# IN*sight*FUL SALIENCY MAPS: INTERACTIONS TO UNDERSTAND MODEL BEHAVIOR

**Swetha Kannan & Katelyn Morrison**
Introduction to Data Visualization
Fall 2021
{swethak,kcmorris}@andrew.cmu.edu

## ABSTRACT

Explainability techniques for image classification such as saliency maps provide insight into which regions of the image influenced the prediction. Saliency techniques can assist data scientists in model debugging by identifying features in images that are not relevant to the image class. However, the static presentation of saliency maps withhold data scientists from further exploring model behavior on out-of-distribution data. Incorporating saliency maps in an interactive tool will enable data scientists to debug their models while understanding how their model behaves on out-of-distribution data. We present a protoype that incorporates saliency maps in an interactive, exploratory tool to showcase novel interactions that data scientists can experience with saliency maps.

## 1 INTRODUCTION

AI is being used for several high-stakes computer vision applications, such as radiology and disaster relief, to help practitioners make quick and accurate decisions. It's important to get vision models correct in order to facilitate this important work. To improve accuracy, we need to understand why the model made a certain prediction and know when it is not reliable. Explainable AI techniques have been developed to provide insight into what contributed to a certain prediction, but these techniques were designed for users with the intuition of data scientists. We present a prototype for several different interactions one can have with saliency maps in hopes to gain insight into how a computer vision model works to improve understanding and collaboration.

**Contributions.** The main contribution of our project is that we've prototyped a webpage that allows users to interact with saliency maps in different ways. This way, the webpage can be used in a research method to gain insight about model behavior and limitations.

## 2 RELATED WORK

With recent advances in artificial intelligence and its use in real-world systems, research communities across AI, visualization and human-computer interaction have been developing new ways to interact with and understand AI. Through user studies, researchers have been able to show the effectiveness of these novel interactions.

### 2.1 INTERACTIVE MACHINE LEARNING

Within recent years, several interactive machine learning tools have been developed and evaluated. A comprehensive survey on interactive machine learning papers identifies exactly what questions the proposed interaction or interactive tool addresses Hohman et al. (2018). Furthermore, an alternative publisher, `Distill.pub`, published interactive articles based on machine learning topics Carter & Olah.

Within interactive machine learning, several contributions are geared specifically towards understanding image classification models. For example, Cabrera et al. (2018) allows users to change

images by erasing different objects within the image in order to understand how image classification models work. This tool provides a table of predictions along with the confidence of the prediction to see how the image modification changed the predictions. More recently, Park et al. (2022) presents a comprehensive framework that visualizes concepts learned by neural networks that allows the user to view the learned concepts in a variety of ways including viewing clusters of learned concepts.

### 2.2 Saliency Maps User Studies

One particular concept in computer vision that has received little attention as of yet is saliency techniques. Saliency techniques are a way to present a visualization of the most salient, or important, features that contributed to the classification of an image Simonyan et al. (2014). Furthermore, Brennen (2020) and Liao et al. (2020) emphasize that these XAI techniques do not address topics that decision makers wish to see to make better decisions such as the limitations of the AI, while Saporta et al. (2021) shows that certain XAI techniques rarely highlight clinically meaningful regions in medical images. Furthermore, Nourani et al. (2019) shows that meaningless explanations, or explanations that highlight unrelated regions of an image, for image classification are not helpful.

One user study evaluates how useful pixel-based saliency maps are as explanations to end users Alqaraawi et al. (2020). Through a between-subject study, the authors evaluate if a pixel-based saliency map, LRP, helps users understand how a convolutional neural network classifies images. They determine this by having the users predict the behavior of the model on other images. Alqaraawi et al. (2020) states there are very few studies evaluating the usefulness and helpfulness of different saliency techniques for image classification through user studies. To our knowledge, we have not seen any contributions propose to explore how interacting with saliency maps impacts the users ability to understand or trust the model.

## 3 Methods

The page is divided into tabs that are meant to introduce new concepts and explorations to the user. Users do not need to explore all tabs to understand the idea behind computer vision as they are each made as stand-alone graphics. Together, however, they inform each other and can give the user a stronger understanding of computer vision models.

We chose a minimalist deign to allow the graphics to speak for themselves. The color we relied on most was blue which provided a good backdrop against the brighter colors of our heat maps: yellow, orange, red, and purple.

### 3.1 *Stylize an Image*

In the first tab, we show users how vision models can develop biases which prevent it from accurately identifying an image.

Out-of-distribution test data can be generated by producing a texture-cue conflict on the original image through neural style transfers using convolutional neural networks Michaelis et al. (2019). This technique is used to create the stylized-imagenet data set created by Geirhos et al. (2018) which motivates the creation of our own out-of-distribution image set. Out-of-distribution images can also be generated using a variety of different augmentation techniques that we did not explore . In Figure 1, the user can choose to make their own out-of-distribution data by stylizing a shape with a conflicting texture.

This tab also features a local explainability technique to allow the data scientist to individually gain insight into how their model performs on out-of-distribution images. A saliency map is generated for the stylized image that is chosen and is overlaid on top of the stylized image. A slider allows for the user to dynamically change the opacity of the saliency map to directly associate regions of the image with regions of the saliency map. To the side of the stylized image, Figure 1 shows an interactive bar chart featuring the top-5 predictions.

**Top-5 Predictions.** The interactive bar chart shows the top-5 predictions for the Inception V3 convolutional neural network Szegedy et al. (2015). These predictions were generated on Google Colab

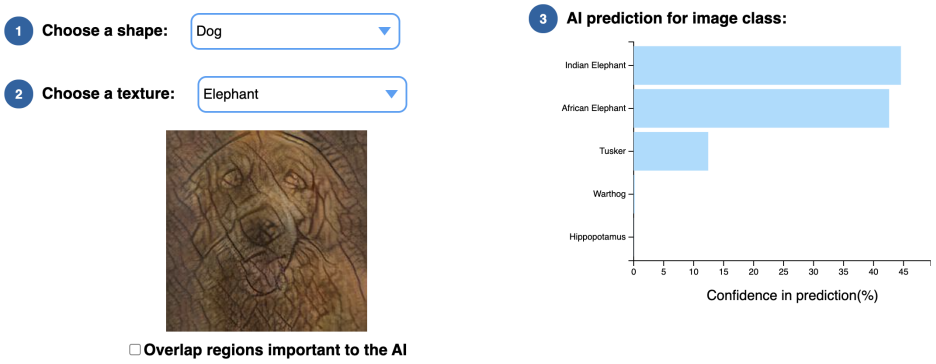


Figure 1: Bar chart on the right reflects the top-5 predictions and confidence for selected image. Predictions are generated from InceptionV3.

by evaluating the model on each image we created for our prototype. We used an ImageNet normalization on all of the images for preprocessing.

### 3.2 *Saliency Similarity*

In this tab, we compare saliency maps to each other by highlighting the similarities.

Visualizing the saliency map for an out-of-distribution image can provide insight into model behavior, however the view is limited in that you cannot easily or directly compare it to a similar image within distribution. The following novel interactions aim to provide deeper insight in model behavior through interactively comparing saliency maps.

***Compare Saliency Maps.*** A stylized image along with its saliency map are presented side-by-side allowing for an indirect comparison between the two. Similar to the *Stylize an Image* tab, a dynamic overlap feature is included. In order to compare directly to an image that is within the distribution, the same features are provided to the side of the stylized image. Showing both of the saliency maps at the same time, seen in Figure 2, allows for extra interactions to take place that are defined below.

***Visualize Saliency Similarity.*** Visualizing the similarity between two saliency maps is currently very difficult and there are few interactions designed that enable this. This feature provides insight into which features in an image are being picked up across stylized versions of that image to better understand the model's behavior. By calculating the intersection between the two saliency maps, we show the colored regions that are the same between two saliency maps. The intersection algorithm compares RGB value for every pixel in both images. As seen in Figure 2, a slider, *delta*, is provided to allow the user to change the range of RGB values included in the intersection generation.

Along with the intersection of the two saliency maps, a user can see the top-$K$ colors that are in the saliency map. The top-$K$ colors are calculated using KMeans clustering algorithm from `sklearn`. By comparing the two pie charts, the user can get a global understanding of how the two saliency maps are different or similar. This is visualization makes it slightly harder to directly interpret, but it can give a more global sense of the image in terms of how many regions were identified as important (yellow).

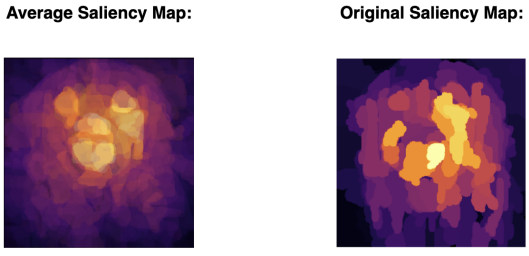


Figure 3: Visualizing average saliency map interactions under the *Saliency Similarity* tab.

***Average Saliency Map.*** There are several ways to stylize and corrupt an image to gain insight into model behavior, but it may be laborious to individually look through all of these saliency maps. The *Average Saliency Map* feature al-

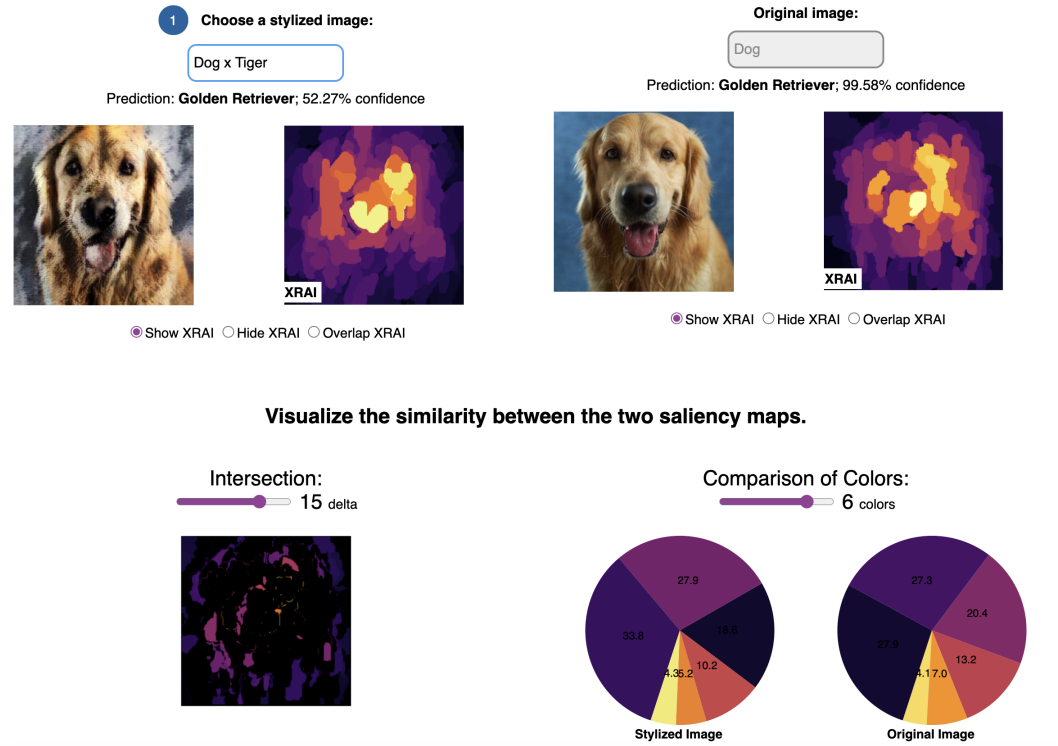


Figure 2: When the *show* option is selected for the *Compare Saliency Maps* feature within the *Saliency Similiarity* tab, two visualizations appear below: the intersection between the two saliency maps and the top-k colors for each saliency map.

lows the user to select stylized images for a given image to compute the average saliency map. The average image is approximated by averaging the pixel intensities across every image in a set. We accomplished by implementing an algorithm available in Python Unknown (2013) using `OpenCV.js`. After generating the average saliency map, the user can indirectly compare this to the saliency map of the original image as seen in Figure 3.

### 3.3 *Model Comparison*

This tab was included to highlight how these interactions can be employed to identify model robustness to corrupted or stylized inputs.

Although there are limited comparison methods offered on this tab currently, we show that when comparing the intersection of the XRAI from the stylized image and the XRAI from the the original image, the ConvMixer Anonymous (2021) uses the same most salient region to classify that it is a golden retriever. Although this is one image for one stylization, this shows that this metric can be used in a larger analysis to compare model robustness.

In Figure 4, when we compare the intersection of the stylized and original saliency maps across three different convolutional

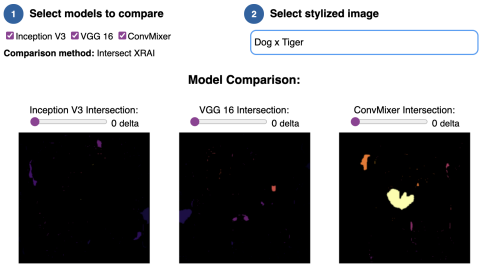


Figure 4: Comparing the intersection of the XRAI saliency map from the stylized image with the XRAI saliency map from the original image in the *Model Comparison* tab.

neural networks, we can see a yellow region for the ConvMixer model while the other two models show mostly a black image. This shows that even though the image was stylized with a conflicting texture, the most important feature that contributed to the classification of the stylized important was the same most important feature that contributed to the classification of the original image.

### 3.4 *Explain an Image*

In this tab, users are able to gain a better understanding of how computers see an image by actively trying to guess how the image's saliency map might look like.

***Draw on Canvas.*** Users are able to draw directly onto a picture to predict what the computer would see. To do this, users are given four colors which are limited for user comfort.

***Dyanamic bar chart.*** To the right of this tab, users can see a dynamically changing bar chart which shows the user how close their predictions are to what the computer actually sees. The chart is updated every time the user draws something new onto the image. If the user accurately predicts when a region is color, for example, yellow (and thus the most important region in the saliency map), then the chart updates to show that the user has guessed correctly. The chart does not update for inaccurate predictions. 5

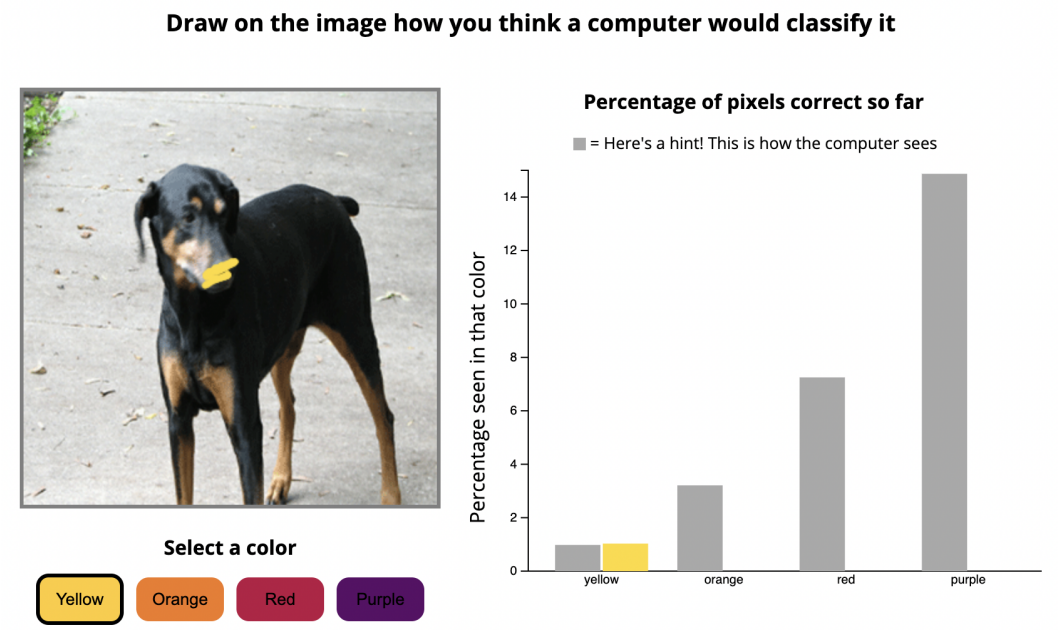


Figure 5: Bar chart on the right reflects what user draws on the image to the left under the *Explain an image* tab.

### 3.5 Accessibility

The interactive charts in our tool are not set up to work using keyboard shortcuts (i.e., tab, space, arrow keys). However, we made sure to include alternative text for all our images and most visualizations used on this site are just that - images. Because of this, a screen reader would still be able to follow along with the graphics on the web page.

To further make sure that the page was accessible, we ran an accessibility audit using the WAVE tool. Through this we were able to make sure that all our inputs were properly labelled and the colors we used were accessible.

# 4 Proposed User Study

The next step of this research project that we were not able to execute involves conducting a user study to evaluate the impact of the interactions in this tool. However, we provide a preliminary scaffold of how we might run a user study for our proposed interactions.

We would recruit two different groups of people: one group consists of data scientists and machine learning engineers while the other group consists of decision makers who collaborate with AI for image classification or object detection tasks. Since these two groups use saliency maps for different reasons, these two groups would go through different tasks. The group of data scientists would be evaluated on how well they understand a model's behavior by simply generating and looking at a saliency map in a jupyter notebook. They will then use our tool for the same model and images. After both tasks, they will be asked some questions to see how well they understand the limitations and behavior of the model. For the group of decision makers, they would be evaluated on how much impact our tool has on calibrating their trust in the model's predictions. As a baseline, we will assess their trust in the model's predictions by showing no saliency map and then showing a static saliency map.

## 4.1 Expected Results

For the data scientist group, we expect that the average saliency map and the intersection of the two saliency maps may be the most helpful debugging tools while the other features may not be as helpful to them. On the other hand, for the decision makers, we expect that the features on the *Stylize an Image* tab and on the *Explain an Image* tab will be most helpful along with the direct comparison on the *Saliency Similarity* tab. Overall, we expect that the presence of interactions for saliency maps will definitely improve their usability and capabilities in the future.

# 5 Future Work & Discussion

This interface is only the beginning of interactive saliency maps. There is plenty room for future work and researchers are encouraged to explore yet more different interactions to have with saliency maps. A good future direction for researchers that want to improve saliency maps for data scientists might want to explore how to incorporate these interactions into a jupyter notebook widget. Researchers that are interested in improving human-AI collaboration for decision makers might want to further explore comprehensive frameworks that can foster interactive saliency maps for decision makers regardless of the image data or task (i.e. classification, object detection).

In terms of our proposed framework, there are many features that we did not have time to include such as adding tooltips for the interactive bar charts or adding more expansive features to the model comparison tab. Another feature that would be recommended to include in the future is to upload you own pre-trained model and set of shapes and textures so the application is unique to an individual's needs.

**Discussion.** Overall, users will walk away from our interactive tool with an insight into the behavior and limitations of current computer vision models. By interacting with saliency maps in a variety of ways, users will gain insight into how to debug their model - potentially by training on more augmented data or maybe modifying the architecture of the model altogether.

This prototype can not only help teach people about saliency maps and but can also help researchers figure what users learn from engaging in the interactions on the web page.

# References

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pp. 275–285. Association for Computing Machinery, Mar 2020. ISBN 9781450371186. doi: 10.1145/3377325.3377519. URL https://doi.org/10.1145/3377325.3377519.

Anonymous. Patches are all you need? Sep 2021. URL https://openreview.net/forum?id=TVHS5Y4dNvM.

Andrea Brennen. What do people really want when they say they want "explainable ai?" we asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pp. 1–7. Association for Computing Machinery, Apr 2020. ISBN 9781450368193. doi: 10.1145/3334480.3383047. URL https://doi.org/10.1145/3334480.3383047.

Ángel Alexander Cabrera, Fred Hohman, Jason Lin, and Duen Horng Chau. Interactive classification for deep learning interpretation. *Demo, IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Shan Carter and Chris Olah. URL http://distill.pub/about/.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. Sep 2018. URL https://openreview.net/forum?id=Bygh9j09KX.

Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, Apr 2020. ISBN 9781450367080. doi: 10.1145/3313831.3376590. URL https://dl.acm.org/doi/10.1145/3313831.3376590.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:97–105, Oct 2019. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5284.

Haekyu Park, Nilaksh Das, Rahul Duggal, Austin P. Wright, Omar Shaikh, Fred Hohman, and Duen Horng Chau. Neurocartography: Scalable automatic visual summarization of concepts in deep neural networks. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2022. doi: 10.1109/TVCG.2021.3114858. URL https://poloclub.github.io/neuro-cartography/.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. *Benchmarking saliency methods for chest X-ray interpretation*. Oct 2021. URL https://www.medrxiv.org/content/10.1101/2021.02.28.21252634v2. Type: article.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034 [cs]*, Apr 2014. URL http://arxiv.org/abs/1312.6034. arXiv: 1312.6034.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *arXiv:1512.00567 [cs]*, Dec 2015. URL http://arxiv.org/abs/1512.00567. arXiv: 1512.00567.

Unknown. python - how to get an average picture from 100 pictures using pil? - stack overflow, 2013. URL https://stackoverflow.com/questions/17291455/how-to-get-an-average-picture-from-100-pictures-using-pil.