Transfer Learning for EDFA Gain Modeling: A Semi-Supervised Approach Using Internal Amplifier Features

Agastya Raj
Computer Science and Statistics, CONNECT
Trinity College Dublin
RAJAG@tcd.ie

Dan Kilper
Engineering, CONNECT
Trinity College Dublin

Marco Ruffini
Computer Science and Statistics, CONNECT
Trinity College Dublin

2024 IEEE Future Networks World Forum (FNWF) | 979-8-3503-7949-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/FNWF63303.2024.11028819

***Abstract*—The gain spectrum of an Erbium-Doped Fiber Amplifier (EDFA) has a complex dependence on channel loading, pump power, and operating mode, making accurate modeling difficult to achieve. Machine Learning (ML) based modeling methods can achieve high accuracy, but they require comprehensive data collection. We present a novel ML-based Semi-Supervised, Self-Normalizing Neural Network (SS-NN) framework to model the wavelength dependent gain of EDFAs using minimal data, which achieve a Mean Absolute Error (MAE) of 0.07/0.08 dB for booster/pre-amplifier gain prediction. We further perform Transfer Learning (TL) using a single additional measurement per target-gain setting to transfer this model among 22 EDFAs in Open Ireland and COSMOS testbeds, which achieves a MAE of less than 0.19 dB even when operated across different amplifier types. We show that the SS-NN model achieves high accuracy for gain spectrum prediction with minimal data requirement when compared with current benchmark methods.**

***Index Terms*—Optical Networks, Machine Learning, Erbium Doped Fiber Amplifier**

I. Introduction

Optical networks play a crucial role in supporting new services, due to their ability to meet the high bandwidth, low latency, and reliability requirements [1]. In addition, they are increasingly important for supporting access and metro optical convergence due to their ability to unify different network layers, enhance efficiency, and meet the growing demands for high-speed connectivity [2]. To transfer data over long distances and across access and metro domains, optical networks are amplified with Erbium-Doped Fiber Amplifiers (EDFAs) to boost optical signals to overcome fiber and link losses. The end-performance metrics such as Optical Signal-to-Noise Ratio (OSNR), depends on the accumulated noise through the network. Thus, characterizing gain spectrum of EDFAs is one of the key factors to design low margin optical networks, and efficient physical layer control and management.

The gain spectrum of an EDFA has a complex dependence on channel loading, pump power, and operating mode, which makes it difficult to achieve high accuracy with a theoretical model. Recently, Machine Learning (ML) techniques such as Neural Networks (NNs) have been used to build EDFA gain models [3], [4]. Other work [5] has produced generalized ML-based EDFA models using training datasets collected from multiple EDFAs of the same make and model, which are shown to achieve lower Mean Absolute Error (MAE) of the gain spectrum prediction across multiple devices of the same make. Although these models achieve high prediction accuracy, they do require a large number of measurements, which can be time-consuming and difficult to obtain if the EDFA is in a live network. Due to the complexity of the model, deep learning methods such as NN also suffer from non-convex training criteria and local minima, which complicate the training process especially with limited number of measurements.

Transfer Learning (TL) is a promising method to reduce the required number of measurements for gain spectrum modeling. TL is a machine learning technique to improve the learning in a new task through the transfer of information from a new domain [6]. Specifically, for modeling the gain spectrum of EDFAs, a base model can be trained on one EDFA which can then be retrained to characterize different devices by using a reduced number of measurements from the new device. Recently, it was demonstrated [7] that a single EDFA model can be transferred between different EDFAs of the same type using only 0.5% of the entire dataset, showcasing the potential for efficient model transfer in this domain. However, the application of transfer learning across amplifiers of different types (i.e., from an EDFA Booster base model towards an EDFA Preamp target model) requires further investigation. In addition, work to date has mostly relied on training data from external features, such as input power levels and output gain spectra, which may not fully capture the complex behavior of EDFAs.

In this paper, we implement and study a novel semi-supervised, self-normalizing NN approach (hereafter referred to as the SS-NN model) that characterizes the wavelength-dependent gain of an EDFA using just 256 labeled measurements along with additional unlabeled data (which are easier to obtain). By incorporating internal EDFA features that are typically available in commercial telecom equipment,

Supported by grants from SFI: 12/RC/2276_p2, 18/RI/5721, and 13/RC/2077_p2

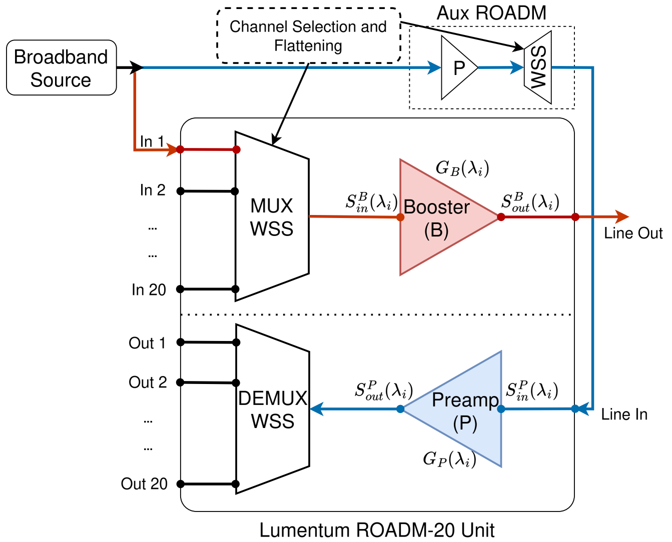

Fig. 1. Measurement setup for the Booster/Pre-amplifier EDFA in COSMOS and Open Ireland testbeds.

our model can be transferred to different EDFA types with only a single new measurement through transfer learning. We have reported our SS-NN model previously [8]. In this paper, we have updated the model to perform better in higher error configurations. Furthermore, we describe in more detail the architecture of the SS-NN model, and report the performance analysis. We evaluate our approach on 22 different EDFAs across the Open Ireland [9] (based in Dublin, Ireland) and PAWR COSMOS [10], [11] (based in Manhattan, USA) testbeds, achieving a MAE within 0.13 dB for same-type transfers and 0.19 dB for cross-type transfers.

II. Measurement Setup and Data Collection

In this section, we describe the experimental setup and data collection strategy of EDFAs from Open Ireland testbed and PAWR COSMOS testbed. Open Ireland testbed [9] is a re-configurable optical-wireless testbed in Dublin, Ireland. PAWR COSMOS testbed [10], [11] is a city-scale optical-wireless programmable testbed deployed in Manhattan, USA.

A. Experimental Setup

We carry out gain spectrum measurements across multiple wavelengths in the C-band from 3 commercial grade Lumentum ROADM-20 units deployed in the Open Ireland testbed and 8 similar units deployed in the PAWR COSMOS testbed. With each Lumentum ROADM-20 unit containing 2 EDFAs, we collect data from 11 Boosters and 11 Pre-amplifier EDFAs in total. Figure 1 shows the experimental topology. An Amplified Spontaneous Emission (ASE) broadband source is used to generate 95 X 50 GHz Wavelength Division Multiplexing (WDM) channels in the C-band according to the International Telecommunication Union (ITU) Dense Wavelength Division Multiplexing (DWDM) 50 GHz grid specification. To ensure consistency, we followed a similar measurement setup and data collection pipeline for both testbeds [12].

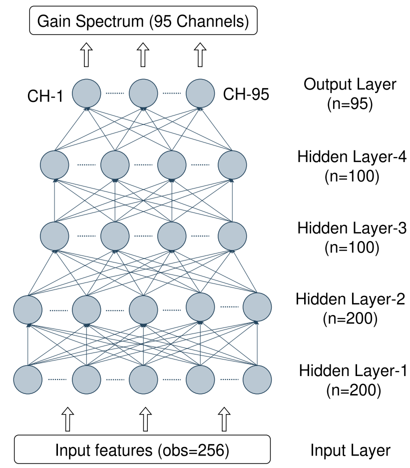

Fig. 2. SS-NN model structure with 5 layers.

In the data collection for Boosters, the MUX Wavelength Selective Switch (WSS) is used to flatten the channels, and control the power and channel loading configuration. For preamps, the broadband source output is connected to Line-IN port of an auxiliary Reconfigurable Optical ADD-DROP Multiplexers (ROADMs), whose DEMUX controls the power and channel loading configuration. The output of this auxiliary ROADM is forwarded to the Line-IN of the ROADM under test. The input and output power spectra for each of the 95 channels are collected through the built-in Optical Channel Monitors (OCMs). Additionally the total input/output power through the EDFAs are collected through built-in Photo-Diodes (PDs).

B. Measurement Configuration

In the Open Ireland testbed, all EDFAs were measured at target gains of 15/20/25 dB, while in the COSMOS testbed, the target gains were 15/18/21 dB for Boosters and 15/18/21/24/27 dB for Pre-Amplifiers in high gain mode with 0 dB gain tilt (we adopt different gain setting to emulate diversity of operation in different networks). The dataset includes 3,168 gain measurements (at multiple wavelengths) for each EDFA, for each given target gain settings, across 95×50 GHz channels in the C-band, for a total of 202,752 measurements from the COSMOS testbed, and 57,024 measurements from the Open Ireland testbed. In addition, measurements for each EDFA are collected under two channel loading modes: Random and Goalpost allocation [12] (i.e., loading groups of channels in different spectrum bands).

III. Model Architecture

In this section, we describe the Semi-Supervised Self-Normalizing Neural Network model for characterizing the gain spectrum of EDFAs.

A. SNN Model

Figure 2 shows the SS-NN model architecture, which consists of an input layer, four hidden layers with 200/200/100/100 neurons, and an output layer. The output layer consists of 95 neurons, predicting the wavelength

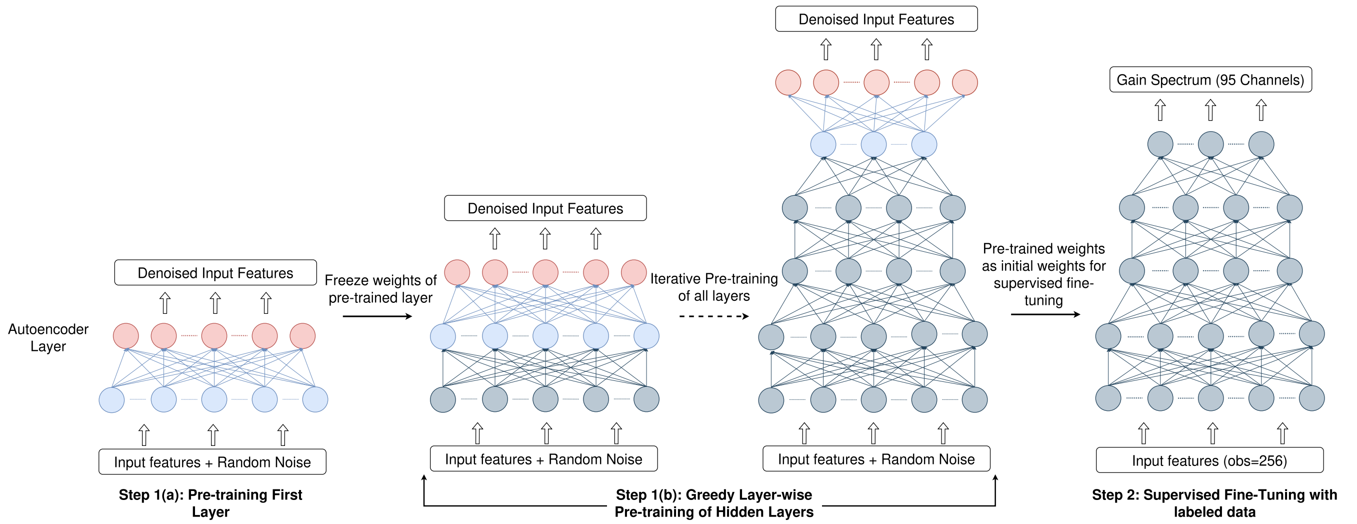

Fig. 3. SS-NN model training framework. Step 1(a) and (b) show the greedy layer-wise pretraining of hidden layers using unsupervised pretraining. This pre-trained model forms the basis for Step 2, where supervised fine-tuning is performed with 256 labeled measurements.

dependent gain output. The input features to the SS-NN model includes the EDFA target gain setting at constant-gain configuration (G_0), total EDFA input power (P_{in}, total EDFA output power (P_{out}, input power spectrum $\overline{P}(\lambda_i) = [P(\lambda_1), P(\lambda_2), P(\lambda_3), ...P(\lambda_{95})]$; and a binary vector indicating the channel-loading configuration denoted by $\overline{C} = [c_i]_{i=1}^{95}$, with

$$c_i = \begin{cases} 1, & \text{if the } i^{th} \text{ wavelength channel is switched on} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In addition, we utilize three additional features related to the value of the internal Variable Optical Attenuator (VOA) in the EDFA, namely total VOA input and output power (P_{in}^V and P_{out}^V), and attenuation (P_{attn}^V). VOAs are an internal component of EDFAs, which indirectly influence the shape of the gain profile by acting on the signal's input powers. This is done to ensure the EDFA operates in its design average inversion for a flat spectrum gain profile which matches the Gain Flattening Filter (GFF) attenuation [13]. The VOA attenuation is controlled automatically in the EDFA based on the model's gain dynamic range, and it grants intrinsic information on the operation of each EDFA. Typically, ML models for gain spectrum prediction rely only on input and output power spectra information to predict the gain spectrum. However, this choice also treats every EDFA like a black box, which leads to poor performance in transfer learning.

The output layer predicts the gain spectrum $\overline{G}(\lambda_i) = [G(\lambda_1), G(\lambda_2), G(\lambda_3), ...G(\lambda_{95})]$.

Typically, batch normalization is used to normalize hidden layer outputs [7]. However, batch normalization does not perform well when training models with lesser data [14]. Given our objective is to utilize minimal additional measurements for model training, and subsequent transfer learning; we utilize Self Normalizing Neural Networkss (SNNs) with Scaled Exponential Linear Unit (SELU) [15] activation function within layers to render the model as self normalizing. This choice enables us to effectively normalize the hidden layer outputs with a small amount of data, while maintaining the benefits of hidden layer normalization and preserving high accuracy. This step is the key enabling factor of our developed NN architecture to achieve effective one-shot training and transferability between models. The SELU activation function is given by:

$$\text{selu}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \le 0 \end{cases} \tag{2}$$

with $\alpha = 1.673$ and $\lambda = 1.050$.

B. Training Process

We use a two-step process to train this model, which includes unsupervised pre-training [16], [17] and supervised fine-tuning [18]. The training process has been selected with 2 key points:

1) Unsupervised pre-training utilizes unlabeled data points, which in this case is the input power spectrum of the measurements. These unlabeled measurements are easier to obtain and can also be simulated in cases of flat spectrum cases. This leads to much lower requirement of labeled measurements which are time consuming.
2) Unsupervised pre-training leads to a better initial weight initialization than random initialization, and captures more intricate dependency between parameters [17]. Additionally, neural networks with pre-training exhibit properties of a regularizer which leads to better generalization [19]. This is especially beneficial for transfer learning from one EDFA to other EDFAs.

Figure 3 shows the training process in detail. In the unsupervised pre-training step, we incrementally initialize the weights of each layer in the model in a greedy manner. First we take 512 unlabeled measurements for each target gain setting. Gaussian noise is added to the measurements. We utilize an auto-encoder layer, with the same number of neurons

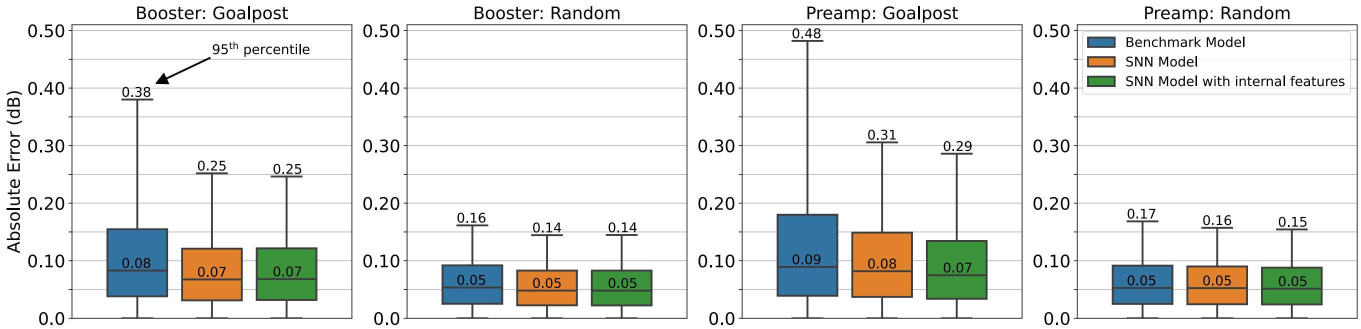

Fig. 4. Boxplot distribution of absolute errors across all 11 Booster and 11 Pre-amplifier EDFAs for goalpost and random channel loading. The boxes denote the inter-quartile range, and the whiskers denote the min/95th percentile

as the dimensions of the feature set, for reconstructing the input features. Starting incrementally from the bottom, the autoencoder layer takes the outputs from the noised inputs, and has the task to predict the denoised inputs. This is achieved by training the layer under test, and autoencoder layer with Mean Squared Error (MSE) to evaluate how good the model is at reconstructing the input even in the presence of noise. Each layer is trained in a greedy manner for 1,800 epochs with a Learning Rate (LR) of $1e-03$, along with MSE loss function. After the weights of one layer are initialized, its weights are frozen for training of subsequent layer. After all the layers are pre-trained in this manner, the weights are frozen and used as the base model for the next step.

Next is the supervised fine-tuning step, where we utilize 512 measurements to train and fine-tune the model. We use fully and randomly loaded measurements for this step. The model is trained using a modified MSE loss function, where the error for any k^{th} measurement is calculated as below:

$$\text{MSE}_k = \frac{1}{\sum_{i=1}^{95} c_i^k} \cdot \sum_{i=1}^{95} c_i^k \cdot \left[g_{\text{pred}}^k(\lambda_i) - g_{\text{meas}}^k(\lambda_i)\right]^2 \tag{3}$$

The model is fine-tuned using Adam Optimizer, with a LR of $1e-03$ over 1,200 epochs, and a gradient clipping threshold of 1.0 for stable training.

1) Usage of less data for training.
2) Better generalization for transfer learning.

C. Training and Test Sets

We compare the SS-NN model with a benchmark state-of-the-art method [7], [12]. For equivalent comparison, we follow the same dataset selection criteria. For each gain setting, we split the dataset into a training/test set ratio of 0.86/0.14. The test set contains 436 gain spectrum measurements per gain setting. This test set contains a mixture of random and goalpost channel loading measurements, which represent a diverse set of channel loading configurations. Note that although the SS-NN model uses less data for training, we allocate a larger portion of training data for the benchmark model, which uses 2,732 measurements per gain setting.

D. SNN Model Performance

We compare the SS-NN model with the benchmark model using the same set of features to highlight the benefits of our approach. Additionally, we demonstrate the advantage of incorporating internal EDFA features by comparing the SS-NN model with and without including these additional features.

Figure 4 shows the distribution of absolute errors of gain spectrum predicted by the benchmark model, SS-NN model using same set of features, and SS-NN model with additional internal VOA features. The errors are calculated across 11 boosters and pre-amplifier EDFAs in the Open Ireland and COSMOS testbeds on the test set with random and goalpost channel configurations. For boosters, the SS-NN model achieves a mean absolute error of 0.07 dB and 0.05 dB under the goalpost and random channel configurations. This is comparable to the performance of the benchmark model which uses a considerably higher number of measurements (8196 measurements), compared to a total of 1,792 measurements utilized by the SS-NN model. Importantly, the SS-NN models exhibit a superior error distribution, with a narrow inter-quartile range, and a 95^{th} percentile error of 0.25/0.14 dB, compared to 0.38/0.16 dB by the benchmark model, across the goalpost/random test sets.

For preamps, the SS-NN model achieves a mean absolute error of 0.08/0.05 dB using the same set of features, and 0.07/0.05 dB using additional internal features across goalpost/random channel configurations. This is marginally better than the benchmark model which achieves a 0.09/0.05 dB error across goalpost/random test sets. Additionally, the distribution of errors for SS-NN models are more stable, with a narrow inter-quartile range, and a 95^{th} percentile error within 0.3 dB across both channel configurations, showing that the SS-NN model generalizes well to unseen channel configurations even when trained with reduced measurements. It should be noted that using additional internal features when directly training EDFA models i.e., training on the source EDFA's measurements without TL does not provide additional performance. Note that since the resolution of the OCM readings is 0.1 dB per channel, the model is achieving the limit of ± 0.05 dB quantization error in some cases.

IV. Transfer Learning

Transfer Learning (TL) is a method to improve the learning in a task through the transfer of information from an existing but related domain. Specifically for ML algorithms, TL can

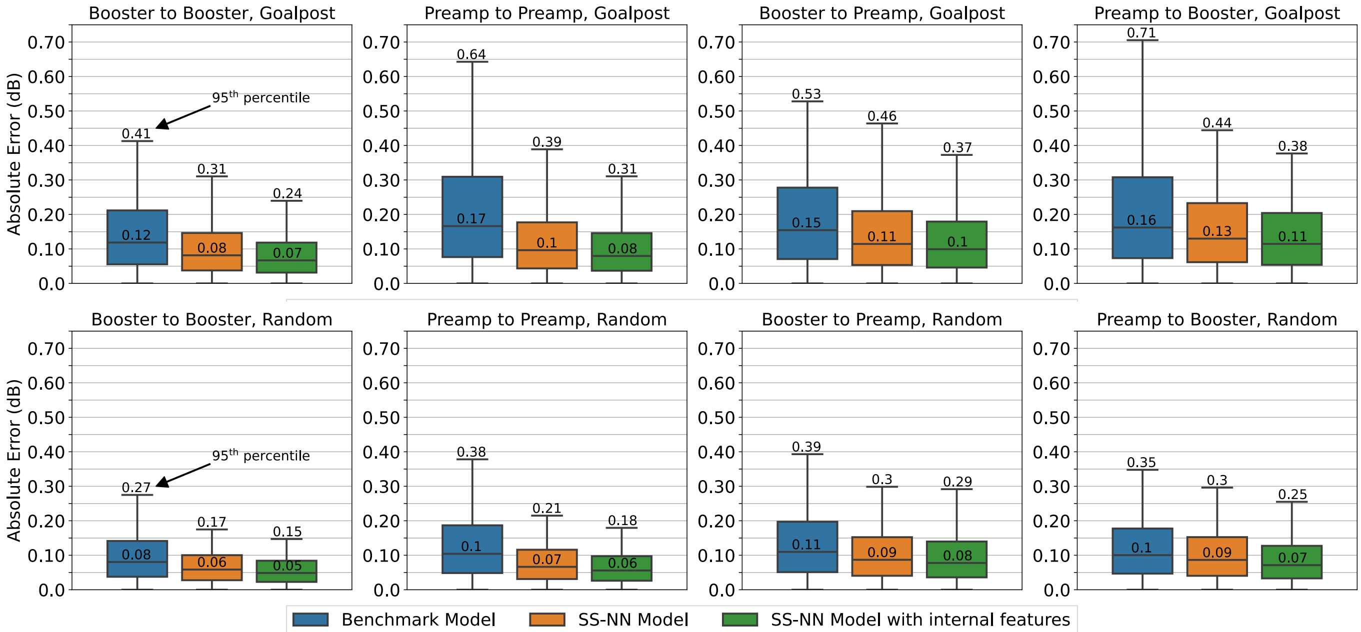

Fig. 5. Boxplot distribution of absolute errors across all 22 EDFAs for (a) Booster to Booster TL, (b) PreAmp to Preamp TL, (c) Booster to Preamp TL and (d) Preamp to Booster TL, for random and goalpost channel loading configurations. The boxes denote the inter-quartile range, and the whiskers denote the min/95th percentile

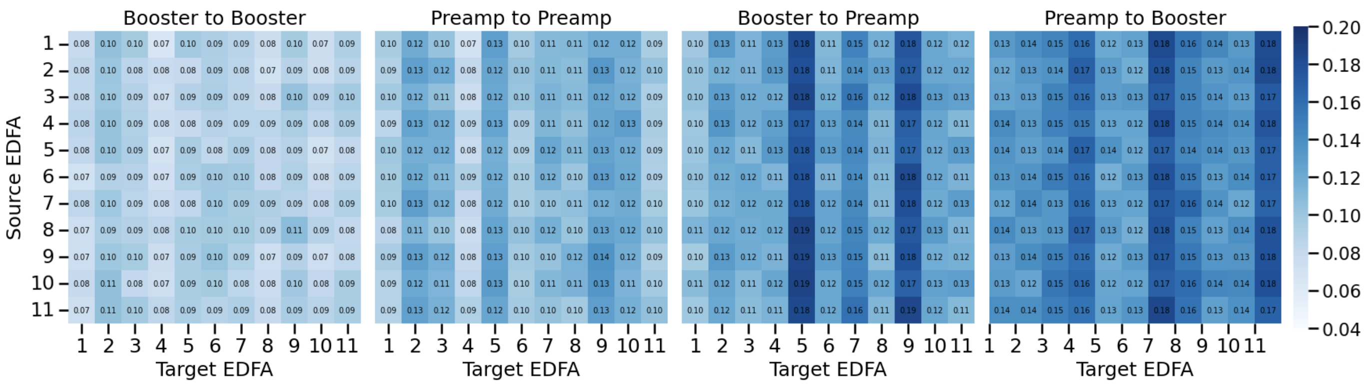

Fig. 6. Transfer Learning MAE matrix of SS-NN model with internal features on random loading. The (i, j) entry corresponds to the TL-based EDFA model, where the i^{th} and j^{th} EDFA serve as the source and target models, respectively. EDFA #1-8 are deployed in COSMOS, while EDFA #9-11 are deployed in Open Ireland.

be used to model a new domain, by transferring a model existing in a related domain [6]. TL is a viable and useful strategy for modeling the gain spectrum in EDFAs, reducing the measurement times. In this section, we show that TL for SS-NN models can be used to model the gain spectrum across different EDFAs with minimal additional data.

A. TL Training Process

To transfer an existing model from a source EDFA to a target EDFA, we re-train the source model using a single fully-loaded measurement for each target gain setting. This model is trained using the Adam Optimizer for 10,000 epochs using the same MSE loss function as Eq. (3) and a gradient clipping threshold of 1.0 for stable training. However, a differential Learning Rate (LR) is applied across layers instead of a flat LR. Specifically, the output layer has a larger LR of $1e-03$ compared to the subsequent hidden layers which have progressively decreasing LRs, with each layer's rate being 10% of the next layer's LR. In this way, the weights of the output layer are modified more aggressively, allowing it to capture the specific characteristics of the target EDFA more effectively. At the same time, the lower levels of the SS-NN model are fine-tuned more gradually to avoid overfitting, ensuring that the model can be generalized to new inputs.

B. Results

Figure 5 shows the boxplot of absolute errors of TL models for all possible source-target model pairs, for both same type transfer (Booster $\rightarrow$ Booster, and Preamp $\rightarrow$ Preamp), as well as cross type transfer (Booster $\rightarrow$ Preamp, Preamp $\rightarrow$ Booster); across random and goalpost channel loading configurations. As earlier, we show the comparison for the benchmark model, SS-NN model using same set of features and SS-NN model using additional internal features.

For the goalpost channel loading configuration, TL based SS-NN models achieve a MAE less than 0.10/0.13 dB for same-type/cross-type transfers. For the random configuration, the SS-NN models achieve a MAE less than 0.07/0.09 dB for same-type/cross-type transfers, respectively. Using additional internal features further improves the performance with a MAE less than 0.10/0.13 dB for same-type/cross-type transfers. The TL based SS-NN models also outperform the benchmark model, with a better error distribution.

The results show that SS-NN based TL models achieve comparable MAE with respect to a directly trained SS-NN model. However, the 95$^{\text{th}}$ percentile error of TL based models is higher than directly trained models. We believe using more measurements for TL, if available, will further improve the performance of TL models in high error configurations. It should also be noted that when directly training a SS-NN model on an EDFA, including additional internal VOA features does not provide much of a performance boost. However, these variables provide a large boost in performance in TL, indicating that these extra variables contain distinctive information about behavior of a particular EDFA, which improves the performance in TL.

Fig. 6 shows the MAE matrices (in dB) of SS-NN model incorporating internal features across 11 EDFAs under goalpost channel loading for both same-type and cross-type transfers. In each matrix entry, entry (i, i) corresponds to a directly trained model (without TL), and entry (i, j) corresponds to the transferred EDFA model where the i^{th} and j^{th} EDFA serve as the source and target models, respectively. The results show that the TL performance on each j^{th} EDFA is similar, irrespective of the source EDFA model used. The SS-NN based TL models achieve a consistent performance for each EDFA, close to its directly-trained counterpart even with different source models. Specifically, the SS-NN based TL models with internal features under goalpost channel loading achieve a per-EDFA MAE less than 0.14 dB for same type transfers, and MAE less than 0.19 dB for cross type transfers.

V. Conclusions

In this paper, we show a novel Semi-Supervised Self-Normalizing Neural Network (SS-NN) architecture to model the wavelength dependent gain of EDFAs. The SS-NN model uses a mix of labeled and unlabeled measurements to predict the gain spectrum in a diverse set of channel configurations and target gain settings with high accuracy. Furthermore, the SS-NN model can be transferred to EDFAs of different types using a single new measurement for each target gain setting with a comparable performance. This demonstrates that a single EDFA can be used to characterize multiple EDFAs using minimal measurements, significantly reducing the amount of data collection. We also find that internal EDFA features provide distinctive information about each EDFA's mechanism. Using these internal features provide enchanced performance in both same-type and cross-type EDFA transfers, showing potential for improvement by incorporating internal features. We aim to analyze other available internal EDFA features, as well as exploring the performance of Transfer Learning when transferring with In-Line Amplifiers (ILAs) and cross-vendor EDFAs.

References

[1] N. Afraz, F. Slyne, H. Gill, and M. Ruffini, "Evolution of access network sharing and its role in 5g networks," *Applied Sciences*, vol. 9, no. 21, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/21/4566

[2] C. Raack, R. Wessälly, D. Payne, and M. Ruffini, "Hierarchical versus flat optical metro/core networks: A systematic cost and migration study," in *2016 International Conference on Optical Network Design and Modeling (ONDM)*, 2016, pp. 1–6.

[3] S. Zhu, C. L. Gutterman, W. Mo, Y. Li, G. Zussman, and D. C. Kilper, "Machine Learning Based Prediction of Erbium-Doped Fiber WDM Line Amplifier Gain Spectra," in *2018 European Conference on Optical Communication (ECOC)*. Rome: IEEE, Sep. 2018, pp. 1–3.

[4] S. Zhu, C. Gutterman, A. D. Montiel, J. Yu, M. Ruffini, G. Zussman, and D. Kilper, "Hybrid Machine Learning EDFA Model," in *Optical Fiber Communication Conference (OFC) 2020*. San Diego, California: Optica Publishing Group, 2020, p. T4B.4.

[5] F. da Ros, U. C. de Moura, and M. P. Yankov, "Machine learning-based EDFA Gain Model Generalizable to Multiple Physical Devices," in *2020 European Conference on Optical Communications (ECOC)*, Dec. 2020, pp. 1–4.

[6] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[7] Z. Wang, D. Kilper, and T. Chen, "Transfer learning-based ROADM EDFA wavelength dependent gain prediction using minimized data collection," in *Optical Fiber Communication Conference (OFC) 2023*, ser. Optical Fiber Communication Conference (OFC) 2023. Optica Publishing Group, 2023, p. Th2A.1.

[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.

[9] CONNECT, "Openireland testbed, funded by science foundation ireland," 2022. [Online]. Available: www.openireland.eu

[10] D. Raychaudhuri, I. Seskar, G. Zussman, T. Korakis, D. Kilper, T. Chen, J. Kolodziejski, M. Sherman, Z. Kostic, X. Gu, H. Krishnaswamy, S. Maheshwari, P. Skrimponis, and C. Gutterman, "Challenge: COSMOS: A city-scale programmable testbed for experimentation with advanced wireless," in *Proc. ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2020.

[11] T. Chen, J. Yu, A. Minakhmetov, C. Gutterman, M. Sherman, S. Zhu, S. Santaniello, A. Biswas, I. Seskar, G. Zussman, and D. Kilper, "A Software-Defined Programmable Testbed for Beyond 5G Optical-Wireless Experimentation at City-Scale," *IEEE Network*, vol. 36, no. 2, pp. 90–99, Mar. 2022.

[12] Z. Wang, D. Kilper, and T. Chen, "An Open EDFA Gain Spectrum Dataset and Its Applications in Data-driven EDFA Gain Modeling," Apr. 2023.

[13] J. Zyskind and A. Srivastava, *Optically amplified WDM networks*, 1st ed. Burlington, MA: Elsevier/Academic Press, 2011.

[14] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[15] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," Sep. 2017.

[16] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006.

[17] J. Ge, S. Tang, J. Fan, and C. Jin, "On the Provable Advantage of Unsupervised Pretraining," Mar. 2023.

[18] N. J. Prottasha, A. A. Sami, M. Kowsher, S. A. Murad, A. K. Bairagi, M. Masud, and M. Baz, "Transfer learning for sentiment analysis using BERT based supervised fine-tuning," *Sensors*, vol. 22, no. 11, 2022.

[19] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 201–208.