Biomedical Data Science
2025, VOL. 5, NO. 3, 174-178
DOI: 10.47297/wspbdsWSP2752-630536.20250503

Knowledge Distillation-Based Model Lightweight for Medical Image Segmentation

Yimeng Xu

College of Psychology and Cognitive Sciences, East China Normal University, Shanghai, 200062, China

ABSTRACT
Accurate delineation of anatomical structures via medical image segmentation constitutes a fundamental prerequisite for clinical diagnostics and therapeutic planning. However, the exorbitant parameter footprint in contemporary deep segmentation architectures imposes prohibitive demands on hardware resources, severely curtailing their viability on edge platforms with computational constraints. This research presents a knowledge distillation-driven framework for model compression, leveraging R2U-Net as the knowledge source and MobileNetV2 as the compact recipient. Through novel boundary-constrained distillation (BCD) incorporating both hard and soft boundary learning mechanisms, we facilitate efficient knowledge transfer while preserving segmentation fidelity. This methodology significantly curtails computational overhead and enhances operational efficiency in resource-limited clinical settings.

KEYWORDS
Knowledge distillation; Medical image segmentation; Lightweight; Neural network

1 Introduction

As science advances, medical imaging - assisted diagnosis is increasingly prevalent. Modern diagnostic workflows leverage multi-parametric imaging systems - including magnetic resonance imaging (MRI), positron emission tomography (PET), and computer tomography (CT) - to generate complementary clinical data. These images enable doctors to make more accurate diagnoses by integrating diverse information.

The primary objective of medical image segmentation involves the accurate identification and extraction of particular structures or areas within medical imaging from surrounding backgrounds, facilitating healthcare professionals in diagnostic assessment and therapeutic planning.Medical image segmentation is a key task in medical image processing with broad applications in radiology, pathology, neuroscience, etc., such as lesion detection, organ positioning, and treatment monitoring. During image segmentation, computer - aided diagnosis and treatment systems can significantly improve disease diagnosis and treatment efficiency. Nevertheless, models with extensive parameter counts inevitably require substantial computational memory, graphics processing capabilities, and energy resources, creating significant implementation barriers for resource-limited edge computing environments such as wearable technology platforms.

This limits the clinical application of image segmentation.

Therefore, reducing model computational volume while maintaining performance stability is of great research significance. Conventional medical image segmentation algorithms, including threshold segmentation, region growing, edge detection, and mathematical morphology, often divide images into regions based on spatial, gray - value, and texture features. These traditional algorithms offer strong interpretability and high computational efficiency but have limited generalization ability and difficulty in clinical application due to their dependence on manually designed features. The field of medical image segmentation has witnessed progressive enhancements in accuracy and computational efficiency, corresponding directly with advancements in artificial intelligence methodologies, particularly Convolutional Neural Network frameworks. Unlike traditional methods, deep - learning - based segmentation does not rely on manually designed features. Neural networks can automatically learn the features needed for segmentation tasks, making deep - learning - based medical image segmentation a hot research topic.

2 Literature review

In the past few years, significant advancements in deep neural architectures have been observed, with convolutional neural networks (CNNs) emerging as the predominant paradigm. Convolutional neural network is a deep learning model mainly used to process high-dimensional data such as images and videos. This system replicates the functioning of the human visual system, employing multiple convolutional and pooling layers to isolate input data features and execute classification or regression functions via fully connected layers. At its core, CNN is centered on the idea of local perception and weight sharing. In the convolutional layer, a set of filters (also known as convolutional kernels) are used to perform convolutional operations on the input local area to extract local features, so that the texture, edges and other low-level features of the image can be captured. Through the stacking of multiple convolutional layers, the network can gradually extract more abstract features, thereby realizing the high-level semantic understanding of the image.

Back in 2015, Long and colleagues suggested the inaugural comprehensive fully convolutional neural network (FCN) for predicting at the pixel scale. FCN accomplishes semantic segmentation at the pixel level by categorizing each pixel in the input image into distinct groups. Differing from conventional convolutional neural networks, FCN substitutes the fully connected layer with a fully convolutional one to preserve the spatial configuration of the input image. The result of the completely convolutional layer is a feature map identical in size to the input image, with each pixel representing a specific region of the input image. Through the stacking of multiple fully convolutional layers, the network can gradually learn the high-level semantic features of the input image and perform pixel-level classification on the feature map. FCN fully utilizes the parallelism of convolutional operations to accelerate the calculation speed of the model and can process input images of any size without being limited by a fixed size. It also retains the spatial structure information of the input image and can better capture the details and context information in the image. FCN is not sensitive to detail information and the segmentation is not fine enough.

Addressing this matter, Ronnerberger and colleagues suggested a U-Net model for image segmentation utilizing FCN. It was the first time that CNN was applied to medical image segmentation tasks. Employing a symmetric encoder-decoder topology, U-Net facilitates precise segmentation by progressively abstracting and reconstructing hierarchical features, fuses shallow and deep features, and achieves more accurate segmentation. The rise of U-Net has significantly advanced the creation of deep learning-based medical image segmentation.

Utilizing U-Net as a foundation, scientists have introduced a range of enhanced algorithms aimed at boosting the precision and effectiveness in segmenting medical images. For example, the DeepLab algorithm introduces atrous convolution and conditional random fields to improve the accuracy and spatial continuity of segmentation. The Mask R-CNN algorithm was proposed in 2017 and combines object detection and segmentation to achieve pixel-level segmentation. The technique has proven effective in segmenting medical images and is capable of concurrently displaying both the object's location and the segmentation mask. In the encoder part, as the depth of the network increases, the receptive field of the convolutional kernel continuously expands, thereby extracting more rich feature information; the decoder part is intended to restore the feature information of the image and fuse the features extracted by the encoder through convolution and upsampling. During the year 2018, Alom and colleagues suggested R2U-Net, which derived from U-Net, utilizing recursive residual convolutional layers for encoding and merging the benefits of U-Net, residual networks, and RCNN. Zhou et al. in 2019 proposed a significant enhancement to U-Net, named U-Net++, through the redesign of skip connections and the consolidation of features across various semantic scales in the decoder.

Although medical image segmentation technology has made remarkable progress in improving the efficiency of medical image analysis, there are still some problems that have not been solved, which limit its wide application in practical applications. For example, the number of model parameters and computational costs are usually too large, making it difficult for the model to be deployed on mobile devices or in clinical environments. This limits the application of real-time medical image segmentation on portable devices because traditional models may require a large amount of computing resources and memory, which does not meet the requirements of the medical field for lightweight and fast response. [10]

Some studies have achieved lightweight by directly designing low-redundancy network models, such as Zhang et al. who combined channel rearrangement with pointwise grouped convolution to propose Shuffle Net; Howard et al. who used depthwise separable convolution with unique network structures to propose Mobile Net for mobile devices; Squeeze Net, etc. Reducing the number of model parameters and the size of the structure can also achieve this goal through model compression techniques, and the commonly used methods include channel pruning, knowledge distillation, and quantization.

Channel pruning is the most direct method. It refers to removing redundant neurons in the neural network, which can improve the generalization ability and performance of the model. The key to channel pruning is to evaluate the parts that can be pruned in the model, which can be divided into structured pruning and unstructured pruning. Quantizing reduces memory requirements and computation complexity by mapping model parameters and activation values to low accuracy, thus increasing energy efficiency. For example, the model is compressed by clustering and sharing weights. The quantization methods include linear and nonlinear quantization, and fixed or variable precision can be used. Knowledge distillation is to construct a small network (student network) and make it learn the "knowledge" of the large network (teacher network). The current innovations mainly focus on redefining knowledge and designing the loss function.

Among them, knowledge distillation, as a new model compression method, has become a research hotspot and focus in the field of deep learning. Hinton's pioneering distillation paradigm enables compact model development through mimicry of complex pre-trained networks' probabilistic outputs. Compared with other model compression technologies, knowledge distillation has unique advantages. Pruning may affect the expression ability of the model; quantization may lead to certain accuracy losses. While knowledge distillation achieves more effective model lightweight by transferring knowledge while maintaining the performance of the model.

To sum up, although convolutional neural networks have achieved great success in image segmentation tasks with excellent feature extraction and expression abilities, they need to be lightweighted for practical applications. At the same time, the feature representation ability of semantic segmentation lightweight networks is limited, which often restricts their performance ceiling. Therefore, it is very crucial to transfer the feature representation ability of the teacher model to the student model through knowledge distillation technology.

3 Method

3.1 Dataset processing and splitting

The study utilizes the ISIC 2019 dataset, which consists of 2594 images of skin lesions, encompassing various diagnostic categories such as melanoma, melanocytic nevus, and others. This dataset provides a robust platform for training and validating our proposed model. The preliminary data processing stage includes standardizing image dimensions to 224x224 pixels, employing data augmentation techniques to counteract overfitting issues, and adjusting pixel intensity values to promote more efficient model training.The dataset was split into training and testing sets using stratified train-test split in an 8:2 ratio, maintaining the original class distribution and preventing sampling bias.

3.2 Teacher and student model construction

Teacher Model (R2U-Net) Model Structure: R2U-Net, an improved U-Net model with a recurrent structure, enhances feature representation by reusing extracted features. It comprises three main components: encoder, decoder, and recurrent structure. The encoder component of the network captures complex features by employing convolutional and pooling operations, thereby diminishing the dimensionality of the feature maps and augmenting the number of channels. Subsequently, the decoder module enhances the spatial resolution through upsampling and convolutional processes. Additionally, the recurrent architecture integrates features extracted at various encoder and decoder stages, thereby amplifying the representational capacity of the features.

Model Training: The training of the teacher model was conducted employing the Adam optimizer, set at a learning rate of 0.001, alongside the cross-entropy loss function as the performance metric. This training regimen spanned over 10 epochs, where each epoch entailed processing the full training dataset, thereby enabling the model to internalize the characteristics of skin lesion images and encode the acquired insights within its parameters.

Student Model (MobileNetV2)

Model Structure: MobileNetV2, a lightweight neural network architecture for mobile and resource-constrained environments, uses depthwise separable convolutions to reduce computational complexity and parameters. It decomposes standard convolutions into depthwise and pointwise convolutions. Additionally, it incorporates linear bottleneck and inverted residual structures for enhanced performance. The student model is based on MobileNetV2, adjusted to match the teacher model's output dimensions.

Model Initialization: To improve initial performance and training stability, the student model was initialized with pre-trained MobileNetV2 weights. Pre-trained on large datasets like ImageNet, these weights provide general feature extraction capabilities.

3.3 Knowledge distillation method

The process of knowledge distillation facilitates the transference of insights from the teacher model to the student model, empowering the student to approximate the teacher's performance with a reduced parameter count.The study employs hard and soft boundary distillation.

Hard Boundary Distillation (Hard BD): Hard boundary distillation targets marginal zones by deploying 3×3×3 convolutional operators for edge feature extraction from real labels. The boundary predictions for both the teacher and student models are derived through an element-wise multiplication with the boundary mask, followed by the computation of their Kullback-Leible (KL) divergence to determine the Hard Boundary Distillation (Hard BD) loss.

Soft Boundary Distillation (Soft BD): Soft BD addresses boundary uncertainty using Hausdorff distance transformations. The loss function considers differences between predicted and real boundaries and penalizes segmentation errors. In the distillation framework, real labels are replaced with teacher predictions, and student predictions are used to calculate the Soft BD loss with a logarithmic adjustment.

Total Loss Function: The total loss function combines Hard BD (LHBD) and Soft BD (LSBD) losses with a linearly decaying weight γ (starting at 1 and ending at 0.5). This balances the two distillation methods, focusing on Hard BD initially and increasing Soft BD weight over training.

3.4 Model training and optimization

Teacher Model Training: The teacher model R2U-Net was trained on the training set using the Adam optimizer with a learning rate of 0.001 for 10 epochs. Cross-entropy loss was calculated and backpropagated to update model parameters, with loss monitoring to ensure effective feature learning. Student Model Training: The student model MobileNetV2 was trained using knowledge distillation. The teacher model was fixed, and the student learned from the teacher's outputs. The Adam optimizer updated student parameters using the defined total loss function. Training lasted 10 epochs, with validation set performance assessment after each epoch.

3.5 Model evaluation and validation

Evaluation Metrics: The student model's performance was evaluated using accuracy, recall, F1-score, confusion matrix, Dice coefficient, Hausdorff distance, and loss. Accuracy measures correct classifications; recall measures positive sample identification; F1-score balances accuracy and recall; the confusion matrix visualizes classification results; Dice coefficient and Hausdorff distance assess segmentation performance, with Dice indicating overlap and Hausdorff measuring boundary precision; loss reflects output differences from real labels.

Validation Set Evaluation: The student model was evaluated on the validation set, comparing its performance with the teacher model to assess knowledge distillation effectiveness.

Test Set Evaluation: Final evaluation on the test set assessed the student model's generalization and practical applicability, providing insights into real-world performance.

4 Conclusion

4.1 Performance metrics

The experimental results indicate that the student model achieves a validation accuracy of 0.6033, with a corresponding F1 score of 0.5410, precision of 0.5672, and recall of 0.6033. These metrics demonstrate that while the student model exhibits slightly reduced performance compared to the teacher model, it maintains a commendable level of accuracy and precision, which is crucial for clinical applications.

4.2 Computational efficiency

Regarding computational effectiveness, the student model markedly surpasses the performance of the teacher model. With only 2, 235, 401 parameters and 2, 235, 401 FLOPs, the student model is substantially lighter, making it suitable for deployment on devices with limited computational resources. The decrease in the intricacy of the model directly stems from the process of knowledge distillation, which efficiently conveys crucial insights from the teacher model, eliminating the necessity for the student model to start learning anew.

4.3 Trade-offs and advantages

This research primarily balances the precision of segmentation with the effectiveness of computation. While the student model does not match the teacher model's performance metrics across the board, it offers a viable alternative that consumes significantly fewer computational resources. This trade-off is particularly advantageous in clinical settings where rapid, accurate segmentation is necessary, but computational resources are limited.

4.4 Distillation effectiveness

The effectiveness of the proposed distillation approach is evident in the student model's performance. The hard and soft boundary distillation strategies collectively contribute to the student model's ability to learn precise boundary delineation and maintain region continuity. This dual focus ensures that the student model not only segments lesions accurately but also does so in a manner that is biologically plausible and clinically meaningful.

5 Conclusion

The research introduces an innovative method for segmenting lightweight medical images through the process of knowledge distillation. By transferring knowledge from a sophisticated teacher model (R2U-Net) to a lightweight student model (MobileNetV2), we achieve a balance between segmentation accuracy and computational efficiency. The student model, despite having fewer parameters and lower computational requirements, demonstrates competitive performance in terms of accuracy and precision, making it suitable for real-world applications where computational resources are constrained.

The proposed method's effectiveness is demonstrated through a comprehensive set of performance metrics, showcasing the student model's capability to segment skin lesions with reasonable accuracy and efficiency. Subsequent research might delve into additional enhancements for the distillation procedure, explore its use in various medical image segmentation activities, and evaluate its practical applicability in real-world situations.

Overall, this research adds to the expanding domain of efficient medical image analysis tools, offering a path toward more accessible and practical computational solutions for clinical diagnostics.

References

[1] Huang L, Ruan S, Denux T. Application of belief functions to medical image segmentation: A review[J]. Information Fusion, 2023, 91: 737-756.

[2] Di Angelo L, Di Stefano P, Guardiani E. An automatic method for feature segmentation of human thoracic and lumbar vertebrae[J]. Computer Methods and Programs in Biomedicine, 2021, 210, 106360.

[3] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:3431-3440.

[4] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI),2015: 234-241.

[5] Huang X. M., He F. Y., Tang X. H., et al. A review of U-Net and its variants for medical image segmentation. Chinese Journal of Biomedical Engineering, 2022,41(05):567-576.

[6] Guerroumi N, Playout C, Laporte C, et al. Automatic segmentation of the scoliotic spine from MR images[C]. 16th IEEE International Symposium on Biomedical Imaging(ISBI). IEEE Computer Society, 2019: 480–484.

[7] Qadir H A, Shin Y, Solhusvik J, et al. Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor cnn always perform better[C]. 2019 13th International Symposium on Medical Information and Communication Technology(ISMICT). 2019: 1-6.

[8] Alom, Md Zahangir, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net)[C]//NAECON 2018-IEEE National Aerospace and Electronics Conference. IEEE, 2018: 228-233.

[9] Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and JianmingLiang. Unet++: Redesigning skip connections to exploit multiscale features inimage segmentation[J]. IEEE transactions on medical imaging, 2019, 39(6): 1856-1867.

[10] Wu, Huisi, Zebin Zhao, Jiafu Zhong, Wei Wang, Zhenkun Wen, and Jing Qin. Polypseg+: A lightweight context-aware network for real-time polyp segmentation[J]. IEEE Transactions on Cybernetics, 2022, 53(4): 2610-2621.

[11] Duan Y. C., Fang Z. Y., Zheng J. B. A review of neural network lightweight. Computer Science and Exploration, 2024:1-25.

[12] M. M. H. Shuvo, S. K. Islam, J. Cheng and B. I. Morshed, "Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review," in Proceedings of the IEEE, vol. 111, no. 1, pp. 42-91, Jan. 2023.

[13] Y. Liu, P. Xiao, J. Fang and D. Zhang, "A Survey on Image Classification of Lightweight Convolutional Neural Network," 2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Harbin, China, 2023, pp. 1-10.

[14] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 535–541.

[15] L. Wang and K. -J. Yoon, "Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3048-3068, 1 June 2022.