nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00711-8

From attribution maps to human-understandable explanations through Concept Relevance Propagation

Received: 7 June 2022

Accepted: 31 July 2023

Published online: 20 September 2023

Check for updates

Reduan Achtibat[1,4], Maximilian Dreyer[1,4], Ilona Eisenbraun[1], Sebastian Bosse[1], Thomas Wiegand[1,2,3], Wojciech Samek[1,2,3] ✉ & Sebastian Lapuschkin[1] ✉

The field of explainable artificial intelligence (XAI) aims to bring transparency to today's powerful but opaque deep learning models. While local XAI methods explain individual predictions in the form of attribution maps, thereby identifying 'where' important features occur (but not providing information about 'what' they represent), global explanation techniques visualize what concepts a model has generally learned to encode. Both types of method thus provide only partial insights and leave the burden of interpreting the model's reasoning to the user. Here we introduce the Concept Relevance Propagation (CRP) approach, which combines the local and global perspectives and thus allows answering both the 'where' and 'what' questions for individual predictions. We demonstrate the capability of our method in various settings, showcasing that CRP leads to more human interpretable explanations and provides deep insights into the model's representation and reasoning through concept atlases, concept-composition analyses, and quantitative investigations of concept subspaces and their role in fine-grained decision-making.

Considerable advances have been made in the field of machine learning (ML), with deep neural networks (DNNs)[1] in particular achieving impressive performances on a multitude of domains[2–4]. However, the reasoning of these highly complex and nonlinear DNNs is generally not obvious[5,6], and, as such, their decisions may be (and often are) biased towards unintended or undesired features[7–10]. This in turn hampers the transferability of ML models to many application domains of interest, for example, due to the risks involved in high-stakes decision-making[5], or the requirements set in governmental regulatory frameworks[11] and guidelines brought forward[12].

To alleviate the 'black box' problem and gain insights into the model and its predictions, the field of explainable artificial intelligence (XAI) has been established. In fact, a multitude of XAI methods have been developed that are able to provide explanations of a model's decision while approaching the subject from different angles, for example, based on gradients[13,14], as modified backpropagation processes[15–18], by probing the model's reaction to changes in the input[19–21] or visualizing stimuli that specific neurons react strongly to[22,23]. The field can roughly be divided into local XAI and global XAI. Methods from local XAI commonly compute attribution maps in input space highlighting input regions or features, which carry some form of importance to the individual prediction process (that is, with respect to a specific sample). However, the visualization of important input regions is often of only limited informative value on its own, as it does not tell us what features in particular the model has recognized in those regions, as Fig. 1 illustrates. Furthermore, attribution maps can be understood as a superposition of many different model-internal decision subprocesses (for example, see ref. 24), working through various transformations of the same input features and culminating in the final prediction. Many intricacies are lost with local explanation

[1]Fraunhofer Heinrich Hertz Institute, Berlin, Germany. [2]Technische Universität Berlin, Berlin, Germany. [3]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany. [4]These authors contributed equally: Reduan Achtibat, Maximilian Dreyer. ✉e-mail: wojciech.samek@hhi.fraunhofer.de; sebastian.lapuschkin@hhi.fraunhofer.de

https://doi.org/10.1038/s42256-023-00711-8

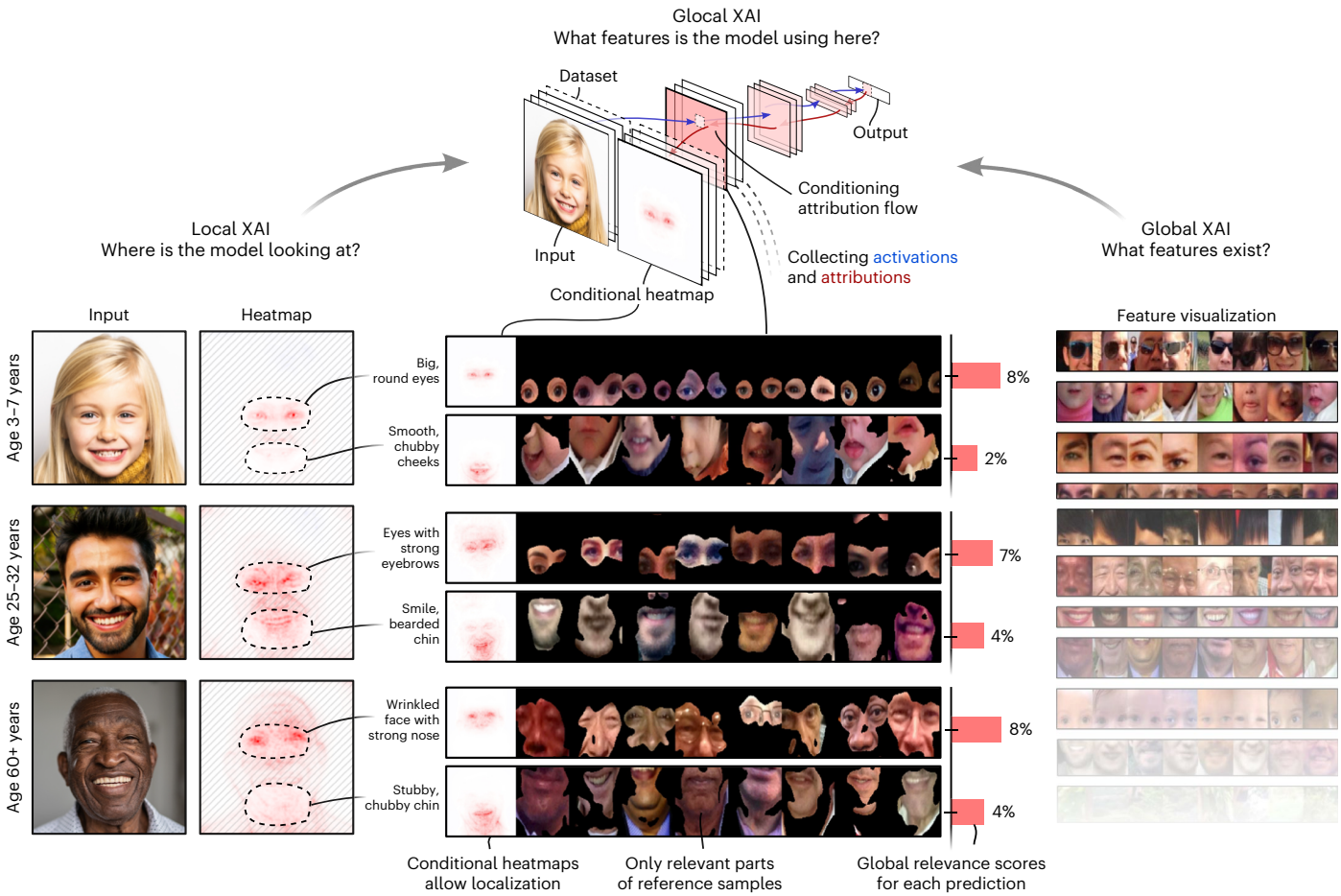

Fig. 1 | Glocal XAI can tell which features exist and how they are used for predictions by unifying local and global XAI. Left: local explanations visualize which input pixels are relevant for the prediction. Here, the model focuses on the eye region for all three predictions. However, what features in particular the model has recognized in those regions remains open for interpretation by the user. Right: by finding reference images that maximally represent particular (groups of) neurons, global XAI methods give insight into the concepts generally encoded by the model. However, global methods alone do not inform which concepts are recognized, used and combined by the model in per-sample inference. Centre: glocal XAI can identify the relevant neurons for a particular prediction (property of local XAI) and then visualize the concepts these neurons encode (property of global XAI). Furthermore, by using concept-conditional explanations as a filter mask, the concepts' defining parts can be highlighted in the reference images, which largely increase interpretability and clarity. Here, the topmost sample has been predicted into age group 3–7 due to the sample's large irides and round eyes, while the middle sample is predicted as 25–32, as more of the sclera is visible and eyebrows are more apparent. For the bottom sample, the model has predicted class 60+ based on its recognition of heavy wrinkles around the eyes and on the eyelids, and pronounced tear sacs next to a large knobby nose. Credit: iStock.com/MStudioImages, iStock.com/LSOphoto, iStock.com/FG Trade.

techniques producing only a singular attribution map in the input space per prediction outcome. The result might be unclear, imprecise or even ambiguous explanations.

Assuming, for example, an image classification setting and an attribution map computed for a specific prediction, it might be clear where (in terms of pixels) important information can be found, but not what this information is, that is, what characteristics of the raw input features the model has extracted and used during inference, or whether this information is a singular characteristic or an overlapping plurality thereof. This introduces many degrees of freedom to the interpretation of attribution maps generated by local XAI, rendering a precise understanding of the models' internal reasoning a difficult task.

Global XAI, however, attempts to address the very issue of understanding the 'what' question, that is, which features or concepts have been learned by a model or have an important role in a model's reasoning in general. Some approaches from this category synthesize example data to reveal the concept a particular neuron activates for[22,23,25–27], but do not inform which concept is in use in a specific classification or how it can be linked to a specific output. From these approaches, we can at most obtain a global understanding of all possible features the model can use, but how these features interact with each other given some specific data sample and how the model infers a decision remain hidden. Other branches of global XAI propose methods, for example, to test a model's sensitivity to a priori known, expected or pre-categorized stimuli[28–31]. These approaches require labelled data, thus limiting, and standing in contrast to, the exploratory potential of local XAI.

Some recent works have begun to bridge the gap between local and global XAI by, for example, drawing weight-based graphs that show how features interact in a global, yet class-specific scale, but without the capability to deliver explanations for individual data samples[32,33]. Others plead for creating inherently explainable models in the hope of replacing black-box models[5]. These methods, however, require either specialized architectures, data and labels, or training regimes (or a combination thereof)[34,35] and do not support the still widely used off-the-mill end-to-end-trained DNN models with their extended explanation capabilities. A detailed discussion of related work can be found in Supplementary Note 1.

https://doi.org/10.1038/s42256-023-00711-8

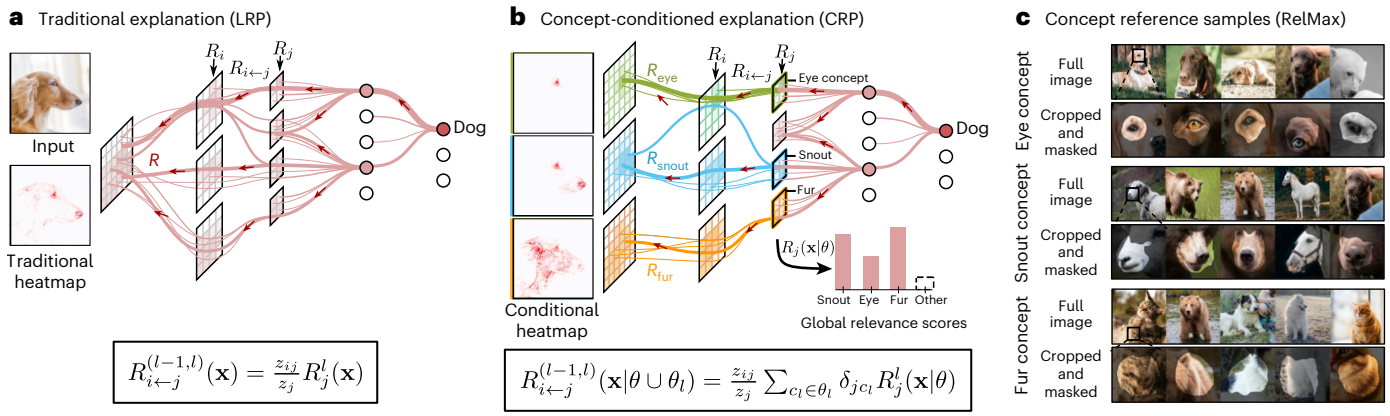

Fig. 2 | Brief overview over the methodological contributions of this work. **a**, Traditional backpropagation-based methods such as LRP propagate relevance scores backwards through the network culminating into a single attribution map. **b**, By conditioning on a concept encoded by a hidden-layer channel of the network, CRP allows to compute concept-conditional explanations. **c**, To provide a semantic meaning for latent model structures, we propose with RelMax to visualize input samples where the latent structure was strongly relevant for a prediction. We can further highlight the semantics by displaying only the relevant input parts according to concept-specific explanations, as introduced in **b**. Credit: iStock.com/Taku_S.

In this work, we connect lines of local and global XAI research by introducing Concept Relevance Propagation (CRP) and Relevance Maximization (RelMax), a set of next-generation XAI techniques that explain individual predictions in terms of localized and human-understandable concepts. In contrast to the related state of the art, CRP and RelMax answer both the 'where' and 'what' questions of ML model inference, thereby providing deep insights into the model's reasoning process. As post hoc XAI methods, CRP and RelMax can be applied to (almost) any ML model with no extra requirements on the data, model or training process. We demonstrate on multiple datasets, model architectures and application domains that CRP-based analyses allow one to (1) gain insights into the representation and composition of concepts in the model as well as quantitatively investigate their role in prediction, (2) identify and counteract Clever Hans filters[8] focusing on spurious correlations in the data, and (3) analyse whole concept subspaces and their contributions to fine-grained decision-making.

Analogously to Activation Maximization (ActMax)[36], our proposed RelMax approach searches for the most important (in terms of relevance, not activation) examples for latent encodings in, for example, the training dataset. Together, CRP and RelMax show their advantages in a conducted user study comparing our proposed techniques with various traditional attribution map-based approaches. Finally, where transparency on unique samples is promptly required, the computational efficiency and ease of application of CRP and RelMax quickly provide valuable insights into the model's representation and decision-making to the human user.

In summary, by lifting XAI to the concept level, CRP and RelMax open up new ways to analyse, debug and interact with ML models, which can be particularly beneficial for safety-critical applications and ML-supported investigations in the sciences.

Methods in brief

This section provides a brief overview over the methodological contributions of this work, that is, CRP and RelMax. A more detailed description of our methods can be found in Methods.

CRP in brief

Layer-wise Relevance Propagation (LRP)[15] is a popular method for explaining the predictions of a neural network by attributing relevance values to individual input dimensions (for example, pixels of images). In this process, relevance is propagated backwards through the network, starting from the output until the input layer (Fig. 2a), which also provides relevance values for each intermediate element of the model (for example, channel of an intermediate layer). As the literature suggests that latent structures of neural networks are encoding abstract human-understandable concepts with distinct semantics, especially in higher layers[23,31,37–40], the channel-wise relevance values can be interpreted as scores quantifying the importance of the corresponding concepts in the inference process.

CRP is an extension of LRP, which disentangles the relevance flows associated with concepts learned by the model via conditional backpropagation. Thus, it allows to compute concept-conditional relevance maps $R(\mathbf{x}|\theta)$, where $\mathbf{x}$ represents the data point the model has predicted for and θ describes a set of conditions (that is, specifying the explained output category (for example, 'dog') and concepts as learned and distinctly encoded by model components (for example, 'fur')), determining the flow of relevance via controlled masking operations in the backwards process (see Methods for technical details). These concept-conditional explanations show us, for example, in which part of the image the concepts (encoded in hidden-layer channels) 'fur' or 'eye' are present (the where question) and how much they contribute to the prediction. For the example in Fig. 2b, it turns out that the concept 'fur' is more relevant than the concept 'eye' for the prediction 'dog', which is not obvious when looking at explanations from LRP (Fig. 2a) or other local attribution methods (for example, refs. 17,41–43), where the contributions of all concepts are superimposed into a single attribution map.

It is noted that condition sets θ can be chosen by the human stakeholder (that is, depending on the task), or as we prefer to do in this paper, they can be configured automatically: per layer we configure θ algorithmically by ranking the network units in descending order of their relevance values for the current prediction, while choosing layer indices uniformly and arbitrarily from the higher, middle or bottom parts of the models throughout the paper for illustration out of simplicity.

RelMax in brief

Although CRP allows to compute concept-specific attribution maps by disentangling the backwards flow, our understanding of the semantics of latent model structures largely remains elusive with local attributions alone. In other words, the channel-wise relevance values and the concept-conditional relevance maps do not provide the full answer to which specific concept a particular channel is actually encoding (the 'what' question). A canonical approach for gaining insight into the meaning and function of latent model structures is ActMax[23,37–39] for generating or selecting samples as representations for concepts

encoded in hidden space. We find, however, that (maximizing) the activation of a latent encoding by a given data point does not always correspond to its utility to the model in an inference context (see, for example, ref. 44 or Supplementary Fig. 7), putting the faithfulness of activation-based example selection for latent concept representation into question.

We therefore introduce RelMax, as an alternative measure to ActMax, with the objective to maximize the relevance criterion for the selection of representative samples for latent model features (see Methods for technical details). For each of the observed concepts, Fig. 2c (right) shows 5 image segments from a holdout set, for which for instance channel 274 of layer features.28 of a pretrained VGG-16 neural network model[45] encoding the concept 'fur' becomes maximally relevant for the prediction of class 'dog'. As relevance is, contrary to activation, directly linked to a model's prediction output, the obtained example sets per latent feature are also highly outcome specific. That is, for a latent feature, we may obtain multiple sets of examples, each illustrating how the model is using a particular feature for the prediction of different outcomes, for example, classes.

Results

In this section, we first present approaches to study the role of learned concepts in individual predictions using our glocal CRP and RelMax-based approach. We then show how understanding of hidden features and their function allows interaction with the model and to test its robustness against feature ablation. Next we study concept subspaces to identify (dis)similarities and roles of concepts in fine-grained decision-making. Finally, we examine the benefits of CRP over traditional local XAI methods in a user study.

More detailed investigations can be found in Supplementary Notes 4, 5 and 8. In addition, Supplementary Note 9 provides an example on how CRP can be leveraged to identify systematically learned biases in male versus female face classification and Supplementary Note 10 demonstrates the applicability of CRP to time-series data.

Understanding concept composition leading to prediction

Attribution maps provide only partial insights into the decision-making process as they show only where the model is focusing on and not which concepts are actually being used. Figure 3a shows an attribution map computed for the prediction 'Northern Flicker'. In this case, the bird's head—in particular the black eye and red stripe—can be identified as the most relevant part of the image. However, it remains unclear from the explanation whether the colour or the shape (or both) of the eye and stripe were the decisive features for the model to arrive at its prediction, and how much these body parts contribute, for example, compared with the bird's feathers. Furthermore, as shown in Supplementary Fig. 1b, attribution maps almost always point to the head or the upper body of a bird, irrespective of the bird explained. Thus, the non-trivial task of interpreting what particular feature of the bird (for example, colour, texture, body part shape or relative position of the body parts) actually led to the decision is put onto the human user, which can result in false conclusions.

By conditioning the explanation on relevant hidden-layer channels via CRP, we can assist in concept understanding and overcome the interpretation gap. Figure 3b shows the result of the CRP analysis. The conditional heatmaps help to localize regions in input space for each relevant concept, and at the same time reveal what the model has picked up in those regions by providing reference samples (that is, explaining by example) via RelMax. Here, the concepts we identified as 'red spot' and 'black eyes' (based on our subjective understanding of the representative examples) can be assigned to the head of the Northern Flicker bird. These concepts have a crucial role in the classification of the bird, although, for example, the 'black eyes' concept naturally also occurs in images of cats and dogs. Furthermore, both 'dots' concepts affecting the prediction can be assigned to the bird's torso and the 'elongated dots and stripes' concept to the bird's wings. Note that CRP also allows to quantitatively determine the individual contribution of each concept to the final classification decision by summation of the conditional relevance scores (Methods). This additional information is very valuable as it indicates, for example, that the dotted texture is the most relevant feature for this particular prediction, or that colour is a very relevant cue (for example, the masked reference samples for channel 10 are all red and for channel 187 contain only black/brown eyes).

The concept atlas shown in Fig. 3c further eases comprehension of the relevant concepts. Technically, the atlas visualizes which concepts are most relevant (and here, second most relevant) in specific input-image regions (for details, see Supplementary Note 2.3.5). By choosing super-pixels as regions of interest, we can aggregate the channel-conditional relevances per super-pixel into regional relevance scores, as discussed in the extended Methods section in Supplementary Note 2.3.2. Here, the concept atlas indicates that the 'red spot' and 'black eye' concepts are most relevant at the bird's head, while the two 'dots' concepts mostly fill the upper body part. Interestingly, a stripe of red colour in the tail feathers of the bird is detected and used by the model, as indicated by the 'red spot' concept being second most relevant in this region. In Supplementary Note 5.1, an alternative way to construct concept atlases using single pixels instead of super-pixels is also discussed. Alternatively to investigating the most relevant channels overall as in Fig. 3b, a region of interest, for example a super-pixel, can be chosen and its most relevant concepts studied. A comparison of relevant concepts regarding two regions of unrelated visual features is shown in Supplementary Figs. 24–26.

With the selection of a specific neuron or concept, CRP allows investigation of how relevance flows from and through the chosen network unit to lower-level neurons and concepts, as is discussed in Methods. This gives information about which lower-level concepts carry importance for the concept of interest and how it is composed of more elementary conceptual building blocks, which may further improve the understanding of the investigated concept and model as a whole. In Fig. 3d, we visualize and analyse the backwards flow of the relevance scores. The graph-like visualization reveals how concepts in higher layers are composed of lower-layer concepts. Here, we show the top-two concepts influencing our concept of choice, the 'animal on branch' concept encoded in features.28 of a VGG-16 model trained on ImageNet. Edges in red colour indicate the flow of relevance with respect to the particular sample from class 'Bee Eater', shown on the far right between the visualized filters with corresponding examples and (multi)-conditional heatmaps. The width of each red edge describes the relative strength of contribution of lower-layer concepts to their upper-layer neighbours. This example shows that different paths in the network potentially activate the filter. Here, concepts that encode feathers, threads or fur together with horizontal structures are responsible for the activation of filter 102 in the observed layer in this particular case. In Supplementary Note 10, an example on time-series data is illustrated.

Understanding concept impact and reach

In this section, we demonstrate how CRP can be leveraged as a human-in-the-loop solution for dataset analysis. In the first step, we uncover a Clever Hans artefact[8], and suppress it by selectively eliminating the most relevant concepts to assess its decisiveness for the recognition of the correct class of a particular data sample. Then, we utilize class-conditional reference sampling (Methods) to perform an inverse search to identify multiple classes making use of the filter encoding the associated concept, both in a benign and a Clever Hans sense.

In Fig. 4a, we analyse a sample of the 'safe' class of ImageNet in a pretrained VGG-16 BN (a VGG-16 with BatchNorm layers) model. Initially, we obtain an input attribution map highlighting a centred horizontal band of the image, where a watermark is located. If we take a closer look at layer features.30 and perform a local analysis (Methods)

https://doi.org/10.1038/s42256-023-00711-8

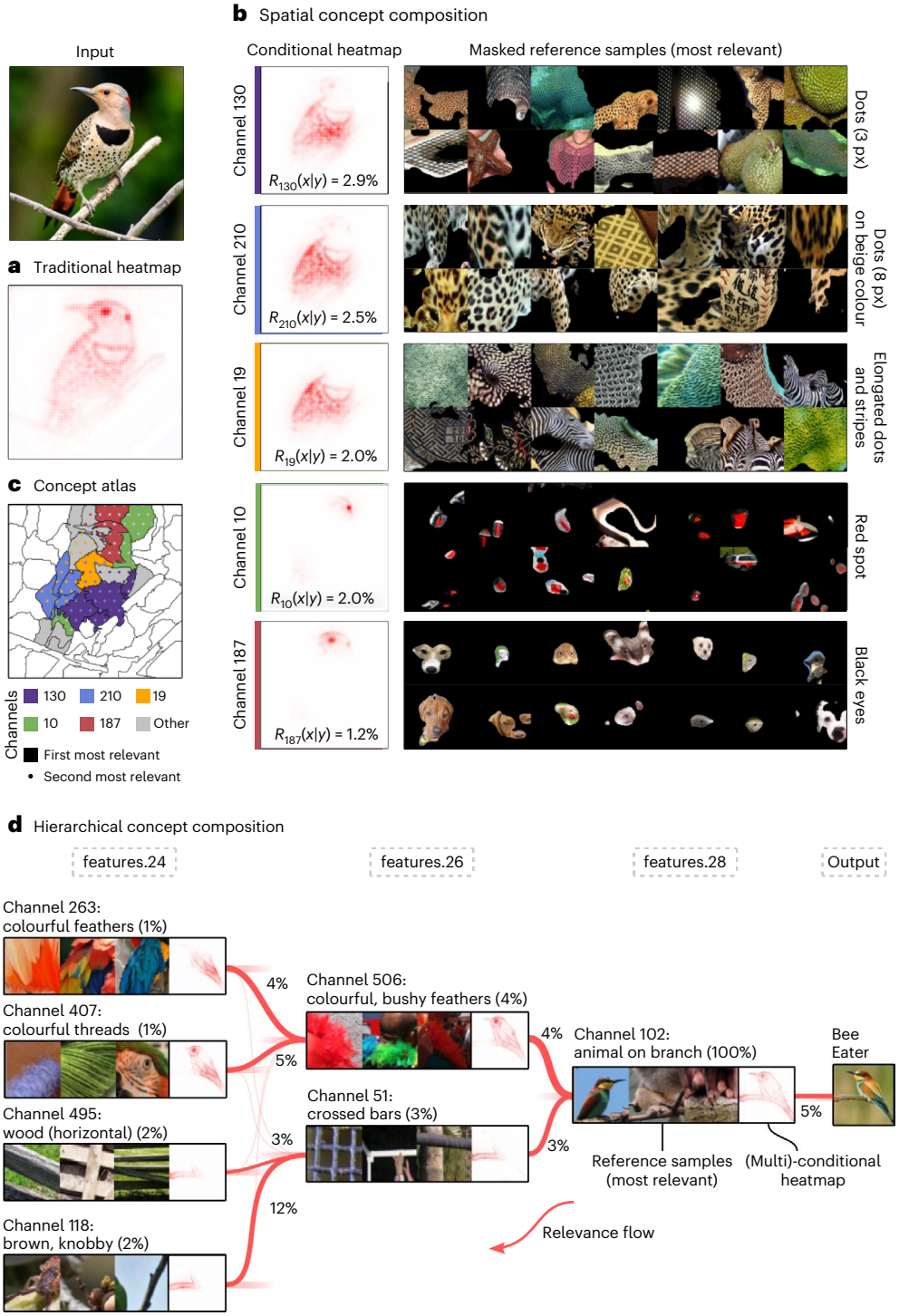

Fig. 3 | Understanding concepts and concept composition with CRP. **a**, Given an input image for inference, **a** constitutes a traditional attribution map indicating that various body parts of the bird are relevant for the prediction. **b**, Channel-conditional explanations computed with CRP help to localize and understand channel concepts by providing masked reference samples (explaining by example with RelMax). **c**, CRP relevances can further be used to construct a concept atlas, visualizing which concepts dominate in specific regions in the input image defined by super-pixels. Here, the most relevant channels in layer layer3.0.conv2 can be identified with concepts 'dots' (channels 210 and 130), 'red spot' (10), 'black eyes' (187) and 'stripes-like' (19). **d**, Concept-composition graphs decompose a concept of interest given a particular prediction into lower-layer concepts, thus improving concept understanding. Shown are relevant (sub)-concepts in features.24 and features.26 for concept 'animal on branch' in features.28 for the prediction of class 'Bee Eater'. The relevance flow is highlighted in red, with the relative percentage of relevance flow to the lower-level concepts. For each concept, the channel is given with the relative global relevance score (with respect to channel 102 in features.28) in parentheses. Following the relevance flow, concept 'animal on branch' is dependent on concepts describing the branch (for example, 'wood (horizontal)' and 'brown, knobby') and colourful plumage (for example, 'colourful feathers' and 'colourful threads'). Additional examples can be found in Supplementary Note 5. Credit: iStock.com/Thomas Marx, iStock.com/erniedecker.

https://doi.org/10.1038/s42256-023-00711-8

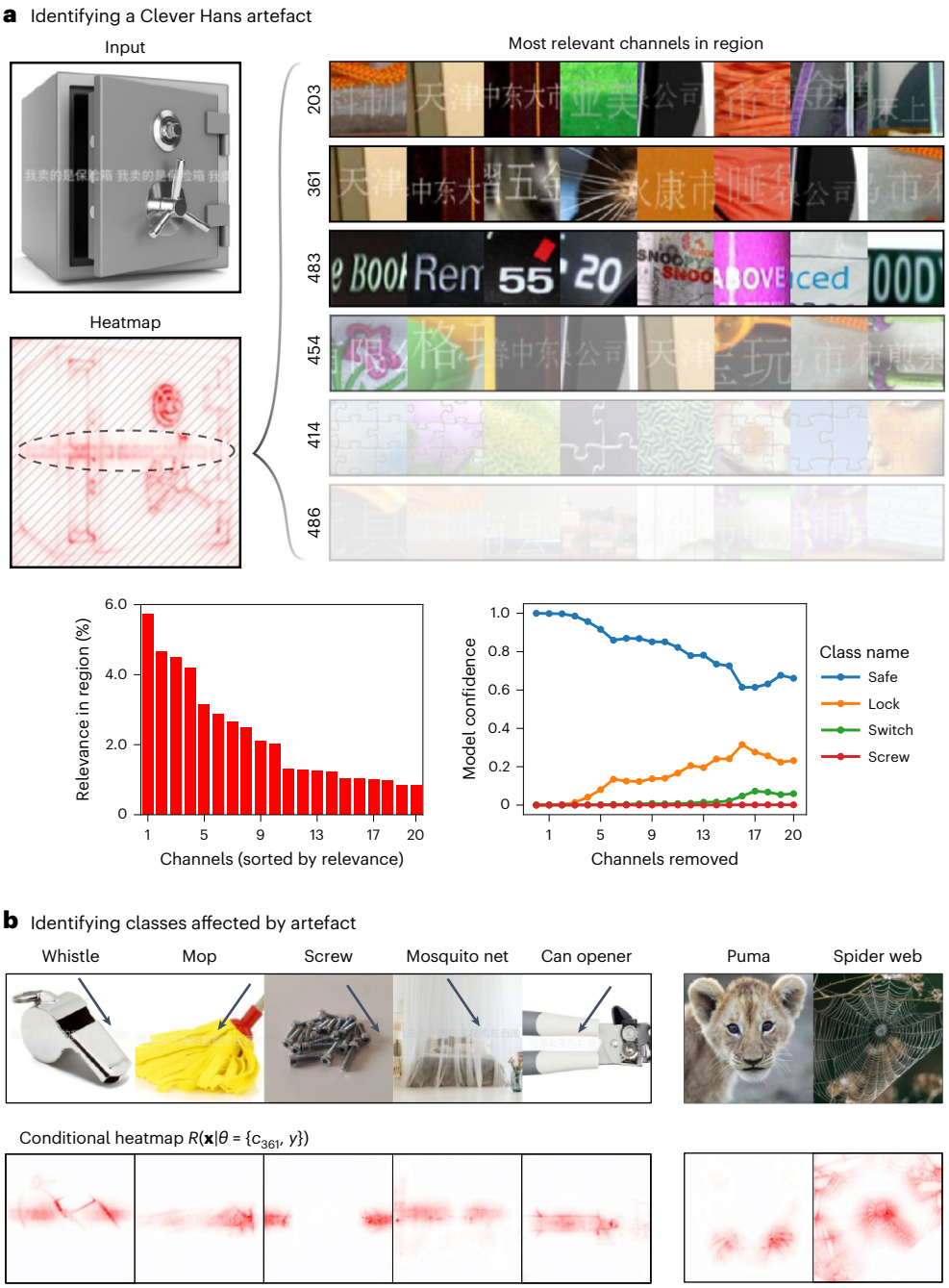

Fig. 4 | From concept-level explanations to model and data debugging. **a**, Local analysis on the attribution map reveals several channels (203, 361, 483, 454, 414, 486 and more) in layer features.30 of a VGG-16 model with BatchNorm pretrained on ImageNet that encode for a Clever Hans feature exploited by the model to detect the safe class. Top left: input image and heatmap. Top right: reference samples $x_{8\,\mathrm{sum}}^{*\,\mathrm{rel}}$ for the six most relevant channels in the selected region in descending order of their relevance contribution. Bottom: relevance contribution of the 20 most relevant filters inside the region (bottom left). These filters are successively set to zero and the change in prediction confidence of different classes is recorded (bottom right). **b**, The previously identified Clever Hans filter 361 has a role for samples of different classes (most relevant reference samples shown). Here, black arrows point to the location of a Clever Hans artefact, that is, a white, delicate font overlaid on images (best to be seen in a digital format). In the case of class 'puma' or 'spiderweb', the channel is used to recognize the puma's whiskers or the web itself, respectively. Below the reference samples, the CRP heatmaps conditioned on filter 361 and the respective true class *y* illustrate which part of their attribution map would result from filter 361. Credit: iStock.com/Andyworks, iStock.com/farakos, iStock.com/GP232, iStock.com/neamov, iStock.com/Stock Depot, iStock.com/t_kimura, shutterstock.com/Peter Zijlstra, shutterstock.com/Ground Picture.

on the watermark, we notice that the five most relevant filters are 203, 361, 483, 454, 414 and 486. Visualizing them using ActMax as illustrated in Supplementary Fig. 46, we conclude that they approximately encode for white strokes. Using our proposed RelMax approach, which uses CRP to identify the most relevant samples, we gain a deeper insight into the model's preferred usage of the filters and discover that the model utilizes them to detect white strokes in 'written characters'. A detailed comparison between ActMax and RelMax can be found in Supplementary Note 4.1. To test the robustness of the model against this Clever Hans artefact, we successively set the activation output map of the 20 most relevant filters activating on the watermark to zero. In Fig. 4a (bottom right), we record the change of classification

confidence of the four classes with the highest prediction confidence for this sample. From the graph, it can be inferred that the Clever Hans filters focusing on the watermark help the model in prediction, but they are not decisive for correct classification. Thus, the model relies on other potential non-Clever Hans features to detect the safe, verifying the correct functioning of the model in cases of samples without watermarks. Another example with strong dependency on Clever Hans artefacts is found in Supplementary Note 8.2.

In an inverse search, we can now explore for which samples and classes these filters also generate high relevance. This allows us to understand the behaviour of the filter in more detail and to find other possible contaminated classes. Figure 4b shows the seven most relevant classes for filter 361. Surprisingly, many classes including 'whistle', 'mop', 'screw', 'mosquito net', 'can opener' and 'safe' (among others) in the ImageNet Challenge 2014 data are contaminated with similar watermarks encoded via filter 361 of features.30, which is used for the correct prediction of samples from those classes. To verify our finding, we locate via CRP the source of the filters' relevance with respect to the true classes in input space and confirm that these filters indeed are used to recognize the characters. This implies that the model has learned a shared Clever Hans artefact spanning over multiple classes to achieve higher accuracy in classification. The high number of contamination of samples with the identified artefactual feature could be explained by the fact that watermarks are sometimes difficult to see with the naked eye (location marked with a black arrow) and thus slip any quality-ensuring data inspection. The impact of this image characteristic can, however, be clearly marked using the CRP heatmap. Although the filter is mainly used to detect characters, there are also valid use cases for the model, such as for the puma's whiskers or the spider's web. This suggests that the complete removal of Clever Hans concepts through pruning may harm the model in its ability to predict other classes that make valid use of the filter, and that a class-specific correction[10] might be more appropriate.

Understanding concept subspaces, (dis)similarities and roles

So far in our experiments, we have treated single filters as functions assumed to (fully) encode the learned concept. Consequently, we have visualized examples and quantified effects based on per-filter granularity. While previous work has suggested that individual neurons or filters often encode for a single human comprehensible concept, it can generally be assumed that concepts are encoded by sets of filters (Supplementary Note 1). The learned weights of potentially multiple filters might correlate and thus redundantly encode the same concept, or the directions described by several filters situated in the same layer might span a concept-defining subspace. In this section, we aim to investigate the encodings of filters of a given neural network layer for similarities in terms of activation and use within the model.

Figure 5a shows an analysis result focusing on a cluster around filter 446 from features.40 of a VGG-16 network with BatchNorm layers trained on ImageNet. The reference samples show various types of typewriter and rectangular laptop keyboard buttons and roofing shingles photographed in oblique perspective, as well as round buttons of typewriters, remote controls for televisions, telephone keys and round turnable dials of various devices and machinery. Thus, the filters around filter 446 seem to cover different aspects of a shared 'button' or 'small tile' concept. The filters located in this cluster have been identified as similar due to their similar activations over sets of analysed reference samples (Methods). Assuming redundancy based on the filter channels' apparently similar activation behaviour, a human could merge them to one encompassing concept, thereby simplifying interpretation by reducing the number of filters in the model. We therefore further investigate the filters 7, 94, 446 and 357 (all showing buttons or keys) to find out (1) whether they encode a concept collaboratively, (2) whether they are partly redundant or (3) whether the cluster serves some discriminative purpose. Figure 5b visualizes the reference samples of these four filters for the most relevant classes 'laptop computer' and 'remote control'. We compute filter activations during a forward pass through the model using instances of both classes as input, as well as filter-conditioned CRP maps for the samples' respective ground-truth class label. Regardless of whether an instance from class 'laptop' or 'remote control' is chosen as input, the activation maps across the observed channels are in part similar per image, for example, they all activate on the centre diagonal part for the left input image. The per-channel CRP attribution map, however, reveals that while all filters react to similar stimuli in terms of activations, the model seems to use the subtle differences among the observed concepts to distinguish between the classes 'laptop' and 'remote control'. In both cases, buttons are striking and defining features, and all observed filters activate for button features. However, when computing the conditional heatmaps with CRP for class 'remote control', the activating filters representing round buttons (filters 7 and 94) dominantly receive positive attribution scores, while filter 357 clearly representing typical keyboard button layouts receives negative relevance scores and filter 446 does not receive any relevance despite being reactive to the given input. For samples of class 'laptop', the computation of relevance scores with respect to their true class yields almost opposite attributions, indicating that filters encoding round buttons and dials (filters 94 and 7) provide evidence against class 'laptop', while the activation of channel 357 clearly speaks for the analysed class as visible in the conditional heatmaps. In both relevance analyses, however, filter 446 receives weak negative to no attributions, presumably as it represents a particular expression of both round and angular buttons that fits (or contradicts) neither of the compared classes particularly well. In fact, filter 446 is highly relevant for class 'typewriter keyboard' instead.

In conclusion, we report that although several filters may show signs of correlation in terms of output activation, they are not necessarily encoding redundant information or are serving the same purpose. Conversely, using our proposed CRP in combination with the RelMax-based process for selecting reference examples representing seemingly correlating filters, we are able to discover and understand the subtleties a neural network has learned to encode in its latent representations. See Supplementary Note 8.4 for additional results in extension to this section.

Human evaluation study

This section presents the results of a human evaluation study, which we performed to assess the practical utility of the CRP and RelMax-based explanations to (non-expert) end users for understanding ML model behaviour. Human participants were asked to decide—based on explanations—whether the model's prediction has been influenced by the presence of a particular and known data artefact or not. We trained two image classifiers, of which one has learned to utilize a data artefact—a thick black border around the image (Fig. 6a). For both models, we then generate explanations (Fig. 6b,c) on images containing the artefact, using the proposed CRP maps with RelMax examples as well as four popular XAI methods, namely, Integrated Gradients (IG)[14], SHapley Additive exPlanations (SHAP)[43], Gradient-weighted Class Activation Mapping (Grad-CAM)[41] and LRP[15]. In the primary task, the participants are asked to assess whether the black border impacts the model prediction according to the explanation (binary answer, yes or no). Furthermore, we ask secondary questions on how confident they are in their answer and about the perceived clarity of the presented explanations. For more details on the study set-up, we refer the reader to Methods.

The results of the study consistently show that participants were reliably able to detect whether the prediction was impacted by the border artefact when exposed to CRP and RelMax explanations (Fig. 6d). CRP shows the highest true positive (TPR) and true negative (TNR) rates of (89.1 ± 2.4)% and (72.6 ± 3.4)%, respectively, and thus results in an accuracy (with respect to the primary task) that is significantly higher

https://doi.org/10.1038/s42256-023-00711-8

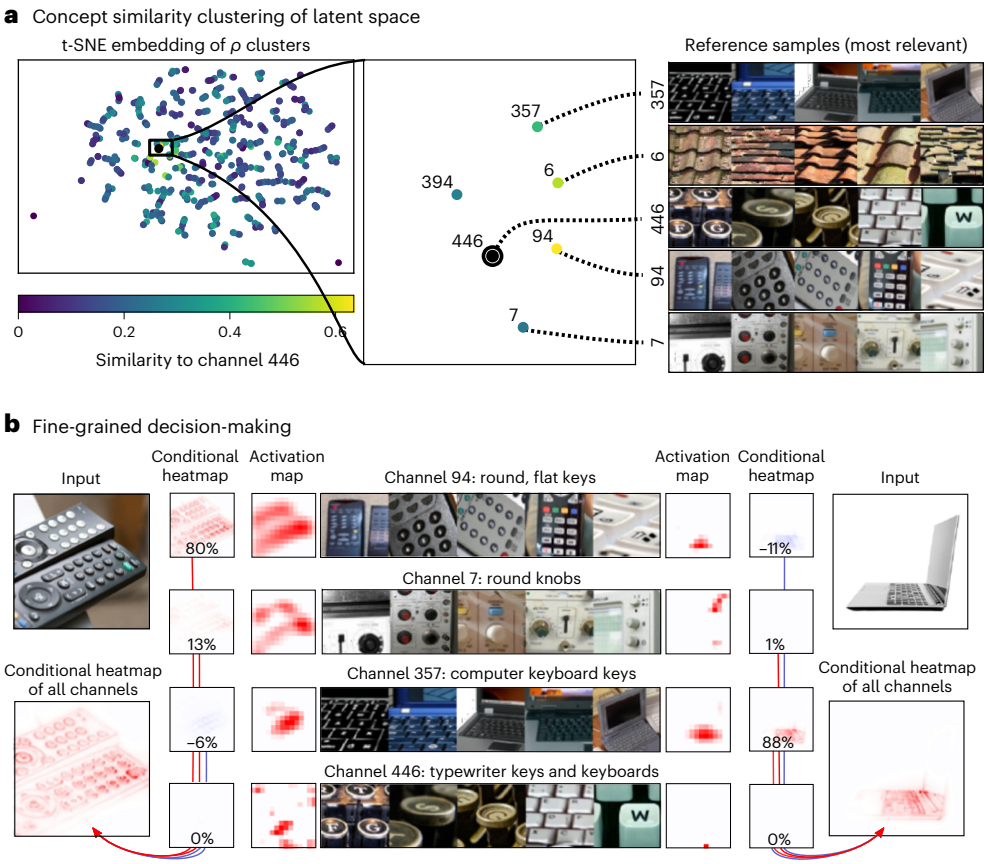

Fig. 5 | Similarity of concepts and analysis of fine-grained decision-making. **a**, Left: channels from layer features.40 of a VGG-16 with BatchNorm, clustered and embedded according to ρ similarity with t-SNE (Methods). Markers are coloured according to their ρ similarity to filter 446. Centre and right: one particular cluster around channel 446 is shown in more detail with five similarly activating channels and their reference images $\mathcal{X}_{8\ \mathrm{sum}}^{\star\ \mathrm{rel}}$ obtained via RelMax. As per the reference images, the overall concept of the cluster seems to be related to keyboard keys, round buttons and rectangular roofing shingles. **b**, Relevance-based investigation of the previously identified similarly activating channels. Centre: reference examples for the identified filters with a similar underlying theme. Left: exemplary input from class 'remote control' with per-channel activation maps and respective ground-truth CRP relevance maps, as well as their aggregation $\theta = \{L: \{y\}, \text{features.40}: \{c_{94}, c_7, c_{357}, c_{446}\}\}$ (bottom left). Right: exemplary input from class 'laptop computer' with per-channel activation maps and respective true class CRP relevance maps, as well as their aggregation. Conditional relevance attributions $R(\mathbf{x}|\theta)$ are normalized with respect to the common maximum amplitude. Similarly activating channels do not necessarily encode redundant information, but might be used by the model for making fine-grained distinctions, which can be observed from the attributed relevance scores. Credit: iStock.com/ezza116, iStock.com/sqback.

than that of all other methods (two-sample t-test P values of less than 8×10^{-4}), as shown in Supplementary Table 3. The study participants obtained the second best results when exposed to Grad-CAM explanations with a TPR and TNR of (74.9 ± 3.3)% and (52.6 ± 3.8)%, respectively. Participants exposed to explanations from IG performed worst with a TPR and TNR of (53.7 ± 3.8)% and (49.7 ± 3.8)%, respectively. It is noted that random guessing would correspond to values of 50%, indicating that the insight obtainable from IG explanations is limited in this study. We note that in general it seems to be easier to detect a model's use of the introduced border artefact through all evaluated XAI approaches than it is to correctly reject an impact of the artefact, as for all methods we observe TPR > TNR.

When inspecting the proclaimed confidence with which the participants made their prediction, a direct link does not seem to exist with the measured prediction performance when assessing the artefact's impact. Interestingly, participants exposed to the IG explanations report the highest confidence (approximately 77%)—while at the same time performing worst in the primary task—followed by CRP (approximately 76% self-rated confidence). These results support the findings of ref. 28 that traditional (that is, single per sample and class, as shown in Fig. 6b) saliency or attribution maps alone might be misleading and insufficient for understanding the reasoning of an ML predictor; for example, see the IG heatmap. Regarding clarity of the explanations as perceived by the participants, the fine-grained attribution maps of IG and LRP receive the highest scores. CRP and RelMax interestingly result in the lowest reported clarity, which might be linked to the more complex nature of the method, potentially leaving some of the participants overwhelmed with the increased amount of information to process, and time required to do so. This result is consistent with the observation of ref. 46 that addressees prefer simple and concise explanations. Despite that, our results demonstrate that our proposed approach is the most effective option for the participants to solve the primary task of the study.

Discussion

In this work, we have introduced CRP, a post hoc explanation method that not only indicates which part of the input is relevant for an individual prediction but also communicates the meaning of involved latent representations by providing human-understandable examples. As CRP combines the benefits of the local and global XAI perspectives, it computes more detailed and contextualized explanations, considerably extending the state of the art. Among its advantages are the high computational efficiency (within the order of a backwards pass to compute near-instantaneous local explanations for the most relevant concepts, and

https://doi.org/10.1038/s42256-023-00711-8

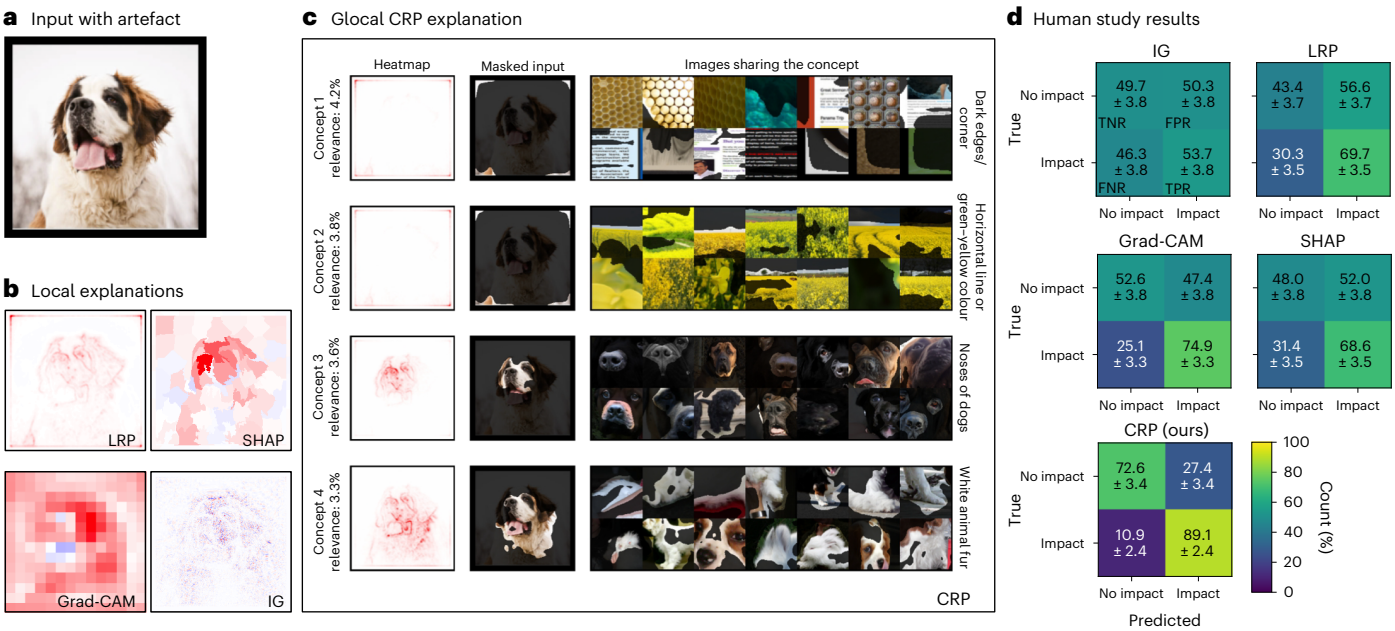

Fig. 6 | User study evaluating the informativeness of different explanation methods with respect to the model's reliance on a border artefact. a, An example input image with an added black border artefact. **b**, Attribution maps computed via the methods LRP, SHAP, Grad-CAM and IG. **c**, An explanation derived with CRP and RelMax. In both **b** and **c**, the model has been trained to be affected by the artefact shown in **a** during inference. **d**, Human evaluation results of participant predictions whether the border artefact has an impact on the prediction outcome, based on the given explanation. Confusion matrices illustrating true positive (TPR), false positive (FPR), true negative (TNR) and false negative (FNR) rates per XAI method. Credit: iStock.com/mauro_grigollo.

a complete concept atlas visualization within the order of seconds) and the out-of-the-box applicability to (almost) any model without imposing constraints on the training process, the data and label availability, or the model architecture. Furthermore, CRP introduces the idea of conditional backpropagation tied to a single concept or a combination of concepts as encoded by the model, within or across layers. Via this ansatz, the contribution of all neurons' concepts in a layer can be faithfully attributed, localized in the input space, and finally their interaction can be studied. As shown in this work, such an analysis allows one to disentangle and separately explain the multitude of in-parallel partial forward processes, which transform and combine features and concepts before culminating into a prediction. Finally, with RelMax, we move beyond the decade-old practice of communicating latent features of neural networks based on examples obtained via maximized activation. In particular, we show that the examples that stimulate hidden features maximally are not necessarily useful for the model in an inference context, or representative for the data the model is familiar and confident with. By providing examples based on relevance, however, the user is presented with data with characteristics that actually have an important role in the prediction process. As the user can select examples with respect to any (that is, not necessarily the ground truth) output class, our approach constitutes a new tool to systematically investigate latent concepts in neural networks.

Our experiments have qualitatively and quantitatively demonstrated the additional value of the CRP approach for common datasets and end-to-end-trained models. Specifically, we showed that reference samples selected with relevance-based criteria, concept heatmaps and atlases, as well as concept-composition graphs, open up the ability to understand model reasoning on a more abstract and conceptual level. These insights then allowed us to identify Clever Hans concepts, to investigate their impact and finally to correct for these misbehaviours. Furthermore, using our relevance-based reference sample sets, we were able to identify concept themes spanned by sets of filters in latent space. Although channels of a cluster have a similar function, they seem to be used by the model for fine-grained decisions regarding details in the data, such as the particular type of buttons to partially decide whether an image shows a laptop keyboard, a mechanical typewriter or a TV remote control. In addition, we have demonstrated the usefulness of CRP in the non-image data domain, where traditional attribution maps are often difficult to interpret and comprehend by the user. Our experiments on time-series data have shown that as long as a visualization of the data can be found, the meaning of latent concepts can be communicated via reference examples. Finally, we did conduct a user study that validates a substantial increase in utility of our glocal CRP and RelMax-based approach above traditional post hoc local XAI methods for understanding a model's inference behaviour by human assessors. For completeness, we make the reader aware of two factors possibly affecting the outcome of our study, namely, the potentially varying degree of technical and in-domain training of the study participants, and the given prior knowledge about the nature of the data artefact potentially affecting the model. Both factors should therefore be addressed and evaluated individually in future work, for example, to assess the potential of (g)local XAI approaches for assessing yet unexplored model behaviour based on feedback for single-instance predictions, across different levels of expert knowledge.

Overall, we believe that the tools we have proposed in this work, and the resulting increase in semantics and detail to be found in sample-specific neural network explanations, will advance the applicability of post hoc XAI to novel or previously difficult to handle models, problems and data domains.

Methods

This section presents the techniques used and introduced in this paper. For a more elaborate introduction and discussion, please refer to Supplementary Notes 2 and 3. For an estimation of run-time requirements, the computational steps involved and guidelines on the interpretation of the output obtained by our techniques, please refer to Supplementary Note 6.

Concept Relevance Propagation

In the following, we introduce CRP, a backpropagation-based attribution method extending the framework of LRP[15]. As such, CRP inherits the basic assumptions and properties of LRP.

LRP revisited. Assuming a predictor with L layers

$$f(\mathbf{x}) = f_L \circ \cdots \circ f_1(\mathbf{x}), \tag{1}$$

LRP follows the flow of activations computed during the forward pass through the model in opposite direction, from the final layer f_L back to the input mapping f_1. Given a particular mapping $f_\cdot(\cdot)$, we consider its pre-activations z_{ij} mapping inputs i to outputs j and their aggregations z_j at j. Commonly in neural network architectures such a computation is given with

$$z_{ij} = a_i w_{ij} \tag{2}$$

$$z_j = \sum_i z_{ij} \tag{3}$$

$$a_j = \sigma(z_j), \tag{4}$$

where a_i are the layer's inputs and w_{ij} its weight parameters. Finally, σ constitutes a (component-wise) nonlinearity producing input activation for the succeeding layer(s). The LRP method distributes relevance quantities R_j corresponding to a_j and received from upper layers towards lower layers proportionally to the relative contributions of z_{ij} to z_j, that is

$$R_{i\leftarrow j} = \frac{z_{ij}}{z_j} R_j. \tag{5}$$

Lower neuron relevance is obtained by losslessly aggregating all incoming relevance messages $R_{i\leftarrow j}$ as

$$R_i = \sum_j R_{i\leftarrow j}. \tag{6}$$

This process ensures the property of relevance conservation between a neuron j and its inputs i, and thus adjacent layers. LRP is mathematically founded in deep Taylor decomposition[47].

Disentangling explanations with CRP. CRP extends the formalism of LRP by introducing conditional relevance propagation determined by a set of conditions θ. Each condition $c \in \theta$ can be understood as an identifier for neural network elements, such as neurons j located in some layer, representing latent encodings of concepts of interest. One such condition could, for example, represent a particular network output to initiate the backpropagation process from. Within the CRP framework, the basic relevance decomposition formula of LRP given in equation (5) then becomes

$$R^{(l-1,l)}_{i\leftarrow j}(\mathbf{x}|\theta \cup \theta_l) = \frac{z_{ij}}{z_j} \sum_{c_l \in \theta_l} \delta_{jc_l} R^l_j(\mathbf{x}|\theta), \tag{7}$$

following the potential for a 'filtering' functionality briefly discussed in ref. 48. Here, $R^l_j(\mathbf{x}|\theta)$ is the relevance assigned to layer output j given from the CRP process performed in upper layers under conditions θ, to be distributed to lower layers. The sum-loop over $c_l \in \theta_l$ then 'selects' via the Kronecker-delta δ_{jc_l} neurons j of which the relevance is to be propagated further, given j corresponds to concepts as specified in set θ_l specific to layer l. The result is the concept-conditional relevance message $R^{(l-1,l)}_{i\leftarrow j}(\mathbf{x}|\theta \cup \theta_l)$ carrying the relevance quantities with respect to the prediction outcome on $\mathbf{x}$ conditioned to θ and θ_l. Note that the sum is not particularly necessary in equation (7), but serves as a means to compare all possible c_l for identity to the current j. In practice, CRP can be implemented efficiently as a single backpropagation step by binary masking of relevance tensors, and is compatible to the recommended rule composites for relevance backpropagation[49,50]. We provide an efficient implementation of CRP based on Zennit[51] at https://github.com/rachtibat/zennit-crp.

The effect of CRP over LRP and other attribution methods is an increase in detail of the obtained explanations. Given a typical image classification convolutional neural network (CNN), one may assume the computation of three-dimensional latent tensors, where the first two axes span the application coordinates of n spatially invariant convolutional filters, which generate output activations stored in the n channels of the third axis. For simplicity, one can further assume that each filter channel is associated with exactly one latent concept. Neurons j can thus be grouped into spatial and channel axes to restrict the application of CRP conditions θ_l to the channel axis only, that is

$$R^{(l-1,l)}_{i\leftarrow(p,q,j)}(\mathbf{x}|\theta \cup \theta_l) = \frac{z_{i(p,q,j)}}{z_{(p,q,j)}} \sum_{c_l \in \theta_l} \delta_{jc_l} R^l_{(p,q,j)}(\mathbf{x}|\theta). \tag{8}$$

Here, the tuple (p, q, j) uniquely addresses an output voxel of the activation tensor $z_{(p,q,j)}$ computed during the forward pass with p and q indicating the spatial tensor positions and j the channel. Figure 2a contrasts the attribution-based explanation with respect to class 'dog' only (which also is possible with LRP and other attribution methods) as $\theta_d = \{\text{dog}^L\}$, to the attributions for, for example, 'dog ∧ fur' as $\theta_{df} = \{\text{dog}^L, \text{fur}^l\}$ (possible with CRP only) by conditionally masking channels responsible for fur pattern representations. Alternatively, conditions can be notated in the form of $\theta_{df} = \{L: \{\text{dog}\}, l: \{\text{fur}\}\}$, to provide a more explicit notation specifying the affiliation of concepts to distinct layers. Here we use the terms 'fur' and 'dog' describing latent or labelled concepts, respectively, as proxy representations for network element identifiers c. We further assume that in any layer l' without explicit designation of conditions all $\delta_\cdot$ operators always evaluate to 1 to not restrict the flow of attributions through these layers.

Due to the conservation property of CRP inherited from LRP, the global relevance of individual concepts to per-sample inference can be measured by summation over input units i as

$$R^l(\mathbf{x}|\theta) = \sum_i R^l_i(\mathbf{x}|\theta), \tag{9}$$

in any layer l where θ has taken full effect. This can easily be extended to a localized analysis of conceptual importance, by restricting the relevance aggregations to regions of interest $\mathcal{I}$

$$R^l_{\mathcal{I}}(\mathbf{x}|\theta) = \sum_{i \in \mathcal{I}} R^l_i(\mathbf{x}|\theta), \tag{10}$$

as also illustrated in Supplementary Fig. 3. In addition, as shown in Supplementary Fig. 4, an aggregation of the relevance messages may be utilized to identify dependencies of a concept c encoded by channels j, to concepts encoded by channels i in a lower layer, in context of the prediction of a sample $\mathbf{x}$ and CRP conditions θ. With an expansion of the indexing of downstream target voxels with respect to equation (9) as

$$R^{(l-1,l)}_{(u,v,i)\leftarrow(p,q,j)}(\mathbf{x}|\theta) = \frac{z_{(u,v,i)(p,q,j)}}{z_{(p,q,j)}} R^l_{(p,q,j)}(\mathbf{x}|\theta), \tag{11}$$

the tuple (u, v, i) addresses the spatial axes with u and v, and the channel axis i at layer $l-1$. An aggregation over spatial axes with

$$R^{(l-1,l)}_{i\leftarrow j}(\mathbf{x}|\theta) = \sum_{u,v} \sum_{p,q} R^{(l-1,l)}_{(u,v,i)\leftarrow(p,q,j)}(\mathbf{x}|\theta) \tag{12}$$

communicates the dependency between channel j to lower-layer channel i, and thus related concepts, in terms of relevance in the prediction context of sample $\mathbf{x}$. Following the LRP methodology, an adaptation of the CRP approach beyond CNN, for example, to recurrent[52] or graph[53]

neural networks, is possible. Further details on our proposed CRP method are given in Supplementary Note 2.

Selecting reference examples

In the following, we discuss the widely used ActMax approach to procuring representations for latent neurons, and present our novel CRP-based RelMax technique to improve concept identification and understanding. An in-depth introduction to all details of our proposed technique is given in Supplementary Note 3, with various modes of application and analyses being discussed in Supplementary Note 4.

Activation Maximization. A large part of feature visualization techniques rely on ActMax, where in its simplest form, input images are sought that give rise to the highest activation value of a specific network unit. Recent work[35,54] proposed to select reference samples from existing data for feature visualization and analysis. In the literature, the selection of reference samples for a chosen concept c manifested in groups of neurons is often based on the strength of activation induced by a sample. For data-based reference sample selection, the possible input space $\mathcal{X}$ is restricted to elements of a particular finite dataset $\mathcal{X}_\mathrm{d} \subset \mathcal{X}$. The authors of ref. 35 assumed convolutional layer filters to be spatially invariant. Therefore, entire filter channels instead of single neurons are investigated for convolutional layers. One particular choice of maximization target $\mathcal{T}(\mathbf{x})$ is to identify samples $\mathbf{x}^\star \in \mathcal{X}_\mathrm{d}$, which maximize the sum over all channel activations, that is

$$\mathcal{T}^{\mathrm{act}}_{\mathrm{sum}}(\mathbf{x}) = \sum_i z_i(\mathbf{x}). \tag{13}$$

resulting in samples $\mathbf{x}^{\star\,\mathrm{act}}_{\mathrm{sum}}$, which are likely to show a channel's concept in multiple (spatially distributed) input features, as maximizing the entire channel also maximizes $\mathcal{T}^{\mathrm{act}}_{\mathrm{sum}}$. However, while targeting all channel neurons, reference samples including both concept-supporting and contradicting features might result in a low function output of $\mathcal{T}^{\mathrm{act}}_{\mathrm{sum}}$, as negative activations are taken into account by the sum. Alternatively, a nonlinearity can be applied on $z_i(\mathbf{x})$, for example, a rectified linear unit (ReLU), to only consider positive activations. A different choice is to define maximally activating samples by observing the maximum channel activation

$$\mathcal{T}^{\mathrm{act}}_{\mathrm{max}}(\mathbf{x}) = \max_i z_i(\mathbf{x}), \tag{14}$$

leading to samples $\mathbf{x}^{\star\,\mathrm{act}}_{\mathrm{max}}$ with a more localized and strongly activating set of input features characterizing a channel's concept. These samples $\mathbf{x}^{\star\,\mathrm{act}}_{\mathrm{max}}$ might be more difficult to interpret, as only a small region of a sample might express the concept.

To collect multiple reference images describing a concept, the dataset $\mathcal{X}_\mathrm{d}$ consisting of n samples is first sorted in descending order according to the maximization target $\mathcal{T}(\mathbf{x})$, that is

$$\mathcal{X}^\star = \{\mathbf{x}^\star_1, \ldots, \mathbf{x}^\star_n\} = \arg\operatorname*{sort}^{\mathrm{desc}}_{\mathbf{x}\in\mathcal{X}_d} \mathcal{T}(\mathbf{x}). \tag{15}$$

Subsequently, we define the set

$$\mathcal{X}^\star_k = \{\mathbf{x}^\star_1, \ldots, \mathbf{x}^\star_k\} \subseteq \mathcal{X}^\star \tag{16}$$

containing the $k \leq n$ samples ranked first according to the maximization target to represent the concept of the filter(s) under investigation. We denote the set of samples obtained from $\mathcal{T}^{\mathrm{act}}_{\mathrm{sum}}$ as $\mathcal{X}_{k\,\mathrm{sum}}^{\star\,\mathrm{act}}$ and the set obtained from $\mathcal{T}^{\mathrm{act}}_{\mathrm{max}}$ as $\mathcal{X}_{k\,\mathrm{max}}^{\star\,\mathrm{act}}$.

Relevance Maximization. We introduce the method of RelMax as a complement to ActMax. Regarding RelMax, we do not search for images that produce a maximal activation response. Instead, we aim to find samples, which contain the relevant concepts for a prediction. To select the most relevant samples, we define maximization targets $\mathcal{T}^{\mathrm{rel}}_{*}(\mathbf{x})$ by using the relevance $R_i(\mathbf{x}|\theta)$ of neuron i for a given prediction, instead of its activation value z_i. Specifically, the maximization targets are given as

$$\mathcal{T}^{\mathrm{rel}}_{\mathrm{sum}}(\mathbf{x}) = \sum_i R_i(\mathbf{x}|\theta) \quad \text{and} \quad \mathcal{T}^{\mathrm{rel}}_{\mathrm{max}}(\mathbf{x}) = \max_i R_i(\mathbf{x}|\theta). \tag{17}$$

By utilizing relevance scores $R_i(\mathbf{x}|\theta)$ instead of relying on activations only, the maximization target $\mathcal{T}^{\mathrm{rel}}_{\mathrm{sum}}$ or $\mathcal{T}^{\mathrm{rel}}_{\mathrm{max}}$ is class-specific (true, predicted or arbitrarily chosen, depending on θ), model-specific and potentially concept-specific (depending on θ), as is also illustrated in Supplementary Fig. 7a. The resulting set of reference samples thus includes only samples that depict facets of a concept that are actually useful for the model during inference (Supplementary Fig. 7b). How differences in resulting reference sets $\mathcal{X}_k^{\star\,\mathrm{act}}$ and $\mathcal{X}_k^{\star\,\mathrm{rel}}$ can occur is depicted in Supplementary Fig. 7c,d. One can see that relevances are not strictly correlated to activations, because they also depend on the downstream relevances propagated from higher layers affected by feature interactions at the current and following layers. For further details and evaluations, we refer the interested reader to Supplementary Notes 3 and 4.

Comparing feature channels with averaged cosine similarity on reference samples

We propose a simple but qualitatively effective method for comparing filters in terms of activations based on reference samples, for grouping similar concepts in CNN layers. Based on the notation in previous sections, $\mathcal{X}^*_{(k,q)}$ denotes a set of k reference images for a channel q in layer l and $\mathbf{z}^l_q(\mathbf{W}, \mathbf{x}_m)$ the ReLU-activated outputs of channel q in layer l for the mth input sample $\mathbf{x}_m$ of the dataset with all required network parameters $\mathbf{W}$ for its computation. Specifically, for each channel q and its associated full-sized (that is, not cropped to the channels' filters' receptive fields; compare with Supplementary Note 3.4) reference samples $\mathbf{x}_m \in \mathcal{X}^{\star\,\mathrm{rel}}_{k,q\,\mathrm{sum}}$, we compute $\mathbf{z}^q_m = \mathbf{z}^l_q(\mathbf{W}, \mathbf{x}_m)$, as well as $\mathbf{z}^p_m = \mathbf{z}^l_p(\mathbf{W}, \mathbf{x}_m)$ for all other channels $p \neq q$, by executing the forward pass, yielding activation values for all spatial neurons for the channels. We then define the averaged cosine similarity ρ_{qp} between two channels q and p in the same layer l as

$$\rho_{qp} = \frac{1}{2}\left(\cos(\phi)_{qp} + \cos(\phi)_{pq}\right) \tag{18}$$

with

$$\cos(\phi)_{qp} = \frac{1}{k} \sum_{\mathbf{x}_m \in \mathcal{X}^{\star\,\mathrm{rel}}_{(k,q)\,\mathrm{sum}}} \frac{\mathbf{z}^q_m \cdot \mathbf{z}^p_m}{||\mathbf{z}^q_m|| \cdot ||\mathbf{z}^p_m||}. \tag{19}$$

Note that we symmetrize ρ_{qp} in equation (18) as the cosine similarities $\cos(\phi)_{qp}$ and $\cos(\phi)_{pq}$ are in general not identical, due to the potential dissimilarities in the reference sample sets $\mathcal{X}^*_{(k,q)}$ and $\mathcal{X}^*_{(k,p)}$. Thus, $\cos(\phi)_{qp}$ measures the cosine similarity between filter q and filter p with respect to the reference samples representing filter q. From equation (18), the resulting symmetric similarity measures $\rho_{qp} = \rho_{pq} \in [0, 1]$ can now be clustered, and visualized via a transformation into a distance measure $d_{qp} = 1 - \rho_{qp}$ serving as an input to t-distributed stochastic neighbor embedding (t-SNE)[55], which visually clusters similar filters together in, typically, $\mathbb{R}^2$. Note that normally, the output value of the cosine distance covers the interval $[-1, 1]$, where for -1 the two measured vectors are exactly opposite to one another, for 1 they are identical and for 0 they are orthogonal. In case output channels of dense layers are analysed, that is, scalar values, the range of output values reduces to the set $\{-1, 0, 1\}$, as both values are either of same or different signs, or at least one of the values is zero. As we process layer activations after the ReLU

nonlinearities of the layer, this yields only positive values for $\mathbf{z}_m^q$ and $\mathbf{z}_m^p$. This results in $\rho_{pq} \in [0, 1]$, and a conversion to a canonical distance measure $d_{qp} \in [0, 1]$.

Human evaluation study details

In the following, we provide further details on the conduction of the human study in the 'Human evaluation study' section. All participants were recruited on the Amazon Mechanical Turk platform, representing people from all backgrounds that do not necessarily have any background knowledge from the field of artificial intelligence. As such, the participants reflect the general non-expert population in interaction with (X)AI. It is noted, however, that on this platform, participants might work on other unrelated studies for several hours, which can have a negative impact on their performance. The study did not consider the sex, gender, race, ethnicity or other socially relevant groupings of the participants, as they were not relevant to the research. Consequently, no corresponding data have been collected.

The study was conducted using a between-subject design from 19–26 September 2022. Each participant was assigned randomly to one of the groups (25 participants per group) associated with one of the XAI methods. The sample size of 25 is chosen such that the differences in terms of accuracy between our method and the other methods become significant (according to two-sample t-test probabilities). For the analysis, we only considered studies fully finished by the participants.

Regarding the computation and visualization of explanations, we used the publicly available ImageNet[56] dataset, and fine-tuned two VGG-16[57] DNNs, with parameters pretrained on ImageNet as obtained from the PyTorch[58] model zoo. The interested reader can find additional details about the design and the evaluation of the conducted study in Supplementary Note 7 and on GitHub (https://github.com/maxdreyer/crp-human-study), providing Python code for generating explanations as well as HTML templates for Amazon Mechanical Turk.

Ethics approval

The Ethics Commission Faculty IV TU Berlin provided guidelines for the study procedure and determined that no protocol approval is required. Informed consent has been obtained from all participants.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The study was conducted using the publicly available ImageNet[56] dataset. Code, models and samples used for the execution of our user study can be found at https://github.com/maxdreyer/crp-human-study. More information about data and models utilized in other experiments can be found in Supplementary Note 12. The license to re-use and reproduce have been granted for the images shown in the figures of this paper and its Supplementary Information to the authors by the respective copyright holders by iStock, Shutterstock, Pixabay and Pexels. Additional results obtained on the openly available benchmark datasets, such as ImageNet or Caltech-UCSD Birds 200, can be found in ref. 59.

Code availability

We provide an open-source CRP toolbox for the scientific community written in Python and based on PyTorch[58] and Zennit[51]. The GitHub repository containing our implementations of CRP and RelMax is publicly available at https://github.com/rachtibat/zennit-crp (ref. 60). All experiments were conducted with Python 3.8, zennit-crp v0.6, Zennit v0.4.6 and PyTorch v1.13.1.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Dai, Z., Liu, H., Le, Q. V. & Tan, M. CoAtNet: marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **34**, 3965–3977 (2021).
3. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
4. Jaderberg, M. et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* **364**, 859–865 (2019).
5. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
6. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J. & Müller, K.-R. Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* **109**, 247–278 (2021).
7. Stock, P. & Cisse, M. Convnets and ImageNet beyond accuracy: understanding mistakes and uncovering biases. In *European Conference on Computer Vision* (eds Ferrari, V. et al.) 498–512 (Springer, 2018).
8. Lapuschkin, S. et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).
9. Schramowski, P. et al. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.* **2**, 476–486 (2020).
10. Anders, C. J. et al. Finding and removing Clever Hans: using explanation methods to debug and improve deep models. *Inf. Fusion* **77**, 261–295 (2022).
11. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-making and a 'right to explanation'. *AI Mag.* **38**, 50–57 (2017).
12. *Communication: Building Trust in Human Centric Artificial Intelligence* COM 168 (Commission to the European Parliament, the Council, the European Economic and Social Committee, the Committee of the Regions, 2019).
13. Morch, N. J. et al. Visualization of neural networks using saliency maps. In *Proc. ICNN'95-International Conference on Neural Networks* 2085–2090 (IEEE, 1995).
14. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks.In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 3319–3328 (PMLR, 2017).
15. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, 0130140 (2015).
16. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. A. Striving for simplicity: the all convolutional net. In *3rd International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) (ICLR, 2015).
17. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 3145–3153 (PMLR, 2017).
18. Murdoch, W. J., Liu, P. J. & Yu, B. Beyond word importance: contextual decomposition to extract interactions from LSTMs. In *6th International Conference on Learning Representations* (ICLR, 2018).
19. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision, Lecture Notes in Computer Science* (eds Fleet, D. et al) 818–833 (Springer, 2014).
20. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?": explaining the predictions of any classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Krishnapuram B. et al.) 1135–1144 (ACM, 2016).
21. Blücher, S., Vielhaben, J. & Strodthoff, N. PredDiff: explanations and interactions from conditional expectations. *Artif. Intell.* **312**, 103774 (2022).

22. Erhan, D., Bengio, Y., Courville, A. & Vincent, P. Visualizing higher-layer features of a deep network. *Univ. Montreal* **1341**, 1 (2009).
23. Olah, C., Mordvintsev, A. & Schubert, L. Feature visualization. *Distill* **2**, 7 (2017).
24. Kindermans, P.-J. et al. Learning how to explain neural networks: PatternNet and PatternAttribution. In *6th International Conference on Learning Representations* (ICLR, 2018).
25. Szegedy, C. et al. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) (ICLR, 2014).
26. Mahendran, A. & Vedaldi, A. Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5188–5196 (IEEE, 2015).
27. Mordvintsev, A., Olah, C. & Tyka, M. Inceptionism: going deeper into neural networks. *Google AI Blog* https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (2015).
28. Kim, B. et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In *Proc. 35th International Conference on Machine Learning* (eds Dy, J. G. & Krause, A.) 2668–2677 (PMLR, 2018).
29. Rajalingham, R. et al. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
30. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: quantifying interpretability of deep visual representations. In *IEEE International Conference on Computer Vision and Pattern Recognition* 3319–3327 (IEEE, 2017).
31. Bau, D. et al. Understanding the role of individual units in a deep neural network. *Proc. Natl Acad. Sci. USA* **117**, 30071–30078 (2020).
32. Hohman, F., Park, H., Robinson, C. & Chau, D. H. P. Summit: scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Trans. Vis. Comput. Graph.* **26**, 1096–1106 (2019).
33. Liu, M. et al. Towards better analysis of deep convolutional neural networks. *IEEE Trans. Vis. Comput. Graph.* **23**, 91–100 (2016).
34. Chen, C. et al. This looks like that: deep learning for interpretable image recognition. *Adv. Neural Inf. Process. Syst.* **32**, 8930–8941 (2019).
35. Chen, Z., Bei, Y. & Rudin, C. Concept whitening for interpretable image recognition. *Nat. Mach. Intell.* **2**, 772–782 (2020).
36. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T. & Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Adv. Neural Inf. Process. Syst.* **29**, 3387–3395 (2016).
37. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A. & Torralba, A. Object detectors emerge in deep scene CNNs. In *3rd International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) (ICLR, 2015).
38. Radford, A., Jozefowicz, R. & Sutskever, I. Learning to generate reviews and discovering sentiment. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1704.01444 (2017).
39. Cammarata, N. et al. Thread: circuits. *Distill* **5**, 24 (2020).
40. Goh, G. et al. Multimodal neurons in artificial neural networks. *Distill* **6**, 30 (2021).
41. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (IEEE, 2017).
42. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning* (ICML, 2017).
43. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
44. Becking, D., Dreyer, M., Samek, W., Müller, K. & Lapuschkin, S. in *xxAI—Beyond Explainable AI* Lecture Notes in Computer Science Vol. 13200 (eds Holzinger, A. et al.) 271–296 (Springer, 2022).
45. Li, C. High quality, fast, modular reference implementation of SSD in PyTorch. *GitHub* https://github.com/lufficc/SSD (2018).
46. Hacker, P. & Passoth, J.-H. Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers* (eds Holzinger, A. et al.) 343–373 (Springer, 2022).
47. Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* **65**, 211–222 (2017).
48. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **73**, 1–15 (2018).
49. Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* Lecture Notes in Computer Science Vol. 11700 (eds Samek, W. et al.) 193–209 (Springer, 2019).
50. Kohlbrenner, M. et al. Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks (IJCNN)* 1–7 (IEEE, 2020).
51. Anders, C. J., Neumann, D., Samek, W., Müller, K.-R. & Lapuschkin, S. Software for dataset-wide XAI: from local explanations to global insights with Zennit, CoRelAy, and ViRelAy. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2106.13200 (2021).
52. Arras, L., Montavon, G., Müller, K.-R. & Samek, W. Explaining recurrent neural network predictions in sentiment analysis. In *Proc. 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (eds Balahur, A. et al.) 159–168 (ACL, 2017).
53. Schnake, T. et al. Higher-order explanations of graph neural networks via relevant walks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7581–7596 (2021).
54. Yeh, C.-K. et al. On completeness-aware concept-based explanations in deep neural networks. *Adv. Neural Info. Processing Syst.* **33**, 20554–20565 (2020).
55. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
56. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
57. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) (ICLR, 2015).
58. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
59. Achtibat, R. et al. From 'where'' to 'what'': towards human-understandable explanations through Concept Relevance Propagation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2206.03208 (2022).
60. Achtibat, R., Dreyer, M. & Lapuschkin, S. rachtibat/zennit-crp: v0.6.0. *Zenodo* https://doi.org/10.5281/zenodo.7962574 (2023).

Acknowledgements

We express our gratitude to A. Angerschmid—associated with the Human-Centered AI Lab at the University of Natural Resources, Vienna, and the Medical University of Graz—for fruitful discussions and feedback.

Author contributions

Conceptualization and methodology: S.L., R.A., M.D., S.B., T.W. and W.S. Design of experiments: R.A., M.D., S.L., S.B., W.S. and T.W. Data analysis: R.A., M.D. and S.L. Software: R.A., I.E., M.D. and S.L.

Supervision and funding acquisition: S.L., W.S. and T.W. Writing—original draft and revision: R.A., M.D., S.L., W.S., I.E., S.B. and T.W.

Funding

Open access funding provided by Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00711-8.

Correspondence and requests for materials should be addressed to Wojciech Samek or Sebastian Lapuschkin.

Peer review information *Nature Machine Intelligence* thanks José Hernández-Orallo, Ribana Roscher and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Prof. Dr.Wojciech Samek, Dr. Sebastian Lapuschkin

Last updated by author(s): June 14, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
x	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	x	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	x	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
x	☐	A description of all covariates tested
x	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	x	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	x	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
x	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
x	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
x	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We provide an open-source CRP toolbox for the scientific community written in Python and based on PyTorch and Zennit [38]. The GitHub repository containing our implementation of CRP and RelMax is publicly available on https://github.com/rachtibat/zennit-crp. We use Python 3.8, zennit-crp v0.6, zennit v0.4.6 and torch v1.13.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The user study was conducted using the publicly available ImageNet [56] dataset. Code, models and samples used for the execution of our user study can be found at https://github.com/maxdreyer/crp-human-study. More information about data and models utilized in other experiments can be found in Supplementary Materials L.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	For the study, sex and gender of the participants were not relevant and thus not considered. Consequently, no corresponding data has been collected.
Reporting on race, ethnicity, or other socially relevant groupings	For the study, race, ethnicity, or other socially relevant groupings of the participants were not relevant and thus not considered. Consequently, no corresponding data has been collected.
Population characteristics	For the study, population characteristics were not relevant and thus not considered. Consequently, no corresponding data has been collected.
Recruitment	Participants have been recruited on the Amazon Mechanical Turk platform.
Ethics oversight	The Ethics Commission Faculty IV Technical University of Berlin provided guidelines for the study procedure and determined that no protocol approval is required.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	
Data exclusions	
Replication	
Randomization	
Blinding	

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is performed in the design of a quantitative experimental study.
Research sample	Participants have been recruited on the Amazon Mechanical Turk platform, representing people from all backgrounds that not necessarily have any background knowledge in the field of Artificial Intelligence. As such, the participants reflect the general non-expert population in interaction with (X)AI.
Sampling strategy	The study has been performed in a between-subject design with 25 random participants per group. The sample size is large enough to test for significant differences between explanation methods (according to two sample t-test probabilities).
Data collection	The data is automatically collected on an online platform, without the need of any researcher to be present.
Timing	Data has been collected throughout the week between September, 19th and September, 26th of 2022.
Data exclusions	We only consider finished studies.
Non-participation	All participants fully performed the study.
Randomization	Participants have been randomly assigned to a group.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	
Data exclusions	
Reproducibility	
Randomization	
Blinding	

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	
Location	
Access & import/export	
Disturbance	

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
[X]	[ ] Antibodies
[X]	[ ] Eukaryotic cell lines
[X]	[ ] Palaeontology and archaeology
[X]	[ ] Animals and other organisms
[X]	[ ] Clinical data
[X]	[ ] Dual use research of concern
[X]	[ ] Plants

Methods

n/a	Involved in the study
[X]	[ ] ChIP-seq
[X]	[ ] Flow cytometry
[X]	[ ] MRI-based neuroimaging

Antibodies

Antibodies used	
Validation	

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	
Authentication	
Mycoplasma contamination	
Commonly misidentified lines (See ICLAC register)	

Palaeontology and Archaeology

Specimen provenance	
Specimen deposition	
Dating methods	

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	
Wild animals	
Reporting on sex	
Field-collected samples	
Ethics oversight	

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	
Study protocol	
Data collection	
Outcomes	

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Novel plant genotypes

Authentication

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Files in database submission

Genome browser session
(e.g. UCSC)

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See Eklund et al. 2016)

Correction

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis