Enhancing Lung Segmentation Under Data Constraints with Transfer Learning

YoungSeok Kim YOUNGSEOK.KIM0301@GMAIL.COM
MODULABS
Seoul, Republic of Korea

Abstract

Chest X-rays (CXR) are crucial for diagnosing lung diseases, yet accurate segmentation remains challenging due to overlapping imaging features and data scarcity. This study evaluates the performance of U-Net, VGG16-U-Net, and ResNet50-U-Net using a limited dataset of 100 CHN chest X-ray images, with pre-trained weights applied to the encoder of each model. Results revealed that VGG16-U-Net achieved strong performance, while ResNet50-U-Net exhibited overfitting despite high training accuracy. These findings emphasize the variability in the effectiveness of transfer learning methods on small datasets, underscoring the importance of understanding the specific characteristics of each model. Importantly, the study highlights the need for research designs tailored to the dataset size and characteristics, offering insights into optimizing model adaptability and generalizability in clinical applications. This work demonstrates the potential of transfer learning while calling for a methodological approach that aligns with the constraints of small-scale datasets.

Keywords: Lung Segmentation, Chest X-ray, Deep Learning, U-Net, Transfer Learning, VGG16, ResNet50, Medical Image Analysis

1. Introduction

Chest X-ray (CXR) is a crucial imaging modality for diagnosing various lung diseases, such as pneumonia, tuberculosis, and pneumothorax (Aberle et al., 2011). However, accurate interpretation of CXR images can be challenging due to the similar radiological features presented by different respiratory diseases (Smith and Doe, 2020). Recently, deep learning-based image segmentation techniques have helped overcome these limitations and improve the accuracy and efficiency of medical image analysis (Litjens et al., 2017; Esteva et al., 2017). In particular, U-Net has emerged as a widely adopted deep learning model in medical image segmentation, effectively extracting features and providing precise segmentation results through its encoder-decoder structure (Ronneberger et al., 2015).

However, the performance of deep learning models is heavily based on large-scale, accurately labeled data. In the medical imaging domain, data collection and labeling are expensive and labor intensive, and data acquisition is further restricted by privacy concerns (**?**). To address this issue, transfer learning has gained significant attention. Transfer learning leverages knowledge from a model pre-trained on a large-scale dataset (e.g. ImageNet) to achieve relatively high performance even with a small dataset (Pan and Yang, 2010; **?**). Recently, studies have reported an improved segmentation performance by applying pre-trained weights to the encoder based on the U-Net architecture (Zhou et al., 2018).

However, in real-world clinical settings, situations often arise where the amount of data is extremely limited, such as with rare diseases or data collection in localized regions. The efficacy of transfer learning in these scenarios requires systematic analysis, which is still lacking. Therefore, this study aims to verify the effectiveness of transfer learning in a limited data environment by comparing the performance of the original U-Net and hybrid U-Nets based on transfer learning (VGG16-U-Net, ResNet50-U-Net) using only 100 CHN CXR images.

2. Methods

2.1. Data

A total of 539 chest X-ray (CXR) images were available for this study. However, considering a scenario with limited training data, 100 training images were restored and extracted to simulate data scarcity. The remaining images were divided into 49 validation sets and 54 test sets for evaluation. All personally identifiable information was de-identified from all images to protect patient privacy. The dataset contains a mix of normal and various lung diseases, though a detailed breakdown of the specific pathologies is unavailable due to the de-identification process. The images have varying resolutions and were captured using different X-ray machines.

2.2. Data Preprocessing

Image normalization and resizing were applied in parallel. Specifically, the image size was unified to 128x128, and pixel values were normalized to between 0 and 1.

2.3. Model Architectures

2.3.1. U-Net

The original U-Net architecture, as proposed by Ronneberger et al. (2015), was utilized in this study. All network weights were initialized randomly to ensure unbiased training and allow the model to learn from scratch.

2.3.2. VGG16-U-Net

The weights of the VGG16 model pre-training on the ImageNet data set were applied to the encoder part (Simonyan and Zisserman, 2014). The decoder part follows the U-Net structure, and the weights were randomly initialized.

2.3.3. ResNet50-U-Net

The weights of the ResNet50 model pre-training on the ImageNet data set were applied to the encoder part (He et al., 2016). The decoder part follows the U-Net structure, and the weights were randomly initialized.

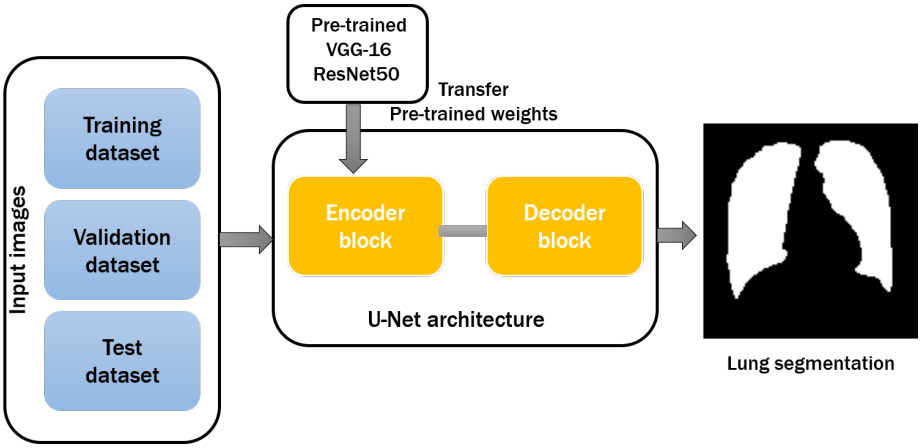


Figure 1: Generic Workflow of Transfer Learning-Based U-Net Approaches for Lung Segmentation. The encoder utilizes pre-trained weights from models (VGG-16 and ResNet50) to extract features, while the decoder reconstructs the segmentation mask through upsampling and skip connections.

2.4. Training Setup

Hyperparameters: Batch size was 16, the initial learning rate was 1e-4, and Adam was used as the optimization algorithm (Kingma and Ba, 2014). The maximum number of epochs was set to 30, and an early stopping technique, which is often used to prevent overfitting, was intentionally not applied in this study. This decision was made to observe the full training trends and performance behaviors of the models over the complete training process without prematurely halting the training based on validation loss.

Loss Function: The loss function used in this study was the Dice Loss, defined as the complement of the Dice Coefficient (Dice). The Dice Coefficient is given by:

$$Dice = \frac{2 \cdot |y_{\text{true}} \cap y_{\text{pred}}| + \epsilon}{|y_{\text{true}}| + |y_{\text{pred}}| + \epsilon},$$

where y_{true} represents the ground truth mask, y_{pred} is the predicted mask, and ϵ is a smoothing term to prevent division by zero. The final loss function was defined as:

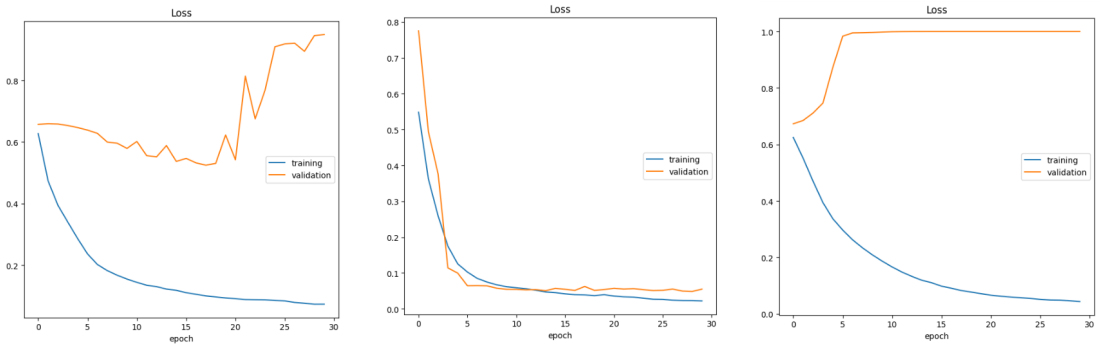
$$\text{Dice Loss} = 1 - Dice$$

This function was used exclusively to optimize the model during training.

Evaluation Metrics: The evaluation metrics included:

- **Accuracy**: The proportion of correctly classified pixels.
- **Dice**: A measure of overlap between the predicted and ground truth masks.
- **Validation Loss**: The average loss calculated on the validation set, used to monitor overfitting during training.

Validation Strategy: A separate validation set of 49 samples was used exclusively to monitor model performance and generalization, ensuring unbiased evaluation during training.



(a) U-Net (b) VGG16-U-Net (c) ResNet50-U-Net

Figure 2: Loss Results: (a) U-Net, (b) VGG16-U-Net, and (c) ResNet50-U-Net.

3. Results

3.1. Overall Findings

The performance of three different models—U-Net, VGG16-U-Net, and ResNet50-U-Net—was evaluated on training, validation, and test datasets. The results are summarized in Tables 1 and 2.

Table 1: Training Performance Comparison of Different Models

Model	Train Accuracy	Train Dice	Validation Accuracy	Validation Dice
U-Net	0.993	0.978	0.977	0.0230
VGG16-U-Net	0.993	0.978	0.973	0.950
ResNet50-U-Net	0.986	0.956	0.737	< 0.001

Table 2: Test Performance Comparison of Different Models

Model	Accuracy	Dice
U-Net	0.741	0.052
VGG16-U-Net	0.975	0.950
ResNet50-U-Net	0.741	< 0.001

3.2. Detailed Results

Training and Validation Performance During the training phase, both U-Net and VGG16-U-Net achieved identical train accuracy (99.3%) and train Dice coefficients (97.8%).

ResNet50-U-Net slightly lagged behind, with a train accuracy of 98.6% and a train Dice coefficient of 95.6%.

For validation, VGG16-U-Net demonstrated superior performance with a validation Dice coefficient of 95.0% and validation accuracy of 97.3%. U-Net showed comparable validation accuracy (97.7%) but had a significantly lower validation Dice coefficient (2.3%). ResNet50-U-Net, while maintaining a moderate train performance, struggled during validation with an accuracy of 73.7% and a Dice coefficient below 0.001, indicating overfitting.

Test Performance The test dataset results further highlighted the effectiveness of VGG16-U-Net. It achieved the highest test accuracy of 97.5% and a Dice coefficient of 95.0%. In comparison, U-Net and ResNet50-U-Net showed identical test accuracies of 74.1%, but their Dice coefficients diverged significantly. U-Net had a Dice coefficient of 5.2%, while ResNet50-U-Net's Dice coefficient was below 0.001, suggesting a failure to generalize to unseen data.

Mask Detection In Figure 3, Under a cut-off value of 0.03, the ResNet50-U-Net model failed to detect the lung region, producing an almost blank mask, while the U-Net model showed faint and incomplete detection, indicating limited performance. In contrast, the VGG16-U-Net model demonstrated accurate and defined segmentation, closely resembling the ground truth mask, highlighting its superior capability for lung region segmentation in this experiment.

(a) Test Image (b) Ground Truth (c) U-Net (d) VGG16-U-Net (e) ResNet50-U-Net

Figure 3: Comparison of segmentation results from different models under a cut-off value of 0.03. (a) Test Image, (b) Ground Truth, (c) U-Net, (d) VGG16-U-Net, and (e) ResNet50-U-Net. The VGG16-U-Net model closely matches the ground truth, while U-Net provides partial detection and ResNet50-U-Net fails to detect the lung region.

4. Discussion

Analysis: Severe overfitting was observed for both U-Net and ResNet50-U-Net due to the limited size of the dataset. In contrast, VGG16-U-Net demonstrated a balance between training and validation performance, attributed to its use of pretrained weights and its relatively simple architecture. This highlights the effectiveness of transfer learning in enhancing generalization performance on small datasets. VGG16, introduced during the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2014), remains a widely used and robust model for image classification and object detection, despite its relatively shallow

depth (Simonyan and Zisserman, 2014). On the other hand, ResNet50, with its deeper architecture and a larger number of trainable parameters, is more prone to overfitting when applied to limited datasets due to its structural complexity (He et al., 2016).

Clinical Implications: This study demonstrated that leveraging transfer learning with pretrained models can achieve reliable performance even with small datasets. This approach holds significant potential for medical image analysis in scenarios with limited imaging data, suggesting its applicability across diverse clinical domains.

Limitations: Transfer learning has the potential to cause overfitting in small datasets, as demonstrated by the ResNet50 model in this study. Although this study did not explore the issue of overfitting in detail, future research should focus on systematically examining the relationship between model size, complexity, and performance. Such investigations are essential for optimizing model architectures and transfer learning strategies in resource-limited data environments.

5. Conclusion

This study highlights the potential of transfer learning in achieving reliable performance for lung segmentation tasks, even with small datasets. By incorporating pretrained encoders, the VGG16-U-Net demonstrated superior generalization performance compared to the original U-Net and ResNet50-U-Net. The relatively simple architecture of VGG16-U-Net, combined with pretrained weights, allowed it to balance training and validation performance effectively, overcoming the limitations of small datasets. In contrast, ResNet50-U-Net suffered from overfitting, underscoring the challenges of using deeper architectures with limited data.

These findings emphasize the importance of selecting appropriate model architectures and leveraging transfer learning to address the constraints of small-scale datasets. Moreover, the study demonstrates the applicability of hybrid U-Net structures to clinical settings, offering a promising solution for medical image analysis in scenarios with restricted imaging data.

Future research should further investigate the impact of model size and complexity on performance, particularly in resource-limited environments. Such studies are crucial for developing optimized transfer learning strategies and ensuring the robustness and generalizability of deep learning models in diverse clinical domains.

Future Research Directions:

- Incorporating advanced data augmentation techniques and cross-validation can help mitigate overfitting and enhance the model's generalization performance (Wong et al., 2016).
- For complex models like ResNet50-U-Net, precise hyperparameter tuning is essential to achieve optimal performance on small datasets.
- Future studies should validate the reproducibility and robustness of these findings using larger and more diverse datasets, including CXR images from multiple institutions. Additionally, exploring other pretrained models, such as Inception or EfficientNet, could provide valuable insights (Szegedy et al., 2016; Tan and Le, 2019).

- Research could be extended to 3D segmentation using CT images, weakly or semi-supervised learning to overcome data scarcity, and uncertainty estimation methods to identify unreliable segmentation regions.

The code used in this study is publicly available at https://github.com/PeterYYong/AIFFEL_quest_rs/blob/main/MainQuest/Quest02/Quest02.ipynb.

References

Denise M. Aberle, Ann M. Adams, Christine D. Berg, William C. Black, Jane D. Clapp, Robert M. Fagerstrom, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 234–241. Springer, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

John Smith and Jane Doe. Radiological challenges in diagnosing respiratory diseases. *Journal of Radiology*, 45:123–130, 2020.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2016. doi: 10.1109/DICTA.2016.7797091.

Zongwei Zhou, M.M. Rizwan Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.