Interpretable artificial intelligence (AI) – saliency maps

Shengping Yang PhD, Gilbert Berdine MD

A*rtificial intelligence has increasingly been used in the biomedical field. Could you please provide an example of such usage and also discuss what can be improved in future applications?*

In a previous article, we introduced the application of Artificial Intelligence (AI) in biomedical research.[1] Artificial intelligence offers inherent advantages in complex data analysis, particularly in making predictions. Unlike linear models commonly used in data analysis, AI, such as deep learning neural networks, can capture non-linear relationships among variables. Additionally, AI is capable of handling large and high-dimensional datasets without requiring extensive data preprocessing or feature engineering. Moreover, AI models scale effectively with increasing data size and complexity. However, despite these advantages, one limitation of AI is its tendency to be perceived as a "black box" with its internal operation being opaque to users. This limitation restricts its applicability in certain fields.

When faced with a new tool, the customer's first question is, "Does it work?" The answer to this question requires a gold standard for a positive and negative outcome that can be compared to the AI predictions. For new pharmaceutical products, the answer to this question is called "efficacy." For a chess AI, the answer to this question is determined by contests between the AI and opponents of different abilities. The opponents may be humans or previous chess AI models. The ability to understand how an AI model arrives at its predictions is called "interpretability." If we encounter someone on a street corner peddling a "miracle" elixir that cures all ailments, we are likely to ask, "What's in it?" or How does it work?" before buying it. Once we have been convinced that something worked in the past (interpretability), we are more likely to accept that it will continue to work in the future if we understand how it works. For AI models, the ability to explain to a lay person how the AI arrived at correct predictions is called "explainability" and is an important prerequisite for acceptance of the AI to give correct predictions when challenged with future unknown inputs. Past performance does not guarantee future results, but we are more likely to accept past performance if we understand how something works and believe the algorithm to be valid for general cases rather than a small subset of cases. Very few people understand how their mobile phone works. However, once the general public is convinced by other users that the phone does, indeed, work (interpretability), the general public accepts the "black box" because the inner workings are electronics rather than magic (explainability).

In this article, we will introduce *saliency maps*, a powerful tool in the field of interpretability and explainability of AI models. Consider an example of an AI model that interprets mammograms (input data) by assigning a score of normal or abnormal (output category). The AI model does not start with an algorithm on how to interpret mammograms. The AI is given a training input set with images known to be normal and abnormal. The AI analyzes the normal and abnormal images and then searches for patterns that distinguish normal from abnormal. Once trained, the AI model is challenged with images that have been interpreted by experts without telling the AI model what category was assigned by the experts. The AI model applies what was learned from the training data to the unknown images and assigns a score of how likely the test image is normal or abnormal. The score is calculated from patterns of color, intensity, and contrast of different locations in the image. All locations are not equally important. Saliency maps are a graphical representation of the importance of each location to the final score. They offer valuable insights into how AI models make predictions or decisions by highlighting the crucial regions or features in the input data that contribute the most to the model's output.[2,3] Developers can use saliency maps to improve the interpretability of AI

Corresponding author: **Shengping Yang**
Contact Information: Shengping.Yang@pbrc.edu
DOI: 10.12746/swrccc.v11i48.1209

models by either improving the internal algorithms or improving the test data set. Saliency maps can improve explainability by helping users better understand how predictions were made. With their wide range of applications across domains such as medical imaging, natural language processing, and autonomous driving, saliency maps have greatly facilitated the interpretation and validation of decisions made by AI models. Understanding the importance of the saliency map has a significant role in enhancing trust, transparency, and accountability of the AI output or predictions, by bridging the gap between the black-box nature of AI and human interpretability.

1. What are saliency maps?

Saliency maps have their origins in the field of computer vision. Just as people focus their attention on the most important part of their visual fields, the AI prioritizes which locations are most important to classify or categorize what abstract object is represented by the image. Saliency maps have been extensively employed to comprehend the visual attention mechanism of deep neural networks. Image classification is one of the key domains in which saliency maps have a significant role. They help in identifying the regions of an image that are crucial for the model's classification decision by highlighting the locations that are most important to determining output category. By analyzing saliency maps, researchers can gain insights into the reasoning process of deep neural networks in image classification tasks.

1.1 Image classification

The goal of image classification is to train an AI model to automatically identify images. By identify, we mean to accurately assign an abstract label or category to images based on their visual characteristics and patterns. For example, does an image represent a human, a dog, or a building? Extending the process to facial recognition, if the image represents a human face, which person is represented by the image? This process involves training the model with a substantial dataset of labeled images, where each image is associated with a specific category. During training, the AI model learns to recognize and extract relevant features from the images, enabling it to make predictions on unseen images. Saliency mapping is a graphical representation of which locations were most important in distinguishing one category from another.

A significant challenge with AI models lies in their lack of explainability of the interpretation algorithms, which limits their applicability in various fields. To paraphrase the physicist Richard Feynman, "If you can't explain something in simple terms, you don't understand it." If the developer of AI cannot explain how the AI makes a determination in simple terms, potential users of the AI are less likely to accept that the AI will continue to make accurate predictions in the future. Visualizing how the AI prioritizes locations using a saliency map facilitates understanding the significant visual features or patterns that the AI relies on for classification. We are willing to accept that the AI can do something we understand very quickly. How many people would agree to let the magician saw them in half without understanding how the illusion works?

1.2 Saliency maps

Saliency maps have emerged as a natural tool to improve the interpretability of AI models. By assigning a relevance score to each pixel or feature in the input image data, saliency maps enable the identification of regions within an image that exert a significant influence on the model's prediction. Incorrect output category might be due to inappropriately giving high priority to background noise or by inappropriately giving low priority to the location corresponding to the needle in the haystack. In essence, saliency maps provide insights into the model's attention and highlight the most critical regions in an image that distinguish one output category from another. By visualizing these salient regions, researchers and practitioners can gain a better understanding of how the AI model arrives at its predictions (explainability) and which specific image features contribute the most to its decision-making process.

The calculation of saliency values in saliency map generation relies on the specific method employed. In general, discreet structures are identified by

continuous regions of similar color and brightness with boundaries defined by contrast with the surrounding pixels. Gradient-based methods and perturbation-based methods are two commonly used approaches to identify locations of interest. However, in this article, we won't delve into the intricate algorithms behind saliency map generation. Instead, we will illustrate how saliency maps are used in interpreting AI models using a classical example. Subsequently, we will apply deep learning methods to diagnose COVID-19 patients and evaluate the potential clinical application of saliency maps in this context.

2. Saliency maps in image classification – a classical example

2.1 The images

In this example, we will utilize AI to classify four distinct images (Figure 1). The images include a mini drone, a dog, an AI generated dog,[4] and the Capitol Dome. Prior to analysis, all the images underwent preprocessing steps such as reshaping and normalization to ensure consistency in the input data.

Figure 1. Images to be classified: Upper left: A drone; Upper right: A dog; Bottom left: An Al generated dog; Bottom right: The Capital Dome.

2.2 The pre-trained deep learning model

The pre-trained Visual Geometry Group 16 (VGG16) model was used for making predictions.[5] The VGG16 is renowned for its depth, consisting of 16 layers, which gives it its name. It is capable of classifying images into 1000 different object categories, such as keyboard, animals, pencil, mouse, and more.

The user output from the AI model are the output scores or probabilities that the image is an example of a pre-defined category. Additional output generated by the deep learning model includes the internal values used to compute the final scores. These internal values were used to calculate the saliency values. These saliency values were then smoothed and normalized to generate the saliency maps. These maps highlight the regions within the input images that are most relevant or influential for the model's predictions, providing valuable insights into the decision-making process of the model.

2.3 The results

2.3.1 Classification probabilities

Table 1 illustrates that a significant portion of the predictions were accurate, with prediction probabilities ranging from approximately 50% to 97%. For instance, although the mini drone being predicted as a warplane is not entirely precise, it is still a noteworthy prediction considering that the VGG16 model may not have a specific category for drones. Similarly, the prediction for the AI-generated dog is reasonable since its accuracy relies on the resemblance of the generated image to an actual dog. To gain deeper insight into the classification process, visual inspections of the saliency maps were conducted to analyze the specific features involved in the predictions.

2.3.2 The saliency maps

The saliency maps (Figure 2) proved to be highly informative as they highlighted the crucial features

Table 1. The Prediction Probabilities for the Top 5 Categories

Drone (Mini 2)		Dog (Havamalt)	
Predicted as:	**Prediction Probability**	**Predicted as:**	**Prediction Probability**
Warplane	52.3%	Papillion	93.2%
Wing	12.0%	Japanese spaniel	2.7%
Airship	6.7%	Chihuahua	0.9%
Projectile	3.8%	Wire-haired fox terrier	0.4%
Airliner	3.5%	Toy terrier	0.4%
AI Generated Dog (Golden Retriever)		**The Capitol Dome**	
Predicted as:	**Prediction Probability**	**Predicted as:**	**Prediction Probability**
Golden retriever	81.0%	Dome	97.3%
Labrador retriever	8.1%	Palace	1.1%
Kuvasz	2.6%	Stupa	0.4%
Dingo	2.5%	Fountain	0.3%
Great Pyrenees	1.0%	Bell cote	0.3%

Figure 2. Saliency maps were overlaid on the 4 images. Upper left: A drone; Upper right: A dog; Bottom left: An AI generated dog; Bottom right: The Capital Dome.

relevant to the classification process. In the case of the drone image, the saliency maps meaningfully emphasized the body, legs, and propellers, while disregarding irrelevant details such as the trees underneath. When examining the saliency maps for the dog images, notable emphasis was observed on the faces (frontal and side view) and partially on the legs, which are essential distinguishing features. Similarly, for the Capitol dome image, the saliency map predominantly highlighted the dome, a significant characteristic of the landmark. These observations suggest that the saliency maps effectively captured and visualized the important features of the image target as well as ignoring the locations determined to be background. Both processes are necessary for the AI model to accurately classify the image category.

3. Diagnosis of COVID-19 using X-ray images

To assess the performance of AI and saliency maps in a biomedical setting, we applied them to a dataset comprising X-ray images from 3,616 COVID-19 patients and 10,192 normal subjects (downloaded from

https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database).

3.1 Prediction model training and making predictions

The dataset included labeled images indicating whether they were obtained from COVID-19 patients or normal subjects. These images were then divided into training (90% of the subjects) and testing (10%) sets. A convolutional neural network model was constructed using the sequential class from the *tensorflow.keras.models* module.[6] The model was compiled with the Adam optimizer, binary cross-entropy loss function, and accuracy metric. Subsequently, the model was trained on the training data using the fit method. During the testing phase, the trained model was used to predict the COVID-19 positivity probability for the testing images. To visualize and gain further insights, two testing samples with prediction probabilities close to 0 and two samples with prediction probabilities close to 100% were randomly selected. Saliency maps were generated for these selected samples to highlight the regions in the images that significantly influenced the model's predictions.

3.2 Results

In the testing set, there were 367 images from COVID-19 patients and 1,014 images from normal subjects. A prediction probability threshold of 50% was used to determine whether an image belonged to a COVID-19 patient (>50%) or a normal subject (≤50%). Among the 367 COVID-19 patients, 321 were correctly predicted as COVID positive, while among the 1,014 normal subjects, 23 were incorrectly predicted as COVID positive. As a result, a total of 321 + 991 = 1,312 (95%) subjects were correctly predicted.

Figure 3 showcases the saliency maps for the randomly selected normal subjects (top 2) and COVID-19 patients (bottom 2). For the normal subjects, only a few features/pixels within the lung area were highlighted as important for making predictions. In contrast, many features/pixels inside the lung area had a significant role in the prediction process for the COVID-19

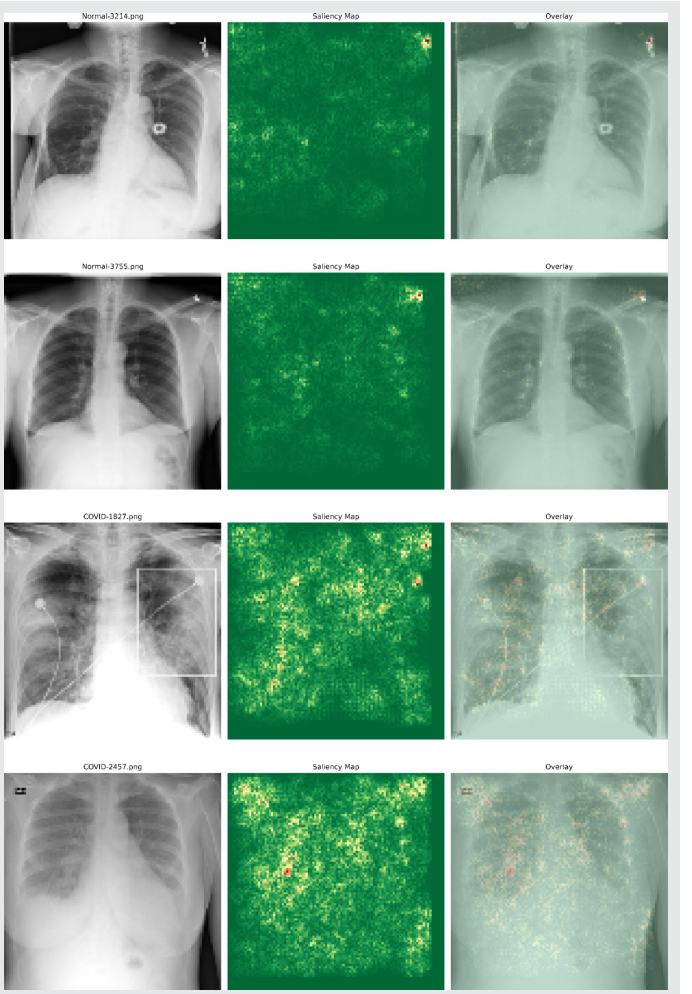

Figure 3. Saliency Maps. Top 2: Two randomly selected images with prediction probabilities for COVID-19 close to 0. Bottom 2: Two randomly selected images with prediction probabilities for COVID-19 close to 100%.

patients. A simple explanation of the AI model might be that too much white in the lung is bad, which seems to be a reasonable generalization. Meanwhile, many features outside the lung areas were considered as important. In the normal images, the AI appeared to be distracted by the "L" marker designating left direction. In the COVID images, the AI appeared to have difficulty distinguishing white in the lung from white in the abdomen. These observations suggest possible improvements to the image pre-processing techniques or training data sets to improve prediction accuracy by

focusing on relevant features within the lung area. Just as important as what features were high priorities were features that were apparently ignored by the AI model. The AI attached low priority to the heart and spine. This priority scheme might work when all the images are from patients with normal hearts and spines but would fail badly if the AI tried to distinguish congestive heart failure or scoliosis from normal. What is left out of the saliency map provides insight on how to improve the training data in order to expand the scope of abnormality that can be detected by the AI model.

The saliency maps could also improve how human experts interpret chest radiographs. More important, the saliency maps could improve how we teach medical students to interpret chest radiographs. Students can be overwhelmed about where to start with CXR interpretation; saliency maps provide insight on where to start and pitfalls to avoid.

4. The limitations of saliency maps

The saliency maps and the overlays with the images provide insight into AI model interpretability and explainability. However, they have certain limitations that are important to consider:

4.1 Interpretation challenges

Interpreting saliency maps can be subjective and challenging, often requiring expertise in relevant fields to make meaningful interpretations. Another article in this issue of the Journal raises concern that AI will steal jobs from physicians.[7] Human experts are going to be necessary for AI training for the foreseeable future. Just as physician assistants and nurse practitioners extend the utility of physicians, interpretive AI models can extend the utility of radiologists. The maps highlight regions considered important for the model's prediction, but they do not necessarily provide a comprehensive understanding of the model's decision-making process. Further in-depth investigations are often necessary to grasp the underlying mechanisms. Sometimes when we do not understand why the AI does what it does, the answer is to improve the AI. However, it is also possible for the AI to uncover paths of analysis not previously considered by human interpreters. Artificial intelligence has revolutionized chess by just such a mechanism.

4.2 Model-specific

Saliency maps are specific to the model architecture and the specific input used. In addition, different models may generate different saliency maps for the same input, making the interpretation less consistent across models.

4.3 High-frequency noise

Saliency maps can be susceptible to high-frequency noise or small perturbations in the input image, which can be misinterpreted as contrast defining boundaries between structures. High frequency noise can have an analogous effect as a pointillist painting, such as "La Grande Jatte" by Georges Seurat, in which viewing up close looks like random dots of paint, but viewing from afar looks like the intended scene on a Sunday afternoon. Small perturbations can cause translation or rotation shifts which, combined with perspective, can significantly alter the perception of the image leading to noisy saliency patterns and/or misleading interpretations.

4.4 Apply primarily to image data

Saliency maps are developed primarily for enhancing the interpretability of image data. They are specifically designed to highlight important regions or features in an image that contribute to the model's prediction. However, there are many other types of high dimensional data, and further developments and adaptations are required to effectively apply saliency maps to those different fields and data types. The image concepts of color, intensity, and contrast can be applied to other types of matrix data in higher dimensions as gradients, local maxima, and local minima. For example, color flow is a technique used in ultrasound imaging to visualize fluid velocity.

In summary, saliency mapping is a tool for improving the interpretability of AI model predictions, for improving

explainability of AI models to consumers of the predictions, and to possibly improve how experts interpret the difficult examples. In biomedical imaging, the best practice may require preliminary sorting of radiographic images by interpretive AI with follow-up review by a human expert for the difficult cases.

Keywords: Artificial intelligence, saliency mapping, medical images, predictions

Article citation: Yang S, Berdine G. Interpretable artificial intelligence (AI) – saliency maps. The Southwest Respiratory and Critical Care Chronicles 2023;11(48):31–37
From: Pennington Biomedical Research Center (SY), Baton Rouge, LA; Department of Internal Medicine (GB), Texas Tech University Health Sciences Center, Lubbock, Texas
Submitted: 7/9/2023
Accepted: 7/12/2023
Conflicts of interest: none

References

1. Yang S, Berdine G. Artificial intelligence in biomedical research. The Southwest Respiratory and Critical Care Chronicles 2023;11(46):62–65. https://doi.org/10.12746/swrccc.v11i46.1139
2. Hou X, Zhang L. Saliency detection: a spectral residual approach. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. June 2007.
3. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. 2013. arXiv preprint arXiv:1312.6034. https://arxiv.org/abs/1312.6034
4. Adobe Firefly – Generative AI for creatives. https://firefly.adobe.com/generate/images. Accessed 7-3-2023.
5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale visual recognition. https://www.robots.ox.ac.uk/~vgg/research/very_deep/ (last access: 2023.07.07).
6. Chollet F. keras. 2015.GitHub. https://github.com/fchollet/keras
7. Peterson CJ. ChatGPT and Medicine: Fears, Fantasy, and the Future of Physicians. The Southwest Respiratory and Critical Care Chronicles 2023;11(48):31–37.