A comparison of deep and shallow residual networks for medical imaging classification

Elaine Yu
Columbia University
New York, NY, USA

Kyle Chard
University of Chicago
Department of Computer Science
Chicago, Illinois, USA

Nathaniel Hudson
University of Chicago
Department of Computer Science
Chicago, Illinois, USA

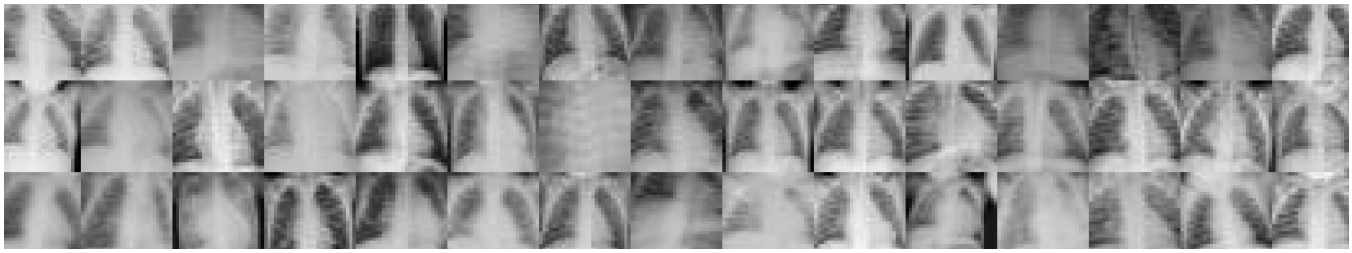


Figure 1. Several sample images of chest scans to diagnose pneumonia from the MedMNIST suite of data-sets.

Abstract

The complexity and parameters of mainstream large models are increasing rapidly. For example, the increasingly popular large language models (e.g., ChatGPT) have billions of parameters. While this has led to performance improvements, the performance gains for simple tasks may be unacceptable for the additional cost. We apply residual networks of three different depths and evaluate them extensively on the MedMNIST pneumonia dataset. Experimental results show that smaller models can achieve satisfactory performance at significantly lower costs than larger models.

Keywords: Machine Learning, Computational Healthcare, Medical Informatics, Trade-Offs

ACM Reference Format:
Elaine Yu, Kyle Chard, and Nathaniel Hudson. 2026. A comparison of deep and shallow residual networks for medical imaging classification . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (SC '23).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

To better understand the relationship between performance, cost (time), and model size, we chose the MedMNIST dataset [2] and the residual network [1]. We test the performance variation and cost of small and large models on our pneumonia classification problem. Specifically, we consider ResNet- 18, ResNet-50, and ResNet-101—models with increasing depth. Finally, we observe the performance of these three models on the pneumonia dataset to see whether we should use large models on our medical dataset.

SC '23, Nov 12–17, 2023, Denver, CO
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

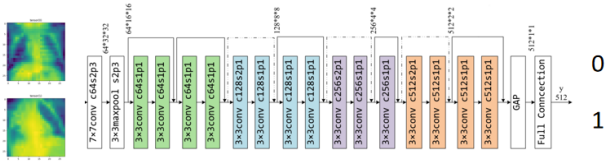


Figure 2. Visualization of the ResNet-18 Architecture.

We want to investigate whether it is possible to use a small model with lower training cost on a relevant and well explored medical dataset. We specifically consider the following research question: How does the reduction in model size affect time to model convergence and resulting model loss and accuracy? Can small models compete with larger ones when training on small datasets?

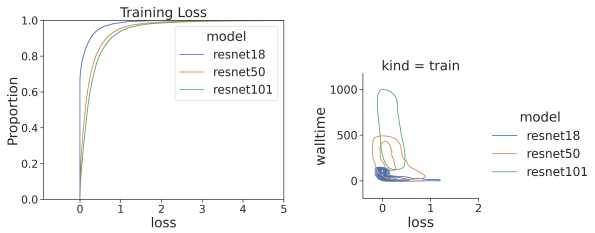
2 Background and Methods

2.1 MedMNIST

MedMNIST is a medical experiment dataset with multiple subsets. The pneumonia dataset in MedMNIST contains 5,856 28×28 grayscale images, which are divided into two categories: those with and those without pneumonia. Examples of these images are shown in Fig. 1 above.

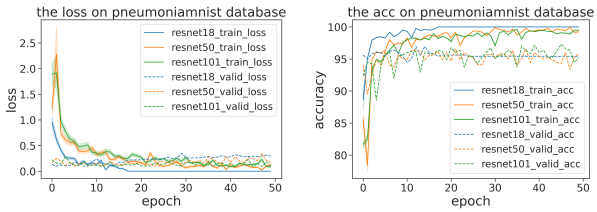
2.2 Residual Network (ResNet) Architecture

The ResNet network was proposed by He et al. [1]. It solves the degradation problem in the deep network, and can artificially make some layers of the neural network skip the connection of neurons in the next layer. The residual network is a very effective network to alleviate the problem of gradient disappearance and gradient explosion, which



(a) Train loss cumulative distribution curve. (b) Train density distribution map.

Figure 3. Performance of three residual networks in training



(a) training loss (b) training accuracy

Figure 4. Training results of three residual networks on pneumonia data.

greatly improves the depth that the network can be effectively trained. It is often used as a benchmark model to evaluate the data classification effect on classification problems. The residual network structure is shown in Fig. 2. Different residual networks have similar structures: the network consists of a total of 5 convolutional groups, each containing 1 or more basic convolutional calculation processes. Each convolutional group contains one downsampling operation to halve the size of the feature map. Downsampling is achieved through the following two methods:

- Maximum pooling, with a step size of 2, only used for the second convolutional group (Conv2x).
- Convolution, with a step size of 2, used for 4 convolution groups except for the second convolution group.

3 Experimental Setup & Results

3.1 Setup

We run the code on a computer equipped with NVIDIA GeForce RTX 4090, and use three ResNet models as a comparison of different deep network models: ResNet-18, ResNet-50, and ResNet-101. Each model is trained for $E = 50$ epochs with a mini-batch size of $B = 32$.

3.2 Results & Discussion

As shown in Figure 4 above, ResNet-18 has faster loss reduction performance and more stable accuracy rate, followed by ResNet-50, and finally ResNet-101. The kernel density estimate plot, shown in Fig. 3(a), shows that ResNet-18 shows greatly reduced walltime needed for training, while also achieving comparable loss when compared to the deeper ResNet-50 and ResNet-101 models. At the same time, as can be seen in Table 1, our results demonstrate that ResNet-18 achieves a higher *test* accuracy than the deeper ResNet models. Interestingly, this seems to be contrary to the generally understood rule that the larger the model, the better the performance.

While some of our observations from our results may, at first, seem unintuitive, there are reasonable explanations. First, it must be reiterated that the MedMNIST data for pneumonia is very small. With just under 6,000 training samples,

Table 1. Performance of three residual networks in test set.

Model	Time(s)	Top-1 Acc.	F1 score
ResNet-18	**131.921**	**0.843**	**0.741**
ResNet-50	475.013	0.835	0.731
ResNet-101	983.782	0.836	0.732

the likelihood of a deeper model memorizing the data and overfitting is increased as a result. It is very possible that this allowed the ResNet-18 model to learn a marginally more general relationship to perform better on the test dataset.

While there is more room for additional analyses and comparisons between these models in terms of their predictive performance, the most compelling finding of this work is that the ResNet-18 model is able to achieve (marginally) superior test accuracy while reducing the walltime needed to train by 72.23% and 86.59% when compared to ResNet-50 and ResNet-101, respectively. This trade-off is important and indicates that bigger models are not always necessary.

4 Conclusions

Our results reveal that on some simple medical datasets, perhaps larger models do not yield significantly higher performance despite the much greater time to train them. After comprehensively considering the running cost and performance, the small model is a better choice. Additional investigation into the effects that dataset size and the specific model architecture used have on task performance.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[2] Jiancheng Yang, Rui Shi, and Bingbing Ni. 2021. MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 191–195.