Pruning nnU-Net with Minimal Performance Loss

Tongyun Yang TONGUYUNYANG@TUDELFT.NL
Yidong Zhao Y.ZHAO-8@TUDELFT.NL
Qian Tao Q.TAO@TUDELFT.NL
Department of Imaging Physics, Delft Unversity of Technology

Editors: Accepted for publication at MIDL 2025

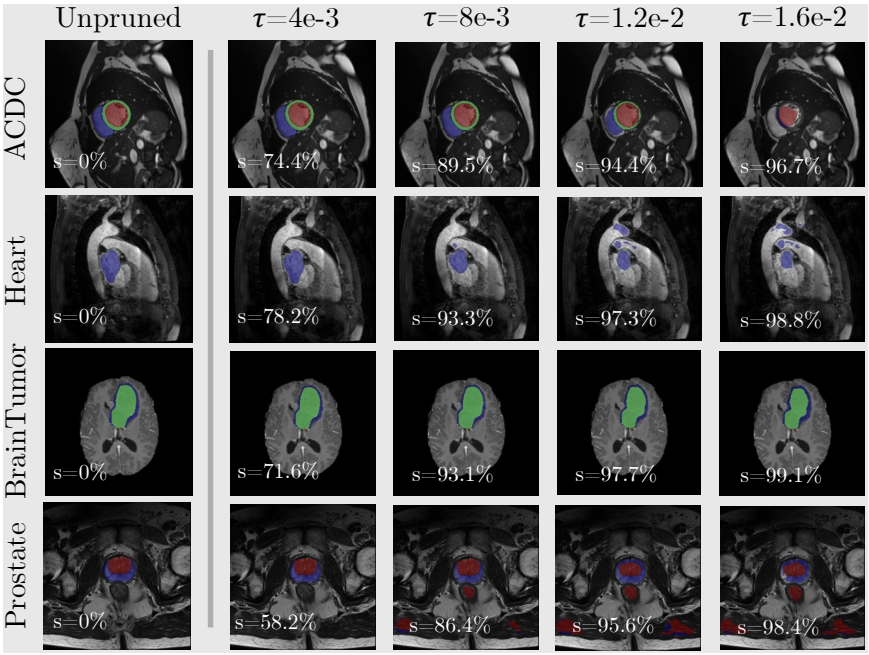

Figure 1: Comparison of the segmentation maps produced by the nnUNet pruned at various magnitude thresholds τ, for four different datasets. Each column corresponds to a different value of τ, and s in each subfigure indicates the resulting weight sparsity.

Abstract

nnU-Net is widely known for its accurate and robust segmentation performance in medical imaging tasks. However, the trained networks are typically heavily parameterized, and the high computational demand limit their deployment on devices with constrained resources. In this paper, we show that more than 80% of the trained nnU-Net weights can be removed without significant performance degradation, maintaining a proxy Dice score of > 0.95. This applies to both 2D and 3D configurations across four different medical image segmentation datasets. Interestingly, we observe that critical weights consistently concentrate near the U-Net encoder and decoder ends, while the bottleneck layers can be heavily pruned. These findings highlight the significant weight redundancy in trained nnU-Net and suggest opportunities for further optimization, to facilitate deployment of the model on devices with limited resources.

Keywords: Medical image segmentation, nnU-Net, weight pruning, sparsity

© 2025 CC-BY 4.0, T. Yang, Y. Zhao & Q. Tao.

1. Introduction

nnU-Net (Isensee et al., 2021) has become the go-to architecture for medical image segmentation tasks, known for its exceptional out-of-the-box performance in a wide range of datasets. Despite its robustness, nnU-Net remains computationally intensive, which limits its feasibility for deployment in devices with restricted resources (Stock et al., 2024). Pruning is a widely used technique that introduces sparsity by removing less significant weights, successfully reducing computational demands in other domains such as computer vision (Tan and Le, 2019) and large language models (Frantar and Alistarh, 2023). Some recent works investigated pruning of the UNet architecture, for example, STAMP (Dinsdale et al., 2022) demonstrated that 85% sparse U-Net could outperform in data-scarce scenarios. Sauron U-Net (Valverde et al., 2024) employed channel pruning based on feature maps and achieved inference speed-ups. However, little effort has been directed towards studying the weight redundancy in the nnU-Net, which is often the default choice, and is auto-configured for any 2D and 3D medical image datasets.

In this paper, we explore the weight redundancy of trained nnU-Nets, for both 2D and 3D datasets, and show that they can be heavily pruned post-training with minimal impact on the performance. Furthermore, we find that critical weights consistently concentrate on specific layers, exhibiting an inverted-U-shaped redundancy pattern. These empirical findings suggest opportunities for further model optimization towards reduced memory and computational footprint.

2. Method

We develop an easily extendable pruning pipeline on top of the nnU-Netv2 pipeline[1]. This pipeline fully leverages the standard nnU-Net preprocessing, postprocessing, and ensembling strategies. We employ post-training pruning, by first training the full model to convergence, and then selectively removing weights based on magnitude. Specifically, for each of the five cross-validation folds, we set weights with magnitude below a fixed threshold τ to zero: $w \leftarrow 0$ if $|w| < \tau$, a straightforward yet highly successful pruning strategy (Han et al., 2015). The pruned network is directly used without retraining.

To evaluate the efficacy of our pruning method, we report the segmentation performance using a *proxy* Dice score, defined as the Dice similarity coefficient (DSC) between predictions from the pruned and unpruned models. This proxy is used because many existing benchmarks do not provide ground truth for the testing set. By using the full nnU-Net as reference, we can effectively measure performance deviation after pruning.

3. Experiments

3.1. Implementation Details

All experiments were conducted on an NVIDIA RTX 4090 card with 64GB RAM. We evaluated on four datasets: Brain Tumor, Prostate, and Heart from the medical segmentation decathlon (MSD) challenge (Antonelli et al., 2022), and the ACDC dataset (Bernard et al., 2018). The nnU-Net auto-configuring pipeline resulted in models of different sizes: 155M

1. https://github.com/MIC-DKFZ/nnUNet

parameters for Brain Tumor, 222M for Prostate, 103M for Heart, and 167M for ACDC, respectively. The code of the pruning pipeline [2] will be made publicly available.

3.2. Results

Figure. 2 (a) reports the segmentation performance of the model after pruning, showing that nnU-Net can be pruned by 80% with minimal performance degradation, i.e. maintaining 0.95 proxy DSC. For the ACDC dataset, which has the ground truth available for the testing set, we also plotted the true Dice score (in pink). Figure. 2 (b) shows the number of weights through the layers and their corresponding pruning ratios. It can be observed that while the parameter count increases in the deeper layers, these deeper layers exhibit a greater sparsity after pruning. This trend is consistent in all four datasets.

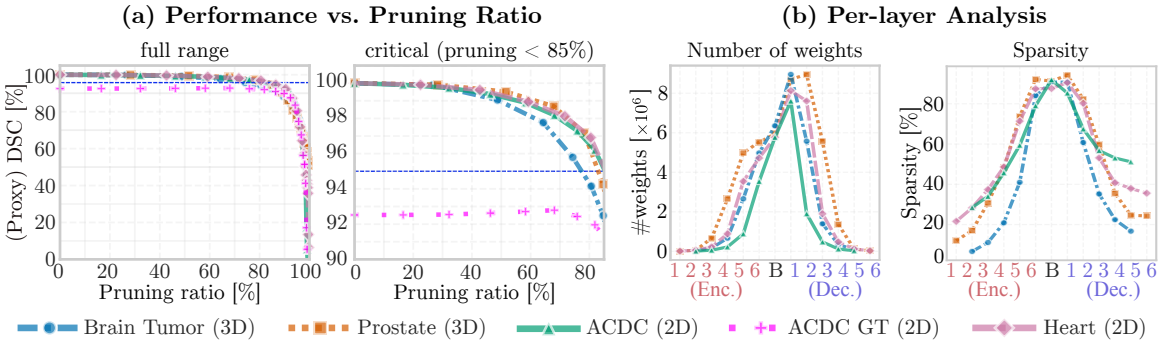

Figure 2: (a): Proxy DSC vs. pruning ratio across all datasets, showing minimal performance degradation up to 80% pruning. For the ACDC dataset where the ground truth (GT) segmentation is available for the testing set, the real DSC score is also plotted (in pink). (b): Weight size and pruning ratios across layers in the encoder (Enc.), bottleneck (B), and decoder (Dec.), revealing increasing sparsity in deeper layers and the bottleneck. The same trend is observed in all datasets.

4. Conclusion

Our results demonstrate that nnU-Net can retain competitive performance after aggressive pruning of up to 80% of weights using a straightforward magnitude-based method. This highlights the substantial weight redundancy in trained nnU-Nets[3]. We further revealed an intriguing inverted-U-shaped weight redundancy pattern of nnU-Net, consistent across four datasets of 2D and 3D images. Our findings suggest opportunities for further optimization, to reduce the energy footprint and facilitate deployment of the model on devices with limited resources.

2. https://github.com/prunennunet/Prune_nnUNet
3. We note that this work only investigated post-training pruning of nnU-Net, and that the weight redundancy does not imply redundancy in the architecture design, which may be essential for effective training and warrants further investigation.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. Stamp: Simultaneous training and model pruning for low data regimes in medical image segmentation. *Medical Image Analysis*, 81:102583, 2022.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR, 2023.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Raphael Stock, Yannick Kirchhoff, Maximilian R Rokuss, Ashis Ravindran, and Klaus Maier-Hein. Segment anything in medical images with nnunet. In *Medical Image Segmentation Challenge*, pages 167–179. Springer, 2024.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Juan Miguel Valverde, Artem Shatillo, and Jussi Tohka. Sauron u-net: Simple automated redundancy elimination in medical image segmentation via filter pruning. *Neurocomputing*, 594:127817, 2024.