Model order reduction for artificial neural networks generated from data driven state space models

Wil Schilders [*]

[*] *Eindhoven University of Technology, Eindhoven, The Netherlands (e-mail: w.h.a.schilders@tue.nl) & TU München – Institute for Advanced Study, Munich, Germany*

1. INTRODUCTION

Model Order Reduction (MOR) for Artificial Neural Networks (ANNs) is an increasingly important field that aims to reduce the complexity and computational cost of ANNs while maintaining their predictive accuracy. This is particularly relevant for scientific machine learning, where neural networks are used in complex, high-dimensional tasks such as solving partial differential equations (PDEs), modelling physical processes, or real-time simulation and control in engineering systems. In this contribution, we provide an overview of the state of the art for MOR applied to ANNs, as well as some ideas we are pursuing for our novel ANNs generated from data-informed state space systems.

2. TECHNIQUES FOR MODEL ORDER REDUCTION IN ANNS

Various techniques have been developed to apply MOR in ANNs, which can be categorized into parameter reduction, layer-wise reduction, and structural simplifications.

2.1 Pruning Methods

Pruning removes unnecessary weights, neurons, or layers from neural networks while maintaining accuracy. Key methods include: *(i)* Magnitude-based pruning, which sets small-magnitude weights to zero; *(ii)* L1/L2 regularization, promoting sparsity by penalizing weight size; *(iii)* Structured pruning, targeting entire neurons, channels, or layers for more efficient models; and *(iv)* the Lottery Ticket Hypothesis, identifying small subnetworks that achieve full model performance when trained separately.

2.2 Low-rank Factorization

Low-rank factorization reduces large weight matrices into products of smaller ones, cutting parameters. Techniques include singular value decomposition (SVD) and advanced tensor factorization methods like Tucker decomposition and tensor train. SVD approximates weight matrices with low-rank representations, while tensor methods decompose higher-order tensors in convolutional layers. These approaches, applied during or after training, are effective for compressing convolutional layers in deep convolutional neural networks (CNNs) for image processing.

2.3 Quantization

Quantization reduces the precision of weights and activations, lowering memory use and computation complexity. Techniques include post-training quantization, where trained weights are mapped to lower precision, like 8-bit integers, without retraining, and quantization-aware training, where the network is trained with quantization constraints to maintain performance. It is commonly used for deploying ANNs on resource-limited devices like mobile phones and edge devices.

2.4 Knowledge Distillation

Knowledge distillation involves training a smaller network (student) to mimic the behavior of a larger network (teacher). The smaller network is trained to replicate the outputs (or feature maps) of the larger network, allowing for substantial reductions in model size while preserving accuracy. This technique is especially popular in reducing the size of very large models (such as BERT or GPT) for practical deployment.

2.5 Neural Network Compression via SVD and PCA

Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) can be applied to the weight matrices of ANNs to reduce their dimensionality. These techniques work by identifying directions of variance in the data (or features) and projecting weights onto a lower-dimensional space, effectively reducing the number of parameters and improving computational efficiency.

2.6 Approximation via Surrogate Modeling

In scientific computing and physical simulations, reduced-order models serve as surrogates for neural networks. They approximate the network's behavior, especially in larger systems like physical process simulations or control applications. Surrogates, such as Polynomial Chaos Expansions or Gaussian Processes, are also used with ANNs to simplify input-output mappings, particularly for real-time settings.

3. MODEL ORDER REDUCTION IN THE CONTEXT OF PHYSICS-INFORMED NEURAL NETWORKS (PINNS)

PINNs represent a growing area where MOR is critically needed. PINNs embed physical laws (governed by PDEs)

DOI: 10.34726/9017

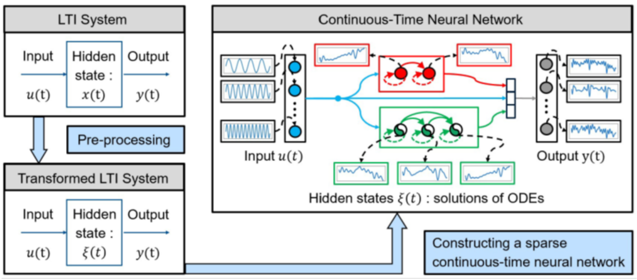

Fig. 1. Overview of our proposed workflow illustrating the systematic construction of continuous-time neural networks from Linear Time-Invariant (LTI) systems.

directly into the neural network loss function, making them suitable for solving complex physical problems like fluid dynamics, structural analysis, and electromagnetism. Key MOR approaches for PINNs include reduced basis methods and Galerkin projection. The former case involves identifying a low-dimensional subspace of the solution space, where solutions to the governing equations can be projected, thus reducing the computational load while maintaining accuracy. In the latter case, an MOR technique is used to project the dynamics of high-dimensional systems into a low-dimensional space by utilizing trial functions that satisfy the governing equations.

4. ANNS CONSTRUCTED FROM DATA-INFORMED STATE SPACE SYSTEMS

In Datar et al. (2024), we developed a systematic approach of constructing continuous-time ANNs for linear dynamical systems, based on original ideas in Meijer (1996). The idea is to first create a state-space system based on the available data, in our case using the so-called MOESP algorithm Verhaegen et al. (1992). Using a sequence of numerical methods, including QR decomposition and the Bartels-Stewart algorithm, the state-space system is transformed into an artificial neural network. The procedure is graphically illustrated in Figure 1. Special about the methodology is that horizontal layers are being formed instead of vertical layers, and that the networks are truly dynamic, i.e. non-recurrent: in the neurons, a first or second order scalar ODE needs to be solved.

This 1-1 relationship between state-space models and artificial neural networks will enable us to translate MOR methods for state-space models into MOR methods for artificial neural networks:

- First we transform the original state-space model into an equivalent ANN
- Next, we apply an arbitrary MOR method to the state-space model
- Then we translate the resulting lower-dimensional state-space model into a smaller ANN
- We then analyse how the smaller ANN can be obtained from the larger ANN, and how to formulate the corresponding MOR method for artificial neural networks.

In the talk, examples will be given of this procedure. It should be noted that the method described in Datar et al. (2024) in principle advocates the use of ANN with neuron activation functions that are special for the underlying problem. In the case described in Meijer (1996), the first and second order ODEs are, in fact, similar to so-called low and high pass filters in electronics, and all applications were also in electronics. Recently, we have also worked on n-body dynamics to predict trajectories and masses of planets, and in that case we use neurons where the 2-body system is solved (Kepler system). This approach is actually also advocated in Ferrari et al. (2014), albeit without mentioning the use of artificial neural networks.

5. CHALLENGES AND FUTURE DIRECTIONS

Despite the progress in MOR for ANNs, several challenges remain:

- **Balancing accuracy and reduction**: Maintaining the accuracy of ANNs while reducing their order is a central challenge. Often, aggressive reductions can lead to significant degradation in performance.
- **Automated MOR techniques**: There is a need for automated techniques that can determine the optimal level of reduction for a given task, without requiring extensive hyperparameter tuning.
- **Generalization of reduced models**: Reduced models often perform well on training data but may generalize poorly to unseen data. Ensuring robust generalization is critical, particularly in safety-critical applications.
- **Integration with scientific machine learning**: As scientific ML grows, integrating MOR techniques seamlessly into hybrid methods (e.g., physics-informed ML, data-driven models) will be essential.

6. CONCLUSION

Model Order Reduction is a rapidly advancing field with significant applications for reducing the computational cost and complexity of ANNs. Techniques such as pruning, quantization, low-rank factorization, and knowledge distillation have made it possible to deploy smaller and more efficient neural networks in real-time and resource-constrained environments. These developments are especially crucial for applications in scientific machine learning, where high-fidelity simulations and real-time control require efficient approximations of large models.

REFERENCES

Datar, C., Datar. A, Dietrich, F., and Schilders, W (2024). Systematic construction of continuous-time neural networks for linear dynamical systems. submitted to *SIAM Journal of Scientific Computing.*

Meijer, P.B.L. (1996). Neural Network Applications in Device and Subcircuit Modelling for Circuit Simulation. PhD thesis, TU Eindhoven.

Verhaegen, M., and Dewilde, P. (1992). Subspace model identification, part 1: The output error state space model identification class of algorithms. *International Journal of Control*, vol. 56, pp. 1187-1210.

Goncalves Ferrari, G., Boekholt, T., and Zwart, S.F.P. (2014). A Keplerian-based Hamiltonian Splitting for Gravitational N-body Simulations. *Mon. Not. R. Astron. Soc.*, 000, pp. 1-13.

DOI: 10.34726/9017