MAESTRO: UNCOVERING LOW-RANK STRUCTURES VIA TRAINABLE DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Neural Networks (DNNs) have been a large driver and enabler for AI breakthroughs in recent years. These models have been getting larger in their attempt to become more accurate and tackle new upcoming use-cases, including AR/VR and intelligent assistants. However, the training process of such large models is a costly and time-consuming process, which typically yields a single model to fit all targets. To mitigate this, various techniques have been proposed in the literature, including pruning, sparsification or quantization of the model weights and updates. While able to achieve high compression rates, they often incur computational overheads or accuracy penalties. Alternatively, factorization methods have been leveraged to incorporate low-rank compression in the training process. Similarly, such techniques (e.g., SVD) frequently rely on the computationally expensive decomposition of layers and are potentially sub-optimal for non-linear models, such as DNNs.

In this work, we take a further step in designing efficient low-rank models and propose MAESTRO, a framework for trainable low-rank layers. Instead of regularly applying a priori decompositions such as SVD, the low-rank structure is built into the training process through a generalized variant of Ordered Dropout. This method imposes an importance ordering via sampling on the decomposed DNN structure. Our theoretical analysis demonstrates that our method recovers the SVD decomposition of linear mapping on uniformly distributed data and PCA for linear autoencoders. We further apply our technique on DNNs and empirically illustrate that MAESTRO enables the extraction of lower footprint models that preserve model performance while allowing for graceful accuracy-latency tradeoff for the deployment to devices of different capabilities.

1 INTRODUCTION

Deep Learning has been experiencing an unprecedented uptake, with models achieving a (super-)human level of performance in several tasks across modalities, giving birth to even more intelligent assistants and next-gen visual perception and generation systems. However, the price of this performance is that models are getting significantly larger, with training and deployment becoming increasingly costly. Therefore, techniques from Efficient ML become evermore relevant (Laskaridis et al., 2022), and a requirement for deployment in constrained devices, such as smartphones or IoT devices.

Typical techniques to compress the network involve *i) quantization*, i.e., reducing precision of the model (Wang et al., 2019) or communicated updates (Seide et al., 2014; Alistarh et al., 2017), *ii) pruning* the model during training, e.g., through Lottery Ticket Hypothesis (LTH) (Frankle & Carbin, 2019), *iii) sparsification* of the network representation and updates, i.e., dropping the subset of coordinates (Suresh et al., 2017; Alistarh et al., 2018) or *iv) low-rank approximation* (*Wang et al., 2021*; *Dudziak et al., 2019*), i.e. keeping the most relevant ranks of the decomposed network. Despite the benefits during deployment, that is a lower footprint model, in many cases, the overhead during training time or the accuracy degradation can be non-negligible. Moreover, many techniques can introduce mutliple hyperparameters or the need to fine-tune to recover the lost accuracy.

In this work, we focus on training low-rank factorized models. Specifically, we pinpoint the challenges of techniques (Wang et al., 2021; 2023) when decomposing the parameters of each layer in low-rank space and the need to find the optimal ranks for each one at training time. To solve this, we adopt and non-trivially extend the Ordered Dropout technique from (Horváth et al., 2021) and apply it to

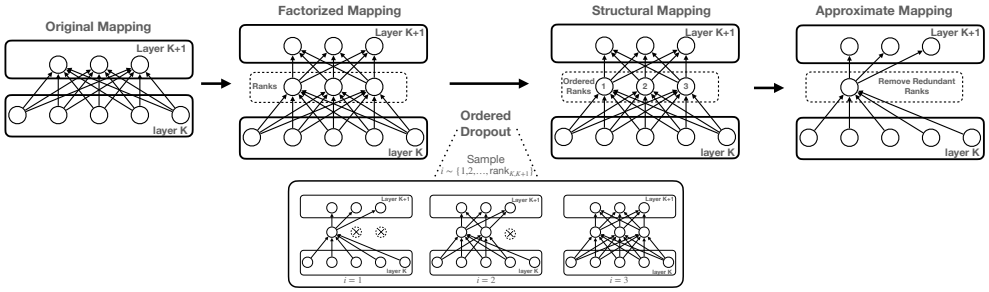


Figure 1: MAESTRO's construction. To obtain low-rank approximation, the given linear map is decomposed and trained with ordered dropout to obtain an ordered representation that can be efficiently pruned.

progressively find the optimal decomposition for each layer of a DNN while training (Fig. 1). Critical differences to prior work include *i)* the non-uniformity of the search space (i.e. we allow for different ranks per layer), *ii)* the trainable aspect of the decomposition to reflect the data distribution, and *iii)* the gains to training and deployment time without sacrificing accuracy. Nevertheless, we also provide a latency-accuracy trade-off mechanism to deploy the model on more constrained devices.

Our contributions can be summarized as follows:

- We propose MAESTRO, a novel layer decomposition technique that enables learning low-rank layers in a progressive manner while training. We fuse layer factorization and an extended variant of the ordered dropout, in a novel manner, by embedding OD directly into the factorized weights. By decomposing layers and training on stochastically sampled low-rank models, we apply ordered importance decomposed representation of each layer. We combine this with a *hierarchical group-lasso* term (Yuan & Lin, 2006) in the loss function to zero out redundant ranks and *progressively shrink* the rank space. This way, we enable computationally efficient training achieved by the proposed decomposition without relying on inexact and potentially computationally expensive decompositions such as Singular Value Decomposition (SVD).
- MAESTRO is a theoretically motivated approach that embeds decomposition into training. First, we show that our new objective is able to recover *i)* the SVD of the target linear mapping for the particular case of uniform data distribution and *ii)* the Principal Component Analysis (PCA) of the data in the case of identity mapping.
- As MAESTRO's decomposition is part of the training procedure, it also accounts for data distribution and the target function, contrary to SVD, which operates directly on learned weights. We show that this problem *already arises* for a simple linear model and empirically generalize our results in the case of DNNs, by applying our method to different types of layers (including fully-connected, convolutional, and attention) spanning across three datasets and modalities. We illustrate that our technique achieves better results than SVD-based baselines at a lower cost.

2 RELATED WORK

The topic of Efficient ML has received a lot of attention throughout the past decade as networks have been getting increasingly computationally expensive. Towards this end, we distinguish between training and deployment time, with the latter having a more significant impact and thus amortizes the potential overhead during training. Nevertheless, with the advent of Federated Learning (McMahan et al., 2017), efficient training becomes increasingly relevant to remain tractable.

Efficient inference. For efficient deployment, various techniques have been proposed that either optimize the architecture of the DNN in a hand-crafted (Howard et al., 2017) or automated manner (i.e. NAS) (Tan & Le, 2019), they remove redundant computation by means of pruning parts of the network (Han et al., 2015; Carreira-Perpinán & Idelbayev, 2018; Frankle & Carbin, 2019; Chen et al., 2021; Sreenivasan et al., 2022; Li et al., 2016; Wen et al., 2016; Hu et al., 2016; Wen et al., 2016; Zhu & Gupta, 2017; He et al., 2017; Yang et al., 2017; Liu et al., 2018; Yu & Huang, 2019b), in a structured or unstructured manner, or utilise low-precision representation (Wang et al., 2019) of the neurons and activations. However, such techniques may involve non-negligible training overheads or lack flexibility of variable footprint upon deployment. Closer to our method, there have been techniques leveraging low-rank approximation (e.g. SVD) for efficient inference (Xue et al., 2013; Sainath et al., 2013; Jaderberg et al., 2014; Wiesler et al., 2014; Dudziak et al., 2019). Last, there is a category of techniques that dynamically resize the network at runtime for compute, memory or energy efficiency, based on early-exiting (Laskaridis et al., 2021) or dynamic-width (Yu et al., 2019) and leverage the accuracy-latency tradeoff.

Efficient training. On the other hand, techniques for efficient training become very relevant nowadays when scaling DNNs sizes Hu et al. (2021) or deploying to embedded devices (Lin et al., 2022), and oftentimes offer additional gains at deployment time. Towards this goal, there have been employed methods where part of the network is masked (Sidahmed et al., 2021) or dropped (Alam et al., 2022; Caldas et al., 2019) during training, with the goal of minimizing the training footprint. Similarly to early-exiting, multi-exit variants for efficient training (Kim et al., 2023; Liu et al., 2022) have been proposed, and the same applies for width-based scaling Horváth et al. (2021); Diao et al. (2021). Last but not least, in the era of transformers and LLMs, where networks have scaled exponentially in size, PEFT-based techniques, such as adapter-based fine-tuning (Houlsby et al., 2019) (such as LoRA (Hu et al., 2021)), become increasingly important and make an important differentiator for tackling downstream tasks.

Learning ordered representation. Originally, Ordered Dropout (OD) was proposed as a mechanism for importance-based pruning for the easy extraction of sub-networks devised to allow for heterogeneous federated training (Horváth et al., 2021). The earlier work that aims to learn ordered representation includes a similar technique to OD—Nested Dropout, which proposed a similar construction, applied to the representation layer in autoencoders (Rippel et al., 2014) to enforce identifiability of the learned representation or the last layer of the feature extractor (Horváth et al., 2021) to learn an ordered set of features for transfer learning. We leverage and non-trivially extend OD in our technique as a means to order ranks in terms of importance in a nested manner during training of a decomposed network that is progressively shrunk as redundant ranks converge to 0. Ranks selection is ensured through hierarchical group lasso penalty, as described in Sec. 3.3. Moreover, contrary to (Horváth et al., 2021), which assumed a uniform width, our formulation allows for heterogeneous ranks per layer. Last, we leverage the ordered representation of ranks at inference time to further compress the model, allowing a graceful degradation of performance as a mechanism for the accuracy-latency trade-off.

3 MAESTRO

In this work, we focus on low-rank models as a technique to reduce the computational complexity and memory requirements of the neural network model. The main challenge that we face is the selection of the optimal rank or the trade-off between the efficiency and the rank for the given layer represented by linear mapping. Therefore, we devise an importance-based training technique, MAESTRO, which not only learns a mapping between features and responses, but also learns the decomposition of the trained network. This is achieved by factorizing all the layers in the network.

3.1 FORMULATION

Low-rank approximation. Our inspiration comes from the low-rank matrix approximation of a matrix $A \in \mathbb{R}^{m \times n}$. For simplicity, we assume that A has rank $r = \min\{m, n\}$ with $k \leq r$ distinct non-zero singular values $\tilde{\sigma}_1 > \tilde{\sigma}_2 > \ldots > \tilde{\sigma}_k > 0$, with corresponding left and right singular vectors $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k \in R^m$ and $\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k \in R^n$, respectively. For such a matrix, we can rewrite its best l-rank approximation as the following minimization problem

$$\min_{U \in \mathbb{R}^{m \times l}, V \in \mathbb{R}^{n \times l}} \left\| \sum_{i=1}^{l} u_i v_i^\top - A \right\|_F^2 \tag{1}$$

where c_i denotes the i-th row of matrix C and $\|\cdot\|_F$ denotes Frobenius norm. We note that Problem (1) is non-convex and non-smooth. However, Ye & Du (2021) showed that the randomly initialized gradient descent algorithm solves this problem in polynomial time. In this work, we consider the best rank approximation across all the ranks that leads us to the following objective

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \frac{1}{r} \sum_{b=1}^{r} \left\| U_{:b} V_{:b}^\top - A \right\|_F^2, \tag{2}$$

where $C_{:b}$ denotes the first b columns of matrix C. This objective, up to scaling, recovers SVD of A exactly, and for the case of distinct non-zero singular values, the solution is, up to scaling, unique Horváth et al. (2021). This formulation, however, does not account for the data distribution, i.e., it cannot tailor the decomposition to capture specific structures that appear in the dataset.

Data-dependent low-rank approximation. Therefore, the next step of our construction is to extend this problem formulation with data that can further improve compression, reconstruction, and

generalization, and incorporate domain knowledge. We assume that data comes from the distribution $x \sim \mathcal{X}$ centered around zero, i.e., $\mathbf{E}_{x\sim\mathcal{X}}[x] = 0$,[1] and the response is given by $y = Ax$. In this particular case, we can write the training loss as

$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}} \mathbf{E}_{x,y\sim\mathcal{X}} \left[\sum_{b=1}^{r} \frac{1}{r} \left\| U_{:b} V_{:b}^{\top} x - y \right\|^2 \right]. \tag{3}$$

It is important to note that the introduced problem formulation (3) is the same as the Ordered Dropout formulation of Horváth et al. (2021) for the neural network with a single hidden layer and no activations, and it can be solved using stochastic algorithms by sampling from the data distribution $\mathcal{X}$ (subsampling) and rank distribution $\mathcal{D}$. However, there is an important distinction when we apply MAESTRO for deep neural networks. While FjORD applies uniform dropout across the width of the network for each layer, we propose to decompose each layer independently to uncover its – potentially different – optimal rank for deployment. We discuss details in the next paragraph.

DNN low-rank approximation. For Deep Neural Networks (DNNs), we seek to uncover the optimal ranks for a set of d linear mappings $W^1 \in \mathbb{R}^{m_1\times n_1}, \ldots, W^d \in \mathbb{R}^{m_d\times n_d}$, where W^i's are model parameters and d is model depth, e.g., weights corresponding to linear layers[2], by decomposing them as $W^i = U^i \left(V^i\right)^{\top}$. We discuss how these are selected in the next section. To decompose the network, we aim to minimize the following objective:

$$\mathbf{E}_{x,y\sim\mathcal{X}} \left[\frac{1}{\sum_{i=1}^{d} r_i} \sum_{i=1}^{d} \sum_{b=1}^{r_i} l(h(U^1\left(V^1\right)^{\top}, \ldots, U^i_{:b}\left(V^i_{:b}\right)^{\top}, \ldots, U^d\left(V^d\right)^{\top}, W^o, x), y) \right], \tag{4}$$

where $r_i = \min\{m_i, n_i\}$, l is a loss function, h is a DNN, and W^o are the other weights that we do not decompose. We note that our formulation aims to decompose each layer, while decompositions across layers do not directly interact. The motivation for this approach is to uncover low-rank structures within each layer that are not affected by inaccuracies from other layers due to multiple low-rank approximations.

3.2 LAYER FACTORIZATION

The following sections discuss how we implement model factorization for different architectures.

FC layers. A 2-layer fully connected (FC) neural network can be expressed as $f(x) = \sigma(\sigma(xW_1)W_2)$, where Ws are weight matrices of each FC layer, and $\sigma(\cdot)$ is any arbitrary activation function, e.g., ReLU. The weight matrix W can be factorized as $UV^{\top}$.

CNN layers. For a convolution layer with dimension, $W \in \mathbb{R}^{m\times n\times k\times k}$ where m and n are the number of input and output channels, and k is the size of the convolution filters. Instead of directly factorizing the 4D weight of a convolution layer, we factorize the unrolled 2D matrix. Unrolling the 4D tensor W leads to a 2D matrix with shape $W_{\text{unrolled}} \in \mathbb{R}^{mk^2\times n}$, where each column represents the weight of a vectorized convolution filter. Factorization can then be conducted on the unrolled 2D matrix; see (Wang et al., 2021) for details.

Transformers. A Transformer layer consists of a stack of encoders and decoders Vaswani et al. (2017). The encoder and decoder contain three main building blocks: the multi-head attention layer, position-wise feed-forward networks (FFN), and positional encoding. We factorize all trainable weight matrices in the multi-head attention (MHA) and the FFN layers. The FFN layer factorization can directly adopt the strategy from the FC factorization. A p-head attention layer learns p attention mechanisms on the key, value, and query (K, V, Q) of each input token:

$$\texttt{MHA}(Q, K, V) = \texttt{Concat}(\text{head}_1, \ldots, \text{head}_p)W^O.$$

Each head performs the computation of:

$$\texttt{head}_i = \texttt{Attention}(QW_Q^{(i)}, KW_K^{(i)}, VW_V^{(i)}) = \texttt{softmax}\left(\frac{QW_Q^{(i)}W_K^{(i)\top}K^{\top}}{\sqrt{d/p}}\right)VW_V^{(i)}.$$

where d is the hidden dimension. The trainable weights $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, i \in \{1, 2, \ldots, p\}$ can be factorized by simply decomposing all learnable weights $W\cdot$ in an attention layer and obtaining $U\cdot V^{\top}\cdot$ Vaswani et al. (2017).

[1] We make this assumption for simplicity. It can be simply overcome by adding a bias term into the model.

[2] We can apply our decomposition on different types of layers, such as Linear, Convolutional and Transformers as shown in Sec. 3.2.

3.3 TRAINING TECHNIQUES

Having defined the decomposition of typical layers found in DNNs, we move to formulate the training procedure of our method, formally described in Algorithm 1. Training the model comprises an iterative process of propagating forward on the model by *sampling a rank* b_i per decomposed layer i up to maximal rank r_i (line 3). We calculate the loss, which integrates an additional *hierarchical group lasso* component (lines 4) and *backpropagate* on the sampled decomposed model (line 5). At the end of each epoch, we *progressively shrink* the network by updating the maximal rank r_i, based on an importance threshold ε_{ps} (line 11). We provide more details about each component below.

Algorithm 1: MAESTRO (Training Process)

Input: epochs E, dataset $\mathcal{D}$, model h parametrized by $U^1 \in \mathbb{R}^{m_1 \times r_1}$, $V^1 \in \mathbb{R}^{n_1 \times r_1}, \ldots, U^d \in \mathbb{R}^{m_d \times r_d}, V^d \in \mathbb{R}^{n_d \times r_d}$, W^o, and hyperparameters λ_{gl}, ε_{ps}

1 **for** $t \leftarrow 0$ **to** $E-1$ **do** *// Epochs*
2 **for** $(x, y) \in \mathcal{D}$ **do** *// Iterate over dataset*
3 Sample $(i, b) \sim \left\{\{(i, b)\}_{b=1}^{r_i}\right\}_{i=1}^{d}$;
4 $L = l(h(U^1(V^1)^\top, \ldots, U^i_{:b}(V^i_{:b})^\top, \ldots, U^d(V^d)^\top, W^o, x), y) +$
  $+\lambda_{gl} \sum_{i=1}^{d} \sum_{b=1}^{r_i} \left(\left\|U^i_{b:}\right\| + \left\|V^i_{b:}\right\|\right)$ *// compute loss*
5 L.backward() *// Update weights*
6 **end**
7 **for** $i \leftarrow 1$ **to** d **do**
8 **for** $b \leftarrow 1$ **to** r_i **do**
9 *// rank importance thresholding*
10 **if** $\left\|V^i_{b:}\right\|\left\|U^i_{b:}\right\| \leq \varepsilon_{ps}$ **then**
11 $r_i = b - 1$ *// progressive shrinking*
12 **break**
13 **end**
14 **end**
15 **end**
16 **end**

Efficient training via sampling. In Sec. 4, we show that for the linear case (3), the optimal solution corresponds to PCA over the linearly transformed dataset. This means that the obtained solution contains orthogonal directions. This property is beneficial because it directly implies that when we employ gradient-based optimization, not only is the gradient zero at the optimum, but the gradient with respect to each summand in Equation (3) is also zero. The same property is directly implied by overparametrization Ma et al. (2018) or strong growth condition Schmidt & Roux (2013). As a consequence, this enables us to sample only one summand at a time and obtain the same quality solution. When considering (4) as an extension to (3), it is unclear whether this property still holds, which would also imply that the set of stationary points of (3) is a subset of stationary points of the original objective without decomposition. However, in the experiments, we observed that sampling is sufficient to converge to a good-quality solution. If this only holds approximately, one could leverage fine-tuning to recover the loss in performance.

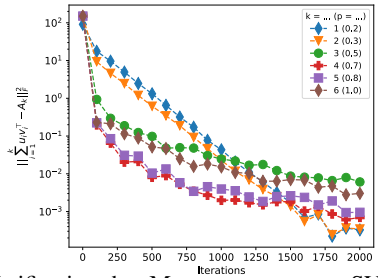
Efficient rank extraction via hierarchical group-lasso. By definition, (3) leads to an ordered set of ranks for each layer. This ordered structure enables efficient rank extraction and selection. To effectively eliminate unimportant ranks while retaining the important ones, thus leading to a more efficient model, we consider Hierarchical Group Lasso (HGL) Lim & Hastie (2015) in the form

$$\lambda_{gl} \sum_{i=1}^{d} \sum_{b=1}^{r_i} \left(\left\|U^i_{b:}\right\| + \left\|V^i_{b:}\right\|\right), \tag{5}$$

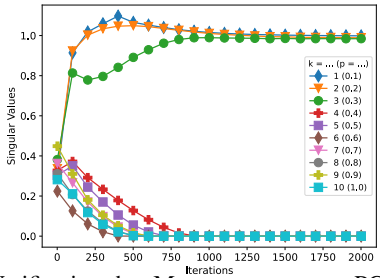
where $C_{b:}$ denotes the matrix that contains all the columns of C except for the first $b-1$ columns.

Progressive shrinking. HGL encourages that unimportant ranks become zero and can be effectively removed from the model. To account for this, for each layer we remove $V^i_{b:}$ and $U^i_{b:}$ (i.e., set $r_i = b-1$) if $\left\|V^i_{b:}\right\|\left\|U^i_{b:}\right\| \leq \varepsilon_{ps}$, where ε_{ps} is a pre-selected threshold – and a hyperparameter of our method.

Initialization. Initialization is a key component of the training procedure He et al. (2015); Mishkin & Matas (2015). To adopt the best practices from standard non-factorized training, we follow a similar approach to Khodak et al. (2021); Wang et al. (2021), where we first initialize the non-factorized model using standard initialization. For initializing factorized layers, we use the Singular Value



(a) Verification that MAESTRO recovers SVD for linear mapping with uniform data. The plot displays the L2 distance between the best rank k and MAESTRO's approximation of mapping A. The target matrix was randomly generated 9×6 matrix with rank 3. p and k represent relative and actual rank.



(b) Verification that MAESTRO recovers PCA for identity mapping. The plot displays the estimates of singular values. The data distribution has only 3 directions. It is expected that the top 3 ranks will converge to value one and the rest to zero. p and k stand for relative and actual rank, respectively.

Figure 2: Empirical showcase of theoretical properties of the MAESTRO's formulation.

Decomposition of the non-factorized initialization – in a full-rank form – to ensure that the resulting product matrix is the same as the original parameter decomposition. In addition, SVD is an optimal decomposition for the linear case with uniform data. However, in contrast with the adaptive baseline method (Wang et al., 2023) we only decompose once, rather than on every training iteration.

3.4 TRAIN-ONCE, DEPLOY-EVERYWHERE

Up until now, we have described how our method works for training low-rank models, which yield computational, memory, network, and energy (Wu et al., 2022) bandwidth benefits during training. At deployment time, one can directly deploy the final model (rank r_i for each layer) on the device, which we acquire from performing a threshold sweep of ε_{ps} over the effective range of rank importance across layers. However, in case we want to run on even more constrained devices, such as mobile (Almeida et al., 2021) or embedded (Almeida et al., 2021) systems, the learned decomposition also gives us the flexibility to further compress the model in a straightforward manner, effectively trading off accuracy for a smaller model footprint. Inspired by Yu & Huang (2019a), we propose to use greedy search. We begin with the current model and compare model performance across various low-rank models, each created by removing a certain percentage of ranks from each layer. We then eliminate the ranks that cause the least decrease in performance. This process is iterated until we reach the desired size or accuracy constraint. To make this approach efficient, we estimate the loss using a single mini-batch with a large batch size, for example, 2048. This also avoids issues with BatchNorm layers; see Yu & Huang (2019a) for details.

In summary, MAESTRO comprises a technique for trainable low-rank approximation during training time that progressively compresses the model, reflecting the data distribution, and a method that enables a graceful trade-off between accuracy and latency for embedded deployment, by selecting the most important parts of the network. We validate these claims in Sec. 5.2 and 5.5, respectively.

4 THEORETICAL GUARANTEES

In this section, we further investigate the theoretical properties of MAESTRO for the linear mappings, i.e., the setup of the problem formulation (3).

Theorem 4.1 (Informal)**.** *Let $A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$ be a SVD decomposition of A. Then, the minimization problem (3) is equivalent to PCA applied to the transformed dataset $x \to \tilde{\Sigma}\tilde{V}^\top x$, $x \sim \mathcal{X}$ projected on the column space of $\tilde{U}$.*

The formal statement can be found in Appendix D. Theorem 4.1 shows that MAESTRO can adapt to data distribution by directly operating on data $x \sim \mathcal{X}$ and also to the target mapping by projecting data to its right singular vectors scaled by singular values. In particular, we show that in the special case, when $\mathcal{X}$ is the uniform distribution on the unit ball, (3), i.e., MAESTRO, exactly recovers truncated SVD of A, which is consistent with the prior results Horváth et al. (2021). In the case A is the identity, it is straightforward to see that MAESTRO is equivalent to PCA. We can see that MAESTRO can efficiently extract low-rank solutions by filtering out directions corresponding to the null space of the target mapping A and directions with no data. We also numerically verify both of the special cases–PCA and SVD, by minimizing (3) using stochastic gradient descent (SGD) with $\mathcal{D}$ being the uniform distribution. These preliminary experiments are provided in Fig. 2a and 2b.

We showed that MAESTRO could recover SVD in a particular case of the linear model and the uniform data distribution on the unit ball. We note that in this case, SVD is optimal, and we cannot acquire better decomposition. Therefore, it is desired that MAESTRO is equivalent to SVD in this scenario. In the more general setting, we argue that MAESTRO decomposition should be preferable to SVD due to the following reasons:

- MAESTRO formulation is directly built into the training and tailored to obtain the best low-rank decomposition, while SVD relies on linearity assumption.
- SVD does not account for data, and even in the linear NN case, the learned singular vectors might exhibit wrong ordering. We demonstrate this issue using a simple example where we take matrix A with rank 3. We construct the dataset $\mathcal{X}$ in such a way that the third singular vector is the most important, the second one is the second, and the first is the third most important direction. Clearly, SVD does not look at data. Therefore, it cannot capture this phenomenon. We showcase that MAESTRO learns the correct order; see Fig. 5 of the Appendix.
- Pre-factorizing models allow us to apply hierarchical group-lasso penalty (Yuan & Lin, 2006) for decomposed weights to directly regularize the rank of different layers.
- SVD is computationally expensive and can only run rarely, while MAESTRO is directly built into the training and, therefore, does not require extra computations. In addition, MAESTRO supports rank sampling so training can be made computationally efficient.

5 EXPERIMENTS

We start this section by describing the setup of our experiments, including the models, datasets and baselines with which we compare MAESTRO. We then compare MAESTRO against the baselines on accuracy and training Multiply-Accumulate operations (MACs) and discuss the results. Subsequently, we analyze the behaviour of our system in-depth and provide additional insights on the performance of our technique, along with an ablation study and sensitivity analysis to specific hyperparameters. Finally, we showcase the performance of models upon deployment and how we can derive a smaller footprint model with some accuracy trade-off, without the need to fine-tune.

5.1 EXPERIMENTAL SETUP

Models & datasets. The datasets and models considered in our experiments span across four datasets, concisely presented along with the associated models on Tab. 1. We have implemented our solution with PyTorch (Paszke et al., 2017)(v1.13.0) trained our models on NVidia A100 (40G) GPUs. Details for the learning tasks and hyperparameters used are presented in the Appendix.

Table 1: Datasets and models for evaluation. The network footprints depict the vanilla variants of the models.

Dataset	Model	# GMACs	# Params (M)	Task
MNIST	LeNet	$2e^{-4}$	0.04	Image classification
CIFAR10	ResNet-18	0.56	11.18	Image classification
CIFAR10	VGG-19	0.40	20.00	Image classification
TinyImageNet	ResNet-50	5.19	53.9	Image classification
Multi30k	6-layer Transformer	1.37	48.98	Translation (en-ge)

Baselines. We have selected various baselines from the literature that we believe are closest to aspects of our system. On the *pruning* front, we compare with the IMP (Paul et al., 2023) and RareGems (Sreenivasan et al., 2022) techniques, themselves based on the LTH (Frankle & Carbin, 2019). On the *quantization* front, we compare with XNOR-Net (Rastegari et al., 2016). With respect to *low-rank* methods, we compare with Spectral Initialisation Khodak et al. (2021), Pufferfish (Wang et al., 2021) and Cuttlefish (Wang et al., 2023).

5.2 PERFORMANCE COMPARISON

We start off by comparing MAESTRO with various baselines from the literature across different datasets and types of models[3]. Results are depicted in Tab. 2 and 3, while additional performance points of MAESTRO for different model footprints are presented in the Appendix F.2 and F.3.

Comparisons with low-rank methods. The low-rank methods we are comparing against are Pufferfish (Wang et al., 2021) and Cuttlefish (Wang et al., 2023). These methods try to reduce training and inference runtime while preserving model accuracy by leveraging low-rank approximations. For ResNet-18, we achieve 94.19$\pm$0.07% for 4.08M parameters and 93.97$\pm$0.25% for 2.19M parameters compared to the 94.17% of Pufferfish at 3.3M parameters. For VGG-19, we achieve +0.41pp (percentage points) higher accuracy compared to Pufferfish and -0.29pp to Cuttlefish at 44.8% and

[3] The operating points we select for MAESTRO are the closest lower to the respective baseline in terms of footprint. Where the result is not present in the Tab. 2, we provide the λ_{gp} value so that it can be referenced from the Appendix, Tab. 11, 12.

Table 2: Maestro vs. baselines on CIFAR10.

Variant	Model	Acc. (%)	GMACs	Params. (M)
Non-factorized	ResNet-18	93.86$_{\pm 0.20}$	0.56	11.17
Pufferfish	ResNet-18	94.17	0.22	3.336
Cuttlefish	ResNet-18	93.47	0.3	3.108
IMP	ResNet-18	92.12	-	0.154
RareGems	ResNet-18	92.83	-	**0.076**
XNOR-Net	ResNet-18	90.06	-	0.349†
MAESTRO† ($\lambda_{gp} = 16e^{-6}$)	ResNet-18	**94.19**$_{\pm 0.07}$	0.39$_{\pm 0.00}$	4.08$_{\pm 0.02}$
MAESTRO† ($\lambda_{gp} = 64e^{-6}$)	ResNet-18	93.86$_{\pm 0.11}$	0.15$_{\pm 0.00}$	1.23$_{\pm 0.00}$
Non-factorized	VGG-19	92.94$_{\pm 0.17}$	0.40	20.56
Pufferfish	VGG-19	92.69	0.29	8.37
Cuttlefish	VGG-19	**93.39**	0.15	2.36
RareGems	VGG-19	86.28	-	5.04
IMP	VGG-19	92.86	-	5.04
XNOR-Net	VGG-19	88.94	-	0.64†
Spectral Init.*	VGG-19	83.27	-	≈ 0.4
MAESTRO† ($\lambda_{gp} = 32e^{-6}$)	VGG-19	93.10$_{\pm 0.10}$	0.13$_{\pm 0.00}$	2.20$_{\pm 0.03}$
MAESTRO† ($\lambda_{gp} = 512e^{-6}$)	VGG-19	88.53$_{\pm 0.13}$	**0.03**$_{\pm 0.00}$	**0.35**$_{\pm 0.00}$

*Results from original work; †: XNOR-Net employs binary weights and activations; although the overall #trainable parameters remain the same as the vanilla network, each model weight is quantized from 32-bit to 1-bit. Therefore, we report a compression rate of 3.125%($^{1}/_{32}$).

Table 3: Maestro vs. baselines on Multi30k.

Variant	Model	Perplexity	GMACs	Params. (M)
Non-factorized	Transformer	9.85$_{\pm 0.10}$	1.370	53.90
Pufferfish*	Transformer	7.34$_{\pm 0.12}$	0.996	26.70
MAESTRO†	Transformer	**6.90**$_{\pm 0.07}$	**0.248**$_{\pm 0.0032}$	**13.80**$_{\pm 0.113}$

*Results from original work; † tuned λ_{gp} from $\{2^i/100; i \in 0, \ldots, 9\}$

Table 4: Ablation study for ResNet18 on CIFAR10

Variant	Acc. (%)	GMACs	Params. (M)
MAESTRO	94.19$_{\pm 0.39}$	0.39$_{\pm 0.0008}$	4.08$_{\pm 0.020}$
w/out GL	94.04$_{\pm 0.10}$	0.56$_{\pm 0.0000}$	11.2$_{\pm 0.000}$
w/out PS	94.12$_{\pm 0.36}$	0.39$_{\pm 0.0010}$	4.09$_{\pm 0.027}$
w/ full-training	94.05$_{\pm 0.32}$	0.39$_{\pm 0.0004}$	4.09$_{\pm 0.032}$

93.2% of the sizes, respectively. Finally, comparing with the spectral initialization (Khodak et al., 2021) for VGG-19, we achieve +5.26pp higher accuracy for 87.5% of parameter size. Detailed results are shown in Tab. 2. This performance benefits also apply in the case of Transformers (Tab. 3), where MAESTRO performs 6% better in terms of perplexity at 25% of the cost (MACs) and 51.7% of the size (parameters) compared to Pufferfish.

Comparisons with pruning methods. The next family of baselines is related to the LTH (Frankle & Carbin, 2019). Specifically, we compare against IMP (Paul et al., 2023) and witness from Tab. 2 that MAESTRO can achieve +1.25pp ($\lambda_{gp} = 128e^{-6}$) and +0.24pp ($\lambda_{gp} = 32e^{-6}$) higher accuracy for ResNet-18 and VGG-19 respectively. Although we cannot scale to the size that RareGems (Sreenivasan et al., 2022) for ResNet-18, the sparsity that they achieve is unstructured, which most modern hardware cannot take advantage of. In contrast, our technique performs ordered structured sparsity, compatibly with most computation targets. On the other hand, for VGG-19, we achieve +6.82pp higher accuracy at 43.6% of the footprint.

Comparisons with quantized models. We also compare against XNOR-Net (Rastegari et al., 2016), which binarizes the network to achieve efficient inference. Training continues to happen in full precision, and inference performance is dependent on the operation implementation of the target hardware. Nonetheless, assuming a compression rate of 3.125%, for the same model size, we achieve +1.08pp ($\lambda_{gp} = 512e^{-6}$) and +2.18pp ($\lambda_{gp} = 256e^{-6}$) higher accuracy on ResNet-18 and VGG-19.

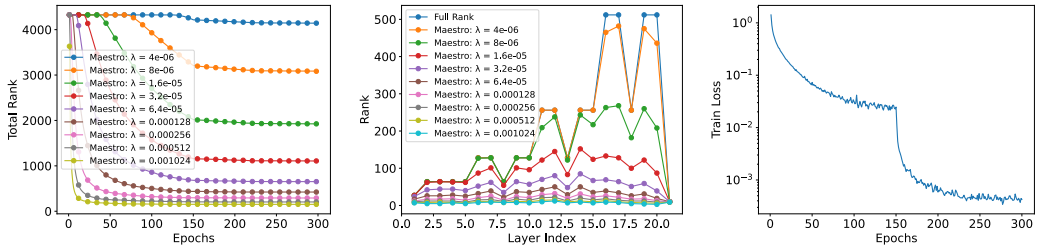
5.3 TRAINING BEHAVIOUR OF MAESTRO

Having shown the relative performance of our framework to selected baselines, we now move to investigate how our method behaves, with respect to its convergence and low-rank approximations.

Model and rank convergence. In Fig. 3, we present the training dynamics for MAESTRO. Fig. 3a illustrates the evolution of total rank throughout the training steps. We observe that the ranks are pruned incrementally. This aligns with the observations made during Pufferfish Wang et al. (2021) training, where the authors suggest warm-start training with full precision to enhance the final model performance. In our situation, we do not need to integrate this heuristic because MAESTRO automatically prunes rank. Fig. 3b reveals the ranks across layers after training. We notice an intriguing phenomenon: the ranks are nested for increasing λ_{gl}. This could imply apart from a natural order of ranks within each layer, a global order. We briefly examine this captivating occurrence in the following section, and we plan to investigate it more thoroughly in future work, as we believe this might contribute to a superior rank selection and sampling process. Lastly, Fig. 3c depicts the progression of training loss. We find that our hypothesis, that sampling does not adversely impact training, is also supported empirically.

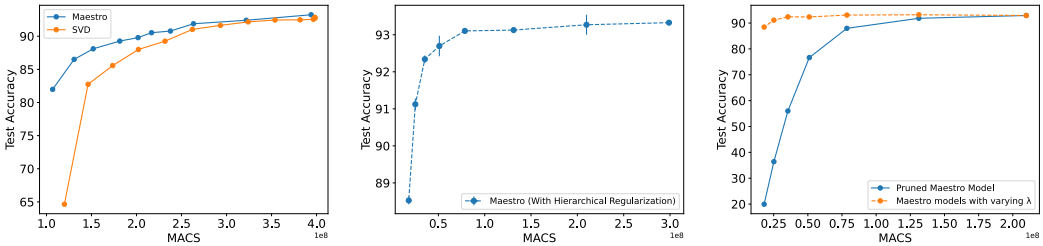
5.4 ABLATION STUDY

In this section, we examine the impact of each component on the performance of MAESTRO. Specifically, we run variants of our method *i)* without the *hierarchical group lasso regularization (HGL)*, *ii)* without progressive *shrinking (PS)*. Additionally, we integrate *iii)* an *extra full low-rank pass* ($b = r_i$) into the training at each step to assess whether extra sampling would be beneficial.



(a) Total rank ($\sum_{i=1}^{d} r_i$). **(b)** Ranks r_i's after training. **(c)** Convergence for $\lambda_{gl} = 0$.

Figure 3: Training dynamics of MAESTRO for ResNet18 on CIFAR10.



(a) MAESTRO vs. SVD. **(b)** Varying HGL. **(c)** Nested MAESTRO.

Figure 4: Accuracy-latency trade-off of MAESTRO under different settings for VGG19 on CIFAR10.

The results are displayed in Tab. 4. As anticipated, our findings confirm that neither inclusion of hierarchical group lasso with a tuned λ_{gl} nor progressive shrinking impair the final performance, but they do significantly enhance the efficiency of MAESTRO. Moreover, sampling more ranks at each training step does not improve the final performance, and, in fact, it hampers training efficiency, making it approximately twice as computationally demanding.

5.5 ACCURACY-LATENCY TRADE-OFF AT TRAINING AND DEPLOYMENT TIME

In Fig. 4, we illustrate various approaches to balance latency (proxied through MACs operations) and accuracy in model training and deployment. Fig. 4a demonstrates how MAESTRO ($\lambda_{gl} = 0$) can be pruned effectively for deployment using the greedy search method discussed in Section 3.4. We contrast this with the greedy pruning of a non-factorized model that has been factorized using SVD. We reveal that this straightforward baseline does not measure up to the learned decomposition of MAESTRO and results in a significant performance decrease. Next, Fig. 4b portrays the final accuracy and the number of model parameters for varying hierarchical group lasso penalties. This leads to the optimal latency-accuracy balance for both training and inference. However, it's crucial to point out that each model was trained individually, while greedy pruning only necessitates a single training cycle. Lastly, we delve into the observation of nested ranks across increasing λ_{gl}. Fig. 4c displays the performance of MAESTRO ($\lambda_{gl} = 0$) across different ranks selected by smaller models MAESTRO ($\lambda_{gl} > 0$). Intriguingly, we observe that MAESTRO ($\lambda_{gl} = 0$) performs very well—for instance, we can decrease its operations in half (and parameters by $10\times$) and still maintain an accuracy of 87.7% without fine-tuning, just by reusing rank structure from independent runs. As aforementioned, we intend to further explore this in the future.

6 CONCLUSION AND FUTURE WORK

In this work, we have presented MAESTRO, a method for trainable low-rank approximation of DNNs that leverages progressive shrinking by applying a generalized variant of Ordered Dropout to the factorized weights. We have shown the theoretical guarantees of our work in the case of linear models and empirically demonstrated its performance across different types of models, datasets, and modalities. Our evaluation has demonstrated that MAESTRO outperforms competitive compression methods at a lower cost. In the future, we plan to expand our technique to encompass more advanced sampling techniques and apply it to different distributed learning scenarios, such as Federated Learning, where data are natively non-independent or identically distributed (non-IID).

REFERENCES

Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=OtxyysUdBE.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. *arXiv preprint arXiv:1809.10505*, 2018.

Mario Almeida, Stefanos Laskaridis, Abhinav Mehrotra, Lukasz Dudziak, Ilias Leontiadis, and Nicholas D Lane. Smart at what cost? characterising mobile deep neural networks in the wild. In *Proceedings of the 21st ACM Internet Measurement Conference*, pp. 658–672, 2021.

Sebastian Caldas, Jakub Konečný, Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements, 2019. URL https://openreview.net/forum?id=SJlpM3RqKQ.

Miguel A Carreira-Perpinán and Yerlan Idelbayev. "learning-compression" algorithms for neural net pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8532–8541, 2018.

Tianyi Chen, Bo Ji, Tianyu Ding, Biyi Fang, Guanyi Wang, Zhihui Zhu, Luming Liang, Yixin Shi, Sheng Yi, and Xiao Tu. Only train once: A one-shot neural network training and pruning framework. *Advances in Neural Information Processing Systems*, 34:19637–19651, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Enmao Diao, Jie Ding, and Vahid Tarokh. Hetero{fl}: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=TNkPBBYFkXg.

Łukasz Dudziak, Mohamed S Abdelfattah, Ravichander Vipperla, Stefanos Laskaridis, and Nicholas D Lane. Shrinkml: End-to-end asr model compression using reinforcement learning. *INTERSPEECH*, 2019.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30k: Multilingual english-german image descriptions. pp. 70–74, 2016.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 1389–1397, 2017.

Samuel Horváth, Aaron Klein, Peter Richtárik, and Cédric Archambeau. Hyperparameter transfer learning with adaptive complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1378–1386. PMLR, 2021.

Samuel Horváth, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34:12876–12889, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014.

Mikhail Khodak, Neil Tenenholtz, Lester Mackey, and Nicolo Fusi. Initialization and regularization of factorized neural layers. *arXiv preprint arXiv:2105.01029*, 2021.

Minjae Kim, Sangyoon Yu, Suhyun Kim, and Soo-Mook Moon. DepthFL : Depthwise federated learning for heterogeneous clients. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pf8RIZTMU58.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Stefanos Laskaridis, Alexandros Kouris, and Nicholas D Lane. Adaptive inference through early-exit networks: Design, challenges and directions. In *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning*, pp. 1–6, 2021.

Stefanos Laskaridis, Stylianos I Venieris, Alexandros Kouris, Rui Li, and Nicholas D Lane. The future of consumer edge-ai computing. *arXiv preprint arXiv:2210.10514*, 2022.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. On-device training under 256kb memory. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

Zicheng Liu, Da Li, Javier Fernandez-Marques, Stefanos Laskaridis, Yan Gao, Łukasz Dudziak, Stan Z Li, Shell Xu Hu, and Timothy Hospedales. Federated learning for inference at anytime and anywhere. *arXiv preprint arXiv:2212.04084*, 2022.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, 2015.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.

Mansheej Paul, Feng Chen, Brett W. Larsen, Jonathan Frankle, Surya Ganguli, and Gintare Karolina Dziugaite. Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xSsW2Am-ukZ.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*, pp. 525–542. Springer, 2016.

Oren Rippel, Michael Gelbart, and Ryan Adams. Learning Ordered Representations with Nested Dropout. In *International Conference on Machine Learning (ICML)*, pp. 1746–1754, 2014.

Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6655–6659. IEEE, 2013.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth annual conference of the international speech communication association*, 2014.

Hakim Sidahmed, Zheng Xu, Ankush Garg, Yuan Cao, and Mingqing Chen. Efficient and private federated learning with partially trainable networks. *arXiv preprint arXiv:2110.03450*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Kartik Sreenivasan, Jy yong Sohn, Liu Yang, Matthew Grinde, Alliot Nagle, Hongyi Wang, Eric Xing, Kangwook Lee, and Dimitris Papailiopoulos. Rare gems: Finding lottery tickets at initialization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Jpxd93u2vK-.

Ananda Theertha Suresh, X Yu Felix, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *International Conference on Machine Learning*, pp. 3329–3337. PMLR, 2017.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Erwei Wang, James J Davis, Ruizhe Zhao, Ho-Cheung Ng, Xinyu Niu, Wayne Luk, Peter YK Cheung, and George A Constantinides. Deep Neural Network Approximation for Custom Hardware: Where we've been, where we're going. *ACM Computing Surveys (CSUR)*, 52(2):1–39, 2019.

Hongyi Wang, Saurabh Agarwal, and Dimitris Papailiopoulos. Pufferfish: communication-efficient models at no extra cost. *Proceedings of Machine Learning and Systems*, 3:365–386, 2021.

Hongyi Wang, Saurabh Agarwal, Yoshiki Tanaka, Eric P Xing, Dimitris Papailiopoulos, et al. Cuttlefish: Low-rank model training without all the tuning. *arXiv preprint arXiv:2305.02538*, 2023.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Simon Wiesler, Alexander Richard, Ralf Schlüter, and Hermann Ney. Mean-normalized stochastic gradient for large-scale deep learning. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 180–184. IEEE, 2014.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022.

Jian Xue, Jinyu Li, and Yifan Gong. Restructuring of deep neural network acoustic models with singular value decomposition. In *Interspeech*, pp. 2365–2369, 2013.

Tien-Ju Yang, Yu-Hsin Chen, and Vivienne Sze. Designing energy-efficient convolutional neural networks using energy-aware pruning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5687–5695, 2017.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439, 2021.

Jiahui Yu and Thomas Huang. Autoslim: Towards one-shot architecture search for channel numbers. *arXiv preprint arXiv:1903.11728*, 2019a.

Jiahui Yu and Thomas S Huang. Universally slimmable networks and improved training techniques. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1803–1811, 2019b.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gMCsAqY7.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.