A Pixel-Level Explainable Approach of Convolutional Neural Networks and Its Application

Haitao Zhang
Lanzhou University
Lanzhou, China
htzhang@lzu.edu.cn

Jing Wang
Lanzhou University
Lanzhou, China
wj2023@lzu.edu.cn

Ziyue Wang
Lanzhou University
Lanzhou, China
2565547858@qq.com

Ziyi Zhao
Lanzhou University
Lanzhou, China
zyzhao21@lzu.edu.cn

Zhuo Cheng*
Jiangxi Normal University
Nanchang, China
zhuo_cheng@126.com

ABSTRACT

Convolutional neural network (CNN) currently has been widely used to undertake the task of image classification. Unfortunately, a trained CNN model is a nonlinear system with high complexity, and the implicit decision knowledge carried by the CNN model is often difficult to be comprehended by humans. A feasible method to make human understanding of decision knowledge is to explain the classification basis of the trained CNN model. In order to solve the problem of insufficient interpretation accuracy of the existing methods, this paper presents a novel pixel-level explainable approach based on a guided symbolic execution strategy. A large number of experiments are conducted on the PyTorch team published CNN models, and the experimental results show that the presented approach is a 100% accurate technique for interpreting classification basis of input images on pixel-level compared the existing explainable methods. In addition, a scheme to enhance the adversarial robustness of CNN models is designed based on the presented explainable approach. The evaluation experiments show that the designed scheme provides an effective way to improve the adversarial robustness of the CNN models, and is a transferable technique in the CNN models that hold different structures.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**; • **Software and its engineering** → *Software design engineering*.

KEYWORDS

convolutional neural network, explanation, symbolic execution

ACM Reference Format:
Haitao Zhang, Jing Wang, Ziyue Wang, Ziyi Zhao, and Zhuo Cheng. 2024. A Pixel-Level Explainable Approach of Convolutional Neural Networks and Its Application. In *39th IEEE/ACM International Conference on Automated Software Engineering (ASE '24), October 27- November 1, 2024, Sacramento, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3691620.3695320

*Zhuo Cheng is the corresponding author.

ASE '24, October 27-November 1, 2024, Sacramento, CA, USA

ACM ISBN 979-8-4007-1248-7/24/10…$15.00
https://doi.org/10.1145/3691620.3695320

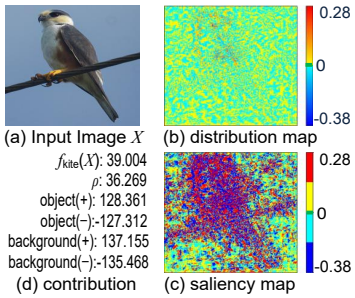

Figure 1: An example for demonstrating the interpretive result of guid-Xmage.

1 INTRODUCTION

Convolutional neural network (CNN), as an effective component with intelligent analysis, has been widely used to undertake the task of image classification. However, the classification knowledge carried by a trained CNN model is often difficult to be comprehended by humans as a result of the intricate network structure and vast parameters. Therefore, a trained CNN model is commonly referred as a black-box system. Due to the black-box property, a trained CNN model is usually considered as an untrustworthy system, which results in the fact that it is difficult to ensure whether a trained CNN model is safe. A feasible method to improve the trustworthiness of a CNN model is to explain the classification basis implied by the pixels of an input image so that the classification knowledge carried by the CNN model is clearly presented. Currently, many explainable methods have been proposed and made great contributions in the area of improving the performance of CNN models such as accuracy and robustness. However, as far as we know, there have no methods which hold 100% interpretive accuracy. This paper presents a novel explainable approach called guid-Xmage to address the problem of lack of interpretive accuracy in the existing methods. Two contributions are made: (*i*) it first introduces the symbolic execution technique into the field of interpreting classification knowledge of the CNN models; and (*ii*) the presented explainable approach

Table 1: Interpretive accuracy of grad-CAM, grad-CAM++, score-CAM and guid-Xmage on the different CNN models.

	AlexNet	VGGNet	ResNet
grad-CAM	62%	55%	62.5%
grad-CAM++	58%	48.5%	56.5%
score-CAM	58.5%	52.5%	57%
guid-Xmage	100%	100%	100%

provides an accurate observation instrument, which is useful to improve the robustness of CNN models in a more accurate way.

2 SYMBOLIC EXECUTION EXPLAINABLE APPROACH AND APPLICATION

The key processes of the presented explainable approach guid-Xmage are as the follows: an input image X is first fed to a target CNN model $f(X)$, and then the numerical relationship (>0 or $=0$) of each `ReLU` function as well as the selecting location of each `max-pooling` are collected according to the execution trails of the input real image X in the target model $f(X)$. Secondly, a symbolic image Z with the same size as the input image X is used to symbolically execute the target model $f(X)$ under the guidance of the collected execution trails. After that, assuming the t-th classification is the given destination of the symbolic execution, a symbolic model $f_t^s(Z)$ corresponding to the target CNN model $f(X)$ is obtained. Then, a weight matrix of pixels Δ_t with respect to the input image X is constructed through extracting the weights of symbolic variables from $f_t^s(Z)$. In the last step, guid-Xmage uses a classification contribution matrix of pixels A_t resulting from the dot-multiplication between the weight matrix Δ_t and the input image X to guide the perturbation computation for generating a minimized adversarial sample. Here, a distribution map and a saliency map is depicted to visualize the classification contribution matrix of pixels A_t. An example shown in Figure 1 is used to demonstrate how guid-Xmage achieves a visual presentation on the influence weights of pixels within an input X in relation to the classification value output by a target CNN model $f(X)$. In Figure 1, (a) is the input image X which is selected from ImageNet and (d) is the classification contribution values of different objects in the input image X. The example takes the top-1 classification branch of the PyTorch team published AlexNet model as interpretive destination. The distribution map (b) and the saliency map (c) consist shades in five colors which is (in descending order of the contributed values): red, yellow, green, cyan, and blue, where red-yellow labelled pixel contributes a positive value, green labelled pixel contributes a value of 0, and cyan-blue labelled pixel contributes a negative value to the classification value, respectively.

Moreover, drawing on the existing adversarial attack methods, a scheme with the ability of constraining the classification behavior of the CNN models is designed. The core idea of the designed scheme is as follows: firstly, keeping the object features of the original training samples X unchanged; secondly, according to the classification contribution matrix of pixels A associated with the top-1 classification result that is outputted by guid-Xmage, the background pixels of the original training samples X are iteratively perturbed in a monotonically decreasing manner under different perturbation strengths in order to generate a series of adversarial samples X' that hold the capability to reduce the background classification contribution value; finally, a set of adversarial samples X' with reduced background classification contribution value is added to the original training samples as a new training set, and then a CNN model with stronger adversarial robustness is obtained by retraining the original CNN model with the new training set.

3 EXPERIMENTS

In the conducted experiments, the PyTorch team published AlexNet, VGGNet and ResNet models trained with the ImageNet (ilsvrc-2012) are considered as target CNN models, and the images from the ImageNet are selected as evaluating dataset. The existing methods that hold the same interpretive purpose as guid-Xmage such as grad-CAM [4], grad-CAM++ [2] as well as score-CAM [5] are considered as the comparison techniques. As shown in Table 1, the experimental results show that guid-Xmage is a 100% accurate technique for interpreting classification bases of input images on pixel-level. In addition, based on the public dataset CIFAR-10, and the CNN models with different structures such as AlexNet, ResNet and DenseNet, experiments are conducted to evaluate the availability and transferability of the designed scheme for improving the adversarial robustness of the CNN models. Experiments show that the designed scheme can effectively improve the robustness of the CNN models against the famous attack methods such as C&W [1] and PGD [3] while keeping the accuracy almost invariant, and is a transferable technique in the different CNN models.

4 CONCLUSION

This paper presents a novel explainable approach guid-Xmage to interpret the classification bases of input images in CNN models. The experimental results show that guid-Xmage holds the 100% interpretive accuracy on pixel-level and also indicate the designed scheme can effectively improve the robustness of the CNN models.

ACKNOWLEDGMENTS

This work is partially supported by Jiangxi Provincial Department of Education Science and Technology Youth Project (Grant No. GJJ210340) and National Natural Science Foundation of China (Grant No. 62102171).

REFERENCES

[1] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy*. IEEE, San Jose, CA, USA, 39–57. https://doi.org/10.1109/SP.2017.49

[2] Aravind Chattopadhay, Anirban Sarkar, Prantik Howlader, and V. N. Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, Lake Tahoe, NV, USA, 839–847. https://doi.org/10.1109/WACV.2018.00097

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083* (2017). https://arxiv.org/abs/1706.06083

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Venice, Italy, 618–626. https://doi.org/10.1109/ICCV.2017.74

[5] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Seattle, WA, USA, 24–25. https://doi.org/10.1109/CVPRW50498.2020.00020