Bandwidth-Scalable Digital Predistortion of Active Phased Array Using Transfer Learning Neural Network

Jalili, Feridoon; Tafuri, Felice Francesco; Jensen, Ole Kiel; Chen, Qingyue; Shen, Ming; Pedersen, Gert F.

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2023.3242648

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 January 2023, accepted 30 January 2023, date of publication 6 February 2023, date of current version 14 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242648

RESEARCH ARTICLE

Bandwidth-Scalable Digital Predistortion of Active Phased Array Using Transfer Learning Neural Network

FERIDOON JALILI[1], FELICE FRANCESCO TAFURI[2], (Member, IEEE), OLE KIEL JENSEN[1], QINGYUE CHEN[1], MING SHEN[1], (Senior Member, IEEE), AND GERT F. PEDERSEN[1], (Senior Member, IEEE)

[1]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[2]Keysight Technologies Inc., Santa Rosa, CA 95403, USA

Corresponding author: Feridoon Jalili (fja@es.aau.dk)

ABSTRACT This paper proposes a transfer learning neural network (TLNN) approach for digital predistortion (DPD) of mm-Wave active phased arrays (APA) operated under variable signal bandwidth regimes. Compared with the conventional artificial neural network (ANN) method, the proposed approach can achieve similar linearization performance with much lower computational complexity by transferring part of a trained model from one bandwidth to another bandwidth. In the recently introduced 5G, the increased signal bandwidth triggers considerable memory effects in the APA. Moreover, dealing with different signal bandwidths typically requires a time-consuming recalculation of the predistorter parameters. In this paper, the authors propose to have those challenges solved by using a DPD model based on the transfer learning method. The proposed approach was validated with over-the-air (OTA) measurements on an APA excited with signals of varying bandwidth, namely from 20 MHz to 100 MHz. Experimental results show a significant reduction in the training time while ensuring good linearization performance. With the applied TLNN DPD, an 8.5 dB improvement of adjacent channel leakage ratio (ACLR) and 8.6% points improvement of error vector magnitude (EVM) is achieved. Under the variable bandwidth regime, the complexity of the DPD model in terms of the number of multiplications is reduced from 199168 to 160. The proposed TLNN DPD proved to be robust concerning variation in the bandwidth of the APA excitation signal.

INDEX TERMS Active phased array (APA), artificial neural networks (ANN), transfer learning (TL), digital pre-distortion (DPD), over-the-air (OTA).

I. INTRODUCTION

Active phased array (APA) transmitters including multiple antennas operating at mmWave frequencies, which are used in the recent wireless communication systems, are facing new challenges in the forms of high bandwidth, high nonlinearity and mutual coupling between antennas together with dynamic change of the bandwidth. Digital predistortion (DPD) techniques based on conventional methods can not easily handle these new challenges without increasing the computational complexity. Together with the wide bandwidth, 5G has introduced a dynamic bandwidth selection that requires the mobile transmitter to quickly adapt to different operating conditions. Dynamic bandwidth selection together with the impact from the transmission channel makes the need for reusing the adjusted parameters defined for calibration, linearization, etc. highly important [1]. The transmission quality of the communication system is to a high degree dependent on how well it can dynamically change the bandwidth and power level with minimum cost in terms of speed and cost. The state-of-the-art (SoA) DPD systems deployed by the industry have excellent performance for

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Ali.

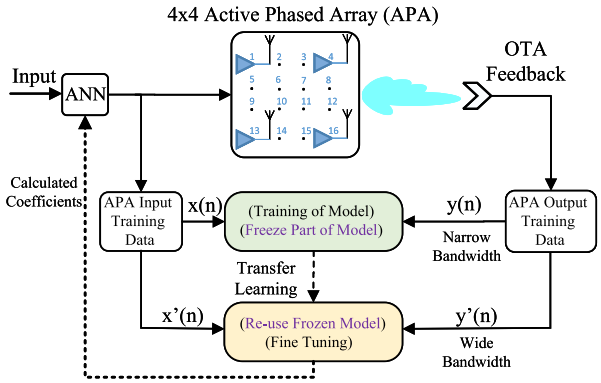


FIGURE 1. TLNN-based linearization model.

relatively steady conditions where the bandwidth and power are not rapidly changed. For cases with a rapid change of transmission parameters and environment, the existing DPD methods need to update a huge amount of coefficients which potentially can make the system complex and slow.

Artificial neural networks (ANN) have been widely used in modeling nonlinear devices because of their good approximation ability to nonlinear functions [2], [3]. For wide bandwidth signals, in particular, the memory effects have a significant impact. There are generally two dynamic neural network structures for taking care of memory effects [4]. The first structure, recurrent neural networks (RNNs), utilizes feed-forward and feedback signal processing and uses output-to-input time-delays lines. In another structure, a time-delay neural network (TDNN), combines I/Q processing with input time-delay lines to handle memory effects. In order to extract amplitude and phase information from modulated complex wave-forms, ANNs need to consider operating with either complex-valued (CV) input signals, weights and activation outputs, or real-valued (RV) double-inputs double-outputs (and real weights and activation outputs), i.e. in the form of multiple I and Q components. CV operation leads to heavy calculations and a longer training phase [5] and therefore the proposed model in this work uses the RV concept. The real-valued time-delay neural networks (RVTDNNs) offer superior performance and easy baseband implementation when used for inverse modeling of PAs with strong nonlinearities and memory effects [6].

However, by increasing the bandwidth and nonlinearity, the RVTDNN requires a higher input dimension, i.e. larger number of IQ data, and more hidden layers which make the model slow. Several works based on transfer learning have been introduced to cope with these challenges [6], [7]. The study of transfer learning is motivated by the fact that one can intelligently apply knowledge learned previously to solve new problems faster or with better solutions [8]. A similar problem also lies in the way of other dense ANN networks with several layers and neurons used in image recognition [9] and channel estimation [10], [11]. In these works, the transfer learning techniques grant the models the ability to rapid image recognition and channel estimation by leveraging prior knowledge. Inspired by these works, this paper investigates applying transfer learning DPD for bandwidth-scalable APAs. Fig. 1 shows the block diagram of the actual transfer learning neural network (TLNN) linearization technique. Part of the narrow bandwidth model from the previous training has been transferred and combined with the fine-tuning layers to make the new model for the wide bandwidth.

This paper is organized as follows: Section I is the introduction. Section II presents the proposed linearization method. The measurement setup is in section III. The optimization of the pre-designed model and the reference model is described in section IV. Section V is about transfer learning implementation. Bandwidth-scalable predistortion results are shown in section VI and finally, the conclusion of this work is presented in section VII.

II. PROPOSED TLNN LINEARIZATION METHOD

This section describes the selected model for linearization, the data structure and architecture of the model together with a complexity analysis of the proposed neural network.

A. SISO MODEL FOR TLNN-BASED LINEARIZATION

Several modified DPD algorithms have been introduced to combat the challenges raised by the recently introduced hardware configuration for 5G mmWave transmitter based on the APA [12], [13], [14]. A single input single output (SISO) model where the entire transmitter has been considered as a two-port system has been presented using an observation receiver in far-field in [15], [16], and [17]. A memory polynomial model (MPM)-based DPD technique based on this SISO model has been used for the linearization of the antenna array in presence of crosstalk. It has been shown that the trained DPD is able to mitigate the impact of cross-talk at PAs outputs, which is also called load modulation, in a limited range of steering angle. The step size for reusing the trained model is dependent on the target specification of linearity and the amount of coupling among the branches of the APA which again is dependent on the size of the array and the distance between the patches [18]. The potential mismatches between PAs can be compensated so that they all exhibit the very same behavior which is presented in [19]. In this way, linearization in all directions can be achieved with a single DPD, in contrast to linearizing the main beam only. However, this approach requires analog circuits for compensating the mismatch in each branch which may introduce high complexity and delay for large arrays and the potential changes in the PAs' behaviors due to crosstalk. In the present work, based on the SISO model, the reference signal for DPD identification is obtained through far-field measurements of an observation antenna placed at the main beam direction, Fig. 1, and the focus here is on the challenges related to high bandwidth and dynamic bandwidth behavior.

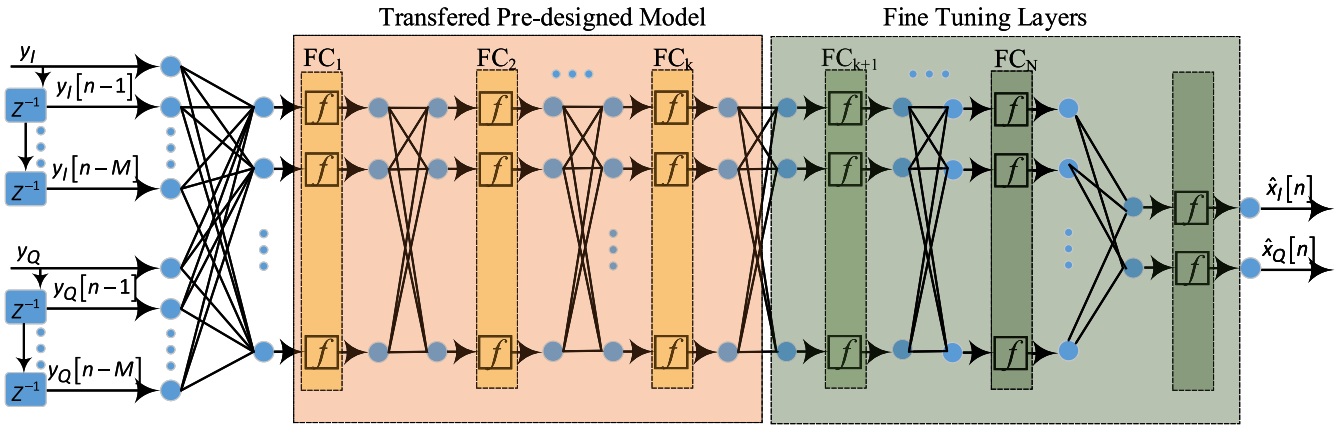


FIGURE 2. The proposed TLNN model based on RVTDNN. The transferred pre-design model is the frozen model from the previous training and is combined with the fine-tuning layers to make the new model.

B. DATA STRUCTURE OF THE MODEL

The data structure of the exploited TLNN is shown in Fig. 2, where $y_I(n)$ and $y_Q(n)$ are the I/Q components of input to the ANN and $\hat{x}_I(n)$ and $\hat{x}_Q(n)$ are the I/Q components of the output of the network. The data format of the source and target datasets is the same, and the inputs and outputs are represented as:

$$Y_n = [y_I(n), y_I(n-1), \ldots, y_I(n-M), \\ y_Q(n), y_Q(n-1), \ldots, y_Q(n-M)] \quad (1)$$

and

$$X_n = \left[\hat{x}_I(n), \hat{x}_Q(n)\right], \quad (2)$$

where M denotes the number of delay lines at the input of the network. The procedure for training is as follows: a set of source datasets, e.g. measured IQ samples of a 5G signal with 20 MHz channel bandwidth, are used for offline training. Part of the network is then used as a transfer learning model for the target dataset, which is a 5G signal that can have the same or different channel bandwidth. As illustrated in Fig. 2, the first k layers of the model, FC_k, are used for extracting the nonlinear characteristics of the APA in low bandwidth cases and are frozen after executing offline training. The output of the frozen layers, T_n, is written as:

$$T_n = f^{frozen}(X_n). \quad (3)$$

Here, $f^{frozen}(.)$ indicates the function representing the frozen layer. The block diagram in Fig. 2 represents a generic implementation of the TL concept.

C. TRANSFER LEARNING DPD ARCHITECTURE

The proposed DPD architecture used in this work is based on RVTDNN, where an arbitrary number of memory taps can be assessed [6]. The same taps configuration is employed between input and feedback signals regardless of the physics to be modeled. The proposed architecture has a fully-connected structure and the input-output relationship between the hidden layers is defined as [21], [22], [23]:

$$\mathbf{y}^{(j)} = f\left(\mathbf{W}\mathbf{x}^{(j-1)} + \mathbf{B}\right), \quad (4)$$

where j is the j-th fully connected layer and $f(.)$ is the activation function and $\mathbf{y}^{(j)}$ is a $P \times 1$ vector representing the output values of the j-th layer, $\mathbf{W}$ is a $P \times Q$ matrix representing the trainable coefficients, $\mathbf{x}^{(j-1)}$ is a $Q \times 1$ vector representing the outputs of the previous layers and $\mathbf{B}$ is a $P \times 1$ vector representing the trainable biases. Thus, the number of outputs of the previous layer is defined by Q, and the number of inputs to the next layer is defined as P. By using the activation function, denoted as f in Fig. 2, any arbitrary nonlinear functions can be fitted. The proposed RVTDNN architecture uses the rectified linear units (ReLU) activation function, which is less computationally expensive than hyperbolic tangent (Tanh) and Sigmoid because it involves simpler mathematical operations [24], [25]. The ReLU activation function is defined as:

$$\sigma_{\mathrm{ReLU}}(x) = \max(0, x) \quad (5)$$

The ReLU activation function introduces nonlinearity by setting negative inputs to 0, which also adds sparsity to the ANN and can simplify the computations.

The fine-tuning layers denoted by z, where $z = N - k$, are defined as transferred layers (TL). The output of the i-th fine-tuning layers, $(TL)_i$, is written as:

$$(TL)_i = f_1\left(w_i^T \cdot (TL)_{i-1} + b_i\right), \quad i = 1, 2, \ldots, z \quad (6)$$

where w_i^T and b_i are the weights and biases of the i-th transfer layer and the final output, Y_n' is defined as:

$$Y_n' = f_2(w_{out}^T \cdot (TL)_z + b_{out}), \quad (7)$$

where w_{out}^T and b_{out} denote the weights and biases of the output layer and $(TL)_z$ is the output of z-th transfer layer. $f_1(.)$ and $f_2(.)$ are the activation functions which can be

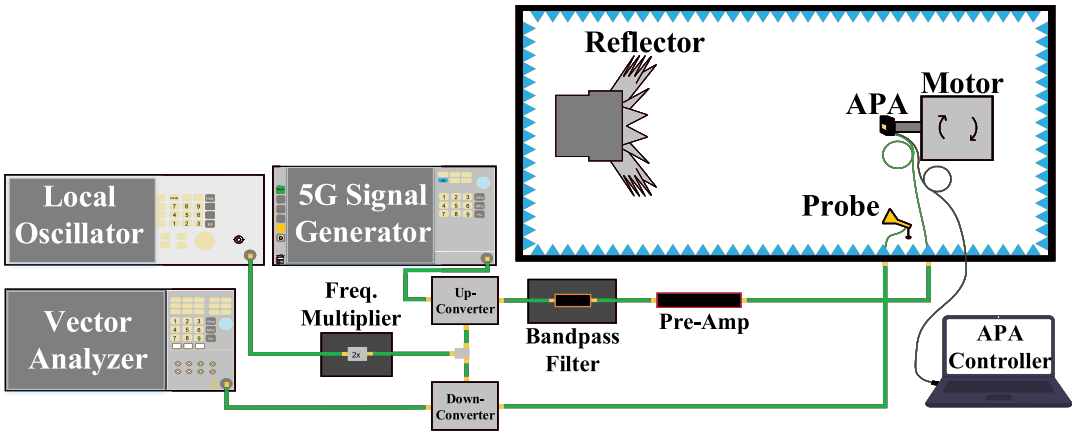


FIGURE 3. Block diagram of measurement setup in compact antenna test range chamber [20].

FIGURE 4. Measurement setup using compact antenna test range chamber.

chosen differently. In the presented work, both activation functions are of the ReLU type. The experimental dataset is divided into a training set and a validation set at 70% and 30%, respectively. The weights and biases of the network are learned by choice of an appropriate loss function. The two most used loss functions for regression tasks are mean square error (MSE) loss and Huber loss. The Huber loss is a robust loss function used for a wide range of regression tasks [26] and it is used for the presented work. The Huber loss function behaves quadratic for small residuals and linearly for large residuals and is defined as [27]:

$$L_\delta(Y'_n, Y_n) = \begin{cases} \frac{1}{2}(Y'_n - Y_n)^2 & for\ \left|Y'_n - Y_n\right| \le \delta \\ \delta\left|Y'_n - Y_n\right| - \frac{1}{2}\delta^2 & otherwise, \end{cases} \quad (8)$$

where δ, set to 1, is the parameter of Huber loss. Y'_n and Y_n denote the observation and prediction values, respectively. Through backward propagation and using the Adam optimization algorithm, the local minimum is approached. The measured data are collected and uploaded using MATLAB. The ANN is built and trained using the Keras 2.3.0-tf package in Python.

D. COMPLEXITY OF THE PROPOSED ANN

The complexity analysis is made with a starting point in Eq. 4, assuming only fully connected layers with equal amounts of neurons and $P = Q$. Between each fully connected layer, there are P^2 multiplications. The number of operations between the input layer and the first hidden layer is $2MP$ multiplications, where M is the number of time delays and P is the number of neurons. There are $2P$ multiplications between the last hidden layer and the output layer. The total amount of multiplications is:

$$C_{\mathrm{m,ANN}} = C_{\mathrm{a,ANN}} = 2MP + (J-1)P^2 + 2P, \quad (9)$$

where the number of hidden layers is defined by J.

III. OTA MEASUREMENTS SETUP

The block diagram of the OTA measurements setup using a compact antenna test range (CATR) is shown in Fig. 3 [20] and the actual laboratory setup is in Fig. 4. The 5G signal generator (R&S SMBV100B) generates the intermediate frequency (IF) signal for transmitter input. It is centered at 3 GHz and generates an up to 100 MHz bandwidth 5G NR signal. The modulation format for the 100 MHz bandwidth is 3GPP downlink OFDM 64-QAM, sub-carrier spacing of

30 kHz and 3168 active sub-carriers. With 3 sub-carrier in each resource block (RB), it ends up to 1056 RB. The sample rate of the transmitter and receiver signals is 400 MHz which gives an oversampling rate of 4. The peak-to-average power ratio (PAPR) of the input signal, after capturing and loading to the generator, is 11.6 dB. A 12.5 GHz continued-wave signal of 12.5 GHz has been generated by the local oscillator (LO) generator (Agilent E3247C) and multiplied to 25 GHz using (MITEQ-MAX2M200400) frequency multiplier. This LO signal is used for up-converting the 3 GHz modulated IF signal to 28 GHz and down-converted it back to 3 GHz. For up-converting the IF signal to the 28 GHz carrier frequency and for down-converting the signal back to IF, two active mixers operating in their highly linear region are utilized (KTX321840 and KRX321840). For selecting the up-converted modulated signal and suppressing the LO leakage and image frequency signals, a 28 GHz band-pass filter is used. The pre-amplifier is a high-power device operating more than 10 dB below its compression point. The output signal from the pre-amplifier is highly linear and the signal power is sufficient to drive the 4 × 4 APA, Amotech AAiPK428GC-A0404 [28], close to its saturated region. The APA device includes four Anokiwave AWMF-0158 transceivers [29] and integrates 16 branches of attenuators, phase shifters and PAs and 16 patch antennas in a 4 × 4 APA. The equivalent isotropic radiated power (EIRP) is 39.8 dBm at an input power of 5 dBm [28]. A host PC is used for capturing and uploading the IQ samples. The measurement setup is power calibrated to keep all other components in their linear operating regions and the only source of nonlinearity is related to the APA. For controlling the main beam of the array the code book and software tools of Amotech have been used. Fig. 5a and Fig. 5b illustrate the amplitude to amplitude (AMAM) gain distortion and the amplitude to phase (AMPM) phase distortion at the APA output. Fig. 5c shows the time-domain compression of the waveform at the APA output. All measurements are based on 100 MHz bandwidth.

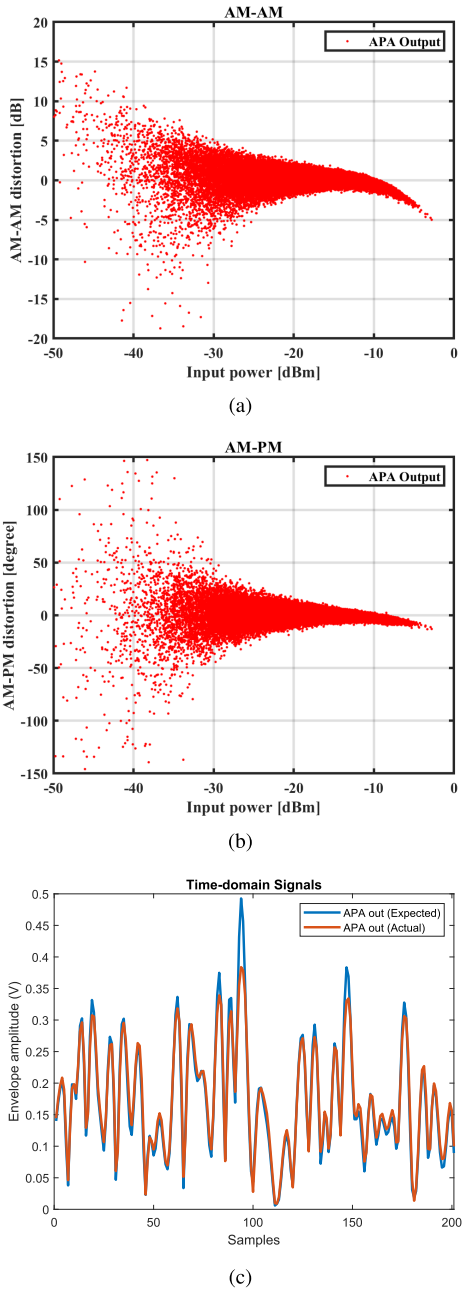
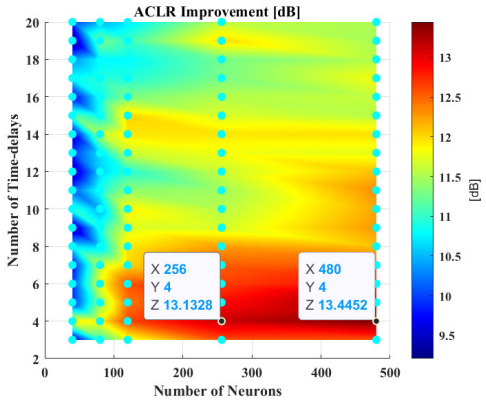


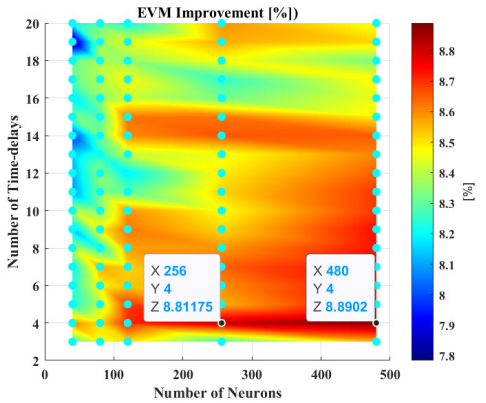
FIGURE 5. APA input-output waveforms. (a): AMAM gain distortion, (b) AMPM phase distortion, (c): Time-domain gain compression.

IV. ANN OPTIMIZATION RESULTS

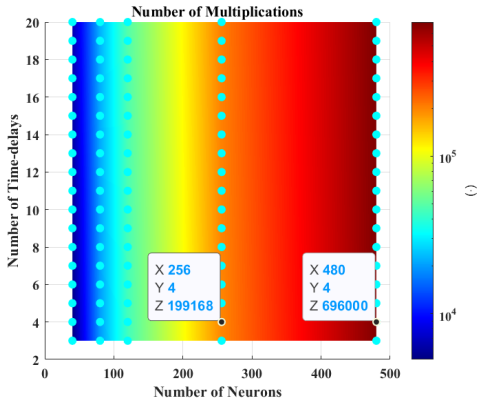
The ANN optimization methodology presented in [30] was used in this paper. The methodology is applied to an ANN model trained using two signal bandwidth values, namely 20 MHz and 100 MHz. When moving from the classical ANN model to the proposed TLNN approach, part of the optimized 20 MHz model will be frozen and used as the pre-design model for TLNN. The results from ANN optimization of 100 MHz bandwidth are used as the benchmark to compare with the results obtained using TLNN. This chapter includes the ANN optimization procedure and verification results carried out for an RF signal bandwidth of 20 MHz. The target of the ANN optimization is to minimize the number of time-delays and the number of neurons while the desired levels of linearization in terms of the adjacent channel leakage ratio (ACLR) and the error vector magnitude (EVM) are maintained. 100 k I/Q samples of the input and output signals are captured, time-aligned and used to train several ANN predistorters. There are four fully connected hidden layers in the model based on the results obtained in [30] where a number above four couldn't improve the linearization performance anymore. The time-delays parameter is swept from 3 to 20 and the neurons are swept from 40 to 480. The optimization results are assessed by constructing the network to use 70 % of the I/Q data for training and 30 % for validation. Fig. 6 shows ANN parameter optimization results of linearization of the narrow-bandwidth signal, where the optimal choice is a



(a)

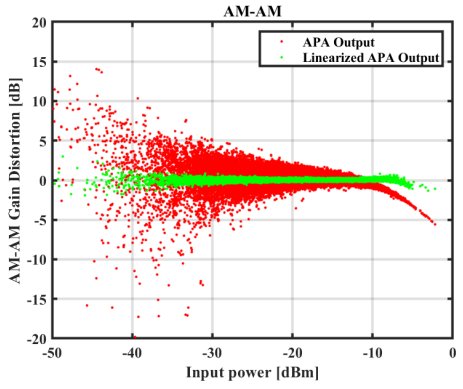


(b)

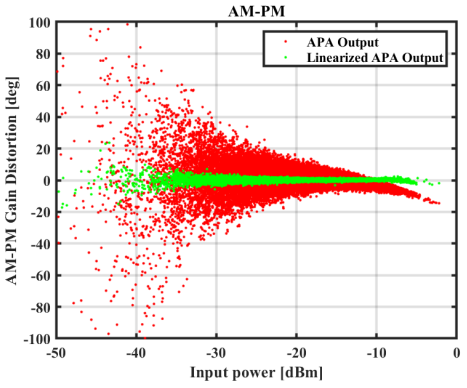


(c)

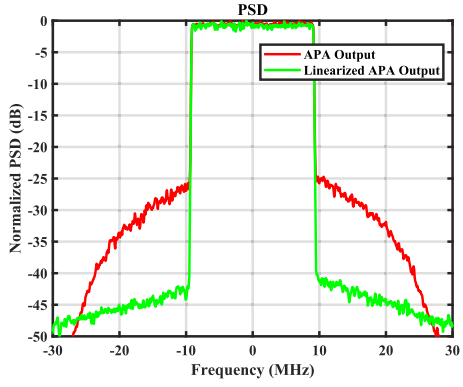
FIGURE 6. ANN parameter optimization results of linearization of the narrow bandwidth signal. (a): The ACLR (average left/right levels) improvements [dB], (b): The EVM improvements [%], (c): The number of required multiplications for each case.



(a)



(b)



(c)

FIGURE 7. ANN-based results for 20 MHz BW using 256 neurons and four time-delays. (a): The AM/AM gain distortion, (b): The AM/PM gain distortion, (c): The power spectral density.

trade-off between the ACLR, the EVM and the number of multiplications. By keeping the number of time delays to 4 and the number of neurons to 256, it is possible to achieve an ACLR improvement of 13.1 dB, as shown in Fig. 6a, and EVM improvement of 8.8 % points, Fig. 6b, while keeping the number of multiplications as low as possible, i.e. app. 199 k, Fig. 6c. Increasing the number of neurons to higher than 256 will lead to ACLR incremental improvements below 0.4 dB and EVM incremental improvements below 0.2 % points, which we consider negligible for the sake of our optimization procedure as shown in Fig. 6a-b. There is a clear indication from Fig. 6c that in a dense network, with several hidden layers, the number of multiplications will increase drastically by the number of neurons. This is in agreement with Equation (9) where the number of multiplications increases approximately as the square of the number of neurons when the number of the hidden layers exceed one. So it is important to keep as low as possible the number of neurons for a dense network with several

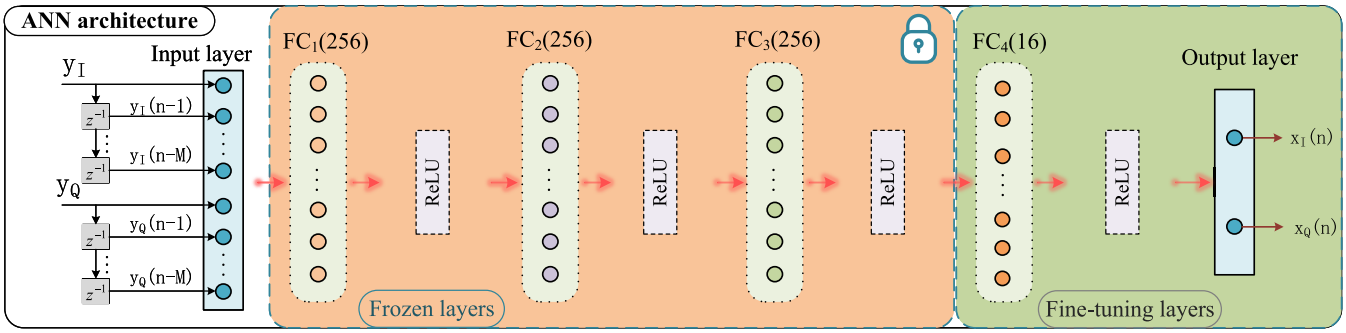


FIGURE 8. The implemented architecture of the TLNN. The transferred pre-design model is the frozen model from the previous training and is combined with the fine-tuning layers to make the new model.

layers, for achieving a lower training time and computational complexity. In [30] a procedure to find the optimal values for the number of layers and the number of neurons has been proposed. The PSD result is in Fig. 7a shows the achieved out-of-band improvement obtained by deploying the proposed optimized ANN-based DPD. Fig. 7b and Fig. 7c show the in-band AM/AM and AM/PM gain distortions related to EVM. These results are perfectly aligned with the expected performance based on the proposed optimization procedure, whose results are summarized in Fig. 6.

V. TRANSFER LEARNING IMPLEMENTATION

For implementing the transfer learning algorithm, the part of the model of 20 MHz bandwidth is copied and used as a transferred pre-design model for linearization of the 100 MHz bandwidth signal. This is done by freezing three hidden layers from the trained model of 20 MHz bandwidth. The frozen layers are then combined with the fine-tuning layers to build the model for 100 MHz bandwidth. The implemented architecture of the TLNN is in shown Fig. 8. Table 1 summarizes the implementation procedure used for the proposed method. Table 2 shows network configuration parameters for regular ANN and TLNN. By using the transfer learning approach, the number of hidden layers is reduced from four to one and the number of neurons is reduced from 256 to 16. Furthermore, the model from one bandwidth is transferred to another bandwidth which means the transferred pre-designed model already includes most of the knowledge of the nonlinear behavioral model of APA.

VI. BANDWIDTH-SCALABLE PREDISTORTION RESULTS

First, the model for the reference 100 MHz bandwidth based on regular ANN, has been optimized using the same procedure as described in section IV. This model is constructed by using four hidden layers with 256 neurons in each. Linearization results of this approach are used for bench-marking of the TLNN-based linearization of the 100 MHz bandwidth. For TLNN, the frozen pre-defined model from 20 MHz training and one fully connected fine-tuning hidden layer are used. This model has been verified with four different sets of neurons, 128, 64, 32 and 16 in the fine-tuning layer. The results for each set of neurons are

TABLE 1. Algorithm used for TLNN training.

Algorithm 1 Training of Regular ANN
i: Generate n samples of IQ data of $x[n]$ and $y[n]$ with 20 MHz BW
ii: Update weights and biases given by Eq. (4) using 70 % of n samples
iii: Continue updating until the minimum cost level is reached
iv: Validate the model using 30 % of n samples
v: If the cost function of validation is ok, then freeze the model
vi: Save the first k layers of the ANN as pre-designed model
vii: Repeat steps i-vi by using 100 MHz BW and save it as a reference model
Algorithm 2 Training of TLNN
i: Generate l samples of IQ data of $x[l]$ and $y[l]$ of 100 MHz BW
ii: combine the pre-design model with the fine-tuning model
iii: Update weights and biases by using 70 % of l samples
iv: Continue updating until the minimum cost level is reached
v: Validate the model using 30 % of l samples
vi: If the cost function of validation is ok, exit the loop
vii: Update network coefficients

TABLE 2. Network configuration parameters for Regular ANN and TLNN.

Model	Regular ANN	**TLNN**
Maximum Epochs	500	200
Minimum Batch Size	500	200
Optimizer	Adam	Adam
Loss Function	Huber loss	Huber loss
Activation Function	RelU	RelU
Initial Learning Rate	0.01	0.01
Early Stop Patience	20 epochs	20 epochs
Minimum Learning Rate	10^{-6}	10^{-6}
Number of delay lines	4	4
Number of hidden layers	4	**1**
Number of Neurons	256 (each hidden layer)	**16**

bench-marked with the regular ANN which has four fully connected layers and 256 neurons in each. The structures of the input and output layers of the networks are the same for both regular and TLNN. The number of multiplications based on Equation (9) for regular ANN and TLNN are given as:

$$C_{\text{m,ANN}} = 2 * 4 * 256 + (4-1) * 256^2 + 2 * 256 = 199168, \quad (10)$$

and for TLNN with 1 hidden layer and 16 neurons, it will result to:

$$C_{\text{m,TNN}} = 2 * 4 * 16 + 2 * 16 = 160 \quad (11)$$

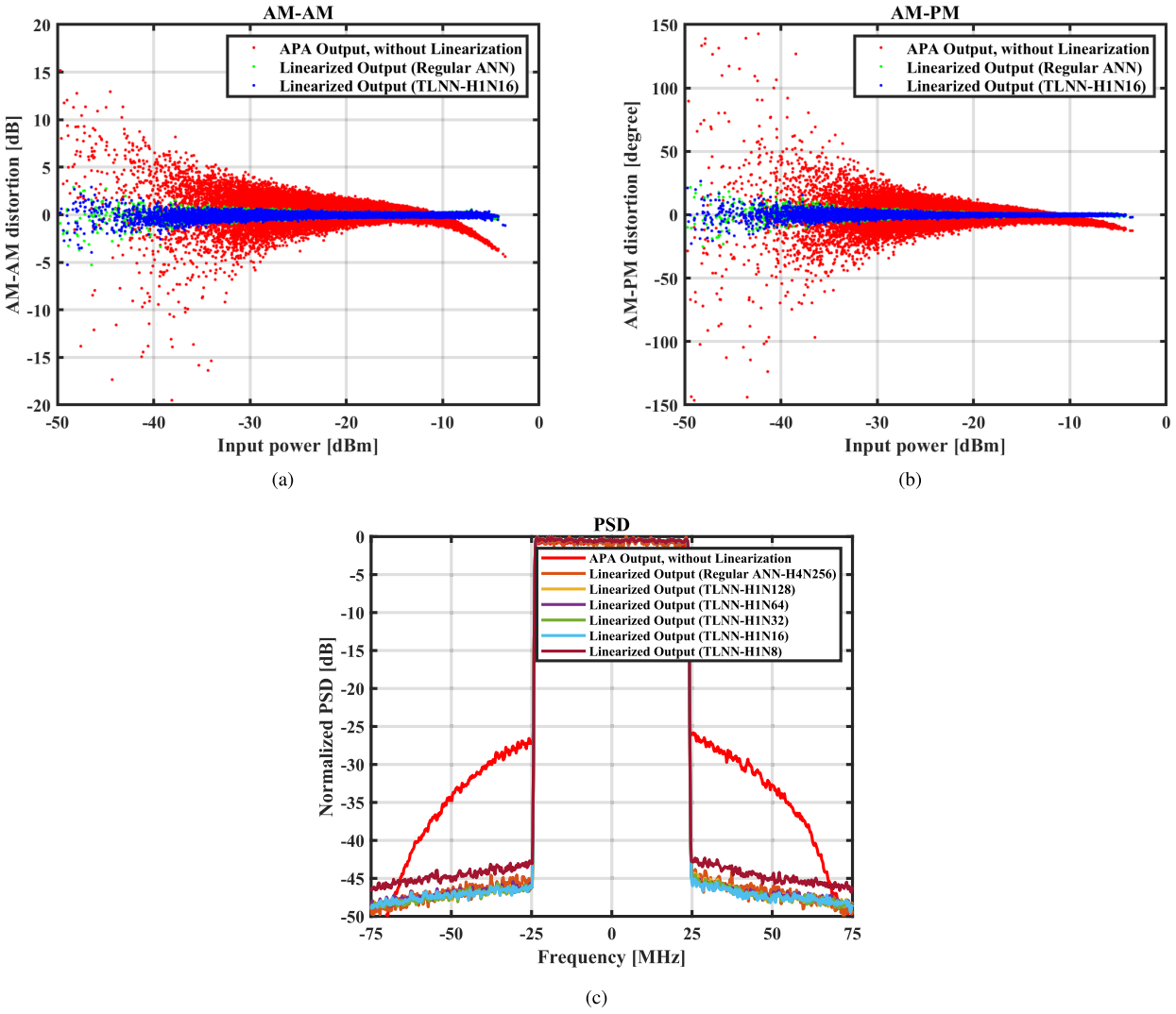


FIGURE 9. Regular ANN vs. TL-ANN for 50 MHz BW. (a): The AM/AM gain distortion, (b): The AM/PM gain distortion, (c) Power spectral density (e.g. TLNN-H1N8, means TLNN with 1 hidden layer and 8 neurons).

TABLE 3. Performance comparison between regular ANN and the proposed TLNN for 50 MHz bandwidth signal. *) ACLR is based on the average of the left and the right sides.

	Number of Multiplications	EVM (without / with DPD), (Improvement)	ACLR *) (without / with DPD), (Improvement)
Regular ANN (**256 Neurons, 4 hidden layers**)	199168	(9.9 / 1.2 %), (8.7 %)	(35.7 / 47.4 dBc), (11.7 dB)
TLNN (**128 Neurons, 1 hidden layer**)	1280	(9.9 / 1.3 %), (8.5 %)	(35.7 / 47.2 dBc), (11.5 dB)
TLNN (**64 Neurons, 1 hidden layer**)	640	(9.9 / 1.4 %), (8.5 %)	(35.7 / 47.1 dBc), (11.4 dB)
TLNN (**32 Neurons, 1 hidden layer**)	320	(9.9 / 1.3 %), (8.6 %)	(35.7 / 47.3 dBc), (11.6 dB)
TLNN (**16 Neurons, 1 hidden layer**)	160	(9.9 / 1.3 %), (8.6 %)	(35.7 / 47.3 dBc), (11.6 dB)
TLNN (**8 Neurons, 1 hidden layer**)	80	(9.9 / 1.6 %), (8.2 %)	(35.7 / 44.7 dBc), (9 dB)

A. DISCUSSION

A comparison of the verification results in terms of AM/AM and AM/PM distortion gains and PSD are illustrated in Fig. 9 and Fig. 10 for respectively 50 MHz and 100 MHz bandwidths. The TLNN linearization can provide the same level of linearity as the regular ANN. Reducing the number

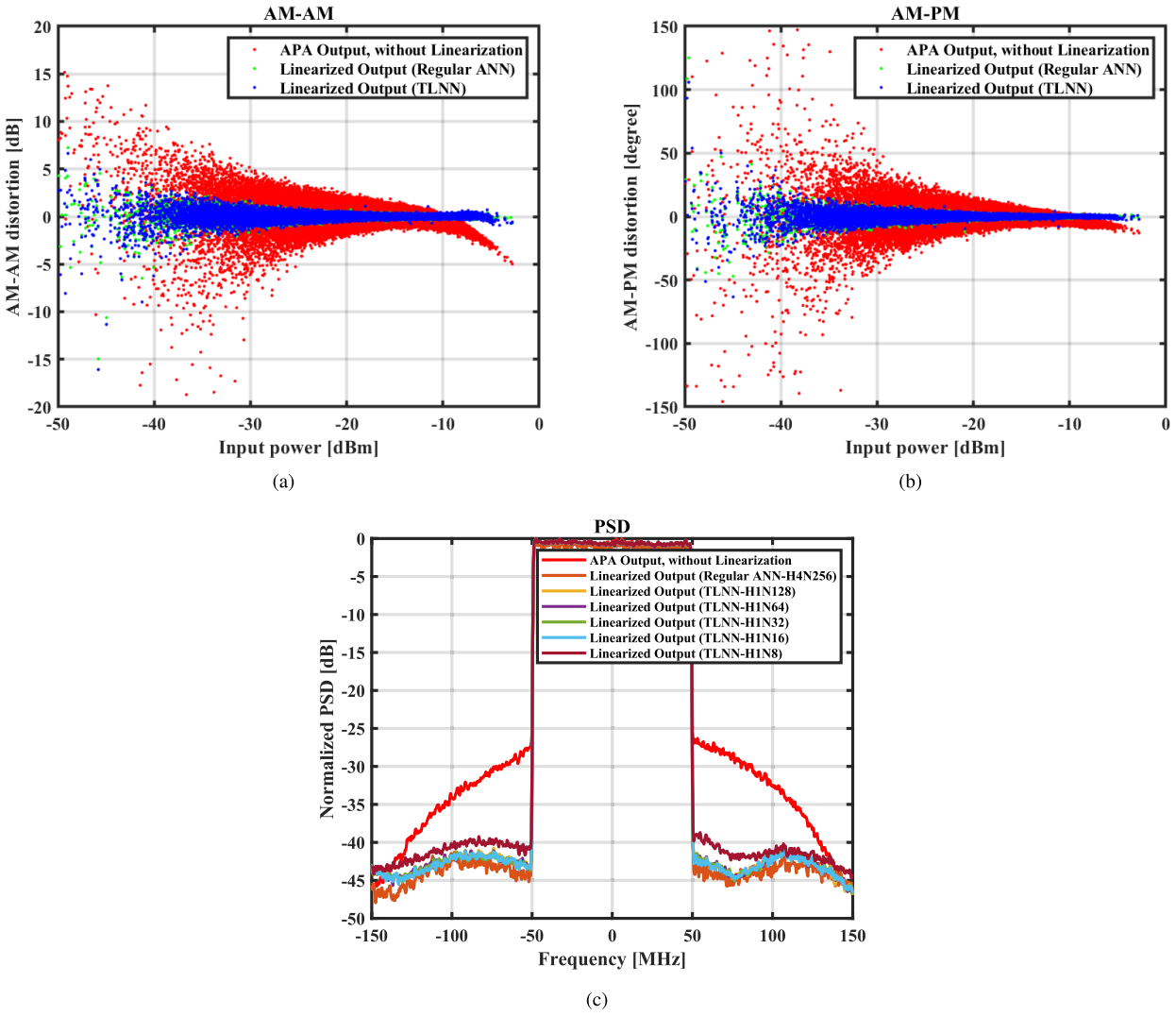


FIGURE 10. Regular ANN vs. TL-ANN for 100 MHz BW. (a): The AM/AM gain distortion, (b): The AM/PM gain distortion, (c) Power spectral density (e.g. TLNN-H1N8, means TLNN with 1 hidden layer and 8 neurons).

TABLE 4. Performance comparison between regular ANN and the proposed TLNN for 100 MHz bandwidth signal. *) ACLR is based on the average of the left and the right sides.

	Number of Multiplications	EVM (without / with DPD), (Improvement)	ACLR *) (without / with DPD), (Improvement)
Regular ANN (**256 Neurons, 4 hidden layers**)	199168	(10.3 / 1.6 %), (8.7 %)	(34.7 / 43.7 dBc), (9 dB)
TLNN (**128 Neurons, 1 hidden layer**)	1280	(10.3 / 1.8 %), (8.5 %)	(34.7 / 43.2 dBc), (8.5 dB)
TLNN (**64 Neurons, 1 hidden layer**)	640	(10.3 / 1.8 %), (8.5 %)	(34.7 / 43 dBc), (8.3 dB)
TLNN (**32 Neurons, 1 hidden layer**)	320	(10.3 / 1.7 %), (8.6 %)	(34.7 / 43.1 dBc), (8.4 dB)
TLNN (**16 Neurons, 1 hidden layer**)	160	(10.3 / 1.7 %), (8.6 %)	(34.7 / 43.7 dBc), (8.5 dB)
TLNN (**8 Neurons, 1 hidden layer**)	80	(10.3 / 2.1 %), (8.2 %)	(34.7 / 41.3 dBc), (6.8 dB)

of neurons to e.g. 8 neurons, results in degradation of the performance in terms of EVM and ACLR. This highlights the role of the fine-tuning layers. Detailed performance comparisons between regular ANN and the proposed TLNN are in Table 3 and Table 4 for respectively 50 MHz and 100 MHz bandwidths. These results show that it is pos-

sible to achieve approximately the same linearization performance compared to regular ANN, i.e. an EVM improvement of 8.6 % points and ACLR improvement of 9 dB, by using TLNN with 16 neurons as is shown in Table 4. Hence the proposed approach proves to be robust versus signal bandwidth and can be used as a bandwidth-scalable linearization technique. On the other hand, TLNN allows for reducing the number of hidden layers (through re-using the frozen model) and the number of neurons which results in the relaxation of the computational complexity in terms of the number of multiplications. The outcomes of the performed linearization experiments can be summarized as follows:

1) By using the SoA conventional RVTDNN approach for linearization of the actual APA, we need an ANN DPD of 256 neurons, 4 hidden layers for 20 MHz signal linearization, and another ANN DPD of 256 neurons, 4 hidden layers multiplications for 100 MHz signal linearization.
2) TLNN approach instead can reuse the model calculated for 20 MHz and need only additional 16 neurons and one layer for 100 MHz signal linearization.
3) For an adaptive DPD, the time to calculate the incremental layers in TLNN is reduced and grants the models the ability to be adaptively re-identified.
4) A clear advantage delivered by the proposed is in terms of the LUT (look-Up Table) size necessary to implement the DPD. Instead of storing two completely different sets of ANN DPD parameters (SoA approach), one for the narrow bandwidth use case and the other for the wide bandwidth use case, system engineers will need to store much fewer parameters for linearizing the wide bandwidth use case, because they can reuse most of the ones calculated for the narrow bandwidth.

A long duration of the algorithm identification will be a problem for an adaptive online NN-based linearization technique. However, using the proposed TLNN, the time to calculate the incremental layers will be reduced and consequently relax the adaptive online processing issue. The HW implementation itself is challenged by the realization of the online OTA feedback receiver. There is a need for a far-filed observation antenna for providing the OTA feedback signal for adaptive online DPD. The feedback signal could be obtained from the receiver antenna of the same device, but the proper implementation techniques are still under discussion in industry and academia. One promising proposal is to use the auxiliary antenna connection (the diversity or MIMO antenna) [31]. However, there may be an issue with the low coupling ratio between the transmitter and these auxiliary antennas.

VII. CONCLUSION

This paper presented a bandwidth-scalable over-the-air DPD of an APA transmitter based on a TLNN method. The proposed methodology allows for reducing the hardware implementation complexity in terms of the number of multiplications while ensuring the same linearization performance as a regular ANN. In the proposed method, part of the model is fixed as a pre-designed model, and then an incremental model component was trained and deployed for fine-tuning the remaining adaptation layers to build the final model. This paper demonstrated how such a TL technique could be used to implement a bandwidth-scalable digital predistorter. The ANN layers identified for one signal bandwidth were reused and enhanced with an incremental neuron layer to allow the ANN predistorter to successfully linearize input signals with wider bandwidths. The proposed linearization technique was validated with measurements on a state-of-the-art 4 × 4 APA and a setup using up- and down-conversion from sub-6 GHz to 28 GHz for verification. Experimental results showed that our optimized ANN-based DPD could linearize a 20 MHz 5G signal with an EVM improvement of 8.8 % points and an ACLR improvement of 13.3 dB. It was also demonstrated that using TL, the same ANN DPD can be reused to linearize a 5G signal with a much wider bandwidth, namely 100 MHz. To do so, only an additional layer of 16 neurons was added on top of the reused ANN DPD. Such an approach allowed us to obtain an EVM improvement of 8.6 % points and an ACLR improvement of 8.5 dB. The multiplications of the ''Frozen layers'' should also be considered when evaluating the complexity of the overall TLNN-based DPD actuator, however, the complexity of TLNN-based DPD model identification is reduced to a factor of 160/199168 compared with the conventional ANN. The reduced complexity allows to bring down the cost of the implementation using digital hardware. Further research is being conducted to make the proposed bandwidth-scalable DPD fully robust concerning the signal bandwidth and other transmitter operating conditions. Our future goal is to enhance the TL methodology to obtain a universal set of parameters that can be fully reused to linearize multiple signal bandwidths. Such a result would allow lowering further the complexity and cost of the DPD implementation on digital hardware. Furthermore, we expect that if the average output power and the peak-to-average power ratio change greatly, the nonlinear characteristics of the power amplifier will also change. An investigation of the capability of TL-based ANN for power-scalable scenarios may also be interesting for future work.

REFERENCES

[1] E. Guillena, W. Li, G. Montoro, R. Quaglia, and P. L. Gilabert, ''Reconfigurable DPD based on ANNs for wideband load modulated balanced amplifiers under dynamic operation from 1.8 to 2.4 GHz,'' *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 1, pp. 453–465, Jan. 2021.

[2] K. Hornik, ''Approximation capabilities of multilayer feedforward networks,'' *Neural Netw.*, vol. 4, no. 2, pp. 251–257, 1991.

[3] F. Jalili, *Digital Predistorion of 5G Millimeter-Wave Active Phased Arrays using Artificial Neural Networks*. Aalborg Universitetsforlag, 2022.

[4] T. Liu, S. Boumaiza, and F. M. Ghannouchi, ''Dynamic behavioral modeling of 3G power amplifiers using real-valued time-delay neural networks,'' *IEEE Trans. Microw. Theory Techn.*, vol. 52, no. 3, pp. 1025–1033, Mar. 2004.

[5] P. L. Gilabert, D. López-Bueno, T. Q. A. Pham, and G. Montoro, "Machine learning for digital front-end: A comprehensive overview," in *Machine Learning for Future Wireless Communications*. 2020, pp. 327–381.
[6] M. Rawat, K. Rawat, and F. M. Ghannouchi, "Adaptive digital predistortion of wireless power amplifiers/transmitters using dynamic real-valued focused time-delay line neural networks," *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 1, pp. 95–104, Jan. 2009.
[7] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.
[8] K. Weiss, T. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, pp. 1–40, May 2016.
[9] P. Smith and C. Chen, "Transfer learning with deep CNNs for gender recognition and age estimation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2564–2571.
[10] Y. Yang, F. Gao, Z. Zhong, B. Ai, and A. Alkhateeb, "Deep transfer learning-based downlink channel prediction for FDD massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7485–7497, Dec. 2020.
[11] W. Alves, I. Correa, N. Gonzalez-Prelcic, and A. Klautau, "Deep transfer learning for site-specific channel estimation in low-resolution mmWave MIMO," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1424–1428, Jul. 2021.
[12] E. Ng, Y. Beltagy, G. Scarlato, A. B. Ayed, P. Mitran, and S. Boumaiza, "Digital predistortion of millimeter-wave RF beamforming arrays using low number of steering angle-dependent coefficient sets," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 11, pp. 4479–4492, Nov. 2019.
[13] C. Fager, T. Eriksson, F. Barradas, K. Hausmair, T. Cunha, and J. C. Pedro, "Linearity and efficiency in 5G transmitters: New techniques for analyzing efficiency, linearity, and linearization in a 5G active antenna transmitter context," *IEEE Microw. Mag.*, vol. 20, no. 5, pp. 35–49, May 2019.
[14] X. Liu, Q. Zhang, W. Chen, H. Feng, L. Chen, F. M. Ghannouchi, and Z. Feng, "Beam-oriented digital predistortion for 5G massive MIMO hybrid beamforming transmitters," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 7, pp. 3419–3432, 2018.
[15] F. Jalili, M. H. Nielsen, M. Shen, O. K. Jensen, J. H. Mikkelsen, and G. F. Pedersen, "Linearization of active transmitter arrays in presence of antenna crosstalk for 5G systems," in *Proc. IEEE Nordic Circuits Syst. Conf. (NORCAS), NORCHIP Int. Symp. System-on-Chip (SoC)*, Oct. 2019, pp. 1–5.
[16] K. Hausmair, U. Gustavsson, C. Fager, and T. Eriksson, "Modeling and linearization of multi-antenna transmitters using over-the-air measurements," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–4.
[17] A. Brihuega, M. Abdelaziz, L. Anttila, M. Turunen, M. Allén, T. Eriksson, and M. Valkama, "Piecewise digital predistortion for mmwave active antenna arrays: Algorithms and measurements," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 9, pp. 4000–4017, Jun. 2020.
[18] F. Jalili, F. F. Tafuri, O. K. Jensen, Y. Li, M. Shen, and G. F. Pedersen, "Linearization trade-offs in a 5G mmWave active phased array OTA setup," *IEEE Access*, vol. 8, pp. 110669–110677, 2020.
[19] C. Yu, J. Jing, H. Shao, Z. H. Jiang, P. Yan, X.-W. Zhu, W. Hong, and A. Zhu, "Full-angle digital predistortion of 5G millimeter-wave massive MIMO transmitters," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 7, pp. 2847–2860, Jun. 2019.
[20] F. Jalili, Y. Zhang, M. Hintsala, O. K. Jensen, Q. Chen, M. Shen, and G. F. Pedersen, "Highly non-linear and wide-band mmWave active array OTA linearisation using neural network," *IET Microw., Antennas Propag.*, vol. 16, no. 1, pp. 62–77, Jan. 2022.
[21] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.
[22] Y. Zhang, Z. Wang, Y. Huang, W. Wei, G. F. Pedersen, and M. Shen, "A digital signal recovery technique using DNNs for LEO satellite communication systems," *IEEE Trans. Ind. Electron.*, vol. 68, no. 7, pp. 6141–6151, Jul. 2021.
[23] J. G. Brask, K. B. Olesen, L. F. Dyring, A. Yadav, F. Jalili, Y. Zhang, and M. Shen, "Deep digital signal recovery for LEO satellite communication in presence of system perturbations," in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 2021, pp. 1–3.
[24] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *J. Mach. Learn. Res.*, vol. 15, 2010.
[25] C. Liu, T. Arnon, C. Lazarus, C. Strong, C. Barrett, and M. J. Kochenderfer, "Algorithms for verifying deep neural networks," 2019, *arXiv:1903.06758*.
[26] G. P. Meyer, "An alternative probabilistic interpretation of the Huber loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5261–5269.
[27] P. J. Huber, "Robust statistics," in *International Encyclopedia of Statistical Science*. Springer, 2011, pp. 1248–1251.
[28] (2019). Amotech. *AAiP28G A0808 EVB Version R01 190321c*. [Online]. Available: https://www.mrc-gigacomp.com/Amotech.php
[29] (2019). Anokiwave. *AWMF-0158 28 GHz Silicon 5G Tx/Rx Quad Core IC*. [Online]. Available: https://www.anokiwave.com/products/awmf-0158/index.html
[30] F. Jalili, Y. Zhang, F. F. Tafuri, O. K. Jensen, Y. Li, Q. Chen, M. Shen, and G. F. Pedersen, "Tuning of deep neural networks for over-the-air linearization of highly nonlinear wide-band active phased arrays," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–4.
[31] X. Liu, W. Chen, L. Chen, F. M. Ghannouchi, and Z. Feng, "Linearization for hybrid beamforming array utilizing embedded over-the-air diversity feedbacks," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 12, pp. 5235–5248, Dec. 2019.

FERIDOON JALILI received the M.Sc. degree in electrical engineering from Aalborg University, Denmark, in 1992, where he is currently pursuing the Ph.D. degree with the Section of Antennas, Propagation and Millimeter-Wave Systems (APMS), Department of Electronic Systems.

He started his industrial work at the Research Department, Terma Electronic, Aarhus, Denmark, where he worked on the design of S-band power amplifiers for ballistic radars. His work in mobile communication started in 1996, as an RF Designer at Maxon Cellular Systems and later on at Ericsson Mobile Communications, Aalborg, Denmark. Starting in 2001, he was a System Architect at Infineon Technologies, responsible for advanced industrial research within mobile technologies. In 2011, he joined Intel Corporation, Aalborg, as a Senior Staff RF System Architect and was in charge of RF front-ends design of mobile platforms for Intel's tier 1 customers. He holds three U.S. patents in this field. His research interests include efficiency improvement of millimeter-wave circuits and systems, with a special focus on digital predistortion techniques for the 5G antenna-array transmitters.

FELICE FRANCESCO TAFURI (Member, IEEE) received the B.Sc. and M.Sc. degrees (cum laude) in electronics engineering from the Polytechnic of Bari, Italy, in 2007 and 2010, respectively, and the Ph.D. degree from the Wireless Communication Program, Aalborg University, Denmark, in 2014, with a dissertation on linearity and efficiency enhancement of mobile communication power amplifiers.

From 2014 to 2018, he was an Industrial Postdoctoral Researcher with Aalborg University and Keysight Technologies, Denmark, working on advanced measurement platforms for envelope-tracking transmitter characterization and modeling. Since 2018, he has been with Keysight Technologies, Spain, as an Application Engineer. His research interests include behavioral modeling and digital predistortion of RF power amplifiers, nonlinear measurement platforms, and applications of digital signal processing to measurement techniques. In 2013, he was a recipient of the European Microwave Conference Young Engineer Prize.

OLE KIEL JENSEN was born in Denmark, in 1955. He received the M.Sc. degree from Aalborg University, Denmark, in 1979. He has been with Aalborg University, since 1980, and has been part-time employed in microwave companies for a few periods. He is currently an Associate Professor. His teaching has been in a wide range of basic and microwave electronics. His research interests include microwave electronics, such as RF-CMOS circuit design and measurement techniques, communication systems, RF-transceiver architectures, modeling, analysis, design, and measurement techniques, and antenna measurement techniques, such as spherical near-field measurements.

QINGYUE CHEN was born in Qingdao, China. He received the B.Eng. degree from the Shandong University of Science and Technology, Qingdao, in 2017. He is currently pursuing the Ph.D. degree in the electromagnetic field and microwave technology with the University of Chinese Academy of Sciences (UCAS), Beijing, China.

He is also a Guest Ph.D. Student with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark. He worked on the linearization of RF-PAs and channel equalization. His research interests include artificial neural networks and their application for communication systems.

MING SHEN (Senior Member, IEEE) was born in Yuxi, China. He received the M.Sc. degree in electrical engineering from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2005, and the Ph.D. degree in wireless communications from Aalborg University, Aalborg, Denmark, in 2010.

He is currently an Associate Professor of RF and mm-wave circuits and systems with the Department of Electronic Systems, Aalborg University. He has 20 years of experience in RF and millimeter wave circuits and systems, including 12 years of experience in CMOS RF/mixed-signal IC design. He is the grant holder and the PI of two Danish national research projects, and the Management Committee Member substitute from Denmark in the EU COST Action IC1301 with the aim to gather international efforts and address efficient wireless power transmission technologies. His current research interests include circuits and antennas for 5G and satellite communications, low-power CMOS RF and millimeter wave circuits and systems, circuits and systems for biomedical imaging, and artificial intelligence. He is a TPC Member of IEEE NORCAS. His Ph.D. thesis was listed at the Spar Nord Annual Best Thesis Nomination. He serves as a Reviewer for IEEE and Kluwer.

GERT F. PEDERSEN (Senior Member, IEEE) was born in 1965. He received the B.Sc.E.E. (Hons.) degree in electrical engineering from the College of Technology in Dublin, Dublin Institute of Technology, Dublin, Ireland, in 1991, and the M.Sc.E.E. and Ph.D. degrees from Aalborg University, Aalborg, Denmark, in 1993 and 2003, respectively.

Since 1993, he has been with Aalborg University, where he is currently a Full Professor, heading the Antennas, Propagation and Millimeter-Wave Systems Laboratory, with 25 researchers. He is also the Head of the Doctoral School on Wireless Communication, with some 40 Ph.D. students enrolled. He has also worked as a Consultant for the development of more than 100 antennas for mobile terminals, including the first internal antenna for mobile phones, in 1994, with the lowest SAR, the first internal triple-band antenna, in 1998, with low-SAR and high-TRP and TIS, and lately, various multi-antenna systems rated as the most efficient on the market. He has worked most of the time with joint university and industry projects and has received more than U.S. $21M in direct research funding. He is also the Project Leader of the RANGE project with a total budget of more than U.S. $8M investigating high-performance centimeter/millimeter-wave antennas for 5G mobile phones. He has been one of the pioneers in establishing over-the-air measurement systems. The measurement technique is now well established for mobile terminals with single antennas and he was chairing the various COST groups with liaison to 3GPP and CTIA for the over-the-air test of MIMO terminals. He is also involved in MIMO OTA measurement. He has published more than 500 peer-reviewed papers, six books, and 12 book chapters, and holds over 50 patents. His research interests include radio communication for mobile terminals especially small antennas, diversity systems, propagation, and biological effects.