One-shot Network Pruning at Initialization with Discriminative Image Patches

Yinan Yang
gr0528xf@ed.ritsumei.ac.jp

Yu Wang
yu.wang@r.hit-u.ac.jp

Ying Ji
jiying@nagoya-u.jp

Heng Qi
hengqi@dlut.edu.cn

Jien Kato
jien@fc.ritsumei.ac.jp

1 Ritsumeikan University
Shiga, Japan
2 Hitotsubashi University
Tokyo, Japan
3 Nagoya University
Nagoya, Japan
4 Dalian University of Technology
Dalian, China

Abstract

One-shot Network Pruning at Initialization (OPaI) is an effective method to decrease network pruning costs. Recently, there is a growing belief that data is unnecessary in OPaI, e.g. [22, 32]. However, we obtain an opposite conclusion by ablation experiments in two representative OPaI methods, SNIP [22] and GraSP [37]. Specifically, we find that informative data is crucial to enhancing pruning performance. In this paper, we propose two novel methods, Discriminative One-shot Network Pruning (DOP) and Super Stitching, to prune the network by high-level visual discriminative image patches. Our contributions are as follows. (1) Extensive experiments reveal that OPaI is data-dependent. (2) Super Stitching performs significantly better than the original OPaI method on benchmark ImageNet, especially in a highly compressed model.

1 Introduction

Since the 1980s, we have known that it is possible to substantially reduce parameters in neural networks without seriously compromising performance [16, 31]. Such pruned neural networks can significantly decrease the computational demands of inference by using specific methods [8, 14]. Among the various pruning methods developed so far, Network Pruning at Initialization (PaI) has attracted considerable attention since it provides a possibility to train sparse networks at lower costs [38]. Specifically, PaI aims to achieve (close to) full accuracy (the accuracy reached by the dense network) by training a sparse subnetwork from a randomly initialized dense network.

In PaI research, the pruning criterion is the key focus [11, 21, 26, 28, 29, 35, 38]. Most PaI works involve iterative pruning processing to improve performance while drastically increasing training costs. In contrast, One-shot Network Pruning at Initialization (OPaI),
Figure 1: **Overview of the Discriminative One-shot Network Pruning (DOP) and Super Stitching.**

1. Cluster segments in trained network’s activation space, extract Discriminative Image Patches. The green is meaningful in network prediction, and the red is meaningless.

2. Using Discriminative Image Patches or Super Stitching to prune unimportant parameters by a specific OPaI algorithm.

Another branch of PaI, attempts to reduce costs by single-step pruning. Specifically, SNIP [22] and GraSP [37], two representative methods of OPaI, use gradient information in the initial network to find subnetworks. Both algorithms employ random mini-batches in the pruning step, however, the data’s role has not been elucidated. Furthermore, despite the lack of extensive experimental evidence, there is a growing belief that data is not essential in OPaI [12, 19, 32], which may impact future OPaI or even PaI research.

This work questions the presumption of data independence in OPaI. To find the answer, we present Discriminative One-shot Network Pruning (DOP), as shown in Fig. 1. Compared to previous studies, we employ discriminative data rather than random mini-batches. As a result, more precise gradient information is retained, so crucial structures and parameters are preserved in the network. To seek critical data in a trained classifier for targeted pruning, we distinctively generate discriminative image patches by Automatic Concept-based Explanation (ACE) [15]. In the ACE algorithm, visual concepts are vital for predicting a certain class and are automatically extracted for each class. The extracted visual concepts are processed into our discriminative image patches for subsequent pruning. Our extensive experiments on the benchmark ImageNet confirmed that discriminative data is significant to OPaI. Moreover, we propose an advanced strategy, Super Stitching, to further enhance the pruning performance. Super Stitching combines dozens of concepts in the same class but with different semantics. Such stitching patches are like chunks of condensed class information that enable pruning algorithms to get competitive outcomes with fewer samples. The experiment results on benchmark ImageNet show that Super Stitching is superior to the original SOTA OPaI method. Our major contributions are:

1. Our experiments show that using discriminative data enhances pruning performance dramatically. We experimentally demonstrate that OPaI is data-dependent, which refreshes the knowledge of OPaI [12, 32].
2. We propose two novel data-dependent OPaI methods, DOP and Super Stitching. Super Stitching significantly improves the pruned network performance, especially in a highly compressed model. We use the One-Shot method with fewer samples to achieve higher or similar results than iterative pruning methods and SOTA. Our research demonstrates that, using informative data, One-Shot PaI can even outperform iterative pruning.

2 Related work

Neural network pruning research started in the late 1980s [31], and it is become increasingly popular until recently [16, 38]. There are three steps in a typical pruning pipeline: (1) pre-training, (2) pruning, and (3) fine-tuning [16, 31]. Pruning has been regarded as a post-processing approach to solving the problem of redundancy [3, 6, 7, 13, 19, 23, 34]. However, a novel pruning paradigm, Network Pruning at Initialization (PaI), is out of this pipeline [22], which trains a sparse subnetwork from a randomly initialized dense network instead of pruning a pre-trained network. PaI seeks to achieve full accuracy at lower training costs. One of the notable PaI efforts is the lottery ticket hypothesis (LTH) [11], which selects subnetwork masks from pre-trained dense networks. This outstanding work sparks an interest in the deep learning community, and numerous related PaI researches have emerged [4, 25, 27, 38, 41, 43].

Most traditional PaI methods, such as LTH, dramatically raise the pruning cost since iterative pruning is needed. A distinctive branch of PaI, One-shot Network Pruning at Initialization (OPaI), solves this problem by computing the importance of each parameter in a neural network in a single step [2, 5, 17, 24, 30, 33, 38, 39, 42]. Two basic OPaI methods are SNIP [22] and GraSP [37]. Concretely, OPaI uses random mini-batches to calculate the connection sensitivity (SNIP) or the gradient signal preservation (GraSP) with respect to the loss for each parameter as an important score. Then, unimportant parameters are masked by a constant value of 0, and a trained sparse network is obtained by regular fine-tuning.

A crucial concern has been raised about whether OPaI is data-dependent. For a long time, although it has been assumed that pruning methods use information from training data to find subnetworks, data in pruning has not received sufficient attention. Specially, Su et al. [32] claim that data is unessential in SNIP and GraSP since the subnetworks generated by the corrupted data (dataset with random labels or pixels) behave similarly to the original one. However, such a perspective is not convincing because their pruning step actually still relies on information from the training set. More explicitly, their experiments cannot be considered fully data-independent, because the corrupted data retain the same distribution as the original one. Recently, it is increasingly accepted that OPaI is independent of data, yet more experiments to support this conclusion are lacking. For instance, [43] and [36] discuss data independence as a limitation of OPaI. [22] develops Layerwise Sparsity Quotas based on the assumption of data-independence. Nevertheless, in numerous benchmark tests, the different samples are one of the reasons for the disparities in outcomes [1, 2, 37], which indicates that data matters in OPaI. To recap, the function of data in OPaI is still unclear and needs to be researched urgently.

In contrast to earlier work that focused on pruning criteria, we investigate whether OPaI is data-dependent. Our work experimentally demonstrates that the improvement of OPaI performance requires informative data. From this viewpoint, we then propose our method, Super Stitching, and show it significantly outperforms the original OPaI approach on bench-
mark ImageNet. Our method beats the SOTA on ImageNet with a 95% sparsity ResNet-50 model.

3 Proposed Approaches

Overview: This section introduces our two approaches. Section 3.1 simply recalls two OPaI methods, SNIP and GraSP. They prune network by random mini-batches at Initialization. Section 3.2 introduces our approach, DOP, which uses Discriminative Image Patches to replace random mini-batches in OPaI. To improve the gradient information quality in pruning, we introduce Super Stitching in Section 3.3.

Notions: We use $T = \{(X, Y)\}$ to denote a certain classification task, where $x_i \in X$ represents a sample, and $y_i \in Y$ represents its label. We consider a neural network $f(T, \theta)$ before pruning for task $T$ with parameters $\theta$, and $\theta_j$ is the parameter of connection $j$ in $f(T, \theta)$. An OPaI pruning criterion $A$ produces $j$’s binary mask, denoted by $m_j \in \{0, 1\}$. Given a sparsity level $\kappa$, the training of $f(T, \theta \odot m)$ following the Minimization Empirical Risk:

$$\arg\min_{m, \theta} \frac{1}{n} \sum^n_i \mathcal{L}( (x_i, f(\theta \odot m)), y_i ), \quad \|m\|_0 \leq \kappa,$$

(1)

where $\mathcal{L}$ denotes the loss function, and $\odot$ denotes the Hadamard product.

3.1 Recall SNIP and GraSP

In SNIP and GraSP, the first step is to sample mini-batches from the training set. The random mini-batch $T^b$ is

$$T^b = \{(x_i, y_i)\}^b_{i=1} \sim T,$$

(2)

which is used for computing gradient information. The important score $s(\theta_j)$ of parameter $\theta_j$ in the two methods are as follows.

**SNIP** computes the connection sensitivity of each parameter $\theta_j$ as an important score. We use the right-hand formula [9, 12, 32] to simplify the calculations:

$$s(\theta_j) = \left. \frac{\partial L(T^b; \theta_j \odot m_j)}{\partial m} \right|_{m=1} = \left. \frac{\partial L(T^b; \theta_j)}{\partial \theta_j} \odot \theta_j \right|.$$  

(3)

**GraSP** uses the Hessian $H$ to preserve the gradient flow in the final sparse network. The score is designed as follows:

$$s(\theta_j) = -\theta_j \left. \frac{\partial L(T^b; \theta_j)}{\partial \theta_j} \right|.$$  

(4)

Once the importance scores of SNIP and GraSP are obtained, then, only the top-$\kappa$ connections are kept. And we finally fine-tunes the pruned network with the mask.

3.2 Discriminative One-shot Network Pruning

To replace $T^b$ with “better data” for pruning, we adopt ACE algorithm [15] to extract concepts from training data. Then these visual concepts are processed into our discriminative
Figure 2: Discriminative image patches extracted by pre-trained ResNet-50 from the ImageNet. We choose the penultimate layer of ResNet-50 for the activation space. Here we show Siberian husky and electric guitar in each class’s Top-5 important concepts based on their TCAV scores. The original images from the training set are shown below, and the discriminative image patches are shown above.

image patches. ACE adopts super-pixel segmentation on the training set, clusters the activations in Euclidean Space, and ranks the concepts by TCAV scores [20]. The higher the TCAV score, the more discriminative the concepts for the recognition of a certain class. Specifically, ACE needs a trained classifier \( F(T) \) with a bottleneck layer \( l \), as shown in Figure 1. We denote \( k \) as a class label in \( T \), and \( X_k \) is all inputs with that given label. We first segment images \( X_k \) by a super-pixel algorithm called SLIC [1]. Then, we let the segments flow to \( F(T) \), pick a bottleneck layer \( l \) in \( F(T) \) as activation space, and cluster similar segments in \( l \) by K-Means. At that point, each harvested cluster plays a specific role in a target class, and we obtain a total of \( p \) clusters, with \( q \) independent segments in each cluster. Finally, we view each cluster with the same semantics as a visual concept patch and unite a segment with a mean image value matrix (with the same size of \( x_i \)) to generate an image so-called discriminative image patch. We denote a set of discriminative image patches as \( C_T \sim \{(c_{k,p,q},y_i)|c_{k,p,q} \in x_i\} \), here \( c_{k,p,q} \) means a discriminative image patch.

Algorithm 1 DOP: Discriminative One-shot Network Pruning

Require: Network \( f(T, \theta) \), loss function \( \mathcal{L} \), sparsity level \( \kappa \), a trained classifier \( F(T) \) with a bottleneck layer \( l \), and an OPaI pruning criterion \( A \)

1: \( C_T \sim \{c_{k,p,q},y_i\} \leftarrow \text{ACE}(F(T), l) \quad \triangleright \text{Create Discriminative Image Patches} \\
2: C_T^b = \{c_{k,p,q},y_i\}_{i=1}^b \sim C_T \quad \triangleright \text{Sample mini-batches} \\
3: S(\theta) = A(C_T^b, f(T, \theta)) \quad \triangleright \text{pruning by } A \text{ with Discriminative Image Patches} \\
4: m \leftarrow 1[\bar{s}_\theta - \bar{s}_\kappa \geq 0] \quad \triangleright \text{Make mask by keeping Top-}\kappa \text{ score} \\
5: f(T, \theta \odot m) \leftarrow \arg\min_\theta \mathcal{L}(T; \theta \odot m) \quad \triangleright \text{Fine-tune the sparse network with mask}

During extracting visual concepts, based on the trained classifier, we derive concept activation vectors (CAVs) [15, 20, 40] as the normal to a hyperplane that separates concept and random samples. By computing the similarity between the loss gradient and the CAV, we can finally estimate the importance rating of each concept conditioned on a certain target label. Specially, the importance rating of a certain concept \( c \), also called the TCAV (Test
with CAVs) score, is defined as follows,

\[ \text{TCAV}_c = \bigg| \frac{\sum_{x_i \in X_k, c \in x_i : \nabla h^l_k(F(c)) \cdot v^l_c > 0}}{|X_k|} \bigg|, \]

(5)

where \( v^l_c \) is the binary linear classifier of \( c \) with random samples at the \( l \) layer of \( F(T) \). \( F(c) \) is \( c \)'s activation in \( l \), and \( \nabla h^l_k(F(c)) \) is the logit for \( c \) towards class \( k \) in \( l \).

Importance of each discriminative image patch is evaluated by the TCAV score. As shown in Figure 2, we extract and rank the discriminative image patches. After this, we sample mini-batches in discriminative image patches and then prune the network using a specific OPaI pruning criterion \( \mathcal{A} \), which is shown in Algorithm 1. We name this pruning algorithm Discriminative One-shot Network Pruning (DOP).

### 3.3 Super Stitching

In addition to DOP, we propose a further data-preprocessing strategy, Super Stitching. This method uses the data to get strengthened gradient information in pruning. Firstly, we define the coverage of valid pixels in each discriminative image patch, \( r(c_{k,p,q}) \), as follows,

\[ r(c_{k,p,q}) = \frac{|c_{k,p,q}|}{|x_i|}, \quad c_{k,p,q} \in x_i, \]

(6)

which is a ratio of the number of valid pixels in the discriminative image patch over the total number of pixels in the image.

**Algorithm 2 Get Super Stitching Image Patch**

**Require:** \( c_k \) with \( p \) concepts and \( q \) segments per concept, threshold \( \sigma \)

1: \( C_k = 0 \), waiting_list = \( \emptyset \)
2: Sort concepts \( c_{k,p} \) by importance in ascending order.
3: while \( r(C_k) < \sigma \) do
   4:     if waiting_list == \( \emptyset \) then
   5:         waiting_list = \( c_k \) \# Ensure waiting list is not empty.
   6:     for idx in \{1, ..., p\} do
   7:         temp = RandomPop(\( c_{k,\text{idx}} \)) \# Random Pop from concept \( c_{k,\text{idx}} \).
   8:         \( C_k += \) temp \# Stitch segment into target image patch.
   return \( C_k \)

Second, we set a hyperparameter threshold \( \sigma \) to ensure that the coverage of valid pixels in each stitching patch is larger than \( \sigma \). Then, we can obtain the Super Stitching discriminative
The results are shown in the Table 1. Sparsity percentage is the proportion of parameters which are equal to 0 in the overall model parameters. Based on SNIP pruning criterion, we show that our DOP dominates the original SNIP at all four sparsity levels {60%, 80%, 90%, 95%} at Top-1 accuracy performances. When the sparsity reaches 90%, the gap has widened
If the OPaI is data-independent, then the same results should be obtained in the ablation experiments. However, we still observe disparity in ablation experiments. Such trends indicate that OPaI is impacted by limited validation information.

The results indicate that, for SNIP and GraSP, the discriminative image patches benefit the pruning performance. The failure of GraSP with DOP at 95% sparsity could be owing to a lack of valid information. We verify the guess in ablation experiments in Section 4.2 and propose Super Stitching in Section 4.3 to solve the problem.

4.2 Ablation Experiments on Discriminative Image Patches

We design three ablation experiments to investigate the function of data in OPaI by gradually changing the content of input data. If the same experimental results are obtained with various input data, then the data-independence can be verified; otherwise OPaI is data-dependent. The three ablation experiments are designed as follows.

**All-One Matrix** explores the case where the image has no content in pruning. In this experiment, we prune the network using all-one matrices with the same size and labels as the original images. If All-One Matrix performs similarly to SNIP and GraSP, it proves the data-independence in OPaI; otherwise, OPaI should be data-dependent. This attempt is similar to the work of Tanaka et al. [35], who devise a fully data-independent iterative pruning method to avoid layer collapse. However, we explore the impact of image content on OPaI, which is pretty different from the purpose of [35].

**Random Segment** considers the segmentation impact on OPaI. We adopt the SLIC algorithm [1] used by ACE to segment the images, and randomly select the same number of segments. We aim to confirm that the original images could include not only discriminative contents, but also image segments that are meaningless for pruning.

**Less Patch** explores the influence of the material number on pruning when informative data become less. Only one discriminative image patch is selected for each class in the pruning step, i.e., a total of 1,000 discriminative image patches.

Except Less Patch, all ablation experiments adopt the same number of materials to guarantee a fair comparison, i.e., 10,000 materials. The results are shown in Figure 4. **All-One Matrix** performs worse in all trails. As the sparsity increases, the gap between **All-One Ma-
<table>
<thead>
<tr>
<th>Method</th>
<th>Material Number</th>
<th>Material Type</th>
<th>Sparsity percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>60%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>GraSP</td>
<td>30,000</td>
<td>image</td>
<td>73.87%</td>
</tr>
<tr>
<td>GraSP with DOP (Ours)</td>
<td>30,000</td>
<td>discriminative patch</td>
<td>74.19%</td>
</tr>
<tr>
<td>GraSP with Super Stitching (Ours)</td>
<td>30,000</td>
<td>stitching patch (\sigma = 0.5)</td>
<td>74.14%</td>
</tr>
<tr>
<td>GraSP with Super Stitching (Ours)</td>
<td>10,000</td>
<td>stitching patch (\sigma = 0.75)</td>
<td>73.94%</td>
</tr>
<tr>
<td>GraSP (Wang et al. [37])</td>
<td>37,500</td>
<td>image</td>
<td>74.02%</td>
</tr>
<tr>
<td>GraSP (Jorge et al. [9])</td>
<td>614,440</td>
<td>image</td>
<td>73.4%</td>
</tr>
<tr>
<td>GraSP (Frankle et al. [12])</td>
<td>10,000</td>
<td>image</td>
<td>73.4%</td>
</tr>
<tr>
<td>FORCE (Jorge et al. [9])</td>
<td>614,440</td>
<td>image</td>
<td>73.4%</td>
</tr>
<tr>
<td>Iter SNIP (Jorge et al. [9])</td>
<td>614,440</td>
<td>image</td>
<td>73.4%</td>
</tr>
<tr>
<td>SynFlow (Hayou et al. [17])</td>
<td>-</td>
<td>all-one matrix</td>
<td>-</td>
</tr>
<tr>
<td>SBP-SR (Hayou et al. [17])</td>
<td>-</td>
<td>image</td>
<td>-</td>
</tr>
<tr>
<td>ProsPr (Alizadeh et al. [2])</td>
<td>-</td>
<td>image</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: DOP and Super Stitching test accuracy of ResNet-50 on ImageNet based on GraSP pruning criterion. The empty means that the description or experiment is missing. At high sparsity, we can observe that Super Stitching has a significant advantage. Also, DOP has advantages at low sparsity.

...and the network pruned by the image or discriminative image patches becomes wider. Random Segment performs similarly to DOP at lower sparsity but worse than DOP at higher sparsity. It verifies our guess that the images include both useful discriminative information and non-useful information for pruning. In Less Patch, the performance is worse than DOP, and GraSP is more sensitive to the amount of informative data than SNIP. GraSP with Less Patch is even worse than All-One Matrix at some sparsity levels. It confirms that our experimental results in Section 4.1, namely, GraSP requires more helpful information at high sparsity to prune.

By gradually changing the data input in the OPaI algorithm, we explicitly show that valid information in the OPaI method directly impacts the pruning performance in the experiments.

### 4.3 Super Stitching Compared to SOTA

We create Super Stitching to further improve the discriminative image patches in OPaI. We aim to use fewer but more informative samples to enhance the gradient flow. In implementation details, we set a coverage threshold \(\sigma \in (0.5, 0.75)\) to balance the overlap and performance. Figure 3 shows the stitched discriminative image patches with \(\sigma = 0.75\). We adopt the GraSP pruning criterion, since GraSP achieves better performance for ResNet-50 on ImageNet than SNIP [9, 12, 37].

The results of Super Stitching experiments are in Table 2. The first part of the table shows our pruning experiment results on ImageNet for ResNet-50 with a fixed random seed. At the second part of the table, we report the results in corresponding papers of ResNet-50 on ImageNet, where FORCE [9], Iter SNIP [9] and SynFlow [35] are iterative pruning algorithms. We notice that SBP-SR [17] claims a higher performance for ResNet-50 on ImageNet recently, which is the SOTA OPaI method. When the sparsity is high, we can observe that Super Stitching has a distinct advantage with fewer samples. Furthermore, DOP has an advantage in low sparsity. It is noticed that DOP and Super Stitching achieved their best results at 60% and 95% sparsity, respectively, outperforming the original SOTA. The results demonstrate that discriminative data in OPaI is able to perform better than iterative...
pruning with less costs.

5 Conclusion

To explore whether data matters in OPaI, we introduce DOP and Super Stitching, two novel data-dependent methods for OPaI based on discriminative image patches. Our research not only reveals that informative data is helpful in OPaI, but it also shows that DOP and Super Stitching can significantly improve pruning performance. This conclusion refreshes our typical views of the OPaI method and provides us with a new route for OPaI advancement. It is especially worth noting that our methods and experiments can help us understand which data are critical for networks at an earlier stage – providing valuable suggestions for PAI improvement.

6 Acknowledgments

This work is supported by the Initiative for Realizing Diversity in the Research Environment (Advanced Type).

References


