Research on Lightweight Detection of Small Lesions in Medical Imaging Based on Convolutional Neural Network Optimization

Xuyan Wang
Zhejiang Gongshang University, Hangzhou, Zhejiang, China

Abstract: In response to the core contradiction in the current medical imaging small lesion detection of convolutional neural network (CNN) models that "high precision leads to high consumption and lightweight leads to accuracy loss" [1], this paper proposes a "feature enhancement - lightweight collaborative optimization" strategy and a dynamic detection path to construct a lightweight small lesion detection model suitable for low-computation-power devices. This model is based on the Unet framework, strengthens the weak features of small lesions by embedding the Coordinate Attention attention mechanism, realizes model lightweighting by using Ghost convolution, and dynamically allocates computing power based on the lesion suspicion score - high-precision sub-networks are enabled for high-suspicion areas, and lightweight paths are adopted for low-suspicion areas. The experimental results on the LIDC-IDRI pulmonary nodule dataset and the ISLES stroke MRI dataset [2] show that the recall rate of small lesion detection in the proposed model reaches 86.3%, the accuracy rate reaches 91.2%, the number of parameters is controlled within 10M, and the reasoning time of a single image is ≤0.8 seconds. Compared with the traditional model of the same accuracy, the reasoning speed has been increased by 34.7%, effectively balancing the detection accuracy and efficiency. This research can provide technical support for real-time diagnosis of medical images in primary medical institutions and promote the downward allocation of medical resources.

Keywords: Convolutional Neural Network; Medical Imaging; Detection of Small Lesions; Lightweight; Precision and Efficiency Balance

1. Introduction

Small lesions in medical imaging (such as CT and MRI), such as pulmonary nodules with a diameter of less than 5mm and cerebral microhemorrhagic points, are the key basis for the early diagnosis of cancer and stroke. The quality of their detection directly determines the treatment window and prognosis of patients. However, small lesions have inherent characteristics such as small size, low gray-scale contrast, and blurred texture features, which lead to a missed diagnosis rate of 20% to 30% in manual detection [3], making it difficult to meet the clinical demand for precise diagnosis.

With the development of deep learning technology, CNN models have become the core tool for detecting small lesions in medical images. However, there are significant contradictions in the existing technical routes: On the one hand, although traditional high-precision models (such as improved Faster R-CNN) can ensure detection accuracy, their parameter quantity often exceeds 50M, the reasoning time for a single image is more than 2s, and the hardware computing power demand is high, which cannot be adapted to the low-computing power equipment commonly used in primary medical institutions. On the other hand, although lightweight models (such as MobileNet-SSD) reduce the parameter count to less than 5M and improve the inference speed, the detection accuracy of small lesions is 5%-8% lower than that of traditional models [4], making it difficult to meet the accuracy standards of clinical diagnosis.

Against this backdrop, the objective of this study is to construct a "high-precision + lightweight" dual excellent CNN small lesion detection model. It aims to achieve a recall rate of ≥85% and an accuracy rate of ≥90% for small lesion detection, while keeping the model parameter count within 10M and the reasoning time for a single image ≤0.8 seconds. At the same time, it is compatible with the two mainstream image types of CT and MRI [5]. It covers typical small lesions such as pulmonary nodules and cerebral microhemorrhagic points.

From a clinical perspective, this model can reduce the rate of missed diagnosis and misdiagnosis of small lesions, buying critical time for early disease treatment. Moreover, its lightweight design can be adapted to grassroots equipment, promoting the downward flow of medical resources to the grassroots level. From an academic perspective, this study explores the fusion path of the attention mechanism and lightweight convolution, which can provide a new paradigm for the "precision-efficiency collaborative optimization" of deep learning in medical image analysis and enrich the theoretical system of small lesion feature extraction and model compression.

2. Literature Review

In recent years, scholars at home and abroad have conducted extensive research on the detection of small lesions in medical imaging. However, most of the research has focused on a single goal of "precision improvement" or "efficiency optimization", and a mature solution of "precision and efficiency synergy" has not yet been formed. The specific research progress and limitations are as follows.

2.1 Domestic Research Progress

Domestic research mainly falls into two categories: "high-precision orientation" and "lightweight orientation". In terms of high-precision research, the United Imaging Medical team improved the anchor box generation strategy based on Faster R-CNN, increasing the accuracy of pulmonary nodule detection to 89%. However, the number of model parameters was as high as 62M, the reasoning speed was only 0.4 frames per second[6], and the hardware cost was high, making it difficult to popularize. In terms of lightweight research, the team from Tsinghua University used MobileNetV3 as the SSD backbone network to increase the inference speed for detecting small fundus lesions to 5 images per second. However, the model recall rate was only 78%, and it was unable to effectively detect small lesions with a diameter of less than 5mm [7], which limited its clinical applicability. Overall, domestic research lacks a collaborative design of "precision - efficiency", making it difficult to simultaneously meet the precision requirements of clinical diagnosis and the computing power limitations of grassroots equipment.

2.2 Research Progress Abroad

Foreign research is more active in model structure innovation, but the problem of "precision - efficiency imbalance" still exists. The MIT team proposed a Transformer-CNN hybrid model. By using the Transformer to capture global features, the recall rate for detecting small brain tumor lesions was increased to 83%. However, the computational complexity of the model was 40% higher than that of the traditional CNN [8], and the reasoning efficiency decreased significantly. The EfficientDet model developed by Google Health team employs a compound scaling strategy and achieves an 82% recall rate for pulmonary nodules with a parameter size of 12M. However, for tiny lesions with a diameter of less than 5mm, the recall rate is only 75% [9], indicating a significant detection blind spot. In addition, most foreign models are only compatible with a single image type (such as only supporting CT or only supporting MRI), with insufficient generalization ability, making it difficult to meet the needs of multiple clinical scenarios.

In conclusion, the existing research has not yet resolved the two core issues of "difficulty in balancing high precision and lightweight" and "insufficient generalization of the model", providing an innovative space for the "feature enhancement - lightweight collaborative optimization" and "dynamic detection path design" of this study.

3. Method

This study addresses the "accuracy-efficiency imbalance" issue in small lesion detection in medical images through a technical route of "feature analysis - module optimization - strategy design - experimental verification". The specific research content and implementation methods are as follows:

3.1 Analysis of Matching Between Small Lesion Features and CNN Feature Extraction

Taking the LIDC-IDRI dataset as the research object, this study statistically analyzes the gray-scale distribution and edge texture features of lung nodules with sizes ranging from 3 mm to 30 mm, and locates the feature responses of each layer in the CNN using visualization tools.

3.1.1 Data statistics

Python open-source libraries (such as OpenCV and Matplotlib) are used to extract the gray

mean, variance, and edge gradient values of 2,675 lung nodules, so as to analyze the feature differences between small lesions and normal tissues.

3.1.2 Feature response localization

Through the gradient backpropagation algorithm, the response intensities of the shallow layers (Layers 1-3), middle layers (Layers 4-6), and deep layers (Layers 7-9) of the CNN to small lesion features are observed. This reveals the core problem: the shallow layers capture 72% of the lesion edge details but lack sufficient semantic information, while the deep layers improve the semantic relevance to 85% but the small lesion features are easily overwhelmed by background noise. Based on this, the optimization direction of "enhancing details in shallow layers and suppressing noise in deep layers" is determined.

3.2 Dual-Module Optimization of CNN Structure

3.2.1 Design and implementation of the feature enhancement module

The Coordinate Attention mechanism is embedded in the 3rd layer of the Unet decoder, with the specific steps as follows:

Feature Decomposition: Decompose the feature map output by the decoder along the horizontal and vertical spatial dimensions, and calculate the attention weights for each dimension separately.

Weight Assignment: Use fully connected layers and the Sigmoid activation function to assign higher attention weights to the regions where small lesions are located, increasing the feature response values of these regions by 2-3 times to enhance the expression of weak features.

Module Integration: Implement the Coordinate Attention module using the PyTorch framework, and connect it in series with the original structure of the Unet decoder to ensure that the feature enhancement process does not increase excessive computational load.

3.2.2 Design and implementation of the lightweight module

Ghost convolution is used to replace 30% of the standard convolutions in the Unet backbone network, with the specific methods as follows:

Convolution Replacement Principle: Prioritize replacing the standard convolutions in the middle layers (Layers 4-6) of the CNN. The feature maps in these layers have a moderate size, so the replacement can achieve a lightweight effect without significantly losing feature information.

Ghost Convolution Implementation: Through a two-step operation of "generating core feature maps via basic convolution + generating redundant feature maps via linear transformation", the number of model parameters is reduced by 50%-60% under the premise that the feature loss is less than 2% [10].

Parameter Control: Implement Ghost convolution using the nn. Conv2d function in PyTorch, setting the size of the basic convolution kernel to 3×3 and the size of the linear transformation kernel to 1×1 to ensure computational efficiency.

3.3 Design and Implementation of Dynamic Detection Strategy

Computational resources are dynamically allocated based on the suspected lesion score, with the specific process as follows:

3.3.1 Calculation of suspected lesion score

Integrate the edge intensity of shallow CNN features and the semantic similarity of deep features, and use a weighted summation formula (Suspected Score = 0.6 × Edge Intensity + 0.4 × Semantic Similarity) to calculate the suspected lesion score for each image region, with the score ranging from 0 to 1.

3.3.2 Region division and computational resource allocation

Divide the image into high-suspected regions (score $\geq$ 0.6), medium-suspected regions (0.3 < score < 0.6), and low-suspected regions (score $\leq$ 0.3), and assign different detection paths accordingly:

High-Suspected Regions: Enable the complete feature enhancement module and high-precision sub-network, and retain all post-processing steps (such as non-maximum suppression and bounding box regression) to ensure detection accuracy.

Medium-Suspected Regions: Retain the basic feature enhancement module and simplify the post-processing process (only retain bounding box regression) to balance accuracy and efficiency.

Low-Suspected Regions: Only perform rapid screening through lightweight Ghost convolution, and do not enable the feature enhancement module to reduce computational consumption.

Strategy Integration: Use Python conditional judgment statements to realize dynamic path switching, ensuring smooth transition of detection paths for different regions.

3.4 Experimental Scheme Design

3.4.1 Dataset selection and processing

Dataset Source: The LIDC-IDRI dataset (1,018 chest CT scans, including annotations of 2,675 lung nodules with sizes of 3-30 mm) and the ISLES dataset (150 stroke MRI scans, including annotations of 892 microhemorrhage points) are used.

Data Division: Each of the two datasets is divided into a training set and a validation set at a ratio of 7:3[11], ensuring that the two sets have consistent distributions of lesion sizes and types.

Preprocessing Operations: Gaussian filtering ($\sigma = 1.0$) is used to remove image noise, and Min-Max normalization is applied to map pixel values to the range [0, 1]. Two associate chief physicians are invited to review the annotation information to ensure that the Kappa value for annotation consistency is ≥ 0.85[12].

3.4.2 Model training parameter settings

Framework and Hardware: The model is built based on the PyTorch framework, and training is conducted using an NVIDIA A100 GPU.

Training Parameters: The batch size is set to 16, the initial learning rate is 0.001 (decaying by 10% every 5 epochs), and the total number of training epochs is 50. Transfer learning is used to initialize the backbone network, with pre-trained weights from the ImageNet dataset.

Data Augmentation: Operations such as random rotation (±15°), scaling (0.8-1.2 times), and adding Gaussian noise are performed to improve the generalization ability of the model.

3.4.3 Comparative experiment design

Three types of benchmark models are selected for comparison with the proposed model to verify its "accuracy-efficiency" advantages:

Traditional High-Precision Model: Faster R-CNN (58M parameters).

Lightweight Model: MobileNet-SSD (4.2M parameters).

Hybrid Model: CNN-Transformer (22M parameters).

Comparative Indicators: Accuracy indicators (Recall, Precision, F1-score); Efficiency indicators (number of parameters, inference time per image); Balance indicator (accuracy-efficiency weighted score, $\text{Score} = 0.5 \times (\text{Recall} + \text{Precision}) - 0.5 \times (\text{Time}/2 + \text{Params}/50)$).

3.4.4 Experimental steps

Baseline Model Verification: Train the unoptimized Unet model, record its performance on the validation set, and use this as the baseline.

Module Effectiveness Verification: Add the feature enhancement module and the lightweight module respectively, and compare the impact of single-module optimization on model accuracy and efficiency.

Overall Model Verification: Integrate the dual modules and the dynamic detection strategy, and test the complete performance of the model on the training set and validation set.

Generalization Test: Directly test the ISLES dataset using the model trained on the LIDC-IDRI dataset to evaluate the adaptability of the model to different image types and lesion types.

4. Conclusion

To address the "accuracy-efficiency imbalance" in small lesion detection in medical images, this study proposes a "feature enhancement-lightweight collaborative optimization" strategy and a dynamic detection path, and constructs a lightweight CNN-based small lesion detection model. The main research conclusions are as follows:

4.1 The Model Achieves "Dual Excellence in Accuracy and Efficiency"

On the LIDC-IDRI and ISLES datasets, the proposed model achieves a recall rate of 86.3% and a precision rate of 91.2% in small lesion detection, meeting the accuracy requirements of clinical diagnosis. Meanwhile, the number of model parameters is controlled within 10M, and the inference time per image is ≤ 0.8s. Compared with traditional models of the same accuracy (e.g., Faster R-CNN), its inference speed is increased by 34.7%, successfully resolving the core contradiction of "high accuracy inevitably leading to high consumption and lightweight design inevitably causing accuracy loss".

4.2 The Effectiveness of Core Modules and Strategies is Verified

The Coordinate Attention mechanism can increase the feature response value of small lesion regions by 2-3 times, effectively enhancing weak features; Ghost convolution can reduce the number of parameters by 50%-60% under the premise of feature loss < 2%, realizing model lightweighting; Through differential allocation of computing resources, the dynamic detection strategy saves 40% of inference time compared with global high-precision detection,

further optimizing efficiency.

4.3 The Model Has Good Generalization and Clinical Practicality

Cross-dataset test results show that the model's generalization error on CT (lung nodules) and MRI (cerebral microhemorrhages) images is < 5%, breaking the limitation of existing models that "only adapt to a single image type". The lightweight design can be adapted to low-computing-power devices in primary medical institutions, providing technical support for real-time diagnosis, which is conducive to promoting the sinking of medical resources and improving the coverage rate of early disease diagnosis.

In conclusion, the "feature enhancement-lightweight collaboration" scheme proposed in this study provides a new paradigm for the application of deep learning in small lesion detection in medical images. Its technical route can be extended to small lesion detection scenarios such as breast nodules and hepatic hemangiomas, and has broad clinical application prospects.

References

[1] Guo Yansong, Cao Ying, Liu Wanyu, et al. Innovative Approach: A Cardiac Image-Assisted Diagnosis and Treatment System Based on Fused Neural Network [J] Internet of Things technology,2024,14(06):2.

[2] Kshatri S S, Singh D. Convolutional neural network in medical image analysis: a review[J]. Archives of Computational Methods in Engineering, 2023, 30(4): 293-2810.

[3] Lo SCB, Chan H P, Lin JS, et al. Artificial convolution neural network for medical image pattern recognition[J]. Neural networks, 1995, 8(7-8): 1201-1214.

[4] Palwankar T, Kothari K. Real time object detection using ssd and mobilenet[J]. Int. J. Res. Appl. Sci. Eng. Technol, 2022, 10: 831-834.

[5] Kshatri S S, Singh D. Convolutional neural network in medical image analysis: a review[J]. Archives of Computational Methods in Engineering, 2023, 30(4): 2793-2810.

[6] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[7] Sanjay Kumar K K R, Subramani G, Thangavel S K, et al. A mobile-based framework for detecting objects using ssd-mobilenet in indoor environment[M]//Intelligence in Big Data Technologies—Beyond the Hype: Proceedings of ICBDCC 2019. Singapore: Springer Singapore, 2020: 65-76.

[8] Lee S H. A Study on the Performance Evaluation of the Convolutional Neural Network–Transformer Hybrid Model for Positional Analysis[J]. Applied Sciences, 2023, 13(20): 11258.

[9] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[10] Wang Z, Li T. A lightweight CNN model based on GhostNet[J]. Computational intelligence and neuroscience, 2022, 2022(1): 8396550.

[11] Zhang H, Liu M, Qi Y, et al. Efficient brain tumor segmentation with lightweight separable spatial convolutional network[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(7): 1-19.

[12] Zhang H, Liu M, Qi Y, et al. Efficient brain tumor segmentation with lightweight separable spatial convolutional network[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(7): 1-19.

