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*—This paper investigates the role of individual gate components in LSTM architectures for digital predistortion (DPD) of power amplifiers...**

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

I. Introduction

- Background of DPD and its importance in PA linearization.
- Motivation for using LSTM-based neural networks.
- Research gap: lack of gate-level analysis.
- Contributions and paper organization.

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) to investigate their impact on the model's performance in the digital predistortion (DPD) task.

1) Training and Validation Loss: Training loss (`TRAIN_LOSS`) and validation loss (`VAL_LOSS`) are key indicators for evaluating the learning ability of the models.

At the start of training (`EPOCH` = 0), the LSTM_CIFG_GeLu model had a training loss of 0.54784819, while the LSTM_CIFG_Relu model had a training loss of 0.53420139. The validation loss of the LSTM_CIFG_GeLu model was 0.43045321, and that of the LSTM_CIFG_Relu model was 0.4108181. As the training progressed, the LSTM_CIFG_GeLu model showed a faster - decreasing trend in both training and validation losses. By the 4th epoch, the training loss of the LSTM_CIFG_GeLu model dropped to 0.08977367, whereas the LSTM_CIFG_Relu model's training loss was 0.01886302. This indicates that the LSTM_CIFG_GeLu model could fit the training data more efficiently during the learning process, and the LSTM_CIFG_Relu model's learning effect was relatively inferior.

The LSTM_CIFG_H model had different performance in terms of training and validation losses. In some epochs, its losses were between those of the LSTM_CIFG_GeLu and LSTM_CIFG_Relu models. For example, at `EPOCH` = 0, its training time was 0.44472913 (compared to 0.47732352 for LSTM_CIFG_GeLu and 0.45121922 for LSTM_CIFG_Relu), which might imply that the activation function affected the computational efficiency. In the long - term training process, the LSTM_CIFG_H model's final convergence effect was not as good as that of the LSTM_CIFG_GeLu model. The training loss of the LSTM_CIFG_GeLu model continued to decline and remained at a low level in the later training stage, while the decline range of the LSTM_CIFG_H model's loss was relatively small, suggesting that its learning effect was also weaker.

2) Generalization Ability Metrics: We evaluated the generalization ability of these models through several metrics, including validation normalized mean - square error (`VAL_NMSE`), validation error vector magnitude (`VAL_EVM`), and validation average adjacent channel leakage ratio (`VAL_ACLR_AVG`).

During the entire training process, the LSTM_CIFG_GeLu model generally had better values for these generalization - related metrics compared to the LSTM_CIFG_Relu model. For instance, in the 3rd epoch

In addition to the LSTM - CIFG models with different activation functions, we also conducted an in - depth comparison among four other 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.

3) Training Loss and Validation Loss Analysis: The training loss (`TRAIN_LOSS`) and validation loss (`VAL_LOSS`) serve as crucial metrics for assessing the learning proficiency of the models. At the very onset of the training process, specifically when `EPOCH` = 0, distinct disparities in the training loss values were observable. The GRU model registered a `TRAIN_LOSS` of 0.44363407. In contrast, the GRU_NEW model exhibited a significantly lower value of 0.2252103. The JANET model's training loss stood at 0.44467418, which was quite close to that of the GRU model. This early divergence in training loss values implies that the GRU_NEW model might have a more favorable initial configuration or a more efficient learning mechanism during the initial phase 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 ability to quickly adapt to the training data and minimize the error between its predictions and the actual labels. It might suggest that the architecture of the GRU_NEW model is better suited to capture the underlying patterns in the dataset or that its optimization algorithm is more effective in finding the optimal parameter values.

The validation loss trends also provided valuable insights into the models' generalization abilities. At the start (`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 an encouraging sign, as it indicated that the model 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. This consistency in low validation loss is a strong indicator of its superior generalization potential. In contrast, the JANET model showed a relatively higher validation loss, which might suggest that it is either overfitting the training data to some extent or is less capable of learning the generalizable features in the dataset. The GRU model, while having a lower validation loss than the JANET model in most epochs, still did not match the performance of the GRU_NEW model in terms of the validation loss trend.

4) 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`).

In the case of `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. Generally, a more negative `VAL_NMSE` value is indicative of better performance, as it implies a smaller normalized mean - square error between the model's predictions and the actual values on the validation set. Thus, the GRU_NEW model started off with a clear advantage in terms of this metric.

As the training progressed, the GRU_NEW model continued to outperform the other models in terms of `VAL_NMSE`. Its `VAL_NMSE` values remained more negative throughout the training epochs, suggesting that it was able to make more accurate predictions on the validation data. The GRU model showed a gradual improvement in its `VAL_NMSE` but could not reach the same level of performance as the GRU_NEW model. The JANET model, on the other hand, had relatively less negative `VAL_NMSE` values, indicating that its predictions on the validation set were less accurate compared to the other two models.

The `VAL_EVM` metric also painted a similar picture. The GRU_NEW model typically had more negative `VAL_EVM` values during the entire training process. A lower `VAL_EVM` value means that the model produced smaller error vector magnitudes on the validation set, which is a sign of more precise predictions. The GRU model's `VAL_EVM` values were higher (less negative) than those of the GRU_NEW model, while the JANET model had the highest `VAL_EVM` values among the three, suggesting that it had the least precise predictions in terms of the error vector magnitude on the validation data.

For the `VAL_ACLR_AVG` metric, which is particularly relevant in the context of power amplifier digital predistortion applications, the GRU_NEW model again demonstrated better performance in numerous epochs. More negative `VAL_ACLR_AVG` values signify lower adjacent channel leakage, which is a desirable characteristic. The GRU_NEW model's ability to achieve more negative `VAL_ACLR_AVG` values indicates its effectiveness in reducing adjacent channel interference, making it potentially more suitable for real - world applications where minimizing such interference is crucial.

5) Training Time and Parameter Count Considerations: The training time (`TIME:`) and parameter count (`N_PARAM`) are two additional aspects that play a vital role in determining the practicality and efficiency of these models. The GRU model had a parameter count of 2798, the GRU_NEW model had 2514, the JANET model had 2513, and the RRU model had 333. At first glance, one might expect the RRU model, with its significantly fewer parameters, to have the lowest computational cost and thus the shortest training time.

However, when looking at the training time at `EPOCH` = 0, the GRU model had a training time of 0.3867887, the GRU_NEW model had 0.40256855, the JANET model had

0.32760531, and the RRU model's data was unfortunately not fully shown in the available output. This indicates that the relationship between parameter count and training time is not as straightforward as one might assume. It is likely that the internal architecture and the complexity of the operations within each model have a significant impact on the training time.

For example, the RRU model, despite having fewer parameters, might have more complex internal computations per parameter update, resulting in a training time that is not necessarily the shortest. On the other hand, the GRU_NEW model, with a relatively lower parameter count compared to the GRU model but a slightly higher training time at the initial epoch, might have more complex optimization algorithms or more intensive data processing steps during training.

In conclusion, among these four models, the GRU_NEW model generally stood out with superior performance in terms of both learning ability, as reflected by the training and validation loss, and generalization ability, as indicated by various validation metrics. Nevertheless, the choice of the most appropriate model is not solely based on performance. Other factors such as the acceptable training time and the available computational resources also need to be carefully considered. For instance, if computational resources are extremely limited, the RRU model with its small parameter count might be a more viable option despite its potentially lower performance. Conversely, if high - accuracy predictions and good generalization are the top priorities, the GRU_NEW model would be the preferred choice.

B. Gate-Level Importance Discussion

Analysis of each gate's contribution to DPD performance.

C. Complexity and Efficiency

Trade-off between performance and parameter count.

VI. Discussion

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

VII. Conclusion and Future Work

Summary of findings and future research directions.

References