Can You Win Everything with A Lottery Ticket?

Anonymous authors
Paper under double-blind review

Abstract

Lottery ticket hypothesis (LTH) has demonstrated to yield independently trainable and highly sparse neural networks (a.k.a. *winning tickets*), whose test set accuracies can be surprisingly on par or even better than dense models. However, accuracy is far from the only evaluation metric, and perhaps not always the most important one. Hence it might be myopic to conclude that a sparse subnetwork can replace its dense counterpart, even if the accuracy is preserved. Spurred by that, we perform the first comprehensive assessment of lottery tickets from diverse aspects beyond test accuracy, including *(i)* generalization to distribution shifts, *(ii)* prediction uncertainty, *(iii)* interpretability, and *(iv)* geometry of loss landscapes. With extensive experiments across datasets {CIFAR-10, CIFAR-100, and ImageNet}, model architectures, as well as tens of sparsification methods, we thoroughly characterize the trade-off between model sparsity and the all-dimension model capabilities. We find that an appropriate sparsity (e.g., $20\% \sim 99.08\%$) can yield the winning ticket to perform comparably or even better **in all above four aspects**, although some aspects (generalization to certain distribution shifts, and uncertainty) appear more sensitive to the sparsification than others. We term it as a `LTH-PASS`. Overall, our results endorse choosing a good sparse subnetwork of a larger dense model, over directly training a small dense model of similar parameter counts. We hope that our study can offer more in-depth insights on pruning, for researchers and engineers who seek to incorporate sparse neural networks for user-facing deployments. Codes are provided in the supplement.

1 Introduction

State-of-the-art pruning techniques are able to remove the majority of weights from deep neural networks (DNNs) almost without compromising the test accuracy (Mozer & Smolensky, 1989; Janowsky, 1989; LeCun et al., 1990b; Han et al., 2016; Guo et al., 2016; Molchanov et al., 2016). The emerging lottery ticket hypothesis (LTH) (Frankle & Carbin, 2019) advocates that dense models contain highly sparse subnetworks, i.e., *winning tickets*, with the same good trainability, expressiveness, and transferability (Morcos et al., 2019a; Chen et al., 2020b;a) compared to their dense counterpart. All these intriguing attributes together with the remarkable efficiency lead to a wide deployment of sparse networks in a resource-constrained real world (Lane & Warden, 2018). However, while many works narrowly refer to model "performance" as its test set accuracy, researchers have been long aware of the more complicated myriad of performance dimensions. Indeed, it remains elusive whether or not there are hidden pitfalls in a winning lottery ticket, besides the test accuracy versus efficiency, i.e., *have we missed or overlooked any unexpected loss along other performance dimensions when we prune a neural network?* This is the central question motivating the current work.

There were preliminary attempts done in earlier literature (Hooker et al., 2019; 2020b; Gui et al., 2019; Ye et al., 2019; Wang et al., 2018; Zhou et al., 2009; Venkatesh et al., 2020; Chen et al., 2020b;a; Koohpayegani et al., 2020; Morcos et al., 2019b; Zhang et al., 2021a; Sakamoto & Sato, 2022; Chen et al., 2022c) trying to address some part of this question. Some researchers advocated the existence of sparse subnetworks (winning tickets) with comparable transferability to the full dense models (Chen et al., 2020b;a; Koohpayegani et al., 2020; Morcos et al., 2019b; Chen et al., 2022a) and adversarial robustness (Chen et al., 2022b; Gui et al., 2019; Ye et al., 2019). Other recent works (Hooker et al., 2019; 2020b) pointed out that sparse networks are brittle to small changes such as natural image corruptions, and might amplify the class imbalance more than

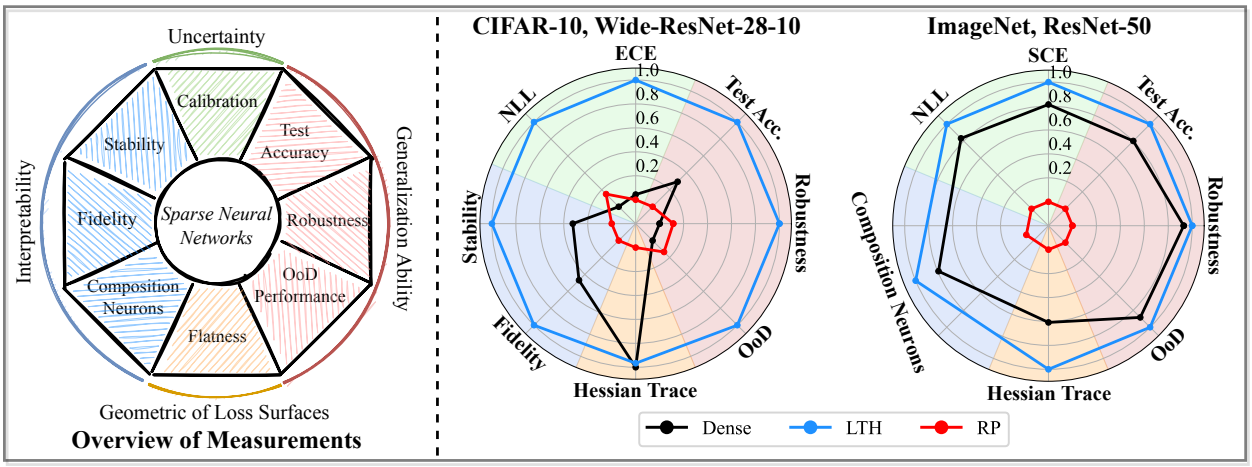

Figure 1: Overall assessments of sparse neural networks. *Left* figure presents an overview of measurements. *Right* figures show achieved full-scale performance, where **outer cycles indicate superior (sparse) networks**. **Dense** is the unpruned full model; LTH (Frankle & Carbin, 2019) denotes the winning ticket identified from the dense network, which is also a `LTH-PASS` here; RP represents a randomly pruned sparse model. All sparse networks on CIFAR/ImageNet only have 32.77%/51.20% parameters of the dense model. Associated with the *left* figure, {Test Accuracy, Robustness, OoD}, {Hessian Trace}, {Fidelity, Stability, Composition Neurons}, {NLL, ECE or SCE} are measurements for generalization ability, geometric of loss surfaces, interpretability, uncertainty respectively. Note that reported numbers of each metric are normalized by subtracting the minimum and dividing the gap between maximum and minimum values.

dense counterparts (Hooker et al., 2020a). Many other important aspects, such as uncertainty, interpretability ,and loss landscape, are not well studied as performance criteria in sparse neural networks, up to our best knowledge. In many applications, we cannot afford to pay any of them as the "hidden price" of sparsification. For example, robustness and interpretability are stipulated by safety-critical scenarios like autonomous cars and medical diagnostic, respectively. With those aspects under-scrutinized, it is hard to draw decisive conclusions on whether sparse winning tickets can become a drop-in replacement for dense ones, despite their appealing accuracy-efficiency trade-offs, and putting off their wider adoption with many open concerns. A comprehensive study into answering this question is thus highly demanded.

To the best of our knowledge, our work for the first time systematically characterizes and quantifies the full-dimension performance of sparse neural networks obtained by LTH and other pruning mechanisms. Specifically, we assess sparsity from its impacts on four carefully picked perspectives (Figure 1), including *generalization* to distribution shifts, *uncertainty* quantification, *interpretability*, and *loss geometry* that locally assess the learned functions. When an identified sparse subnetwork can be separately trained, to match all the above four aspects as the full dense model can do - we name it a `LTH-PASS`. Our contributions are outlined:

- ⋆ We define a more rigorous notion, `LTH-PASS`, which requires located subnetworks to match **all measured aspects** with their dense networks, i.e. with the same or even better ability to generalize to various shifted data distributions, to quantify uncertainty, to be interpreted (especially by neuron explanations), and to preserve learned functional approximation as indicated by loss landscape geometry, in addition to the unimpaired test accuracy. Such `LTH-PASS` is identified at $20\% \sim 99.08\%$ sparsity from diverse scenarios.
- ⋆ The excessive sparsification can often deteriorate any of the above aspects, and some aspects are far more sensitive to sparsity than others, e.g., the generalization to distribution shifts (e.g., with natural corruptions or adversarial perturbations) and the uncertainty quantification.
- ⋆ We also observe that it is advantageous to choose a sparse subnetwork (i.e., `LTH-PASS`) of a larger dense model, than directly using a small dense network with similar parameter counts (Figure A19), along all those performance dimensions. That implies the role of sparsity as a sophisticated, comprehensive regularization affecting multiple aspects of neural networks, rather than just ad-hoc reduction of parameters.
- ⋆ The above insights are drawn from extensive experiments across multiple datasets: CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009) and ImageNet (Deng et al., 2009); using diverse dense model architectures:

ResNets (He et al., 2016) and Wide-ResNets (Zagoruyko & Komodakis, 2016); as well as performing tens of representative sparsification regimes such as magnitude pruning (Han et al., 2016), lottery ticket hypothesis (Frankle & Carbin, 2019), random pruning, pruning at initialization (Lee et al., 2019; Wang et al., 2020; Tanaka et al., 2020), and dynamic sparse training (Evci et al., 2020; Liu et al., 2021b). We hope our benchmarking efforts to motivate more future studies as a common ground for comparison.

2 Related Works

2.1 Pruning, Lottery Ticket Hypothesis, and Dynamic Sparsity

Weight pruning can effectively eliminate redundancy in deep neural networks (LeCun et al., 1990b; Han et al., 2016) and obtain storage and computational savings. In general, it contains the following iterative cycles: *(a)* training the dense neural networks for at least several iterations; *(b)* removing unnecessary weights according to certain criteria and deriving subnetworks; *(c)* fine-tuning obtained sparse model to recover accuracy. Different sparsity patterns may be pursued, from unstructured (Han et al., 2015; LeCun et al., 1990a; Han et al., 2016) to structured sparsity (Liu et al., 2017; He et al., 2017; Zhou et al., 2016), the former being more flexible while the latter is often treated as more hardware friendly.

One of the mainstream pruning techniques is magnitude-based, which zeroes out a percentage of model weights by thresholding their magnitudes (Han et al., 2015; 2016). Later methods (Blalock et al., 2020) perform thresholding based on gradients (Molchanov et al., 2016; 2019) or hessian (LeCun et al., 1990a; Hassibi & Stork, 1992; Hassibi et al., 1993) based measures, instead of raw element magnitudes. The iterative pruning fashion (Han et al., 2016; Zhu & Gupta, 2017; Tan & Motani, 2020; Liu et al., 2019c) is often adopted for ameliorating performance degradation. Other pruning strategies formulate pruning as optimization objectives, by incorporating sparsity-promoting regularization (Liu et al., 2017; He et al., 2017; Zhou et al., 2016) or by constrained optimization (Boyd et al., 2011; Ouyang et al., 2013; He et al., 2017; Luo et al., 2017; Yu et al., 2018; Aghasi et al., 2017; Serra et al., 2020; ElAraby et al., 2020; Serra et al., 2021).

Lottery ticket hypothesis (LTH) (Frankle et al., 2019) recently emerges to investigate the independent trainable, extremely sparse neural networks from scratch, which are capable of recovering or even surpassing the original dense network's performance. Weight rewinding techniques (Renda et al., 2020; Frankle et al., 2020a) help scale up LTH for large networks and large-scale datasets. The intriguing properties of LTH received wide attention and have been broadly explored in various contexts, such as image classification (Frankle & Carbin, 2019; Liu et al., 2019b; Wang et al., 2020; Evci et al., 2019; Frankle et al., 2020b; Savarese et al., 2020; You et al., 2020; Ma et al., 2021; Chen et al., 2020a; 2022c;a), object detection (Girish et al., 2020), natural language processing (Gale et al., 2019; Yu et al., 2020; Prasanna et al., 2020; Chen et al., 2020b;c), generative adversarial networks (Chen et al., 2021e; Kalibhat et al., 2020; Chen et al., 2021a), graph neural networks (Chen et al., 2021b), reinforcement learning (Yu et al., 2020), and life-long learning (Chen et al., 2021c). Most of them leverage unstructured iterative magnitude pruning (Han et al., 2016; Frankle & Carbin, 2019) to identify the *winning tickets*, which we also follow in this work.

To save resources at training stages, SNIP (Lee et al., 2019), GraSP (Wang et al., 2020), and SynFlow (Tanaka et al., 2020) can be introduced to obtain high-quality sparse subnetworks before the training process starts, i.e., at the random initialization, based on several salience criteria. Another related field is dynamic sparse training (DST) (Mocanu et al., 2018; Liu et al., 2020) which trains sparse neural networks from scratch by optimizing the sparse connectivity and model parameters simultaneously. Numerous approaches (Mocanu et al., 2016; Evci et al., 2019; Mostafa & Wang, 2019; Dettmers & Zettlemoyer, 2019; Liu et al., 2021a; Dettmers & Zettlemoyer, 2019; Evci et al., 2020; Jayakumar et al., 2020; Raihan & Aamodt, 2020; Liu et al., 2021b) study such dynamic sparsity, often matching state-of-the-art training performance (Liu et al., 2021b).

2.2 Measurements of Sparse Neural Networks

Although the test set accuracy is often the core interest, more researchers (Hooker et al., 2019; 2020b; Gui et al., 2019; Ye et al., 2019; Wang et al., 2018; Zhou et al., 2009; Venkatesh et al., 2020; Chen et al., 2020b;a; Koohpayegani et al., 2020; Morcos et al., 2019b; Zhang et al., 2021a; Sakamoto & Sato, 2022; Chen et al.,

Table 1: Details of training configurations for experiments with OMP, LTH, RP, PI approaches.

Dataset	Learning Rate	Batch Size	Epochs	Optimizer	Momentum	Weight Decay
CIFAR-10/100	0.1; ×0.1 at 91,136 epoch	128	182	SGD	0.9	1×10^{-4}
ImageNet	0.4; ×0.1 at 30,60,80 epoch; linearly warmup 5 epochs	1024	90	SGD	0.9	1×10^{-4}

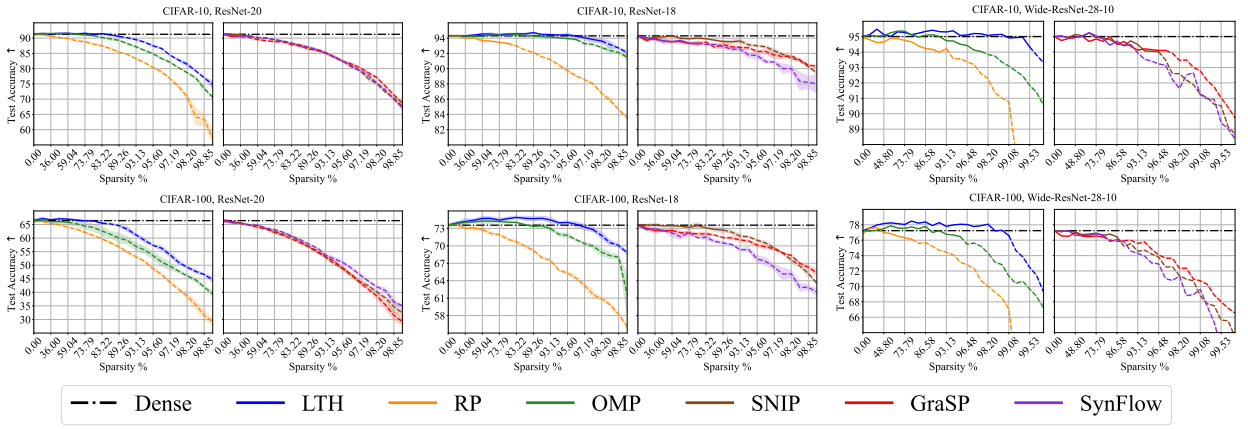

Figure 2: **Test accuracy** (↑) of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR datasets. The sparsity levels in X-axis are obtained from iteratively pruning with a ratio of 20%, i.e., $(1-0.8^n) \times 100\%$ and n is the number of pruning rounds. **Dense** (balck dashed lines) denotes the unpruned dense models. **↑/↓ indicate a better model should have a lager/smaller measurement**. Curves with errors (shadow regions) are the average across three independent runs, with standard deviations: same hereinafter. Each curve is divided into the region I (solid lines) of winning tickets; the region II of degraded subnetworks marked by dash lines. Regions I and II are separated by the extreme sparsity defined as the maximum sparsity when the subnetwork is at most 1% test accuracy drop compared to its dense counterpart. More detailed can be found in Table A3. The majority of our investigations are conducted on the high quality winning tickets from region I.

2022b) start to examine and characterize the impact of pruning from more perspectives beyond that. (1) Compression w.r.t. *fairness*: (Hooker et al., 2019; 2020b; Paganini, 2020) demonstrates compression may amplify existing algorithmic bias on the underrepresented long-tail of the data distribution, which is at odds with fairness objectives, and potentially results in disparate treatments of protected attributes (Zink & Rose, 2020). (2) Compression w.r.t. *robustness*: (Gui et al., 2019; Ye et al., 2019) show that with an appropriate sparsity, pruned subnetworks are capable of maintaining unimpaired adversarial robustness and standard accuracy. (Hooker et al., 2019; 2020b) tell a different story that compressed models are more sensitive and brittle to shifted data distributions such as natural corrupted samples (Hendrycks & Dietterich, 2019). (3) Compression w.r.t. *privacy*: (Wang et al., 2018; Zhou et al., 2009) enable sparse models to obtain a strong differential-privacy guarantee. (4) Compression w.r.t. *transferability*: extensive investigations (Chen et al., 2020b;a; Koohpayegani et al., 2020; Morcos et al., 2019b) indicate that there exist high quality subnetworks with competitive or even enhanced transferability across diverse datasets. (5) Compression w.r.t. *uncertainty*: (Venkatesh et al., 2020) integrates a suite of calibration strategies into existing pruning procedures, and locates reliable subnetworks with improved uncertainty.

3 Preliminary

Network. We use the official ResNet-20s (R20s), ResNet-18 (R18), ResNet-50 (R50) (He et al., 2016), and Wide-ResNet-28-10 (WR28-10) (Zagoruyko & Komodakis, 2016) as the original (unpruned) dense networks. $f(x;\theta)$ represents the output of a model with parameters $\theta \in \mathbb{R}^d$ and on input images x. Similarly, subnetworks extracted from the dense model θ can be depicted as $m \odot \theta$, where $m \in \{0,1\}^d$ is a pruning binary mask and $\odot$ denotes the element-wise product. Note that pruning is mainly conducted over networks without counting their classification heads.

Pruning Methods. To find the subnetworks $m \odot \theta$, we leverage several classical pruning approaches: (1) *one-shot magnitude pruning* (OMP) by removing a portion of weights with the globally smallest magnitudes (Han

et al., 2016); (2) *the lottery ticket hypothesis* (LTH) (Frankle & Carbin, 2019) with iterative weight magnitude pruning (IMP) (Han et al., 2016). Following the LTH's standard routines, it iteratively prunes the 20% of remaining weight with the globally smallest magnitudes and rewinds model weights to the same random initialization (Frankle & Carbin, 2019) or early training epochs (Frankle et al., 2020b; Chen et al., 2020a). In our case, weights are rewound to the $3_{\text{rd}}/15_{\text{th}}$ epoch for CIFAR and ImageNet experiments respectively, following the setups in Chen et al. (2020a); (3) *random pruning* (RP) which usually serves as a necessary baseline for the sanity check (Frankle & Carbin, 2019); (4) *pruning at initialization* (PI) mechanisms. Some representative methods, SNIP (Lee et al., 2019), GraSP (Wang et al., 2020), and SynFlow (Tanaka et al., 2020) are selected, which locates sparse subnetworks at random initialization via certain salience criterion; (5) *dynamic sparse training (DST)*. We choose the top-performing algorithm, RigL (Evci et al., 2020; Liu et al., 2021b), which starts from a random sparse network and encourages the connectivity to evolve dynamically based on a grow-and-prune strategy. All results and analyses about RigL are referred to Appendix A2.4.

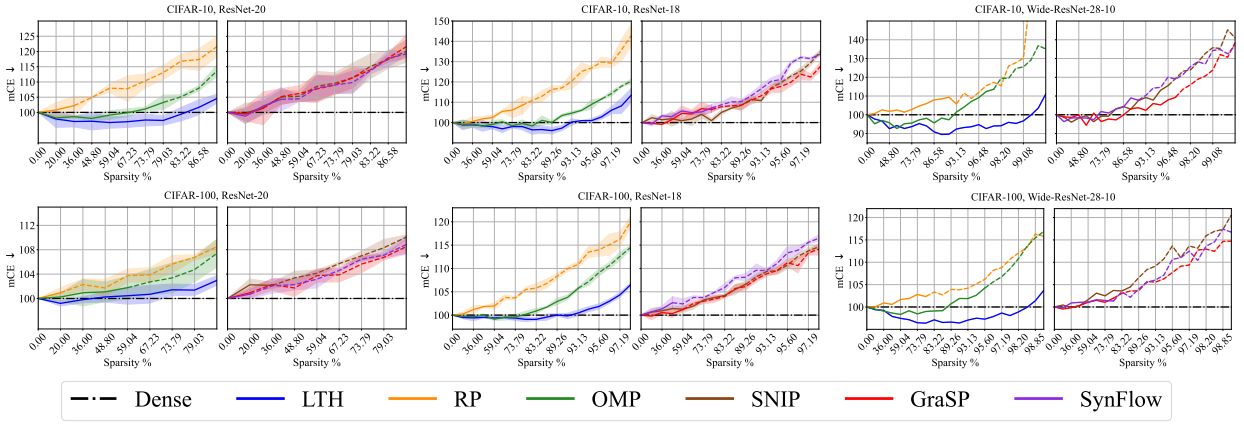

Figure 3: **Natural corruption robustness**, i.e. mCE (↓), of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR-10-C and CIFAR-100-C datasets.

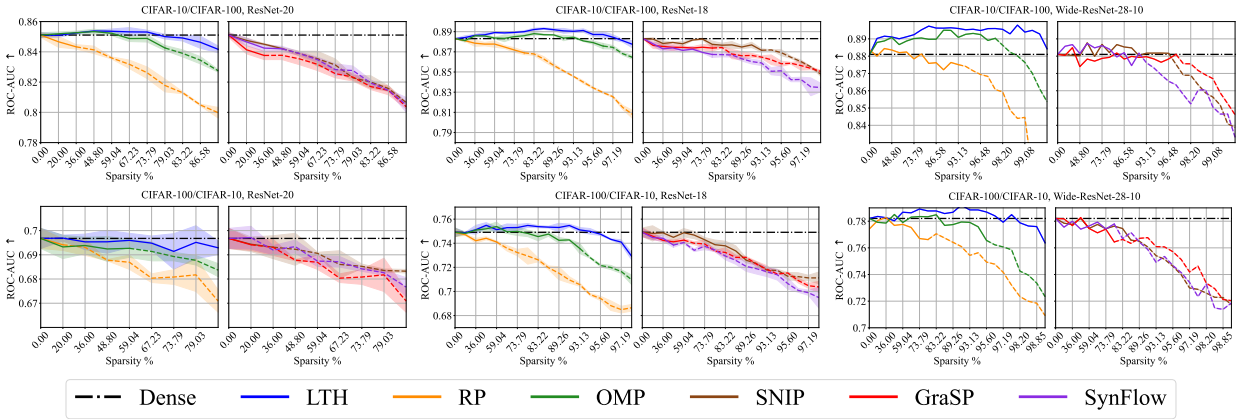

Figure 4: **Out-of-distribution** (OoD) detection performance, i.e., ROC-AUC (↑), of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR datasets. Titles are formed by in-distribution dataset / out-of distribution dataset, architecture.

Implementation details. Experiments are conducted on CIFAR-10 (C10), CIFAR-100 (C100) (Krizhevsky & Hinton, 2009), and ImageNet (IMG) (Deng et al., 2009). For a fair comparison, we follow the standard implementations and hyperparameters in (Renda et al., 2020) for OMP, LTH, RP, and PI experiments, as shown in Table 1. All RigL experiments follow the recent SOTA training configurations (Liu et al., 2021b). More details can be found in the Appendix A1.

4 What is Lost or Gained after Pruning?

In this section, we comprehensively investigate the full-dimension performance of sparse neural networks from LTH and other pruning algorithms, including (i) generalization to distribution shifts, (ii) uncertainty and reliability, (iii) interpretability, and (iv) geometry of loss landscapes. If a sparse subnetwork $m \odot \theta$ can be trained from the random initialization and match the dense model results in aspects $(i - iv)$, it is a `LTH-PASS`.

In summary, `LTH-PASS` broadly exists for diverse network architectures and datasets at $20 \sim 99.08\%$ *sparsity.*

4.1 Generalization to Distribution Shifts

The generalization ability of (sparse) neural networks is often considered equal to their training-test accuracy gap, while the test sets are i.i.d. selected from the same underlying distribution as the training set. While sparse neural networks can often achieve unimpaired test set accuracy, we broaden the scope of generalization ability by considering manipulated or shifted data distributions. Specifically, we will examine their generalization to *natural corruptions*, *adversarial perturbations* and *out-of-distribution (OoD) data* performance. The main takeaways are summarized below.

Takeaways: ❶ With appropriate sparsity levels, e.g., $59.04\% \sim 99.08\%$, subnetworks from LTH enjoy better generalization than its dense models, on (shifted) data distributions. ❷ Regardless of pruning methods, sparse neural networks are relatively more brittle to natural corruptions, compared to the other data distribution shifts.

Specifically, we quantify the generalization ability of network pruning in four main aspects:

• (Clean) generalization gap: $\frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{(x,y)\in\mathcal{D}_{\text{val}}}\delta(f(x;\theta)=y) - \frac{1}{|\mathcal{D}_{\text{train}}|}\sum_{(x,y)\in\mathcal{D}_{\text{train}}}\delta(f(x;\theta)=y)$, where $f(x;\theta)$ is the model's output and $\delta(\cdot)$ is the indicator function. $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$, x and y denotes the training data, testing data, input sample, and its corresponding label. Empirically, for well-trained models (i.e., $\sim$ zero training error), *the test set accuracy* is adopted to represent the generalization ability on original test sets, which is the conventional metric to evaluate the quality of sparse neural networks.

• Natural corruption robustness: $\text{mCE} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}[\sum_{s=1}^{5}\mathcal{E}_{s,c}^{m\odot\theta})/(\sum_{s=1}^{5}\mathcal{E}_{s,c}^{\theta})]$. Following the standard setup in Hendrycks & Dietterich (2019), we use the mean corruption error (mCE) to indicate model robustness to different natural corruptions, where $\mathcal{E}_{s,c}^{m\odot\theta}$ and $\mathcal{E}_{s,c}^{\theta}$ are the top-1 error of dense model θ and its sparse subnetwork $m \odot \theta$, respectively. $\mathcal{C}$ is the set of corruptions such as noise, blur, weather, digital process, and each corruption type $c \in \mathcal{C}$ has five corruption severity levels (i.e., $1 \leq s \leq 5$). Note that all corrupted samples are never shown in the training stage. CIFAR-10/100-C and ImageNet-C (Hendrycks & Dietterich, 2019) are adopted in our experiments. More details are included in Appendix A1.

• Adversarial robustness: testing accuracy on adversarial perturbed images, i.e., robust accuracy. We choose the classical adversarial attack, i.e., Fast Gradient Sign Method (FSGM) (Goodfellow et al., 2014), to generate adversarial samples as $x + \epsilon \times \text{sgn}(\nabla_x \mathcal{L}(f(x;\theta), y))$, where $\mathcal{L}$ is the empirical loss and ϵ (in our case, $\epsilon = \frac{8}{255}$) is the predefined magnitude of perturbations.

• Out-of-distribution (OoD) performance: ROC-AUC[1] as the standard metric is utilized to gauge obtained subnetworks. Since deep neural networks suffer from overconfident predictions on out-of-distribution data (Nguyen et al., 2015; Hendrycks & Gimpel, 2016), it is valuable to investigate whether this issue will be amplified or diminished by introduced model sparsity. Following Hendrycks & Gimpel (2016); Hendrycks et al. (2018); Hein et al. (2019); Augustin et al. (2020), for CIFAR-10 experiments, CIFAR-100 (Krizhevsky & Hinton, 2009) is regarded as the OoD dataset; for CIFAR-100 experiments, CIFAR-10 is selected as the OoD dataset; for ImageNet experiments, ImageNet-O (Hendrycks et al., 2019) is the OoD dataset.

Experimental observations. We present the results of **test accuracy**, **natural corruption robustness**, **OoD performance**, and **adversarial robustness** in Figure 2, 3, 4, and A15, respectively. Additional

[1]ROC-AUC stands for the area under the receiver operating characteristic (ROC) curve, in which we adopt the prediction confidence as the threshold.

results of (IMG, R50), (C10, VGG-11), and (C100, VGG-11) can be found in A12, A13, and A14, respectively. Several consistent findings can be drawn:

① *Superior sparse models?* Winning tickets broadly exist with unimpaired generalization on multiple (shifted) data distributions. Specifically, the matched or even outperformed performance can be achieved by winning tickets at sparsity $\{83.22 \sim 86.58\%, 93.13\% \sim 98.20\%, 94.50\% \sim 99.53\%, 67.23\% \sim 83.22\%, 93.13\% \sim 97.75\%, 98.20\% \sim 99.08\%, 59.04\% \sim 89.26\%\}$[2] on the original, corrupted, adversarial perturbed, out-of-distribution data. We see the extreme sparsity of winning tickets alters substantially given different evaluation metrics like natural corruption robustness, which suggests the limitation of considering the clean test set accuracy as the only quality measurement for pruned subnetworks.

② *Sensitivity metrics?* In general, natural corruption robustness (mCE) is the most sensitive measure to pruning since found winning tickets via mCE have smaller extreme sparsity. It suggests excessively pruned models are relatively more fragile to natural corruptions, such as blur, noise, fog, etc., which coincides with the findings in Hooker et al. (2020b).

③ *Data or model dependent?* Regarding to investigated generalization on various (shifted) data distributions, on the same dataset, more overparameterized models (e.g., WR28-10 v.s. R20s) are more amenable to be sparsified; with the same network, dataset contains more classes (e.g., C100 v.s. C10) is more intractable for pruning. Similar observations also presented in Morcos et al. (2019a).

④ *Superior pruning methods?* With low sparsity levels (e.g., $\leq 48.80\%$), OMP appears comparable generalization ability to LTH, and all PI algorithms perform no better than random pruning. At higher sparsity levels, although PI especially GraSP shows moderate advantages of generalization on the original test set and out-of-distribution data, subnetworks from PI are similarly vulnerable to corrupted or perturbed samples as randomly pruned models.

4.2 Uncertainty and Reliability

Confidence calibration uncovers the prediction uncertainties and the model reliability (Guo et al., 2017; Quiñonero-Candela et al., 2006; DeGroot & Fienberg, 1983; Venkatesh et al., 2020). It advocates the classification models must not only be accurate but also should reveal the true correctness likelihood (i.e., when the predictions are likely to be incorrect) (Guo et al., 2017), especially in safety/security-critical scenarios like self-driving vehicles (Bojarski et al., 2016) and automated health care (Jiang et al., 2012). In this section, we investigate whether pruning hurts or benefits confidence calibration. The main takeaways are summarized below.

Takeaways: ❶ Sparse networks from LTH with $48.80\% \sim 99.53\%$ sparsity, are capable of maintaining or enhancing both generalization and uncertainty performance, compared to dense models. ❷ In general, uncertainty measures are more sensitive to pruning than generalization metrics.

Several classical and representative evaluation metrics (Guo et al., 2017; Venkatesh et al., 2020) are used in our experiments:

• Expected calibration error (ECE): $\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$, where n and $|B_m|$ is the number of total samples and samples in the bin B_m, respectively. ECE (Pakdaman Naeini et al., 2015) is a widely adopted metric to approximate the difference in expectation between confidence and accuracy (i.e., miscalibration). Specifically, it partitions predictions into M equally-spaced bins, and then calculate a weighted average of the accuracy/confidence discrepancy in each of these bins.

• Static calibration error (SCE): $\mathrm{SCE} = \frac{1}{nC} \sum_{c=1}^{C} \sum_{m=1}^{M} |\sum_{i \in B_m} \mathbf{1}(y_i = c) - \mathrm{conf}(B_m)|$, where C is the total number of classes and y_i denotes the label of sample i. SCE bins the predictions separately for each class probability. Unlike ECE that only considers the highest probability, SCE (Gweon & Yu, 2019) treats all probabilities in a multi-class regime equally. We adopt it for ImageNet experiments with $1,000$ classes.

[2]Results are produced by the configurations of {(C10,R20s), (C10,R18), (C10,WR28-10), (C100, R20s), (C100,R18), (C100,WR28-10), (IMG,R50)}. Such narrative style is adopted hereinafter.

• Negative log likelihood (NLL): NLL $= -\sum_{(x,y)\in\mathcal{D}_{\text{val}}} \log(\hat{p}(y|x))$ as another standard measure of the calibration quality (Hastie et al., 2001; LeCun et al., 2015), is minimized in expectation if and only if the prediction distribution $\hat{p}(Y|X)$ recovers the ground truth conditional distribution $p(Y|X)$.

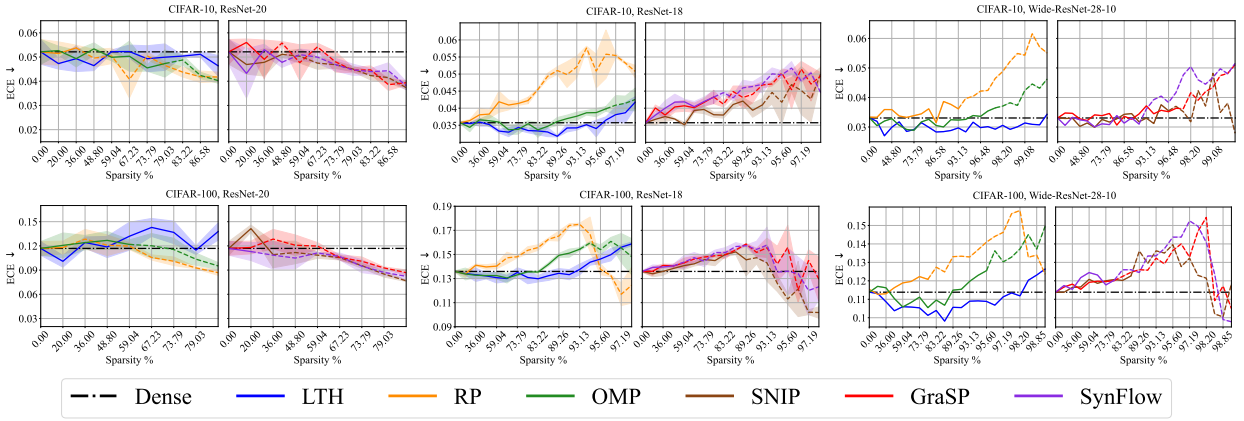

Figure 5: **Expected calibration error** (ECE ↓) of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR datasets. More results can be found in Figure A13 and A14.

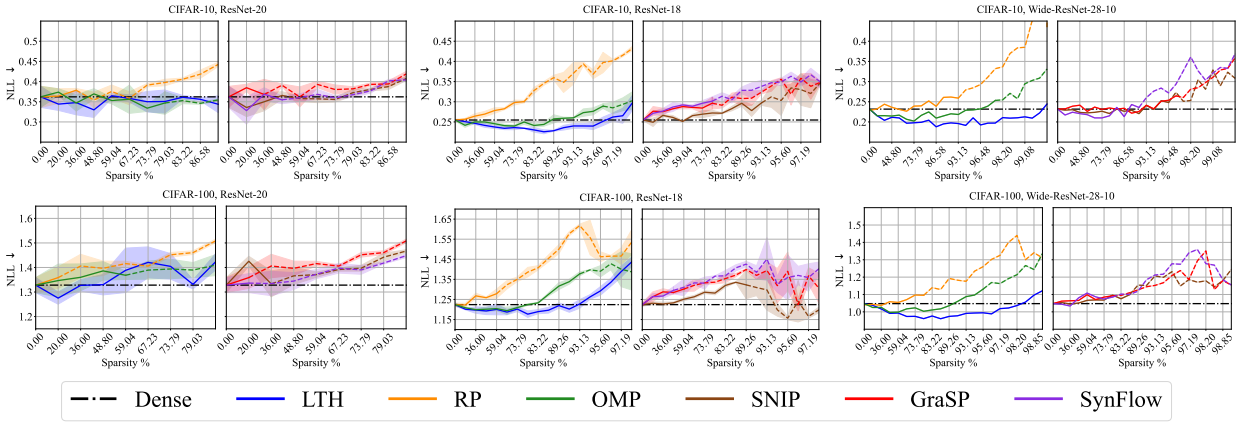

Figure 6: **Negative log likelihood** (NLL ↓) of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR datasets, which is normalized by dense networks.

Experimental observations. Main results of **expected calibration error** and **negative log likelihood** are presented in Figure 5 and 6. More results of (IMG, R50), (C10, VGG-11), and (C100, VGG-11) are included in Figure A12, A13, and A14 respectively. We observe that:

① *Superior sparse models?* Winning tickets in terms of uncertainty metrics can be found at sparsity {89.26%, 96.48 ∼ 97.75%, 99.53%, 48.80%, 93.13%, 98.20%, 48.80% ∼ 73.09%}.

② *Sensitivity metrics?* Generally, ECE and NLL metrics show a similar sensitivity to pruning. While uncertainty measures are more sensitive than generalization measures, it is within expectation since investigated pruning algorithms are mainly designed to avoid generalization drops.

③ *Superior pruning methods?* At high sparsity ratios (e.g., $\geq 67.23\%$), RP and PI have outperformed ECE than LTH in cases (C10/C100,R20s), yet at the price of high degraded generalization ability. Similar things happened in exorbitantly sparsified models in cases (C100,R18/WR28-10).

4.3 Interpretability

In this section, we investigate whether and which sparse neural networks are able to maintain the interpretability, compared to their dense counterpart. We assess these subnetworks from both macro and micro views. The former quantitatively evaluates the explainability from the functional representation perspective,

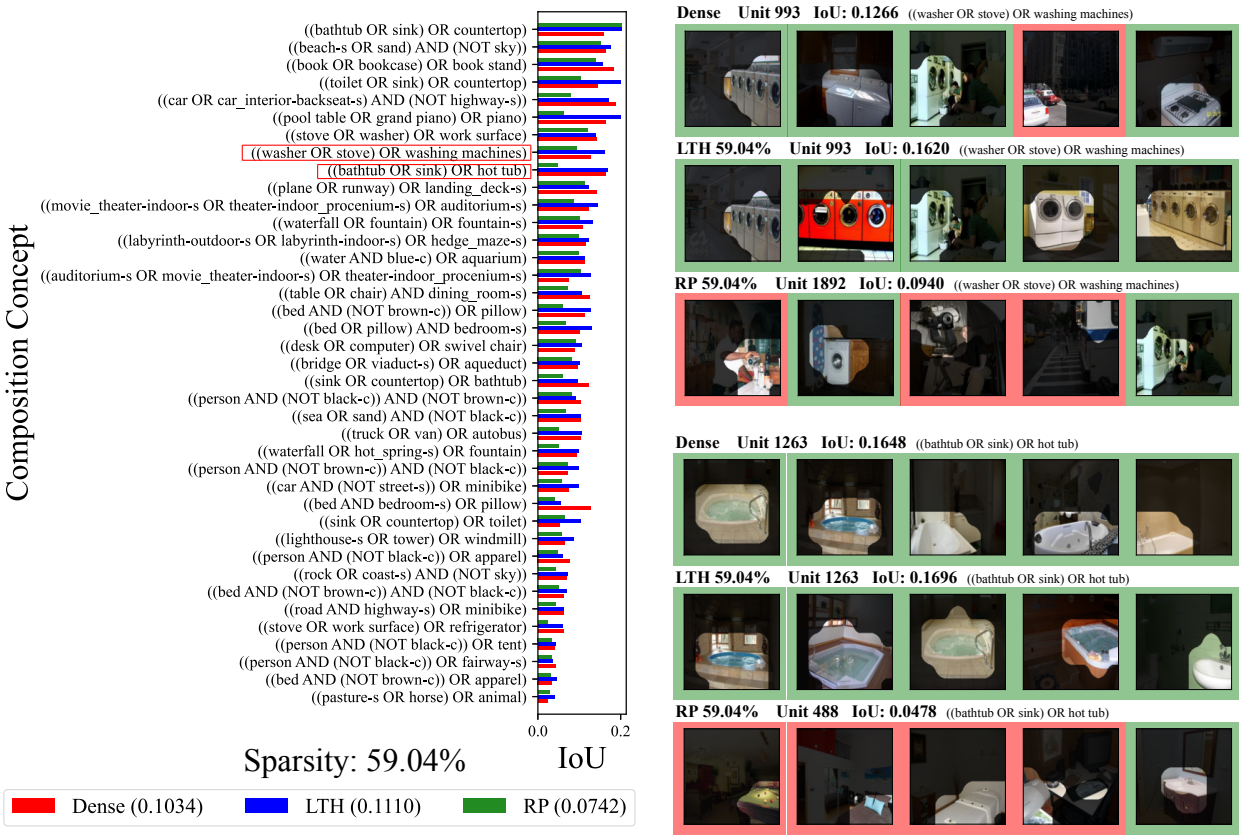

Figure 7: Results of **composition neurons**. (*Left*) Maximum IoUs (↑) of the intersection compositional concepts (top 70) by dissecting dense (ave. IoU: 0.1034), winning tickets (ave. IoU: 0.1110), and randomly pruned networks (ave. IoU: 0.0742) at the 59.04% sparsity. (*Right*) Neuron explanations for two of top IoU concepts in the left figure's red box. For each concept and neuron, top-5 IoU samples are presented. Green border indicates objects are **coincided** with the concept. Red border means they are **unrelated**.

with the most commonly used metrics, i.e., *fidelity* (Plumb et al., 2020; 2018; Ribeiro et al., 2016) and *stability* (Plumb et al., 2020; Alvarez-Melis & Jaakkola, 2018; Ghorbani et al., 2019). The latter performs the NetDissect procedure (Bau et al., 2017) for explaining neurons' behavior by identifying compositional logical concepts (Mu & Andreas, 2020). The main takeaways lie below.

Takeaways: ❶ Although LTH has superior generalization and uncertainty performance than other pruning methods, it is much more difficult to be interpreted by linear explainers. ❷ When we dissect sparse neural networks, LTH shows a significantly enhanced interpretability in terms of neuron behaviors, even compared to its dense counterpart.

• Fidelity and stability: $\mathcal{F} = \mathbb{E}_{x\in\mathcal{D}_{\text{val}}}[\mathbb{E}_{x'\sim\mathcal{N}_x}[(g(x') - f(x'))^2]], \mathcal{S} = \mathbb{E}_{x\in\mathcal{D}_{\text{val}}}[\mathbb{E}_{x'\sim\mathcal{N}_x}[\|e(x,f) - e(x',f)\|_2^2]]$. *Fidelity* $\mathcal{F}$ and *stability* $\mathcal{S}$ focus on local explanations for semantic features, which attempts to predict how the model's output would change if the input samples were perturbed. Following the classical routines in LIME (Ribeiro et al., 2016) to compute the metrics, we first perturb each input images x and build its neighborhood set $\mathcal{N}_x$ with a size of $1,000$ samples. Then, we generate a class of interpretable functions $\mathcal{G} := \{g_x \in \mathcal{G} | x \in \mathcal{D}_{\text{val}}\}$, where g_x is a linear function obtained form a regression to the corresponding model's output on $\mathcal{N}_x$. In the above formulation, $f(\cdot)$ denotes the target model we want to interpret. $e(x,f), e(x',f)$ are the learned weights of linear models g_x and $\tilde{g}_x$. Both g_x and $\tilde{g}_x$ are trained on $\mathcal{N}_x$, while each training sample $\hat{x} \in \mathcal{N}_x$ is weighted by the Hamming distance of $(\hat{x},x)$ and $(\hat{x},x')$ (Ribeiro et al., 2016), respectively. All our experiments follows the same implementation in (Plumb et al., 2020). Intuitively, $\mathcal{F}$ quantifies how accurately the explainer g_x models the target network f in a neighborhood $\mathcal{N}_x$; $\mathcal{S}$ measures the degree to which the explanation changes across points in $\mathcal{N}_x$.

• Composition neurons. Following (Mu & Andreas, 2020), we consider each individual neuron $f_n(x) \in \mathbb{R}$ of the model's output $f(x) \in \mathbb{R}^{d_2}$, and its activation on concrete input images. A good explanation of neuron f_n is a description (e.g., category or property) which locates the same inputs for which f_n activates. Specifically, we search the most appropriate *compositional concepts* $\mathcal{K}$ with the largest IoU (i.e., the intersection over union) for neuron f_n, i.e., $\text{IoU}(n, \mathcal{K}) = [\sum_x \mathbb{1}(M_n(x) \wedge \mathcal{K}(x))]/[\sum_x \mathbb{1}(M_n(x) \vee \mathcal{K}(x))]$, where $M_n(x)$ is a binary mask generated by thresholding the continuous neurons activation of $f_n(x)$. $\mathcal{K}$ consists of several pre-defined atomic concepts $\mathcal{K}_i$ ($1 \leq i \leq t$, t is 3 in our experiments) from the ADE20k (Zhou et al., 2017) and Broden (Bau et al., 2017) datasets. Each atomic concept is an image segmentation mask, and can be combined via disjunction (OR), conjunction (AND), and negation (NOT) operations, as shown in Figure 7. All procedures of our experiments follow the default configuration in (Bau et al., 2017; Mu & Andreas, 2020), and models for NetDissect are (sparse) R50 trained on ImageNet.

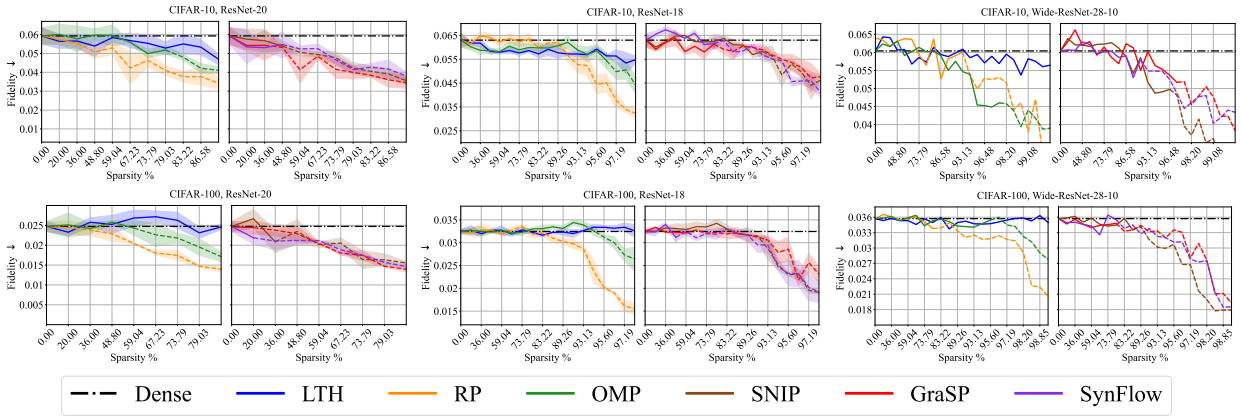

Figure 8: **Fidelity** (↓) of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR.

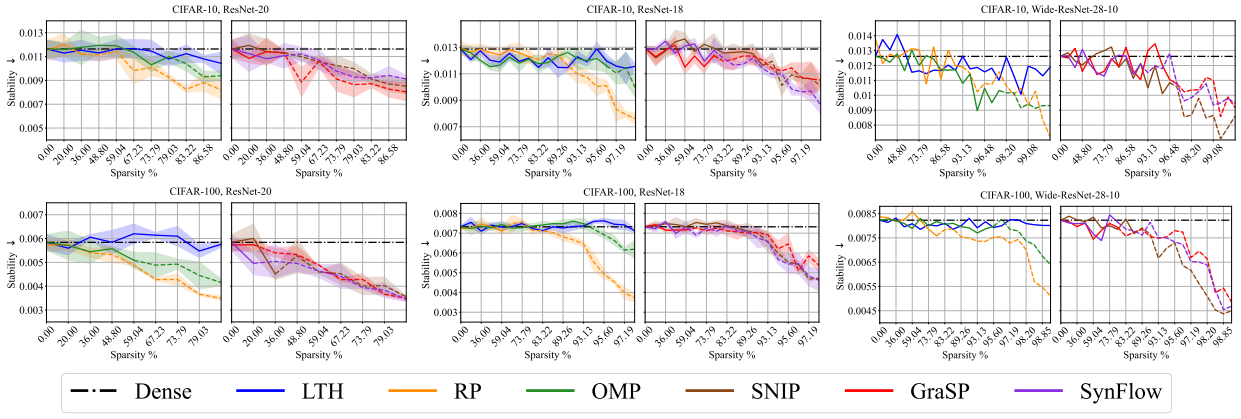

Figure 9: **Stability** (↓) of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR.

Experimental observations. According to main results in Figure 8, 9, and 7, we summarize important observations below. Additional experimental results are presented in Appendix A2. Figure A13 and A14 collect the evaluation on (C10, VGG-11) and (C100, VGG-11), respectively. Figure A16 and A17 display the extra composition neuron results of subnetworks at 67.23% sparsity. Figure A18 reports the IoU distribution of all 2048 neurons of diverse sparse neural networks.

① *Linear interpretability.* On CIFAR, we can locate winning tickets with a range of sparsity from 20% to 99.53%. Meanwhile, although with an inferior generalization, RP, OMP, and PI algorithms have consistently better linear interpretability. It suggests that the critical sparse topology mined by LTH potentially offers a more non-linear functional representation.

② *Neuron behaviors.* Based on NetDissect (Bau et al., 2017; Mu & Andreas, 2020) results in Figure 7 and A18, we find winning tickets (LTH) at 59.04% sparsity to achieve competitive generalization and neuron

interpretability, compared to the unpruned model. Coherently with the observations in (Mu & Andreas, 2020), when a neuron is active, the more generalizable the sparse network is, the more interpretable that neuron is (with a large IoU). Qualitative visual results in Figure 7 also imply the winning tickets from LTH succeed in finding more interpretable true positives than the corresponding dense model and randomly pruned networks.

4.4 Geometry of Loss Landscapes

The geometry of loss surfaces (e.g., *flatness*) reflects the learned functional approximation of derived subnetworks, which provides various insights to assess the sparse model's generalization ability and understand its behaviors such as transferability (Liu et al., 2019a). Some other works (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Jiang et al., 2019) show that the loss landscapes of well-generalizing models are relatively "flat" respect to model weights. Similarly, (Wu et al., 2020; Moosavi-Dezfooli et al., 2019; Chen et al., 2021d) claim that a flatter adversarial loss landscape with respect to model inputs enhances the robustness generalization. Our main takeaways can be summarized as follows.

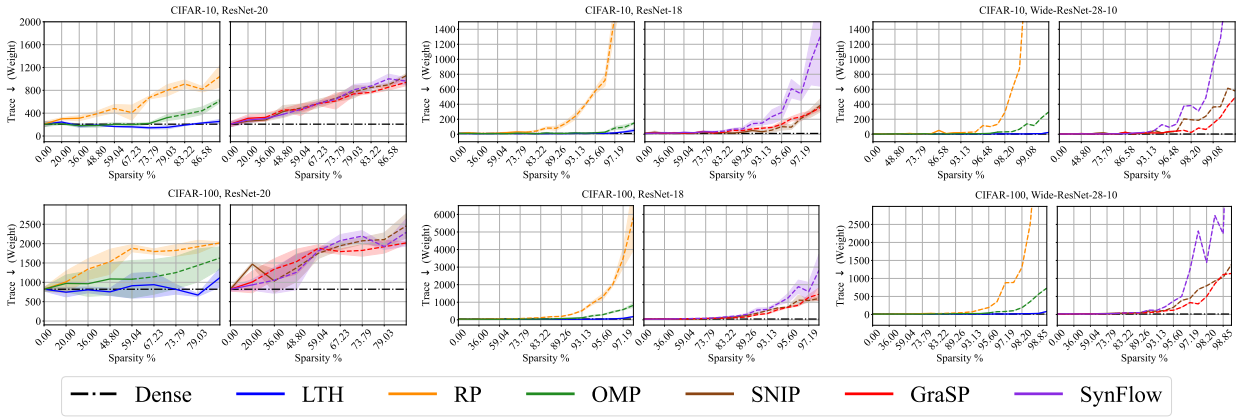

Figure 10: **Trace** (↓) for the weight flatness of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR.

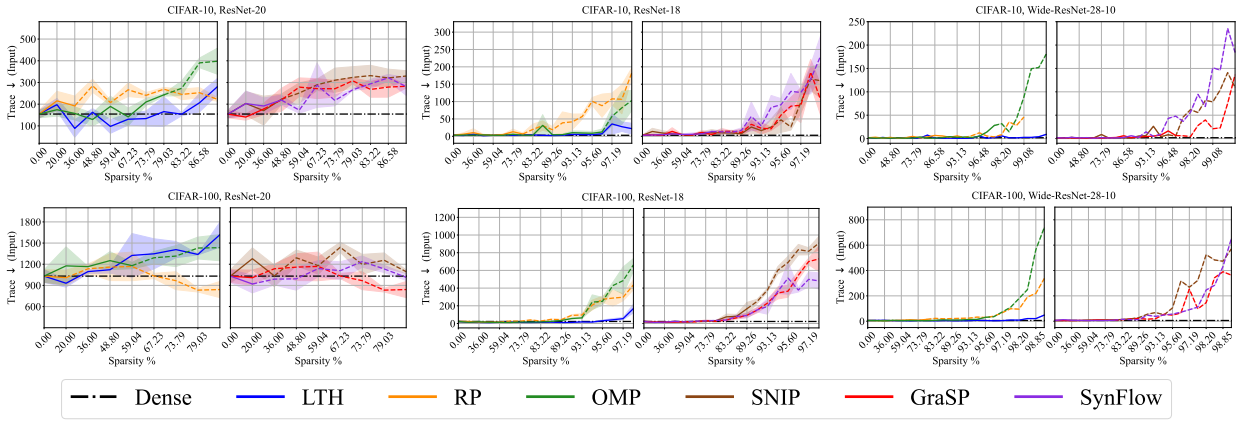

Figure 11: **Trace** (↓) for the input flatness of diverse subnetworks with a range of sparsity from 0.00% to 99.53% on CIFAR.

Takeaways: ❶ Winning tickets (LTH) with 20.00% ∼ 99.53% sparsity exist, achieving unscathed generalization ability, uncertainty, interpretability, and the loss landscape geometry. ❷ Besides pruning methods, network backbones and dataset scale also play non-negligible roles in the learned loss geometry of sparse neural networks.

Here we compute the Hessian trace via the PyHessian library (Yao et al., 2020) to measure the *weight/input flatness* of diverse sparse neural networks.

Experimental observations. Comprehensive results are collected in Figure 10 and 11, together with Figure A12, A13 and A14 which record additional results of (IMG, R50), (C10, VGG-11), and (C100, VGG-11) respectively. We observe:

① *Superior sparse models?* In terms of learned (local) functional approximation, winning tickets at sparsity {83.22% $\sim$ 86.58%, 96.48% $\sim$ 97.75%, 99.41% $\sim$ 99.53%, 20.00% $\sim$ 79.03%, 97.19%, 98.85%, 59.04% $\sim$ 93.13%} are exist.

② *Data or model dependent?* For the small backbone, e.g., R20s on CIFAR, different sparse neural networks from diverse pruning approaches have analogous flatness with respect to both weight and input spaces. On large-scale ImageNet with R50 in Figure A12, sparse subnetworks from LTH have much smaller Hessian traces of both model weights and input, which implies that LTH indeed identifies *flatter* local optimals (Zhang et al., 2021b).

③ *Superior pruning methods?* For R18 and WR28-10 on CIFAR, LTH, OMP, SNIP, and GraSP show better weight flatness than RP and SynFlow, which provides a possible explanation of their generalization performance gap. However, as for input flatness, randomly pruned networks surprisingly appear similar quality approximations, compared to dense and other sparsified models.

4.5 Extra Experimental Investigations

Extra network backbone. In this section, we evaluate an additional network architecture, VGG-11 (Simonyan & Zisserman, 2014), beyond ResNets. Results on CIFAR-10 and CIFAR-100 are collected in Figure A13 and A14 respectively. We observe that winning lottery tickets exist at (98.20%, 96.48%, 98.20%, 98.20%) and (96.48%, 89.26%, 96.48%, 95.60%) sparsity levels in terms of generalization ability (test accuracy, natural corruption robustness, adversarial robustness, out-of-distribution detection performance), (98.20%, 98.20%) and (86.58%, 94.50%) sparsity levels in terms of uncertainty (ECE, NLL), (98.20%, 98.20%) and (91.41%, 96.48%) sparsity levels in terms of interpretability (fidelity, stability), (98.20%) and (93.13%) sparsity levels in terms of loss surfaces' geometric, for (C10, VGG-11) and (C100, VGG-11) respectively. Therefore, `LTH-PASS` is identified in these two cases with the sparsity of 96.48% and 86.58%.

Comparison with a smaller dense network. As demonstrated in Figure A19, we observe the `LTH-PASS` at the 79% sparsity outperforms the small-dense baseline with similar parameter counts, by significant performance margins along all four evaluation dimensions. It suggests the sparsity functions as a comprehensive regularization and influences diverse aspects of neural networks, far beyond a simple reduction of network capacity (i.e., parameter counts).

5 Conclusion and Discussion of Broad Impact

In this paper, we perform an exhaustive screening test on the performance of various sparse neural networks, along diverse dimensions far beyond test-set accuracy. We believe that our compelling empirical results offer many in-depth insights of understanding network pruning, and endorse the wider adoption of (properly chosen) sparse neural networks in place of dense ones. Our future works will extend the similar screening to other model compression methods such as quantization.

We do not think this scientific research places a substantial risk of societal harm. The potential societal impact is that, with the assistance of our comprehensive assessment, it may be possible to establish accurate, robust, reliable, interpretable sparse networks with reduced energy and financial costs.

References

Alireza Aghasi, Afshin Abdi, Nam Nguyen, and Justin Romberg. Net-trim: Convex pruning of deep neural networks with performance guarantee. *Advances in Neural Information Processing Systems*, 30, 2017.

David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*, 2018.

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pp. 228–245. Springer, 2020.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

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

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. *arXiv preprint arXiv:2012.06908*, 2020a.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. *arXiv preprint arXiv:2007.12223*, 2020b.

Tianlong Chen, Yu Cheng, Zhe Gan, Jingjing Liu, and Zhangyang Wang. Ultra-data-efficient gan training: Drawing a lottery ticket first, then training it toughly. *arXiv preprint arXiv:2103.00397*, 2021a.

Tianlong Chen, Yongduo Sui, Xuxi Chen, Aston Zhang, and Zhangyang Wang. A unified lottery ticket hypothesis for graph neural networks, 2021b.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Long live the lottery: The existence of winning tickets in lifelong learning. In *International Conference on Learning Representations*, 2021c.

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

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Yang Zhang, Shiyu Chang, and Zhangyang Wang. Data-efficient double-win lottery tickets from robust pre-training, 2022a. URL https://arxiv.org/abs/2206.04762.

Tianlong Chen, Zhenyu Zhang, pengjun wang, Santosh Balachandra, Haoyu Ma, Zehao Wang, and Zhangyang Wang. Sparsity winning twice: Better robust generalization from more efficient training. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=SYuJXrXq8tw.

Tianlong Chen, Zhenyu Zhang, Yihua Zhang, Shiyu Chang, Sijia Liu, and Zhangyang Wang. Quarantine: Sparsity can uncover the trojan attack trigger for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 598–609, June 2022c.

Xiaohan Chen, Yu Cheng, Shuohang Wang, Zhe Gan, Zhangyang Wang, and Jingjing Liu. Earlybert: Efficient bert training via early-bird lottery tickets. *arXiv preprint arXiv:2101.00063*, 2020c.

Xuxi Chen, Zhenyu Zhang, Yongduo Sui, and Tianlong Chen. {GAN}s can play lottery tickets too. In *International Conference on Learning Representations*, 2021e. URL https://openreview.net/forum?id=1AoMhc_9jER.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983. ISSN 00390526, 14679884. URL http://www.jstor.org/stable/2987588.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Tim Dettmers and Luke Zettlemoyer. Sparse networks from scratch: Faster training without losing performance. *arXiv preprint arXiv:1907.04840*, 2019.

Mostafa ElAraby, Guy Wolf, and Margarida Carvalho. Identifying critical neurons in ann architectures using mixed integer programming. *arXiv preprint arXiv:2002.07259*, 2020.

Utku Evci, Fabian Pedregosa, Aidan Gomez, and Erich Elsen. The difficulty of training sparse neural networks. *arXiv preprint arXiv:1906.10732*, 2019.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning*, 2020.

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

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. The lottery ticket hypothesis at scale. *arXiv preprint arXiv:1903.01611*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020a.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=Hkl1iRNFwS.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

Sharath Girish, Shishira R Maiya, Kamal Gupta, Hao Chen, Larry Davis, and Abhinav Shrivastava. The lottery ticket hypothesis for object recognition. *arXiv preprint arXiv:2012.04643*, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Shupeng Gui, Haotao Wang, Haichuan Yang, Chen Yu, Zhangyang Wang, and Ji Liu. Model compression with adversarial robustness: A unified optimization framework. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Yiwen Guo, Anbang Yao, and Yurong Chen. Dynamic network surgery for efficient dnns. In *Advances in neural information processing systems*, pp. 1379–1387, 2016.

Hyukjun Gweon and Hao Yu. How reliable is your reliability diagram? *Pattern Recognition Letters*, 125:687–693, 2019. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2019.07.012. URL https://www.sciencedirect.com/science/article/pii/S016786551930203X.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1135–1143. Curran Associates, Inc., 2015.

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

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Babak Hassibi, David Stork, and Gregory Wolff. Optimal brain surgeon: Extensions and performance comparisons. *Advances in neural information processing systems*, 6, 1993.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

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*, pp. 770–778, 2016.

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*, 2017.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2020a.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020b.

Steven A Janowsky. Pruning versus clipping in neural networks. *Physical Review A*, 39(12):6600, 1989.

Siddhant Jayakumar, Razvan Pascanu, Jack Rae, Simon Osindero, and Erich Elsen. Top-kast: Top-k always sparse training. *Advances in Neural Information Processing Systems*, 33, 2020.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila OhnoMachado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 2012.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Neha Mukund Kalibhat, Yogesh Balaji, and Soheil Feizi. Winning lottery tickets in deep generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations, 2020.

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

Nicholas D. Lane and Pete Warden. The deep (learning) transformation of mobile and embedded computing. *Computer*, 51(5):12–16, 2018. doi: 10.1109/MC.2018.2381129.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*, pp. 598–605. Morgan-Kaufmann, 1990a. URL http://papers.nips.cc/paper/250-optimal-brain-damage.pdf.

Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pp. 598–605, 1990b.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. Snip: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1VZqjAcYX.

Hong Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Towards understanding the transferability of deep representations. *arXiv preprint arXiv:1909.12031*, 2019a.

Shiwei Liu, Decebal Constantin Mocanu, Amarsagar Reddy Ramapuram Matavalam, Yulong Pei, and Mykola Pechenizkiy. Sparse evolutionary deep learning with over one million artificial neurons on commodity hardware. *Neural Computing and Applications*, 2020.

Shiwei Liu, Decebal Constantin Mocanu, Yulong Pei, and Mykola Pechenizkiy. Selfish sparse rnn training. *arXiv preprint arXiv:2101.09048*, 2021a.

Shiwei Liu, Lu Yin, Decebal Constantin Mocanu, and Mykola Pechenizkiy. Do we actually need dense over-parameterization? in-time over-parameterization in sparse training. *arXiv preprint arXiv:2102.02887*, 2021b.

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*, pp. 2736–2744, 2017.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations*, 2019b.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019c.

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*, pp. 5058–5066, 2017.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Good students play big lottery better. *arXiv preprint arXiv:2101.03255*, 2021.

Decebal Constantin Mocanu, Elena Mocanu, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. A topological insight into restricted boltzmann machines. *Machine Learning*, 104(2-3):243–270, 2016.

Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9(1):2383, 2018.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11264–11272, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

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 32*, 2019a.

Ari S. Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers, 2019b.

Hesham Mostafa and Xin Wang. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *International Conference on Machine Learning*, 2019.

Michael C Mozer and Paul Smolensky. Using relevance to reduce network size automatically. *Connection Science*, 1(1):3–16, 1989.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *arXiv preprint arXiv:2006.14032*, 2020.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Hua Ouyang, Niao He, Long Tran, and Alexander Gray. Stochastic alternating direction method of multipliers. In *International Conference on Machine Learning*, pp. 80–88. PMLR, 2013.

Michela Paganini. Prune responsibly. *arXiv preprint arXiv:2009.09936*, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 04 2015.

Gregory Plumb, Denali Molitor, and Ameet Talwalkar. Model agnostic supervised local explanations. *arXiv preprint arXiv:1807.02910*, 2018.

Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems*, 33, 2020.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. When bert plays the lottery, all tickets are winning. *arXiv preprint arXiv:2005.00561*, 2020.

Joaquin Quiñonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 1–27, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Md Aamir Raihan and Tor M Aamodt. Sparse weight activation training. *arXiv preprint arXiv:2001.01969*, 2020.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Keitaro Sakamoto and Issei Sato. Analyzing lottery ticket hypothesis from pac-bayesian theory perspective. *arXiv preprint arXiv:2205.07320*, 2022.

Pedro Savarese, Hugo Silva, and Michael Maire. Winning the lottery with continuous sparsification. In *Advances in Neural Information Processing Systems 33 pre-proceedings*, 2020.

Thiago Serra, Abhinav Kumar, and Srikumar Ramalingam. Lossless compression of deep neural networks. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pp. 417–430. Springer, 2020.

Thiago Serra, Xin Yu, Abhinav Kumar, and Srikumar Ramalingam. Scaling up exact neural network compression by relu stability. *Advances in Neural Information Processing Systems*, 34, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Chong Min John Tan and Mehul Motani. DropNet: Reducing neural network complexity via iterative pruning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9356–9366. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/tan20a.html.

Hidenori Tanaka, Daniel Kunin, Daniel LK Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems 33 pre-proceedings*, 2020.

Bindya Venkatesh, Jayaraman J Thiagarajan, Kowshik Thopalli, and Prasanna Sattigeri. Calibrate and prune: Improving reliability of lottery tickets through prediction calibration. *arXiv preprint arXiv:2002.03875*, 2020.

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

Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and Philip S. Yu. Private Model Compression via Knowledge Distillation. *arXiv e-prints*, art. arXiv:1811.05072, November 2018.

Dongxian Wu, Yisen Wang, and Shu-tao Xia. Revisiting loss landscape for adversarial robustness. *arXiv preprint arXiv:2004.05884*, 2020.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 581–590. IEEE, 2020.

Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

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

Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S Davis. Nisp: Pruning networks using neuron importance score propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9194–9203, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pp. 12356–12367. PMLR, 2021a.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks, 2021b. URL https://openreview.net/forum?id=8pz6GXZ3YT.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017. doi: 10.1109/CVPR.2017.544.

Hao Zhou, Jose M Alvarez, and Fatih Porikli. Less is more: Towards compact cnns. In *European Conference on Computer Vision*, pp. 662–677. Springer, 2016.

Shuheng Zhou, Katrina Ligett, and Larry Wasserman. Differential privacy with compression. In *2009 IEEE International Symposium on Information Theory*, pp. 2718–2722. IEEE, 2009.

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.

Anna Zink and Sherri Rose. Fair regression for health care spending. *Biometrics*, 76(3):973–982, Jan 2020. ISSN 1541-0420. doi: 10.1111/biom.13206. URL http://dx.doi.org/10.1111/biom.13206.