This report has been submitted for ITU AI challenge: Multi-environment automotive QoS prediction

Predicting Cellular QoS in V2X via Domain Adaptation

Qixin Zhang, Wei Ye, Steven Sleder
University of Minnesota, Twin Cities, USA
{zhan8548, ye000094, slede001}@umn.edu

***Abstract*—In this study, we leverage the Berlin V2X dataset [1], a comprehensive collection of high-resolution wireless measurements from multiple vehicles and radio access technologies, to tackle the challenge [2] of predicting Quality of Service (QoS) in cellular communication for vehicle-to-anything (V2X) applications. The dataset spans diverse urban environments in Berlin, enabling a broad range of machine learning investigations in the domain of V2X communication. Our primary objective is to predict QoS parameters, with a focus on 'datarate,' in the cellular domain. To enhance model generalization across different operators, we employ domain adaptation techniques, treating one operator as the source domain and another as the target domain. Our approach involves feature selection using domain adaptation and the training of predictive models with adjusted parameters. The results demonstrate the effectiveness of our methodology, achieving an impressive R-squared (R2) score of 0.835 on the target domain.**

***Index Terms*—LTE, V2X, Domain Adaptation**

I. Introduction

Driven by the convergence of cutting-edge technologies and communication networks, modern multi-environments are witnessing transformations in transportation systems. At the heart of this revolution is vehicle-to-everything (V2X) communication, a paradigm that enables vehicles to interact not only with each other, but with infrastructure, pedestrians and other road users. V2X communication holds promise for safer roads, more efficient traffic management, and enabling autonomous driving systems.

In this context, the dataset in paper written by R. Hernangómez et al. [3] constitutes an important milestone, offering researchers a unique opportunity to explore the complexities of V2X communication. This dataset contains high-resolution wireless measurements collected from various environments in Berlin, covering both cellular and sidelink radio access technologies. It provides a rich and comprehensive resource for studying the complexity of vehicle communication systems with data collected from multiple vehicles over several days.

Amidst the broad promise of V2X communications, a central and critical challenge emerges: Quality of Service (QoS) prediction for cellular communications. In this challenge, we focus our attention on predicting the "data rate" of downlink cellular communications. "Data rate" is a key metric that directly impacts the efficiency and reliability of various V2X applications, including real-time traffic management, autonomous vehicle navigation, and multimedia streaming.

The importance of accurately predicting "data rates" cannot be overemphasized. It forms the basis for QoS configuration, ensuring that data transfer between vehicles and infrastructure meets the needs of safety-critical and time-sensitive applications. Accurate forecasts enable traffic controllers to make informed decisions, self-driving cars to navigate seamlessly, and passengers to enjoy uninterrupted in-vehicle service.

This study strives to address this fundamental challenge by employing machine learning techniques on the "Berlin V2X" dataset. We delve into the mathematics of predicting "data rates" by creating predictive models that capture the complex relationship between different sets of input features and desired outputs. Our approach involves rigorous feature selection, model building using random forest regressors, and exploration of domain adaptation to account for variation between different operators.

The importance of this research goes beyond its immediate application. It lays the foundation for the development of robust adaptive V2X communication systems, facilitating safer and more efficient mobility. Furthermore, it demonstrates the potential of machine learning to solve complex real-world challenges, emphasizing the importance of data-driven approaches in shaping the future of vehicular communications.

In the following sections, we provide a detailed overview of the problem, describe our approach, present experimental results, and discuss our findings and their wider implications.

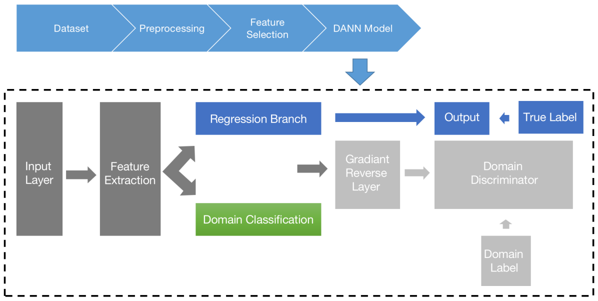

Fig. 1. An image of a galaxy

Advised by Zhi-Li Zhang

II. PROBLEM STATEMENT

In the era of smart transportation systems, the seamless exchange of data between vehicles and their surrounding environment has become paramount. V2X communication represents a cornerstone of this transformation, promising to revolutionize road safety, traffic efficiency, and the development of autonomous vehicles. Within this context, the "Berlin V2X" dataset offers a unique glimpse into the intricacies of vehicular communication across diverse urban landscapes.

At the heart of V2X communication is the QoS, a critical factor that directly influences the reliability and efficiency of data transmission. Specifically, in this study, we focus on the prediction of QoS for cellular communication, with a particular emphasis on predicting the 'datarate' for downlink cellular data transmission.

The 'datarate' metric is of utmost importance in V2X communication. It quantifies the rate at which data can be transmitted from the cellular network infrastructure to the vehicle, impacting the ability to support applications such as real-time navigation, traffic management, and infotainment services. Accurate prediction of 'datarate' is essential for ensuring a seamless and responsive V2X ecosystem.

Mathematically, our goal can be formalized as follows:

Given a domain D, where $D \subset \mathbb{R}^n$, we seek to establish a mathematical function $f(X; D) : \mathbb{R}^n \to \mathbb{R}$ that can approximate the relationship between the input features X and the target variable "datarate," denoted as Y:

$$f(X; D) \approx Y$$

In this formulation, $f(X; D)$ is a function that maps the input features X, which belong to $\mathbb{R}^n$, to a real number in $\mathbb{R}$, and the function's behavior depends on the specific domain D over which it operates.

We are interested in finding the optimal parameters θ that minimize the following loss function:

$$\theta^* = \arg\min_{\theta} L(f(X; D), Y)$$

where θ^* represents the optimal parameters, and L is the loss function that quantifies the dissimilarity between the predictions $f(X; D)$ and the actual target values Y. The argmin function returns the set of parameters θ that minimizes this loss.

Essentially, our goal is to develop a predictive model to estimate "data rate" from diverse input characteristics, spanning wireless communication parameters, geographic data, and environmental factors. Addressing the complexity of QoS prediction in cellular-based V2X networks involves several key challenges:

1) **Complex V2X Communication**: V2X networks are intricate and dynamic, influenced by vehicle mobility, changing environments, and varying radio access technologies. Accurately modeling the relationship between inputs and "data rate" is challenging.
2) **High-dimensional Feature Space**: The "Berlin V2X" dataset offers a wealth of features, from communication metrics to geographic and environmental data. Navigating this high-dimensional space to identify crucial predictors of "data rate" is daunting.
3) **Interaction of Functions**: Understanding how different functions interact and impact QoS is critical. Features may exhibit dependencies and nonlinear relationships, necessitating advanced modeling techniques.
4) **Dynamic Environments**: V2X communications occur in dynamic environment settings with rapid changes. Accurate predictions must adapt to these fluctuations, making modeling challenging.
5) **Generality across Operators**: The dataset spans multiple operators, each with unique network characteristics. Achieving a model that generalizes across operators is crucial, emphasizing the need for domain adaptation techniques.
6) **Real-world Implications**: Beyond theoretical models, accurate QoS predictions have practical implications, including improved traffic management, safer autonomous driving, and enhanced mobility. The translation of research findings into real-world solutions is of paramount importance.

Addressing these challenges requires a robust approach that not only leverages advanced machine learning techniques but also considers the complexity and diversity of V2X communication environments. In the following sections, we detail our approach, present experimental results, and discuss our findings, all with the aim of improving the understanding and applicability of QoS prediction in V2X communication networks.

In addition to the standard regression task, we employ a Domain Adversarial Neural Network (DANN) model [4], as shown in Fig. 1 to address domain adaptation challenges. The DANN model can be defined as follows:

$$\mathcal{L}_{\text{DANN}} = \lambda\mathcal{L}_{\text{task}} - \mathcal{L}_{\text{domain}}$$

where $\mathcal{L}_{\text{DANN}}$ is the loss for the DANN model, λ is a trade-off parameter, $\mathcal{L}_{\text{task}}$ is the loss for the regression task, and $\mathcal{L}_{\text{domain}}$ is the domain adversarial loss.

We will explain the detailed usage and architecture of the DANN model in the methodology section, and use R^2 to evaluate our result. The mathematical formula for calculating R^2 score is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where n represents the total number of samples, Y_i represents the true 'datarate' value for sample i, $\hat{Y}_i$ is the predicted 'datarate' value for sample i, and $\bar{Y}$ denotes the mean of the true 'datarate' values.

III. METHODOLOGY

In this section, we provide a comprehensive overview of the methodology employed to address the challenges of QoS prediction in the context of cellular communication for V2X networks. Our approach leverages advanced machine learning techniques, including the use of a Domain-Adversarial Neural Network (DANN) to adapt the model to operator-specific variations, as shown in Fig. 1.

A. Data Preprocessing

We commence the methodology with data preprocessing, a crucial step in ensuring data quality and compatibility with our predictive model. The following steps summarize our data preprocessing procedure:

Filtering and Subset Selection: We narrow down the data to include only 'downlink' measurements with 'datarate' as the measured QoS parameter. This selection focuses our analysis on the specific aspect of cellular communication that is most relevant to V2X applications.

Handling Missing Data: Missing data is a common challenge in real-world datasets. To address this, we employ mean imputation, filling in missing values with the mean of their respective columns. This approach ensures that we maintain data completeness while minimizing the introduction of bias.

Domain Splitting: Since the dataset contains data from multiple operators, we divide it into source and target domains based on the 'operator' column. One operator serves as the source domain, and the other as the target domain. This separation forms the basis for domain adaptation, a key aspect of our methodology.

B. Feature Selection

The next phase of our methodology involves feature selection, aimed at identifying the most influential predictors for 'datarate' and improving model interpretability. We employ a random forest classifier to rank the importance of features, allowing us to select a subset of high-impact features for modeling.

The mathematical representation of feature selection involves calculating feature importances (FI) using the random forest classifier. These importances are computed based on the Gini impurity:

$$FI = \sum_{i}^{n} \frac{FV_i}{\sum_{j}^{n} FV_j} \times G_i$$

where FI represents the feature importance, FV_i denotes the feature value of feature i, G_i is the Gini impurity for feature i, and n is the total number of features.

The subset of selected features is chosen based on their importance scores. After screening, we found that resource block number, Modulation Coding Scheme, signal to noise ratio, and especially transporation block size have a very high impact on datarateI. Selected features include a combination of physical layer parameters, radio resource management metrics, and context information.

C. Model Architecture

Our choice of model architecture is tailored to the Berlin V2X dataset and the challenges it presents for predicting QoS in cellular communication within the context of V2X applications. The architecture consists of the following components:

TABLE I
FEATURE IMPORTANCE RESULTS

Feature	Importance
SCell_Downlink_TB_Size	0.5521
ping_ms	0.1711
PCell_Downlink_TB_Size	0.1473
jitter	0.0968
SCell_SNR_1	0.0017
PCell_RSRP_2	0.0015
PCell_RSSI_max	0.0013
PCell_Downlink_Num_RBs	0.0013
windSpeed	0.0011
...	...
visibility	0.0000
PCell_MCC	0.0000
PCell_MNC_Digit	0.0000
SCell_MCC	0.0000
SCell_MNC_Digit	0.0000

Shared Feature Extractor for Context The shared feature extractor, represented mathematically as f_{shared}, is designed to capture essential context-specific features from the data. It comprises a dense layer with 100 neurons and a Rectified Linear Unit (ReLU) activation function:

$$f_{\text{shared}}(X) = \text{ReLU}(W_{\text{shared}} \cdot X + b_{\text{shared}})$$

where X is the input feature vector, W_{shared} and b_{shared} are the weight matrix and bias vector for the shared feature extractor, respectively.

This layer focuses on extracting critical context information from the input data.

Regression Head for QoS Prediction The regression head, denoted as $f_{\text{regression}}$, immediately follows the shared feature extractor and is responsible for predicting the 'datarate' QoS parameter. It comprises another dense layer with 100 neurons and a ReLU activation function:

$$f_{\text{regression}}(X) = \text{ReLU}(W_{\text{regression}} \cdot X + b_{\text{regression}})$$

where X is the input feature vector, $W_{\text{regression}}$ and $b_{\text{regression}}$ are the weight matrix and bias vector for the regression head, respectively.

This layer deciphers the complex relationships between various physical layer parameters, radio resource management metrics, and QoS, crucial for accurate 'datarate' predictions.

Domain Classification Head for Operator-Specific Adaptation To adapt the model to the characteristics specific to each operator, we introduce a domain classification head, represented as f_{domain}. This component comprises a dense layer with 100 neurons and a ReLU activation function, followed by a binary classification output layer with a sigmoid activation function:

$$f_{\text{domain}}(X) = \text{Sigmoid}(W_{\text{domain}} \cdot X + b_{\text{domain}})$$

where X is the input feature vector, and W_{domain} and b_{domain} are the weight matrix and bias vector for the domain classification head, respectively.

The domain classification head distinguishes between the source and target domains, enabling domain adaptation.

D. Model Training and Domain Adaptation

Our model training and domain adaptation strategy is designed to address the challenges posed by the Berlin V2X dataset and operator-specific variations:

Data Normalization for Environment Heterogeneity Given the significant heterogeneity in multi-environments, we begin training by normalizing both input features and target values. This normalization step ensures that the model can effectively learn patterns from diverse contexts, maintaining data consistency and scale invariance.

Domain Adaptation for Operator Variations To address operator-specific variations, we employ Domain-Adversarial Neural Network (DANN) techniques. DANN introduces a gradient reversal layer, mathematically defined as $GRL(\alpha)$, that enables the model to learn domain-invariant features while minimizing operator-specific variations:

$$GRL(\alpha)(X) = -\alpha \cdot X$$

where X represents the input, and α is a scaling factor that modulates the gradient flow.

During training, the gradient reversal layer helps the model generalize effectively across operators, reducing the impact of operator-specific variations.

Loss Functions and Optimization for Robust Learning To optimize our model, we utilize the Adam optimizer with a learning rate of 0.05. Our model incorporates multiple loss functions (See Table.II) to achieve robust learning: - Mean Squared Error (MSE) for the regression task (QoS prediction):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

- Binary Cross-Entropy (BCE) for domain classification:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\Big(D_i \cdot \log(\hat{D}_i) + (1 - D_i) \cdot \log(1 - \hat{D}_i)\Big)$$

where N is the total number of samples, Y_i represents the true 'datarate' value, $\hat{Y}_i$ is the predicted 'datarate' value, D_i is the true domain label (0 for source, 1 for target), and $\hat{D}_i$ is the predicted domain label.

IV. EVALUATION

A. Experiment Setup

Hardware The experiments in this study were conducted on a MacBook Pro (2021) laptop. This MacBook Pro is equipped with a powerful hardware configuration that includes an Intel Core i7 processor, ample RAM, and a dedicated GPU. This hardware configuration ensures efficient execution of deep learning tasks and model training.

Implementation The code for this experiment was implemented using TensorFlow, a popular deep learning framework.

B. Result

After training the model, we evaluate its performance in predicting the 'datarate' QoS parameter for cellular communication in the Berlin V2X dataset. To quantify the model's effectiveness, we employ the coefficient of determination, commonly known as R^2 score(See Table.II). The R^2 score measures the proportion of the variance in the target variable that is predictable from the input features. A higher R^2 score indicates a better fit of the model to the data with 1.0 representing a perfect fit. Conversely, an R^2 score of 0.0 suggests that the model's predictions are equivalent to simply using the mean of the target values, indicating poor predictive performance.

TABLE II
LOSS VALUES AND R2 SCORE

Epoch	Loss	DANN Loss	Elapsed Time (s)
1	1.8318	1.8262	1.02
2	0.1370	0.1370	0.69
3	0.1205	0.1205	0.69
4	0.1338	0.1338	0.70
5	0.1134	0.1134	0.69
6	0.1288	0.1288	0.69
7	0.1226	0.1226	0.69
8	0.1262	0.1262	0.69
9	0.1104	0.1104	0.69
10	0.1064	0.1064	0.71
...	...	...	...
61	0.1042	0.1042	0.69
62	0.0953	0.0953	0.69
63	0.0978	0.0978	0.69
64	0.1056	0.1056	0.69
R2 Score	0.83526049422022		
Total Time	45.70733690261841 seconds		

In our experiments, the model achieved an R^2 score of 0.835 on the target domain, highlighting its ability to accurately predict 'datarate' QoS in the context of V2X communication.

The R^2 score provides valuable insights into how well our model generalizes and captures the underlying relationships between input features and QoS, thereby demonstrating its efficacy in addressing the challenges posed by the Berlin V2X dataset.

Comparison After training the model, we evaluated its performance in predicting the 'datarate' QoS parameter for cellular communication in the Berlin V2X dataset. The model achieved a remarkable R^2 score of 0.938 on the target domain, indicating strong predictive capabilities. However, a closer look at the MSE loss value revealed a extremely large value of 30,413,578,057,921. This apparent contradiction arises because the MSE loss is sensitive to outliers and may be influenced by extreme predictions. The high R^2 score suggests that the model without DANN may be overfitting to the training data, meaning it is too closely tailored to the idiosyncrasies of the training set. While it performs well on the training data, it may struggle to generalize to new, unseen data points. In summary, the model without DANN achieved a high R^2 score but exhibited signs of overfitting, which can limit its utility when faced with diverse and complex real-world data.

V. CONCLUSION

In this study, we have addressed the critical task of QoS prediction for cellular communication in V2X environments. Our approach involved feature selection, model creation, and domain adaptation, showcasing the adaptability of the model across different operators. The achieved R2 score of 0.835 reflects the model's strong predictive performance. This research not only provides valuable insights into V2X communication but also highlights the importance of feature selection and domain adaptation in real-world applications.

In conclusion, our work contributes to the foundation of future vehicular communication systems, paving the way for more efficient and reliable transportation networks. Further research in this domain can build upon these findings to enhance the performance and adaptability of V2X communication systems in multi-environments.

REFERENCES

[1] "Berlinv2x github repository," https://github.com/fraunhoferhhi/BerlinV2X, accessed: September 4, 2023.

[2] "Itu ai challenge," https://challenge.aiforgood.itu.int/match/matchitem/80, accessed: September 4, 2023.

[3] R. Hernangómez, P. Geuer, A. Palaios, D. Schäufele, C. Watermann, K. Taleb-Bouhemadi, M. Parvini, A. Krause, S. Partani, C. Vielhaus, M. Kasparick, D. F. Külzer, F. Burmeister, F. H. P. Fitzek, H. D. Schotten, G. Fettweis, and S. Stańczak, "Berlin v2x: A machine learning dataset from multiple vehicles and radio access technologies," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–5.

[4] Y. Ganin and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2017.