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

Ajay Jaiswal1, Tianlong Chen1, Justin F. Rousseau1, Yifan Peng2, Ying Ding1, Zhangyang Wang1‡

1The University of Texas at Austin, 2Weill 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 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 (e.g., indistinguishable color, irregular shape, small size etc.).

Figure 1 illustrates 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 disproportionaly 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.

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 ~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 \( \kappa \) (i.e., number of non-zero weights), NN pruning can be written as a constrained optimization problem:

\[
\min_{w} L(w; D) = \min_{w} \frac{1}{n} \sum_{i=1}^{n} l(w; (x_i, y_i)), \quad (1)
\]

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

where, \( l(\cdot) \) is a standard loss function, \( w \) 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
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)
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 \(L_{\text{localization}}\) to fit on our labeled training data. In order to identify HTLs \(\tilde{D}_{\text{HTL}}\) 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 \(\{\tilde{y}_i\}_{i=1}^{n}\) corresponding to \(n\) input training images \(\{\tilde{x}_i\}_{i=1}^{n}\). 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 \(L_{\text{localization}}\) to fit on our labelled training data \(D\). Next, we generate pseudo-labels for \(k\) unlabelled pathology images. In order to identify HTLs \(\hat{D}_{\text{HTL}}\) 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 drop 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 \(L_{\text{classification}}\). 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 1: Supervised pruning assisted localization**

1. **Input:** Data \(D = \{x_i\}_{i=1}^n\), Localization labels \(Y = \{y_i\}_{i=1}^n\), Model = \(f(\theta_{\text{init}}, \cdot)\), Pruning algorithm = \(P\), threshold\(D_{\text{HTL}} = \tau\)
2. **Learn** \(f(\theta_{\text{intermediate}}, \cdot)\) by minimizing \(\sum_{i=1}^n L_{\text{localization}}(\theta_{\text{init}}, D, Y)\)
3. **Prune** \(f(\theta_{\text{intermediate}}, \cdot)\) by \(p\%\) using \(P\) to get \(f(\theta_{\text{intermediate}} \odot m, \cdot)\) where \(m \in \{0, 1\}^{\theta_{\text{init}}}\) is prune mask
4. **Identify** \(D_{\text{HTL}} = \{\tilde{x}_i, \tilde{y}_i\}\) st. \(\text{IoU}(f(\theta_{\text{intermediate}} \odot m, \tilde{x}_i) - \text{IoU}(f(\theta_{\text{intermediate}}, \tilde{x}_i)) > \tau)\)
5. **Fine-tune** \(f(\theta_{\text{intermediate}}, \cdot)\) \(\rightarrow f(\theta_{\text{final}}, \cdot)\) by minimizing \(\sum_{i=1}^{\|D_{\text{HTL}}\|} L_{\text{localization}}(\theta_{\text{init}}, \hat{D}_{\text{HTL}}, Y_{\text{HTL}})\)
6. **Return** \(f(\theta_{\text{final}}, \cdot)\)
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
Table 1. Percentage drop in IoU of training samples in ISIC-2017 dataset when the trained network is pruned by $p\%$ 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.

<table>
<thead>
<tr>
<th>Prune Ratio</th>
<th>Supervised 25%</th>
<th>50%</th>
<th>75%</th>
<th>99%</th>
<th>Semi-supervised 25%</th>
<th>50%</th>
<th>75%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Foreground</td>
<td>12.41% (↓)</td>
<td>26.64% (↓)</td>
<td>42.19% (↓)</td>
<td>86.15% (↓)</td>
<td>9.72% (↓)</td>
<td>15.33% (↓)</td>
<td>43.01% (↓)</td>
</tr>
<tr>
<td></td>
<td>Background</td>
<td>0.95% (↓)</td>
<td>6.03% (↓)</td>
<td>17.11% (↓)</td>
<td>48.30% (↓)</td>
<td>0.26% (↓)</td>
<td>4.87% (↓)</td>
<td>11.32% (↓)</td>
</tr>
</tbody>
</table>

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.

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-
our proposed method achieve
larger terms. More precisely, our proposed method achieves
localization metrics IoU and DICE across all three train-
sistent and significant performance gain for two popular
weakly-supervised settings. It can be clearly observed that
in independent runs with seed 10, 20, 30) against three afore-
mentioned baselines for supervised, semi-supervised, and
weakly-supervised settings. It can be clearly observed that
our pruning-assisted algorithms 1, 2, and 3, provide consis-
tent and significant performance gain for two popular
localization metrics IoU and DICE across all three training
settings. More precisely, our proposed method achieve

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervised</th>
<th>Semi-supervised</th>
<th>Weakly-supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B(IoU)</td>
<td>F(IoU)</td>
<td>Mean(IoU)</td>
</tr>
<tr>
<td>Baseline 1</td>
<td>0.894</td>
<td>0.654</td>
<td>0.774</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>0.891</td>
<td>0.637</td>
<td>0.764</td>
</tr>
<tr>
<td>Baseline 3</td>
<td>0.892</td>
<td>0.663</td>
<td>0.775</td>
</tr>
<tr>
<td>Baseline 4</td>
<td>0.883</td>
<td>0.660</td>
<td>0.772</td>
</tr>
<tr>
<td>Baseline 5</td>
<td>0.878</td>
<td>0.659</td>
<td>0.769</td>
</tr>
<tr>
<td>Baseline 6</td>
<td>0.880</td>
<td>0.665</td>
<td>0.773</td>
</tr>
<tr>
<td>Our Method</td>
<td>0.890</td>
<td>0.681</td>
<td>0.785</td>
</tr>
</tbody>
</table>

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.

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 achieves 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 ∼0.9% and ∼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,
and share only ∼HTLs subset which again points out the model’s difficulty to implicitly impacted by pruning and contribute which have the least share in the data (interestingly, patients belonging to the age group <30-40 years constitute the second largest proportion within 30-40 years constitute the second largest 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. 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 ∼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 (∼30.95%), but they are least impacted by pruning and share only ∼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 (∼23.79%), are heavily impacted by pruning and contribute ∼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 ∈ {0%, 20%, 70%, 95%} on foreground, with marginal impact on background IoU. In our experiments, we observed that p has minimal sensitivity to foreground performance, and we achieve approximately similar performance gain of ∼2.812±0.369 for p ∈ {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


[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.


[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.


