Lung Segmentation: the COVID-QU-Ex Dataset

Pham Hai Nam - BI12 307

March 19, 2024

Abstract

The COVID-QU-Ex dataset is a comprehensive collection of 33,920 chest X-ray (CXR) images meticulously curated by researchers from Qatar University. This dataset encompasses various respiratory conditions, including COVID-19, non-COVID infections (viral or bacterial pneumonia), and normal cases. The standout feature of the COVID-QU-Ex dataset is the inclusion of ground-truth lung segmentation masks, facilitating precise delineation of lung regions within each image. This annotation process ensures unprecedented accuracy and reliability in the analysis of lung pathology, making it a milestone achievement in medical image segmentation research. Furthermore, the dataset's training, validation, and test sets enable researchers to develop and evaluate machine learning algorithms for the diagnosis of respiratory diseases. In this study, we explore the dataset's composition, present visualizations, and evaluate the performance of a ResNet-50 model on the COVID-QU-Ex dataset. Our results demonstrate high accuracy and promising performance in identifying COVID-19 cases, underscoring the dataset's significance in advancing research on respiratory disease diagnosis.

1 Introduction

The COVID-QU-Ex dataset is an incredible collection of 33,920 chest X-ray (CXR) images painstakingly put together by researchers from Qatar University. Their aim in creating this dataset is to help advance research in the field of respiratory diseases, with a specific focus on COVID-19.

When it comes to medical diagnostics, chest X-rays are a commonly used tool to assess the condition of the lungs and provide valuable insights into respiratory infections like COVID-19. The COVID-QU-Ex dataset serves as a valuable resource for researchers, offering them a vast array of chest X-ray images to develop and evaluate machine learning algorithms, computer-aided diagnostic systems, and other innovative solutions driven by artificial intelligence. By leveraging these tools, medical professionals can potentially improve the accuracy and speed of COVID-19 and respiratory illness diagnoses, leading to better outcomes for patients and more effective public health interventions.

Having comprehensive datasets like COVID-QU-Ex is of utmost importance for researchers. They can use these datasets to train and validate AI models, paving the way for the creation of robust and reliable tools to diagnose and manage respiratory diseases. Moreover, this dataset holds the potential to deepen our understanding of the radiological manifestations of COVID-19, helping researchers identify specific patterns or features associated with the disease.

It's important to remember that while chest X-rays provide valuable information, a definitive diagnosis of COVID-19 typically requires additional tests such as polymerase chain reaction (PCR) tests or antigen tests, which directly detect the presence of the SARS-CoV-2 virus or its genetic material.

The COVID-QU-Ex dataset represents a significant contribution to the scientific community's relentless efforts in combating the COVID-19 pandemic. By fostering research and innovation in the diagnosis and treatment of respiratory diseases, it plays a crucial role in our collective fight against this global health crisis.

2 Project study

2.1 Dataset - COVID-QU-Ex

Notably, the COVID-QU-Ex dataset encompasses a diverse array of conditions, including 11,956 cases of COVID-19, 11,263 instances of non-COVID infections such as viral or bacterial pneumonia, and 10,701 normal cases. However, what sets this dataset apart is the inclusion of ground-truth lung segmentation masks for the entire repository, facilitating precise delineation of lung regions within each image. This dataset represents a crucial resource for medical professionals, data scientists, and researchers alike, providing invaluable insights into the spectrum of lung pathologies.

This meticulous annotation process ensures unprecedented accuracy and reliability in the analysis of lung pathology, marking a significant leap forward

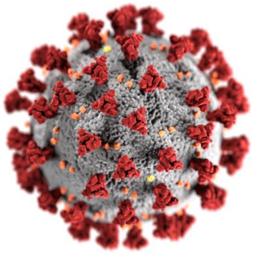

Figure 1: COVID-19 example

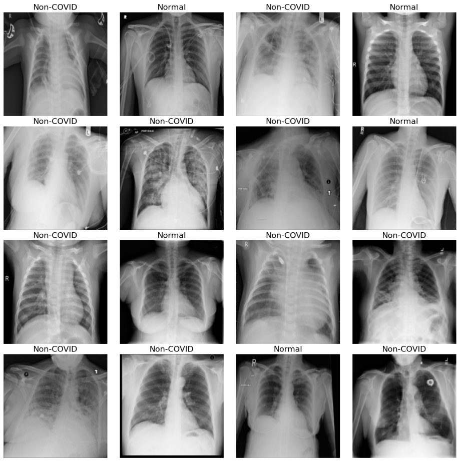

Figure 2: Dataset Visualizations of COVID-QU-Ex dataset

in medical image segmentation research. Indeed, the COVID-QU-Ex dataset represents a milestone achievement, standing as the largest lung mask dataset ever curated.

The training, validation, and test datasets consist of a total of 21,715, 5,417, and 6,788 images, respectively. Each dataset is divided into three categories: COVID-19, non-COVID (potentially including other types of infections), and normal cases.

- Training dataset: 21,715 images, distributed among COVID-19 (7,658 images), non-COVID (7,208 images), and normal cases (6,849 images).
- Validation dataset: 5,417 images, distributed similarly among the three categories.
- Test dataset: 6,788 images, again distributed among COVID-19, non-COVID, and normal cases.

Table 1: Dataset Composition

	COVID-19	**Non-COVID**	**Normal**
Training images	7658	7208	6849
Validation images	1903	1802	1712
Test images	2395	2253	2140

2.2 Deep learning model

ResNet-50 is a deep convolutional neural network architecture that was introduced by Microsoft Research in the paper titled ”Deep Residual Learning for Image Recognition” by Kaiming He et al. It is part of the ResNet (Residual Network) family of models, which are known for their ability to train very deep neural networks effectively.

The ”50” in ResNet-50 refers to the number of layers in the network. It consists of 48 convolutional layers and 1 fully connected layer. ResNet-50 employs residual connections, which allow the network to learn residual functions with respect to the layer inputs, thereby mitigating the vanishing gradient problem. This enables the training of much deeper networks compared to previous architectures.

ResNet-50 has achieved state-of-the-art performance on various computer vision tasks, including image classification, object detection, and image segmentation. It has been widely used in both academic research and industrial applications due to its effectiveness and efficiency.

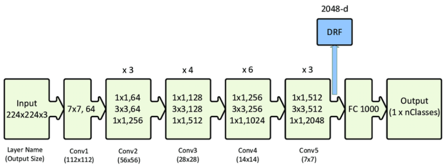

Figure 3: ResNet-50 Achitecture

3 Model Evaluation

3.1 Confusion Matrix

After training, the model's performance is evaluated on the validation dataset. The provided classification report shows the precision, recall, and F1-score for each class (COVID-19, non-COVID, and normal), along with overall accuracy and macro and weighted averages.

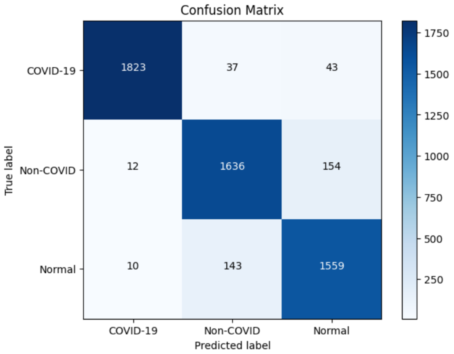

Figure 4: Confusion Matrix

- **Precision**: The proportion of true positive predictions among all positive predictions.
- **Recall**: The proportion of true positive predictions among all actual positive instances.
- **F1-score**: The mean of precision and recall, providing a balanced measure between the two.
- **Support**: The number of true instances for each class in the validation dataset.

The model demonstrates high performance across all classes, with F1-scores ranging from 0.90 to 0.97. Notably, the model achieves a particularly high

Table 2: Classification Report on Validation Dataset

	Precision	**Recall**	**F1-score**	**Support**
COVID-19	0.99	0.96	0.97	1903
Non-COVID	0.90	0.91	0.90	1802
Normal	0.89	0.91	0.90	1712
Accuracy			0.93	5417
Macro avg	0.93	0.93	0.93	5417
Weighted avg	0.93	0.93	0.93	5417

F1-score of 0.97 for COVID-19 classification, indicating strong performance in identifying COVID-19 cases. Overall accuracy on the validation dataset is reported as 93%, indicating that the model's predictions align with the ground truth labels for a majority of the cases.

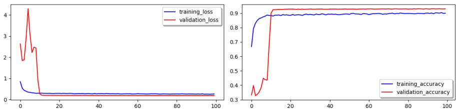

Figure 5: Training and Validation Loss

Figure 5, the plot on the left side illustrates the training and validation loss. As the model trains, the loss initially decreases rapidly and then levels off, indicating that the model is effectively learning from the training data. It's a positive sign that the validation loss follows a similar pattern, decreasing alongside the training loss. This suggests that the model is able to generalize well to new, unseen data.

On the right side, the plot displays the training and validation accuracy. The training accuracy increases quickly and reaches a high level, remaining relatively stable afterward. Similarly, the validation accuracy also rises rapidly and stays close to the training accuracy. This alignment between the two accuracy curves implies that the model is not overfitting significantly and is capable of accurately classifying both the training and validation data.

In both plots, the convergence of the training and validation metrics, along with the small gaps between them, indicates that the model performs well and generalizes effectively to the validation dataset. However, it's worth noting that the initial spikes in the graphs suggest some volatility during the early stages of training. This volatility could be attributed to factors such as the learning rate chosen at the beginning or the random initialization of the model's weights.

Most of the images in figure 6 are labeled as "Non-COVID [OK]" indicating that the individual in the X-ray does not have COVID-19, according to the label. A couple of the X-rays are labeled as "Normal [NO→Non-COVID]" which seems to imply a normal finding, not indicative of COVID-19 infection. There is one image labeled "COVID-19 [NO→Non-COVID]" which might suggest a

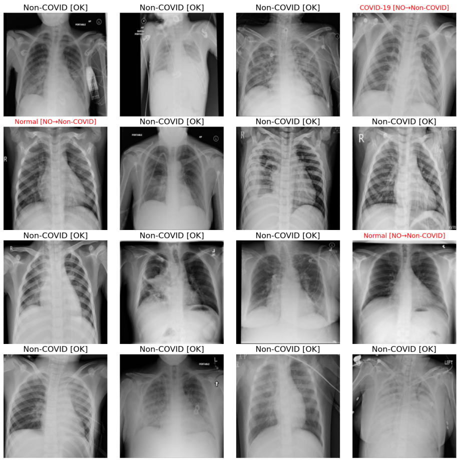

Figure 6: Result images

case that was initially suspected as COVID-19 but was later determined to be non-COVID.

Chest X-rays are a useful tool in the diagnostic process, particularly for respiratory illnesses such as COVID-19, because they can reveal signs of pneumonia or other changes in the lung that may be associated with the infection. However, it's important to note that the diagnosis of COVID-19 typically requires a combination of clinical evaluation, imaging, and testing (like PCR or antigen tests), as many conditions other than COVID-19 can cause changes in the lung visible on an X-ray.

4 Conclusion

In summary, the COVID-QU-Ex dataset represents a major breakthrough in the fields of medical imaging and artificial intelligence. With its extensive collection of chest X-ray images and accurate lung segmentation masks, this dataset provides researchers with a valuable tool to develop and evaluate machine learning algorithms for diagnosing respiratory diseases, especially COVID-19.

Furthermore, the performance evaluation of the ResNet-50 model on the COVID-QU-Ex dataset showcases promising results, with high precision, recall, and F1-scores across all categories. The model's ability to accurately classify COVID-19 cases holds significant implications for swift and effective disease diagnosis.

Overall, the combination of the COVID-QU-Ex dataset and advanced deep learning models like ResNet-50 contributes to the ongoing global efforts to combat the COVID-19 pandemic and enhance healthcare outcomes. By fostering collaboration among medical professionals, data scientists, and researchers, this dataset opens the door to innovative solutions addressing the challenges posed by respiratory diseases, ultimately leading to improved patient care and more effective public health interventions.