

Accelerating Neural ODEs Using Model Order Reduction

Mikko Lehtimäki, Lassi Paunonen, Marja-Leena Linne

Abstract—Embedding nonlinear dynamical systems into artificial neural networks is a powerful new formalism for machine learning. By parameterizing ordinary differential equations (ODEs) as neural network layers, these Neural ODEs are memory-efficient to train, process time-series naturally and incorporate knowledge of physical systems into deep learning models. However, the practical applications of Neural ODEs are limited due to long inference times, because the outputs of the embedded ODE layers are computed numerically with differential equation solvers that can be computationally demanding. Here we show that mathematical model order reduction methods can be used for compressing and accelerating Neural ODEs by accurately simulating the continuous nonlinear dynamics in low-dimensional subspaces. We implement our novel compression method by developing Neural ODEs that integrate the necessary subspace-projection and interpolation operations as layers of the neural network. We validate our model reduction approach by comparing it to two established acceleration methods from the literature in two classification tasks. In compressing convolutional and recurrent Neural ODE architectures, we achieve the best balance between speed and accuracy when compared to the other two acceleration methods. Based on our results, our integration of model order reduction with Neural ODEs can facilitate efficient, dynamical system-driven deep learning in resource-constrained applications.

Index Terms—Acceleration, Compression, Discrete Empirical Interpolation Method, Neural ODE, Proper Orthogonal Decomposition, Reduced Order Model

I. INTRODUCTION

As Deep Learning (DL) reaches and surpasses human performance in domain-specific applications, there is increased demand for including DL-based algorithms into internet of things and consumer appliances, such as mobile phones or wearable devices [1]. While progress in DL has relied on increasingly powerful hardware, consumer devices may contain only limited computational capacity and be battery-powered, making resource efficiency of DL-based algorithms important. An interesting new development in DL research is Artificial Neural Networks (ANNs) that employ dynamical systems, replacing traditional discrete layers with a continuous-time layer in the form of ordinary differential equations (ODEs) [2], [3], [4], [5]. In these Neural ODEs, the continuous formalism allows flexible processing of time-series and irregularly sampled data, such as medical and physiological signals [6]. Moreover, Neural ODEs are useful for resource-constrained and embedded applications because they are very memory

and parameter efficient [5]. The ODE layer also facilitates engineering physical details, such as energy conservation laws or spectral properties, into neural networks [4]. However, often a big computational bottleneck in Neural ODEs is the dynamical system layer, since propagating data through the system requires many evaluations using numerical ODE solvers. Reducing the computational cost of the ODE layer is the main motivation of our work.

In this work we show that Neural ODEs can be accelerated by compressing them using model order reduction (MOR) methods. The MOR approach is based on projecting dynamical systems onto low-dimensional subspaces. Here, we develop MOR methods that are integrated into Neural ODEs. In this manner, we lower the required storage size, memory consumption, multiply-adds and nonlinear activation count needed to compute predictions from input data. The resulting compressed Neural ODEs can be deployed for real-time computing and devices where energy efficiency is paramount. In order to validate the performance of our MOR method, we compare it to two established compression methods from the literature. Our results demonstrate that MOR is the most suitable method for accelerating Neural ODEs, because it achieves the best balance between speedup and accuracy of compressed models. We show this result in two different classification tasks that use different Neural ODE architectures, a convolutional and a recurrent neural network (RNN).

Compressing ANNs is one of the principal ways of converting high-performing trained networks into more efficient networks, since good accuracy can often be recovered without long training times, while the size of the compressed network can be chosen in a flexible manner. Several neural network compression methods have been proposed for accelerating ANNs [7], [8], [9]. These include singular value decomposition (SVD) based weight truncation [10], [11] and neuron pruning with an importance score based on zero activations [9], which we implement for comparison to our MOR method. Many prominent existing approaches rely on assigning importance scores to neurons based on their connection weights or output values. However, we argue that such methods are non-optimal for compressing Neural ODEs, as it is difficult to quantify the importance of neurons in the ODE layer based on criteria such as the output of the layer alone. This is because the final state of the nonlinear ODE system gives no information about the dynamics of the actual computations. In contrast, MOR methods are designed precisely for the approximation of state trajectories and input-output relationships in dynamical systems, hence they can overcome the shortcomings of existing compression methods

Mikko Lehtimäki and Marja-Leena Linne are with the Faculty of Medicine and Health Technology, Tampere University, Finland. Lassi Paunonen is with the Faculty of Information Technology and Communication Sciences, Tampere University, Finland.

Correspondence: mikko.lehtimaki@tuni.fi

when aiming to accelerate Neural ODEs.

Our approach to the acceleration problem draws inspiration from recent progress in a similar challenge faced in computational neuroscience. While the brain itself is deemed to be very power efficient, computational models of the brain in realistic scales are limited to simulating parts of the mouse brain, and even this may require the use of supercomputers due to the computational burden [12]. To overcome the computational resource requirements in simulations of the brain, MOR methods have been employed successfully for reducing single neuron [13], [14], synapse [15] and neuronal population [16], [17] models. Additionally, it has been suggested that biological and artificial neural computation may be intrinsically low-dimensional, which further motivates our subspace projection method [18], [19].

Our MOR approach for compressing Neural ODEs is based on the Proper Orthogonal Decomposition (POD) [20] with the Discrete Empirical Interpolation Method (DEIM) [21], a variant of the Empirical Interpolation Method [22]. Using the POD-DEIM method, we derive reduced order models (ROMs) that can be simulated efficiently in low-dimensional subspaces, even in the presence of nonlinear activation functions. In the context of Neural ODEs, our method compresses the ODE block by removing activation functions and introducing an interpolation operation in the projection subspace. In fully connected ANN layers, an entire row of the weight matrix is pruned for each removed activation function. Furthermore, columns of the weight matrix are removed by a projection operation. In convolutional layers, the reduced model computes convolutions only at selected interpolation points so that every kernel has its own set of evaluation coordinates. We integrate the subspace projection and interpolation steps of POD-DEIM into the Neural ODEs as layers, and this introduces additional operations into the neural network. However, with proper network design these steps can actually further reduce the overall number of multiply-adds in the Neural ODE. The POD-DEIM reduction is applied after training and the reduced model can be fine-tuned for increased accuracy. In summary, Neural ODEs allow us to bridge a gap between ANN research and control theory research, so that a substantial amount of previously unemployed prior knowledge in model reduction can be utilized in ANN compression and acceleration. For example, analytical optimality results and error bounds exist for our MOR algorithms [20], [23].

In Section II we review previous work in compressing and accelerating neural networks in general and Neural ODEs specifically. In Section III we present our proposed model order reduction approach and show how to formulate it in the context of continuous neural networks. In the same section we introduce two established acceleration methods that we compare our method to in benchmark problems. In Section IV we provide theoretical compression ratios for the chosen methods and show actual accuracy and wall-time metrics of compressed Neural ODEs in two different classification tasks, one using a convolutional and the other using a recurrent ODE architecture. We discuss the success of the model reduction approach and the significance of our work in Section V with future suggestions and conclude in Section VI.

II. RELATED WORK

Several studies have reported that ANNs contain structures and redundancies that can be exploited for reducing memory and power usage [7]. Here we focus on structural acceleration approaches that modify trained networks to achieve a more efficient architecture, and leave efforts such as low precision algebra, quantization of weights [24], binarization [25], hashing [26], vectorization [27] and frequency space evaluation [28] out of the review since those are complementary to the approaches presented here. Moreover, several hardware accelerators have been proposed (for example [29]), and those will not be addressed here. Furthermore, computational bottlenecks in ANNs have also been addressed by first reducing data dimension and then training simpler models. These techniques range from feature engineering to data dimensionality reduction. However, in this work our focus is on compressing trained *networks*, and hence *data* compression is considered complementary to our approach. Structural compression methods can be further grouped into several categories.

a) Pruning weights: Prior work has addressed weight pruning [30], [31] and enforcing weight sparsity [32], [33] during training to obtain weight matrices that require less storage space and memory than dense weight matrices (curiously, pruning weights is an analogy to synapse pruning in the brain, where). Several methods to evaluate weight importance have been proposed, see for example a recent review addressing 81 pruning studies [7] and compares the achieved compression rates and speedups for several ANN models. It is common to include pruning in the training loop, because altering weights after training leads to accuracy loss. In order to achieve significant acceleration with weight-pruning methods the use of special software, masking strategies, sparse algebra or hardware accelerators is recommended [34], [35].

b) Decomposition: Decompositions and low-rank matrix factorizations have been used for compress network weights so that linear operations in a layer are computed efficiently [36], [10], [37], [38], [11]. Decomposition is based on the observation that weight matrices, especially as their size increases, are seldom full rank. The idea is that a fully connected weight matrix or a tensorized convolutional kernel can be decomposed into compact low-rank matrices that require less storage space, memory and multiply-adds during training and testing. An approximation of the original operation is then obtained as a product of the low-rank tensors. In addition to compressing linear operations, prior work has also addressed decomposition by taking the nonlinear response into account [39]. The cost of decomposition approaches is that a single layer is replaced with multiple smaller ones, which trades parallelism for forward operations, canceling out some of the obtained acceleration.

c) Pruning neurons: Weight pruning and decomposition approaches maintain the structure of the compressed network in that the number of nonlinear activation functions is not changed. Eliminating activation functions leads to acceleration as an entire row of weights from a fully connected layer can be removed [8], [40]. As neuron pruning changes the number of neurons in a layer, the next layer must take this

into account, for example by deleting columns (inputs) from the weight matrix in the case of fully connected layers. Alternatively, interpolation can be used to approximate the original response, possibly introducing additional computation. In fully connected architectures large compression rates in storage space and memory are obtained when rows or columns are deleted from fully connected layers. Importance scores, such as percent of zero activations have been developed for identifying prunable neurons [9]. Pruning and quantization have been combined with Huffman coding into a compression framework that delivers memory and energy efficiency [41]. An optimization approach has been used to enforce sparse columns in a fully connected layer so that the corresponding input neurons can be pruned [40].

d) Pruning filters: Convolutional Neural Networks (CNNs) have attracted a lot of attention in the compression literature. This is not surprising given their good rate of success in real-life tasks and their high computational cost. Convolutional operations make the bulk of modern image and video processing networks and are used in many other applications, such as sequence processing. It is possible to approach CNN compression by analysing either the convolutional filters or feature maps obtained by applying the filters on input data [42], [43], [44]. The filters can be compressed with decomposition methods, similarly to fully connected layers. Alternatively, entire kernels can be pruned from filters. Pruning kernels is very effective as intermediate feature maps are eliminated. The number of methods and criteria available for identifying pruning targets in CNNs is very high [44]. In convolutional networks, neuron pruning corresponds to skipping a kernel evaluation at a single spatial location. However, such an operation is rarely supported by deep learning frameworks, and pruning in convolutional networks has focused on eliminating parts of or entire filters or feature maps.

In the context of Neural ODEs, a few studies have addressed improving inference and training times. One approach focuses on learning simple dynamics that do not burden ODE solvers as much as stiff systems [45], [46], [47]. Additionally, training times have been reduced by further improving the adjoint method, for example by relaxing error criteria [48] while inference times have been accelerated by introducing hypersolvers - neural networks that solve Neural ODEs [49]. These methods are complementary to ours, since we aim at compressing the learned architecture with MOR methods. Our approach is comparable to pruning neurons based on an importance score [8], [9], so that a number of rows from the weight matrix and nonlinear activation functions can be removed altogether. The POD-DEIM method improves on the existing methods of determining neuron importance, as it accounts for the complete dynamics of the ODE block. Additionally, POD-DEIM can be combined with many of the prior methods such as quantization of weights.

III. METHODS

In this section, we present the theory of the continuous interpretation of ANNs. We then describe our model order reduction method and show how to formulate it in the ANN

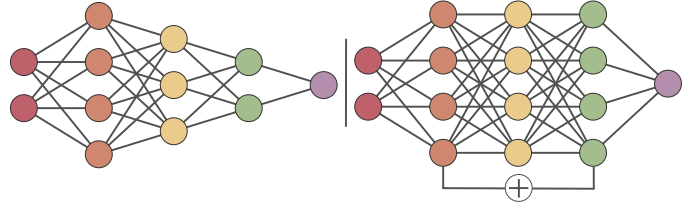


Fig. 1. A plain discrete feedforward network on the left and a residual network on the right. The defining feature of residual networks is that the output from an earlier layer skips layers and is added directly to a later layer.

and Neural ODE context. Finally, we present two other ANN acceleration methods from the literature, which we use as benchmarks to our method, and their implementation in the Neural ODE setting.

A. Continuous networks

Since the success of Residual Neural Networks (ResNets) [50], a continuous interpretation of neural networks has gained traction. The ResNet architecture utilizes skip connections to deal with the problem of vanishing gradients, allowing the training of extremely deep neural networks. In a ResNet block, the output is the sum of the usual feedforward operations on the data and the unprocessed data entering the block. The ResNet architecture is illustrated in Fig. 1 on the right, with a plain feedforward network on the left. If the hidden layers of the plain network implement a nonlinear function $x_{k+1} = f(x_k)$, the skip connection of the ResNet implement $x_{k+1} = x_k + f(x_k)$. By introducing a constant $h = 1$ to obtain $x_{k+1} = x_k + hf(x_k)$, the resemblance of the skip connection to Euler's formula for solving Ordinary Differential Equations (ODEs) is seen [2], [4], [3]. This has lead to the continuous dynamical system interpretation of ANNs.

In the continuous interpretation of ANNs, a set of layers is replaced by one layer that parameterizes a dynamical system as a system of ODEs with solution $x(t)$. We assume this ODE system has the form

$$x'(t) = f(Ax(t) + b) + Zu(t), \quad (1)$$

where $x(t) \in \mathbb{R}^n$ is the state of the ODE at time t , $A \in \mathbb{R}^{n \times n}$ is a weight matrix, $b \in \mathbb{R}^n$ is a bias vector, $u(t) \in \mathbb{R}^i$ is a time-dependent input, $Z \in \mathbb{R}^{n \times i}$ is an input matrix and $f: \mathbb{R}^n \mapsto \mathbb{R}^n$ is a vector-valued function $f(\chi) = [f_a(\chi_1), f_a(\chi_2), \dots, f_a(\chi_n)]^T$ where the activation function $f_a(\chi_i)$ is commonly the hyperbolic tangent. Here, A, b, Z are learned parameters of the ODE. In feed-forward networks there are typically no time-dependent inputs and hence $Z = 0$. On the other hand, in RNNs where $x'(t)$ is the hidden state, the input data enters the system through $Z \neq 0$. The output of the layer is $x(T_{end})$, which is the state of the dynamical system at the user specified final time T_{end} . Initial values of the ODE system can be obtained in two ways, either as the output of the layer preceding the ODE or set explicitly. The former is more common in feed-forward architectures and the latter in RNNs that receive time-

dependent input data. The output $x(T_{end})$ of the ODE layer is computed using numerical methods, taking discrete or adaptive steps with an ODE solver. Neural ODEs use the adjoint method of calculating gradients [5] and have enabled parameterizing ODEs as several different ANN operations, or chains of them. The primary restriction is that the output size of the ODE layer must match the input-size of the layer. Overall, it is possible to train a wide variety of ANN architectures for different tasks as continuous networks [5], [6].

Training Neural ODEs with the **adjoint method** is memory efficient compared to deep discrete networks [5], as the backpropagation algorithm does not need to store intermediate ODE states to calculate gradients on the loss function. However, the use of the adjoint method is not required for training ODE networks. Parameterizing an ODE in place of several discrete layers may also lead to parameter efficiency and correspondingly require less storage space and memory, since the continuous layer can replace several independent discrete layers. Other benefits of Neural ODEs include allowing the use of adaptive ODE solvers for speed versus accuracy tuning, and enabling ANNs to flexibly process continuous and irregularly sampled time-series data [5], [6]. Neural ODEs have also improved on **existing ANN-based density estimation models** [5]. However, the dynamics learned by Neural ODEs may be unnecessarily complex [47], [46] and result in stiff systems [51]. Often the ODE block is the most computationally demanding part of the network, since many numerical evaluations of the state of the ODE are needed to obtain the output. Hence, Neural ODEs require more time to evaluate data, both in training and testing, compared to traditional discrete networks [46].

It is possible to parameterize the ODE block so that the model in training has favourable properties that facilitate learning. **Antisymmetric networks** are a step toward this direction, since they guarantee that the state $x(t)$ of the ODE system does not diverge far from or decay to the origin as $t \mapsto \infty$ [4]. This prevents the gradient of the loss function from vanishing or exploding and makes training the network well-posed. Antisymmetric networks use an antisymmetric weight matrix $A = W - W^T$ that by definition has eigenvalues λ_i so that for all i , $\text{Re}(\lambda_i(A)) = 0$. Furthermore, a small shift of the eigenvalues by γ may be applied so that $\text{Re}(\lambda_i(A - \gamma I)) = -\gamma < 0$, which improves behavior of the ODE system in the presence of noisy data [4]. Notice that in practical applications t is finite, hence a small γ will not make the gradients of the loss function vanish. In [52] it is demonstrated that the shifted antisymmetric weight matrix $A - \gamma I$ gives the hidden state of recurrent neural networks a favourable property of long-term dependence on the inputs to the system, which helps classifying data with temporal relationships. In this work, we implement an ODE-RNN as

$$\begin{aligned} A &= W - W^T - \gamma I, \\ x'(t) &= \tanh(Ax(t) + b) + Zu(t) \end{aligned} \quad (2)$$

where the weight matrix A gives the network the desired properties that facilitate learning and during training the parameters W are learned. Additionally, in Equation (2) Z is the input weight matrix, b is a bias vector, $x(t)$ is the hidden state of the

RNN and $u(t)$ is the time-dependent input to the system. In Section IV, we will demonstrate compression and acceleration of this ODE-RNN. Such networks make an interesting model compression target since their architecture gives the system temporal memory capacity that the compressed model must retain.

B. Model Order Reduction

A key contribution of our work is the formalization of MOR in the context of Neural ODEs and the realization that the necessary online operations can be expressed as layers of ANNs. A powerful method for MOR of general nonlinear systems is Proper Orthogonal Decomposition (POD) [20] coupled with the Discrete Empirical Interpolation Method (DEIM) [21], [22], developed in the fields of systems and control theory. POD-DEIM has been used for reducing models in computational neuroscience [13], [16], where architectures and nonlinearities resemble those found in ANNs. In order to compress Neural ODEs, we construct reduced order models (ROMs) of the ODE block in trained Neural ODEs with the POD-DEIM method. POD approximates the original system of dimension n in a low-dimensional subspace via projection. A k -dimensional POD basis with orthonormal column vectors $V_k \in \mathbb{R}^{n \times k}$, where $k < n$, is computed using Singular Value Decomposition (SVD) of snapshots $V \Sigma V^T = \text{SVD}(Y)$ [53]. V_k is then the first k left singular vectors of the snapshot matrix, equaling the first k columns of V . The snapshot vectors $Y = [y_1, y_2, \dots, y_s]$ are values of the solution to the original system and they are collected using numerical simulation of the ODE block. ANNs provide a very natural setting for gathering snapshots, since we have access to training data that can be propagated through the trained network and the ODE block of Neural ODEs. However, it is important that the snapshots are collected after the model is trained, so that the snapshots reflect true dynamics of the ODE layer.

A reduced state vector $V_k^T x(t) = \tilde{x}(t) \in \mathbb{R}^k$ is obtained by a linear transformation and at any time t an approximation of the original, full-dimensional state vector can be computed with $x(t) \approx V_k \tilde{x}(t)$. Projecting the system in Equation (1) onto V_k by Galerkin projection results in a reduced system

$$\tilde{x}'(t) = V_k^T f(AV_k \tilde{x}(t) + b) + V_k^T Z u(t), \quad (3)$$

where we can precompute the transformation of the weight and input matrices into matrices $\tilde{A} = AV_k$, $\tilde{Z} = V_k^T Z$ that are used in place of the original operations in the online (prediction) stage. The reduced POD model approximates the original system optimally in the sense that the POD subspace contains the largest possible amount of energy [20]. However, the nonlinear form of the equation prevents computational savings as the number of neurons has not been reduced. The size of A is still $n \times k$ and $f(\cdot)$ is computed in the original dimension n . This is known as the lifting bottleneck in reducing nonlinear models.

Efficient evaluation of the nonlinear term can be achieved with DEIM [22], [21]. DEIM extends the subspace projection approach with an interpolation step for nonlinear functions. To

Algorithm 1 Discrete Empirical Interpolation Method**INPUT:** $\{u_l\}_{l=1}^m \subset \mathbb{R}^n$ linearly independent**OUTPUT:** $\vec{p} = [p_1, \dots, p_m], P \in \mathbb{R}^{n \times m}$

```

1:  $p_1 = \operatorname{argmax}(|u_1|)$ 
2:  $U = [u_1], P = [e_{p_1}], \vec{p} = [p_1]$ 
3: for  $l = 2$  to  $m$  do
4:   solve  $(P^T U)c = P^T u_l$  for  $c$ 
5:    $p_l = \operatorname{argmax}(|u_l - U c|)$ 
6:    $U \leftarrow [U \ u_l], P \leftarrow [P \ e_{p_l}], \vec{p} \leftarrow [\vec{p} \ p_l]$ 
7: end for

```

construct an interpolated approximation of the nonlinear term, we use

$$\tilde{f}(x, t) \approx U_m (P_m^T U_m)^{-1} P_m^T f(x, t), \quad (4)$$

where the DEIM basis vectors $U_m = [u_1, u_2, \dots, u_m]$, $m < n$ are the first m left singular vectors of the snapshot matrix of nonlinear vector-valued function outputs $F = [f(x_1, t_1), f(x_2, t_2), \dots, f(x_s, t_s)]$ computed via SVD, $P_m^T f(x, t) := f_m(x, t)$ is a nonlinear function with m components chosen from f according to DEIM determined interpolation points $\vec{p} = p_1, \dots, p_m$ and $P_m = [e_{p_1}, e_{p_2}, \dots, e_{p_m}]$ with e_{p_i} being the standard basis vector i of \mathbb{R}^n . Together POD and DEIM form a ROM

$$\tilde{x}'(t) = \underbrace{V_k^T U_m (P_m^T U_m)^{-1}}_N f_m(\tilde{A} \tilde{x}(t) + b) + \tilde{Z} u(t), \quad (5)$$

where $N \in \mathbb{R}^{k \times m}$ can be precomputed offline. Now in the online prediction phase only m nonlinear functions are evaluated. Due to the structure of $f_m : \mathbb{R}^m \mapsto \mathbb{R}^m$ we can then further compress \tilde{A} so that only m rows at indices \vec{p} remain. Correspondingly, we only select m elements from the bias vector at indices \vec{p} . Thus we have obtained

$$\tilde{x}'(t) = N f_m(\tilde{A}_m \tilde{x}(t) + b_m) + \tilde{Z} u(t) \quad (6)$$

where $\tilde{A}_m \in \mathbb{R}^{m \times k}$ and $b_m \in \mathbb{R}^m$. Note that the dimension k of the projection subspace does not need to equal the dimension m of the nonlinear component, although in practice it is common to choose $m = k$. This also means that the number of activation functions can be controlled to individual neuron level and independently of the size of the POD subspace where forward propagations are computed.

Algorithm 1 shows the process of determining the DEIM interpolation points and matrix P . The ordered linearly independent basis vector set given as input is the basis U_m that is obtained from snapshots of the nonlinear function. The argmax -function returns the index of the largest value in a given vector and e_i is the i :th standard basis vector. For more details see the original reference [21]. Therein an error bound for the DEIM approximation in Euclidean space is also given. For an error estimate of the reduced state in POD-DEIM models specifically, see [23].

An essential part of our work is building the POD-DEIM online operations $V_k^T x$, $V_k \tilde{x}$ and Equation (6) into the Neural

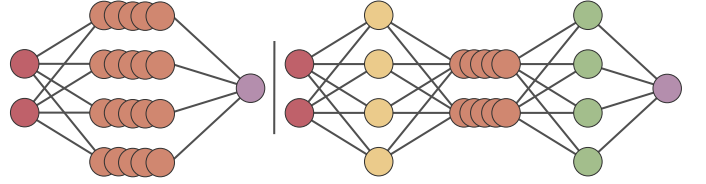


Fig. 2. A Neural ODE on the left with the continuous block indicated in orange color. A POD-DEIM reduced network on the right illustrates that the ODE block is evaluated in a low-dimensional subspace, with linear transformations around the ODE block. The networks have equal inputs and approximately equal outputs.

ODE. Fig. 2 illustrates the compressed Neural ODE. Consider a network with an input layer, an ODE block, and a readout layer, as on the left of Fig. 2. We add two linear operators, one for projecting the ODE block into a low-dimensional subspace and one for transforming the result of the ODE block back to the original space, as seen on the right of Fig. 2. These layers implement the computations $V_k^T x$ and $V_k \tilde{x}$, respectively. The cost of these operations is offset by the cheaper evaluation of the ODE layer according to Equation (6), which typically makes the bulk of the computational cost of Neural ODEs. Furthermore, if the operations around the ODE block are linear, the transformations can be computed into the existing weight matrices, resulting in even cheaper online evaluation.

The DEIM method is known to exhibit lower accuracy in the presence of noise or perturbations in the snapshots. The input function $Zu(t)$ in ODE-RNNs can cause such non-smooth behavior and a large number of timesteps taken is sensitive to instabilities. Hence for compressing ODE-RNNs, we use an oversampling strategy that stabilizes the method as shown in [54]. In this oversampled DEIM (ODEIM), the number of nonlinear sampling points in \vec{p} is chosen higher than the dimension m of the DEIM subspace, so that the interpolation task turns into a regression problem. In ODEIM, the matrix inverse of Equation (5) is replaced with the Moore-Penrose pseudoinverse $(P^T U_m)^\dagger$. We choose $m + 1$ interpolation points, which has a slightly negative effect on the acceleration but a significant positive effect on accuracy of the compressed model.

The POD-DEIM method has tunable parameters that affect the performance of the reduced model. The dimension k of the linear projection matrix and the number of interpolation points m is set typically according to the decay of the singular values of the snapshot matrices Y, F . The snapshot collection is done separately for the full model and the nonlinear function outputs, however in the case of ANNs these are typically the same, since purely linear parts in activation functions are rare. The number of snapshots can be controlled in two ways, the number of samples of training data to use and the number of snapshots of model dynamics to save per sample. Computing the SVD of the snapshot matrix is the most computationally demanding step of the MOR pipeline, with memory being the typical bottleneck.

C. Benchmark compression methods

We compare the performance of the POD-DEIM method to two other ANN compression methods presented in the literature. To the best of our knowledge, no other studies using these methods to compress Neural ODEs have been published before.

The first method compresses the weight matrix of the ODE block into two low-rank matrices [10], [11] using SVD based truncation. This reduces the overall number of multiply-adds needed to evaluate the layer but does not change the number of activation functions. Given a fully connected layer with activation $f(Ax + b)$ and $A \in \mathbb{R}^{n \times l}$, we apply SVD to the weight matrix to obtain $V, \Sigma, \Psi^T = A$. Then, we only keep the first $k < n$ singular values by truncation so that $\Sigma_k = \text{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k \times k}$ and the first k left and right singular vectors so that $V_k = \{v_1, \dots, v_k\} \in \mathbb{R}^{n \times k}$, $\Psi_k = \{w_1, \dots, w_k\} \in \mathbb{R}^{l \times k}$. An approximation to A is then obtained as $\hat{A} = V_k \Sigma_k \Psi_k^T$. In ANNs we can implement the approximation in two consequent fully connected layers that only use activation and bias after the last layer. The connection weight of the first layer is $\Sigma_k \Psi_k^T \in \mathbb{R}^{k \times l}$ and the second weight is V_k , which changes the parameter count from nl to $k(n + l)$. Notice that in a single-layer ODE block we have $l = n$. This method is theoretically sound for compressing the ODE block. It retains the most parameters out of the compression methods we test, meaning that it is expected to yield low to moderate acceleration.

The second method we used for comparison is based on computing an importance score for neurons, called average percent of zeros (APoZ) method, following [9]. There it is suggested to prune neurons that have the largest number of zero activation values over the training or testing data set. The corresponding rows of the weight, bias and input matrices are then removed altogether. Moreover, with Neural ODEs specifically, for each removed row of the weight matrix we also remove the corresponding column. The importance score of the c :th neuron in a layer is defined as

$$\text{APoZ}_c = \frac{\sum_k^N \sum_j^M f(O_{c,j}(k) = 0)}{NM}, \quad (7)$$

where N is the number of examples used to calculate the score and M is the dimension of the output feature map O . In Neural ODEs it is common to use hyperbolic tangent activations, hence we adapt the scoring so that neurons with lowest absolute activation magnitudes are eliminated. We compute the score after the last timestep of the ODE block, as this is the value that gets propagated to the following layers in the network. We only use training data for computing the score. This method does not account for the dynamics of the ODE block, as scoring is computed only at the last timestep of the ODE block, hence we evaluate the performance in detail here. Overall, this method applies the largest amount of compression out of the methods we test here.

IV. RESULTS

We implemented two Neural ODE models and trained them on different classification tasks; one Neural ODE with a convolutional ODE block to classify digits of the MNIST dataset

TABLE I
THEORETICAL LAYER SIZES

Model	ODE weights	ODE activations	Preceding weights	Following weights
Original model	n^2	n	ni	on
POD-DEIM, best case	$2k^2$	k	ki	ok
POD-DEIM, worst case	$2k^2$	k	$n(i + k)$	$n(o + k)$
APoZ, best case	k^2	k	ki	ok
APoZ, worst case	k^2	k	ki	$n(o + 1)$
SVD	$2kn$	n	ni	on

and one ODE-RNN to classify digits in the pixel-MNIST task, where the network is given one pixel at a time as input. We trained both models using data-augmentation with random rotations and affine translations in order to prevent overfitting. Next, we compressed the models using POD-DEIM, APoZ trimming and SVD truncation and evaluated their performance directly after compression, after 3 epochs of fine-tuning and after excessive fine-tuning on training data. The performance of each compressed model was compared to the respective original model using Top 1 accuracy, Top 3 accuracy and wall-time as metrics. We restrict fine-tuning to the layers following the ODE block in order to get a better view of how well the compressed blocks maintain their computational capacity. Our results below have been implemented using the PyTorch machine learning framework [55]. To implement Neural ODEs, we additionally used the TorchDiffEq [5] and Torchdyn [56] python packages. Execution times are true wall-times realized on an Intel Xeon E5-2680 v3 2.5GHz processor and measured as the time it takes to classify every item in the test dataset.

In Table I, we show the number of parameters in a theoretical Neural ODE and after it has been compressed with each compression method. We consider a nonlinear ODE block of the form $x'(t) = f(Ax(t))$ where the weight matrix has dimensions $A \in \mathbb{R}^{n \times n}$, hence there are n activation functions. The ODE block is preceded by a layer with weight matrix $W_1 \in \mathbb{R}^{n \times i}$ and followed by a layer with weight matrix $W_2 \in \mathbb{R}^{o \times n}$. In Table I, k denotes the dimension of the compressed ODE block and $n \times i, o \times n$ are the number of weights in the layers preceding and following the ODE block, assuming those layers exist. With regards to the model speed, the size of the ODE layer is significantly more important than the sizes of the surrounding layers. This is because the ODE function is evaluated multiple times; at least once for each intermediate time point in $[T_{start}, T_{end}]$. The architecture of the original ANN affects the achieved compression. The POD-DEIM and APoZ methods can reduce the ODE layer to a dimension that is independent of the original size n . However, for POD-DEIM the best case is realized when the preceding and following operations are linear, whereas the APoZ method reaches the best case even if they are nonlinear. With the APoZ method, some nonlinear operations, such as max pooling, require that we maintain a lookup table of pooled indices or insert the k compressed values to their respective indices in an n size vector, which is the worst case result for the method. SVD truncation is not affected by the surrounding layers.

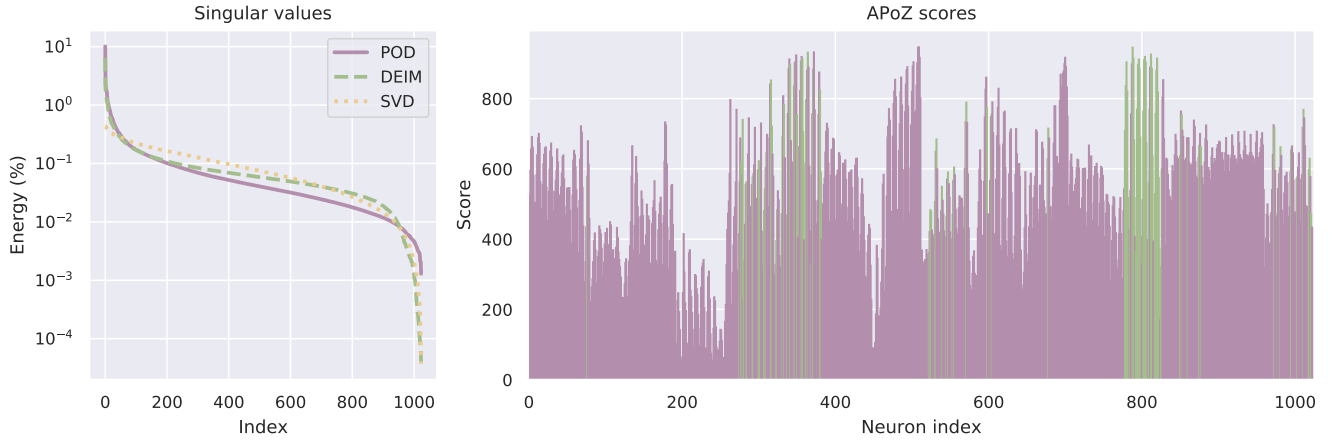


Fig. 3. Metrics for the convolutional Neural ODE. Left: Singular values POD (purple) and DEIM (green) snapshot matrices and singular values of the weight matrix of the ODE block (yellow). Right: histogram of APoZ scores, with the first 100 DEIM indices indicated in green color.

A. Convolutional Neural ODE

We illustrate the performance of reduced models on the MNIST dataset of handwritten digits [57]. For our first task we train a convolutional Neural ODE that uses 16 convolutional kernels and \tanh -activation in the ODE block. Before the ODE block we use a convolutional layer with 3×3 kernels and rectified linear unit (ReLU) activations that outputs 16 channels into 3×3 max pooling. After the ODE block we use 3×3 max pooling into a linear readout layer with ReLU activations. The ODE block propagates the data for $t = [0, 1]$ with timestep $dt = 0.1$ using a fourth order fixed step Runge-Kutta solver. We use the adjoint method for training the ODE weights [5]. In order to implement model order reduction and model compression of the convolutional ODE block, the convolution operator is needed in matrix form. After training, we replaced the original convolution operator with an equivalent operation implemented as a linear layer. The conversion yields a sparse Toeplitz matrix. With this operation, 16 convolution channels correspond to 1024 neurons and activation functions. All reported wall times are obtained with this conversion and compressed models are compared to these times. The reported wall times are medians of 10 consecutive evaluations of the entire validation data set. The trained network achieves 97.9% top-1 and 99.7% top-3 accuracies on held-out test dataset of 10 000 MNIST images and it takes 11.2 seconds to run the model on the test dataset on CPU.

In order to determine the suitability of our chosen compression methods, we looked at several metrics of the trained Neural ODEs, namely decay of singular values in the POD-DEIM snapshot matrices Y, F and the weight matrix A . Moreover, we looked at the distribution of APoZ scores and their relation to DEIM interpolation indices. Fig. 3 shows these metrics for the convolutional Neural ODE. The left plot shows sorted singular values of the solution snapshot matrix Y (POD) and F (DEIM) of the ODE block, gathered from training data, in purple and green colors, respectively. The logarithmic y-axis shows the magnitude of a given singular value divided by the sum of all singular values. Both POD

and DEIM singular values collapse rapidly, indicating that a small number of singular vectors can span a space where the majority of dynamics are found. This is an important indicator for the suitability of the POD-DEIM method. In the same plot, the yellow plot shows the singular values of the weight matrix A of the ODE block. These values are useful for determining a rank for the SVD based truncation method that we compare against POD-DEIM. Singular values of A also decay exponentially, meaning that a good low-rank approximation to A can likely be found. The right plot of Fig. 3 shows APoZ scores for each neuron in the ODE block, with low-scoring neurons being the least important. Additionally, we have indicated the first 100 interpolation points p_i chosen by the DEIM method in green color. All neurons seem to contribute to the model, with only a small number of relatively low-scoring neurons seen, which could make the APoZ method challenging to apply in the Neural ODE context. Some consensus between the methods is seen, as neurons with low APoZ scores are not in the top DEIM selections either.

We wanted to find out how the POD-DEIM method compares to existing model compression methods when compressing only the ODE block of a convolutional network. The results on the MNIST dataset are seen in Fig. 4. The top-1 accuracy is seen on the left, top-3 accuracy in the middle and factor of speedup on the right. The x-axis shows different dimensions of compressed models, and the hue indicates the compression method. The lighter stacked bars indicate results when the models are fine-tuned for 3 epochs. After these few epochs, the training loss starts decreasing very slowly. Results in Fig. 4 are relative to the performance of the original convolutional Neural ODE model. Table II and Table III present the absolute performance values of compressed models without fine-tuning and with 3 epochs of fine-tuning. Results for fine-tuning until convergence are seen in Table IV. Each column in the tables shows the best performance among all models in bold font.

For POD-DEIM, we have chosen $k = m$ and use dimension k compressed models, which corresponds in number of

activation functions to dimension k APoZ trimmed models. In Fig. 4 we can immediately see that in terms of accuracy the SVD truncation method performs the best, but yields very little acceleration. The APoZ method is the opposite, resulting in the fastest models with the least accuracy. Meanwhile, the POD-DEIM method achieves compressed models that are almost as fast as trimmed networks. The extra projection operations around the ODE block in POD-DEIM as well as the interpolation operation make the total runtime slower compared to APoZ. In terms of accuracy the POD-DEIM almost matches the SVD truncation method. It is interesting to note that the POD-DEIM method performs better without fine-tuning than the APoZ does with three epochs of fine-tuning. At dimension 500, the POD-DEIM method is two times faster to run than the original model, while achieving 97.0% of the accuracy of the original model (that is, 95.4% top-1 accuracy). At dimension 100, the POD-DEIM model is over 4 times faster than the original model, reaching 73.1% top-1 accuracy without fine-tuning. SVD truncation achieves speedup following the theoretical result in Table I, which estimates the break even with the original model at $k = 512$.

B. Recurrent Neural ODE

For our second study, we designed a task in which Neural ODEs have an advantage over discrete networks. ODE-RNNs [6] and antisymmetric RNNs [4], [52] are one such example. Using a continuous hidden state, ODE-RNNs enable flexible inference on continuous-time data even with irregularly sampled or missing values and antisymmetric models have spectral properties that are favourable for RNNs, as described in Section III. We implemented a shifted antisymmetric ODE-RNN for the pixel MNIST task. In this task, the network sees a pixel of a handwritten digit at a time and after iterating over each pixel outputs the class of the image. The length of the input sequence is 784 steps, corresponding to a flattened MNIST image, and the network sees each pixel for $dt = 0.1$ seconds. Following [52], we initialize the ODE weights from zero mean, unit variance normal distribution as well as train and test the network with the Euler discretization method. We use 512 hidden units with a hyperbolic tangent nonlinearity and the shifted antisymmetric weight matrix as shown in Equation 2. The ODE block is followed by a linear readout layer. Compared to the convolutional network, this model is smaller in neuron count but needs to incorporate the time-dependent input data. The trained network achieves 96.2% accuracy on held-out MNIST test data in 41.5 seconds.

Fig. 5 shows compression metrics for the ODE-RNN. The left plot shows sorted relative singular value magnitude of the solution snapshot matrices Y (POD) and F (DEIM) and the weight matrix A of the ODE block in purple, green and yellow, respectively. For this model, the singular values of the snapshot matrices as well as the weight matrix also decay rapidly. However, with regards to POD-DEIM, the singular values of the snapshot matrices do not decay as fast as those obtained from the convolutional model and there is a considerable amount of total energy contained in the singular values until rank 500. This indicates that the dynamics of the ODE block are more

difficult to approximate in low dimensions than those learned by the convolutional model. Such a result is expected, because the ODE-RNN has a time-dependent input contributing to the hidden state and the forward propagations are computed for a longer time span than in our earlier example, allowing for richer dynamics. On the other hand, singular values of the shifted antisymmetric weight matrix decay more rapidly than in the earlier task, indicating suitability of SVD-based truncation of weights. The APoZ scores on the right of Fig. 5 show the same pattern as earlier, where it is not straightforward to detect the most important neurons conclusively. This is expected, given that in the context of ODE blocks, the output activity at last time step is not a conclusive measure of neuron importance.

We then evaluated the selected compression methods on the pixel MNIST task using our ODE-RNN. The results relative to the performance of the original model are seen in Fig. 6. Absolute results of compressed models without fine-tuning are given in Table V, with 3 epochs of fine-tuning in Table VI and with exhaustive tuning in Table VII. This task is more challenging for the POD-DEIM and APoZ methods than the convolutional network, but the SVD truncation method maintains good accuracy. As indicated by the singular values of the snapshot matrices, the overall performance of the POD-DEIM method on this model is not as satisfying as with the convolutional model. On this smaller model the POD-DEIM method is faster than the original model until dimension 350, while the APoZ method remains faster at all times. This is also the dimension at which the POD-DEIM method still recovers satisfactory accuracy without any fine-tuning, achieving higher accuracy than the APoZ method with fine-tuning. Interestingly, at this point and at higher dimensions, exhaustive fine-tuning consistently recovers over 4% better performance from the POD-DEIM method than the APoZ method, despite us only tuning the readout layer. Considering the trade-off between run time and accuracy, POD-DEIM shows a good compromise, although with exhaustive fine-tuning the APoZ method may be a better fit for compressing the ODE-RNN to a low dimension, since APoZ yields the best absolute acceleration. At very low dimensions, the dynamics of the APoZ method also seem easier to fine-tune than the POD-DEIM reduced model, as more accuracy is recovered in 3 fine-tuning epochs. A minor amount of acceleration can also be obtained with the SVD truncation method without loss in accuracy.

V. DISCUSSION

In this work we have studied compression of artificial neural networks that contain ordinary differential equations as layers (Neural ODEs). Inspired by recent progress in computational neuroscience [17], we compressed Neural ODEs using mathematical model order reduction methods and showed how these methods can be built directly into the Neural ODE architecture. After training the networks, we evaluated their accuracies as well as execution speeds. Although in the literature many compression methods are applied in the training loop, the training of Neural ODEs is slower than discrete networks to begin with, making it attractive to compress the network in

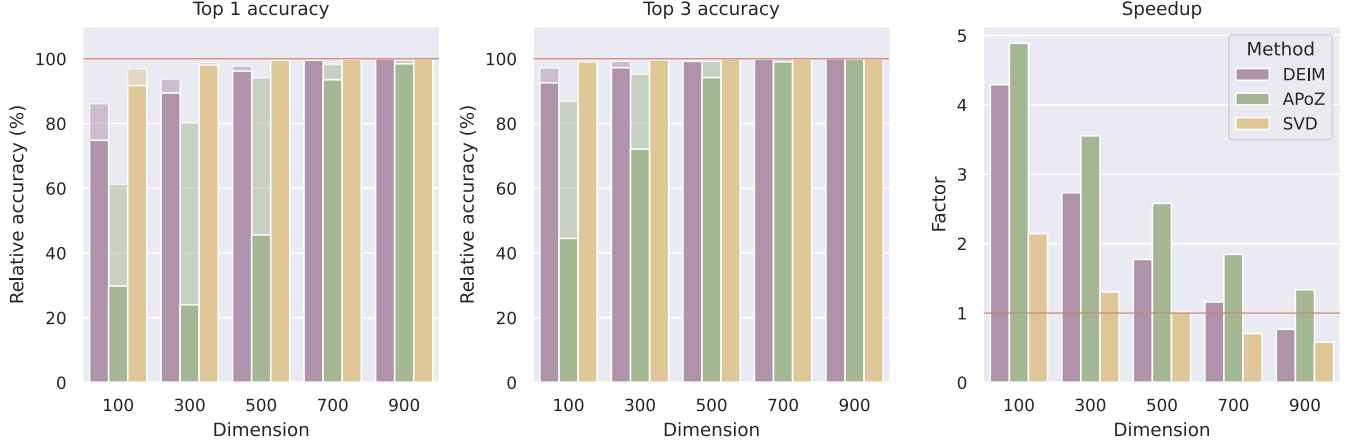


Fig. 4. Relative performance of reduced convolutional Neural ODEs on the MNIST dataset. All values are relative to those obtained from the original model, with the red horizontal line indicating equal prediction accuracy or wall-time with the original model. Lighter stacked bars indicate results for fine-tuned models. Left: Top 1 accuracies, middle: Top 3 accuracies, right: factor of speedup.

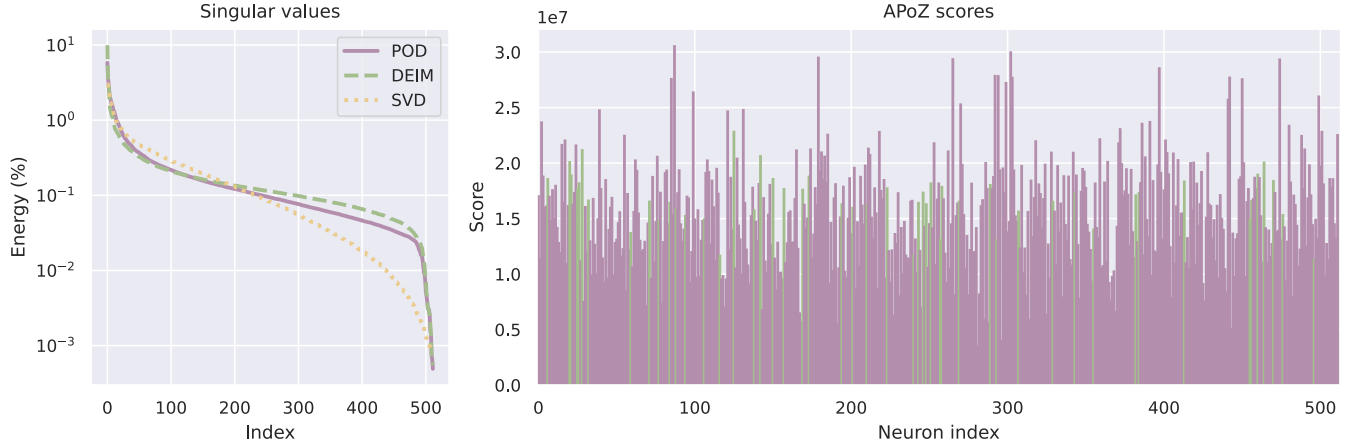


Fig. 5. Metrics of the ODE-RNN. Left: Singular values of POD (purple) and DEIM (green) snapshot matrices and singular values of the weight matrix of the ODE block (green). Right: histogram of APoZ scores, with the first 50 DEIM indices indicated in green color.

a separate step after training. We also assessed fine-tuning the compressed network to see how much accuracy can be recovered from each method. We focused on the runtime aspect of compressed models, since Neural ODEs are already very memory and parameter efficient by design.

We argued that some existing neuron pruning methods that measure neuron importance based on the final output of the layer only, such as the average percent of zeros (APoZ) approach, are not optimal for compressing Neural ODEs. The reason is that these methods do not account for the dynamics of the ODE block when identifying pruning targets. Here we showed how to formulate the Proper Orthogonal Decomposition with Discrete Empirical Interpolation Method (POD-DEIM) into compressing Neural ODEs. This method is developed for accurately approximating dynamical systems in low dimensions and the continuous-time network formalism makes it possible to use POD-DEIM in deep learning

applications. Additionally, we compared the POD-DEIM and APoZ methods to Singular Value Decomposition (SVD) based truncation, which is applicable to both discrete and continuous networks. The performance was measured with two different architectures, a convolutional Neural ODE and a recurrent ODE architecture, which were trained on the MNIST digit classification task.

Based on our results, the proposed approach for Neural ODE compression is the POD-DEIM method. While the SVD based truncation is the best at retaining accuracy, it yields significantly less acceleration than the APoZ and POD-DEIM methods. Moreover, with a small number of fine-tuning epochs the POD-DEIM method may reach the same accuracies as the SVD method. Even with exhaustive tuning, the APoZ methods lags slightly behind the other two in accuracy. What is more, in this study we only fine-tuned the layers following the ODE block, keeping results more indicative of raw performance

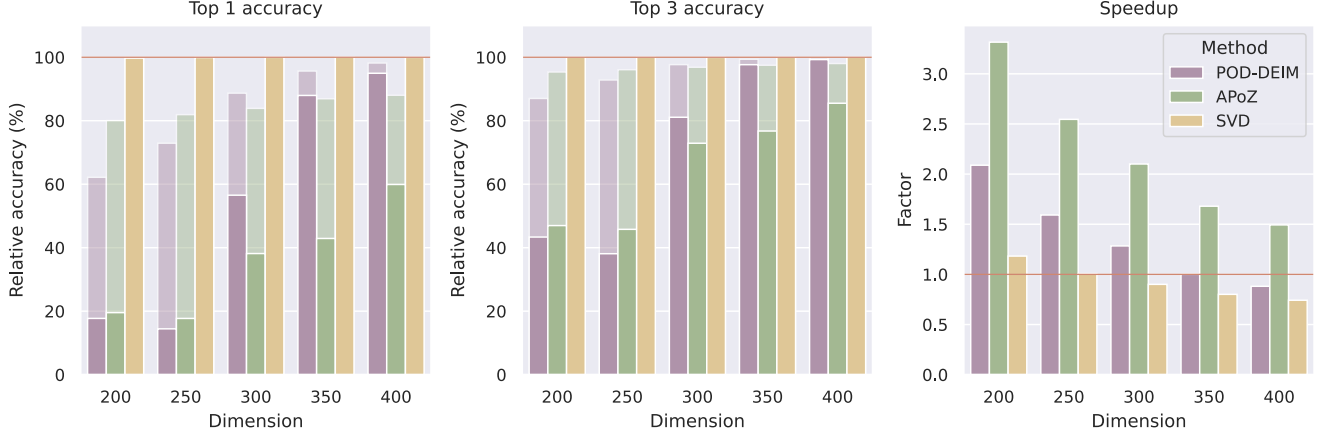


Fig. 6. Relative performance of reduced ODE-RNNs on the pixel MNIST task. All values are relative to those obtained from the original model, with the red horizontal line indicating equal prediction accuracy or wall-time with the original model. Lighter stacked bars indicate results for fine-tuned models. Left: Top 1 accuracies, middle: Top 3 accuracies, right: factor of speedup.

TABLE II
ABSOLUTE PERFORMANCE OF REDUCED CONVOLUTIONAL NEURAL ODES WITHOUT FINE-TUNING.

Method	Dimension	Top 1 (%)	Top 3 (%)	Runtime (s)
Original model	1024	97.9	99.7	11.2
DEIM	100	73.1	92.3	2.6
DEIM	300	87.5	97.0	4.1
DEIM	500	94.2	98.9	6.3
DEIM	700	97.4	99.6	9.7
DEIM	900	97.8	99.7	14.6
APoZ	100	29.2	44.4	2.3
APoZ	300	23.5	71.9	3.2
APoZ	500	44.6	93.9	4.3
APoZ	700	91.5	98.8	6.1
APoZ	900	96.3	99.6	8.4
SVD	100	89.8	98.8	5.2
SVD	300	96.0	99.4	8.6
SVD	500	97.4	99.8	11.1
SVD	700	97.7	99.7	15.9
SVD	900	97.9	99.8	19.3

TABLE III
ABSOLUTE PERFORMANCE OF REDUCED CONVOLUTIONAL NEURAL ODES, 3 FINE-TUNING EPOCHS.

Method	Dimension	Top 1 (%)	Top 3 (%)	Runtime (s)
Original model	1024	97.9	99.7	11.2
DEIM	100	84.3	96.9	2.6
DEIM	300	91.8	99.0	4.1
DEIM	500	95.7	99.4	6.3
DEIM	700	97.4	99.7	9.7
DEIM	900	97.8	99.7	14.6
APoZ	100	59.9	86.6	2.3
APoZ	300	78.5	94.9	3.2
APoZ	500	92.1	98.9	4.3
APoZ	700	96.2	99.6	6.1
APoZ	900	97.4	99.7	8.4
SVD	100	94.8	99.4	5.2
SVD	300	96.8	99.6	8.6
SVD	500	97.5	99.7	11.1
SVD	700	97.7	99.8	15.9
SVD	900	97.9	99.7	19.3

of the methods. If the entire model were tuned, POD-DEIM and SVD truncation have more parameters available for fine-tuning than the APoZ method, allowing for greater accuracy. On the other hand, the small number of parameters is what makes APoZ compressed models the fastest out of the three methods, hence it should be favored if there are no restrictions on fine-tuning and retraining time or if processing time or power restrictions prevent use of the POD-DEIM method. Moreover, compared to POD-DEIM, the APoZ method is more data and memory efficient, since it does not need snapshots of trajectories of the dynamics nor does it compute large matrix decompositions. Overall, we found that POD-DEIM gives the best compromise between speed and accuracy of compressed models.

We expect our results with the POD-DEIM method motivate machine learning researchers to look into the possible applications of other MOR methods developed in control theory and related fields. For example, we assessed here only the original DEIM [21] and the ODEIM [54] methods, although in recent years many advanced versions have been developed. These

include novel interpolation point selection methods [58] and methods with adaptive basis and interpolation point selection capabilities [59], [60], [61] that allow the model to change its dimension or projection subspace online. It is worth noting that the POD-DEIM method aims at approximating the dynamics of the original system and is not directly aware of classification or other performance indicators of the ANN. A potential future development could explore connecting the low-dimensional dynamics to downstream network performance.

A limitation of the present work is that we could only study relatively small models, due to known challenges with training very large Neural ODEs [47]. In small models the expected gains from compression methods are lesser than in large models. Moreover, we only discussed computation times on CPUs, as on GPUs the differences between original and compressed models were minor due to the small size of the models. While the computational speedup from POD-DEIM needs to be verified on GPUs and extremely large models, theoretical reductions in parameter and activation function count, comparable results from existing literature as well as

TABLE IV
ABSOLUTE PERFORMANCE OF REDUCED CONVOLUTIONAL NEURAL
ODEs, FINE-TUNING UNTIL CONVERGENCE.

Method	Dimension	Top 1 (%)	Top 3 (%)	Runtime (s)
Original model	1024	97.9	99.7	11.2
DEIM	100	86.5	97.4	2.6
DEIM	300	93.0	99.0	4.1
DEIM	500	96.1	99.4	6.3
DEIM	700	97.6	99.7	9.7
DEIM	900	97.9	99.7	14.6
APoZ	100	63.2	88.9	2.3
APoZ	300	82.5	96.1	3.2
APoZ	500	93.0	99.2	4.3
APoZ	700	96.4	99.6	6.1
APoZ	900	97.5	99.6	8.4
SVD	100	95.5	99.4	5.2
SVD	300	96.9	99.6	8.6
SVD	500	97.4	99.7	11.1
SVD	700	97.8	99.7	15.9
SVD	900	97.8	99.7	19.3

TABLE V
ABSOLUTE PERFORMANCE OF REDUCED ODE-RNNs, WITHOUT
FINE-TUNING.

Method	Dimension	Top 1 (%)	Top 3 (%)	Runtime (s)
Original model	512	96.2	99.6	41.5
DEIM	200	17.0	43.2	19.9
DEIM	250	13.9	38.0	26.1
DEIM	300	54.4	80.8	32.3
DEIM	350	84.6	97.3	41.5
DEIM	400	91.3	98.9	47.1
APoZ	200	18.8	46.8	12.5
APoZ	250	17.0	45.6	16.3
APoZ	300	36.7	72.6	19.8
APoZ	350	41.3	76.5	24.7
APoZ	400	57.6	85.2	27.8
SVD	200	95.9	99.6	35.1
SVD	250	96.1	99.6	41.2
SVD	300	96.1	99.6	46.1
SVD	350	96.1	99.6	51.8
SVD	400	96.1	99.6	55.9

TABLE VI
ABSOLUTE PERFORMANCE OF REDUCED ODE-RNNs, 3 FINE-TUNING
EPOCHS

Method	Dimension	Top 1 (%)	Top 3 (%)	Runtime (s)
Original model	512	96.2	99.6	41.5
DEIM	200	59.8	86.7	19.9
DEIM	250	70.1	92.5	26.1
DEIM	300	85.3	97.3	32.3
DEIM	350	92.0	99.0	41.5
DEIM	400	94.4	99.4	47.1
APoZ	200	77.0	95.0	12.5
APoZ	250	78.8	95.7	16.3
APoZ	300	80.7	96.5	19.8
APoZ	350	83.6	97.1	24.7
APoZ	400	84.7	97.6	27.8
SVD	200	96.1	99.6	35.1
SVD	250	96.2	99.7	41.2
SVD	300	96.3	99.6	46.1
SVD	350	96.2	99.7	51.8
SVD	400	96.3	99.6	55.9

TABLE VII
ABSOLUTE PERFORMANCE OF REDUCED ODE-RNNs, FINE-TUNING
UNTIL CONVERGENCE

Method	Dimension	Top 1 (%)	Top 3 (%)	Runtime (s)
Original model	512	96.2	99.6	41.5
DEIM	200	66.2	87.9	19.9
DEIM	250	78.1	95.1	26.1
DEIM	300	85.3	97.5	32.3
DEIM	350	90.6	98.6	41.5
DEIM	400	95.4	99.6	47.1
APoZ	200	81.5	96.5	12.5
APoZ	250	81.8	97.0	16.3
APoZ	300	84.0	97.4	19.8
APoZ	350	86.1	97.9	24.7
APoZ	400	87.4	98.3	27.8
SVD	200	96.1	99.6	35.1
SVD	250	96.3	99.7	41.2
SVD	300	96.2	99.6	46.1
SVD	350	96.3	99.6	51.8
SVD	400	96.3	99.6	55.9

our results on the CPU indicate that the method should perform well.

Our results were obtained with fixed-step ODE solvers. We found that compressing Neural ODEs with the POD-DEIM and APoZ methods sometimes induces stiffness to the low-rank models (data not shown). This could cause adaptive ODE solvers to be slower, mitigating some of the speedup from compression. In our findings, fine-tuning usually restored the number of forward steps taken by adaptive solvers to the level of the original model or better. However, systematic results of this behavior were not sought here, and the fixed step solver was a way to control variability in the results. A future study is needed to quantify the effect of stiffness that may rise as an effect of compressing Neural ODEs.

VI. CONCLUSION

In this study, we addressed speeding up and compressing artificial neural networks with continuous layers. Our results are significant and novel, as we found that our POD-DEIM method, which has not previously been used to compress ANNs or Neural ODEs, provides more attractive results speed and accuracy wise than two established compression methods from the literature. The reason is that POD-DEIM properly

considers the trajectory of the ODE layer in an efficient manner. Additionally, our results show that existing approaches for compressing neural networks may under-perform, as they do not account for the dynamics of embedded ODE.

We expect our acceleration method to be useful when Neural ODEs are deployed on low-power hardware or battery-powered devices, for example wearable devices, medical monitoring instruments or smart phones. The dynamical system formalism for neural networks also allows other model order reduction methods to be applied in deep learning, which will provide for interesting future studies.

ACKNOWLEDGMENT

This research has received funding from the Academy of Finland (decision no. 297893 M.-L.L. and decision no. 310489 L.P.) and the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement Nos. 785907 (Human Brain Project SGA2) and 945539 (Human Brain Project SGA3). M.L. acknowledges the support from Tampere University Graduate School and The Finnish Foundation for Technology Promotion.

Neural network illustrations generated partly with <http://alexlenail.me/NN-SVG/index.html>

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
- [2] E. Weinan, "A proposal on machine learning via dynamical systems," *Communications in Mathematics and Statistics*, vol. 5, no. 1, pp. 1–11, 2017.
- [3] Y. Lu, A. Zhong, Q. Li, and B. Dong, "Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3276–3285.
- [4] E. Haber and L. Ruthotto, "Stable architectures for deep neural networks," *Inverse Problems*, vol. 34, no. 1, p. 014004, 2017.
- [5] R. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, 2018, pp. 6571–6583.
- [6] Y. Rubanova, R. T. Chen, and D. K. Duvenaud, "Latent ordinary differential equations for irregularly-sampled time series," in *Advances in Neural Information Processing Systems*, 2019, pp. 5320–5330.
- [7] D. Blalock, J. J. Gonzalez Ortiz, J. Frankle, and J. Gutttag, "What is the state of neural network pruning?" in *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds., vol. 2, 2020, pp. 129–146.
- [8] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in Neural Information Processing Systems*, vol. 28, pp. 1135–1143, 2015.
- [9] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang, "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures," *arXiv preprint arXiv:1607.03250*, 2016.
- [10] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Advances in Neural Information Processing Systems*, 2014, pp. 1269–1277.
- [11] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [12] H. Markram *et al.*, "Reconstruction and simulation of neocortical microcircuitry," *Cell*, vol. 163, no. 2, 2015.
- [13] A. Kellems, S. Chaturantabut, D. Sorensen, and S. Cox, "Morphologically accurate reduced order modeling of spiking neurons," *Journal of Computational Neuroscience*, vol. 28, no. 3, 2010.
- [14] B. Du, D. Sorensen, and S. Cox, "Model reduction of strong-weak neurons," *Frontiers in Computational Neuroscience*, vol. 8, no. 164, 2014.
- [15] M. Lehtimäki, L. Paunonen, S. Pohjolainen, and M.-L. Linne, "Order reduction for a signaling pathway model of neuronal synaptic plasticity," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7687–7692, 2017.
- [16] M. Lehtimäki, L. Paunonen, and M.-L. Linne, "Projection-based order reduction of a nonlinear biophysical neuronal network model," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 1–6.
- [17] M. Lehtimäki, I. Seppälä, L. Paunonen, and M.-L. Linne, "Accelerated simulation of a neuronal population via mathematical model order reduction," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2020, pp. 118–122.
- [18] R. Chaudhuri, B. Gerçek, B. Pandey, A. Peyrache, and I. Fiete, "The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep," *Nature Neuroscience*, vol. 22, no. 9, pp. 1512–1520, 2019.
- [19] U. Cohen, S. Chung, D. Lee, and H. Sompolsky, "Separability and geometry of object manifolds in deep neural networks," *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.
- [20] J. Lumley, G. Berkooz, and P. Holmes, "The Proper Orthogonal Decomposition in the analysis of turbulent flows," *Annual Review of Fluid Mechanics*, vol. 25, pp. 539–575, 1993.
- [21] S. Chaturantabut and D. Sorensen, "Nonlinear model reduction via discrete empirical interpolation," *SIAM Journal on Scientific Computing*, vol. 32, no. 5, pp. 2737–2764, 2010.
- [22] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, "An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations," *Comptes Rendus Mathématique*, vol. 339, no. 9, pp. 667–672, 2004.
- [23] S. Chaturantabut and D. Sorensen, "A state space error estimate for pod-deim nonlinear model reduction," *SIAM Journal on Numerical Analysis*, vol. 50, no. 1, pp. 46–63, 2012.
- [24] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in *Google Research*, 2011.
- [25] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training deep neural networks with binary weights during propagations," in *Advances in Neural Information Processing Systems*, 2015, pp. 3123–3131.
- [26] W. Chen, J. Wilson, S. Tyree, K. Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," in *International conference on machine learning*. PMLR, 2015, pp. 2285–2294.
- [27] J. Ren and L. Xu, "On vectorization of deep convolutional neural networks for vision tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.
- [28] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through FFTs," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.
- [29] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," *ACM SIGARCH Computer Architecture News*, vol. 42, no. 1, pp. 269–284, 2014.
- [30] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems*, 1990, pp. 598–605.
- [31] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in *Advances in Neural Information Processing Systems*, 1993, pp. 164–171.
- [32] D. Yu, F. Seide, G. Li, and L. Deng, "Exploiting sparseness in deep neural networks for large vocabulary speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4409–4412.
- [33] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Interspeech*, 2013, pp. 2365–2369.
- [34] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 243–254, 2016.
- [35] A. Parashar, M. Rhu, A. Mukkara, A. Puglielli, R. Venkatesan, B. Khailany, J. Emer, S. W. Keckler, and W. J. Dally, "SCNN: An accelerator for compressed-sparse convolutional neural networks," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 2, pp. 27–40, 2017.
- [36] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua, "Learning separable filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2754–2761.
- [37] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up convolutional neural networks with low rank expansions," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.
- [38] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," *arXiv preprint arXiv:1412.6115*, 2014.
- [39] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1943–1955, 2015.
- [40] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 677–684.
- [41] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.
- [42] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1389–1397.
- [43] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.
- [44] V. Lebedev and V. Lempitsky, "Speeding-up convolutional neural networks: A survey," *Bulletin of the Polish Academy of Sciences, Technical Sciences*, vol. 66, no. 6, 2018.
- [45] E. Dupont, A. Doucet, and Y. W. Teh, "Augmented Neural ODEs," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.
- [46] J. Kelly, J. Bettencourt, M. J. Johnson, and D. K. Duvenaud, "Learning differential equations that are easy to solve," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell,

- M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4370–4380.
- [47] C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. Oberman, “How to train your neural ODE: the world of Jacobian and kinetic regularization,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 3154–3164.
 - [48] P. Kidger, R. T. Chen, and T. Lyons, ““Hey, that’s not an ODE”: Faster ODE adjoints with 12 lines of code,” *arXiv preprint arXiv:2009.09457*, 2020.
 - [49] M. Poli, S. Massaroli, A. Yamashita, H. Asama, and J. Park, “Hy-persolvers: Toward fast continuous-depth models,” *arXiv preprint arXiv:2007.09601*, 2020.
 - [50] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
 - [51] S. Massaroli, M. Poli, M. Bin, J. Park, A. Yamashita, and H. Asama, “Stable neural flows,” *arXiv preprint arXiv:2003.08063*, 2020.
 - [52] B. Chang, M. Chen, E. Haber, and E. H. Chi, “AntisymmetricRNN: A dynamical system view on recurrent neural networks,” in *International Conference on Learning Representations*, 2019.
 - [53] L. Sirovich, “Turbulence and the dynamics of coherent structures. I-III,” *Quarterly of Applied Mathematics*, vol. 45, no. 3, pp. 561–590, 1987.
 - [54] B. Peherstorfer, Z. Drmac, and S. Gugercin, “Stability of discrete empirical interpolation and gappy proper orthogonal decomposition with randomized and deterministic sampling points,” *SIAM Journal on Scientific Computing*, vol. 42, no. 5, pp. A2837–A2864, 2020.
 - [55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshine, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.
 - [56] M. Poli, S. Massaroli, A. Yamashita, H. Asama, and J. Park, “Torchdyn: A neural differential equations library,” *arXiv preprint arXiv:2009.09346*, 2020.
 - [57] Y. LeCun, “The MNIST database of handwritten digits,” <http://yann.lecun.com/exdb/mnist/>, 1998.
 - [58] Z. Drmac and S. Gugercin, “A new selection operator for the Discrete Empirical Interpolation Method—improved a priori error bound and extensions,” *SIAM Journal on Scientific Computing*, vol. 38, no. 2, pp. A631–A648, 2016.
 - [59] B. Peherstorfer, D. Butnaru, K. Willcox, and H. Bungartz, “Localized discrete empirical interpolation method,” *SIAM Journal on Scientific Computing*, vol. 36, no. 1, 2014.
 - [60] B. Peherstorfer and K. Willcox, “Online adaptive model reduction for nonlinear systems via low-rank updates,” *SIAM Journal on Scientific Computing*, vol. 37, no. 4, 2015.
 - [61] S. Chaturantabut, “Temporal localized nonlinear model reduction with a priori error estimate,” *Applied Numerical Mathematics*, vol. 119, pp. 225–238, 2017.