A MODIFIED PERTURBED SAMPLING METHOD FOR LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATION

Sheng Shi[†] Xinfeng Zhang^{*} Wei Fan[†]

[†] AI Laboratory, Lenovo Research, Beijing 100094, China ^{*} University of Chinese Academy of Sciences, Beijing 100049, China

ABSTRACT

Explainability is a gateway between Artificial Intelligence and society as the current popular deep learning models are generally weak in explaining the reasoning process and prediction results. Local Interpretable Model-agnostic Explanation (LIME) is a recent technique that explains the predictions of any classifier faithfully by learning an interpretable model locally around the prediction. However, the sampling operation in the standard implementation of LIME is defective. Perturbed samples are generated from a uniform distribution, ignoring the complicated correlation between features. This paper proposes a novel Modified Perturbed Sampling operation for LIME (MPS-LIME), which is formalized as the clique set construction problem. In image classification, MPS-LIME converts the superpixel image into an undirected graph. Various experiments show that the MPS-LIME explanation of the black-box model achieves much better performance in terms of understandability, fidelity, and efficiency.

Index Terms— Explainable AI, Local Fidelity, Feature Correlations, Perturbed sampling, Clique

1. INTRODUCTION

Artificial Intelligence (AI) [1] [2] [3] has gone from a science-fiction dream to a critical part of our everyday life. Notably, deep learning has achieved superior performance in image classification and other perception intelligence tasks. Despite its outstanding contribution to the progress of AI, deep learning models remain mostly black boxes, which are extremely weak in explaining the reasoning process and prediction results. Nevertheless, many real-world applications are mission-critical, and users concern about how the AI solution is arriving at its decisions and insights. Therefore, model transparency and explainability are essential to ensure AI's broad adoption in various vertical domains.

There has been a recent surge in the development of explainable AI techniques [4] [5] [6]. Among them, the post hoc techniques for explaining black-box models in a human-understandable manner have received much attention in the research community [7] [8] [9]. Model-agnostic is the prominent characteristic of these methods, which generate

perturbed samples of a given instance in the feature space and observe the effect of these perturbed samples on the output of the black-box classifier. In [7], the authors proposed the Local Interpretable Model-agnostic Explanation (LIME), which explains the predictions of any classifier faithfully by fitting a linear regression model locally around the prediction. The sampling operation for LIME is a random uniform distribution, which is straightforward but defective, ignoring the correlation between features. Proper sampling operation is especially essential in natural image recognition because the visual features of natural objects exhibit a strong correlation in the spacial neighborhood, rather than a complete uniform distribution. In some cases, when most uniformly generated samples are unrealistic about the actual distribution, false information contributors lead to poorly fitting of the local explanation model.

In this paper, we propose a Modified Perturbed Sampling method for LIME (MPS-LIME), which takes into full account the correlation between features. We convert the superpixel image into an undirected graph, and then the perturbed sampling operation is formalized as the clique set construction problem. We perform various experiments on explaining Google's pre-trained Inception neural network [10]. The experimental results show that the MPS-LIME explanation of the black-box model can achieve much better performance than LIME in terms of understandability, fidelity, and efficiency.

2. MPS-LIME EXPLANATION

In this section, we first introduce the interpretable image representation and the modified perturbed sampling for local exploration. Then we present the explanation system of MPS-LIME.

2.1. Interpretable Image Representation

An interpretable representation should be understandable to observers, regardless of the underlying features used by the model. Most image classification tasks represent the image as a tensor with three color channels per pixel. Considering the poor interpretability and high computational complexity

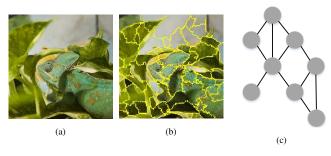


Fig. 1. (a) Pixel-based image; (b) Superpixel image; (c) Constructing a graph of all superpixel blocks

of the pixel-based representation, we adopt a superpixel based interpretable representation. Each superpixel, as the primary processing unit, is a group of connected pixels with similar colors or gray levels. Superpixel segmentation is dividing an image into some non-overlapping superpixels. More specifically, we denote $x \in \mathbb{R}^d$ be the original representation of an image, and binary vector $x' \in \{0,1\}^{d'}$ be its interpretable representation where 1 indicates the presence of original superpixel and 0 indicates an absence of original superpixel.

2.2. A Modified Perturbed Sampling for Local Exploration

In order to learn the local behavior of image classifier f, we generate a group of perturbed samples of a given instance, x, by activating a subset of superpixels in x. For the images, especially natural images, superpixel segments often correspond to the coherent regions of visual objects, showing strong correlation in a spacial neighborhood. If the activated superpixels come from an independent sampling process, we may lose much useful information to learn the local explanation models. The perturbed sampling operation in the standard implementation of LIME is to draw nonzero elements of x' uniformly at random. This approach is at risk of ruining the learning process of local explanation models, since the generated samples may ignore the correlation between superpixels.

In this section, we propose a modified perturbed sampling method, which takes into full account the correlation among superpixels. Firstly, we convert the superpixel segments into an undirected graph. Specifically, as shown in Figure 1, the superpixel segments are represented as vertices of a graph whose edges connect to only those adjacent segments. Considering a graph G=(V,E), where V and E are the sets of vertices and undirected edges, with cardinalities |V|=d' and |E|, a subset of V can be represented by a binary vector $z' \in \{0,1\}^{d'}$, where 1 indicates that vertice is in the subset.

The modified perturbed sampling operation is formalized as finding the clique C ($C \subseteq V$), where every two vertices are adjacent. Since the cardinality of maximum clique of the constructed graph is 3, the clique C consists of three subset $C = C_1 \cup C_2 \cup C_3$. The three subsets are as follows: C_1

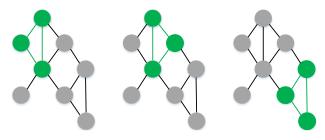


Fig. 2. The clique C_3 which is the subset that contains three vertices, where every two vertices are adjacent (marked green)

is the subset that only contains one vertice. C_2 is the subset that only contains two vertices that are connected by an edge. C_3 is the subset that contains three vertices, and every two vertices are adjacent (Figure 2). In this paper, we use the Depth-First Search (DFS) method to get the clique C. Algorithm 1 shows a simplified workflow diagram.

Algorithm 1 DFS(graph, V, v, clique, visited, start, path, n)

```
1: visited[v] \leftarrow True
 2: if n==0 then
 3:
      visited[v] \leftarrow False
 4:
      if graph[v][start]==1 then
 5:
         c=copy.deepcopy(path)
         clique.append(c)
 6:
 7:
         path.pop()
         return clique
 8:
 9:
         return clique
10:
      end if
11:
12: end if
   for i in range(V) do
      if visited[i]==False and graph[v][i]==1 then
14:
15:
         path.append(i)
         pp=DFS(graph, V, v, clique, visited, start, path, n-1)
16:
         for node in path do
17:
            if visited[node]==False then
18:
19:
               path.remove(node)
20:
            end if
         end for
21:
      end if
22:
23: end for
24: visited[v]=False
25: return clique
```

Since there is a strong correlation between the adjacent superpixel image segments, the clique C set construction can take into full account the various types of neighborhood correlation. Moreover, the number of perturbed samples of MPS-LIME is much smaller than that in the current implementation of LIME, which significantly reduces the runtime.

2.3. Explanation System of MPS-LIME

The goal of the explanation system is to identify an interpretable model over the interpretable representation that is locally faithful to the classifier. We denote the original image classification model being explained by f, and the interpretable model by g. This problem can be formalized as an optimization problem:

$$\xi(x) = argmin \quad L(f, g, \pi_x) + \Omega(g),$$
 (1)

where the locality fidelity loss $L(f, g, \pi_x)$ is calculated by the locally weighted square loss:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} e^{(-D(x, z)^2 / \sigma^2)} (f(z) - g(z'))^2.$$
 (2)

The database Z is composed of perturbed samples $z' \in \{0,1\}^{d'}$ which are sampled around x' by the method described in Section 3.2. Given a perturbed sample z', we recover the sample in the original representation $z \in \mathbb{R}^d$ and get f(z). Moreover, $\pi_x(z)$ is the L_2 distance function to capture locality.

Algorithm 2 shows a simplified workflow diagram of MPS-LIME. Firstly, MPS-LIME gets the superpixel image by using the segment method. Then it converts the superpixel image segments into an undirected graph. The database Z is constructed by finding the clique of an undirected graph, which is solved by the DFS method. Finally, MPS-LIME gets the g by using the K-LASSO method, which is the same as that in LIME [7].

Algorithm 2 Modified Perturbed Sampling Method for Local interpretable model-agnostic explanation (MPS-LIME)

```
Require: Classifier f, Instance x, Length of explanation K

1: get superpixel image x' by segment method

2: get f(x') by classifier f

3: convert the superpixel image segments into an undirected graph

4: initial Z \leftarrow \{\}

5: construct the clique C by DFS method

6: for z' \in C do

7: get z by recovering z'

8: Z \leftarrow Z \cup (z'_i, f(z_i), \pi_x(z_i))

9: end for

10: get \omega by \leftarrow K-Lasso(Z, K)

11: return \omega
```

3. EXPERIMENTAL RESULTS

In this section, we perform various experiments on explaining the predictions of Google's pre-trained Inception neural network [10]. We compare the experimental results between LIME and MPS-LIME in terms of understandability, fidelity, and efficiency.

3.1. Measurement criterion of interpretability

Fidelity, understandability, and efficiency are three important goals for interpretability [11] [12]. An explainable model with good interpretability should be faithful to the original model, understandable to the observer, and graspable in a short time so that the end-user can make wise decisions. Mean Absolute Error (MAE) and Coefficient of determination \mathbb{R}^2 are two import measures of fidelity. MAE is the absolute error between the predicted value and true value, which can reflect the predictive accuracy well,

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_{true(i)} - y_{pred(i)}|.$$
 (3)

 \mathbb{R}^2 is calculated by Total Sum of Squares (SST) and Error Sum of Squares (SSE):

$$R^{2} = 1 - SSE/SST$$

$$SSE = \sum_{i=1}^{n} (y_{true(i)} - y_{pred(i)})^{2}$$

$$SST = \sum_{i=1}^{n} (y_{true(i)} - y_{mean(i)})^{2},$$
(4)

where y_{true} is the true value, y_{pred} is the predicted value and y_{mean} is the mean value of true value. The best R^2 is 1.0. The closer the score is to 1.0, the better the performance of fidelity is to explainer.

3.2. Google's Inception neural network on Image-net database

We explain image classification predictions made by Google's pre-trained Inception neural network [10]. The first row in Figure 3 shows six original images. The second row and third row are the superpixels explanations by LIME and MPS-LIME, respectively. The explanations highlight the top 5 superpixel segments, which have the most considerable positive weights towards the predictions (K=5).

Table 1 lists the MAE of LIME and MPS-LIME. We find some of the predictive probability values of LIME is bigger than 1.0. This is because LIME adopts a sparse linear model to fit the perturbed samples, and has no more constraints such as the probability values distribution should range between 0 and 1. Comparing to LIME, we can see that MPS-LIME provides better predictive accuracy than LIME. Besides, R^2 of LIME and MPS-LIME are listed in Table 1. The closer the score is to 1.0, the better the performance of fidelity is to an explainer. The R^2 of MPS-LIME is around 0.9, which is much bigger than LIME. By comparing the MAE and R^2 of two algorithms, we can conclude that MPS-LIME has better fidelity than LIME.

Efficiency is highly related to the time necessary for a user to grasp the explanation. The runtime of LIME and MPS-LIME are shown in Table 2, which shows that the runtime of MPS-LIME is nearly half as the runtime of LIME. We can conclude from the above results that MPS-LIME not only has a higher fidelity but also take less time than LIME.

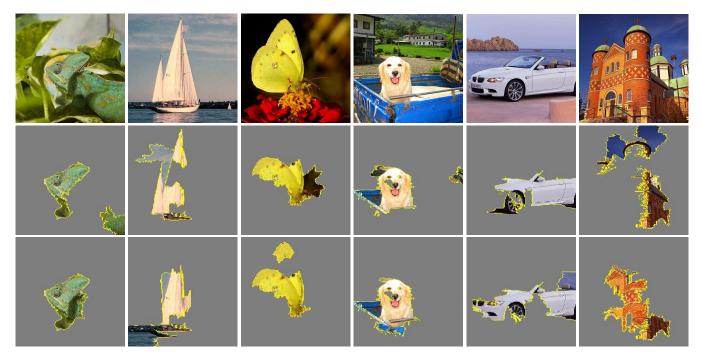


Fig. 3. Explaining image classification predictions made by Google's Inception neural network. The first row shows 6 original images, and the top 1 class predicted of the original images are $African\ chameleon(p=0.9935)$, yawl(p=0.6077), $sulphur\ butterfly(p=0.9431)$, $golden\ retriever(p=0.5641)$, convertible(p=0.9356), castle(p=0.7646). The second row shows the superpixels explanations by LIME (K=5). The third row shows the superpixels explanations by MPS-LIME (K=5).

Table 1. The MAE of LIME and MPS-LIME on Google's pretrained Inception neural network

	true prob (Inception)	pred prob	MAE	R^2
LIME	$p_{chameleon} = 0.9935$	1.6285	0.6350	0.6885
MPS-LIME	Penameteon 0.0000	0.9783	0.0152	0.8944
LIME	$p_{yawl} = 0.6077$	0.8291	0.2214	0.4531
MPS-LIME	Fyawi 3.33.	0.5973	0.0104	0.9825
LIME	$p_{butterfly} = 0.9431$	1.6668	0.7237	0.636
MPS-LIME	Fourier jig 5.5 -5-	0.9284	0.0147	0.9222
LIME	$p_{retriever} = 0.5641$	0.4822	0.0819	0.6958
MPS-LIME	Pretriever 0.0011	0.5568	0.0073	0.9304
LIME	$p_{convertible} = 0.9356$	1.2854	0.3498	0.7407
MPS-LIME	Peontertible 3.0333	0.9203	0.0153	0.9925
LIME	$p_{castle} = 0.7646$	1.0166	0.2520	0.3535
MPS-LIME	Feasite 011010	0.7531	0.0115	0.9155

Table 2. The runtime of LIME and MPS-LIME on Google's pre-trained Incention neural network

trained inception neural network										
	img1	img2	img3	img4	img5	img6				
LIME	232.20	230.45	245.36	264.51	223.79	226.58				
MPS-LIME	91.02	113.85	109.29	154.57	117.21	152.84				

4. CONCLUSION AND FUTURE WORK

The sampling operation for local exploration in the current implementation of LIME is a random uniform sampling, which possibly generates unrealistic samples ruining the learning of local explanation models. In this paper, we propose a modified perturbed sampling method MPS-LIME, which takes into full account the correlation between features. We convert the superpixel image into an undirected graph, and then the perturbed sampling operation is formalized as the clique set construction problem. We perform various experiments on explaining the random-forest classifier and Google's pre-trained Inception neural network. Various experiment results show that the MPS-LIME explanation of multiple black-box models can achieve much better performance in terms of understandability, fidelity, and efficiency.

There are some avenues of future work that we would like to explore. This paper only describes the modified perturbed sampling method for image classification. We will apply the similar idea to text processing and structural data analytics. Besides, we will improve other post hoc explanations techniques that rely on input perturbations such as SHAP and propose a general optimization scheme.

5. REFERENCES

- [1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016, http://www.deeplearningbook.org.
- [2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning, Springer, 2009, www.web.stanford.edu/~hastie/ElemStatLearn.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017.
- [4] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing, "Harnessing deep neural networks with logic rules," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, 2016.
- [5] Yin Lou, Rich Caruana, and Johannes Gehrke, "Intelligible models for classification and regression," in *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, 2012, pp. 150–158.
- [6] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu, "Interpretable convolutional neural networks," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 8827–8836.
- [7] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, "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, August 13-17, 2016*, 2016, pp. 1135–1144.
- [8] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, "Model-agnostic interpretability of machine learning," *CoRR*, vol. abs/1606.05386, 2016.
- [9] Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, Eds., Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017.
- [10] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in

- 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 1–9.
- [11] Christoph Molnar, "Interpretable machine learning: A guide for making black box models explainable," in *Lulu, 1st edition, March 24, 2019; eBook, 2019.*
- [12] S. Ruping, "Learning interpretable models (phd thesis)," in *Technical University of Dortmund*, 2006.