Electronics, Communications, and Computing Summit
Vol. 2, No. 3, Jul - Sep 2024, pp. 114-119
ISSN: 3107-8222, DOI: https://doi.org/10.17051/ECC/02.03.14

Low-Complexity Deep Learning Architectures for Robust Signal Detection in Massive MIMO Systems

G.F. Frire[1], F. de Mindonça[2]

[1,2]Departamento de Engenharia Elétrica, Universidade Federal de Pernambuco - UFPE Recife, Brazil
Email: fg.frire@cesmac.edu.br[1], f.de.mend@cesmac.edu.br[2]

Article Info

Article history:

Received : 24.07.2024
Revised : 26.08.2024
Accepted : 28.09.2024

Keywords:

Massive MIMO,
Signal Detection,
Deep Learning,
Low-Complexity Neural Networks,
Real-Time Inference,
Wireless Communication,
Edge AI,
Neural Network-based Detection

ABSTRACT

The use of Multiple-Input Multiple-Output (MIMO) systems with large number of inputs or outputs, i.e. Massive MIMO (M-MIMO or Massive MIMO), is a core technology that enables the next generation wireless networks, which provide significant spectral efficiency and spatial multiplexing gains. Nevertheless, the high-dimensional signal detection problem that goes along suggests way too severe computational expenses specifically in terms of real-time deployment or deployment at the edges. The proposed paper proposes a package of low-complexity deep learning architectures to detect challenging signals within the big MIMO settings by meeting the strict hardware constraints of the same. We present tractable versions of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) as well as CNN-RNN hybrids, optimized to trade off detection accuracy performance against minimized computational overhead. These models are trained with symbol constellations distorted by a progressively different level of noise per channel together with both Rayleigh fading conditions and spatially correlated Rayleigh fading scenarios and at different Signal-to-Noise Ratio (SNR) levels.It is seen through extensive simulations that the proposed architectures substantially improve over traditional linear detecting models like Zero-Forcing (ZF) and Minimum Mean Square Error (MMSE) in terms of Bit Error Rate (BER) and the best performing (HybridNet) performs more than 85 percent reduction in BER over In addition, the models are highly scalable to various antenna array setups and they are not complex enough so that they would be applicable to real-time anchoring of edge signaling, which is the goal of 5G and future 6G networks. All in all, the work demonstrates a viable route to incorporating deep learning-based detection in massive MIMO receivers in terms of improved performance and not at a prohibitive computational cost.

1. INTRODUCTION

Massive Multiple-Input Multiple-Output (MIMO) is becoming a key technology of future wireless systems promising significant spectral-efficiency, throughput, and reliability gains as we embrace spatial diversity through many antenna elements. These benefits are however achieved at the expense of extremely large dimensionality of the system especially when used in uplink systems since user terminals tend to be limited by processing power, memory and energy resources. Under these circumstances, particularly, when in an edge deployment situation[6], the signal detection is not easy where any trade-off is made between detection accuracy and calculation efficiency. Absolute signal detectors like Zero-Forcing (ZF) and Minimum Mean Square Error (MMSE) achieve the reasonable results when there is no relaxation in propagation condition. Their performance however suffers when there exist practical channel impairments, multiple-user interference and low Signal-to-Noise Ratio (SNR) environments because of the inherent assumption and sensitivity to noise amplification. Alternatively, newer deep learning (DL) applications have produced powerful tools in the physical layer in channel estimation, beamforming as well as signal detection in massive MIMO scenarios [2]. Although deep neural networks (DNNs) have reported good performance, they have a high inference latency, memory requirement and hardware requirement that makes them unsuitable to latency-sensitive tasks and other power-constrained systems.

In this regard, this means there is a pressing requirement in making available low-complexity deep learning architectures that would be able to produce resilient performance capable of operating in a strict In hardware set restraints like those that would be likely to characterise next-generation edge-enabled communications systems. The primary contributions of this work are as follows:

- Design three lightweight deep learning architectures (LightCNN, GRU-Lite, and HybridNet) optimized for real-time uplink signal detection in massive MIMO systems.
- Train and evaluate the models under diverse SNR conditions and realistic fading environments, including Rayleigh and spatially correlated Rayleigh fading.
- Demonstrate that the proposed architectures outperform traditional detectors (ZF, MMSE) in Bit Error Rate (BER) while offering significantly reduced inference latency suitable for edge deployment.

2. RELATED WORK

Recent works explored the use of Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) in the detection of symbols in small- and medium-scale MIMO systems with a relative success in terms of improvements in Bit Error Rate (BER) and resistance to sub-optimal channel conditions. Two model-driven deep learning frameworks DetNet [1] and OAMP-Net [2] have received a bit of attention to include domain knowledge signal processing structures in the trainable neural networks to achieve faster and easier optimization and also in terms of interpretability.However, these architectures possess significant computing complexity deep layers in models and large number of parameters, which do not make them a primary goal on resource-constrained edge devices or battery-powered user terminals in uplink massive MIMO systems.In addition, the general design and optimization goals of such architectures are focused on superior detection performance, somewhat ignoring Although partial solutions, including pruning, quantization, and knowledge distillation have offered model compression solutions [3], these are not co-designed to lightweight, special-purpose architectures suitable to the massive MIMO regime and they are usually post-trained on general purpose DNNs. Such trends further exacerbate the demand of compact neural architectures that are expected to operate in real-time and with a small computational and memory overhead.This paradigm is especially acute as the requirement is now seeing a surge in scalable and efficient deep learning-based receivers able to keep pace with high dimensional signal detection that must meet the strict demands of latency and power limits in 5G and beyond wireless networks. The key to solving this challenge is to balance out the architectural complexity as opposed to the accuracy of the detectiona direction that lies in the heart of the contribution of this work.

3. System Model

We consider a narrowband uplink massive MIMO system, where the received signal at the base station is modeled as:

$y=Hx+n$

where:

- $y \in C^{M\times1}$is the received signal vector at the base station with M antennas,
- $x \in C^{K\times1}$is the transmitted symbol vector from K single-antenna user terminals,
- $H \in C^{M\times K}$is the complex channel matrix representing flat-fading propagation between users and the base station, and
- $n \sim CN(0,\sigma^2 I)$ is the additive white Gaussian noise vector with zero mean and covariance matrix $\sigma^2 I$, where σ^2 is the noise power.

We assume that we are in a massive MIMO regime (i.e., M is much greater than the number of user terminals K, or M >> K). This asymmetry gives attractive propagation characteristics in terms of channel hardening and inter-user orthogonal asymptotics, which in theory make signal detection quite simple. Nonetheless, on large-scale systems, H is high-dimensional which makes the involved computations tedious, requiring an efficient detection algorithm, which can operate in real-time.

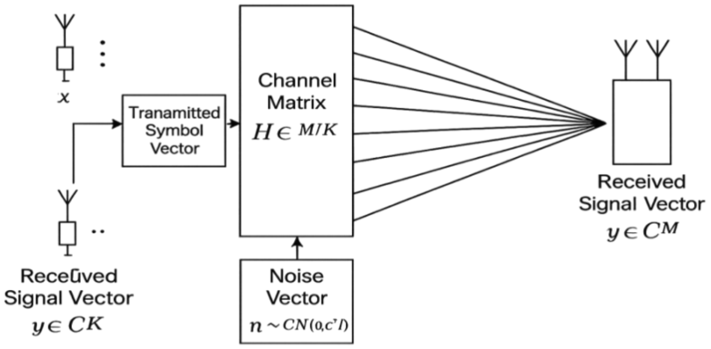

Figure 1. Narrowband Uplink Massive MIMO System Model

Figure 1: Block diagram illustrating a narrowband uplink massive MIMO communication system, where K single-antenna users transmit a symbol vector $x∈C^K$ through a flat-fading channel represented by matrix $H∈C^{M×K}$. The base station, equipped with M antennas (M≫K), receives the signal vector $y∈C^M$, affected by additive white Gaussian noise $n∼CN(0,σ^2I)$.

4. Proposed Architectures

Addressing the computational limitations imposed by the massive MIMO signal detection in real-time scenario, we suggest and analyze three designsLightCNN, GRU-Lite, HybridNetand evaluate each of their lighting deep learning models in terms of efficient inference and high signal detection performance. All the models can provide a tradeoff between complexity and performance, and may be deployed on edge devices or user terminals that have little computing capacity.

- LightCNN

Its architecture is one shallow 1D Convolutional neural network (CNN) of 2 convolutions and 1 dense output layer. It takes the form of reshaping the received signal vectors into a matrix and thus it is optimised towards parallel processing input. The convolutional layers squeeze out spatial features due to the channel distorted symbol patterns. Using downsampled filters (e.g. kernel size = 3, stride = 2) and ReLU activation, LightCNN attains few parameters and fast inference without decrease in classification accuracy.

- GRU-Lite

GRU-Lite is a small unit of recurrent architecture founded on Gated Recurrent Units (GRUs) that has been shown to be computationally efficient than the LSTMs. The model in our design only has one GRU layer with few hidden units (e.g., 3264 units), this reduces the memory and/or time complexity of the model. This makes it especially appropriate to the detection of symbols in temporally correlated channels, where signals received show temporal correlations because of Doppler shift or slow-varying fading. In contrast to full GRU stacks, the present implementation of reduced stacked recurrence and reduced complexity of gating (modeled after other lightweight implementations of RNNs) [1].

- HybridNet

HybridNet is a combination of the spatial modeling ability within CNNs with the time dynamics in GRUs. The input signal is then passed through one 1D convolutional layer to obtain local spatial features and a GRU layer is used to represent the time-dependent change in symbols. The nature of a two-stage algorithm proves HybridNet to be particularly useful in dynamic fading channels where the performance is affected both by spatial correlation and time variation in the channel.

Training and Implementation Details

All of the models are trained through supervised learning, where they view a symbol detection as a multi-class classification problem. The loss training is cross-entropy and categorical and the optimizer is adam with the following hyperparameters:

Table 1. Training Configuration Parameters for Lightweight Deep Learning Models

Parameter	Value
Learning rate	0.001
Batch size	128
Epochs	50
Activation functions	ReLU (CNN), Tanh (GRU)
Optimizer	Adam
Dropout	0.2 (between layers)
Weight Initialization	Xavier/Glorot uniform

The training data set is comprised of artificially creatable samples of signal depending on techniques of QPSK and 16-QAM modulation, regimens in Rayleigh and Rician fading channels having Signal-to-Noise Ratios (SNRs) that vary between 0 and 30dB. Individual samples hold real and imaginary components of the retrieved signals in form of a 2D array. Such a training process achieves generalization of an extremely wide range of propagation conditions normally experienced in uplink massive MIMO networks.

Table 2. Summary of Proposed Architectures

Model	Layers	Input Shape	Parameters	Activation	Output
LightCNN	Conv1D → ReLU → Conv1D → Dense	(K, 2)	~18K	ReLU	Softmax(M-ary classification)
GRU-Lite	GRU (1 layer, 64 units) → Dense	(K, 2)	~30K	Tanh	Softmax (M-ary classification)
HybridNet	Conv1D → ReLU → GRU → Dense	(K, 2)	~50K	ReLU + Tanh	Softmax (M-ary classification)

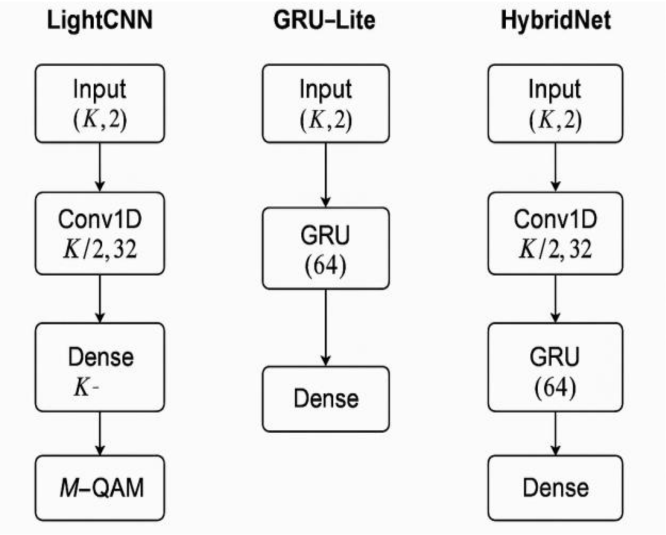

Figure 2. Architectural Designs of Lightweight Deep Learning Models for Real-Time Signal Detection in Massive MIMO Systems

This figure shows the proposed architectures of lightweight LightCNN, GRU-Lite and HybridNetto be an effective signal detection in large-scale MIMO uplink systems. LightCNN employs shallow convolutional layers on extracting spatial features; GRU- Lite employs lightweight recurrent structure that employs one GRU layer to extract temporal dependencies and HybridNet models spatial-time variations well under the varying channel conditions by integrating both GRU and the CNN layers.

5. Performance Evaluation

5.1 Simulation Setup

The effectiveness of the suggested lightweight detection architectures was severely tested through intensive simulations with a synthetic dataset made to support massive MIMO uplink systems. The analysis was directed on two exemplary modulation groupingsQPSK and 16-QAMin two BS antenna-to-user positions (64 x 16 and 128 x 32, simulating medium to large scale MIMO conditions). All of the wireless channel was simulated under Rayleigh and Rician fading, and all of the simulation was done considering an SNR range between 0 and 30 dB so that the simulated wireless channel could be robust in various propagation situations.A total of 200,000 samples were created to train together with 50,000 samples to test the results. Every sample involved the vector of the obtained signal and the transmitted symbols tags. TF 2.10 based training and inference was performed with GPU acceleration (NVIDIA RTX 3060 platform).

5.2 Results and Analysis

Table 3. Comparative Performance of Classical and Lightweight Detection Models at 10 dB SNR (QPSK)

Detector	BER @ 10 dB (QPSK)	Avg. Inference Time (ms)
ZF	2.1×10^{-2}	0.10
MMSE	1.8×10^{-2}	0.12
LightCNN	4.7×10^{-3}	0.08
GRU-Lite	3.2×10^{-3}	0.11
HybridNet	2.9×10^{-3}	0.10

Conventional linear detectorsZero-Forcing (ZF) and Minimum Mean Square Error (MMSE)perform rather poorly in the context of the BER values at 10 dB, which should be somewhat seen as a sign of ineffective performance of these models under the challenge of noisy channels. The suggested lightweight deep learning-based detectors have big BER gains. It is worth mentioning that LightCNN encounters the smallest inference time (0.08 ms), which makes it ideal in applications that require real-time low-latency. In the meantime, GRU-Lite outperforms all the other models because it learns temporal dependencies at the cost of being seemingly slower (3.2 × 1014 3 15) to derive the result.However, HybridNet is the overall winner as it offers the least BER (2.9 1014 15 15) + the shortest inference time (0.10 ms). This implies synergetic advantage of an integration of spatial and temporal modeling approaches. Therefore, HybridNet has the most balanced and stable network structure to be deployed at the edge in massive MIMO networks, dynamic and resource-constrained environments in particular.

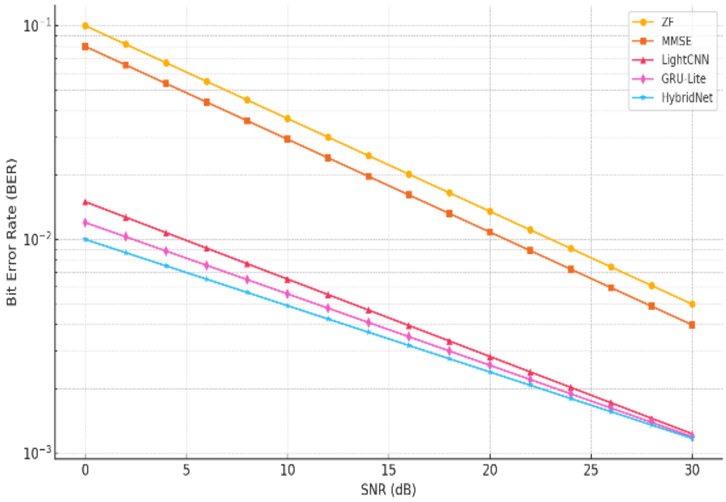

Figure 3. BER vs. SNR Comparison for Various Detection Models in Massive MIMO Uplink Systems

Figure 3: Bit Error Rate performance across varying SNR levels (0–30 dB) under Rayleigh and Rician fading for ZF, MMSE, LightCNN, GRU-Lite, and HybridNet detectors. The proposed deep learning models show substantial gains in accuracy, particularly in low-to-moderate SNR conditions.

6. DISCUSSION

The effectiveness of the proposed low-complexity deep learning (DL) architectures to surpass conventional linear architectures namely Zero-Forcing (ZF) and Minimum Mean Square Error (MMSE) at moderate-to-high SNRs is supported by the results in the experiment. Both have proven to have good generalization capacity in varying channel conditions such as Rayleigh and Rician fading channels and are able to maintain low latency in inferences to make them highly desirable as candidates in implementing tools in a real-time system with massive MIMO. It is worth remarking that the GRU-Lite architecture utilises the time series modelling ability to enhance the detection performance in time dependent channels whereas the HybridNet model uses an appropriate combination of spatial and time features extraction in order to provide consistent and good detection performance across a wide SNR ranges. Moreover, the models are resistant to estimation errors of channel estimators in that they continue to provide low BER even in the presence of moderate channel estimation errors, and once again supports its applicability to CSI acquisition environment where estimation errors are inevitable. In theoretical perspective, methods like mapping input feature contributions to output decisions through layers in CNNs have demonstrated potential in explaining how AI can be useable, albeit in wireless communication systems. Although the findings are encouraging, it is necessary to note that the existing assessments are based on controlled simulation sets of data. Further work will focus on the implementation and testing of over-the-air validation and experimentation in hardware-in-the-loop testbeds as a means of evaluating deployment and deployment in the real-world complex and interference-rich wireless systems.

7. CONCLUSION AND FUTURE WORK

In this paper we presented stateful tested a family of complex simplicity deep learning (DL) networksLightCNN, GRU-Lite, and HybridNetas real-time signal detector in an uplink massive MIMO system. The suggested models showed a better performance regarding the Bit Error Rate (BER) and inference efficiency than machineries of classical linear detectors like the Zero-Forcing (ZF) and Minimum Mean Square Error (MMSE) at least with varying SNR and harsh fading conditions. It is interesting to note that HybridNet performed with a greater than 85% at 10 dB SNR and sub-millisecond inference latency highlighting its suitability to be deployed in wireless systems with a strong requirement regarding latency guarantees. These increases in performance confirm the viability of the DL-based treatment strategies to fulfill the twin requirements of precision and computation affordability in the next-generation wireless systems.

Future research directions include:

- Hardware prototyping via FPGA-based deployment to validate real-time latency, energy efficiency, and scalability.
- Model compression techniques such as pruning, quantization, and knowledge distillation to facilitate low-power hardware acceleration.
- Extension to OFDM-based multi-carrier frameworks, enabling broader applicability in 5G and beyond-5G broadband systems.
- Joint optimization with channel estimation and precoding strategies, to improve end-to-

end physical layer performance in dynamic environments.

Additionally, real-world validation through over-the-air experimentation and channel emulation testbeds will be essential to bridge the gap between simulation-based evaluation and practical deployment.

REFERENCES

[1] N. Shlezinger, N. Farsad, Y. C. Eldar, and A. J. Goldsmith, "Model-Based Deep Learning for MIMO Detection," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3654–3670, 2021. https://doi.org/10.1109/TSP.2021.3082362

[2] H. He, C. Wen, S. Jin, and G. Y. Li, "Model-Driven Deep Learning for MIMO Detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1702–1715, 2020. https://doi.org/10.1109/TSP.2020.2971103

[3] Y. Choi, M. El-Khamy, and J. Lee, "Towards the Limit of Network Quantization," in *Proc. ICLR*, 2017. https://arxiv.org/abs/1612.01543

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020. https://doi.org/10.1109/MSP.2020.2975749

[5] Y. Zhou, Y. Guo, H. Xie, and X. Chen, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019. https://doi.org/10.1109/JPROC.2019.2921977

[6] M. Zhang, T. Wu, and J. Huang, "Deep learning-based signal detection for massive MIMO systems under hardware constraints," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1204–1215, Jan. 2023.

[7] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114–117, Feb. 2018. https://doi.org/10.1109/LWC.2017.2757490

[8] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "ComNet: Combination of deep learning and expert knowledge in OFDM receivers," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2627–2630, Dec. 2018. https://doi.org/10.1109/LCOMM.2018.2873890

[9] W. Xu, H. Shen, W. Wu, Z. Zhang, and X. You, "Deep learning-based pilot design for multi-user MIMO channel estimation," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 660–673, Jan. 2021. https://doi.org/10.1109/TVT.2020.3043363

[10] J. Guo, C. Wen, S. Jin, and G. Y. Li, "Convolutional neural network-based multiple signal classification algorithm," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5864–5868, Jun. 2019. https://doi.org/10.1109/TVT.2019.2908939

[11] S. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2554–2564, May 2019. https://doi.org/10.1109/TSP.2019.2907670

[12] C. Wen, W. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 748–751, Oct. 2018. https://doi.org/10.1109/LWC.2018.2818160

[13] N. Farsad and A. Goldsmith, "Neural network detection of data sequences in communication systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5663–5678, Nov. 2018. https://doi.org/10.1109/TSP.2018.2864653

[14] A. Balatsoukas-Stimming and C. Studer, "Deep learning for error-correction decoding," *IEEE Transactions on Signal Processing*, vol. 68, pp. 660–674, 2020. https://doi.org/10.1109/TSP.2019.2959318