Generating Explanations for Chest Medical Scan Pneumonia Predictions

Samson Qian
Halicioğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92093
saqian@ucsd.edu

Abstract

With the spread of COVID-19, significantly more patients have required medical diagnosis to determine whether they are a carrier of the virus. COVID-19 can lead to the development of pneumonia in the lungs, which can be captured in X-Ray and CT scans of the patient's chest. The abundance of X-Ray and CT image data available can be used to develop a high-performing computer vision model able to identify and classify instances of pneumonia present in medical scans. Predictions made by these deep learning models can increase the confidence of diagnoses made by analyzing minute features present in scans exhibiting COVID-19 pneumonia, often unnoticeable to the human eye. Furthermore, rather than teaching clinicians about the mathematics behind deep learning and heat maps, we introduce novel methods of explainable artificial intelligence (XAI) with the goal to annotate instances of pneumonia in medical scans exactly as radiologists do to inform other radiologists, clinicians, and interns about patterns and findings. This project explores methods to train and optimize state-of-the-art deep learning models on COVID-19 pneumonia medical scans and apply explainability algorithms to generate annotated explanations of model predictions that are useful to clinicians and radiologists in analyzing these images.

1. Introduction

In late 2019, instances of pneumonia in Wuhan, China was found to have been caused by a viral infection, now referred to as COVID-19. Over the year 2020, the virus had spread rapidly across various parts of the world and infected millions of people [1]. Pneumonia is one of the most common developments of COVID-19 and can cause symptoms such as coughing, high fever, and breathing difficulty [2]. These cases of pneumonia, if not identified and treated quickly, gradually worsen over time and can potentially be fatal. Increased digitization and technological capabilities have resulted in the creation of large labeled datasets of medical imaging, such as CT and X-Ray scans. Artificial intelligence and machine learning models can be useful in learning the shapes and patterns of pneumonia present in chest medical scans taken of patients. Hand-labeling regions of pneumonia on these images is often time-consuming, expensive, and prone to human error. Rather than investing large amounts of effort in manually labeling existing medical images, it is more efficient to apply mathematical algorithms to examine AI model architectures and generate explanations.

1.1 Objectives

Complex artificial intelligence models are often regarded as black-boxes, where a prediction is outputted without explanation. The primary objective in building explainable computer vision models is to accurately identify pneumonia instances in chest medical scans and generate annotated regions that explain the prediction. Such a framework places more trust in deep learning algorithms and can aid clinicians' and radiologists' work in analyzing regions of pneumonia while diagnosing patients. Furthermore, the generated explanations can provide further insight into regions of pneumonia than existing hand-labeled data.

2. Methods

2.1 Training & Optimizing Computer Vision Models

Convolutional neural networks (ConvNets) are neural networks with layers that apply convolutions to parts of images to learn image features of various classes. Extensive prior research has been done to improve these models' performance by experimenting with the number of neurons, types of layers, and activation functions. Two types of architectures, VGG16 [3] and ResNet50 [4], displayed in Figure 1, are state-of-the-art models in learning complex features to analyze medical images. These models are accompanied by pre-trained weights which contain information already previously learned by the model on large image datasets. The training is performed through batch gradient descent to optimize the cross-entropy loss between the two classes of images, with and without pneumonia.

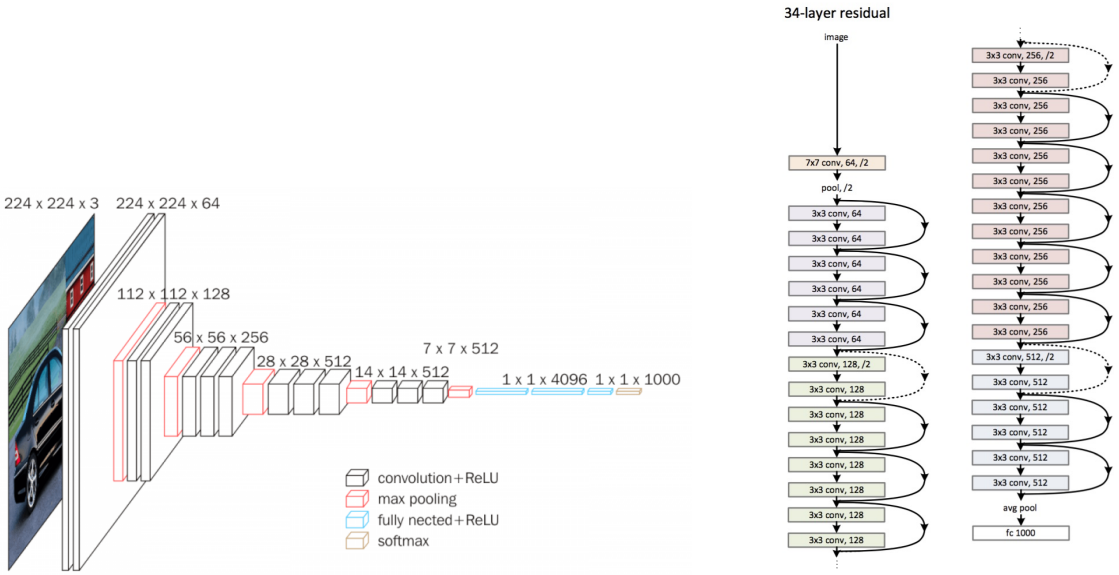

Figure 1. VGG16 (Left) and ResNet50 (Right) Architectures

The trained model performs with approximately 92% accuracy, with a 0.90 F_1-score across images with and without pneumonia. There may still be methods to further improve model

accuracy, precision, and recall scores, but this study will focus more on generating explanations for model predictions. The trained VGG16 and ResNet50 models perform sufficiently in distinguishing between medical scans with and without pneumonia, but only output a binary classification. It is necessary to examine model weights to understand predictions made.

2.2 Explainability Methods & Generated Annotated Explanations

The purpose of implementing explainability methods is to generate a likelihood heatmap on inputted images to the model to identify regions that contribute to model predictions. This can be done by, for example, propagating through the model's weights and identifying important features, or by testing regions of images to understand the model's behavior. These techniques can produce a model that additionally outputs an annotated explanation, which can be used to understand the prediction made and increase confidence and transparency.

2.2.1 Layer-wise Relevance Propagation

Layer-wise relevance propagation (LRP) is an explainability method that examines the trained model’s learned weights and identifies which image features contribute most to predictions [5]. This is done by propagating through the layers of the deep learning model and identifying which weights are the most for a certain image. In essence, LRP is a top-down method that examines the structure and activations of the neural network to generate explanations of important image features and regions that are most important for predicting on a certain image.

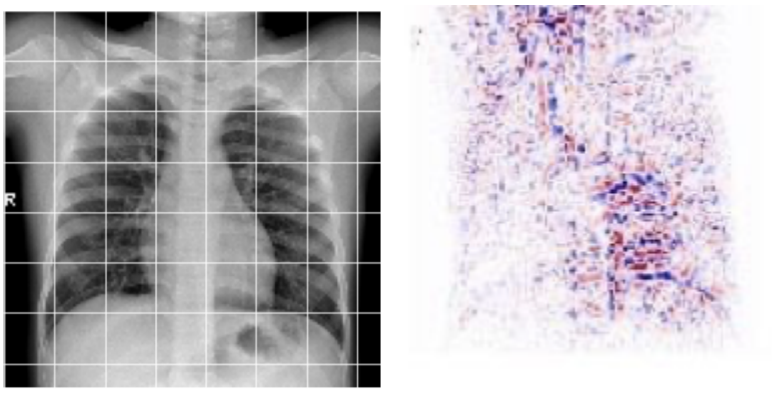

Figure 2. LRP Output for Pneumonia X-Ray Scan

Figure 2 displays how the LRP algorithm propagates through the model's weights and finds corresponding image features that are impacted by the most significant weights. The output is a form of heatmap that intensifies in regions where weights are larger and contribute more to a prediction made. The heatmap output explains regions of the image that are most important to the model when performing forward-propagation and outputting a prediction. This heatmap is useful for clinicians in observing high-intensity areas of the scan where pneumonia may appear.

2.2.2 Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations (LIME) is an explainability method that identifies parts of an image that contribute the greatest to the model's output. It does so by segmenting and modifying existing regions of the image and quantifying the impact that these changes have on the model [6]. Regions with changes that greatly affect the model's output are important to making predictions, as opposed to regions whose changes do not have an effect. The LIME method is a bottom-up method that starts with manipulating the inputted image's pixels and examining the model's behavior on such changes.

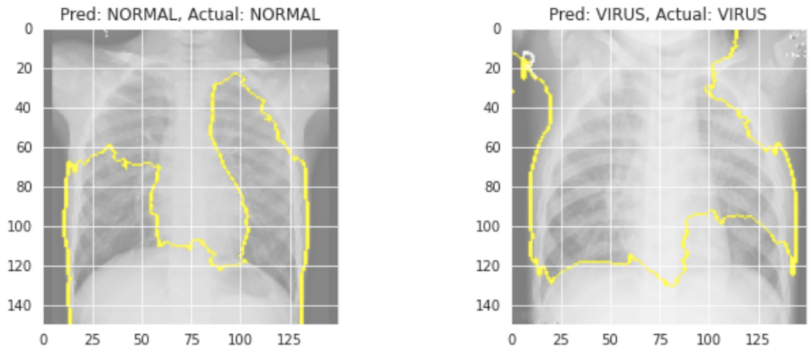

Figure 3. LIME Output for Normal (Left) vs. Pneumonia (right) X-Ray Scan

The LIME output for both X-ray images with and without instances of pneumonia is shown in Figure 3. The outlined regions contain the most information for the model, so it is likely that instances of pneumonia occur in these regions. Although this method may not exactly outline the mask of pneumonia, it provides a specific region for clinicians to focus on when examining medical scans.

2.2.3 Gradient-weighted Class Activation Mapping

The Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm examines the neuron activations in the trained neural network while making predictions on an inputted image [7]. This method is similar to the LRP algorithm because it examines the inner structure of the network and attempts to find the features in the image with the greatest activation of the model's neurons.

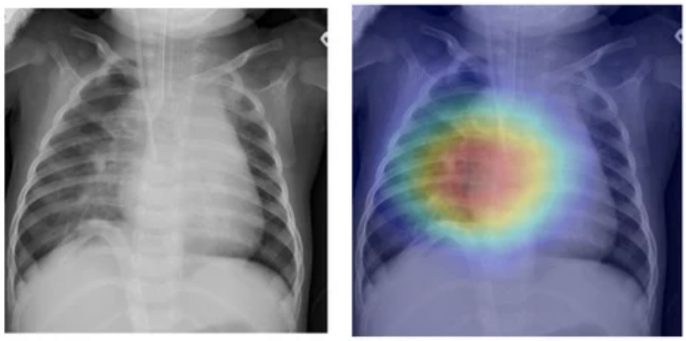

Figure 4. Grad-CAM Output for Pneumonia X-Ray Scan

The output of Grad-CAM on an X-ray image with an instance of pneumonia is shown in Figure 4. The Grad-CAM, in contrast with LIME, outputs a form of heatmap similar to LRP. This is because the algorithm examines the areas of the image that maximizes the activations of the neurons. The intensities of the heatmap show the relative neuron activations, which represents the importance of making a prediction.

3. Conclusion

This study is potentially useful to clinicians analyzing COVID-19 pneumonia present in medical scans and identifying the regions in which pneumonia appears. The presented framework not only generates a classification output, produced by a model trained on large amounts of image data, but also highlights regions of each image responsible for the prediction made. Many other applications are possible for clinicians and radiologists performing medical diagnoses on patients for different types of illnesses. The trained model and explainability algorithms can assist clinicians and radiologists in their analysis.

The rapid generation and availability of data will likely lead to the incorporation of machine learning methods to assist with traditional processes. With more patients contracting viruses such as COVID-19 and requiring treatment, artificial intelligence is useful in speeding up pneumonia identification. A common concern of complex AI models is the lack of transparency because locating pneumonia is as important as the actual classification. Without explanations, false positive and false negative diagnoses can be dangerous. However, with added regional explanations by the AI model and analysis by trained clinicians, the efficiency and confidence of the diagnosis process can be significantly improved.

4. Acknowledgments

The methods and algorithms described and developed in this paper were guided by Dr. Michael Pazzani, Distinguished Scientist of the Halicioğlu Data Science Institute at the University of California, San Diego. This research project was conducted at the Halicioğlu Data Science Institute for a research grant funded by the National Science Foundation.

5. References

[1] Preventing the spread of the coronavirus - Harvard Health, (n.d.). https://www.health.harvard.edu/diseases-and-conditions/preventing-the-spread-of-the-coronavirus.

[2] COVID-19 basics - Harvard Health, (n.d.). https://www.health.harvard.edu/diseases-and-conditions/covid-19-basics.

[3] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, (2015), https://arxiv.org/abs/1409.1556.

[4] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, (2015), https://arxiv.org/abs/1512.03385.

[5] W. Samek, A. Binder, G. Montavon, S. Bach, K. Müller, Evaluating the visualization of what a Deep Neural Network has learned, (2015), https://arxiv.org/abs/1509.06321.

[6] M. Tulio Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, (2016), https://arxiv.org/abs/1602.04938.

[7] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, (2019), https://arxiv.org/abs/1610.02391.