Evolving Local Interpretable Model-agnostic Explanations for Deep Neural Networks in Image Classification

Bin Wang, Wenbin Pei, Bing Xue and Mengjie Zhang
School of Engineering and Computer Science, Victoria University of Wellington
Wellington, New Zealand
{bin.wang,wenbin.pei,bing.xue,mengjie.zhang}@ecs.vuw.ac.nz

ABSTRACT

For deep convolutional neural networks (deep CNNs), a severe drawback is the poor interpretability. To address this drawback, this paper proposes a novel genetic algorithm-based method for the first time to automatically evolve local interpretable explanations that can assist users to decide whether to trust the predictions of deep CNNs. In the experiments, the results show that the evolved explanations can explain the predictions of deep CNNs on images by successfully capturing meaningful interpretable features.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

Explainable Deep learning, Evolutionary Deep Learning, Image Classification

ACM Reference Format:
Bin Wang, Wenbin Pei, Bing Xue and Mengjie Zhang. 2021. Evolving Local Interpretable Model-agnostic Explanations for Deep Neural Networks in Image Classification. In *Proceedings of the Genetic and Evolutionary Computation Conference 2021 (GECCO '21).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Deep convolutional neural networks (CNNs) have achieved the state-of-the-art classification accuracies on image classification tasks that even humans feel very hard or impossible to achieve. However, deep CNNs are very complex and not easy for humans to understand. In fact, it is essential for users to understand the reasons behind the predictions made by deep CNNs, especially when it comes to crucial decision makings, such as medical diagnosis and self-driving cars. Consequently, a new research area — *Explainable Deep Learning* (EDL), has resurged in recent years, which is to learn the explanations of the black-box deep learning models. One of the most successful methods of the local approximation is called Local Interpretable Model-agnostic Explanations (LIME) [5]. The major drawback of LIME is the high computational cost because of the process of sampling a large set of perturbed images around a local instance. Another restraint of LIME is that the number of superpixels selected is fixed, which needs to be manually fine-tuned. In this paper, to overcome the aforementioned limitations of LIME, we propose a novel Genetic Algorithm (GA) based method as a local approximation method of EDL, named E-LIME.

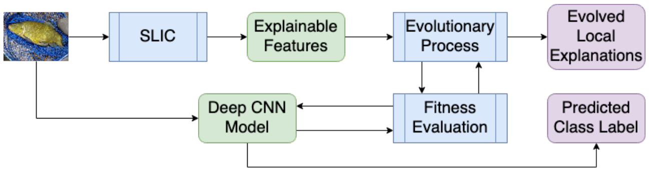


Figure 1: Overall framework.

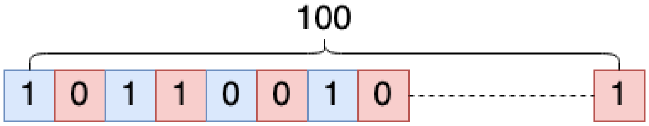


Figure 2: Encoding strategy.

2 THE PROPOSED METHOD: E-LIME

The overall framework of E-LIME is outlined in Figure 1. An image of a tencha, which is a dark olive or brown fish with a distinct brick red eye, is used as a sample image to be classified and explained by the proposed method. On the top row of the figure, the image of a tencha is processed by Simple Linear Iterative Clustering (SLIC) [1]. A set of superpixels are generated as the interpretable features. A number of binary vectors are randomly generated, forming the initial population of GA, where the length of those binary vectors is equal to the number of interpretable features. Shown in Figure 2, in each binary vector, 1 means the corresponding feature being selected, and 0 means not being selected. After the population initialisation, GA operators are applied to evolve the binary vectors, which are decoded to interpretable features. The fitness value is obtained by evaluating decoded interpretable features on the deep CNN model, which is the

GECCO '21, July 10–14, 2021, Lille, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Parameter settings

Parameter	Value
E-LIME hyper-parameters	
number of superpixels ns	100
GAs parameters	
crossover rate	0.9
mutation rate	0.2
population size	100
number of generations	50

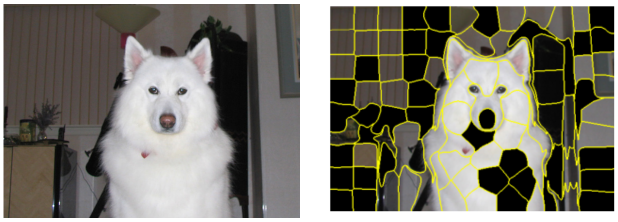
model to be explained. At the end of the evolutionary learning process, the best individual represents the evolved subset of the interpretable features, which are the evolved local explanations. Besides the local explanations, the predicted class label is also the output of the framework achieved by performing a prediction of the deep CNN model on the input image of a tench. Both of them are presented to the end-users, who can check the predicted label and decide whether to trust the output of the deep CNN model or not based on the evolved local explanations.

3 EXPERIMENT DESIGN

The ImageNet dataset [4] is selected as the benchmark dataset to evaluate the proposed method. The deep CNN models, which the proposed method is supposed to explain, are trained on the training set of the ImageNet dataset. The trained model then performs predictions on an image from the test set. ResNet [3], which is one of the most successful deep CNN models in recent years, is chosen as the deep CNN model to be explained. However, the proposed method can be applied to explaining any modern deep learning models.

The parameter settings are listed in Table 1. The GA parameters are set according to the EC community convention recommended in [2]. There is only one hyper-parameter specific to E-LIME, which is the number of superpixels ns, i.e. the number of interpretable features. The value of ns is set to 100 in the experiments on the ImageNet dataset.

4 RESULT ANALYSIS



(a) (b)

Table 2: Results

Image	Original fitness (probability)	Fitness (probability)		Training time (Seconds)	
		Best	**Mean ± Std**	**Shortest**	**Mean ± Std**
Samoyed	0.8241	0.9814	0.9585 ± 0.0165	20.1235	22.2195 ± 0.9624

1: Std means standard deviation.

Figure 3a shows the original image of a samoyed. In Figure 3a, a samoyed has a white and dense coat, with a pair of upright and prick ears, brown eyes and nose. Besides, this samoyed wears a red collar in its neck, and we can see a corner of the collar. On the left side of this image, there is a cabinet, where a box and a vase of flowers are above the cabinet. In addition, a hanger is behind this samoyed. After using E-LIME to classify and explain this image, as can be seen from Figure 3b, one of the most informative features is extracted as a local explanation, i.e. the shape and size of ears. As we know, samoyeds have upright and prick ears, and the size of each ear is usually larger than cats'.

Apart from that, other informative features related to eyes and nose are also extracted as local explanations. These extracted explanations could give users the confidence to believe in the prediction (i.e. a dog). Moreover, most of the background information, which may interfere with the prediction, is filtered, such as a vase of flowers and the hanger. This could further improve the probability of the prediction and convince users to believe this prediction (with higher confidence). Accordingly, after presenting the prediction as well as its local explanations to users, it is relatively easier to decide whether the prediction is trustworthy or not.

According to Table 2, E-LIME improves the probability on the image of Samoyed (increase by 0.1344, i.e. 13.44%), and thereby increases the confidence of users when determining whether the prediction is trustworthy or not. Table 2 also reports the training time of E-LIME. As can be seen from Table 2, the proposed method is more time-efficient than LIME (**10 minutes**).

5 CONCLUSIONS

The overall goal of the paper is to propose a novel GA-based method for evolving local interpretable model-agnostic explanations. The goal has been achieved and the proposed E-LIME has demonstrated its effectiveness and efficiency based on the experimental results.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34, 11 (2012), 2274–2282.

[2] Thomas Back. 1996. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms.* Oxford university press.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.* 1097–1105.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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.* 1135–1144.