Discrimination-aware Channel Pruning for Deep Neural Networks

Zhuangwei Zhuang, Mingkui Tan†, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, Jinhui Zhu

Background and Motivation
Channel pruning reduces the model size and speeds up the inference by removing redundant channels directly. Existing methods include:

• Training-from-scratch methods: select channels to minimize the cross-entropy loss with sparsity regularization [1].
• Reconstruction-based methods: select channels to minimize the reconstruction error of feature maps between the pruned model and a pre-trained model [2].

Limitations of existing channel pruning methods:
• Training-from-scratch methods: are difficult to converge.
• Reconstruction-based methods: ignore the discriminative power.
• Both methods result in apparent drop in accuracy.

Our solution: propose a discrimination-aware channel pruning (DCP) scheme to choose channels with true discriminative power.

Contribution
We propose a discrimination-aware channel pruning (DCP) scheme to choose channels with true discriminative power.

We formulate the channel selection problem as an ε₂p-norm constrained optimization problem and propose a greedy method to solve the resultant optimization problem using SGD.

Extensive experiments demonstrate the effectiveness of DCP.

Problem Definition
Channel Pruning prunes those redundant channels in W to save the model size and accelerate the inference speed in Eq. (1):

\[ \mathbf{O}_{j,k} = \sum_{i \in \mathcal{A}} X_{i,k} \cdot W_{j,k,i} \]

where \( X_{i,k} \) is the input feature map, \( W_{j,k,i} \) denotes the parameters, and \( \mathbf{O}_{j,k} \) is the output feature map.

\[ \mathbf{F}_{\mathcal{P}} = \left\{ W_{j,k,i} \right\}_{j,k} \] is the selected channel subset.

\[ \mathbf{F}_{\mathcal{P}} \] is complementary set of \( \mathcal{A} \).

Algorithm 1 Discrimination-aware channel pruning (DCP)

Input: Pre-trained model \( M \), training data \( \{x_n, y_n\}_{n=1}^{N} \), and parameters \( \{\alpha_j\}_{j=1}^{J} \).

Output: Optimal selected channel subset \( \mathcal{P} \) and model parameters \( W_{\mathcal{P}} \).

1. Construct loss \( \mathcal{L}_2 \) to layer \( \mathcal{A} \) as in Figure 1.
2. Learn \( \mathbf{F}_0 \) and \( \mathbf{F}_1 \) of time-wise \( M \) with \( \mathcal{L}_2 \) and \( \mathcal{L}_1 \) for \( (\mathbf{F}_0, \mathcal{P}, \mathbf{F}_1) \) do
3. Do Channel Selection for layer \( \mathcal{A} \) using Algorithm 2.
4. end

DPC introduces P discrimination-aware losses \( \mathcal{L}_{S,j} \) (cross-entropy loss) and updates the model to increase the discriminative power of intermediate layers.

DPC performs channel pruning with \( (P + 1) \) stages.

Greedy Algorithm
Algorithm 2 Greedy algorithm for channel selection

Input: Training data, model \( M \), parameters \( \alpha_j \), and \( \epsilon \).

Output: Selected channel subset \( \mathcal{P} \) and model parameters \( W_{\mathcal{P}} \).

while (stopping conditions are not achieved) do

Compute gradient of \( \mathcal{L}_S \) w.r.t. \( W_j \).

Find the channel \( \mathcal{P} \) with the max gradient norm \( \mathcal{L}_S(W_j) \).

Let \( \mathcal{P} = \mathcal{P} \cup \{j\} \).

Solve Problem (4) to update \( W_j \).

end while

Instead of solving problem (3), DCP uses a greedy algorithm to optimize \( W \) w.r.t. \( \mathcal{P} \) selected channels by minimizing:

\[ \mathcal{L}(W) = \sum_{j \in \mathcal{P}} \mathcal{L}_S(W_j) \]

where \( \mathcal{L}_S \) denotes the submatrix indexed by \( \mathcal{P} \) which is the complementary set of \( \mathcal{P} \).

Exploring pruning rate and \( \lambda \)

Table 5: Training results on ResNet-56 with different pruning rates. We report the top-1 and top-5 error (%) on ILSVRC-12.

Table 4: Comparisons on ResNet-18 and ResNet-50 with different pruning rates. We report the top-1 and top-5 error (%) on ILSVRC-12.

Table 3: Comparisons on CIFAR-10 with different pruning rates. We report the top-1 and top-5 error (%) on ILSVRC-12.

Table 2: Comparisons on CIFAR-10 and ILSVRC-12.

Table 1: Comparisons on CIFAR-10. We display the results as reported.

Table 5: Effect of \( \epsilon \) for channel selection over VGGNet on CIFAR-10.

Results on CIFAR-10 and ILSVRC-12

Effect of the Stopping Condition

• Given a predefined parameter \( \alpha_j \), Algorithm 2 will be stopped if \( \epsilon \) is not increased.
• Since \( \mathcal{L} \) is convex, \( \mathcal{L}(W_{\mathcal{P}}) \) will monotonically decrease with iteration index \( t \) in Algorithm 2. The number of selected channels can be automatically determined by following stopping condition:

\[ \left| \mathcal{L}(W_{\mathcal{P}}^{t-1}) - \mathcal{L}(W_{\mathcal{P}}^t) \right| < \epsilon \]

Table 5: Effect of \( \epsilon \) for channel selection over VGGNet on CIFAR-10.

Effect of the Stopping Condition

• A smaller \( \epsilon \) leads to better performance of the pruned model.

Visualisation of Feature Maps

• Feature maps of the pruned channels are less informative.

References


Contact Information

• Correspondence to Prof. Mingkui Tan
• Email: mingkui@scut.edu.cn
• School: South China University of Technology

Effect of the Stopping Condition

Table 5: Effect of \( \epsilon \) for channel selection over VGGNet on CIFAR-10.

Convergence of Discrimination-Aware Channel Pruning

Figure 1: The architecture of discrimination-aware channel pruning.

Proposition 1. (Convexity of the loss function) \( \mathcal{L}(W_{\mathcal{P}}) \) is convex w.r.t. \( W \).

Greedy Algorithm

Algorithm 2 Greedy algorithm for channel selection

Input: Training data, model \( M \), parameters \( \alpha_j \), and \( \epsilon \).

Output: Selected channel subset \( \mathcal{P} \) and model parameters \( W_{\mathcal{P}} \).

while (stopping conditions are not achieved) do

Compute gradient of \( \mathcal{L}_S \) w.r.t. \( W_j \).

Find the channel \( \mathcal{P} \) with the max gradient norm \( \mathcal{L}_S(W_j) \).

Let \( \mathcal{P} = \mathcal{P} \cup \{j\} \).

Solve Problem (4) to update \( W_j \).

end while

Instead of solving problem (3), DCP uses a greedy algorithm to optimize \( W \) w.r.t. \( \mathcal{P} \) selected channels by minimizing:

\[ \mathcal{L}(W) = \sum_{j \in \mathcal{P}} \mathcal{L}_S(W_j) \]

where \( \mathcal{L}_S \) denotes the submatrix indexed by \( \mathcal{P} \) which is the complementary set of \( \mathcal{P} \).

Exploring pruning rate and λ

Table 4: Comparisons on ResNet-18 and ResNet-50 with different pruning rates. We report the top-1 and top-5 error (%) on ILSVRC-12.

Table 5: Training results on ResNet-56 with different pruning rates. We report the top-1 and top-5 error (%) on ILSVRC-12.

Table 3: Comparisons on CIFAR-10. We display the results as reported.

Table 2: Comparisons on CIFAR-10 and ILSVRC-12.

Table 1: Comparisons on CIFAR-10. We display the results as reported.