Artificial Intelligence 314 (2023) 103823

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

$\mathcal{G}$-LIME: Statistical learning for local interpretations of deep neural networks using global priors

Xuhong Li [a], Haoyi Xiong [a,*], Xingjian Li [a], Xiao Zhang [b], Ji Liu [a], Haiyan Jiang [a], Zeyu Chen [a], Dejing Dou [a]

[a] *Baidu Inc., Beijing, China*
[b] *Tsinghua University, Beijing, China*

ARTICLE INFO

Article history:
Received 14 September 2021
Received in revised form 5 August 2022
Accepted 9 November 2022
Available online 14 November 2022

Keywords:
Explainable AI (XAI)
Interpretable deep learning

ABSTRACT

To explain the prediction result of a Deep Neural Network (DNN) model based on a given sample, LIME [1] and its derivatives have been proposed to approximate the local behavior of the DNN model around the data point via linear surrogates. Though these algorithms interpret the DNN by finding the key features used for classification, the random interpolations used by LIME would perturb the explanation result and cause the instability and inconsistency between repetitions of LIME computations. To tackle this issue, we propose $\mathcal{G}$-LIME that extends the vanilla LIME through high-dimensional Bayesian linear regression using the sparsity and informative global priors. Specifically, with a dataset representing the population of samples (e.g., the training set), $\mathcal{G}$-LIME first pursues the global explanation of the DNN model using the whole dataset. Then, with a new data point, $\mathcal{G}$-LIME incorporates an modified estimator of ElasticNet-alike to refine the local explanation result through balancing the distance to the global explanation and the sparsity/feature selection in the explanation. Finally, $\mathcal{G}$-LIME uses Least Angle Regression (LARS) and retrieves the solution path of a modified ElasticNet under varying ℓ_1-regularization, to screen and rank the importance of features [2] as the explanation result. Through extensive experiments on real world tasks, we show that the proposed method yields more stable, consistent, and accurate results compared to LIME.

1. Introduction

While deep neural networks (DNNs) [3] have achieved remarkable performance in a variety of tasks, explaining their behaviors remains challenging due to their hierarchical non-linear nature in a black-box fashion [4]. The lack of interpretability raises a severe issue on the trustworthiness of DNNs in decision-critical domains, such as autonomous driving, healthcare, and financial services. Implementing unreliable decisions made by black-box models without further interpretations or explanations[1] may expose the risk of adverse societal consequences related to privacy or fairness [5–7].

* Corresponding author.
E-mail addresses: lixuhong@baidu.com (X. Li), xionghaoyi@baidu.com (H. Xiong), lixingjian@baidu.com (X. Li), xzhang19@mails.tsinghua.edu.cn (X. Zhang), liuji04@baidu.com (J. Liu), jianghaiyan01@baidu.com (H. Jiang), chenzeyu01@baidu.com (Z. Chen), doudejing@baidu.com (D. Dou).

[1] In this work, we use the terms “interpretation” and “explanation” interchangeably.

https://doi.org/10.1016/j.artint.2022.103823
0004-3702/© 2022 Elsevier B.V. All rights reserved.

Existing interpretation algorithms To explain the prediction results of DNN models based on given samples, recent studies [1,8–12] majorly focus on interpolating and approximating the local regions and/or global regimens of DNNs in the feature domain. In addition, they use simple and self-interpretable models such as linear and tree-based models. In these studies, the contribution or importance of every feature for classifying either a *specific sample* or *the whole dataset* could be obtained as the results for *local* and *global* interpretations respectively.

For example, LIME [1] incorporated a new concept, namely *local fidelity*, to enable the local explanation. Given a data point, interpretable surrogate models, such as linear combinations of features or linear regression models, have been proposed in LIME to approximate the DNN model behaviors with small random perturbations for interpolation in the input feature domains. While such strategies work to a wide range of models, the random perturbation often causes inconsistent interpretations as the explanation results for the same data point on the same model might be different in two independent trials [9,13–16], leading to unstable explanation results at times. Some comparison examples are shown later in Fig. 8. Thus, there needs a way to *stabilize the results of LIME under the perturbation caused by random interpolations and finally improve the consistency of interpretation results*.

In addition to local approximations, one could also interpret DNN models in a global manner. To achieve the goal, [17,10,11] propose to run LIME [1] or gradient-based interpretations [18–20] for every sample in a large dataset, and then simply aggregate local approximations, so as to pursue the global interpretation. While the global interpretation provides some trends of DNNs' behaviors in the overall feature space, it fails to sketch the behaviors of DNNs especially in most local areas. For example, a global linear model is incapable of characterizing the multiple linear subregions [21–23] that co-exist in a DNN. Thus, there needs a way to *balance the local and global interpretations for better explanation*.

Furthermore, with a linear surrogate model, LIME [1] optionally uses LASSO [24] to first select a subset of features and later assign weights to selected features for regression, in an ad-hoc manner. While subset selection in regression has been proven to be NP-hard [25,26] for achieving the optimal solution, there needs an *end-to-end approach that ensure the sparsity or control the number of nonzero weight coefficients* [27] in the results of interpretations, especially for high-dimensional data (e.g., images and/or texts). Of-course, there exist other ways to interpret the prediction results in addition to linear models, such as model reconciliation [28], sample contribution to predictions [12], rule/example-based explanations [29], and perception of raw features with logic [30], which are not in the scope of our work.

Our contributions In this work, we reformulate the original linear model based local interpretation problem using the high-dimensional Bayesian linear regression framework with a global prior [31]. Specifically, we consider that the local interpretation result for every sample independently distributes over a prior multivariate Gaussian distribution with a center location at the global interpretation result of the model [32]. Endowed by a Ridge-type estimator [33], the use of global priors with ℓ_2-regularization could lower the variances of estimation through constraining the solution around an informative global prior. Furthermore, to improve the feature selection in the interpretation result, an ℓ_1-regularization is incorporated to penalize the number of nonzero elements in the regression in a noisy setting [34]. In this way, we obtain a novel estimator for LIME in an ElasticNet-style [35], namely *Modified ElasticNet*, which incorporates both ℓ_2 and ℓ_1-regularization for the global prior and sparsity respectively. We show the $\mathcal{G}$-LIME explanation results for visual tasks based on superpixels in Fig. 1, where superpixels have been gathered from the image to form the features for the explanation results. In summary, we make contributions as follows.

- This work studies the linear local surrogate [1] to interpret the classification results of DNN models for input samples, with respect to the consistency between local and global interpretations of the model [17,10,11]. In this way, we reformulate the research problem in a Bayesian linear regression framework, where the prior probability distribution for the regression is defined by the global interpretation. To the best of our knowledge, this work is the first to balance the local and global interpretations of DNNs for linear surrogate models in a Bayesian framework.
- We propose $\mathcal{G}$-LIME, an efficient and effective sparse Bayesian estimator of LIME with the global interpretation as prior. Specifically, with aggregated global interpretations, given an individual data point, $\mathcal{G}$-LIME interprets prediction results of DNN models using a modified ElasticNet estimator, where an ℓ_2-regularizer is used to constrain the local interpretation around the informative global prior and an ℓ_1-regularizer is used to control the sparsity of the explanation result. Furthermore, the least angle regression (LARS) [36] is used to retrieve the solution path of the modified ElasticNet varying ℓ_1-regularization, so as to provide an end-to-end explanation endowed with the sparsity and global priors.
- We carry out extensive experiments and evaluate $\mathcal{G}$-LIME in comparison with LIME and other interpretability baselines. The experiments are based on real-world datasets including computer vision (CV), natural language processing (NLP), and structural datasets with commonly-used deep models such as ResNet [37], EfficientNet [38], AlexNet [39] and the standard multilayer perceptron (MLP) models. Specifically, we compare $\mathcal{G}$-LIME with LIME and other interpretation algorithms using local fidelity evaluation, deletion/insertion experiments, and the stability evaluation. Results clearly demonstrate the effectiveness and efficiency of $\mathcal{G}$-LIME in providing more stable and accurate explanations. The ablation studies show that the use of global prior could significantly improve and refine the local explanation results.

The rest of this manuscript is organized as follows. Section 2 reviews preliminaries in interpretation algorithms for DNN models and common ways to evaluate the explanation results. In Section 3, we present the design of $\mathcal{G}$-LIME, including

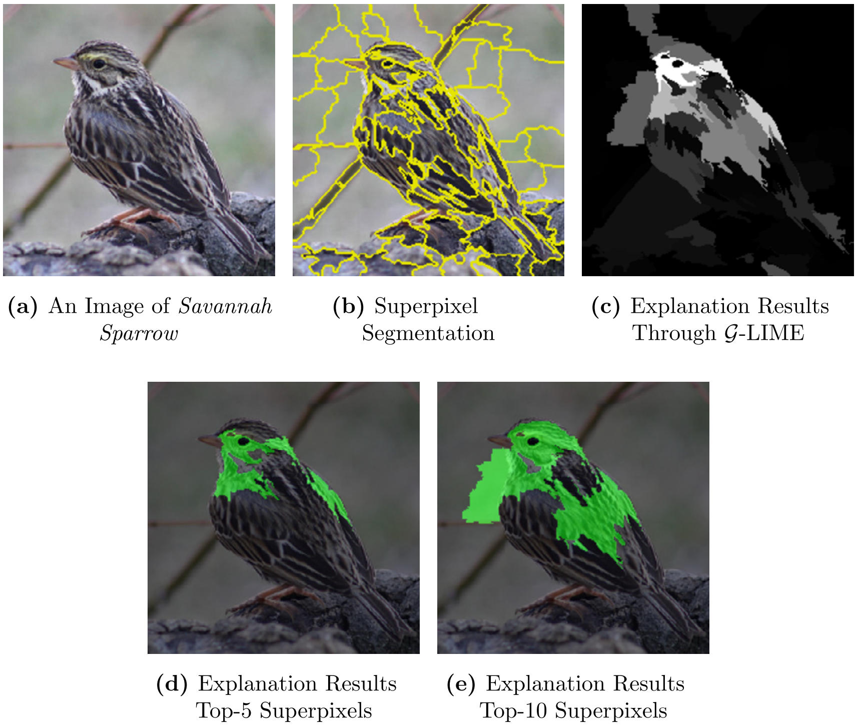

(a) An Image of *Savannah Sparrow*
(b) Superpixel Segmentation
(c) Explanation Results Through $\mathcal{G}$-LIME
(d) Explanation Results Top-5 Superpixels
(e) Explanation Results Top-10 Superpixels

Fig. 1. Illustration of an image with its superpixel segmentation, explanation results with different presentations. The image is correctly classified into the "Savannah Sparrow" class with 75% probability. $\mathcal{G}$-LIME explanation results based on superpixels are presented in gray-scale with the attributed values to the superpixels (see the middle figure). For a better visualization, the explanation results are also shown by highlighting the top superpixels in green (see the right two figures). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

algorithm design, analysis, and adaptation to computer vision (CV) tasks. In Section 4, we present experiments and results to evaluate $\mathcal{G}$-LIME with comparisons and analysis. In Section 5, we conclude this paper.

2. Preliminaries and related works

In this section, we review algorithms that locally and globally explains DNN models' behaviors and the evaluation methods.

2.1. Local explanations

Local Interpretable Model-agnostic Explanations (LIME) [1] searches an interpretable model, usually a linear one, to approximate the output of a deep model for an individual data point, such that LIME obtains a weighted linear combination of features including the importance of every feature for classifying the data point. For the local approximation, LIME introduces the random interpolation in the feature domain to generate samples around the data point, that are used for approximating the linear model. Specifically, given a d-dimensional vector as the input data point $\boldsymbol{x}$, LIME generates n random samples (denoted as a $n \times d$ matrix as $\boldsymbol{X}$) around $\boldsymbol{x}$ through injecting small random noises to $\boldsymbol{x}$, then LIME classifies all these n samples using the DNN model and obtains the vector of classification results $\boldsymbol{y}$ for every random sample in $\boldsymbol{X}$ accordingly. Finally, LIME interprets the local behavior of the DNN model around the sample $\boldsymbol{x}$ through the Ridge regression

$$\boldsymbol{\beta}^{\text{LIME}} \leftarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min}\, (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the weighted linear combination of features to be optimized, and λ is the tuning parameter for regularization. In the vanilla LIME, samples are weighted and the sample weights can be represented by a diagonal matrix $\boldsymbol{D} \in \mathbb{R}_+^{n\times n}$. For the reason that $\boldsymbol{D}$ can be written as $\sqrt{\boldsymbol{D}}^\top \sqrt{\boldsymbol{D}}$ and integrated into the inner product and that our approach does not change the sample weights, we omit $\boldsymbol{D}$ for simplicity. Note that to lower the complexity of regression in high-dimensional settings, LIME first selects a subset of features for performing above regression (in Eq. (1)) through variables/predictors selection based on LASSO [24].

The closed-form solution for the Ridge regression can be deduced by simply taking the first-order derivative of the objective function to zero:

$$\boldsymbol{\beta}^{\text{LIME}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2}$$

In this way, we consider $\boldsymbol{\beta}^{\text{LIME}}$ obtained in Eq. (2) as the exact solution of LIME for the local interpretation of the DNN model around the data point $\boldsymbol{x}$, while the tuning parameter λ is used to ensure the non-singularity of the regression and would affect the results of interpretations.

Related local explanations A number of interpretation algorithms have been proposed recently, mainly focusing on improving the stability of LIME. For example, ALIME [40] proposed to weight the generated samples by using an auto-encoder which, however, requires the training set and one additional training process. This would be difficult if the training set is not accessible. Anchors [41] proposed to find a subset of input features that are considered to be important by the model with much more computation. Instead of using the synthetically generated samples to train the linear model, D-LIME [14] proposed to find a relevant cluster of the local data point from the training set. This improves the stability of explanations but it is very challenging for visual tasks where the input features are not explicitly shared by different images. Moreover, there are several others [42,43,16] that focus on improving the explanation quality of LIME.

Our proposed $\mathcal{G}$-LIME differs from them mainly in the integration of sparsity and global prior into the one single estimator of the modified ElasticNet. $\mathcal{G}$-LIME is suitable for all tasks and does not require the access to the training set. We believe our approach and theirs provide improvements in orthogonal directions for better explaining the DNN models. In this work, to demonstrate the effectiveness and efficiency from our proposed approach $\mathcal{G}$-LIME, we directly compare with LIME and related variants, e.g. Anchors.

Besides LIME variants, Integrated Gradient (IG) [20] introduced two axioms violated by LIME to improve the reliability of explanations, SHAP [44] incorporated a game-theoretic framework, and Grad-CAM [18] was specifically designed for convolutional networks. See the recent surveys [45,46] for more explanation algorithms. To show the advantage of our methods, we also compare with SmoothGrad [19], IG [20] and Grad-CAM [18]. We will show that the two pixel-level explanations (SmoothGrad and IG) provide explanations that do not cover all important features (low insertion scores, as introduced later) while the low-resolution explanation Grad-CAM provides less discriminative rankings for the important features (low deletion scores, as introduced later).

2.2. Global explanations through aggregations of local explanations

In addition to the local interpretations, the LIME explanations have been frequently used for global interpretations [11, 10], aiming to obtain a weighted linear combination of features that represents the feature importance for classifying a large set of samples based on a DNN model.

Let $\boldsymbol{B} \in \mathbb{R}^{m \times d}$ be the LIME explanations of m samples, where each row is a linear weight vector $\boldsymbol{\beta}$. Then we briefly recall four global aggregation approaches: LIME-SP [1], Averaged-Importance [10], Homogeneity [10] and NormLIME [11], as follows.

- *LIME-SP:* Due to the limited budget, LIME-SP [1] searches a subset from m samples via a sub-modular pick (SP) algorithm. We consider all the m samples and do not need the sample-picking strategy. So LIME-SP aggregates local interpretations as follows,

$$\boldsymbol{\beta}_j^{\text{SP}} = \sqrt{\sum_{i=1}^{m} |B_{ij}|}, \tag{3}$$

  where $|\cdot|$ refers to the absolute value. $|\boldsymbol{\beta}_j^{\text{SP}}|$ increases when the corresponding feature tends to be more influential on the DNN model's output, or the feature occurs more often on various data points.
- *Averaged-Importance:* LIME-SP would be biased towards features that always occur. For example, in NLP tasks, common words occur more often than other words, but they should not be assigned a large value of importance. Averaged-Importance [10] copes with this problem by simply performing an average operation,

$$\boldsymbol{\beta}_j^{\text{AVG}} = \frac{\sum_{i=1}^{m} |B_{ij}|}{\sum_i \mathbb{1}_{B_{ij} \neq 0}}. \tag{4}$$

- *NormLIME:* For one specific data point, LIME interpretations are reasonable because the obtained linear parameters can be compared among all features. However, for a set of data points, comparing the parameters on the same feature is less plausible because different data points have variant scales/norms of the linear parameters. To this end, [11] proposed to normalize the LIME interpretation before averaging,

$$\boldsymbol{\beta}_j^{\text{N}} = \frac{1}{\sum_i \mathbb{1}_{B_{ij} \neq 0}} \sum_{i=1}^{m} \frac{B_{ij}^2}{\|B_i\|_1}, \tag{5}$$

  where $\|B_i\|_1$ is the ℓ_1 norm of the i-th sample's LIME explanation.

- *Homogeneity:* Homogeneity-weighted global importance [10] aims at quantifying the homogeneity per feature by Shannon entropy and re-scaling the LIME-SP result,

$$\boldsymbol{\beta}_j^{\mathrm{H}} = \left(1 - \frac{H_j - H_{min}}{H_{max} - H_{min}}\right) \boldsymbol{\beta}_j^{\mathrm{SP}}, \tag{6}$$

  where $H_j = -\sum_{c'} p_j(c') \log p_j(c'))$ summed over all classes, $p_j(c) = \frac{\sqrt{\sum_{i \in S_c} |\boldsymbol{B}_{ij}|}}{\sum_{c'} \sqrt{\sum_{i \in S_{c'}} |\boldsymbol{B}_{ij}|}}$, and S_c is the set of all data points that are labeled to the class c.

2.3. Explanation method evaluations

To evaluate the explanation methods, the trustworthiness and the stability are usually assessed [47,13,48,49,46]. The intuition is that the explanation method should be trustworthy and faithful to the original model behaviors, otherwise the explanation result is not valuable and helpful; and that the explanation method should be stable when explaining similar inputs. We summarize several common evaluation methods below, note that in this paper, we mainly consider evaluating the explanation methods that provide explanations of feature importance for the original model.

- Local fidelity (R^2 score and mean square error). LIME [1] measures the local fidelity for explanation method evaluation, which indicates the R^2 score between original model outputs and surrogate (linear) model outputs, or the mean square error between the two [48].
- Deletion/Insertion. RISE [50] proposes to *delete* the features from an original data according to the explanation results and compute the probability degradation with respect to the DNN model. The model quickly fails to do the right prediction if the most important features are removed, so the degradation speed, measured by the area under curve (AUC), indicates the quality of explanation results. Similarly, RISE also proposes to *insert* the features from an average data and computes the AUC accordingly. The model performance degrades when removing important features because of two reasons: the important information is removed or the data distribution is changed. For these reasons, [51,11] proposed to additionally retrain the model to eliminate the factor of data distribution changes.
- Stability [13,48]. The local explanation methods give inconsistent results when the number of interpolated input samples is small [1,13,52]. We will show an illustration later in Fig. 8. The motivation is to encourage stable explanation methods that provide consistent results when similar inputs are given. This can be quantitatively measured by the stability metric [13,48].
- We evaluate the proposed explanation method $\mathcal{G}$-LIME with metrics mentioned above while there are several other evaluation methods including infidelity and sensitivity evaluations [53], BAM [54] and so on.

In this paper, we validate our proposed method $\mathcal{G}$-LIME using local fidelity, deletion/insertion and stability evaluations.

3. $\mathcal{G}$-LIME

Our work considers the original LIME with linear surrogate models (Eq (1)) as an application of Ridge estimator. In order to improve the stability of LIME, we propose to use the sparsity and the informative global priors via a modified ElasticNet estimator. Sparsity is favored through LASSO by LIME too, while LIME uses LASSO to first select a subset of features and then assign weights to selected features by regression. We achieve this in an end-to-end manner, as shown later in Eq (7). The ℓ_1 regularization term for encouraging sparsity and the ℓ_2 for reducing variance are used together. More importantly, we note that the ℓ_2 does not always introduce bias because we integrate the global priors into the ℓ_2 term. In this way, compared to LIME, our proposed method gets additional global information from the aggregated local explanations. This does not only alleviate the randomness of LIME explanations, but also improves the faithfulness of explanations (in condition that the global prior is informative and well exploited). To this end, we propose to improve LIME using the sparsity and informative global priors via a modified ElasticNet estimator.

3.1. Main framework

$\mathcal{G}$-LIME starts by pursuing the global interpretation $\boldsymbol{\beta}^{\mathrm{global}}$, and then computes the explanation through a modified ElasticNet estimator with $\boldsymbol{\beta}^{\mathrm{global}}$ as a global prior and fast feature screening via LARS. The three processes are shown in Algorithm 1 and listed as follows.

1. *Global Interpretation Pursuit* - As shown in lines 7–12 of Algorithm 1, given a DNN model for interpretation and a dataset representing the population of samples, $\mathcal{G}$-LIME first obtains the local interpretation for every sample in the dataset using LIME via linear local surrogate. Then, $\mathcal{G}$-LIME aggregates these local interpretations and pursues the global interpretation $\boldsymbol{\beta}^{\mathrm{global}}$ of the DNN model through NormLIME [11] and Averaged-Importance [10]. For the design of these two algorithms, please refer to preliminaries introduced in Section 2.2.

Algorithm 1: $\mathcal{G}$-LIME Main Framework.

```
 1 /* Initialization */
 2 Input: $\boldsymbol{D}$ the dataset for global prior pursuit
 3 Input: model$(\cdot)$ the model for interpretation
 4 Input: $\boldsymbol{x}$ the sample for interpretation
 5 Input: $\lambda_2^*$ the regularization effects of $\ell_2$-regularization
 6 Set $\boldsymbol{B}$, $\boldsymbol{X}$ and $\boldsymbol{y}$ to empty sets
 7 /* Global Prior Pursuit */
 8 for each $\boldsymbol{x}'$ in $\boldsymbol{D}$ do
 9     $\boldsymbol{\beta}' \leftarrow \mathrm{LIME}(\mathrm{model}, \boldsymbol{x}')$
10     $\boldsymbol{B} \leftarrow \boldsymbol{B} \cup \boldsymbol{\beta}'$
11 end
12 $\boldsymbol{\beta}^{\mathrm{global}} \leftarrow \mathrm{NormLIME}(\boldsymbol{B})$ or AverageImportance$(\boldsymbol{B})$
13 /* Building the Modified ElasticNet Estimator */
14 $\boldsymbol{X} \leftarrow \mathrm{RandomInterpolations}(\boldsymbol{x})$
15 for each $\boldsymbol{x}''$ in $\boldsymbol{X}$ do
16     $\boldsymbol{y} \leftarrow \boldsymbol{y} \cup \mathrm{model}(\boldsymbol{x}'')$
17 end
18 $\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2) \leftarrow \mathrm{argmin}_{\beta} \|\boldsymbol{X}\beta - \boldsymbol{y}\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta - \boldsymbol{\beta}^{\mathrm{global}}\|_2^2$
19 /* Screening Variables for Explanations */
20 Solution_Path $\leftarrow \mathrm{LARS}(\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2^*))$ with varying $\lambda_1$
21 return Solution_Path
```

2. <u>*Modified ElasticNet Estimator*</u> - As shown in lines 13–18 of Algorithm 1, given the DNN model and the data point $\boldsymbol{x}$ for interpretation as well as the global interpretation $\boldsymbol{\beta}^{\mathrm{global}}$, $\mathcal{G}$-LIME first follows the standard operations of LIME – the algorithm first randomly interpolates the feature domains around $\boldsymbol{x}$ and obtain the random samples in the matrix $\boldsymbol{X}$, then $\mathcal{G}$-LIME tests every sample in $\boldsymbol{X}$ and restores the corresponding responses from the DNN in the vector $\boldsymbol{y}$. With $\boldsymbol{\beta}^{\mathrm{global}}$ as the global prior, random samples (around $\boldsymbol{x}$) in $\boldsymbol{X}$ and the DNN responses in $\boldsymbol{y}$, $\mathcal{G}$-LIME surrogates the local interpretation of the DNN model around $\boldsymbol{x}$ using a modified ElasticNet estimator $\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2)$ incorporating both ℓ_1 and ℓ_2 regularizers, to enjoy sparse and informative global priors in Bayesian linear regression.
3. <u>*Fast Feature Screening for Interpretation via LARS*</u> - As shown in lines 19–21 of Algorithm 1, with an appropriate tuning λ_2^* to ℓ_2-regularization as the input, $\mathcal{G}$-LIME uses LARS to output the solution path of the Modified ElasticNet – a series of $\boldsymbol{\beta}^{\mathcal{G}}(t)$ for an increasing t from 0^+ to a large positive number, which corresponds to $\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2^*)$ with a decreasing λ_1 from $\infty \to 0^+$. Then, $\mathcal{G}$-LIME screens features and ranks the features by their importance through a standard statistical variable selection procedure over the solution path [55,2]. The combination of the top ranked features is considered as the interpretation results.

In the following, we present the design and analysis of the modified ElasticNet estimator and fast screening algorithms.

3.2. Modified ElasticNet estimator

As was mentioned, with the input data point $\boldsymbol{x}$ and the DNN model for interpretation, $\mathcal{G}$-LIME follows the standard implementation of LIME and randomly interpolates feature domains around $\boldsymbol{x}$ to obtain the matrix of random samples $\boldsymbol{X}$ and the corresponding DNN outputs in $\boldsymbol{y}$. Then, with the global interpretation $\boldsymbol{\beta}^{\mathrm{global}}$ as the prior, $\mathcal{G}$-LIME pursues the local interpretation of the model around $\boldsymbol{x}$ via a modified ElasticNet estimator as follows

$$\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2) \leftarrow \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \left\|\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{global}}\right\|_2^2 , \tag{7}$$

where λ_1 and λ_2 are two tuning parameters to adjust the sparsity and the $\boldsymbol{\beta}^{\mathrm{global}}$-centered Gaussian prior accordingly.

3.3. Fast feature screening for explanations

To screen the possible local interpretations with varying strength of ℓ_1-regularization or using different subsets of feature selection in an extremely fast manner, $\mathcal{G}$-LIME adopts Least Angle Regression (LARS) [36] to retrieve the complete paths of coefficients. With an appropriate tuning parameter λ_2^* for ℓ_2-regularization, we rewrite the modified ElasticNet estimator into a ℓ_1-constrained quadratic minimization form, as follows,

$$\begin{aligned} \boldsymbol{\beta}^{\mathcal{G}}(t) \leftarrow \underset{\boldsymbol{\beta}}{\arg\min}\ & \boldsymbol{\beta}^{\top}\left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda_2^* I\right)\boldsymbol{\beta} - 2\left(\boldsymbol{X}^{\top}\boldsymbol{y} + \lambda_2^*\boldsymbol{\beta}^{\mathrm{global}}\right)^{\top}\boldsymbol{\beta}, \\ & s.t.\ \|\boldsymbol{\beta}\|_1 \le t\ . \end{aligned} \tag{8}$$

According to the Karush–Kuhn–Tucker conditions in ElasticNet estimator [35,56], for any $\lambda_1 > 0$, there exists a specific t such that $\boldsymbol{\beta}^{\mathcal{G}}(t) = \boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2^*)$. In this way, LARS can output a series of $\boldsymbol{\beta}^{\mathcal{G}}(t)$ with an increasing t from 0^+ to a large number, which are equivalent to the series of $\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2^*)$ with a decreasing λ_1 from $\infty \to 0^+$. In this way, we can screen the selection of features in the local interpretations obtained by $\mathcal{G}$-LIME.

Finally, we follow the standard operation proposed for ℓ_1-regularized solution path [55] to screen and rank the features by their importance – i.e., the order of every feature's linear regression coefficient turns to non-zero in the solution path of the modified ElasticNet with increasing t. The top ranked features are combined as the final interpretation result based on $\mathcal{G}$-LIME.

3.4. Algorithm analysis

We analyze the proposed algorithms from three perspectives as follows.

3.4.1. Interpretation as a Bayesian inference

As early as [57], ℓ_q-norm regularized least square estimators have been considered as performing Bayesian inference based on certain prior distributions, such as Gaussian prior for Ridge and Laplacian prior for Lasso. We here follow the Bayesian analysis in [35] to connect $\mathcal{G}$-LIME via the modified ElasticNet estimator with Bayes. Our analysis conditional on $\boldsymbol{X}$, which was obtained through random interpolation around the data point $\boldsymbol{x}$.

In this way, we can model the DNN outputs $\boldsymbol{y}$ as $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{\beta}$ refers to the random variable of the local interpretation result and σ^2 refers to the variance of random noises in observation. Such noises should be caused by the higher-order responses of DNNs under random interpolation in the feature domain, as the linear surrogate can only approximate the first-order derivative of the DNN output for the local interpretation. In this way, the square error term coincides the logarithm of the likelihood $\mathcal{L}(\boldsymbol{X}, \boldsymbol{y}, \sigma^2 | \boldsymbol{\beta})$ as follows,

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{y}, \sigma^2 | \boldsymbol{\beta}) \propto \exp\left(-\frac{\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2}{2\sigma^2}\right) . \tag{9}$$

Further, we could model the prior probability distribution of $\boldsymbol{\beta}$ based on the global interpretation $\boldsymbol{\beta}^{\mathrm{global}}$ as $\pi(\boldsymbol{\beta} | \boldsymbol{\beta}^{\mathrm{global}})$, such that

$$\pi(\boldsymbol{\beta} | \boldsymbol{\beta}^{\mathrm{global}}) \propto \exp\left(-C_1 \|\boldsymbol{\beta}\|_1 + C_2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{global}}\|_2^2\right) , \tag{10}$$

where C_1 and C_2 are two constants defining the Gaussian and Laplacian priors. Thus, with appropriate settings of λ_1 and λ_2 with respect to σ^2, C_1 and C_2, the estimator of $\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2)$ (defined in Eq (7)) is equivalent to the result of Bayesian inference that maximizes the posterior probability combining the likelihood term $\mathcal{L}(\boldsymbol{X}, \boldsymbol{y}, \sigma^2 | \boldsymbol{\beta})$ and the prior $\pi(\boldsymbol{\beta} | \boldsymbol{\beta}^{\mathrm{global}})$, as follows

$$\boldsymbol{\beta}^{\mathrm{mENet}}(\lambda_1, \lambda_2) := \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\arg\max} \left\{ \mathcal{L}(\boldsymbol{X}, \boldsymbol{y}, \sigma^2 | \boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta} | \boldsymbol{\beta}^{\mathrm{global}}) \right\} . \tag{11}$$

In this way, we can conclude that $\mathcal{G}$-LIME interpret the DNN around a sample through a Bayesian inference that models the interpretation result following a joint prior distribution of Laplacian and $\boldsymbol{\beta}^{\mathrm{global}}$-centered Gaussian.

3.4.2. Trade-off between local interpretation and global prior

The original LIME regularizes the estimation of local interpretation using a simple ℓ_2-norm regularizer. $\mathcal{G}$-LIME uses both ℓ_1 and ℓ_2 regularization to the estimation, while the tuning parameter λ_2 controls the balance between the local and global interpretations.

Note that the tuning parameter λ_2 is non-negative, and changing its value will produce different effects. When $\lambda_2 \to \infty$, the interpretation totally ignores the local individuality, reserving only the global information. When $\lambda_2 \to 0$, the interpretation leans towards a sparse approximation to the local approximation of LIME (depends on the strength of ℓ_1-regularization).

3.4.3. Algorithmic complexity

Compared to LIME, the increased computations of $\mathcal{G}$-LIME mainly comes from the global aggregations. Fortunately, the global aggregations need to be performed only *once* and can be prepared *offline* in advance. Furthermore, aggregating the interpretations from all the data points over the whole dataset is not always needed, especially when the dataset is large-scale. Alternatively, it is possible to aggregate a subset of data points for global information, so as to lower computation cost upon the budget limit.

In addition to the estimation of global priors, the computation of $\mathcal{G}$-LIME is passing one round of feature screening algorithm LARS (Equation (8)).

3.5. Adaptation of $\mathcal{G}$-LIME to computer vision (CV) tasks through clustering superpixels

In previous subsections, we have introduced core algorithms design and analysis of $\mathcal{G}$-LIME to explain the prediction results of multivariate models, where the inputs of the model are supposed to be structural data in $\mathbf{R}^d$, i.e., with d variables/features. However, for nonstructural datasets such as images, the number of dimensions for every sample depends on the size the image, while pixels or superpixels in different images could not be considered as distinct variables in a multivariate system with a fixed number of dimensions. Thus, for adaptation to CV tasks, there needs to first (1) *discover variables and formulate the linear surrogate models* for every image, and (2) *construct and reconfigure the global prior* subject to the number of variables.

3.5.1. Variable discovery and linear surrogate formulation

To discover variables (i.e., $\boldsymbol{X}$) for interpretations based on arbitrary image datasets, $\mathcal{G}$-LIME considers the common visual objects/patterns in the images as variables. Given any image, $\mathcal{G}$-LIME follows the default setting in LIME with QuickShift [58] for segmenting the image into superpixels, and sets the number of superpixels as the variable dimension d in the linear surrogate model for the interpretation. Then, $\mathcal{G}$-LIME generates N interpolations in d dimensions (denoted as $\boldsymbol{X}$) towards the given image and collect N responses (denoted as $\boldsymbol{y}$) from the DNN model accordingly. With $\boldsymbol{X}$ and $\boldsymbol{y}$, $\mathcal{G}$-LIME can either build the linear surrogate models for LIME listed in Eq. (1), or established the proposed modified ElasticNet estimator listed in Eq. (8) for advanced interpretation through incorporating with the global prior (introduced below).

Algorithm 2: Superpixel Clustering.

```
/* Initialization */
Input: $\boldsymbol{D}$ the dataset for global prior pursuit
Input: N_Clusters the number of superpixel clusters for global prior
Set $\boldsymbol{F}$ to an empty set
/* Superpixel and Feature Extraction */
for each image in $\boldsymbol{D}$ do
    $\mathcal{S} \leftarrow$ Segment(image)
    for each superpixel $s$ in $\mathcal{S}$ do
        $\boldsymbol{f} \leftarrow$ ComputeSuperpixelFeature($s$)
        $\boldsymbol{F} \leftarrow \boldsymbol{F} \cup \boldsymbol{f}$
    end
end
/* Feature Vectors Clustering */
Cluster_Centers $\leftarrow$ KMeans($\boldsymbol{F}$, N_Clusters)
return Cluster_Centers
```

3.5.2. Global prior construction and mapping

As was mentioned in Section 2.2, the global prior could be obtained through aggregating the local explanations based on the set of samples representing the whole population (e.g., ImageNet). However, the dimensions of the local explanations actually vary on the number of superpixels in images and it is difficult to aggregate vectors of different dimensions. In this way, as shown in Fig. 2, $\mathcal{G}$-LIME proposes to map local explanations of different dimensions to the global prior for $\mathcal{G}$-LIME in three steps as follows.

(a) *Superpixel Extraction and Clustering.* Given the superpixels extracted from every image in ImageNet, $\mathcal{G}$-LIME transforms every superpixel into a feature vector using the first layer output of ImageNet-pretrained models (i.e., ResNet-101 [37] or MobileNet [59]), and groups the feature vectors of all superpixels into K clusters using K-Means, where we set $K = 100$ in our experiments. The clustering on ImageNet would be more robust and less biased/overfit towards any specific target dataset. Moreover, this is an *off-line* step that needs to be done once. With the computed clusters, we can perform the following two steps.

(b) *Mapping and Merging Local Explanations for Global Prior.* Given the local explanation obtained by LIME for every image in the dataset, $\mathcal{G}$-LIME first maps the local explanation to a vector of K dimensions, where every dimension refers to a cluster of superpixels. Specifically, the algorithm assigns every dimension in the local explanation to one of the K dimensions through searching the nearest cluster center in feature vectors. Finally, with the K-dimension vector for every local explanation in the dataset, $\mathcal{G}$-LIME aggregates these K-dimension vectors using Algorithms in Section 2.2, to obtain the global prior (of K dimensions).

(c) *Mapping Global Prior for $\mathcal{G}$-LIME Explanation.* Given a new image for explanation, $\mathcal{G}$-LIME first screens the superpixels in the image (i.e., d superpixels), then maps the global prior (a K-dimension vector) to a d-dimension vector through matching every superpixel in the image to the K clusters via the nearest cluster center search.

Fig. 3 shows three clusters that show human-understandable semantics obtained from ImageNet. Each row presents one cluster with five images from ImageNet and five another from CUB-200-2011. The first row shows a kind of layer-over-layer pattern: in ImageNet, this pattern can be water waves, plaids in t-shirt or snow layers; in CUB, this pattern principles reside in wings. The second row obviously shows red colors. In fact, there are also black, yellow, blue colors in the obtained

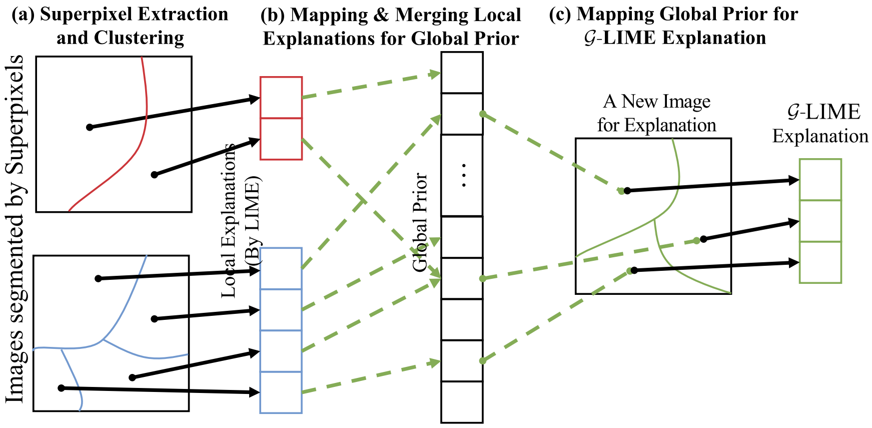

(a) Superpixel Extraction and Clustering
(b) Mapping & Merging Local Explanations for Global Prior
(c) Mapping Global Prior for $\mathcal{G}$-LIME Explanation
Images segmented by Superpixels
Local Explanations (By LIME)
Global Prior
A New Image for Explanation
$\mathcal{G}$-LIME Explanation

Fig. 2. Global Prior Construction and Mapping.

Table 1
Baseline algorithms and configurations. "**x**" (referring to the aggregator for Global Priors) could be "N" for NormLIME, "AVG" for Averaged-Importance, "SP" for LIME-SP, "H" for Homogeneity, as introduced in Subsection 2.2.

	Algorithms	Remarks
LIME	Feature Selection (LASSO) + Feature Importance (Ridge)	Vanilla LIME
LIME-Path	ElasticNet for Feature Selection and Importance Ranking	$\mathcal{G}$-LIME using Zero as Global Priors
$\mathcal{G}$-LIME-**x** (**Ours**)	Feature Selection (LASSO) + Feature Importance (Ridge with Global Prior)	With Global Prior in ℓ_2-penalty
$\mathcal{G}$-LIME-**x**-Path (**Ours**)	Modified ElasticNet	End-to-end Approach

clusters. The third rows show the patterns of grass, furs and earth. All of them exhibit that the clusters are obtained through low-level local features because features from the first layer of ResNet-101 are used for K-Means clustering.

In this way, the algorithm aggregates the local explanations of different dimensions into the global prior and maps it to a d-dimensional vector for $\mathcal{G}$-LIME explanation, according to the number of variables d desired. The source code is available at https://github.com/PaddlePaddle/InterpretDL, where the K-Means clusters can be automatically downloaded.

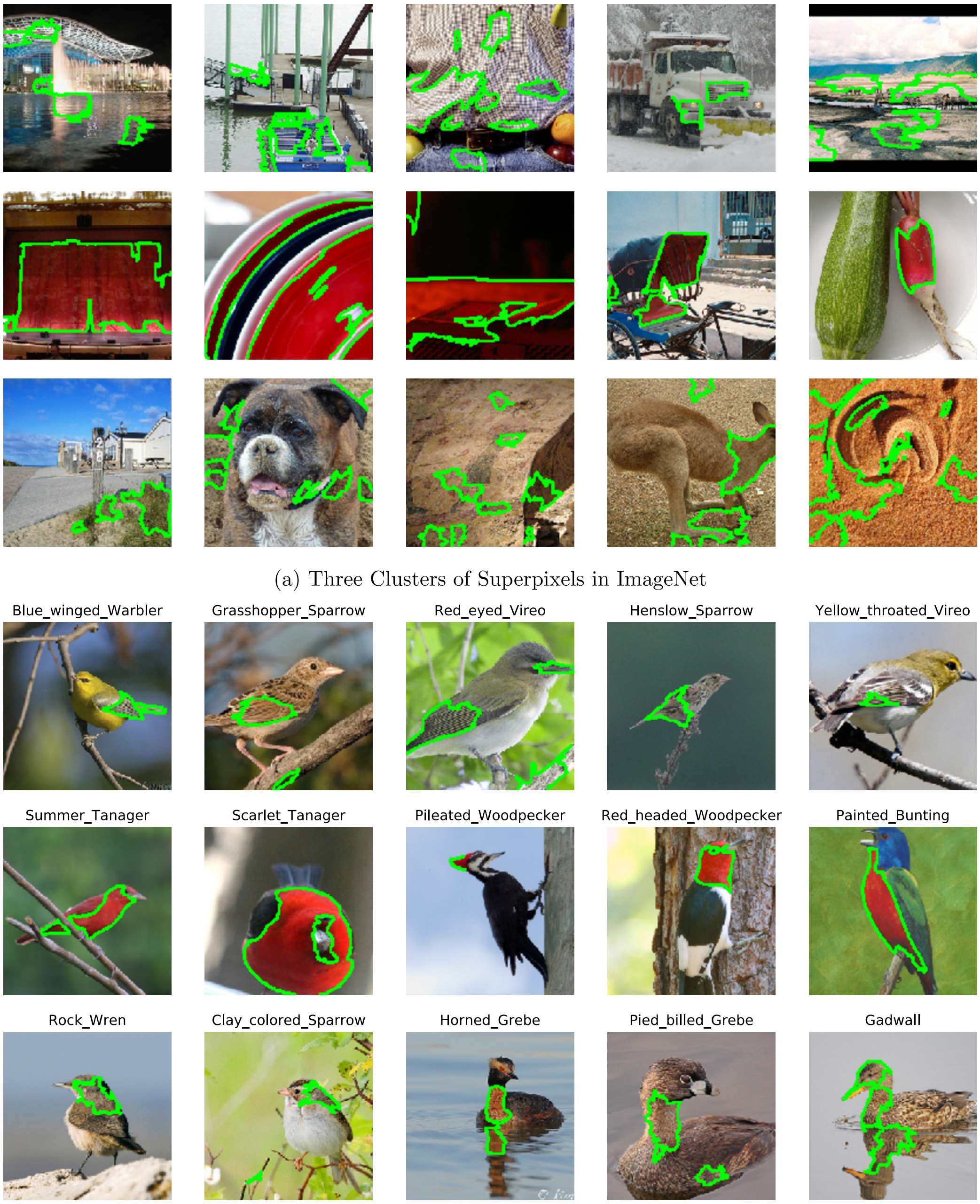
4. Experiments and analyses

In this section, we show the experimental results assessing the quality of $\mathcal{G}$-LIME explanations in comparison with vanilla LIME. Further ablation studies and applicability analyses are also provided.

4.1. Experiment setup

We describe the experiment setup to compare LIME and the proposed $\mathcal{G}$-LIME, where the global priors are obtained with the aggregation methods as introduced in Subsection 2.2. We conduct the evaluation experiments using the metrics presented in Subsection 2.3, on the CUB-200-2011 dataset [60] with the DNN model ResNet101 [37], introduced as follows. Note that more datasets and models are tested in the applicability analyses.

Datasets To validate the effectiveness of $\mathcal{G}$-LIME, we conduct experiments on CUB-200-2011 [60]. CUB-200-2011 contains images 200 categories of birds with 5990 and 5790 images for training and test. Each image contains only one birds and is annotated with both image-wise and pixel-wise labels by human experts. Note that we use the classification labels for training the deep neural networks to address the image classification task. In addition to computer vision tasks, we also evaluate $\mathcal{G}$-LIME using NLP and structural datasets and report results in Subsection 4.5.



(a) Three Clusters of Superpixels in ImageNet

(b) Three Clusters of Superpixels in CUB-200-2011

Fig. 3. Visualization of the same three clusters. Each row shows the same cluster with five images from ImageNet (top) and five images from CUB-200-2011 (bottom).

Deep neural network Regarding the networks, we carry out experiments and present the results in Subsection 4.2 and 4.4 using ResNet101 [37], AlexNet [39] and EfficientNet [38]— three most used architectures in practice. All these models have been initialized with the pre-trained weights on ImageNet [61] and then fine-tuned on CUB-200-2011 with classification labels, following the standard fine-tuning procedure. In addition to Convolutional Neural Networks, we also evaluate $\mathcal{G}$-LIME using Multi-Layer Perceptrons (MLPs) and NLP models, and report results in Subsection 4.5.

Nomenclature For clarity, we introduce the corresponding name and its math formula of each interpretation approach in Table 1. We recall the abbreviations that are used for the global priors: “N” for NormLIME, “AVG” for Averaged-Importance, “SP” for LIME-SP, “H” for Homogeneity, as introduced in Subsection 2.2.

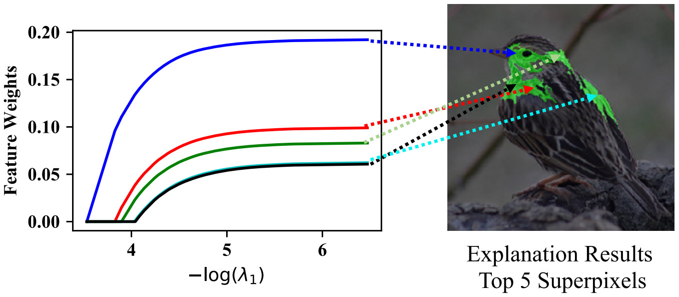
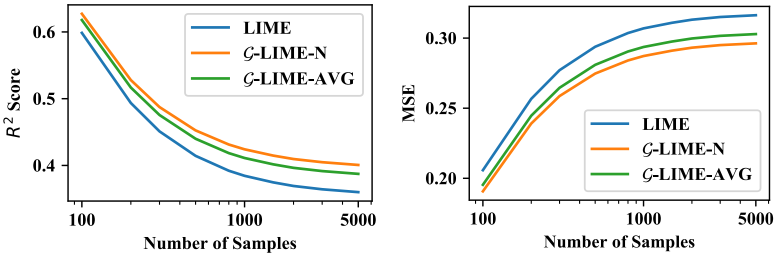


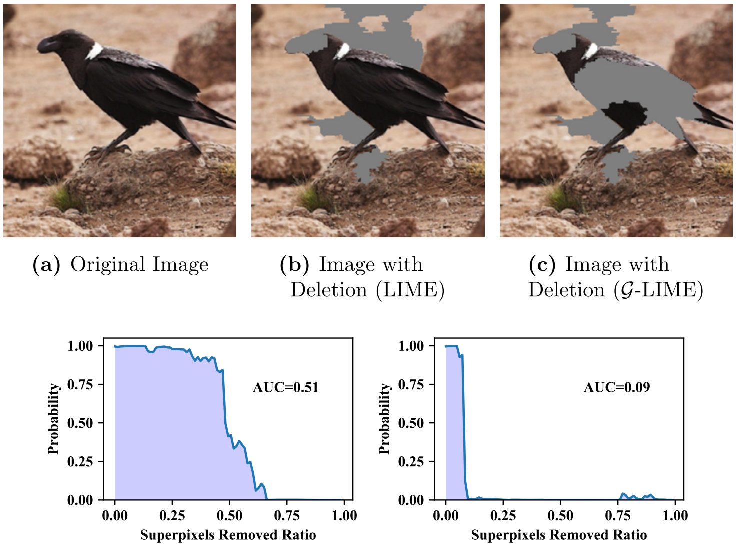
Fig. 4. Illustration of the top 5 features during the solution path of $\mathcal{G}$-LIME modified ElasticNet estimator and the corresponding to the superpixels in the visualization. Note that when λ_1 (controlling the sparsity) decreases, more features emerge and form the solution path.



(a) R^2 score ↑

(b) MSE ↓

Fig. 5. Local fidelity comparison, where higher is better for R^2 score, and lower is better for MSE.



(a) Original Image

(b) Image with Deletion (LIME)

(c) Image with Deletion ($\mathcal{G}$-LIME)

(d) Deletion (LIME)

(e) Deletion ($\mathcal{G}$-LIME)

Fig. 6. Illustration of an image with deletion evaluations on the LIME and $\mathcal{G}$-LIME explanations. We present the images during deletion in (b) and (c) according to the LIME and $\mathcal{G}$-LIME explanations respectively, where the top 5 superpixels are removed in both images. The probability change during deletion and the area under curve (AUC) are also recorded in (d) and (e). The overall comparison between LIME and $\mathcal{G}$-LIME is reported in Table 2.

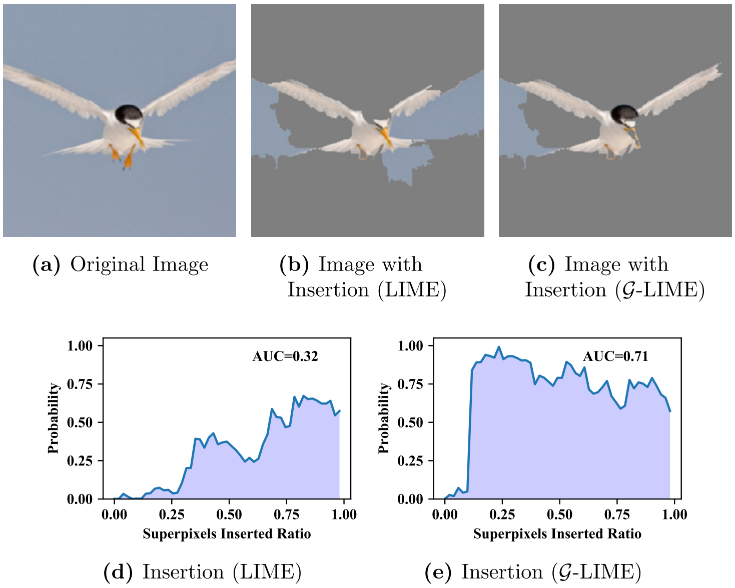
4.2. Experiment results

We first show a visual example in Fig. 4 that maps the solution path of $\mathcal{G}$-LIME to the visual explanation, for a better understanding of the $\mathcal{G}$-LIME explanation results. To evaluate the trustworthiness of $\mathcal{G}$-LIME, we compute the local fidelity [1,48], deletion/insertion [50] and stability [13,48].

Local fidelity Local fidelity measures the (dis)similarity between the surrogate linear model and the DNN model through R^2 scores and MSE, indicating the faithfulness of the surrogate model to the DNN model in a local region. Fig. 5 shows

Table 2
Comparison results of deletion evaluations (lower is better ↓) using images in the test set of CUB-200-2011. For comparison, global explanation algorithms GP-N and GP-AVG obtain 0.3724 and 0.3635 deletion scores respectively.

Explanation Method	Number of Samples		
	100	1000	3000
LIME	0.2470	0.1679	0.1581
LIME-Path	0.2451	0.1683	0.1574
$\mathcal{G}$-LIME-N	0.2256	0.1615	0.1544
$\mathcal{G}$-LIME-N-Path	0.2270	0.1624	0.1546
$\mathcal{G}$-LIME-AVG	**0.2186**	**0.1586**	**0.1524**
$\mathcal{G}$-LIME-AVG-Path	0.2249	0.1597	0.1527


(a) Original Image
(b) Image with Insertion (LIME)
(c) Image with Insertion ($\mathcal{G}$-LIME)

(d) Insertion (LIME)
(e) Insertion ($\mathcal{G}$-LIME)

Fig. 7. Illustration of an image with insertion evaluations on the LIME and $\mathcal{G}$-LIME explanations. We present the images during deletion in (b) and (c) according to the LIME and $\mathcal{G}$-LIME explanations respectively, where the top 5 superpixels are inserted in both images. The probability change during insertion and the area under curve (AUC) are also recorded in (d) and (e). The overall comparison between LIME and $\mathcal{G}$-LIME is reported in Table 3.

Table 3
Comparison results of insertion evaluations (higher is better ↑) using images in the test set of CUB-200-2011. For comparison, global explanation algorithms GP-N and GP-AVG obtain 0.3049 and 0.3143 insertion scores respectively.

Explanation Method	Number of Samples		
	100	1000	3000
LIME	0.3665	0.5415	0.5692
LIME-Path	0.3782	0.5457	0.5731
$\mathcal{G}$-LIME-N	0.4074	0.5655	0.5843
$\mathcal{G}$-LIME-N-Path	0.4085	0.5617	0.5825
$\mathcal{G}$-LIME-AVG	**0.4093**	**0.5671**	**0.5846**
$\mathcal{G}$-LIME-AVG-Path	0.4029	0.5635	0.5831

the results of local fidelity for comparing LIME and $\mathcal{G}$-LIME. Both metrics are averaged over all images and measured with varying the number of generated data points. We can observe a clear gap between LIME and $\mathcal{G}$-LIME from Fig. 5 with both metrics favoring $\mathcal{G}$-LIME, and different global aggregation methods do not yield large differences. The local fidelity evaluation validates that $\mathcal{G}$-LIME is locally more faithful to model's behaviors than LIME. We will show in the ablation studies that $\mathcal{G}$-LIME with any informative global prior yields more faithful explanations than vanilla LIME.

Deletion/insertion Linear models used by LIME and $\mathcal{G}$-LIME are incapable to fully and globally explain deep models. For further supporting the trustworthiness of $\mathcal{G}$-LIME, we conduct evaluation experiments of measuring the deletion and in-

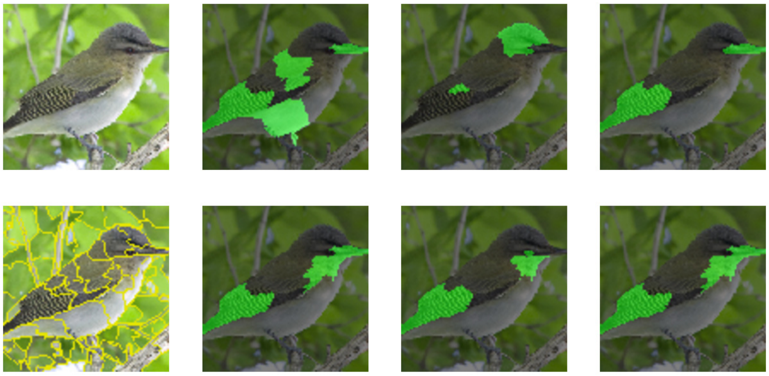


Fig. 8. Stability investigation for LIME and $\mathcal{G}$-LIME explanations using a small number of samples (100). We repeat the random sampling process five times (corresponding to the right five columns) and obtain the explanations of LIME (upper row) and $\mathcal{G}$-LIME (bottom row). For clarity, we highlight the three most important superpixels given by explanations, which coincide to the visual segmentation ground truth [62]. Visually, the $\mathcal{G}$-LIME explanations are quite stable. We report the quantitative comparisons in Table 4.

Table 4
Stability evaluations (lower is better ↓) for LIME and $\mathcal{G}$-LIME explanations using small numbers of samples (≤ 1000), measured by the stability metric following [13,48]. We repeat the random sampling process five times and report the average results.

Explanation Method	Number of Samples			
	100	200	500	1000
LIME	0.7404	0.2204	0.0404	0.0121
$\mathcal{G}$-LIME-N	**0.7027**	**0.2095**	**0.0382**	**0.0112**
$\mathcal{G}$-LIME-AVG	0.7193	0.2141	0.0390	0.0114
$\mathcal{G}$-LIME-SP	0.7302	0.2172	0.0395	0.0115
$\mathcal{G}$-LIME-H	0.7316	0.2177	0.0396	0.0116

sertion scores [50]. Fig. 6 illustrates the deletion evaluation process and results, where the probability for the ground truth class quickly decays with $\mathcal{G}$-LIME explanation results, indicating that the important features are found and ranked high by $\mathcal{G}$-LIME. Similar processes are performed for insertion evaluations, see Fig. 7. We report the overall results of deletion and insertion evaluations in Table 2 and Table 3 respectively. Global explanation results are also reported in the captions. The deletion and insertion scores are generally coherent, and the scores achieved by $\mathcal{G}$-LIME are clearly better than LIME does. Meanwhile, $\mathcal{G}$-LIME-Path gets very similar results to $\mathcal{G}$-LIME and thus we mainly report the results of $\mathcal{G}$-LIME later.

Stability evaluation We show in Fig. 8 the (in)consistency of LIME and $\mathcal{G}$-LIME explanations respectively when the number of generated samples is small. Following the stability metric [13,48], we compute the explanation changes between similar examples, measured by the expectation of ℓ_2 distance over the neighborhood, so as to quantitatively validate $\mathcal{G}$-LIME. We report the numerical experimental results in Table 4, showing that $\mathcal{G}$-LIME gets better stability scores than LIME, while the difference becomes very small when the number of samples increases.

4.3. Comparison with prevailing and SOTA interpretability methods

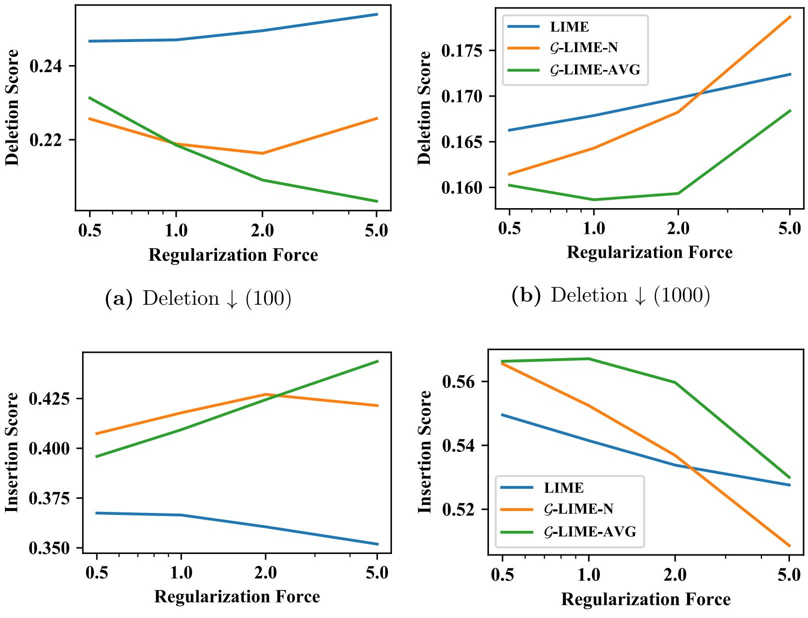
In the previous subsection, we presented the experiment results of $\mathcal{G}$-LIME and the vanilla LIME for demonstrating the direct improvements over LIME, with full evaluation metrics, including local fidelity, deletion/insertion scores and stability. In this subsection, we show the comparison across prevailing and state-of-the-art interpretability methods, e.g., IG [20], SmoothGrad [19], Grad-CAM [18] and Anchors [41]. We have also discussed other explanation algorithms [42,43,16,14,40] in Section 2.1.

While the local fidelity and stability evaluations are only available for surrogate model based methods, such as LIME, we compare these algorithms with $\mathcal{G}$-LIME using deletion/insertion scores. Experiment setups are the same as previous, with a fine-tuned ResNet101 on CUB-200-2011. From the previous results, we generate 3000 neighbor points to compute the LIME and $\mathcal{G}$-LIME explanations. For straightforward comparisons, we also report the metric named the area between perturbation curves (ABPC) [63], which computes the difference between insertion and deletion scores for each sample. We also average the results of 1000 samples from the test set of CUB-200-2011. Experiment results are shown in Table 5.

We notice from the results that pixel-level explanations (IG and SmoothGrad) get very low deletion scores but low insertion scores too. Low deletion scores may benefit from the fine-grained level of explanations. Low insertion scores indicate that they are not able to cover all the important features. Grad-CAM, however, gets very high insertion score while the deletion score is the worst among those algorithms. This is probably caused by the low-resolution explanations produced

Table 5
Comparison with other explanation algorithms, measured in deletion/insertion scores and area between perturbation curves (ABPC) score. Each method has been repeated three times except Grad-CAM and Anchors, because Grad-CAM does not involve any randomness and that Anchors has been done once for its slow convergence.

	Deletion ↓	Insertion ↑	ABPC ↑
Grad-CAM [18]	0.1706	0.6423	0.4445
Anchors [41]	0.2418	0.5836	0.3073
IG [20]	**0.0605** ± 0.0001	0.5004 ± 0.0003	0.4492 ± 0.0003
SmoothGrad [19]	0.0762 ± 0.0001	0.4649 ± 0.0020	0.3954 ± 0.0022
LIME [1]	0.1501 ± 0.0010	0.5762 ± 0.0019	0.4061 ± 0.0010
$\mathcal{G}$-LIME-N (**Ours**)	0.1253 ± 0.0017	0.6347 ± 0.0027	0.4998 ± 0.0039
$\mathcal{G}$-LIME-AVG (**Ours**)	0.1230 ± 0.0022	**0.6539** ± 0.0023	**0.5155** ± 0.0019



(a) Deletion ↓ (100)

(b) Deletion ↓ (1000)

(c) Insertion ↑ (100)

(d) Insertion ↑ (1000)

Fig. 9. Deletion/Insertion evaluation scores with varying the regularization effects for ℓ_2 norm (LIME) and distance to global prior ($\mathcal{G}$-LIME). We present results of two cases where the number of samples is 100 for (a,c) and 1000 for (b,d), and two global priors, i.e., NormLIME [11] and Average [10].

by Grad-CAM, which usually depends on the end layer's feature map. Anchors, with the objective of high coverage of important features, get higher insertion scores than LIME. The two variants of $\mathcal{G}$-LIME get a reasonable deletion score and the highest insertion scores among all algorithms. More importantly, as for the ABPC score, $\mathcal{G}$-LIME shows clear improvements over others, including fine-grained explanations.

4.4. Ablation studies

We present the ablation studies on the choice of ℓ_2 regularization effects λ_2, as well as the choice of global priors. Note that the number of generated samples for computing explanation results is also an important hyper-parameter. However, in practice, we find that the evaluation metric score is always better when the number of samples is higher, and we have reported the results using different numbers of samples in most cases, so here we do not study this essential factor that is clear and presented.

Regularization effects The ℓ_2 regularization effect controls the norm of linear weights for LIME and the distance to the global prior for $\mathcal{G}$-LIME. Fig. 9 shows the evolution of varying the regularization effects from 0.5 to 5.0, where $\mathcal{G}$-LIME is generally better than LIME as previously presented. More important, we find that when the number of samples is small (100), the regularization effect favors a large value; when the number of samples is relatively large (1000), it is better to choose a relatively small value. It aligns with the pattern of regularization usage, but $\mathcal{G}$-LIME benefits more from the introduction of global priors, where in a large range of regularization effects, $\mathcal{G}$-LIME gives better scores than LIME does. Furthermore, according our experiments across datasets, the conventional value (1.0) can be safely used as regularization force for an

Table 6
Comparison among LIME and four variants of $\mathcal{G}$-LIME using different global priors. The results reported here are using 1000 generated samples to compute the explanations with a conventional regularization effect (i.e., 1.0).

	R^2 score ↑	MSE ↓	Deletion ↓	Insertion ↑
LIME	0.3845	0.2409	0.1679	0.5415
$\mathcal{G}$-LIME-N	**0.4242**	**0.2254**	0.1615	0.5655
$\mathcal{G}$-LIME-AVG	0.4111	0.2305	**0.1586**	**0.5671**
$\mathcal{G}$-LIME-SP	0.4023	0.2340	0.1615	0.5510
$\mathcal{G}$-LIME-H	0.4010	0.2345	0.1662	0.5413

Table 7
Evaluation of LIME and $\mathcal{G}$-LIME when explaining the prediction results of AlexNet [39] and EfficientNet [38].

Network	Explanation Method	R^2 score ↑	MSE ↓	Deletion ↓	Insertion ↑
AlexNet	LIME	0.3308	0.3273	0.2898	0.3470
	$\mathcal{G}$-LIME-N	**0.3483**	**0.3186**	0.2591	0.3612
	$\mathcal{G}$-LIME-AVG	0.3421	0.3217	**0.2559**	**0.3641**
EfficientNet	LIME	0.3061	0.3437	0.2531	0.5514
	$\mathcal{G}$-LIME-N	**0.3401**	**0.3267**	0.2234	0.5679
	$\mathcal{G}$-LIME-AVG	0.3269	0.3334	**0.2155**	**0.5872**

appropriate global prior. Both NormLIME [11] and Average [10] are good choices as global priors over the vanilla LIME. We will study the difference among the global priors in the following paragraph.

Global priors The global aggregation of local explanations aims to approximate a global explanation for the model. While most global aggregations are meaningful, they lead to various priors for our proposed $\mathcal{G}$-LIME, where NormLIME [11] and Average [10] are more appropriate in this framework. In most cases, each of the four priors introduced in Subsection 2.2 improves the explanation quality (local fidelity, deletion/insertion scores, and stability) over the vanilla LIME. We show the results in Table 6. We recall that the results of stability have been reported in Table 4.

While the studied four global priors have subtle difference, they all show the effectiveness of $\mathcal{G}$-LIME, integrating global priors into the local ones within a Bayesian framework. We also find that $\mathcal{G}$-LIME with NormLIME as global prior obtains the best local fidelity (as well as the stability) and $\mathcal{G}$-LIME with Average gets the best deletion/insertion scores.

4.5. Evaluations for applicability analyses

In addition to ResNet-101, we conduct experiments using AlexNet [39], a standard convolutional network, and EfficientNet [38], an automatically searched network, to evaluate $\mathcal{G}$-LIME explanations on CUB-200-2011. We report the results of local fidelity and deletion/insertion scores for these two networks in Table 7, where $\mathcal{G}$-LIME improves all the metrics.

We further explore the applicability of $\mathcal{G}$-LIME on two other visual datasets, StanfordCars [64] and OxfordFlowers [65], which focus on images of car models and flowers respectively. For each of them, we fine-tune the ImageNet-pretrained ResNet-101 on the training set, and evaluate LIME and $\mathcal{G}$-LIME explanations of the fine-tuned model on the validation set. Results on StanfordCars and OxfordFlowers are shown in the first two blocks of Table 8, which validate the improvements from $\mathcal{G}$-LIME. Experiments on the visual datasets, including Table 7, also confirm the effectiveness and efficiency of the superpixel clustering process for global priors.

Moreover, we extend the application of $\mathcal{G}$-LIME to the structural and text datasets. GiveMeSomeCredits[2] is a structural dataset for building a model that helps banks to make the best financial decisions, where we trained an MLP model for evaluating our explanation methods. IMDb movie review [66] is a text dataset for binary sentiment classification, where we trained an LSTM classifier with Glove [67] embeddings and performed the evaluation experiments. Note that for structural and text datasets, clustering process is not needed to generate global priors because the features or words are naturally shared by all data samples. We use the same explanation evaluation methods introduced in Subsection 4.2 and show the results in Table 8. In all datasets and tasks, the local fidelity (both R^2 score and MSE) favors $\mathcal{G}$-LIME over LIME, either using NormLIME or Averaged-Importance as global prior for $\mathcal{G}$-LIME. Similar results have been obtained for deletion/insertion scores.

[2] https://www.kaggle.com/c/GiveMeSomeCredit.

Table 8
Evaluation of LIME and $\mathcal{G}$-LIME on various datasets and tasks. Cars [64] and Flowers [65] are of visual data, Credits (GiveMeSomeCredits) is of structural data, IMDb [66] is of textual data.

Dataset	Explanation Method	R^2 score ↑	MSE ↓	Deletion ↓	Insertion ↑
Cars	LIME	0.3846	0.3039	0.2460	0.2860
	$\mathcal{G}$-LIME-N	**0.4294**	**0.2817**	0.2373	0.2846
	$\mathcal{G}$-LIME-AVG	0.4144	0.2891	**0.2230**	**0.2924**
Flowers	LIME	0.4207	0.2784	0.4132	0.3600
	$\mathcal{G}$-LIME-N	**0.4633**	**0.2581**	0.4104	0.3782
	$\mathcal{G}$-LIME-AVG	0.4539	0.2626	**0.4035**	**0.3807**
Credits	LIME	0.2595	0.3867	0.1501	0.0945
	$\mathcal{G}$-LIME-N	**0.3066**	**0.2843**	0.1444	0.1124
	$\mathcal{G}$-LIME-AVG	0.2783	0.2961	**0.1425**	**0.1148**
IMDb	LIME	0.5382	0.2290	0.5036	0.2111
	$\mathcal{G}$-LIME-N	**0.5942**	**0.2020**	**0.4951**	0.2172
	$\mathcal{G}$-LIME-AVG	0.5842	0.2082	0.4960	**0.2173**

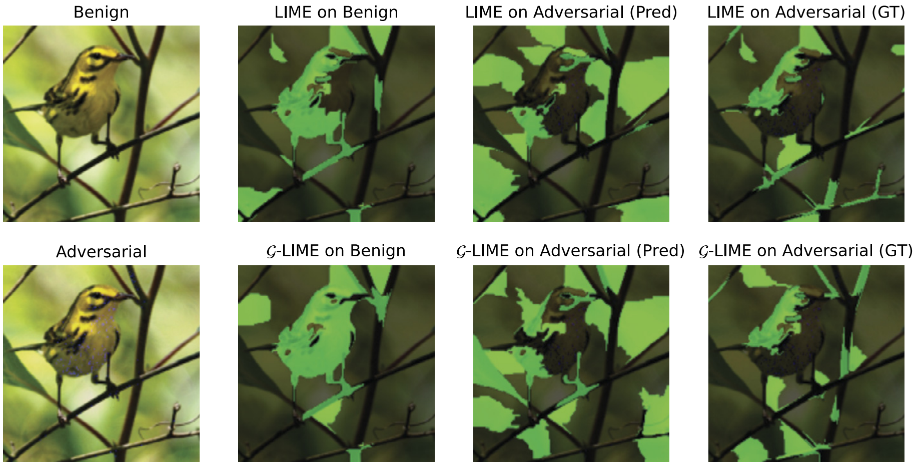

Benign
LIME on Benign
LIME on Adversarial (Pred)
LIME on Adversarial (GT)
Adversarial
$\mathcal{G}$-LIME on Benign
$\mathcal{G}$-LIME on Adversarial (Pred)
$\mathcal{G}$-LIME on Adversarial (GT)

Fig. 10. LIME and $\mathcal{G}$-LIME explanations on benign and adversarial examples. The benign one is correctly and confidently predicted by the model. When explaining the adversarial example, we show both the results using the predicted class (the wrong one fooled by the adversarial attack) and the ground truth class.

4.6. Analyses on adversarial examples

Recent works show that adversarial robustness is shown to be related to model interpretability [68–70], and that explanations can be used to detect adversarial attacks [71–74]. Here we provide some results and analyses on the adversarial examples using LIME and $\mathcal{G}$-LIME.

By using the projected gradient descent as the attacker (with the default hyperparameters used in [75]), the model was completely fooled by the adversarial examples. In such scenario, we visualize LIME and $\mathcal{G}$-LIME explanations on the benign and adversarial examples. As shown in Fig. 10, we observe that the adversarial attacks affect a lot the explanation results of both LIME and $\mathcal{G}$-LIME. Both of them are not able to locate the object in the image, and thus would have similar performances if used to detect adversarial examples. Here $\mathcal{G}$-LIME is no longer predominant and we argue that this is because the global prior does not provide information on the adversarial examples and the sparsity is affected by the attack too.

4.7. Discussions and limitations

$\mathcal{G}$-LIME shares many of the same assumptions as LIME including a linear surrogate around perturbations of a local data point, and assumes local explanations should be aligned with the global one. Though, this seems to be a reasonable assumption in practice, it still lacks of theoretical supports. Second, $\mathcal{G}$-LIME still suffers from the randomness as LIME, while the variance is significantly reduced thanks to the integration of sparsity and global priors. In addition, compared to other post-hoc local explanations, $\mathcal{G}$-LIME requires generating global explanations from a number of local data points (albeit this can be cached). Last, faced with adversarial examples, $\mathcal{G}$-LIME does not show any advantages over the vanilla LIME.

5. Conclusion

In this paper, we present $\mathcal{G}$-LIME – a local interpretation algorithm that explains the prediction results of DNN models for given samples, using sparse and informative global priors. Specifically, with a dataset representing the population of samples (e.g., the training set), $\mathcal{G}$-LIME first pursues the global explanation of the DNN model using the whole dataset. Given a new data point and the prediction result of a DNN model for interpretation, $\mathcal{G}$-LIME proposes a modified ElasticNet [35] to regularize the estimator of local explanations, using (1) ℓ_1-regularization for sparsity and (2) ℓ_2-regularization centered at the global explanation to provide informative global priors. Least Angle Regression (LARs) has been used to rank the importance of every feature in the explanation over the path of ℓ_1-regularization. Extensive experiments have been done to demonstrate the advantages of $\mathcal{G}$-LIME on a wide range of DNN models for computer visions, NLP, and structural datasets.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Haoyi Xiong reports financial support was provided by National Key R&D Program of China (No. 2021ZD0110303).

Acknowledgement

The first two authors of this work were supported in part by National Key R&D Program of China under the grant No. 2021ZD0110303.

References

[1] Marco Tulio Ribeiro, Sameer Singh, 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, 2016, pp. 1135–1144.
[2] Vivekananda Roy, Sounak Chakraborty, et al., Selection of tuning parameters, solution paths and standard errors for bayesian lassos, Bayesian Anal. 12 (3) (2017) 753–778.
[3] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[4] Cynthia Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.
[5] Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models, arXiv preprint, arXiv:1708.08296, 2017.
[6] Filip Karlo Došilović, Mario Brčić, Nikica Hlupić, Explainable artificial intelligence: a survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.
[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al., Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[8] Xuan Liu, Xiaoguang Wang, Stan Matwin, Improving the interpretability of deep neural networks with knowledge distillation, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2018, pp. 905–912.
[9] David Alvarez-Melis, Tommi S. Jaakkola, On the robustness of interpretability methods, arXiv preprint, arXiv:1806.08049, 2018.
[10] Ilse van der Linden, Hinda Haned, Evangelos Kanoulas, Global aggregations of local explanations for black box models, in: Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval at ACM SIGIR, 2019.
[11] Isaac Ahern, Adam Noack, Luis Guzman-Nateras, Dejing Dou, Boyang Li, Jun Huan, Normlime: a new feature importance metric for explaining deep neural networks, arXiv preprint, arXiv:1909.04200, 2019.
[12] Kjersti Aas, Martin Jullum, Anders Løland, Explaining individual predictions when features are dependent: more accurate approximations to Shapley values, Artif. Intell. 298 (2021) 103502.
[13] David Alvarez Melis, Tommi Jaakkola, Towards robust interpretability with self-explaining neural networks, in: Advances in Neural Information Processing Systems, 2018, pp. 7775–7784.
[14] Muhammad Rehman Zafar, Naimul Mefraz Khan, Dlime: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems, arXiv preprint, arXiv:1906.10263, 2019.
[15] Naman Bansal, Chirag Agarwal, Anh Nguyen, Sam: the sensitivity of attribution methods to hyperparameters, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8673–8683.
[16] Zhengze Zhou, Giles Hooker, Fei Wang, S-lime: stabilized-lime for model explanation, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 2429–2438.
[17] Mark Ibrahim, Melissa Louie, Ceena Modarres, John Paisley, Global explanations of neural networks: mapping the landscape of predictions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 279–287.
[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359.
[19] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, Martin Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint, arXiv:1706.03825, 2017.
[20] Mukund Sundararajan, Ankur Taly, Qiqi Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning (ICML), 2017.
[21] Thiago Serra, Christian Tjandraatmadja, Srikumar Ramalingam, Bounding and counting linear regions of deep neural networks, in: International Conference on Machine Learning, PMLR, 2018, pp. 4558–4566.
[22] Thiago Serra, Srikumar Ramalingam, Empirical bounds on linear regions of deep rectifier networks, in: AAAI, 2020, pp. 5628–5635.
[23] Xiao Zhang, Dongrui Wu, Empirical studies on the properties of linear regions in deep neural networks, in: International Conference on Learning Representations (ICLR), 2020.
[24] Robert Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B 58 (1) (1996) 267–288.
[25] William J. Welch, Algorithmic complexity: three np-hard problems in computational statistics, J. Stat. Comput. Simul. 15 (1) (1982) 17–25.

[26] Paolo Victor, T. Redondo, Optimal variable subset selection problem in regression analysis is np-complete, Philipp. Stat. 68 (1) (2019) 41–50.

[27] Niall Hurley, Scott Rickard, Comparing measures of sparsity, IEEE Trans. Inf. Theory 55 (10) (2009) 4723–4741.

[28] Sarath Sreedharan, Tathagata Chakraborti, Subbarao Kambhampati, Foundations of explanations as model reconciliation, Artif. Intell. 301 (2021) 103558.

[29] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, Mark Neerincx, Evaluating XAI: a comparison of rule-based and example-based explanations, Artif. Intell. 291 (2021) 103404.

[30] Richard Evans, Matko Bošnjak, Lars Buesing, Kevin Ellis, David Pfau, Pushmeet Kohli, Marek Sergot, Making sense of raw input, Artif. Intell. 299 (2021) 103521.

[31] Daniela M. Witten, Robert Tibshirani, Covariance-regularized regression and classification for high dimensional problems, J. R. Stat. Soc., Ser. B, Stat. Methodol. 71 (3) (2009) 615–636.

[32] Matthias W. Seeger, Bayesian inference and optimal design for the sparse linear model, J. Mach. Learn. Res. 9 (Apr 2008) 759–813.

[33] Ryan Tibshirani, Modern regression 1: ridge regression, in: Data Mining, vol. 36, 2013, pp. 462–662.

[34] Charles Dossal, Marie-Line Chabanol, Gabriel Peyré, Jalal Fadili, Sharp support recovery from noisy random measurements by ℓ_1-minimization, Appl. Comput. Harmon. Anal. 33 (1) (2012) 24–43.

[35] Hui Zou, Trevor Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc., Ser. B, Stat. Methodol. 67 (2) (2005) 301–320.

[36] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al., Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[38] Mingxing Tan, Quoc V. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning (ICML), 2019.

[39] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), 2012.

[40] Sharath M. Shankaranarayana, Davor Runje, Alime: autoencoder based approach for local interpretability, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2019, pp. 454–463.

[41] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, Anchors: high-precision model-agnostic explanations, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[42] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, Madeleine Udell, “Why should you trust my explanation?” understanding uncertainty in lime explanations, arXiv preprint, arXiv:1904.12991, 2019.

[43] Giorgio Visani, Enrico Bagli, Federico Chesani, Optilime: optimized lime explanations for diagnostic computer algorithms, arXiv preprint, arXiv:2006.05714, 2020.

[44] Scott M. Lundberg, Su-In Lee, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, vol. 30, 2017.

[45] Mengnan Du, Ninghao Liu, Xia Hu, Techniques for interpretable machine learning, Commun. ACM 63 (1) (2019) 68–77.

[46] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, Dejing Dou, Interpretable deep learning: interpretations, interpretability, trustworthiness, and beyond, arXiv preprint, arXiv:2103.10689, 2021.

[47] Finale Doshi-Velez, Been Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint, arXiv:1702.08608, 2017.

[48] Gregory Plumb, Maruan Al-Shedivat, Angel Alexander Cabrera, Adam Perer, Eric Xing, Ameet Talwalkar, Regularizing black-box models for improved interpretability, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.

[49] Alon Jacovi, Yoav Goldberg, Towards faithfully interpretable nlp systems: how should we define and evaluate faithfulness?, arXiv preprint, arXiv:2004.03685, 2020.

[50] Vitali Petsiuk, Abir Das, Kate Saenko, Rise: randomized input sampling for explanation of black-box models, arXiv preprint, arXiv:1806.07421, 2018.

[51] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, Been Kim, A benchmark for interpretability methods in deep neural networks, in: Advances in Neural Information Processing Systems, 2019, pp. 9737–9748.

[52] Amirata Ghorbani, Abubakar Abid, James Zou, Interpretation of neural networks is fragile, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3681–3688.

[53] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, Pradeep Ravikumar, On the (in) fidelity and sensitivity for explanations, arXiv preprint arXiv:1901.09392, 2019.

[54] Mengjiao Yang, Been Kim, Benchmarking attribution methods with relative feature importance, arXiv, 2019, arXiv–1907.

[55] Mee Young Park, Trevor Hastie, ℓ_1-regularization path algorithm for generalized linear models, J. R. Stat. Soc., Ser. B, Stat. Methodol. 69 (4) (2007) 659–677.

[56] Stephen Boyd, Stephen P. Boyd, Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[57] Wenjiang J. Fu, Penalized regressions: the bridge versus the lasso, J. Comput. Graph. Stat. 7 (3) (1998) 397–416.

[58] Andrea Vedaldi, Stefano Soatto, Quick shift and kernel methods for mode seeking, in: European Conference on Computer Vision, Springer, 2008, pp. 705–718.

[59] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[60] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-UCSD birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[61] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Fei-Fei Li, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[62] Riccardo Guidotti, Evaluating local explanation methods on ground truth, Artif. Intell. 291 (2021) 103428.

[63] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2016) 2660–2673.

[64] Jonathan Krause, Michael Stark, Jia Deng, Fei-Fei Li, 3D object representations for fine-grained categorization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 554–561.

[65] Maria-Elena Nilsback, Andrew Zisserman, Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE, 2008, pp. 722–729.

[66] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, Christopher Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, Association for Computational Linguistics, June 2011, pp. 142–150.

[67] Jeffrey Pennington, Richard Socher, Christopher D. Manning, Glove: global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[68] Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, Somesh Jha, Robust attribution regularization, Adv. Neural Inf. Process. Syst. 32 (2019).

[69] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, Luca Daniel, Proper network interpretability helps adversarial robustness in classification, in: International Conference on Machine Learning, PMLR, 2020, pp. 1014–1023.

[70] Adam Noack, Isaac Ahern, Dejing Dou, Boyang Li, An empirical study on the relation between network interpretability and adversarial robustness, SN Comput. Sci. 2 (1) (2021) 1–13.

[71] Guanhong Tao, Shiqing Ma, Yingqi Liu, Xiangyu Zhang, Attacks meet interpretability: attribute-steered detection of adversarial samples, Adv. Neural Inf. Process. Syst. 31 (2018).

[72] Tianyu Pang, Chao Du, Yinpeng Dong, Jun Zhu, Towards robust detection of adversarial examples, Adv. Neural Inf. Process. Syst. 31 (2018).

[73] Alexey Ignatiev, Nina Narodytska, Joao Marques-Silva, On relating explanations and adversarial examples, Adv. Neural Inf. Process. Syst. 32 (2019).

[74] Gihyuk Ko, Gyumin Lim, Unsupervised detection of adversarial examples with model explanations, arXiv preprint, arXiv:2107.10480, 2021.

[75] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint, arXiv:1706.06083, 2017.