Multi-Span Optical Power Spectrum Evolution Modeling using ML-based Multi-Decoder Attention Framework

Agastya Raj[*,(1)], Zehao Wang[(2)], Frank Slyne[(1)], Tingjun Chen[(2)], Dan Kilper[(1)], Marco Ruffini[(1)]

[(1)] CONNECT Centre, School of Computer Science and Statistics and School of Engineering, Trinity College Dublin, Ireland *rajag@tcd.ie

[(2)] Duke University, Department of Electrical and Computer Engineering, Durham, NC, USA

Abstract *We implement a ML-based attention framework with component-specific decoders, improving optical power spectrum prediction in multi-span networks. By reducing the need for in-depth training on each component, the framework can be scaled to multi-span topologies with minimal data collection, making it suitable for brown-field scenarios. ©2024 The Author(s)*

Introduction

Today's network operators depend on advanced networks, which include reconfigurable optical Add-Drop Multiplexer (ROADM) systems using flex-grid Dense Wavelength Division Multiplexing (DWDM), for high-speed and low latency services. As these networks become more configurable and adaptable, accurate estimation of optical link performance, including power spectrum evolution and Optical Signal-to-Noise Ratio (OSNR), has become crucial[1]. DWDM signals experience different propagation characteristics through various components, for example due to the wavelength dependent gain of Erbium-Doped Fiber Amplifiers (EDFAs) and losses induced by fibers and Wavelength Selective Switches (WSSs). The challenge largely arises from the difficulty in developing accurate models of devices like EDFAs, where different units can show significant variations, for example, in their wavelength-dependent gain.

Traditional monolithic end-to-end system models, when applied to a large network, require extensive in-field data collection and may overlook the nuanced interactions between components, which are important for understanding power evolution through the network[2],[3]. In addition, these models are limited in the sense that any network topology change will require training of new models, with extensive collection of new data[4]. It has been shown that individual components such as EDFAs can be characterized using Machine Learning (ML) to predict optical power spectrum. However, in multi-span networks involving multiple such components, a direct cascade of these models perform poorly, resulting in high error accumulation[5],[6]. Recently, a Cascaded Learning (CL) framework was proposed using component-level models for each EDFA in a multi-span network, utilizing additional end-to-end measurements[7]. This method enabled models trained in the lab to be applied in the field as a scalable approach for large networks. However, this process requires detailed characterization of each EDFA before deployment in order to achieve low end-to-end prediction errors. This motivates the challenge of finding the right balance of lab data collection and field measurements–minimizing the quantity and complexity of the field measurements while achieving high accuracy.

In this paper, we introduce a novel approach by interpreting a network as a series of data points from Optical Channel Monitors (OCMs) deployed within in-line ROADMs. We propose a sequential Multi-Decoder Attention Model (MDAM) that leverages intermediate data to accurately emulate power spectrum evolution across a network. This is achieved by encoding the input signal with a shared attention-based Long Short-Term Memory (LSTM) encoder, and utilizing component-specific decoders to predict the power spectrum at each network node, improving power spectrum evolution prediction in multi-span networks. Instead of cascading discrete component-models, the encoder maintains a shared information layer through the model, while decoders enable accurate predictions for specialized component modeling in a unified and scalable framework.

Another key issue in modeling power spectrum evolution is the limited number of measurements that might be available for green field and brown field scenarios[1]. In brown field scenarios, working with fewer measurements is the key because of the high cost/complexity associated with live network probing, as these could introduce impairments on existing live channels. In green field scenarios, using less training data speeds up the process of accurately modeling each device. To enable high adaptability and ensure minimal data collection in these scenarios, we introduce a novel Transfer Learning (TL) process. We first develop a base model of encoder and device-specific decoders within a controlled laboratory setup, featuring a single physical device for each component type. TL is then applied to the other devices in a multi-span topology to scale up the model. This strategy can, for example, be used to first carry out in-depth data collection in a laboratory environment, and reduce the number of data points from the live network. This greatly simplifies model training and deployment, requiring in-depth characterization of only single physical devices to predict power spectra for unseen, multi-span networks.

In order to study the performance of MDAM and

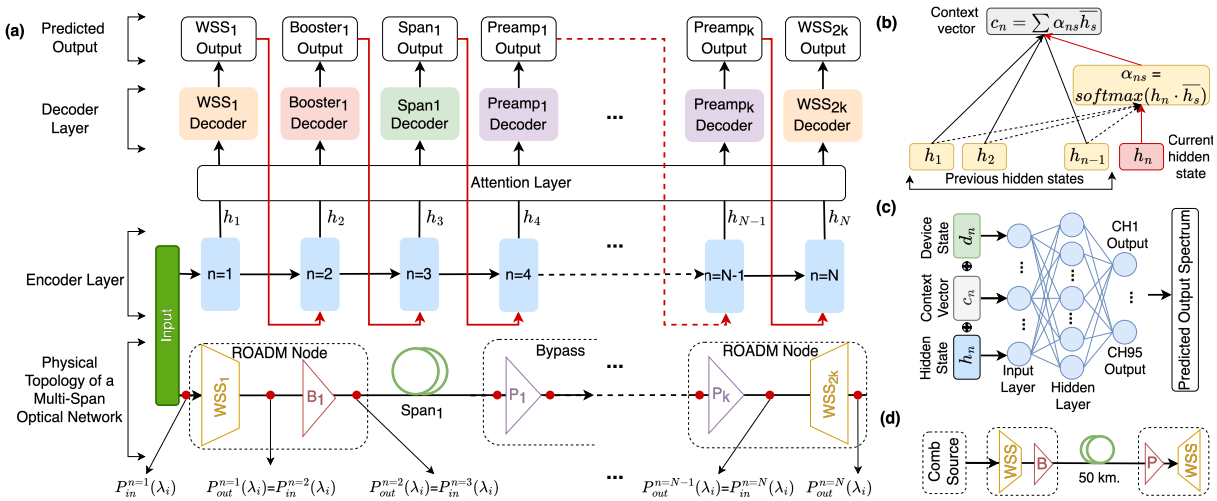

Fig. 1: (a) Multiple-Decoder Attention Model (MDAM) architecture, (b) Dot-product attention layer mechanism for n^{th} component, (c) Component-specific decoder model structure, (d) Base model training setup

End-component absolute prediction error: Mean/95$^{\text{th}}$ percentile (dB)	Random			Goalpost		
	Direct Cascade	Benchmark	MDAM	Direct Cascade	Benchmark	MDAM
Topo. #1 (6-span, 234 km): 40 km–40 km–40 km–*32 km*–*32 km*–50 km	1.73/4.39	0.16/0.43	0.16/0.27	2.13/2.60	0.59/1.48	0.20/0.49
Topo. #2 (4-span, 234 km): 40 km–*72 km*–*72 km*–50 km	0.61/2.03	0.16/0.42	0.14/0.21	1.03/1.28	0.58/1.43	0.24/0.47

Tab. 1: Mean/95$^{\text{th}}$ percentile absolute error of end-component predictions for different topologies (*italics* spans indicate field fibers)

its ability to carry out transfer learning across different networks, we carry out experiments across two different testbeds. A first model pre-training in OpenIreland[8] is followed by a transfer towards two multi-span target topologies, over 234 km of fiber, in the COSMOS PAWR Testbed[9]. Experimental results show that our approach achieves a 50-fold reduction in the amount of training data with respect to a state of the art benchmark model[7], while also improving prediction accuracy.

Model Architecture

The data from OCMs in a transmission system with N components comprises a sequence of power spectrum values for each component[10] $\overline{P}(\lambda_i) = [P_0(\lambda_i), P_1(\lambda_i), P_2(\lambda_i), ...P_N(\lambda_i)]$, which defines the environment. $P_0(\lambda_i)$ is the input power spectrum at first ROADM's EDFA, and $P_n(\lambda_i), n \in \{0, 1, \cdots, N\}$ denotes the power spectrum after component n. Given the initial power spectrum P_0, our objective is to predict power spectra after transmission through each component in the network. Since the output of each component serves as an input to the next component, this can be treated as an auto-regressive sequential modeling problem. For any n^{th} component, the input features $P_{in}(\lambda_i)$ and output features $P_{out}(\lambda_i)$ can be defined as $P_{n-1}(\lambda_i)$ and $P_n(\lambda_i)$ respectively.

We employ a sequential model architecture using a shared encoder and multi-decoder model with attention layer (refer to Fig. 1(a)). The encoder is a 3 layer LSTM model with a hidden size of 100 units each, with a dropout of 20% applied at each layer to reduce overfitting. Given the substantial variations in the physical behavior of different network devices, a single shared output layer proves inadequate for accurately modeling each component. Instead, we create component-specific decoders for each component used, namely, Booster, Preamp, Span and WSS. As shown in Fig. 1(c), decoders are shallow neural networks consisting of a single hidden layer of 100 neurons, and 95 neurons in the final layer, predicting the channel-wise power output. A deeper encoder helps the model accommodate diverse components across longer networks, while shallow decoders are sufficient to model a single component[11]. Dot-product attention is implemented to avoid error accumulation and dynamically focus on signal's evolution through the network, without adding parameter overhead[12]. As the signal progresses, at any step n, the encoder processes the output from previous step, generating a hidden state vector h_n. The attention layer (refer Fig. 1(b)) computes attention scores for all previous components using dot product attention where $score(n) = h_n \cdot \overline{h_s}$, with $\overline{h_s} = [h_1, h_2, \cdots h_{n-1}]$ denoting all the previous hidden states. Softmax is applied to these scores to derive attention weights $\alpha_{\mathbf{ns}}$, and a context vector c_n is calculated as a weighted sum of all previous hidden states weighted by their respective attention weights, give by $c_n = \sum \alpha_{\mathbf{ns}} \overline{h_s}$. This context vector is concatenated along with the current hidden state and corresponding device configuration features d_n(such as EDFA target gain, span length and WSS attenuation), to get the final output vector $o_n = h_n \oplus c_n \oplus d_n$. This output vector is then passed to the corresponding decoder to get the predicted power spectrum for that component. This predicted power spectrum serves as the input for next component step.

We develop the shared encoder and base models for four decoders types, each corre-

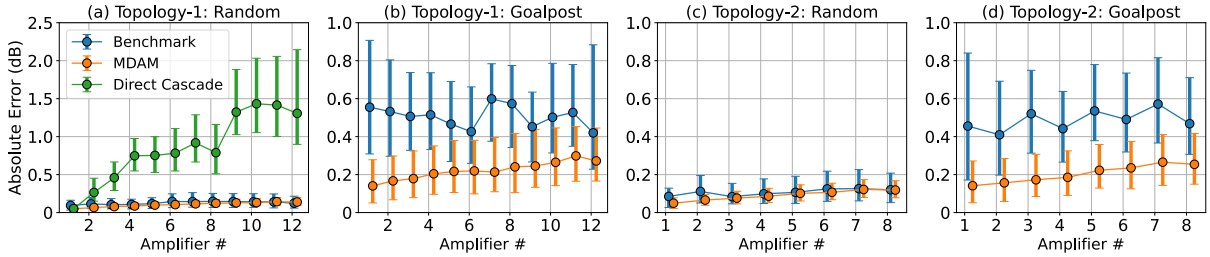

Fig. 2: Comparison of absolute error distribution in power evolution predictions for Benchmark vs MDAM model for Topology-1 in (a) Random, (b) Goalpost, and for Topology-2 in (c) Random and (c) Goalpost configurations. Markers denote the median, and the whiskers denote the inter-quartile range (25^{th}-75^{th} percentile). *(Comparison with Direct Cascade model shown only in (a) for clarity)*

sponding to a different component type (Booster, Preamp, Span, and WSS). This is implemented in a laboratory environment in the OpenIreland Testbed[8], shown in Fig. 1(d), where we carry out a total of 3,168 power spectrum measurements. The model undergoes a two-phase training process: **a. Teacher-Forcing Phase**: for the initial 3,000 epochs, training uses ground-truth power spectra to accelerate learning and reduce error propagation[13]. **b. Auto-Regression Phase**: the model is trained in auto-regressive mode for 9,000 epochs. Throughout both processes, Stochastic Gradient Descent (SGD) optimizer is used to optimize the total weighted Mean Absolute Error (MAE) across all components[14], with an initial rate of 1e-03, decaying exponentially every 1,000 epochs. The encoder and decoders are preserved separately for subsequent transfer learning.

Experimental Setup and Transfer Learning

To evaluate the performance of MDAM, we employ the PAWR COSMOS testbed in Manhattan, USA[9] as the target network, conducting transfer learning with base models developed in OpenIreland and then collect the performance results. We set up two multi-span topologies: one 6-span link with 12 EDFAs and one 4-span link with 8 EDFAs. The span configuration for the two topologies is summarized in Table 1, with total lengths of 234 km (including 64 km of Manhattan field fibers). For both topologies, all booster and pre-amplifier EDFAs are set to high gain mode with 18 dB gain and zero gain tilt. A comb source is used to generate 95$\times$50 GHz Wavelength Dependent Multiplexing (WDM) channels in the C-band, and channel configurations and spectrum flattening are managed at the initial MUX, with the signal traversing all spans and being dropped at the end DEMUX. We consider three types of channel loading configurations with different number of loaded channels: *(i)* Fixed (fully/half loaded), *(ii)* Random (loading with randomly selected channels), and *(iii)* Goalpost (groups of channel loading in different spectral bands). For transfer learning, we employ the pre-trained base encoder and replicate component specific decoders–assigning them to corresponding devices in the target network. In total, we use 48 measurements (fixed and random loading) for transfer learning into the COSMOS target network, using the same two-phase training process described in the previous section. However, we use a reduced initial learning rate of 1e-5 and a lower gradient clipping threshold of 0.5.

Results

We compare our model against a Direct Cascade of individually trained component-level models for each EDFA in the network, and the benchmark CL-Model[7]; with the test set consisting of 658 random and 27 goalpost data points. Table 1 summarizes the mean and 95^{th} percentile absolute errors of end-EDFA power spectrum prediction for two topologies across random and goalpost configurations. It can be seen that Direct Cascade suffers high accumulation of errors, while CL-Model and MDAM achieves a similar MAE for random configuration. However, MDAM displays improved predictions in extreme/edge cases, showing a lower 95^{th} percentile error. Moreover, MDAM outperforms the benchmark model in goalpost configuration, showing a more stable distribution of errors across diverse channel configurations. This improvement is particularly significant, given the reduction in measurements from over 160,000 in the direct cascade model and the benchmark study[7] to 3,216 in this work (of which only 48 is in the target network).

Fig. 2 shows the distribution of absolute prediction errors across both topologies for random and goalpost configurations. MDAM displays a reduced error accumulation through the network and a more stable error distribution. Especially in the goalpost scenario, MDAM outperforms the benchmark model with a consistent median absolute error$<$0.3 dB through all components. Note that MDAM jointly predicts power spectrum for all components along the network, other methods require separate models to be trained for individual devices. Additionally, MDAM can be fine-tuned with added measurements for network expansions, while other models require complete retraining for new network configurations.

Conclusion

We demonstrate a scalable ML-based model pre-trained on devices of the same manufacturer that can be generalized and transferred to a larger network. Our results show improved performance with respect to a state-of-the-art benchmark model, while achieving a 50-fold reduction in training data.

Acknowledgements

The work was supported by SFI grants 12/RC/2276-p2, 22/FFP-A/10598, 18/RI/5721 and 13/RC/2077-p2, and NSF grants CNS-1827923, OAC-2029295, CNS-2112562, CNS-2211944, and CNS-2330333.

References

[1] Y. Pointurier, "Design of low-margin optical networks", *Journal of Optical Communications and Networking*, vol. 9, no. 1, A9–A17, 2017. DOI: 10.1364/JOCN.9.0000A9.

[2] L. E. Kruse, S. Kühl, and S. Pachnicke, "Exact component parameter agnostic qot estimation using spectral data-driven lstm in optical networks", in *Optical Fiber Communication Conference (OFC) 2022*, Optica Publishing Group, 2022, Th1C.1. DOI: 10.1364/OFC.2022.Th1C.1.

[3] N. Morette, H. Hafermann, Y. Frignac, and Y. Pointurier, "Machine learning enhancement of a digital twin for wavelength division multiplexing network performance prediction leveraging quality of transmission parameter refinement", *Journal of Optical Communications and Networking*, vol. 15, no. 6, pp. 333–343, Jun. 2023. DOI: 10.1364/JOCN.487870.

[4] G. Liu, K. Zhang, X. Chen, *et al.*, "Hierarchical learning for cognitive end-to-end service provisioning in multi-domain autonomous optical networks", *Journal of Lightwave Technology*, vol. 37, no. 1, pp. 218–225, 2019.

[5] Z. Wang, E. Akinrintoyo, D. Kilper, and T. Chen, "Optical signal spectrum prediction using machine learning and in-line channel monitors in a multi-span roadm system", in *2022 European Conference on Optical Communication (ECOC)*, 2022, pp. 1–4.

[6] S. Kamel, H. Hafermann, D. Le Gac, *et al.*, "Osnr prediction for optical links via learned noise figures", in *2021 European Conference on Optical Communication (ECOC)*, 2021, pp. 1–4. DOI: 10.1109/ECOC52684.2021.9605932.

[7] Z. Wang, Y.-K. Huang, S. Han, T. Wang, D. Kilper, and T. Chen, "Multi-Span Optical Power Spectrum Prediction using ML-based EDFA Models and Cascaded Learning", en, in *Optical Fiber Communication Conference (OFC) 2024*.

[8] A. Raj, Z. Wang, F. Slyne, T. Chen, D. Kilper, and M. Ruffini, "Self-normalizing neural network, enabling one shot transfer learning for modeling edfa wavelength dependent gain", in *49th European Conference on Optical Communications (ECOC 2023)*, vol. 2023, 2023, pp. 748–751. DOI: 10.1049/icp.2023.2325.

[9] T. Chen, J. Yu, A. Minakhmetov, *et al.*, "A software-defined programmable testbed for beyond 5g optical-wireless experimentation at city-scale", *IEEE Network*, vol. 36, no. 2, pp. 90–99, 2022. DOI: 10.1109/MNET.006.2100605.

[10] Z. Wang, D. C. Kilper, and T. Chen, "Open edfa gain spectrum dataset and its applications in data-driven edfa gain modeling", *Journal of Optical Communications and Networking*, vol. 15, no. 9, pp. 588–599, Sep. 2023. DOI: 10.1364/JOCN.491901.

[11] X. Kong, A. Renduchintala, J. Cross, Y. Tang, J. Gu, and X. Li, "Multilingual Neural Machine Translation with Deep Encoder and Multiple Shallow Decoders", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021, pp. 1613–1624. DOI: 10.18653/v1/2021.eacl-main.138.

[12] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[13] R. J. Williams and D. Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks", *Neural Computation*, vol. 1, no. 2, pp. 270–280, Jun. 1989, ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.2.270.

[14] N. S. Keskar and R. Socher, "Improving generalization performance by switching from adam to sgd", *arXiv preprint arXiv:1712.07628*, 2017.