One-shot Network Pruning at Initialization with Discriminative Image Patches

Yinan Yang[1]
gr0528xf@ed.ritsumei.ac.jp
Yu Wang[2]
yu.wang@r.hit-u.ac.jp
Ying Ji[3]
jiying@nagoya-u.jp
Heng Qi[4]
hengqi@dlut.edu.cn
Jien Kato[1]
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.[12, 32]. However, we obtain an opposite conclusion by ablation experiments in two representative OPaI methods, SNIP [20] 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),

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

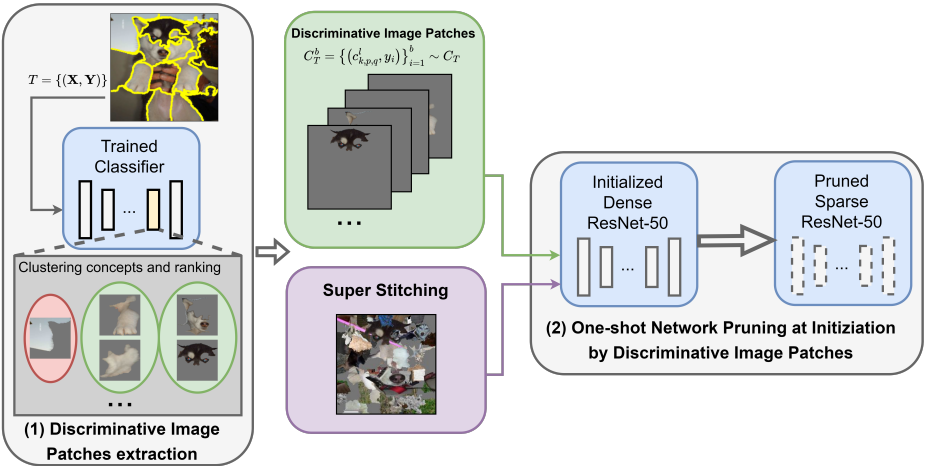

Figure 1: **Overview of the Discriminative One-shot Network Pruning (DOP) and Super Stitching.** (1) Cluster segments in trained network's activation space, extract Discriminative Image Patches. The green is meaningful in network prediction, and the red is meaningless. (2) Using Discriminative Image Patches or Super Stitching to prune unimportant parameters by an 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 signifiant 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 [38], 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 training set. Recently, it is increasingly accepted that OPaI is independent of data, yet more experiments to support this conclusion are lacking. For instance, [12] and [19] discuss data independence as a limitation of OPaI. [36] 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 [9, 12, 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 $T = \{(\mathbf{X}, \mathbf{Y})\}$ to denote a certain classification task, where $x_i \in \mathbf{X}$ represents a sample, and $y_i \in \mathbf{Y}$ represents its label. We consider a neural network $f(T, \theta)$ before pruning for task T with parameters θ, and θ_j is the parameter of connection j in $f(T, \theta)$. An OPaI pruning criterion $\mathcal{A}$ produces j's binary mask, denoted by $m_j \in \{0, 1\}$. Given a sparsity level κ, the training of $f(T, \theta \odot m)$ following the Minimization Empirical Risk:

$$\arg\min_{m,\theta} \frac{1}{n} \sum_{i}^{n} \mathcal{L}\left(\left(x_i, f\left(\theta \odot m\right)\right), y_i\right), \quad \|m\|_0 \leq \kappa, \tag{1}$$

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

$$T^b = \{(x_i, y_i)\}_{i=1}^{b} \sim T, \tag{2}$$

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

SNIP [22] computes the connection sensitivity of each parameter θ_j as an important score. We use the right-hand formula [9, 12, 32] to simplify the calculations:

$$s(\theta_j) = \left| \frac{\partial L(T^b; \theta_j \odot m_j)}{\partial m} \Big|_{m=1} \right| = \left| \frac{\partial L(T^b; \theta_j)}{\partial \theta_j} \odot \theta_j \right|. \tag{3}$$

GraSP [32] uses the Hessian H to preserve the gradient flow in the final sparse network. The score is designed as follows:

$$s(\theta_j) = -\theta_j (H \frac{\partial L(T^b; \theta_j)}{\partial \theta_j})_j. \tag{4}$$

Once the importance scores of SNIP and GraSP are obtained, then, only the top-κ 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

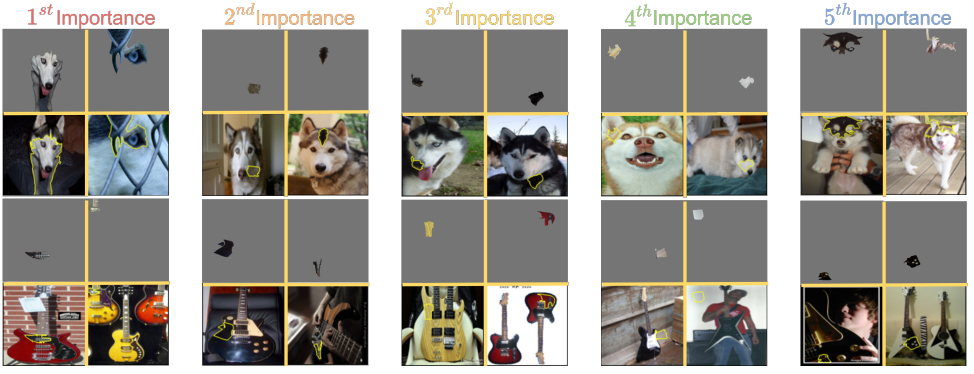

Figure 2: Discriminative image patches extracted by pre-trained ResNet-50 from the ImageNet. We choose the penultimate layer of ResNet-50 for the activation space. Here we show Siberian husky and electric guitar in each class's Top-5 important concepts based on their TCAV scores. The original images 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 such a 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 κ, a trained classifier $F(T)$ with a bottleneck layer l, and an OPaI pruning criterion $\mathcal{A}$

1: $C_T \sim \left\{\left(c_{k,p,q}, y_i\right)\right\} \leftarrow \text{ACE}\left(F(T), l\right)$ ▷ Create Discriminative Image Patches
2: $C_T^b = \left\{\left(c_{k,p,q}, y_i\right)\right\}_{i=1}^{b} \sim C_T$ ▷ Sample mini-batches
3: $S(\theta) = \mathcal{A}\left(C_T^b, f(T,\theta)\right)$ ▷ pruning by $\mathcal{A}$ with Discriminative Image Patches
4: $m \leftarrow \mathbb{1}\left[s_\theta - \tilde{s}_\kappa \geq 0\right]$ ▷ Make mask by keeping Top-κ score
5: $f(T, \theta \odot m) \leftarrow \arg\min_\theta \mathcal{L}(T; \theta \odot m)$ ▷ 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

Figure 3: Super Stitching with $\sigma = 0.75$. From left to right: zebra, goldfish, Siberian husky, ambulance, cash machine, and window screen.

with CAVs) score, is defined as follows,

$$\mathrm{TCAV}_c = \frac{\left|\left\{x_i \in X_k, c \in x_i : \bigtriangledown h_k^l\left(F_l\left(c\right)\right) \cdot v_c^l > 0\right\}\right|}{|X_k|}, \tag{5}$$

where v_c^l is the binary linear classifier of c with random samples at the l layer of $F(T)$. $F_l(c)$ is c's activation in l, and $\bigtriangledown h_k^l(F_l(c))$ 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,

$$r\left(c_{k,p,q}\right) = \frac{\left|c_{k,p,q}\right|}{|x_i|}, \quad c_{k,p,q} \in x_i, \tag{6}$$

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 σ

```
ℂ_k = 0, waiting_list = ∅
Sort concepts c_{k,p} by importance in ascending order.
while r(ℂ_k) < σ do
    if waiting_list == ∅ then
        waiting_list = c_k                    ▷ Ensure waiting list is not empty.
    for idx in {1, ..., p} do
        temp = RandomPop(c_{k,idx})           ▷ Random Pop from concept c_{k,idx}.
        ℂ_k += temp                           ▷ Stitch segment into target image patch.
return ℂ_k
```

Second, we set a hyperparameter threshold σ to ensure that the coverage of valid pixels in each stitching patch is larger than σ. Then, we can obtain the Super Stitching discriminative

Sparsity percentage (Baseline)	60% 76.47%	80%	90%	95%
SNIP	74%	70.94%	61.06%	36.43%
SNIP with DOP (Ours)	**74.29%**	**71.15%**	**64.12%**	**48.14%**
GraSP	73.87%	71.14%	67.07%	**61.76%**
GraSP with DOP (Ours)	**74.19%**	**71.76%**	**67.65%**	60.02%

Table 1: Top-1 Test Accuracy of ResNet-50 on ImageNet.

image patches in the same class from Algorithm 2, noted as $\mathbb{C}_k$. Since we keep the same position of patches as the original images, when stitching a new patch into $\mathbb{C}_k$, overlap may occur. So, we sort all the patches belonging to different concepts in ascending order by their importance. Then, we randomly select a single segment from each concept in turn, and stitch them into $\mathbb{C}_k$. In this way, we make important patches as visible as possible. This process will be repeated until the threshold σ is exceeded. In the Figure 3 we show several examples of Super Stitching with $\sigma = 0.75$. This algorithm ensures that the coverage of all samples obey a Gaussian distribution.

Finally, we sample mini-batches from Super Stitching discriminative image patches $\mathbb{C}_T^b = \{(\mathbb{C}_k, y_i)\}_{i=1}^b \sim \mathbb{C}_T$ and prune the network.

4 Experiments

We evaluate DOP and Super Stitching on benchmark ImageNet-1k [10] with ResNet-50 architecture [18]. Our results provide clear evidence that OPaI is data-dependent, and pruning based on discriminative image patches improves the performance of OPaI.

4.1 DOP: Pruning ResNet-50 with Varying Levels of Sparsity

We first compare the performances of DOP with the original OPaI methods at different sparsity levels. We prune an initialized ResNet-50 network by discriminative image patches. Those discriminative image patches are from a pre-trained ResNet-50 on ImageNet. To implement SNIP and GraSP, we adopt the open codes of Wang *et al*. [37] and Su *et al*. [32].

Based on the TCAV score, we select the Top-5 important concepts of each class. In the SNIP-based comparison experiments, we select 10,000 materials, i.e., 10 per class. Our sampling strategy follows [12], since [22] do not prune ResNet-50 on ImageNet. In the GraSP-based comparison experiments, we adopt 30,000 materials, i.e., 30 per class, closed to [37]'s implementation (250 batch size with 150 iterations). To sum up, for SNIP with DOP, we randomly select 2 discriminative image patches per concept. For GraSP with DOP, we select 6 discriminative image patches. These discriminative image patches use the same data augmentation as the training set, following [22, 37]. We train the pruned network 120 epochs. The batch size is 128. The optimization is SGD with a learning rate of 0.1 at the beginning, multiplying 0.1 at epochs $[30, 60, 90]$, a momentum of 0.9, and a weight decay of 0.0001.

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

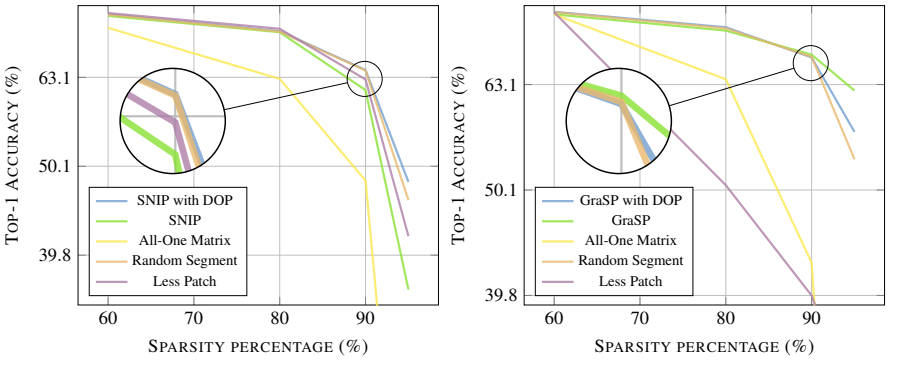

Figure 4: Ablation experiments at DOP. **Left:** SNIP with DOP. **Right:** GraSP with DOP. If the OPaI is data-independent, then the same results should be obtained in the ablation experiments. However, we still observe disparity in ablation experiments. Such trends indict that OPaI is impacted by limited validation information.

to about 3%. Moreover, based on GraSP pruning criterion, our DOP has a slight advantage at 90% sparsity and fails at 95% sparsity.

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 with a different pruning criterion 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-*

Method	Material Number	Material Type	Sparsity percentage			
			60%	80%	90%	95%
GraSP	30,000	image	73.87%	71.14%	67.07%	61.76%
GraSP with DOP (Ours)	30,000	discriminative patch	**74.19%**	**71.76%**	67.65%	60.02%
GraSP with Super Stitching (Ours)	30,000	stitching patch $\sigma = 0.5$	74.14%	71.57%	67.59%	**62.70%**
GraSP with Super Stitching (Ours)	10,000	stitching patch $\sigma = 0.75$	73.94%	71.65%	**68.02%**	62.06%
GraSP (Wang *et al.* [37])	37,500	image	74.02%	72.06%	68.14%	-
GraSP (Jorge *et al.* [9])	614,440	image	-	-	65.4%	46.2%
GraSP (Frankle *et al.* [12])	10,000	image	73.4%	71.0%	67%	-
GraSP (Hayou *et al.* [17])	-	image	-	-	66.41%	62.1%
FORCE (Jorge *et al.* [9])	614,440	image	-	-	64.9%	59.0%
Iter SNIP (Jorge *et al.* [9])	614,440	image	-	-	63.7%	54.7%
SynFlow (Hayou *et al.* [17])	-	all-one matrix	-	-	66.2%	62.05%
SBP-SR (Hayou *et al.* [17])	-	image	-	-	67.02%	62.66%
ProsPr (Alizadeh *et al.* [2])	-	image	-	-	66.86%	59.62%

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.

trix 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 valuablae suggestions for PaI improvement.

6 Acknowledgments

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

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[2] Milad Alizadeh, Shyam A. Tailor, Luisa M Zintgraf, Joost van Amersfoort, Sebastian Farquhar, Nicholas Donald Lane, and Yarin Gal. Prospect pruning: Finding trainable weights at initialization using meta-gradients. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AIgn9uwfcD1.

[3] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020.

[4] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16306–16316, 2021.

[5] Tianyi Chen, Bo Ji, Tianyu DING, Biyi Fang, Guanyi Wang, Zhihui Zhu, Luming Liang, Yixin Shi, Sheng Yi, and Xiao Tu. Only train once: A one-shot neural network training and pruning framework. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=p5rMPjrcCZq.

[6] Jian Cheng, Pei-song Wang, Gang Li, Qing-hao Hu, and Han-qing Lu. Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology & Electronic Engineering*, 19(1):64–77, 2018. ISSN 2095-9184. doi: 10.1631/fitee.1700789.

[7] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136, 2018. ISSN 1053-5888. doi: 10.1109/msp.2017.2765695.

[8] Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro*, 41(2):29–35, 2020. ISSN 0272-1732. doi: 10.1109/mm.2021.3061394.

[9] Pau de Jorge, Amartya Sanyal, Harkirat Behl, Philip Torr, Grégory Rogez, and Puneet K. Dokania. Progressive skeletonization: Trimming more fat from a network at initialization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=9GsFOUyUPi.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, page 248–255, 2009. ISSN 1063-6919. doi: 10.1109/cvpr.2009.5206848.

[11] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

[12] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ig-VyQc-MLK.

[13] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv*, 2019.

[14] Trevor Gale, Matei Zaharia, Cliff Young, and Erich Elsen. Sparse gpu kernels for deep learning. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press, 2020. ISBN 9781728199986.

[15] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

[17] Soufiane Hayou, Jean-Francois Ton, Arnaud Doucet, and Yee Whye Teh. Robust pruning at initialization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=vXj_ucZQ4hA.

[18] 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.

[19] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.*, 22(241):1–124, 2021.

[20] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[21] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

[22] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. SNIP: SINGLE-SHOT NETWORK PRUNING BASED ON CONNECTION SENSITIVITY. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1VZqjAcYX.

[23] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJlnB3C5Ym.

[24] Ekdeep Singh Lubana and Robert Dick. A gradient flow framework for analyzing network pruning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rumv7QmLUue.

[25] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6682–6691. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/malach20a.html.

[26] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJGCiw5gl.

[27] Ari Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32, 2019.

[28] Michael C. Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In *Advances in neural information processing systems*, pages 107–115, 1989.

[29] Ben Mussay, Margarita Osadchy, Vladimir Braverman, Samson Zhou, and Dan Feldman. Data-independent neural pruning via coresets. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gmHaEKwB.

[30] Wimmer Paul, Mehnert Jens, and Condurache Alexandru. Cops: Controlled pruning before training starts. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[31] R. Reed. Pruning algorithms-a survey. *IEEE Transactions on Neural Networks*, 4(5): 740–747, 1993. ISSN 1045-9227. doi: 10.1109/72.248452.

[32] Jingtong Su, Yihang Chen, Tianle Cai, Tianhao Wu, Ruiqi Gao, Liwei Wang, and Jason D Lee. Sanity-checking pruning methods: Random tickets can win the jackpot. *Advances in Neural Information Processing Systems*, 33:20390–20401, 2020.

[33] Arvind Subramaniam and Avinash Sharma. N2nskip: Learning highly sparse networks using neuron-to-neuron skip connections. In *BMVC*, 2020.

[34] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12): 2295–2329, 2017. ISSN 0018-9219. doi: 10.1109/jproc.2017.2761740.

[35] Hidenori Tanaka, Daniel Kunin, Daniel L Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in Neural Information Processing Systems*, 33:6377–6389, 2020.

[36] Artem Vysogorets and Julia Kempe. Connectivity matters: Neural network pruning through the lens of effective sparsity. *arXiv*, 2021.

[37] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgsACVKPH.

[38] Huan Wang, Can Qin, Yue Bai, Yulun Zhang, and Yun Fu. Recent advances on neural network pruning at initialization. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5638–5645. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/786. URL https://doi.org/10.24963/ijcai.2022/786. Survey Track.

[39] Paul Wimmer, Jens Mehnert, and Alexandru Paul Condurache. Dimensionality reduced training by pruning and freezing parts of a deep neural network, a survey. *arXiv*, 2022.

[40] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20554–20565. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ecb287ff763c169694f682af52c1f309-Paper.pdf.

[41] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G. Baraniuk, Zhangyang Wang, and Yingyan Lin. Drawing early-bird tickets: Toward more efficient training of deep networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxsrgStvr.

[42] Shunshi Zhang and Bradly C. Stadie. One-shot pruning of recurrent neural networks by jacobian spectrum evaluation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1e9GCNKvH.

[43] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. *Advances in neural information processing systems*, 32, 2019.