Sustainable Machine Intelligence Journal

Journal Homepage: sciencesforce.com/smij

SUSTAINABLE MACHINE INTELLIGENCE JOURNAL

SCIENCES FORCE
International Academic Publisher

Sustain. Mach. Intell. J. Vol. 8 (2024) 46–55

Paper Type: Original Article

VisionCam: A Comprehensive XAI Toolkit for Interpreting Image-Based Deep Learning Models

Walid Abdullah [1],*, Ahmed Tolba [1], Ahmed Elmasry [1] and Nihal N. Mostafa [2]

[1] Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt; Emails: waleed@zu.edu.eg; a.tolba24@fci.zu.edu.eg; a.elmasry24@fci.zu.edu.eg.

[2] Department of Computer Science,Misr higher institute for computer and commerce, Egypt; nihal.nabil@fci.zu.edu.eg.

Received: 20 Feb 2024 **Revised:** 08 May 2024 **Accepted:** 06 Jun 2024 **Published:** 11 Jun 2024

Abstract

Artificial intelligence (AI), a rapidly developing technology, has revolutionized various aspects of our lives. However many AI models' complex inner workings are still unknown, frequently compared to a "black box." Particularly in crucial fields, this lack of explainability (XAI) reduces responsible AI research and reduces public confidence, and is accompanied by a growing demand for transparency and interpretability in AI decision-making. In response, this paper introduces a Python Extensible Toolkit for Explainable AI (XAI), This toolkit comprises nine state-of-the-art techniques for explaining AI models (especially deep learning models) decisions in image processing: GradCAM, GradCAM++, GradCAMElementWise, HriesCAM, RespondCAM, ScoreCAM, SmoothGradCAM++, XgradCAM, and AblationCAM. Each tool offers unique insights into model decision-making processes of deep learning models that work with image data, addressing various aspects of interpretability. Through case studies, we demonstrate the toolkit's impact on improving transparency and interpretability in AI systems that analyze visual information. The source code for the VisionCam toolkit is accessible at https://github.com/VisionCAM.

Keywords: Explainable AI, XAI Toolkit, Interpretability, Machine Learning, Convolution Neural Network (CNN).

1 | Introduction

Artificial intelligence (AI) models such as deep learning models, have achieved remarkable success in various image-based applications [1]. Convolutional neural networks (CNNs) and other sophisticated models have revolutionized image reprocessing tasks such as medical diagnostics [2], autonomous driving [3], and security surveillance [4]. Despite their high performance, these models often operate as "black boxes," making it difficult to understand how they arrive at specific decisions or predictions [5]. This opacity poses significant challenges, especially in high-stakes fields such as healthcare where understanding the rationale behind AI decisions is crucial for trust, accountability, and effective human-AI collaboration [6].

The primary motivation for developing this XAI (Explainable AI) toolkit is to address the need for interpretability in image-based AI systems. Users in a variety of fields need tools that make these models' decision-making processes more explainable, making them more transparent and understandable [7]. By

Corresponding Author: waleed@zu.edu.eg

https://doi.org/10.61356/SMIJ.2024.8290

offering a suite set of state-of-the-art techniques specifically designed for image data, the VisionCam toolkit aims to provide the necessary explanations to directly address several key challenges in the deployment of image-based AI systems:

- Trust and Acceptance: “black-box” models can lead to a lack of trust in their image classification or object detection outputs, particularly in critical applications like medical diagnosis or autonomous vehicles.
- Model Biases Debugging: If a model consistently misclassifies a particular type of image, it's difficult to identify and rectify potential biases in the training data without explainability tools.
- Model Improvement: Understanding how models arrive at image-based decisions is crucial for improving their performance.

In this work, we produce a new XAI toolkit named VisionCam. This toolkit comprises nine state-of-the-art techniques specifically designed for explaining AI decisions while working with image data: GradCAM, GradCAM++, GradCAMElementWise, HriesCAM, RespondCAM, ScoreCAM, SmoothGradCAM++, XgradCAM, and AblationCAM. Each model aimed at providing clear and actionable insights into machine learning model behavior and decision-making processes addresses various aspects of interpretability, and addresses the scientific problem of bridging the gap between AI's predictive capabilities and its interpretability with human understanding, thus facilitating broader acceptance and deployment of AI technologies across diverse domains [8]. This can contribute to scientific discovery in image analysis in several ways as follows:

- Enhanced Transparency: Our toolkit empowers users to understand how deep learning models arrive at image-based decisions. This transparency fosters trust in image recognition and object detection systems, especially in critical applications like medical diagnosis or autonomous vehicles.
- Responsible AI Development: By promoting user understanding of model behavior in image analysis, our toolkit facilitates the responsible development of AI for image tasks. This ensures models are unbiased, fair, and aligned with human values, mitigating potential issues in image classification or object detection.
- Improved Human-AI Collaboration: Our toolkit bridges the gap between humans and deep learning models for image analysis. Users gain insights into how models interpret visual information, enabling effective collaboration for tasks like image segmentation and classification.
- Educational Value: The VisionCam toolkit serves as a valuable resource for researchers and practitioners working with explainable AI in image analysis. It provides a platform to explore and understand diverse XAI techniques specifically designed for image data.

The rest of the paper is structured as follows: Section 2 presents the most recent XAI methods and toolkits, and their features. The material and techniques used in these toolkits are described in detail in Section 3. The experimental Setup is shown in Section 4; The Experimental analysis and Discussion are presented in Section 5. Section 6 presents the implications of this work. The conclusion and future directions are given in Section 7.

2 | Related Work

The field of explainable artificial intelligence (XAI) has garnered significant attention over recent years, particularly in the context of deep learning models, which are often criticized for their black-box nature. Numerous techniques and toolkits have been developed to address the need for interpretability and transparency, especially for image-based AI models. These toolkits provide various methods to explain the decision-making processes of these models, but each has its limitations. Below, we review some prominent image-focused XAI toolkits, their features, and their gaps, followed by a comparison with our VisionCam toolkit, which includes nine state-of-the-art techniques specifically designed for image data. Table 1 presents some of the related toolkits and their techniques and methods.

Table 1. Examples of the Existing XAI Toolkits**.**

Title	Description	Limitations
SHAP [9]	A versatile toolkit offering model-agnostic explanations for various machine learning models. primarily designed for tabular and text data, it also has inspired the development of interpretable models for image data	- For image data, SHAP might require additional processing steps to translate explanations into a visually interpretable format - SHAP explanations can be difficult to interpret directly for complex image data
LIME [10]	Similar to SHAP, it is a popular tool that explains individual predictions of any classifier by approximating the model locally with an interpretable model.	- LIME's explanations might be less intuitive for complex deep-learning models used in image analysis - The quality of explanations can vary, sometimes leading to unstable or less accurate interpretations for image-based models.
AIX360 [11]	An open-source toolkit from IBM that offers a variety of XAI methods	- Though it is not specifically tailored for images
Integrated Gradients [12]	Is specifically designed for explaining deep learning models.	- integrated gradients might struggle with models having many layers, potentially leading to noisy explanations for complex image recognition tasks
Captum [13]	A comprehensive toolkit offering various XAI techniques, including Grad-CAM which is a popular method for visualizing the image regions. Designed for easy integration with PyTorch models	- Captum offers a limited number of XAI techniques specifically designed for images compared to VisionCam - Limited support for non-PyTorch frameworks, reducing its flexibility for users working with other deep learning libraries
DALEX [14]	It is a package in R and Python that provides a set of tools for visualizing and explaining predictive models. Offers functionalities for a broad range of models and data types	- DALEX's visualizations are more suited for tabular data, with limited capabilities for detailed image data interpretations. - It lacks advanced image-specific methods provided by images XAI toolkits.

VisionCam offers several advantages over existing toolkits. it is a Comprehensive toolkit that integrates nine diverse XAI techniques specifically designed for image data. This diversity allows users to choose the most appropriate explanation method for their specific image analysis task. In addition, VisionCam leverages visualization techniques like heat maps and saliency maps. This specialization, combined with user-friendly features and comprehensive support for major deep learning frameworks prioritizes user experience. It offers different ways to select the target layer within the model for explanation generation to enhance the user experience.

3| Materials and Methodology

VisionCam toolkit comprises nine distinct XAI tools, these models can easily integrate with deep learning models. each tailored to address specific challenges in deep learning model interpretability. offering a comprehensive suite of techniques for visualizing and understanding image-based deep-learning model decisions.

3.1| Software Architecture

VisionCam toolkit is designed with a good architecture enabling the ease of deep learning model integrations and producing the outputs to the users and stakeholders, increasing flexibility and ease of use. Figure 1 shows the pipeline architecture of the VisionCam toolkit.

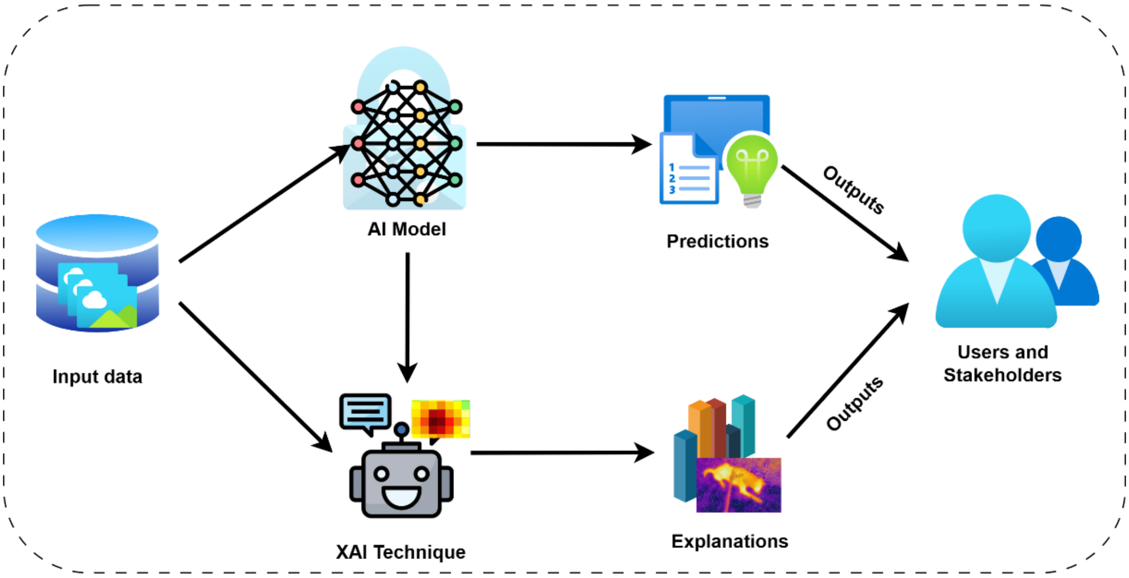


Figure 1. The Pipeline Architecture of VisionCam Toolkit.

3.2 | XAI Techniques

XAI techniques This component houses the core functionalities of the toolkit. It implements the various techniques (e.g., GradCAM, RespondCAM) as independent modules. Each module receives a user-specified target layer or function to identify the relevant feature representations for explanation generation. The following is a brief description of all used XAI techniques:

- GradCAM (Gradient-weighted Class Activation Mapping): Visualizes the regions of an input image that contribute most to the model's prediction for a specific class [15].
- GradCAM++: An extension of GradCAM that addresses limitations in the original method, particularly for the rectified linear unit (ReLU) activation functions [16].
- GradCAMElementWise: Generates element-wise attribution maps, providing a better understanding of the contributions of individual pixels [17].
- HiResCAM (High-Resolution Class Activation Mapping): Achieves higher resolution visualizations compared to GradCAM, especially for models with smaller receptive fields[18].
- RespondCAM: A model-agnostic technique that utilizes gradients to compute saliency maps, explaining the model's response to a specific input [19].
- ScoreCAM: Integrates gradient information with the model's final score to generate class activation maps [20].
- SmoothGradCAM++: Enhances the stability of GradCAM by averaging gradients computed with noisy inputs [21].
- XGradCAM (Axiom-based Grad-CAM): Extends GradCAM to handle models with multiple branches or outputs [22].

- AblationCAM: Estimates the contribution of each feature by systematically removing parts of the input and observing the change in the model's output [23].

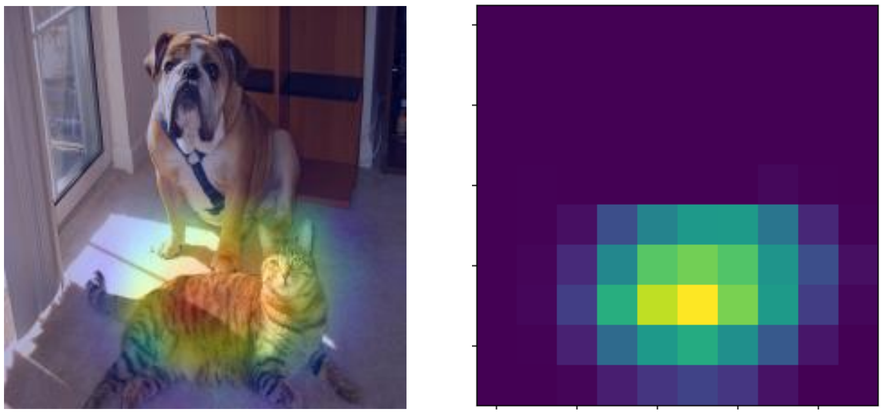
These methods leverage various strategies, including gradient-based visualization, class activation mapping, and sensitivity analysis, to highlight the salient features and regions contributing to AI predictions. The implementation of each method is meticulously crafted to ensure efficiency, extensibility, and compatibility with popular deep-learning models such as CNN.

3.3 | Visualization Tools

The visualization tools component takes the explanation data generated by the XAI technique and renders it into human-interpretable visualizations. for image data, VisionCam toolkit techniques can provide the user with two methods of visualization:

- Heatmaps: are designed to represent data values in a matrix format using colors, allowing for a quick visual understanding of data density or intensity across a given area.
- Saliency maps: are specifically used to highlight the most important features or regions of an input (such as an image) that have the greatest influence on a model's prediction.

These visualization tools are then displayed for user explanation, highlighting the areas with the most significant influence on the model's prediction. Figure 2 present an example of Heatmap and Saliency map visualization for a detected object (Tiger Cat) with an image.



(B) (A)

Figure 2. Example of visualization tools: (A) Heatmap; (B) Saliency map.

4| Experimental Setup

4.1| Utilized Tools and Technologies

In developing the VisionCam toolkit, a variety of tools and technologies have been used as follows:

- Programming Language (Python v.3.10)
- Deep Learning Frameworks (TensorFlow v.2.15)
- Visualization Libraries (Matplotlib, OpenCV)

4.2| Experimental Setting

Users can interact with the VisionCam toolkit with a user-friendly experience designed for integration with a popular deep learning framework (TensorFlow). The workflow typically involves the following steps:

1. Installing the toolkit (if it isn't installed).
2. Model Loading: Users load their pre-trained image-based models into the toolkit.
3. Data Input: Users input the image data they wish to analyze.
4. Method Selection: Users select from the nine available XAI methods based on their specific needs. Parameter Configuration: Users can configure various parameters for the chosen methods to customize the explanations such as specify the target layer.
5. Explanation Generation: The toolkit computes explanations and generates visualizations (heatmaps, saliency maps) depicting the model's decision-making process.
6. Analysis and Export: Users can analyze the generated explanations, and gain insights into the model's behavior, compare different methods, and export the visualizations for further use.

In this example, we will use a model trained on the popular ImageNet dataset and apply several of our toolkit's techniques to explain its predictions. In this example, we want the model to illustrate why it thinks that the image label is “tiger_cat”.

Step 1: Installing the VisionCam toolkit
First, If VisionCam is not already installed within your environment, you can use the Python Package Index (PyPI) for installation [24]. Here's the command to install it.

```
!pip install visioncam==0.0.1
```

Step 2: Model Loading and Image Input
load the Deep learning model such as the pre-trained CNN model and the input image of a tabby cat. The model is an Xception, a widely used architecture known for its robust performance on image classification tasks.

```
import keras
from keras.applications import Xception
model = Xception(weights="imagenet")
preds = model.predict(image)      #image is holds the image input
```

Step 3: XAI Technique Choice from VisionCam toolkit.
Users select from the nine available XAI methods based on their specific needs, in the following code snippet, GradCam is selected to identify the image regions contributing most to the model's prediction of a malignant tumor. Users can follow this format for selecting any specific XAI method from VisionCam (“ from *visioncam.methods_name* import *Method_Name* “).

```
from visioncam.gradcam import GradCAM
grad_cam = GradCAM(model , class_index['tiger_cat '] )
```

Step 4: Parameter Configuration: Users can configure various parameters for the chosen methods to customize the explanations such as specifying the target layer.

```
last_conv_l_name = "block14_sepconv2_act"
grad_model = gradcam = GradCAM(model, target_layer= last_conv_l_name)
```

Step 5: Explanation Generation: The toolkit computes explanations and generates a visual saliency map depicting the model's decision-making process. This saliency map highlights the areas with the most significant influence on the model's prediction

```
#preparing image for pre-trained model
```

```
img_array = preprocess_input(image)
grad_heatmap = grad_cam.compute_cam_features(img_array)
# Plot the heatmap
plt.imshow(grad_heatmap)
# Display heatmap on the original image to produce the saliency map
XgradCAM.save_and_display_gradcam('/content/sample.jpg', grad_heatmap)
```

5| Experimental Analysis and Discussion

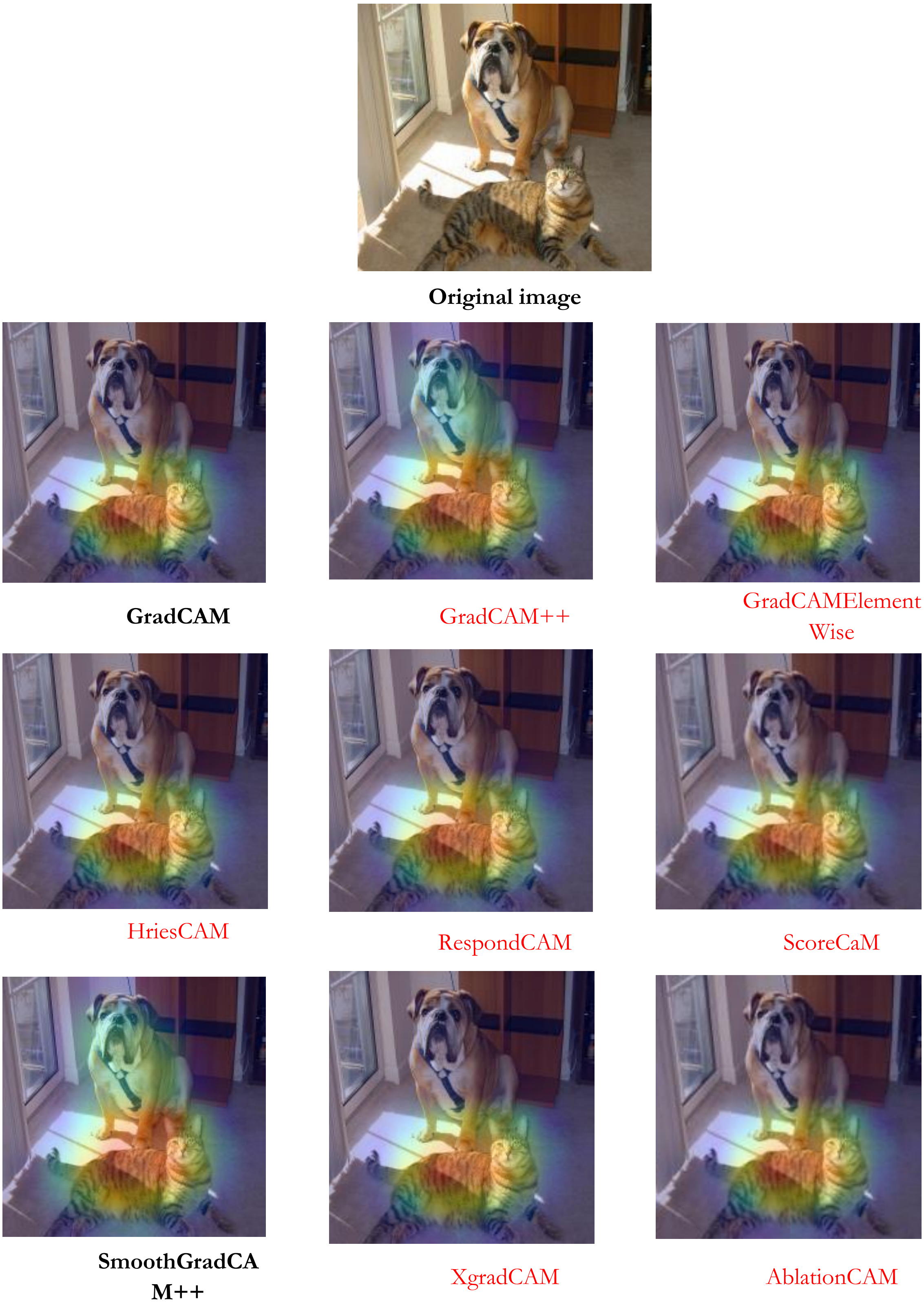
VisionCam toolkit offers invaluable insights into the decision-making processes of diverse AI models. Through visually compelling demonstrations and quantitative analyses, we elucidate how each tool in our toolkit can be utilized to uncover hidden patterns, identify potential biases, and improve the overall transparency of AI systems. to demonstrate the major functions and the efficiency of our XAI toolkit in practice, in section 4.1, we provide a detailed example by employing the transfer learning model Xception a pre-trained convolutional neural network (CNN) for image classification. Then XAI tool should illustrate why the model thinks that the image label is “tiger_cat”. It should compute and generate a saliency map, which highlights the areas with the most significant influence on the model's prediction users can analyze the generated explanations, gain insights into the model's behavior, compare different methods, and export the visualizations for further use. Figure 3 presents various images produced by various VisionCam toolkits' XAI techniques.

6| Implications

VisionCam's focus on producing a set of diverse and user-friendly XAI techniques for image data has the potential to significantly impact the field of computer vision and holds significant implications across diverse domains. Here's The major impact of this toolkit on both research and practical applications in various fields:

VisionCam can facilitate the exploration of novel research questions related to the exploration of model bias and explainability in image analysis models. For instance, researchers can investigate how different XAI techniques influence human understanding of model behavior or how explainability impacts trust and acceptance of AI models in specific image analysis tasks. Furthermore, the availability of multiple XAI techniques allows for comparative studies on their effectiveness and reliability, leading to deeper insights into the strengths and weaknesses of each method. It can also help in Improving the existing research: By providing clear explanations for model decisions, it can significantly improve the pursuit of existing research questions. Researchers can gain deeper insights into model behavior, leading to more robust and trustworthy findings in areas like medical image analysis or autonomous vehicle perception. In addition, it helps in machine learning model validation and performance improvement. By providing detailed visual explanations, researchers can better validate the behavior of their models, ensuring that they operate as intended and identifying areas for improvement.

Our XAI toolkit also can help in Building Trust in AI Models: providing explanations for model predictions is crucial for real-world applications where transparency and user confidence are paramount, such as healthcare diagnostics or autonomous systems. In medical imaging, for example, the toolkit aids researchers in verifying model predictions, potentially leading to more accurate and reliable diagnostic tools. Furthermore, the toolkit's influence extends to commercial settings, where it has been adopted in various industries such as Healthcare and Autonomous Vehicles by providing interpretable diagnostic tools, enhancing the reliability of AI-driven diagnoses, and improving the safety and transparency of their autonomous systems.



Original image

GradCAM

GradCAM++

GradCAMElement Wise

HriesCAM

RespondCAM

ScoreCaM

**SmoothGradCA
M++**

XgradCAM

AblationCAM

Figure 3. Original image and different explanations produced and generated using Various VisionCam toolkit's XAI techniques.

7 | Conclusion and Future Work

The expanding field of deep learning has revolutionized image analysis tasks. However, the lack of explainability in image recognition and object detection models hinders trust, hinders model improvement, and makes it difficult to identify potential biases. In this work, we have introduced a comprehensive XAI toolkit tailored specifically for image-based AI models called VisionCam. This toolkit integrates nine state-of-the-art techniques—GradCAM, GradCAM++, GradCAMElementWise, HriesCAM, RespondCAM, ScoreCAM, SmoothGradCAMPlusPlus, XgradCAM, and AblationCAM—to address the critical need for interpretability and transparency in complex AI systems. Each method offers unique advantages and perspectives, allowing users to gain deeper insights into their models' decision-making processes. VisionCam is a user-friendly toolkit that empowers researchers, developers, and practitioners to gain valuable insights into how deep learning models arrive at image-based decisions. This fosters transparency, facilitates collaboration between humans and AI, and paves the way for the responsible development of explainable AI in image analysis.

Finally, as AI continues to advance, the need for explainable models will become increasingly important. Future developments of our toolkit will focus on Expanding XAI Techniques and working with different data types.

Acknowledgments

The authors are grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributions

All authors contributed equally to this work.

Funding

This research was conducted without external funding support.

Data Availability

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Shinde, P.P. and S. Shah. A review of machine learning and deep learning applications. in 2018 Fourth international conference on computing communication control and automation (ICCUBEA). 2018. IEEE.

[2] Yadav, S.S. and S.M. Jadhav, Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big data, 2019. **6**(1): p. 1-18.

[3] Uçar, A., Y. Demir, and C. Güzeliş, Object recognition and detection with deep learning for autonomous driving applications. Simulation, 2017. **93**(9): p. 759-769.

[4] Sabri, Z.S. and Z. Li, Low-cost intelligent surveillance system based on fast CNN. PeerJ Computer Science, 2021. **7**: p. e402.

[5] Loyola-Gonzalez, O., Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE access, 2019. **7**: p. 154096-154113.

[6] Castelvecchi, D., Can we open the black box of AI? Nature News, 2016. **538**(7623): p. 20.

[7] Arrieta, A.B., et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 2020. **58**: p. 82-115.

[8] Choung, H., P. David, and A. Ross, Trust in AI and its role in the acceptance of AI technologies. International Journal of Human–Computer Interaction, 2023. **39**(9): p. 1727-1739.

[9] Lundberg, S.M. and S.-I. Lee, A unified approach to interpreting model predictions, in Advances in neural information processing systems. 2017. https://arxiv.org/abs/1705.07874.

[10] Ribeiro, M.T., S. Singh, and C. 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. p. 1135-1144. https://paperswithcode.com/method/lime.

[11] Arya, V., et al., Ai explainability 360 toolkit, in Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD). 2021. p. 376-379. https://aif360.res.ibm.com/.

[12] Ma, X., et al., Adversarial generation of real-time feedback with neural networks for simulation-based training, in arXiv preprint arXiv:1703.01460. 2017. https://arxiv.org/abs/1703.01460.

[13] Kokhlikyan, N., et al., Captum: A unified and generic model interpretability library for pytorch, in arXiv preprint arXiv:2009.07896. 2020. https://captum.ai/.

[14] Biecek, P., DALEX: Explainers for complex predictive models in R, in Journal of Machine Learning Research. 2018. p. 1-5. https://dalex.drwhy.ai/.

[15] Selvaraju, R.R., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. in Proceedings of the IEEE international conference on computer vision. 2017.

[16] Chattopadhay, A., et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. in 2018 IEEE winter conference on applications of computer vision (WACV). 2018. IEEE.

[17] Pillai, V. and H. Pirsiavash. Explainable models with consistent interpretations. in Proceedings of the AAAI Conference on Artificial Intelligence. 2021.

[18] Draelos, R.L. and L. Carin, Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. arXiv preprint arXiv:2011.08891, 2020.

[19] Han, Y. and F. Meng. A Saliency-based Weakly-supervised Network for Fine-Grained Image Categorization. in 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2020.

[20] Wang, H., et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.

[21] Omeiza, D., et al., Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. arXiv preprint arXiv:1908.01224, 2019.

[22] Fu, R., et al., Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. arXiv preprint arXiv:2008.02312, 2020.

[23] Ramaswamy, H.G. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. in proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020.

[24] VisionCami. 2024, Pypi. https://pypi.org/project/visioncam/0.0.1/.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.