Generating visual explanations from deep networks using implicit neural representations

Michal Byra[1,2,3], Henrik Skibbe[2]
[1]Institute of Fundamental Technological Research, Polish Academy of Sciences, Poland
[2]RIKEN Center for Brain Science, Japan
[3]Samsung AI Center Warsaw, Poland
mbyra@ippt.pan.pl, henrik.skibbe@riken.jp

Abstract

Explaining deep learning models in a way that humans can easily understand is essential for responsible artificial intelligence applications. Attribution methods constitute an important area of explainable deep learning. The attribution problem involves finding parts of the network's input that are the most responsible for the model's output. In this work, we demonstrate that implicit neural representations (INRs) constitute a good framework for generating visual explanations. Firstly, we utilize coordinate-based implicit networks to reformulate and extend the extremal perturbations technique and generate attribution masks. Experimental results confirm the usefulness of our method. For instance, by proper conditioning of the implicit network, we obtain attribution masks that are well-behaved with respect to the imposed area constraints. Secondly, we present an iterative INR-based method that can be used to generate multiple non-overlapping attribution masks for the same image. We depict that a deep learning model may associate the image label with both the appearance of the object of interest as well as with areas and textures usually accompanying the object. Our study demonstrates that implicit networks are well-suited for the generation of attribution masks and can provide interesting insights about the performance of deep learning models.

1. Introduction

Neural networks have achieved remarkable performance in various computer vision problems. However, explaining deep learning models in a way that humans can easily understand is essential for various applications, especially in medical fields [38]. Despite their excellent performance, deep neural networks struggle with the well-known 'black box' problem, which undermines confidence in the network's predictions. Methods for explainable artificial intelligence have been intensively studied to help better understand the logic behind the network's predictions. Explainable methods have the potential to detect subtle classification errors, enabling the addressing of unexpected and unwanted behaviors of the networks, and consequently help develop more efficient and trustworthy deep learning models.

Attribution methods constitute an important area of explainable deep learning. The attribution problem involves finding parts of the network's input that are the most responsible for the model's output. In studies on convolutional networks, attribution methods commonly compute saliency maps that highlight input image regions important for the output. Saliency maps assign a score related to prediction importance to each pixel of the input image. The most basic attribution approach is based on perturbing input image pixels and determining the effect of that change on the prediction. Clearly outlining the regions important for the prediction, pointing out the desired objects, is vital for increasing confidence in deep learning methods. However, as highlighted in previous studies, perturbation-based methods are associated with several challenges [10]. Firstly, perturbing all possible combinations of image pixels is infeasible from the computational point of view. As a remedy, the importance mapping is commonly treated as an optimization problem, with carefully designed loss functions and constraints ensuring plausible perturbations. For instance, to avoid adversarial effects, the extremal perturbations technique aims to generate perturbations that have a specific smoothness and size [10]. Secondly, attribution methods usually aim to determine small perturbations that have a potentially large impact on the network's prediction. However, as presented in recent studies, multiple independent explanations might exist for a single image, and determining them might provide additional insight about the deep learning model [23, 27].

Implicit neural representations (INRs) have recently gained attention in computer vision and medical image analysis (see the Related Work section for a list of potential applications). INRs serve as a continuous, nonlinear, and coordinate-based approximation of the target quantity obtained through a multi-layer perceptron (MLP). Due to this flexibility, implicit networks are especially well-suited for representing complex mappings, for instance representing objects of variable geometry [20]. Moreover, implicit networks can leverage custom objective loss functions for optimization, jointly addressing various tasks such as image reconstruction and inpainting [19]. Due to this versatility, implicit networks have been used to address various complex problems, often requiring case-by-case optimization, which would be difficult to tackle using standard optimization algorithms or convolutional networks that demand large volumes of training data [31].

In this work, we explore the use of implicit networks for explainable deep learning. As far as we know, implicit networks have not been yet used to generate visual explanations. Our main contributions are as follows:

- We demonstrate that INRs constitute a good framework for generating explanations. In comparison to attribution techniques based on standard optimization procedures, implicit networks provide a convenient way to consider non-linear and continuous relationships between the input image pixel coordinates and their importance for the model's prediction. Moreover, implicit networks can be trained using complex custom loss functions, enabling the association of the attribution mapping problem with other computer vision tasks, such as segmentation.
- We use INRs to reformulate and extend the extremal perturbations technique [10]. Originally, this technique was introduced to determine attribution masks obtained with an optimization procedure to output a mask that is smooth and covers a specific pre-defined area. However, the optimization procedure has to be repeated for each area constraint to determine mask expansion, which leads to masks that are not continuous with respect to the area constraint. Here, we present that this problem can be mitigated with properly conditioned implicit networks, see Fig. 1.
- We present an iterative method based on INRs that can be used to generate multiple explanations for the same input image. This is achieved by utilizing a segmentation-related loss function for the training of the implicit network, which ensures that the newly generated attribution masks do not overlap with the previous explanations. We demonstrate that this approach can provide useful insights about mechanisms guiding network predictions.

2. Related Work

Explainable artificial intelligence is an intensive area of research in computer vision and medical fields. Below, we discuss the prototypical attribution methods for the selected families of techniques. For a more detailed description of the attribution methods, we refer to one of the recent review papers [2, 25, 35].

2.1. Attribution methods

2.1.1 Perturbation-based methods

This family of attribution methods aims to occlude the input image with different types of perturbations and then assess the resulting change in the model's output. For instance, the model-agnostic meaningful perturbations technique optimizes a spatial perturbation mask indicating the image region that maximally affects the output of the model [11]. The extremal perturbations method extends the meaningful perturbations framework by introducing additional mask smoothing factors and an area loss function [10]. Leveraging the meaningful perturbation approach, a U-Net-like masking model was trained on ImageNet to generate attribution masks [7]. While fast at inference, this method requires large volumes of training data, making it infeasible to apply for small datasets, or associating the mask generation problem with other tasks. Moreover, the RISE technique probes the deep learning model with randomly masked versions of the input image to determine image regions important for the predictions and derive a saliency map [22].

2.1.2 Activation-based methods

These attribution techniques utilize network weights and activations at specific layers to generate saliency maps. For instance, the prototypical Class Activation Map (CAM) technique combines the weights of the last layer with its activations to compute a low-resolution saliency map [46]. Various approaches have been proposed to improve the CAM method, including the popular GradCAM algorithm [26], as well as other extensions, such as AblationCAM [24], CAMERAS [14], GradCAM++ [6], and Score-CAM [39], to name a few. Leveraging CAM techniques and perturbation-based approaches, in OptiCAM, the saliency map is optimized on a per-case basis by combining the weights of the model and the activations via a standard numerical procedure [45].

2.1.3 Propagation-based methods

This family of techniques generates saliency maps based on gradients back-propagated from the selected model's output to the input. In the most basic approach, the back-propagated gradient is considered as a saliency map [28]. In

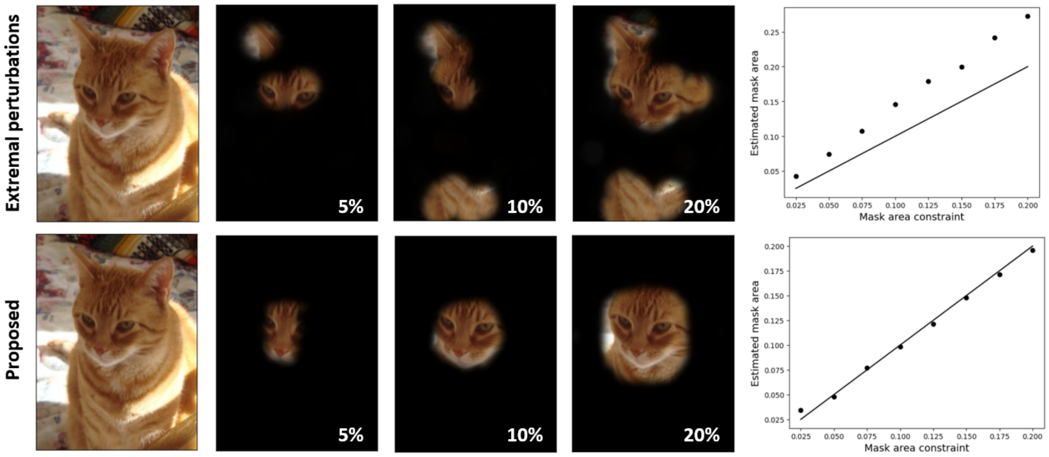

Figure 1. A comparison between the extremal perturbations technique and the proposed attribution method based on implicit networks, which due to the conditioning mechanism ensures more continuous and well-behaved attribution mask with respect to the mask area constraint. Percentage indicates the area of the attribution mask.

the guided back-propagation technique, the backward pass for the ReLU activation functions is modified to enhance the gradient flow and, consequently, the saliency map. DeconvNet utilizes deconvolution operations to map prediction-related activations back through the network to the input image space [33]. More recent methods are based on layer-wise relevance propagation [3] or various approaches to gradient computations, such as in SmoothGrad [32].

2.2. Implicit neural representations

Coordinate based implicit networks have proven to be an efficient methods for various problems encountered in computer vision, such as view synthesis [16], image reconstruction [31], signal processing [42], shape modelling [43], image stylization [9] and image generation [29], to name a few applications. In medical image analysis, neural implicit segmentation functions have been proposed for cardiac segmentation in magnetic resonance imaging [34], as well as for image registration [40], image decomposition [5] and vascular modelling [1]. In the context of the model interpretability, various studies have been conducted to better understand the mechanics of implicit networks, for example by interpreting implicit networks as Fourier series [4] or examining INRs using neural tangent kernel [13, 44]. For more applications of INRs, we refer to review papers [18, 41].

3. Methods

3.1. Perturbation analysis with implicit networks

In this section, we describe our approach to the attribution mask generation with INRs in the context of the extremal perturbations technique.

Let Φ stand for the deep learning model we wish to examine with the attribution method. Given a color image $I \in \mathbb{R}^2 \to \mathbb{R}^3$, let $\Phi(I) \in [0, 1]$ indicate the post-softmax probability that the input image belongs to the category of interest. In addition, let $\mathbf{x} \in \mathbb{R}^2 \to [0, 1]^2$ stand for the coordinates of the input image pixels defined on a normalized 2D grid, with $x \in [0, 1]^2$ indicating the 2D coordinate of a single pixel. The perturbation analysis deals with finding a mask M which assigns to each input pixel a value $M(x) \in [0, 1]$, where $M(x) = 1$ indicates that a pixel is important for the prediction and $M(x) = 0$ otherwise. Ideally, the mask should point out a coherent part of the image that contributes to the prediction. To assess the importance of the region corresponding to the mask, we modify the input image according to the following equation:

$$\hat{I} = M \otimes I + (1 - M) \otimes I', \tag{1}$$

where $\otimes$ is the Hadamard product and I' stands for the perturbed image, commonly obtained using either the Gaussian blur perturbation or the fade-to-black zero matrix perturbation.

Following the extremal perturbation method, we consider finding the mask M as an optimization problem, associated with the minimization of the following loss function [10]:

$$\mathcal{L}_{\text{ext}}(M) = -\Phi\big(M \otimes I + (1 - M) \otimes I'\big) + \lambda_r R_a(M). \tag{2}$$

The term $R_a(M)$ is a regularization function that enforces the mask area to be equal to a, and λ_r stands for the weighting parameter. To constrain the area, the values of the mask M are vectorized and sorted in increasing order to form

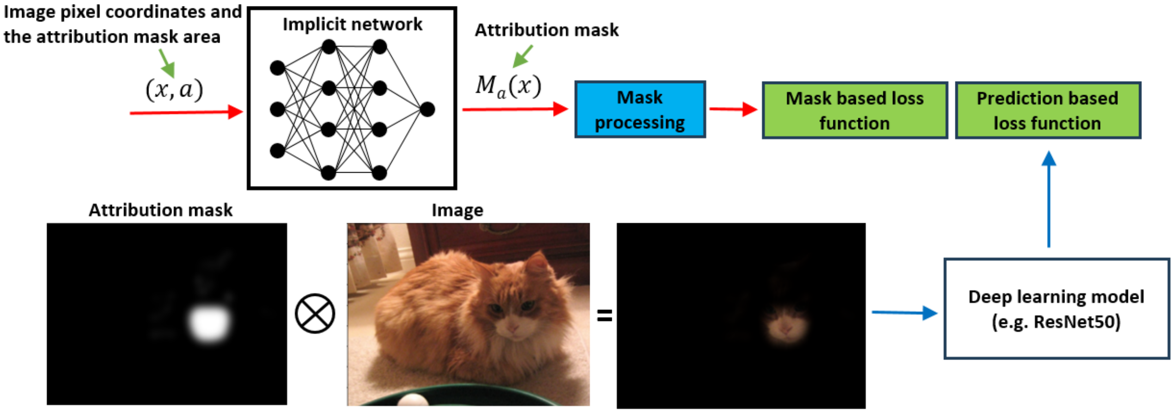
Image pixel coordinates and the attribution mask area
(x, a)
Implicit network
Attribution mask
$M_a(x)$
Mask processing
Mask based loss function
Prediction based loss function
Attribution mask
Image
Deep learning model (e.g. ResNet50)

Figure 2. Scheme illustrating the method proposed in this study. We used a coordinate-based implicit network to compute an attribution mask of size specified by the area parameter. For visualizations, we present blacked-out masks. In implementations, the network processed blurred images, see eq. 1.

a vector $\text{vecsort}_M \in [0,1]^{|M|}$, with $|M|$ indicating the number of the mask pixels. To constrain the mask area to $a \in [0,1]$, an auxiliary vector $r_a \in [0,1]^{|M|}$ is introduced, which consists of $(1-a)|M|$ zeros followed by $a|M|$ ones. Then the regularization function is expressed in the following way:

$$R_a(M) = \frac{1}{|M|}\sum_i \left(\text{vecsort}_M(i) - r_a(i)\right)^2 . \quad (3)$$

In this work, we utilize coordinate-based implicit networks to represent the attribution masks. The network has the following general architecture:

$$f_l(x,c) = \begin{cases} \text{FE}([x,a]), & l = 0 \\ \rho\left(W^{(l)} f_{l-1}(x,c) + b^{(l)}\right), & l \in \{1,...,L-1\} \\ \sigma\left(W^{(l)} f_{l-1}(x,c) + b^{(l)}\right), & l = L \end{cases} \quad (4)$$

where x and a indicate the input 2D coordinates and a vector of area value parameters used to condition the network. ρ and σ stand for the ReLU and sigmoid activation functions, respectively. $W^{(l)}$ and $b^{(l)}$ correspond to the weight and bias of the l-th layer. FE is the Fourier encoding utilized to compensate for the frequency bias resulting from the utilization of the ReLU activation functions [37]. We found that inputting the area condition parameters together with the coordinates worked well in our experiments. This approach directly relates the coordinates with the condition parameters and ensures a certain level of smoothness in the output function.

To obtain a smooth attribution mask M having area parameter of a, we imitate the approach from the original study and process the output of the implicit network with a filter based on the radial basis function, $M_a = \text{Filter}(f_L(x,a))$ [10]. The resulting mask is a function of the area parameter. Next, we search for the smallest mask area according to the following formula:

$$a^* = \min\left\{a : \Phi\big(M_a \otimes I + (1-M_a)\otimes I'\big) \geq \Phi_0\right\}, \quad (5)$$

where Φ_0 stands for a post-softmax probability threshold corresponding to a correct prediction. M_{a^*} indicates the extremal attribution mask, corresponding to input image area that is sufficient for the network to provide an accurate prediction. Our approach is depicted in Fig. 2.

3.2. Generating multiple explanations

Multiple explanations might be present for a single image, which may occur in the case of occluded objects or images presenting multiple objects [23, 27]. A deep learning model may also assign varying levels of importance to different segments of an object, making predictions based solely on the presence of specific object parts. We propose an iterative algorithm based on implicit networks, which can be used to generate multiple explanations. Given an attribution mask obtained using the perturbation analysis, we tackle the problem of whether we can generate a subsequent attribution mask, which does not overlap with the first one and similarly presents an image region important for the model's prediction. Our approach works in an iterative manner; given a baseline attribution mask M^b, we train an implicit network using the following loss function:

$$\mathcal{L}_{\text{mlt}}(M, M^b) = \mathcal{L}_{\text{ext}}(M) + \lambda_d \mathcal{L}_{\text{dice}}(M, M^b), \quad (6)$$

where $\mathcal{L}_{\text{dice}}(M, M^b)$ is a soft Dice-based loss function,

Algorithm 1 Generating multiple attribution masks

```
Input: Deep learning model $\Phi$, input image $I$, perturbed input image $I'$, initial baseline attribution mask $M^0$, number of attribution masks to generate $N$.
for $n \leftarrow 1$ to $N$ do
    $M^b = \sum_{i=0}^{n} M^i$
    $M^b = \text{clamp}(M^b, 0, 1)$
    Train the implicit network from scratch using loss function from eq. 6, utilizing $\Phi$, $I$, $I'$ and $M^b$
    Use the trained implicit network to generate the attribution mask $M_n$
end for
Output: Set of attribution masks $\{M^n\}_{n=0}^{N}$.
```

equal to 0 when M and M^b do not overlap and 1 vice versa [17]. λ_d stands for the weighting parameter. By utilizing a Dice score-based loss function, we ensure that the new mask is actively pushed to avoid overlapping with the baseline mask M^b. Algorithm 1 depicts our iterative approach to attribution mask generation. To determine a subsequent baseline mask, we combine the previously computed masks into a single new mask M^b, and then re-train the implicit network using eq. 6.

3.3. Evaluations

3.3.1 Attribution mask evaluation

Explainable methods typically serve as a tool for visual inspection of mechanisms governing the prediction generation process. Quantitative evaluation remains challenging as the attribution masks may depend on the performance of the deep learning model and its internal biases. Moreover, multiple visually plausible explanations for a model's prediction may co-exist for a single input image. Common approaches to saliency map evaluation include the pointing game metric and various overlap scores designed to compare the saliency map area with the reference object segmentation. In this work, we used the precision score for the evaluations, which can be expressed with the following equation:

$$\text{Precision} = \frac{|M \cap S|}{|M|}, \tag{7}$$

with M and S indicating the attribution mask and the reference segmentation, respectively. By using the precision score, we aim to evaluate if the computed attribution mask is within the reference segmentation mask. Moreover, we introduce a hit rate metric, which we define as a percentage of attribution masks presenting a precision score above 0.5 (at least half of the mask within the reference segmentation), which we believe is more suited for evaluations than the pointing game metric that may produce zero scores both for masks that point out borders of the object and regions slightly outside the reference segmentation. Also, as stated in [10], a single attribution mask is not suited for the pointing game metric, as the mask commonly does not present a single spatial point corresponding to the maximal saliency score.

In addition, in this study, we trained the implicit network five times for each input image to evaluate the variability of the mask generation procedure with respect to the network weights initialization. Given the five attribution masks, we determined the mean performance scores for each image.

3.3.2 Datasets

We used the ImageNet-S$_{50}$ validation dataset, which includes detailed semantic segmentations for 752 ImageNet images corresponding to 50 categories [12]. For the evaluation, we examined the ResNet50 model from the PyTorch model zoo [21]. Moreover, we also employed the 2007 PASCAL VOC test dataset [8]. To assess the attribution techniques, we generated rectangular segmentations based on the reference bounding box annotations. For the evaluation, we utilized the ResNet50 model from the TorchRay library, which was pre-trained to classify 20 PASCAL VOC categories [10]. Following the TorchRay library, we used 2230 images corresponding to difficult cases. Unless explicitly stated, all visualizations presented in this study were generated for the PASCAL dataset.

3.4. Implementation details

The perturbed images were generated with a Gaussian filter. We used the same MLP architecture for all experiments in this study. Each implicit network included five fully connected hidden layers, each with 256 neurons. In addition, we used the Fourier input mapping with six frequencies and 128 components to encode both the spatial coordinates and the area constraint parameter [37]. Following the extremal perturbation study, we investigated the area constraint parameter in range of [0.025, 0.2], which was scaled before training to [0, 1] to match the range of the coordinates [10]. Adam optimizer with a learning rate of 0.0001 was used to train the networks on a single NVidia 4090 GPU [15]. Each network was trained for 4000 epochs, with each epoch corresponding to a batch of all pixel coordinates and the scaled area parameter uniformly sampled from [0, 1]. λ_r and λ_d were set to 1. After the training, to determine the attribution mask for the extremal perturbation technique and the proposed method, we examined area parameters equal to {0.025, 0.05, 0.1, 0.2} [10]. These two attribution methods were additionally compared with the GradCAM and RISE techniques [22, 26], which correspond to popular activation-based and perturbation-based approaches. To generate binary attribution masks

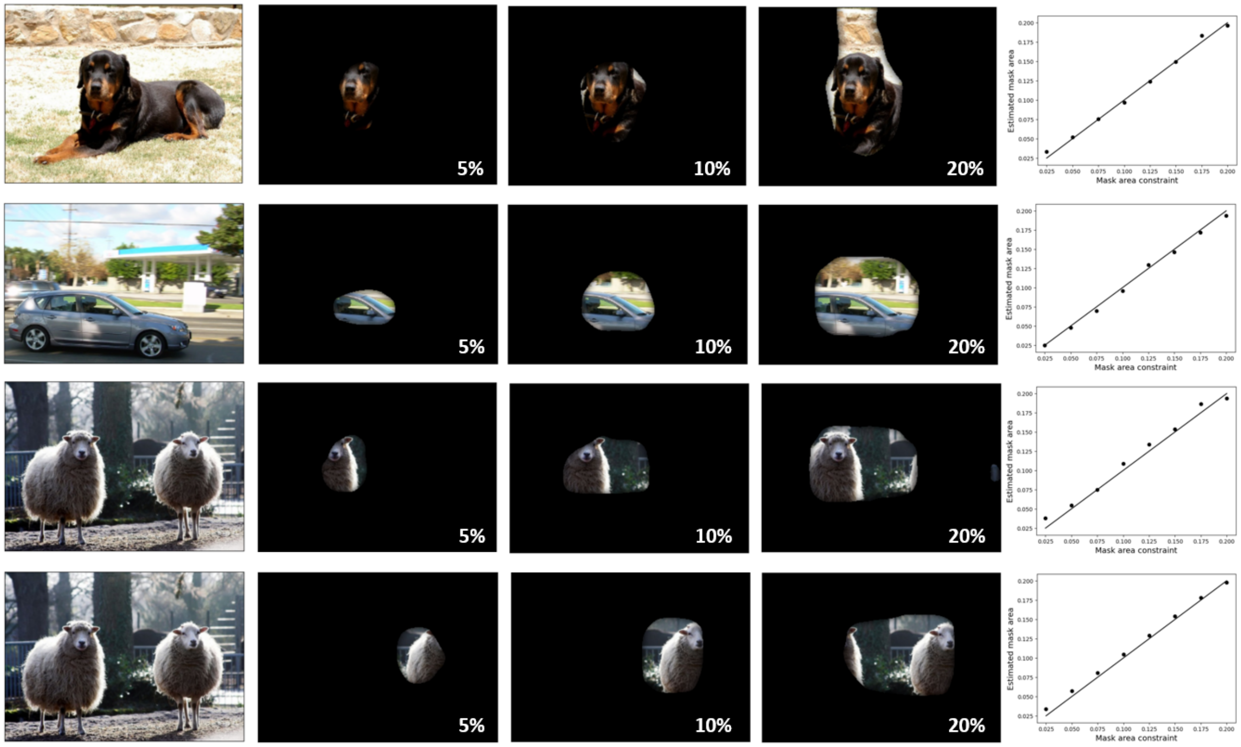

Figure 3. Illustration of the attribution masks generated with the proposed method. We found that an implicit network could converge to solutions presenting different visual explanations, depending on the network weight initialization. Percentage indicates the area of the attribution mask.

for the saliency maps computed with the GradCAM and RISE techniques, we applied thresholding as in [46]. The saliency maps scaled to range of [0,1] were thresholded with manually selected cut-off values of 0.2 and 0.5 for the ImageNet and PASCAL VOC datasets, respectively. All computations were performed using PyTorch 2.1.2 in Python [21]. Implementation of the proposed method is available at github.com/mbyr/INR-EXP.

4. Results

4.1. Perturbations with implicit networks

Fig. 1 visually compares the proposed INR-based method with the extremal perturbations technique. Due to the conditioning mechanism, our approach determines smoother and more continuous attribution masks with respect to the area constraint parameter. In contrast, the extremal perturbations technique computes each attribution mask from scratch for each area parameter, resulting in less spatially continuous masks as different regions can be selected from run to run. Moreover, we found that our method achieved better monotonic correspondence between the area constraint and the actual calculated mask area. In addition, Fig. 3 presents several more examples illustrating the performance of the proposed method. Here, the images from the last two rows demonstrate that the determined explanations may not be unique, as different network weight initialization may result in plausible but non-overlapping attribution masks.

Qualitative comparison between the proposed method and several other popular attribution techniques is presented in Fig. 4. Here, we can notice the diversity between the obtained explanations, associating importance to various parts of the input image. For example, for the results in the second row, obtained for the ’cow’ category, saliency maps highlight different parts of the cow’s face. Quantitative performance scores are depicted in Table 1. For the ImageNet dataset, featuring detailed semantic segmentations, our method outperformed the other techniques, achieving a mean precision score of 0.68. For the VOC dataset, which includes rough bounding box segmentations, we obtained a mean precision score of 0.44 for our method, which was better compared to the RISE and GradCAM techniques but slightly worse than for the extremal perturbation algorithm, 0.46. However, additional statistical analysis based on the t-test ($p < 0.05$) presented that there were no significant difference between our method and the extremal perturbation technique on the VOC dataset. In addition, we evalu-

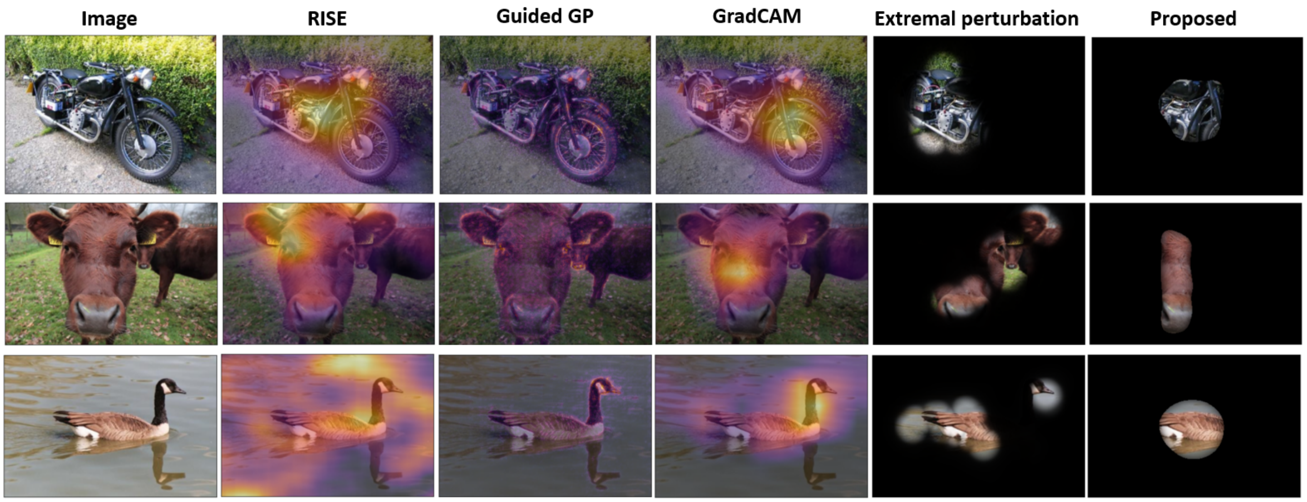

Figure 4. Qualitative comparison of several attribution methods.

Table 1. Precision scores (hit rates) were determined for the investigated methods using the ImageNet and PASCAL VOC datasets. For the ImageNet dataset, which features detailed semantic segmentations, our method outperformed the other techniques. For the VOC dataset, which includes rough bounding box segmentations, our method achieved slightly worse results than the extremal perturbation technique, but the difference was not statistically significant.

Method	ImageNet	VOC
RISE	0.36 (0.28)	0.39 (0.37)
GradCAM	0.58 (0.60)	0.38 (0.37)
Extremal perturbation	0.63 (0.71)	**0.46** (0.47)
Proposed, mean	**0.68** (0.73)	**0.44** (0.44)

ated the training time of the implicit network on our GPU, which was equal to approximately three minutes.

4.2. Multiple explanations

Fig. 5 illustrates the first three explanations obtained with the proposed iterative approach. The method determined non-overlapping visual explanations related to different parts of the input image. For example, in the second case, to predict the ’plane’ category, the network associates importance with different parts of the plane. Fig. 6 shows an image with five explanations, demonstrating an important finding: the network not only associates the ’boat’ category with different parts of the boat but also with surroundings commonly accompanying the boat, such as water and blue sky. Notice that for the fifth explanation, the proposed method highlighted both the sky and the water, suggesting that these two regions might be visually compared within the network to provide correct prediction.

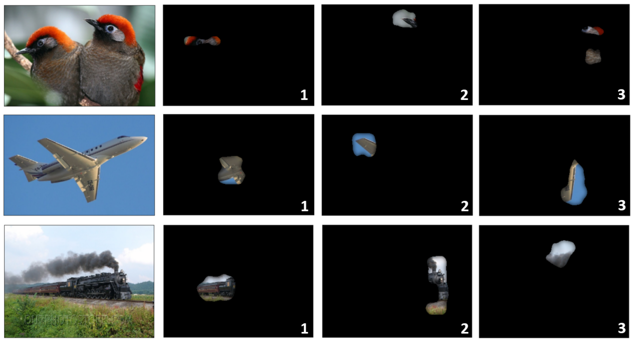

Figure 5. We used implicit networks to generate multiple non-overlapping attribution masks, separately highlighting different parts of the input image that are important for the prediction.

We evaluated the consecutive explanations obtained for the ImageNet dataset, with results presented in Table 2. The precision score gradually decreased from 0.68 in the first iteration to 0.30 by the third iteration. This result clearly demonstrates that subsequent explanations tend to be more often associated with the borders of the reference segmentation mask or even with regions outside the mask. However, this was not always the case. Some attribution masks generated during the second and third iterations showed larger overlaps with the reference mask, evidenced by a maximum precision score of 0.73 (mean score over the maximal precision value of all three subsequent explanations).

5. Discussion

In this work, we proposed a novel model-agnostic approach to attribution mask generation. Given an input im-

Table 2. Performance on the ImageNet dataset. Precision scores (hit rates) were computed for the subsequent attribution masks generated with the proposed iterative procedure (Algorithm 1). Max indicates the performance when selecting the mask with the highest overlap over the three iterations.

Method	Precision (hit rate)
Iteration #1	0.68 (0.72)
Iteration #2	0.42 (0.42)
Iteration #3	0.30 (0.24)
Combination, max	0.73 (0.77)

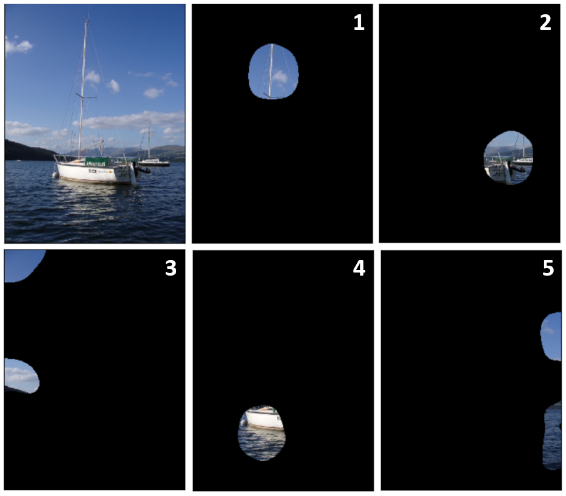

Figure 6. The proposed technique for iterative explanation generation can be used to provide important insights about the performance of the deep learning models. This example shows that the network associated the ’boat’ category not only with the parts of the boat, but also with the sky, clouds and water.

age and a deep learning model, we trained an implicit coordinate-wise network to output the attribution mask highlighting image regions important for the prediction. By using INRs, we extended the extremal perturbations technique. Conditioning the implicit network resulted in attribution masks that are well-behaved with respect to the imposed area constraints. Next, we developed a novel approach for multiple explanation generation. We modified the loss function to train implicit networks to generate non-overlapping attribution masks, pointing out different important parts of the input image. This way, we found that a prediction model may associate the image label with both the appearance of the object of interest as well as with areas and textures usually accompanying the object (e.g., boat vs water). Such findings, demonstrating different operational mechanisms behind the models, are crucial for ensuring robustness in applications.

Utilization of implicit networks offers several advantages for explainable deep learning. Firstly, our study shows that, compared to standard attribution methods, various conditioning mechanisms can be considered in the optimization of the coordinate-wise implicit network, enabling control over the attribution mask generation procedure. Secondly, the training of the implicit network can be performed using various custom loss functions, opening new and interesting directions for the development of explainable methods. Aside from the approach presented in this work, the attribution mask generation procedure could be jointly performed with the input image regression or coordinate-wise image perturbation.

There are several limitations to this work. Firstly, our method explains a ’black-box’ deep learning model by training another neural network, which can itself be considered as a ’black-box’ model. Therefore, the proposed method might be considered less trustworthy than other attribution techniques. Secondly, training the implicit network requires far more time compared to techniques such as GradCAM, which are based on a single forward/backward pass. However, this limitation could be mitigated by using meta-learning-based weight initialization [30, 36]. Thirdly, while framing the attribution mask generation problem as an optimization task has several advantages, it also presents several challenges, such as the network divergence issue or the requirement to balance loss function components. Although we examined our approach using two datasets, it might be more difficult to converge the network for high-resolution images due to, for instance, GPU memory constraints.

6. Conclusion

We believe that our study presents several novel and interesting insights about the explainability of deep learning models. Our work demonstrates that implicit networks are well-suited for the generation of attribution masks. We devised an algorithm that can be used to provide multiple visual explanations, improving the understanding of a network’s performance. In the future, we plan to examine different conditioning mechanisms. For example, it would be interesting to utilize implicit networks to associate the attribution mask generation process with other tasks, such as image decomposition or coordinate-wise image perturbation.

Acknowledgments

This work was partially supported by the programs for Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) and Multidisciplinary Frontier Brain and Neuroscience Discoveries (Brain/MINDS 2.0) from the Japan Agency for Medical Research and Development AMED (JP15dm0207001,JP23wm0625001).

References

[1] Dieuwertje Alblas, Christoph Brune, Kak Khee Yeung, and Jelmer M Wolterink. Going off-grid: continuous implicit neural representations for 3d vascular modeling. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 79–90. Springer, 2022. 3

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805, 2023. 2

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 3

[4] Nuri Benbarka, Timon Höfer, Andreas Zell, et al. Seeing implicit neural representations as fourier series. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2041–2050, 2022. 3

[5] Michal Byra, Charissa Poon, Tomomi Shimogori, and Henrik Skibbe. Implicit neural representations for joint decomposition and registration of gene expression images in the marmoset brain. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 645–654. Springer, 2023. 3

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 2

[7] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017. 2

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 5

[9] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *European Conference on Computer Vision*, pages 636–654. Springer, 2022. 3

[10] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019. 1, 2, 3, 4, 5

[11] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 2

[12] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 5

[13] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 3

[14] Mohammad AAK Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16327–16336, 2021. 2

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[18] Amirali Molaei, Amirhossein Aminimehr, Armin Tavakoli, Amirhossein Kazerouni, Bobby Azad, Reza Azad, and Dorit Merhof. Implicit neural representation in medical imaging: A comparative survey. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2381–2391, 2023. 3

[19] Seonghyeon Nam, Marcus A Brubaker, and Michael S Brown. Neural image representations for multi-image fusion and layer separation. In *European conference on computer vision*, pages 216–232. Springer, 2022. 2

[20] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[21] 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 *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5, 6

[22] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2, 5

[23] Luc-Etienne Pommé, Romain Bourqui, and Romain Giot. H^2o: Heatmap by hierarchical occlusion. In *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, pages 111–118, 2023. 1, 4

[24] Harish Guruprasad Ramaswamy et al. 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*, pages 983–991, 2020. 2

[25] A Saranya and R Subhashini. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, page 100230, 2023. 2

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 5

[27] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One explanation is not enough: structured attention graphs for image classification. *Advances in Neural Information Processing Systems*, 34:11352–11363, 2021. 1, 4

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[29] Rajhans Singh, Ankita Shukla, and Pavan Turaga. Polynomial implicit neural representations for large diverse datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2051, 2023. 3

[30] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147, 2020. 8

[31] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2, 3

[32] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[33] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 3

[34] Nil Stolt-Ansó, Julian McGinnis, Jiazhen Pan, Kerstin Hammernik, and Daniel Rueckert. Nisf: Neural implicit segmentation functions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 734–744. Springer, 2023. 3

[35] Karolina Szczepankiewicz, Adam Popowicz, Kamil Charkiewicz, Katarzyna Nałecz-Charkiewicz, Michał Szczepankiewicz, Sławomir Lasota, Paweł Zawistowski, and Krystian Radlak. Ground truth based comparison of saliency maps algorithms. *Scientific Reports*, 13(1):16887, 2023. 2

[36] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 8

[37] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 4, 5

[38] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020. 1

[39] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 2

[40] Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. Implicit neural representations for deformable image registration. In *International Conference on Medical Imaging with Deep Learning*, pages 1349–1359. PMLR, 2022. 3

[41] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library, 2022. 3

[42] Dejia Xu, Peihao Wang, Yifan Jiang, Zhiwen Fan, and Zhangyang Wang. Signal processing for implicit neural representations. *Advances in Neural Information Processing Systems*, 35:13404–13418, 2022. 3

[43] Guandao Yang, Serge Belongie, Bharath Hariharan, and Vladlen Koltun. Geometry processing with neural fields. *Advances in Neural Information Processing Systems*, 34:22483–22497, 2021. 3

[44] Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19228–19238, 2022. 3

[45] Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability. *arXiv preprint arXiv:2301.07002*, 2023. 2

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2, 6