Model-Centric versus Data-Centric Machine Learning for Soft-Failure Cause Identification in Optical Networks

Lareb Zar Khan[(1,*)], João Pedro[(2,3)], Nelson Costa[(2)], Antonio Napoli[(4)], Nicola Sambo[(1)]

[(1)] Scuola Superiore Sant'Anna, Pisa, Italy, [(2)] Infinera Unipessoal Lda, Carnaxide, Portugal, [(3)] Instituto de Telecomunicações IST, Lisboa, Portugal, [(4)] Infinera, Munich, Germany
([*]larebzar.khan@santannapisa.it)

Abstract *We compare model-centric and data-centric machine learning (ML) approaches to address the issue of insufficient training data for ML-based failure identification. The results suggest that a data-centric approach can improve classification accuracy by up to 7.1% on under-represented failures, albeit at a higher computational cost. ©2023 The Author(s)*

Introduction

Failure management has emerged as one of the key use-cases for machine learning (ML) applications in optical networks[1]–[3], which have been extensively investigated over the past few years. Traditional ML is mostly used in existing approaches, which usually revolves around model-centric techniques in which the "best" model for a given dataset is produced. However, in the applied ML research, there has been a recent shift in focus towards data-centric ML[4] which involves a systematic and algorithmic increase in the quantity and/or quality of the training dataset for a given model. Data augmentation, a data-centric approach, has been investigated in recent works[5],[6] to improve ML-based failure management in optical networks. Yet, none of them has provided a direct comparison to its counterpart model-centric approaches in terms of classification accuracy and computational cost. Therefore, it is not possible to say which approach is superior when applied to optical networks, particularly for failure management.

This paper aims to fill that research gap by providing a direct comparison between model-centric and data-centric approaches for dealing with the imbalanced training problem in ML-based soft-failure cause identification, which is a typical use-case within failure management. Imbalanced training is very common in failure cause identification because different failures occur with different frequencies. This results in an unequal number of observations for each failure within the training dataset and, thus, training on such datasets results in a comparatively poor performance on less frequent failure classes. To address this issue, in the model-centric approach, we modified the loss function of a neural network, and in the data-centric approach, using SMOTE-TOMEK technique, we generated synthetic data by utilizing the experimental data. To the best of our knowledge, these specific approaches have been investigated for the first time in the context of failure identification in optical networks. The results obtained on these approaches show that a data-centric approach may perform better in classification, though at the expense of higher computational cost.

Experimental Testbed Setup

For this study, we collected data from an experimental testbed, shown in Fig. 1. It consisted of a single link carrying a 100 Gb/s coherent signal traversing over four spans, each 80 km long. The power attenuation experienced along the fiber was compensated using a series of Erbium-doped fiber amplifiers (EDFAs). A bandwidth variable-wavelength selective switch (BV-WSS) was placed at the end of span-2 to emulate different soft-failures. Tab. 1 lists the failures that have been considered, including the assigned labels and the corresponding system configurations.

Tab. 1: Considered soft-failures

Label	Soft-Failure	Filter Bandwith (GHz)	Attenuation (dB)	Central Frequency (THz)
SF_0	Filter Tightening	26	0	192.3
SF_1	Attenuation	37.5	6	192.3
SF_2	Filter Tightening + Attenuation	26	6	192.3
SF_3	Filter Tightening + Filter Shift	26	0	192.32
SF_4	Filter Shift	37.5	0	192.32

We collected coherent receiver data, specifically the optical signal-to-noise ratio (OSNR) and bit error rate (BER). During normal operation, the central frequency (f_c) of the BV-WSS was 192.3 THz, while its (extra) attenuation and bandwidth

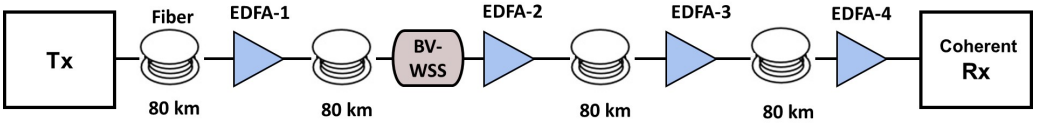

Fig. 1: Experimental testbed setup

were 0 dB and 37.5 GHz, respectively. Removing duplicate samples created an imbalanced distribution of failure classes (SF_0, SF_1, SF_2, SF_3, and SF_4) in the training dataset with 143, 98, 634, 164, and 208 samples, respectively, making it suitable for this study.

Model-Centric vs. Data-Centric Approaches

To address the problem of imbalanced failure classes, we investigated focal loss and SMOTE-TOMEK as reference model-centric and data-centric approaches, respectively.

A) Focal Loss

Focal loss (FL)[7] is a modified version of the cross-entropy (CE) loss which is a commonly used loss function in ML for classification problems like soft-failure cause identification. The CE loss for multi-class classification is given as

$$CE(p, y) = -\sum_{i=0}^{N-1} y_i \log(p_i), \tag{1}$$

where N is the number of classes i.e., 5 in this case, p_i is the predicted probability for class i and y_i is the ground truth label for class i (1 if the sample belongs to class i, and 0 otherwise). For the true (actual) class t, Eq. (1) can be expressed as

$$CE(p_t) = -\log(p_t), \tag{2}$$

where p_t is the predicted probability for the true class t. If we modulate CE loss in Eq. (2) using $(1-p_t)^\gamma$ term in which γ is a focusing parameter, then the resultant loss function is the FL function,

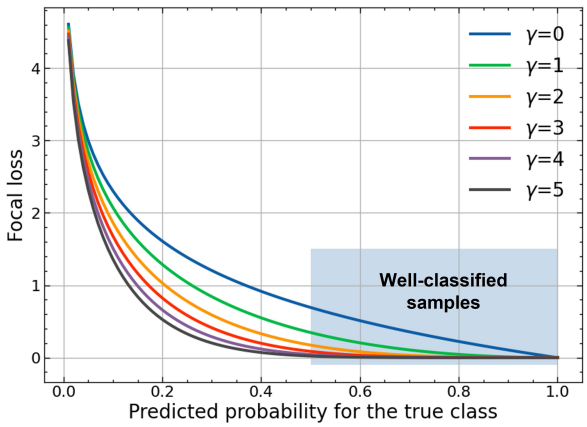

Fig. 2: Focal loss vs. predicted probability for the true class

given as

$$FL(p_t) = -(1-p_t)^\gamma \log(p_t). \tag{3}$$

Fig. 2 shows the FL as a function of predicted probability for different γ values, where γ = 0 corresponds to CE loss function. A sample is considered as well-classified if $p_t \geq 0.5$, and all such samples fall in the blue shaded region highlighted in Fig. 2. For imbalanced training datasets, most of the samples belong to the majority classes. As there are enough samples from these classes, the ML model can learn the underlying patterns well and, therefore, classify these failures accurately. But, for the minority classes, there are comparatively fewer samples, which makes it difficult for the ML model to learn their underlying patterns. This is because the overall training loss is dominated by the samples from the majority classes, which are typically well-classified. As a result, the ML classifier does not perform as well on minority classes as it does on majority classes.

The modulating factor (controlled by γ) down-weights the contribution of well-classified samples and focuses more on hard-to-classify samples which are usually from minority classes. For a given dataset, a suitable value of γ can be tuned and in our case, γ = 2 has been used.

B) SMOTE-TOMEK

SMOTE-TOMEK[8] is a combination of two different approaches i.e., synthetic minority oversampling technique (SMOTE)[9] and TOMEK-links[10]. SMOTE generates synthetic samples of minority classes by interpolation between a randomly selected sample from the minority class and its randomly selected nearest neighbor from the same class. To reduce the number of misclassifications, SMOTE is followed by the identification and removal of TOMEK links (i.e., pairs of samples that are closest to each other but belong to different classes). Hence, it is a hybrid approach that combines over- and under-sampling, and the process is repeated until the desired proportion of minority class samples is obtained.

Results and Discussion

A neural network (NN) was chosen as the ML classifier for this investigation. It consisted of five

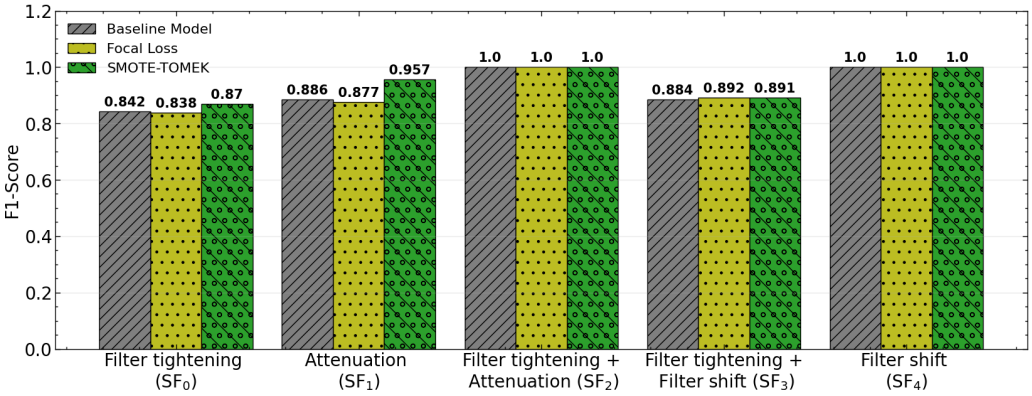

Fig. 3: Results on test dataset for each failure class in terms of F1-Score

layers in total: an input layer with 2 neurons, three hidden layers with 183, 186, and 63 neurons, and an output layer with 5 neurons. *tanh* was used as an activation function in the hidden layers, and *softmax* was used in the output layer. The dropout rate was 0.05456, the learning rate was 0.002574, and the batch size was 8. All these hyperparameters were tuned using Bayesian Optimization. This NN was used as a baseline against which the performance of model-centric and data-centric approaches was compared.

Fig. 3 shows the results on test dataset for each soft-failure in terms of F1-score, which is one of the appropriate evaluation metrics in the case of imbalanced datasets. In our training dataset, SF_0, SF_1, and SF_3 have BER and OSNR values in the same range, making them inseparable, resulting in a comparatively poor performance on these three failure classes. Moreover, these three failures have fewer samples than SF_2 and SF_4 .

As it is clear from Fig. 3, a significant performance improvement on these minority classes has been achieved using SMOTE-TOMEK. F1-score improved from 0.842 to 0.87 for SF_0 (2.8% improvement), from 0.886 to 0.957 for SF_1 (7.1% improvement), and from 0.884 to 0.891 for SF_3 (0.7% improvement) with no negative impact on majority classes (SF_2 and SF_4). In contrast, with focal loss, NN failed to improve overall performance on minority failure classes, suggesting that a data-centric approach may outperform model-centric approaches in this scenario. However, this only describes one aspect of the performance. In order to have a complete comparison, we considered performance in terms of training and computational time as well. Tab. 2 shows the average training time over 100 training iterations (not epochs) as well as computational time on a Intel(R) Core(TM) i7-12700H @ 2.30 GHz with NVIDIA GeForce RTX 3060 Laptop GPU. It should be noted that computational time is only relevant to the SMOTE-TOMEK where we modified the training datasets. The impact of modifying the loss function reflects already in the training time, and additional computation is not required.

Tab. 2: Computational cost comparison

	Computational Time (s)	Average Training Time (s)	Total Time (s)
Baseline Model	N/A	19.62	19.62
Focal Loss	N/A	21.71	21.71
SMOTE-Tomek	391 x 10^{-3}	31.82	31.82391

The average training time for baseline NN was 19.62 seconds, which increased slightly with the modification of loss function (i.e., focal loss). However, with the SMOTE-TOMEK approach, we increased both the quality (as indicated by the improved F1-scores) and the quantity (2.53-fold increase in this case) of data by adding synthetic samples. Due to this increase in training dataset size, the longer time was expected as now NN has to deal with many more samples during each training epoch. SMOTE-TOMEK's computational time was around 391 x 10^{-3} seconds, which is insignificant as compared to its training time.

Conclusions

We investigated the potential of data-centric ML against model-centric ML for addressing the issue of insufficient data for some failures within the training dataset for ML-based soft-failure cause identification. The obtained results indicate that a data-centric approach can significantly improve classification accuracy on under-represented failure classes with an improvement of up to 7.1% being observed. However, this improvement is accomplished at the expense of longer computational and training times. Based on the scenario and acceptable trade-offs, the most suitable approach can be chosen.

Acknowledgements

This work has been funded by EU H2020 Marie Skłodowska-Curie Actions ITN project MENTOR (GA 956713) and the Horizon Europe SEASON project (GA 101096120).

References

[1] F. Musumeci, C. Rottondi, G. Corani, S. Shahkarami, F. Cugini, and M. Tornatore, "A tutorial on machine learning for failure management in optical networks", *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4125–4139, 2019. DOI: 10.1109/JLT.2019.2922586.

[2] F. N. Khan, Q. Fan, C. Lu, and A. P. T. Lau, "An optical communication's perspective on machine learning and its applications", *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 493–516, 2019. DOI: 10.1109/JLT.2019.2897313.

[3] L. Z. Khan, A. Triki, M. Laye, and N. Sambo, "Optical network alarms classification using unsupervised machine learning", in *2022 27th OptoElectronics and Communications Conference (OECC) and 2022 International Conference on Photonics in Switching and Computing (PSC)*, 2022, pp. 1–3. DOI: 10.23919/OECC/PSC53152.2022.9849872.

[4] D. Zha, Z. P. Bhat, K.-H. Lai, *et al.*, *Data-centric artificial intelligence: A survey*, 2023. arXiv: 2303.10158 [cs.LG].

[5] L. Z. Khan, J. Pedro, N. Costa, L. De Marinis, A. Napoli, and N. Sambo, "Data augmentation to improve performance of neural networks for failure management in optical networks", *Journal of Optical Communications and Networking*, vol. 15, no. 1, pp. 57–67, 2023. DOI: 10.1364/JOCN.472605.

[6] C. Xing, C. Zhang, B. Ye, *et al.*, "Failure data augmentation for optical network equipment using time-series generative adversarial networks", in *Optical Fiber Communication Conference (OFC) 2023*, Optica Publishing Group, 2023, M3G.4. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=OFC-2023-M3G.4.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, 2018. arXiv: 1708.02002 [cs.CV].

[8] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and smote approaches for machine fault classification with an imbalanced dataset", *Sensors*, vol. 22, no. 9, 2022, ISSN: 1424-8220. DOI: 10.3390/s22093246. [Online]. Available: https://www.mdpi.com/1424-8220/22/9/3246.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002. DOI: 10.1613/jair.953. [Online]. Available: https://doi.org/10.1613%2Fjair.953.

[10] I. Tomek, "Two modifications of cnn", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, 1976. DOI: 10.1109/TSMC.1976.4309452.