



Context-based image explanations for deep neural networks

Sule Anjomshoae^{a,*}, Daniel Omeiza^b, Lili Jiang^a

^a Department of Computing Science, Umeå University, Sweden

^b Department of Computer Science, University of Oxford, UK

ARTICLE INFO

Article history:

Received 9 June 2021

Received in revised form 13 August 2021

Accepted 10 September 2021

Available online 22 September 2021

Keywords:

DNNs

Explainable AI

Contextual importance

Visual explanations

ABSTRACT

With the increased use of machine learning in decision-making scenarios, there has been a growing interest in explaining and understanding the outcomes of machine learning models. Despite this growing interest, existing works on interpretability and explanations have been mostly intended for expert users. Explanations for general users have been neglected in many usable and practical applications (e.g., image tagging, caption generation). It is important for non-technical users to understand features and how they affect an instance-specific prediction to satisfy the need for justification. In this paper, we propose a model-agnostic method for generating context-based explanations aiming for general users. We implement partial masking on segmented components to identify the contextual importance of each segment in scene classification tasks. We then generate explanations based on feature importance. We present visual and text-based explanations: (i) saliency map presents the pertinent components with a descriptive textual justification, (ii) visual map with a color bar graph showing the relative importance of each feature for a prediction. Evaluating the explanations using a user study ($N = 50$), we observed that our proposed explanation method visually outperformed existing gradient and occlusion based methods. Hence, our proposed explanation method could be deployed to explain models' decisions to non-expert users in real-world applications.

© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Deep neural networks (DNN) have improved the accuracy of prediction tasks in many areas from computer vision to natural language processing. However, the inability of DNNs to show their reasoning is limiting the wide adoption of these models in real-world applications. Given that, there has been an increasing interest in explaining and understanding classifiers retrospectively and examining what they have learned during training. Most of the existing approaches [20,32,31,24] explain how a model determines its final output by where the network is 'looking'. These often produce intuitive visualizations that are intended for expert users to interpret network representations and evaluate the correctness of a model. For the non-experts who do not have sufficient knowledge of computer vision systems, a justification model that can provide intelligible explanations to non-experts would be more useful [6,14,16]. Generally, the non-technical users would be interested in knowing instance-specific explanations instead of how the model makes its decision. Thus, the utility of explanations varies

with stakeholders and the necessary stakeholders' requirements should be considered [25].

We propose a context-based justification method detailing how visual evidence is compatible with a classifier's output. Context-based justification is an approach that examines the relationship between an input and output by modifying an input image and observing its effect on the output. Image is segmented into several interpretable regions to measure this effect. In this work, we demonstrate our approach using both manual annotation and automated segmentation with their semantic categories. Semantic categories provides the properties of the image scene needed to assess the effect of each component in scene classification tasks. Scene classification or scene recognition in this context involves assigning a label such as a playground, kitchen, beach to an input image based on the image's overall content. The overall idea is to map all the components to a semantic space based on their contextual importance.

Context provides critical information about particular scene; such as objects in an image, their arrangement, relative physical size to other objects, and location. Contextual information provide important cues for a model to learn during training and also use during prediction. While some features have more influence on the outcome than others, the influence of each individual component is also dependent on other components. We extract the degree of importance by estimating how each feature is contributing to a prediction. This is performed through

* Corresponding author.

E-mail addresses: sulea@cs.umu.se (S. Anjomshoae), daniel.omeiza@cs.ox.ac.uk (D. Omeiza), lili.jiang@cs.umu.se (L. Jiang).

a systematic masking and observation of the prediction score. We assess the effect of context in scene prediction and use it as justification. We note that sometimes the terms *explanation* and *justification* are used interchangeably in the domain of explainable AI.

The contributions of this paper on visual explanations research are summarized as follows:

- We propose an algorithm for generating context-based explanations that can be applied to DNNs and other black-box models that output probabilities.
- We present visual and text-based explanations that aim to improve intelligibility for non-expert users.
- Using a user study, we present evaluation results of the different aspects of visual explanation.

The rest of the paper is organized as follows: Section 2 discusses the relevant work. Section 3 introduces the contextual importance (*CI*) algorithm as an explanation method, and describes the adaptation of *CI* for visual explanations. Section 4 describes the experiments with *CI* for image explanations. Section 5 provides details on the evaluation method. Section 6 presents the results of the user study. Results are discussed in Section 7, and finally, Section 8 concludes the paper.

2. Related work

Researchers have been focusing on integrating explanation facilities into computer vision tasks (e.g., image captioning, visual question answering, object localization) [32,30,29]. Generally, these explanation models can be categorized as intrinsically interpretable models and the post-hoc interpretability [8]. Both categories produce saliency maps which simplify the image representations, making it easy to analyze. Saliency maps help to highlight the most important features for a prediction.

Interpretable models are usually specific to a certain model which focus on the internal functioning of the model. Interpretable models analyse the interaction between neurons and what each neuron has learned. Post-hoc interpretability methods explain instance-specific predictions on the basis of how each feature (or a set of features) influences the final outcome. In general, both approaches can have some limitations and strengths. The results of interpretable models are directly explainable without requiring another model to generate explanations. However, they are limited to a specific learning model. Whereas the post-hoc interpretability methods are generally model-agnostic but they may be limited in their approximate nature [12].

Some of the intrinsically interpretable models applies gradients and decomposition based methods to understand the internal structure of complex black-box models. Gradient methods highlight the unit changes and emphasize the important features or regions in an image. In this way, it is possible to learn the prototypes that have the highest probability to be predicted as a certain class of a trained DNN. Considering this, Li et al. [18] proposed an interpretable neural network architecture whose predictions are based on the similarity of an input to a small set of prototypes learned during training. As presented by Nguyen et al. [23], the activation maximization method synthesizes an image that highly activates a neuron and reveals the features learned by each neuron in an interpretable way. Some works proposed visual analysis by clustering important neurons based on the features and the interactions between them [19,15]. Furthermore, decomposition methods such as layer-wise relevance methods are presented to analyze which pixels are contributing to what extent to a classification result [5]. Some post-hoc explanations also proposed using gradients to create saliency maps. One way to visualize saliency maps is by going backward through the inverted network from an output of interest. It highlights the discriminative features of the image with respect to the given class [30]. Another method uses class activation mapping with the gradients of a target input in the last convolutional layer to produce a rough

localization map highlighting the important regions [28]. This method is further developed for explaining occurrences of multiple object instances in a single image [9].

Moreover, other post-hoc explanation methods suggested modifying the input image and observing the effect on the prediction. Zeiler and Fergus [32] proposed the occlusion sensitivity method to monitor the prediction score for a specific class by masking different parts of the input image with a grey square. Results are then visualized as saliency maps to show which parts of an image are most important for the classification. Another method that uses prediction score is LIME (Local Interpretable Model-Agnostic Explanation). This method generates explanations by approximating a black-box model by perturbing an input image. They adopt SLIC (Simple Linear Iterative Clustering) superpixel method to partition image into smaller regions and use these parts to perturb the image [1]. Then, they present the superpixels with the highest positive weights as an explanation [27]. Although these methods seem intuitive, they are inefficient in terms of salient features representation. In our approach, we propose observing the prediction result using semantic occlusion and measuring the effect of each semantic category on the output. The effects are then visualized to explain the main contributors. The pixel-level segmentation allows us to produce not only saliency maps but also visual and text-based explanations which is one possible means to address the non-expert user requirements.

3. Contextual importance (*CI*) and visual explanations

This section describes the motivation for contextual importance as an explanation method, and the adaptation of this method for generating visual explanations.

3.1. Motivation

In this study, we propose an approach for generating visual explanations for DNNs through the contextual importance method. This concept originates from the idea that the set of input features forms the context, and given a context, the importance of feature is dependent on other feature values [13]. Therefore, contextual importance indicates the degree of significance of a feature value (or set of feature values) when changes are made to that particular value(s) while the rest of the input values remain constant.

Feature importance usually signifies a measure for how much one feature affects the outcome when taking into account the whole dataset. Building an explanation method is simple in the case of a purely linear model, where every feature's importance is constant and irrelevant to the other feature values. Neural networks are mainly useful for tasks where linear models are not sufficiently expressive. Considering post-hoc explanations of non-linear methods, the feature importance might be specific for the current set of input values.

When dealing with non-linear models, it becomes non-trivial how to define feature importance. Most current model-agnostic methods tend to use local gradients as an indicator of the contextual importance. Rather than observing how much an output value changes with fixed amount of small perturbations to the current input value of the studied feature(s), we study how much perturbations over the whole range of possible feature values affects the output range. contextual importance (*CI*) was initially proposed for a tabular data type (to see implementation for tabular data [3,2]). Though we have experimented *CI* with MNIST dataset [4], its potential for more complex image explanations has yet to be investigated.

3.2. Generating visual explanations

For an image, the collection of pixels corresponds to a feature. Thus, the image is deemed a single variable with various "interpretable regions/ features". One way of parsing image into interpretable regions

is using segmentation methods. To measure the contextual importance, we turn off each region (i.e., setting the pixel to 0) and monitor the effect on the prediction score as illustrated in Fig. 1. Formally, the contextual importance is defined as follows:

$$CI_j(C, x) = \frac{Cmax_j(C, x) - out_j(C, x)}{Cmax_j(C, x) - Cmin_j(C, x)} \quad (1)$$

$CI_j(C, x)$ signifies the contextual importance (CI) of a feature for a particular class index j . For (C, x) , C denotes the context after each interpretable region is masked from x by setting pixel values of that particular cluster to zero. Given (C, x) , we get the prediction score $out_j(C, x)$ by model f for each region (Algorithm 1, line 1). Then, we find how the prediction score drops over the perturbed input and identify the $Cmax_j(C, x)$ and $Cmin_j(C, x)$ among these predicted scores. Consequently, $Cmax_j(C, x)$ and $Cmin_j(C, x)$ are the maximum and minimum prediction scores (for a specific class index) after all clusters have been individually masked, while $out_j(C, x)$ is the prediction score of a particular class index with a particular cluster of pixels masked. So, contextual importance is calculated for each cluster based on $out_j(C, x)$ (where a particular region masked).

Algorithm 1 Calculating Contextual Importance

Given: Input image x , model f , class index j

Require: Cluster set (C, x) for perturbing the image

- 1: **Run** f with (C, x) to get $out_j(C, x)$
- 2: **Find** $Cmin_j(C, x)$ and $Cmax_j(C, x)$
- 3: **Calculate** $CI_j(C, x)$ using (Eq. 1)
- 4: **Return** $CI_j(C, x)$

Output: Visual and text explanations

For the segmentation part, as an alternative to SLIC superpixel method which was previously adopted by LIME explanations, we demonstrate contextual importance using AMR (Adaptive Morphological Reconstruction) segmentation [17]. AMR helps to reduce over-segmentation and increase the chances that boundaries of importance are extracted. Fig. 2 shows side by side comparison of the explanations generated based on these two segmentation methods. The visual explanations show the regions of those with a high significance for a model to



Fig. 2. Contextual importance using different segmentation methods. AMR is better at extracting the region of interest in a complex background than the SLIC superpixel method. This resulted in generating more detailed explanations with AMR segmentations.

make a certain prediction. AMR seems to be more aligned with object boundaries which results in more clear visual explanations while SLIC segmentation often fails to extract boundaries that affect the area shown as an explanation.

Although the visual explanations provided based on AMR seem more reasonable, it is still challenging to identify coherent regions due to the broad diversity and ambiguity of visual patterns in images. However, a good visual explanation should ideally capture fine-grained details of an object while localizing it in the image which is found rather difficult for current automatic vision systems. Generally, semantic segmentation methods aim to address this issue.

Next, we proceed with experimenting with semantic segmentation. Semantic segmentation allows generating more detailed explanations using semantic categories. Once we know the contextual importance of each semantic category in an image, this can be presented in several possible ways depending on the end-user requirements. In this work, we present visual and textual justifications as follows:

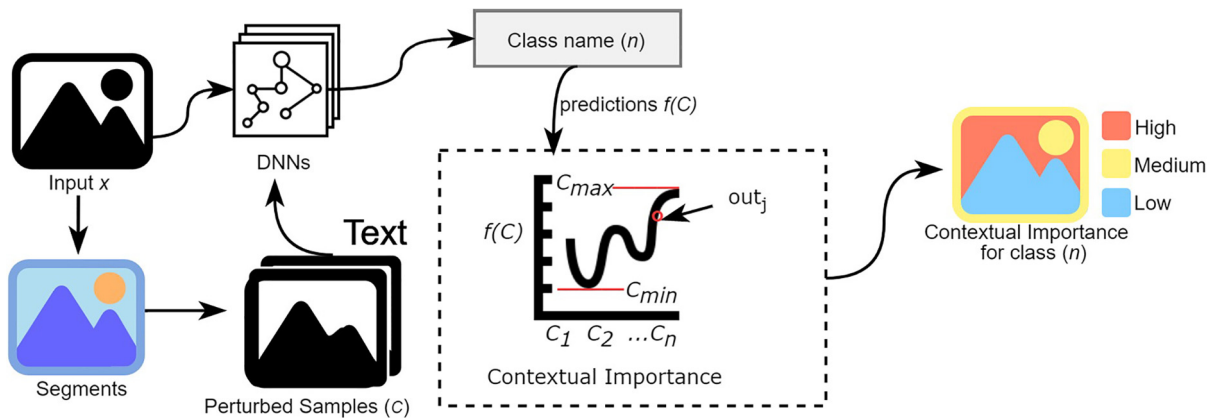


Fig. 1. The process of extracting contextual importance of each feature in an image. An input image is clustered into interpretable regions, then perturbed samples are created by turning off each region at a time. The model is run for each perturbed sample for the predicted class (n). The contextual importance of each region is calculated based on the obtained prediction scores and represented visually and in text forms.

- **Saliency map:** This explanation presents the most relevant parts of an image for a prediction and all other parts with lower CI values (≤ 0.5) greyed out.
- **Textual justifications:** This explanation type describes the effect of a set of features on the prediction class based on the degree of importance in three levels (i.e., important (>0.75), not as important but contributing ($0.5 \leq CI \leq 0.75$), and not relevant (<0.5)).
- **Visual map with graph:** This explanation type provides a more detailed visual representation to look into how much each feature is affecting the prediction. The color map is created based on the degree of importance levels same as the textual explanations. The color bar graph indicates the extent a category is contributing to prediction with their respective CI values.

4. Experiments

This section presents the visual and text explanations based on image annotations and automated semantic segmentation. We first experiment with manually annotated images to generate more detailed explanations. For the best exhibition of the method, we focus on the effect of context in a scene classification problem with deep neural networks. Following that, we present CI 's potential for contrastive explanations to generate “why?—why not?” explanations. In some contexts, contrastive explanations could be more useful for model debugging. Lastly, we experiment with an automated semantic segmentation method to assess whether the explanations generated based on

the semantic segmentation model are similar to explanations generated from annotations.

4.1. Explanations using image annotations

For this experiment, we selected the PASCAL VOC 2010 dataset to produce fine-grained visual justifications for scene classification. The dataset provides annotations for the whole scene where every pixel is labeled with a semantic category. It contains 460 labels for semantic segmentation and object detection [22]. Given the labels within the image scene, the aim is semantically mapping the essential components according to the contextual importance for a prediction.

The experiment is performed on the GoogLeNet convolutional neural network with Places365 extension for scene recognition tasks [33]. The network classifies images into 365 unique scene categories such as beach, canyon, dining hall, and ski slope.

Considering that different users might prefer different explanation types, we present explanations in various ways detailing how the visual evidence is compatible with the predicted class. Fig. 3(a) presents a saliency map to display the most relevant parts of an image needed by the model for prediction. The grey portion of the saliency map represents the less important regions required by a model to classify an input image correctly. Fig. 3(b) presents the textual explanations with the effect of a set of features on the prediction class. We create text-based justifications using textual templates where the content is adjusted based on visual evidence in an image.

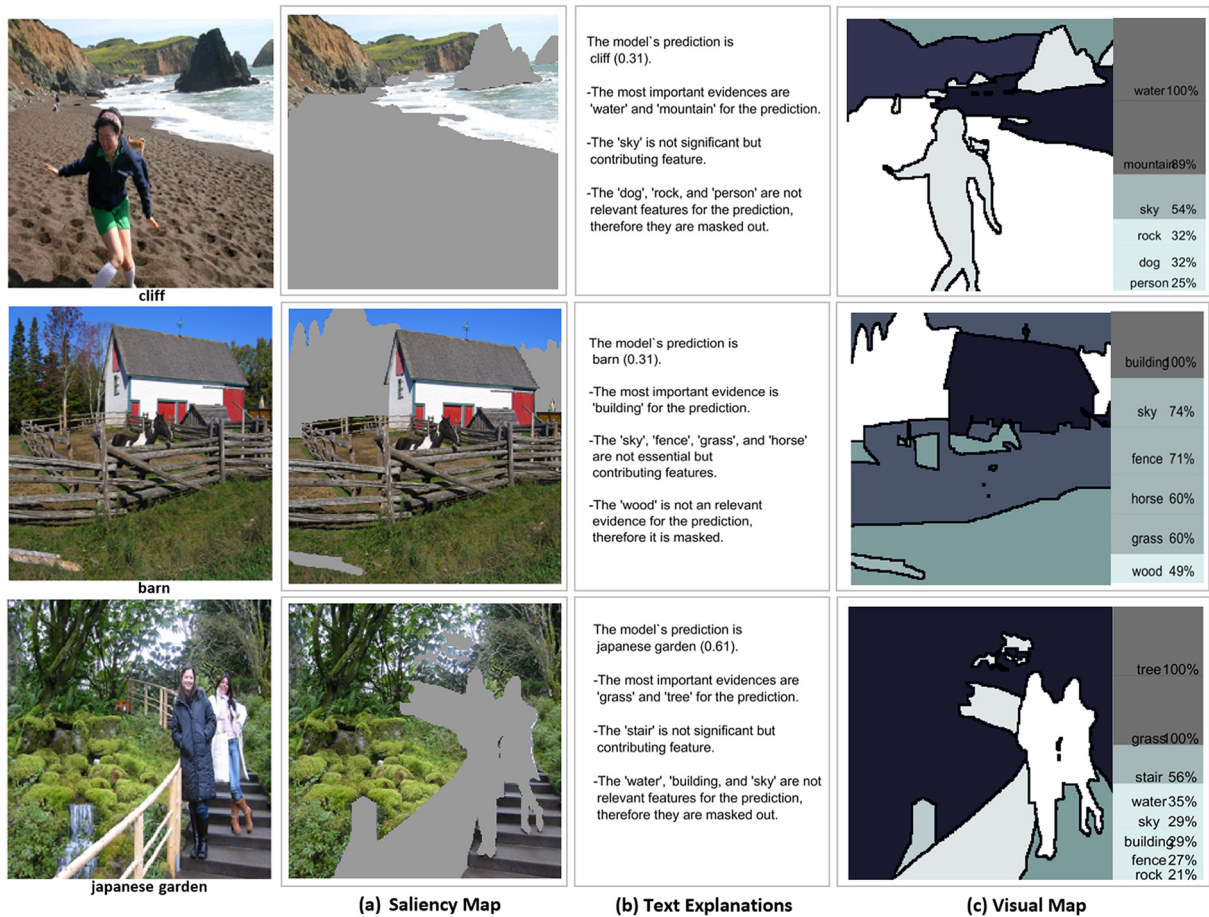


Fig. 3. Context-based justifications (a) Saliency map is showing only the components with the highest influence on the prediction. The features with low contextual importance are masked out. (b) Text-based justifications are listing the features and their effect on the output. (c) Visual map with stacked bar graph indicating the contextual importance of each feature and their values.

Fig. 3(c) demonstrates the visual map showing the relative importance of each component for a particular prediction. High importance features are shown in darker, medium importance is shown in gray, and low importance is presented in white. Moreover, the visual map is paired with a stacked bar graph to provide details of the features affecting the outcome. Each feature is represented in its corresponding color based on the visual map, also containing their respective contextual importance values. We removed the features with the lower importance close to zero from the graph to reduce redundancy in presentation. Note that the importance of all the features does not add to 100 because the contextual importance shows the degree of influence of each feature relative to others, as well as it is not the distribution of the probability score.

4.1.1. Visual explanations for counterfactual cases

Humans generally state and request contrastive facts to distinguish between similar examples. Thus, explanations are usually in the form of “Why this?—Why not that?”, asking a contrasting case that did not happen, even if it is not implied in the question [21]. Recently this type of explanation gained attention in explaining machine learning predictions. It is expected that counterfactual explanations would make the model change its explanation and produce class discriminative visualization for differing classes. This would make the model more trustworthy and help to assess whether the model is working as expected.

The contextual importance can be computed for any model that outputs a prediction score for all classes. This makes it possible to generate explanations for other likely classes as well. Contextual importance over the perturbation variable is computed to explain the model's prediction results for a counterfactual case (e.g., what if B instead of C). We are not only interested in visualizing the features that are contributing but also looking for ‘missing features’ that discriminate a class from other likely classes. As shown in Fig. 6, given an input image with a prediction class, we find predictions for the counterfactual case. The importance of the features that are contributing to the prediction class are computed based on the same equation (Eq. (1)). The visualizations demonstrate the most critical features contributing to different classes and help identify class discriminative features.

4.2. Explanations using automated semantic segmentation

In this section, we show the implementation of contextual importance using an automated semantic segmentation. The Deeplab v3+ network trained on a road-segmentation dataset for autonomous driving is used for the segmentation task [10,7]. The dataset contains street-view images captured while driving and provides 32 semantic category labels including road, tree, and building. Fig. 4 compares the segmentation results with the ground truth. Visual comparisons show that the semantic segmentation results match well for classes with large areas such as buildings, roads and the sky, while some finer objects on the scene such as pole and sign-symbol are not as precise.

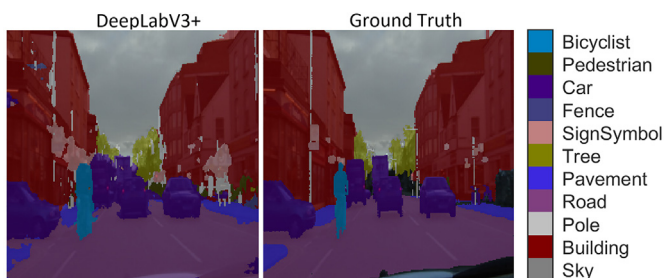


Fig. 4. Ground truth comparison of segmentation results.

Moreover, we measured the amount of overlap per class using three segmentation metrics; the intersection-over-union (IoU) [26], Dice similarity coefficient (DC) [34], and contour matching (CM) [11]. Although each function gives a slightly different measurement, overall they confirm the visual results as shown in Table 1. Sky, road, and building categories have high scores, while categories such as pavement, tree, and car have lower scores.

Next, we generated CI explanations based on these automated semantic segmentation results. GoogLeNet Places360 model is used to make predictions, then we identified how each feature is contributing to a prediction. Fig. 5 displays the visual and text explanations generated based on the automated semantic segmentation and compares with explanations generated from ground truth annotations. Furthermore, we conducted a similarity assessment through a user study to evaluate these explanations. The results of this study is presented in Section 6.

5. Evaluation method

In order to evaluate the proposed explanation method, we conducted a user study to assess the quality of different explanation methods and obtain users' views on this. Our user study evaluated the main idea of our approach, that is, whether explanations based on coherent region identification are more acceptable than other methods for the general users. We evaluated whether these explanations can influence an end user's confidence in a model. Moreover, we investigated human perception on model predictions, and their preference on different explanation presentation forms. Finally, we assess the effectiveness of proposed explanations on automatically segmented image through a similarity study.

5.1. User study

The survey study explored the following structured questions: (i) How fine-grained details would affect humans' understanding of an image? (ii) How do explanations and prediction score affect the users' confidence in a model? (iii) How machine classification compares with human classification? (iv) What is the user preference among different explanation types, and (v) How well does the user think the automated segmentation explanations match ground truth explanations?

There were five sections in the survey questionnaire as follow:

- **Comparison of algorithms:** Given a prediction class, participants were asked to select an option (among four options) that best describes the prediction. Each option represented a visual explanation obtained from different explanation algorithm. We provided a visual comparison of four saliency maps (CI-Pascal VOC, Grad-Cam, Occlusion Sensitivity, and CI-SLIC) to observe whether the important features are pointing to similar regions as in our results. Grad-cam uses the gradient of the last classification score for the last convolutional feature map where the parts with the highest gradient has the most affect on a prediction score. The occlusion sensitivity measures

Table 1
Similarity measurements.

	IoU	DC	CM
Sky	.958	.978	.837
Building	.861	.925	.676
Pole	.127	.225	.446
Road	.812	.896	.567
Pavement	.665	.799	.773
Tree	.594	.746	.736
Sign-Symbol	.294	.455	.501
Car	.738	.849	.621
Pedestrian	.039	.075	.150
Bicyclist	.641	.781	.606

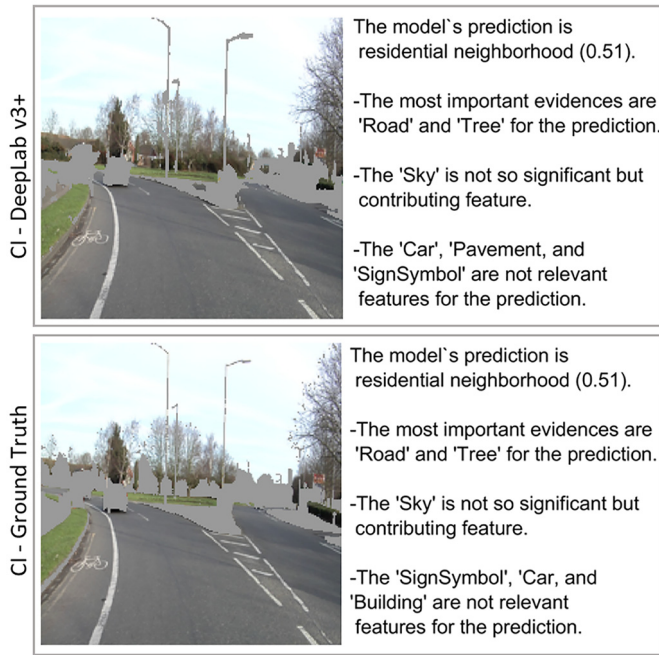


Fig. 5. Visual and text explanation comparison between DeepLab V3+ semantic segmentation and ground truth annotations.

DNN's sensitivity by masking small areas of the input image. We also compare with CI-SLIC superpixel segmentation to observe the difference with CI-Pascal VOC annotations and the affect on human decision. Fig. 7 shows examples of these comparisons.

- **Confidence:** This section assess users' confidence and observes whether their decision changes based on explanations with and without prediction scores provided. Given two visual explanations, participants were asked to specify the model in which they have more confidence in. Subsequently, they were presented with the visual explanations but with prediction scores and were asked to indicate again the model they place more confidence in. For this experiment, we compared GoogLeNet with Resnet on PASCAL VOC 2020 dataset. We took the instances in which both models made the same prediction, then ran our explanation method on both models. An example of this is shown in Fig. 13.
- **Perception:** This section focuses on questioning whether explanations can extract enough class discriminative features which justify the model's prediction and assess if the human perception of scenes is in line with a system's (or model's) judgment (or classification). To evaluate this, we selected two prediction class one being the correct class for the shown image explanation and the other one is the

second or another nearest category that does not produce the same explanation result. An example of this is shown in Fig. 6. The explanations for the campsite being the predicted class and lawn is given as the second option to the participants. From Fig. 6, it can be seen that the two classes produce different explanations relevant to their respective category. Given that we aimed to assess whether human can conclude the same prediction from visual explanations.

- **Presentation:** This section gets insights about users' view on the different explanation forms (e.g., visual, text). Participants were shown both visual and text explanation forms, and were asked to indicate their preference. These explanation types are demonstrated in Fig. 3.
- **Similarity assessment:** This part of the study assesses how similar the explanations generated from automated segmentation to ground truth explanations are through a similarity score (i.e., Fig. 5). The participants were asked to assign a similarity score between 0 and 10 by inspecting the grey portion of both images to observe any similar pattern, and also the texts of both images to see if the explanations were referencing the same features. To justify the scores they provided, they were also asked to provide a reason for their choice of score for each of the 10 scenarios.

In each section, participants were presented with a set of images that represents a visual explanation for a prediction and subsequently asked to respond to some questions. The option orders were mixed in each question to eliminate potential bias.

For this study, 50 participants were recruited through the Prolific platform (www.prolific.co). Participants were between age 18 and 54. 14 of them have between 1 and 10 years learning/work experience in an AI related field while 36 of them did not have any experience.

To ensure the evaluation quality, we only included individuals who have completed at least 5 surveys on the Prolific platform and have received at least a 95% approval rate. Overall, the participants took about 10 minutes on average to complete the survey and each participant was compensated with £2 for their time.

6. Results

6.1. Quantitative Results

From the comparison of the different explanation algorithms, most of the participants indicated that the contextual importance algorithm on Pascal VOC produced the best explanations (Mean = 7.34, SD (standard deviation) = 1.48) in the 10 questions in this section of the survey. The mean counts and standard deviations were: Grad-CAM (Mean = 1.04, SD = 1.12), CI-SLIC (Mean = 0.86, SD = 0.73), occlusion sensitivity (Mean = 0.76, SD = 0.66), and CI-Pascal VOC (Mean = 7.34, SD = 1.48) (see Fig. 8).

Assessing the confidence of the participants on the models based on the presence and absence of prediction scores, there seemed to be no

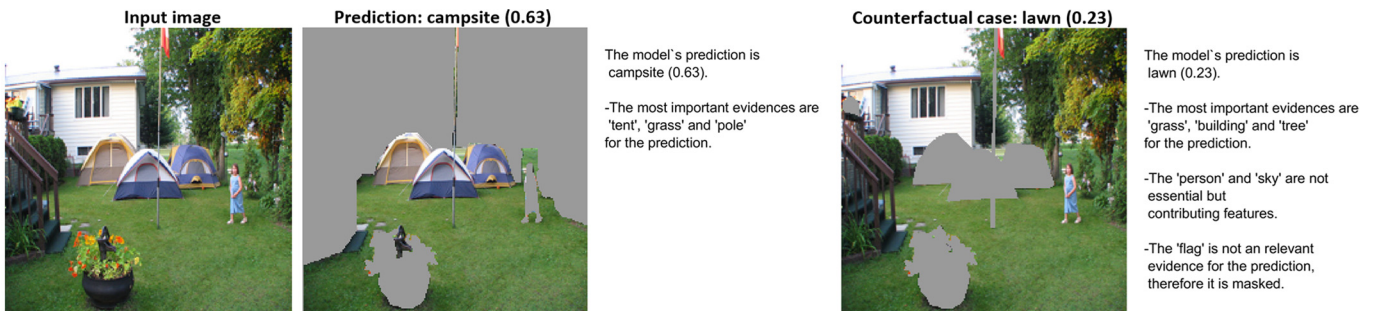


Fig. 6. Counterfactual explanations with CI show class discriminative visualization for differing classes. The campsite is the predicted class and the lawn is the counterfactual case. Visual explanations show the features that are most relevant to their respective category.

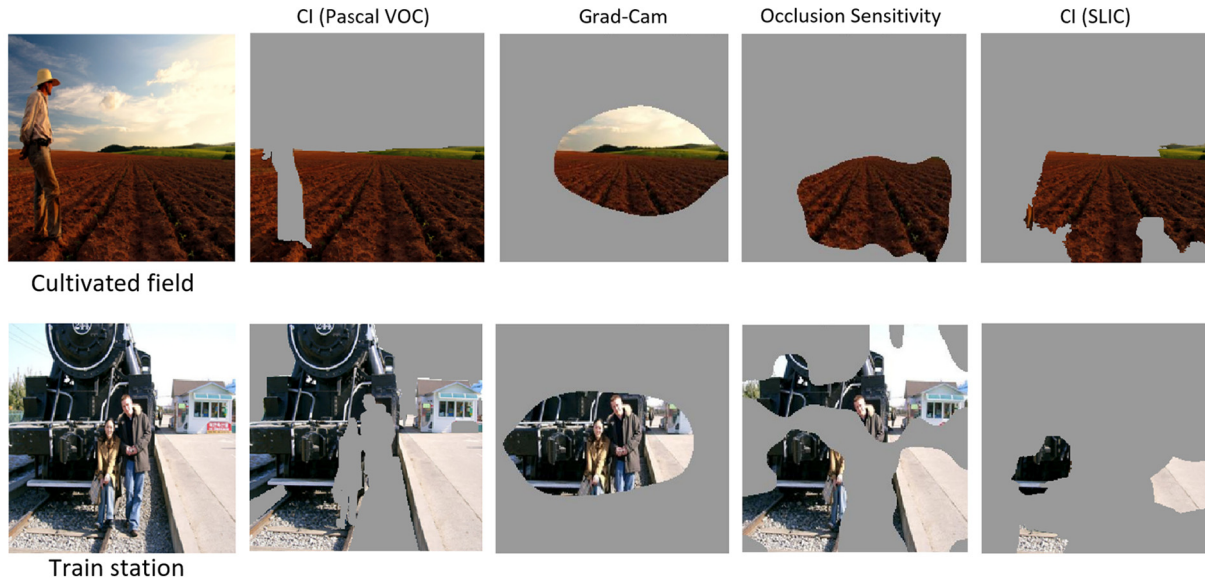


Fig. 7. Visual comparison of contextual importance (Pascal VOC-annotations) method with Grad-Cam [28], Occlusion Sensitivity [32], and contextual importance based on SLIC superpixel segmentations.

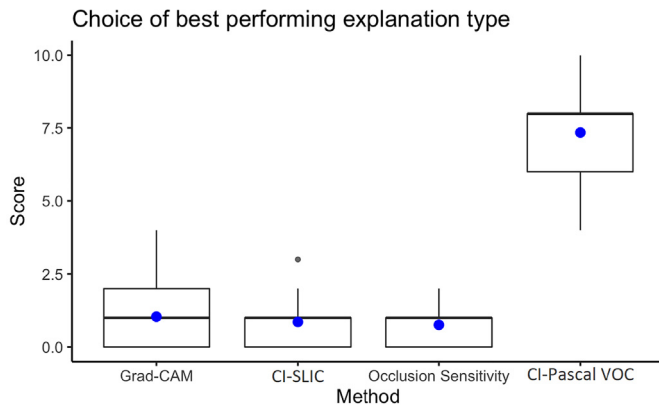


Fig. 8. Comparison of the performance of different explanation types. Based on the participants' judgement, CI algorithm produced the best explanations.

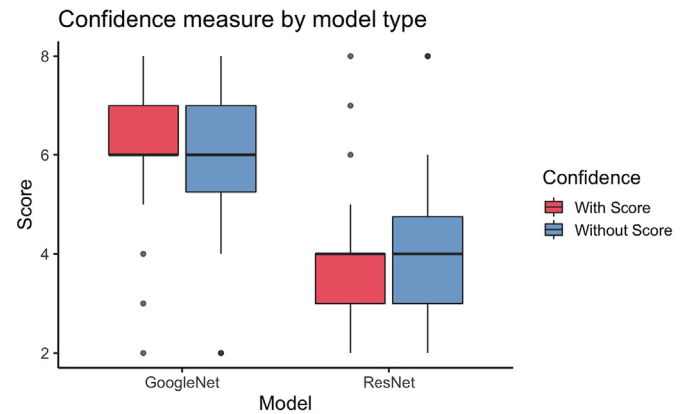


Fig. 9. Confidence measures based on CI explanations by model types. GoogleNet's predictions significantly increased the confidence of the participants higher than ResNet.

significant difference within models. However, across models, GoogleNet's predictions significantly increased the confidence of the participants higher than ResNet. The results for the explanations with prediction scores were: GoogleNet (Mean = 6.04, SD = 1.34), ResNet (Mean = 3.96, SD = 1.34). The results for the explanations without prediction scores were: GoogleNet (Mean = 6.18, SD = 1.12), ResNet (Mean = 3.82, SD = 1.12) as shown in Fig. 9.

Fig. 10 shows the participants' perception of model predictions based on CI explanations. The statistic for correct prediction (true positive) for the 10 samples was (Mean = 8.02, SD = 1.35) while that for wrong prediction (False positive) was (Mean = 1.98, SD = 1.35).

Participants were also asked to indicate their preferred explanation presentation forms. The different presentation forms included saliency map, textual explanation, visual map, and saliency map that contained texts. Results showed that most people preferred explanations presented in the form of saliency maps. Descriptive statistic of the result were: SaliencyMap (Mean = 2.12, SD = 1.79), Text (Mean = 0.22, SD = 0.51), visual map (Mean = 0.94, SD = 1.33), Saliency with text (Mean = 1.72, SD = 1.69) (see Fig. 11).

Finally, participants were asked to provide a similarity score (from 0 to 10) for ten different explanation cases shown to them. This scores shows how similar the explanations generated from automated

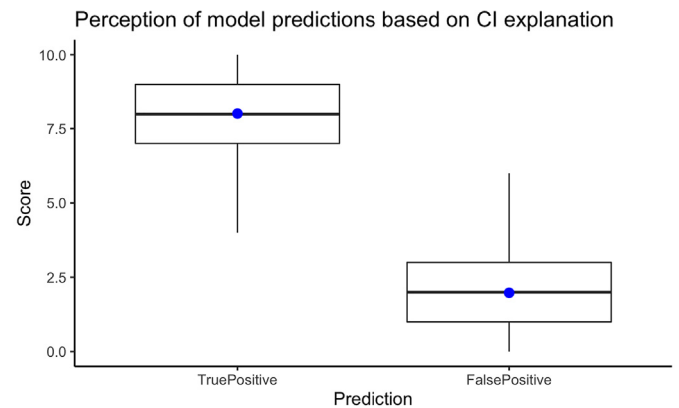


Fig. 10. Users' perception of model predictions based on CI explanation. There were very few false positives (or mis-identifications).

segmentation are to those generated from ground truth. As can be seen in Fig. 12, most of the participants agreed that the two explanations were very similar. The mean similarity score across the 10 question was

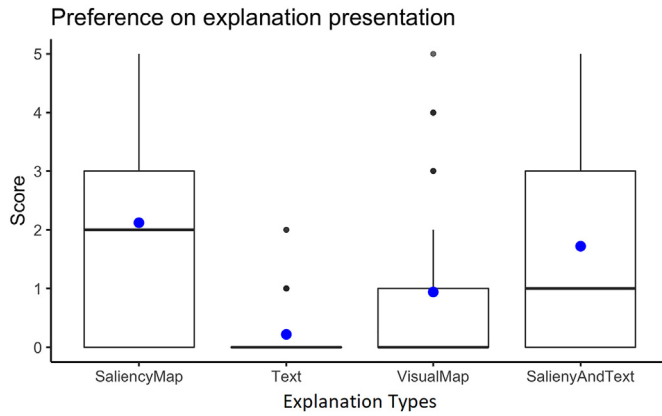


Fig. 11. Users' preference on explanation types. Explanations in saliency map forms were the most preferred.

6.93. The highest value was 8.20 and the lowest value was 4.43. The standard deviation across the ten questions was 1.12.

6.2. Qualitative results: reflections

The participants were asked to provide free comments on the explanation forms provided during the survey. The responses showed that many of the participants appreciated explanations containing visual cues.

"Visual map is more friendly for end-user."—P1. *"I most often chose the pictorial version because I liked it the most."*—P2

However, some think that the visual cues are not realistic enough and are sometimes misleading. They think that the visual explanations would be enhanced if visual cues are complemented with intelligible text and/or audio explanations.

"It is easier to have the image in front of with an explanation of why areas have been made grey or left in colour. This allows for a larger variety of people to understand the images."—P3

"Visual map shares similar information to saliency map with additional text-based explanation. If the picture is simple, the first option is the best. But if it is more complicated - the second option is better."—P4

For the similarity assessment part, in addition to the score provision, the participants were asked to state the reasons for the similarity score they provided for each of the 10 questions. Performing an inductive analysis from their free text responses, we discovered that most of the

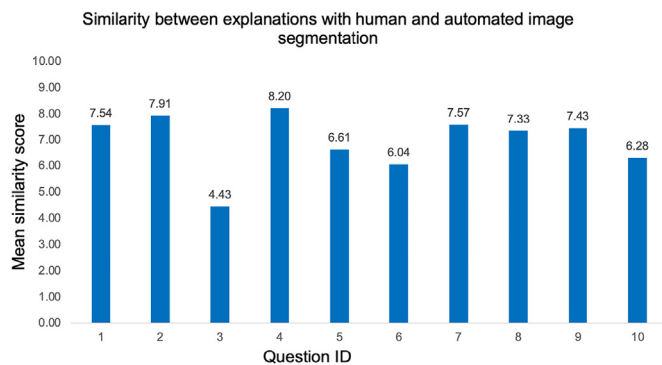


Fig. 12. Similarity scores show how similar the explanations generated from automated segmentation are to ground truth explanations; this result shows mean similarity scores for the 10 samples presented to the participants.

participants did not find any key difference between the two explanations. The response from of the participants who provided a score of 10 is provided below:

"The text is exactly the same. Both methods have the same areas greyed out."—P5

However, participants generally provided a low score for one of the scenarios where the image presented was that of a gas station. The response from one of the participants who provided a score of 1 is provided below:

"They are similar but not very similar, especially as for a gas station the most important evidence must include pavement, whereas the other doesn't. So the option on the right actually considers what's needed for a gas station."—P6

We however think that there is no significant difference the explanations provided from automatically segmented image and annotated images.

7. Discussion

Based on the visual comparison of the explanation algorithms, it is observed that each model has shown differing regions relevant for a prediction. In general, the grad-cam algorithm seems to be pointing to a single region even though the area contains irrelevant features. In the train station example, the category 'people' has a negative relevance as shown by our method and occlusion sensitivity while grad-cam included it as highly relevant for the prediction. Overall the CI results appear to be closer to the occlusion sensitivity results as they show similar regions as relevant. Moreover, CI is able to indicate the degree of importance of each region through text explanations and visual maps. As for the superpixel segmentation results (CI-SLIC), it gives reasonable results when the image contains simple visual patterns, while in complex background, explanations are rather unclear. This is because the explanations generated depend on the outcome of superpixel segmentation.

In general, the saliency maps produced by grad-cam and occlusion methods are showing where the neural network is 'looking' at but leaving out lots of important information and losing details significantly. They give considerable insight into the network learning. However, their results of spotting a region in a formless way provide less information than knowing the exact shape and size of the object in an image. This is particularly important in real-world applications such as in medical imaging or automotive driving where the pixel-level determination of the object is crucial.

This is further supported by the significant difference between CI and other explanation algorithms obtained from the user study. The results from the user study show that people have a higher preference for CI on Pascal VOC explanations (see Fig. 8). This could be due to high-resolution explanations generated through employing annotated images. Although the complete segmentation of complex shapes in the physical world is still a challenging task, it is clear that users would favour such explanations.

Furthermore, the results from the confidence assessment disclose that there is a relationship between the performance of the explanation and the architecture of the model it is applied on. This is evidenced in Fig. 13 where the participants preferred explanations applied on GoogLeNet for the same input. It is also interesting that the effect of the presence of the models' prediction scores was insignificant. This hints that the accuracy of a model may not generally improve peoples' confidence but explanations do.

It is expected that good explanations should be confined with the features specific to the category predicted. We assess whether human subjects can correctly identify the predicted category through the explanations. The results show that CI explanations can match humans'

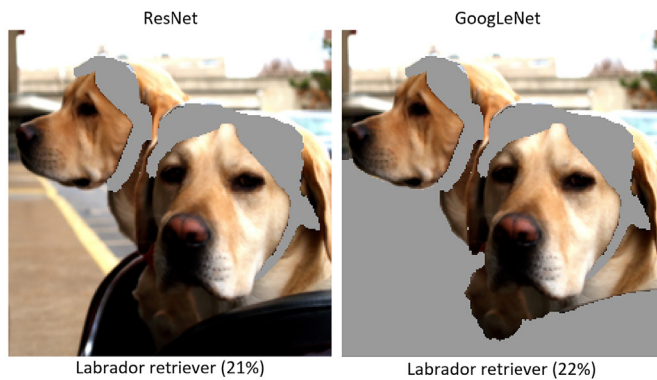


Fig. 13. Comparison of models and explanations with their prediction scores.

perception of scenes in the absence of the low importance features in an input sample.

Regarding the form of explanation presentation, while many of the participants preferred saliency map with text explanations, more participants preferred only saliency map explanations (see Fig. 11). It is also necessary to consider including what the network is ‘seeing’ along with where it is ‘looking’ at. With our explanation method, this issue is partly addressed by identifying class-relevant features using semantic categories. In this regard, the improvements in semantic segmentation algorithms for deep learning could provide critical explanation capabilities compared to current saliency maps. We also recommend that the option to choose a combination of explanation presentation forms be provided to users in real situations.

8. Conclusion and future work

We presented context based visual explanation method for DNNs which can also be applied to other machine learning models for generating instance-level explanations. We have shown the implementation with both annotated and automatically segmented (semantically) images. Semantic information gives the properties of the image scene which enables us to assess the importance of each component for a prediction in a given context. We visualized the significance of each component contributing to the prediction score in a way that is easily understandable by common users. We generated visual and text-based explanations using saliency maps, a color bar graph, and descriptive phrases listing the features and their importance. These explanations could be extended to designing dynamic templates and visualizations by taking user's characteristics into account and testing the usability of these explanations in user studies.

The visual comparisons with three methods have shown varying saliency maps. Unlike these methods, our results present justifications beyond the vague representation of the important parts of an image. It uses semantic categories to distinguish the degree of importance of different parts of an image. We confirmed these findings using a user study and our claim about the elegance of the *CI* algorithm was validated. The results motivate current saliency map methods towards specifying not only where the network is pointing but also what it is seeing.

As the limitation of the work, the examples used are not those in which explanations are critical. In the future work, we will extend *CI* to be able to generalise for different input images (whether already semantically segmented or not). We will also apply *CI* to real-world critical examples (e.g., autonomous driving) and conduct a more immersive evaluation with different stakeholder categories in a lab setting.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

We would like to thank the reviewers for their helpful comments. We would also like to thank Monika Jingar and Zahoor Ul Islam for their feedback on the survey. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Daniel Omeiza thanks the UK's Engineering and Physical Sciences Research Council (EPSRC) RoboTIPS project: Developing Responsible Robots for the Digital Economy, grant reference EP/S005099/1.

References

- [1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (11) (2012) 2274–2282.
- [2] S. Anjomshoe, K. Främling, A. Najjar, Explanations of black-box model predictions by contextual importance and utility, *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer 2019, pp. 95–109.
- [3] S. Anjomshoe, T. Kampik, K. Främling, Py-ciu: a python library for explaining machine learning predictions using contextual importance and utility, *IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI)*, 2020.
- [4] S. Anjomshoe, L. Jiang, K. Främling, Visual explanations for DNNs with contextual importance, *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer 2021, pp. 83–96.
- [5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PLOS ONE* 10 (7) (2015) e0130140.
- [6] O. Biran, K. McKeown, Justification narratives for individual classifications, *Proceedings of the AutoML Workshop at ICML*, vol. 2014 2014, pp. 1–7.
- [7] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high-definition ground truth database, *Pattern Recognit. Lett.* 30 (2) (2009) 88–97.
- [8] V. Buhrmester, D. Münch, M. Arens, Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey, 2019 arXiv:1911.12116 (arXiv preprint).
- [9] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE 2018, pp. 839–847.
- [10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *Proceedings of the European Conference on Computer Vision (ECCV)* 2018, pp. 801–818.
- [11] G. Csúrk, D. Larus, F. Perronnin, F. Meylan, What is a good evaluation measure for semantic segmentation? *Bmvc*, vol. 27, 2013 10–5244.
- [12] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, *Commun. ACM* 63 (1) (2019) 68–77.
- [13] K. Främling, *Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère*, Institut National de Sciences Appliquées de Lyon, Ecole Nationale Supérieure des Mines de Saint-Etienne, France, 1996 PhD Thesis.
- [14] L.A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, *European Conference on Computer Vision*, Springer 2016, pp. 3–19.
- [15] F. Hohman, H. Park, C. Robinson, D.H. Chau, Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations, 2019 arXiv:1904.02323 (arXiv preprint).
- [16] D.H. Park, L.A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: justifying decisions and pointing to the evidence, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, pp. 8779–8788.
- [17] T. Lei, X. Jia, T. Liu, S. Liu, H. Meng, A.K. Nandi, Adaptive morphological reconstruction for seeded image segmentation, *IEEE Trans. Image Process.* 28 (11) (2019) 5510–5523.
- [18] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [19] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, S. Liu, Towards better analysis of deep convolutional neural networks, *IEEE Trans. Vis. Comput. Graphics* 23 (1) (2016) 91–100.
- [20] A. Mahendran, A. Vedaldi, Visualizing deep convolutional neural networks using natural pre-images, *Int. J. Comput. Vis.* 120 (3) (2016) 233–255.
- [21] T. Miller, Explanation in artificial intelligence: insights from the social sciences, *Artif. Intell.* 267 (2019) 1–38.
- [22] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2014, pp. 891–898.
- [23] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, *Advances in Neural Information Processing Systems* 2016 3387–3395.

- [24] D. Omeiza, S. Speakman, C. Cintas, K. Weldermariam, Smooth Grad-cam++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models, 2019 [arXiv:1908.01224](#) (arXiv preprint).
- [25] D. Omeiza, H. Webb, M. Jirotko, L. Kunze, Explanations in Autonomous Driving: A Survey, 2021 [arXiv:2103.05154](#) (arXiv preprint).
- [26] M.A. Rahman, Y. Wang, Optimizing intersection-over-union in deep neural networks for image segmentation, International Symposium on Visual Computing, Springer 2016, pp. 234–244.
- [27] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you? Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM 2016, pp. 1135–1144.
- [28] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 618–626.
- [29] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, Proceedings of the 34th International Conference on Machine Learning – Vol. 70, JMLR.org 2017, pp. 3145–3153.
- [30] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2013 [arXiv:1312.6034](#) (arXiv preprint).
- [31] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for Simplicity: The All Convolutional Net, 2014 [arXiv:1412.6806](#) (arXiv preprint).
- [32] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, European Conference on Computer Vision, Springer 2014, pp. 818–833.
- [33] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1452–1464.
- [34] A.P. Zijdenbos, B.M. Dawant, R.A. Margolin, A.C. Palmer, Morphometric analysis of white matter lesions in MR images: method and validation, IEEE Trans. Med. Imaging 13 (4) (1994) 716–724.