

Toward Compact Deep Neural Networks via Energy-Aware Pruning

Seul-Ki Yeom

Nota AI GmbH

Winterfeldstrasse 21, 10781 Berlin, Germany

skyeom@nota.ai (corresponding author)

Kyung-Hwan Shim

Nota AI

4F, Daedong Bldg, 36 Seolleung-ro 92-gil, Gangnam-gu, Seoul

kyunghwan.shim@nota.ai

Jee-Hyun Hwang

Nota AI

4F, Daedong Bldg, 36 Seolleung-ro 92-gil, Gangnam-gu, Seoul

hwanggh96@gmail.com

Abstract

Despite of the remarkable performance, modern deep neural networks are inevitably accompanied with a significant amount of computational cost for learning and deployment, which may be incompatible with their usage on edge devices. Recent efforts to reduce these overheads involves pruning and decomposing the parameters of various layers without performance deterioration. Inspired by several decomposition studies, in this paper, we propose a novel energy-aware pruning method that quantifies the importance of each filter in the network using nuclear-norm (NN). Proposed energy-aware pruning leads to state-of-the-art performance for Top-1 accuracy, FLOPs, and parameter reduction across a wide range of scenarios with multiple network architectures on CIFAR-10 and ImageNet after fine-grained classification tasks. On toy experiment, despite of no fine-tuning, we can visually observe that NN not only has little change in decision boundaries across classes, but also clearly outperforms previous popular criteria. We achieve competitive results with 40.4/49.8% of FLOPs and 45.9/52.9% of parameter reduction with 94.13/94.61% in the Top-1 accuracy with ResNet-56/110 on CIFAR-10, respectively. In addition, our observations are consistent for a variety of different pruning setting in terms of data size as well as data quality which can be emphasized in the stability of the acceleration and compression with negligible accuracy loss. Our code is available at <https://github.com/nota-github/nota-pruning-rank>.

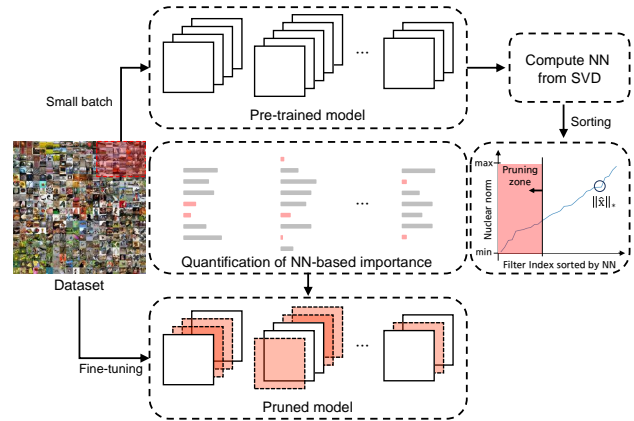


Figure 1. Framework of the proposed method for pruning. After flattening/concatenating each filter maps for all inputs, SVD is applied to retrieve nuclear-norm values. Then the actual pruning process takes place as presented in the bottom-most column, according to the ordering calculated for each layer.

1. Introduction

Deep Neural Networks (DNNs) have achieved great successes in various applications such as image classification [43], detection [44], and semantic segmentation [46]. However, these modern networks require significant computational costs and storage, making it difficult to deploy in real-time applications without the support of a high-efficiency Graphical Processing Unit (GPU). To address this issue, various network compression methods such as prun-

ing [10, 33, 12, 53], quantization [17, 27], low-rank approximation [21, 5], and knowledge distillation [13, 37] are constantly being developed.

Among diverse network compression strategies, network pruning has steadily grown as an indispensable tool, aiming to remove the least important subset of network units (i.e. neurons or filters) in the structured or unstructured manner. For network pruning, it is crucial to decide how to identify the “irrelevant” subset of the parameters meant for deletion. To address this issue, previous researches have proposed specific criteria such as Taylor approximation, gradient, weight, Layer-wise Relevance Propagation (LRP), and others to reduce complexity and computation costs in the network. Recently several studies, inspired by low-rank approximation which can efficiently reduce the rank of the corresponding matrix, have been started from the viewpoint of pruning [28, 26]. Indeed, pruning and decomposition have a close connection like two sides of the same coin from perspective of compression [26]. For more details, related works are introduced in Section 2.

The concept of the decomposition-based compression studies proposes that the network is compressed by decomposing a filter into a set of bases with singular values on a top- k basis, in which singular values represent the importance of each basis [48]. In other word, we can say that decomposition allows to optimally conserve the *energy*, which can be a summation of singular values [1], of the filter in the network. From the macroscopic point of view, we here believe that the energy-aware components could be used as an efficient criterion to quantify the filters in the network.

We propose an energy-aware pruning method that measures the importance scores of the filters by using energy-based criterion inspired by previous filter decomposition methods. More specifically, we compute nuclear-norm (NN) derived from singular values decomposition (SVD) to efficiently and intuitively quantify the filters into an energy cost. Our experimental results show that the NN based pruning can lead the state-of-the-art performance regardless of network architectures and datasets, assuming that the more/less energy contains, the better/worse filter stands for. We prune the filters with the least energy throughout the network. A detailed description of the overall framework of our energy-aware pruning process is shown in Fig. 1.

To summarize, our main contributions are:

- We introduce a novel energy-aware pruning criterion for filter pruning which remove filters with lowest nuclear-norm that can be quantified which lead to efficiently reduce network complexity. Results prove the efficiency and effectiveness of our proposed method through extensive experiment.
- Nuclear-norm based energy-aware pruning achieves state-of-the-art performances with similar compression

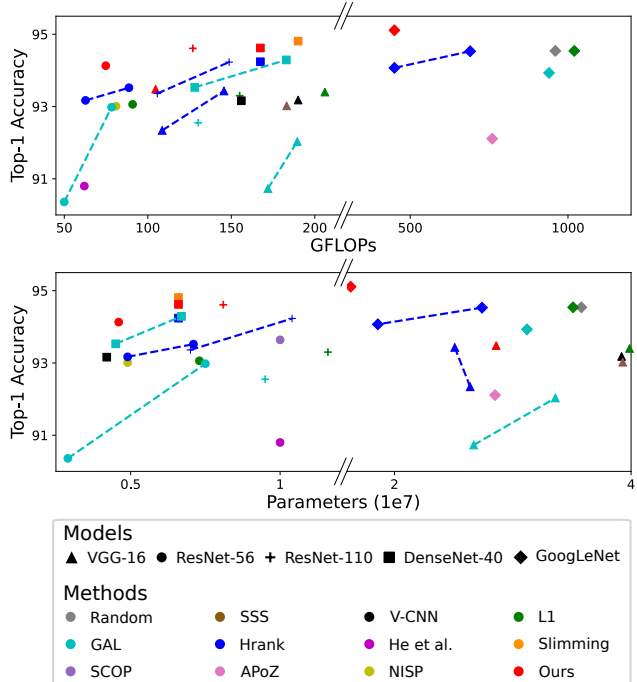


Figure 2. Comparing between accuracy and FLOPs (top) and accuracy and total number of remained parameters (bottom) with five network architectures (VGG-16, ResNet-56, ResNet-110, DenseNet-40, and GoogLeNet) on CIFAR-10 dataset. Top-left is better performance.

ratio over a variety of existing pruning approaches [11, 12, 16, 28, 30, 32, 34, 50, 51] on all kinds of network architectures, as shown in Figure 2.

- Furthermore, the proposed NN-based pruning approach can lead high stability over the quality and quantity of the data, which is great beneficial to practical industry aspect. This property of the proposed method is described in detail in 4.5.

The rest of the paper is organized as follows. Section 2 summarizes related works for network compression. Section 3 describes the details of the proposed pruning method. The experimental results are illustrated and discussed in Section 4. And Section 5 gives a conclusion and an outlook to future work.

2. Related Works

Filter Decomposition. Filter decomposition approaches decompose network matrices into several bases for vector spaces to estimate the informative parameters of the DNNs with low-rank approximation/factorization, thus reducing computation cost of the network [25] such as SVD [5], CP decomposition [21], Tucker decomposition [19], and others, [18] suggests methods to approximate convolutional operations by representing the weight matrix as smaller bases set

of 2D separable filters without changing the original number of filters. In [40], Principal Component Analysis (PCA) was applied on max-pooled and flattened feature maps, to compute the amount of information to be preserved in each layer among all layers, enabling integration with each other.

Filter Pruning. Network filter pruning removes redundant or non-informative filters which are less-informative for performance from the given model at once (one-shot pruning) or iteratively (iterative pruning). The most network filter pruning techniques make filters sparse by removing connections and adopt an appropriate criterion for discriminating whether it is crucial or not. Obviously it is a critical point to decide how to quantify the importance of the filters in the current state of the model for deletion. In previous studies, pruning criteria have been typically proposed based on the magnitude of 1) mostly weights with l_1 / l_2 -norm [7, 23], 2) gradients [41], 3) Taylor expansion / 2^{nd} partial derivative (a.k.a. Hessian matrix) [22, 36], 4) Layer-wise relevance propagation (LRP) [49], and 4) other criteria [50, 32]. For more detail in magnitude-based pruning, please refer to [49].

Pruning by decomposition. Concurrently with our work, there is a growing interest in compressing DNNs motivated by decomposition in terms of pruning as well as fusion approach [24, 26, 47, 28]. Due to the close connection between two different compression methods, those works demonstrate that decomposition-based approach can enhance the performance for pruning in efficiently compressing the model even in the filter level. [24] proposes a hardware-friendly CNN model compression framework, PENNI, which applies filter decomposition to perform a small number of basis kernel sharing and adaptive bases and coefficients with sparse constraints. [26] proposes a unified framework that allows to combine the pruning and the decomposition approaches simultaneously using group sparsity. [47] proposed Trained Ranking Pruning (TRP) which integrates low-rank approximation and regularization into the training process. In order to constrain the model into low-rank space, they adopt a stochastic sub-gradient descent optimized nuclear-norm regularization which is utilized as a different purpose from our proposed method. Similarly to our work, [28] proposes a high rank-based pruning method as a criterion by computing the full-rank of each feature map from SVD layer-by-layer, which leads to inconsistent rank order regardless of batch size.

3. Method

3.1. Preliminaries

From a pre-trained CNN model, we first define trainable parameters, weights as $\mathbf{W}_l = \{\mathbf{w}_l^1, \mathbf{w}_l^2, \dots, \mathbf{w}_l^{c_l}\} \in \mathbb{R}^{c_l \times c_{l-1} \times k \times k}$, where c_{l-1} and c_l denote the number of the input and output channels and k is the the height/width of

the squared kernel at l th convolutional layer. Please note that for the sake of the simplicity, we omit biases term here.

Pruning has been started with a pretrained full-size network $f(\mathcal{X}; \mathbf{W})$ which is overparameterized throughout the network. For DNN, our original objective function is to minimize our loss given dataset and parameters \mathbf{W} .

$$\min_{\mathbf{W}} \mathcal{L}(\mathcal{Y}, f(\mathcal{X}; \mathbf{W})) \quad (1)$$

where $\mathcal{X} \in \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_N\}$ and $\mathcal{Y} \in \{\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_N\}$ represent a set of paired training inputs and its labels, respectively. N denotes the total number of batches.

In order to get structured pruning, sparsity regularization is added in Equation 1 as follows,

$$\min_{\mathbf{W}} \mathcal{L}(\mathcal{Y}, f(\mathcal{X}; \mathbf{W})) + \lambda \mathcal{R}(\mathbf{W}) \quad (2)$$

where \mathcal{R} denotes sparsity regularization function and λ indicates a regularization factor. Here, the main issue of the pruning is how to define \mathcal{R} function under the given constraints.

Algorithm 1 Energy-Aware Pruning

- 1: **Input:** pre-trained model f , training data \mathcal{X} , pruning ratio r , and pruning threshold t
 - 2: **while** t not reached **do**
 - 3: // Assess network substructure importance;
 - 4: **for all** BN layer **in** f **do**
 - 5: **for all** channels **in** BN layer **do**
 - 6: ▷ compute equation 4 and 5
 - 7: **end for**
 - 8: **end for**
 - 9: // Identify and remove least important filters in groups of r ;
 - 10: ▷ remove r channels with the lowest $\|\hat{\mathbf{x}}\|_*$ from f
 - 11: ▷ remove its corresponding connections of each removed channel
 - 12: **if** desired **then**
 - 13: // Optional fine-tuning to recover performance;
 - 14: ▷ fine-tune f' on \mathcal{X}
 - 15: **end if**
 - 16: **end while**
 - 17: **return** pruned model f'
-

3.2. Energy-based Filter Pruning Approach

We define a \mathcal{R} function by adopting an energy-aware pruning criterion. Our hypothesis is that the more energy a filter has, the larger amount of information it contains. In other words, we could define an regularization function that can minimize the difference between the energies from the pre-trained model and the pruned model. Therefore, in

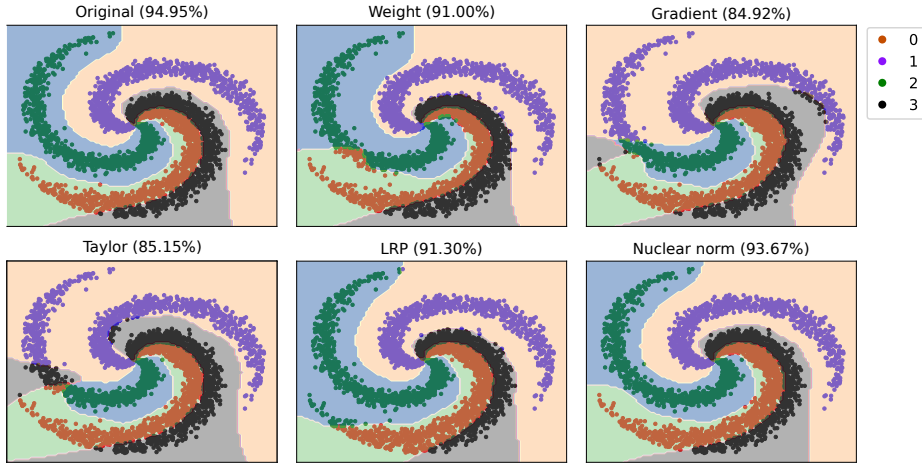


Figure 3. Qualitative Comparison of the impact of the pruning criteria – Original model, Weight, Gradient, Taylor, LRP, and Nuclear-norm (from left top to right bottom) –on the decision function with toy dataset ($k = 4$). Scores in bracket indicate accuracy after pruning 33.3% filters of the original model followed by no fine-tuning.

terms of energy efficiency, \mathcal{R} in Equation 2 can be defined as

$$\mathcal{R}(\mathbf{W}) = |\mathbf{E}(\mathcal{X}; \mathbf{W}) - \mathbf{E}(\mathcal{X}; \mathbf{W}')| \quad (3)$$

where $\mathbf{E}(\cdot) = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_l\}$ indicate total amount of energy in the network. And each \mathbf{e}_l denotes the amount of energy at layer l and is computed on the corresponding feature map using our criterion which will be discussed thoroughly afterwards. Additionally, we introduce a pruning mask $\mathbf{M} \in \{0, 1\}^{c_l}$ which determines if a filter is remained or pruned during feed-forward propagation such that when \mathbf{M} is vectorized: $\mathbf{W}' = \mathbf{W} \odot \mathbf{M}$, where \odot is an element-wise multiplication between \mathbf{W} and \mathbf{M} . And here, we assume that each \mathbf{e}_l can be approximated by $\mathbf{e}_l \approx \|\mathbf{w}_l\|_*$ computed by decomposition approach. Here, we adopt the decomposition approach, SVD, to quantify filter-wise energy consumption. SVD is the basis for many related techniques in dimensionality reduction used to obtain reduced order models (ROMs). For pruning, SVD helps finding the best k -dimensional perpendicular subspace with respect to the dataset in each point. Especially, the singular values plays an important role in algebraic complexity theory. That is, the singular value represents the energy of each rank-one matrix. Singular values represent the importance of its associated rank-one matrix.

A previous research showed that filter pruning and decomposition are highly related from the viewpoint of compact tensor approximation [26]. There is the hinge point between both strategies in investigating a compact approximation of the tensors despite of the usage of different operation in a variety of the application scenarios. Decomposition is done to quantify the energy on the output channels in batch normalization (BN) layers. Additional to the efficient trade-

off of channel-level sparsity, BN provides normalized values of the internal activation using mini-batch statistics to any scale [32]. This process is achieved by applying 3D filters $\hat{\mathbf{x}}_l \in \mathbb{R}^{c_l \times h \times w}$, where h and w denote the height and width at l^{th} BN layer, respectively. The superscript l in \mathbf{w}_l is omitted for readability. Based on $\hat{\mathbf{x}}$, we first reshape the original 3D tensor into a 2D tensor $\hat{\mathbf{x}} \in \mathbb{R}^{c_l \times hw}$

From the SVD, a channel output at l^{th} layer can be decomposed as follow,

$$\hat{\mathbf{x}} = \mathbf{U} \mathbf{S} \mathbf{V}^T = \sum_{i=1}^N \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad (4)$$

where \mathbf{U} and \mathbf{V} denote the left and right singular vector matrix respectively and \mathbf{S} indicates the diagonal matrix of singular values σ_n where $\mathbf{S} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_N)$.

$$\|\hat{\mathbf{x}}\|_* = \sum_{i=1}^N \sigma_i \quad (5)$$

$\|\hat{\mathbf{x}}\|_*$ denotes *nuclear-norm*, the sum of the singular values which can represent the energy of the model [38]. Here, based on our hypothesis, a useful rule of thumb for the efficient filter pruning is to optimally preserve the energy throughout the network. In this respect, based on equation 5, we can not only evaluate the distribution, but also estimate the contribution of the feature spaces simultaneously, which can be applicable for a pruning criterion. Additionally, it provides necessary and sufficient conditions for rank consistency while minimizing the loss of the model [2]. For this reason, it leads to achieve the consistent results regardless data quality as well as data quantity.

The procedure based on the pruning method is outlined in Algorithm 1,

4. Experiments

4.1. Experimental Setup

Models and Dataset We demonstrate the effectiveness of the proposed energy-aware pruning with nuclear-norm on four types of pre-trained feed-forward deep neural network architectures from various perspective comparison studies: 1) simple CNNs (VGG-16 [39] on CIFAR-10 [20]), 2) Residual networks (ResNet-56 and ResNet-110 [8] on CIFAR-10 and ResNet-50 on ImageNet [4]), 3) Inception networks (GoogLeNet [42] on CIFAR-10), 4) Dense networks (DenseNet-40 [15] on CIFAR-10). The resolution of each image is 32×32 (CIFAR-10) and 224×224 (ImageNet) pixels, respectively.

Implementation details We conduct all pruning experiments on *Pytorch 1.6* under *Intel(R) Xeon(R) Silver 4210R CPU 2.40GHz* and *NVIDIA RTX 2080Ti with 12GB* for GPU processing. After one-shot pruning, we adopt the Stochastic Gradient Descent (SGD) algorithm as an optimization function. For both the CIFAR-10 and ImageNet, over-parameterized models are pruned at a time and fine-tuned by using 200 epochs with early stopping with 0.01 initial learning rate, scheduled by using cosine scheduler. Cross entropy is selected as a loss function. And the momentum and the weight decay factor are 0.9 and 5×10^{-4} , respectively. And we set the fine-tuning batch size of 128. For pruning, we adopt the built-in function *torch.nn.utils.prune* in *Pytorch* throughout the experiments.

Evaluation metrics For a fair competition, we measure Top-1 accuracy (CIFAR-10 and ImageNet) and Top-5 accuracy (ImageNet only) of the pruned network as baselines. Also, we computed the Floating point operations (FLOPs) as well as total remained number of parameters (params) to precisely compare the efficiency of the proposed criterion in terms of computational efficiency.

4.2. Results on Toy experiment

First, we start by comparing the properties and effectiveness of the several pruning criteria on toy dataset. In addition to our proposed criterion (i.e. nuclear-norm), we also evaluate against pruning methods that use various property important based pruning criteria on the toy dataset: weight [23], gradient [41], Taylor [36], and layer-wise relevance propagation (LRP) [49]. We generated 4-class toy datasets from *Scikit-Learn*¹ toolbox.

Each generated consists of 1000 training samples per class in 2D domain. We firstly construct a simple model and train the model. The model we constructed is stacked with a sequence of three consecutive ReLU-activated dense layers with 1000 hidden neurons each. We have also added a Dropout function with the probability of 50%. For the toy experiment, all structures are as follows,

Table 1. Pruning results of five network architectures on CIFAR-10. Scores in brackets of “FLOPs” and “Params” denote the compression ratio of FLOPs and parameters in the compressed models.

Criterion	Pruned	Gap	FLOPs (%)	Params (%)
VGG-16-BN				
L1 [23]	93.40	0.15	206.00M (34.3)	5.40M (64.0)
Variational CNN [51]	93.18	-0.07	190.00M (39.4)	3.92M (73.3)
SSS [16]	93.02	-0.94	183.13M (41.6)	3.93M (73.8)
GAL-0.05 [30]	92.03	-1.93	189.49M (39.6)	3.36M (77.6)
GAL-0.1 [30]	90.73	-3.23	171.89M (45.2)	2.67M (82.2)
HRank-53 [28]	93.43	-0.53	145.61M (53.5)	2.51M (82.9)
HRank-65 [28]	92.34	-1.62	108.61M (65.3)	2.64M (82.1)
Propose method	93.48	-0.48	104.67M (66.6)	2.86M (80.9)
ResNet-56				
L1 [23]	93.06	0.02	90.90M (27.6)	0.73M (14.1)
NISP [50]	93.01	-0.25	81.00M (35.5)	0.49M (42.4)
GAL-0.6 [30]	92.98	-0.28	78.30M (37.6)	0.75M (11.8)
GAL-0.8 [30]	90.36	-2.90	49.99M (60.2)	0.29M (65.9)
He <i>et al.</i> [12]	90.80	-2.00	62.00M (50.6)	N/A
HRank-29 [28]	93.52	0.26	88.72M (29.3)	0.71M (16.8)
HRank-50 [28]	93.17	-0.09	62.72M (50.0)	0.49M (42.4)
SCOP [45]	93.64	-0.06	N/A (56.3)	N/A (56.0)
Propose method	94.13	0.87	74.83M (40.4)	0.46M (45.9)
ResNet-110				
L1 [23]	93.30	-0.20	155.00M (38.7)	1.16M (32.6)
GAL-0.5 [30]	92.55	-0.95	130.20M (48.5)	0.95M (44.8)
HRank-41 [28]	94.23	0.73	148.70M (41.2)	1.04M (39.4)
HRank-58 [28]	93.36	-0.14	105.70M (58.2)	0.70M (59.2)
Propose method	94.61	1.11	126.96M (49.8)	0.81M (52.9)
GoogLeNet				
Random	94.54	-0.51	0.96B (36.8)	3.58M (41.8)
L1 [23]	94.54	-0.51	1.02B (32.9)	3.51M (42.9)
APoZ [14]	92.11	-2.94	0.76B (50.0)	2.85M (53.7)
GAL-0.5 [30]	93.93	-1.12	0.94B (38.2)	3.12M (49.3)
HRank-54 [28]	94.53	-0.52	0.69B (54.9)	2.74M (55.4)
HRank-70 [28]	94.07	-0.98	0.45B (70.4)	1.86M (69.8)
Propose method	95.11	0.06	0.45B (70.4)	1.63M (73.5)
DenseNet-40				
Network Slimming [32]	94.81	-0.92	190.00M (32.8)	0.66M (36.5)
GAL-0.01 [30]	94.29	-0.52	182.92M (35.3)	0.67M (35.6)
GAL-0.05 [30]	93.53	-1.28	128.11M (54.7)	0.45M (56.7)
Variational CNN [51]	93.16	-0.95	156.00M (44.8)	0.42M (59.7)
HRank-40 [28]	94.24	-0.57	167.41M (40.8)	0.66M (36.5)
Propose method	94.62	-0.19	167.41M (40.8)	0.66M (36.5)

- Dense (1000) → ReLU → Dropout (0.5) → Dense (1000) → ReLU → Dense (1000) → ReLU → Dense (k)

The model which takes 2D inputs will take an output which is the same number of classes (i.e. = 4). We then sample a number of new datapoints (unseen during training) for the computation of the pruning criteria. For pruning, we remove a fixed number of 1000 of 3000 hidden neurons with the least relevance for prediction according to each criterion. This is equivalent to removing 1000 learned filters from the model. After pruning, we observed the changes in the decision boundary area and re-evaluated classification accuracy on the original 4000 training samples with pruned model. Please note that after pruning, we directly show the decision boundary and accuracy as it is without fine-tuning step.

Figure 3 shows the data distributions of the generated multi-class toy datasets to see the qualitative impact to the

¹https://scikit-learn.org/stable/datasets/toy_dataset.html

Table 2. Pruning results on ResNet-50 with ImageNet. Scores in brackets of “FLOPs” and “Params” denote the compression ratio of FLOPs and parameters in the compressed models.

Criterion	ResNet-50					
	Top-1 Acc (%)		Top-5 Acc (%)		FLOPs ↓ (%)	Params ↓ (%)
	Pruned	Gap	Pruned	Gap		
He <i>et al.</i> [12]	72.30	-3.85	90.80	-1.40	2.73B (33.25)	N/A
ThiNet-50 [34]	72.04	-0.84	90.67	-0.47	N/A (36.8)	N/A (33.72)
SSS-26 [16]	71.82	-4.33	90.79	-2.08	2.33B (43.0)	15.60M (38.8)
SSS-32 [16]	74.18	-1.97	91.91	-0.96	2.82B (31.0)	18.60M (27.0)
GAL-0.5 [30]	71.95	-4.20	90.94	-1.93	2.33B (43.0)	21.20M (16.8)
GAL-0.5-joint [30]	71.80	-4.35	90.82	-2.05	1.84B (55.0)	19.31M (24.2)
GAL-1 [30]	69.88	-6.27	89.75	-3.12	1.58B (61.3)	14.67M (42.4)
GAL-1-joint [30]	69.31	-6.84	89.12	-3.75	1.11B (72.8)	10.21M (59.9)
GDP-0.5 [29]	69.58	-6.57	90.14	-2.73	1.57B (61.6)	N/A
SFP [9]	74.61	-1.54	92.06	-0.81	2.38B (41.8)	N/A
AutoPruner [33]	74.76	-1.39	92.15	-0.72	2.09B (48.7)	N/A
FPGM [10]	75.59	-0.56	92.27	-0.60	2.55B (37.5)	14.74 (42.2)
Taylor [35]	74.50	-1.68	N/A	N/A	N/A (44.5)	N/A (44.9)
RRBP [52]	73.00	-3.10	91.00	-1.90	N/A	N/A (54.5)
GDP-0.6 [29]	71.19	-4.96	90.71	-2.16	1.88B (54.0)	N/A
HRank-74 [28]	74.98	-1.17	92.33	-0.54	2.30B (43.7)	16.15M (36.6)
HRank-71 [28]	71.98	-4.17	91.01	-1.86	1.55B (62.1)	13.77M (46.0)
HRank-69 [28]	69.10	-7.05	89.58	-3.29	0.98B (76.0)	8.27M (67.5)
SCOP [45]	75.26	-0.89	92.53	-0.34	1.85B (54.6)	12.29M (51.8)
Propose method	75.25	-0.89	92.49	-0.37	1.52B (62.8)	11.05M (56.7)
	72.28	-3.87	90.93	-1.93	0.95B (76.7)	8.02M (68.6)

models’ decision boundary when removing a fixed set of 1000 neurons as selected among the considered pruning criteria. This demonstrates how the toy models’ decision boundaries change under influence of pruning with all five criteria. We can observe that both the Taylor and gradient measures degrade the model significantly whereas weight and LRP preserve the decision boundary from the pruned models reasonably except for the area where classify between 1) class No. 0 (brown) and class No. 2 (green) and 2) between class No. 0 and class No. 3 (black). On the other hand, we can clearly see that in contrast to the other property importance based pruning criteria, nuclear-norm significantly classify multi-classes even after pruning process, thus allows to safely remove the unimportant (w.r.t. classification) elements. As we can see in Figure 3, NN-based pruning results in only minimal change in the decision boundary, compared to the other criteria. Furthermore, nuclear-norm can successfully preserve original accuracy of 94.95% up to 93.67% whereas 91.00% of weight, 84.92% of gradient, 85.15% of Taylor expansion, and 91.30% of LRP.

4.3. Results on CIFAR-10

To prove the expandability of the proposed nuclear-norm based pruning approaches on the various deep learning-related modules, such as residual connection or inception module, we compress several popular DNNs, including VGG-16, ResNet-56/110, GoogLeNet, and DenseNet-40. Due to the different original performance of each literature,

we then report the performance gap between their original model and the pruned model. All results are presented in Table 1 on the CIFAR-10 dataset.

VGG-16. We first test on the basic DNN architecture, VGG-16, which is commonly used as a standard architecture. It can verify the efficiency of the proposed pruning method on the consecutive convolutional block. For a fair comparison study, we adopt several conventional importance-based methods – L1 [23], HRank [28], SSS [16], Variational CNN *et al.* [51], and GAL [30] – in this experiment. We reached initial Top-1 accuracy of 93.96% with 313.73 million of FLOPs and 14.98 million of parameters. VGG-16 consists of 13 convolutional blocks with 4224 convolutional filters and 3 fully-connected layers. In terms of complexity, VGG-16 with batch normalization contains 313.73 million of FLOPs and 14.98 million of parameters initially.

The proposed nuclear-norm based pruning method outperforms previous conventional pruning approaches, especially on the performance and the FLOPs as well as parameter reduction. Most of the conventional pruning approaches could compress more than 70% of the parameters, while they could not accelerate the VGG-16 model effectively. On the other hand, the proposed method could yield a highly accelerated model but with a tiny performance drop. To be more specific, GAL [30] accelerates the baseline model by 45.2% and 39.6% while it compresses 82.2% and 77.6% of the model with 90.73% and 92.03% of the performance. However, the proposed method yields

the pruned model with 66.6% reduced FLOPs (104.67M) and 80.9% reduced parameters (2.86M) with only 0.48% of accuracy drop from scratch, which outperforms in all of the aspects (performance, acceleration, and compression). Compared to the recent property importance-based method, HRank, which also uses the rank property for pruning, the proposed method achieves the competitive performance acceleration (93.48% vs. 92.34% and 104.67M vs. 108.61M) but with a similar compress ratio.

ResNet-56/110 The residual connection of the ResNet consists of an element-wise add layer, requiring the same input shape. For this reason, pruning on ResNet needs to be carefully managed compared to pruning other conventional sequential model. To equalize those inputs of the element-wise add operation of the ResNet, we prune common indices of the connected convolutional layer. By using the nuclear-norm based pruning method and the above pruning strategy, we could yield a faster and smaller model than the other approaches.

Initial Top-1 accuracies of ResNet-56 / 110 are 93.26 / 93.50% with 125.49 / 252.89 million of FLOPs and 0.85 / 1.72 million of parameters, respectively. Compared to the baseline ResNet-56 model and the compressed model by previous pruning approaches, the pruned model with the proposed method achieves 0.87% higher performance but with similar compression and acceleration rate (40.4% of FLOPs and 45.9% of parameters). Most of the conventional pruning approaches could not exceed the performance of the original model except HRank (93.52% of Top-1 accuracy). However, the compression and acceleration ratio of Hrank is comparatively low (29.3% of FLOPs and 16.8% of parameters). On the other hand, the proposed method could exceed the original performance (94.13%) with similar or more acceleration and compression rate (40.4% of FLOPs and 45.9% of parameters reduced).

Furthermore, the compressed ResNet-110 also outperforms the baseline model by 1.11% with 40.8% of acceleration rate and 52.9% of compression rate. Similar to ResNet-56, the NN based pruning method achieves the highest performance on ResNet-110 with a similar acceleration and compression ratio. On the other hand, the conventional pruning approaches yield around 92.55% - 94.23% of Top-1 accuracies while the pruned model contains around up to 0.70 - 1.16 million of compressed parameters and 105.70 - 155 million of accelerated FLOPs. Similar to the compressed model of the proposed method, HRank also outperforms the baseline accuracy, but with the larger and slower model compared to our method. In conclusion, the compressed model of the proposed method outperforms the baseline of both ResNet-56/110, which has the potential to be compressed or accelerated more without performance deterioration.

GoogLeNet Unlike the residual connection, the input

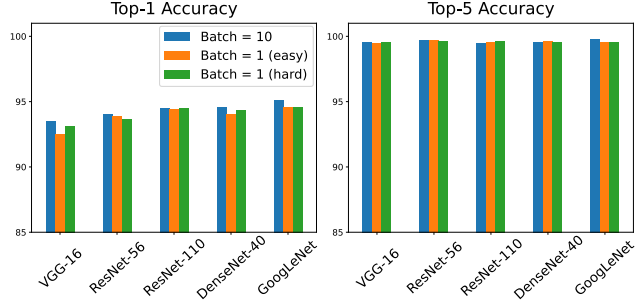


Figure 4. Comparison study of Top-1 and Top-5 accuracies with 1) small (=batch of 10), 2) easy (=batch of 1), 3) hard (=batch of 1) dataset with five different neural network architectures.

kernel size of the concatenation module does not have to be equivalent, therefore, coping with the inception module is relatively straightforward. We initially achieved Top-1 accuracy of 95.05%, 1.52 billion of FLOPs, and 6.15 million of parameters. The proposed nuclear-norm based method greatly reduces the model complexity (70.4% of FLOPs and 73.5% of parameters) while it outperforms the baseline model (95.11% vs. 95.05%).

GoogLeNet with the proposed pruning approach could yield the highest performance (95.11%) with the most limited number of parameters (73.5%). HRank reaches the performance of 94.07%, while it accelerates around 70.4%, but the proposed method returns 1.04% higher performance and prune an additional 0.23M of the parameters. The performance and the complexity of the nuclear-norm based pruning method indicate that the GoogLeNet can be compressed and accelerated more with tolerable performance drop. It demonstrates its stability to compress and accelerate the inception module without performance degradation.

DenseNet-40 The original model contains 40 layers with a growth rate of 12, it achieves 94.81% on the CIFAR-10 dataset with 282.00M of FLOPs and 1.04M of parameters. The channel-wise concatenation module of the DenseNet-40 is also treated similarly to the inception module of GoogLeNet. We followed the global pruning ratio of HRank. As a result, the proposed method could outperform by 0.38% with the same amounts of FLOPs and parameters. The compressed model could not exceed the performance of Network slimming, however, the FLOP compression rates of the proposed model could be accelerated by 22.59M.

4.4. Results on ImageNet

We also test the performance with our proposed criterion on ImageNet with a popular DNN, ResNet-50. Comparison of pruning ResNet-50 on ImageNet by the proposed method and other existing methods presented can be seen in the Table 2 where we report Top-1 and Top-5 accuracies, as well as FLOPs and parameters reduction. Initial performance of ResNet-50 on ImageNet is 76.15% and 92.87%

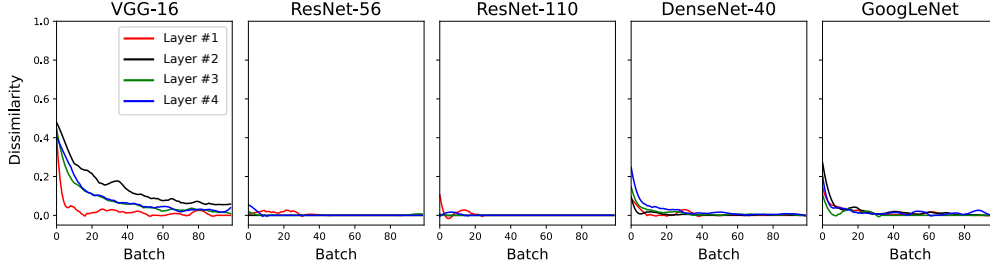


Figure 5. Results of Kendall tau distance between filter ranking lists of two neighbour batch sizes. Here, values with y-axis is close to 0 when paired observations between two neighbour batches have a similar rank order and vice versa.

of Top-1 and Top-5 accuracies with 4.09 billion of FLOPs and 25.50 million of parameters. Compare with other existing pruning methods, it is clearly observed that our proposed method achieves better performance in all aspects. By pruning 62.8% of FLOPs and 56.7% of parameters from original ResNet-50 we only lose 0.89% and 0.37% in Top-1 and Top-5 accuracies while compressing 2.69 \times of FLOPs and 2.30 \times of parameters at the same time. When compressing the model aggressively, we could achieve 72.28% and 90.93% of Top-1 and Top-5 accuracies while reducing 76.7% of FLOPs and 68.6% of parameters which still represent a reasonable result.

4.5. Ablation study

We further conduct two additional ablation studies in the perspectives of the data quality and quantity to see whether our proposed method also yields stable performance regardless of two properties for the practical industry issue. These would be the critical points when you encounter 1) lack of data, 2) dataset with overconfidence or uncertainty for the efficient pruning. We test on two more scenarios with modern neural network architectures to see the effect of rank consistency.

Results in data quality First, we see if our proposed method can achieve reasonable performances regardless of data quality. These results demonstrate that the performance of nuclear-norm based pruning is stable and independent of the data quality. Among the first 10 batches, we select a single batch of samples with 1) the lowest loss (called “easy” samples) and 2) the highest loss (called “hard” samples). In the previous pruning or neural architecture search (NAS) literatures, they use a small proxy dataset for searching and pruning the models, which means that it also gives a great impact with respect to pruning efficiency [3].

Figure 4 shows comparison results of the Top-1 and Top-5 accuracy across small-batch (= 10), easy (= 1) and hard (= 1) samples on five different network architectures. We can observe that by using only a batch with easy as well as hard samples, our first ablation study found no significant differences across three different conditions (i.e. small-batch vs. easy vs. hard). This experiment result demonstrates that

competitive performance can be produced by NN based filter pruning regardless without considering data quality for the efficient pruning.

Results in data quantity From the practical point of view, compared to ImageNet, PASCAL VOC [6], and COCO [31], most of the private dataset have a smaller amount of data quantity which might be not guaranteed to be optimal for efficient pruning. In this manner, one of the interesting points in the pruning community is to see how large the amount of dataset we need for the proper pruning in terms of data quantity. Therefore, to evaluate the stability of the proposed criterion by data quantity, we perform a statistical test on 4 convolutional layers at regular intervals, called Kendall tau distance, to measure the pairwise similarity of two filter ranking lists of neighbour batches based on nuclear-norm to see the evolutionary change in increasing batch size. The equation for Kendall tau distance can be expressed as follows:

$$K(\tau^1, \tau^2) = \frac{1}{n \times (n-1)} \sum_{(j,s), j \neq s} K_{js}^*(\tau^1, \tau^2) \quad (6)$$

where $K_{js}^*(\tau^1, \tau^2)$ is assigned to 0 if x_j, x_s are in the same order in τ^1 and τ^2 and 1 otherwise.

We empirically observe that the ranking order generated by the proposed criterion is stable and independent of the data quantity. Figure 5 shows the similarity between neighbour of batches with Kendall tau distance. Here, we can observe that for ResNet-56/110, DenseNet-40, and GoogLeNet, there is a very close similarity of ranking order before batch of ten which means the proposed method extracts stable ranking order indices layer-wisely, whereas VGG-16 observes the higher similarity between neighbour of batches after batch indices of 50 which indicates that it needs more data to get the stable ranking order.

5. Conclusion

Behind the remarkable growth of modern deep neural networks, millions of trainable parameters remain an unsolved problem. After training, extremely high cost for in-

ference time remains one of the main issues in the entire machine learning applications. In this paper, we propose a novel energy-aware criterion which prunes filters to reduce network complexity using nuclear-norm motivated by decomposition/approximation based approaches. Empirically, we demonstrated that the proposed criterion outperforms prior works on a variety of DNN architectures in terms of accuracy, FLOPs as well as number of compressed parameters. Furthermore, it can be applicable for the specific scenarios which limit on data quantity (e.g. pruning after transfer learning and few-shot learning which small amount of dataset are required) and data quality (e.g. consisting of over-confident/uncertainty data)

For the further research, more experiments can be done on 1) an unified framework which pruning is followed by decomposition of pretrained models to simultaneously achieve small drop in accuracy (by pruning) and reduced FLOPs and parameters for the fast inference time (by decomposition) 2) eXplainable Artificial Intelligence (XAI) approach using our proposed method.

References

- [1] Saieed Akbari, Ebrahim Ghorbani, and Mohammad Reza Oboudi. Edge addition, singular values, and energy of graphs and matrices. *Linear Algebra and its Applications*, 430(8-9):2192–2199, 2009.
- [2] Francis R. Bach. Consistency of trace norm minimization. *J. Mach. Learn. Res.*, 9:1019–1048, 2008.
- [3] Xiyang Dai, Dongdong Chen, Mengchen Liu, Yinpeng Chen, and Lu Yuan. DA-NAS: data adapted pruning for efficient neural architecture search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVII*, volume 12372 of *Lecture Notes in Computer Science*, pages 584–600. Springer, 2020.
- [4] 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*, pages 248–255. Ieee, 2009.
- [5] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems*, pages 1269–1277, 2014.
- [6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
- [7] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1135–1143, 2015.
- [8] 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*, pages 770–778, 2016.
- [9] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*, 2018.
- [10] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2019.
- [11] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2019.
- [12] 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*, pages 1389–1397, 2017.
- [13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [14] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures, 2016.
- [15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [16] Zehao Huang and Naiyan Wang. Data-driven sparse structure selection for deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 304–320, 2018.
- [17] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.
- [18] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014.
- [19] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- [20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [21] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Osledets, and Victor Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*, 2014.
- [22] Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems (NIPS)*, pages 598–605, 1989.

- [23] 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.
- [24] Shiyu Li, Edward Hanson, Hai Li, and Yiran Chen. Penni: Pruned kernel sharing for efficient cnn inference. *arXiv preprint arXiv:2005.07133*, 2020.
- [25] Yawei Li, Shuhang Gu, Luc Van Gool, and Radu Timofte. Learning filter basis for convolutional neural network compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5623–5632, 2019.
- [26] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8015–8024. IEEE, 2020.
- [27] Darryl Lin, Sachin Talathi, and Sreekanth Annapureddy. Fixed point quantization of deep convolutional networks. In *International conference on machine learning*, pages 2849–2858, 2016.
- [28] Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. Hrank: Filter pruning using high-rank feature map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2020.
- [29] Shaohui Lin, Rongrong Ji, Yuchao Li, Yongjian Wu, Feiyue Huang, and Baochang Zhang. Accelerating convolutional networks via global & dynamic filter pruning. In *IJCAI*, pages 2425–2432, 2018.
- [30] Shaohui Lin, Rongrong Ji, Chenqian Yan, Baochang Zhang, Liujuan Cao, Qixiang Ye, Feiyue Huang, and David Doermann. Towards optimal structured cnn pruning via generative adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2799, 2019.
- [31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.
- [32] 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 2017, Venice, Italy, October 22-29, 2017*, pages 2755–2763. IEEE Computer Society, 2017.
- [33] Jian-Hao Luo and Jianxin Wu. Autopruner: An end-to-end trainable filter pruning method for efficient deep model inference. *Pattern Recognition*, page 107461, 2020.
- [34] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*, pages 5058–5066, 2017.
- [35] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Froisio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [36] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *International Conference of Learning Representation (ICLR)*, 2016.
- [37] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.
- [38] Rowayda A. Sadek. SVD based image processing applications: State of the art, contributions and research challenges. *CoRR*, abs/1211.7102, 2012.
- [39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [40] Xavier Suau, Luca Zappella, and Nicholas Apostoloff. Filter distillation for network compression, 2019.
- [41] Xu Sun, Xuancheng Ren, Shuming Ma, and Houfeng Wang. meprop: Sparsified back propagation for accelerated deep learning with reduced overfitting. In *International Conference on Machine Learning (ICML)*, pages 3299–3308, 2017.
- [42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [43] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.
- [44] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [45] Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chun-jing XU, Chao Xu, and Chang Xu. Scop: Scientific control for reliable neural network pruning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10936–10947. Curran Associates, Inc., 2020.
- [46] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.
- [47] Yuhui Xu, Yuxi Li, Shuai Zhang, Wei Wen, Botao Wang, Yingyong Qi, Yiran Chen, Weiyao Lin, and Hongkai Xiong. TRP: trained rank pruning for efficient deep neural networks. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 977–983. ijcai.org, 2020.
- [48] Jian Xue, Jinyu Li, and Yifan Gong. Restructuring of deep neural network acoustic models with singular value decomposition. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 2365–2369, 2013.

- [49] Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition*, 2021.
- [50] 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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9194–9203. IEEE Computer Society, 2018.
- [51] Chenglong Zhao, Bingbing Ni, Jian Zhang, Qiwei Zhao, Wenjun Zhang, and Qi Tian. Variational convolutional neural network pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2780–2789, 2019.
- [52] Yuefu Zhou, Ya Zhang, Yanfeng Wang, and Qi Tian. Accelerate cnn via recursive bayesian pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3306–3315, 2019.
- [53] Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. Discrimination-aware channel pruning for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 875–886, 2018.