videos with different lengths impaired by diverse KPIs?"*
- Q2:*"What performance can the transformer-based model achieve in predicting the QoE of videos watched on different devices?"*
- Q3:*"How does the transformer-based model perform compared to the ITU-T P.1203 model?"*

Fig. 1 illustrates the workflow process required to collect the KPIs and prepare the data for the proposed transformer-based QoE prediction model to capitalize on its deep learning architecture structure. The first step requires KPIs to be collected every second during the video streaming session. The collected KPIs are then encoded in a specific scheme that the transformer network can easily understand. Next, the encoded data is sequentalized to create a timeline of the observed KPIs. Data preprocessing steps may be needed to adapt the sequential data to a form that can be input for training the transformer-based model.

Simultaneously, the overall quality of the video is rated by a pool of subjects, whose scores are averaged in the mean opinion score (MOS). In the figure, we are considering the case in which the same video is watched and rated on different devices, which may influence the end user's QoE. The MOS score is the prediction variable involved in the training process of the transformer-based model.

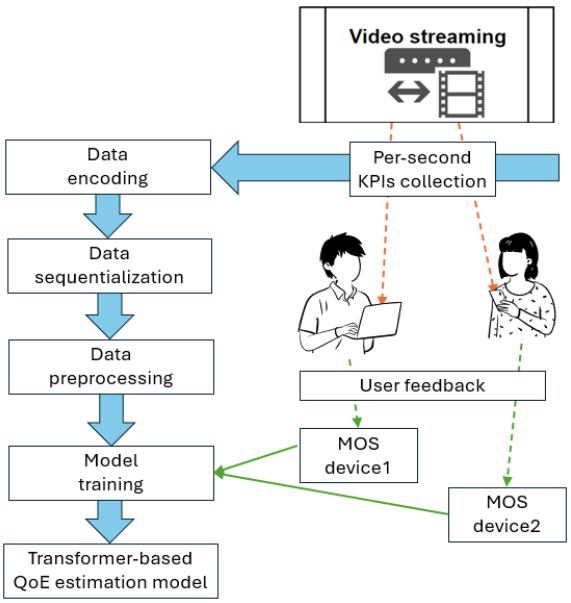

Fig. 1. The proposed workflow process.

The final output of our research study is a transformer-based model that estimates the QoE of video sequences streamed over the Internet in terms of MOS scores. In the next sections, we describe the implementation details of the model and we present the achieved QoE estimation results compared to the state-of-the-art ITU-T P.1203 model.

IV. Model Implementation

A. Datasets

We have considered two open datasets[1] from the ITU-T P.1203 standardization procedure (P.NATS) [9]. We refer to these datasets as in [9], i.e., TR04 and TR06. These datasets were created to explore the impact of different HAS-typical conditions (Hypothetical Reference Circuits, HRCs) on user QoE. Each HRC includes diverse combinations of quality switches between different resolutions and bitrates, initial loading delay, and stalling events. In particular, the TR04 dataset includes 20 HRCs, which were tested using 3 different 60-second-long video contents, for a total of 60 test video sequences. The TR06 dataset, instead, includes 11 HRCs, which were tested using 2 different 180-second-long video contents, for a total of 22 test video sequences. The overall 82 test video sequences included in these datasets were used in a subjective assessment [9], where people watched and rated the perceived video quality on PC and mobile devices. Thus, the datasets also include the MOS rated for each test video sequence for PC and mobile devices. The subjective scores

[1]https://github.com/itu-p1203/open-dataset

TABLE I
QUALITY LEVELS OF THE TEST VIDEO SEQUENCES.

Q	*Bitrate (kbps)*	*Resolution - height (px)*
7	10000	1080
6	2500	1080
4	500	480
2	150	240

were provided using the Absolute Category Rating (ACR) scale, which ranges from 1 (Bad) to 5 (Excellent). Then, the MOS assumes continuous values between 1 and 5.

B. Data encoding and sequentialization

The Transformers, as explained in section II, have been developed to perform well on long-time sequential data. Thus, the data included in the considered datasets was first encoded and then sequentalized.

Data encoding concerned the creation of a tuple for each second of a test video sequence. We define the tuple as

$$t^d_{v,n} = (Q, D, S), \tag{1}$$

where d identifies the dataset (TR04 or TR06), v identifies the number of the test video sequence, and n identifies the second of the video ranging from 1 to N^d_v, i.e., the length in second of video v from dataset d. The value of the tuple identifies the KPI values collected for that second of the video, which include the quality level of the video frame Q in terms of resolution and bitrate as summarized in Table I, the occurrence of initial delay D (1 in case of delayed frame, 0 otherwise), and the occurrence of stalling S (1 in case of stalling event, 0 otherwise).

Data sequentialization concerns the concatenation of the tuples created for each test video sequence into a sequential data type, as follows

$$SQ^d_v = [t^d_{v,1} \oplus t^d_{v,2} \oplus \ldots \oplus t^d_{v,N^d_v}]. \tag{2}$$

Each of the created SQ^d_v is accompanied by two MOS values concerning the QoE rated on PC devices (MOS_{PC}) and mobile devices (MOS_M). These MOSes are the QoE values that must be predicted by the proposed Transformer-based model trained on the input sequential data.

C. Data preprocessing

Some preprocessing steps were required to make the input sequential data appropriate for the Transformer. Transformers require input data sequences to be of the same length. Thus, data padding was performed to standardise the length of each sequential data. Indeed, the duration of test video sequences could overcome the duration of the considered video contents (60 s for TR04 and 180 s for TR06) because they also include the occurrence of initial delay and buffering events. The zero-pad approach was used, i.e., each SQ^d_v shorter than the longest SQ^d_v was extended by adding the necessary number of tuples with values of Q, D, and S set to 0.

Transformers, likewise, deep learning neural networks, need to be trained with a large amount of data. Moreover, to avoid biased results, the input labelled data should be uniformly distributed between the complete MOS range from 1 to 5. Therefore, we adopted the SMOGN data augmentation technique for unbalanced regression problems [23]. SMOGN employs the SmoteR interpolation method [24] to interpolate examples that are close, thereby reducing the risk of interpolating examples that, despite being among the seed example's nearest neighbours, are excessively distant. Using a Random Search approach, we have found that triple the size of the original dataset (we tested increasing the size of the original dataset from 2 to 10 times) allowed the proposed model to achieve the greatest QoE estimation performance.

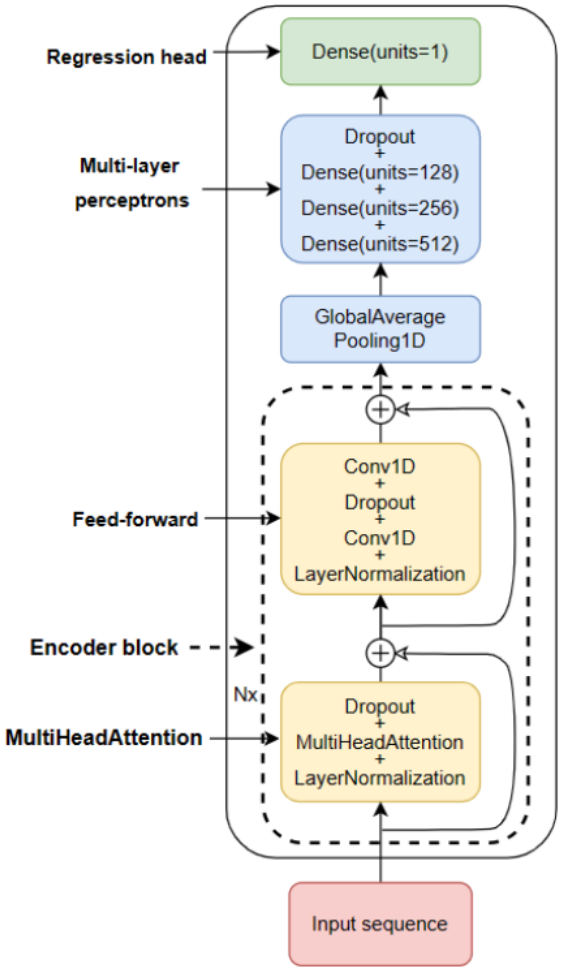

Fig. 2. The architecture of the proposed Transformer-based model.

D. Proposed transformer-based model

As anticipated in Section III, the main novelty of transformer deep learning architectures concerns the multi-head attention layer within the encoder block, which focuses on the different parts of the input data sequence to identify correlations between data variations. Moreover, each encoder block repeats its computation N times to draw connections between any part of the input sequence. The redundancy generated by this approach generates similar output sequences,

helping the regression task to predict the possible output. Thus, we rely on these characteristics of the transformer architecture to achieve accurate and robust MOS regression predictions.

The proposed transformer-based network, illustrated in Fig. 2, is composed of 4 main blocks: the Encoder block, the GlobalAveragePooling1D, the Multi-layer perceptions (MLP), and the Regression head. The encoder block, identified with a dashed line, includes a Multi-Head Attention (MHA) module and a Feed-forward module. The MHA module comprises a normalization layer, an MHA layer (composed of 4 attention heads, each of size 256), and a dropout layer with a dropping rate of 0.15. The MHA module is followed by a skip connection, which adds the MHA module's output to its input. This allows the gradient to propagate directly through the network, bypassing the intervening layers.

The following Feed-forward module applies non-linear transformations to the token representations received by the MHA module, enabling the model to capture complex patterns and relationships within the data. It contains a normalization layer, two 1-dimensional convolution layers (Conv1D), each containing 8 filters with a kernel size equal to 1, and a dropout layer with a dropping rate of 0.15. The optimal number of stacked encoder blocks N was found to be 4. Similarly to the MHA module, the Feed-forward module is also followed by a skip connection. These skip connections allow the preservation of important information across layers of the transformer encoder block [25]. The encoder block is followed by the GlobalAveragePooling1D layer, which reduces the encoder block's output to an array of features for each data point. The MLP module consists of 3 fully connected layers, including 512, 256, and 128 neuron units, respectively, and a final dropout layer with a dropping rate of 0.25. Each dense layer of the MLP is activated by the Rectified Linear Unit (ReLU) activation function with an L2 kernel regularization function to prevent over-fitting. Finally, the Regression head module outputs the QoE prediction in terms of the MOS.

The values of all the mentioned hyper-parameters, number of layers, and number of stacked encoder blocks were defined using a Random Search approach to achieve the best QoE estimation performance [26]. The proposed neural network was trained using the Mean Square Error (MSE) loss function and the Adam optimization method [27]. The neural network training followed a 5-fold cross-validation process, and the datasets were divided with a $70\%/30\%$ splitting rate for training and validation. The early-stop function was used for training, which allowed the network to converge in 40 epochs.

V. Results

Table II summarizes the QoE estimation performance of the proposed transformer-based model in terms of root mean square error (RMSE), Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC). Moreover, for comparison, we have reported the performance achieved by the official ITU-T P.1203 standard model (mode 0), which is based on the entire set of 30 datasets created by the P.NATS group rather than just the two open datasets considered in this study. Unfortunately, the rest of the datasets are not open. The Device column identifies the type of device on which the videos were watched before being rated by the users during the subjective test. Both TR04 and TR06 datasets include test video sequences that were watched and rated on both devices.

TABLE II
QoE estimation performance of the proposed transformer-based model compared to the state-of-the-art ITU-T P.1203 model in terms of RMSE, PCC, and SCC.

Device	Model	Dataset	RMSE	PCC	SCC
Mobile	P.1203	TR04	0.3850	0.9118	0.8858
		TR06	0.3964	0.9195	0.8994
		TR04+TR06	0.3881	0.9092	0.8869
	Prop.	TR04	0.3552	0.9230	0.8920
		TR06	0.2855	0.9540	0.9525
		TR04+TR06	0.3549	0.9267	0.9152
PC	P.1203	TR04	0.5257	0.8783	0.8235
		TR06	0.3595	0.9548	0.9206
		TR04+TR06	0.4867	0.9014	0.8654
	Prop.	TR04	0.3934	0.9148	0.8714
		TR06	0.3031	0.9589	0.9377
		TR04+TR06	0.3698	0.9289	0.9043
PC + Mobile	P.1203	TR04	0.4608	0.8854	0.8472
		TR06	0.3784	0.9286	0.8998
		TR04+TR06	0.4402	0.8959	0.8692
	Prop.	TR04	0.3987	0.9093	0.8733
		TR06	0.3344	0.9456	0.9442
		TR04+TR06	0.3873	0.9192	0.8972
Cross-Device	Prop.	Trn: Mob. TR06 Val: PC TR06	0.3426	0.9536	0.9541
		Trn: Mob. TR04 Val: PC TR04	0.4006	0.9151	0.8786
		Trn: PC TR06 Val: Mob. TR06	0.3608	0.9566	0.9634
		Trn: PC TR04 Val: Mob. TR04	0.4055	0.9051	0.8538

Firstly, we answer research questions Q1 and Q3 from Section III. It can be seen that the proposed model outperforms the P.1203 model in predicting the MOS rated on both Mobile and PC devices. This is true when considering the single datasets (TR04 and TR06) as well as the union of the 2 datasets (TR04+TR06). In particular, a relevant performance increase is observed for the MOS prediction of Mobile devices on the TR06 dataset (0.2855 vs. 0.3964 of RMSE, 0.954 vs. 0.9195 of PCC) and PC devices on the TR04 (0.3934 vs. 0.5257 of RMSE, 0.9148 vs. 0.8783 of PCC) and TR04+TR06 (0.3698 vs. 0.4867 of RMSE, 0.9289 vs. 0.9014 of PCC) datasets. When considering the union of Mobile and PC datasets, our model still achieves better performance than P.1203 for all datasets in terms of the 3 considered metrics.

Secondly, we answer research question Q2. To test the robustness of the proposed model to device and dataset changes, we have computed cross-device performance, i.e., the capability of the model to estimate the QoE of a video watched on a device different from that used for training the model. Thus, in this case, we have trained on the Mobile dataset and validated on the PC dataset, and vice versa. Cross-device performance is higher on the TR06 dataset than on TR04. In particular, for TR06, when training on Mobile and validating on PC, an RMSE of 0.3426 and a PCC of 0.9536 is achieved.

When training on PC and validating on Mobile, slightly lower performance results are achieved (RMSE of 0.3608 and PCC of 0.9566). For TR04, when training on Mobile and validating on PC (and vice versa), an RMSE and PCC of about 0.40 and 0.91 are achieved, respectively. These results are comparable to those achieved by the proposed model and the P.1203 model on the complete datasets (PC+Mobile) that were previously discussed. This proves the robustness of the proposed model in predicting the QoE for different devices than those used for training the model.

VI. Conclusion

This paper has explored the potential of transformer learning architectures to implement a QoE estimation model for video streaming services. A workflow process including data collection, data encoding, and data sequentalization steps has been proposed to prepare the data in the appropriate form required by the transformer's learning processes.

The QoE estimation performance of the proposed model has been measured and compared with that of the standard ITU-T P.1203 by considering two open datasets, including a total of 82 videos impaired by video quality switching, delay, and stalling events, which were watched and rated on mobile and PC devices. The proposed model has shown superior performance in the QoE estimation than P.1203 in terms of RMSE, PCC, and SRCC. In particular, relevant performance enhancements were observed for the MOS prediction of Mobile devices on the TR06 dataset (RMSE lowered by 0.11) and PC devices on the TR04 (RMSE lowered by 0.13) and TR04+TR06 (RMSE lowered by 0.117) datasets. Better performance was also achieved when considering the union of Mobile and PC datasets (RMSE lowered by 0.08 on average).

Moreover, we have proved cross-device QoE prediction performance, i.e., robustness to device and dataset changes. When training on Mobile and validating on PC (and vice versa) for dataset TR06, an RMSE of 0.3426 (0.3608) was achieved. When training on Mobile and validating on PC, and vice versa, for dataset TR04, an RMSE of 0.40 was achieved. These RMSE results are comparable and even lower (for dataset TR06) than those achieved by the P.1203 model on the complete datasets (PC+Mobile).

References

[1] P. Le Callet, S. Möller, and A. Perkis. (2012) Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013.

[2] S. Vlahovic, M. Suznjevic, and L. Skorin-Kapov, "A survey of challenges and methods for Quality of Experience assessment of interactive VR applications," *Journal on Multimodal User Interfaces*, pp. 1–35, 2022.

[3] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous prediction of streaming video QoE using dynamic networks," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 1083–1087, 2017.

[4] M. Ghosh, D. C. Singhal, and R. Wayal, "DeSVQ: Deep learning based streaming video QoE estimation," in *Proc. of the 23rd Int. Conf. on Distributed Computing and Networking*, 2022, pp. 19–25.

[5] T. Abar, A. Ben Letaifa, and S. El Asmi, "Chapter Five - User behavior-ensemble learning based improving QoE fairness in HTTP adaptive streaming over SDN approach," ser. Advances in Computers, A. R. Hurson, Ed. Elsevier, 2021, vol. 123, pp. 245–269.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] M. D. Deepa *et al.*, "Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 7, pp. 1708–1721, 2021.

[8] K. Irie, A. Zeyer, R. Schlüter, and H. Ney, "Language Modeling with Deep Transformers," in *Proc. Interspeech 2019*, 2019, pp. 3905–3909.

[9] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software," in *9th ACM Multimedia Systems Conference*, 2018.

[10] G. Bingöl, S. Porcu, A. Floris, and L. Atzori, "QoE Estimation of WebRTC-based Audiovisual Conversations from Facial Expressions," in *2022 16th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, 2022, pp. 577–584.

[11] C. Marche, L. Serreli, and M. Nitti, "Analysis of feedback evaluation for trust management models in the Internet of Things," *IoT*, vol. 2, no. 3, pp. 498–509, 2021.

[12] C. Marche and M. Nitti, "A binary trust game for the internet of things," *IoT*, vol. 2, no. 1, pp. 50–70, 2021.

[13] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "A survey on parametric QoE estimation for popular services," *Journal of Network and Computer Applications*, vol. 77, pp. 1–17, 2017.

[14] N. Barman and M. G. Martini, "QoE Modeling for HTTP Adaptive Video Streaming–A Survey and Open Challenges," *IEEE Access*, vol. 7, pp. 30 831–30 859, 2019.

[15] ITU, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport." Recommendation ITU-T P.1203, 2017.

[16] G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis, and P. I. Lazaridis, "A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction," *IEEE Access*, vol. 10, pp. 19 507–19 538, 2022.

[17] M. Seufert and I. Orsolic, "Improving the Transfer of Machine Learning-Based Video QoE Estimation Across Diverse Networks," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2023.

[18] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan, "DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3210–3223, 2020.

[19] R. Babaria, S. C. Madanapalli, H. Kumar, and V. Sivaraman, "FlowFormers: Transformer-based Models for Real-time Network Flow Classification," in *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, 2021, pp. 231–238.

[20] W. Shen, X. He, C. Zhang, X. Zhang, and J. Xie, "A transformer-based user satisfaction prediction for proactive interaction mechanism in DuerOS," in *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 2022, pp. 1777–1786.

[21] Z. Li and L. Yang, "DCVQE: A Hierarchical Transformer for Video Quality Assessment," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 2562–2579.

[22] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[23] P. Branco, L. Torgo, and R. P. Ribeiro, "SMOGN: a pre-processing approach for imbalanced regression," in *Proc. of the First Int. Workshop on Learning with Imbalanced Domains: Theory and Applications*, ser. Proc.s of Machine Learning Research, P. B. Luís Torgo and N. Moniz, Eds., vol. 74. PMLR, 22 Sep 2017, pp. 36–50.

[24] L. Torgo, P. Branco, R. P. Ribeiro, and B. Pfahringer, "Resampling strategies for regression," *Expert Systems*, vol. 32, pp. 465 – 476, 2015.

[25] F. Liu, X. Ren, Z. Zhang, X. Sun, and Y. Zou, "Rethinking Skip Connection with Layer Normalization," in *International Conference on Computational Linguistics*, 2020.

[26] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations*, 2015.