Gate-Level Analysis of LSTM Structures for Digital Predistortion of RF Power Amplifiers

Xiaodong Liu,Qiyuan Wu,Rui Huang,Yukai Jin
School of Communication Engineering
Heriot-Watt University of Electronic Science and Technology
Email: qw2011@hw.ac.uk

***Abstract*—Digital predistortion (DPD) remains a practical and widely adopted solution for mitigating the nonlinear distortion introduced by radio-frequency (RF) power amplifiers (PAs). As communication systems continue to push toward wider bandwidths and higher efficiency, accurately modeling PA behavior becomes increasingly important. In recent years, recurrent neural networks (RNNs), particularly long short–term memory (LSTM) structures, have shown promising results in capturing the dynamic and nonlinear characteristics of PAs.**

Most existing work, however, focuses on the overall performance of these networks and pays relatively little attention to the role played by individual gate components inside the LSTM unit. A deeper understanding of these mechanisms is valuable, especially when considering model simplification or hardware implementation. In this paper, we take a closer look at LSTM architectures from a gate-level perspective and investigate how different gate configurations influence DPD performance. By examining several LSTM variants with modified or simplified gate structures, we aim to provide practical insights into how each gate contributes to learning PA behavior and how lightweight recurrent architectures can be designed for RF linearization tasks.

***Index Terms*—Digital Predistortion, LSTM, Gate Analysis, Power Amplifier, Neural Networks**

I. Introduction

Power amplifiers (PAs) are essential components in wireless transmitters, but their nonlinear characteristics can lead to significant in-band distortion and unwanted spectral emissions. These effects become more pronounced in modern systems operating with high bandwidth and high efficiency. Digital predistortion (DPD) is widely used to compensate for these impairments, allowing the PA to operate efficiently while still meeting linearity requirements.

Traditional behavioral models, such as memory polynomials or Volterra-based methods, provide a useful framework but often struggle to capture the strong memory effects that appear in wideband scenarios. Motivated by this limitation, researchers have increasingly explored machine-learning-based approaches for DPD. Among them, recurrent neural networks (RNNs), and LSTM networks in particular, have demonstrated strong capability in modeling the time-varying and nonlinear characteristics of PAs.

Although LSTMs have been successfully applied to DPD, most studies focus primarily on performance comparisons against classical models or other neural network architectures. The internal structure of the LSTM itself is rarely examined in detail within the context of PA modeling. Standard LSTM units consist of several gating mechanisms that regulate how information is stored, updated, and passed through time. Understanding whether all gates are equally necessary, or whether certain components can be simplified without sacrificing performance, is an important question—especially for hardware-oriented or low-power DPD deployments.

In this work, we explore the behavior of LSTM structures at the gate level. By modifying specific gates and constructing several LSTM variants, we investigate how each mechanism contributes to learning PA characteristics. This analysis allows us to better understand the internal dynamics of LSTMs when applied to DPD and provides guidance for designing more efficient recurrent models.

The main contributions of this paper are summarized as follows:

- We examine LSTM behavior from a gate-level perspective and analyze the contribution of individual gates to PA behavioral modeling.
- Several LSTM variants with reduced or modified gating structures are developed to investigate the necessity and influence of different gates.
- A unified evaluation framework is established to compare the modeling capability, generalization behavior, and parameter efficiency of the proposed models.
- The study provides practical insights that may benefit the design of lightweight recurrent structures for real-time or hardware-oriented DPD implementations.

The rest of the paper is organized as follows. Section II reviews related work on neural-network-based DPD and gated recurrent architectures. Section III introduces the proposed LSTM variants and analysis framework. Section IV describes the experimental setup. Section V presents the results and discussion. Section VI concludes the paper and outlines potential future work.

II. Related Work

A. Neural Networks for PA Behavioral Modeling

Review of traditional and deep learning DPD models.

B. Gated Recurrent Architectures in DPD

Overview of LSTM, GRU, and RRU structures; discussion of prior work.

III. Proposed Methodology

A. Problem Formulation

Define the DPD modeling objective and data representation.

B. Model Variants Design

Describe different LSTM gate modifications: NFG, NIG, NIAF, CIFG, etc.

C. Evaluation Framework

Metrics: NMSE, EVM, ACLR, N_PARAM; training setup and comparison design.

IV. Experimental Setup

Hardware, dataset preprocessing, hyperparameters, and configuration.

V. Results and Analysis

A. Performance Comparison

Quantitative results for all models.

In this experiment, we compared LSTM - CIFG models with different activation functions (ReLU, GeLU, and H) in the digital predistortion (DPD) task.

1) Training and Validation Loss: Training loss (`TRAIN_LOSS`) and validation loss (`VAL_LOSS`) are vital for assessing model learning. At the start (`EPOCH` = 0), LSTM_CIFG_GeLu had a training loss of 0.54784819 and a validation loss of 0.43045321. In contrast, LSTM_CIFG_Relu had a training loss of 0.53420139 and a validation loss of 0.4108181.

As training advanced, LSTM_CIFG_GeLu's losses decreased more rapidly. By the 4th epoch, its training loss dropped to 0.08977367, while LSTM_CIFG_Relu's was 0.01886302. This shows LSTM_CIFG_GeLu's better learning efficiency.

LSTM_CIFG_H's losses fell between the other two in some epochs. At `EPOCH` = 0, its training time of 0.44472913 (versus 0.47732352 for LSTM_CIFG_GeLu and 0.45121922 for LSTM_CIFG_Relu) suggests activation functions impact computational efficiency. Overall, LSTM_CIFG_H's convergence was weaker than LSTM_CIFG_GeLu's.

2) Generalization Ability Metrics: We used validation normalized mean - square error (`VAL_NMSE`), validation error vector magnitude (`VAL_EVM`), and validation average adjacent channel leakage ratio (`VAL_ACLR_AVG`) to evaluate generalization.

LSTM_CIFG_GeLu usually outperformed LSTM_CIFG_Relu. In the 3rd epoch, LSTM_CIFG_GeLu's `VAL_NMSE` was - 4.873387 and `VAL_EVM` was - 5.04421603, while LSTM_CIFG_Relu's were - 6.554629 and - 6.77155761 respectively. LSTM_CIFG_H also lagged behind LSTM_CIFG_GeLu in these metrics.

In conclusion, LSTM_CIFG_GeLu showed better learning and generalization, with LSTM_CIFG_Relu and LSTM_CIFG_H having relatively weaker performance.

B. Performance Comparison

Quantitative results for all models.

In our study, we carried out an in - depth comparison among four models: GRU, GRU_NEW, JANET, and RRU. Each of these models was configured with a hidden size of 15 and a frame length of 200, yet they differed in their parameter counts, which could potentially influence their performance, computational cost, and generalization capabilities.

1) Training and Validation Loss Analysis: Training loss (`TRAIN_LOSS`) and validation loss (`VAL_LOSS`) are fundamental metrics for evaluating the learning proficiency of the models.

At the start of training (`EPOCH` = 0), clear disparities in training loss values were observable. The GRU model registered a `TRAIN_LOSS` of 0.44363407. In sharp contrast, the GRU_NEW model exhibited a significantly lower value of 0.2252103. The JANET model's training loss stood at 0.44467418, relatively close to that of the GRU model. This early divergence in training loss values implies that the GRU_NEW model might possess a more favorable initial configuration or a more efficient learning mechanism right from the start of training.

As the training epochs advanced, these differences became even more pronounced. By the 4th epoch, the GRU_NEW model demonstrated an impressive decline in its training loss, reaching 0.00509711. This was substantially lower than the 0.02428841 recorded by the GRU model and the 0.26742271 of the JANET model. Such a rapid decrease in the training loss of the GRU_NEW model indicates its strong ability to adapt to the training data and minimize the error between its predictions and the actual labels.

Regarding validation loss, at `EPOCH` = 0, the GRU model had a `VAL_LOSS` of 0.2803984, the GRU_NEW model had 0.15722473, and the JANET model had 0.36345455. The lower validation loss of the GRU_NEW model in the early stage was a positive sign, suggesting that it was not only performing well on the training data but also had a good start in generalizing to unseen data. Throughout the training process, the GRU_NEW model maintained a relatively low validation loss compared to the other models, highlighting its superior generalization potential.

2) Generalization Ability Metrics Evaluation: To comprehensively assess the generalization ability of these models, we examined several key metrics, namely the validation normalized mean - square error (`VAL_NMSE`), validation error vector magnitude (`VAL_EVM`), and validation average adjacent channel leakage ratio (`VAL_ACLR_AVG`).

C. Performance Comparison

Quantitative results for all models.

In our study, we carried out an in - depth comparison among four models: GRU, GRU_NEW, JANET, and RRU. Each of these models was configured with a hidden size of 15 and a frame length of 200, yet they differed in their parameter counts, which could potentially influence their performance, computational cost, and generalization capabilities.

1) Training and Validation Loss Analysis: Training loss (TRAIN_LOSS) and validation loss (VAL_LOSS) are fundamental metrics for evaluating the learning proficiency of the models.

At the start of training (EPOCH = 0), clear disparities in training loss values were observable. The GRU model registered a TRAIN_LOSS of 0.44363407. In sharp contrast, the GRU_NEW model exhibited a significantly lower value of 0.2252103. The JANET model's training loss stood at 0.44467418, relatively close to that of the GRU model. This early divergence in training loss values implies that the GRU_NEW model might possess a more favorable initial configuration or a more efficient learning mechanism right from the start of training.

As the training epochs advanced, these differences became even more pronounced. By the 4th epoch, the GRU_NEW model demonstrated an impressive decline in its training loss, reaching 0.00509711. This was substantially lower than the 0.02428841 recorded by the GRU model and the 0.26742271 of the JANET model. Such a rapid decrease in the training loss of the GRU_NEW model indicates its strong ability to adapt to the training data and minimize the error between its predictions and the actual labels.

Regarding validation loss, at EPOCH = 0, the GRU model had a VAL_LOSS of 0.2803984, the GRU_NEW model had 0.15722473, and the JANET model had 0.36345455. The lower validation loss of the GRU_NEW model in the early stage was a positive sign, suggesting that it was not only performing well on the training data but also had a good start in generalizing to unseen data. Throughout the training process, the GRU_NEW model maintained a relatively low validation loss compared to the other models, highlighting its superior generalization potential.

2) Generalization Ability Metrics Evaluation: To comprehensively assess the generalization ability of these models, we examined several key metrics, namely the validation normalized mean - square error (VAL_NMSE), validation error vector magnitude (VAL_EVM), and validation average adjacent channel leakage ratio (VAL_ACLR_AVG).

For VAL_NMSE, at EPOCH = 0, the GRU model had a value of - 1.4295063, the GRU_NEW model had - 3.9386165, and the JANET model had - 0.3064511. A more negative VAL_NMSE value generally indicates better performance. Thus, the GRU_NEW model started with an advantage. As training progressed, the GRU_NEW model

VI. Discussion

Interpretation of results, implications, and relation to prior work.

VII. Conclusion and Future Work

Summary of findings and future research directions.

References