A Comparative Analysis of Transfer Learning Methods for Medical Super-Resolution

Lingxiao Tang

School of Civil and Environmental Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China

23s054018@stu.hit.edu.cn

Abstract. Precisely and rapidly acquiring high-resolution medical images with a limited radiation dose is of great importance in clinical diagnosis. However, due to the challenges in gathering medical images and the high costs of developing a deep learning model from scratch, the super-resolution of medical images is still a tough problem. To deal with above, transfer learning is a favorable technique that leverages a previously trained model for a different but related task, significantly cutting down on the amount of training data and time needed for the downstream task. In this study, a pre-trained U-shaped Network++ (UNet++) model is initially trained using the ImageNet dataset to learning the general information. Subsequently, multiple finetuning methods are adopted to transfer the natural images super-resolution to the super-resolution of skin lesion images. 1279 manually annotated skin lesion images are selected for training and testing. The results indicate that transfer learning not only improve the convergence speed of deep learning models but also generate clearer high-resolution skin lesion images. Compared to the normal training group, the Mean Square Error (MSE) value was dropped by 24% to 70%. Notably, BitFit might lead to worse performance.

Keywords: Transfer learning; image super-resolution; deep learning.

1. Introduction

Restricted by the capabilities of photographic apparatus [1], communication bandwidth [2], scanning time and human safety concerns [3], the extracted images always contain noise that fail to meet the requirement of various downstream tasks. The Super-Resolution (SR) technique is being extensively applied in various applications within the field of Computer Vision (CV). It is intended to convert blurred, low-resolution, indistinct images into the high-resolution images. The existing SR approaches are grouped into three main types, including interpolation-based, construction-based and machine learning-based [4]. Among them, machine learning-based approaches are currently most prevalent thanks to their robust nonlinear representation capabilities and effective adaptability.

At present, digital medical imaging techniques such as Positron Emission Tomography (PET) scan, ultrasonography and Magnetic Resonance Imaging (MRI) are essential during medical consultation, offering clear and reliable reference for making accurate diagnose evaluations. Therefore, the quality of the medical images may affect the final diagnosis. Unlike natural image SR, medical images contain abundant physiological details [5]. It is important to ensure the exact and legible representation of details, and retain vital elements like organ tissue structures and lesion annotations. Additionally, processing speed is also a significant factor in medical SR. Since medical institutions often deal with large volume of images for a single patient. Park et al. proposed a modified U-shaped Network (Unet) model to train on the brain Computed Tomography (CT) images with suspected Parkinson's disease, forming a mathematical mapping between thinner-slice and thick-slice images [6]. This technique reduces artifacts and the radiation dose while better assisting doctors in identifying small lesions. Gu et al. presented Medical Images SR using Generative Adversarial Networks (MedSRGAN), which integrates with attention mechanism and adversarial loss [7]. It has been proven effective in reconstructing high-resolution medical image with superior texture details and realistic patterns, making MedSRGAN a valuable tool for improving medical imaging while maintaining low radiation doses.

Developing a deep learning model from sketch is time-consuming and labor-intensive. It demands not only extensive computational costs but also a meticulous finetuning process to achieve effective performance. Due to the limitations imposed by data privacy laws, coupled with high acquisition and processing costs, the datasets of medical images are typically scarce and fail to encompass the full range of pathological and clinical conditions. Thus, existing literature in CV tends to focus on the same network structures with different weights, which is also called transfer learning. A model with prior can be finetuned to satisfy the needs of specific downstream tasks, thus improving the performance and generalizability of deep learning networks [8]. At the same time, it can be run on lightweight mobile devices such as miniature unmanned aerial vehicle and cellphone with limited training time and processing power [9]. Transferring weights from natural images to medical images can alleviate the issue of restricted medical image datasets. However, there is lack of literature on this topic.

In this study, a deep learning SR network initially trained on large-scale natural image dataset International Skin Imaging Collaboration (ISIC 2016), and various finetuning methods are adopted to transfer the pre-trained model to medical images. The training process and performance of different transfer methods are compared and analyzed.

2. Method

2.1. Dataset

The ISIC2016 used in this study is the benchmark dataset for the classification and segmentation of skin lesion tissues, which was introduced by the ISIC in 2016 [10]. It consists of 1279 manually annotated skin pathology images, with 900 used to training and the remaining 379 for testing. This dataset is highly valuable for evaluating the capabilities of neural network algorithms in the automated diagnosis of skin lesions.

The images in the ISIC2016 have varying resolutions. For optimal use in image super-resolution tasks, all images are resized to 128×128 by the transform functions in Torchvision toolbox. Then, the low-resolution images are generated by employing the bicubic interpolation on the aforementioned images, which resolution is set as 32×32.

The pre-trained weight parameters are derived from the ILSVRC2012[11] dataset. The ILSVRC2012 dataset is a subset of ImageNet designed for image classification and object detection. It includes over 1.4 million natural images in total, with 1.2 million for training, 50000 for validation and 150000 for testing. Following the preprocessing method applied to the ISIC2016 dataset, natural images are resized before being fed into the network.

2.2. Architecture of the Network

Fig.1 depicts the overall architecture of UNet++. It builds upon the UNet by introducing redesigned skip pathway, dense skip connection and deep supervision [12]. Similar to UNet, encoder and decoder are two key components of UNet++, both constructed using convolutional neural networks (CNN). The encoder extracts the features from input images and transform them into feature maps at various scales through multiple CNN layers. The obtained feature maps are passed to the decoder to gradually restores the spatial resolution, ultimately generating prediction results of the same size as the original input images. The dash line in the U-shape structure represents the dense skip connection, aimed at reinforcing the linkage between encoder and decoder modules. To be specific, a single feature extractor is shared among all sub-layers in UNet++ and features at different levels are reconstructed by separate decoding paths. Taking into account that the goal of this paper is to improve image resolution rather than classification and segmentation, the deep supervision module is abundant and thus has been removed. In addition, a PixelShuffle layer and a convolutional layer are included following the decoder.

2.3. Transfer Learning Methods

As the parameters within the deep learning networks increased by orders of magnitude, the financial and computational costs of training a model from the beginning become very high. Therefore, transfer learning is an efficient way to address the above issues. In this study, several finetuning methods are selected to evaluate the effectiveness of transferring super-resolution of nature images to medical images.

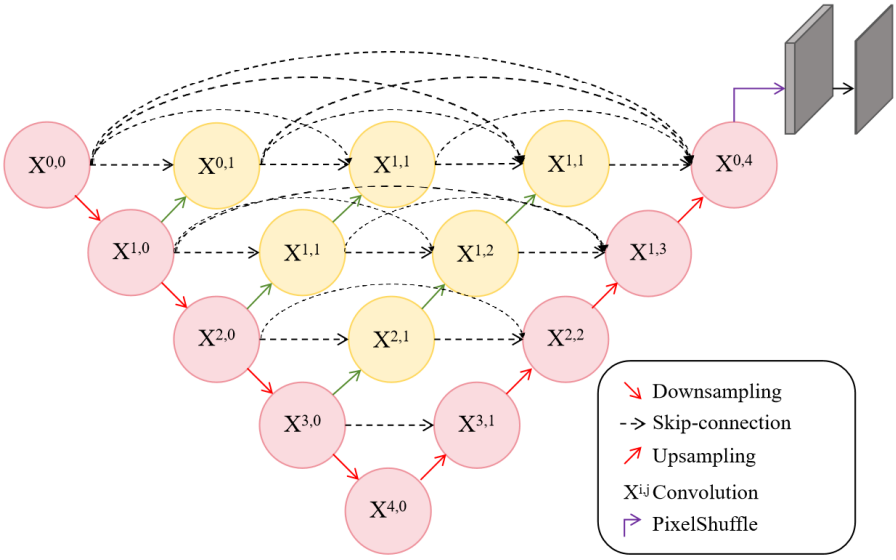

Fig. 1 Architecture of UNet++ [12].

Full-finetuning: Full-finetuning is the most widely used approach, containing the adjustment of all parameters of a previously trained model to fit a specific task. The previously trained model is initially trained on massive datasets, such as ImageNet. For downstream applications, the pre-trained parameters are firstly loaded and the model is further trained with the specific dataset. The parameters of all layers are fine-tuned to capture the nuances of the downstream task. Although full-finetuning require substantial computational resources and time, it improves the performance by fully utilizing the potential of the previously trained model and optimizing it with the characteristics of downstream requirement.

BitFit: BitFit is a relatively simple finetuning method where all weight terms are frozen, allowing only the bias terms to be updated when applied to downstream tasks [13]. It can be applied to various deep learning models, offering practical solutions for large-scale deployment and applications on edge devices due to its small parameter size.

AdaptFormer: Adapt-based methods introduce adapter to the convolution or transformer layers and only fine tune the adapter for model optimization. AdaptFormer incorporates AdaptMLP modules, only adding less than 2% extra parameters, with the remaining over 98% parameters shared across different tasks [14]. AdaptMLP consists of two sub-branches. The Multilayer Perceptron (MLP) layer in the left pathway maintains similarity to the original Vision Transformers (ViT) network, whereas the right pathway includes both downsample and upsample operations. It's plug-and-play and adaptable for integration into existing popular deep learning networks. Considering that UNet++ used in this study exhibits significant structure differences compared to the Transformer, the adapter cannot be directly integrated into the MLP layer that follows the Transformer Block. Thus, the adapter is added to each convolutional layer of the encoder.

3. Experiments and Results

3.1. Experimental Configurations

To better guide gradient descent, mean square error (MSE) is chosen as the loss function and its mathematical representation is listed in Eq. (1).

$$loss = \frac{1}{N}\sum_{i=1}^{N}(I_i - R_i)^2 \quad (1)$$

where N describes the complete count of samples, I_i stands for the reconstructed results of the deep learning networks, R_i is the reference high-resolution image.

The adaptive moment estimation (Adam) with β_1=0.9 is applied in optimization. After extensive testing, all groups are trained for 25 epochs using a learning rate (LR) of 1.0×10^{-3}. Moreover, the Graphics Processing Unit (GPU) receives 24 data in a single batch to enhance computational efficiency and stability. The training and assessment are performed on a server installed with a Xeon(R) CPU of DDR4 system memory and two parallel NVIDIA RTX3090 GPUs with 48GB in total of DDR6x graphics memory. The system operates on Windows 10 with Python 3.7 and utilizes Pytorch 1.13.0 to accelerate the training process on the GPU.

3.2. Results

Fig.2 exhibits the variation of loss value for different finetuning methods, and the loss value for each group is recorded starting from the first epoch. It can be clearly seen that the group without transfer learning experience a more pronounced gradient descent phase compare to those with finetuning modification. After 15 epochs, the results of UNet++ on both training as well as test set show convergence, and the minimum loss values are presented in Table 1. The outcomes of full-finetuning method demonstrate that finetuning operation actually improves the effectiveness of deep learning networks while achieving rapid convergence. In terms of the specific data, the MSE for training set and test set decreased from 1.17×10^{-3} and 1.39×10^{-3} to 3.56×10^{-4} and 4.38×10^{-4}, respectively. Among the groups with finetuning techniques, the full-finetuning shows the best performance, followed by AdaptFormer (8.88×10^{-4} and 1.37×10^{-3}), with BitFit performed the worst (2.35×10^{-3} and 3.48×10^{-3}). Simply focusing on statistical metrics is insufficient, Fig. 3 showcases the reconstructed images post super-resolution. In comparison, it reveals that the reconstructed images from the normal training are quite blurry, whereas those obtained through transfer learning clearly show the skin lesions. Noticeably, the reconstructed images in BitFit group differ in color from the reference high-resolution images, showing a somewhat darker. Among these finetuning methods, AdaptFormer produced the best super-resolution results, followed by Full fine-tuning and BitFit.

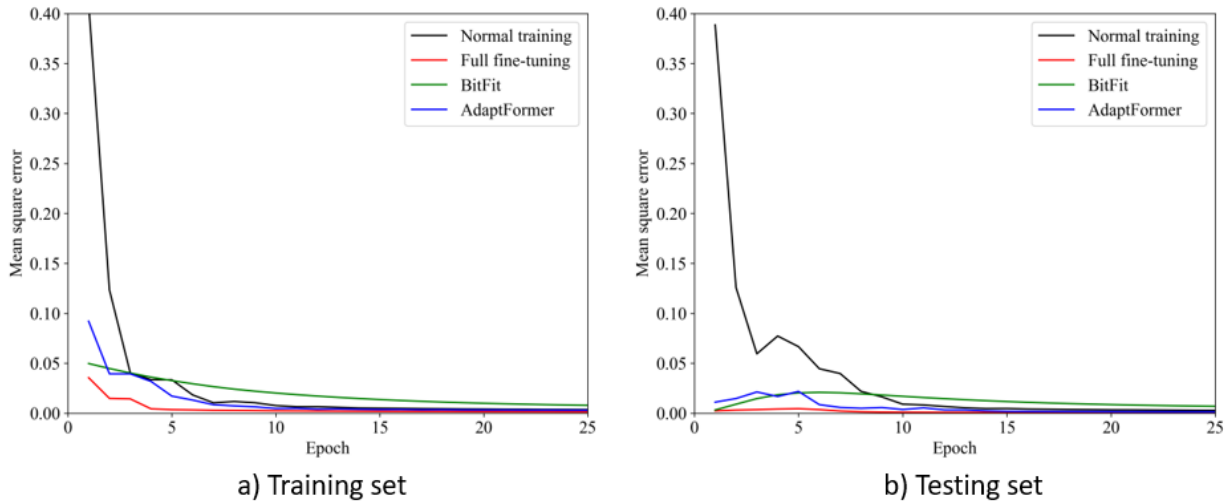

Fig. 2 Result of MSE value during training process (Figure Credit: Original).

Table 1. Quantitative results of the super-resolution tasks measured by MSE

Dataset	Transfer Methods			
	None	Full fine-tuning	BitFit	AdaptFormer
Training set	1.17e-3	3.56e-4	2.35e-3	8.88e-4
Test set	1.39e-3	4.38e-4	3.48e-3	1.37e-3

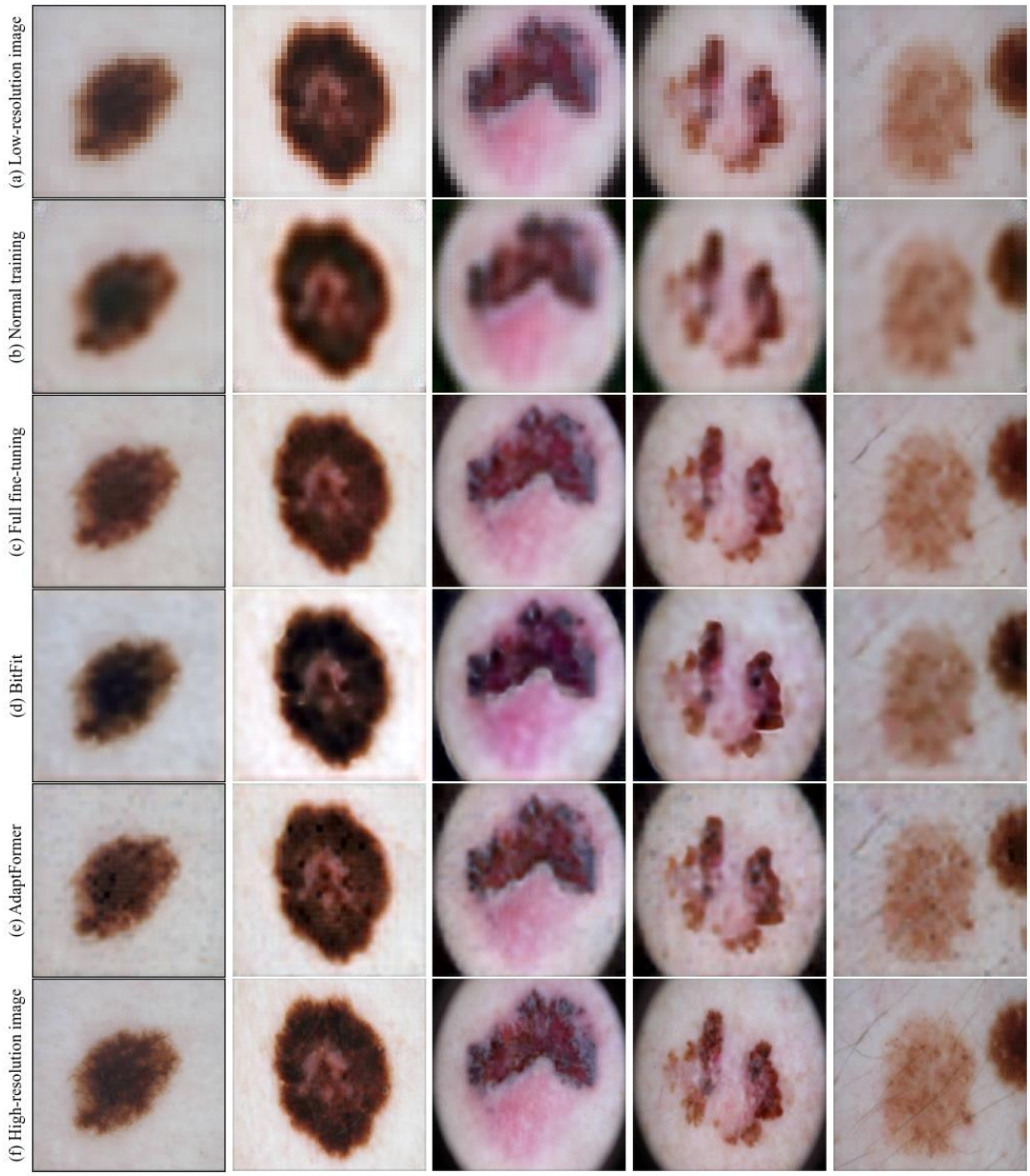

Fig. 3 Visual evaluation of the reconstructed medical results (Figure Credit: Original).

4. Discussion

In scenarios where both accuracy and speed are important, transfer learning methods can better meet practical application requirements. Specifically, Full-finetuning provides the superior performance, but it involves finetuning all parameters during transfer learning, which consumes a large amount of time and GPU memory in the backpropagation process. This method is suitable for scenarios that demand the extreme level of precision. BitFit performs the worst because it only finetunes the bias terms and fails to extract the features represented by the weight terms. It's extremely

low computational cost is ideal for application in lightweight devices. The performance of AdaptFormer falls between the two. Due to its unique plug-and-play unit, it can be applied to a broader range of deep learning frameworks. From the comparison of SR images between AdaptFormer and BitFit, it can be seen that bias terms tend to extract detail information such as brightness and hue in medical image super-resolution task.

However, it should be noted that the ISIC2016 dataset only cover a subset of medical image types and do not comprehensively represent all possible medical imaging scenarios. The finetuning approaches chosen in this study are not exhaustive, thus further research could investigate a broader range of fine-tuning techniques. After transfer learning, the model shows splendid results on the target dataset but suffers a significant drop in performance on the original dataset. This aspect is also worth investigating.

5. Conclusion

This study transfers a natural image super-resolution pre-trained UNet++ model to a skin lesion dataset. A total of three finetuning approaches are chosen, including Full-finetuning, BitFit and AdaptFormer. The trustworthy and resultful outcomes of UNet++ refer to the viability of using the transfer learning approach for SR reconstructing improved detail in low-resolution images obtained from clinical practice. In terms of performance, Full-Finetuning is the best, followed by AdaptFormer, with BitFit being the least effective. Nevertheless, when considering both accuracy and inference speed, AdaptFormer and BitFit are more appropriate choices. Future investigations will center on the performance on the original dataset after finetuning.

References

[1] Mingfeng Jiang, Minghao Zhi, Liying Wei, et al. FA-GAN: Fused attentive generative adversarial networks for MRI image super-resolution. Computerized Medical Imaging and Graphics, 2021, 92: 101969.

[2] Zhisheng Lu, Juncheng Li, Hong Liu, et al. Transformer for single image super-resolution. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022: 457-466.

[3] Lepcha Dawa Chyophel, Goyal Bhawna, Dogra Ayush, et al. Image super-resolution: A comprehensive review, recent trends, challenges and applications. Information Fusion, 2023, 91: 230-260.

[4] Jiang Mingfeng, Zhi Minghao, Wei Liying, et al. FA-GAN: Fused attentive generative adversarial networks for MRI image super-resolution. Computerized Medical Imaging and Graphics, 2021, 92: 101969.

[5] Li Yufei, Bruno Sixou, and Francois Peyrin. A review of the deep learning methods for medical images super resolution problems. Irbm, 2021, 42(2): 120-133.

[6] Park Junyoung, Hwang Donghwi, Kim Kyeong Yun, et al. Computed tomography super-resolution using deep convolutional neural network. Physics in Medicine & Biology, 2018, 63(14): 145011.

[7] Gu Yuchong, Zeng Zitao, Chen Haibin, et al. MedSRGAN: medical images super-resolution using generative adversarial networks. Multimedia Tools and Applications, 2020, 79: 21815-21840.

[8] Iman Mohammadreza, Hamid Reza Arabnia, Khaled Rasheed. A review of deep transfer learning and recent advancements. Technologies, 2023, 11(2): 40.

[9] Voghoei Sahar, Tonekaboni Navid Hashemi, Wallace Jason, et al. Deep learning at the edge. International Conference on Computational Science and Computational Intelligence, 2018: 895-901.

[10] Gutman David, Codella Noel, Celebi Emre, et al. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv, 2016: 1605.01397.

[11] Deng Jia, Dong Wei, Socher Richard, et al. Imagenet: A large-scale hierarchical image database. IEEE conference on computer vision and pattern recognition. 2009: 248-255.

[12] Zhou Zongwei, Rahman Siddiquee Md Mahfuzur, Tajbakhsh Nima, et al. Unet++: A nested u-net architecture for medical image segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018: 3-11.

[13] Zaken Elad Ben, Shauli Ravfogel, Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv, 2021: 2106.10199.

[14] Chen Shoufa, Ge Chongjian, Tong Zhan, et al. Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems, 2022, 35: 16664-16678.