

---

# Relational Local Explanations

---

Vadim Borisov\*  
University of Tübingen

Gjergji Kasneci  
University of Tübingen

## Abstract

The majority of existing post-hoc explanation approaches for machine learning models produce independent per-variable feature attribution scores, ignoring a critical characteristic, such as the inter-variable relationship between features that naturally occurs in visual and textual data. In response, we develop a novel model-agnostic and permutation-based feature attribution algorithm based on the relational analysis between input variables. As a result, we are able to gain a broader insight into machine learning model decisions and data. This type of local explanation measures the effects of interrelationships between local features, which provides another critical aspect of explanations. Experimental evaluations of our framework using setups involving both image and text data modalities demonstrate its effectiveness and validity.

## 1 Introduction

The increasing reliance on machine learning (ML) models in various domains of our daily life has brought a need for explaining the inner workings and decision-making processes of these models [1]. This is particularly relevant for deep convolutional neural networks (CNNs) in visual domains, which have demonstrated superior performance on tasks such as object detection [2], segmentation [3], and classification [4]. Also, in the natural language processing (NLP) domain, self-attention models [5], specifically deep Transformer-based models, have achieved state-of-the-art results on tasks such as text summarization [6] and sentiment analysis [7].

As a result, it is necessary to have confidence that black-box ML models are functioning as intended, and explanations that include *inter-variable relational information* can help achieve this. Moreover, the interpretability of ML models is a vital aspect for numerous applications, particularly those involving life-critical uses such as healthcare and autonomous driving [8, 9].

Furthermore, in accordance with the General Data Protection Regulation (GDPR) [10] and California Consumer Privacy Act (CCPA) [11], it is essential for real-world applications to not only provide accurate and reliable predictions but also to provide transparent and easily understood explanations for automated decision systems. Additionally, there is a practical need for model-agnostic feature methods that can be used with any machine learning system. Last but not least, from a practical industrial perspective, there is a need for model-agnostic feature methods which can work with any ML system [12].

**Motivation.** Although numerous feature attribution approaches exist, the vast majority of them work with each input variable *independently*, ignoring the crucial property such as the relationship that intrinsically exists in the many homogeneous data formats such as visual and textual.

Another issue with the state-of-the-art feature attribution approaches is that many local explanation methods “corrupt” a data sample to obtain local approximations of it [13, 14], as a result, it leads to the out-of-the-distribution problem [15]. Further discussion of this topic is provided in Section 2 and Section 5.

---

\*Corresponding author: vadim.borisov@uni-tuebingen.de

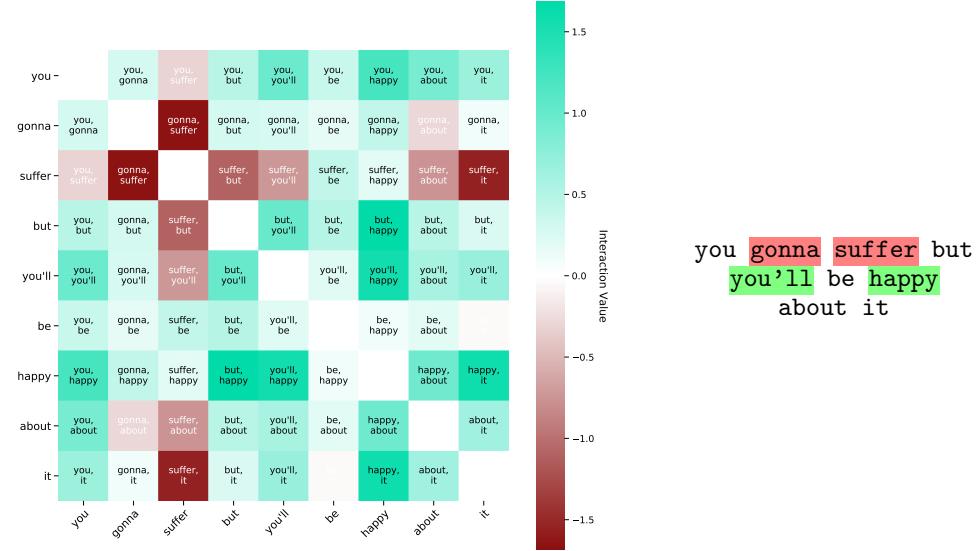


Figure 1: An example of the relational local explanation (*left*) and standard local explanation (*right*) for textual data from the proposed RLE framework, where green color indicates positive influence and red negative. It can be seen that the relational local explanation allows the analysis of the pairwise influence of each word. For the task we select a pre-trained *DistilBERT* model [16] for the sentiment analysis task. For more results, please refer to the Sec. 4 and Appendix A.

In response, we propose a new framework for post hoc feature attribution that provides *relational local explanations*. Our framework represents homogeneous data as a graph and leverages the edge information to identify the relationship between features.

The proposed approach presents a double-view on the feature attributions: (1) General local explanations as coefficients of how particular variables (words or a group of pixels) influence the decision of the given ML model positively or negatively, w.r.t other variables. (2) Relational local explanations in the form of attention matrices, where the relationship between each input variable and other variables is represented as a coefficient. This type of explanation answers an important question - *How strongly is this variable related to all other variables?* By that, we add another layer of depth to the explanation.

**Contributions.** Below, we list the main contributions of our work are:

- We highlight the importance of relational interactions between input features for local explanations. Since visual and textual data types are “compositional” per nature i.e. the “regional” information between variables naturally exists, it is crucial not only to understand what variable is important but also to spotlight and quantify the most critical combinations of them in a given data sample.
- We develop a novel model-agnostic local feature attribution technique - coined relational local explanations (RLE) - and formally describe it. To the best of our knowledge, this is the first model-agnostic local explanation algorithm based on the relationships between input variables.
- We extensively evaluate the proposed approach on image and text datasets and empirically show that it produces explanations that are superior to those produced by state-of-the-art attribution techniques.
- We open-sourced the RLE implementation <https://github.com/unnir/rle>.

The remainder of this work is organized as follows. In Section 2, we give a short overview of the related methods for explaining machine learning models. Section 3 presents the proposed RLE algorithm. After, in Section 4 we visually and empirically compare our algorithm against other state-of-the-art feature attribution approaches. Section 5 discusses the properties and limitations of the proposed method before we conclude in Section 6.

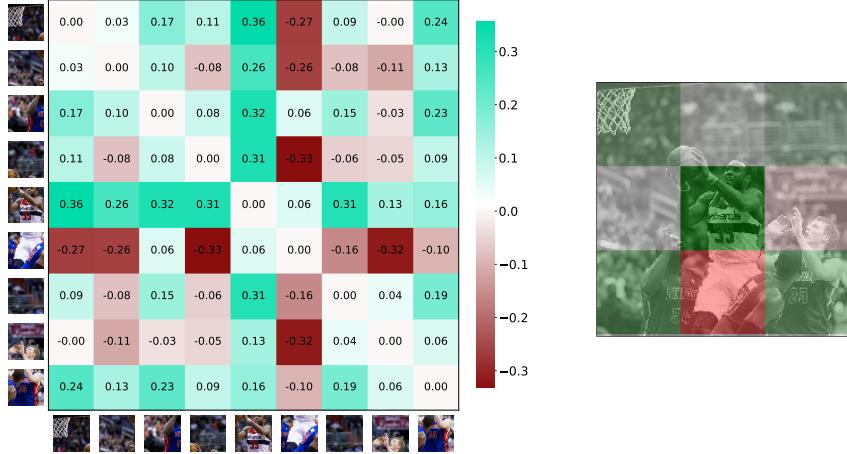


Figure 2: An example of the relational local explanation (*left*) and standard local explanation (*right*) for visual data from the proposed RLE framework where green color indicates positive influence, and red negative. The relational local explanation can be used for a deeper feature analysis of the image data. For the task we select a pre-trained ResNet-50 model [4] and an image with a class basketball from the ImageNet data set [24], e.g., to uncover a combination of patches that is the most important to a model. For more results, please refer to Appendix. 4.

## 2 Related Work

In recent years, there have been several studies that have focused on methods for explaining feature interactions and adjacency. Lundberg et al. [17] propose an efficient local explanation method based on the SHAP framework [18] for decision tree-based models, which allows for the direct measurement of local feature interaction effects. Cui et al. [19] propose a probabilistic estimation method to assess the joint effect of two input features and the sum of their marginal effects in order to evaluate global feature pairwise interactions.

A number of studies have also explored feature interaction approaches specifically for deep neural networks (DNNs). For example, Greenside et al. [20] explore interactions between variables using deep feature interaction maps by calculating the difference between the attributions of two variables. Singh et al. [21] present the generalization of the Contextual Decomposition [22] to explain interactions for dense DNNs and CNNs.

More recently, Janizel et al. [23] propose an efficient method for feature interaction local explanation for DNNs called Integrated Hessians (IH). This method is based on an enhancement of the Integrated Gradients (IG) approach [14] and has been shown to produce trustworthy results. However, from a practical perspective, the Hessian matrix is significantly larger than gradient matrices and requires a sufficient memory size.

**Limitations of prior approaches.** Despite the progress made in understanding feature relationships through previous approaches, a major limitation of these methods is that they are often tied to specific model architectures or data types, making them not fully model agnostic. In addition, perturbation-based algorithms such as LIME [13], SHAP [18], and IG [14] rely on altering the data sample in order to provide explanations, while our method aims to preserve as much information as possible by only altering the global structure. We discuss this issue in Section 5.

## 3 Relational Local Explanation (RLE) Framework

This section introduces the Relational Local Explanation (RLE) algorithm by first discussing its main components. In addition, we present how the RLE approach can be utilized for visual and textual data modalities.

---

**Algorithm 1** Relational Local Explanations (RLE)

---

**Require:** ML model  $f$ , Instance to explain  $\mathbf{x}_o$ , Number of permutations  $n$

```

 $\mathcal{X}' \leftarrow \{\}$                                 ▷ New auxiliary data set
for  $i \in \{1, 2, 3, \dots, n\}$  do
     $\mathbf{x}_i^p \leftarrow \text{permute}(\mathbf{x}_o)$           ▷ Replace and shuffle the instance to explain
     $\mathcal{G}_i \leftarrow \text{Graph}(\mathbf{x}_i^p)$           ▷ Get the graph structure of patches
     $\mathbf{A}_i \leftarrow \text{AdjacencyMatrix}(\mathcal{G}_i)$     ▷ Get the adjacency matrix
     $\mathbf{x}' \leftarrow \text{Lower}(\mathbf{A}_i)$               ▷ Get the lower triangle
     $\mathcal{X}' \leftarrow \mathcal{X}' \cup \langle \mathbf{x}', f(\tilde{\mathbf{x}}_i) \rangle$ 
end for
 $w \leftarrow \text{LinearModel}(\mathcal{X}')$            ▷ Train an explainable-by-design surrogate model
return  $w$ 

```

---

### 3.1 Formal definitions

Before proceeding to the description of the proposed method, we introduce central definitions of our study. The definitions are based on existing works [18, 14].

**Definition 1 (Local Explanation)** A feature attribution function can be seen as  $\phi(f, \mathbf{x}, c_{\mathbf{x}}) \in \mathbb{R}^n$ , where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a black-box model and  $\mathbf{x} \in \mathbb{R}^n$  is an input sample belonging to a class  $c_{\mathbf{x}} \subset \mathbb{R}$ . The output of  $\phi$  is an explanation representation vector  $\mathbf{e}_{\mathbf{x}} \in \mathbb{R}^n$ .

Each element of  $\mathbf{e}_{\mathbf{x}}$  is an importance score corresponding to a feature value in  $\mathbf{x}$ . A large positive or negative value in  $\mathbf{e}_{\mathbf{x}}$  indicates that a corresponding feature greatly influences the outcome. Features with values close to zero in  $\mathbf{e}_{\mathbf{x}}$  have little impact. Note that there are explanation methods that do not require a class specification; thus, for simpler and more general notation, a feature scoring function has the form  $\phi(f, \mathbf{x})$ .

Taking the description of local explanations, we can extend it to the definition of relational local explanations.

**Definition 2 (Relational Local Explanation)** A relational local explanation function can be seen as  $\Psi(f, \mathbf{x}, c_{\mathbf{x}}) \in \mathbb{R}^{n \times n}$ , where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a black-box model as above and  $\mathbf{x} \in \mathbb{R}^n$  is an input sample belonging to a class  $c_{\mathbf{x}} \subset \mathbb{R}$ . The output of the  $\Psi$  is an relational explanation representation in a form of a adjacency matrix  $\mathbf{A}_{\mathbf{x}}$ .

The relational local explanation matrix  $\mathbf{A}_{\mathbf{x}}$  contains in each cell  $\mathbf{A}_{\mathbf{x}}(i, j)$  the relational interaction between the  $i$ 'th and  $j$ 'th input feature. Note that  $\mathbf{A}_{\mathbf{x}}$  is symmetric, i.e.,  $\mathbf{A}_{\mathbf{x}} = \mathbf{A}_{\mathbf{x}}^\top$ , and thus, the mean value of column or row elements corresponds to average local importance for the corresponding feature. Formally, let  $\bar{\mathbf{A}}_{\mathbf{x}}$  denote the vector of mean values of the rows of matrix  $\mathbf{A}_{\mathbf{x}}$ . Then:

$$\bar{\mathbf{A}}_{\mathbf{x}} \triangleq \mathbf{e}_{\mathbf{x}}. \quad (1)$$

The symmetry property of relational local explanations is based on the assumption that the association between two variables has to be symmetrical. This was also indicated in the previous related works [23].

### 3.2 RLE : The proposed framework

Our approach follows common strategies for the generation local explanation proposed in LIME [13], SHAP [18], and Anchors [25], since they have a solid theoretical foundation [26] and established reputation in the ML community [1, 27]. The main idea of the RLE algorithm is to generate  $n$  local permutations of a data sample to explain, then construct corresponding graph representations and adjacency matrices of the relationships between input features from the shuffled data. Thus we obtain a new data set of local relations between features. Next, a linear model (that is explainable by design) is fitted to the new data set - using information from the adjacency matrices to get the corresponding coefficients, which can be utilized for the relational local explanation.

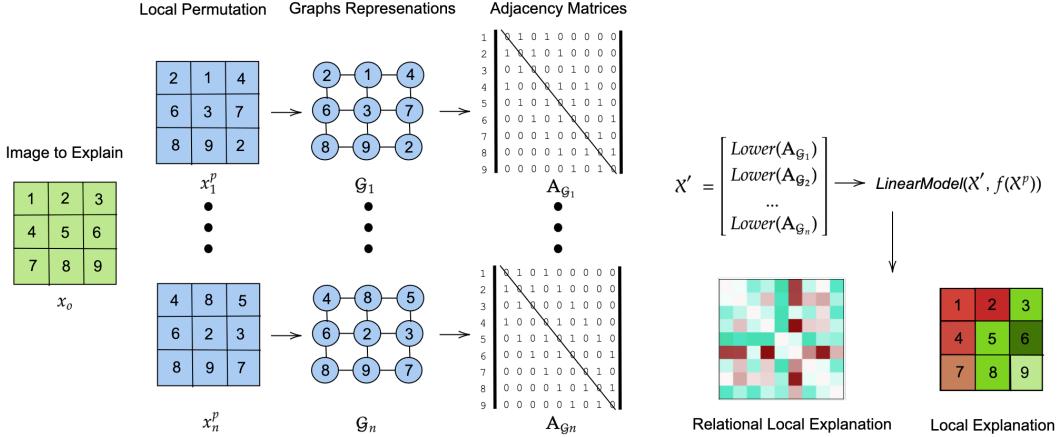


Figure 3: A relational local explanation of a data sample given a vision model using the RLE algorithm. Where  $f$  is a black-box machine learning model to explain,  $x_o$  is a sample of interest,  $\mathcal{G}_i$  is a graph of representation of a perturb image sample  $x_i^p$ .

Formally, for a given black-box model  $f : \mathcal{X} \rightarrow \mathcal{Y}$  with  $\mathcal{X} \subseteq \mathbb{R}^n$ ,  $\mathcal{Y} \subset \mathbb{R}$ , we may learn an interpretable *surrogate* model  $g$ , which is a local approximation of  $f$  for a given perturbation of a particular input  $x_o \in \mathcal{X}$ . For this purpose, we first divide a data sample (image, text) into discrete elements of pixel patches for visual data, and groups of words for textual data. Then the chosen data sample  $x_o$  is perturbed  $n$  times to generate  $x_{oi}^p$ ,  $i = 1..n$  and we randomly replace a single element (i.e., patch/group)  $p_{oi}$  from  $x_o$  with another randomly selected element from  $x_o$ . After an undirected graph representation of the shuffled sample is received  $\mathcal{G}_{x_o}^p$ , where each vertex is a discrete element  $p$  (e.g., superpixel or word), and the edge is the connection between them. Further, an adjacency matrix  $\mathbf{A}_{\mathcal{G}_i}$  for each  $x_i^p$  is obtained. Since *adjacency matrices* for undirected graphs are symmetric, only the lower triangle is utilized  $Lower(\mathbf{A}_{\mathcal{G}_i})$  for the next step. The key idea is to permute a data sample and keep local features since the strong perturbation may lead to the out-of-distribution problem [15]; we examine this issue in detail in Section 5.

These procedures yield a new data set  $\mathcal{X}' = \{Lower(\mathbf{A}_{\mathcal{G}_i}), f(x_i^p)\}_{i=1}^n$ . We then learn a sparse linear regression  $g_{w_{x_o}}(x^p) = \mathbf{w}_{x_o}^\top x^p$  using the local data set  $\mathcal{X}'$  by optimizing the following loss function with  $\Omega(\cdot)$  as a measure of complexity.

$$w_{x_o} = \underset{\mathbf{w}}{\operatorname{argmin}} \mathcal{L}(f, g) + \Omega(\mathbf{w}), \quad (2)$$

where  $\mathcal{L}(f, g)$  is the mean squared loss,

$$\mathcal{L}(f, g) = \frac{1}{n} \sum_{i=1}^n \left[ f(x_i^p) - g(x_i^p, x_o) \right]^2. \quad (3)$$

The RLE algorithm yields  $g_{w_{x_o}}(x')$ , which approximates the complex model  $f(x')$  around  $x_o$ . In case  $g$  is a linear model, the components of the weight vector  $\mathbf{w}_{x_o}$  indicate the relative influence of the relationship between features values of  $x_o$  based the sample  $\mathcal{X}'$  and can be used as the relational local explanation of  $f(x_o)$ . The full approach is summarized in Algorithm 1.

The following subsections discuss how the proposed algorithm can be adapted to the visual and textual data modalities.

### 3.3 Relational local explanations for multidimensional visual data

In the case of visual data, we divide an image into patches to disrupt the spatial layout of local image regions, this is a common approach for the local explanation methods [13, 18]. Formally, given an input image  $I$ , we first uniformly partition the image into  $N \times N$  sub-regions denoted by  $R_{i,j}$ ,

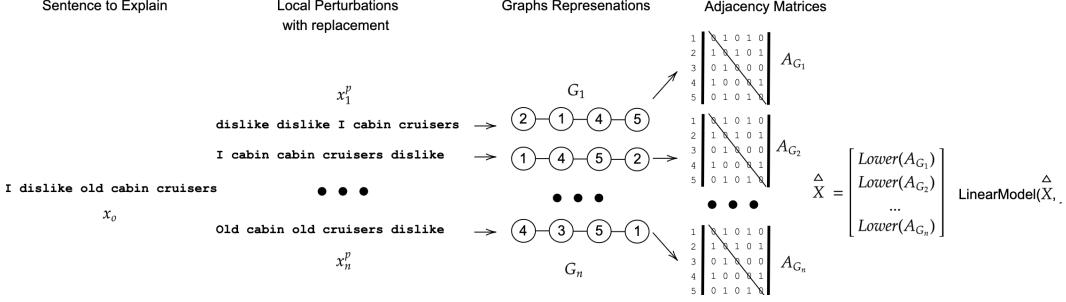


Figure 4: Relational local explanation of a data sample given a textual model using the RLE algorithm. Where  $f$  is a black-box machine learning model to explain,  $x_o$  is a sentence of interest,  $\mathcal{G}_i$  is a graph of representation of a perturb sentence  $x_i^p$ .

where  $i$  and  $j$  are the horizontal and vertical indices respectively and  $1 \leq i, j \leq N$ . The procedure is presented in Fig. 3. Following that, a single patch  $R_{i,j}$  represents a feature for the RLE algorithm.

Note that obtained graph representations for patched images do not consider the neighbor direction, e.g., a patch connected from the top, bottom, right, or left. To include the directional information, we may use a simple one-hot-encoding encoding technique for the adjacency matrix  $\mathbf{A}_{\mathcal{G}_i}$ . In practice, we observe that a meaningful relational local explanation can be constructed even without the one-hot-encoding step.

We argue that local details are much more important than a global structure for fine-grained image recognition, as it is these details that distinguish between different classes. In most cases, various fine-grained categories tend to share similar global structures and vary only in specific local details [28]; therefore, by random permutation, a given computer vision black-box model is forced to focus on the local details, and it favors the most distinguished areas. Multiple studies support our argument: the jigsaw puzzle pretext task for Self-Supervised Learning (SSL) approaches [29], a regularization scheme for Variational Autoencoders (VAEs) [30], and last but not least for Vision Transformers (ViTs) an image is divided into patches as well [31, 32]. Besides, graph-based representation of visual data is a common approach for many downstream tasks [33].

### 3.4 Relational local explanations for textual data

For relational local explanations of the textual models, in particular self-attention-based models [5], we represent a sentence as a graph  $\mathcal{G}_i$ , where each word is expressed as a node. Then, as we presented before, we permute the sentence by chaining the word order with replacement  $n$  times. The whole procedure is illustrated in Fig. 4. This operation allows learning bidirectional context for a word. As a result, each position grasps directional context from both “directions”.

The permutation idea was successfully used for training an XLNet [34] and GReAT [35] Transformer models; thus, it shows that transformer models are able to understand the semantics of a shuffled sentence. Furthermore, multiple studies [36, 37] made the same observation - a sentence can be seen as a graph, where words correspond to nodes and the computation of an attention score is the assignment of a weight to an edge between two words.

## 4 Experiments

We evaluate the RLE framework on several data sets by visually comparing them with state-of-the-art explanation methods with the goal of ensuring that our method fulfills its purpose. First, in Section 4.1 we present a visual comparison of state-of-the-art methods and the proposed framework. After, we compare local explanation for a textual data given a pre-trained DistilBERT model [16], furthermore, we demonstrate relational local explanation for a sentence in Section 4.2. Next, Section 4.3 presents a quantitative benchmark analysis of the RLE method with a comparison to selected baselines. For more visual experiments and detailed reproducibility details, please refer to the supplemental materials.

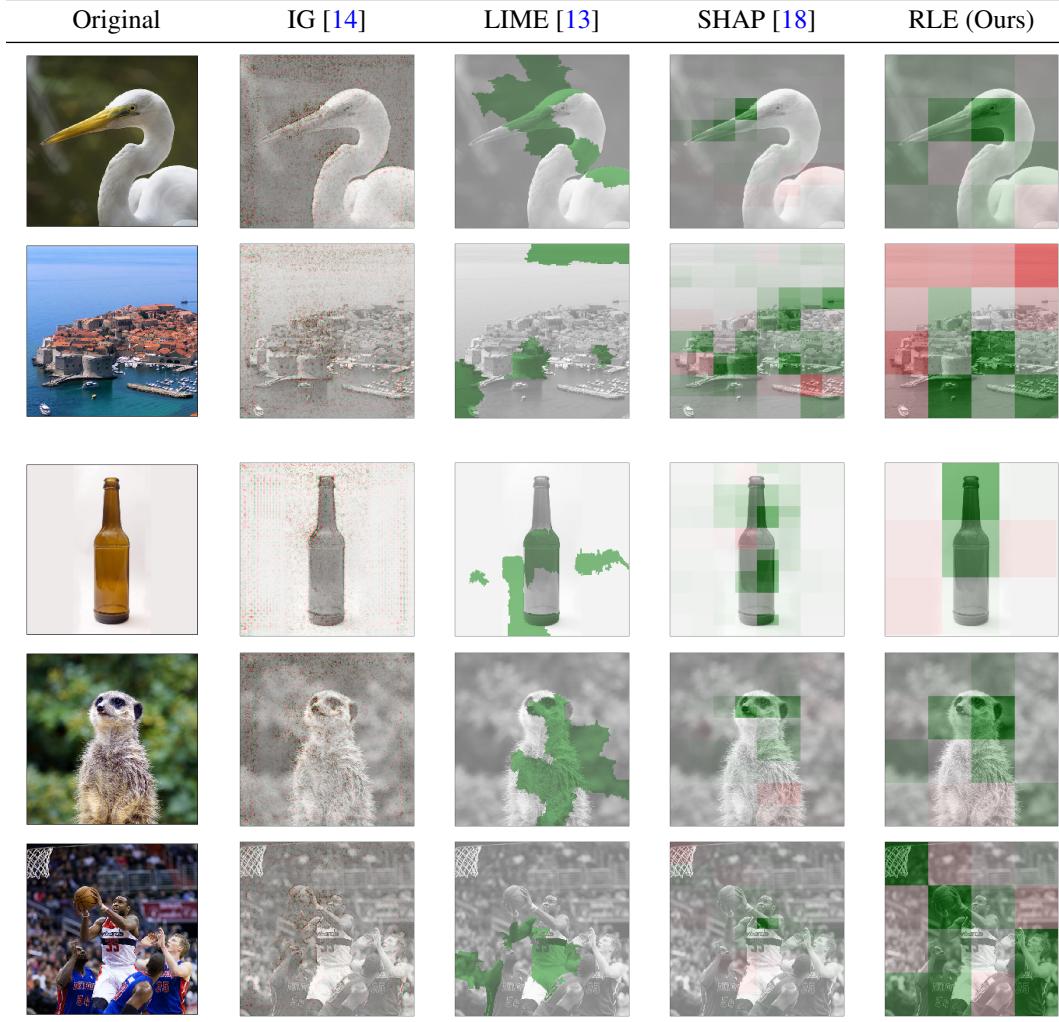


Table 1: A Comparison of different state-of-the-art local explanation methods for a pre-trained ResNet-50 model [4] given random samples from the ImageNet data set [24]. Name of the original classes according to the selected model (from top to bottom): American egret, seashore, bottle, marmot, and basketball.

#### 4.1 Visual analysis on image data

In our first experiment, a qualitative visual evaluation on images from the ImageNet data set [24] is performed for selected baseline: IG [14], LIME [13], SHAP [18], and the proposed RLE framework. We color the explanations from all baselines to illustrate that they distinguish between input variables that positively (green) and negatively (red) contribute to the CNN model estimations for a given class. The results are summarized in Table. 1. Also, an example of a relational local explanation for an image is depicted in Fig. 2.

#### 4.2 Comparison on textual data

For the evaluation of the proposed method on the textual data, we utilize a pre-trained self-attention-based DistilBERT model for the sentiment analysis task [16] from the open-source Hugging Face library [38]. The results are summarized in the Table 2. Also, we demonstrate relational local explanations analysis using the RLE framework (Fig. 1) for the same transformer model and compare it to results from the IH method in Fig. 5.

| Method            | Local Explanation   |
|-------------------|---|
| IG [14]           | You gonna suffer but you'll be happy about it             |
| LIME [13]         | You gonna suffer but you'll be happy about it             |
| SHAP [18]         | You gonna suffer but you'll be happy about it             |
| IH [23]           | You gonna suffer but you'll be happy about it             |
| <b>RLE (ours)</b> | You gonna suffer but you'll be happy about it             |
| IG [14]           | You might be interested this product performs well        |
| LIME [13]         | You might be interested this product performs well        |
| SHAP [18]         | You might be interested this product performs well        |
| IH [23]           | You might be interested this product performs well        |
| <b>RLE (ours)</b> | You might be interested this product performs well        |
| IG [14]           | The idea is nicely presented, but it has some limitations |
| LIME [13]         | The idea is nicely presented, but it has some limitations |
| SHAP [18]         | The idea is nicely presented, but it has some limitations |
| IH [23]           | The idea is nicely presented, but it has some limitations |
| <b>RLE (ours)</b> | The idea is nicely presented, but it has some limitations |

Table 2: A comparison of state-of-art feature attribution approaches to the presented RLE algorithm. given a pre-trained DistilBERT model [16] for the sentiment analysis task. We highlight the most important words according to each feature attribution method. For more results, please refer to the supplementary materials.

### 4.3 Quantitative comparison

In order to quantitatively evaluate our novel explanation framework, we utilize the well-accepted measure in the ML community - Iterative Removal Of Features (IROF) [39]. The IROF measure is fully described in the supplementary materials. The full definition of the mesure is in the Appendix C. This technique was featured in multiple studies before [40]. We compare against this study baselines: IG [14], LIME [13], and SHAP [18]. We also introduce the random baseline as the “sanity check”, which assigns variable importance randomly. Notably, the authors of [41, 42] show that this primitive baseline can outperform some of the commonly used explanation approaches based on saliency maps in ablation tests. Results are in Table 3.

### 4.4 Reproducibility Details

In this subsection, we briefly introduce the main frameworks used in this study; further details about all experiments, such as hyperparameters for each baseline and experiment, are provided in the supplementary materials. We selected official implementation for the LIME, SHAP, and IH baselines, and for the IG baseline, we employed the Captum library [43]. The graph structure was analyzed using the NetworkX library [44]. For all experiments we use a single NVIDIA 2080TI GPU with 12 GB of memory. For future comparison, we also open-source the code for the RLE framework for PyTorch [45] models and publish it online.

## 5 Discussion and Future Work

**Experimental results.** In our challenging experiments with multiple data modalities, local explanations from the RLE framework show competitive performance against selected feature attribution baselines. Overall, our quantitative experimental results resemble similar image areas or words from highlighting by other state-of-the-art non relational explanation methods. In qualitative experiments, the proposed approach shows the best results on the IROF measure [39]. For more results, please refer to the supplementary material.

| Method            | IROF [39]         |
|-------------------|-------------------|
| Random            | 0.179±0.18        |
| IG [14]           | 0.211±0.23        |
| LIME [13]         | 0.421±0.19        |
| SHAP [18]         | 0.368±0.24        |
| <b>RLE (ours)</b> | <b>0.434±0.23</b> |

Table 3: A quantitative comparison of selected baselines on the fifty random images from ImageNet data set [24] given a pre-trained ResNet-50 model [16]. The top result is bold, whereas the second result is underlined.

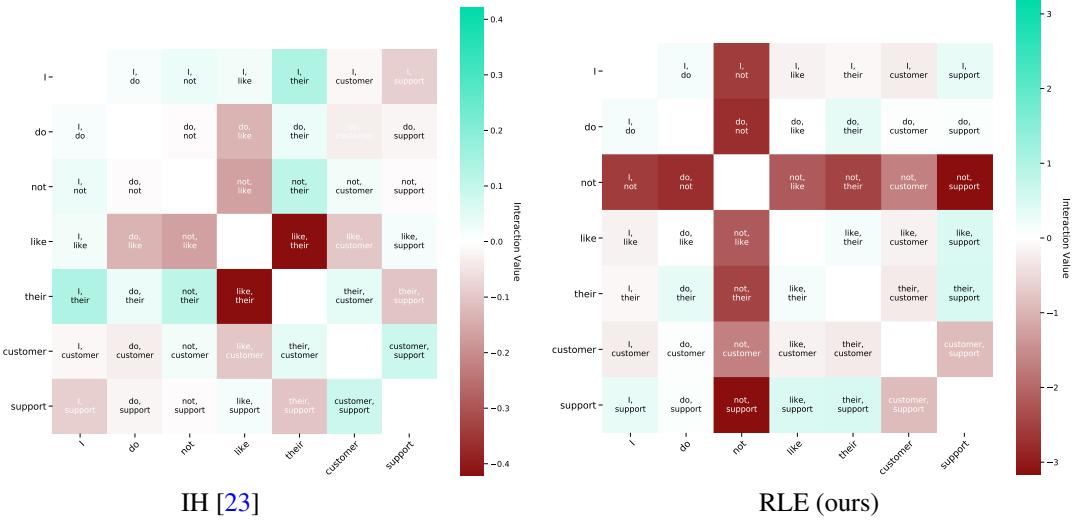


Figure 5: A comparison of the relational local explanations from IH [23] and RLE methods for a sentence “I do not like their customer support”, given a pre-trained DistilBERT model [16] for the sentiment analysis task. According to the IH method the most negative pair is “like, their”, where our proposed approach shows the most negative word is “not” with two pairs “not, support” and “do, not”. Appendix A presents more results.

**Permutation step.** One of the core steps of the proposed algorithm is the random permutation with a replacement - we refer to it as a *weak perturbation*. In comparison to other perturbation-based feature attribution methods which use a *strong perturbation*, e.g., perturb a data sample by adding random noise [13, 14] or removing parts of information [18], the RLE framework preserves the local attribution unchanged, only shuffling the global structure. Moreover, local details are more important than a global structure for deep neural network models, as shown for vision and textual modalities [28, 34, 46]. Another known issue related to the strong perturbation approaches for the local approximations, this type of perturbation leads to the out-of-the-distribution problem [47], which creates the vulnerability to adversarial attacks [15].

**Evaluation measure for relational local explanations.** Another point of the work’s continuation with relational local explanations is the absence of evaluation technique. For future work, a trustworthy and plausibility measure is needed; this is challenging since there is no access to the ground truth. On the flip side, with an unambiguous measure, a possible strategy would involve direct optimization over it.

**Ensembling of feature attributions.** To improve the robustness of local and relational explanations, unsupervised ensemble techniques can be applied to the outputs of multiple runs of the explanation algorithm. By aggregating the outputs of multiple runs of the algorithm, we can effectively reduce the impact of any individual run that may have produced biased or inaccurate explanations. This approach has been shown to be effective at improving the robustness of explanations in a variety of contexts [48].

**RLE limitations.** The proposed approach does not currently support the quantification of higher-order interactions between features for the relational local explanations. A more complex graph-based representation can be utilized for this task for future work. Furthermore, patches of image data should have adequate local information, and the adequacy depends on the resolution of the images. In our experiments using the ImageNet data set - an image usually cropped into the  $224 \times 224 \times 3$  format, we observe that we can operate with up to 36-49 patches (depending on the content of an image). Lastly, the current implementation of the RLE framework cannot be utilized on tabular modality since tabular data has no spatial relationships [49].

## 6 Conclusion

This paper introduces the RLE (Relational Local Explanations) method, a novel, model-agnostic approach for generating local explanations that addresses a common challenge in post hoc explana-

tions, which is the interpretability of inter-variable relationships. This method provides a qualitative measure of how different feature attributions interact with each other, which is useful for various data types and problems where knowledge of the relationships between features is necessary. In addition, the RLE framework also offers standard feature attributions as local explanations, which provide insight into the specific contributions of each feature towards the final prediction made by a machine learning model. Through extensive visual and quantitative experiments, we demonstrate that the proposed RLE method performs comparably to state-of-the-art methods for generating comprehensive (relational) local explanations. These results suggest that RLE may be a valuable tool for practitioners seeking to understand and improve the performance of their machine learning models.

## References

- [1] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019. (cited on p. 1, 4)
- [2] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020. (cited on p. 1)
- [3] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021. (cited on p. 1)
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. (cited on p. 1, 3, 7, 14, 15, 18, 19)
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. (cited on p. 1, 6)
- [6] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021. (cited on p. 1)
- [7] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021. (cited on p. 1)
- [8] Iam Palatnik De Sousa, Marley Maria Bernardes Rebuzzi Vellasco, and Eduardo Costa Da Silva. Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors (Basel, Switzerland)*, 19(13), 2019. (cited on p. 1)
- [9] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *arXiv preprint arXiv:2012.14261*, 2020. (cited on p. 1)
- [10] Council of the European Union European Parliament. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016. *Official Journal of the European Union*, 2016. (cited on p. 1)
- [11] CA OAG. Ccpa regulations: Final regulation text. *Office of the Attorney General, California Department of Justice*, 2021. (cited on p. 1)
- [12] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020. (cited on p. 1)
- [13] Marco Tulio 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*, pages 1135–1144, 2016. (cited on p. 1, 3, 4, 5, 7, 8, 9, 13, 15, 16)

- [14] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. (cited on p. 1, 3, 4, 7, 8, 9, 13, 15, 16)
- [15] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. (cited on p. 1, 5, 9)
- [16] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. (cited on p. 2, 6, 7, 8, 9, 14, 16, 17)
- [17] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019. (cited on p. 3)
- [18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. (cited on p. 3, 4, 5, 7, 8, 9, 13, 15, 16)
- [19] Tianyu Cui, Pekka Marttinen, and Samuel Kaski. Learning pairwise interactions with bayesian neural networks. *arXiv preprint arXiv:1901.08361*, 2019. (cited on p. 3)
- [20] Peyton Greenside, Tyler Shimko, Polly Fordyce, and Anshul Kundaje. Discovering epistatic feature interactions from neural network models of regulatory dna sequences. *Bioinformatics*, 34(17):i629–i637, 2018. (cited on p. 3)
- [21] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018. (cited on p. 3)
- [22] W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv preprint arXiv:1801.05453*, 2018. (cited on p. 3)
- [23] Joseph D Janizek, Pascal Sturmels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021. (cited on p. 3, 4, 8, 9, 13, 14, 16, 17)
- [24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. (cited on p. 3, 7, 8, 13, 15, 18, 19)
- [25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. (cited on p. 4)
- [26] Damien Garreau and Ulrike Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International Conference on Artificial Intelligence and Statistics*, pages 1287–1296. PMLR, 2020. (cited on p. 4)
- [27] Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*, 2020. (cited on p. 4)
- [28] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019. (cited on p. 6, 9)
- [29] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. (cited on p. 6)

- [30] Saeid Asgari Taghanaki, Mohammad Havaei, Alex Lamb, Aditya Sanghi, Ara Danielyan, and Tonya Custis. Jigsaw-vae: Towards balancing features in variational autoencoders. *arXiv preprint arXiv:2005.05496*, 2020. (cited on p. 6)
- [31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. (cited on p. 6)
- [32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. (cited on p. 6)
- [33] Alberto Sanfeliu, René Alquézar, J Andrade, Joan Climent, Francesc Serratosa, and J Vergés. Graph-based representations and techniques for image processing and image analysis. *Pattern recognition*, 35(3):639–650, 2002. (cited on p. 6)
- [34] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. (cited on p. 6, 9)
- [35] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022. (cited on p. 6)
- [36] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. (cited on p. 6)
- [37] Chaitanya Joshi. Transformers are graph neural networks. *The Gradient*, 2020. (cited on p. 6)
- [38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. (cited on p. 7, 14)
- [39] Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. *arXiv preprint arXiv:2003.08747*, 2020. (cited on p. 8, 14)
- [40] Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020. (cited on p. 8)
- [41] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. Evaluating feature importance estimates. *arXiv preprint arXiv:1806.10758*, 2018. (cited on p. 8)
- [42] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, pages 18770–18795. PMLR, 2022. (cited on p. 8)
- [43] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. (cited on p. 8, 13)
- [44] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008. (cited on p. 8)

- [45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. (cited on p. 8, 14)
- [46] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021. (cited on p. 9)
- [47] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. *Advances in Neural Information Processing Systems*, 34, 2021. (cited on p. 9)
- [48] Vadim Borisov, Johannes Meier, Johan van den Heuvel, Hamed Jalali, and Gjergji Kasneci. A robust unsupervised ensemble of feature-based explanations using restricted boltzmann machines. *arXiv preprint arXiv:2111.07379*, 2021. (cited on p. 9)
- [49] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*, 2021. (cited on p. 9)
- [50] Naman Bansal, Chirag Agarwal, and Anh Nguyen. Sam: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8673–8683, 2020. (cited on p. 13)
- [51] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. (cited on p. 14)

## A Additional Experiments

This section present experimental results on visual (Table 4) and textual (Table 5) data. Additionally, we show relational local explanations for several visual and text samples in Figures 6, 7, 8, 9, 10, 11, 12. We compare results from the proposed algorithm to the baselines of this study: IG [14], LIME [13], SHAP [18], and IH [23].

## B Further Reproducibility Details

**Hyperparameters.** We select similar hyperparameters for each baseline to have a fair evaluation; for image data, the number of perturbations (auxiliary samples)  $n$  is set to 5000, and for textual, we set  $n$  to 2000. In our experiments, we observe that a higher number of perturbation steps leads to better quality local explanations. This was also observed in [50]. The rest of the hyperparameters default to a selected package.

**RLE plotting function.** For the relational local explanation visualization we apply a plotting function from the IH [23] official implementation.<sup>2</sup> From the open-source library Captum [43], we utilize plotting function for visualization feature attribution maps on visual data.<sup>3</sup>

**Data sets.** For image data, we utilize samples from ImageNet data set [24] provided by the official python package of the SHAP algorithm [18].<sup>4</sup>

<sup>2</sup>[https://github.com/suinleelab/path\\_explain](https://github.com/suinleelab/path_explain)

<sup>3</sup><https://captum.ai/api/utilities.html>

<sup>4</sup><https://shap.readthedocs.io/en/latest/generated/shap.datasets.imagenet50.html>

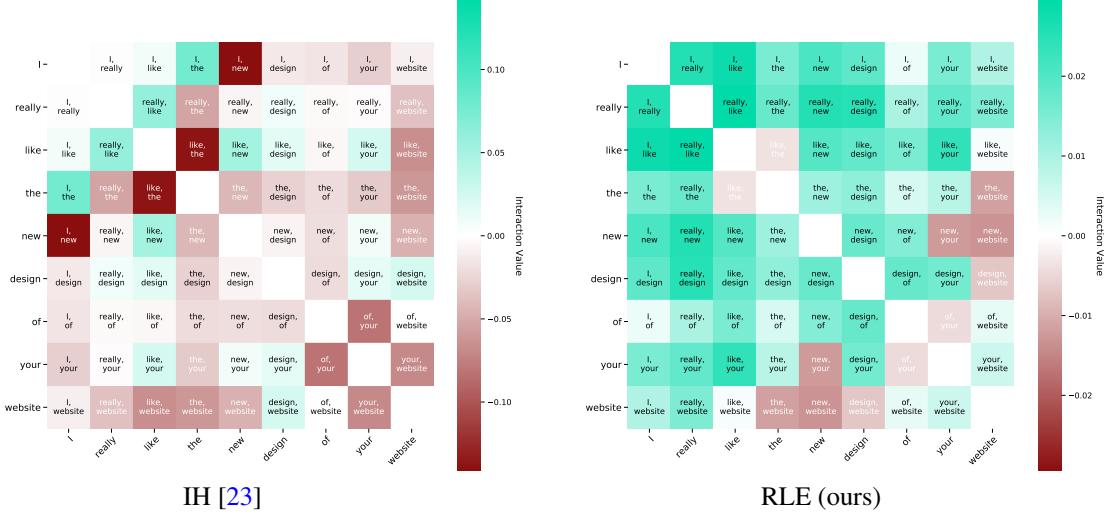


Figure 6: A comparison of the relational local explanations from IH [23] and RLE methods for a sentence “I really like the new design of your website”, given a pre-trained DistilBERT model [16] for the sentiment analysis task.

**Pre-trained models.** In this work, we employ the pre-trained ResNet-50 model [4] from the *torchvision* package [45].<sup>5</sup>. We utilize the pre-trained DistilBERT model [16] from the *HugginFace* library [38].<sup>6</sup>

For even better reproducibility, we also report the used package versions in `requirements.txt` file. It can be found in the corresponding code repository of the RLE algorithm.

## C The IROF Measure

**Choice of the quantitative measure.** In our work, we select the IROF framework [39], since it allows for an efficient and fairly evaluation of feature attribution methods for the visual data. In comparison to popular approaches for single-pixel-based evaluation of local explanations (e.g., DAUC, IAUC), the chosen evaluation framework uses the super-pixel approach. Since the influence of a single pixel is minimal, the unsupervised grouping of a pixel into local regions allows us for a more fair comparison of the feature attribution methods.

The IROF approach has several steps: First, the image is divided into coherent segments and bypasses the input features’ inter-dependency. According to the creators of the IROF measure, we use the SLIC method for unsupervised image segmentation [51].

Formally, the IROF measure is defined as follows:

$$\text{IROF}(e_j) = \frac{1}{N} \sum_{n=1}^N \text{AOC} \left( \frac{f(x_n^0)_y}{f(x_n^0)_y} \right)_{l=0}^L \quad (4)$$

where  $e_j$  is a local feature attribution map,  $N$  is the number of super-pixels,  $x^0$  is an image to explain,  $f$  is a black-box model,  $y$  is a target, and  $L$  represents the class score based on how many segments of the image were removed. Also, the proposed measure utilized the area over the curve (AOC) function. The higher the IROF score, the more plausible the local explanations, i.e., the more information was collected.

<sup>5</sup><https://pytorch.org/vision/stable/models.html>

<sup>6</sup><https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english>

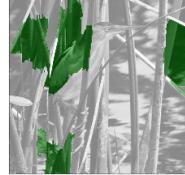
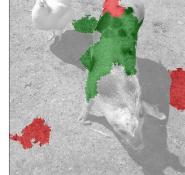
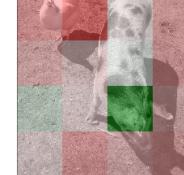
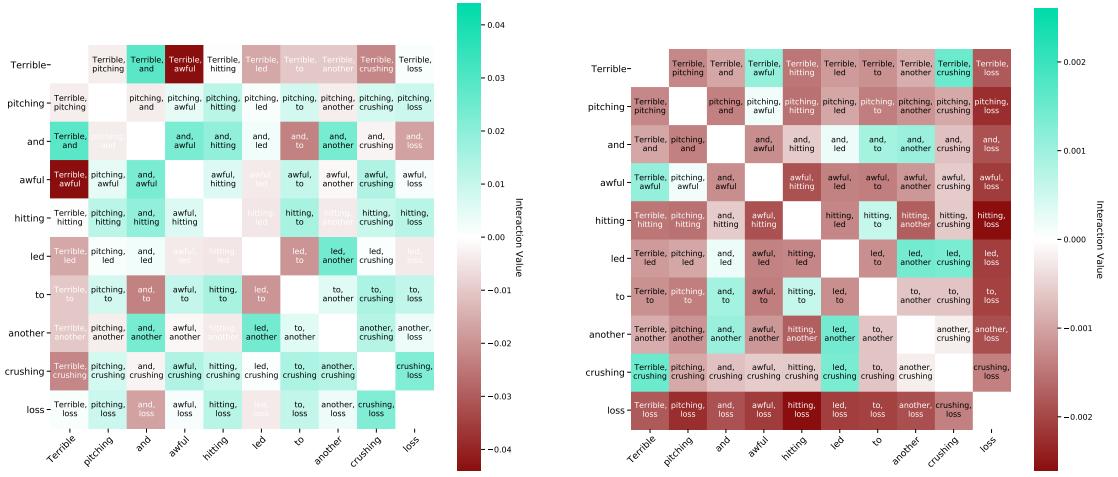
| Original  | IG [14]   | LIME [13]   | SHAP [18]  | RLE (Ours)  |
|---|---|---|--|---|
|    |    |    |    |    |
|    |    |    |    |    |
|    |    |    |    |    |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Table 4: A Comparison of different state-of-the-art local explanation methods for a pre-trained ResNet-50 model [4] given random samples from the ImageNet data set [24]. Name of the original classes according to the selected model (from top to bottom): bittern, Indian elephant, hog, dowitcher, cardigan, and desktop computer. The explanations from the RLE method are less noisy and rank almost all parts of the image with either positive (green) or negative (red) influence.

| Method  | Local Explanation   |
|---|---|
| IG [14]<br>LIME [13]<br>SHAP [18]<br>IH [23]<br><b>RLE (ours)</b> | The new <b>design</b> is <b>awful</b><br>The new design is <b>awful</b><br>The new design <b>is</b> <b>awful</b><br>The new design <b>is</b> <b>awful</b><br>The new design is <b>awful</b>   |
| IG [14]<br>LIME [13]<br>SHAP [18]<br>IH [23]<br><b>RLE (ours)</b> | I <b>love</b> you and I <b>hate</b> you<br>I <b>love</b> you and I <b>hate</b> you<br>I <b>love</b> you and I <b>hate</b> <b>you</b><br>I <b>love</b> <b>you</b> and I <b>hate</b> you<br>I <b>love</b> you and I <b>hate</b> you   |
| IG [14]<br>LIME [13]<br>SHAP [18]<br>IH [23]<br><b>RLE (ours)</b> | I'm <b>not</b> <b>sure</b> if I <b>like</b> the new <b>design</b><br>I'm <b>not</b> <b>sure</b> if I <b>like</b> the new design<br>I'm <b>not</b> <b>sure</b> if I like the new design<br>I'm <b>not</b> <b>sure</b> if I <b>like</b> the <b>new</b> design<br>I'm <b>not</b> <b>sure</b> if I like the <b>new</b> design   |
| IG [14]<br>LIME [13]<br>SHAP [18]<br>IH [23]<br><b>RLE (ours)</b> | I <b>really</b> like the <b>new</b> design of <b>your</b> <b>website</b><br>I <b>really</b> like the new design of your website<br>I <b>really</b> like the new design of your <b>website</b><br>I <b>really</b> like the <b>new</b> design of your <b>website</b><br>I <b>really</b> like the new design of your website   |
| IG [14]<br>LIME [13]<br>SHAP [18]<br>IH [23]<br><b>RLE (ours)</b> | The <b>bed</b> was <b>super</b> <b>comfy</b> . The <b>chair</b> wasn't bad, <b>either</b><br>The <b>bed</b> was <b>super</b> <b>comfy</b> . The chair wasn't bad, <b>either</b><br>The bed <b>was</b> <b>super</b> <b>comfy</b> . The chair <b>wasn't</b> bad, <b>either</b><br>The bed was super <b>comfy</b> . The <b>chair</b> wasn't bad, either<br>The bed was super <b>comfy</b> . The chair <b>wasn't</b> <b>bad</b> , either  |
| IG [14]<br>LIME [13]<br>SHAP [18]<br>IH [23]<br><b>RLE (ours)</b> | Terrible <b>pitching</b> and <b>awful</b> <b>hitting</b> led to <b>another</b> crushing <b>loss</b><br>Terrible <b>pitching</b> and <b>awful</b> <b>hitting</b> led to another <b>crushing</b> <b>loss</b><br>Terrible <b>pitching</b> and <b>awful</b> <b>hitting</b> led to <b>another</b> crushing loss<br>Terrible <b>pitching</b> and <b>awful</b> <b>hitting</b> led to <b>another</b> crushing loss<br>Terrible <b>pitching</b> and <b>awful</b> <b>hitting</b> <b>led</b> to another crushing <b>loss</b> |

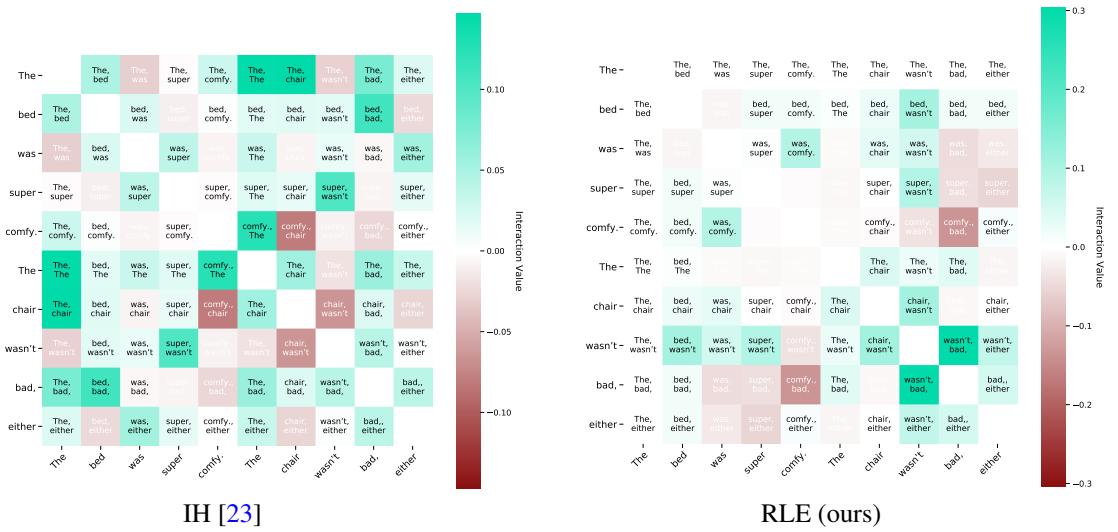
Table 5: A comparison of state-of-art feature attribution approaches to the presented RLE algorithm given a pre-trained DistilBERT model [16] for the sentiment analysis task. We highlight the most important words according to each feature attribution method, where the green and red colors indicate the positive and negative impact, respectively.



IH [23]

RLE (ours)

Figure 7: A comparison of the relational local explanations from IH [23] and RLE methods for a sentence “Terrible pitching and awful hitting led to another crushing loss”, given a pre-trained DistilBERT model [16] for the sentiment analysis task.



IH [23]

RLE (ours)

Figure 8: A comparison of the relational local explanations from IH [23] and RLE methods for a sentence “The bed was super comfy. The chair wasn’t bad, either”, given a pre-trained DistilBERT model [16] for the sentiment analysis task.



Figure 9: An example of the relational local explanation (*left*) and standard local explanation (*right*) for visual data from the proposed RLE framework where green color indicates positive influence, and red negative. For the task we select a pre-trained ResNet-50 model [4] and an image with a class `desktop computer` from the ImageNet data set [24], e.g., to uncover a combination of patches that is the most important to a model.

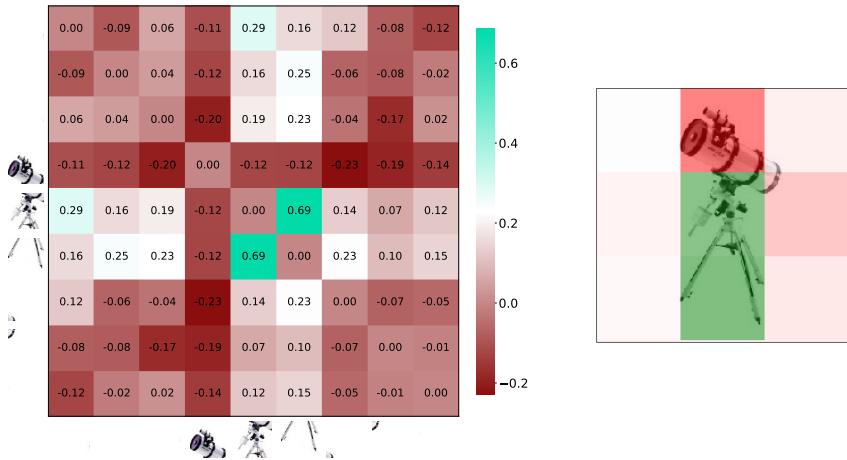


Figure 10: An example of the relational local explanation (*left*) and standard local explanation (*right*) for visual data from the proposed RLE framework where green color indicates positive influence, and red negative. For the task we select a pre-trained ResNet-50 model [4] and an image with a class `tripod` from the ImageNet data set [24], e.g., to uncover a combination of patches that is the most important to a model.

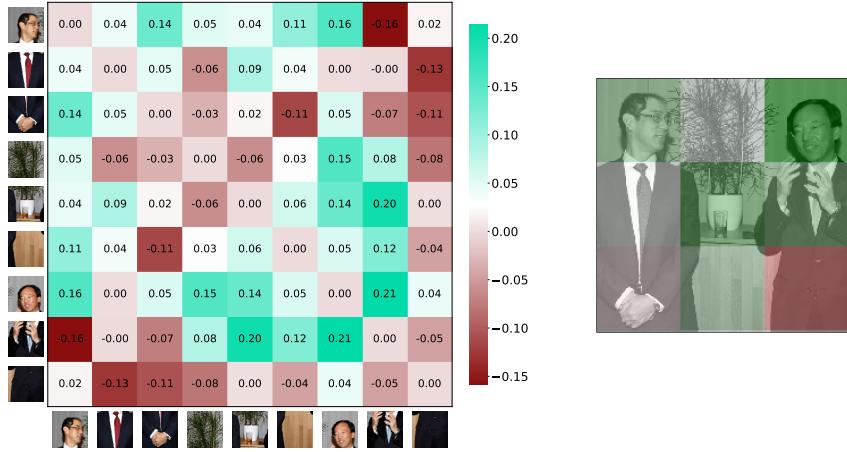


Figure 11: An example of the relational local explanation (*left*) and standard local explanation (*right*) for visual data from the proposed RLE framework where green color indicates positive influence, and red negative. For the task we select a pre-trained ResNet-50 model [4] and an image with a class *groom* from the ImageNet data set [24], e.g., to uncover a combination of patches that is the most important to a model.

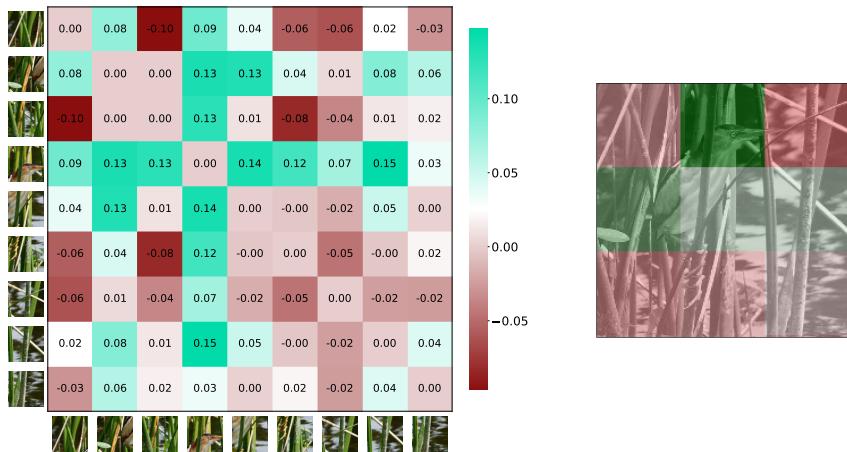


Figure 12: An example of the relational local explanation (*left*) and standard local explanation (*right*) for visual data from the proposed RLE framework where green color indicates positive influence, and red negative. For the task we select a pre-trained ResNet-50 model [4] and an image with a class *bittern* from the ImageNet data set [24], e.g., to uncover a combination of patches that is the most important to a model.