Data-Efficient Double-Win Lottery Tickets from Robust Pre-training

Tianlong Chen [1] **Zhenyu Zhang** [1] **Sijia Liu** [2 3] **Yang Zhang** [3] **Shiyu Chang** [4] **Zhangyang Wang** [1]

Abstract

Pre-training serves as a broadly adopted starting point for transfer learning on various downstream tasks. Recent investigations of *lottery tickets hypothesis* (LTH) demonstrate such enormous pre-trained models can be replaced by extremely sparse subnetworks (a.k.a. *matching subnetworks*) without sacrificing transferability. However, practical security-crucial applications usually pose more challenging requirements beyond standard transfer, which also demand these subnetworks to overcome adversarial vulnerability. In this paper, we formulate a more rigorous concept, **Double-Win Lottery Tickets**, in which a located subnetwork from a pre-trained model can be independently transferred on diverse downstream tasks, to reach **BOTH** the same standard and robust generalization, under **BOTH** standard and adversarial training regimes, as the full pre-trained model can do. We comprehensively examine various pre-training mechanisms and find that robust pre-training tends to craft sparser double-win lottery tickets with superior performance over the standard counterparts. For example, on downstream CIFAR-10/100 datasets, we identify double-win matching subnetworks with the standard, fast adversarial, and adversarial pre-training from ImageNet, at $89.26\%/73.79\%$, $89.26\%/79.03\%$, and $91.41\%/83.22\%$ sparsity, respectively. Furthermore, we observe the obtained double-win lottery tickets can be more **data-efficient** to transfer, under practical data-limited (e.g., 1% and 10%) downstream schemes. Our results show that the benefits from robust pre-training are amplified by the lottery ticket scheme, as well as the data-limited transfer setting. Codes are available at https://github.com/VITA-Group/Double-Win-LTH.

[1]Department of Electrical and Computer Engineering, University of Texas at Austin [2]Michigan State University [3]MIT-IBM Watson AI Lab [4]University of California, Santa Barbara. Correspondence to: Zhangyang Wang <atlaswang@utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

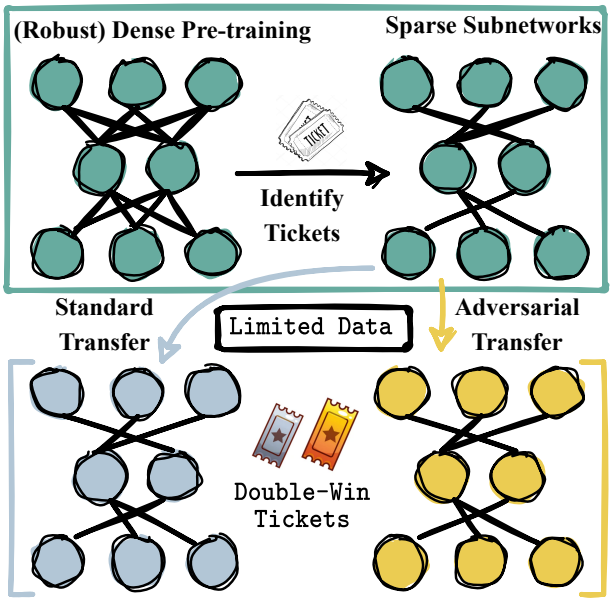

Figure 1. Overview of our work paradigm: we investigate the existence of double-win lottery tickets drawn from robust pre-training in the scenario of transfer learning, with the full training data and the limited training being available, respectively.

1. Introduction

The lottery tickets hypothesis (LTH) (Frankle & Carbin, 2018) demonstrates that there exist subnetworks in dense neural networks, which can be trained in isolation from the same random initialization and match the performance of the dense counterpart. We call such subnetworks as winning tickets. Unlike the conventional pipeline (Han et al., 2016) of model compression that follows the train-compress-retrain process and aims for efficient inference, the LTH sheds light on the potential for more computational savings by training a small subnetwork from the start if only we had known which subnetwork to choose. However, finding these intriguing subnetworks is quite costly since the current most effective approach, iterative magnitude pruning (IMP) (Frankle & Carbin, 2018; Han et al., 2016), requires multiple rounds of burdensome (re-)training, especially for large models like BERT (Devlin et al., 2018). Fortunately, recent studies (Chen et al., 2020b;a) provide a remedy by leveraging the popular paradigm of pre-training and fine-tuning, which first identifies critical subnetworks (a.k.a. pre-trained tickets) from standard pre-training and then transfers

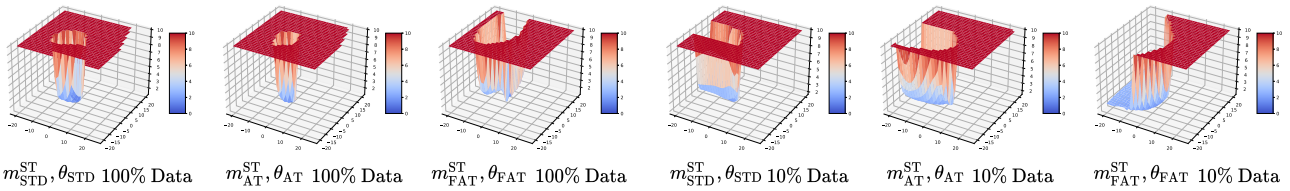

Figure 2. Loss landscape visualization of subnetworks (73.79% sparsity) from diverse adversarial fine-tuning schemes. Each sparse network is first identified by IMP and standard training on the pre-training task with standard θ_{STD}, fast adversarial θ_{FAT}, or adversarial θ_{AT} pre-training, respectively. Then, they are fine-tuned from the corresponding (robust) pre-training on downstream CIFAR-10 with 100% or 10% training data.

to a range of downstream tasks. The demonstrated *universal transferability* across various datasets and tasks, shows the positive sign of replacing the gigantic pre-trained models with a much smaller subnetwork while maintaining the impressive downstream performance and leading to substantial memory/computation reductions. Meantime, the extraordinary cost of both pre-training and finding pre-trained tickets can be amortized by reusing and transferring to diverse downstream tasks.

Nevertheless, in practical settings, the deployed models usually ask for strong robustness, which is beyond the scope of standard transfer, e.g., for safety-critical applications like autonomous cars and face recognition. Therefore, a more challenging requirement arises, which demands the located subnetworks can effectively transfer in both standard and adversarial training schemes (Madry et al., 2017). Thus, it is a new perspective to investigate the transferability of pre-trained tickets across diverse training regimes, differing from previous works (Chen et al., 2020b;a) on transferring downstream datasets and tasks. This inspires us to propose a new hypothesis of lottery tickets. Specifically, when an identified sparse subnetwork from pre-training can be independently trained (transferred) on diverse downstream tasks, to match the same accuracy and robustness, under both standard and adversarial training regimes, as the full pre-trained model can do – we name it a **Double-Win Lottery Ticket** illustrated in Figure. 1.

Meanwhile, inspired by (Salman et al., 2020), which suggests that robust pre-training shows better transferability for dense models, we examine (1) whether this appealing property still holds under the lens of sparsity; (2) how can robust pre-training benefits our double-win tickets compared to its standard counterpart. To address such curiosity, we comprehensively investigate representative robust pre-training approaches besides standard training, including fast adversarial (FAT) (Wong et al., 2020) and adversarial (AT) (Madry et al., 2017) pre-training. Our results reveal the prevailing existence of double-win tickets with different pre-training, and suggest the subnetworks obtain from AT pre-trained models consistently achieve superior generalization and robustness, under both standard and adversarial transfer learning, when the typical full training data is available for downstream tasks.

Yet, another critical constraint in real-world scenarios is the possible scarcity of training data (e.g., due to the difficulty of data collection and annotation). What makes it worse is that satisfactory adversarial robustness intrinsically needs more training samples (Schmidt et al., 2018). Our proposed double-win tickets from robust pre-training tackle this issue by leveraging the crafted sparse patterns as an inductive prior, which (i) is found to reduce the sample complexity (Zhang et al., 2021b) and brings data efficiency (Chen et al., 2021a); (ii) converges to a flatter loss landscapes with improved robust generalization as advocated by (Wu et al., 2020; Hein & Andriushchenko, 2017), particularly for data-scarce settings shown in Figure 2. To support these intuitions, extensive experiments about few-shot (or data-efficient) transferability are evaluated with only 10% or 1% data for adversarial downstream training (Jiang et al., 2020). In what follows, we summarize our **contributions** in order to bridge LTH and its practical usage in the data-limited and security-crucial applications:

- We define a more rigorous notion of double-win lottery tickets, which requires the sparse subnetworks found on pre-trained models to have the same transferability as the dense pre-trained ones: in terms of both accuracy and robustness, under both standard and adversarial training regimes, and towards a variety of downstream tasks. We show such tickets widely exist.
- Using IMP, we find double-win tickets broadly across diverse downstream datasets and at non-trivial sparsity levels $79.03\% \sim 89.26\%$ and $83.22\% \sim 96.48\%$ sparsity, using the fast adversarial (FAT) and adversarial (AT) pre-training. In general, subnetworks located from the AT pre-trained model have superior performance than FAT and standard pre-training.
- We further demonstrate the intriguing property of double-win tickets in the data-limited transfer settings (e.g., 10%, 1%). In this specific situation, FAT can surprisingly find higher-quality subnetworks with small sparsity while AT overtakes in a larger sparsity range.
- We show that adopting standard or adversarial training in the process of IMP makes no significant difference for the transferability of identified subnetworks on downstream tasks.

2. Related Works

The lottery tickets hypothesis (LTH). The (LTH) (Frankle & Carbin, 2018) points out the existence of sparse subnetworks which are capable of training from scratch and match or even surpass the performance of the full network. (Frankle et al., 2019; Renda et al., 2020) further scale up LTH to larger datasets and networks by weight rewinding techniques that re-initialize the subnetworks to the weight from the early training stage instead of scratch. Follow-up researchers have explored LTH in various fields, including image classification (Frankle & Carbin, 2018; Liu et al., 2019; Wang et al., 2020; Evci et al., 2019; Ma et al., 2021; You et al., 2020), natural language processing (Gale et al., 2019; Yu et al., 2020; Chen et al., 2020d;b), vision+language multi-modal tasks (Gan et al., 2021), graph neural networks (Chen et al., 2021b), generative adversarial networks (Chen et al., 2021e;a), reinforcement learning (Yu et al., 2020) and life-long learning (Chen et al., 2021c). Most existing works of LTH identify subnetworks by resource-consuming (iterative) weight magnitude pruning (Han et al., 2016; Frankle & Carbin, 2018). Studies about the transferability of the subnetworks provide a potential offset to the computationally expensive process of finding high-quality subnetworks. (Chen et al., 2020b; Desai et al., 2019; Morcos et al., 2019; Mehta, 2019) investigate the transferability across different datasets (i.e., *dataset transfer*), while other pioneers study the transferability of pre-trained tickets from supervised and self-supervised vision pre-training (Chen et al., 2020a) across diverse downstream tasks like detection and segmentation (i.e., *task transfer*). These two transfer capabilities form the core target of the pre-training / fine-tuning paradigm. In this paper, we take a leap further to meet more practical requirements by designing the concept of double-win tickets. It examines the transferability across different downstream training regimes, including standard and adversarial transfer, data-rich and data-scarce transfer. To our best knowledge, this *training schemes transfer* has never been explored in the LTH literature, offering a new view to analyze beneficial properties of pre-trained tickets.

Adversarial training and robust pre-training. Deep neural networks are vulnerable to imperceivable adversarial examples (Szegedy et al., 2013), which limits their applications in security-crucial scenarios. To tackle this limitation, massive defense methods were proposed (Goodfellow et al., 2014; Kurakin et al., 2016; Madry et al., 2017), while many of them, except adversarial training (Madry et al., 2017), were later found to provide false security from obfuscated gradients caused by input transformation (Xu et al., 2017; Liao et al., 2018; Guo et al., 2017; Dziugaite et al., 2016) and randomization (Liu et al., 2018b;a; Dhillon et al., 2018). Besides, several works that focus on certified defenses (Cohen et al., 2019; Raghunathan et al., 2018), aim to provide a theoretical guarantee of robustness yet lack scalability. Nowadays, adversarial training (AT) (Madry et al., 2017) remains one of the most effective approaches and numerous following works endeavor to improve its performance (Zhang et al., 2019b; Chen et al., 2021d) and computation efficiency (Shafahi et al., 2019b; Zhang et al., 2020), while it may suffer from overfitting issues. Particularly, (Zhang et al., 2019a; Shafahi et al., 2019a; Wong et al., 2020) point out the overfitting phenomenon in several fast adversarial training methods, where sometimes the robust accuracy against a PGD adversary suddenly drops to nearly zero after some training. (Andriushchenko & Flammarion, 2020) suggests it can be mitigated by performing local linearization to the loss landscape in those "fast" AT. Another reported *robust overfitting* (Rice et al., 2020) seems to raise a completely new challenge for the classical AT (not fast), which can be alleviated by early stopping and smoothening (Chen et al., 2021d). Meantime, several pioneering efforts have been made to obtain models that are both compact and robust to adversarial attacks (Gui et al., 2019; Sehwag et al., 2020; Fu et al., 2021), spurious features (Zhang et al., 2021a), and input corruptions (Diffenderfer et al., 2021)

Although the standard pre-training is commonly used in both areas of computer vision (He et al., 2019b; Girshick et al., 2014) and natural language process (Devlin et al., 2018), such as the supervised ImageNet and self-supervised BERT (Devlin et al., 2018) pre-training, there exist only few investigations of robust pre-training. The work (Chen et al., 2020c) for the first time demonstrates that adversarial pre-training can speed up and improve downstream adversarial fine-tuning. Latter works (Jiang et al., 2020; Salman et al., 2020) show extra benefits of enhanced dataset transferability and data efficiency from adversarial pre-training. All the above studies were only conducted with dense networks.

3. Preliminary

Networks. Aligned with previous work of pre-trained tickets (Chen et al., 2020a), we consider the official ResNet-50 (He et al., 2016) as the unpruned dense model, and formulate the output of the network as $f(x;\theta)$, where x is the input images and $\theta \in \mathbb{R}^d$ is the network parameters. In the same way, a subnetwork is a network $f(x; m \odot \theta)$[1] with a binary pruning mask $m \in \{0,1\}^d$, where $\odot$ is the element-wise product. In our experiment, we sparsify the major part of the dense network, leaving the task-specific classification head out of the scope of pruning.

Adversarial training (AT). The classical AT (Madry et al., 2017) remains one of the most effective approaches to tackle the vulnerability for small perturbations and build a robust model, in which the standard empirical risk mini-

[1]For simplicity purpose, we use $f(x;\theta)$ to denote a network or its output in different contexts.

mization is replaced by a robust optimization, as depicted in equation 1:

$$\min_{\theta} \mathbb{E}_{(x,y)\in\mathcal{D}} \max_{\|\delta\|_p\leq\epsilon} \mathcal{L}\big(f(x+\delta;\theta),y\big) \tag{1}$$

where the perturbation is constrained in an ℓ_p norm ball with the radius equals to ϵ, and input data x with its associated label y are sampled from the training set $\mathcal{D}$. To solve the inner maximization problem, projected gradient descent (PGD) (Madry et al., 2018) is frequently adopted and believed to be the strongest first-order adversary, which works in an iterative fashion as equation 2:

$$\delta^{t+1} = \text{proj}_{\mathcal{P}}\Big(\delta^t + \alpha\cdot\text{sgn}\big(\nabla_x\mathcal{L}(f(x+\delta^t;\theta),y)\big)\Big) \tag{2}$$

where δ^t is the generated perturbation, t denotes the number of iterations, α represents the step size, and sgn is a function that returns the sign of its input. Besides, (Wong et al., 2020) proposes a fast adversarial training method and claimed that adversarial training with Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014), which is the single-step variant of PGD, can be as effective as PGD-based adversarial training once combined with random initialization. In the following context, we will refer standard empirical risk minimization process to standard training (**ST**) and robust optimization to adversarial training (**AT**) or fast adversarial training (**FAT**) according to the number of PGD steps. We remark that FAT alone may cause the issue of robust catastrophic overfitting (Andriushchenko & Flammarion, 2020) when the train-time attack strength grows. Thus, an early-stopping policy (Rice et al., 2020), which was also suggested by (Andriushchenko & Flammarion, 2020), is adopted to mitigate such catastrophic overfitting.

Pruning algorithms. For a dense neural network $f(x;\theta)$, we adopt the unstructured iterative magnitude pruning (IMP) (Frankle & Carbin, 2018; Han et al., 2016) to identify the subnetworks $f(x;m\odot\theta)$, which is a standard option for mining lottery tickets (Frankle & Carbin, 2019). More precisely, starting from the pre-trained weights θ_p as initialization, we follow the circle of prune-rewind-retrain to locate subnetworks, in which we prune $p\%$ of the remaining weight with the smallest magnitude and rewind the weights of the subnetwork to their values from θ_p. We repeat the prune-rewind-retrain process until the desired sparsity.

In our experiments, we choose a precise $p\% = 20\%$ (Frankle & Carbin, 2019; Chen et al., 2020a) and consider three initialization: the standard pre-trained[2] ResNet-50 θ_{STD}, the PGD-based adversarial pre-trained[3] ResNet-50 θ_{AT} and fast adversarial pre-trained[4] ResNet-50 by θ_{FAT}. All the models are pre-trained with the classification task on the

[2] https://pytorch.org/vision/stable/models.html

[3] https://github.com/microsoft/robust-models-transfer

[4] https://github.com/locuslab/fast_adversarial/tree/master/ImageNet

Table 1. Summary of our setups.

Source domain: finding subnetworks via pruning with pre-trained weights θ_p			
Target domains: evaluating transferability of $f(x;m\odot\theta_p)$ across training schemes			
Training scheme	Standard Training	Adversarial Training	
Evaluation metrics	SA	SA	RA
Double-Win Tickets if and only if	winning ✓	winning ✓	winning ✓

ImageNet source dataset (Krizhevsky et al., 2012). It is worthy to mention that all pruning are applied to the source dataset (or pre-training task) only, since our main focus is investigating the mask transferability cross training schemes of subnetworks obtained from pre-training.

Downstream datasets, training and evaluation. After producing subnetworks from the pre-training task on ImageNet by IMP, we implement both standard and adversarial transfer on three downstream datasets: CIFAR-10 (Krizhevsky & Hinton, 2009), CIFAR-100 (Krizhevsky & Hinton, 2009), and SVHN (Netzer et al., 2011). For adversarial training, we train the network against ℓ_∞ adversary of 10-steps Projected Gradient Descent (PGD-10) with $\epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$. On CIFAR-10/100, we train the network for 100 epochs with an initial learning rate of 0.1 and decay by ten times at $50, 75$th epoch. As for SVHN, we start from 0.01 learning rate and decay by a cosine annealing schedule for 80 epochs. Moreover, an SGD optimizer is adopted with 5×10^{-4} weight decay and 0.9 momentum. And we use a batch size of 128 for all downstream experiments. To evaluate the downstream performance of subnetworks, we report both Standard Testing Accuracy (**SA**) and Robust Testing Accuracy (**RA**), which are computed on the original and adversarial perturbed test images respectively. During the inference, we generate the adversarial test images by PGD-20 attack with other hyper-parameters kept the same as in training (Chen et al., 2021d). More details are in Sec. A1.

Double-Win lottery tickets. Here we introduce formal definitions of our double-win tickets:

▷ *Matching subnetworks* (Chen et al., 2020b;a; Frankle et al., 2020). A subnetwork $f(x;m\odot\theta)$ is matching for a training algorithm $\mathcal{A}_t^{\mathcal{T}}$ if its performance of evaluation metric $\epsilon^{\mathcal{T}}$ is no lower than the pre-trained dense network $f(x;\theta_p)$ that trained with the same algorithm $\mathcal{A}_t^{\mathcal{T}}$, namely:

$$\epsilon^{\mathcal{T}}\Big(\mathcal{A}_t^{\mathcal{T}}\big(f(x;m\odot\theta)\big)\Big) \geq \epsilon^{\mathcal{T}}\Big(\mathcal{A}_t^{\mathcal{T}}\big(f(x;\theta_p)\big)\Big) \tag{3}$$

▷ *Winning Tickets* (Chen et al., 2020a; Frankle et al., 2020). If a subnetwork $f(x;m\odot\theta)$ is matching with $\theta=\theta_p$ for a training algorithm $\mathcal{A}_t^{\mathcal{T}}$, then it is a winning ticket for $\mathcal{A}_t^{\mathcal{T}}$.

▷ *Double-Win Lottery Tickets*. When a subnetwork $f(x;m\odot\theta)$ is a winning ticket for standard training under metric **SA** and for adversarial training under both metrics **SA** and **RA**, we name it as a double-win lottery ticket, as demonstrated in Table 1.

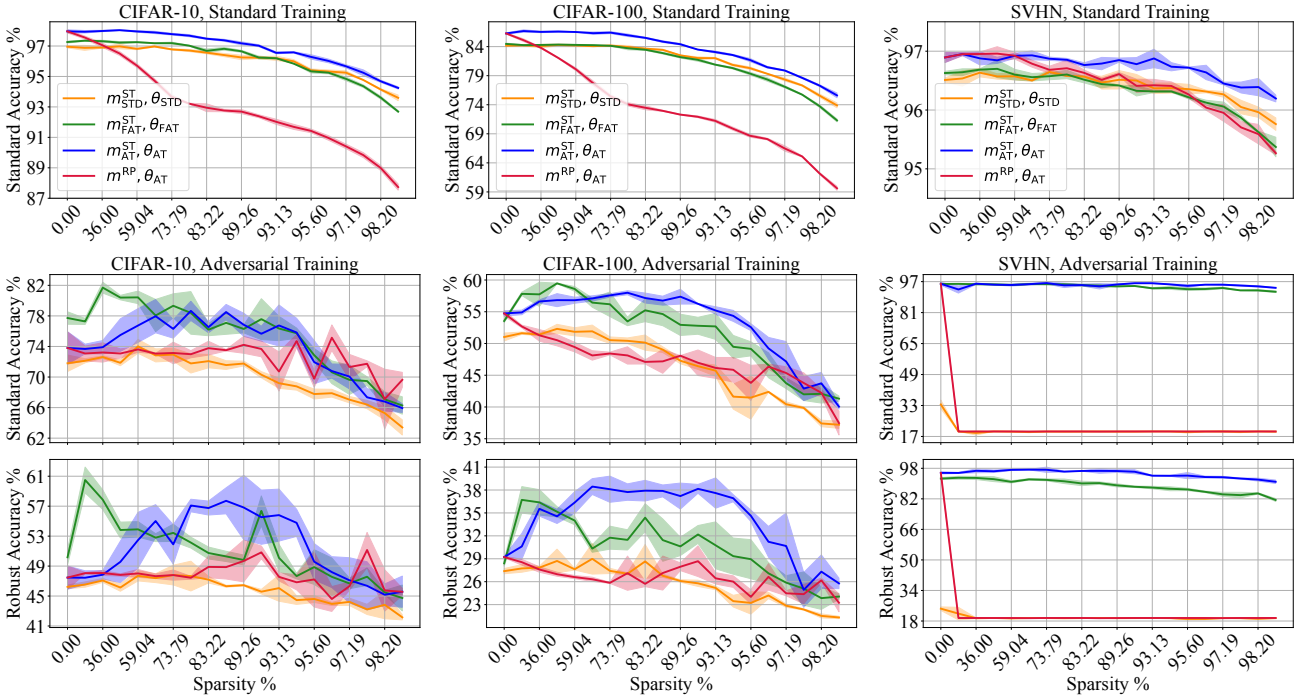

Figure 3. Comparison results of the subnetworks that are fine-tuned on three downstream datasets (i.e., CIFAR-10, CIFAR-100 and SVHN) under both standard and adversarial training regimes. For standard training, we report the standard accuracy; while for adversarial training, both standard and robust accuracy are presented. Orange, Green and Blue represent the performance of subnetworks generated from IMP on pre-trained ImageNet classification (m^{ST}) with standard re-training and different pre-trained weights (i.e. standard θ_{STD}, fast adversarial θ_{FAT}, and adversarial θ_{AT} pre-training, respectively) while Red stands for random pruning with adversarial pre-trained weight. The solid line and shading area are the mean and standard deviation of standard/robust accuracy.

4. Drawing Double-Win Lottery Tickets from Robust Pre-training

In this section, we evaluate the quality of subnetworks $f(x; m \odot \theta)$ on multiple downstream tasks under both standard and adversarial training regimes. Before that, we extract desired subnetworks via IMP on the ImageNet classification tasks. During the process, the pre-trained weights θ_p are treated as initialization for rewinding, and standard training (ST) or adversarial training (AT) is adopted for re-training the sparse model on the pre-trained task. In the downstream transferring stage, subnetworks start from mask $m_p^{\mathcal{P}}$ and pre-trained weights θ_p, where $p \in \{\mathrm{STD}, \mathrm{FAT}, \mathrm{AT}\}$ stands for {standard, fast adversarial, adversarial} pre-training, and the pruning method $\mathcal{P} \in \{\mathrm{ST}, \mathrm{AT}, \mathrm{RP}, \mathrm{OMP}\}$ which represents {IMP with standard (re-)training, IMP with adversarial (re-)training, random pruning, one-shot magnitude pruning} (Han et al., 2016) on the pre-training task (RP or OMP indicates there is no re-training). In the following content, Section 4.1 shows the existence of double-win lottery tickets from diverse (robust) pre-training with impressive transfer performance for both standard or adversarial training; Section 4.2 investigates the effects of standard or adversarial re-training on the quality of derived double-win tickets. All experiments have three independent replicates with different random seeds and the mean results and standard deviation are reported.

4.1. Do Double-Win Lottery Tickets Exist?

To begin with, we validate the existence of double-win lottery tickets drawn from diverse (robust) pre-training and source ImageNet dataset. We consider the sparsity masks from IMP with standard re-training m^{ST} and random pruning m^{RP} on the pre-training task, together with three different pre-trained weights, i.e. standard θ_{STD}, fast adversarial θ_{FAT}, and adversarial weights θ_{AT}. As demonstrated in Figure 3, we adopt two downstream fine-tuning receipts, i.e., standard training (report SA) and adversarial training (report SA/RA)), simultaneously. **Note that all presented numbers here are subnetwork's sparsity levels.** Several consistent observations can be drawn:

① Double-win lottery tickets generally exist from various pre-training, showing unimpaired performance on diverse downstream tasks for both standard and adversarial transfer. To account for fluctuations, we consider the performance of subnetworks is *matching* when it's within one standard deviation of the unpruned dense network. The **extreme sparsity levels** of subnetworks drawn from $\{\theta_{\mathrm{STD}}, \theta_{\mathrm{FAT}}, \theta_{\mathrm{AT}},\}$

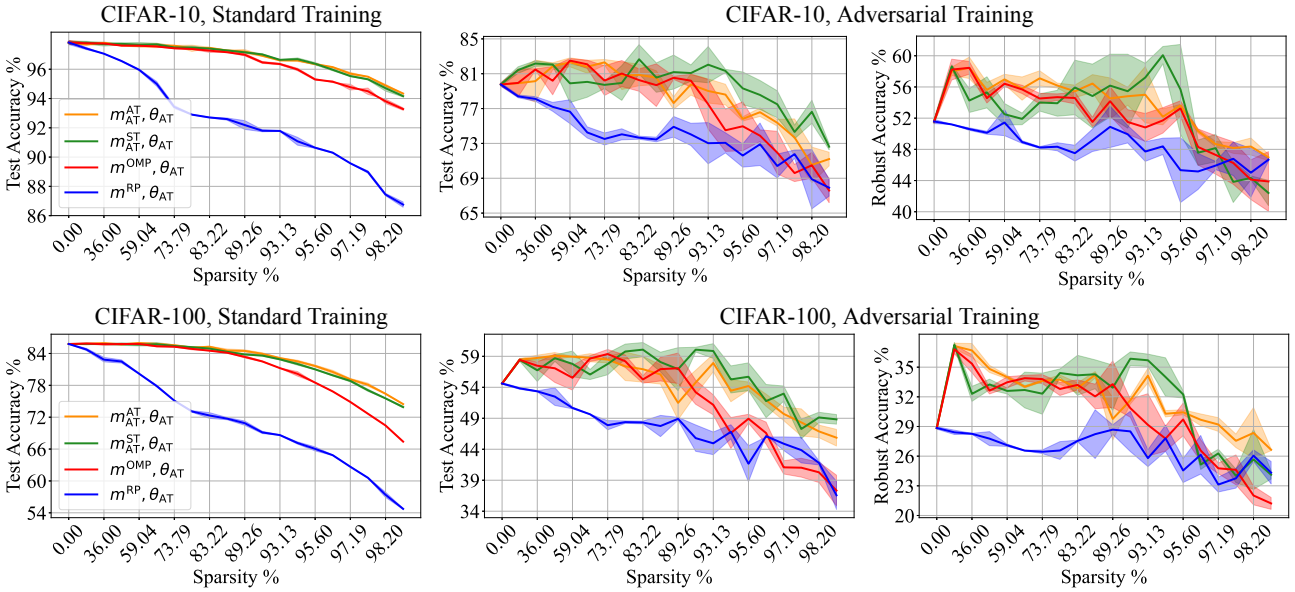

Figure 4. Comparison results of the subnetworks that are independently trained on three downstream datasets (i.e., CIFAR-10, CIFAR-100, SVHN) under both standard and adversarial training regimes. For standard training, we report the standard accuracy; while for adversarial training, both standard and robust accuracy are presented. Orange, Green, red and Blue represent the performance of subnetworks generated by IMP with standard re-training ($m^{\text{ST}}_{\text{AT}}$), adversarial re-training ($m^{\text{AT}}_{\text{AT}}$) on ImageNet classification task, one shot magnitude pruning (m^{OMP}) and random pruning (m^{RP}), together with the adversarial pre-training (θ_{AT}) as initialization. The solid line and shading area are the mean and standard deviation of standard/robust accuracy.

are $\{89.26\%, 89.26\%, 91.41\%\}$, $\{73.79\%, 79.03\%, 83.22\%\}$, $\{0.00\%, 79.03\%, 96.48\%\}$ with matching or even superior standard and robust performance under both training regimes (standard and adversarial) on CIFAR-10, CIFAR-100 and SVHN, respectively. All these double-win tickets surpass randomly pruned subnetwork by a significant performance margin. It demonstrates the superior performance of double-win tickets is not only from reduced parameter counts but also credits to the located sparse structural patterns.

② Subnetworks identified from adversarial pre-training consistently outperform the ones from fast adversarial and standard pre-training across all three downstream classification tasks, which is aligned with the result in (Salman et al., 2020). Taking the extreme sparsity as an indicator, the adversarial pre-training finds double-win lottery tickets to the extreme sparsity of $83.22\% \sim 96.48\%$ while fast adversarial, standard pre-training reach the extreme sparsity level of $20.00\% \sim 89.26\%$ and $0.00\% \sim 89.26\%$. This suggests that the adversarial pre-trained model can serve as a desirable starting point for locating high-quality double-win tickets to cover both standard and adversarial downstream transferability. Note that here all downstream transferring can access full training data, i.e., *data-rich* fine-tuning.

③ Along with the increase of sparsity, we notice that the performance improvements from adversarial pre-training θ_{AT} (i) remain stable in the standard transfer (the first row in Figure 3) even at extreme sparsity like 98.56%; (ii) first increase then diminish in adversarial transfer after 95.60% sparsity. It suggests that double-win tickets from adversarial pre-training are more sensitive to the aggressive sparsity in the scenario of adversarial transfer learning.

④ The comparison results among different pre-training varies with the training regime of downstream tasks. Take the result on CIFAR-100 as an example, the subnetworks drawn from fast adversarial pre-training shows superior performance than the ones from standard training in the range of sparsity level from $0.00\% \sim 98.56\%$ under adversarial training. While for the standard training regime, fast adversarial and standard pre-training locate subnetworks with similar performance across the sparsity level from 0.00% to 95.60%. The inferior performance of standard pre-training suggests that the vanilla lottery tickets that only focus on the standard training regime and use standard test accuracy as the only evaluation metric, is insufficient in practical security-crucial scenarios. Thus we take adversarial transfer into consideration and propose the concept of double-win lottery tickets to improve the original LTH.

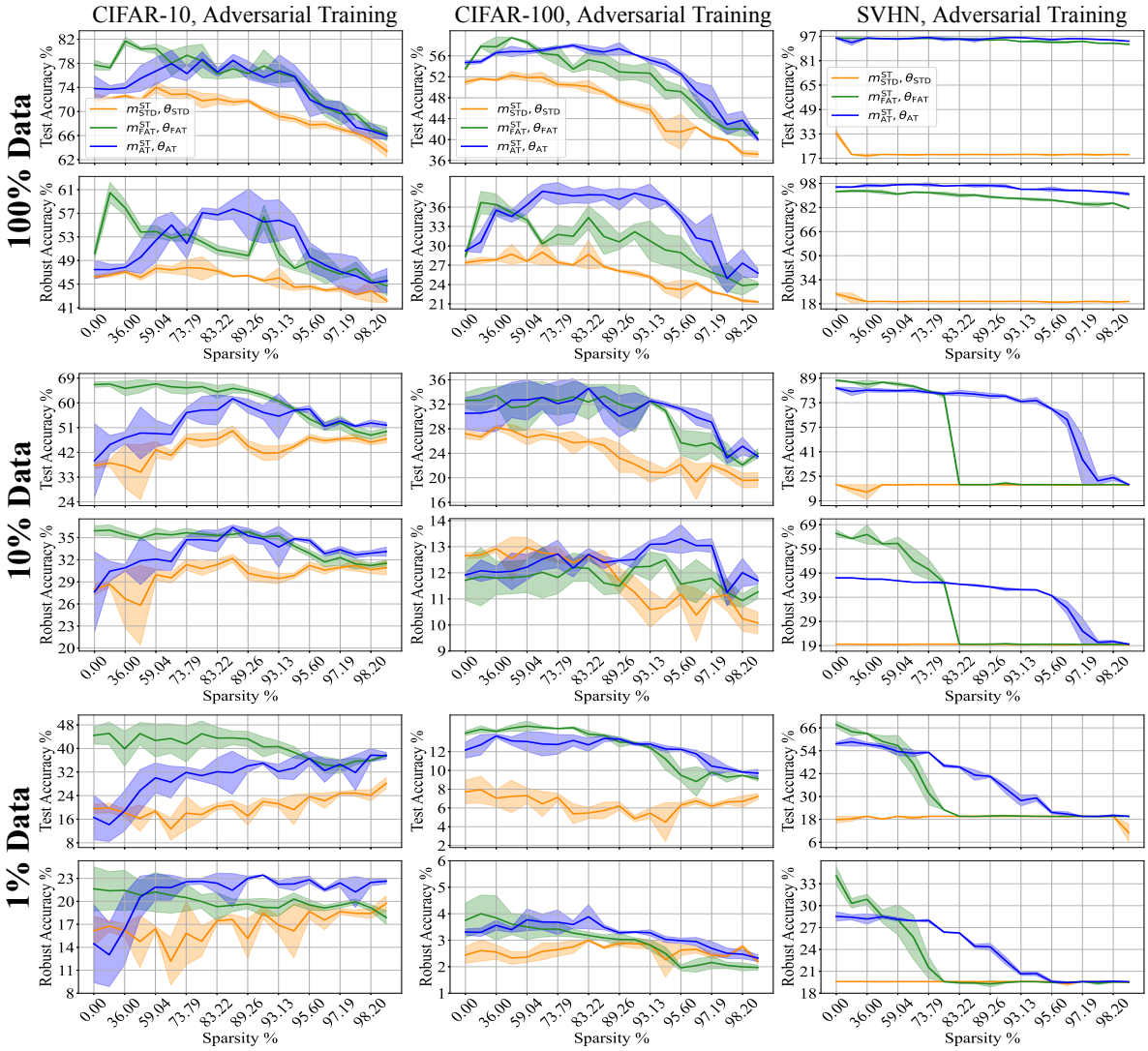

Figure 5. Data-efficient transfer results of double-win tickets from adversarial pre-training on three downstream datasets (i.e. CIFAR-10, CIFAR-100, and SVHN) with 100%, 10% and 1% training data. Both standard and robust accuracy are reported. Orange, Green and Blue represent the performance of subnetworks located from IMP together with standard training on ImageNet classification (m^{ST}) with different pre-trained weights (i.e. standard θ_{STD}, fast adversarial θ_{FAT}, and adversarial pre-training θ_{AT}, respectively). The solid line and shading area are the mean and standard deviation of standard/robust accuracy.

4.2. Do training regimes on source domain affect the located subnetworks?

During the ticket finding on pre-trained tasks, we can adopt standard re-training and adversarial re-training after each IMP pruning process. Intuitively, adversarial re-training should be able to maintain more information from adversarial pre-training, and lead to better transfer performance on downstream tasks (Salman et al., 2020). However, our experiment results surprisingly challenge this "common sense". Specifically, we choose the adversarial pre-trained weight (θ_{AT}[5]) as the initialization and compare four types of pruning and re-training methods on the pre-trained tasks, i.e., IMP with standard training ($m^{\text{ST}}_{\text{AT}}$), IMP with adversarial training ($m^{\text{AT}}_{\text{AT}}$), one-shot magnitude pruning (OMP) (m^{OMP}) and random pruning (m^{RP}).

As shown in Figure 4, the extreme sparsity of double-win lottery tickets on {CIFAR-10, CIFAR-100} is (91.41%, 83.22%), (91.41%, 83.22%), (89.26%, 79.03%), (20%, 0%) for $(m^{\text{AT}}_{\text{AT}}, \theta_{\text{AT}})$, $(m^{\text{ST}}_{\text{AT}}, \theta_{\text{AT}})$, $(m^{\text{OMP}}, \theta_{\text{AT}})$ and $(m^{\text{RP}}, \theta_{\text{AT}})$, respectively. IMP with standard and adversarial training shows similar performance, and both of them are better than OMP and random pruning. It suggests that the re-training regimes during IMP doesn't make an significant impact on the downstream transferablity of subnetworks. Due to the heavy computational cost of adversarial training

[5]The official robust model (ℓ_∞ adversary with $\epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$) on ImageNet zoo at https://github.com/MadryLab/robustness.

and the inferior performance of OMP, we consider IMP with standard training as our major pruning method and investigate the data efficiency property of subnetworks in following sections.

5. Double-Win Tickets with Robust Pre-training Enables Data-Efficient Transfer

In this section, we further exploit the practical benefits of double-win lottery tickets by assessing the data-efficient transferability under limited training data schemes (e.g., 1% and 10%). All subnetworks are drawn from IMP with standard re-training on the pre-training task. We consider three different pre-trained weights (i.e., standard θ_{STD}, fast adversarial θ_{FAT}, and adversarial pre-training θ_{AT}). The results are included in Figure 5, from which we find:

① θ_{FAT} and θ_{AT} significantly outperform θ_{STD} on both data-rich and data-scarce transfer for all three datasets. It evidences that the robust pre-training improves data-efficient transfer. While on the challenging SVHN downstream dataset with limited training data (i.e., 10% and 1%), the performance of θ_{FAT} and θ_{AT} degrades to θ_{STD}'s level at large sparsity 83.22% and 97.19% respectively.

② For data-limited transferring, sparse tickets derived from robust pre-training $\{\theta_{\text{FAT}}$,$\theta_{\text{AT}}\}$ surpass their dense counterpart by up to $\{0.75\%$,$22.53\%\}$ SA and $\{0.79\%$,$8.97\%\}$ RA, which indicates the enhanced data-efficiency also comes from appropriate sparse structures. The consistent robustness gains under unseen transfer attacks in Section A2, also exclude the possibility of obfuscated gradients.

③ In general, when subnetworks are trained with only 10% or 1% training samples available, those drawn from fast adversarial pre-trained weight θ_{FAT} show superior performance at middle sparsity levels, with performance improvements up to $\{30.97\%$, 2.42%, $10.05\%\}$ SA and $\{8.36\%$, 0.70%, $18.49\%\}$ RA compared with θ_{AT} for CIFAR-10, CIFAR-100 and SVHN, respectively. But with the increase of sparsity, adversarial pre-trained weight θ_{AT} overtakes θ_{FAT} and dominates the larger sparsity range.

To understand the counter-intuitive results that subnetworks from weak robust pre-training θ_{FAT} perform better than the ones from strong robust pre-training θ_{AT} at middle sparsity levels such as 73.79% particularly for data-limited transferring, we visualize the training trajectories along with loss landscapes through tools in (Li et al., 2018). We take robustified subnetworks with 73.79% sparsity on CIFAR-10 as an example. As shown in Figure 6 and A9, for the results on the original test data (columns: a,b,c), the loss contour of θ_{FAT}

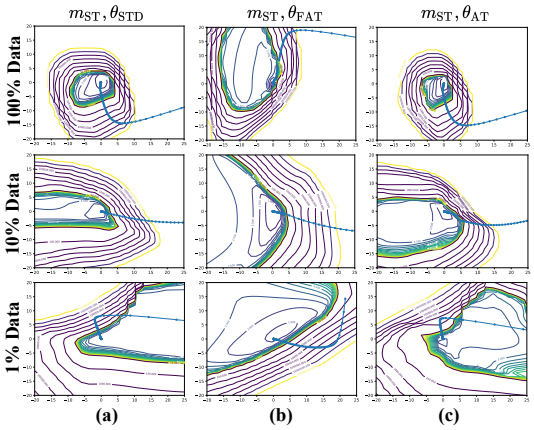

Figure 6. Visualization of loss contours and training trajectories of subnetworks located by IMP with standard re-training m^{ST} at 73.79% sparsity. Each subnetwork is adversarial trained with 100%, 10% or 1% training data on CIFAR-10. We compare three pre-training (i.e., standard θ_{STD}, fast adversarial θ_{FAT}, and adversarial pre-training θ_{AT}). The original test set is used.

is smoother/flatter than θ_{AT} and θ_{STD}, i.e., the basin with converged minimum has larger area in terms of the same level of loss like the 2.000 contour in the middle row's plots of Figure 6. A smoother/flatter loss surface is often believed to indicate enhanced standard (Keskar et al., 2017; He et al., 2019a) and robust generalization (Wu et al., 2020; Hein & Andriushchenko, 2017). It offers a possible explanation of θ_{FAT}'s superior performance to θ_{AT}'s by up to 8.98% and 9.62% SA improvements for data-limited transferring with 10% and 1% training samples. Moreover, loss geometric on attacked test data (Fig. A9) reveals similar conclusions.

6. Analyzing Properties of Double-Win Tickets

Relative similarity. In Fig. A8, we report the relative similarity (i.e., $\frac{|m_i \cap m_j|}{|m_i \cup m_j|}$) between binary pruning masks m_i and m_j, which denotes the degree of overlapping in sparse patterns located from different pre-trained models. We observe that subnetworks from different pre-training has distinct sparse patterns. Specifically, the relative similarity is less than 20.00% when the sparsity of subnetworks reaches 73.79% and the more sparsified, the larger differences arise.

Structural patterns. Meanwhile, we calculate the number of completely pruned (zero) kernels and visualize the kernel-wise heatmap of subnetworks with an extreme sparsity of 97.19%. As depicted in Figure 7, the subnetworks from the standard pre-trained model have the largest number of zero kernels, which roughly reveals the most clustered sparse patterns. And the subnetworks from robust pre-training are less clustered, especially for θ_{FAT}. We notice that these zero kernels are mainly distributed in the front/later residual blocks for subnetworks from θ_{AT}/θ_{FAT}, where they scatter evenly across all blocks. Typically, subnetworks with more zero kernels may have a stronger potential for hardware speedup (Elsen et al., 2020).

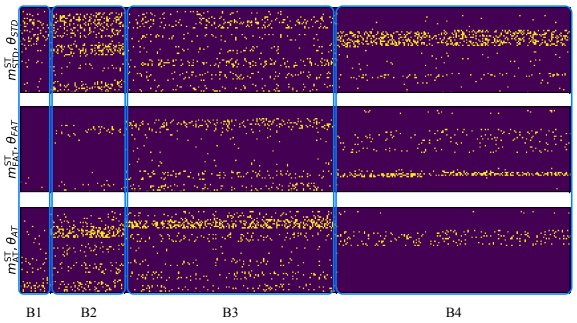

Figure 7. Kernel-wise heatmap visualizations of sparse masks drawn from three different pre-training, i.e., $m^{\text{ST}}_{\text{STD}}$, $m^{\text{ST}}_{\text{FAT}}$, and $m^{\text{ST}}_{\text{AT}}$ at 97.19% sparsity. The bright dots (●) are the completely pruned (zero) kernels and the dark dots (●) stand for the kernels with at least one remaining weight. B1 $\sim$ B4 represent the four residual blocks in ResNet-50.

7. Conclusion and Limitation

In this paper, we examine the lottery tickets hypothesis in a more rigorous and practical scenario, which asks for competitive transferability across both standard and adversarial downstream training regimes. We name these intriguing subnetworks as double-win lottery tickets. Extensive results reveal that double-win matching subnetworks derived from robust pre-training enjoy superior performance and enhanced data efficiency during transfer learning. However, the current investigations are only demonstrated in computer vision, and we leave the exploration in other fields such as natural language processing to future work.

Acknowledgment

Z.W. is in part supported by an NSF RTML project (#2053279).

References

Andriushchenko, M. and Flammarion, N. Understanding and improving fast adversarial training. *arXiv preprint arXiv:2007.02617*, 2020.

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Carbin, M., and Wang, Z. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. *arXiv preprint arXiv:2012.06908*, 2020a.

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Wang, Z., and Carbin, M. The lottery ticket hypothesis for pre-trained bert networks. *arXiv preprint arXiv:2007.12223*, 2020b.

Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L., and Wang, Z. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020c.

Chen, T., Cheng, Y., Gan, Z., Liu, J., and Wang, Z. Ultra-data-efficient gan training: Drawing a lottery ticket first, then training it toughly. *arXiv preprint arXiv:2103.00397*, 2021a.

Chen, T., Sui, Y., Chen, X., Zhang, A., and Wang, Z. A unified lottery ticket hypothesis for graph neural networks, 2021b.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Long live the lottery: The existence of winning tickets in lifelong learning. In *International Conference on Learning Representations*, 2021c.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021d. URL https://openreview.net/forum?id=qZzy5urZw9.

Chen, X., Cheng, Y., Wang, S., Gan, Z., Wang, Z., and Liu, J. Earlybert: Efficient bert training via early-bird lottery tickets. *arXiv preprint arXiv:2101.00063*, 2020d.

Chen, X., Zhang, Z., Sui, Y., and Chen, T. Gans can play lottery tickets too. In *International Conference on Learning Representations*, 2021e. URL https://openreview.net/forum?id=1AoMhc_9jER.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Desai, S., Zhan, H., and Aly, A. Evaluating lottery tickets under distributional shifts. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., and Anandkumar, A. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

Diffenderfer, J., Bartoldson, B., Chaganti, S., Zhang, J., and Kailkhura, B. A winning hand: Compressing deep networks can improve out-of-distribution robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

Dziugaite, G. K., Ghahramani, Z., and Roy, D. M. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Elsen, E., Dukhan, M., Gale, T., and Simonyan, K. Fast sparse convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14629–14638, 2020.

Evci, U., Pedregosa, F., Gomez, A., and Elsen, E. The difficulty of training sparse neural networks. *arXiv preprint arXiv:1906.10732*, 2019.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.

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

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*, 2019.

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.

Fu, Y., Yu, Q., Zhang, Y., Wu, S., Ouyang, X., Cox, D., and Lin, Y. Drawing robust scratch tickets: Subnetworks with inborn robustness are found within randomly initialized networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Gale, T., Elsen, E., and Hooker, S. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Gan, Z., Chen, Y.-C., Li, L., Chen, T., Cheng, Y., Wang, S., and Liu, J. Playing lottery tickets with vision and language. *arXiv preprint arXiv:2104.11832*, 2021.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gui, S., Wang, H., Yang, H., Yu, C., Wang, Z., and Liu, J. Model compression with adversarial robustness: A unified optimization framework. *Advances in Neural Information Processing Systems*, 32, 2019.

Guo, C., Rana, M., Cisse, M., and Van Der Maaten, L. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

He, H., Huang, G., and Yuan, Y. Asymmetric valleys: Beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 2553–2564, 2019a.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Girshick, R., and Dollár, P. Rethinking imagenet pre-training. In *Proceedings of the IEEE international conference on computer vision*, pp. 4918–4927, 2019b.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pp. 2266–2276, 2017.

Jiang, Z., Chen, T., Chen, T., and Wang, Z. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D., and Smelyanskiy, M. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 1778–1787, 2018.

Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 369–385, 2018a.

Liu, X., Li, Y., Wu, C., and Hsieh, C.-J. Adv-bnn: Improved adversarial defense through robust bayesian neural network. *arXiv preprint arXiv:1810.01279*, 2018b.

Liu, Z., Sun, M., Zhou, T., Huang, G., and Darrell, T. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.

Ma, H., Chen, T., Hu, T.-K., You, C., Xie, X., and Wang, Z. Good students play big lottery better. *arXiv preprint arXiv:2101.03255*, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Mehta, R. Sparse transfer learning via winning lottery tickets. *arXiv*, abs/1905.07785, 2019.

Morcos, A., Yu, H., Paganini, M., and Tian, Y. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. In *Advances in Neural Information Processing Systems 32*, 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Raghunathan, A., Steinhardt, J., and Liang, P. S. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pp. 10877–10887, 2018.

Renda, A., Frankle, J., and Carbin, M. Comparing rewinding and fine-tuning in neural network pruning. In *8th International Conference on Learning Representations*, 2020.

Rice, L., Wong, E., and Kolter, J. Z. Overfitting in adversarially robust deep learning. *arXiv preprint arXiv:2002.11569*, 2020.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? *arXiv preprint arXiv:2007.08489*, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, 2018.

Sehwag, V., Wang, S., Mittal, P., and Jana, S. Hydra: Pruning adversarially robust neural networks. *Advances in Neural Information Processing Systems*, 33:19655–19666, 2020.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3358–3369, 2019a.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3358–3369, 2019b.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Wang, C., Zhang, G., and Grosse, R. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgsACVKPH.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Wu, D., Wang, Y., and Xia, S.-t. Revisiting loss landscape for adversarial robustness. *arXiv preprint arXiv:2004.05884*, 2020.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

You, H., Li, C., Xu, P., Fu, Y., Wang, Y., Chen, X., Baraniuk, R. G., Wang, Z., and Lin, Y. 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.

Yu, H., Edunov, S., Tian, Y., and Morcos, A. S. Playing the lottery with rewards and multiple languages: lottery tickets in rl and nlp. In *8th International Conference on Learning Representations*, 2020.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Painless adversarial training using maximal principle. *arXiv preprint arXiv:1905.00877*, 2 (3), 2019a.

Zhang, D., Ahuja, K., Xu, Y., Wang, Y., and Courville, A. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pp. 12356–12367. PMLR, 2021a.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019b.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning*, pp. 11278–11287. PMLR, 2020.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks, 2021b. URL https://openreview.net/forum?id=8pz6GXZ3YT.

A1. More Implementation Details

To identify subnetworks in the pre-trained models, we consider both standard and adversarial re-training for IMP, in which we remove 20% parameters with the lowest magnitude for each pruning step and fine-tune the network for 30 epochs with a fixed learning rate of 5×10^{-4}. And we use an SGD optimizer with the weight decay and momentum kept to 1×10^{-4} and 0.9, respectively. The batch size equals 2048 for all experiments of IMP on the pre-training task with ImageNet. And for adversarial training, we apply PGD-3 with $\epsilon = \frac{8}{255}$ and $\alpha = \frac{2}{255}$ against the ℓ_∞ adversary.

A2. More Experiments Results

Table A2. Transfer attack performance of subnetworks located from adversarial pretraining $\theta_{\rm AT}$ through IMP with standard re-training $m^{\rm ST}$. And the subnetworks are trained with PGD-10 on CIFAR-10. We report the accuracy on attacked test sets, which are generated from an unseen robust model, together with the vanilla robust accuracy.

Sparsity (%)	Transfer Attack Accuracy (%)	Robust Accuracy (%)
0	55.54	47.11
89.26	62.79	56.00
93.13	60.09	60.05
95.60	56.49	50.85

Excluding obfuscated gradients. Table A2 demonstrates that the sparse subnetworks consistently outperform the dense counterpart under transfer attack from an unseen robust model, which is aligned with the vanilla robust accuracy. This piece of evidence excludes the possibility of gradient masking for our obtain RA improvements.

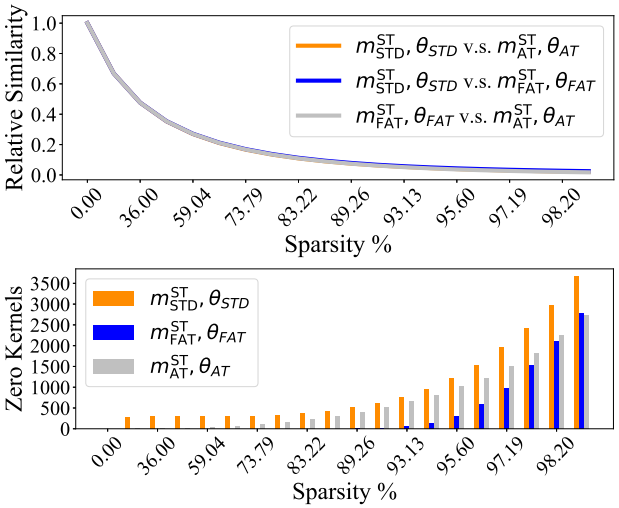

Figure A8. The statistic of subnetworks drawn from three different pre-training, i.e. $(m^{\rm ST}_{\rm STD}, \theta_{\rm STD})$, $(m^{\rm ST}_{\rm FAT}, \theta_{\rm FAT})$ and $(m^{\rm ST}_{\rm AT}, \theta_{\rm AT})$. (Top): The relative mask similarity between subnetworks from different pre-training. (Bottom): The number of completely pruned (zero) kernels in these subnetworks.

Relative similarity. To measure the overlapping level in sparse patterns drawn from different pre-trained models, we adopt the relative similarity (i.e., $\frac{|m_i \cap m_j|}{|m_i \cup m_j|}$) between binary pruning masks m_i and m_j. As shown in Fig. A8, subnetworks from different pre-training share remarkably heterogeneous sparse structures. For instance, the relative similarity is less than 20.00% when the sparsity of subnetworks reaches 73.79% and the more sparsified, the larger differences arise.

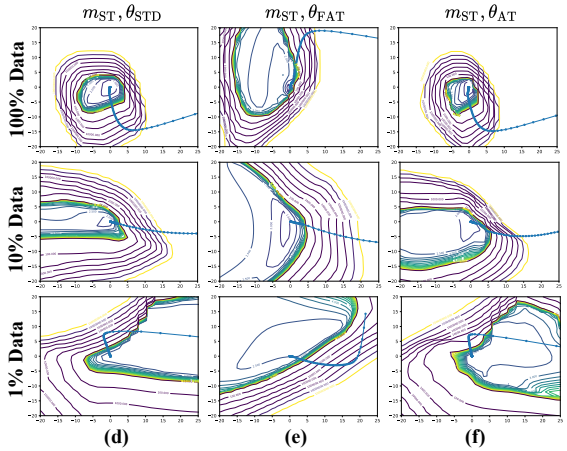

Figure A9. Visualization of loss contours and training trajectories of subnetworks located by IMP with standard re-training $m^{\rm ST}$ at 73.79% sparsity. Each subnetwork is adversarial trained with 100%, 10% and 1% training data on CIFAR-10, respectively. We compare three pre-training (i.e., standard $\theta_{\rm STD}$, fast adversarial $\theta_{\rm FAT}$, and adversarial pre-training $\theta_{\rm AT}$). Columns (d,e,f) stand for the results on attacked test data by PGD-20.

Extra loss surface visualizations. As shown in Figure A9, consistent observations with Figure 6 can be drawn.

More datasets and tasks. We conduct additional experiments of classification on (i) CUB-200 birds (more classes and higher resolution), (ii) VisDA17 ($4 \sim 5$ times bigger than CIFAR), and (iii) instance segmentation on VOC. Results of 60 sparse models are collected in Figure A10, showing consistent conclusion that robust pre-training helps.

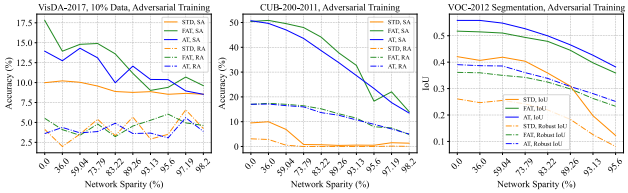

Figure A10. Results on more datasets and tasks.

A3. Boarder Impact

Although our work makes great contributions to efficient machine learning and security-critical applications, it still has potential negative social impacts when it is abused by malicious attackers. Specifically, our methods may speed up and robustify attackers' harmful algorithms or software. One possible solution is to issue a license and limit the blind distribution of our proposals.