Knowledge Distillation-Based Compression Model for QoT Estimation of an Unestablished Lightpaths

Knowledge Distillation-Based Compression Model for QoT Estimation of an Unestablished Lightpaths

Fehmida Usmani[1], **Ihtesham Khan**[2], **Muhammad Umar Masood**[2], **Arsalan Ahmad**[1], **Vittorio Curri**[2]

[1]*National University of Sciences & Technology (NUST), Islamabad, Pakistan*
[2]*Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy*
e-mail: fusmani.dphd18seecs@seecs.edu.pk

ABSTRACT A precise Quality-of-transmission (QoT) estimation of a Lightpath (LP) before its deployment is a key step in effective network design and resource utilization. Deep neural network-based methods have recently shown promising results for QoT estimation tasks. However, these methods contain a large number of parameters and require heavy computational resources for accurate predictions. To this end, we propose a novel Knowledge distillation (KD) based compression method to obtain a compact and more accurate model for QoT estimation. Our simulation results demonstrate that the model trained using KD significantly improves accuracy with reduced parameters and computational complexity. To the best of our knowledge, this is the first time that the knowledge distillation technique has been used to estimate the QoT of an unestablished LP.
Keywords: Quality-of-transmission, Machine learning, Knowledge distillation.

1. Introduction

The estimation of the QoT of an LP is vital to the effective design and operation of optical networks. The adoption of Machine Learning (ML) approaches for QoT estimation is a contemporary substitute to analytical models, like the Gaussian Noise (GN) model, which applies conservative ways to cater to the model deficiencies and adjust generalizations [1]–[3]. Most of these data-driven strategies are based on deep learning models, which involve multiple neural network layers to discover the underlying patterns of the data. Generally, deep learning-based complex models with a large number of parameters demonstrate excellent performance in terms of accuracy and generalization, but they are challenging to deploy in a real operational network due to their large storage requirements and enormous computational complexity. Additionally, training these huge models is time-consuming. To tackle this problem, the authors in [4] first presented compression as a solution to this problem to transfer knowledge from the larger model to the smaller model without degrading the performance. Recently, Knowledge distillation (KD) has gained a lot of attention from the research community; it aims to transfer knowledge from a large model (teacher model) that performs best into a smaller model (student model) in terms of size, computational resources, and prediction performance [5]. Fig. 1 depicts the general framework of KD between teacher-student networks. The basic idea is that the student model imitates the teacher model to achieve better performance. The key components of the KD system include teacher architecture, the KD algorithm, and student architecture. KD is a promising solution to reduce the model complexity while keeping its generalization capabilities as much as possible. The effectiveness of the KD technique is demonstrated in a wide range of applications such as machine translation quality estimation [6], wind-power estimation [7], and acoustic-event detection [8]. Generally, for QoT estimation tasks, previous works employ similar models for the training and deployment stages even though both stages have pretty different requirements [9]. We can train the complex model to extract important knowledge, but it is not necessarily required to deploy the same model in a real operational network because it utilizes a significant amount of computational resources and takes more time to make predictions.

In this direction, we propose the novel KD-based framework to classify LP QoT as good or bad before deployment. The basic idea is to directly distill the larger QoT model into a smaller, lightweight student model with different architecture. To the best of our knowledge, it is the first time the KD approach has been proposed for a QoT estimation task. The main contribution of our work is to develop a novel KD-based framework for classifying the LP QoT into good or bad before deployment. We propose a response-based KD approach that mimics the teacher model's final layer outputs for knowledge distillation.

2. Simulation and Dataset Generation

This work considers a software-defined optical network with an Optical line system (OLS) serving as the edges and Re-configurable optical add-drop multipliers (ROADMs) acting as the nodes. The OLSs assumed are running at their optimal operating point, and the amplifier's noise figure and ripple gain are the sole variables accounting for the physical layer's perturbation behavior—the spectral load changes, which causes the gain-ripples to oscillate. OLS controllers can thus ensure that they operate at the nominal operating point even with considerable working point variability. During transmission, the LPs are affected by several impairments, but Amplified-spontaneous noise-(ASE) and Non-linear interference (NLI) is particularly notable. Statistically, independent ASE noise is introduced at each In-line amplifier (ILA), and it builds up as the signal travels

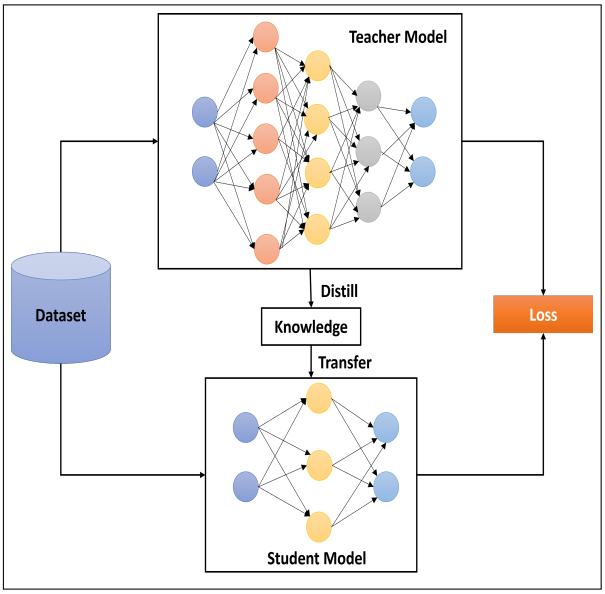

Figure 1: Knowledge Distillation Framework.

Model	Number of epochs	Accuracy (EU)	Accuracy (USA)
Teacher	100	0.979	0.887
Student	90	0.993	0.899
Teacher	90	0.969	0.877
Student	80	0.992	0.887
Teacher	80	0.957	0.866
Student	70	0.987	0.872
Teacher	70	0.938	0.857
Student	60	0.970	0.856
Teacher	60	0.915	0.849
Student	50	0.965	0.850
Teacher	50	0.890	0.837
Student	40	0.947	0.845
Teacher	40	0.878	0.817
Student	30	0.943	0.831
Teacher	30	0.844	0.817
Student	20	0.921	0.827

TABLE I: Detailed performance analysis of teacher-student models.

across the network. However, there is a statistically significant relationship between the NLI of each span and another The QoT metric is expressed as the GSNR of each LP traversing across the OLS taking into account both ASE and NLI, using $\mathrm{GSNR_i} = \frac{P_{\mathrm{S,i}}}{P_{\mathrm{ASE}(f_i)}+P_{\mathrm{NLI,i}}(f_i)}$, where for the ith channel with central frequency f_i, $P_{\mathrm{S,i}}$ is the signal launch power, $P_{\mathrm{ASE}(f_i)}$ is the amplified spontaneous emission while $P_{\mathrm{NLI,i}}(f_i)$ is fiber nonlinear interference. Furthermore, the overall $GSNR$ of a given LP traversing across the OLS: is given by $\frac{1}{GSNR} = \sum_n \frac{1}{GSNR_n}$ where n is the number of OLSs, the LP passed along a specific path. The ASE and NLI over the particular path are taken into account by the GSNR metric. The GSNR precisely determines the Bit error rate (BER) by examining the transceiver's back-to-back profile. The simulation scenario considers a grid size of 50 GHz with 76 C-band channels. Due to a lack of processing resources, only 76 channels with a total bandwidth of approximately 4 THz are considered. A root-raised cosine filter shapes the 32 GBaud signals generated by the transmitter. Operating in a constant output power mode of 0 dBm per channel, an Erbium-doped fiber amplifier (EDFA) maintains the launch power of the signal at 0 dBm. The EDFA noise figure remains constant between 3.5 dB and 4.5 dB, while the ripple gain varies by no more than 1 dB. All links are expected to use conventional single-mode fiber (SSMF). Also included are fiber impairments such as attenuation (α) = 0.2dB/km and dispersion (D) = 16ps/nm/km.

The physical layer abstraction is provided by an open-source GNPy package that is utilized to simulate the scenario and generate synthetic datasets [10]. The GNPy package generates physical layer network models utilizing an end-to-end simulation environment. Two unique network topologies are used to build the synthetic dataset: the European Union (EU) network and the United States (US) network. Regarding fiber and optical network elements, both networks are identical. However, they differ concerning the amplifier's sensitive characteristics (noise and ripple gain) and fiber insertion losses. The spectral load realization for each simulated link in a dataset is a subset of 2^{76}, where 76 is the number of channels. We evaluated 3000 realizations of arbitrary traffic flows, including between 34% and 100% of the total operating bandwidth for each source-to-destination (s $\rightarrow$ d). The dataset for this study contains 6 pairs (s $\rightarrow$ d) pairs from the EU network and 11 (s $\rightarrow$ d) pairs from the USA network. Thus, the EU network topology generates 18,000 realizations, while the US network topology generates 33,000.

3. Knowledge Distillation Framework for QoT Estimation

This work estimates the QoT of an unestablished LP before its deployment. The more extensive Artificial neural network (ANN) model (teacher) is trained to extract the underlying relationship in the data and then employ the KD training approach to transfer the learned knowledge from the larger ANN model to the smaller ANN model (student), which is more appropriate for deployment in a real operational network.

Both the teacher and student models are well-trained on the large dataset obtained from the EU and US networks. The input feature space for teacher and student ANN models includes power, number of spans, ASE noise, and NLI for classifying LP into good or bad QoT. The proposed framework is developed using the high-level Keras Application program interface, built on the TensorFlow platform. Several sections of the proposed framework are described below.

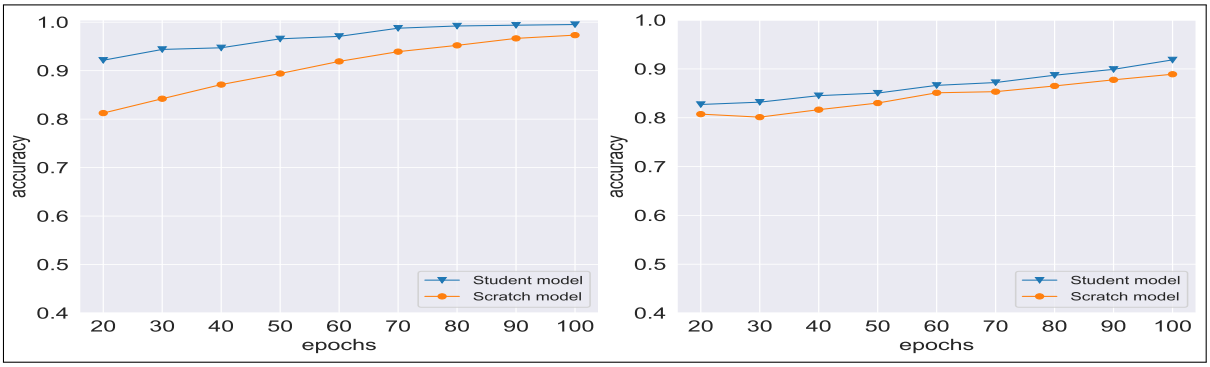

Figure 2: Performance comparison of student model trained with KD approach and model trained from scratch using conventional training scheme on EU (left) and US (right) network.

ANN teacher Model: The proposed ANN teacher model consists of an input layer with 306 neurons, three hidden layers with 128, 64, and 64 neurons, respectively, and an output layer with two neurons. Each hidden layer employs a ReLU-based activation function, whereas the softmax activation function is implemented at the output layer for the QoT score prediction of the LP. The model is trained for 100 epochs on 10,000 data samples obtained from the EU network and 21,000 samples from the US network using an adaptive learning rate optimizer (ADAM) and cross-entropy loss function.

ANN student Model: The proposed lightweight ANN-based student model consists of an input layer with 306 neurons, one hidden layer with 32 neurons, and an output layer with two neurons. The ReLU-based activation is applied in the hidden layer and softmax is used in the output layer. The student model is trained using the knowledge distillation technique (described in next section) with 90 epochs.

Knowledge Distillation: The proposed KD framework works as follows: Initially, the dataset related to LP QoT is fed to the huge teacher model to pre-train it to learn the underlying data representations and to produce the prediction results. The learned knowledge is transferred to a lightweight student model in the next step. An intuitive method to transfer the learned knowledge or generalizability of a bigger teacher model to a simpler student model is to use the teacher model's class probabilities as "soft targets" to train the student model. With soft targets, we get significantly better distillation results than hard targets. The hard target provides information only about the predicted label, whereas the soft target gives information about the predicted probabilities of all the given classes, which enhances the distillation performance significantly [5]. We propose applying a response-based KD approach for the given LP QoT estimation. This approach's key idea is to accurately mimic the teacher model's final-layer outputs. The distillation loss for the response-based KD approach is defined as follows:

$$\mathrm{L_D}(z_{\mathrm{teacher}}, z_{\mathrm{student}}) = \mathcal{L}_{\mathrm{DL}}(z_{\mathrm{teacher}}, z_{\mathrm{student}}) \ , \tag{1}$$

where z_{teacher} and z_{student} indicate the logits of teacher and student and $\mathcal{L}_{\mathrm{DL}}$ represents the divergence loss of logits, respectively. It is to note that logits represent the output of the last fully connected layer of the ANN. We compute the soft targets that give us the probabilities for each class to which the input belongs. The softmax function is applied to compute soft targets as : $\mathrm{p_i} = \frac{exp(z_\mathrm{i}/T)}{\sum_j exp(z_\mathrm{j}/T)}$, where z_i denotes the logits for i-th class, and T is the temperature factor introduced to regulate the significance of each soft target [5]. When T is set to a larger value, it generates a softer probability distribution across classes. This equation is used to transfer the knowledge learned from the teacher to a student model. The student model is trained using the soft targets produced by the teacher model. Soft targets carry useful hidden knowledge from the teacher model. In light of this, distillation loss for soft logits is redefined as follows:

$$\mathrm{L_D}(p_{\mathrm{teacher}}, p_{\mathrm{student}}) = \mathcal{L}_{\mathrm{DL}}(p_{\mathrm{teacher}}, p_{\mathrm{student}}) \ , \tag{2}$$

To compute divergence loss $\mathcal{L}_{\mathrm{DL}}(p_{\mathrm{teacher}}, p_{\mathrm{student}})$, we optimize Eq. 2 to match the student and teacher logits. We use the cross entropy loss for the student model which computes the difference between true label (y) and the soft logits of the student model as follows: $\mathcal{L}_{\mathrm{student}}(y, (p_{\mathrm{student}}))$. Our knowledge distillation solution involves increasing the temperature T of the final softmax layer of the teacher model until it generates the appropriate soft targets. The smaller student model is then trained to mimic the soft targets using the same value of T. We set the T value to 10 for our simulation.

4. Performance Evaluation

In this section, we assess the performance of the KD training approach to estimate QoT. Our proposed method allows the operator to predict the QoT (GSNR) state of the forthcoming LP before its deployment. The proposed solution functions as a binary classifier and performs the LP classification based on the GSNR estimation. Eq. 4.

Model	Number of Parameters	Train time (sec)	Prediction time (sec)
Teacher	145,784	47.122	0.607
Student	96,416	20.525	0.408

TABLE II: Performance comparison of teacher-student models

The $OSNR$ sensitivity threshold at the receiver taken into consideration in this research is based on [11] where $\mathrm{GSNR} > OSNR_{\mathrm{Rx}} \longrightarrow 1$ $(otherwise\ 0)$. The performance of the proposed framework is evaluated on EU and US networks using an accuracy metric. We consider 8000 test samples acquired from the EU network and 12000 test samples from the US network. Firstly, we compare the performance of the teacher model with the student model, which is trained using the KD approach. Our simulation results are reported in Tab. I. We can see that the student model can achieve better accuracy with fewer epochs than the teacher model for both considered networks, i.e., the EU and US. We further compare the performance of the teacher and student model in terms of several parameters and train and prediction time, as given in Tab. II. Our student model has fewer parameters, leading to less training and prediction time. It is noted that the train and test time is computed for 100 epochs. The results reported in Tab. I and Tab. II validate the effectiveness of the KD approach for the given QoT scenario. The student model is more compact, fast, and accurate than the teacher model. To further analyze the performance of a proposed KD approach, we compare the performance of a student model with the scratch model, which is considered a baseline model trained from scratch while using the traditional training method. Our proposed student model and the scratch model use the same ANN architecture. In Fig. 2, we plot the accuracy achieved by the student and scratch model against the number of epochs. We varied the number of epochs from 20 to 100. As we increase the number of epochs, the performance of both models increases. As we can see, our proposed student model achieves 99% accuracy with 100 epochs for the EU network and 91% accuracy for the US network. Analyzing the results shows that the model trained with the KD approach outperforms the model trained from scratch using the conventional training approach.

5. Conclusion and Future Work

This work explored the novel ML model training framework, which utilizes the KD teacher-student approach to train the compact, fast, and more accurate model. We validated this approach's effectiveness for classifying LP QoT into good or bad. It is demonstrated in the results that the proposed lightweight KD-based model achieves significantly better results than the teacher model and the scratch model. Furthermore, it is a more suitable model for deployment in real-time networks operation because of its small size, better accuracy, and faster evaluation speed. For future work, it is worthwhile to investigate the transfer learning approach with KD for QoT estimation to fully realize this approach's potential.

REFERENCES

[1] I. Khan, M. Bilal, and V. Curri, "Assessment of cross-train machine learning techniques for qot-estimation in agnostic optical networks," OSA Continuum **3**, 2690–2706 (2020).

[2] T. Panayiotou, S. P. Chatzis, and G. Ellinas, "Performance analysis of a data-driven quality-of-transmission decision approach on a dynamic multicast-capable metro optical network," JOCN **9**, 98–108 (2017).

[3] F. Usmani, I. Khan, M. Siddiqui, M. Khan, M. Bilal, M. U. Masood, A. Ahmad, M. Shahzad, and V. Curri, "Cross-feature trained machine learning models for QoT-estimation in optical networks," Optical Engineering **60**, 125106 (2021).

[4] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Knowledge Discovery and Data Mining,* (2006).

[5] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," ArXiv **abs/1503.02531** (2015).

[6] A. Gajbhiye, M. Fomicheva, F. Alva-Manchego, F. Blain, A. Obamuyide, N. Aletras, and L. Specia, "Knowledge distillation for quality estimation," in *Findings,* (2021).

[7] H. Chen, "Knowledge distillation with error-correcting transfer learning for wind power prediction," (2022).

[8] G. Cerutti, R. Prasad, A. Brutti, and E. Farella, "Compact recurrent neural networks for acoustic event detection on low-energy low-complexity platforms," IEEE Journal of Selected Topics in Signal Processing **14**, 654–664 (2020).

[9] S. Aladin, A. V. S. Tran, S. Allogba, and C. Tremblay, "Quality of transmission estimation and short-term performance forecast of lightpaths," Journal of Lightwave Technology **38**, 2807–2814 (2020).

[10] A. Ferrari, M. Filer, K. Balasubramanian, Y. Yin, E. Le Rouzic, J. Kundrat, G. Grammel, G. Galimberti, and V. Curri, "Gnpy: an open source application for physical layer aware open optical networks," JOCN **12** (2020).

[11] "Cisco transceiver modules - cisco 400g digital coherent optics qsfp-dd optical modules data sheet," (2021).