An Optimization Method for Coupled-Line Bandpass Filters Using Transformer-Based Estimator and Multilayer Perceptron

Kai-Da Xu, *Senior Member, IEEE*, Jing Tian, Yijun Cai, Daotong Li, *Senior Member, IEEE*, Wen Wu, and Qiang Chen, *Senior Member, IEEE*

***Abstract*—A novel method for estimation and optimizing structural parameters of coupled-line bandpass filters (BPFs) using transformer-based estimator (TBE) and multilayer perceptron (MLP) is proposed. Once trained, the TBE can quickly obtain the predicted values of the BPF structural parameters from desired S-parameters, while the trained MLP can replace the time-consuming electromagnetic (EM) simulation process, and establish the mapping from structural parameters to S-parameters. In order to obtain the optimal structural parameters, the trained MLP is combined with genetic algorithm (GA) for fast optimization. To demonstrate the effectiveness of the proposed method, a BPF using three pairs of coupled-line microstrip structure is designed and fabricated. The experimental results demonstrate that the proposed method can quickly and accurately obtain the optimal structural parameters from the desired frequency response of the coupled-line BPF.**

***Index Terms*—Bandpass filters, genetic algorithm, multilayer perceptron, structural parameters, transformer.**

I. Introduction

BANDPASS filters (BPFs) with excellent frequency response characteristics are extensively used in modern radio frequency and wireless communication systems. Among various types of BPFs, the coupled-line microstrip structure is a good candidate due to its compact size and ease of fabrication [1]. In the traditional approach, the design of coupled-line BPF heavily relies on deriving impedance equations and simulating circuit models using electromagnetic (EM) simulation software. However, this method is highly dependent on the designers' expertise, and optimizing the structural parameters is time-consuming and energy-intensive, even with the aid of optimization algorithms available in EM simulation software.

In recent years, deep learning has emerged a powerful alternative to traditional EM modeling in the design of microwave components. This is primarily due to its exceptional data processing capabilities. Consequently, there has been significant research on investigating the applications of neural networks (NNs) in microwave components [2], [3], [4], [5]. For instance, a modified NN learning process is utilized for the microwave circuit design [6], instead of employing conventional solution-searching optimization. Moreover, to enhance optimization speed and accuracy, a combination of NN and filter coupling matrix has been investigated for filter design [7], [8]. In [9] and [10], an NN inverse modeling technique is proposed in order to solve the non-uniqueness issue of input-output relationship. Additionally, the work presented in [11] introduces a novel adjoint NN technique based on sensitivity analysis to generate accurate parametric models with minimal training data. A unique eigenmode-based NN approach is suggested in [12] as an alternative to traditional structural adjustment and optimization design process. Furthermore, a deep NN model capable of solving higher-dimensional microwave modeling problems is proposed in [13]. The relationship between coupling matrix and structural parameters can be expressed when utilizing a deep Q-network [14].

The above-mentioned NN models for the design of BPFs mostly use multilayer NN architectures to establish the mapping between the BPF's coupling matrix and structural parameters. However, extracting the coupling matrix during the filter design process is relatively complicated, while obtaining S-parameters from ideal circuit of the coupled-line BPF is simple and straightforward. Therefore, it is valuable to investigate the relationship between S-parameters and structural parameters of BPFs using NN models, enabling fast and accurate optimization to realize the final filter design.

In this brief, a new method using two deep NNs simultaneously is proposed for structural parameters optimization and automatic adjustment of BPFs. Compared to the 3-layer NN models in [7] and [9], our proposed deep NNs possess higher accuracy when processing high-dimensional data. The first NN, i.e., transformer-based estimator (TBE), is utilized to establish a mapping from S-parameters to structural parameters, which facilitates estimating the predicted values of the structural parameters on the frequency response of ideal

Manuscript received 24 January 2024; revised 25 February 2024; accepted 18 March 2024. Date of publication 21 March 2024; date of current version 4 September 2024. This work was supported in part by NSAF Joint Fund under Grant U2130102. This brief was recommended by Associate Editor H. Jiang. *(Corresponding authors: Yijun Cai; Kai-Da Xu.)*

Kai-Da Xu and Jing Tian are with the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Department of Communications Engineering, Tohoku University, Sendai 980-8579, Japan (e-mail: kaidaxu@ieee.org).

Yijun Cai is with the Fujian Provincial Key Laboratory of Optoelectronic Technology and Devices, Xiamen University of Technology, Xiamen 361024, China (e-mail: yijuncai@foxmail.com).

Daotong Li is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China.

Wen Wu and Qiang Chen are with the Department of Communications Engineering, Tohoku University, Sendai 980-8579, Japan.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2024.3380412.

Digital Object Identifier 10.1109/TCSII.2024.3380412

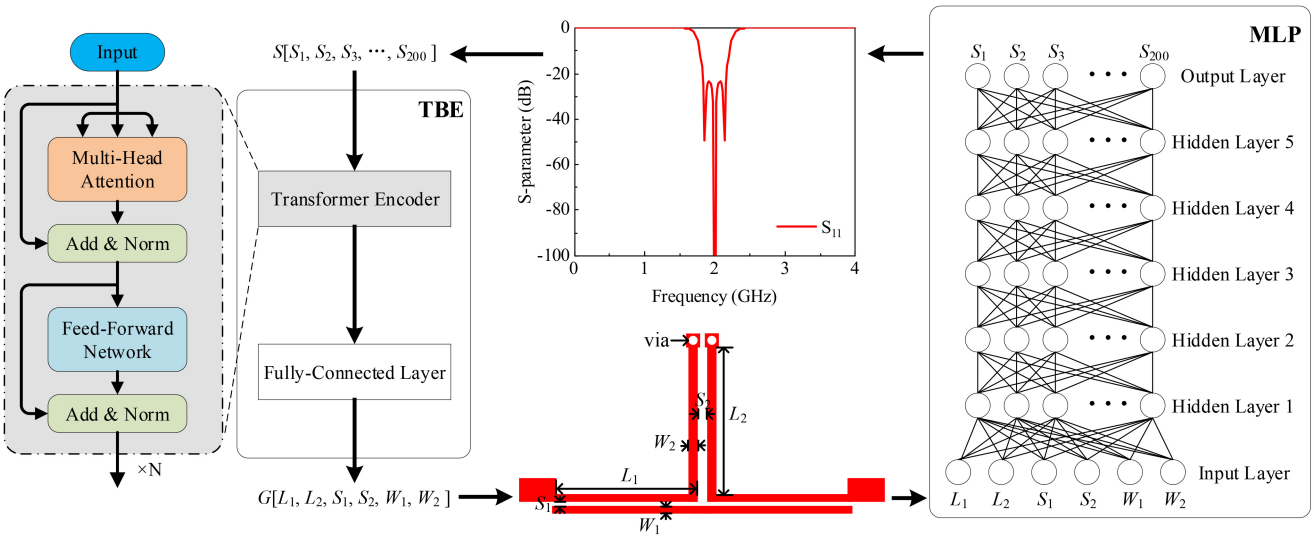

Fig. 1. Schematic of TBE and MLP architecture for coupled-line BPF design.

circuit. The second NN, i.e., multilayer perceptron (MLP), is then employed to establish a mapping from structural parameters to S-parameters. By integrating the MLP with the genetic algorithm (GA), the optimal structural parameters can be obtained. To demonstrate the effectiveness and accuracy of the proposed idea, a coupled-line BPF is designed and fabricated, whose simulation and measurement are in good agreement.

II. BPF Structural Parameter Optimization

The proposed TBE and MLP architecture is shown in Fig. 1, where the TBE consists of transformer encoder and fully-connected layer. The transformer encoder contains a stack of N identical layers, and each layer has two sub-layers, where the first sub-layer is a multi-head attention and the second one is a simple fully-connected feed-forward network [15]. The multi-head attention includes several attention layers running in parallel, and the output matrix of an attention layer is as follows [15],

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are the matrices of query, keys, and values, respectively. They are obtained by applying a linear transformation to the input of self-attention. The d_k is the dimension of matrices Q and K. We calculate the dot products of the matrix Q with matrix K^T, divide each entry of matrix by $\sqrt{d_k}$, and apply a softmax function to obtain the weight values of all matrix entries, where the softmax function serves to compress the values in each row of a matrix into the interval [0, 1], guaranteeing that the sum of these values in any given row equals 1. This normalization process transforms raw scores into probability distributions, where each matrix entry represents the likelihood or confidence associated with a specific entry in the row. Then, these weights are multiplied with matrix V to obtain the attention coefficient. Finally, the corresponding multi-head attention output matrix can be described as,

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{2a}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_i, \ldots, \text{head}_M)W^L \tag{2b}$$

where W_i^Q, W_i^K, and W_i^V are weight matrices for matrices Q, K, and V, respectively. There are M heads for multi-head attention, where the head$_i$ represents the ith head attention. The data from input is passed through computations involving M attention layers, leading to the generation of M output matrices. These matrices are then concatenated together and passed through a linear layer to generate the final output, where the W^L is weight matrix for linear layer. The feed-forward network is composed of two fully-connected layers, where the first layer utilizes the Gaussian error linear units (GELU) [16] as an activation function and the second layer does not have the activation function. The input data passes through the transformer encoder without any change in its dimension, resulting in identical dimension to the output data. By adding a fully-connected layer, we can reduce the dimension of output data generated by the transformer encoder. Through the function of the above layers, the TBE finally realizes the mapping from S-parameters to structural parameters.

The MLP is composed of seven fully-connected layers, including an input layer, an output layer, and five hidden layers. The activation function used for each hidden layer is rectified linear unit (ReLU). The relationship between the input and output of the MLP can be expressed by the following equation,

$$\begin{aligned} y_k = \sum_{i_5=1}^{N_5} w_{i_5}^{N_5} \max\Bigg(0, \sum_{i_4=1}^{N_4} w_{i_4}^4 \max\Bigg(0, \sum_{i_3=1}^{N_3} w_{i_3}^3 \max \\ \Bigg(0, \sum_{i_2=1}^{N_2} w_{i_2}^2 \max\Bigg(0, \sum_{i_1=1}^{N_1} w_{i_1}^1 \max\Bigg(0, \sum_{i_0=1}^{N_0} w_{i_0}^0 x_i + b_{i_0}^0\Bigg) \\ + b_{i_1}^1\Bigg) + b_{i_2}^2\Bigg) + b_{i_3}^3\Bigg) + b_{i_4}^4\Bigg) + b_{i_5}^5 \end{aligned} \tag{3}$$

where y_k is the output of the kth (k=1, 2, …, N_6) neuron in the output layer, x_i is the input of the ith (i=1, 2, …, N_0) neuron in the input layer. The N_0, N_1, N_2, N_3, N_4 and N_5 are

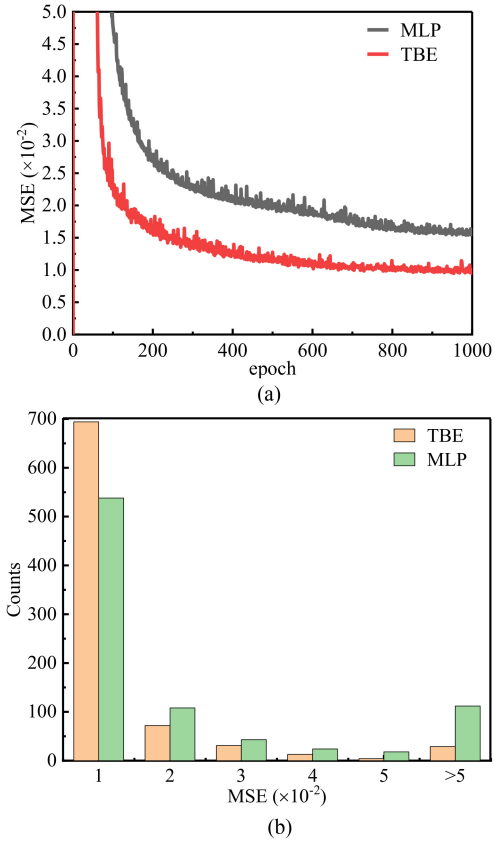

Fig. 2. (a) MSE curves of TBE and MLP. (b) MSE distributions of testing samples.

the number of neurons in the input layer and the five hidden layers, respectively, and N_6 is the number of neurons in the output layer. $w_{i_0}^0$, $w_{i_1}^1$, $w_{i_2}^2$, $w_{i_3}^3$, $w_{i_4}^4$, $w_{i_5}^5$ and $b_{i_0}^0$, $b_{i_1}^1$, $b_{i_2}^2$, $b_{i_3}^3$, $b_{i_4}^4$, $b_{i_5}^5$ are the weights and bias of each layer, respectively. The MLP can establish the mapping from structural parameters to S-parameters, which is utilized to replace the complicated EM simulation process.

Firstly, the performance of TBE is evaluated and compared with that of classical MLP in predicting structural parameters. The mean square error (MSE) loss function is utilized to calculate the error (loss values) between the predicted results and the actual values. This loss function can be represented by the following equation,

$$\mathrm{MSE}(\boldsymbol{y}', \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}'_i - \boldsymbol{y}_i)^2 \tag{4}$$

where n is the number of samples, $\boldsymbol{y}'$ is the predicted results by the model, and $\boldsymbol{y}$ is the desired values. As shown in Fig. 2(a), the MSE using TBE (N = 3 and M = 8) is 0.97×10^{-2}, much smaller than that using classical MLP with seven layers (numbers of neurons from the input layer to the output layer: $N_0 = 200$, $N_1 = 800$, $N_2 = 1200$, $N_3 = 800$, $N_4 = 400$, $N_5 = 200$, and $N_6 = 6$) in the testing dataset. The histogram of MSE count distributions in Fig. 2(b) illustrates that the TBE supports more testing samples with a lower MSE than the MLP. This implies the higher prediction accuracy in predicting structural parameters can be obtained using TBE.

Subsequently, the proposed TBE and MLP are utilized to estimate and optimize the structural parameters of a coupled-line BPF. The procedure for estimating and optimizing is as follows.

1) The EM simulation software is used for coupled-line BPF modeling (see layout in Fig. 1). By adjusting the structure of the BPF within a certain range, the corresponding S-parameters are obtained through EM simulation. Each group of structural parameters and corresponding S-parameters is taken as a group of data to form a dataset.

2) The TBE (i.e., left side of Fig. 1) and MLP (i.e., right side of Fig. 1) are trained and validated with identical training dataset and testing dataset. The distinction of the trained data between TBE and MLP lies in their different input and output. The TBE takes S-parameters ($S[S_1, S_2, S_3, \cdots, S_{200}]$) as the input and then produces structural parameters ($G[L_1, L_2, S_1, S_2, W_1, W_2]$), where these S-parameters are obtained from the magnitude values (unit: dB) of S_{11} parameters across 200 frequency points, sampled within the frequency range from 0.02 to 4 GHz with increments of 0.02 GHz. In contrast, the MLP takes structural parameters as input and produces S-parameters. During the training and testing process, the MSE loss function is utilized to calculate the error between the predicted results and the real values. Then, the Adam optimizer [17] is applied to update the weights and bias of TBE and MLP in the training process.

3) Before commencing the training process of the NN models, it is vital to set the essential parameters such as batch size, learning rate, and the number of training iterations. Subsequently, the TBE and MLP are trained until the loss values reach the designated threshold on the training dataset or until the specified number of training iterations is attained, thereby saving the models. The trained models can then be utilized to make predictions on structural parameters and S-parameters accordingly.

4) By inputting the desired S-parameters into the trained TBE, one group of structural parameters is obtained from the output. These predicted structural parameters are then fed into the trained MLP to predict the S-parameters. The objective function F of GA is determined by calculating the mean absolute error (MAE) between the predicted S-parameters and the desired S-parameters. The function F can be expressed as follows,

$$\mathrm{F}(\boldsymbol{s}, \boldsymbol{s}') = \frac{1}{m}\sum_{i}^{m}|s_i - s'_i| \tag{5}$$

where the multidimensional vector $\boldsymbol{s}$ and $\boldsymbol{s}'$ are the desired S-parameters and the predicted S-parameters, respectively, and the m is the element number of each vector.

5) The predicted values of the structural parameters are obtained from the TBE, and they are utilized as the starting points. These values are disturbed to generate population samples. Then, the fitness value of individual in the population is calculated based on the objective function F of GA. The individual with the smallest fitness value is selected for crossover, mutation, and recombination, resulting in the generation of new individual in the population. The new individual will inherit the excellent features but with better performance than the previous generation. In each iteration of

TABLE I
RANGES FOR THE STRUCTURAL PARAMETERS

Para.	L_1	L_2	S_1	S_2	W_1	W_2
Range (mm)	24.9-28.9	24.6-28.6	0.28-0.68	1.42-1.82	0.1-0.46	0.18-0.58

TABLE II
NUMBER OF NEURONS IN EACH LAYER OF THE MLP MODEL

Layer	Input Layer	Hidden Layer 1	Hidden Layer 2	Hidden Layer 3
No.	6	200	400	800
Layer	Hidden Layer 4	Hidden Layer 5	Output Layer	
No.	1200	800	200	

the GA, the trained MLP is invoked to obtain the predicted values of S-parameters until the objective function F reaches the set threshold, and finally the obtained structural parameters will be the optimal structural parameters.

III. EXAMPLE OF BPF DESIGN

To validate the proposed method, a BPF using three pairs of coupled-lines is designed and the specifications are as follows: Center frequency $f_0 = 2$ GHz, bandwidth of $1.73 \sim 2.27$ GHz, and return loss (RL) of over 20 dB within passband. The experimental environment is AMD Ryzen 7 3700X 8-Core Processor at 3.59 GHz, 16 GB memory, Windows 10 operating system, PyCharm for running the NN models and GA, and HFSS for EM simulation.

The layout of the BPF is shown in Fig. 1, where the diameter of two vias is both 0.6 mm. Due to the structure symmetry, the two horizontal pairs of coupled-lines have the identical length, width and gap. There are several structural parameters that need to be estimated and optimized: the lengths of coupled-lines L_1 and L_2, the gaps of coupled-lines S_1 and S_2, and the widths of coupled-lines W_1 and W_2. The values of the coupled-line BPF structural parameters are adjusted using the EM simulation software to obtain corresponding S-parameters, i.e., the Step 1) of the optimization procedure in Section II. The ranges for all structural parameters are indicated in Table I.

The number of neurons in each layer of the MLP model is shown in Table II, where the dimensions of the input and output data determine the numbers of neurons in input and output layers, while the numbers of neurons in hidden layers are obtained from experiments. According to the Steps 2) and 3) of the optimization procedure, the batch size is set to 32 and the learning rate is set to 0.00001. The TBE and MLP are trained on the training dataset for 1000 and 1500 times, respectively. During the training process, the operating frequency and bandwidth of BPF are not the features captured by NN, so the training process is still valid if the operating frequency or bandwidth is changed. To demonstrate the predictive accuracy of the TBE and MLP, the models are validated using the test dataset. The average error value of the TBE is 0.97%, and the average error value of the MLP is 3.37%.

According to the Step 4) of the optimization procedure, the desired S-parameters (obtained from the ideal circuit of the coupled-line BPF using ADS software) are placed into the TBE to obtain the predicted structural parameters, i.e., predicted values shown in Table III. They are then applied to the coupled-line BPF, and the S-parameters obtained through EM simulations (i.e., predicted S-parameters) are compared with the desired S-parameters, as shown in Fig. 3. It is observed that the S-parameter curve obtained by EM simulations has a certain frequency offset with the desired S-parameter curve, and one transmission pole is missing within the passband. Therefore, further optimization through MLP using GA is needed to improve the performance of the coupled-line BPF.

TABLE III
PREDICTED AND OPTIMAL VALUES OF THE STRUCTURAL PARAMETERS

Para.	L_1	L_2	S_1	S_2	W_1	W_2
Predicted values	26.9	26.6	0.6	1.52	0.24	0.45
Optimal values	26	26.2	0.56	1.37	0.22	0.29

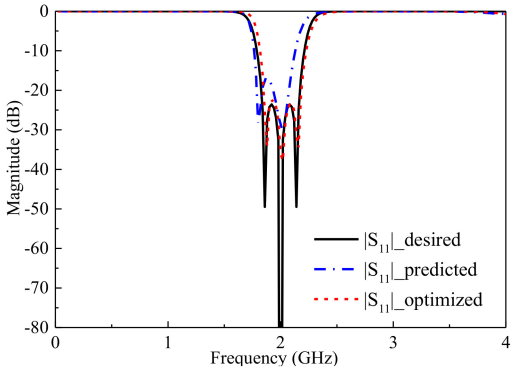

Fig. 3. Comparisons among the desired S-parameter, predicted S-parameter and optimized S-parameter.

According to the Step 5) of the optimization procedure, the initial sample number of GA is set as 200, which is obtained by changing $\pm 10\%$ on the basis of the predicted values from TBE. After 100 iterations of GA, the optimized structural parameters are obtained, i.e., optimal values shown in Table III. Subsequently, these optimized structural parameters are applied to the physical coupled-line BPF, and the corresponding S-parameter results are shown in Fig. 3. As can be seen, the optimized S-parameter curve is in good agreement with the desired counterpart. Furthermore, the proposed method can be also employed for the optimization of the BPF with higher order.

To prove that the desired S-parameter curve in Fig. 3 is not a special case for the proposed MLP using GA, 1000 new samples with different data are extracted by adding noise (i.e., perturbation) to the desired S-parameter curve and then placed into the traditional MLP and the MLP using GA, respectively, for comparison. The corresponding MAEs are obtained, 3.73% for the traditional MLP and 1.78% for the MLP using GA. Obviously, the MAE of the MLP using GA is much lower than that of the traditional MLP, indicating that our proposed method has higher prediction accuracy.

The proposed method for training the NN and optimizing structural parameters takes only 3 hours, much less time than both of the HFSS optimization method and the method in [9], which take 43 hours and 6 hours, respectively. It indicates that

TABLE IV
ACCURACY COMPARISONS WITH SOME OTHER NN MODELS

	Filter structure	No. of input parameters	Filter order	NN Architecture	Training error (%)	Testing error (%)
[7]	Waveguide	4	4	Feedforward	1.15	1.16
[10]	Microstrip	51	8	Feedforward	1.94	2
[13]	Multicoupled Cavity	35	4	Feedforward	1.22	1.88
This work	Coupled-line	200	3	TBE+MLP	0.76	0.97

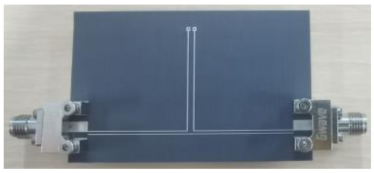

Fig. 4. Photograph of the fabricated BPF.

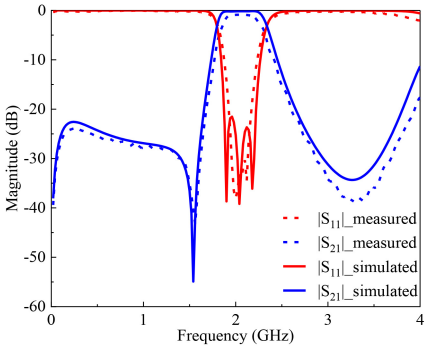

Fig. 5. Simulated and measured results of the coupled-line BPF using optimization method.

the proposed method possesses higher efficiency in optimizing the structural parameters of the BPF. Furthermore, as compared to other reported works, the proposed NN models exhibit lower training and testing errors, as tabulated in Table IV.

For demonstration, a filter example is also fabricated on a substrate with a relative permittivity of $\varepsilon_r = 2.65$ and a thickness of $h = 1$ mm, whose photograph is displayed in Fig. 4. Fig. 5 presents a performance comparison of frequency responses between the EM simulation by HFSS software and the measurement by vector network analyzer. As can be seen, the measured bandwidth is $1.86 \sim 2.28$ GHz with center frequency at 2.07 GHz. Within the passband, the RL is greater than 25 dB. The measured results satisfy the design specifications, with a slight deviation that may be due to the fabrication tolerance.

IV. CONCLUSION

This brief proposes a method for estimating and optimizing the structural parameters of coupled-line BPFs. The trained TBE can quickly predict the values of structural parameters from the S-parameters of coupled-line BPF ideal circuit. This is because the TBE captures the intricate relationship among the parameters of coupled-line BPFs, which are often nonlinear and complex. For the trained MLP, it can achieve fast prediction of the S-parameters from the structural parameters. Based on the predicted values of the structural parameters using the TBE, the GA is utilized to invoke the MLP for optimizing these structural parameters, ultimately obtaining the optimal ones. To verify the proposed method, a BPF using three pairs of coupled-lines is designed and its structural parameters are successfully estimated and optimized using the proposed TBE and MLP. Finally, the optimized structural parameters are applied to the coupled-line BPF, and the corresponding filter prototype is fabricated, whose measured results are in good agreement with the design specifications.

REFERENCES

[1] K. D. Xu, D. Li, and Y. Liu, "High-selectivity wideband bandpass filter using simple coupled lines with multiple transmission poles and zeros," *IEEE Microw. Wireless Compon. Lett.*, vol. 29, no. 2, pp. 107–109, Feb. 2019.
[2] Q.-J. Zhang, K. C. Gupta, and V. K. Devabhaktuni, "Artificial neural network for RF and microwave design—From theory to practice," *IEEE Trans. Microw. Theory Techn.*, vol. 51, no. 4, pp. 1339–1350, Apr. 2003.
[3] J. E. Rayas-Sánchez, "EM-based optimization of microwave circuits using artificial neural networks: The state-of-the-art," *IEEE Trans. Microw. Theory Techn.*, vol. 52, no. 1, pp. 420–435, Jan. 2004.
[4] W. Liu, W. Na, L. Zhu, and Q.-J. Zhang, "A review of neural network based techniques for nonlinear microwave device modeling," in *Proc. IEEE MTT-S Int. Conf. Numer. Electrom. Multiphys. Model. Optim. (NEMO)*., Beijing, China, 2016, pp. 1–2.
[5] J. Jin, F. Feng, and Q.-J. Zhang, "An overview of neural network techniques for microwave inverse modeling," in *Proc. IEEE Int. Symp. Radio Freq. Integr. Techn.*, Hualien, Taiwan, 2021, pp. 1–2.
[6] M. M. Vai, S. Wu, B. Li, and S. Prasad, "Reverse modeling of microwave circuits with bidirectional neural network models," *IEEE Trans. Microw. Theory Techn.*, vol. 46, no. 10, pp. 1492–1494, Oct. 1998.
[7] Y. Wang, M. Yu, H. Kabir, and Q. J. Zhang, "Effective design of cross-coupled filter using neural networks and coupling matrix," in *Proc. IEEE MTT-S Int. Microw. Symp. Dig.*, San Francisco, CA, USA, 2006, pp. 1431–1434.
[8] J.-J. Sun, X. Yu, and S. Sun, "Coupling matrix extraction for microwave filter design using neural networks," in *Proc. IEEE Int. Conf. Comput. Elect.*, Chengdu, China, 2018, pp. 1–2.
[9] H. Kabir, Y. Wang, M. Yu, and Q.-J. Zhang, "Neural network inverse modeling and applications to microwave filter design," *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 4, pp. 867–879, Apr. 2008.
[10] C. Zhang, J. Jin, W. Na, Q.-J. Zhang, and M. Yu, "Multivalued neural network inverse modeling and applications to microwave filters," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 8, pp. 3781–3797, Aug. 2018.
[11] S. A. Sadrossadat, Y. Cao, and Q.-J. Zhang, "Parametric modeling of microwave passive components using sensitivity-analysis-based adjoint neural-network technique," *IEEE Trans. Microw. Theory Techn.*, vol. 61, no. 5, pp. 1733–1747, May 2013.
[12] M. Ohira, A. Yamashita, Z. Ma, and X. Wang, "A novel eigenmode-based neural network for fully automated microstrip bandpass filter design," in *Proc. IEEE MTT-S Int. Microw. Symp.*, Honolulu, HI, USA, 2017, pp. 1628–1631.
[13] J. Jin, C. Zhang, F. Feng, W. Na, J. Ma, and Q.-J. Zhang, "Deep neural network technique for high-dimensional microwave modeling and applications to parameter extraction of microwave filters," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 10, pp. 4140–4155, Oct. 2019.
[14] M. Ohira, K. Takano, and Z. Ma, "A novel deep-Q-network-based fine-tuning approach for planar bandpass filter design," *IEEE Microw. Wireless Compon. Lett.*, vol. 31, no. 6, pp. 638–641, Jun. 2021.
[15] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762v5*.
[16] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2020, *arXiv:1606.08415v4*.
[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, p. 13.