Structured Sparsification with Joint Optimization of Group Convolution and Channel Shuffle

Xin-Yu Zhang[*1] **Kai Zhao**[*1] **Taihong Xiao**[2] **Ming-Ming Cheng**[1] **Ming-Hsuan Yang**[2]

[1]TKLNDST, CS, Nankai University
[2]University of California, Merced

Preliminary version

Abstract

Recent advances in convolutional neural networks (CNNs) usually come with the expense of excessive computational overhead and memory footprint. Network compression aims to alleviate this issue by training compact models with comparable performance. However, existing compression techniques either entail dedicated expert design or compromise with a moderate performance drop. In this paper, we propose a novel structured sparsification method for efficient network compression. The proposed method automatically induces structured sparsity on the convolutional weights, thereby facilitating the implementation of the compressed model with the highly-optimized group convolution. We further address the problem of inter-group communication with a learnable channel shuffle mechanism. The proposed approach can be easily applied to compress many network architectures with a negligible performance drop. Extensive experimental results and analysis demonstrate that our approach gives a competitive performance against the recent network compression counterparts with a sound accuracy-complexity trade-off.

1 INTRODUCTION

Convolutional Neural Networks (CNNs) have made significant advances in a wide range of vision and learning tasks [Russakovsky et al., 2015, Long et al., 2015]. However, the performance gains usually entail a heavy computational cost, which makes the deployment of CNNs on portable devices difficult. To meet the memory and computational constraints in real-world applications, numerous model compression techniques have been developed.

Existing network compression techniques are mostly based on weight quantization [Chen et al., 2015, Courbariaux et al., 2016, Rastegari et al., 2016, Wu et al., 2016], knowledge distillation [Hinton et al., 2015, Chen et al., 2017, Yim et al., 2017], or network pruning [Li et al., 2017, He et al., 2017, Liu et al., 2017, Molchanov et al., 2019, Tang et al., 2020]. Weight quantization methods use low bit-width numbers to represent weights and activations, which usually bring a moderate performance degradation. Knowledge distillation schemes transfer knowledge from a large teacher network to a compact student network, which are typically susceptible to the teacher/student network architectures [Mirzadeh et al., 2020, Liu et al., 2019b]. Closely related to our work, network pruning approaches reduce the model size by removing a proportion of model parameters that are considered unimportant. Notably, filter pruning algorithms [Li et al., 2017, Liu et al., 2017, Molchanov et al., 2019, Tang et al., 2020] remove the entire filters and result in structured architectures that can be readily incorporated into modern BLAS libraries.

Identifying unimportant filters is critical to the filter pruning methods. It is well-known that the weight norm can serve as a good indicator of the corresponding filter importance [Li et al., 2017, Liu et al., 2017]. Filters corresponding to smaller weight norms are considered to contribute less to the outputs. Furthermore, the L_1 regularization can be used to promote sparsity [Liu et al., 2017]. However, there are still several issues in the existing pruning methods: 1) pruning a large proportion of convolutional filters will result in severe performance degradation; 2) pruning alters the input/output feature dimensions, and thus meticulous adaptation is required to handle network architectures with shortcut connections (*e.g.,* residual connections [He et al., 2016] and dense connections [Huang et al., 2017]).

Before presenting the proposed method, we briefly introduce the group convolution (GroupConv) [Russakovsky et al., 2015], which plays an important role in this work. For the vallina convolution operation, the output features are densely-connected with the input features, while for the

[*]Xin-Yu Zhang (xinyuzhang@mail.nankai.edu.cn) and Kai Zhao contribute equally to this work.

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

GroupConv, the input features are equally split into several groups and transformed within each group independently. Essentially, each output channel is connected with only a proportion of the input channels, which leads to sparse neuron connections. Therefore, deep CNNs with GroupConvs can be trained on less powerful GPUs with smaller amount of memory.

In this work, we propose a novel approach for network compression where unimportant neuron connections are pruned to facilitate the usage of GroupConvs. Nevertheless, converting vallina convolutions into GroupConvs is a challenging task. First, not all sparse neuron connectivities correspond to valid GroupConvs, while certain requirements must be satisfied, *e.g.,* mutual exclusiveness of different groups. To guarantee the desired structured sparsity, we impose *structured regularization* upon the convolutional weights and zero out the sparsified weights. Another challenge is that stacking multiple GroupConvs sequentially will hinder the inter-group information flow. Zhang et al. [2018] propose the ShuffleNet with a *channel shuffle* mechanism, *i.e.,* gathering channels from distinct groups, to ensure the inter-group communication, though the order of permutation is handcrafted. However, we solve the problem of channel shuffle in a learning-based scheme. Concretely, we formulate the learning of channel shuffle as a linear programming problem, which can be solved by efficient algorithms like the network simplex method [Bonneel et al., 2011]. Since the structured sparsity is induced among the convolutional weights, our method is nominated as *structured sparsification*.

The proposed structured sparsification method is designed for three goals. (a) Our approach can better handle those network architectures with shortcut connections. A wide range of backbone architectures are amenable to our method without the need for any special adaptation. (b) Our method is capable of achieving *high compression rates*. In modern efficient network architectures, the complexity of 3×3 convolutions is highly compressed, while the computation bottleneck becomes the point-wise convolutions (*i.e.,* 1×1 convolutions) [Zhang et al., 2018]. For example, the point-wise convolutions occupy 81.5% of the total FLOPs in the MobileNet-V2 [Sandler et al., 2018] backbone and 93.4% in ResNeXt [Xie et al., 2017]. Our method is applicable to all convolution operators so that a high compression rate is reachable. (c) Our approach brings *negligible performance drop*. As all filters are preserved under our methodology, we retain stronger representational capacity of the compressed model and achieve better accuracy-complexity trade-off than the pruning-based counterparts (see Fig. 3).

The main contributions of this work are as follows.

- We propose a *learnable channel shuffle mechanism* in which the permutation of the convolutional weight norm matrix is learned via linear programming;
- We formulate a novel *structured sparsification* framework for efficient network compression, which unifies weight pruning, GroupConv, and the learnable channel shuffle;
- The experimental results on the CIFAR-10 and ImageNet datasets demonstrate that the proposed structured sparsification performs well against the concurrent filter pruning approaches with a balanced trade-off of accuracy and complexity.

2 RELATED WORK

Network Compression. Compression methods for deep models can be broadly categorized based on *weight quantization*, *knowledge distillation*, or *network pruning*. Closely related to our work are the network pruning approaches based on *filter pruning*. It is well-acknowledged that filters with smaller weight norms are considered to make negligible contribution to the outputs and can be pruned. Li et al. [2017] prune filters according to the L_1 norm of the convolutional weights, while Liu et al. [2017] prune models with batch normalization [Ioffe and Szegedy, 2015] by removing the channels with smaller batch-norm scaling factors. An L_1 regularization term is further imposed on these scaling factors to promote sparsity.

However, removing those filters corresponding to smaller weight norms may significantly reduce the representational capacity. Instead, we propose a structured sparsification method that enforces structured sparsity among neuron connections and merely removes certain unimportant connections while the entire filters are preserved. As a result, the network capacity is less affected than the pruning-based approaches [Li et al., 2017, Liu et al., 2017, Molchanov et al., 2019, Tang et al., 2020]. Furthermore, our method does not alter the input/output dimensions, and can be easily incorporated into numerous backbones.

Group Convolution. Group convolution (GroupConv) is introduced in the AlexNet [Russakovsky et al., 2015] to overcome the GPU memory constraints. GroupConv partitions the input features into mutually exclusive groups and transforms the features within each group in parallel. Compared with the vallina (*i.e.,* densely connected) convolution, a GroupConv with G groups can reduce the computational cost and number of parameters by a factor of G. The ResNeXt [Xie et al., 2017] designs a multi-branch architecture by employing GroupConvs and defines the *cardinality* as the number of parallel transformations, which is simply the group number in each GroupConv. If the cardinality equals to the number of channels, GroupConv becomes the *depthwise separable convolution*, which is widely used in recent lightweight neural architectures [Howard et al., 2017, Sandler et al., 2018, Zhang et al., 2018, Ma et al., 2018, Chollet, 2017].

However, the aforementioned methods all treat the cardi-

Preliminary version

nality as a hyper-parameter, and the connectivity patterns between consecutive features are hand-crafted as well. On the other hand, there is also a line of research focusing on learnable GroupConvs [Huang et al., 2018, Wang et al., 2019, Zhang et al., 2019]. Both CondenseNet [Huang et al., 2018] and FLGC [Wang et al., 2019] pre-define the cardinalities of GroupConvs and learn the connectivity patterns. We note that the work by Zhang et al. [2019] learns the cardinality and neuron connectivity simultaneously. Essentially, this dynamic grouping convolution is modeled by a binary relationship matrix $\boldsymbol{U}$ where U_{ji} indicates the connectivity between the ith input channel and the jth output channel. To guarantee that the resulting operator is a valid GroupConv, the relationship matrix is constructed using a Kronecker product of several binary symmetric 2×2 matrices, which is a sufficient but unnecessary condition. Consequently, the space of all valid GroupConvs is not fully exploited.

In contrast to the prior art [Zhang et al., 2019], our method can learn the optimal connection over all possible neuron connectivity patterns, thus resulting in better model structures and performance. The connection between two layers is represented by the composition of a permutation matrix (over the permutation set) and the convolutional weights. The importance of each neuron connection is quantified by the corresponding entry of the weight norm matrix.

Channel Shuffle Mechanism. The ShuffleNet [Zhang et al., 2018] combines the channel shuffle mechanism with GroupConv for efficient network design, in which channels from different groups are gathered so as to facilitate the inter-group communication. Without channel shuffle, stacking multiple GroupConvs will eliminate the information flow among different groups and weaken the representational capacity. Different from the hand-crafted counterpart [Zhang et al., 2018], the proposed channel shuffle operation is learnable over the space of all possible channel permutations. Furthermore, without bells and whistles, our channel shuffle only involves a simple permutation along the channel dimension, which can be conveniently implemented by an index operation.

Neural Architecture Search. Neural Architecture Search (NAS) [Zoph and Le, 2017, Baker et al., 2017, Zoph et al., 2018, Real et al., 2019, Wu et al., 2019] aims to automate the process of designing neural architectures within certain budgets of computational resources. Existing NAS algorithms are developed based on reinforcement learning [Zoph and Le, 2017, Baker et al., 2017, Zoph et al., 2018], evolutionary search [Real et al., 2019], and differentiable approaches [Liu et al., 2019a, Wu et al., 2019]. Our method can be viewed as a special case of hyper-parameter (*i.e.,* cardinality) optimization and neuron connectivity search. However, different from existing approaches evaluated on numerous architectures, the proposed method can determine the compressed architecture in one single training pass. We highlight that the efficiency of our approach is in accordance with the aim of neural architecture search.

3 STRUCTURED SPARSIFICATION

3.1 OVERVIEW

The structured sparsification method is designed to zero out a proportion of the convolutional weights so that the vanilla convolutions can be structured into group convolutions (GroupConvs) via a learned permutation matrix. We adopt the "train, compress, finetune" pipeline, in a way similar to the recent pruning approaches [Liu et al., 2017]. Concretely, we first train a large model under the structured regularization, then compress vanilla convolutions into GroupConvs under a certain criteria, and finally finetune the compressed model to recover accuracy. To this end, three issues need to be addressed: 1) how to learn the connectivity patterns (Sec. 3.2); 2) how to design the structured regularization (Sec. 3.3); 3) how to decide the grouping criteria (Sec. 3.4). Additional details of our pipeline are presented in Sec. 3.5.

3.2 LEARNING CONNECTIVITY WITH LINEAR PROGRAMMING

Let $\boldsymbol{F} \in \mathbb{R}^{C^{\text{in}} \times H \times W}$ be the input feature map, where C^{in} denotes the number of input channels. We apply a vallina convolution[1] with weights $\boldsymbol{W} \in \mathbb{R}^{C^{\text{out}} \times C^{\text{in}} \times K \times K}$ to $\boldsymbol{F}$, *i.e.,* $\boldsymbol{O} = \boldsymbol{W} * \boldsymbol{F}$, where $\boldsymbol{O} \in \mathbb{R}^{C^{\text{out}} \times H \times W}$ with C^{out} denoting the number of output channels. Each entry of $\boldsymbol{O}$ is a weighted sum of a local patch of $\boldsymbol{F}$, namely,

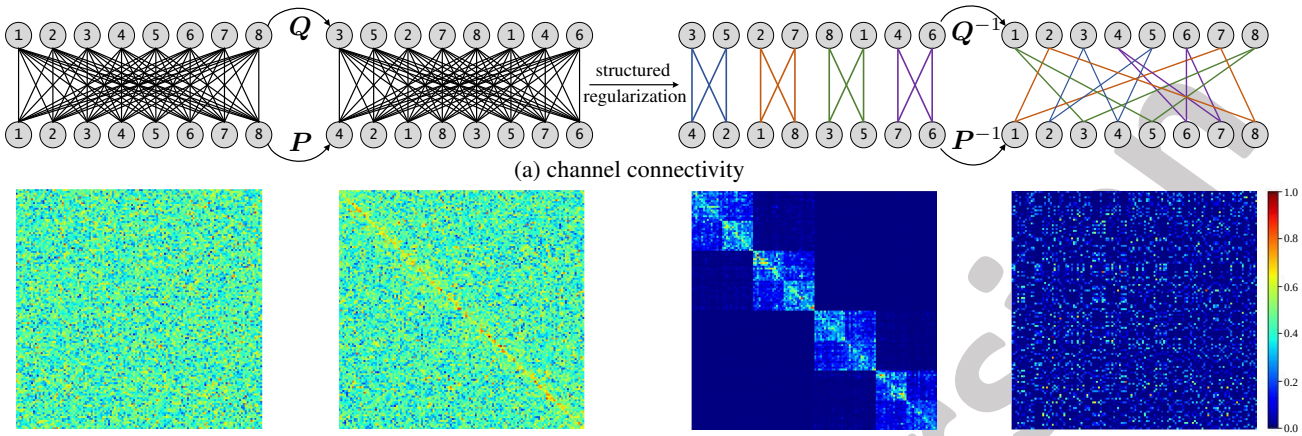
$$O_{j,p,q} = \sum_{i,k,l} W_{j,i,k,l} F_{i,p+k,q+l}. \tag{1}$$

In Eq. (1), the ith channel of $\boldsymbol{F}$ relates to the jth channel of $\boldsymbol{O}$ via weights $\boldsymbol{W}_{j,i,:,:}$. Motivated by the norm-based importance estimation in filter pruning [Li et al., 2017, Liu et al., 2017], we quantify the importance of the connection between $\boldsymbol{F}_i$ and of $\boldsymbol{O}_j$ by $\|\boldsymbol{W}_{j,i,:,:}\|$. Thus, the importance matrix $\boldsymbol{S} \in \mathbb{R}^{C^{\text{out}} \times C^{\text{in}}}$ can be defined as the norm along the "kernel size" dimensions of $\boldsymbol{W}$, *i.e.,* $S_{j,i} = \|\boldsymbol{W}_{j,i,:,:}\|$.

Next, we extend our formulation to GroupConvs with cardinality G. A GroupConv can be considered as a convolution with sparse neuron connectivity, in which only a proportion of input channels is visible to each output channel. Without loss of generality, we assume both C^{in} and C^{out} are divisible

[1] For simplicity, we omit the bias term from Eq. (1), and assume the convolution operator is of stride 1 with proper paddings.

Preliminary version



(a) channel connectivity

(b) weight norm matrix

Figure 1: Illustration of the learnable channel shuffle mechanism. The original convolutional weights (first column) are shuffled along the input/output channel dimensions in order to solve Eq. (4). The structured regularization is imposed upon the permuted weight norm matrix (second column) to promote the structured sparsity, and connections with small weight norms are discarded (third column). In the original ordering of channels, a structurally sparse connectivity pattern is learned (fourth column), and notably every valid connectivity pattern can be possibly reached in this manner.

by G, and Eq. (1) can be adapted as

$$O_{j,p,q} = \sum_{i=(n-1)m+1}^{nm} \sum_{k,l} W_{j,i,k,l} F_{i,p+k,q+l}, \qquad (2)$$

where $n = \text{ceil}(jG/C^{\text{out}})$ indicates the jth output channel belongs to the nth group, and $m = C^{\text{in}}/G$ is the number of input channels within each group. Clearly, the valid entries of $\boldsymbol{W}$ form a block diagonal matrix with G equally-split blocks at the input/output channel dimensions. Thus, the GroupConv module requires $C^{\text{in}}C^{\text{out}}K^2/G$ parameters and $C^{\text{in}}C^{\text{out}}K^2HW/G$ FLOPs for processing the feature $\boldsymbol{F}$, and the computational complexity is reduced by a factor of G compared with the vanilla counterpart.

We note that if a vanilla convolution operator can be converted into GroupConv without affecting its functional property (we call such convolution operators *groupable*), the convolutional weights $\boldsymbol{W}$ must be block diagonal after certain permutations along the input/output channel dimensions. Due to the positive definiteness property of the norm, a necessary and sufficient condition of a convolution operator $\boldsymbol{W}$ being groupable is that

$$\begin{aligned} &\exists\, \boldsymbol{P} \in \mathcal{P}^{C^{\text{out}}} \text{ and } \boldsymbol{Q} \in \mathcal{P}^{C^{\text{in}}}, \\ &\text{s.t. } \boldsymbol{PSQ} \text{ is block diagonal with equally-split blocks,} \end{aligned} \qquad (3)$$

where $\mathcal{P}^N$ denotes the set of $N \times N$ permutation matrices. Here, the permutation matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ shuffle the channels of the input and output features, and thus determine the connectivity pattern between $\boldsymbol{F}$ and $\boldsymbol{O}$ (see Fig. 1).

However, a randomly initialized and trained convolution operator by no means can be groupable unless sparsity constraints are imposed. To this end, we resort to permuting $\boldsymbol{S}$ so as to make $\boldsymbol{S}' = \boldsymbol{PSQ}$ "as block diagonal as possible". The next question is how to rigorously characterize the term "as block diagonal as possible". Here, we assume both C^{in} and C^{out} are powers of 2, where the most widely-used backbone architectures (*e.g.,* VGG [Simonyan and Zisserman, 2015] and ResNet [He et al., 2016]) satisfy this assumption[2]. Then, the potential cardinality is also a power of 2. As the cardinality grows, more and more non-diagonal blocks are zeroed out (see Fig. 2(c)). As illustrated in Fig. 2(b), we define the cost matrix $\boldsymbol{R}$ to progressively penalize the non-zero entries of the non-diagonal blocks. Finally, we utilize $\boldsymbol{S}' \otimes \boldsymbol{R}$ as a metric of the "block-diagonality" of the matrix $\boldsymbol{S}'$, where $\otimes$ indicates element-wise multiplication and summation over all entries, *i.e.,* $\boldsymbol{A} \otimes \boldsymbol{B} = \sum_{i,j} A_{i,j}B_{i,j}$. Therefore, we can give the optimal connectivity pattern between the adjacent layers by optimizing the following:

$$\begin{aligned} &\min_{\boldsymbol{P},\boldsymbol{Q}} \ \boldsymbol{PSQ} \otimes \boldsymbol{R} \\ &\text{s.t. } \boldsymbol{P} \in \mathcal{P}^{C^{\text{out}}} \text{ and } \boldsymbol{Q} \in \mathcal{P}^{C^{\text{in}}}. \end{aligned} \qquad (4)$$

However, the optimization over the set of permutation matrices is a non-convex and NP-hard problem, which requires combinatorial search. To overcome the difficulty, we relax the feasible space to its convex hull and alternatively optimize $\boldsymbol{P}$ and $\boldsymbol{Q}$. The Birkhoff-von Neumann theorem [Birkhoff, 1946] states that the convex hull of the set of permutation matrices is the set of doubly-stochastic matrices[3],

[2] Similar reasoning can be applied if both C^{in} and C^{out} have many factors of 2. (See appendix A in the supplementary material.)

[3] Doubly-stochastic matrices are non-negative square matrices whose rows and columns sum to one.

Preliminary version

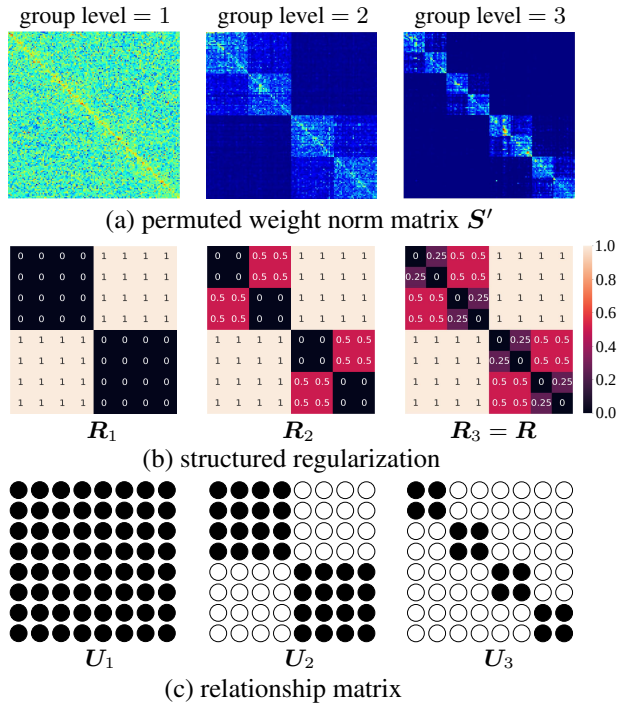


Figure 2: Illustration of the structured regularization matrix $\boldsymbol{R}_g$ and the relationship matrix $\boldsymbol{U}_g$ corresponding to the group level g. (a) Heat map of the permuted weight norm matrix $\boldsymbol{S}'$. Non-diagonal blocks of the weight norm are sparsified. (b) Structured regularization matrix $\boldsymbol{R}_g$. The regularization coefficient decays exponentially as the group level grows. A special case of the decay rate of 0.5 is demonstrated. Besides, the matrix $\boldsymbol{R}_g$ depends on the current group level g, and when the maximal possible group level is achieved, the matrix $\boldsymbol{R}_g$ becomes the cost matrix $\boldsymbol{R}$ in Eq. (4); (c) Relationship matrix $\boldsymbol{U}_g$. The entries of the permuted weight norm matrix corresponding to the zero entries of the relationship matrix will be zeroed out during grouping.

known as the *Birkhoff polytope*:

$$\mathcal{B}^N = \{\boldsymbol{X} \in \mathbb{R}_+^{N\times N} : \boldsymbol{X}\mathbf{1}_N = \mathbf{1}_N,\ \boldsymbol{X}^\top \mathbf{1}_N = \mathbf{1}_N\}, \quad (5)$$

where $\mathbf{1}_N$ denotes the column vector composed of N ones. Then, we alternatively optimize $\boldsymbol{P}$ and $\boldsymbol{Q}$ on the Birkhoff polytope until convergence. For example, when optimizing $\boldsymbol{P}$, the objective can be formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{P}}\ \boldsymbol{P} \otimes \boldsymbol{R}\boldsymbol{Q}^\top \boldsymbol{S}^\top \\ &\text{s.t. } \boldsymbol{P} \in \mathcal{B}^{C^{\text{out}}}. \end{aligned} \quad (6)$$

Similarly, we can have the objective for optimizing $\boldsymbol{Q}$.

Note that the Birkhoff polytope is a *simplex*. Therefore, the linear programming problem in Eq. (6) can be solved by efficient linear programming algorithms such as the network simplex method [Bonneel et al., 2011]. In addition, the theory of linear programming guarantees that at least one of the solutions is achieved at the vertex of the simplex, and the vertices of the Birkhoff polytope are precisely the permutation matrices [Birkhoff, 1946]. Thus the solution to Eq. (6) is a naturally a permutation matrix without the need of any post-processing operation. Furthermore, Eq. (6) can be viewed equivalently as the *optimal transport* problem, where an established computation library[4] is available for efficient linear programming.

3.3 STRUCTURED REGULARIZATION

Permutation alone does not suffice to induce structurally sparse convolutional weights, and we still need to impose a sparsity regularization to achieve the desired sparsity structure. Inspired by the sparsity-inducing penalty in Liu et al. [2017], we impose the structured L_1 regularization on the permuted weight norm $\boldsymbol{S}' = \boldsymbol{PSQ}$. Since the cardinality should be a power of 2, suppose cardinality $= 2^{g-1}$ and here g is defined as the group level, as shown in Fig. 2. Given the group level g, the structured L_1 regularization is formulated as[5] $\mathcal{L}_{\text{reg}} = \boldsymbol{S}' \otimes \boldsymbol{R}_g$, where $\boldsymbol{R}_g$ denotes the structured regularization matrix as illustrated in Fig. 2(b). Intuitively, the proposed structured regularization aims to zero out the non-diagonal entries so as to make the permuted weight norm matrix $\boldsymbol{S}'$ as much sparsified as possible. Furthermore, the regularization coefficient decays exponentially as the group level grows as we desire balanced cardinality distribution among the network. In the end, the overall loss becomes

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda \mathcal{L}_{\text{reg}}, \quad (7)$$

where $\mathcal{L}_{\text{data}}$ denotes the regular data loss (standard classification loss in the following experiments) and λ is the balancing scalar coefficient.

3.4 CRITERIA OF LEARNING CARDINALITY

As the sparsity changes during the training process, we need to determine the cardinality based on the structurally sparsified convolutional weights. Since the weight norms corresponding to the valid connections constitute at least a certain proportion of the total weight norms, we set a threshold p to determine the group level g. Motivated by Zhao et al. [2020], the group level g can be determined by

$$g = \max\{g : \boldsymbol{S}' \otimes \boldsymbol{U}_g \geq p \sum_{i,j} S_{i,j},\ g = 1, 2, \cdots\}, \quad (8)$$

where p is a threshold set to 0.9 in all of the experiments and $\boldsymbol{U}_g$ is the relationship matrix [Zhang et al., 2019] as illus-

[4] https://github.com/rflamary/POT/

[5] For simplicity, we here compute the regularization of a single convolutional layer. In the experiments, the regularization is the summation of those of all the convolution layers.

Preliminary version

Algorithm 1: Training Pipeline.

Initially update the permutation matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$.
for $t := 1$ *to #epochs* **do**
  Train for 1 epoch with the structured regularization;
  Solve Eq. (4) to update the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$;
  Determine the current group levels g by Eq. (8);
  Update the structured regularization matrices;
  Adjust the coefficient λ.
end

trated in Fig. 2(c) and specifies the valid neuron connections at group level g.

3.5 IMPLEMENTATION DETAILS

Overall Configuration. Our implementation is based on the PyTorch library [Paszke et al., 2019]. The proposed method is applied to the ResNet [He et al., 2016] and DenseNet [Huang et al., 2017] families, and evaluated on the CIFAR-10 [Krizhevsky et al., 2009] and ImageNet [Russakovsky et al., 2015] datasets. For the CIFAR-10 dataset, we follow the common practice of data augmentation [He et al., 2016, Liu et al., 2017, Xie et al., 2017]: zero-padding of 4 pixels on each side of the image, random crop of a 32×32 patch, and random horizontal flip. For fair comparisons, we utilize the same network architecture as Liu et al. [2017], and the model is trained on a single GPU with a batch size of 64. For the ImageNet dataset, we adopt the standard data augmentation strategy [Simonyan and Zisserman, 2015, He et al., 2016, Xie et al., 2017]: image resize such that the shortest edge is of 256 pixels, random crop of a 224×224 patch, and random horizontal flip. The overall batch size is 256, which is distributed to 4 GPUs. For both datasets, we employ the SGD optimizer with momentum 0.9.

Training Protocol. For the first stage, we train a large model from scratch with the structured regularization as described in Sec. 3.3. At the end of each epoch, we update the permutation matrices as in Sec. 3.2, determine the current group levels as in Sec. 3.4, adjust the structured regularization matrices accordingly, and search for the coefficient λ to meet the desired compression rate as shown in the supplementary material (see appendix B). We train with a fixed learning rate of 0.1 for 100 epochs on the CIFAR-10 dataset and 60 epochs on the ImageNet dataset and exclude the weight decay due to the existence of the structured regularization. The training pipeline is summarized in Alg. 1.

Finetune Protocol. The remaining parameters are restored from the training stage and the compressed model is finetuned with an initial learning rate of 0.1. We finetune for 160 epochs on the CIFAR-10 dataset and the learning rate decays by a factor of 10 at 50% and 75% of the total epochs.

Table 1: Network compression results on the CIFAR-10 [Krizhevsky et al., 2009] dataset. "Baseline" means the network without compression. The percentages in our method indicate the compression rate (measured by the reduction of "#Params."), while those in other methods indicate the pruning ratio.

Methods	#Params.(10^5) ↓	FLOPs (10^7) ↓	Acc.(%) ↑
ResNet-20			
Baseline	2.20	3.53	91.70 (±0.12)
Slimming-40%	1.91 (±0.00)	**3.10** (±0.02)	91.74 (±0.35)
StrucSpars-20%	**1.76** (±0.00)	3.18 (±0.07)	**91.79** (±0.23)
Slimming-60%	1.36 (±0.02)	**2.24** (±0.01)	89.68 (±0.38)
StrucSpars-40%	**1.31** (±0.01)	2.58 (±0.00)	**91.42** (±0.04)
ResNet-56			
Baseline	5.90	9.16	93.50 (±0.19)
Slimming-60%	4.15 (±0.03)	5.75 (±0.10)	93.10 (±0.25)
StrucSpars-30%	4.08 (±0.05)	7.17 (±0.20)	**94.19** (±0.16)
StrucSpars-50%	**2.96** (±0.03)	**4.81** (±0.03)	93.70 (±0.06)
Slimming-80%	2.33 (±0.04)	**3.50** (±0.02)	91.01 (±0.02)
StrucSpars-60%	2.34 (±0.08)	4.20 (±0.08)	**93.48** (±0.13)
StrucSpars-70%	**1.80** (±0.00)	3.52 (±0.16)	93.25 (±0.02)
ResNet-110			
Baseline	11.47	17.59	94.62 (±0.22)
Slimming-40%	9.24 (±0.03)	12.55 (±0.00)	94.49 (±0.12)
StrucSpars-20%	9.12 (±0.06)	14.76 (±0.02)	**94.78** (±0.11)
StrucSpars-40%	**6.69** (±0.24)	**11.60** (±0.01)	94.55 (±0.18)
Slimming-60%	8.15 (±0.03)	**10.66** (±0.00)	94.29 (±0.11)
StrucSpars-30%	7.89 (±0.03)	12.47 (±0.01)	**94.69** (±0.08)
StrucSpars-60%	**5.41** (±0.02)	**10.66** (±0.01)	94.42 (±0.04)

On the ImageNet dataset, the learning rate is decayed according to the cosine annealing strategy [Loshchilov and Hutter, 2017] within 120 epochs. For both datasets, a standard weight decay of 10^{-4} is adopted to prevent overfitting.

4 EXPERIMENTS AND ANALYSIS

In this section, we present the experimental results on the CIFAR-10 and ImageNet datasets. In addition, we carry out ablation studies to demonstrate the effectiveness of components of the proposed method. Please refer to appendix C in the supplementary material for more detailed experimental results, and our source codes are available at https://github.com/Sakura03/StrucSpars.

4.1 RESULTS ON CIFAR-10

We first compare our proposed method with the Network Slimming [Liu et al., 2017] approach on the CIFAR-10 dataset. The Network Slimming approach is a representative filter pruning method that compresses CNNs by pruning less important filters. As the experimental results on the

Preliminary version

CIFAR-10 dataset are somewhat random, we repeat the train-compress-finetune pipeline for 10 times and record the mean and standard deviation (std). As shown in Tab. 1, the proposed structured sparsification performs favorably under various compression rates. For ResNet-110, with 60% parameters compressed, the structured sparsification can still achieve a 94.42% top-1 accuracy which is nearly equal to the performance of the baseline method without compression. Compared with the Network Slimming, our method consistently performs better, especially under high compression rates. Experiments on the CIFAR-10 dataset demonstrate that our method is able to compress CNNs with negligible performance drop and favorable accuracy against pruning methods such as Network Slimming.

4.2 RESULTS ON IMAGENET

Tab. 2 shows the evaluation results of the proposed method against the recent representative network pruning approaches, including ThiNet [Luo et al., 2017], Slimming [Liu et al., 2017], NISP [Yu et al., 2018], BN-ISTA [Ye et al., 2018], FPGM [He et al., 2019], Taylor [Molchanov et al., 2019], ABCPrunner [Lin et al., 2020b], HRank [Lin et al., 2020a], Hinge [Li et al., 2020], DMC [Gao et al., 2020], DSA [Ning et al., 2020], and SCOP [Tang et al., 2020]. Overall, the structured sparsification method performs favorably against the previous network compression methods under different settings. These performance gains achieved by our method can be attributed to the fact that discarding the entire filters will negatively affect the representational strength of the network model, especially when the pruning ratio is high, *e.g.,* 50%. In contrast, our method removes only a proportion of neuron connections and preserves all of the filters, thereby making a mild impact on the model capacity. In addition, it is known that pruning neuron connections would eliminate the information flow and affect performance. To alleviate this issue, the learnable channel shuffle mechanism assists the information exchange among different groups, thereby minimizing the potential negative impact.

4.3 ABLATION STUDIES

Accuracy *v.s.* Complexity. As shown in Fig. 3, the proposed structured sparsification is designed to make sound accuracy-complexity trade-off. On the ImageNet [Russakovsky et al., 2015] dataset, a slight top-1 accuracy drop of 0.28% is compromised for about 25% complexity reduction on the ResNet-50 backbone, and an accuracy loss of 1.02% for about 60% reduction on ResNet-101. Furthermore, high compression rates can be achieved in our methodology while maintaining competitive performance. It is worth noticing that our method achieves an accuracy of 72.47% with only about 20% complexity of the ResNet-50 backbone, which performs favorably against the pruning methods with two times complexity.

Table 2: Network compression results on the ImageNet [Russakovsky et al., 2015] dataset. The center-crop validation accuracy is reported. "Baseline" means the network without compression. The percentages in the table have the same meaning as those in Tab. 1.

Methods	#Params.(10^6) ↓	GFLOPs ↓	Acc.(%) ↑
ResNet-50			
Baseline	25.6	4.14	77.10
NISP-A	18.6	≈2.97	72.75
Slimming-20%	17.8	2.81	75.12
Taylor-19%	17.9	2.66	75.48
FPGM-30%	N/A	**2.39**	75.59
StrucSpars-35%	**17.2**	3.12	**76.82**
ThiNet-30%	16.9	≈2.62	72.04
NISP-B	14.3	≈2.29	72.07
ABCPrunner-80%	11.8	1.89	73.86
Taylor-28%	14.2	2.25	74.50
DSA-50%	N/A	2.07	74.69
Hinge-46%	N/A	≈1.93	74.70
FPGM-40%	N/A	1.93	74.83
HRank-36%	16.2	2.30	74.98
StrucSpars-65%	**10.3**	**1.67**	**75.10**
ThiNet-50%	12.4	≈1.83	71.01
Taylor-44%	7.9	1.34	71.69
HRank-46%	13.8	1.55	71.98
Slimming-50%	11.1	1.87	71.99
StrucSpars-85%	**5.6**	**0.90**	**72.47**
ResNet-101			
Baseline	44.5	7.87	78.64
FPGM-30%	N/A	**4.55**	77.32
Taylor-25%	31.2	4.70	77.35
StrucSpars-40%	**26.7**	5.05	**78.16**
BN-ISTA-v1	17.3	3.69	74.56
BN-ISTA-v2	23.6	4.47	75.27
ABCPrunner-80%	17.7	3.16	75.82
Taylor-45%	20.7	**2.85**	75.95
Slimming-50%	20.9	3.16	75.97
SCOP-B	18.8	3.13	77.36
DMC-56%	N/A	3.46	77.40
StrucSpars-65%	**16.5**	2.98	**77.62**
Taylor-60%	13.6	1.76	74.16
ABCPrunner-50%	12.9	1.98	74.76
StrucSpars-80%	**10.6**	**1.70**	**75.73**
DenseNet-201			
Baseline	20.0	4.39	77.88
Taylor-40%	**12.5**	**3.02**	76.51
StrucSpars-38%	13.1	3.53	**77.43**
Taylor-64%	**9.0**	2.21	75.28
StrucSpars-60%	9.2	**2.10**	**75.86**

Learned Channel Shuffle Mechanism. We evaluate the effectiveness of our learned channel shuffle mechanism on the ResNet backbone with a compression rate of 65%. We use the following five settings for performance evaluation:

1. FINETUNE: The preserved parameters after compression are restored and the compressed model is finetuned. For the other four settings, the parameters of the compressed

Preliminary version

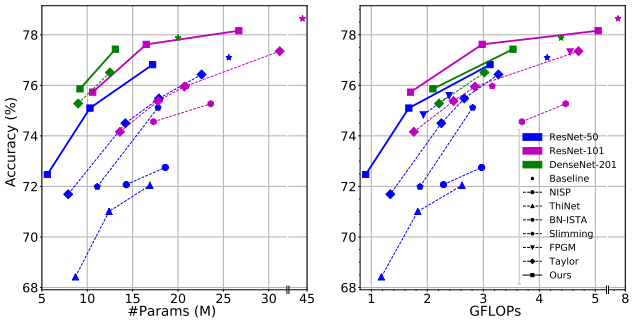


Figure 3: Accuracy-complexity Trade-off on the ImageNet [Russakovsky et al., 2015] dataset. (Upper left is better.)

model are re-initialized for the finetune stage.

2. FROMSCRATCH: We keep the learned channel connectivity, *i.e.,* $\boldsymbol{P}$ and $\boldsymbol{Q}$, from the training stage, and train the model from randomly re-initialized weights.
3. SHUFFLENET: The same channel shuffle operation in the ShuffleNet [Zhang et al., 2018] is adopted. Specifically, if a convolution is of cardinality G and has $G \times N$ output channels, then the channel shuffle operation is equivalent to reshaping the output channel dimension into (G, N), transposing and flattening it back. Compared with SHUFFLENET, the way of channel shuffle is learned rather than pre-defined in our method, *i.e.,* FINETUNE and FROMSCRATCH.
4. RANDOM: The permutation matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ are randomly given, independent of the training process.
5. NOSHUFFLE: The channel shuffle operations are removed, *i.e.,* $\boldsymbol{P}$ and $\boldsymbol{Q}$ are identity matrices.

The results are demonstrated in Tab. 3. First, the finetuned models perform slightly better than those trained from scratch, which implies that the preserved parameters take an essential role in the final performance. Furthermore, the model with learned channel shuffle mechanism, *i.e.,* neuron connectivity, performs the best among all settings. The channel shuffle mechanism in the ShuffleNet [Zhang et al., 2018] is effective as it outperforms the no-shuffle counterpart. However, it can be further improved by a learning-based strategy. Interestingly, the random channel shuffle scheme performs the worst, even worse than the no-shuffle scheme. This implies that learning the channel shuffle operation is a challenging task, and randomly gathering channels from different groups gives no benefits.

Table 3: Ablation study of different channel shuffle operations on the ImageNet dataset [Russakovsky et al., 2015].

Config.	**ResNet-50**-65%		**ResNet-101**-65%	
Acc.	Top-1	Top-5	Top-1	Top-5
FINETUNE	75.10	92.52	77.62	93.72
FROMSCRATCH	75.02	92.46	77.14	93.53
SHUFFLENET	74.97	92.41	76.91	93.38
RANDOM	69.45	89.45	73.16	91.44
NOSHUFFLE	73.30	91.39	75.31	92.64

Table 4: Wall-time acceleration of the structured sparsification method.

Model	GFLOPs	Avg. Runtime (ms)	FPS
ResNet-50	4.14	80.2	12.4
StrucSpars-35%	3.12	68.2	14.7
StrucSpars-65%	1.67	61.3	16.3
StrucSpars-85%	0.90	53.5	18.7

Wall-time Acceleration We measure the wall-time of the ResNet-50 backbone and the compressed variants on a single core of the Intel E5-2603 v4@1.70GHz CPU. As in Tab. 4, we report the average runtime and the frames per second (FPS) of different models when processing the 224×224 images. It can be seen that our method can result in actual wall-time acceleration in the real-world scenarios.

5 FUTURE WORK

We discuss three potential directions for future work along the line of our work.

(i) Data-Driven Structured Sparsification. In this work, the structured regularization is uniformly imposed on the convolutional weights, thus the learned cardinality distribution is prone to uniformity. Besides, the structured regularization is calculated independently of the data loss (see Eq. (7)), so the gradients of the structured regularization are not guided by the task information, thus leading to task-agnostic cardinalities of the compressed models. Nevertheless, it is possible to obtain model structures with better performance if the structured regularization could be guided by the back-propagated signals of the data loss. The optimization-based meta-learning techniques [Finn et al., 2017] can be exploited for this purpose.

(ii) Progressive Sparsification Solution. Typically, finetune-free compression techniques are desired in practical applications [Cheng et al., 2018]. Therefore, the sparsified weights can be removed progressively during training, and the architecture search as well as model training can be completed in a single training pass.

(iii) Combination with Filter Pruning Techniques. As the entire feature maps are reserved in our approach, the reduction of memory footprint is limited. This issue can be addressed by combining with the filter pruning

Preliminary version

techniques, which is non-trivial as uniform filter pruning is required within each group. It is of great interest to exploit group sparsity constraints [Yoon and Hwang, 2017] to achieve such uniform sparsity.

6 CONCLUSION

In this work, we propose a structured sparsification method for efficient network compression, where the structurally sparse representations of the convolutional weights are induced and the inter-group information flow is facilitated by the learnable channel shuffle. The compressed model can be readily incorporated in modern deep learning frameworks thanks to their support for the group convolution. The proposed approach is flexible with special network structures and highly compressible with negligible performance degradation, as validated on the CIFAR-10 and ImageNet datasets.

References

Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing Neural Network Architectures using Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*, 2017. 3

Garrett Birkhoff. Three Observations on Linear Algebra. *Univ. Nac. Tacuman, Rev. Ser. A*, 5:147–151, 1946. 4, 5

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement Interpolation Using Lagrangian Mass Transport. In *ACM Transactions on Graphics (SIGGRAPH Asia)*, pages 1–12, 2011. 2, 5

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning Efficient Object Detection Models with Knowledge Distillation. In *Neural Information Processing Systems (NeurIPS)*, pages 742–751, 2017. 1

Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. Compressing Neural Networks with the Hashing Trick. In *International Conference on Machine Learning (ICML)*, pages 2285–2294, 2015. 1

Jian Cheng, Pei-song Wang, Gang Li, Qing-hao Hu, and Han-qing Lu. Recent Advances in Efficient Computation of Deep Convolutional Neural Networks. *Frontiers of Information Technology & Electronic Engineering*, 19(1): 64–77, 2018. 8

François Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017. 2

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*, 2016. 1

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017. 8

Shangqian Gao, Feihu Huang, Jian Pei, and Heng Huang. Discrete Model Compression with Resource Constraint for Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1899–1908, 2020. 7

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 4, 6

Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2019. 7

Yihui He, Xiangyu Zhang, and Jian Sun. Channel Pruning for Accelerating Very Deep Neural Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1398–1406, 2017. 1

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *Neural Information Processing Systems Workshops*, 2015. 1

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. 2

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 1, 6

Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger. CondenseNet: An Efficient DenseNet using Learned Group Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2752–2761, 2018. 3

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, page 448–456, 2015. 2

Preliminary version

Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009. 6

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning Filters for Efficient ConvNets. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 3

Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. Group Sparsity: The Hinge Between Filter Pruning and Decomposition for Network Compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8018–8027, 2020. 7

Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. HRank: Filter Pruning using High-Rank Feature Map. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1538, 2020a. 7

Mingbao Lin, Rongrong Ji, Yuxin Zhang, Baochang Zhang, Yongjian Wu, and Yonghong Tian. Channel Pruning via Automatic Structure Search. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 673–679, 2020b. 7

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable Architecture Search. In *International Conference on Learning Representations (ICLR)*, 2019a. 3

Yu Liu, Xuhui Jia, Mingxing Tan, Raviteja Vemulapalli, Yukun Zhu, Bradley Green, and Xiaogang Wang. Search to Distill: Pearls are Everywhere but not the Eyes. *arXiv preprint arXiv:1911.09074*, 2019b. 1

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning Efficient Convolutional Networks through Network Slimming. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2736–2744, 2017. 1, 2, 3, 5, 6, 7

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 6

Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A Filter Level Pruning Method for Deep Neural Network Compression. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5058–5066, 2017. 7

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ShuffleNetv2: Practical Guidelines for Efficient CNN Architecture Design. In *European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 2

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. Improved Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020. 1

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance Estimation for Neural Network Pruning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11264–11272, 2019. 1, 2, 7

Xuefei Ning, Tianchen Zhao, Wenshuo Li, Peng Lei, Yu Wang, and Huazhong Yang. DSA: More Efficient Budgeted Pruning via Differentiable Sparsity Allocation. In *European Conference on Computer Vision (ECCV)*, 2020. 7

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-performance Deep Learning Library. In *Neural Information Processing Systems (NeurIPS)*, pages 8026–8037, 2019. 6

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. In *European Conference on Computer Vision (ECCV)*, pages 525–542, 2016. 1

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized Evolution for Image Classifier Architecture Search. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 4780–4789, 2019. 3

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal on Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2, 6, 7, 8

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 2

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 4, 6

Preliminary version

Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chunjing Xu, Chao Xu, and Chang Xu. SCOP: Scientific Control for Reliable Neural Network Pruning. *Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 7

Xijun Wang, Meina Kan, Shiguang Shan, and Xilin Chen. Fully Learnable Group Convolution for Acceleration of Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9049–9058, 2019. 3

Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10734–10742, 2019. 3

Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized Convolutional Neural Networks for Mobile Devices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4820–4828, 2016. 1

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017. 2, 6

Jianbo Ye, Xin Lu, Zhe Lin, and James Z Wang. Rethinking the Smaller-Norm-Less-Informative Assumption in Channel Pruning of Convolution Layers. In *International Conference on Learning Representations (ICLR)*, 2018. 7

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4133–4141, 2017. 1

Jaehong Yoon and Sung Ju Hwang. Combined Group and Exclusive Sparsity for Deep Neural Networks. In *International Conference on Machine Learning (ICML)*, pages 3958–3966, 2017. 9

Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S Davis. NISP: Pruning Networks using Neuron Importance Score Propagation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9194–9203, 2018. 7

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, 2018. 2, 3, 8

Zhaoyang Zhang, Jingyu Li, Wenqi Shao, Zhanglin Peng, Ruimao Zhang, Xiaogang Wang, and Ping Luo. Differentiable Learning-to-Group Channels via Groupable Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3542–3551, 2019. 3, 5

Kai Zhao, Xin-Yu Zhang, Qi Han, and Ming-Ming Cheng. Dependency Aware Filter Pruning. *arXiv preprint arXiv:2005.02634*, 2020. 5

Barret Zoph and Quoc V Le. Neural Architecture Search with Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*, 2017. 3

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning Transferable Architectures for Scalable Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8697–8710, 2018. 3

Preliminary version