axioms

MDPI

Article

Interpretable Model-Agnostic Explanations Based on Feature Relationships for High-Performance Computing

Zhouyuan Chen, Zhichao Lian * and Zhe Xu

School of Cyberspace Security, Nanjing University of Science and Technology, Nanjing 214400, China; gaobanzou@njust.edu.cn (Z.C.); 123127211612@njust.edu.cn (Z.X.)
* Correspondence: newlzcts@njust.edu.cn

Abstract: In the explainable artificial intelligence (XAI) field, an algorithm or a tool can help people understand how a model makes a decision. And this can help to select important features to reduce computational costs to realize high-performance computing. But existing methods are usually used to visualize important features or highlight active neurons, and few of them show the importance of relationships between features. In recent years, some methods based on a white-box approach have taken relationships between features into account, but most of them can only work on some specific models. Although methods based on a black-box approach can solve the above problems, most of them can only be applied to tabular data or text data instead of image data. To solve these problems, we propose a local interpretable model-agnostic explanation approach based on feature relationships. This approach combines the relationships between features into the interpretation process and then visualizes the interpretation results. Finally, this paper conducts a lot of experiments to evaluate the correctness of relationships between features and evaluates this XAI method in terms of accuracy, fidelity, and consistency.

Keywords: interpretability; model-agnostic explanations; feature relationship; super pixel

MSC: 68T07

Citation: Chen, Z.; Lian, Z.; Xu, Z. Interpretable Model-Agnostic Explanations Based on Feature Relationships for High-Performance Computing. *Axioms* **2023**, *12*, 997. https://doi.org/10.3390/axioms12100997

Academic Editor: Oscar Humberto Ross

Received: 11 July 2023
Revised: 26 September 2023
Accepted: 27 September 2023
Published: 23 October 2023

1. Introduction

In recent years, deep learning has developed rapidly in image processing [1,2], natural language processing [3,4], speech recognition [5,6] and other related fields and has shown to surpass human capacity in all walks of life, which makes people increasingly reliant on decisions made by AI. To meet the increasing requirements of people in healthcare [7,8] and the precision industry, the complexity of the model has also increased significantly. But the more complex the model is, the more difficult it is for people to understand its structure and the more difficult it is to explain why it makes this decision, and this creates a problem: people begin to distrust the decisions made by the model. To solve the above problems, XAI has become a hot field. And through methods in this field, people can think more about the reasons why models have this effect. Such thinking is conducive to preserving important features in images, and people can use these important features to train other models, which can reduce computational cost. This thinking can help better understand the model and improve the service quality of the model.

In the field of image processing, XAI methods can be divided into ante hoc interpretability and post hoc interpretability. Ante hoc usually involves data preprocessing and model selection; the purpose of the former is to show the distribution of features, and the purpose of the latter is to explain the decision-making process by constructing a structurally interpretable model, such as a linear model [9] and a decision tree [10,11]. Post hoc interpretability is to make visible an analysis of the decision-making process of the model or to analyze the importance of features [12–14]. Although the former is relatively simple and the cost is relatively small, the structure of many deep learning models is very

complex and generally unknown to users, so post hoc interpretability has more advantages in the field of deep learning.

Post hoc interpretability also contains two categories: global and local. Global interpretation focuses on the operating principle of the model when it works, including activation maximization [15,16] and knowledge distillation [17,18], while the focus of local interpretation is on the impact of the sample itself when the model makes a decision, including at the pixel level [19,20], concept level [21,22] and picture level [23,24]. Although the former can effectively improve the transparency of the model, the latter can often be more easily understood by users and is easier to mine because of the direct relationship between features.

In the field of XAI, especially for white-box models, there are already some methods for mining the relationship between features. These methods can partly solve the problem of only considering the importance of features and ignoring the relationship between features [25], and at the same time, these methods' results are not intuitive, and their methods cannot adapt to the research of XAI methods around deep learning models to a certain extent. And for black-box models, relationships between features are effectively mined in the XAI methods for tabular data or text data, but the manner in which to obtain and evaluate relationships cannot be well applied to image data [26,27]. Therefore, the idea of this paper is to combine mining relationships between features with LIME (local interpretable model-agnostic explanations), which is an XAI method based on black-box that obtains explanations by locally approximating simple models, while using LIME [12] to obtain important superpixel blocks as feature blocks, obtain relationships between feature blocks, and optimize the visualization effect of the results, that is, visualize the importance of features and the relationship between features at the same time. The contributions of this paper can be summarized as follows:

(1) We analyze the shortcomings of existing XAI methods for obtaining features' relationships. And then, to solve these problems, we propose an interpretation method based on masking to consider relationships between features in the process of interpreting a mode, which makes the interpretation more complete and improves the credibility of the interpretation.
(2) We perform a lot of experiments in this paper, and the results prove the correctness of relationships between features obtained in this paper and show that our method achieves higher accuracy, fidelity, and consistency compared to LIME.

2. Related Work

The proposed method's characteristics are mainly reflected in two aspects: (1) methods based on black boxes or white boxes; and (2) methods to obtain relationships between features. So, we analyze the existing methods from these two aspects:

In terms of explaining black-box models, J. H. Friedman proposed partial dependence plots (PDPs) in 2001 [28]. This method can show the marginal effect of one or two features on the prediction results of models, that is, the probability of a specific category under different feature values of a feature, thus showing that the relationship between the target and the feature is linear, monotonic, or more complex. In 2016, Marco et al. proposed the LIME method. This method focuses on training a local proxy model to interpret a single prediction. It selects an interesting instance and uses it as the input of the original model, then perturbs this instance to generate a new data set, which is composed of perturbed samples and corresponding predictions of the black-box model. Finally, on this new data set, LIME trains an interpretable model that is weighted by the distance between disturbed instances and interesting instances. At the same time, Marco et al. also proposed the S-LIME method, which uses the hypothesis testing framework based on the central limit theorem to determine the number of disturbance points required to ensure the stability of the result interpretation rather than using just random disturbance. After that, on the basis of LIME, in 2018, Marco and others proposed the model-agnostic method Anchors [29], which is based on finding a minimum subset of features. As long as any instance has this feature

subset, the prediction result of the black box is the same, independent of other features, and this subset can be used as an explanation. It can also be regarded as the anchor of the black-box model to accurately explain the relatively complex local black-box prediction model. The Shapley value can be understood as a method of allocating expenses according to the contribution of players to total expenses. In XAI, features are players, and the model prediction is the total expenses. Because the difference between the prediction and the average prediction can be perfectly distributed between the characteristics through this method, it has become very popular as a way to explain the black-box model prediction. In 2016, Lundberg et al. proposed the SHAP method, which replaced the method of weighting samples according to their proximity to the original instance in LIME with the method of weighting the samples according to the weights obtained by the alliance in the Shapley value estimation. In 2020, Messalas et al. proposed a MASHAP method [30]. It first builds a global proxy model on the interested instance, then transfers the proxy model as an original model to the Tree SHAP method, and then generates an explanation. Because this method simplifies the original model in the SHAP method, it also achieves faster results than SHAP and LIME. However, in the above methods, the relationship between features is not well considered for image data, which also reduces the credibility of the interpretable algorithm. Therefore, in this paper, we consider introducing the method of calculating the relationship between features at the image data level to mine the relationship between features and improve the interpretation reliability.

There are also many studies in XAI on obtaining relationships between features. In terms of white-box models, Wang et al. [31] proposed spatial activation concept vector, which considers the spatial location relationship. Ge et al. [32] demonstrated the relationships between features by extracting important visual concepts related to a specific category and representing the image as a structured visual concept map. They proposed a visual reasoning explanation framework (VRX) that can obtain structural concept graphs similar to that shown Figure 1. And the colors of components' scores from high to low are: blue, green, and pink.

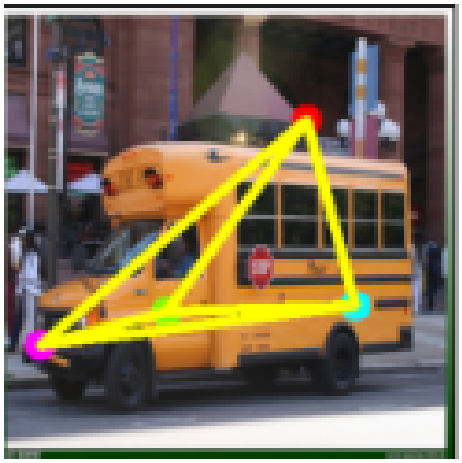

Figure 1. The result of VRX.

However, the available models of these methods are limited to some extent, and they cannot directly show how relationships between features impact the models' predictions, so the degree of visualization is also limited. In terms of black-box models, there are also many characteristic relationship calculation methods for the classic XAI methods based on black-box models. For example, for LIME, Zoumpolia et al. proposed the GLIME method [25], which relies on the combination of LIME and the graphical least absolute shrinkage and selection operator to generate the undirected Gaussian graph model. In addition, regularization reduces the small partial correlation coefficient to zero to provide a more sparse and interpretable graphical interpretation. For the Shapley value, KJERSTI et al. proposed a method [26] that extends the kernel SHAP method to deal

with dependency features. These two kinds of methods can effectively determine the relationship between features through experiments, and KJERSTI proves the correctness of the found relationship through experimental comparison. Although it effectively solved the problem of the universality of the use model and the problem of non-intuitive results, these methods can only be limited to text and tabular data. It has two problems: first, tabular and text data can easily change in numerical value to affect the model prediction, while for the pixel or pixel block in the picture as a feature, it cannot be simply changed to observe the impact on the model prediction. Generally, for the feature of a picture, there are only two possibilities: existence and non-existence. The second is that the features in the tabular and text data can be artificially set so that the features have relationships, such as Gaussian, but the picture is difficult to make, which makes it a big problem to design an indicator to verify the correctness of relationships.

Therefore, in this paper, we draw on some ideas of finding feature association from XAI methods based on black-box models for text and tabular data, combine the idea of masking features with the LIME method, and propose an interpretable method based on black-box that can obtain the relationship between features through the combination masking of feature blocks for image data, which not only makes the relationship between features more intuitive but also improves the universality of the method.

3. Methods

The core idea of this paper is to obtain the direct impact of the relationship between feature blocks on the model's decision by observing the influence of the combined masking of superpixel blocks on the model's output and combining the relationship between feature blocks with the importance of the feature itself to optimize the selection process of important feature blocks to improve the credibility and stability of the explanation. In Section 3.1, the overall architecture of the local interpretable model-agnostic explanation approach based on feature relationships is introduced. Section 3.2 describes the specific implementation of obtaining the relationship between features. Section 3.3 introduces the method of optimizing the selection sequence of important features.

3.1. The Overall Architecture

The structure of this method is shown in Figure 2. After obtaining the segmented superpixel blocks in LIME, namely feature blocks, we calculate the relationship between two superpixel blocks by the combined mask of superpixel blocks and then obtain the feature correlation matrix from these relationships. Then, we obtain the importance of feature blocks from LIME and combine it with the association size between feature blocks to rearrange feature blocks. The feature blocks for interpretation are re-selected, the feature blocks are used as vertices, and the feature association size is used as edge weights for visualization. Compared with the traditional methods based on the black-box approach, which assume that features are independent, this method takes into account the relationship between features, making the interpretation more accurate. Compared with the method of finding feature correlation in the methods based on the white-box method, this method is more versatile because it is based on the black-box method.

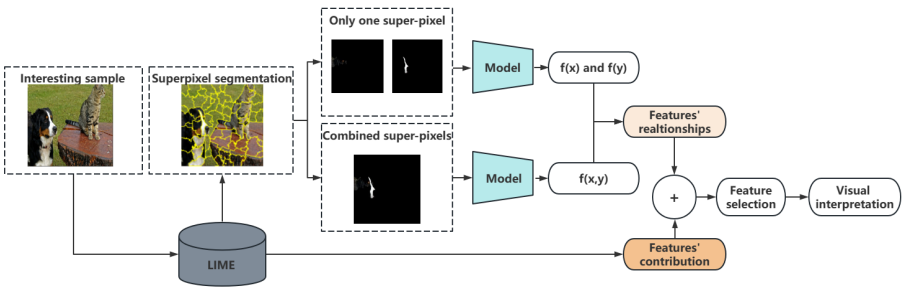

Figure 2. Interpretable model-agnostic explanations based on feature relationships.

3.2. Acquisition of Relationship between Features

The specific steps of LIME are shown in Figure 3.

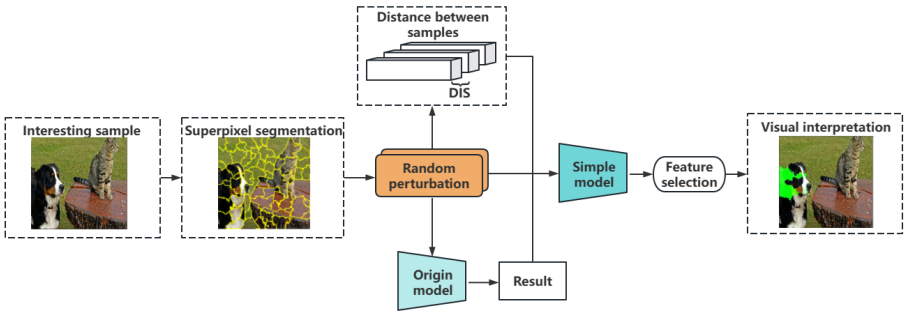

Figure 3. The framework of LIME.

From Figure 3, it is easy to see that LIME is a proxy model method. It will first generate a new data set by perturbing the interested instance and then calculate the distance between them and the interested instance by using a similar distance measurement. Here, the distance can also be converted into similarity, and the results of the original model for the perturbed instance can also be obtained. Finally, a simple interpretable model, such as a linear model, can be trained by using the disturbance data set, distance weight, and results of the original model, which also means that the features will be regarded as independent relationships while ignoring the correlation between features. At the same time, because the internal structure of the model is invisible, it is difficult to find the interaction size of features in the prediction process. Therefore, this paper uses the idea of tabular and text to obtain the correlation between features through control variables. In tabular and text data, data can be perturbed to change the value of features, while in image data, the presence and absence of features can be controlled by occlusion. First, preserve the feature blocks i and j of interest, obtain the prediction results $f_x(i)$ and $f_x(j)$ of the model for the specified class in the case of only feature block i and only feature block j, and then reserve both feature blocks i and j to obtain the prediction results $f_x(i \cup j)$ of the model for the specified class. Since there are no other feature blocks in this method, the impact of the relationship between i and j on the model decision is

$$\sigma_x(i,j) = f_x(i \cup j) - f_x(i) - f_x(j), \tag{1}$$

where $\sigma_x(i,j)$ represents the direct impact of the relationship between feature i and feature j on the model result when the model classification result is x; $f_x(i)$ represents the probability of the result being class x when there is only feature i; and $f_x(j)$ represents the probability of the result being class x when there is only feature j. The specific masking method is shown in Figure 4.

(a) (b) (c)

Figure 4. The method of masking super pixels. (**a**) means preserving the feature blocks i. (**b**) means preserving the feature blocks j. (**c**) means preserving the feature blocks i and j.

Through cyclic calculation, which only requires n^2 time complexity, the direct influence of the pairwise relationship between all feature blocks in the graph on model prediction can be obtained.

3.3. Optimization of Important Features Selection Method

As mentioned in the previous section, LIME will treat each feature as independent, so it is obviously not comprehensive to consider the importance of the feature itself and ignore relationships between features to select important features for interpretation, and the credibility will also be greatly reduced. At the same time, incorporating the relationship between features into the important feature selection process can effectively reduce the impact of the randomness of the LIME algorithm itself, especially when the relationship between features is large. Therefore, the feature importance of this paper is calculated as follows:

$$Con_x(i) = ConSelf_x(i) + \sum_{j=1 \ and \ j!=i}^{k} \frac{\sigma_x(i,j)}{2}, \tag{2}$$

where $Con_x(i)$ represents the contribution of feature i to the model's prediction of the result as class x, that is, the importance of feature i; $ConSelf_x(i)$ represents the importance of feature i obtained by LIME; and $\sigma_x(i,j)$ represents the direct impact of the relationship between feature i and j on the model's prediction result.

4. Results

This paper will select common evaluation indicators in the evaluation of XAI algorithms to evaluate this algorithm, namely, sensitivity [33], fidelity/accuracy [34], and stability/consistency [35], and mainly compare them with classic XAI methods based on black-box models and analyze them based on classic XAI methods based on black-box models. This paper selects InceptionV3 and ResNet50 as deep neural network models for research and conducts experiments on ImageNet data sets.

4.1. Analysis of Results

As shown in Figure 5, setting the number of features required for interpretation K = 3. Compared to LIME, the algorithm in this paper can visualize the relationships between important features. The blue line indicates that the relationships between feature blocks play a direct and positive role in the model prediction results, while the red line indicates that the relationships between feature blocks have a direct and negative impact on the model prediction results; the stronger the effect, the wider the line. Compared to the result in Figure 1 which can only display features that are related, our method has a better visual effect.

Figure 5. Presentation of interpretation results.

In Figure 6, we construct an undirected graph by using superpixels as vertices and the strength of the relationships between superpixels as edge weights, making the relationships between superpixels more intuitive, especially when the superpixels are close.

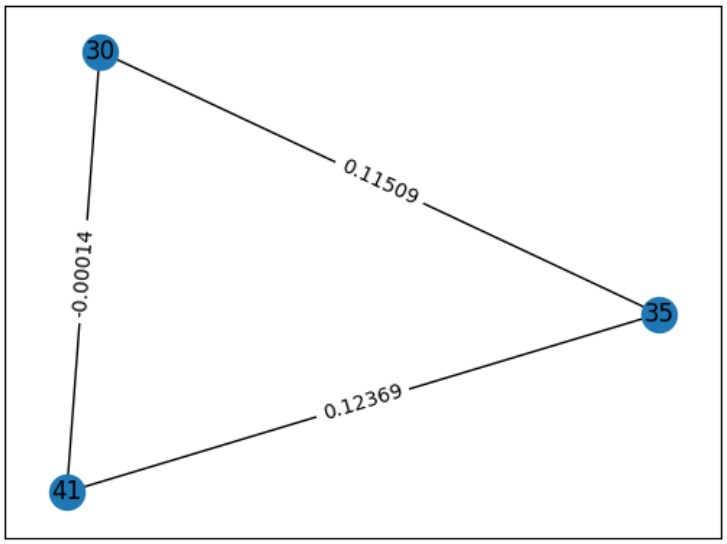

Figure 6. Linear graph of features' relationships.

Combining the features' relationships matrix in Table 1, it can be found that the method in this paper can obtain a relationship size between feature blocks in the image compared to the feature relationships found in the XAI method based on white boxes, and this relationship shows the direct impact of the relationship between two features on the model prediction results rather than simply showing the positional relationship or the degree of correlation between the features. This makes the interpretation more intuitive and more consistent with the general idea of the black-box interpretable method. For example, the relationship between feature 24 and feature 35 increases the probability of the model predicting the result class by 0.32375.

Table 1. Matrix constructed by features' relationships.

	V24	**V30**	**V35**	**V37**	**V41**
V24	/	−0.00048	0.32375	−0.00120	0.00125
V30	−0.00048	/	0.11509	−0.00073	−0.00014
V35	0.32375	0.11509	/	0.04726	0.12369
V37	−0.00120	−0.00073	0.04726	/	−0.00080
V41	0.00125	−0.00014	0.12369	−0.00080	/

4.2. Fidelity/Accuracy Analysis

Currently, the methods used to prove the fidelity/accuracy of XAI algorithms are mainly implemented based on the idea of perturbation. Therefore, the evaluation of accuracy in this article is based on the idea derived from the SHAP method: subtracting the main effect of the feature from the total effect to obtain the pure interaction effect to obtain the feature association size. This uses the following formula:

$$\sigma'_x(i,j) = f_x(Sub) + f_x(Sub \cup \{i,j\}) - f_x(Sub \cup \{i\}) - f_x(Sub \cup \{j\}), \quad (3)$$

where Sub is a subset of pixel blocks; i and j are feature blocks; f_x is the predicted result of the model; and σ'_x is the features' relationship based on the above idea.

The main purpose of this article is to obtain the pairwise low-order relationship between superpixel blocks. Therefore, to minimize the impact of the high-order relationship between superpixel blocks on the effect, the S in the experimental section will contain

fewer non-important superpixel blocks with a small spatial correlation (relatively distant). And we use the results obtained from Formula (3) as a benchmark to calculate the similarity between our method and it:

$$Similarity = \frac{\sigma'_x(i,j) - \sigma_x(i,j)}{\sigma'_x(i,j)}, \tag{4}$$

where $\sigma_x(i,j)$ is the correlation between features obtained by this paper's method, σ'_x is the size of the relationship between features obtained during validation, and the results are shown in Table 2.

Table 2. Results of fidelity/accuracy.

Model	Resnet50	InceptionV3
S = 1	78.01%	76.22%
S = 2	77.89%	75.49%
S = 3	78.85%	76.09%

From the experiments, it can be seen that the relationships between features obtained by the method in this paper and the relationships between features obtained by Formula (3) are highly similar, so it means our method can obtain a result similar to that of the SHAP method. Instead of being essentially the same, it may be because using black pixel blocks during occlusion results in artifacts in the image, and there may be a higher-order relationship between pixel blocks, resulting in changes in the output of the model. However, after eliminating this effect as much as possible, experiments can demonstrate the correctness of the relationship obtained in this article.

4.3. Stability/Consistency Analysis

This section focuses mainly on verifying whether the interpretation results of the algorithm will be the same and whether the algorithm can achieve better stability/consistency for the same input sample with constant parameters. The method used in this paper is calculating the similarity of the selected feature set. First, we compare with LIME and define $T1/T$ as an evaluation index, where T1 is the one we believe to be the most accurate, and T is the number of experiments. The results are as follows.

In Table 3, it can be seen that when the important feature selection method in this article is used to replace the important feature selection method in LIME, the proportion of interpretation results consistent with the standard increases.

Table 3. Results of stability/consistency.

Features	K = 2		K = 3		K = 4	
Times	**N = 100**	**N = 1000**	**N = 100**	**N = 1000**	**N = 100**	**N = 1000**
LIME	31.21%	66.97%	19.66%	57.80%	12.46%	45.87%
Proposed	37.64%	70.64%	24.74%	61.39%	15.26%	47.62%

Also, we chose quantitative testing with concept activation vectors (TCAV) [21] for the second comparative experiment. This method clusters superpixels and then uses concept vector scores to select important concepts. And for the same, we choose the one with the highest number of occurrences as the correct result and calculate the result through $T1/T$. The results are as follows.

In Table 4, it can be seen that our method obtains more stable results than LIME and TCAV.

Finally, we choose randomized input sampling for explanation (RISE) [36] for comparative experiments. This method generates multiple masks through Monte Carlo sampling and then weights the masks to obtain the results. As the result of this method is heat maps,

we convert the results obtained by our method into heat maps through the superpixels' weights. And we choose the structural similarity index (SSIM) as our evaluation index. SSIM calculates the similarity score between two images by comparing their similarities in brightness, contrast, and structure. And the results are as follows.

Table 4. Evaluating stability/consistency through feature selection.

	TCAV	LIME	Proposed
K = 1	66.97%	73.61%	77.19%
K = 2	41.32%	66.97%	70.64%
K = 3	17.64%	57.80%	61.39%

In Table 5, it can be seen that, in terms of selecting features, our method can achieve higher stability/consistency. And by evaluating the similarity of heat maps, our methods can obtain higher structure similarity index measure (SSIM) scores than RISE and obtain similar results as LIME. From the above experiments, it can be seen that our method obtains better stability/consistency, which proves that when considering the correlation between features, the stability/consistency of the interpretable algorithm can be effectively improved.

Table 5. Evaluating stability/consistency through heat maps.

	RISE	LIME	Proposed
N = 500	44.62%	93.52%	95.83%
N = 1000	65.49%	95.37%	96.24%
N = 2000	70.89%	97.92%	98.03%
N = 5000	80.76%	99.04%	98.97%

4.4. Sensitivity Analysis

This section mainly focuses on whether the algorithm is sensitive to parameters when replacing the important feature selection method in LIME, that is, whether the interpretation results of the interpretable algorithm will significantly change when the parameters change. The key parameter studied in this section is the number of neighborhood data generated, N. For comparison purposes, this paper still selects LIME as the standard to change the number N of neighborhood data, which are N = 100, N = 500, N = 1000, N = 3000, and N = 5000, including 500 test images. Each image is repeated to obtain 100 interpretation results. Similarly, the sequence that occurs most when N = 5000 is considered the standard interpretation. At the same time, compare the interpretation results of algorithms under different N conditions. In order to control other conditions, the number of selected features K is set to 3, and the sensitivity is reflected by the difference in the proportion of correct explanations that can be obtained under different N conditions.

In Table 6, it can be seen that when the important feature selection method in LIME is replaced, the algorithm in this article can achieve a sensitivity that is basically similar to that of LIME. And when N is large enough, the proportion of correct results changes less as N changes. But generally, the algorithm in this article is more insensitive than LIME.

Table 6. Results of sensitivity.

Features	**InceptionV3**		**Resnet50**	
Methods	**LIME**	**Proposed**	**LIME**	**Proposed**
N = 100	19.66%	24.74%	20.34%	24.26%
N = 500	47.89%	53.52%	43.68%	49.70%
N = 1000	57.80%	61.39%	59.63%	60.66%
N = 3000	62.53%	64.62%	63.37%	66.41%
N = 5000	73.34%	74.51%	76.27%	78.64%

5. Conclusions

This paper proposes a local interpretable model-agnostic explanation method based on feature relationships that aims to directly quantify and visualize the impact of the relationship between features on model prediction. It uses the combination mask of superpixel blocks to obtain the pairwise relationship between features. It not only introduces the relationship between features into the XAI methods based on black-box models and image processing but also effectively improves the generality of searching for the relationship between features in XAI methods. In addition to enriching the interpretation results of the XAI methods, helping users better understand the decision of the model, improving the credibility of the algorithm, and reducing computational cost to realize high-performance computing, the experiments also prove that the method is more stable and consistent than LIME to a certain extent.

Author Contributions: Conceptualization, Z.C. and Z.L.; methodology, Z.C.; validation, Z.C.; formal analysis, Z.C. and Z.X.; writing—original draft preparation, Z.C. and Z.L.; writing—review and editing, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: https://github.com/kousaska-kanade/FLIME.git (accessed on 25 September 2023).

Acknowledgments: This work was supported by the Key R & D Program of Jiangsu (BE2022081).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25, Proceedings of Twenty-Sixth Conference on Neural Information Processing Systems (NIPS 2012), Lake Tahoe, NV, USA, 3–8 December 2012*; Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q., Eds.; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2012.
3. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
4. Yuan, Y.; Zhou, X.; Pan, S.; Zhu, Q.; Song, Z.; Guo, L. A Relation-Specific Attention Network for Joint Entity and Relation Extraction. In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021.
5. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.
6. Ho, N.H.; Yang, H.J.; Kim, S.H.; Lee, G. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access* **2020**, *8*, 61672–61686. [CrossRef]
7. Li, L.; Qin, L.; Xu, Z.; Yin, Y.; Wang, X.; Kong, B.; Xia, J. Regarding artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* **2020**, 201178. [CrossRef]
8. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef] [PubMed]
9. Nelder, J.A.; Wedderburn, R.W. Generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1972**, *135*, 370–384. [CrossRef]
10. Zhang, Q.; Yang, Y.; Ma, H.; Wu, Y.N. Interpreting cnns via decision trees. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
11. Wan, A.; Dunlap, L.; Ho, D.; Yin, J.; Lee, S.; Jin, H.; Gonzalez, J.E. NBDT: Neural-backed decision trees. *arXiv* **2020**, arXiv:2004.00221.
12. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should i trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
13. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
14. Pawelczyk, M.; Broelemann, K.; Kasneci, G. Learning model-agnostic counterfactual explanations for tabular data. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020.
15. Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
16. Nguyen, A.; Yosinski, J.; Clune, J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv* **2016**, arXiv:1602.03616.

17. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and global knowledge distillation for detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
18. Kang, Z.; Zhang, P.; Zhang, X.; Sun, J.; Zheng, N. Instance-conditional knowledge distillation for object detection. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16468–16480.
19. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
20. Iwana, B.K.; Kuroki, R.; Uchida, S. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Republic of Korea, 27–28 October 2019.
21. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
22. Yao, L.; Li, Y.; Li, S.; Liu, J.; Huai, M.; Zhang, A.; Gao, J. Concept-Level Model Interpretation From the Causal Aspect. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 8799–8810. [CrossRef]
23. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
24. Kim, B.; Khanna, R.; Koyejo, O.O. Examples are not enough, learn to criticize! criticism for interpretability. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
25. Zhang, Q.; Cao, R.; Shi, F.; Wu, Y.N.; Zhu, S.C. Interpreting CNN knowledge via an explanatory graph. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
26. Dikopoulou, Z.; Moustakidis, S.; Karlsson, P. GLIME: A new graphical methodology for interpretable model-agnostic explanations. *arXiv* **2017**, arXiv:2107.09927.
27. Aas, K.; Jullum, M.; Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.* **2021**, *298*, 103502. [CrossRef]
28. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
29. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
30. Messalas, A.; Aridas, C.; Kanellopoulos, Y. Evaluating MASHAP as a faster alternative to LIME for model-agnostic machine learning interpretability. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Busan, Republic of Korea, 19–22 February 2020.
31. Wang, A.; Lee, W.N. Exploring Concept Contribution Spatially: Hidden Layer Interpretation with Spatial Activation Concept Vector. *arXiv* **2022**, arXiv:2205.11511.
32. Ge, Y.; Xiao, Y.; Xu, Z.; Zheng, M.; Karanam, S.; Chen, T.; Wu, Z. A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
33. Tomsett, R.; Harborne, D.; Chakraborty, S.; Gurram, P.; Preece, A. Sanity checks for saliency metrics. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 6021–6029. [CrossRef]
34. Alvarez Melis, D.; Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7786–7795.
35. Montavon, G.; Samek, W.; Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [CrossRef]
36. Petsiuk, V.; Das, A.; Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv* **2018**, arXiv:1806.07421.