Neural Network on the Edge: Efficient and Low Cost FPGA Implementation of Digital Predistortion in MIMO Systems

Yiyue Jiang*, Andrius Vaičaitis†, Miriam Leeser*, John Dooley†
* Department of Electrical and Computer Engineering, Northeastern University, Boston, USA
† Department of Electronic Engineering, Maynooth University, Maynooth, Ireland

***Abstract*—Base stations in cellular networks must operate linearly, power efficiently, and with ever increasing flexibility. Recent FPGA hardware advances have demonstrated linearization using neural networks, however the latency introduced by these solutions is a concern. We present a novel hardware implementation for a low digital cost, high throughput pipelined Real Valued Time Delay Neural Network (RVTDNN) structure with a hardware-efficient activation function. Network training times are reduced by minimizing the training signal samples used, based on a biased probability density function (pdf). The design has been experimentally validated using an AMD/Xilinx RFSoC ZCU216 board and surpasses the data throughput of conventional RVTDNN-based DPD while using a fraction of their hardware utilization.**

***Index Terms*—Digital predistortion, Time Delay Neural Networks, FPGA, power amplifiers, MIMO**

Power Amplifiers (PAs) are core components for cellular network base stations. As PA bandwidths become wider and more power efficient, their behavior becomes more non-linear. To maintain the power efficiency but linearize the nonlinear PA output, Digital PreDistortion (DPD) techniques are employed.

FPGA-based implementations for DPD have been published for polynomial-based structures and neural networks [1]–[4]. This work extends the state-of-the-art by implementing a neural network solution on FPGA fabric with lower latency. Previous works have avoided investigating how the coefficient training is to be carried out at the network edge. Implementations of DPD with changing coefficients on FPGA fabric have been performed by training offline [5]. In this paper, we present an augmented neural network structure with advanced training methodologies to reduce the training time. One focus is on the minimisation of training time through intelligent selection of training samples. A second focus is more efficient digital instantiations for the activation functions to speed up the forward path calculation of DPD.

The contributions of this work are: 1) A procedure for reducing the number of training samples used while retaining important signal characteristics to enable real-time training; 2) an efficient FPGA implementation of predistortion using RVTDNN that makes use of an accurate activation function; and 3) the application of training and DPD to multiple PAs for MIMO operation.

RVTDNN FOR DPD

a) Training for DPD: To train the RVTDNN to perform DPD, the Indirect Learning Architecture (ILA) is used combined with biased pdf training. We do training on the RFSoC ARM processor and inference on its FPGA fabric. This allows us to rapidly adapt to changes in the current environment and the devices' region of operation. To speed up the training phase of RVTDNN on the ARM processor, the training data set is reduced without losing its range of characteristics. In this paper, we use a subset of four PAs' averaged output signal samples guided by the biased pdf. This method first bins the inputs to RVTDNN in hardware in a computationally efficient way. Subsequently the linear and nonlinear parts of the training data are separated according to a threshold, and random samples are selected from each bin in the linear part as a fraction of the bin size while also keeping previous samples according to the memory tap size. In addition, all samples from the signal's nonlinear part and their previous samples (according to the memory tap size) are selected. Fig. 1. shows the biased pdf based sample selection used in a two layer RVTDNN where the subsampling fraction of linear samples is chosen to be 0.3 and the threshold between linear and nonlinear samples is 0.7. The yellow part shows the samples retained for training. After biased pdf sample selection, the training sample size is reduced from 204.8K to 64.7K in this example.

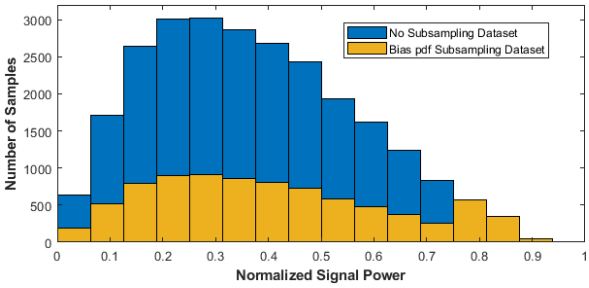

Fig. 1. pdf of Biased Subsampling

b) Pipelined RVTDNN: The RVTDNN digital implementation involves linear operations (weighted multiplications and additions) for neural net outputs, while it also makes use of Look-up Tables (LUTs) for nonlinear activation functions. To use the hardware resources efficiently, the hidden layer neuron structure makes use of the DSP48 slices available in the FPGA fabric. Fig. 2 shows that in a single neuron, the inputs after delay taps are multiplied by the corresponding weights and accumulated in sequence to be fed into the activation function block. The nonlinear hyperbolic tangent sigmoid function used

978-3-9819263-7-8/DATE23/© 2023 EDAA

is defined by:

$$tansig(n) = \frac{2}{1+\exp(-2n)} - 1 \quad (1)$$

We multiply the output of the hidden layer neuron using a shift operation, and use LUTs and adders to implement the exponential function and get the 16-bit fixed-point output. By using this pipelined architecture and LUT-based nonlinear activation function, for N hidden neurons with M delay taps, the design requires only $2N(M+1)$ DSP48 slices.

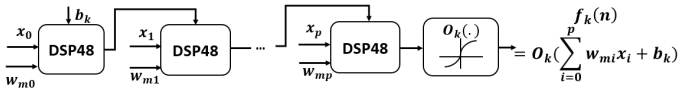

Fig. 2. A Single Hidden Neuron Structure

c) RFSoC Architecture: Our MIMO DPD system is designed based on the AMD/Xilinx RFSoC ZCU216 board to test the performance. This RFSoC evaluation board with 16 pairs of integrated 14-bit ADCs and DACs, a quad core ARM processor and FPGA fabric is an excellent target for verifying this subsample-trained and pipelined-RVTDNN based DPD. The SoC architecture is shown in Fig. 3. The signal to be transmitted, sent from the host computer, will go through the Processor System (PS) side DDR memory and then be stored in the external DDR4 on the Programmable Logic (PL) side. The RF front end DAC will transmit the signal by reading DDR4 memory repeatedly. After the signal passes through the four power amplifiers, the four PAs' outputs will be written back to the PS and used in the DPD training phase.

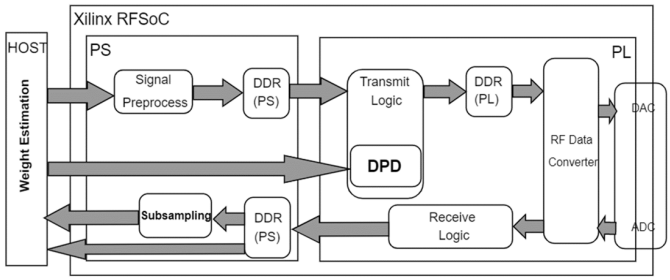

Fig. 3. SoC Architecture

RESULTS

DPD testing targets four Skyworks SKY66297-11 4W PAs with the AMD/Xilinx RFSoC ZCU216 board. A customized 5MHz LTE signal with peak-cancellation crest factor reduction (PC-CFR) is used for a 16-neuron hidden layer, 2-neuron output layer RVTDNN. For the activation functions, the hidden layer uses the hyperbolic tangent sigmoid function, while the output layer uses a purely linear function.

We set the sampling fraction to 0.8 and threshold to 0.4 for the biased pdf-based subsampling function. The signal quality is measured by Normalised Mean Square Error (NMSE) and Adjacent Channel Leakage Ratio (ACLR). The NMSE and ACLR from RVTDNN with bias pdf subsampling method is -24.32 dB and -37.60 dB, while without subsampling it is -23.04 dB and -37.55 dB. The biased pdf-based subsampling RVTDNN can offer a similar or even slightly better result in terms of accuracy while using less training time. The signal without DPD only gets NMSE and ACLR as -15.77 dB and -32.17 dB. The improvement is because the biased pdf-based subsampling method reduces the training data size while making the neural network more focused on the nonlinear characteristics. The DPD spectrum performance is shown in Fig. 4, which shows the averaged AM/AM curve and spectral response of all the PAs. The hardware utilization of pipelined RVTDNN is much smaller than the conventional RVTDNN structure with a similar operating frequency of around 270 MHz. The pipelined RVTDNN uses only 287 DSP slices, 2385 LUTs, and 2597 flip flops, while the conventional structure uses 700 DSP slices, 6061 LUTs, and 4732 flip flops.

CONCLUSION

Achieving a more computationally efficient implementation for multi-path digital pre-distortion is critically important for future cellular network base stations. In this paper, we present a neural network structure that improves performance compared to previous neural network based DPD instantiations on FPGAs. In addition, a novel approach to selecting samples to reduce the training time has been presented that preserves the non-linear samples.

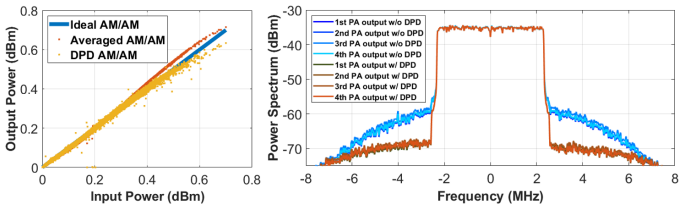

Fig. 4. DPD performance

ACKNOWLEDGMENT

This publication has emanated from research conducted with the financial support of MathWorks, Xilinx and the Science Foundation Ireland (SFI) under Grant Number 18/CRT/6222.

REFERENCES

[1] Z. Han, M. Loughman, Y. Jiang *et al.*, "Computationally efficient look-up-tables for behavioral modelling and digital pre-distortion of multi-standard wireless systems," in *Cognitive Radio Oriented Wireless Networks*, 2022.

[2] S. A. Juárez-Cázares, A. Meléndez-Cano *et al.*, "FPGA-based modeling and design methodology of a digital pre-distortion system for power amplifier linearization," in *Int'l Conf on Mechatronics, Electronics and Automotive Engineering (ICMEAE)*, 2016.

[3] S. Yeşil, C. Şen, and A. Ö. Yılmaz, "Experimental analysis and FPGA implementation of the real valued time delay neural network based digital predistortion," in *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2019, pp. 614–617.

[4] R. S. N. Ntouné, M. Bahoura, and C.-W. Park, "Power amplifier behavioral modeling by neural networks and their implementation on FPGA," in *IEEE Vehicular Technology Conference*, 2012.

[5] W. Li, E. Guillena, G. Montoro, and P. L. Gilabert, "Fpga implementation of memory-based digital predistorters with high-level synthesis," in *2021 IEEE Topical Conference on RF/Microwave Power Amplifiers for Radio and Wireless Applications (PAWR)*, 2021, pp. 37–40.