Addressing Traffic Prediction Uncertainty in Multi-Period Planning Optical Networks

Tania Panayiotou and Georgios Ellinas

Department of Electrical and Computer Engineering and the KIOS Research and Innovation Center of Excellence (KIOS CoE), University of Cyprus, Nicosia, 1678, Cyprus

{*panayiotou.tania, gellinas*}*@ucy.ac.cy*

Abstract: Deep-quantile regression is leveraged to capture traffic prediction uncertainty over future network planning intervals. We show that quantile predictions, acting as discriminative margins, result to significant spectrum savings compared to empirically estimated myopic margins considered.

1. Introduction

Predictive multi-period network planning, leveraging the capabilities of software defined networking (SDN), has recently emerged as a promising approach towards effectively dealing with network overprovisioning that is present in static optical networks [1, 2]. In the heart of multi-period planning is traffic analysis and specifically machine learning (ML), with proven capabilities on sufficiently modeling the highly non-linear nature of network traffic [1, 3, 4]. Network traffic modeling is in essence a time-series forecasting problem in which past traffic observations are exploited to predict future traffic demand for proactive decision-making (e.g., proactive network optimization). Hence, deep neural networks (DNNs) with recurrent units, specifically designed to model sequential data of high dimensionality, are today amongst the most promising and widely adopted traffic prediction models, exhibiting competitive performance accuracies on real-world data [1, 3].

Commonly, regression is applied in traffic prediction analysis, with the model of choice trained to minimize a least squares loss function. The resulting model is, in essence, an approximation of the relationship between the input (i.e., past traffic observations) and output (i.e., future traffic demand) variables. The output of the model (i.e., the prediction) is an estimate or an approximation, containing some uncertainty that is represented by the variability around the mean response value. Evidently, model accuracy, even though important, is not a sufficient measure of how much we can trust such point-based predictions for decision-making. In fact, prediction uncertainty, resulting from the errors in the model itself and the uncertainty (e.g., noise) over the input variables, may result in erroneous prediction-driven resource allocation decisions, especially when the predictions highly underestimate the true traffic demand. While prediction uncertainty is considered in a few works in the literature, this is done through the estimation of myopic empirical margins [1], largely ignoring the fact that diverse input patterns are subject to different levels of uncertainty. Another approach, indirectly dealing with prediction uncertainty, suggests on-line provisioning the unpredictable traffic in a best-effort approach to mitigate quality-of-service (QoS) violations [5]. Given the existing gap in the literature on fine dealing with traffic prediction uncertainty, the main contribution of this work is to primarily provide a mathematical framework capable of measuring the certainty level of model predictions towards appropriate resource allocation decisions.

In ML, several approaches exist capable of achieving this objective, such as conducting Bayesian or Monte Carlo (MC) dropout inference and through the estimation of conditional quantile functions (e.g., deep quantile regression) [6, 7]. In this work, as a first step towards capturing traffic prediction uncertainty, a deep quantile regression framework is adopted in which margins are learned and inferred in a discriminative fashion over the traffic inputs. The proposed framework is evaluated and compared with a number of baseline approaches on a real-word dataset, demonstrating significant network savings (up to 82%), with, however, a small penalty on the unpredictable traffic. Unpredictable traffic can be, however, successfully handled on-line with a reduced operational overhead (i.e., up to 72%) compared with the case where least squares model predictions are directly driving network planning, without any margin consideration.

2. Deep Quantile Regression for Network Traffic Prediction

Deep quantile regression is applied to approximate the conditional quantile function $Q_Y(q|\mathbf{x})$, also known as the q-quantile, where $0 < q < 1$, $\mathbf{x} \in X$ is a r.v. representing the past traffic observations, and $y \in Y$ is a r.v. representing the true value of the next (future) traffic observation. By definition, a q-quantile is learned to return a value $\hat{y}$, such that the probability that the true value y of the next traffic observation, given $\mathbf{x}$, will be less than or equal to $\hat{y}$, is equal to q [8]. In essence, q-quantile determines the percentage of a population that is above or below a certain threshold. Hence, for our use case it is important that q is high enough (e.g., above 0.9) to approximate an upper prediction bound $\hat{y}$ for each input $\mathbf{x}$ (i.e., to increase the probability that sufficient spectrum is proactively

allocated to serve the future traffic). In a deep learning framework, a q-quantile is estimated by minimizing the asymmetrically weighted sum of absolute errors [6, 8]:

$$L_q = \frac{1}{n}\sum_{\tau=1}^{T} \rho_q(y_\tau - \hat{Q}_Y(q|\mathbf{x}_\tau)), \quad \text{where} \quad \rho_q(z) = \begin{cases} qz, & \text{if } z \geq 0, \\ (q-1)z, & \text{if } z < 0, \end{cases} \tag{1}$$

where T is the number of input traffic patterns, and $\hat{Q}_Y(q|\mathbf{x}_\tau)$ is an approximation of the q-quantile returning traffic prediction $\hat{y}_\tau$ given $\mathbf{x}_\tau$, $\forall\ \tau = 1,..,T$. In this work, the q-quantile is parameterized by a recurrent DNN model trained to estimate $\hat{Q}_Y(q|X,\theta)$, where θ are the unknown parameters of the recurrent DNN model. Specifically, the model is optimized to minimize the loss function of Eq. (1), given a training dataset $D = (X,Y) = \{\mathbf{x}_\tau, y_\tau\}_{\tau=1}^{T}$.

3. Dataset Preprocessing

Dataset D was created according to the real traffic traces of the 12-node Abilene backbone network (http://sndlib.zib.de/home.action). This dataset provides bit-rate information for every pair of nodes in the network (i.e. traffic demand matrices) for 5-minute intervals over a 6 month period. For simplicity, in our problem setting, a different dataset D_s is created for each source node s in the network, with the traffic patterns in each dataset representing the aggregated bit-rate from each node for 5-minute intervals over the entire dataset. Given the aggregated bit rates, input traffic patterns, for each dataset D_s are created as: $\mathbf{x}_\tau = [\mathbf{f}_{\tau-1}, \mathbf{f}_{\tau-2}, ..., \mathbf{f}_{\tau-w}]$, where $\mathbf{f}_{\tau-i} = [f^1_{\tau-i}, f^2_{\tau-i}, ..., f^k_{\tau-i}]$ $\forall i = 0,1,..,w$ and $\forall \tau = 1,2..,T$, with τ representing a network planning interval, w is the number of previous planning intervals considered, and $\mathbf{f}_\tau$ is a vector consisting of k traffic fluctuations in τ. Hence, ground-truths of each D_s are created as: $y_\tau = \max\{\mathbf{f}_\tau\}$, $\forall \tau = 1,2..,T$,, targeting the prediction of the maximum traffic demand that may occur in τ.

Since in the original dataset bit-rate information is provided in 5-minute intervals, for the creation of the datasets we assume that fluctuations occur in the same time scale. Specifically, network reconfigurations occur every 30 minutes, resulting in $k = 6$ fluctuations within each planning interval. For the input patterns, fluctuations of the previous $w = 6$ planning intervals are considered. In total, each dataset D_s consists of $T = 800$ sequential (in time) patterns, spanning days $1-17$ in the original Abilene dataset. Before model training, input patterns are scaled into the range $[0,1]$. Eighty percent of the patterns are used for model training/validation and the rest are used for testing purposes (i.e., the last 160, spanning 3.5 days ahead).

4. Model Training and Evaluation

For each node, various recurrent DNN models were trained over their corresponding datasets D_s. Specifically, DNNs with simple recurrent units (RNNs), with long short-term memory (LSTMs) units, and with gated recurrent units (GRUs) were examined over the same set of hyperparameters. A modified implementation of [9] was used that included the quantile loss function. All DNNs achieved similar accuracies with GRUs slightly outperforming the rest. Hence, the rest of the analysis follows the GRU model, initialized with 1 layer with 25 hidden units (up to 4 layers were considered with no significant improvements). For model training, the Adam algorithm was used, with 0.001 learning rate, 32 batch size, and for 500 epochs. Indicatively, Fig. 1 provides GRU training evolution (up to 30 epochs) over various loss functions utilizing the dataset for the ATLAM5 node. For all loss functions, the GRU model converged to a sufficiently low validation accuracy (0.11 for mean squared error (MSE) (benchmark approach), 0.09 for 0.9-quantile, 0.06 for 0.95-quantile after 500 epochs). Similar training behaviors were observed for all datasets, resulting also to sufficiently low accuracies within their corresponding test patterns.

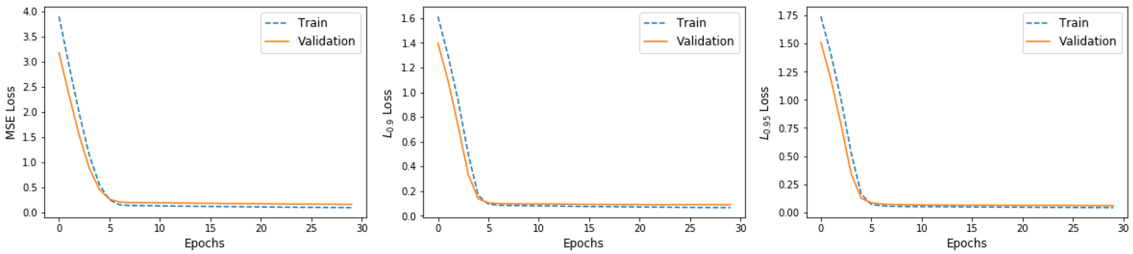

Fig. 1. small DNN-GRU training evolution for MSE, 0.9-quantile, and 0.95-quantile loss functions.

Importantly, we validated the q-quantiles over all the test datasets, resulting to 0.87 and 0.91 validation accuracies for the 0.9- and 0.95-quantiles, respectively. These results were obtained by comparing the true demand with the q-quantile predictions to compute the percentage of true values that fall below the upper quantile predictions (i.e., to validate that the quantiles are sufficiently approximated). According to the results, both quantiles are well approximated even though improvements may be possible by fine-tuning DNN model hyperparameters (this is out of the scope of this work). Note that, for the MSE loss function, 45% of the predictions fall below the true values, whereas for the 0.9- and 0.95-quantiles this percentage is equal to 13% and 9%; an indicator that quantile predictions lead to fewer QoS violations when considered for spectrum allocation (SA) decisions.

5. Network Performance Evaluation

The impact of each prediction approach on the SA decisions is evaluated on the 24-node 43-link USNET [4] assuming an elastic optical network (EON) operating with BPSK, QPSK, 8-QAM, and 16-QAM modulation formats. Each fiber link has a capacity of 320 frequency slots (FSs), spaced at 12.5GHz, and with a baud rate of 10.5Gbaud. For each one of the 12 GRU models, representing a source node, a destination node d is randomly generated. The routing and spectrum (RSA) problem is solved according to the κ-shortest paths algorithm ($\kappa = 5$) followed by the first-fit SA scheme. Specifically, for the SA decisions (i.e., number of FSs to be allocated) the following network planning approaches are examined:

• **Static with Empirical Margin (EM)**: For each $s-d$ pair, the RSA is solved to serve the traffic demand over all future planning intervals. Hence, the maximum bit-rate predicted by the MSE-based GRU model over all patterns in test dataset D_s^t is considered. To account for prediction uncertainty, the maximum bit-rate is increased prior to SA by the *empirical margin* computed as: $m = \max_{\hat{y}_\tau < y_\tau} \{y_\tau - \hat{y}_\tau\}_{\tau:1,\ldots,|D_s^t|}$.

• **Multi-period with EM**: For each $s-d$ pair, the RSA is solved for every planning interval $\tau \in D_s^t$, where the bit-rate $\hat{y}_\tau$ of the MSE-based GRU model is considered. To account for prediction uncertainty, each $\hat{y}_\tau$ is increased prior to SA by empirical margin m.

• **Multi-period with q-Quantile Margin (q-QM)**: For each $s-d$ pair, the RSA is solved for every planning interval $\tau \in D_s^t$, where the bit-rate $\hat{y}_\tau$ predicted by the q-quantile-based GRU model is considered. As quantile predictions already account for the prediction uncertainty, no additional margin is considered.

For all scenarios, bit-rates are multiplied by 50 to bring the spectrum demands in reasonable levels, given the capacity of USNET. The spectrum demands (in FSs) are computed according to a conventional distance adaptive modulation scheme and the SAs are evaluated against the true fluctuations in D_s^t. Table 1 summarizes the results for various metrics, i.e., the number of unutilized FSs and the unserved traffic (in Gbps) averaged over all the fluctuations and $s-d$ pairs. All cases resulted in zero blocking. Clearly, multi-period approaches significantly

Table 1. Evaluation of SA decisions for various network planning schemes.

	Static with EM	Multi-period with EM	Multi-period with 0.9-QM	Multi-period with 0.95-QM
Av. no. of Unutilized FSs	13.1	7.8	1.4	1.6
Av. Unserved Traffic in Gbps	0	0	0.76	0.78

outperforms the static scheme in unutilized FSs, with q-QM achieving up to 89% and 82% spectrum savings compared to the static and multi-period with EM cases, respectively. While the q-QM approach results in some unserved traffic, this was successfully provisioned on-line [5]. On the average, 1.5 lightpaths were provisioned on-line per planning interval, in contrast to 5.5 lightpaths when considering a multi-period planning scheme without any margin (i.e., considering the MSE-based GRU predictions). Hence, q-quantiles result in 72% reduction on the operational overhead compared to the case where prediction uncertainty was completely ignored. Comparatively, 0.9- and 0.95-quantiles perform similarly, indicating that q is sufficiently high for our dataset.

6. Conclusions

Considering the certainty level of traffic predictions for SA decisions leads to significant capacity and operational savings. Quantile regression is a promising approach towards this direction. Accounting for various QoS requirements and comparisons with other ML methods (e.g., MC dropout) are planned as future work.

Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 739551 (KIOS CoE - TEAMING) and from the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

References

1. S. Troia, et. al., "Deep learning-based traffic prediction for network optimization," *ICTON*, 2018.
2. T. Panayiotou, et. al., "A data-driven bandwidth allocation framework with QoS considerations for EONs," *JLT*, **37**(9), 1853, 2019.
3. A. Azzouni, et. al., "NeuTM: A neural network-based framework for traffic matrix prediction in SDN," *NOMS*, 2018.
4. Y. Xiong, et. al., "Lightpath management in SDN-based elastic optical networks with power consumption considerations," *JLT* **36**(9), 1650, 2018.
5. R. Alvizu, et. al., "Energy efficient dynamic optical routing for mobile metro-core networks under tidal traffic patterns," *JLT*, **35**(2), 325, 2017.
6. F. Rodrigues, et. al., "Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems", *IEEE Trans. Neural Netw. Learn. Syst.*, **31**(12), 5377, 2020.
7. T. Panayiotou, et. al., "Deep quantile regression for QoT inference and confident decision making", *ISCC*, 2021.
8. R. Koenker, "Fundamentals of quantile regression", *In Quantile Regression*, Cambridge University Press, 2005.
9. J. Yoon, "Time-series prediction with RNN, GRU, LSTM and attention", [Online] https://github.com/jsyoon0823/Time-series-prediction.