Discrimination Assessment for Saliency Maps

Ruiyi Li[1,2], Yangzhou Du[2], Zhongchao Shi[2], Yang Zhang[2], and Zhiqiang He[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences
liruiyi18g@ict.ac.cn
[2] Lenovo Group, Beijing 100094, China
{liry8,duyz1,shizc2,zhangyang20,hezq}@lenovo.com

Abstract. Saliency methods can effectively mark which patterns in the input have higher impacts in model decision, and highlight the relationship between the features and inference results. However, different saliency maps have different performance in classification tasks. Through experiments, we find that some saliency maps show more discriminative ability, while others do not. Saliency methods with higher discrimination ability will be more helpful to human while making final decision, for example, the dominant features of malignant area are expected to be identified in medical diagnosis. In this work, a method is proposed to evaluate whether the saliency methods can provide effective discriminant information. In addition to giving intuitive judgment, we will also introduce a quantitative measurement method. We regard the saliency map as a weighting vector in class discrimination, which is analogue to the projection direction of Linear Discriminant Analysis (LDA) [12], and measure the discriminant ability of saliency map by comparing the difference between the vector direction of saliency map and the projection direction of LDA. Through this metric, our experiments will present the ranking of popular saliency map methods, in terms of discriminative ability.

Keywords: Saliency Methods · Discriminant Analysis · Interpretability.

1 Introduction

With the development of machine learning, the neural network model has shown unparalleled performance in more and more tasks, and even reached and exceeded the human level in some fields. However, compared to classical machine learning methods, such as linear model, decision tree, support vector machine, it is difficult to understand how neural networks make decisions. In mission-critical tasks, if the inference result generated by the algorithm cannot be explained, people will have a big concern to use it. Therefore, in mission-critical scenarios, such as autonomous driving and medical diagnosis, the application of neural networks is strictly limited.

In the existing work, a large number of researchers have gradually begun to pay attention to the interpretability of models. People often call algorithmically transparent models interpretable, while opaque models are considered as "black

boxes". What is transparency? It is difficult to assess comprehensively the interpretability of a model by the definition of each parameter in the model or the complexity of the model.

In Lipton's article [5], interpretability is the way to trust models. The interpretable definition of the model can be divided into the following aspects. The first is simulation, that is, whether human can manually simulate machine learning model. The second is decomposability, that is, whether each part has an intuitive explanation. The third is security, that is, whether the model can prevent deception and whether decision-making does not depend on the wrong information. We believe that the saliency map can get the decision basis of the model, that is, the ability to explain whether the model is safe or not is the most important.

In the neural network model, we can see that many researchers successfully deceive the model by adding noise, which leads to the wrong results of the model. Just as X. Yuan et al. did in the adversarial examples [13]. Although the accuracy of models and other indicators sometimes exceed human standards, it is still difficult for people to trust the results of models, especially in some areas related to human security. This is also one reason why some people call the neural network model as alchemy.

It is hoped that an explanation-based tool can be used to clarify the basis for the model to make decisions and to help model designers to improve their models, eliminating biased information and other unintended effects learned by the model. In tasks such as image recognition, the saliency method becomes a popular tool that highlights the feature information of the input that is highly correlated with the model and obtains the underlying semantic patterns within the model.

However, because interpretability itself is difficult to define, there are few suitable indicators to evaluate these saliency methods. Sanity Checks [1] argues that some explanatory methods are so biased towards human intuition that they are misguided that the results are independent of data and machine learning models. In this work, we analyze it from another angle. We focus on a class of discriminant saliency methods. How do we distinguish two or more different classes in our daily life? If professionals tell you all the characteristics of each class, it may be difficult for laymen to remember and understand. However, just tell you the main difference between the two classes, it's easy to get an intuitive impression and trust that explanation.

Therefore, we hope that we can evaluate whether saliency maps contain enough discriminant information to make it easy for people to determine whether a neural network model is trustworthy. The most intuitive understanding of discriminant information is, how do we decide that a number is 1 instead of 7? It's the short horizontal line in the top area of the digit. However, different neural network models have different decision-making bases, and saliency maps are not necessarily intuitive. We want to know a way to quantify this discriminant information, so we measure quantitatively by comparing it with LDA.

In the projection direction of LDA, the inter-class distance can be maximized and the intra-class divergence can be minimized, and the optimal projection vectors for distinguishing different classes can be obtained. We use saliency maps generated by interpretive methods as projection vectors to distinguish categories, at the same time, the projection vector of LDA is used as ground-truth to compare the discriminating ability of different saliency maps. However, such methods have certain limitations, that is, we assume that the samples are largely linearly separable.

At the same time, we visualize the samples, saliency maps and LDA projection vectors, and more intuitively observe the difference of discriminant information produced by different saliency methods.

Our contributions

1. We figure out different saliency map methods have different performance in terms of discrimination. In discriminant tasks, in order to help users better, we should choose a saliency method with strong discriminant ability.
2. In order to quantify the discrimination ability of saliency map, we consider it as a projection vector to distinguish different categories, analogous to the projection vector of LDA. At the same time, we use the projection direction of LDA as the ground-truth, and measure the difference between the two vector directions to obtain the discriminant ability of saliency map.
3. Through our metric, We find that some saliency maps can only provide little discriminant information, while others have stronger discriminant ability. The experimental results on data sets such as MNIST [4] show that Grad-CAM [11] is superior to other saliency map methods.
4. At the same time, the experimental results in different neural network structures show unique explanatory information. AlexNet can be based on more different information when making decisions.

2 Related Work

In the research of the interpretability of the neural network model, a large number of researchers have done excellent work, hoping to open the black box of the neural network, and promote the progress of human society with the powerful computational performance and predictive ability of the model.

2.1 Saliency Method

In the neural network model, we put the data into the model to get the results, how to know which information is the most important in the input data, and has the greatest impact on the judgment of the model? The saliency map method is to find an effective way of this part of the information. Among the saliency map methods, some directly start from the data and use different input data combination methods to determine which data is sensitive to the model. SHAP [6] uses

the direct method to find the most important information in the input, which originates from the classical game theory method. LIME [8] pays attention to the explanation of local models. It believes that no matter how complex the models are, there is always a linear function that can approximate the model in a certain local area. LIME seeks for a reasonable linear explanation of this part.

Others use the flow of data in the model to determine which inputs the model is sensitive to. LRP [10] pushes the results back to the input data layer by layer, so as to get an effective saliency map. CAM [15] and GradCAM [11] believe that the parameters of the neural network model contain explanatory intuitive information, so the feature of the neural network model is combined with the parameters to get the interpretation of the model in judgment.

2.2 Visualization of Features

Other methods do not focus on explaining a single input sample, but rather on seeing intuitively what patterns the model learns and stores for judgment. Interpreting CNN Knowledge Via An Explanatory Graph [14] wants to know what models each neuron in the model represents.

Some cognitive neurological experiments have now demonstrated that artificially constructed neural network models are very similar to real animal brains, and that images can be used to stimulate specific groups of neurons in James' experiments [2]. And in another highly enlightening experiment, Ponce et al.[7] constructed a framework that combines depth-generated neural networks and genetic algorithms to synthesize images that maximize the activation of animal neurons, in which animal memories can even be seen.

2.3 Evaluation Method

There are now some models for evaluating saliency map methods. AOPC [9] examines the impact on the results by evaluating whether the weight ranking of all points in the saliency map is correct and eliminating the highest weight points in turn. Sanity Checks [1] evaluates whether interpretation is independent of data and model. Some saliency map methods are not sensitive to data, some are similar to edge detection, but they are not helpful to interpret models.

3 Methodology

Some existing saliency map methods can already provide some intuitive information: the neural network model depends on which part of the input to judge. However, such explanatory information can sometimes be very vague, and it is difficult to present enough information in some categories and similar tasks. For example, the difference between plastic buckets and bags in automatic driving tasks can lead to serious consequences. So clear information is needed to distinguish between these two categories, especially when such differences can have serious consequences.

Therefore, it is very important to measure whether a saliency assessment method can provide enough discriminatory information. How to determine the difference between the two categories? LDA can provide a projection vector in the binary classification problem. This projection vector can point out which input information is more helpful to distinguish the two categories, which provides us with a simple ground-truth.

Mapping the input information to a vector that maximizes the distance between classes and minimize divergence within class can give us enough discriminant information, especially when two simple classes are discriminated. Then the projection vector of LDA is taken as the basic method of measurement.

To formalize the problem, the *input* information is a vector $x \in \mathbb{R}^d$, a *classification model* describes a function $S : \mathbb{R}^d \to \mathbb{R}^C$, where C is the number of classes in the classification work. An explanation method that can generate saliency maps is described as a function $Out_{maps} : E(S(x), x)$, where Out_{maps} is the saliency maps showing weights of input information and E is the explanation method.

Next, the projection vector of LDA is taken as ground-truth, and the discriminant information of the explanation method is measured by Pearson correlation coefficient [3] and Cosine similarity.

In order to better describe discriminant information in saliency maps, all samples need to be counted and analyzed. At the same time, in order to avoid training errors, all the pre-processing and model training parameters are unified to obtain more accurate discriminant assessment.

4 Experiments

In the experiment, the saliency map methods and the neural network model are evaluated respectively. In the first part, we mainly show whether different saliency map methods can mine discriminant information in the model. Although some saliency map methods can clearly mark out which information in the input plays a decisive role in the model determination, it is difficult to distinguish such information from other categories of saliency maps. We may know that this part is really useful, but if this part of the information of all samples is useful, it is difficult to obtain enlightening knowledge.

In the second part, we mainly evaluate whether different models can provide enough discriminant information. When carrying out discriminatory tasks, we can choose a model with more discriminatory information or construct a Neural Network with sufficient discriminatory information to complete the corresponding tasks.

In order to observe the information of saliency maps more intuitively, we only focus on the impact of the information on the output, regardless of whether it is positive or negative. Therefore, it is a very straightforward and simple way to take absolute values of all the weight information in the saliency map in the experiment. At the same time, all figures are displayed in pseudo-color, with blue as the minimum and yellow as the maximum.

4.1 Saliency Method Assessment

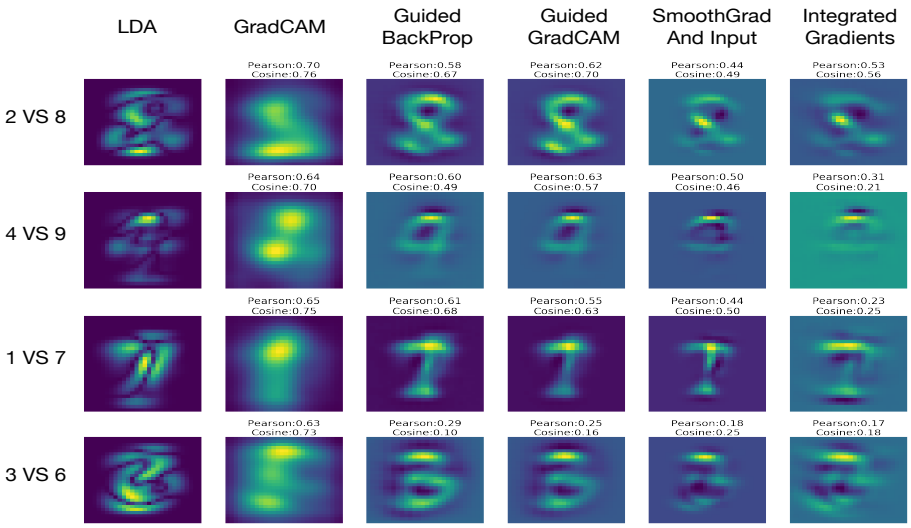

Fig. 1. *Evaluation of different saliency map methods.* Four pairs of digits were used to compare the results. GradCAM performed well in all tests, but the other saliency map methods were not effective. Especially when the category difference is small, such as 3 and 6, it is difficult to provide enough discriminant information.

Different saliency map methods provide completely different saliency information, among which GradCAM can provide more discriminant information. At the same time, the other saliency map methods focus on the generality of categories, and it is difficult to reflect the difference information between categories, so they have a low score in the evaluation. When the difference between categories is small, GradCAM can provide the discriminant information steadily, while the other methods are difficult to distinguish the difference. Category information is displayed in Fig.1.

The Fig.2 shows the statistical results of all samples, the yellow line is the median and the green triangle is the mean of the data. It can be seen that there is a big difference between GradCAM method and other saliency map methods, GradCAM displays more discriminatory information, while Integrated Gradients was the worst performers.

4.2 Neural NetWork Assessment

Four different Neural Network models, AlexNet, ResNet, Vgg16 and DenseNet, are used in the experiment. We hope to find out whether there are enough discriminant information in different network models.

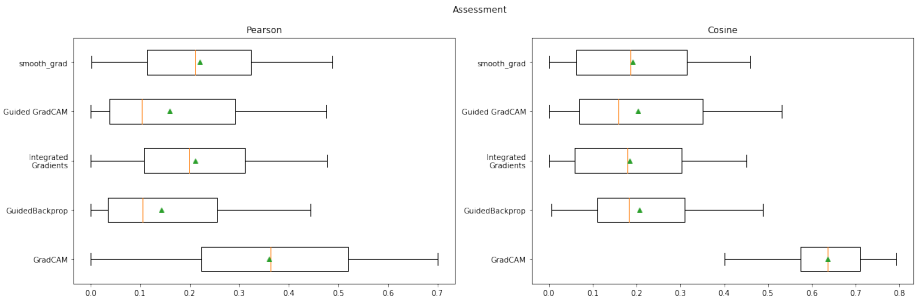

Fig. 2. *The evaluation values of different saliency map methods.* Each saliency map in the sample is evaluated separately, and the results are taken as absolute values to get the above illustrations. The rectangle in the box-plot marks the range of the first quartile and the third quartile, and the line segment in the rectangle is the median of the data.

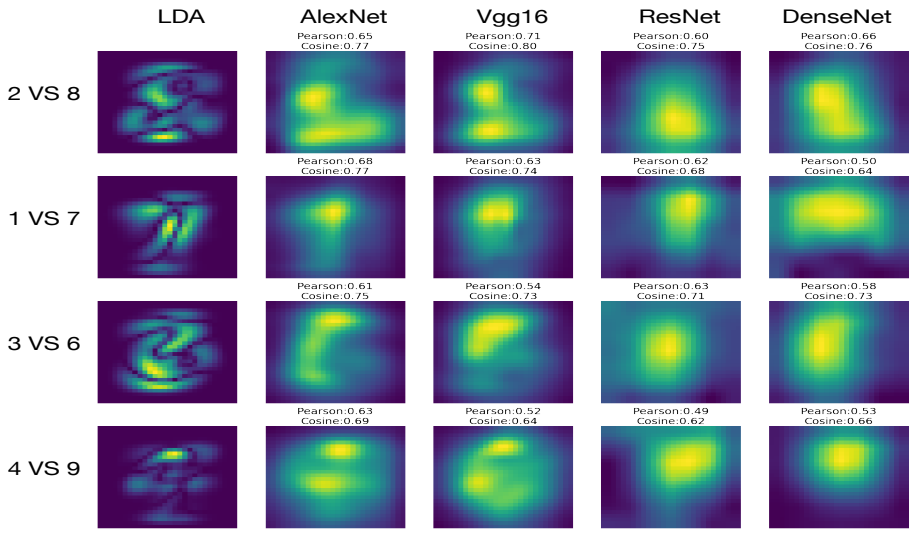

Fig. 3. *Evaluate different neural network models.* Four different neural network models were compared in the experiment, and the best gradCAM method in 4.1 assessments was used to explain the model. The same four different categories were used for the experiment.

In the Fig.3, it is showed that different network models show completely different discriminant information although they have similar measurements. AlexNet and Vgg16 are similar, while ResNet and DenseNet fully show the interpretation information of another style. Although their scores are similar, they make decisions based on different regions.

5 Conclusions And Futrue Work

In this paper, we have quantitatively evaluated the discriminant information provided by different saliency map methods. We can see that some saliency map methods can provide some more intuitive explanations, but such explanations lack sufficient discriminant information and it is hard to believe that the models have found the differences between different categories.

At the same time, experiments on various neural network models show that they predict based on different information. We have reason to believe that such information reflects the difference of model structure to some extent. (The exact same parameters and data were used in the experiment, the only difference is that the model structure is different.)

However, due to the difficulty of multi-classification tasks and the huge amount of computation, only a part of the MNIST data set is used in this experiment. In future experiments, we need to validate our ideas in more data sets and large tasks. We hope to find a saliency method, which can provide clear discriminant information and improve the interpretation ability of the model. Let the excellent performance of the neural network model be applied to more valuable tasks.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 9505–9515. Curran Associates, Inc. (2018), http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf
2. Bashivan, P., Kar, K., DiCarlo, J.J.: Neural population control via deep image synthesis. Science **364**(6439), eaav9436 (2019)
3. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
4. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist **2**, 18 (2010)
5. Lipton, Z.C.: The mythos of model interpretability **61**(10) (2016)
6. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

7. Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., Livingstone, M.S.: Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell **177**(4), 999–1009 (2019)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier (2016)
9. Samek, W., Binder, A., Montavon, G., Bach, S., Mller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks & Learning Systems **28**(11), 2660–2673 (2016)
10. Samek, W., Wiegand, T., Mller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models (2017)
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization (2016)
12. Webb, A.R.: Linear Discriminant Analysis (2003)
13. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems (2019)
14. Zhang, Q., Cao, R., Feng, S., Ying, N.W., Zhu, S.C.: Interpreting cnn knowledge via an explanatory graph (2017)
15. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Computer Vision and Pattern Recognition (2016)