A generalizable saliency map-based interpretation of model outcome

Shailja Thakur
Electrical and Computer Engineering
University of Waterloo
s7thakur@uwaterloo.ca

Sebastian Fischmeister
Electrical and Computer Engineering
University of Waterloo
sfischme@uwaterloo.ca

***Abstract*—One of the significant challenges of deep neural networks is that the complex nature of the network prevents human comprehension of the outcome of the network. Consequently, the applicability of complex machine learning models is limited in the safety-critical domains, which incurs risk to life and property. To fully exploit the capabilities of complex neural networks, we propose a non-intrusive interpretability technique that uses the input and output of the model to generate a saliency map. The method works by empirically optimizing a randomly initialized input mask by localizing and weighing individual pixels according to their sensitivity towards the target class. Our experiments show that the proposed model interpretability approach performs better than the existing saliency map-based approaches methods at localizing the relevant input pixels.**

Furthermore, to obtain a global perspective on the target-specific explanation, we propose a saliency map reconstruction approach to generate acceptable variations of the salient inputs from the space of input data distribution for which the model outcome remains unaltered. Experiments show that our interpretability method can reconstruct the salient part of the input with a classification accuracy of 89%.

***Index Terms*—Interpretability, deep learning, Images**

I. Introduction

Recent advances in the field of deep neural networks have led to widespread applicability of artificially intelligent systems in the field of computer vision for the task of object detection [1], image classification [2], segmentation [3], image captioning [4], visual question-answer [5]. Despite the significant advances in the speed and accuracy of neural networks, the complexity of the models makes the human-level understanding of the model's decision-making a challenging problem. Notably, the highly non-linear interactions between the layers of the network make the outcome unintuitive and unpredictable. As a result of their inexplicable nature, their applicability remains limited in the domain of safety-critical systems (medicine, automotive, robotics, finance, nuclear) where a decision based on the outcome can lead to fatal consequences.

Some of the research in the direction of explainable AI elucidate instances reflecting on the unpredictable nature of complex machine learning systems. For instance, in [6], the author shows how bias manifested in the machine learning algorithm through data leads it to misconstrued the characteristics of the snow for that of *husky*. Another work by Stock et al. [7] demonstrates the ImageNet [2] bias introduced in the ResNet [8] model. As a consequence of the bias, the model prefers the image of a black person with a basketball for the class basketball, and Asians in red dress for the *ping pong* class. Athalye et al. [9] show the sensitivity of the model can lead to misclassification. In the paper, the authors demonstrate that adding imperceptible perturbation to the input causes the model to misclassify the image of the turtle to the class of *rifle*.

All the above examples scenarios show the unpredictable nature of model prediction in the presence of uncertainty. Consequently, such advanced AI systems cannot be reliably used for decision making in critical systems that demand explanation and verification.

As a consequence of the black-box nature of complex AI systems, many possible solutions for understanding and interpreting the complex machine learning models have been developed in the last couple of years. One such technique is to visualize the activations of the individual layers of the network [10]. However, for a particular image, this method can only tell apart what neurons are important for classifying the input to a certain class. Salient map-based methods exist [11] that localize the input pixels sensitive to the classification of the input to an output class. However, one of the limitations of these techniques is that they are intrusive; they require access to the network parameters and gradients flowing through the network to localize the important input pixels. Thus, there is a need for a *non-intrusive explanation technique* for target-specific model outcomes.

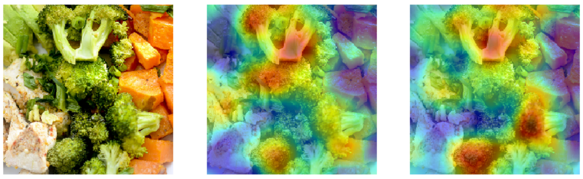


Fig. 1. Saliency map generated for the image containing *broccoli*. The two saliency map for *broccoli* highlights non-overlapping regions of the image as important for *broccoli* classification.

Furthermore, explanation using perturbation based techniques lacks consistency in the explanation. As shown in Figure 1, multiple iterations of explaining the same image result in contradictory salient region detection for the target

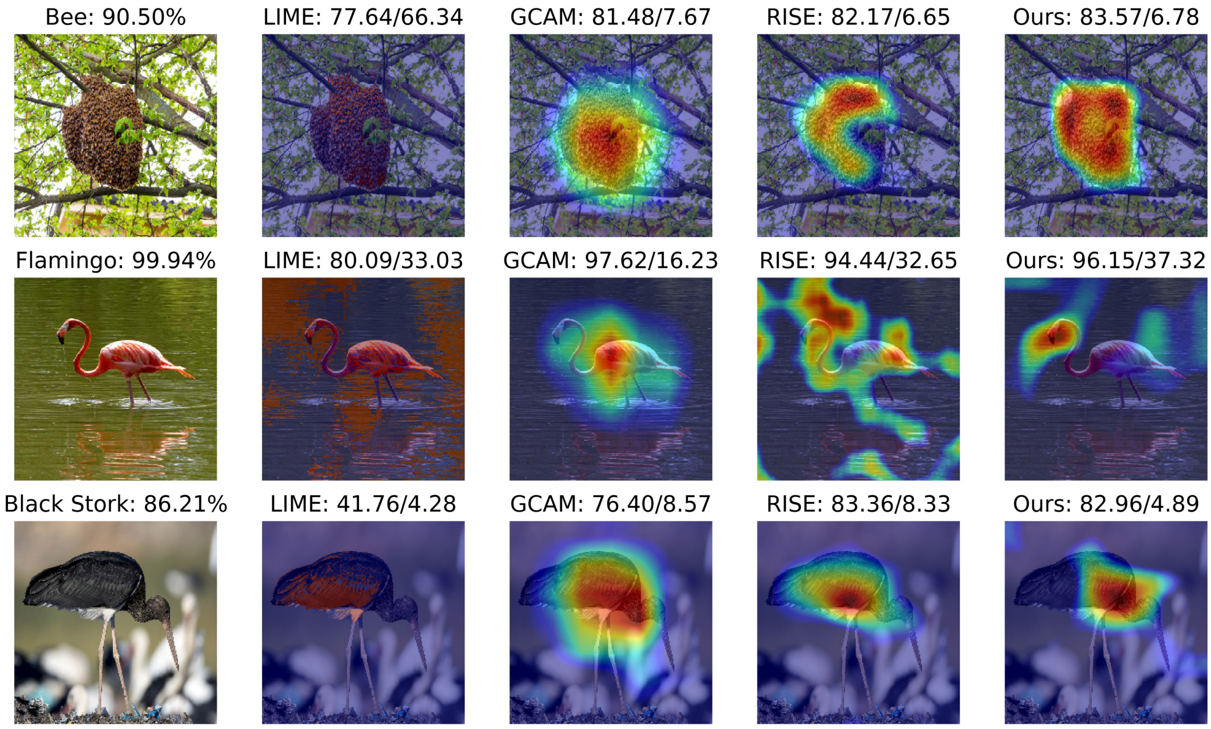


Fig. 2. Saliency map generated for the target-specific image classification using our approach, RISE [12], GCAM [13], and LIME [6]. The first column shows the input image along with the top predicted class of the model outcome and the accuracy of classification. Second column onwards shows the saliency map overlapped with the input image and the AUC scores (%) of insertion/deletion metrics [12] where a higher value is considered good for insertion, and a lower value is considered good for deletion.

class. Therefore, there is also a need for a global perspective on the explanation for a target class.

A. *Contribution*

Our contribution to the paper is two-fold. First, we propose a non-intrusive interpretability technique by generating a saliency map based target-specific model outcome explanation. And second, we propose a method for generating *alternate explanations* for the part of the input, which is salient for target-specific classification.

Inspired by the work of Petsuik et al. [12], we propose a non-intrusive explainability technique that leverages the input and the output of the model to generate a class-specific saliency map. Our method differs from RISE [12] in the sense that for a target class, instead of performing MCMC sampling [14] of input subsamples, we generate a class-specific saliency map by optimizing a randomly initialized mask. We use an *empirical risk minimization* approach with a randomly initialized mask to locate the optimal set of input pixels sensitive for the classification of the input to the target class. Therefore, if for the masked input, the confidence of the model in the most probable class is given by $p \in [0, 1]$, then the optimal pixels for the input is empirically located by randomly retaining $p\%$ of the unmasked pixels (pixels with value 1) and $(1 - p)\%$ of the masked pixels (pixels with value zero) followed by weighing the pixels using the class score. Figure 2 shows a visual comparison of the saliency for example images from ImageNet [2] using our approach, RISE [12], GCAM [13], and LIME [6].

To generalize the explanation for the target class using the saliency map, we propose a technique to identify the variations of the pixels in the salient regions of the input for which the model prediction remains unaltered. The hypothesis for finding variations of the salient region comes from the analogy that the model is invariant to small perturbations in the input. Thereby, the approach helps identify variations (changes in colour intensities, object rotation, or inversion) in the salient region of the input space for which the model classification remains unaltered [15]. To generate alternative explanations for the salient regions of the input, we use an image completion technique [16], [17] that uses the neighbouring pixels features to reconstruct the pixels in the salient regions of the input. Using this approach, we can find an exhaustive and contextually similar set of transformations for the pixels in the semantic regions classified to the same output class as the original input image.

II. Related Work

With the increasing applicability of complex machine learning models, the need for an explainable and verifiable AI is increasing. In an attempt to justify models' outcomes, a variety of techniques have been proposed over the years. A majority of the explainability techniques use either a *third-party explainer*: a separate model for explaining the outcomes of the base model, or is based on justifying the base model outcome using techniques such as input perturbation and network parameters.

One of the earliest *third-party explainer* [6] attempts to provide a linear interpretation within the local neighbourhood of the data point. However, the approach is not effective at explaining non-linear models. [18] is another technique that trains two models for providing textual as well as visual justification for the visual question answering task and activity recognition task. However, these approaches are costly to implement because of the reliance on the availability of large human-annotated ground-truth explanations.

Within the realm of justifiable models, a variety of explainability approaches have been developed. Some of the approaches [10], [11], [19], [20] attempts to synthesize input images that result in a high activation score for particular neurons. Another approach by [21] generates a target-specific saliency map by taking the global average pooling of the feature maps at the layer before the fully connected layer. GradCAM [13] is a generalized version of CAM that, in addition to the feature map weights, feeds the class gradient to the fully connected layer to assign importance to each of the input pixels. However, [13], [21] can only be applied to limited network architectures with global average pooling. Another work by Zhang et al. [22] proposes to use a backpropagation scheme to generate an attention map by propagating the signal downward through the network hierarchy using a winner-take-it-all strategy. A few techniques examine the relationship between input and output to learn a perturbation mask by backpropagating the error signal [23].

Despite the ability of the techniques [13], [21], [23], [24] to justify the model's decision, the methods mentioned above have limitations. The methods [13] are constrained by the use of network parameters such as gradients flowing through the network and network layer weights. While techniques such as [13], [22] require a specific kind of network architecture, in some cases [10], the method requires access to intermediate layers of computation for visualizing the features at several layers. Furthermore, the techniques can explain only a particular input at a time, without considering the possible variants (rotation, inversion, deformations) of the image. A work by Kim et al. [25] proposes a technique to provide an explanation that is representative of user-defined concepts, but the manually generated concepts limit the technique. Our work is an extension of the work by Petsuik et al. [12] to localize and generalize the salient pixels of the target class using the input and the output of the model. We obtain a saliency map (using N less than that of [12]) by empirically optimizing the pixels important for target-specific classification. Also, we propose an approach to provide a global perspective on the explanation of the outcome using a reconstruction technique, which generates possible variations of the salient pixels of the input.

III. Proposed Technique

We propose a non-intrusive explainability technique by localizing and generalizing the parts of input important for target-specific classification. To localize the pixels sensitive for target-specific classification, we extend [12] to generate a saliency map. Second, to obtain a global perspective on the saliency map of the model outcome, we propose a salient region reconstruction approach that reconstructs the input image with alternate variations of the pixels from the salient region of the input, all of which classifies to the target class.

A. Saliency Map Generation

Given an image I of dimension $H \times W$ from the space of images $\mathcal{I} = \{I : \Lambda^{H \times W} \to \mathbb{R}^3\}$ that maps each pixel coordinate to three color values, a target class $c \in \mathcal{C}$, a classifier $f : \mathcal{I} \to \mathbb{R}^{\mathcal{C}}$, which maps inputs from the input space, $\mathcal{I}$ to a vector of real numbers signifying the strength of the classifier in the output classes, $\mathcal{C}$, a random initial mask, $M_0 = \{\Lambda^{h \times w} \to [0, 1], h < H, w < W\}$. M_0 is composed of a set of unmasked and masked pixels such that $\Lambda^{h \times w} = \{\Lambda_{\text{on}}^{h1 \times w1} \cup \Lambda_{\text{off}}^{h2 \times w2}\}$, and n_1 and n_2 are the number of pixels in the Λ_{on} and Λ_{off} sets. Thus, given a binary mask M_0, if $\lambda \in \Lambda_{\text{on}}$, then $M_0(\lambda) = 1$ else $M_0(\lambda) = 0$.

We use the *empirical risk minimization* approach to find a mask M with a minimal set of unmasked pixels $\Lambda_{\text{on}}^{\text{min}}$. The mask is such that the model's confidence in class c using the masked version of the input departs from the confidence with that of the actual input by no more than a predefined threshold δ.

$$\begin{aligned} \underset{\Lambda_{\text{on}}}{\text{minimize}} \quad & M \\ \text{subject to} \quad & |f(I \odot M) - f(I)| \leq \delta \end{aligned} \tag{1}$$

In the equation 1, δ is a threshold determined empirically and $\odot$ is the element wise multiplication of the input with the mask. The key to a good estimate of the saliency map, which satisfies the optimization condition in Equation 1, is to iteratively update the initial mask M_0 by selectively filtering the masked and unmasked pixels from Λ_{on} and Λ_{off} based on the prediction probability $p = f(I \odot M)$ of the class c, where M is the upsampled version of the mask at the i^{th} iteration. M is upsampled using bilinear interpolation to the size of the input image as shown in the second column of Figure 3. The bilinear interpolation is a common resizing technique in computer vision that helps avoid the inclusion of unwanted artifacts in the mask during empirical optimization by blurring out the edges, as shown in figure 3, thus, eliminating misclassifications.

Algorithm 1 shows the saliency map generation method. Based on the prediction probability of the target class c, the algorithm randomly retains $p\%$ of the unmasked pixels, Λ_{on} and $(1 - p)\%$ of the masked pixels, Λ_{off}. The subset of the pixels preserved from Λ_{on} and Λ_{off} forms the new set of unmasked pixels Λ_{on}, which are sensitive for the classification of input I to the target class c. However, the mask update is likely to saturate if the change in the prediction probability p of the target class is negligible. To overcome the issue of saturation, we add a regularizer, which penalizes the set of unmasked pixels according to two factors: the change in the prediction probability of the target Δp and the total variation

Algorithm 1 Algorithm for generating saliency map

```
Input: Input image $I \in \mathcal{I}$, Target class $c \in \mathcal{C}$
Output: saliency map $M$
    Initialisation : $M_0 \in \mathbb{R}^{h \times w}$
for $i = 1$ to $N$ do
    $M \leftarrow$ upsample $M_{i-1}$
    compute $p_i = f(I \odot M)$
    $\Lambda_1 \leftarrow$ randomly select $n_1 p_i$ pixels of $\Lambda_{\text{on}}$
    $\Lambda_2 \leftarrow$ randomly select $n_2(1 - p_i)$ pixels of $\Lambda_{\text{off}}$
    $\Lambda_{\text{on}} \leftarrow \Lambda_1 \cup \Lambda_2$
    $\Lambda_{\text{off}} \leftarrow \Lambda \setminus \Lambda_{\text{on}}$
    if ($\lambda \in \Lambda_{\text{on}}$) then
        $M_i(\lambda) \leftarrow M_{i-1}(\lambda)\Delta p$
    end if
    if ($\lambda \in \Lambda_{\text{off}}$) then
        $M_i(\lambda) \leftarrow 0$
    end if
    $V = \sum_{x,y} ||M_{i-1}^{xy+1} - M_{i-1}^{xy}||^2 + \sum_{x,y} ||M_{i-1}^{x+1y} - M_{i-1}^{xy}||^2$
    $M_i \leftarrow M_i + \eta V M_{i-1}$
end for
return $M$
```

of the mask V where Δ denotes the change in the value from the previous iteration. If the mask is invariant to small changes in the classification accuracy, then the regularizer will heavily penalize unmasked pixel values by adding a large negative penalty. On the other hand, if the change in Δp is significant, the regularizer adds a low penalty to the updated mask. This way, M captures the pixels sensitive towards the classification of I to the target class c. In the update equation for the mask, η is a constant that determines the amount of change in the total variation of the mask to retain.

B. Variations of the salient region of the input

We propose an approach to generate the acceptable variations for the salient region (the highlighted region in the second column of Figure 3) of the input image as *alternate explanations* for the model outcome. To achieve this, we use an image completion technique proposed by Pathak et al [17], which reconstructs a patch of the input image by iteratively backpropagating the error in a generator.

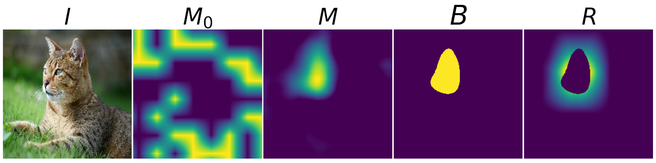


Fig. 3. From left to right, I: input image, M: saliency map, B: bounding box, R: reconstruction mask.

Let B be a binary bounding box for the saliency mask M of the input image I, as shown in Figure 3 third from left, where pixels inside the box are set to one, and the pixels outside are set to zero. The reconstruction mask, R, is obtained by inverting B, followed by convolving B with a kernel of size (s, s) such that the weights assigned to pixels are inversely proportional to their distance from the bounding box. As pixels near the box are more important for the reconstruction of missing pixels from the box, the mask R is such that it assigns more importance to the pixels in the vicinity of the box than to the pixels far away. Let G be a generator that learns an encoding d_z of the input image distribution $d_{\mathcal{I}}$ in the latent space (z).

The objective is to reconstruct the corrupt image $((\mathbf{1} - R) \odot I)$ using G. As the corrupt image is not a sample from the input distribution $d_{\mathcal{I}}$; therefore, G will be poor at recognizing the patterns of the missing part of the image. Therefore, we use the image reconstruction technique as described in [16], where the authors use the back-propagation technique with the generator to find an encoding $(z^{'})$ for the missing part of the image that is closest in encoding to the input I while being confined to the learned manifold (z). The objective function for learning the encoding $(z^{'})$ comprises context loss and discriminative loss.

Context Loss is used to reconstruct the missing part of the original image given the corrupt image by measuring a squared error between the corrupt image and the reconstructed image.

$$L_{cxt}(z) = (I \odot (\mathbf{1} - R)) - (G(z) \odot (\mathbf{1} - R)) \qquad (2)$$

Discriminative Loss is used for measuring the authenticity of the generated images by feeding them to the discriminator D, which returns the confidence in $G(z)$ being real.

$$L_{dis}(z) = -D(G(z)) \qquad (3)$$

The overall loss for learning the encoding for the missing salient region of the input is as follows,

$$L(z^{'}) = \underset{z}{\text{argmin}}\{L_{cxt}(\boldsymbol{z}|I, R) + L_{dis}(z)\} \qquad (4)$$

Using the learned encoding, G$(z^{'})$ generates the image that is approximately close to the missing salient part of the image. The image reconstructed using the generated image is given by,

$$I_{\text{rec}} = (I \odot (\mathbf{1} - B)) + (B \odot G(z^{'})) \qquad (5)$$

We repeat image reconstruction for k randomly sampled noise vectors z to generate k reconstructed variants of the salient region of the input image I where z is a sample from a gaussian distribution with mean zero and variance one. The generated images will be such that their encoding will lie in the vicinity of the learned manifold in the latent space (z). However, if the encoding $z^{'}$ fails to capture the context of the salient region of the input using the evidence from the local neighbourhood of pixels, then some of the reconstructed images will not be contextually similar to the original input salient region. The selected set of I_{rec} for which the model prediction remains unaltered is considered *alternative explanations* for the salient regions of the input image.

IV. Evaluation

This section gives an overview of the evaluation metrics, models, and datasets used to describe the results of the evaluation of the proposed approach.

A. Model and Data Description

We evaluate the efficacy of the proposed saliency map based explainability approach using a range of publically available open image datasets such as ImageNet [2] and MS-COCO (Microsoft-Common Object in Context) [26]. ImageNet is a repository of 15 M high-resolution images gathered from more than 20 K categories. MS-COCO, on the other hand, is a significantly smaller database of images but with more number of instances per category. The dataset has 330 K images with more than 80 object categories. The MS-COCO dataset is used for object detection, object segmentation, and image captioning.

We use pre-trained models: VGG16 [27], Inception V3 [28], and ResNet50 [8] as base models for image classification. The pre-trained models are loaded with ImageNet weights and accepts inputs of size 224 × 224. The models differ in terms of their network structure and number of trainable parameters. VGG16 is a convolutional neural network trained on ImageNet dataset achieving top-5 accuracy of 92.0% on ImageNet. ResNet50 is a 50 layer network achieving 93.29% accuracy on the ImageNet. Inception V3 is another model that achieves an accuracy of 94.4% on the ImageNet. Relying on pre-trained models [8], [27], [28] for image classification helps avoid the common training pitfalls such as model over-fitting, skewed data distribution, right model selection, and insufficient resources (such as GPUs) for training.

B. Evaluation Metrics

1) Insertion and Deletion Metrics: Motivated by [12], we use the metrics of insertion and deletion to evaluate our saliency map approach. In the deletion metric, the deletion of salient pixels from the input causes the model to drop the probability of the target class. And in the insertion metric, the insertion of pixels from the relevant region of the input causes the model to increase the probability of target class. We capture the sensitivity of the model to the removal and insertion of pixels from the relevant region of the input using an average AUC (Area Under the Curve) score. Thus, during deletion, as the relevant input pixels are deleted from the masked input, the AUC curve for the model will shrink to a thin area, thus, dropping the average AUC score, indicating the right explanation for the model decision. Similarly, during the insertion, as the pixels from the relevant region of the input are added to the masked input, the AUC curve expands to cover the large area under the probability curve, thus increasing the average AUC score.

2) Pointing Game: The pointing game [22] metric is a method of evaluating the class discriminative nature of saliency-map based approaches. Given an annotated segmentation box for an instance of an object and the corresponding saliency map, the method measures the number of pixels on the saliency map of the input that lies on the annotated box of the object instance. To measure the overlap, we calculate the fraction of the area (A_s) of the saliency map that overlap with the annotated segmentation box (A_t) of the image using an IOU (Intersection Over Union) score (%) = $\frac{\sum(A_s \cap A_t)}{\sum(A_s \cup A_t)}$. The numerator is the sum of pixels values in the union of A_s and A_t, and the denominator is the sum of pixels values in the intersection of A_t and A_s. A high IOU score (typically $\geqslant 50\%$) means that a large fraction of the salient region of the input overlaps with the annotated box, indicating a good explanation.

C. Evaluation using insertion and deletion metrics

Given a pre-trained classifier, we evaluate the class discriminative capability of saliency-map based approaches [6], [12], [13], [20] using the quantitative measures of insertion and deletion metric. For the base models: VGG16 [27], ResNet50 [8], and Inception V3 [28] and the datasets: ImageNet [2] and MS-COCO [26], we report the average AUC score of insertion and deletion on a set of images. The test images are randomly selected from the test set of the datasets. For each technique, Table I shows the mean AUC score of the insertion and deletion metrics for each technique across all the models and all the datasets. We also show the standard deviation of the AUC of the insertion and deletion metrics for our approach. From the table, it is evident that our approach outperforms other saliency-map based approaches in localizing the pixels sensitive for the classification of the input to the target class across both the datasets and all the base models.

D. The pointing game metrics

We evaluate the localization capability of saliency map-based approaches for target-specific objects using the pointing game. For the ImageNet dataset, we report the average IOU score for a set of test images across all the models. The IOU score is calculated for the saliency map of the images using the enclosed area of the bounding box, A_s and the annotated segmentation box, A_t for the target-specific input images. Table II shows the mean IOU score (%) for the test set of images across all the models and both the datasets. The table also shows the standard deviation of the IOU score for our approach. From the table, it is evident that the performance of our saliency map approach across the models is at least 5% more accurate than [12] and 20% more accurate than [6] at localizing the pixels sensitive towards target-specific classification. Note that we omit the evaluation of the gradient backpropagation technique using the pointing game as the technique highlights only the edges of the target object, which is insufficient for evaluation against this metric.

E. Convergence

We show that our approach converges to localize pixels sensitive for the target class in less than half the number of iterations compared to [12]. To show this, we calculated the saliency map of a set of randomly selected images from class *honey bee* using our approach and RISE [12] for $N = 5000$.

TABLE I

MEAN AUC SCORE(%) USING INSERTION (INS) AND DELETION (DEL) METRICS

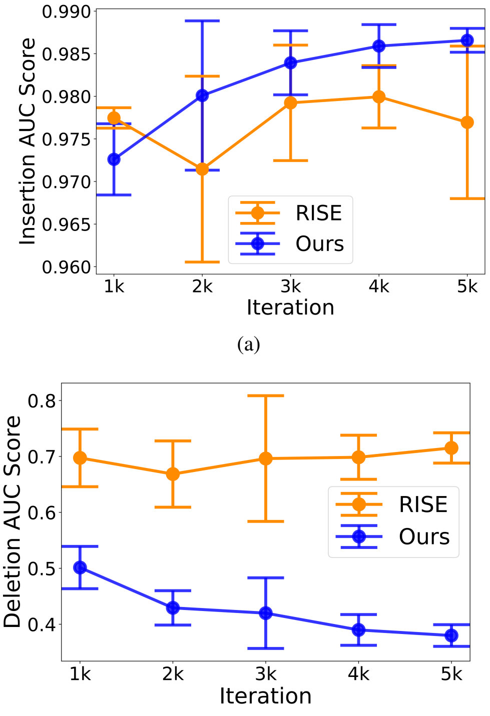
Model	Dataset	Ours		RISE [12]		GCAM [13]		LIME [6]		Guided Backprop [20]	
		Ins	Del	Ins	Del	Ins	Del	Ins	Del	Ins	Del
ResNet50 [8]	MS-COCO [26]	**75.53/0.02**	**1.80/0.001**	73.71	4.14	55.08	6.95	46.27	7.02	38.96	4.25
	ImageNet [2]	**63.16/0.004**	**11.48/0.05**	60.32	13.04	58.57	18.22	45.89	15.86	49.37	3.54
VGG16 [27]	MS-COCO [26]	**64.64/0.001**	**4.49/0.003**	62.88	5.00	40.03	10.07	37.94	7.06	38.96	4.25
	ImageNet [2]	58.72/0.03	12.23/0.1	**59.47**	**12.66**	51.59	15.75	45.75	16.23	49.20	3.13
InceptionV3 [28]	MS-COCO [26]	**66.25/0.001**	**5.88/0.005**	65.87	5.00	62.43	4.83	53.59	6.41	39.62	6.00
	ImageNet [2]	**67.43/0.004**	**6.34/0.003**	65.12	10.25	60.99	11.31	45.76	13.32	49.12	4.27

TABLE II

MEAN IOU SCORE(%) USING POINTING GAME METRIC

Model	Ours	RISE [12]	GCAM [13]	LIME [6]
ResNet50 [8]	**79.01/0.02**	74.9	69.11	57.29
VGG16 [27]	78.0/0.001	**81.12**	62.31	51.17
InceptionV3 [28]	**76.0/0.002**	63.23	57.54	48.21

The criteria for choosing a particular image category includes the presence of at least one distractor object, which makes the target class discrimination difficult. Figure 4(a) and 4(b) shows the insertion and deletion metric of the saliency map at every iteration. The error bar at each iteration shows the standard deviation of the average AUC. For every 1000^{th} iteration, we report the average AUC of insertion and deletion of the pixels using the saliency map generated using both the approaches. From Figure 4(a), we observe that the accuracy increases over iterations and then approaches a point after which updating the saliency mask does not improve the performance of insertion accuracy. This point is the point of convergence. The figure shows that our approach reaches the point of convergence faster than [12]. Furthermore, unlike [12], the monotonically increasing and decreasing curve of the mean AUC of insertion and deletion metric shows that the saliency map over iterations generated using our approach is more reliable for decision-making.



(a)

(b)

Fig. 4. The figure shows the AUC score of insertion 4(a) and deletion 4(b) for the saliency map of an input image using our approach and RISE [12] over the iterations.

F. Evaluation of the variations of the salient region

In this section, we evaluate the generalizability of our saliency map approach on an input image of a *lynx*. The image is chosen such that the presence of distractor objects makes the discrimination of the target class against other classes challenging using a saliency mask. For instance, the image of *lynx* has a background whose pattern and colour match with that of the lynx, which contributes to reducing the target accuracy to 62.15%. However, we found that the results of the evaluation followed a similar trend for other images evaluated from the test set of the ImageNet dataset.

1) Classification Accuracy of reconstructed images: Given a saliency map for an input image of *Lynx*, we report the classification accuracy of a subset of reconstructed images of lynx that are correctly classified as *Lynx* by ResNet50. To obtain the reconstructed images, first, we created a reconstruction mask R by convolving the bounding box for the saliency map of *Lynx* with a kernel of size $(15, 15)$. We also trained a GAN [29] as the generator (G) using a set of 40K training images from the ImageNet dataset. The input images were downsampled to a size of (64×64) to speed up the training process of G. The training is followed by feeding G with the input image, the reconstruction mask and a batch of 64 random noise vector z to generate a batch of 64 images. The generated images are used to reconstruct a set of 64 images using Equation 5.

Figure 5 shows a subset of the reconstructed images that are correctly classified to the target class Lynx with an average accuracy of 64%. The figures are pixelated because the reconstructed images are resized from size $(64{\times}64)$ to size (224×224) where (64×64) is the size of generator output and (224×224) is the size of the saliency map. From the figure, we observe that most of the reconstructed parts of the

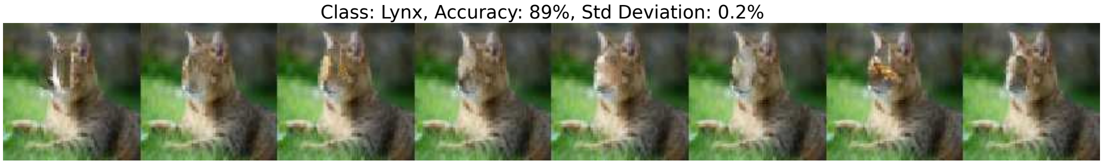


Fig. 5. Reconstructed images for the image of a *Lynx* with saliency map (M) as shown in Figure 3

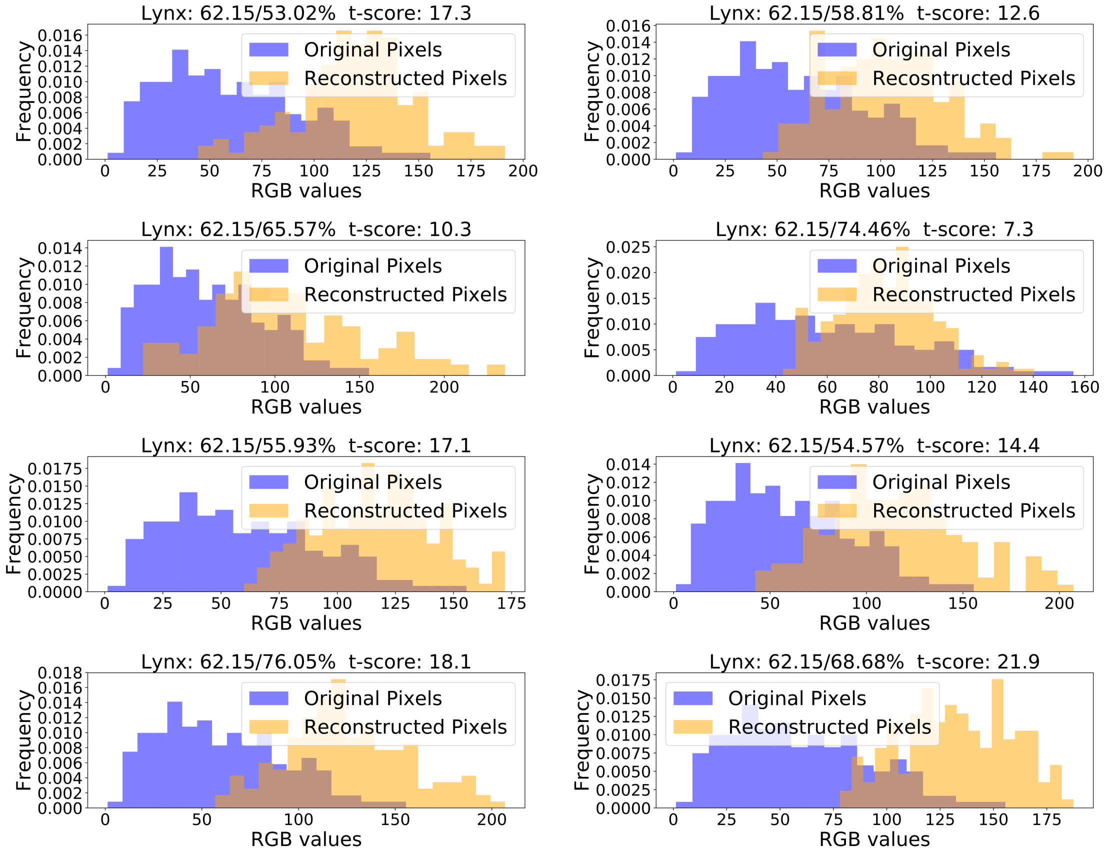


Fig. 6. Figures showing the histogram of the reconstructed salient pixels and the original salient pixels. The title of the sub-figures show the accuracy of the class *Lynx* for input image/reconstructed image.

images (face including eyes, nose, mouth, left side of the face) contain a blob, which is black in the vicinity of the mouth and the lower part of the nose. The presence of a consistent feature across the reconstructed images shows that the area around the nose and the mouth of the image is essential for classifying the image to the class of *Lynx*.

Figure 6 shows the histograms of the reconstructed salient region of the input and the original salient region of the input. For each of the reconstruction, the figure also shows the accuracy of the target class *Lynx* and the t-score. The t-score is a measure to tell apart the difference between the reconstructed pixels and the original pixels; thus, the higher the t-score value, the larger the differences. Based on the t-score, it is evident that the reconstructed pixels classified to the target class *Lynx* have variations that are significantly different from the original pixels. Thus, these reconstructed pixels of the salient region of the input are the variations, which form alternate explanations for the target class.

2) *Impact of varying sizes of bounding boxes:* We show that the size of the bounding box enclosing the salient region of the input influences the quality of the reconstructed images and hence the alternate explanations. The idea is that as the size of the salient region to reconstruct shrinks, the evidence from the neighbourhood to reconstruct missing pixels increases, thereby generating contextually similar images. We reconstructed the input image using the varying sizes of the bounding boxes by reducing the bounding box by a factor of $\alpha = 0.1$ until half the original size. From Figure 7, it is evident that the number of reconstructed images correctly classified to the target class *Lynx* increases as the size of the bounding box decreases. The most number of correct classifications are observed with the bounding box that is half the size of the original salient region. Equivalently, as shown in Figure 8, the loss incurred by the generator during the reconstruction of the images decreases

as the size of the bounding box decreases. This shows that a carefully chosen value of α helps generate contextually similar images. For the image of lynx, a value of 8α, which retains 80% of the salient region in the bounding box, generates reconstructed images with an average classification accuracy of 63.4%.

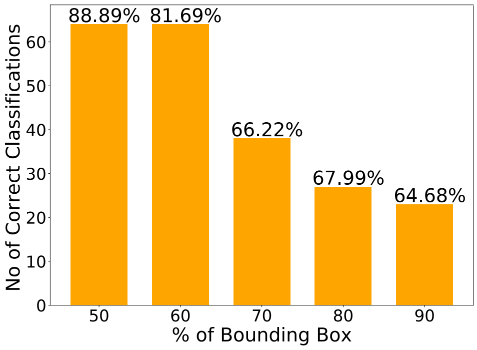


Fig. 7. Figure shows the number and accuracy of correct classifications using the reconstructed images over different sizes of bounding boxes.

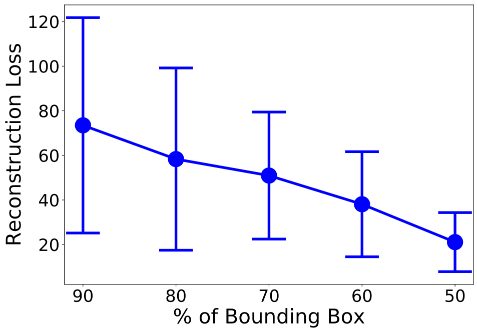


Fig. 8. Figure shows the reconstruction loss over the different sizes of bounding boxes.

V. Conclusion and Future Work

In this paper, we proposed a generalizable saliency map-based explainability technique for explaining the target-specific outcome of a model. We evaluated our saliency map based explanation approach using two datasets: ImageNet and MS-COCO, and against a set of models: VGG16, ResNet50, Inception V3. We also compared our approach against other saliency map-based techniques. Experiments show that our method performs better than the existing explainability approaches across most of the tested scenarios. Furthermore, to evaluate the generalizability of our approach, we evaluated the correctness of the alternate explanations. The accuracy of the reconstructed input using the saliency map as the mask shows that the approach can find explanations that are relevant to the classification of input to the target class.

As part of future work, we will extend our non-intrusive saliency map-based approach for explaining the mode outcome for the input of type time-series as well sequences.

References

[1] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013.

[2] D. J., D. W., S. R., L. L.-J., L. K., and F.-F. L., "Imagenet: A large-scale hierarchical image database," 2009.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2014.

[4] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemela, , and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," 2015.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," 2016.

[7] P. Stock and M. Cissé, "Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism," 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[9] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," 2018.

[10] J. Yosinski, J. Clune, A. M. Nguyen, T. J. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015.

[11] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2014.

[12] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," 2018.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, , and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," 2017.

[14] B. Walsh, "Markov chain monte carlo and gibbs sampling," 2004.

[15] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng, "Measuring invariances in deep networks," 2009.

[16] R. A. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with perceptual and contextual losses," 2016.

[17] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," 2016.

[18] D. H. Park, L. A. Hendricks, Z. Akata, B. S. Anna Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," 2018.

[19] A. M. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," 2016.

[20] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2015.

[21] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." *CVPR*, 2016.

[22] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," 2016.

[23] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," 2017.

[24] D. Piotr and G. Yarin, "Real time image saliency for black box classifiers," 2017.

[25] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Vigas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," ser. Proceedings of Machine Learning Research, 2018.

[26] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," 2014.

[27] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," 2014.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," 2014.