
Sanity-Checking Pruning Methods: Random Tickets can Win the Jackpot

Jingtong Su^{1,*} Yihang Chen^{2,*} Tianle Cai^{3,4,*}

Tianhao Wu² Ruiqi Gao^{3,4} Liwei Wang⁵ Jason D. Lee³

¹Yuanpei College, Peking University

²School of Mathematical Sciences, Peking University

³Department of Electrical Engineering, Princeton University

⁴Zhongguancun Haihua Institute for Frontier Information Technology

⁵Key Laboratory of Machine Perception, MOE, School of EECS, Peking University

Abstract

Network pruning is a method for reducing test-time computational resource requirements with minimal performance degradation. Conventional wisdom of pruning algorithms suggests that: (1) Pruning methods exploit information from training data to find good subnetworks; (2) The architecture of the pruned network is crucial for good performance. In this paper, we conduct sanity checks for the above beliefs on several recent unstructured pruning methods and surprisingly find that: (1) A set of methods which aims to find good subnetworks of the randomly-initialized network (which we call “initial tickets”), hardly exploits any information from the training data; (2) For the pruned networks obtained by these methods, randomly changing the preserved weights in each layer, while keeping the total number of preserved weights unchanged per layer, does not affect the final performance. These findings inspire us to choose a series of simple *data-independent* prune ratios for each layer, and randomly prune each layer accordingly to get a subnetwork (which we call “random tickets”). Experimental results show that our zero-shot random tickets outperform or attain a similar performance compared to existing “initial tickets”. In addition, we identify one existing pruning method that passes our sanity checks. We hybridize the ratios in our random ticket with this method and propose a new method called “hybrid tickets”, which achieves further improvement.²

1 Introduction

Deep neural networks have achieved great success in the overparameterized regime [44, 35, 8, 7, 2]. However, overparameterization also leads to excessive computational and memory requirements. To mitigate this, network pruning [33, 22, 15, 14, 6] has been proposed as an effective technique to reduce the resource requirements with minimal performance degradation.

One typical line of pruning methods, exemplified by retraining [14], can be described as follows: First, find a subnetwork of original network using pruning methods (which we call the “pruning step”), and then, retrain this subnetwork to obtain the final pruned network (which we call the “retraining step”). (Figure 1). Different pruning methods are mainly distinguished by *different criterion for finding the subnetwork, different weights to assign to the subnetwork and retraining scheme*.

*Equal Contribution, reverse alphabetical order.

²Our code is publicly available at <https://github.com/JingtongSu/sanity-checking-pruning>.

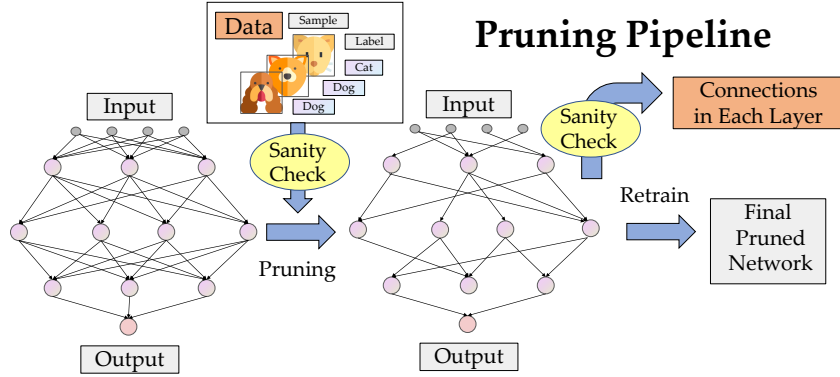


Figure 1: **A typical pipeline of obtaining the final pruned network.** The first step is a (usually) data-dependent pruning step that finds a subnetwork of the original network. The second step is a retraining step that trains the subnetwork to achieve minimal performance drop compared to the full network. Our sanity checks are applied to the data used in pruning step and the layerwise structure, i.e., connections in each layer, of the subnetwork obtained in the pruning step.

Generally, there are two common beliefs behind this pruning procedure. First, it is believed that pruning methods *exploit the information from training data in the pruning step* to find good subnetworks. Therefore, there are only a few *data-independent* pruning methods [34], and it is not clear whether these methods can get similar (or even better) performance than the popular data-dependent pruning methods. Second, it is believed that *the architecture of the pruned network* is crucial for obtaining the final pruned network with good performance [14]. As a result, most of the papers in the pruning literature usually use “random pruning”, i.e., randomly selecting some number of weights/neurons to be pruned according to the pruning ratio, as a weak baseline to show that the pruning methods outperform the trivial random architecture. However, we put into question whether these two beliefs really hold. To evaluate them, we propose two sets of sanity check approaches which examine whether the *data used in the pruning step* and *the structure of pruned network* are essential for the final performance of pruned networks for *unstructured* pruning methods (See Section 2.2)³.

To evaluate the dependency on the data, we propose to corrupt the data *in the pruning step* while using the original data in the retraining step, and see if the subnetworks pruned using the corrupted data can achieve similar performance as those pruned using the true data. To test the importance of the architecture of the pruned subnetwork, we introduce a novel attack on the network architecture which we call *layerwise rearrange*. This attack rearranges all the connections of neurons *layerwise* and thus totally destroys the structure obtained in pruning step. It is worth noting that, after applying layerwise rearranging, the number of preserved weights in *each layer* stays unchanged. In contrast, if we rearrange the connections in the entire subnetwork, the resulting subnetwork is basically the same as that obtained by random pruning. We refer the readers to Figure 1 for an illustration of where we deploy our sanity checks.

We then apply our sanity checks on the pruning methods in four recent papers from ICLR 2019 and 2020 [23, 9, 41, 37]. We first classify the subnetworks found by these methods into “initial tickets”, i.e., the weights before retraining are set to be the weights at initialization (this concept is the same as “winning tickets” in [9]), and “partially-trained tickets”, i.e., the weights before retraining are set to be the weights from the middle of pretraining process⁴ (See Section 4.1 for the formal definition and Figure 2 for an illustration). The effectiveness of “initial tickets” is highlighted by “the lottery ticket hypothesis (LTH)” [9], and the methods on “partially-trained tickets” make some modifications on the original method in [9].

We show that both of the beliefs mentioned above are not necessarily true for “initial tickets”. Specifically, the subnetworks can be trained to achieve the same performance as initial tickets even when they are obtained by using corrupted data in the pruning step or rearranging the initial tickets layerwise. This surprising finding indicates that true data may not be necessary for these methods and the architecture of the pruned network may only have limited impact on the final performance.

³Due to space limitation, we review some related works in Appendix F.

⁴When clear from the context, we may use “tickets” to refer to the subnetworks obtained by certain methods or directly refer to the methods.

Inspired by the results of sanity checks, we propose to choose a series of simple *data-independent* pruning ratios for each layer, and randomly prune each layer accordingly to obtain the subnetworks (which we call “random tickets”) at initialization. This zero-shot method pushes beyond the one-shot pruning-at-initialization methods [23, 41]. Concretely, our random tickets are drawn without any pretraining optimization or any data-dependent pruning criterion, and thus we save the computational cost of pruning. Experimental results show that our zero-shot random tickets outperforms or attains similar performance compared to all existing “initial tickets”. Though we only try pruning ratios in a few simple forms without much tuning, the success of random tickets further suggests that our layerwise ratios can *serve as a compact search space for neural architecture search* [47, 48, 36, 25, 4].

In addition, we also find a very recent pruning method in ICLR 2020 [37] for obtaining “partially-trained tickets” can pass our sanity checks. We then hybridize the insights for designing random tickets with partially-trained tickets and propose a new method called “hybrid tickets”. Experimental results show that the hybrid tickets can achieve further improve the partially-trained tickets.

In summary, our results advocate for a re-evaluation of existing pruning methods. Our sanity checks also inspire us to design data-independent and more efficient pruning algorithms.

2 Preliminaries

2.1 Notations

We use $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ to denote the training dataset where \mathbf{x}_i represents the sample and y_i represents the label. We consider a neural network with L layers for classification task on \mathcal{D} . We use vector $\mathbf{w}_l \in \mathbb{R}^{m_l}$ to denote the weights in layer l where m_l is the number of weights in layer l . Then the network can be represented as $f(\mathbf{w}_1, \dots, \mathbf{w}_L; \mathbf{x})$ where \mathbf{x} is the input to the network. The goal of classification is minimizing an error function ℓ over the outputs of network and target labels, i.e., $\sum_{i=1}^n \ell(f(\mathbf{w}_1, \dots, \mathbf{w}_L; \mathbf{x}_i), y_i)$, and the performance is measured on a test set.

2.2 Unstructured Pruning (Pruning Individual Weights)

One major branch of network pruning methods is unstructured pruning, and it dates back to Optimal Brain Damage [22] and Optimal Brain Surgeon [16], which prune weights based on the Hessian of the loss function. Recently, [13] proposes to prune network weights with small magnitude and incorporates the pruning method to “Deep Compression” pipeline to obtain efficient models. Several other criterions and settings are proposed for unstructured pruning [6, 39, 31]. Apart from unstructured pruning, there are structured pruning methods [20, 29, 26, 21, 24, 42, 1] which prune at the levels of convolution channels or higher granularity. In this paper, we mainly focus on *unstructured pruning*.

Unstructured pruning can be viewed as a process to find a mask on the weights which determines the weights to be preserved. Formally, we give the following definitions for a clear description:

Mask. We use $\mathbf{c}_l \in \{0, 1\}^{m_l}$ to denote a mask on \mathbf{w}_l . Then the pruned network is given by $f(\mathbf{c}_1 \odot \mathbf{w}_1, \dots, \mathbf{c}_L \odot \mathbf{w}_L; \mathbf{x})$.

Sparsity and keep-ratios. The sparsity p of the pruned network calculates as $p = 1 - \frac{\sum_{l=1}^L \|\mathbf{c}_l\|_0}{\sum_{l=1}^L m_l}$, where $\|\cdot\|_0$ denotes the number of nonzero elements. We further denote p_l as the “keep-ratio” of layer l calculated as $\frac{\|\mathbf{c}_l\|_0}{m_l}$.

2.3 Pruning Pipeline and Tickets

In Figure 1, we show a typical pipeline of network pruning in which we first prune the full network and obtain a subnetwork, then retrain the pruned subnetwork. The processes of pruning and retraining can be repeated iteratively [13], which requires much more time. In this paper, we focus on *one shot pruning* [14], which does not iterate the pipeline.

We evaluate several recently proposed pruning methods which can be classified as follows using the names “tickets” adopted from “Lottery Ticket Hypothesis” [9]:

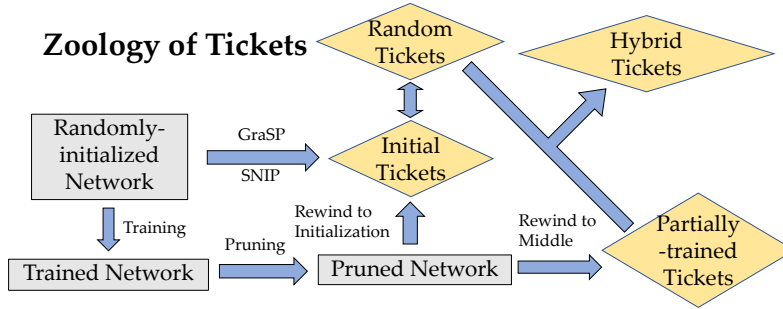


Figure 2: **Zoology of the tickets considered in this paper.** Initial tickets and partially trained tickets are introduced in Section 2.3, random tickets are proposed in Section 4.2, and hybrid tickets are proposed in Section 5.2.

- **Initial tickets:** This kind of methods aim to find a subnetwork of the randomly-initialized network which we call “*initial tickets*” that can be trained to reach similar test performance of the original network. Initial tickets include pruning at initialization [23, 41] and the method proposed in [9].
- **Partially-trained tickets:** Different from initial tickets, partially-trained tickets are constructed by first training the network, pruning it, and then *rewinding* the weights to some middle stage [37]. Initial tickets found by the method in [9] can be viewed as rewinding to epoch 0.

We illustrate these two kinds of tickets in Figure 2 together with our proposed new tickets that will be introduced in later sections.

3 Sanity Check Approaches

In this section, we propose sanity check approaches for testing if some common beliefs of pruning methods truly hold in several recently proposed methods.

3.1 Checking the Dependency on Data

Popular pruning methods usually utilize data in the pruning step to find a good pruned subnetwork that can be trained to perform as well as the full network on the same data (Figure 1).

However, there is no clear evidence showing that the data used in pruning step is important. Furthermore, from the few existing data-independent pruning methods [34], it is also hard to judge whether data-independent methods have the potential to attain similar performance as popular data-dependent pruning methods. Therefore, we ask the following question:

Is data information useful for finding good subnetworks in the pruning step?

To answer this question, we introduce two operations, using random labels and random pixels, to corrupt the data [44]. Note that different from [44] in which these two operations are used to demonstrate whether overparameterized network can converge on bad data, we apply these operations on the data used in the *pruning step* to check *if corrupted data still leads to good subnetworks*. In addition to these strong corruptions, we also provide a weaker check that only reduces the number of data but not corrupts the data. Concretely, we define the two operations as follows:

- **Random labels:** In the pruning step, we replace the dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ with $\hat{\mathcal{D}} = \{\mathbf{x}_i, \hat{y}_i\}_{i=1}^n$ where \hat{y}_i is randomly generated from a uniform distribution over each class.
- **Random pixels:** In the pruning step, we replace the dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ with $\hat{\mathcal{D}} = \{\hat{\mathbf{x}}_i, y_i\}_{i=1}^n$ where $\hat{\mathbf{x}}_i$ is the randomly shuffled⁵ \mathbf{x}_i and shuffling of each individual sample is independent.
- **Half dataset:** We reduce the number of data used in the pruning step. For convenience, in this setting, we simply take half of the data in the dataset for the pruning step.

⁵By shuffling, we mean generating a random permutation and then reordering the vector by this permutation.

3.2 Checking the Importance of the Pruned Network’s Architecture

In unstructured pruning, the architecture, or the connections of neurons, of the pruned subnetwork has long been considered as the key to make the pruned network able to achieve comparable performance to the original network [14]. As a result, random pruning, i.e., randomly selecting $(1 - p) \cdot \sum_{l=1}^L m_l$ weights to preserve according to a given sparsity level p , is widely used as a baseline method to show the effectiveness of pruning methods with more carefully designed criterion.

However, we cast doubt on whether these individual connections in the pruned subnetwork is crucial, or if there any intermediate factors that determine the performance of pruned network. Concretely, we answer the following question:

To what extent does the architecture of the pruned subnetwork affect the final performance?

Towards answering this question, we propose a novel layerwise rearranging strategy, which keeps the number of preserved weights in each layer but *completely destroys the network architecture (connections) found by the pruning methods* for each individual layer. As an additional reference, we also introduce a much weaker layerwise weight shuffling operation, which only shuffle the unmasked weights but keep the connections.

- **Layerwise rearrange:** After obtaining the pruned network, i.e., $f(\mathbf{c}_1 \odot \mathbf{w}_1, \dots, \mathbf{c}_L \odot \mathbf{w}_L; \mathbf{x})$, we randomly rearrange the mask \mathbf{c}_l of *each layer, independently*, into $\hat{\mathbf{c}}_l$, and then train on the network with the rearranged masks $f(\hat{\mathbf{c}}_1 \odot \mathbf{w}_1, \dots, \hat{\mathbf{c}}_L \odot \mathbf{w}_L; \mathbf{x})$. We give an illustration of this operation in Appendix A.
- **Layerwise weights shuffling:** Layerwise weights does not change the masks but shuffles the *unmasked weights* in each layer independently.

With the approaches above, we are ready to perform sanity-check on existing pruning methods.

4 Case Study: Initial Tickets

In this section, we will apply our sanity checks to “initial tickets” defined in Section 2.3. Surprisingly, our results suggest that the final performance of the retrained “initial tickets” does not drop when using corrupted data including random labels and random pixels in the pruning step. Moreover, the layerwise rearrangement does not affect the final performance of “initial tickets” either. This finding further inspires us to design a *zero-shot* pruning method, dubbed “random tickets”.

4.1 Initial Tickets Fail to Pass Sanity Checks

We conduct experiments on three recently proposed pruning methods that can be classified as “initial tickets”. We first briefly describe these methods:

- **SNIP [23] and GraSP [41]:** These two methods prune the network at initialization by finding the mask using different criterion. Specifically, SNIP leverages the notion of connection sensitivity as a measure of importance of the weights to the loss, and then remove unimportant connections. GraSP aims to preserve the gradient flow through the network by maximizing the gradient of the pruned subnetwork. The merit of these methods is that they can save the time of training by pruning at initialization compared to methods that require pretraining[13].
- **Lottery Tickets (LT) [9]:** The pruning procedure of Lottery Tickets follows a standard three-step-pipeline: training a full network firstly, using magnitude-based pruning [13] to obtain the ticket’s architecture, and *resetting the weights to the initialization* to obtain the initial ticket.⁶

After obtaining the initial tickets, we train them by standard methods and evaluate the test accuracy of the final models. We follow a standard training procedure as [23, 41, 17] *throughout the entire paper* on ResNet [19] and VGG [38]. The detailed setting can be found in Appendix B.

⁶In the main body of our paper, we focus on one-shot pruning methods since: 1. SNIP and GraSP are both pruning methods without significant training expenses, and 2. iterative pruning cannot consistently outperform one shot pruning, and even when iterative pruning is better, the gap is small, as shown in [27] and [37]. The discussions and experiments about Iterative Magnitude Pruning are deferred to Appendix D, from which we conclude the iterative procedure hardly helps initial tickets.

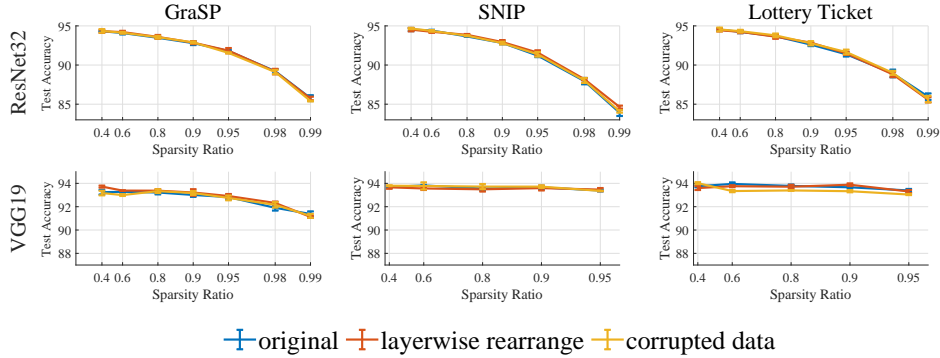


Figure 3: Sanity checks of the initial tickets of ResNet32 and VGG19 on CIFAR10.

To sanity-check these three methods and answer the motivating questions proposed in Section 3, we apply the two strongest checks in our toolbox.

We first check if the pruning procedure can utilize information on the training data. Towards this goal, we use *the combination of random label and random pixel corruptions* in Section 3 in the pruning step. We then train these tickets on the *original CIFAR-10 dataset*. Surprisingly, all initial tickets generated by the corrupted dataset behave as well as the originally picked tickets (Figure 3).⁷ That is to say, *even corrupted data can be used to find good initial tickets*.

Given the weak dependency on the data used in the pruning step, we then investigate whether the architecture (connections between neurons) of the initial tickets matter. We then apply our *layerwise rearrange* attack, which totally removes the dependency of the learned connections in each layer on the initial tickets. We find that the performance of initial tickets does not drop when applying the layerwise rearrange attack (Figure 3). This contradicts the belief that the pruned structure is essential for obtaining a subnetwork with good performance.⁸ We also conduct experiments on the CIFAR-100 dataset. The results are deferred to Appendix E.

4.2 Random Tickets Win the Jackpot

The results of sanity check on initial tickets suggest that:

- Data may not be important for finding good initial tickets.
- The connections of neurons in individual layers in initial tickets can be totally rearranged without any performance drop, indicating that only the number of remained connections or keep-ratio (recall the definition in Section 2.2) in each layer matters.

Based on these observations, we only need to find a series of good *keep-ratios* and randomly prune the network accordingly. We call the subnetworks obtained by this rule “random tickets”. As illustrated in the Ticket Zoo in Figure 2, random tickets can be viewed as a special kind of initial tickets since they use the weights at initialization. Meanwhile, applying layerwise rearrangement to initial tickets obtained by existing pruning methods can transform these initial tickets to random tickets with the layerwise keep-ratios determined by the initial tickets.

Though we can obtain the keep-ratios from existing initial tickets in the above way, we are more willing to push even further to directly design a series of *simpler but effective* keep-ratios. Towards this goal, we first investigate the ratios of the initial tickets obtained using GraSP, SNIP and Lottery Tickets on both VGG and ResNet architectures. From these initial tickets, we extract the following principle of keep-ratios that leads to good final performance:

⁷We omit results with significantly high variances or are trivial, i.e. have a 10% accuracy. This kind of omitting is used on figures shown throughout our paper.

⁸In the paper [32]’s Appendix A2 part, the authors report applying layerwise rearrange can hurt the performance when the pruning ratio is high. However, in their setting they use a “late resetting” trick which is kind of similar to the Learning Rate Rewinding as we will discuss in the next section of our paper. So it is not surprising to see that Figure A1 in [32] show evidence that the method may pass the layerwise rearrange check.

Table 1: Test accuracy of pruned VGG19 and ResNet32 on CIFAR-10 and CIFAR-100 datasets. In the full paper, the bold number indicates the average accuracy is within the best confidence interval.

Dataset	CIFAR-10			CIFAR-100		
	90%	95%	98%	90%	95%	98%
Pruning ratio						
VGG19 (Full Network)	93.70	-	-	72.60	-	-
LT [9]	93.66±0.08	93.39±0.13	10.00±0.00	72.58±0.27	70.47±0.19	1.00±0.00
SNIP [23]	93.65±0.14	93.39±0.08	81.37±18.17	72.83±0.16	71.81±0.11	10.83±6.74
GraSP [41]	93.01±0.16	92.82±0.20	91.90±0.23	71.07±0.31	70.14±0.21	68.34±0.20
Random Tickets (Ours)	93.77±0.10	93.42±0.22	92.45±0.22	72.55±0.14	71.37±0.09	68.98±0.34
ResNet32 (Full Network)	94.62	-	-	74.57	-	-
LT [9]	92.61±0.19	91.37±0.28	88.92±0.49	69.63±0.26	66.48±0.13	60.22±0.60
SNIP [23]	92.81±0.17	91.20±0.19	87.94±0.40	69.97±0.17	64.81±0.44	47.97±0.82
GraSP [41]	92.79±0.24	91.80±0.11	89.21±0.26	70.12±0.15	67.05±0.39	59.25±0.33
Random Tickets (Ours)	92.97±0.05	91.60±0.26	89.10±0.33	69.70±0.48	66.82±0.12	60.11±0.16

Table 2: Test accuracy of pruned VGG19 and ResNet32 on Tiny-Imagenet dataset.

Network	VGG19			ResNet32		
	90%	95%	98%	90%	95%	98%
Pruning ratio						
VGG19/ResNet32 (Full Network)	61.57	-	-	62.92	-	-
SNIP [23]	61.65±0.17	58.72±0.20	9.37±13.67	54.53±0.34	47.68±0.20	32.09±1.87
GraSP [41]	60.71±0.25	59.11±0.03	57.08±0.16	56.23±0.09	51.36±0.37	39.43±1.52
Random Tickets (Ours)	60.94±0.38	59.48±0.43	55.87±0.16	55.26±0.22	51.41±0.38	43.93±0.18

Keep-ratios decrease with depth. For all these methods on both VGG and ResNet, we observe that the keep-ratios of initial tickets have the common behavior that the keep-ratio of a deeper layer is lower, i.e., the keep-ratios decay as a function of depth. Moreover, the keep-ratios of VGG decays faster than ResNet for all three algorithms.⁹

We also observe that apart from the overall trend of declining in layerwise keep-ratios, the keep-ratios in some special layers behave differently. For example, the keep-ratio of the downsampling layers in ResNet are significantly larger than their neighboring layers.

Smart ratios. Based on these observations, we propose *smart-ratios*, a series of keep-ratios that take the form of simple decreasing functions. Specifically, for an L -layer ResNet, we set the keep-ratio of the linear layer to 0.3 for any sparsity p , and let the keep-ratio of a given layer l , i.e., p_l (Section 2.2), to be proportional to $(L - l + 1)^2 + (L - l + 1)$. For VGG, we divide $(L - l + 1)^2 + (L - l + 1)$ by l^2 for a faster decrease. The details can be found in Appendix C.

We note that we choose our ratio just by trying a few decreasing functions *without careful tuning*, but already get good performance. We also conduct ablation studies on ascending keep-ratios and different rates of decaying keep-ratios. Experimental results suggest that the descending order and, upon that, the descent rate of keep-ratios is crucial for the final performance of random tickets. This partially explains why the baseline of randomly pruning over the whole network other than using a tailored layerwise ratio usually performs poorly. Due to space limitation, we put the detailed ablation study in the Appendix. Moreover, we point out that as good keep-ratios produce good subnetworks, one may use keep-ratios as a compact search space for Neural Architecture Search, which may be an interesting future direction.

The test accuracy on CIFAR-10/CIFAR-100 and Tiny-Imagenet datasets are shown in Table 1 and 2. As reported, our random tickets with smart ratios surpass several carefully designed methods on several benchmarks and sparsity. And we note that when the sparsity is very high, some layers might retain only few weights when using SNIP or LT, so that the test accuracy drops significantly [17]. Recently [40] also observe this phenomenon and propose a pruning method without data and be able to avoid the whole-layer collapse.

⁹Similar trends are also reported in Figure 3 of [41], Figure 1b of [32] and Figure 11 of [11].

5 Case Study: Partially-trained Tickets

In this section, we study pruning methods in a very recent ICLR 2020 paper [37], which is classified as partially-trained tickets (Section 2.3). These methods can pass our sanity checks on data dependency and architecture. We then combine our insights of random tickets with these partially-trained tickets and propose a method called “hybrid tickets” (Figure 2) which further improves upon [37].

5.1 Partially-trained Tickets Pass Sanity Checks

Different from initial tickets, [37] propose a *learning rate rewinding* method that improves beyond *weights rewinding* [10]. The methods used in [37] include weights rewinding and learning rate rewinding. Both methods first fully train the unpruned network (which we call pretraining) and generate the mask by magnitude pruning [14]. Then the two methods for retraining are

- **Weights rewinding:** First rewind the unmasked weights of the pruned subnetwork to their values at some middle epoch of the pretraining step and then retrain the subnetwork using the same training schedule as the pretraining step at that epoch.
- **Learning rate rewinding:** Retrain the masked network with its learned weights using the original learning rate schedule.

These two methods can be classified as “partially-trained tickets” as they use weights from partially-trained (or actually for the latter, fully-trained) network. And as pointed out in [37], learning rate rewinding usually surpasses weights rewinding, so we mostly focus on learning rate rewinding.

Similar to Section 4.1, we conduct sanity checks on learning rate rewinding method. We show the results of two weaker attacks, layerwise weight shuffling and half dataset (see Section 3) since if learning rate rewinding can pass these checks, i.e., its performance degrades under these weaker attacks, it will naturally pass the previous sanity check with strong attacks used in Section 4.1.

As shown in Table 3, both modifications do impair the performance. We note that the degradation of test accuracy on VGG is not such significant as that on ResNet, which is understandable since the number of parameters of VGG is huge and the gaps between different methods are small. When using layerwise weight shuffling, the test accuracy is reduced to being comparable to the case when the weights are rewound to the initialization (LT). When using half dataset in the pruning step, the test accuracy also drops compared to learning rate rewinding while it is higher than LT (Section 4.1). We conclude from these observations that the partially-trained tickets can pass our sanity checks.

5.2 Hybrid Tickets

As the results of the sanity checks on learning rate rewinding suggest, this method truly encodes information of data into the weights, and the architectures of the pruned subnetworks cannot be randomly changed without performance drop.

On another side, the success of random tickets shed light on using smarter keep-ratios to attain better performance. Therefore, we combine our “smart ratios” (Section 4.2) and the idea of learning rate rewinding. Concretely, we first pretrain the network as in learning rate rewinding, and then, while pruning the pretrained network, we use magnitude-based pruning *layerwise* and the keep-ratios are determined by the *smart-ratios*. In other words, for a given layer l , we only keep those weights whose magnitude are in the largest p_l portion in layer l , where p_l is the smart-ratio of layer l . And finally, the pruned network is retrained with a full learning-rate schedule as learning rate rewinding does.

We call the pruned subnetworks obtained by this hybrid method “hybrid tickets” (See Figure 2). The test accuracy on CIFAR-10 and CIFAR-100 is shown in Table 4. As reported, our hybrid tickets gain further improvement upon learning rate rewinding, especially at high sparsity for both datasets and network architectures. At the same time, hybrid tickets can avoid pruning the whole layer [17] which may happen for learning rate rewinding.

6 Conclusion and Discussion

In this paper, we propose several sanity check methods (Section 3) on unstructured pruning methods that test whether the data used in the pruning step and whether the architecture of the pruned

Table 3: Sanity-check on partially-trained tickets on CIFAR-10 dataset.

Network	VGG19			ResNet32		
	90%	95%	98%	90%	95%	98%
Pruning ratio						
VGG19/ResNet32 (Full Network)	93.70	-	-	94.62	-	-
LT [9]	93.66±0.08	93.39±0.13	10.00±0.00	92.61±0.19	91.37±0.28	88.92±0.49
Shuffle Weights	93.54±0.04	93.33±0.10	10.00±0.00	92.38±0.36	91.29±0.28	88.52±0.47
Half Dataset	93.79±0.14	93.53±0.13	10.00±0.00	93.01±0.18	92.03±0.21	89.95±0.08
Learning Rate Rewinding [37]	94.14±0.17	93.99±0.15	10.00±0.00	94.14±0.10	93.02±0.28	90.83±0.22

Table 4: Test accuracy of partially-trained tickets and our hybrid tickets of VGG19 and ResNet32 on CIFAR-10 and CIFAR-100 datasets.

Dataset	CIFAR-10			CIFAR-100		
	90%	95%	98%	90%	95%	98%
Pruning ratio						
VGG19 (Full Network)	93.70	-	-	72.60	-	-
OBD [22]	93.74	93.58	93.49	73.83	71.98	67.79
Learning Rate Rewinding [37]	94.14±0.17	93.99±0.15	10.00±0.00	73.73±0.18	72.39±0.40	1.00±0.00
Hybrid Tickets (Ours)	94.00±0.12	93.83±0.10	93.52±0.28	73.53±0.20	73.10±0.11	71.61±0.46
ResNet32 (Full Network)	94.62	-	-	74.57	-	-
OBD [22]	94.17	93.29	90.31	71.96	68.73	60.65
Learning Rate Rewinding [37]	94.14±0.10	93.02±0.28	90.83±0.22	72.41±0.49	67.22±3.42	59.22±1.15
Hybrid Tickets (Ours)	93.98±0.15	92.96±0.13	90.85±0.06	71.47±0.26	69.28±0.40	63.44±0.34

subnetwork are essential for the final performance. We find that one kind of pruning method, classified as “initial tickets” (Section 2.3) hardly exploit any information from data, because randomly changing the preserved weights of the subnetwork obtained by these methods layerwise does not affect the final performance. These findings inspire us to design a *zero-shot data-independent* pruning method called “random tickets” which outperforms or attains similar performance compared to initial tickets. We also identify one existing pruning method that passes our sanity checks, and hybridize the random tickets with this method to propose a new method called “hybrid tickets”, which achieves further improvement. Our findings bring new insights in rethinking the key factors on the success of pruning algorithms.

Besides, a concurrent and independent work [11] got a similar conclusion to our layerwise rearrange sanity check. As a complementary to our results, the results on ImageNet in [11] shows our finding can be generalized to large-scale datasets. Both our works advocate to rigorously sanity-check future pruning methods and take a closer look at opportunities to prune early in training.

Acknowledge

We thank Jonathan Frankle, Mingjie Sun and Guodong Zhang for discussions on pruning literature. YC and TW are partially supported by the elite undergraduate training program of School of Mathematical Sciences in Peking University. TC and RG are supported in part by the Zhongguancun Haihua Institute for Frontier Information Technology. LW was supported by National Key R&D Program of China (2018YFB1402600), Key-Area Research and Development Program of Guangdong Province (No. 2019B121204008) and Beijing Academy of Artificial Intelligence. JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0303, the Sloan Research Fellowship, and NSF CCF 2002272.

Broader Impact

We investigate into neural network pruning methods, which is an important way to reduce computational resource required in modern deep learning. It can potentially help enable faster inference, conserve less energy, and make AI widely deployable and assessable in mobile and embedded systems. Our experiments can also provide more insight for neural networks in deep learning algorithms, which could potentially lead to the development of better deep learning algorithms.

References

- [1] Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems*, pages 2270–2278, 2016.
- [2] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
- [3] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Gutttag. What is the state of neural network pruning? *arXiv preprint arXiv:2003.03033*, 2020.
- [4] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations*, 2019.
- [5] Miguel A Carreira-Perpinán and Yerlan Idelbayev. “learning-compression” algorithms for neural net pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8532–8541, 2018.
- [6] Xin Dong, Shangyu Chen, and Sinno Jialin Pan. Learning to prune deep neural networks via layer-wise optimal brain surgeon. pages 4857–4867, 2017.
- [7] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.
- [8] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.
- [9] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.
- [10] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. *arXiv preprint arXiv:1912.05671*, 2019.
- [11] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? *arXiv preprint arXiv:2009.08576*, 2020.

- [12] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.
- [13] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
- [14] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.
- [15] Babak Hassibi, David G. Stork, and Gregory Wolff. Optimal brain surgeon and general network pruning. pages 293 – 299 vol.1, 02 1993.
- [16] Babak Hassibi and David G Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pages 164–171, 1993.
- [17] Soufiane Hayou, Jean-Francois Ton, Arnaud Doucet, and Yee Whye Teh. Pruning untrained neural networks: Principles and analysis. *arXiv preprint arXiv:2002.08797*, 2020.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [19] 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.
- [20] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1389–1397, 2017.
- [21] Zehao Huang and Naiyan Wang. Data-driven sparse structure selection for deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 304–320, 2018.
- [22] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.
- [23] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019.
- [24] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.
- [25] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.
- [26] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2736–2744, 2017.
- [27] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.
- [28] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018.
- [29] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*, pages 5058–5066, 2017.
- [30] Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. *arXiv preprint arXiv:2002.00585*, 2020.

- [31] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference.
- [32] Ari Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. In *Advances in Neural Information Processing Systems*, pages 4933–4943, 2019.
- [33] 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.
- [34] Ben Mussay, Margarita Osadchy, Vladimir Braverman, Samson Zhou, and Dan Feldman. Data-independent neural pruning via coresets. In *International Conference on Learning Representations*, 2020.
- [35] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann Lecun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv: Learning*, 2018.
- [36] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pages 4095–4104, 2018.
- [37] Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. In *International Conference on Learning Representations*, 2020.
- [38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [39] Suraj Srinivas and R Venkatesh Babu. Generalized dropout. *arXiv preprint arXiv:1611.06791*, 2016.
- [40] Hidenori Tanaka, Daniel Kunin, Daniel LK Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *arXiv preprint arXiv:2006.05467*, 2020.
- [41] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020.
- [42] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *Advances in neural information processing systems*, pages 2074–2082, 2016.
- [43] 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.
- [44] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.
- [45] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, pages 3592–3602, 2019.
- [46] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.
- [47] Barret Zoph and Quoc Le. Neural architecture search with reinforcement learning. 2016.
- [48] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

A Illustration of Layerwise Rearrange

The operation layerwise arrangement is illustrated in Figure 4.

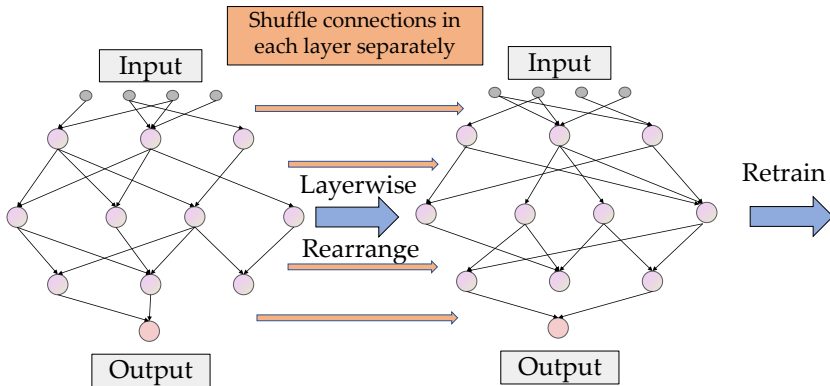


Figure 4: **Layerwise rearrange.** This operation keeps the number of preserved weights in each layer but totally destroys the network architecture (connections) found by the pruning methods for each individual layer.

B Experiment Settings

In all experiments throughout the paper, we use three benchmark image classification datasets, i.e., CIFAR-10, CIFAR-100 and Tiny-ImageNet¹⁰ with VGG [38] and ResNet [19]. The CIFAR-10 dataset consists of 50,000 32x32 color (three-channel) training examples and 10,000 test examples, and has 10 classes of labels. The CIFAR-100 dataset consists of 50,000 32x32 color training examples and 10,000 test examples. Different from CIFAR-10, CIFAR-100 has 100 classes of labels with 500/100 training/test examples for each class. The Tiny-Imagenet dataset consists of 100,000 64x64 color training examples and 10,000 test examples. It has 200 classes, and each class has 500 training examples with 50 validation examples.

Following the setting in [41, 17], we double the number of filters in each convolutional layer of all the ResNet architectures throughout the paper in order to make comparisons with these baseline algorithms. The pruned network is trained with Kaiming initialization [18], using SGD for 160 epochs for CIFAR-10/100, and 300 epochs for Tiny-ImageNet, with an initial learning rate of 0.1 and batch size 64. The learning rate is decayed by a factor of 0.1 at 1/2 and 3/4 of the total number of epochs, together with a weight decay factor of 1e-4. Moreover, we run each experiment for 3 trials and report the best test accuracy and the corresponding standard deviations.

Our code is based on the released code of [41, 27] (<https://github.com/alecwangcq/GraSP>, <https://github.com/Eric-mingjie/rethinking-network-pruning>).

C Details of Random Tickets and Ablation Studies

In this section, we first describe the detailed process of generating random tickets. Then we provide ablation studies on the performance of different smart ratios; and the performance of our proposed methods (random tickets, hybrid tickets) on different network architectures.

C.1 Generation of Random Tickets

We first note that in different pruning methods, the number of weights retained in linear layer is chosen differently. In SNIP [23] and GraSP [41], the linear layer is pruned together with other

¹⁰Can be downloaded from <http://cs231n.stanford.edu/tiny-imagenet-200.zip>

Table 5: Ablation study of different keep-ratios on CIFAR-10 dataset.

Network	VGG19			ResNet32		
	90%	95%	98%	90%	95%	98%
VGG19/ResNet32 (Full Network)	93.70	-	-	94.62	-	-
Random Tickets (Ours)	93.77±0.10	93.42±0.22	92.45±0.22	92.97±0.05	91.60±0.26	89.10±0.33
Balanced	91.30±0.22	89.79±0.22	86.31±0.32	91.78±0.17	90.06±0.16	86.52±0.36
Ascending	91.79±0.24	90.29±0.17	87.22±0.30	10.00±0.00	10.00±0.00	10.00±0.00
Linear Decay	93.83±0.23	93.34±0.12	92.32±0.14	92.68±0.03	91.60±0.23	88.54±0.14
Cubic Decay	93.90±0.15	93.63±0.29	92.66±0.28	92.74±0.13	91.56±0.23	89.14±0.16

convolutional layers, while in Lottery Tickets [9], the linear layer is fully preserved. For simplicity, we fix the keep-ratio of linear layer to be 30%¹¹ for all experiments on smart ratio.

The smart ratios for the convolutional layers are designed as follows. As mentioned in Section 4.2, for an L -layer ResNet¹², we set the keep-ratio of l -th layer to be proportional to $(L-l+1)^2 + (L-l+1)$. Concretely, this is achieved by following steps:

1. Calculate the number of weights to be retained as $(1-p) \sum_{i=1}^L m_i$.
2. Set the keep-ratio of layer l as $(L-l+1)^2 + (L-l+1)$.
3. Linearly scale the keep-ratio of each layer such that the final number of retained weights is equal to $(1-p) \sum_{i=1}^L m_i$.

For an L -layer VGG net, the process is similar except that we add an extra heterogeneous constant penalty for each layer to make the smart ratios decay faster. Specifically, we replace $(L-l+1)^2 + (L-l+1)$ with $\frac{(L-l+1)^2 + (L-l+1)}{l^2}$ in the second step.

We further note that, for several small sparsity parameters p (e.g., $p = 0.9$), linearly scaling the keep-ratio as in step 3 may result in a keep-ratio greater than 1. Under these circumstances, we simply set these layers' keep-ratios to be 1 and move the extra retained parameters to the deeper layer immediately behind them.

C.2 Ablation Studies on Smart Ratio

We conduct experiments on CIFAR-10 dataset to compare several types of keep-ratios including ascending, balanced, and descending.

1. **Ascending keep-ratio:** Reverse our smart ratio.
2. **Balanced keep-ratio:** Set the keep-ratio of each layer to be $1-p$, where p is the target sparsity.
3. **Linear decay:** Set the keep-ratio of l -th convolutional layer to be proportional to $L-l+1$.
4. **Cubic decay:** Set the keep-ratio of l -th convolutional layer to be proportional to $(L-l+1)^3$.

For all the ratios, the keep-ratio of linear layer is uniformly set to be 30%. The results can be found in Table 5. We observe that the balanced and ascending keep-ratios lead to much worse performance than the descending keep-ratios, while descending keep-ratios with different decaying speeds are comparative on CIFAR10. As mentioned in Section 4.2, we did not tune the ratio so much, and the results in Table 5 suggest that to get a better performance, one may use different keep-ratios for different sparsity to generate high-performance random tickets.

C.3 Ablation Studies on Different Architectures

We test our proposed random tickets and hybrid tickets on more architectures including ResNet20/56 and VGG11/16 on CIFAR-10/100 datasets. The comparison of our random tickets and those baseline methods with VGG11/16 and ResNet20/56 on CIFAR-10/100 can be found in Table 6, while the results of our hybrid tickets can be found in Table 7. These results show that our smart ratio can generalize at both ResNet/VGG architectures with different depths.

¹¹30% is similar to those keep-ratios of linear layers obtained by SNIP and GraSP, and we also find that the choice of keep-ratio of linear layer hardly affects the final performance.

¹² L denotes the total number of convolutional and linear layers in a network.

Table 6: Test accuracy of pruned VGG11/16 and ResNet20/56 on CIFAR-10 and CIFAR-100 datasets.

Dataset	CIFAR-10			CIFAR-100		
	90%	95%	98%	90%	95%	98%
VGG11 (Full Network)	92.05	-	-	69.09	-	-
LT [9]	91.50±0.27	90.35±0.12	86.71±0.34	68.58±0.18	67.42±0.32	65.23±0.55
SNIP [23]	91.29±0.06	90.77±0.35	89.16±0.07	68.65±0.58	67.48±0.35	64.05±0.37
GraSP [41]	90.48±0.08	89.70±0.12	88.46±0.31	66.11±0.18	64.47±0.56	62.07±0.18
Random Tickets (Ours)	91.67±0.24	90.86±0.26	89.11±0.16	68.82±0.47	66.90±0.09	63.85±0.19
VGG16 (Full Network)	93.77	-	-	73.24	-	-
LT [9]	93.56±0.17	93.23±0.08	36.27±45.5	72.32±0.11	70.65±0.18	64.88±0.72
SNIP [23]	93.47±0.16	93.13±0.05	92.05±0.02	72.41±0.18	71.18±0.26	67.25±0.37
GraSP [41]	93.03±0.16	92.71±0.33	91.79±0.11	71.08±0.43	69.99±0.41	67.68±0.29
Random Tickets (Ours)	93.80±0.06	93.31±0.17	92.13±0.15	72.42±0.14	71.12±0.28	68.17±0.36
ResNet20 (Full Network)	94.37	-	-	73.45	-	-
LT [9]	91.69±0.06	90.12±0.12	86.92±0.21	67.25±0.44	63.47±0.10	52.40±0.42
SNIP [23]	91.76±0.06	89.91±0.07	85.28±0.07	67.21±0.38	61.88±0.23	47.25±0.77
GraSP [41]	91.64±0.17	90.24±0.15	86.60±0.05	67.53±0.46	63.60±0.06	53.82±0.38
Random Tickets (Ours)	91.88±0.01	90.13±0.10	86.66±0.09	67.49±0.35	63.42±0.18	54.62±0.32
ResNet56 (Full Network)	94.49	-	-	76.94	-	-
LT [9]	92.70±0.42	92.14±0.23	89.59±0.40	71.50±0.44	69.11±0.43	54.24±0.56
SNIP [23]	93.37±0.11	43.05±19.67	<i>10.00±0.00</i>	63.01±4.15	6.52±3.86	<i>1.00±0.00</i>
GraSP [41]	93.30±0.15	92.22±0.10	22.14±13.56	73.64±0.19	70.03±0.51	10.50±9.22
Random Tickets (Ours)	93.53±0.11	92.24±0.21	89.80±0.24	72.87±0.66	70.88±0.39	63.60±0.58

Table 7: Test accuracy of partially-trained tickets and our hybrid tickets of VGG11/16 and ResNet20/56 on CIFAR-10 and CIFAR-100 datasets.

Dataset	CIFAR-10			CIFAR-100		
	90%	95%	98%	90%	95%	98%
VGG11 (Full Network)	92.05	-	-	69.09	-	-
Learning Rate Rewinding [37]	92.23±0.07	91.81±0.22	88.71±0.36	69.41±0.29	68.70±0.25	66.64±0.36
Hybrid Tickets (Ours)	92.21±0.13	91.61±0.12	91.04±0.21	69.60±0.34	69.08±0.41	67.45±0.02
VGG16 (Full Network)	93.77	-	-	73.24	-	-
Learning Rate Rewinding [37]	94.01±0.12	93.70±0.22	<i>10.00±0.00</i>	74.14±0.17	72.82±0.35	65.51±1.41
Hybrid Tickets (Ours)	94.01±0.11	93.89±0.22	93.31±0.04	73.81±0.28	72.88±0.44	70.75±0.15
ResNet20 (Full Network)	94.37	-	-	73.45	-	-
Learning Rate Rewinding [37]	93.46±0.15	92.12±0.14	88.78±0.23	70.45±0.10	65.27±1.26	54.77±0.55
Hybrid Tickets (Ours)	93.18±0.19	92.13±0.12	89.06±0.17	68.83±0.31	65.77±0.84	57.74±0.59
ResNet56 (Full Network)	94.49	-	-	76.94	-	-
Learning Rate Rewinding [37]	94.57±0.34	94.13±0.26	92.35±0.14	75.69±0.71	68.22±0.46	57.82±0.48
Hybrid Tickets (Ours)	94.10±0.34	93.39±0.13	92.02±0.13	73.37±1.12	71.20±0.90	65.44±1.70

D Discussions of Iterative Magnitude Pruning (IMP)

In [9], the authors propose to find initial tickets with *extremely high sparsities* by iteratively throwing out weights with the smallest magnitude since one-shot LT method can find initial tickets with a certain layer being fully pruned. This method is named as iterative magnitude pruning (IMP) and is regarded as a standard pruning method such as in [37] and [10]. In this section, however, we show that the iterative procedure itself **cannot** help initial tickets to pass our sanity-checks although it can help these tickets to be trainable at a high level of sparsity.

In both [9] and [27], the authors admit that initial tickets found in VGG and ResNet architectures with CIFAR-10 and more complicated datasets are robust to re-initialization, and thus is not efficient. However, since this phenomenon is verticle to the property we discuss throughout our paper, i.e., whether the iterative procedure can help to find a good structure or to encode data information into

initial tickets, for the integrity of our work, we apply our sanity-checks on iteratively-found initial tickets and show the experiment results in Figure 5.

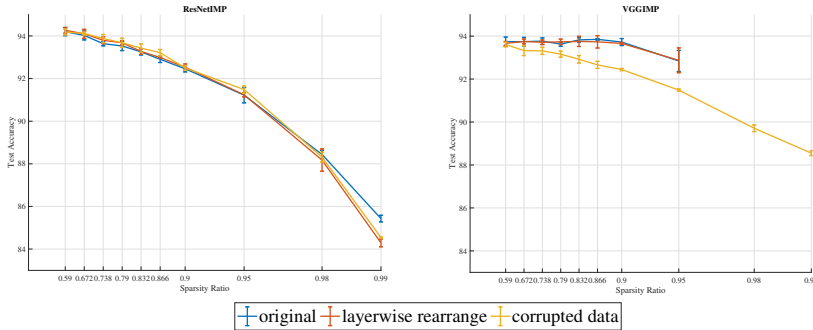


Figure 5: Sanity checks of iteratively-found initial tickets of ResNet32 and VGG19 on CIFAR10.

The results are similar to those of one-shot LT initial tickets, that is iteratively-found initial tickets cannot pass sanity-checks. One interesting thing we observe from the figure is that although on VGG the *corrupt data* check causes a performance gap, the tickets obtained by it is much more stable when the sparsity is high.¹³ This suggests there is still room to explore how to use the data effectively to get trainable sparse initial tickets.

Moreover, we investigate the efficiency of iterative tickets including iterative LT and iterative LR rewinding. The results can be found in Table 8. From the table we can get the following several conclusions.

- **The iterative procedure even hardly helps when finding initial tickets of modern architectures as ResNet and VGG.** Compared to LT, Iterative LT only gains a little improvement or even performs much *worse* than LT. Furthermore, our Random Tickets can behave better than both Iterative LT and LT. This observation further strengthens our statement that initial tickets’ performance to a large extent depend on the *layerwise keep-ratios* obtained. At the same time, this observation suggests that it may be difficult to answer the question of *whether we can find a kind of initial ticket that can pass sanity-checks*. The exploration will be left as future works.
- **The iterative procedure helps a LOT when finding partially-trained tickets.** Compared to Learning Rate Rewinding, Iterative LR rewinding can benefit from the iterative process and thus becomes a hard-to-beat sota to now.
- **Iterative Hybrid Tickets can also benefit from the iterative procedure.** Due to the simple structure of our smart ratio, the Iterative Hybrid Tickets (IHT) cannot surpass the Iterative LR rewinding method. However, note that the performance gap is small, we assert that this phenomenon together with our sanity-checks on LR rewinding suggests there *must be some information encoded in both weights and partially-trained tickets’ structure*, and that’s the reason causes the different effectiveness of the iterative procedure on initial/partially-trained tickets. This sheds light that if we can find a better way to determine the layerwise keep-ratios explicitly, we can find a better pruning method instead of keeping weights of largest magnitude globally to implicitly find those keep-ratios.

E Sanity-Checks on More Complicated Dataset

Also for the integrity of our work, in this section we provide results of applying our sanity-checks using the CIFAR-100 dataset with ResNet32/VGG19 architectures on initial tickets. The results are presented in Figure 6.

Obviously, as we expect, we can see conclusions on initial tickets still hold on this more complicated dataset: initial tickets all failed to pass our sanity-checks, suggesting the *generalizability across datasets* of our checking methods.

¹³As a comparison, the original VGGIMP’s 0.98 sparsity result is 78.85 ± 9.76 .

Table 8: Test accuracy of pruned VGG19 and ResNet32 on CIFAR-10 and CIFAR-100 datasets, indicating the effectiveness of the iterative procedure on both LR Rewinding and Hybrid Tickets.

Dataset	CIFAR-10			CIFAR-100		
	90%	95%	98%	90%	95%	98%
VGG19 (Full Network)	93.70	-	-	72.60	-	-
LT [9]	93.66±0.08	93.39±0.13	10.00±0.00	72.58±0.27	70.47±0.19	1.00±0.00
Random Tickets (Ours)	93.77±0.10	93.42±0.22	92.45±0.22	72.55±0.14	71.37±0.09	68.98±0.34
Iterative LT [9]	93.72±0.16	92.84±0.49	78.85±9.80	72.74±0.36	71.18±0.12	67.69±0.50
Learning Rate Rewinding [37]	94.14±0.17	93.99±0.15	10.00±0.00	73.73±0.18	72.39±0.40	1.00±0.00
Iterative LR rewinding	94.13±0.09	94.16±0.19	93.95±0.14	74.53±0.09	74.52±0.07	72.90±0.11
Iterative Hybrid Tickets (Ours)	93.95±0.12	94.15±0.20	93.87±0.24	74.24±0.04	74.19±0.06	72.63±0.28
ResNet32 (Full Network)	94.62	-	-	74.57	-	-
LT [9]	92.61±0.19	91.37±0.28	88.92±0.49	69.63±0.26	66.48±0.13	60.22±0.60
Random Tickets (Ours)	92.97±0.05	91.60±0.26	89.10±0.33	69.70±0.48	66.82±0.12	60.11±0.16
Iterative LT [9]	92.46±0.15	91.22±0.36	88.44±0.19	69.13±0.21	66.69±0.45	59.53±0.57
Learning Rate Rewinding [37]	94.14±0.10	93.02±0.28	90.83±0.22	72.41±0.49	67.22±3.42	59.22±1.15
Iterative LR rewinding	94.86±0.07	94.71±0.08	93.43±0.15	74.57±0.08	72.59±0.12	67.17±0.72
Iterative Hybrid Tickets (Ours)	94.73±0.04	94.35±0.07	93.45±0.12	72.89±0.31	71.23±0.11	67.81±0.09

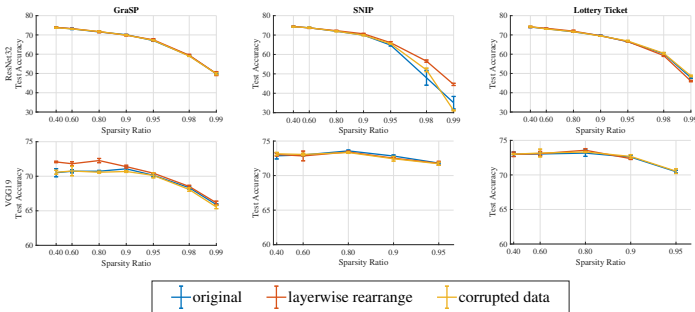


Figure 6: Sanity checks of the initial tickets of ResNet32 and VGG19 on CIFAR100.

F Other Related Works

We add some related works on network pruning. Most pruning methods are applied to pre-trained networks and thus *require training the full network* [33, 22, 15, 6]. Another line of pruning methods prune the models during training [28, 5, 46, 12]. However, these methods do not save many resources as they require the whole network during training time [17].

Towards a better understanding of network pruning, [27] evaluates several *structured pruning methods* and concludes that training a large, over-parameterized model is often not necessary to obtain an efficient final model. [3] releases a library, ShrinkBench, for comprehensively evaluating pruning methods. [32] studies the transfer of pruned networks across datasets. [30] gives a theoretical understanding of the lottery tickets hypothesis. [45] shows the importance of setting weights to zero, the signs of weights and masking. [43] shows the tickets can be obtained in an earlier phase than that in [9].