Blending Pruning Criteria for Efficient Convolutional Neural Networks

Zhongzhan Huang[*]
Tsinghua University
Haidian, Peking
hzz_dedekinds@foxmail.com

Wei He[*]
Nanyang Technological University
Singapore
wei005@ntu.edu.sg

Mingfu Liang
Northwestern University
Evanston, Illinois
mingfuliang2020@u.northwestern.edu

Senwei Liang
Purdue University
West Lafayette, USA
liang339@purdue.edu

Haizhao Yang
Purdue University
West Lafayette, USA
yang1863@purdue.edu

Abstract

Network Pruning is an effective acceleration method for large-scale deep Convolutional Neural Networks (CNNs). Recently, various pruning criteria have been proposed to remove redundant filters of CNNs under different circumstances. However, the rank of pruned filters according to their “importance" may be inconsistent among different pruning criteria, *e.g.*, it might happen that one filter could be important according to a certain criterion, while it is useless according to another one, which indicates that each criterion may be only a partial view on the comprehensive “importance”. From this motivation, we propose an ensemble framework for integrating the existing filter pruning criteria by examining the criteria diversity. The proposed framework contains two stages: criteria clustering and filters importance calibration. In the first stage, we condense the pruning criteria via layerwise clustering based on the rank of “importance” score. In the second stage, in each cluster, we propose a calibration factor to adjust their significance for each selected ensemble candidates and search the optimal ensemble criterion via Evolutionary Algorithm (EA). Quantitative results on CIFAR-100 and ImageNet benchmarks show that our framework outperforms the state-of-the-art baselines, regrading to the compact model performance after pruning.

1 Introduction

Deep Convolutional Neural Networks (CNNs) have been the prevailing methods in computer vision and brought remarkable improvement to various tasks (1; 2; 3; 4). However, as the CNNs are normally over-parameterized and cumbersome, it is challenging for the model deployment on the devices with limited resources, and the acceleration during the inference stage becomes necessary. Network Pruning, one of the critical directions in Network Compression, aims at eliminating the unimportant parameters or operations without compromising on the model performance. Aiming to remove the

[*] Equal contribution

Preprint. Under review.

entire filter in CNNs, Filter Pruning methods (5; 6; 7; 8; 9), which are more practical to deploy and easily to be implemented, excel in this area recently. Generally, the workflow of filter pruning can be mainly divided into three steps: (1) Normal Training: train the original network on a specific dataset from scratch. (2) Pruning: prune the insignificant network components such as neurons, filters, based on a well-handcrafted criterion, where the score magnitude under the criterion reflects their "importance". (3) Fine-tuning: recover the performance loss caused by the removal of the components to a certain extent. Among all steps, an effective pruning criterion plays an essential role in one filter pruning algorithm.

Conventional pruning methods mainly concentrate on designing a better criterion to increase the viable prune ratio without harming the performance, but few of them have introspected the actual correlations among them. Recent work in (10) reveals that some of the filter pruning criteria, *e.g.*, L1-Norm (5), L2-Norm (11), FPGM (9) and Fermat (10), have a substantial similarity on the "importance" index of pruned filters in most layers. That is, some pruning criteria incline to remove alike filters, though their considerations are from different perspective. From this motivation, we further extend the comparison to more criteria. In Fig. 1, we demonstrate the filters rank similarity under assorted state-of-the-art criteria and their variants using the Spearman's correlation (12) coefficient (Sp), which is a non-parametric correlation measurement of rankings, and able to assess the relationship between two variables using a monotonic function. Specifically, for two sequences $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, .., y_n\}$, if the rank of X and Y are $\{x_{1}\prime, x_{2}\prime, \cdots, x_{n}\prime\}$ and $\{y_{1}\prime, y_{2}\prime, \cdots, y_{n}\prime\}$ respectively, the definition of the Sp between them is

$$\rho(X, Y) = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{n(n^2-1)}, \tag{1}$$

where $d_i = x_{i}\prime - y_{i}\prime$. In general, if the Sp is above 0.8, there is a strong confidence that the two sequences of variables are highly similar. Fig. 1 indicates an empirical fact that although each criterion is competent for filter pruning, the Sp correlation on their measured "importance" rank has a discrepancy. For example, in the first layer, Taylor BN_γ (13) has negative Sp value with other criteria, *i.e.*, it tends to remove different filters comparing to other criteria. Additionally, in the intermediate layers, this discrepancy becomes obvious on more criteria, excluding the four criteria mentioned in (10). These inconsistencies imply that some of the pruning criteria may prune the potentially significant filters, *e.g.*, a filter may be viewed as an important filter by one pruning criterion whereas another criterion may judge it unnecessary. Hence, each criterion may only be a partial view of the comprehensive "importance" of filters. Thus, inspired by this phenomenon, we consider a problem: can we introduce one criterion that integrates both the criteria diversity and their advantages as much as possible?

To solve this problem, we propose an ensemble framework to layerwise integrate the existing filter pruning criteria by examining the criteria diversity on their "importance" measurements, filters "importance" rank, and the discrepancy on their similarity. The proposed framework contains two stages: criteria clustering and filters importance calibration. Since the searching space of the candidate combination pools will be extremely large, especially when the network is particularly deep, *e.g.*, ResNet152 has over a hundred layers, finding the solution for the ensemble criterion among them is non-trivial. Therefore, to ease the above problem and to ensure the variance on ensemble candidates, we condense the pruning criteria via layerwise clustering based on the rank of "importance" score during the first stage. Then, in each cluster, we propose a calibration factor to adjust their significance for the ensemble. In the second stage, we introduce an Evolutionary Algorithm (EA) to optimize the combination of the calibration factor and the ensemble candidates sampled from each cluster and ultimately generate the optimal ensemble criterion.

Our contributions are summarized as follows:

- To the best of our knowledge, our work is the first to explore the correlation among some existing filter pruning criteria and propose a simple-yet-effective ensemble framework to integrate different pruning criteria and generate an optimal criterion.
- Compared to the conventional filter pruning methods, our method searches for the optimal pruning criteria automatically without manual interaction and empirical knowledge prior. The comprehensive experiments on benchmarks exhibit that our method can further outperform the current state-of-the-art methods. Especially, for ResNet56 pruning in CIFAR-100, our pruned model can exceed the original model performance.

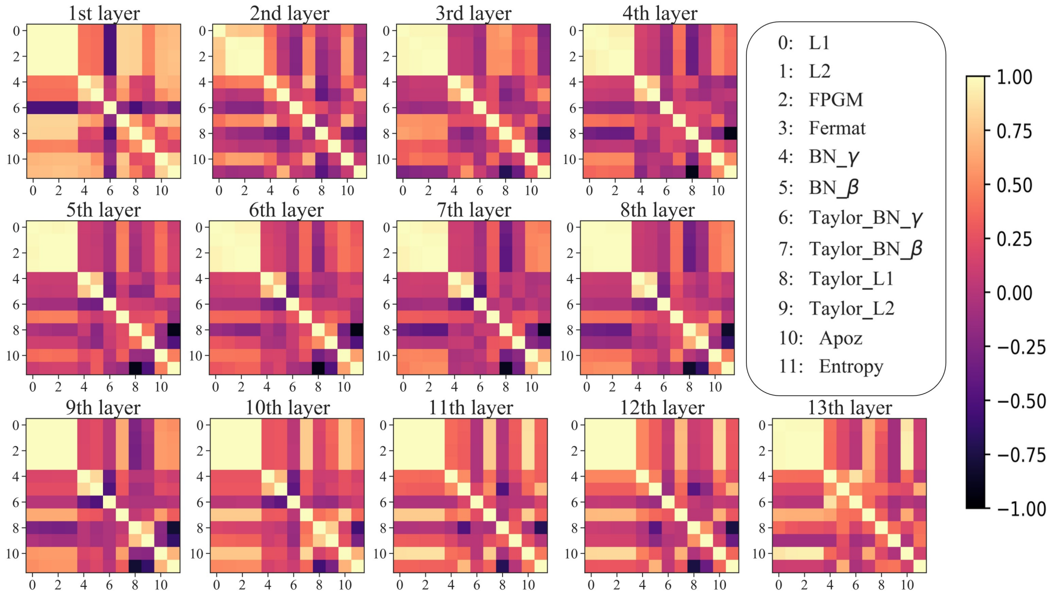

Figure 1: The Spearman's rank correlation coefficient (Sp) of 12 pruning criteria in VGG16 (14). The color of each pair of pruning criteria represents the value of Sp, and the lighter the color (close to 1), the stronger similarity.

- The proposed ensemble pruning framework has large potentials and flexibility as each novel filter pruning criterion can be readily complementary to our framework, which is non-trivial and insightful for future works on network pruning.

2 Related Works

Filter Pruning. Filters pruning focuses on removing the insignificant filters from the original network, so as to perform fast inference without largely sacrificing the accuracy. Therefore, it is crucial to construct an effective criterion to evaluate the "importance" or contribution of the filters. Previously proposed criteria can be mainly categorized into four folds: (1) Filter-based criteria, which consider the property among filters as the importance indicator. The scoring metrics can be the filter's L1-Norm (5) or L2-Norm (11). The underlying assumption of these methods is that the small norm parameters play a less informative role on the final prediction. Recent FPGM (9) and Fermat (10) advance the local filter property for the correlation of multiple filters in a layer. (2) Batch Normalization (BN) based criteria, which estimate channel importance via the value of the parameter inside each BN layer. Network Slimming (7) and SSS (15) consider taking the magnitude of the scaling factor γ to reveal the importance of their corresponding filter. (3) Feature maps activation based criteria, which take the activation value in each feature map as a proxy for its importance, for instance, the average percentage of zeros (APoZ) and the in-between entropy on the channel maps (16; 17). (4) First-order Taylor based criteria, which estimate the filter's contribution with respect to the cost function and the scoring function is designed based on the Taylor expansion (18; 13).

(19) proposed a framework to adaptively select pruning criterion for each layer to better suit the filter distribution across layers. Our framework differs from (19) in two folds: (1) for each layer, we integrate different criteria based on the similarity of pruned filter selection, while (19) used only one pruning criterion for each layer; (2) according to the findings on (10) and the experiments (see Fig. 1), for L1-norm (5), L2-Norm (11) and FPGM (9) used in (19), these three criteria tend to remove the consistent set of filters in most layers. Thus, although one of them is selected, the strength of different criteria are not sufficiently utilized.

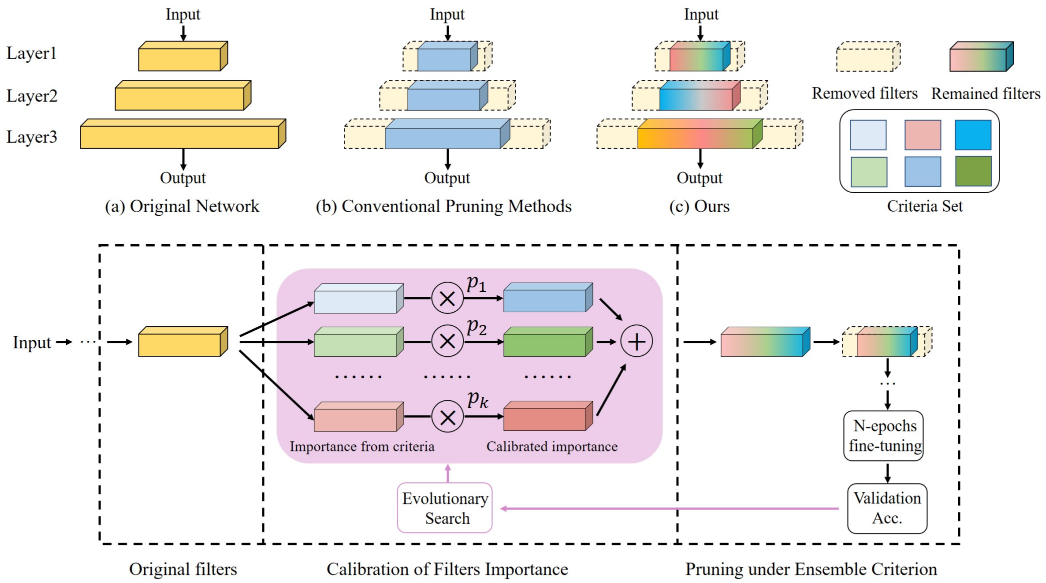

Figure 2: Overview of the ensemble framework for pruning. (a)-(c) show three layers within the original unpruned network, the pruned network under certain conventional filter pruning criterion (denoted in blue) and the pruned network under the ensemble criterion. The three-step ensemble process in one layer is illustrated in the dash-line bounding box, where the notation " $\oplus$ " denotes the element-wise addition and the notation " $\otimes$ " denotes the multiplication between each filter score and the corresponding calibration factor p.

Evolutionary Algorithms (EA) in Network Compression and NAS. Recently, the Evolutionary Algorithms (EA) (20) and their variants are widely used in Network Compression and Neural Architecture Search (NAS) areas as the EA can flexibly solve the multi-objective optimization problem and combinatorial optimization with conflicting objectives. In Network Compression, MetaPruning (21) applied an evolutionary search to find the high accuracy pruned network under the soft or hard constraints. (22) leveraged an Evolution Strategy (ES) algorithm to find a good solution for multi-objective optimization problems. In NAS, (23) modified the EA to search for high-performance neural network architectures for large and realistic classification dataset. (24) introduced the novel aging evolution such that the tournament selection can be biased to choose the younger genotypes. In this paper, our target is totally different from all the previous work utilizing the EA. Our work leverages the EA to search for the ensemble of filter pruning criteria and our empirical results demonstrate the effectiveness.

3 Proposed Method

In this section, we introduce our proposed ensemble framework for filter pruning. Given a CNN, our method adaptively generates an integrated criterion to identify the model redundancy layerwise. The proposed method consists of two stages: in the first stage, we divide different pruning criteria via clustering. In the second stage, we propose the calibration factors to combine criteria sampled from each cluster. Furthermore, the heuristic Evolutionary Algorithm (EA) is applied to optimize the calibration factors and to search the optimal combination of criteria. The details of these two stages are illustrated in Algorithm 1.

For notation, suppose that we are given I criteria, *e.g.*, L1-Norm, L2-Norm and FPGM, and the overall criteria set is denoted as $U_{criteria} = \{f_i\}_{i=1}^{I}$, where f_i is the mapping to filter importance score under $criterion_i$. Consider a L-layer network, the filters set in l-th convolution layer is denoted as $F^l = \{\mathbf{F}_i^l\}_{i=1}^{\lambda_l}$, where λ_l is the number of filters in layer l. The filters importance score $\mathbf{S}_i^l \in [0, 1]^{\lambda_l}$ under $criterion_i$ is calculated by $f_i(\mathbf{F}_1^l, \cdots, \mathbf{F}_{\lambda_l}^l)$, where $i = 1, \cdots, I$. Each component of $\mathbf{S}_i^l$

represents the importance score of the corresponding filter given by the $criterion_i$. For each criterion, the larger the score, the more important the filter is.

Algorithm 1 All in One: Ensemble Framework for Filter Pruning

```
Input: Unpruned model $\Phi$ with $L$ convolution layers; Criteria Set $U_{criteria} = \{f_i\}_{i=1}^{I}$, $I$ is the number of included criteria; Convolution filters in layer $l$: $F^l = \{\mathbf{F}_i^l\}_{i=1}^{\lambda_l}, l = 1, 2, \cdots, L$; Number of clusters each layer: $K$;
Output: Pruned model $\phi_{\text{ensemble}}$ over the optimal integrated criterion $S_{\text{ensemble}}$.
                                        ▷ Calculation of Importance
for $l$ from 1 to $L$ do
    for $i$ from 1 to $I$ do
        $\mathbf{S}_i^l \leftarrow f_i(\mathbf{F}_1^l, \mathbf{F}_2^l, \cdots, \mathbf{F}_{\lambda_l}^l)$
    end for
    $S_{Criteria}^l \leftarrow \{\mathbf{S}_1^l, \mathbf{S}_2^l, \cdots, \mathbf{S}_I^l\}$
end for
                                        ▷ Clustering Criteria
for $l$ from 1 to $L$ do
    for i from 1 to $I$ do
                                        ▷ Calculate the Spearman's correlation matrix
        $Sp_{ij}^l \leftarrow \rho(\mathbf{S}_i^l, \mathbf{S}_j^l), j = 1, \cdots, I$
    end for
    $(C_1^l, \cdots, C_K^l) \leftarrow K$-Means$(Sp_{1:}^l, \cdots, Sp_{I:}^l)$
end for
Obtain K-Means results: $(C_1^{:}, \cdots, C_K^{:})$
                                        ▷ Evolutionary Search
EA Hyperparameters: Population size $\mathcal{N}$, Number of iterations $\mathcal{I}$, Mutation Probability $\mathcal{M}$, Crossover Probability $\mathcal{C}$, Drop Ratio $\mathcal{D}$, Finetune Epochs $\mathcal{E}$.
$\mathcal{G}_0 \leftarrow (\mathbf{P}_0, \mathbf{S}_0)_{\mathbf{P}_0 \sim \mathcal{U}[0,1], \mathbf{S}_0 \sim (C_1^{:}, \cdots, C_K^{:})}$
$\mathbf{P}_0, \mathbf{S}_0 \in \mathbb{R}^{\mathcal{N} \times L \times K}$
for $iter$ from 1 to $\mathcal{I}$ do
    $\mathcal{G}_{Crossover} \leftarrow$ Crossover$(\mathcal{G}_{iter-1}, \mathcal{C})$
    $\mathcal{G}_{Mutation} \leftarrow$ Mutation$(\mathcal{G}_{Crossover}, \mathcal{M})$
    $\mathcal{G}_{Drop} \leftarrow$ Drop$(\mathcal{G}_{Mutation}, \mathcal{D})$
    $\mathcal{G}_{iter} \leftarrow \mathcal{G}_{Drop}$
    $\phi_{iter} \leftarrow$ Ensemble$(\mathcal{G}_{iter})$
    Acc$_{\phi_{iter}} \leftarrow$ Finetune$(\phi_{iter}, \mathcal{E})$
end for
$\mathbf{P}_{topk}, \mathbf{S}_{topk} \leftarrow$ TopK(Acc$_{\phi_{\mathcal{I}}}$)
$\phi_{\text{ensemble}}, S_{\text{ensemble}} \leftarrow$ Final$(\mathbf{P}_{topk}, \mathbf{S}_{topk})$
return $\phi_{\text{ensemble}}, S_{\text{ensemble}}$
```

3.1 Criteria Clustering

We first consider the selection complexity on the ensemble of K candidate criteria among N given criteria on a L-layer CNN. When $N = 12, K = 6$, and $L = 13$, we have $\left(C_{12}^{6}\right)^{13} \approx 3^{38}$ combinations. For the commonly used model, such number of selection will be extremely large. In addition, from Fig. 1, we observe that some of the filter pruning criteria have a strong similarity on the rank of the criteria score. As a result, they tend to prune a similar set of filters in one convolution layer. Moreover, in traditional implementation of the ensemble method (25; 26), its capability of achieving greater performance than an individual method comes from the diversity and effectiveness of the candidate methods that will be integrated. Given these above points, we cluster the given criteria set in each layer based on the Spearman's correlation matrix $\mathbf{Sp}$ before the ensemble, which is able to decrease the search space on the criteria selections efficiently.

When the K candidate criteria are obtained from each cluster in one layer, using rule of product and Arithmetic Mean-Geometric Mean Inequality, the upper bound of the search space is $\left(\frac{N}{k}\right)^k$. And

$$\left(\frac{N}{k}\right)^k = \prod_{j=0}^{k-1}\left(\frac{N}{k}\right) \leq \prod_{j=0}^{k-1}\left(\frac{N-j}{k-j}\right) = \frac{\prod_{j=0}^{k-1}(N-j)}{k!} = C_N^k. \tag{2}$$

When $k \in [1, N]$ is selected appropriately, we have $\left(\frac{N}{k}\right)^k \ll C_N^k$. If we sample the candidate criteria from different clusters, the Sp value between them should be relatively small (as shown in Fig. 3), *i.e.*, their filter importance rank would be dissimilar. Therefore, this clustering of criteria can not only maintain the criteria diversity, but also it can reduce the search space via selecting the number of clusters. Before clustering, we assign $criterion_i$ with a correlation vector,

$$Sp_{i:}^l = (\rho(\mathbf{S}_i^l, \mathbf{S}_1^l), \rho(\mathbf{S}_i^l, \mathbf{S}_2^l), \cdots, \rho(\mathbf{S}_i^l, \mathbf{S}_I^l)), \tag{3}$$

where $i = 1, 2, \cdots, I$ and ρ is Spearman's correlation defined in Eq. 1. Subsequently, we conduct K-Means to cluster the correlation vectors of the criteria into K clusters and obtain K clustering sets $\{C_1^l, \cdots, C_K^l\}$. The two criteria in the same cluster have similar correlation vectors in the sense of Euclidean distance. We want to point out that the two criteria in the same cluster have large Sp correlation value (see Fig. 3), which indicates the rank of these two criteria is similar. Therefore, we are able to sample criteria from each cluster whose in-between Sp is relatively small and indicate the adequate diversity of basic criteria for the ensemble.

3.2 Filters Importance Calibration

From Sec. 3.1, we obtained K clustering sets $(C_1^l, C_2^l, \cdots, C_K^l)$ in layer l. To filter out the similar criteria, we sample the distinctive criterion score $\mathbf{S}_{i_k}^l$ from each cluster C_k^l as the candidate for ensemble, where $i_k \in \{1, 2, \cdots, I\}$, $k = 1, 2, \cdots, K$. Thus, to combine those selective criteria and integrate their importance measurements to identify the redundancy, we calibrate their filters importance with the introduced filter importance calibration factors $p_k^l \in [0, 1]$. When filters in different layer extract multi-level features, we conduct layerwise criteria ensemble to adaptively discover their importance:

$$S_{\text{ensemble}}^l = \sum_{k=1}^{K} p_k^l \mathbf{S}_{i_k}^l. \tag{4}$$

The larger the value of p_k^l reveals that the cluster k is much significant for pruning filters in this layer.

We denote $\phi(\{S_{\text{ensemble}}^l\}_{l=1}^L)$ as the network after discarding filters according to the ensemble score. As the pruning objective is to remove redundancy without harming the model performance too much, therefore, our ensemble framework for filter pruning tends to discover the ensemble criterion (see Eq. 4), such that its pruned network maximize the accuracy after N-epoch fine-tuning in the validation set, *i.e.*,

$$\max_{\left\{p_k^l, \mathbf{S}_{i_k}^l; k=1,\cdots,K, l=1,\cdots,L\right\}} \text{Accuracy } \left(\phi\left(\left\{S_{\text{ensemble}}^l\right\}_{l=1}^L\right), N\right). \tag{5}$$

Since the objective function (5) is not differentiable, we consider using the Evolutionary Algorithm (EA) oriented by the validation accuracy to optimize the problem, as the superiority of EA in solving the related problems has been mentioned in Sec. 2. In the evolutionary search, the optimization fitness is the model evaluation result after N-epochs fine-tuning over part of the training set. The validation set and train set are split from the original training set, the splitting details are discussed in Sec. 4. To be specific, each evolution gene consists of the calibration combination and the pruned network in terms of the corresponding calibrated criterion. All calibration factors are initiated from the uniform distribution $\mathcal{U}[0, 1]$. Though the criteria in each cluster possess high similarity, they also have their ability to probe the network redundancy individually. To avoid sticking to one criterion that gives high ensemble accuracy, during crossover in EA, we give other criteria in the same cluster the opportunity to be selected again via random sampling in each evolutionary process.

After iterations of crossover, mutation and drop, the TopK genes with the highest validation accuracy are considered the potential optimal calibrated pruning criterion. Then, under these ensemble results, we can obtain the gene with the highest accuracy after the final one-shot pruning and fine-tuning, and the criterion under this gene is considered the optimal filter importance measurement to discard unimportant filters in this architecture.

4 Experiments

In this section, we evaluate the effectiveness of our proposed method over different image classification benchmarks.

Criteria. To compare, we implement four categories baselines of the state-of-the-art filters pruning criteria and their variants. We provide a more detailed baseline criteria overview in our supplementary. To briefly introduce, we consider:

- Filter based criteria: L1-Norm (5); L2-Norm (11); FPGM (9); Fermat (10).
- Batch Normalization (BN) based criteria: BN_γ scale (7); BN_β scale.

Table 1: Quantitative results on CIFAR-100 dataset

Model	Criterion	Pruned/Finetuned Acc.(%)	Acc.↓(%)	Model	Criterion	Pruned/Finetuned Acc.(%)	Acc.↓(%)
VGG16[*]	L1-Norm	15.76/71.29	0.93	ResNet56[†]	L1-Norm	52.16/69.43	0.07
	L2-Norm	16.24/71.32	0.90		L2-Norm	50.75/69.45	0.05
	Apoz	5.69/70.91	1.31		Apoz	2.03/63.68	5.82
	BN_γ	15.87/71.26	0.96		BN_γ	29.35/69.39	0.11
	BN_β	6.92/71.45	0.77		BN_β	22.44/69.32	0.18
	Entropy	11.80/71.09	1.13		Entropy	17.37/69.14	0.36
	FPGM	15.91/71.37	0.85		FPGM	51.20/69.50	0.00
	Fermat	15.39/71.39	0.83		Fermat	**51.45**/69.47	0.03
	Taylor L1-Norm	1.29/70.35	1.87		Taylor L1-Norm	15.77/69.27	0.23
	Taylor L2-Norm	16.19/71.19	1.03		Taylor L2-Norm	39.22/69.23	0.27
	Taylor_BN_γ	7.46/71.19	1.03		Taylor_BN_γ	30.41/69.34	0.16
	Taylor_BN_β	7.10/71.16	1.06		Taylor_BN_β	25.42/69.34	0.16
	Ours	**16.67/71.68**	**0.54**		**Ours**	40.38/**69.82**	**-0.32**

[*] VGG16 original Acc.: 72.22%
[†] ResNet56 original Acc.: 69.5%

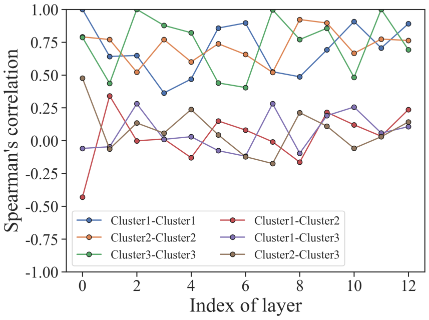

Figure 3: Average Sp of the criteria score between two clusters

Table 2: Quantitative results on ImageNet dataset

Model	Criterion	Pruned Finetuned Acc.(%)	Acc.↓(%)
ResNet34[*]	L1-Norm	59.08/72.76	1.03
	L2-Norm	61.02/72.77	1.02
	Apoz	4.70/72.19	1.60
	BN_γ	19.02/72.71	1.08
	BN_β	4.12/72.59	1.20
	Entropy	25.84/72.57	1.22
	FPGM	**62.28**/72.78	1.01
	Fermat	44.63/72.80	0.99
	Taylor L1-Norm	25.71/72.67	1.12
	Taylor L2-Norm	46.43/72.67	1.12
	Taylor BN_γ	27.47/72.65	1.14
	Taylor BN_β	18.00/72.67	1.12
	Ours	61.33/**72.85**	**0.94**

[*] ResNet34 original Acc.: 73.79%

- Feature maps activation based criteria: Average Percentage of Zeros (APoZ) (16)); Entropy (17).
- First-order Taylor based criteria (18; 13): Taylor L1-Norm; Taylor L2-Norm; Taylor BN_γ; Taylor BN_β.

Dataset. We conduct experiments on CIFAR-100 (27) and ImageNet (28) datasets. CIFAR-100 has 50k train images and 10k test images of size 32 by 32 from 100 classes. 10% of train images are split for validation and the remaining for training. ImageNet comprises 1.28 million train images and 50k validation images from 1000 classes. 50k out of 1.28 million train images (50 images in each class) are used for sub-validation. The cropping size 224 by 224 is used in our ImageNet experiments. Adopting the same predefined pruning configuration (5), we evaluate our method on VGG (14) and ResNet (1).

Results & Analysis. In Table. 1 and Table. 2, we present the quantitative comparison on CIFAR-100 and ImageNet, where the average accuracy over three repeated experiments are attached (denoted as Acc.). Comparing to the baselines, our method performs the best in the same settings.

According to Fig. 2, the optimization fitness in the evolutionary search needs N-epochs fine-tuning over part of the training set. Is N-epoch fine-tuning necessary in evolutionary search? To illustrate, we take VGG16 on CIFAR-100 and ResNet34 on ImageNet as examples. First, we calculate the Pearson correlation coefficient (R^2) between the accuracy of the pruned model without fine-tuning and the best accuracy after completed fine-tuning. In Fig. 4 (a) and (c), the value of R^2 indicates that the accuracy between the pruned model accuracy and the best accuracy does not have a strong linear relationship. Therefore, the accuracy of the pruned model without fine-tuning is not suitable as a metric for our Evolutionary process in Sec 3.2. However, after several epochs fine-tuning (as shown in Fig. 4 (b) and (d)), the R^2 improve significantly and it means that a certain amount of fine-tuning is necessary.

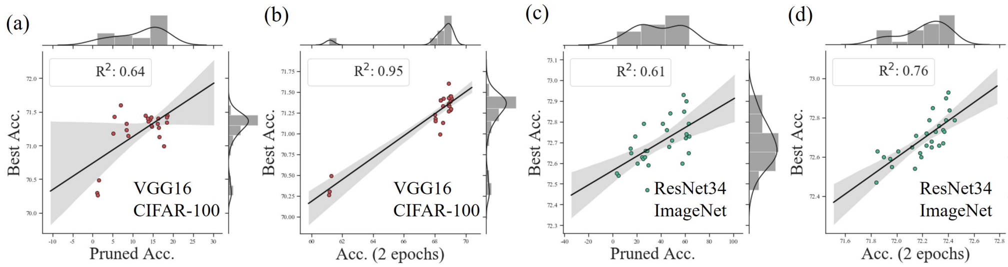

Figure 4: The R^2 between the Acc. on different fine-tuend epochs. (a) and (c): best fine-tuned Acc. vs pruned Acc.; (b) and (d): best fine-tuned Acc. vs Acc. after 2 epochs fine-tuning.

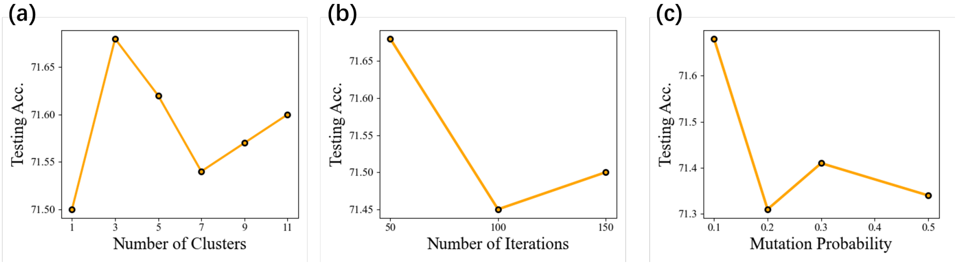

Figure 5: The testing accuracy of VGG16 after fine-tuning . (a) Comparison of the performance under different number of clusters. Note that when the number of clusters equals to 1, we choose one criterion at each layer and when the number of clusters equals to number of criteria, we choose all criteria (b) Comparison of the performance over different number of iteration in EA. (c) Comparison of the performance over mutation probability in EA.

5 Implementation Details

For the normal training and the fine-tuning, we use the SGD optimizer with momentum and weight decay parameter 0.9 and 0.0001 respectively. For fine-tuning, the learning rate is started at 0.001. On CIFAR-100, we fine-tune the pruned model for 40 epochs in batch size 64. On ImageNet, we fine-tune for 20 epochs with a mini-batch size of 256. For the evolutionary search setting, we consider the criteria as follows: L1-Norm, L2-Norm, FPGM, Fermat, BN_γ scale, BN_β scale, Entropy, Taylor L1-Norm, Taylor L2-Norm, Taylor BN_γ, Taylor BN_β. In each evolution iteration, we set the mutation probability to 0.1, crossover constant to 0.8 and drop probability to 0.05. In CIFAR-100, the population size is 20, number of iterations is 50 and the drop ratio is 0.08. In ImageNet, the population size is 10, number of iterations is 30 and the drop ratio is 0.1. After pruning the network based on the weighted sum of the scores, we fine-tune the pruned network on the validation split for 3 epochs in CIFAR-100 experiments and 1 epoch for ImageNet. Finally, the network is pruned under the optimal ensemble criterion obtained by EA, where the fine-tuning epochs is 40 on CIFAR-100 and 20 on ImageNet over the whole training set.

6 Ablation Studies

In this section, to understand the performance of our method in different settings, we conduct the following ablation experiments on the number of clusters used in Sec. 3.1 and the hyperparameters of Sec. 3.2. The ablation results are shown in Fig. 5 and the details are provided in supplementary. The search space of EA is related to the number of clusters. In each layer l, the larger the number of clusters K^l, the harder the optimization of the calibration factors p^l. From our experiments, the best performance appears when $K^l = 3$ for VGG16. Also, we compare the performance with different hyperparameters of EA, including the number of evolution iterations and mutation ratio. As we see in Fig. 5(b), the increment of iterations is not sufficient to the increment of performance. Therefore, we choose the relative small number of iterations and use top-K strategy. From Fig. 5(c), we observe that the large mutation probability may harm the search.

7 Conclusion

In this paper, we propose an ensemble framework for filter pruning, which includes two stages: to reduce the search space for criteria selection and the requirement for ensemble method, we first conduct clustering on the given criteria. In the second stage, we formulate the criteria ensemble problem as an optimization problem on filters importance calibration via Evolutionary algorithm. The comprehensive experiments on benchmarks exhibit that our criterion can outperform the current state-of-the-art criteria. Besides, we explore the correlation among existing filter pruning criteria and provides a way to obtain effective criteria without manual efforts. For the future work, we will incorporate more exploitation on our framework, *e.g.*, to add a constraint on computational cost for optimization during EA, to choose the layerwise number of clusters adaptively and to adopt the iterative or automatic pruning pipeline.

Broader Impact

We can conclude the broader impact of our work as follows:

- Our work is a general ensemble framework for different filter pruning criteria, which is flexible and easy-to-implement. Hence, it can be further extended by adding more novel criteria in the future and achieve much better performance.
- The effectiveness of leveraging the criteria diversity may potentially prompt the community to self-examine every novel pruning criterion and also inspire the researcher in this area to design an effective-yet-discrepant criterion.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[4] Z. Huang, S. Liang, M. Liang, and H. Yang, "Dianet: Dense-and-implicit attention network," *arXiv preprint arXiv:1905.10671*, 2019.

[5] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[6] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1389–1397, 2017.

[7] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2736–2744, 2017.

[8] Z. Zhuang, M. Tan, B. Zhuang, J. Liu, Y. Guo, Q. Wu, J. Huang, and J. Zhu, "Discrimination-aware channel pruning for deep neural networks," in *Advances in Neural Information Processing Systems*, pp. 875–886, 2018.

[9] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang, "Filter pruning via geometric median for deep convolutional neural networks acceleration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4340–4349, 2019.

[10] Z. Huang, X. Wang, and P. Luo, "Convolution-weight-distribution assumption: Rethinking the criteria of channel pruning," *arXiv preprint arXiv:2004.11627*, 2020.

[11] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2019.

[12] P. Sedgwick, "Spearman's rank correlation coefficient," *Bmj*, vol. 349, p. g7327, 2014.

[13] P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz, "Importance estimation for neural network pruning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11264–11272, 2019.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] Z. Huang and N. Wang, "Data-driven sparse structure selection for deep neural networks," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 304–320, 2018.

[16] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang, "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures," *arXiv preprint arXiv:1607.03250*, 2016.

[17] J.-H. Luo and J. Wu, "An entropy-based pruning method for cnn compression," *arXiv preprint arXiv:1706.05791*, 2017.

[18] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, 2016.

[19] Y. He, Y. Ding, P. Liu, L. Zhu, H. Zhang, and Y. Yang, "Learning filter pruning criteria for deep convolutional neural networks acceleration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies–a comprehensive introduction," *Natural computing*, vol. 1, no. 1, pp. 3–52, 2002.

[21] Z. Liu, H. Mu, X. Zhang, Z. Guo, X. Yang, K.-T. Cheng, and J. Sun, "Metapruning: Meta learning for automatic neural network channel pruning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3296–3305, 2019.

[22] F. E. F. Junior and G. G. Yen, "Pruning deep neural networks architectures with evolution strategy," *arXiv preprint arXiv:1912.11527*, 2019.

[23] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin, "Large-scale evolution of image classifiers," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2902–2911, JMLR. org, 2017.

[24] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4780–4789, 2019.

[25] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[26] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[27] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," tech. rep., Citeseer, 2009.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.