Performance Evaluation of Combination Methods of Saliency Mapping Algorithms

Ian E. Nielsen, Jacob R. Epifano, Ghulam Rasool, Nidhal C. Bouaynaya and Ravi P. Ramachandran
Department of Electrical and Computer Engineering
Rowan University, Glassboro, NJ, 08028, USA
nielseni6@students.rowan.edu, epifanoj0@students.rowan.edu, rasool@rowan.edu, bouaynaya@rowan.edu, ravi@rowan.edu

***Abstract*—The multilayer and nonlinear structure of complex machine learning models is widely referred to as the black box because they are not transparent and hence, their decisions are difficult to explain. Interpretability techniques are used to understand the black box nature of these complex models by facilitating a comprehension of how and why a decision is taken. One technique is the saliency map which processes an image to depict each pixel's visual significance as a gray scale value. This paper (1) examines and evaluates several recent saliency mapping algorithms with respect to how effective they are at interpreting the results of a deep convolutional neural network and (2) configures and assesses novel combinational techniques. Both qualitative and quantitative metrics generally show that combination methods give better results.**

I. Introduction

The decisions of complex machine learning models like deep neural networks are difficult to explain since one can only directly observe their inputs and outputs. Not knowing what the algorithm of these models is doing is widely referred to as the black box phenomenon [1]. Interpretability techniques for machine learning models are useful tools for understanding what goes on inside this black box. Diverse applications of interpretability include the medical field [2][3][4], natural language processing [5], finance and genomics [6].

A common approach to model interpretation involves calculating an attribution score for each feature based on its contribution to the classification decision of the model. For the purposes of this paper, the features are the pixels since the application is image processing. The attribution scores reveal the extent to which each pixel contributes to the overall label prediction of a classifier. These scores are displayed as a saliency map (or image map) in which each pixel is set to a gray scale value based on its score. The saliency map constitutes the attribution scores for each pixel The term saliency map is also referred to as a sensitivity map, pixel attribution or feature attribution. Throughout this paper, the term "saliency map" will be used. Saliency mapping seeks to provide a visual representation of how each pixel affects the soft-max predictions or class activation scores of a neural network.

Multiple methods for producing saliency maps have been developed including SmoothGrad [7], Integrated Gradients [8] and Guided Backprop [9]. In this paper, methods are combined to produce better results. The goals of this paper are to (1) evaluate these three existing methods of saliency mapping and (2) configure combinations of each to produce better results. Each technique is assessed using qualitative and quantitative metrics. By evaluating the combinations as well as prior methods, the aim is to provide ample evidence that these methods are accurate and valuable for understanding the decisions produced by neural networks.

II. Methods

This section gives a description of three developed methods and the new combination techniques of this paper.

A. SmoothGrad

The SmoothGrad method [7] uses the "Vanilla Gradient", which assigns attribution scores (denoted as $M_c(x)$) to each pixel by taking the partial derivative of the class activation score for class c (denoted as $S_c(x)$ with respect to the image x [10] as given in Eq. (1).

$$M_c(x) = \frac{\partial S_c(x)}{\partial x} \tag{1}$$

This establishes the gradient saliency map $M_c(x)$. Equation (2) [7] defines the smooth gradient $\hat{M}_c(x)$ as

$$\hat{M}_c(x) = \frac{1}{n}\sum_{i=1}^{n} M_{ci}(x + N(0, \sigma^2)) \tag{2}$$

where M_{ci} denotes the ith noisy sample of $M_c(x)$ formed by adding Gaussian noise with mean 0 and a standard deviation of σ to $M_c(x)$. A total of n noisy samples are averaged to form the smooth gradient or SmoothGrad saliency map $\hat{M}_c(x)$. The implementation in this paper uses 15 noisy samples and $\sigma = 0.15$ as given in Eq. (2).

B. Integrated Gradients

The Integrated Gradients approach [8] uses multiple images along a straight line path between a baseline image x' and the input image x in order to compute the saliency map as given in Eq. (3). The gradient of each of these images along the path is computed. The Integrated Gradient is defined to be the path integral of the gradients along this straight line path [8]. The integrated gradient along the ith dimension (or equivalently the ith pixel) is defined in Eq. (3)

$$\text{IntegratedGrads}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial S_c(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{3}$$

where $\frac{\partial S_c(x)}{\partial x_i}$ is the gradient along the ith dimension and $\times$ denotes the cross product. This method requires no modification to the original network and is simple to implement in that the integral is approximated as the sum of the gradients at points occurring at sufficiently small intervals along the path between x' and x [8]. The implementation in this paper uses 50 steps to approximate the integral given in Eq. (3).

C. Guided Backprop

Guided Backprop builds on prior work in back-propagation [11] and "Deconvnet" [12]. Guided Backprop and back-propagation calculate attribution scores layer by layer starting at the output and working back through the layers of a neural network. In contrast to back-propagation, Guided Backprop incorporates "Deconvnet" into its calculation at specific layers. Guided Backprop is defined in Eq. (4) and Eq. (5) [9].

$$R_i^l(x) = (f_i^l(x) > 0) \cdot (R_i^{l+1}(x) > 0) \cdot R_i^{l+1}(x) \tag{4}$$

$$R_i^{l+1}(x) = \frac{\partial f^{out}(x)}{\partial f_i^{l+1}(x)} \tag{5}$$

These equations define how the attribution scores are acquired where R represents the saliency map, x is the input image, l denotes the layer of the neural network being calculated and i is the individual neuron in that layer. In both equations f is the activation score of each neuron, which is stored during the forward pass of the image x through the network. Also, f^{out} is the activation score at the output of the network and is equivalent to S_c as defined in the SmoothGrad method (see Eq. (1)).

Equation (4) describes the guided back-propagation for calculating the saliency map at layer l. The map is obtained by multiplying the saliency map at layer $l+1$ by f at layer l (with all negative values replaced by zero) and R values at layer $l+1$ (with all negative values replaced by zero). Equation (5) is the forward pass that finds the saliency map at layer $l+1$ as the partial derivative of the output layer neuron activation with respect to the neuron activation at layer $l+1$ for each pixel of the image x. Note that this equation is analogous to the "Vanilla Gradient" from Eq. (1), except that this approach takes a layer by layer approach. The method allows for the approximation of the gradient at any layer.

D. SmoothGrad For Noise Tunnelling

This method of combining SmoothGrad with other methods (often referred to as noise tunneling) provides an average of multiple saliency maps generated using the same mapping algorithm over several noisy inputs [7]. It follows the same basic premise as Eq. (2), except that M_c is replaced with another saliency map. Equations (6) and (7) show how this is done for Integrated Gradients and Guided Backprop respectively. These two combination methods are called Smoothed Integrated Gradients and Smoothed Guided Backprop respectively. Each of these methods simply takes an average of the attribution scores from n noisy samples. An issue with computation time arises due to the need for multiple passes through the neural network and is a relatively intensive approach like Integrated Gradients.

$$\hat{\text{IntegratedGrads}}_i(x) = \frac{1}{n}\sum_1^n \text{IntegratedGrads}_i(x + N(0, \sigma^2)) \tag{6}$$

$$\hat{R}_i^l(x) = \frac{1}{n}\sum_1^n R_i^l(x + N(0, \sigma^2)) \tag{7}$$

E. Combination Methods Based on Gradient Replacement

For the combinational methods proposed in this paper, the idea is to take the integral of any saliency mapping method to be combined. This is equivalent to combining Integrated Gradients with SmoothGrad and Guided Backprop as described by Eqs. (8) and (9) respectively. Note that Eq. (8) differs from Eq. (3) in that the former uses the smooth gradient which is not used by the latter. These two methods are called Integrated SmoothGrad and Integrated Guided Backprop respectively. The approaches have the same issue as those in Section II-D in that multiple passes through the network are needed. However, replacing the gradient in Integrated Gradients can improve the results obtained since there is an error introduced by every gradient approximation. The motivation is that there is less error introduced by gradient replacement resulting in more accurate saliency maps.

$$\text{IntSmoothGrads}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \hat{M}_c(x' + \alpha \times (x - x'))d\alpha \tag{8}$$

$$\text{IntegratedGBP}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} R_i^l(x' + \alpha \times (x - x'))d\alpha \tag{9}$$

III. Experimental Protocol

Assessing the accuracy of interpretability techniques is an ongoing challenge largely due to the fact that there is no ground truth for saliency maps to compare against. In order to provide a broad evaluation, both qualitative and quantitative assessments are performed. Qualitative evaluation is important since interpretability is largely about expressing the results of a complex learning model to a human in a form one can understand. Quantitative assessment is also significant as one must assess whether the saliency maps are accurate in depicting which pixels are relatively more important than others. Both assessments use images from the MNIST handwritten digit database [13], CIFAR10 [14] and Tiny ImageNet [15].

In papers that propose interpretability techniques, qualitative evaluation has been the primary tool for assessing saliency maps [7][8][9][10][16][17]. In this paper, the qualitative evaluation focuses on two aspects, namely, visual coherence and edge differentiation. Visual coherence assesses how accurately the map highlights the object of interest and how much spillover there is into the background. Edge differentiation

MNIST Model Layers

Layer Name	Output Size (Input 28x28x1)	Model
conv1	28x28x32	32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1)
SELU	28x28x32	
max pool2d	14x14x32	kernel_size=2, stride=2, padding=0, dilation=1
conv2	14x14x64	64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1)
SELU	14x14x64	
max pool2d_2	7x7x64	kernel_size=2, stride=2, padding=0, dilation=1
flatten	3136	reshape(x.size(0), -1)
linear1	128	
SELU	128	
dropout	128	p = 0.5
linear2	10	
softmax	10	

Fig. 1. Model architecture for training and evaluation on the MNIST dataset.

assesses how visibly contrasted the edges of the object are in the saliency map. This is valuable since some objects may not be highlighted by the saliency map but may be recognizable if the edges of the object are still clearly defined.

Quantitative evaluation provides an objective measure of how much the highlighted pixels contribute to the final classification score of the neural network. There currently exist multiple quantitative evaluation techniques which seek to evaluate the accuracy of saliency maps [3][18][19][20]. There is currently no consensus on which quantitative evaluation technique is the most accurate. In this paper, the evaluation uses RemOve And Retrain (ROAR) [18].

The class activation scores of the output layer of the image classifier is used to assess the saliency maps. The model with architecture given in Fig. 1 is used for MNIST. The ResNet18 architecture [21] is used for CIFAR10 and Tiny ImageNet. The output layer of the neural network outputs a classification score or softmax output. The label with the highest associated class activation score is chosen as the predicted output label of each classifier. The score depicts the confidence that the classifier has in each prediction. This allows one to assess whether removing the pixels which are important (according to the saliency map) brings about a change in class activation scores. The setup replaces each pixel with the image mean. This degraded image results in a drop in the scores. Similarly, a drop in scores results when the image is degraded based on replacing each pixel by a random value. The removed information in the degraded images are not accounted for in the training of the model. The class activation scores of the degraded image are not accurate. Hence, the model is retrained according to the ROAR method.

The model is trained on ten different sets of data, each with varying levels of noise described by a parameter t starting at 0.0 (no pixels removed) to $t = 0.9$ (90% of the pixels removed). The models trained for the lower noise levels of $t = 0.0$ to $t = 0.4$ had a validation set accuracy of above 98% for the MNIST dataset. The validation set accuracy decreased to 97% ($t = 0.5$), 96% ($t = 0.6$), and 95% ($t = 0.7$). A larger drop of 91% occurred at $t = 0.8$ and a significantly diminished accuracy of 74% was achieved at $t = 0.9$.

The removed pixels were chosen at random, meaning that the retrained model scores should be minimally affected by the removal of random pixels, but should still be affected by the selective removal of significant pixels. For MNIST, retraining of the model produced a class activation score of 1.0 for degradation levels from $t = 0.1$ to $t = 0.8$. For a degradation level of $t = 0.9$, the model has less confidence, mainly due to the lack of information provided by the high level of degradation of the images during both training and evaluation. When evaluating the saliency maps, one must take into account these scores so that one does not attribute score changes which occur during random degradation to the degradation determined by the map. The same phenomenon is observed for the CIFAR10 and Tiny ImageNet data.

	MNIST	CFAR10	Tiny ImageNet
Vanilla Gradient	Very Low	Medium	Very Low
Grad $\times$ Input	Very High	Very Low	Low
SmoothGrad	Medium	High	Medium
Integrated Gradients	Very Low	Very Low	Medium
Guided Backprop	Low	Low	High
Smoothed Integrated Gradients	High	Low	Medium
Smoothed Guided Backprop	Very Low	Medium	Very High
Integrated SmoothGrad*	Medium	Low	Medium
Integrated Guided Baxkprop*	Low	Very Low	Medium

TABLE I
VISUAL COHERENCE OF SALIENCY MAPPING METHODS ON ALL THREE DATASETS. EACH METHOD IS EVALUATED ON A SCALE FROM VERY LOW TO VERY HIGH. THE $\star$ INDICATES THE NEW PROPOSED METHODS.

IV. RESULTS AND DISCUSSION

Results are presented for the (1) Vanilla Gradient (defined in Eq. (1)), (2) Grad x Input, (3) the benchmark methods and combination techniques and (4) the new proposed combination methods Integrated SmoothGrad and Integrated Guided Backprop. The Grad x input approach is a commonly used technique which is an elementwise product of the Vanilla Gradient and the input image [22].

A. Qualitative Evaluation

Qualitative evaluation largely focuses on the human interpretability aspect of the attribution map as exemplified by visual coherence and edge differentiation. A "good" saliency map is depicted by both these aspects. A strong visual coherence is achieved when the object of interest is clearly differentiated and recognizable to a human and that the important pixels are concentrated within that object with sharp edges and little to no spillover into the background. A strong edge differentiation is achieved if the observer can easily differentiate the object of interest from the rest of the image.

A visual coherence evaluation of each saliency method on all three datasets can be found in Table I. The visual coherence of each method varies from dataset to dataset. This is likely due to differences in the dataset images and differences in the model architecture. Overall, Smoothgrad performed the best on average over the three datasets.

The evaluation of edge differentiation is depicted in Table II. There is less variation within the results of this evaluation. The largest differences are between MNIST and the other two datasets. This is likely due to the different model architecture used on MNIST. The SmoothGrad, Smoothed Integrated Gradients and Smoothed Guided Backprop approaches performed the best on average.

	MNIST	CFAR10	Tiny ImageNet
Vanilla Gradient	Very Low	Low	Very Low
Grad $\times$ Input	Very High	Low	Low
SmoothGrad	Medium	Medium	High
Integrated Gradients	Medium	Low	Low
Guided Backprop	Very Low	Medium	Very High
Smoothed Integrated Gradients	High	High	Medium
Smoothed Guided Backprop	Low	High	Very High
Integrated SmoothGrad*	Medium	Low	Medium
Integrated Guided Baxkprop*	Low	Very Low	Low

TABLE II
EDGE DIFFERENTIATION OF SALIENCY MAPPING METHODS ON ALL THREE DATASETS. EACH METHOD IS EVALUATED ON A SCALE FROM VERY LOW TO VERY HIGH. THE $\star$ INDICATES THE NEW PROPOSED METHODS.

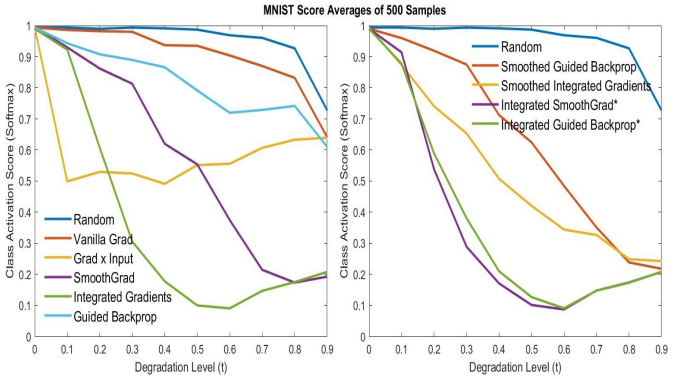

Fig. 2. Average of the softmax class activation scores for the ground truth class label versus image degradation level for 500 MNIST images. Individual (left) and combinatorial (right) methods with a random map as a baseline. The $\star$ indicates the new proposed methods.

B. Quantitative Evaluation

For quantitative evaluations of saliency maps, one must define what different results mean on a retrained model. In an ideal scenario, a saliency map perfectly ranks each pixel according to its relative contribution to the class activation score and therefore, every degraded image would ideally get a class activation score of 0.0. This is the case even though the images with larger amounts of degradation have imperfect scores. The worst case scenario for saliency map accuracy would be a completely random map which should give approximately a class activation score of 1.0 on the MNIST dataset. Due to the decrease in activation scores at the higher levels of degradation for a random map, the higher levels of degradation will provide relatively less accurate results compared to the lower levels of degradation. A random map score in each of the graphs in Figs. 2, 3 and 4 can be used as a reference against the scores for each saliency mapping method. In obtaining the results, the model is retrained for each level of degradation. A highly accurate saliency map should continually decrease class activation scores as the degradation level increases. As the random map scores decrease, there is more error introduced into the evaluation at that level of degradation. The maps which decrease class activation scores the fastest and to the largest extent are considered the most accurate according to the proposed assessment.

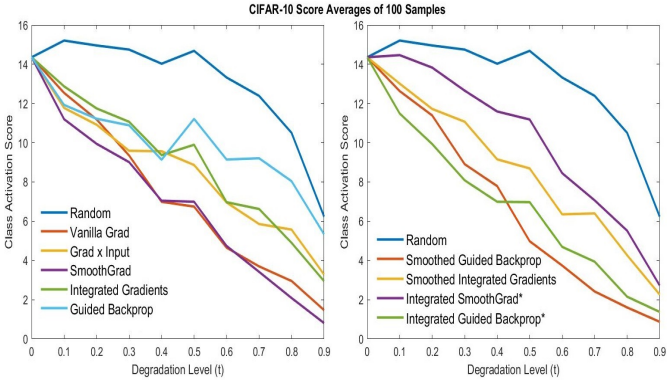

Fig. 3. Average of ResNet18 class activation scores for the ground truth class label versus image degradation level for 100 CIFAR10 images. Individual (left) and combinatorial (right) methods with a random map as a baseline. The $\star$ indicates the new proposed methods.

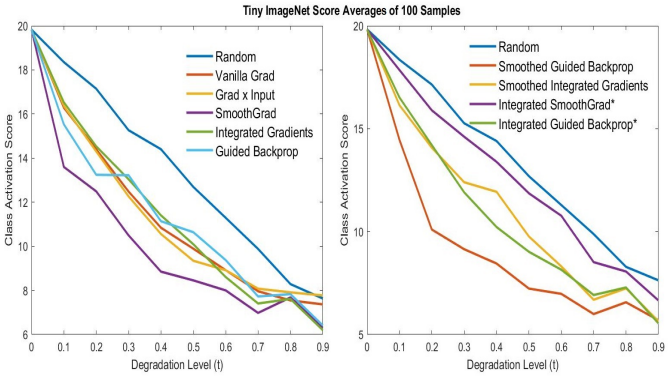

Fig. 4. Average of ResNet18 class activation scores for the ground truth class label versus image degradation level for 100 Tiny ImageNet images. Individual (left) and combinatorial (right) saliency mapping methods with a random map as a baseline. The $\star$ indicates the new proposed methods.

Based on the results from Figs. 2, 3 and 4, the methods which perform the best on average for all three datasets (for the ROAR evaluation) are SmoothGrad, Smoothed Guided Backprop and Integrated Guided Backprop.

V. SUMMARY AND CONCLUSIONS

There is minimal correlation between the qualitative and quantitative assessments of each saliency mapping method. There is also some variation among these methods with regard to how well they perform on each dataset. The SmoothGrad and Smoothed Guided Backprop approaches performed well on both the qualitative and quantitative evaluations. The Smoothed Integrated Gradients and Integrated Guided Backprop performed well on the qualitative and quantitative assessments respectively. The evaluation of these recent saliency mapping techniques and their combinations provides a reference for further research into new applications.

VI. ACKNOWLEDGEMENT

This work was supported by the U.S. Department of Education Graduate Assistance in Areas of National Need (GAANN) Grant Number P200A180055.

VII. References

1) P. W. Koh and P. Liang, "Understanding blackbox predictions via influence functions", *International Conference on Machine Learning*, Vol. 70. 2017.
2) M. C Hughes, H. M. Elibol, T. McCoy, R. Perlis and F. Doshi-Velez, "Supervised topic models for clinical interpretability", arXiv:1612.01678, 2016.
3) B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)", arXiv:1711.11279, 2017.
4) S. Pereira, R. Meier, R. McKinley, R. Wiest, V. Alves, C. A Silva and M. Reyes, "Enhancing interpretability of automatically extracted machine learning features: Application to a rbm-random forest system on brain lesion segmentation", *Medical Image Analysis*, 2018.
5) A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?", arxiv:2004.03685v3, 2020.
6) B. Shickel1 and P. Rashidi, "Sequential Interpretability: Methods, Applications, and Future Direction for Understanding Deep Learning Models in the Context of Sequential Data". arxiv:2004.12524, 2020.
7) D. Smilkov, N. Thorat, B. Kim, F. Viegas and M. Wattenberg, "Smoothgrad: Removing noise by adding noise", arXiv:1706.03825, 2017.
8) M. Sundararajan, A. Taly and Q. Yan, "Axiomatic attribution for deep networks", *International Conference on Machine Learning*, August 2017.
9) J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for simplicity: The all convolutional net", arXiv:1412.6806, 2014.
10) K. Simonyan, A. Vedaldi and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps", arXiv:1312.6034, 2013.
11) D. Erhan, Y. Bengio, A. Courville and P. Vincent, "Visualizing higher-layer features of a deep network", University of Montreal, vol. 1341, no. 3, 2009.
12) M. D. Zeiler, G. W. Taylor and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning", *IEEE International Conference on Computer Vision*, 2011.
13) Y. LeCun, C. Cortes and C. J. C. Burges, *MNIST handwritten digit database*, http://yann.lecun.com/exdb/mnist/#:~:text=THE%20MNIST%20DATABASE%20%20%20CLASSIFIER%20%20,al.%20CVPR%202012%20%2021%20more%20rows%20
14) A. Krizhevsky, "Learning multiple layers of features from tiny images", https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf, 2009.
15) Y. Le and X. Yang, "Tiny imagenet visual recognition challenge", https://tiny-imagenet.herokuapp.com/, 2015.
16) M. D Zeiler and R. Fergus, "Visualizing and understanding convolutional networks", *European conference on computer vision*, Springer, 2014.
17) R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh and D. Batra, "Grad-cam: Why did you say that?", arXiv:1611.07450, 2016.
18) S. Hooker, D. Erhan, P.-J. Kindermans and B.n Kim, "A benchmark for interpretability methods in deep neural networks", in *Advances in Neural Information Processing Systems*, 2019.
19) J. Zhang, Z. Lin, J. Brandt, X. Shen and S. Sclaroff, "Top-down neural attention by excitation backprop", European Conference on Computer Vision (ECCV), 2016.
20) M. Yang and B. Kim, "Bim: Towards quantitative evaluation of interpretability methods with ground truth", arXiv:1907.09701, 2019.
21) K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", 2015.
22) A. Shrikumar, P. Greenside and A. Kundaje, "Learning important features through propagating activation differences", *International Conference on Machine Learning*, 2017.