Transactions on Computational and Scientific Methods | Vol. 5, No. 5, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Channel Pruning for Lightweight UNet in Medical Image Segmentation

Thayer Elwood
Missouri Western State University, St. Joseph, USA
telwood331@missouriwestern.edu

Abstract: UNet has become a cornerstone in medical image segmentation due to its strong performance, yet its high computational complexity poses challenges for real-world deployment. This paper presents a structured channel pruning method that removes less important convolutional channels to reduce model size and accelerate inference. The approach maintains the original architecture while significantly improving computational efficiency, enabling real-time performance on edge devices and in clinical settings with limited resources. Experiments on benchmark datasets, including ISIC 2018 and BraTS, demonstrate that the pruned UNet achieves substantial reductions in parameters and inference time with minimal impact on segmentation accuracy. This work offers a practical solution for compressing UNet models without redesigning the network or relying on specialized hardware.

Keywords: Channel pruning; UNet; Medical image segmentation;

1. Introduction

The rapid advancement of deep learning has brought significant breakthroughs in the field of medical image analysis, particularly in the area of image segmentation. Among the various architectures proposed, UNet [1] and its enhanced version, UNet++ [2], have become the backbone of many state-of-the-art solutions for tasks such as organ delineation, lesion detection, and tumor segmentation. These networks are designed with intricate encoder-decoder structures and skip connections, enabling them to capture both low-level details and high-level semantic information from medical images.

Despite their impressive accuracy, the practical deployment of UNet and UNet++ in clinical settings and on edge devices remains challenging. The main obstacles are the large number of parameters and the considerable computational resources required during inference. In many real-world scenarios, such as point-of-care diagnostics or mobile health applications, computational efficiency and memory footprint are as critical as segmentation accuracy. Therefore, there is a pressing need to develop strategies that can reduce the complexity of these networks without sacrificing their effectiveness.

One promising approach to address this issue is channel pruning, which targets the removal of less significant channels in convolutional layers. By systematically identifying and discarding redundant channels, it is possible to simplify the network structure and accelerate inference. Among various criteria for channel importance, the L1-norm [3] of convolutional weights has gained attention due to its simplicity and

effectiveness. The L1-norm provides a straightforward metric for ranking channel significance, making it a practical choice for large-scale network pruning.

In this work, we investigate the impact of L1-norm-based channel pruning on both UNet and UNet++ architectures. Our objective is to achieve substantial reductions in model size and computational demands, while maintaining high segmentation performance. We conduct comprehensive experiments on publicly available medical image datasets, evaluating how different pruning ratios affect both quantitative metrics and qualitative outcomes. Through these experiments, we aim to provide a clear understanding of the trade-offs involved in model compression and to offer guidance for deploying efficient segmentation networks in resource-limited environments.

The remainder of this paper is organized as follows: Section 2 reviews related work in model compression and pruning techniques; Section 3 details the proposed L1-norm-based pruning methodology; Section 4 presents experimental results and analysis; and Section 5 concludes with insights and directions for future research.

2. Related Work

2.1 Model Compression in Deep Neural Networks

The growing complexity of deep neural networks has prompted extensive research into model compression techniques. Early efforts primarily focused on weight quantization and parameter sharing to reduce memory usage and computation. Quantization methods, for example, map high-precision weights to lower bit representations, allowing models to run efficiently on hardware with limited resources. Other approaches, such as knowledge distillation, train smaller student networks to mimic the behavior of larger teacher models, thus achieving compactness while preserving performance. These strategies, while effective, often require specialized training or hardware support.

2.2 Channel Pruning Techniques

Channel pruning [4][5] has emerged as a practical strategy for reducing the computational burden of convolutional neural networks (CNNs). Unlike unstructured pruning, which removes individual weights and leads to sparse matrices, channel pruning eliminates entire feature maps, resulting in structured and hardware-friendly reductions. Various criteria have been proposed for channel selection, including sensitivity analysis [6], Taylor expansion [7], and norm-based metrics [8]. Among these, the L1-norm of convolutional filters is widely adopted due to its simplicity and interpretability. By ranking channels based on the sum of absolute weights, less important channels can be systematically removed, leading to more efficient models without the need for extensive retraining.

At its core, channel pruning seeks to eliminate redundant or less informative channels (also referred to as feature maps) from convolutional layers. The intuition behind this approach is that not all channels contribute equally to the network's decision-making process; some may encode similar or even irrelevant information. By systematically removing these less important channels, the model can achieve substantial reductions in parameter count and computational operations (FLOPs), often with minimal impact on overall performance. The process of channel pruning typically involves three main stages: importance evaluation, channel selection, and network fine-tuning.

The first stage, importance evaluation, is critical to the success of pruning. A variety of criteria have been proposed to assess the significance of each channel. One of the most straightforward and widely used methods is based on the magnitude of the filter weights, such as the L1-norm or L2-norm. The L1-norm

approach sums the absolute values of the weights in each channel, assuming that channels with smaller sums are less important. Alternatively, the L2-norm considers the Euclidean norm of the weights. Other, more sophisticated criteria involve data-driven measures, such as the average activation of each channel across a dataset, sensitivity analysis using Taylor expansion, or the impact of channel removal on the loss function. Some recent techniques leverage attention mechanisms or reinforcement learning to adaptively determine which channels to prune, further improving the efficiency and effectiveness of the pruning process.

After determining the importance of each channel, the next step is channel selection and removal. This can be performed globally across the entire network or locally within individual layers. Global pruning considers the importance scores of all channels in the network and prunes the least important ones, regardless of their layer, while local pruning applies a fixed pruning ratio to each layer independently. The choice between global and local pruning depends on the specific architecture and the desired balance between model compactness and accuracy. Once the channels to be pruned are identified, the corresponding filters are removed from the affected layers, and the input dimensions of subsequent layers are adjusted accordingly. Special care must be taken in architectures with skip connections or feature concatenations, such as UNet and UNet++, to ensure that the pruned network remains structurally valid and that feature maps can still be properly merged.

The final stage of channel pruning involves fine-tuning the pruned network. This step is essential to recover any potential loss in accuracy due to the structural changes introduced by pruning. Fine-tuning is typically performed with a lower learning rate and may involve additional regularization or data augmentation to help the network adapt to its new, more compact form. In practice, channel pruning is often applied iteratively, with repeated cycles of pruning and fine-tuning, to achieve the desired trade-off between efficiency and accuracy. Recent advances in channel pruning have also explored joint optimization with other model compression techniques, such as quantization or knowledge distillation, to further enhance the deployment potential of deep neural networks in clinical settings.

In summary, channel pruning techniques offer a powerful and flexible means of optimizing deep learning models for medical image segmentation. By carefully evaluating channel importance and strategically removing redundancy, these methods enable the construction of lightweight, high-performance networks that are well-suited for real-world medical applications. The ongoing development of more adaptive and data-driven pruning strategies continues to push the boundaries of what is possible in efficient deep learning, paving the way for broader adoption of AI-assisted diagnostics and interventions.

2.3 UNet and UNet++ Architectures

UNet and UNet++ are two of the most influential architectures in the field of medical image segmentation, particularly known for their ability to efficiently capture both local and global contextual information. These architectures are designed to address the challenges posed by biomedical images, such as limited data availability, complex anatomical structures, and the need for precise boundary delineation. Below, we provide a comprehensive and original discussion of their designs, key features, and the motivations behind their architectural choices.

UNet, first introduced by Ronneberger et al. in 2015, is characterized by its symmetric encoder-decoder structure, which enables effective learning of spatial hierarchies. The encoder path, also known as the contracting path, consists of repeated applications of two 3×3 convolutional layers followed by a rectified linear unit (ReLU) activation and a 2×2 max pooling operation with stride 2 for downsampling. This process gradually reduces the spatial dimensions while increasing the number of feature channels, allowing the network to capture increasingly abstract and high-level representations. In contrast, the decoder path, or expansive path, mirrors the encoder but replaces pooling operations with up-convolutions (transposed

convolutions) to restore the original spatial resolution. At each step in the decoder, the feature maps are concatenated with the corresponding feature maps from the encoder via skip connections. These skip connections play a vital role in preserving fine-grained spatial information and enabling precise localization, which are critical for accurate segmentation of medical images where boundaries between different tissues can be subtle.

A defining characteristic of UNet is its ability to make efficient use of limited annotated data. The architecture can be trained end-to-end from relatively few images and still achieve high segmentation accuracy, largely due to its extensive use of data augmentation and the strong inductive bias introduced by its symmetric structure and skip connections. Additionally, the use of small convolutional kernels and deep supervision at multiple scales allows UNet to capture both fine details and global context, making it highly adaptable to various biomedical segmentation tasks, such as cell tracking, organ delineation, and lesion detection.

Building upon the foundation laid by UNet, UNet++ introduces a more sophisticated approach to feature fusion and multi-scale representation. The primary innovation in UNet++ is the redesign of skip pathways, which are now composed of a series of nested, dense convolutional blocks. Unlike the direct skip connections in the original UNet, UNet++ utilizes a series of intermediate convolutional layers to bridge the semantic gap between encoder and decoder feature maps. This nested structure allows for more gradual and effective integration of low-level and high-level features, resulting in improved segmentation accuracy, especially in cases where the boundaries between structures are ambiguous or the objects exhibit significant scale variation.

In UNet++, each skip pathway is constructed as a dense convolutional block, where the output at each depth is connected to all subsequent nodes within the same skip pathway. This design not only enhances gradient flow during training but also enables the network to aggregate features at multiple semantic levels. Furthermore, UNet++ supports deep supervision by attaching auxiliary segmentation heads at different depths within the decoder, which encourages the network to learn robust representations at various scales and improves convergence during training. The flexibility of UNet++ allows it to be adapted to different computational budgets by pruning or reconfiguring the nested skip pathways, making it suitable for both high-performance and resource-constrained environments.

Both UNet and UNet++ have demonstrated outstanding performance in a wide range of medical image segmentation challenges, including but not limited to brain tumor segmentation, retinal vessel extraction, and lung nodule detection. Their architectures are highly modular, facilitating easy adaptation to three-dimensional data, multi-modal inputs, or integration with attention mechanisms and other advanced modules. Despite their similarities, UNet++ typically achieves higher segmentation accuracy than the original UNet, particularly on complex datasets, due to its enhanced feature fusion strategy and deep supervision. However, this comes at the cost of increased computational complexity and memory usage, which may require additional optimization for deployment on hardware-limited systems.

In summary, UNet and UNet++ represent two landmark architectures in medical image segmentation, each with unique design philosophies tailored to the challenges of biomedical data. UNet's simplicity, effectiveness, and efficiency make it a popular baseline for many applications, while UNet++ pushes the boundaries of segmentation performance through sophisticated multi-scale feature aggregation and deep supervision. Their widespread adoption and continued evolution underscore their importance in advancing the state of the art in automated medical image analysis.

2.4 Pruning in Semantic Segmentation Networks

While pruning techniques have been widely studied in classification networks, their application to semantic segmentation architectures is less explored. The unique demands of segmentation—such as preserving spatial resolution and fine-grained details—make pruning more challenging in this context. Some recent works have adapted channel pruning and other compression techniques to segmentation models, demonstrating that significant reductions in model size and inference time are possible with minimal loss in accuracy. Nevertheless, the trade-off between efficiency and segmentation quality remains an open question, particularly for architectures as complex as UNet and UNet++.

2.5 Summary

In summary, existing research highlights the potential of channel pruning, particularly L1-norm-based methods, for optimizing deep neural networks. However, there is a gap in systematically evaluating these techniques within advanced segmentation architectures like UNet and UNet++. This study aims to address this gap by providing a comprehensive analysis of L1-norm-based channel pruning applied to both networks, with a focus on practical deployment in medical imaging scenarios.

3. Methodology

3.1 Overview

This section presents a comprehensive overview of the proposed channel pruning methodology tailored for UNet and UNet++ architectures, with a particular focus on leveraging the L1-norm criterion to assess channel importance. The goal of this approach is to systematically streamline the network by eliminating redundant or less informative channels within convolutional layers. By doing so, the method aims to achieve a significant reduction in both model size and computational burden, which is particularly beneficial for deploying deep learning models in real-time or resource-constrained medical environments. Importantly, the pruning process is carefully designed to preserve the segmentation accuracy and ensure that the essential features required for precise delineation of anatomical structures are retained.

The proposed methodology is structured into three distinct yet interconnected stages: channel importance evaluation, pruning strategy implementation, and fine-tuning of the pruned network. In the first stage, the importance of each channel is quantitatively assessed using the L1-norm of the corresponding filter weights. This criterion provides an efficient and interpretable measure of how much each channel contributes to the overall feature representation. Channels with lower L1-norm values are considered less significant and are flagged as potential candidates for removal.

The second stage involves the actual implementation of the pruning strategy. Based on the importance scores obtained from the first stage, a predetermined proportion of the least important channels are systematically pruned from each convolutional layer. This process is conducted with careful attention to the architectural characteristics of UNet and UNet++, such as skip connections and feature concatenations, to ensure structural compatibility and seamless information flow throughout the network. The pruning can be performed either globally across the entire model or locally within individual layers, depending on the desired balance between efficiency and accuracy.

The final stage is the fine-tuning of the pruned network. Following the removal of redundant channels, the network is retrained on the original dataset with a reduced learning rate. This step is crucial for enabling the model to adapt to its new, more compact structure and to recover any potential loss in segmentation

performance caused by pruning. Fine-tuning helps to recalibrate the remaining weights and ensures that the pruned model maintains high accuracy in segmenting complex medical images.

Overall, the outlined approach provides a practical and effective framework for optimizing UNet-based segmentation models. By systematically identifying and removing unnecessary channels, the method not only enhances computational efficiency but also facilitates the deployment of deep learning solutions in clinical workflows where speed and resource utilization are critical. The modular nature of the methodology allows for easy adaptation to different network architectures and pruning criteria, making it a versatile tool for model compression in a wide range of medical imaging applications.

3.2 Channel Importance Evaluation Using L1-Norm

Identifying which channels to prune is a critical step in model compression. In this work, the L1-norm of convolutional weights is employed as the primary criterion for evaluating channel importance. This approach is favored for its simplicity, computational efficiency, and proven effectiveness in various pruning studies.

- *Theoretical Basis*
  The L1-norm, defined as the sum of the absolute values of a channel's weights, provides a direct measure of the overall magnitude of the filter. The underlying assumption is that channels with smaller L1-norms contribute less to the learned feature representation and, therefore, can be considered less important for the network's predictive performance. This is particularly relevant in deep networks where over-parameterization often leads to redundancy.

- *Computation Process*
  For each convolutional layer in the network, the L1-norm is calculated for every output channel. Given a convolutional kernel tensor W with dimensions $C_{out} \times C_{in} \times K \times K$, the L1-norm for the i-th output channel is computed as follows:

$$L1(i) = \sum_{j=1}^{C_{in}} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} |W_{i,j,k_1,k_2}| \qquad (1)$$

  This operation is repeated for all output channels across all convolutional layers targeted for pruning.The resulting L1-norm values are then used to rank the channels within each layer.

- *Selection of Pruning Candidates*
  After ranking, a pruning threshold or ratio is determined based on the desired level of model compression. Channels with the lowest L1-norm values—those deemed least significant—are selected for removal. The pruning ratio can be set globally (the same proportion for all layers) or locally (different proportions per layer), depending on the specific requirements and sensitivity of each layer.
  The use of the L1-norm offers several benefits:
  Simplicity: It does not require additional training or complex computations, making it easy to implement.
  Interpretability: The magnitude of weights has a clear, intuitive relationship with feature importance.
  Efficiency: The calculation is straightforward and can be performed as a post-processing step after initial training.

- *Limitations and Considerations*
  While the L1-norm is effective, it does have limitations. It does not directly account for the interdependence between channels or the potential impact on downstream layers. In some cases, removing a channel with a low L1-norm might still affect the network disproportionately if it carries unique information. To address this, the pruning process is typically followed by fine-tuning, allowing the network to recover and redistribute representational capacity.

- *Comparison with Other Criteria*
  Alternative methods for channel importance evaluation include the L2-norm, first-order Taylor expansion, and data-driven approaches such as sensitivity analysis. The L2-norm, for example, considers the Euclidean magnitude of weights, while Taylor-based methods estimate the impact of pruning on the loss function. Although these techniques may offer marginal improvements in certain scenarios, the L1-norm remains a popular choice due to its balance of effectiveness and computational simplicity.

In the context of UNet and UNet++, the L1-norm-based evaluation is systematically applied to all convolutional layers except those involved in the final output mapping. Special care is taken in the encoder-decoder architecture to maintain the integrity of skip connections and concatenations, ensuring that the pruned network structure remains functional and compatible with the original design.

By leveraging the L1-norm as a channel importance metric, the proposed methodology achieves an effective reduction in network redundancy with minimal manual intervention, paving the way for efficient and scalable model compression in medical image segmentation tasks.

3.3 Channel Pruning Strategy

Channel pruning is a crucial strategy for compressing convolutional neural networks, especially in medical image segmentation tasks where computational efficiency and model size are important considerations. The central idea of channel pruning is to identify and remove redundant or less important channels from convolutional layers, thereby reducing the number of parameters and the computational cost without significantly sacrificing segmentation performance. Among various criteria, the L1-norm of convolutional filters is widely used to evaluate channel importance due to its simplicity and effectiveness. Specifically, the L1-norm of each output channel is calculated by summing the absolute values of its weights, and channels with the smallest L1-norms are considered less informative. This approach is based on the assumption that filters with smaller weight magnitudes contribute less to the network's representational power. After ranking all channels in each layer by their L1-norms, a predefined proportion of the least important channels are pruned. This process not only removes the corresponding filters in the current layer but also requires adjusting the input channels of subsequent layers to maintain structural consistency.
The channel pruning workflow typically begins with training the original network, such as UNet or its variants, to achieve satisfactory segmentation accuracy. Once a well-trained baseline model is obtained, the L1-norm of each channel is computed across all convolutional layers. Channels are then ranked, and a pruning mask is generated according to the desired pruning ratio. Careful structural adjustment is necessary, particularly for architectures with skip connections or feature concatenations, as in UNet and UNet++. In such cases, pruning must ensure that the number of channels matches across encoder and decoder paths, and concatenated feature maps remain compatible. After pruning, the network is fine-tuned with a lower learning rate to help recover any performance loss and adapt to the new, more compact architecture. This fine-tuning step is essential for restoring segmentation accuracy and ensuring the pruned model remains effective for clinical applications.

While channel pruning based on the L1-norm is straightforward and computationally efficient, it does have certain limitations. This method does not directly consider the interactions between channels or the dynamic behavior of channels during inference, which may limit its effectiveness in some scenarios. Moreover, aggressive pruning may result in the loss of crucial structural information, especially in complex segmentation tasks. Nevertheless, empirical results show that moderate channel pruning can significantly reduce model size and inference time with minimal impact on segmentation accuracy. Visualizations of L1-norm distributions typically reveal that pruned channels have much lower norms, supporting the validity of

the criterion, and feature map analysis demonstrates that important anatomical details are largely preserved after pruning. In summary, channel pruning using the L1-norm offers a practical balance between efficiency and accuracy, making it a valuable technique for deploying deep segmentation networks in real-time and resource-limited medical environments. Future work may explore adaptive pruning strategies, integration with other compression techniques, and extension to advanced network architectures.Fine-tuning the Pruned Model

3.4 Fine-tuning the Pruned Model

Fine-tuning is an essential stage in the channel pruning process, serving as a critical step to restore and potentially even enhance the performance of the pruned model. When channels are removed from the network, the representational capacity of the model is inevitably altered, which can lead to a temporary drop in segmentation accuracy or an increased risk of overfitting or underfitting. To address these challenges, the pruned model must undergo a carefully designed fine-tuning phase, during which it is retrained on the original training dataset.

During fine-tuning, the learning rate is typically reduced compared to the initial training phase. This lower learning rate helps stabilize the optimization process, allowing the model to gradually adapt to its new, more compact architecture without causing large fluctuations in the learned parameters. The fine-tuning process enables the remaining channels and weights to compensate for the information loss resulting from pruning, effectively redistributing the representational burden across the network. In many cases, this adaptation not only helps recover lost accuracy but can also improve the model's generalization ability by encouraging it to focus on the most salient features.

In addition to adjusting the learning rate, other training strategies may be employed during fine-tuning to maximize recovery. Data augmentation techniques, such as random cropping, flipping, and intensity variations, can be used to expose the network to a broader range of examples and prevent overfitting. Regularization methods, including dropout and weight decay, may also be applied to further enhance generalization. Moreover, early stopping based on validation performance can be implemented to avoid over-training and ensure optimal convergence.

Fine-tuning is especially important in medical image segmentation tasks, where the precise delineation of anatomical boundaries is critical and minor errors can have significant clinical implications. By retraining the pruned model on the original dataset, the network can relearn the subtle patterns and contextual cues necessary for accurate segmentation. This process ensures that the pruned model maintains high performance not only on the training data but also on unseen test samples, thereby supporting its robust deployment in real-world medical settings.

In summary, fine-tuning acts as a bridge between the initial pruning operation and the final deployment of the compressed model. It plays a pivotal role in restoring segmentation accuracy, enhancing generalization, and ensuring that the pruned network remains reliable and effective for clinical applications. The combination of pruning and fine-tuning thus offers a powerful framework for developing lightweight, high-performance models suitable for practical use in medical image analysis.

4. Experiments

4.1 Experimental Setup

To validate the effectiveness of the proposed L1-norm-based channel pruning method, extensive experiments were conducted on two widely used medical image segmentation datasets. All experiments

were implemented in PyTorch and performed on a workstation equipped with an NVIDIA RTX 3090 GPU. The main training settings are summarized in Table 1.

Table 1: Training Hyperparameters

Parameter	Value
Batch Size	256
Initial Learning Rate	1e-3
Optimizer	Adam
Learning Rate Decay	StepLR, 0.1
Epochs	200
Loss Function	Dice Loss + BCE Loss

4.2 Datasets

ISIC 2018 Skin Lesion Segmentation Dataset: Contains 2,594 dermoscopic images with pixel-level lesion annotations, characterized by diverse lesion shapes and ambiguous boundaries.

MICCAI 2015 BraTS Brain Tumor Segmentation Dataset: Consists of multi-modal MRI images with finely annotated tumor regions, suitable for evaluating the model's ability to segment complex structures.
All images were resized to 256×256, normalized, and augmented using random rotations and horizontal flips to enhance generalization.

4.3 Evaluation Metrics

Dice Coefficient: Measures the overlap between predicted and ground truth segmentation masks.Intersection over Union (IoU): Evaluates the ratio of the intersection to the union of predicted and true regions.Parameter Count: Indicates the total number of trainable parameters, reflecting model compactness.Inference Time: Average time required for a single forward pass, reflecting efficiency.FLOPs: Floating-point operations per forward pass, indicating computational complexity.

4.4 Experimental Procedure

Baseline Training: Both UNet and UNet++ were trained from scratch on the selected datasets to establish baseline performance.

Channel Importance Evaluation: The L1-norm of each channel in the convolutional layers was calculated and ranked.

Pruning Implementation: Various pruning ratios (20%, 40%, 60%) were tested, removing channels with the lowest L1-norm scores.

Structural Adjustment: Compatibility of skip connections and concatenation operations was ensured after pruning.

Fine-tuning: Pruned models were fine-tuned with a lower learning rate to recover potential accuracy loss.
Performance Evaluation: The pruned and fine-tuned models were evaluated on the test set using all metrics.

4.5 Quantitative Results

Table 2: UNet Results on ISIC 2018.

Pruning Ratio	Dice (%)	IoU (%)	Inference Time (ms)	FLOPs (G)
0%	89.5	81.2	45	29.8
20%	89.2	80.8	38	24.1
40%	88.3	80.1	28	17.7
60%	86.7	78.0	19	11.5

Table 3: UNet++ Results on BraTS

Pruning Ratio	Dice (%)	IoU (%)	Inference Time (ms)	FLOPs (G)
0%	90.2	82.4	56	36.7

Pruning Ratio	Dice (%)	IoU (%)	Inference Time (ms)	FLOPs (G)
20%	89.8	81.9	47	29.2
40%	89.0	81.0	33	21.6
60%	87.4	79.1	23	13.9

4.6 Ablation Study

To further analyze the impact of different pruning strategies, several ablation studies were conducted:
1.Pruning Location Comparison
We compared pruning applied to the encoder, decoder, and the entire network. Results are shown in Table 4.
2.Pruning Criterion Comparison
We compared L1-norm and L2-norm based pruning under the same pruning ratio (40%). Results are in Table 5.
3.Effect of Fine-tuning
We compared performance with and without fine-tuning after pruning. Results are in Table 6.

Table 4: Effect of Pruning Location on UNet Performance (ISIC 2018)

Pruning Location	Dice (%)	Params (M)
Encoder	88.6	21.2
Decoder	88.1	20.8
Entire Network	88.3	18.6

Table 5: Comparison of L1 and L2 Norm Pruning (40% Pruning, UNet, ISIC 2018)

Criterion	Dice (%)	IoU (%)	Params (M)
L1-norm	88.3	80.1	18.6
L2-norm	88.1	79.9	18.7

Table 6: Effect of Fine-tuning after Pruning (40% Pruning, UNet, ISIC 2018)

Fine-tuning	Dice (%)	IoU (%)
No	83.7	75.3
Yes	88.3	80.1

4.7 Discussion

The experimental results demonstrate that L1-norm channel pruning substantially reduces model size and computational cost with minimal loss in segmentation accuracy. Ablation studies confirm the robustness of the proposed approach across different pruning locations and criteria. Fine-tuning after pruning is essential for restoring performance. Overall, the method is well-suited for deployment in resource-constrained or real-time clinical environments.

5. Conclusions

This study proposes a practical channel pruning strategy to optimize UNet for efficient medical image segmentation. By eliminating redundant channels while preserving the network's structural integrity, the method reduces computational demands and accelerates inference, making it suitable for deployment in real-time and resource-constrained environments. Experimental results confirm that the pruned model retains competitive segmentation accuracy while achieving notable improvements in speed and compactness. Future work will explore adaptive pruning techniques, integration with other model compression strategies, and applications to 3D segmentation and multi-modal medical imaging.

References

[1] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer international publishing, 2015: 234-241.

[2] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4. Springer International Publishing, 2018: 3-11

[3] Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient convnets[J]. arXiv preprint arXiv:1608.08710, 2016.

[4] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[J]. Advances in neural information processing systems, 2015, 28.

[5] He Y, Zhang X, Sun J. Channel pruning for accelerating very deep neural networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1389-1397.

[6] LeCun Y, Denker J, Solla S. Optimal brain damage[J]. Advances in neural information processing systems, 1989, 2.

[7] Molchanov P, Tyree S, Karras T, et al. Pruning convolutional neural networks for resource efficient inference[J]. arXiv preprint arXiv:1611.06440, 2016.

[8] Persand K, Anderson A, Gregg D. Taxonomy of saliency metrics for channel pruning[J]. IEEE Access, 2021, 9: 120110-120.