Attend Who is Weak: Pruning-assisted Medical Image Localization under Sophisticated and Implicit Imbalances

Ajay Jaiswal[1], Tianlong Chen[1], Justin F. Rousseau[1], Yifan Peng[2], Ying Ding[1], Zhangyang Wang[1] ‡
[1]The University of Texas at Austin, [2]Weill Cornell Medicine

Abstract

Deep neural networks (DNNs) have rapidly become a de facto choice for medical image understanding tasks. However, DNNs are notoriously fragile to the class imbalance in image classification. We further point out that such imbalance fragility can be amplified when it comes to more sophisticated tasks such as pathology localization, as imbalances in such problems can have highly complex and often implicit forms of presence. For example, different pathology can have different sizes or colors (w.r.t.the background), different underlying demographic distributions, and in general different difficulty levels to recognize, even in a meticulously curated balanced distribution of training data. In this paper, we propose to use pruning to automatically and adaptively identify hard-to-learn (HTL) training samples, and improve pathology localization by attending them explicitly, during training in supervised, semi-supervised, and weakly-supervised settings. Our main inspiration is drawn from the recent finding that deep classification models have difficult-to-memorize samples and those may be effectively exposed through network pruning [15] - and we extend such observation beyond classification for the first time. We also present an interesting demographic analysis which illustrates HTLs ability to capture complex demographic imbalances. Our extensive experiments on the Skin Lesion Localization task in multiple training settings by paying additional attention to HTLs show significant improvement of localization performance by ∼2-3%.

1. Introduction

In the past decade, deep learning advancements have significantly influenced numerous medical imaging applications such as automated pathology diagnosis, detection, localization, and registration [27, 47, 16, 31, 30, 2, 17, 12, 19]. The success of these applications has motivated several researchers in the community to develop large-scale public datasets that can improve task performance. Although these real-world datasets have helped to build high-quality deep learning solutions, they usually suffer from class imbalance problems that can go way beyond the common perception of different training sample numbers across categories. Indeed, **imbalance in medical imaging datasets can have highly complex and subtle forms of presence**, e.g., different pathology can have different colors and sizes, and can be difficult to recognize even in meticulously curated training data. Additionally, it can further have many implicit imbalances based on gender, race, ethnicity, and demographics of individuals and can be very difficult to account for. While there have been numerous efforts to handle imbalance during DNNs training using data-level approaches such as oversampling, undersampling, synthetic sampling [23, 22, 3, 32, 33, 5], or cost-sensitive learning-based approaches [28, 44, 45, 39], these efforts primarily rely on the assumption of known class distribution and overlook the more complicated forms of imbalance. The generalization ability of DNNs can suffer significantly when data imbalance is overlooked during training, resulting in poor sensitivity towards minorities and substandard performance [40, 14]. Many works [43, 1, 29, 42, 10, 41] have recently observed that DNNs tend to prioritize learning simple patterns. More concretely, the DNN optimization is content-aware, taking advantage of patterns shared by more training examples, and therefore inclined towards memorizing the majority samples. Since minority samples are underrepresented in the training set, they tend to be *poorly-memorized*, and more prone to be *easily-forgotten* by the model. In the **context of image classification**, a recent empirical finding by [15] observed that network pruning which usually removes the smallest magnitude weight in a trained DNN, disproportionately impact various classes and samples, hurting poorly-memorized samples more. In simple words, minority samples are not "memorized well" and suffer significantly as a consequence of network pruning.

Inspired by this observation, in this paper, we attempt to ask an interesting question: *Can we identify instances which are difficult to memorize by DNNs and can be representative of complex and implicit imbalance?*. We for the **first time**, study network pruning impact on the spatial memorization/forgetting effect. We **go beyond** image classification to explore DNNs sensitivity towards *instance-level spatial region imbalance*, on the real-world skin lesion lo-

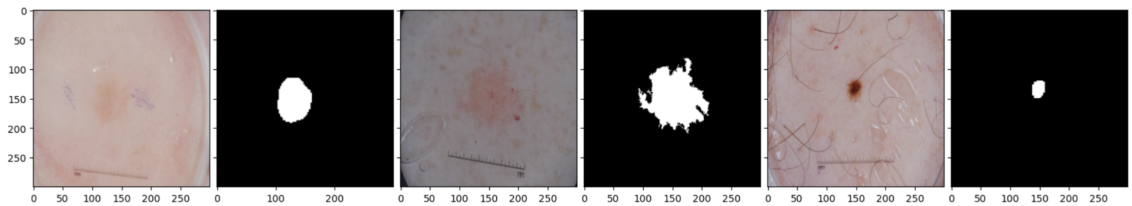

Figure 1. Localization as foreground vs background classification. Spatial Imbalance can be subtle and complex (eg. indistinguishable color, irregular shape, small size etc.).

calization task(S-LLT). Figure 1 illustrate the S-LLT task as foreground vs background classification. We observed that pruning of a trained localization model has a varying impression on spatial memorization, where it significantly impacts foreground performance while having a marginal impact on the background (Figure 2). Considering foreground as the representative of pathology and region of interest, we propose using drop-in localization (foreground) performance as a proxy to identify training instances which are poorly-memorized and can encode complex imbalance – we call them "***hard-to-learn***" (HTLs) due to their high sensitivity to pruning, and show that by explicitly and adaptively paying additional attention to them during training, we can achieve notable performance gain in their localization. Interestingly, an in-depth analysis of HTLs using demographic attributes such as *gender and age* reveals that **pruning impact some demographics more significantly than others**, diligently eliciting the complex and subtle imbalances in the data, going beyond class distribution. Additionally, our work demonstrates, for the first time, that pruning can elicit the demographic bias of trained models using real-world dataset instead of curated datasets such as CIFAR and CelebA. We selected S-LLT as our evaluation task (although it can be easily adapted to any localization task) considering the technique-friendly availability of ISIC 2017, 2018 datasets [7, 6], which will allow us to show the effectiveness of our technique in a supervised, semi-supervised, or unsupervised setting along with a detailed demographic study of HTLs. Our primary contributions can be summarized as:

- We propose **pruning** as an indicator to expose the spatial weakness of a trained localization model and show the existence of "*hard-to-learn*" training examples. For the first time, we reveal that pruning disproportionately impacts the foreground and background classes, where the foreground performance of some training examples can drop by a much larger margin than the background, indicating their high sensitivity to pruning.
- Tailored for the localization problem, we present **three novel HTL mining strategies** in the supervised, semi-supervised, and weakly-supervised settings using ground truth labels, pseudo-labels, and saliency maps respectively. We additionally show that by attending HTLs by fine-tuning, we can significantly improve localization performance.
- We have conducted extensive experiments and ablation studies to understand the specialty of HTLs on the S-LLT. Additionally, we provide an interesting **demographics analysis** of HTLs and illustrate our method's ability to capture complex implicit imbalances. Moreover, our extensive experiments show a significant and consistent **performance gain** of $\sim$2-3% IoU across different settings for S-LLT.

2. Methodology

2.1. Network pruning

The fundamental hypothesis behind the NN pruning is that DNNs are overparameterized, and a comparatively smaller network (sparse network) can be used to achieve a similar level of performance. Provided a dataset $D = \{(x_i, y_i)\}_{i=1}^n$, and a preferred sparsity level κ (i.e, number of non-zero weights), NN pruning can be written as a constrained optimization problem:

$$\min_w L(\mathbf{w}; D) = \min_w \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{w}; (x_i, y_i)), \quad (1)$$

$$s.t.\ \mathbf{w} \in \mathbb{R}^m, ||w||_0 \leq \kappa \quad (2)$$

where, $l(\cdot)$ is a standard loss function, $\mathbf{w}$ is a set of parameters of NN, m is the total number of parameters, and $||\cdot||_0$ is L_0 norm. The traditional approach to minimize the above equation is by adding sparsity enforcing penalty terms or saliency-based methods. Saliency-based methods solve the above equation by removing redundant parameters in the NN using a good criterion. Popular criterion includes magnitude-based weight pruning (i.e., weight below a certain threshold is redundant) [11, 9], or hessian of loss wrt. weights (i.e., the higher the value of hessian, the higher the parameter importance) [25, 13].

In this work, we have used magnitude-based unstructured weight pruning due to its simplicity and keep the focus on the "*forgetting*" behavior of sparse NNs [15]. To avoid

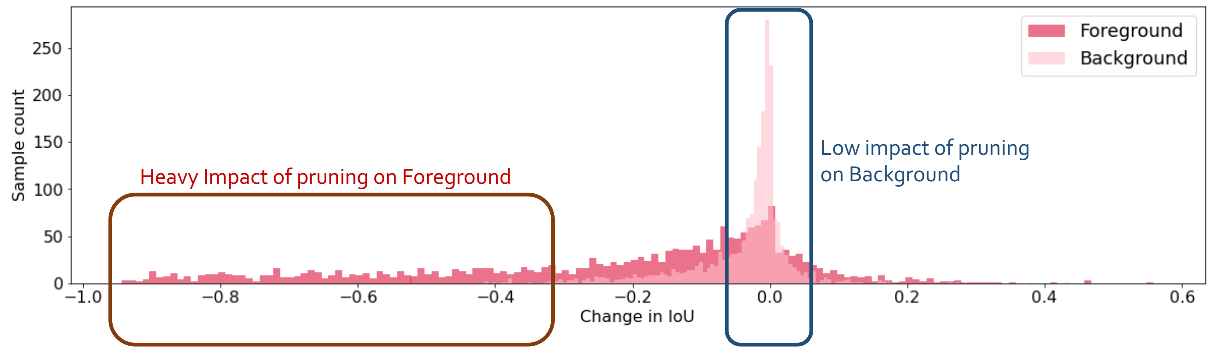

Figure 2. Impact of magnitude-based pruning (70%) of a trained U-Net based localization network on the IoU of foreground and background classes in the supervised setting. Pruning heavily impact the foreground IoU compared to background IoU.

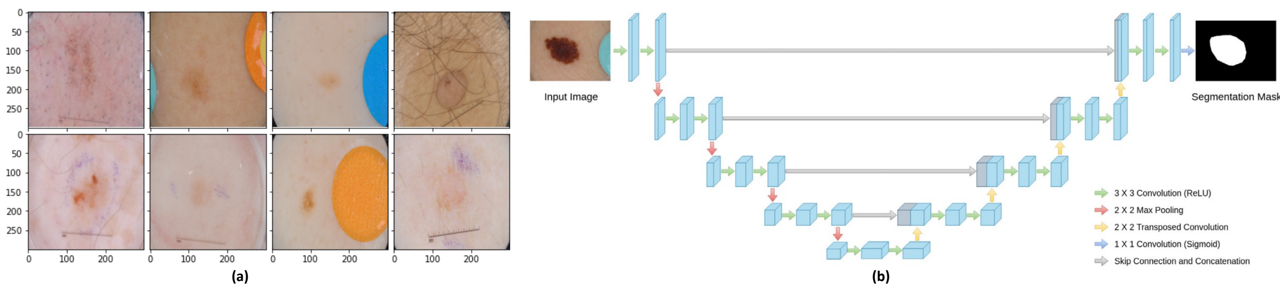

Figure 3. **(a)** Randomly sampled examples of HTLs identified by pruning U-Net based localization network by 70% in supervised setting. **(b)** Classical U-Net architecture used in our experiments for our Lesion localization. Our simple model choice ensures retaining focus on highlighting the relation between network pruning and implicit complex imbalance in dataset.

possible confusion, we do **NOT** use pruning for any model efficiency purpose. In our framework, pruning would be better described as "selective brain damage". It is mainly used for effectively spotting HTLs not yet well memorized and learned by the current model.

2.2. Hard-to-Learn (HTL) Instances

DNNs can be compressed to significantly huge levels at startlingly little loss of test accuracy using various pruning methods [25, 38, 20, 8, 36, 26, 18]. Recently, some works have identified the deeper connection of pruning with generalization/memorization, beyond considering it just as an ad-hoc compression tool [21, 46]. The most relevant work by [15] used pruning as a mean to expose the weakness of a trained model in generalization. More specifically, [15] identified that pruning a trained image classifier, produces a non-uniform impact on long-tail less frequent instances. In this paper, we study this observation for the **first time** to identify "*easily forgotten*" training examples for pathology localization in supervised, semi-supervised, and weakly-supervised settings. Using S-LLT as our experimental task, we identified that pruning disproportionately impacts the foreground and background class, hurting the foreground significantly (Figure 2). We observed that foreground performance of some training examples drops by a large margin, indicating their high sensitivity to pruning. We term these most impacted images as "***hard-to-learn***" (HTLs).

We would like to highlight that our idea of HTLs is a **bold attempt** to explore beyond the class-wise label imbalance. Even artificially class-balanced datasets such as CIFAR-10/100 and ImageNet. have many hidden inherent forms of imbalance such as class-level difficulty variations or instance-level feature distribution, which reflect in the performance of trained DNNs. Since HTLs are label-agnostic and completely rely on DNN's memorization ability and learning patterns, it is applicable to various more complicated forms of imbalance in real data, such as complex attribute imbalances [35] and demographic imbalances [37, 24]. Our in-depth analysis for S-LLT using ISIC-2017 validates the ability of HTLs to capture implicit demographic (gender and age) imbalances in the real-world dataset, which provide an opportunity to explicitly pay attention to them during training. Figure 3(a) presents some sampled examples of HTLs identified by pruning U-Net based localization network by 70% in supervised setting.

2.3. Mining HTLs

Medical imaging real-world datasets exhibit subtle forms of imbalances where various feature attributes have very different frequencies (eg. pathology color, size, and shape)

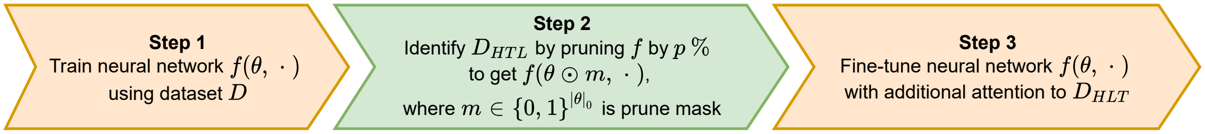

Figure 4. An overview of our Pruning Assisted NN training paradigm for localization.

Algorithm 1: Supervised pruning assisted localization

Input: Data $D = \{x_i\}_{i=1}^n$, Localization labels $Y = \{y_i\}_{i=1}^n$, Model = $f(\theta_{init}, \cdot)$, Pruning algorithm = P, $thresold_{HTL} = \tau$
Learn $f(\theta_{intermediate}, \cdot)$ by minimizing $\sum_{i=1}^n L_{localization}(\theta_{init}, D, Y)$
Prune $f(\theta_{intermediate}, \cdot)$ by $p\%$ using P to get $f(\theta_{intermediate} \odot m, \cdot)$ where $m \in \{0,1\}^{\|\theta\|_0}$ is the prune mask.
Identify $\tilde{D}_{HTL} = \{\{\tilde{x}_i, \tilde{y}_i\}$ st. $IoU(f(\theta_{intermediate} \odot m, \tilde{x}_i)) - IoU(f(\theta_{intermediate}, \tilde{x}_i)) > \tau\}$
Fine-tune $f(\theta_{intermediate}, \cdot) \rightarrow f(\theta_{final}, \cdot)$ by minimizing $\sum_{i=1}^{\|\tilde{D}_{HTL}\|} L_{localization}(\theta_{init}, \tilde{D}_{HTL}, Y_{HTL})$
Return $f(\theta_{final}, \cdot)$

and instance-level difficulty variations. Broadly speaking, such imbalances are not only limited to the standard majority versus minority class but extend to implicit forms based on gender, race, ethnicity, as well as demographics of individuals and can be very difficult to account for. In this section, we present **three different HTLs** (inherently capturing subtle imbalances in a label-agnostic way) **mining strategies** for localization in supervised, semi-supervised, and weakly-supervised settings.

2.3.1 Supervised Setting:

Our supervised setting considers the availability of segmentation masks ($\{y_i\}_{i=1}^n$) corresponding to every training image ($\{x_i\}_{i=1}^n$). We first train a U-Net model (Figure 3) using the supervised cross-entropy loss ($L_{localization}$) to fit on our labeled training data. In order to identify HTLs ($\tilde{D}_{HTL}$) from the training data, we prune the trained network using a pruning algorithm P by $p\%$ and look for instances which are highly sensitive to pruning (i.e., observed significant drop in foreground IoU performance). The complete supervised pruning assisted localization process is summarized in Algorithm 1. The final network is generated by fine-tuning with additional attention to spotted HTLs.

2.3.2 Semi-supervised Setting:

Our semi-supervised setting considers the availability of segmentation masks ($\{y_i\}_{i=1}^n$) corresponding to n input training images ($\{x_i\}_{i=1}^n$). Additionally, it make use of k pathology images ($\{\tilde{x}_i\}_{i=1}^k$) for which no segmentation mask is available. We first train a U-Net model (Figure 3) using the supervised cross-entropy loss ($L_{localization}$) to fit on our labelled training data (D). Next, we generate pseudo-labels for k unlabelled pathology images. In order to identify HTLs ($\tilde{D}_{HTL}$) from the training data, we prune the trained network using a pruning algorithm P by $p\%$ and look for pseudo-labels instances which are highly sensitive to pruning (i.e., observed significant change in bounding box generated using [17]). The complete semi-supervised pruning-assisted localization process is summarized in Algorithm 2. The final network is generated by fine-tuning with additional attention to spotted HTLs.

2.3.3 Weakly-supervised Setting:

In our weakly-supervised setting, we do not use any segmentation label corresponding to the input training dataset. Instead, we propose to use high-level classification labels to train our U-Net backbone using an additional MLP layer and supervised classification loss ($L_{classification}$). We have used ISIC-2018 dataset, which provides 10,015 images without segmentation masks divided into eight different clinical scenarios and evaluated performance on the ISIC-2017 test set with segmentation labels. In this setting, we pass the feature tensor from the last convolutional layer of the U-Net model trained with classification loss to GradCAM++[4] and generate a bounding box using [17]. To identify HTLs, we compare the bounding boxes generated for input images before and after pruning the backbone network by $p\%$ using the pruning algorithm P. We summarize the complete weakly-supervised pruning-assisted localization process in Algorithm 3. The final network is generated by fine-tuning with additional attention to HTLs.

2.4. Unified Pipeline

Our completed pruning-assisted localization pipeline is presented in Figure 4. Given a neural network $f(\theta, \cdot)$, we first train f using the training dataset D. Next, we identify

Algorithm 2: Semi-supervised pruning assisted localization

Input: Data $D = \{x_i\}_{i=1}^{n}$, Localization labels $Y = \{y_i\}_{i=1}^{n}$, Unlabelled Data $\tilde{D} = \{\tilde{x}_i\}_{i=1}^{k}$, Model = $f(\theta_{init}, \cdot)$, Pruning algorithm = P, $thresold_{HTL} = \tau$
Learn $f(\theta_{intermediate}, \cdot)$ by minimizing $\sum_{i=1}^{n} L_{localization}(\theta_{init}, D, Y)$
Generate localization pseudo-labels $\tilde{Y} = f_{\theta_{intermediate}}(\tilde{x_i})$ for $i = 1, 2, ..., k$
Prune $f(\theta_{intermediate}, \cdot)$ by $p\%$ using P to get $f(\theta_{intermediate} \odot m, \cdot)$ where $m \in \{0, 1\}^{\|\theta\|_0}$ is the prune mask.
Identify $\tilde{D}_{HTL} = \{\{\tilde{x}_i, \tilde{y}_i\}$ st. $BB(f(\theta_{intermediate} \odot m, \tilde{x_i})) - BB(f(\theta_{intermediate}, \tilde{x_i})) > \tau\}$
Fine-tune $f(\theta_{intermediate}, \cdot) \rightarrow f(\theta_{final}, \cdot)$ by minimizing $\sum_{i=1}^{\|\tilde{D}_{HTL}\|} L_{localization}(\theta_{init}, \tilde{D}_{HTL}, \tilde{Y}_{HTL})$
Return $f(\theta_{final}, \cdot)$

Algorithm 3: Weakly-supervised pruning assisted localization

Input: Data $D = \{x_i\}_{i=1}^{n}$, Classification labels $Y = \{y_i\}_{i=1}^{n}$, Heatmap Generator = $GradCAM_L(\cdot)$, Model = $f(\theta_{init}, \cdot)$, Pruning algorithm = P, $thresold_{HTL} = \tau$
Learn $f(\theta_{intermediate}, \cdot)$ by minimizing $\sum_{i=1}^{n} L_{classification}(\theta_{init}, D, Y)$
Generate heatmap as $Y_{CAM} = GradCAM_L(f(\theta_{intermediate}, x_i))$ for $i = 1, 2, ..., n$, where L represents tensor output of L-th layer of $f(\theta_{intermediate}, \cdot)$. Note that heatmap is an indicator of localization.
Prune $f(\theta_{intermediate}, \cdot)$ by $p\%$ using P to get $f(\theta_{intermediate} \odot m, \cdot)$ where $m \in \{0, 1\}^{\|\theta\|_0}$ is the prune mask.
Generate heatmap $\tilde{Y}_{CAM} = GradCAM_L(f(\theta_{intermediate} \odot m, x_i))$ for $i = 1, 2, ..., n$
Identify $\tilde{D}_{HTL} = \{\{\tilde{x}_i, \tilde{y}_i\}$ st. $BB(Y_{CAM}(x_i)) - BB(\tilde{Y}_{CAM}(x_i)) > \tau\}$
Fine-tune $f(\theta_{intermediate}, \cdot) \rightarrow f(\theta_{final}, \cdot)$ by minimizing $\sum_{i=1}^{\|\tilde{D}_{HTL}\|} L_{classification}(\theta_{init}, \tilde{D}_{HTL}, Y_{HTL})$
Return $f(\theta_{final}, \cdot)$

HTLs using methods proposed in section 2.3 which highlights the weakness of our trained network $f(\theta, \cdot)$. Finally, we fine-tune our network f by paying additional attention to HTLs using weighted cross-entropy loss as:

$$l_n = -\sum_{i=1}^{C} w_c \times \log \frac{\exp(\tilde{x}_{n,c})}{\exp(\sum_{i=1}^{C} \tilde{x}_{n,i})} \times \tilde{y}_{n,c} \qquad (3)$$

where $\tilde{x} \in \tilde{D}_{HTL}$, $\tilde{y}$ is the target, w is the weight of class c, C is the number of classes (i.e, foreground, background), and n is n-th training example. Our extensive experimental analysis indicates that the fine-tuned network achieve significantly high foreground performance gain on the ISIC 2017 test set across all three training paradigms.

3. Experimental Settings

Dataset Details: Our experiments used skin lesion localization (S-LLT) as our evaluation task and acquired dermoscopic images from the ISIC-2017 [7] and ISIC-2018 [6] challenge. The ISIC-2017 dataset consists of 2000, 150, and 600 lesion images in JPEG format for training, validation, and test along with the corresponding expert-annotated binary segmentation mask images in PNG format. It additionally provides demographic metadata entries of age and sex for the patients which we have used to validate our method's ability to elicit complex and sophisticated demographic imbalances. For our semi- and weakly- supervised task we have used ISIC-2018 dataset, which provides 10,015 images without segmentation masks divided into eight different clinical scenarios. Although our method can be adapted to any task, ISIC datasets provide metadata information (demographic details such as gender and age) along with high-quality segmentation annotations as well as unannotated images with classification labels. This facilitates a unique opportunity to effectively evaluate the benefits of our proposed method (Section 2.3) in multiple training settings (i.e., supervised, semi-supervised, and weakly-supervised) along with demographic analysis.

Training and Fine-tuning Details: In our experiments, all models are trained using similar settings and seed values (10, 20, 30). We have used an SGD optimizer with a momentum of 0.9 and weight decay of $2e^{-4}$. The initial learning rate is set to 0.1, and the networks are trained for 100 epochs with a batch size of 64. The learning rate decays by a factor of 10 at the [20, 50, 80]th epoch during the training. We have used standard augmentation techniques to flip, rotate, and mirror the images during training. For pruning, we have used a global unstructured-magnitude based pruning and prune ratio of 70% across all experiments. During the fine-tuning stage, we started with a smaller learning rate of 0.01 and retrained the network with HTLs for 20 epochs with a decay at the 15th epoch. All our models have

Prune Ratio	Supervised				Semi-supervised			
	25%	50%	75%	99%	25%	50%	75%	99%
Foreground	12.41% (↓)	26.64% (↓)	42.19% (↓)	86.15% (↓)	9.72% (↓)	15.33% (↓)	43.01% (↓)	81.98% (↓)
Background	0.95% (↓)	6.03% (↓)	17.11% (↓)	48.30% (↓)	0.26% (↓)	4.87% (↓)	11.32% (↓)	47.71% (↓)

Table 1. Percentage drop in IoU of training samples in ISIC-2017 dataset when the trained network is pruned by $\mathbf{p}\%$ using unstructured magnitude-based pruning in supervised and semi-supervised settings. Disproportionate impact of pruning can be clearly observed for the background and foreground classes.

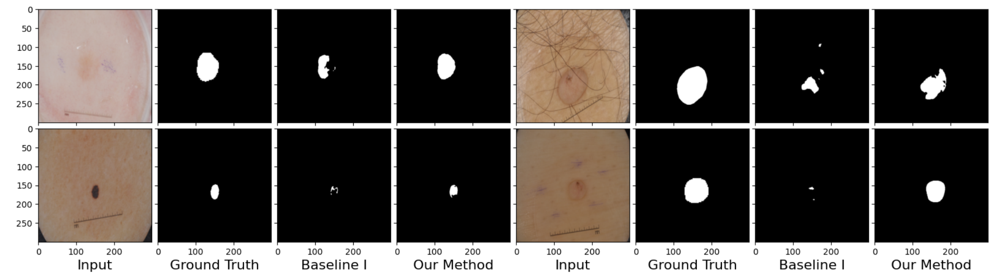

Figure 5. Examples of the visualization of segmentation mask generated by Baseline 1 and our method on HTL samples. With additional attention to HTL samples in fine-tuning, the model is able to generate better segmentation masks for HTLs improving overall performance.

been trained using 4 Quadro RTX 5000 GPUs and we have evaluated our models using foreground and background IoU scores against different baselines.

Additional Implementation Utility Details : In all our experiments (except ablation), we have used unstructured magnitude-based pruning, where we find and remove the least salient connections (weight magnitude) in the model wherever they are. For bounding box generation in Algorithm 2 and 3, we pass the feature tensor of the last convolution layer to GradCAM++[4] to extract the heatmaps. We further scale the heatmap intensity to the range [0-255] and use an ad-hoc threshold (pixel value = 180) to binarize the heatmap. In last, we followed the pseudocode proposed in [17] to create the bounding box. We have used two popular evaluation metrics (IoU and DICE) to compare our method performance against different baselines.

Baseline Comparison: In our experiments, the first baseline is a U-Net architecture trained to perform the S-LLT. We have adapted the original vanilla U-Net (**Baseline 1**) version proposed in [34] to avoid any design overhead and highlight the importance of HTLs. Our second baseline is the top-performing architecture from ISIC-2017 challenge leaderboard (**Baseline 2**). Our third baseline uses focal loss [28], which has been one default choice to handle imbalance (**Baseline 3**). In our fourth baseline, we randomly sample exactly the same number of instances (not specifically picked HTLs) and fine-tune our network similar to our proposed method, to validate the significance of identifying and using HTLs (**Baseline 4**). Next, in our fifth baseline, we randomly sampled exactly the same number of instances following the class distribution in S-LLT dataset where minority classes are sampled with higher probability to fine-tune our network (**Baseline 5**). Lastly, our final baseline randomly sampled exactly the same number of instances following the demographics distribution (gender) to fine-tune our network (**Baseline 6**). The performance comparison of all baselines compared to our HTL-based fine-tuning is reported in Table 2, which clearly unveil the effectiveness of our method. Note that the *main goal of our work* is to elicit the effectiveness of network pruning in identifying complex implicit imbalances in medical datasets, and propose a simple and unified approach to identify data samples suffering from imbalance during training, rather than proposing a task-specific novel class imbalance algorithm.

4. Main Results and Discussion

In this section, we present a comprehensive analysis of the impact of pruning in eliciting the weakness of trained localization models, and its high sensitivity towards the foreground performance compared to the background. Additionally, we provided a performance comparison of our proposed pruning-assisted localization algorithms 1,2, and 3 against several baselines. Finally, we provide an interesting observation of demographic bias captured by our pruning-based HTLs, where some demographics have a higher impact of pruning compared to others. Last, we have conducted an ablation study to illustrate that our observations are agnostic to various pruning methods, and HTLs identified by any pruning methods have similar benefits.

How does pruning impact foreground vs background? We find that pruning consistently amplifies the disparate treatment of foreground performance for all levels of compression we consider. Figure 2 illustrate when a U-Net based trained localization network is pruned by 70% us-

Method	Supervised				Semi-supervised				Weakly-supervised			
	B(IOU)	F(IoU)	Mean(IoU)	DICE	B(IoU)	F(IoU)	Mean(IoU)	DICE	B(IoU)	F(IoU)	Mean(IoU)	DICE
Baseline 1	0.894	0.654	0.774	0.855	0.901	0.661	0.781	0.865	0.622	0.420	0.521	0.637
Baseline 2	0.891	0.637	0.764	0.849	0.876	0.642	0.759	0.813	0.568	0.390	0.479	0.596
Baseline 3	0.892	0.663	0.775	0.863	0.906	0.669	0.788	0.872	0.614	0.411	0.513	0.644
Baseline 4	0.883	0.660	0.772	0.852	0.894	0.662	0.778	0.866	0.609	0.407	0.508	0.641
Baseline 5	0.878	0.659	0.769	0.844	0.900	0.647	0.774	0.859	0.610	0.399	0.504	0.635
Baseline 6	0.880	0.665	0.773	0.859	0.891	0.671	0.781	0.871	0.619	0.413	0.516	0.640
Our Method	0.890	0.681	0.785	0.871	0.904	0.695	0.800	0.884	0.643	0.438	0.542	0.657
Std	$\pm$0.002	$\pm$0.001	$\pm$0.001	$\pm$0.003	$\pm$0.001	$\pm$0.001	$\pm$0.001	$\pm$0.002	$\pm$0.006	$\pm$0.008	$\pm$0.007	$\pm$0.005

Table 2. Performance comparison of our method against different baselines using unstructured magnitude-based pruning with prune ratio of 70% on ISIC-2017 test set. Our method significantly improves the Foreground performance compared to all baselines.

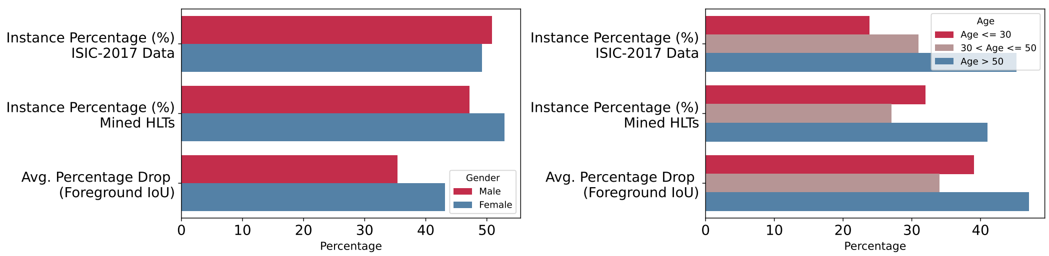

Figure 6. (a) Gender distribution, (b) Age distribution of instances in ISIC-2017, and mined HTLs using unstructured pruning (70%) of trained U-Net in supervised setting. Clearly, pruning differently impacts different age and gender groups' foreground performance.

ing magnitude-based unstructured pruning. We observe that nearly $\sim$18% instances lose their foreground performance (IoU) significantly by $>=$40%, compared to $\sim$1% losing their background performance (IoU) by $>=$40%. Additionally, Table 1, presents a detailed analysis of the disproportionate impact of pruning with varying thresholds $p \in \{25\%, 50\%, 75\%, 99\%\}$ on the foreground and background IoU in supervised and semi-supervised settings. It can be observed that across all the pruning thresholds, the foreground suffers more than the background in both training settings. Note that at very high sparsity (eg. 99%), the network performance becomes significantly low, and even the background performance drop significantly by $> 47\%$, making it unsuitable for HTL mining. Based on our experiments, we recommend 60-80% pruning ratio for unstructured magnitude-based pruning.

How does HTL-aware training benefit overall performance? In our work, we hypothesized HTLs to be representative of prevalent complex and implicit imbalance in the medical imaging dataset. In order to justify the benefits of workload to capture HTLs, it is important to illustrate how they can help in improving localization performance. Table 2 presents the comparison of our method (mean across 3 independent runs with seed 10, 20, 30) against three aforementioned baselines for supervised, semi-supervised, and weakly-supervised settings. It can be clearly observed that our pruning-assisted algorithms 1, 2, and 3, provide consistent and significant performance gain for two popular localization metrics IoU and DICE across all three training settings. More precisely, our proposed method achieve significant gain (IoU) of $+2.7\%, +3.4\%$, and $+2.1\%$ over baseline 1 in supervised, semi-supervised, and weakly-supervised settings for desired foreground class. Our method performance is significantly high when compared with ISIC leaderboard performance (Baseline 2). Focal loss proposed in [28] has been a very popular choice to handle imbalance but it is limited by the requirement of class distribution and the inability to capture subtle implicit imbalances. Compared to focal loss based baseline 3, our method achieves $+1.0\%, +1.2\%, and + 2.9\%$ better performance for three training settings.

To confirm that HTLs are special instances, our Baseline 4 randomly samples exactly same number of training instances in the dataset followed by fine-tuning using the same training protocols as HTLs. Table 2 elucidate the importance of HTLs when compared to baseline 4. Next, to our surprise Baseline 5, which samples fine-tuning instances following the class distribution, performs significantly worse than Baseline 4, which randomly samples HTLs without following any class distribution. An in-depth analysis reveals that although minority classes *dermatofibroma* and *vascular lesion* corresponds to only $\sim 0.9\%$ and $\sim 1\%$ in the dataset, they have **almost perfect performance** on the test set (98.7%, and 98.6% respectively). *Sampling additional samples to fine-tune these classes doesn't help (possibly leads to over-fitting on minority samples).* This clearly bolsters our motivation that the imbalances are not only limited to the standard majority versus minority class but extend to implicit and subtle forms such as feature attributes (e.g., pathology color, size, and shape) and based on demographics (e.g., gender, race,

Method	Full Network	20%	70%	95%
Random Pruning	0.654	0.431 (↓0.223)	0.218 (↓0.436)	0.022 (↓0.632)
Unstructured MB-Pruning	0.654	0.580 (↓0.074)	0.409 (↓0.245)	0.296 (↓0.358)
Structured MB-Pruning	0.654	0.533 (↓0.121)	0.387 (↓0.267)	0.215 (↓0.439)

Table 3. Impact of various pruning methods with varying sparsity on foreground performance in supervised setting.

ethnicity), and can be very difficult to account for. Lastly, the proportion of diagnoses associated with males and females is 50.81% and 49.19% in S-LLT dataset. Our last Baseline 6, which randomly samples following the gender distribution, has marginal benefit over Baseline 4 (+0.5% foreground performance), which further suggests the complexity of imbalance. Finally, Figure 5 illustrates a visualization of the segmentation mask generated by Baseline 1 and our method on some randomly selected samples from HTL mining. With additional attention to HTL samples in fine-tuning, the UNet-based localization model can generate better segmentation masks for HTL instances, improving the overall performance.

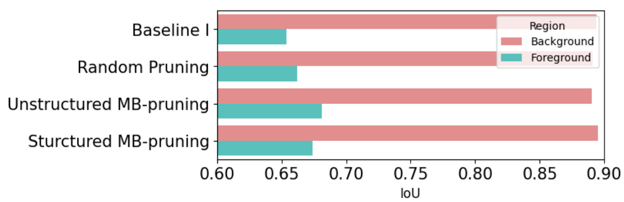

Figure 7. Localization performance of different pruning methods in supervised setting (HTLs mined with prune ratio 70%).

How do HTLs capture subtle demographic imbalance? HTLs are our bold attempt toward exploring complex and implicit imbalances which go beyond the class distribution in localization settings. In this section, we present an interesting in-depth demographic analysis of HTLs and validate that our pruning-assisted HTL mining indeed captures the demographic bias of the trained U-Net localization network. Figure 6 illustrates the gender and age group distribution of data points in ISIC-2017 dataset and mined HTLs. It can be clearly observed that although ISIC-2017 is fairly balanced from the gender perspective (50.82% and 49.18% for males and females), mined HTLs subset has a skewed gender distribution favoring females by $\sim$ 4%. Moreover, the average **drop** in foreground performance for females is >8% compared to men, which is an indicator of the model **favoritism** towards learning instances belonging to men patients. In addition, from the age group perspective, it can be observed that patients belonging to the age group within 30-40 years constitute the second largest proportion in data ($\sim$30.95%), but they are least impacted by pruning and share only $\sim$26.52% in HTLs subset and have a minimal drop in foreground IoU performance. However, interestingly, patients belonging to the age group <=30 years which have the least share in the data ($\sim$23.79%), are heavily impacted by pruning and contribute $\sim$32.11% in the HTLs subset which again points out the model's difficulty to learn patients belonging to this age group. This analysis is a **strong** indicator of pruning ability to uncover demographic bias in medical imaging datasets in localization tasks.

Impact of pruning algorithms on performance: To investigate that our observation of the disproportionate impact of pruning is agnostic to pruning methods, we carry out ablation studies on random pruning, unstructured magnitude-based pruning, and structured magnitude-based pruning. Table 3 illustrates the disproportionate impact of pruning methods with varying pruning ratio $p \in \{0\%, 20\%, 70\%, 95\%\}$ on foreground, with marginal impact on background IoU. In our experiments, we observed that p has minimal sensitivity to performance, and we achieve approximately similar performance gain of $\sim$2.812%±0.369 for $p \in \{50\%, 60\%, 70\%, 80\%\}$ in the supervised setting. Note that a large value of p will lead the pruned network to forget a lot of information, and due to overparametrization of DNNs, a small value of p will have no effect. Note that in all our experiments, we have used a pruning ratio of 70% considering its slightly better performance. Lastly, Figure 7 illustrates the localization performance of different pruning methods in supervised settings. Clearly, it can be observed that without hurting background IoU, all the pruning methods help in improving the foreground IoU significantly.

5. Conclusion

Contrary to the popular usage of pruning as an ad-hoc compression tool, in this paper, we present pruning as a technique to expose the weakness of a trained localization model and show the existence of "textithard-to-learn" training examples. We present three HTL mining strategies in supervised, semi-supervised, and weakly-supervised settings using ground truth labels, pseudo-labels, and saliency maps. We experimentally show that by attending HTLs during fine-tuning, we can significantly improve localization performance. Lastly, we present an interesting demographic analysis which illustrates HTLs ability to capture complex demographic imbalances. Our future work will aim for more theoretical understanding of the HTLs and their significance.

Acknowledgment

This work is supported by the National Library of Medicine under Award No. 4R00LM013001 and National NSF AI Center at UT Austin.

References

[1] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML 2017*, pages 233–242. PMLR, 2017.

[2] Enes Ayan and Halil Murat Ünver. Diagnosis of pneumonia from chest x-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–5. Ieee, 2019.

[3] Gustavo E. A. P. A. Batista, Ronaldo Cristiano Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD*, 6:20–29, 2004.

[4] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *WACV 2018*, pages 839–847, 2018.

[5] N. Chawla, K. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[7] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[9] Yiwen Guo, Anbang Yao, and Yurong Chen. Dynamic network surgery for efficient dnns. *Advances in neural information processing systems*, 29, 2016.

[10] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pages 4006–4016. PMLR, 2020.

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

[12] Yan Han, Chongyan Chen, Liyan Tang, Mingquan Lin, Ajay Jaiswal, Ying Ding, and Yifan Peng. Using radiomics as prior knowledge for abnormality classification and localization in chest x-rays. *ArXiv*, abs/2011.12506, 2020.

[13] Babak Hassibi, David G. Stork, and Gregory J. Wolff. Optimal brain surgeon and general network pruning. *IEEE International Conference on Neural Networks*, 1993.

[14] Robert C. Holte, Liane Acker, and Bruce W. Porter. Concept learning and the problem of small disjuncts. In *IJCAI*, 1989.

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

[16] Ajay Jaiswal, Kumar Ashutosh, Justin F. Rousseau, Yifan Peng, Zhangyang Wang, and Ying Ding. Ros-kd: A robust stochastic knowledge distillation approach for noisy medical imaging. 2022.

[17] Ajay Jaiswal, Tianhao Li, Cyprian Zander, Yan Han, Justin F. Rousseau, Yifan Peng, and Ying Ding. Scalp - supervised contrastive learning for cardiopulmonary disease classification and localization in chest x-rays using patient metadata. *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1132–1137, 2021.

[18] Ajay Jaiswal, Haoyu Ma, Tianlong Chen, Ying Ding, and Zhangyang Wang. Training your sparse neural network better with any mask. *arXiv preprint arXiv:2206.12755*, 2022.

[19] Ajay Jaiswal, Liyan Tang, Meheli Ghosh, Justin F Rousseau, Yifan Peng, and Ying Ding. Radbert-cl: Factually-aware contrastive learning for radiology report classification. In *Machine Learning for Health*, pages 196–208. PMLR, 2021.

[20] Ajay Kumar Jaiswal, Haoyu Ma, Tianlong Chen, Ying Ding, and Zhangyang Wang. Spending your winning lottery better after drawing it. *arXiv preprint arXiv:2101.03255*, 2021.

[21] Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *International Conference on Machine Learning*, pages 4927–4939. PMLR, 2021.

[22] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *SIGKDD Explor.*, 6:40–49, 2004.

[23] Miroslav Kubát and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, 1997.

[24] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[25] Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *NIPS*, 1989.

[26] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

[27] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *CVPR*, 2018.

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[29] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[30] Mehdi Habibzadeh Motlagh, Mahboobeh Jannesari, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, and Iman Hajirasouliha. Breast cancer

histopathological image classification: A deep learning approach. *BioRxiv*, page 242818, 2018.

[31] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports, 2017.

[32] Clifton Phua, Damminda Alahakoon, and Vincent Cheng-Siong Lee. Minority report in fraud detection: classification of skewed data. *SIGKDD Explor.*, 6:50–59, 2004.

[33] Mayank Raj, Ajay Jaiswal, Rohit R.R, Ankita Gupta, Sudeep Kumar Sahoo, Vertika Srivastava, and Yeon Hyang Kim. Solomon at SemEval-2020 task 11: Ensemble architecture for fine-tuned propaganda detection in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1802–1807, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[35] Nikolaos Sarafianos, Xiang Xu, and I. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. *ArXiv*, abs/1807.03903, 2018.

[36] Hidenori Tanaka, Daniel Kunin, Daniel LK Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *preprint arXiv:2006.05467*, 2020.

[37] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1), Apr. 2022.

[38] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. *arXiv preprint arXiv:2002.07376*, 2020.

[39] Tao Wang, Zhenxing Qin, Shichao Zhang, and Chengqi Zhang. Cost-sensitive classification with inadequate labeled data. *Information Systems*, 37(5):508–516, 2012.

[40] Gary M. Weiss. Mining with rarity: a unifying framework. *SIGKDD*, 6:7–19, 2004.

[41] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, ZongYuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021.

[42] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pages 10789–10798. PMLR, 2020.

[43] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[44] Shichao Zhang. Cost-sensitive classification with respect to waiting cost. *Knowledge-Based Systems*, 23(5):369–378, 2010.

[45] Shichao Zhang. Decision tree classifiers sensitive to heterogeneous costs. *Journal of Systems and Software*, 85(4):771–779, 2012.

[46] Zhenyu Zhang, Xuxi Chen, Tianlong Chen, and Zhangyang Wang. Efficient lottery ticket finding: Less data is more. In *International Conference on Machine Learning*. PMLR, 2021.

[47] Steven Zhou, Yixin Zhuang, and Rusong Meng. Multi-category skin lesion diagnosis using dermoscopy images and deep cnn ensembles. *línea], ISIC Chellange*, 2019.