Pattern Recognition Letters 150 (2021) 228–234

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Perturbation-based methods for explaining deep neural networks: A survey

Maksims Ivanovs*, Roberts Kadikis, Kaspars Ozols

Institute of Electronics and Computer Science, Dzerbenes str.14, Riga LV-1006, Latvia

ARTICLE INFO

Article history:
Received 16 November 2020
Revised 5 May 2021
Accepted 18 June 2021
Available online 21 July 2021

Edited by Maksims Ivanovs

MSC:
41A05
41A10
65D05
65D17

Keywords:
Deep learning
Explainable artificial intelligence
Perturbation-based methods

ABSTRACT

Deep neural networks (DNNs) have achieved state-of-the-art results in a broad range of tasks, in particular the ones dealing with the perceptual data. However, full-scale application of DNNs in safety-critical areas is hindered by their black box-like nature, which makes their inner workings nontransparent. As a response to the black box problem, the field of explainable artificial intelligence (XAI) has recently emerged and is currently rapidly growing. The present survey is concerned with perturbation-based XAI methods, which allow to explore DNN models by perturbing their input and observing changes in the output. We present an overview of the most recent research focusing on the differences and similarities in the applications of perturbation-based methods to different data types, from extensively studied perturbations of images to the just emerging research on perturbations of video, natural language, software code, and reinforcement learning entities.

1. Introduction

Artificial Intelligence (AI), in particular its subfield Machine Learning (ML), plays an increasingly prominent role in science, technology, industry, and public life. Success of ML is to a large extent due to advances in research on deep neural networks (DNNs). Originally inspired by the structure and functionality of the brain, DNNs consist of multiple layers of nonlinear units known as artificial neurons. While the intricate architectures of state-of-the-art DNN models resemble biological neural systems only remotely, they have recently demonstrated remarkable results in solving various problems that used to be perceived as amenable solely to human intelligence, from recognising speech [73] and visual objects [34] with high accuracy to playing chess and go [61] and driving autonomous vehicles [32]. Unfortunately, many of the most efficient ML methods also tend to be the least transparent ones [48]; that is particularly true for DNNs, which operate as black boxes, the inner workings of which are still poorly understood. As a consequence, DNN-based AI systems cannot be fully trusted yet with making decisions in safety-critical or high-stakes applications such as medicine, smart assistance to elderly and disabled people, finance, criminal justice, or navigation of unmanned vehicles in complex environment [6,53].

In response to the pressing need to make AI systems more transparent and therefore more trustworthy, a new field of study, most frequently referred to as explainable AI (XAI) or interpretable AI, has started to emerge. The present survey is concerned with a particular paradigm in XAI research, perturbation-based methods. These methods investigate properties of DNNs by perturbing the input of a model, e.g. by occluding part of the input image with a mask or replacing a word in a sentence with its synonym, and observing the changes in the output of the model. XAI research on perturbations has begun with the seminal study by Zeiler and Fergus [70], in which simple grey masks were applied to images. Since then, image perturbation methods have developed to make their output more accurate [25,27,52,55,55] as well as faster [18,69]. Furthermore, in recent studies, perturbations have also been applied to other input data types such as videos [43,47], natural language [45], software code [13], and reinforcement learning (RL) entities [31,37,54]. There are good reasons for interest in them, as they offer several advantages: thus, observing the input-output relationship is a natural and intuitive way of exploring black box-like models; perturbations allow to analyse models dynamically rather than treating them as invariant objects; finally, they are normally applicable to any DNN regardless of the architecture of a model. As a result, perturbation-based methods is

* Corresponding author.
E-mail address: maksims.ivanovs@edi.lv (M. Ivanovs).

https://doi.org/10.1016/j.patrec.2021.06.030

one of the most popular and arguably promising approaches in XAI research.

XAI studies employing perturbation-based methods are typically focused on a particular input data type, e.g. images (most commonly by far), video, or text. While such a focus is natural for an empirical study concerned with a specific problem, it may not capture the general characteristics of using perturbations to better understand DNNs, i.e. characteristics that do not depend on the modality of the input data. The opposite is true as well: while some studies do highlight the differences in applying perturbations e.g. to videos vs images [43] or text vs images [45], such comparisons are usually limited in their scope. The goal of the present survey is to outline both input modality-agnostic properties and input modality-dependent specifics of perturbation methods; rather than attempting to cover the whole of the vast and diverse range of existing ML methods, we limit the scope of the survey to perturbation-based methods specifically for DNNs.

The survey is organised as follows. In Section 2, we discuss the black box problem inherent to DNNs and challenges for the application of deep learning methods to real-world tasks that it entails. In Section 3, we provide an overview of the field of explainable AI, which to a large extent emerged as a response to the black box problem. In Section 4, the central section of the article, we survey perturbation-based methods. Finally, in Section 5, we offer conclusions and directions for future work.

2. The black box problem

DNN models are essentially complex nonlinear functions $f: X \to Y$ mapping from an input space X to an output space Y. They are aptly characterised as black boxes, because the respective mapping function in their case is obtained by means of an opaque learning process [27] and therefore is opaque itself as well. The term 'black box' is the opposite of the 'white box' (or the 'glass box' [36]), which refers to a fully transparent system [67][1]. While the transparency of 'white box' models makes them comprehensible (at least for experts) and therefore trustworthy, the black box-like nature of the DNNs posits the problem of (the lack of) trust. In particular, one cannot be sure whether a DNN model makes a correct prediction on the previously unseen dataset because it did learn the salient features of the data or, just the opposite, because it just memorised some irrelevant idiosyncrasies in the training dataset that also happen to occur in the test dataset. In the former case, the model can credibly generalise in the real world, whereas in the latter case it relies on the spurious correlation and therefore fails to be of good use [42]. To exemplify, a model that has learnt salient features classifies an image of a plane correctly because it may recognise that the plane has wings, fuselage, cockpit, etc. A remarkable example of the opposite case is provided in [42]: a classifier identified images of horses in PASCAL VOC 2007 dataset [23] with unusually high precision, as it learned a certain copyright tag that was added to some of the horse images in the dataset. As a consequence, it would misclassify the image of the horse without the tag as well as wrongly classify an image of e.g. a car as that of a horse if the 'horse' source tag was added to it.

The lack of detailed understanding of what knowledge DNNs possess makes it difficult to fully rely on them in many safety-critical domains such as navigation of autonomous cars and medicine, where they could otherwise be of much help. The concerns that lack of transparency of DNNs can lead to undesirable consequences are well justified. Thus, the general public and experts alike are increasingly apprehensive that AI systems can inherit a bias against people of a particular race, ethnic background, or gender that has seeped into the training data [15,74]. One of the first and most notorious examples of that was Google Photo's ML algorithm labelling the photos of black people as those of gorillas [9]. Although Google resolved the problem after it attracted public attention,[2] it has reoccurred recently in less conspicuous, bet perhaps even more alarming form: another product of the same company, Google Vision Cloud, labelled an image of a black person holding a thermometer as a 'gun' image, whereas a similar image with a light-skinned person was labelled as an 'electronic device' image [38]. It is not hard to imagine the possible dramatic consequences of applying a DNN model suffering from such a bias to criminal profiling or other safety-critical tasks in an uncontrolled manner. Furthermore, a bias in a trained DNN model may be not as glaring as in the above examples or less accessible to public scrutiny, thus remaining unnoticed and resulting in unfair and discriminatory decisions regarding e.g. one's finances (e.g. application for a mortgage), health, or legal status.

Yet another obstacle for deploying DNNs in fail-safe tasks is that they may be subject not only to unintentional mistakes, but also to malicious attacks: thus, convolutional neural networks, which are ubiquitously used in various computer vision tasks, can be 'fooled' into misclassifying an object by modifying an input image so subtly that the changes in it would seem to be of little consequence to a human observer [6]. Such adversarial attacks pose a threat per se as well as undermine decision-makers' trust in DNN-based systems in general: indeed, if putting several innocuously-looking black and white stickers on a physical traffic Stop sign can 'fool' a state-of-the-art DNN into classifying it as a Speed Limit 45 sign [24], it might be rather problematic to persuade lawmakers to entrust a driverless vehicle controlled by similar DNN models with full autonomy on the streets. As Ribeiro et al. [55] observe, 'if the users do not trust a model or a prediction, they will not use it'; to overcome their hesitancy, a plausible explanation of the model may be helpful.

The concerns for the possible impact of the lack of transparency in deep learning applications have also been reflected in the law. In EU, the EU General Data Protection Regulation (GDPR) stipulates the legal right of end-users to explanations of algorithmic decisions affecting them including explanations of the decisions made by AI systems [30]. Furthermore, according to GDPR, end-users have the right to know whether the decision-making algorithm was non-discriminatory and impartial. That entails the obligation on ML experts' part to be able to explain to non-experts how exactly their AI system reached its decision from the input data [16]

Last but not least, from the point of view of AI researchers, the black box problem causes the feeling of intellectual dissatisfaction and hinders further development of deep learning methods, because, as it has been pointed out [70], our lack of understanding of how and why DNN models work reduces the development of better models to a mere trial-and-error approach.

3. Explainable artificial intelligence

3.1. Terminology and scope

In response to the black box problem, the scope of ML research has expanded, comprising not only such traditionally topical is-

[1] Arguably, it would be more precise to use the term 'grey box' to describe DNNs, as 'black box' implies that the system is intransparent and its insides might be completely inaccessible [67], whereas 'grey box' refers to a system that is less transparent than 'white box', but for which we still have some limited knowledge of its inner workings and fundamental principles it is built upon [12,40]. However, taking into account the prevalence of the term 'black box' in ML literature as well as the connotations of mystery evoked by 'black' but not by 'grey', it is unlikely that the term 'grey box' will gain currency in the AI community.

[2] By disabling the labelling of any image as a gorilla, chimpanzee, or monkey [65], which was hardly the most convincing solution.

sues as performance of models, improvement of algorithms, and increase in training and inference speed [58], but also explanations of how DNN models infer their predictions. There is quite a lot of variation in referring to field of study dealing with the (lack of) transparency of AI methods. The most frequently used terms are explainable AI and interpretable AI; some other terms in use are intelligible intelligent systems, context-aware systems, and software learnability [1], responsible AI [8], and safe AI [6]. All in all, a comprehensive survey of terminology [50] lists 14 relevant terms; as a consequence of such diversity, the meaning of different terms often overlap, and there are no commonly accepted strict definitions of the terms. Thus, while 'explainable' and 'interpretable' are often used interchangeably in this context [14,22,50], some preference for the latter term is reported to have been observed in the ML community [2]; furthermore, some authors clearly distinguish between these terms [21,28]. In the present survey, we prefer to refer to explainability, especially as that allows for a convenient and pithy abbreviation, XAI. However, in general, we perceive 'explainable' and 'interpretable' and their derivatives as largely equivalent in meaning and interchangeable in use.

In addition to the ambiguity in terminology, it is also difficult to draw a clear demarcation line between XAI and other subfields of ML research, as much of research on ML is at least to some extent concerned with understanding AI better. To provide an example of a borderline case, while research on adversarial attacks on DNNs [29,64] is typically presented as a field of study in its own rather than part of XAI or even contrasted with it [25], there are also examples to the contrary: thus, a recent study [16] uses adversarial stimuli to visually explain recurrent neural networks.

Regardless of the disagreements about the terminology and the scope of the XAI field, research on XAI is developing at a fast pace, which is demonstrated by the exponential increase in the number of publications in the last decade [2,8]. As XAI research is concerned not only with AI algorithms per se, but also involves human users, the field of XAI is multidisciplinary, comprising not only ML, but also aspects of visual analytics, human-computer interaction (HCI), and psychology [50].

3.2. Related work

Several surveys of XAI have recently been published. Some of them emphasise the interconnection between ML methods and human factor-related methods in XAI: thus, Mohseni et al. [50] aims at capturing the multidisciplinary nature of the field; Abdul et al. [1] focuses on the cross-links between HCI and XAI; Choo and Liu [17] discusses applications of visual analytics to the design of explainable deep learning systems; [35] surveys visual analytics tools for DL applications; Bhatt et al. [11] is concerned with whether and how organisations use XAI for the benefit of stakeholders. Other surveys are more focused on ML methods per se: thus, Guidotti et al. [33] present a comprehensive overview and classification of XAI techniques, and there is a number of other general surveys of the ML methods in XAI [2,8,51,68]. Furthermore, there are surveys with a more narrow scope, e.g outlining the use of XAI methods in a particular domain such as medicine [5,66] or natural sciences [57], or concerned with XAI in particular subfields of deep learning such as convolutional neural networks [72] or RL [53].

In the present survey, we do not attempt to comprise the multidisciplinary nature of XAI research, but rather focus specifically on ML methods in XAI. We further narrow the scope of our work by focusing on perturbation-based methods and their application to DNNs rather than to the whole range of ML methods. To the best of our knowledge, there has been only one XAI survey [56] specifically concerned with perturbation-based methods. We aim at extending its contribution both by covering more recent publications, as recent research significantly advances the state of the art of research on perturbation-based methods, and highlighting the differences and similarities in applying perturbations to the different types of the input data.

4. Perturbation-based methods

4.1. Overview

Perturbation-based methods aim at exploring DNNs by modifying the input of a model, be it pixels in an image, words in a text, or similar elements of some other data type, and observing the changes in the output. The observations of the changes in the output are expected to indicate which parts of the input are particularly important for the inference. The importance of the perturbed element is estimated by comparing the output with the element is present with the output in its absence: it is considered important if its removal changes the output considerably. To exemplify, in case of image classification, considerable changes would make the classifier to assign an image to a different class than previously. Generally speaking, this approach to explaining ML systems is known as input attribution, and perturbation-based methods (as well as gradient-based methods, which are briefly outlined below for the sake of comparison) are therefore attribution methods. The main challenge for attribution methods is the combinatorial explosion that would occur if one attempted at going through all elements of the input as well as all their possible combinations to observe how each of them would change the output [39]. While a consequential removal of all possible combinations of elements from a particular input item would allow to compute the part of the input (e.g. a region of the image) with the decidedly highest impact on the output, it is just not feasible in practice due to the enormous cost in terms of time and required computational power.

There are two main approaches to tackling the complexity issue. First, it is possible to compute the importance of the element using the gradients of a model gradients as a proxy [7,60]; this approach is known as gradient-based attribution. For that purpose, a modification of backpropagation algorithm is typically used, as it allows to retrace the flow of the information in a DNN from the output back to the input. Deploying modifications of backpropagation algorithm is computationally efficient, as that requires only a single forward and backward pass through the DNN. However, gradient-based attribution methods tend to be noisy, especially in case of large DNNs [7]; furthermore, some backpropagation-based methods produce the same saliency map regardless of network parameters, which implies that they actually perform partial image recovery and cannot pass a sanity check [3]. As a consequence, while backpropagation-based methods may outline average properties of a model, they often fail to capture its more refined characteristics [25]. The second approach is to use perturbation-based methods; in that case, the main challenge is to select plausible groups of elements in the input and apply perturbations to test them in a reasonable period of time to obtain an approximation to an optimal solution [43]. A remarkable advantage of perturbation-based methods is their dynamism: while many other approaches investigate the model as an invariant object [45], perturbations allow to query it repeatedly and develop and test hypotheses about it 'on the fly'. Furthermore, they can be applied to any model regardless of its architecture: while model-dependent methods (e.g. many of backpropagation-based methods) require access to the internal information of the model to generate explanations and therefore are restricted in their use as well as often offer only low-resolution results [69], perturbation-based methods do not suffer from such shortcomings. As model-agnostic methods, they do not need to use the internal information of a model and can therefore be used to explain the predictions of almost all types of ML models including DNNs [69].

Table 1
Summary of perturbation-based methods.

Method	Data Type	Dataset	Main Evaluation Methods	Applications
Occlusion [70]	images	ImageNet [20]	qualitative analysis	understanding DNN inference
LIME [55]	images	custom	human experiments	assessing and improving untrustworthy models, understanding DNN inference
RISE [52]	images	PASCAL VOC07 [23], MSCOCO2014 [44], ImageNet [20]	pixel deletion and insertion scores, pointing game	explaining image captioning models
Meaningful perturbations [27]	images	ImageNet [20]	weakly supervised localisation, pointing game	understanding DNN inference
Real-time saliency [18]	images	ImageNet [20], CIFAR-10 [41]	weakly supervised localisation, saliency metric for bounding boxes	fast saliency detection
Extremal perturbations [25]	images	PASCAL VOC07 [23], MSCOCO2014 [44], ImageNet [20]	pointing game	understanding DNN inference
MFPP [69]	images	PASCAL VOC07 [23], MSCOCO2014 [44]	pointing game	understanding DNN inference
Temporal Masks [47]	video	20BN-Something-something-V2 [46], KTH actions dataset [59]	drop ratio and difference for normalised 'freeze' and 'reverse' perturbations	comparison between video classification by 3D CNNs and C-LSTM
Perturbation-based video explanation [43]	video	UCF101-24 [62], EPIC-Kitchens [19]	pointing game, temporal pointing game	comparison between video classification by 3DCNNs and CNN-RNN
NLIZE [45]	natural language	wordNet [49]	qualitative analysis	visualisation system
AutoFocus [13]	software code	custom: web-crawled code from GitHub	qualitative analysis	visualisation
Saliency method for Atari agents[31]	RL entities	Atari env [10]	human experiments	visualisations to help non-experts understand RL agents
Object saliency maps for deep RL networks[37]	RL entities	Atari env [10]	human experiments	visualisations to help non-experts understand RL agents
SARFA [54]	RL entities	Atari env [10], custom chess dataset, Minigo env*	human experiments	visualisations to help understand RL agents

* https://github.com/tensorflow/minigo

4.2. Evaluation methods

The results obtained with perturbation-based methods are typically some representations of saliency: thus, in case of images, they are saliency maps that highlight the importance of the parts of the image for the output of the model. Saliency maps are also used for videos and RL entities, whereas for textual data such as natural language or software code, saliency can be represented as different colours or varying brightness of words [45] or code tokens [13].

There are different ways to evaluate obtained saliency representations (see Table 1). The simplest approach is qualitative analysis - evaluation of the morphology and granularity of a saliency map performed by the authors themselves. While such analysis is invariably performed in all surveyed studies - after all, it is only natural that the authors evaluate the quality of the obtained saliency representations and share their insights with the reader - it is usually only the first evaluation step. To make human evaluation of saliency representations more objective, a group of respondents can be engaged. In some studies, e.g. [31,37,55], respondents are members of general public, which allows to verify that the output of the method is comprehensible for non-specialists, whereas in others they are experts: thus, in Liu et al. [45] linguists evaluated natural language perturbation model; in Puri et al. [54], expert-level chess players evaluated whether the method correctly indicated the saliency of chess pieces in puzzles. Qualitative human evaluation of saliency representations is the main approach for perturbations of textual data and RL entities, whereas saliency maps obtained by perturbing images and videos allow to use more elaborate quantitative metrics. The most popular of them (see e.g. [25,69]) is pointing game [71], the accuracy of which is given by $\frac{1}{N}\frac{\#Hits}{\#Hits+\#Misses}$, N being the number of relevant categories in the dataset, and a hit occurring when the highest saliency point lies within the human-annotated bounding box of an object. In [43], spatial pointing game was extended to temporal dimension for evaluating perturbations of video. Another common evaluation metric (see e.g. [18,27] is weakly supervised object localisation, which considers as hit such a localisation of the bounding box of the most salient areas of the obtained map that it overlaps significantly (e.g., $\geq$0.5) with the human-annotated bounding box. To obviate the need for human involvement, Petsiuk et al. [52] introduce two automatic evaluation metrics, deletion and insertion. Deletion measures a decrease in the probability of the predicted class as more and more salient pixels are removed, whereas the insertion metric measures the increase in probability as more and more pixels are added. Some other metrics of interest are found in [18], where the tightest rectangular crops that contain the entire salient regions were found and fed to the classifier to verify whether it is able to recognise the classes of interest, and in [47], where the results of 'freeze' perturbations, which remove motion data through time, were compared with the results of 'reverse' perturbations, which inverse the sequential order of the frames, after normalising both.

4.3. Perturbations of different input types

4.3.1. Images

Perturbations of images are performed by removing or inserting information by means of applying occlusion masks, blurring, or replacing parts of an image. That can be done pixel-wise or patch-wise, which will result in different granularity of the ob-

tained saliency map [69]. Thus, pixel-wise perturbations are spatially discrete and therefore represent saliency more accurately in terms of location, yet their higher granularity leads to worse representation of the semantics of salient objects [69]. Patch-wise methods smooth over fine details of saliency yet deliver more comprehensible saliency maps, as the boundaries in the maps correspond better to object boundaries [69]. In both cases, the main challenge is to find an appropriate scope and shape of the perturbations.

Some landmarks in the development of perturbation-based methods for images are as follows. Occlusion [70] produces changes in the output of a classifier by perturbing input images by sliding a gray square over them. A similar approach is employed in RISE [52]: input images are occluded with random occlusion patterns that are produced by sampling small (7x7 pixels) binary masks and then upsampling them to larger resolution with bilinear interpolation. The main limitation of both Occlusion and RISE is that they do not consider the morphology of the objects in the image and therefore yield only approximate results. LIME [55] employs occlusions of superpixels and approximates networks with linear models; as superpixels are coarse-grained, obtained saliency maps suffer from rather low precision. To improve the spatial accuracy, the method of meaningful perturbations [27] introduces optimisation of the shape of perturbation masks so that they would blur the input image as little as possible while still decreasing class score as much as possible. Real-time saliency [18] develops the work of [27] further and offers a fast (single forward-pass) method of obtaining optimal perturbation masks by generating them with a second neural network. The method of extremal perturbations [25] is concerned with finding such a perturbation mask for a given area that has a maximal effect on the output of the DNN in comparison with all other masks possible for that area. To optimise the masks, [26] also introduce the smooth max operator. Another recent approach, MFPP [69], employs morphological fragmentation to divide the input images into multiscale fragments and produces a perturbation mask by randomly masking some of them. Currently, state-of-the-art results of image perturbations are obtained with the methods in [25,69], as they allow to obtain more refined saliency maps than methods in previous work. Furthermore, MFPP [69] outperforms the method of extreme perturbations [25] in terms of speed.

4.3.2. Video

Due to the nature of the video data, perturbations of video have not only a spatial, but also a temporal dimension, which inflates searching space and increases requirements for processing time and computational capacity [43]. In contrast to the notable progress in research on image attribution, there are only few studies on attribution methods for videos. The general direction of work is to adapt existing image attribution approaches for videos input [43]. Most of the studies published so far, e.g. [4,60,63] employ backpropagation-based methods. Recently, perturbation-based methods have been applied to the video data in [43,47]. Mänttäri et al. [47] extend the concept of meaningful perturbation introduced by Fong and Vedaldi [27] to the temporal dimension to identify the temporal part of a sequence that has the greatest impact on the output of the DNN classifier. Li et al. [43] build on the work of Fong and Vedaldi [26], extending the extremal perturbation method from the two dimensions of images to the three dimensions of video.

4.3.3. Natural language

Applications of perturbations to the textual data are currently still at an early stage of development. One of the likely reasons for that is that perturbations of text involve various data-specific challenges such as the discrete nature of words: thus, while perturbing a single pixel in the image is unlikely to affect the decision of a classifier, small alterations of words can dramatically change the meaning of a sentence and therefore the output of a model [45]. In [45], perturbations are applied to language inference models, which are concerned with investigating whether the relationship between two sentences is that of entailment (one sentence can be inferred from the other), contradiction (one sentence contradicts the other), or a neutral one (sentences inform about different or unrelated things) [45]. Language inference can be investigated with attention mechanisms, i.e. by exploring how alignment between words in different sentences influences a prediction [45]. Liu et al. [45] have developed a visual analytics environment, in which NLP (Natural Language Processing) experts can explore how perturbations of sentence input, attention, or prediction (performed by a DNN) affect each other. Due to to the dynamic nature of the perturbation paradigm, the users can develop intuitions, transform them into hypotheses, and immediately test them. More specifically, the environment enables the automated or user-guided perturbations of the input sentence (i.e., replace words) or attention (i.e., alter the alignment between sentences) inside the model, and the perturbation of the prediction (i.e., adjust the prediction by making updates to the model) [45]. All that can done by employing an automated sentence perturbation scheme that replaces nouns and verbs by their synonyms in the WordNet [49], a lexical database standardly used in NLP research. While the specific setup described in [45] is suitable only for the natural language inference task, the authors point out that due to the modular design of the environment and similarities among NLP models it can be extended to deal with other tasks as well.

4.3.4. Software code

While software code is textual in its nature and therefore to some extent resembles natural language data discussed in *4.2.3*, we make it a subject of a separate section, in particular as code is different from other text types by the virtue of not only describing, but also actually *doing* something. The application of neural networks to the analysis of the software code is a subfield of ML that has only started to develop; as a consequence, the first studies exploring the application of XAI methods to the DNNs performing such tasks have appeared only recently. Bui et al. [13] apply AutoFocus, a framework for rating and visualizing the importance of input elements, to the attention networks trained for algorithm classification, i.e., networks that classify the algorithm implemented in a given program. Perturbations of the input of the attention networks investigate attribution of the elements of software code, as they allow to explore correlation between the perturbed elements of the code (statements in case of this study) and the class of algorithms to which the networks assign the input code. Attention scores of individual statements can be used for visualizing a program and help the users of the framework to understand the program without the need to peruse it [13]. Similarly as with perturbations of other data types, the key challenge is to optimally determine the size of the element to be perturbed. Thus, [13] observe that while the perturbations deployed in their study delete one statement of the code at a time, another, perhaps even more promising approach is to increase the size of the perturbation mask, i.e. delete multiple code elements at once. Presumably, that would identify the minimal input (i.e. amount of code) required for correct algorithm classification.

4.3.5. Reinforcement learning entities

XAI research for understanding RL entities, i.e. RL agents and environment, has just started to emerge [53]. The temporal and interactive nature of RL systems presents a challenge for XAI methods, as DNNs in RL have to select sequential actions whose effects can interact over long periods of time. Methods for RL-based agents are mainly inspired by XAI methods in computer vision.

Thus, to explain behaviour of RL agents, Greydanus et al. [31] perturb different parts of the input frames by applying Gaussian blur and generate saliency maps by computing differences in a policy (actor) distribution and a value (critic) estimate between the original and perturbed state. By applying their method to several games on Atari 2600 platform, Greydanus et al. [31] demonstrate what parts of the RL environment well-performing RL agents attend to, how the agents improve during the training, and how to detect situations when the agent makes seemingly 'right' (i.e. high-reward) decisions for wrong reasons. In another study, Iyer et al. [37] obtain saliency maps by masking objects in the original image with background colour and calculating the difference between the respective action values. The task that [37] apply their method to is Atari games as well.

As [54] point out, approaches in [31,37] have two major limitations. First, they highlight not only the features of interest, but also those that are not salient for explaining the RL agent, but whose perturbations nevertheless affect its actions [54]. Second, they also highlight features that are not salient at all to the action of the agent that one is interested in [54]. To deal with these shortcomings, Puri et al. [54] propose a perturbation-based approach SAFRA (Specific and Relevant Feature Attribution) for generating saliency maps for RL agents. SAFRA addresses the above limitations by means specificity and relevance. Specificity refers to focusing on the effect of perturbation only on the action-value of the action of interest, whereas relevance downweights the non-salient features that alter the expected rewards of actions other than the action to be explained. Puri et al. [54] use SARFA to explain the actions of the RL agents is board games (Chess and Go) and Atari games (Breakout, Pong and Space Invaders); they report that SARFA allows to obtain more focused and accurate interpretations for all of these applications than those in [31,37].

5. Conclusions and future outlook

Perturbation-based methods is a promising and rapidly developing XAI research paradigm. Perturbations allow to peer into the black box of DNN models by investigating input-output relationship and observing to which part of the input a model attributes particular importance. Perturbation-based methods have several advantages over another group of attribution methods, gradient-based methods: in particular, they are less noisy and more reliable. However, they also face the challenge of the combinatorial complexity explosion, as it is not possible to sample all possible perturbations of the input. Therefore, one of the main directions in the further development of perturbation-based methods are experimentations to find an optimal scope of perturbations. That holds true across all input data domains that we surveyed; however, there are also domain-specific challenges. First, while it is easier to determine the scope of perturbation for images, other data types, such as natural language or software code, present greater challenges due to the discrete nature of their elements. Second, whenever the data have a temporal dimension, which is the case of video and RL entities, there are additional challenges such as dealing with increase in dimensionality and the difficulties in generating saliency maps with a temporal dimension. As a consequence, research on perturbation-based methods for image classification has advanced much more than in other data domains, where it has just started to emerge. Yet another consequence is that research in other domains in many cases has borrowed approaches that have been already applied to images, a trend which is likely to continue in the future. To further develop cross-domain applications of methods, studies empirically comparing applications of perturbations to different data domains are needed.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is the result of activities within the 'Programmable Systems for Intelligence in Automobiles' (PRYSTINE) project, which has received funding from ECSEL Joint Undertaking under grant agreement No. 783190 and from specific national programs and/or funding authorities.

References

[1] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–18.
[2] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[3] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Proceedings of the Advances in Neural Information Processing Systems, 2018, pp. 9505–9515.
[4] S. Adel Bargal, A. Zunino, D. Kim, J. Zhang, V. Murino, S. Sclaroff, Excitation backprop for RNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1440–1449.
[5] M.A. Ahmad, C. Eckert, A. Teredesai, Interpretable machine learning in healthcare, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018, pp. 559–560.
[6] Amodei et al.(2016)Amodei, Olah, Steinhardt, Christiano, Schulman, and Mané D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in ai safety, arXiv preprint arXiv:1606.06565(2016).
[7] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Gradient-based attribution methods, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 169–191.
[8] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (Xai): concepts, taxonomies, opportunities and challenges toward responsible ai, Inf. Fusion 58 (2020) 82–115.
[9] BBC News, 2015, Google apologises for photos app's racist blunder, 2015. https://www.bbc.com/news/technology-33347866.
[10] M.G. Bellemare, Y. Naddaf, J. Veness, M. Bowling, The arcade learning environment: an evaluation platform for general agents, J. Artif. Intell. Res. 47 (2013) 253–279.
[11] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J.M. Moura, P. Eckersley, Explainable machine learning in deployment, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 648–657.
[12] M. Büchi, W. Weck, The Greybox Approach: When Blackbox Specification Hide too much, Citeseer, 1999.
[13] N.D. Bui, Y. Yu, L. Jiang, Autofocus: interpreting attention-based neural networks by code perturbation, in: Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), IEEE, 2019, pp. 38–41.
[14] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: a survey on methods and metrics, Electronics 8 (8) (2019) 832.
[15] S. Cave, K. Dihal, The whiteness of ai, Philos. Technol. (2020) 1–19.
[16] I. Chalkiadakis, A brief survey of visualization methods for deep learning models from the perspective of explainable AIU, 2018.
[17] J. Choo, S. Liu, Visual analytics for explainable deep learning, IEEE Comput. Graph. Appl. 38 (4) (2018) 84–92.
[18] P. Dabkowski, Y. Gal, Real time image saliency for black box classifiers, in: Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 6967–6976.
[19] D. Damen, H. Doughty, G.M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al., Scaling egocentric vision: the epic-kitchens dataset, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 720–736.
[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, IEEE, 2009, pp. 248–255.
[21] Doran et al.(2017)Doran, Schulz, and Besold D. Doran, S. Schulz, T.R. Besold, What does explainable ai really mean? A new conceptualization of perspectives, arXiv preprint arXiv:1710.00794(2017).
[22] F.K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: a survey, in: Proceedings of the 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.

[23] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VoC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[24] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.
[25] R. Fong, M. Patrick, A. Vedaldi, Understanding deep networks via extremal perturbations and smooth masks, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2950–2958.
[26] R. Fong, A. Vedaldi, Explanations for attributing deep neural network predictions, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 149–167.
[27] R.C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.
[28] Gilpin et al.(2018)Gilpin, Bau, Yuan, Bajwa, Specter, and Kagal L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: an approach to evaluating interpretability of machine learning, arXiv preprint arXiv:1806.00069(2018).
[29] Goodfellow et al.(2014)Goodfellow, Shlens, and Szegedy I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572(2014).
[30] B. Goodman, S. Flaxman, European union regulations on algorithmic decision–making and a 'right to explanation', AI Mag. 38 (3) (2017) 50–57.
[31] S. Greydanus, A. Koul, J. Dodge, A. Fern, Visualizing and understanding Atari agents, in: Proceedings of the International Conference on Machine Learning, 2018, pp. 1792–1801.
[32] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, J. Field Robot. 37 (3) (2020) 362–386.
[33] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. (CSUR) 51 (5) (2018) 1–42.
[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[35] F. Hohman, M. Kahng, R. Pienta, D.H. Chau, Visual analytics in deep learning: an interrogative survey for the next frontiers, IEEE Trans. Vis. Comput. Graph. 25 (8) (2018) 2674–2693.
[36] Holzinger et al.(2017)Holzinger, Plass, Holzinger, Crisan, Pintea, and Palade A. Holzinger, M. Plass, K. Holzinger, G.C. Crisan, C.-M. Pintea, V. Palade, A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop, arXiv preprint arXiv:1708.01104(2017).
[37] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, K. Sycara, Transparency and explanation in deep reinforcement learning neural networks, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 144–150.
[38] N. Kayser-Bril, Google Apologizes After Its Vision AI Produced Racist Results, AlgorithmWatch, 2020. https://algorithmwatch.org/en/story/google-vision-racism/
[39] Khakzar et al.(2019)Khakzar, Baselizadeh, Khanduja, Rupprecht, Kim, and Navab A. Khakzar, S. Baselizadeh, S. Khanduja, C. Rupprecht, S.T. Kim, N. Navab, Improving feature attribution through input-specific network pruning, arXiv (2019) arXiv–1911.
[40] M.E. Khan, F. Khan, et al., A comparative study of white box, black box and grey box testing techniques, Int. J. Adv. Comput. Sci. Appl 3 (6) (2012).
[41] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).
[42] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever Hans predictors and assessing what machines really learn, Nat. Commun. 10 (1) (2019) 1–8.
[43] Li et al.(2020)Li, Wang, Li, Huang, and Sato Z. Li, W. Wang, Z. Li, Y. Huang, Y. Sato, A comprehensive study on visual explanations for spatio-temporal networks, arXiv preprint arXiv:2005.00375(2020).
[44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[45] S. Liu, Z. Li, T. Li, V. Srikumar, V. Pascucci, P.-T. Bremer, Nlize: a perturbation–driven visual interrogation tool for analyzing and interpreting natural language inference models, IEEE Trans. Vis. Comput. Graph. 25 (1) (2018) 651–660.
[46] Mahdisoltani et al.(2018)Mahdisoltani, Berger, Gharbieh, Fleet, and Memisevic F. Mahdisoltani, G. Berger, W. Gharbieh, D. Fleet, R. Memisevic, Fine-grained video classification and captioning, arXiv preprint arXiv:1804.09235 5(6) (2018).
[47] Mänttäri et al.(2020)Mänttäri, Broomé, Folkesson, and Kjellström J. Mänttäri, S. Broomé, J. Folkesson, H. Kjellström, Interpreting video features: a comparison of 3d convolutional networks and convolutional LSTM networks, arXiv preprint arXiv:2002.00367(2020).
[48] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decis. Support Syst. 51 (4) (2011) 782–793.
[49] G.A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (11) (1995) 39–41.
[50] Mohseni et al.(2018)Mohseni, Zarei, and Ragan S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, arXiv (2018) arXiv–1811.
[51] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2018) 1–15.
[52] Petsiuk et al.(2018)Petsiuk, Das, and Saenko V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, arXiv preprint arXiv:1806.07421(2018).
[53] Puiutta and Veith(2020) E. Puiutta, E. Veith, Explainable reinforcement learning: a survey, arXiv preprint arXiv:2005.06247(2020).
[54] N. Puri, S. Verma, P. Gupta, D. Kayastha, S. Deshmukh, B. Krishnamurthy, S. Singh, Explain your move: Understanding agent actions using specific and relevant feature attribution, in: Proceedings of the International Conference on Learning Representations, 2019.
[55] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data Mining, 2016, pp. 1135–1144.
[56] M. Robnik-Šikonja, M. Bohanec, in: Human and Machine Learning, Springer, 2018, pp. 159–175.
[57] R. Roscher, B. Bohn, M.F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, IEEE Access 8 (2020) 42200–42216.
[58] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2016) 2660–2673.
[59] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., volume 3, IEEE, 2004, pp. 32–36.
[60] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad–cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[61] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., A general reinforcement learning algorithm that masters chess, Shogi, and go through self-play, Science 362 (6419) (2018) 1140–1144.
[62] Soomro et al.(2012)Soomro, Zamir, and Shah K. Soomro, A.R. Zamir, M. Shah, Ucf101: a dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402(2012).
[63] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, R. Poppe, Saliency tubes: visual explanations for spatio-temporal convolutions, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1830–1834.
[64] Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199(2013).
[65] The Guardian, 2018, Google's solution to accidental algorithmic racism: ban gorillas, 2018. https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people.
[66] Tjoa and Guan(2019) E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): towards medical XAI, arXiv preprint arXiv:1907.07374(2019).
[67] N. Weiner, Cybernetics: or the control and communication in the animal and the machine: or control and communication in the animal and the machine, 1961.
[68] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable ai: a brief survey on history, research areas, approaches and challenges, in: Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2019, pp. 563–574.
[69] Yang et al.(2020)Yang, Zhu, Ye, Fwu, You, and Zhu Q. Yang, X. Zhu, Y. Ye, J.-K. Fwu, G. You, Y. Zhu, Mfpp: morphological fragmental perturbation pyramid for black-box model explanations, arXiv preprint arXiv:2006.02659(2020).
[70] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[71] J. Zhang, S.A. Bargal, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top-down neural attention by excitation backprop, Int. J. Comput. Vis. 126 (10) (2018) 1084–1102.
[72] Q.-s. Zhang, S.-C. Zhu, Visual interpretability for deep learning: a survey, Front. Inf. Technol. Electron. Eng. 19 (1) (2018) 27–39.
[73] Zhang et al.(2019)Zhang, Cui, Finkler, Saon, Kayi, Buyuktosunoglu, Kingsbury, Kung, and Picheny W. Zhang, X. Cui, U. Finkler, G. Saon, A. Kayi, A. Buyuktosunoglu, B. Kingsbury, D. Kung, M. Picheny, A highly efficient distributed deep learning system for automatic speech recognition, arXiv preprint arXiv:1907.05701(2019).
[74] J. Zou, L. Schiebinger, Ai can be sexist and Racistit's time to make it fair, 2018.