Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks

BOSCH

Jörg Wagner[1,2], Jan Mathias Köhler[1], Tobias Gindele[1], Leon Hetzel[1], Jakob Thaddäus Wiedemer[1], Sven Behnke[2]

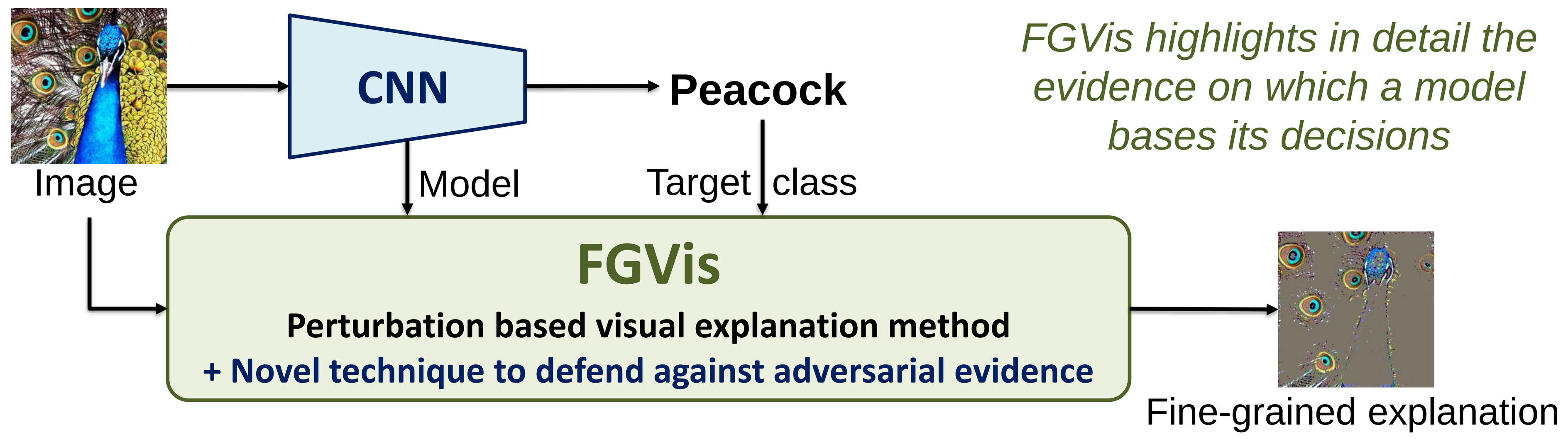
[1]Bosch Center for Artificial Intelligence [2]University of Bonn

UNIVERSITÄT BONN

1 Overview

Motivation: A better understanding of the decision-making process of a CNN is required to provide hints for improving it. This allows to uncover and understand failure cases, limitations of the model, and shortcomings of the training data.

Fine-grained visual explanation method (FGVis):



FGVis highlights in detail the evidence on which a model bases its decisions

Contributions:

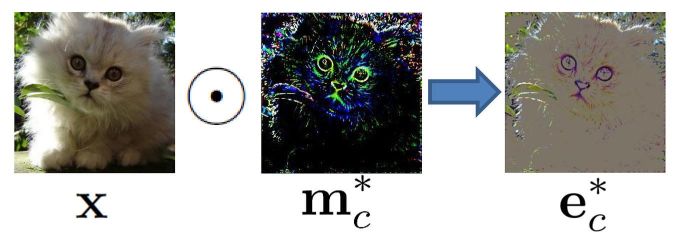
- A method (**FGVis**) to generate **fine-grained explanations** in the image space.
- A **novel technique for defending against adversarial evidence**, which does not depend on human-tuned parameters.
- **Interpretable** and **class discriminative** explanations, visualizing **detailed evidence**.

2 Perturbation based explanation methods

An explanation $\mathbf{e}_c^*$ is computed by perturbing the input image $\mathbf{x}$.

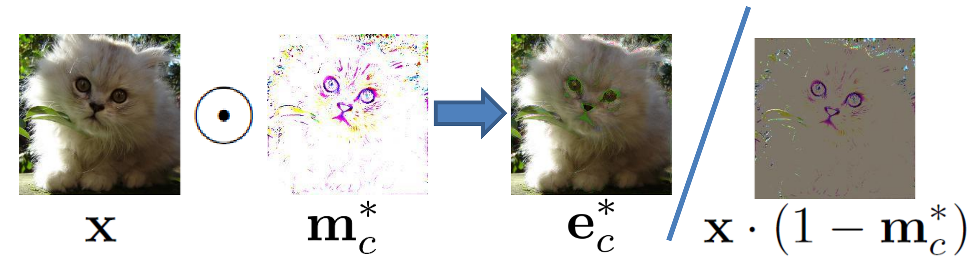
Mask $\mathbf{m}_c^*$ based perturbation: $\mathbf{e}_c^* = \mathbf{x} \cdot \mathbf{m}_c^*$ | c : Target class of the explanation

Preserving explanation:



$$\mathbf{m}_c^* = \underset{\mathbf{m}_c \in [0,1]^{3\times H\times W}}{\arg\min} \; \lambda \cdot \|\mathbf{m}_c\|_1 - \mathrm{CNN}_c(\mathbf{e}_c)$$

Deleting explanation:



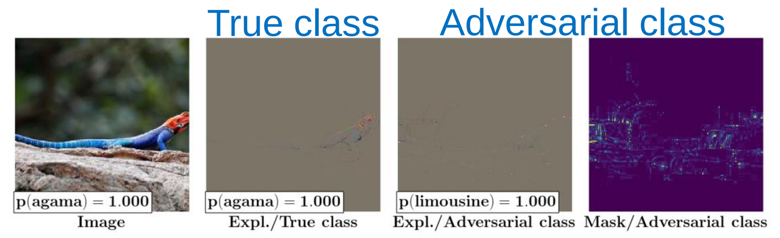
$$\mathbf{m}_c^* = \underset{\mathbf{m}_c \in [0,1]^{3\times H\times W}}{\arg\max} \; \lambda \cdot \|\mathbf{m}_c\|_1 - \mathrm{CNN}_c(\mathbf{e}_c)$$

Probability of target class c for explanation $\mathbf{e}_c$

Perturbation based explanations represent valid model inputs and are thus testable

3 Defending against adversarial evidence

Drawback of perturbation based methods: Adversarial evidence, i.e. faulty evidence due to artefacts introduced in the optimization of the explanation.



Model	*VGG16*	*AlexNet*	*ResNet50*	*GoogleNet*
Accuracy	100.0 %	100.0 %	100.0 %	100.0 %

Accuracy of adversarial expl. on 1000 random images

Without defense the optimization introduces adversarial evidence

Novel adversarial defense:

- **Idea:** The features in an explanation should be a subset of the image features.
- **Corresponding optimization constraint:**

$$\begin{cases} 0 \le h_i^l(\mathbf{e}_c) \le h_i^l(\mathbf{x}), & \text{if } h_i^l(\mathbf{x}) \ge 0, \\ 0 \ge h_i^l(\mathbf{e}_c) \ge h_i^l(\mathbf{x}), & \text{otherwise}, \end{cases}$$

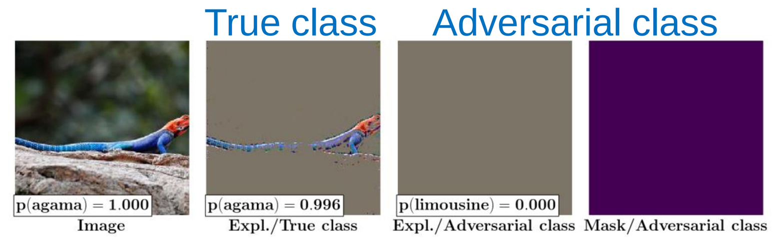
  - $h_i^l(\cdot)$: Activation of the i-th neuron in the l-th layer.
  - The constraint is applied after each nonlinearity-layer (e.g.: ReLU-Layer).

- **Implemented via gradient clipping:**

Indicator function

$$\gamma_i^l = \bar{\gamma}_i^l \cdot \mathbb{1}[h_i^l(\mathbf{e}_c) \le \max(0, h_i^l(\mathbf{x}))] \cdot \mathbb{1}[h_i^l(\mathbf{e}_c) \ge \min(0, h_i^l(\mathbf{x}))]$$

Updated error-signal back-propagated through the l-th layer



Model	*VGG16*	*AlexNet*	*ResNet50*	*GoogleNet*
Accuracy	0.2 %	0.0 %	0.0 %	0.0 %

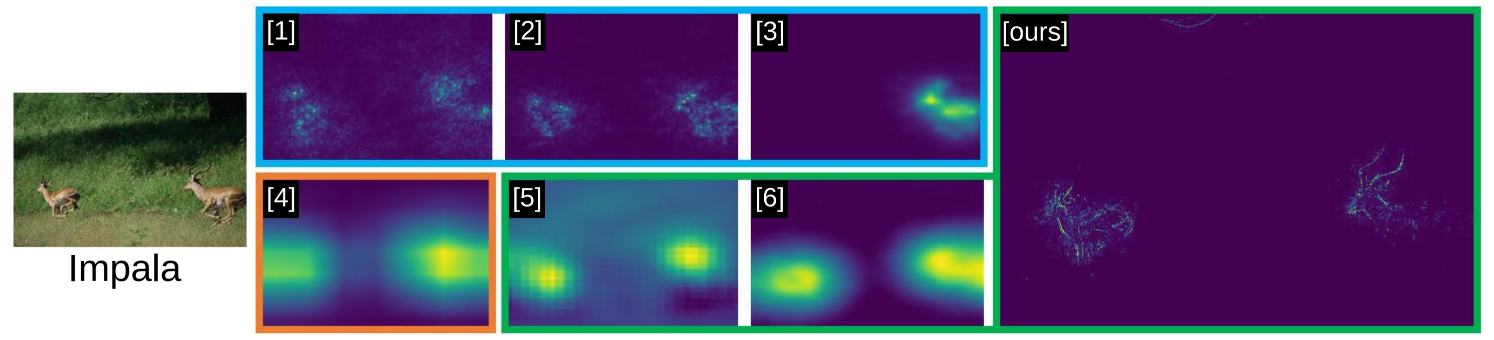
Accuracy of adversarial expl. on 1000 random images

Our defense prevents the hallucination of adversarial evidence

Our defense does not depend on human-tuned parameters and enables an explanation which is both fine-grained and preserves the characteristics of the image

4 Experiments

Qualitative comparison with other methods



Backpropagation based methods [1,2,3]

Activation based method [4]

Perturbation based methods [5, 6, ours]

FGVis generates the most fine-grained explanation mask

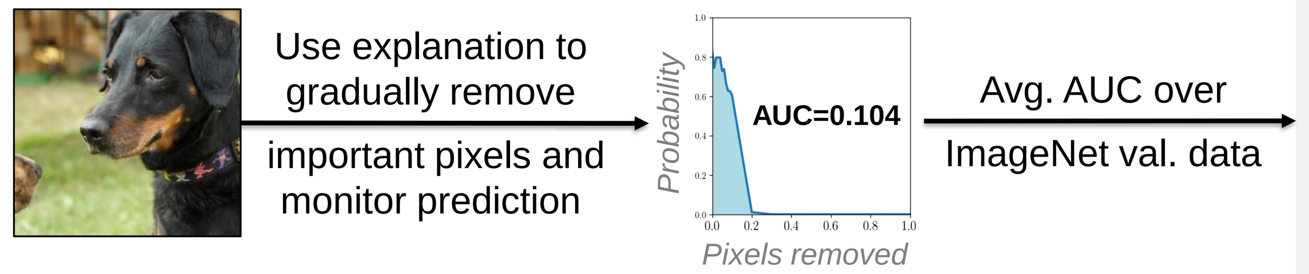
5 Experiments

Class discriminative / fine-grained

Explanation masks for images with multiple objects computed for two different target classes.

FGVis produces class discriminative explanations even when objects partially overlap

Faithfulness of explanations: How accurate does an explanation represent the true evidence on which a model bases its decision?



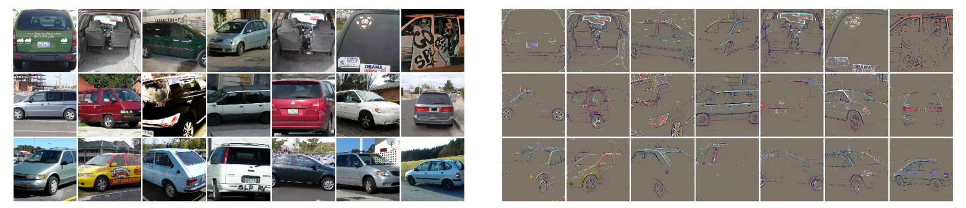
Method	*ResNet50*	*VGG16*
Grad-Cam [4]	0.1232	0.1087
Sliding Window [5]	0.1421	0.1158
LIME [8]	0.1217	0.1014
RISE [7]	0.1076	0.0980
FGVis [ours]	**0.0644**	**0.0636**

Deletion metric [7]

Color bias of *VGG16* trained on *ImageNet*:

Yellow is dominant in most explanations of the class *school bus*

FGVis can provide a first indication for the importance of different colors



Explanations of class *minivan* focus on edges, not consistently preserving the color

Quantitative verification: Ratio of maintained true classifications after swapping the color channels

Class	BGR	RBG	GRB	Avg. RBG, GRB
school bus	100 %	9.5 %	7.1 %	8.3 %
minivan	100 %	71.4 %	95.2 %	83.3 %

6 References

[1] Karen Simonyan *et al.* (ICLR, 2014). Deep inside convolutional networks: Visualising image classification models and saliency maps.

[2] Jost T. Springenberg *et al.* (ICLR, 2015). Striving for simplicity: The all convolutional net.

[3] Jianming Zhang *et al.* (ECCV, 2016). Top-down neural attention by excitation backprop.

[4] Ramprasaath R. Selvaraju (ICCV, 2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization.

[5] Matthew D. Zeiler and Rob Fergus (ECCV, 2014). Visualizing and understanding convolutional networks.

[6] Ruth C. Fong and Andrea Vedaldi (ICCV, 2017). Interpretable explanations of black boxes by meaningful perturbation.

[7] Vitali Petsiuk *et al.* (BMVC, 2018). Rise: Randomized input sampling for explanation of black-box models.

[8] Marco T. Ribeiro *et al.* (SIGKDD, 2016). Why should I trust you?: Explaining the predictions of any classifier.