FunnyBirds: A Synthetic Vision Dataset for a Part-Based Analysis of Explainable AI Methods

Robin Hesse\(^1\)  Simone Schaub-Meyer\(^{1,2}\)  Stefan Roth\(^{1,2}\)

\(^1\)Department of Computer Science, TU Darmstadt  \(^2\)hessian.AI

Abstract

The field of explainable artificial intelligence (XAI) aims to uncover the inner workings of complex deep neural models. While being crucial for safety-critical domains, XAI inherently lacks ground-truth explanations, making its automatic evaluation an unsolved problem. We address this challenge by proposing a novel synthetic vision dataset, named FunnyBirds, and accompanying automatic evaluation protocols. Our dataset allows performing semantically meaningful image interventions, e.g., removing individual object parts, which has three important implications. First, it enables analyzing explanations on a part level, which is closer to human comprehension than existing methods that evaluate on a pixel level. Second, by comparing the model output for inputs with removed parts, we can estimate ground-truth part importances that should be reflected in the explanations. Third, by mapping individual explanations into a common space of part importances, we can analyze a variety of different explanation types in a single common framework. Using our tools, we report results for 24 different combinations of neural models and XAI methods, demonstrating the strengths and weaknesses of the assessed methods in a fully automatic and systematic manner.

1. Introduction

Even though deep learning models have achieved breakthrough results in computer vision, their inner workings remain largely opaque. As a result, deep networks sometimes receive only limited user trust and cannot be applied blindly in safety-critical domains. To overcome this issue, a growing interest in the field of explainable artificial intelligence (XAI) has emerged, attempting to explain the inner workings of deep neural models in a human-comprehensible way. However, since there are generally no ground-truth explanations, evaluating XAI methods remains an open challenge. In fact, a third of XAI papers lack sound quantitative evaluation [38], while other work has limited comparability [31] or problematic evaluation protocols [23, 38].

To overcome the issue of missing ground-truth explanations, automatic evaluations are often done via proxy tasks that adhere to the idea of performing image interventions by removing certain input features to then measure the resulting impact on the model output [21, 23, 55]. As image interventions are non-trivial to perform on existing vision datasets, they are usually applied on a pixel level, e.g., masking out single pixels [21, 23, 27, 52, 55]. However, this and related approaches share several downsides. First, performing interventions, as well as evaluating explanations on a pixel level, is disconnected from the downstream task of providing human-understandable explanations since humans perceive images in concepts rather than pixels. Second, existing automatic evaluation protocols are developed for specific explanation types, e.g., pixel-level attribution maps, and thus, cannot be extended to other explanation approaches. For each input image, we render the corresponding part annotation and part interventions (e.g., removed beak). The interventions are used to estimate ground-truth importance scores of each input part, which are then compared to the estimated part importances from the explanation method under inspection (here [44]). We evaluate multiple dimensions of explainability to draw a more conclusive picture. Please refer to Sec. 4.1 for details.

Dataset and code available at github.com/visinf/funnybirds/.
types like prototypes [12]. Third, by performing unrealistic interventions in image space, e.g., masking out pixels, they introduce domain shifts compared to the training distribution [10, 23, 28], which can cause the model to behave unexpectedly, and thus negatively affects the evaluation.

In this work, we address the above and more challenges to contribute an important step toward a more rigorous quantitative evaluation of XAI methods by proposing a thorough, dedicated evaluation/analysis tool. We do so by building a fully controllable, synthetic classification dataset consisting of renderings of artificial bird species. This approach to analyzing XAI methods is analogous to controlled laboratory research, where we have full control over all variables, eliminating the potential influence of irrelevant factors, and therefore, providing clearer evidence of the observed behavior [5]. Our proposed dataset allows us to make the following main contributions: (1) We cover a wide range of dimensions of explainability by considering a collection of different evaluation protocols. (2) We allow to automatically compare various explanation types in a shared framework. (3) We avoid the out-of-domain issue of previous image-space interventions by introducing semantically meaningful interventions at training time. (4) We reduce the gap between the downstream task of human comprehension and XAI evaluation by proposing metrics that operate at a semantically meaningful part level rather than the semantically less meaningful pixel level. (5) We automatically analyze the coherence of explanations. (6) We analyze 24 different combinations of existing XAI methods and neural models, highlighting their strengths and weaknesses as well as identifying new insights that may be of general interest to the XAI community.

2. Related Work

XAI methods. In this work, we focus on image classification explanations for the outcome of one particular input using deep neural networks (DNNs). Perturbation-based explanation methods [1, 11, 33, 44, 47, 63, 68, 69] estimate the impact of pixels/regions by measuring the output change occurring when repeatedly perturbing inputs or neurons. Backpropagation-based explanation methods [4, 20, 27, 53, 55, 57, 59] use importance signals, e.g., gradient information, to estimate the importance of each input feature. Both types of methods generate post-hoc explanations for already trained networks. On the other hand, intrinsically explainable models [3, 6, 8, 9, 12, 39] consider explainability already in the design process of the model itself. For example, B-cos networks [9] enforce weight-input alignment to highlight the most important input regions. ProtoPNet [12] computes the feature similarity between test image patches and prototypes to help explain why a decision was made based on this patch looking like that prototype. BagNets [6] use a linear classifier on top of each feature extracted from small image patches to highlight task-important regions. \( \alpha \)-DNNs [27] remove the bias term in ReLU networks to allow for an efficiently computable closed-form solution of Integrated Gradients [59].

Evaluating XAI methods. A sound evaluation of XAI methods is essential to compare different approaches and unveil their strengths and weaknesses. Nauta et al. [38] argue that explainability is a multi-facet concept and propose various properties that describe different aspects of explanation quality. Below, we outline existing evaluation protocols for three of these dimensions that have received particularly much attention in related work and that allow for a well-defined automatic evaluation:

Correctness denotes the faithfulness of an explanation w.r.t. the model [38]. A common way to measure the correctness of attribution methods is the Incremental Deletion protocol [8, 10, 21, 23, 27, 28, 38, 39, 52, 55, 60], where input pixels are incrementally removed in the order of their relative importance to measure the impact on the output. Similarly, in the Single Deletion protocol [3, 14, 38, 41, 67], individual (input) features are deleted to compute the correlation between feature importance and the actual output change. A drawback of the above protocols is the domain shift occurring when introducing image interventions, which can cause undesirable model behavior influencing the metrics [10, 23, 28]. To reduce this, masked pixels can be inpainted with a generative model [10] or text classification datasets can be augmented with training inputs that have similar alterations as the evaluation protocol [23]. However, none of these approaches achieve an exact alignment of the train and test domain as we do in this work.

Completeness measures the degree to which an explanation describes all aspects of the model’s decision [38]. In the Preservation/Deletion Check protocols [10, 13, 16, 38, 45], important features are preserved/deleted to test if the output prediction stays the same/changes. However, as above, these protocols introduce unrealistic domain changes. In the Controlled Synthetic Data Check protocol [2, 13, 30, 34, 38, 50], a synthetic dataset is used to evaluate if the explanation of a model with almost perfect accuracy aligns with the dataset generation process. While closely related to our work, existing synthetic datasets for evaluating XAI methods are very small-scale, only evaluate a single dimension of explainability, and do not allow for diverse image-space interventions that are crucial for evaluating correctness.

Contrastivity measures the degree to which an explanation is discriminative w.r.t. different outputs [38]. It can be measured with the Target Sensitivity protocol [38] that evaluates if the explanations for two different targets are different [45] and sometimes, even more, whether the explanation aligns with the image region of the correct target [8, 19, 65]. However, these evaluations can require accurate object annotations, which are not given for most classi-
Besides above issues when arranging images in a grid as in [8, 9, 46].

Human studies. Besides above automatic evaluations, there have been efforts to assess the quality of XAI methods with human subjects [15, 40, 51, 54]. Human studies nicely reflect the actual downstream tasks of explanations, i.e., helping humans to understand models, and they can be applied to explanations that go beyond simple heat maps [31]. Downsides of user studies include larger cost and human biases [36]. Further, humans are not capable of directly evaluating if explanations are correct w.r.t. a complex, deep model. With our custom analysis, we aim to reduce the gap between the automatic evaluation and human studies while addressing the described drawbacks of human assessment.

Synthetic datasets. Synthetic datasets have an important and long history in many other disciplines of computer vision [7, 18, 29, 42, 48]. Richter et al. [48] utilize a video game to create a large-scale segmentation dataset. The “Flying Chairs” dataset [18] and MPI Sintel [7] are popular synthetic datasets used for the training and evaluation of optical flow methods. CLEVR [29] was introduced to test different visual reasoning abilities. All these works advanced their specific sub-fields, showing that synthetic datasets are a valuable tool, even in the absence of full realism.

3. The FunnyBirds Dataset

We propose a synthetic vision dataset that is developed to automatically and quantitatively analyze XAI methods. To this end, we made a multitude of carefully considered design choices, which we summarize in the following.

Since XAI methods in computer vision have been developed mainly in the context of classification to date [38], we focus on classification and propose a fine-grained bird species dataset inspired by the CUB-200-2011 dataset [61] commonly used by XAI methods [12, 22, 26, 39, 66]. Our dataset consists of 50,500 images (50k train, 500 test) of 50 synthetic bird species. We found 500 test images to be sufficient to produce stable results while allowing efficient evaluations with limited hardware (see supplement).

One particularly important desideratum behind our design process is the notion of concepts that we define as mental representations of psychological entities that are crucial to categorization [35, 37, 49, 64]. For instance, humans might categorize a bird as a flamingo when observing the concepts curved beak and pink wings. Similarly, they would communicate their reasoning process in terms of concepts [49]. Hence, we argue that concepts are essential in the context of XAI for humans, and thus, we design our dataset with a particular emphasis on the notion of concepts. To derive a precise level of granularity for the concepts in our dataset, we enforce them to be as granular as possible, while corresponding to an existing individual word, and being attached to the bird’s body. This avoids “unrealistic” scenarios where the removal of one concept leaves another concept unattached (the body of the bird is never removed) and eliminates overly fine-grained concepts. Following these rules, our FunnyBirds consist of the five human comprehensible concepts beak, wings, feet, eyes, and tail, to which we will refer as parts in the following.

We manually design 4 beaks, 3 eyes, 4 feet, 9 tails, and 6 wings of predefined varied shape and/or color (see supplement). Every FunnyBird class consists of a unique combination of these parts. We randomly sample 50 classes out of the 2,592 combinations as our dataset. Using these classes, we follow the data-generation process outlined in Fig. 2 by adding class-specific parts to a neutral body model to obtain an instance of a FunnyBird. To reflect real-world challenges, we randomly add background objects and change the illumination and viewpoint. We also explored more photo-realistic simulations with real background images and textured parts. However, with neural models learning to ignore these “distractions”, our quantitative analysis was largely unaffected. We thus omit this in pursuit of simplicity. We verified the solvability of our dataset with a human expert who achieved an accuracy of 97.4% on the proposed test set. To ensure that removing any part does not result in another valid class, each complete FunnyBird instance contains all five parts. To ensure that images with removed parts are not out-of-distribution, we include images with missing parts in the training set. As these images can no longer be associated with one specific class, we use a multi-label classification training scheme where all classes that contain all of the remaining parts are considered valid (with equal probability).

Our data-generation process has several important implications that make it particularly well suited for evaluating and analyzing XAI methods. First, based on the assumption that humans think in concepts rather than pixels, we argue that a concept-based dataset and the resulting evaluations are better aligned with downstream tasks of human comprehension than pixel-based evaluations. Second, we have pixel-accurate annotations of each object and bird part in the scene, which is required for various automatic evaluation
protocols and tailored analyses. Third, having full control over the data-generation process allows applying realistic image-space interventions, such as removing parts, which can be used to estimate ground-truth importance scores.

4. Evaluation Protocols

We will now outline how our dataset and its functionalities can be used for the quantitative evaluation and analysis of XAI methods. We first propose a general, multi-dimensional analysis framework, named FunnyBirds framework, that allows to evaluate a wide range of existing XAI methods in a common framework. Second, we propose a more tailored way of analysis where our dataset’s capabilities are used to gain deeper insights into specific methods similar to what human studies could do.

4.1. The FunnyBirds framework

The multi-dimensional FunnyBirds analysis framework for XAI methods entails six evaluation protocols for three explainability dimensions, i.e., completeness, correctness, and contrastivity. They were chosen for their popularity in related work [38], ability to be automatically measurable, and compatibility with our dataset. Most of our protocols are inspired by well-established evaluation practices and follow accepted assumptions. Our dataset generation process ensures that all conducted interventions can be considered in-domain and semantically meaningful, eliminating common downsides [23, 28] of several existing evaluation protocols. We thus propose to build our evaluation framework around so-called interface functions that implement more general properties, such as which object parts are important. They can be derived from a variety of explanation types, and thus, used to evaluate distinct explanations in a common framework. This allows us to not only compare a spectrum of existing methods, but also to potentially include future, novel explanation types. Our evaluation framework currently uses the following interface functions:

- PI(·) – The part importance score for each part. For example, the summed attribution within that part.
- P(·) – The parts that are considered to be important, derived from the output of an XAI method. For attribution methods, a part can be considered important if its part importance is at least 1% of the total attribution.

Note that different explanation types vary in their suitability for differing downstream tasks, and thus, also in their suitability for above interface functions. Furthermore, just as humans differ in interpreting various explanations, there may not be a single correct way of implementing the interface functions (see supplement). Therefore, comparing different explanation types is not absolutely fair and should be done with caution. Nonetheless, we argue that estimating important image regions and their relative importance are fairly basic challenges in XAI, and thus, appropriate downstream tasks, respectively, interface functions.

Framework. For a simple comparison and visualization, we design all metrics to be in the range [0, 1], with higher values indicating better scores. For each dimension (as described below), we report a single score comprised of all included protocols (see Tab. 1). Additionally, we report the average of all our explanation quality dimensions, i.e., completeness, correctness, contrastivity, as a single score that summarizes the overall quality of an XAI method for the inspected model. We let \( D \) denote our dataset of size \( N \), consisting of images \( x_n \) with class label \( c_n \): \( f \) is the inspected model with \( f(x_n) \) denoting the predicted class or the logit of the target class, and \( e_f(x_n) \) is the explanation of \( f(x_n) \) w.r.t. the target class \( c_n \).

4.1.1 Accuracy and background independence

Accuracy (A). An overly simple model may be explainable but not solve the task at hand. To detect such cases, our framework reports the standard classification accuracy.

Background independence (BI). Similarly, another trivial solution to achieve high explanation scores is a model that is sensitive to the entire image. An explanation would consequently highlight the entire image. To detect this case, we report the background independence as the ratio of background objects that are unimportant, i.e., that, when removed, cause the target logit to drop less than 5%.
4.1.2 Completeness

Controlled synthetic data check (CSDC). Our transparent data generation process allows us to know the subsets of parts \( \{ P'_{c_n,i} \mid i = 1, \ldots \} \) that are sufficient to correctly categorize an image of class \( c_n \), containing parts \( P_{c_n} \). A model that accurately captures the essence of our dataset, \( i.e., \), has an accuracy close to 1.0, must thus consider at least one of these sets \( P'_{c_n,i} \) \[13,34,38,50\]. To ensure that this assumption holds and CSDC is faithful \( w.r.t. \) the model, we exclude misclassified samples from this protocol. We measure the CSDC completeness of an explanation as the set overlap between the parts that are estimated to be important by the explanation \( i.e., \), \( P(ef(x_n)) \) where \( P(\cdot) \) is the interface function giving the important parts) and the sufficient part set that has the largest overlap with \( P(ef(x_n)) \), normalized by the size of the sufficient part set:

\[
\text{CSDC} = \frac{1}{N} \sum_{n=1}^{N} \max_i \frac{|P(ef(x_n)) \cap P'_{c_n,i}|}{|P'_{c_n,i}|}. \tag{1}
\]

Preservation check (PC). If an explanation is complete, preserving only the parts of the input that are estimated to be important by the explanation should still result in the same classification output \[13,16,38\]. To verify this, we first pass the original image through the inspected model to obtain the classification and the corresponding explanation. Afterward, we delete all parts of the image that are estimated to not be important by the explanation, \( i.e., \), the inverse of \( P(\cdot) \), and measure how often the class prediction remains unchanged. We define our PC score as

\[
\text{PC} = \frac{1}{N} \sum_{n=1}^{N} \left| f(x'_n) = f(x_n) \right|, \tag{2}
\]

with \( \lfloor \cdot \rfloor \) denoting the Iverson bracket \[32\] and \( x'_n \) denoting the image arising by removing all bird parts but \( P(ef(x_n)) \) from \( x_n \). \( I.e., \), the parts estimated to be important remain.

Deletion check (DC). Similarly, if an explanation is complete, deleting all the parts of the input that are estimated to be important by the explanation should result in a different classification output \[16,38,45\]. Thus, the deletion check is the inverse of PC and defined as

\[
\text{DC} = \frac{1}{N} \sum_{n=1}^{N} \left| f(x''_n) \neq f(x_n) \right|, \tag{3}
\]

with \( x''_n \) denoting the image arising by removing the estimated important parts \( P(ef(x_n)) \) from \( x_n \).

Distractibility (D). As over-complete explanations would achieve high completeness scores, we argue that completeness should be considered together with a metric that counts over-completeness (similar to precision and recall). To this end, we introduce a novel over-completeness score that measures how many actually unimportant input parts are also estimated to be unimportant by the explanation. In detail, we successively remove each background object and bird part from the input and consider it to be unimportant if the logit target output drops less than 5% compared to the output for the original image. Next, we measure the overlap between parts \( P(ef(x_n)) \) estimated to be important by the explanation and parts \( P'(ef(x_n)) \) that are actually unimportant, normalized by the number of unimportant parts. As the overlap should be small for good evaluations, we subtract the result from 1:

\[
D = 1 - \frac{1}{N} \sum_{n=1}^{N} \frac{|P(ef(x_n)) \cap P'(ef(x_n))|}{|P(ef(x_n))|}. \tag{4}
\]

The final completeness score \( (\text{Com.}) \) is computed as the mean of the averaged three completeness metrics \( \text{CSDC}, \text{PC}, \text{and DC} \), and the over-completeness \( D \). By having complementing metrics, we can also automatically estimate the threshold \( t \) of \( P(\cdot) \), such that \( t \) optimally trades off the completeness and over-completeness metrics.

4.1.3 Correctness

Single deletion (SD). In the single deletion protocol, we measure the rank correlation between the explanation’s importance score of each part and the actual amount that the model’s logit of the target class changes when removing that
part [3, 14, 38]. Specifically, we define SD as

$$SD = \frac{1}{2} + \frac{1}{2N} \sum_{n=1}^{N} \rho \left( \text{PI}(f(x_n)), f(x_n) - \{f(x''_n)\} \right),$$

with $\rho$ denoting the Spearman rank-order correlation coefficient, $\text{PI}(\cdot)$ the part importance interface function, and $\{f(x''_n)\}$ the target scores of the images obtained by removing any individual bird part from $x_n$.

4.1.4 Contrastivity

Target sensitivity (TS). If an explanation is sensitive to a target class, it should highlight image regions that are relevant for the respective target [8, 9, 38]. To measure target sensitivity, we take an input sample and select two classes $\hat{c}_1$ and $\hat{c}_2$ that each have exactly two non-overlapping common parts with the actual class of input sample $x_n$. We choose two parts because this is the floor of half the 5 bird parts available in each image. Afterward, we compute the explanations $f(x_n, \hat{c}_1)$ with respect to $\hat{c}_1$ and $\hat{c}_2$ and evaluate if they correctly highlight the parts belonging to each respective class:

$$TS = \frac{1}{2N} \sum_{n=1}^{N} \left[ \text{PI}(f(x_n, \hat{c}_1)) > \text{PI}(f(x_n, \hat{c}_2)) \right] + \left[ \text{PI}(f(x_n, \hat{c}_2)) > \text{PI}(f(x_n, \hat{c}_1)) \right],$$

with $\text{PI}(\cdot)$ and $\text{PI}''(\cdot)$ denoting the summed part importances of the two parts belonging to class $\hat{c}_1$, respectively $\hat{c}_2$. To ensure that our protocol is faithful w.r.t. the model, we only consider images where removing parts belonging to $\hat{c}_1$ causes a larger model output drop for $\hat{c}_1$ than removing parts belonging to $\hat{c}_2$, and vice versa.

4.2. The FunnyBirds custom analysis

As it is almost impossible to analyze all aspects of an explanation with a single framework, we also show how full control over our dataset enables more tailored and systematic custom evaluations. Similar to recent human studies [31], we evaluate the coherence of the assessed methods, i.e., how reasonable an explanation is for a human. To the best of our knowledge, this is one of the first methods for evaluating coherence in a fully automatic and systematic manner. We give two illustrative examples for implementing such an evaluation, but other instantiations are possible.

ProtoPNet. ProtoPNet [12] promotes explainability by communicating which patches of the training data, called prototypes, look similar to patches in a test image. However, this only increases interpretability if the detected similarities are compatible with human notions of similarity. To quantitatively verify this in an automatic fashion, we test if relevant part types (e.g., beak or wing) in the detected patch of the test image are also contained in the corresponding prototype identified by ProtoPNet. This assumes that different part types, i.e., beak and wings, do not look similar to most humans, and therefore, are not helpful similarities for explainability. While debatable for real images, we argue that this is a reasonable assumption for our dataset.

Counterfactual visual explanations. The objective of counterfactual visual explanations (CVE) [22] is to highlight image regions that discriminate the class $c$ of a query image from the class $c'$ of a distractor image. To this end, CVE extracts convolutional neural network (CNN) feature maps from the query and distractor images to then perform feature swaps between the two feature maps that, when fed to the CNN classifier, lead to the strongest class change (from $c$ to $c'$). The corresponding image locations of the feature swaps are then highlighted to yield the CVE. We argue that this kind of explanation is only meaningful for humans if the swaps are semantically correct, i.e., when a beak is swapped with another beak but not a foot. Using our proposed dataset, we conduct a systematic study and measure the average number of swaps that do not touch the same bird part in both images, and thus, can be considered semantically incorrect. We evaluate four different setups of increasing difficulty where we only use aligned, respectively unaligned, viewpoint pairs with classes that are different in only one, respectively multiple parts.

Limitations. As with controlled laboratory research [5], our dataset does not reflect all real-world challenges, and thus, is not fully representative of natural images. However, we argue that, as long as XAI methods are challenged to truthfully explain models trained even on our controlled dataset, our proposed method is a valuable tool to faithfully assess the quality of XAI methods. Nonetheless, for future XAI methods we suggest a two-stream evaluation that utilizes our proposed framework and custom analysis for theoretically grounded assessments of the explanation quality in addition to benchmarks that work on natural image data, e.g., Quantus [25]. Further, while the dimensions of explainability evaluated in our work can be considered relevant for many applications, there are dimensions that we omit here (see supplement). However, our framework can easily be extended with additional protocols if needed.

5. Results and Interpretation

5.1. FunnyBirds framework

Results for our multi-dimensional FunnyBirds framework across a broad variety of different XAI methods and three different backbones are given in Fig. 3. For detailed results of each individual evaluation protocol, please refer to the supplement. We plot qualitative explanations for a random image and give the evaluation scores from Sec. 4.1 averaged over the entire test set. From this, we can make
several interesting findings and derive various hypotheses:

(1) Overall, BagNet [6] achieves the highest explainability score, which could be due to its internal model structure that is better aligned with the explanatory power of an attribution map, i.e., pixels are processed more independently.

(2) On average, Integrated Gradients (IG) [59] performs best among model-agnostic methods for different backbones, which is in line with the literature when considering gradient-based approaches [56, 59]. However, LIME [47] and IG achieve similar completeness, with LIME lacking the ability to communicate meaningful importance scores, which impairs practicability for certain downstream tasks.

Interestingly, taking the absolute value of IG impairs the ability to communicate meaningful importance scores, which impairs practicability for certain downstream tasks.

(3) Explanation scores for model-agnostic XAI methods are consistently better for VGG16 [58] than for ResNet-50 [24], indicating that different backbones vary in their explainability. We also observe that the relative ordering of different XAI methods changes across different backbones (e.g., RISE [44] and IxG [56]), showing the importance of evaluating future XAI methods on different backbones.

(4) IxG [56] and Grad-CAM [11, 53] seem to work better on CNN-based architectures [24, 58]. Methods developed for vision transformers, i.e., Rollout [1] and Chefer LRP [11], provide strong explanations, with Chefer LRP achieving the second highest explanation score.

Rollout

bagged networks (BagNet) [6] and contextualized attention maps (ProtoPNet) [12] are consistently better for VGG16 [58] than for ResNet-50 [24], indicating that different backbones vary in their explainability. We also observe that the relative ordering of different XAI methods changes across different backbones (e.g., RISE [44] and IxG [56]), showing the importance of evaluating future XAI methods on different backbones.

(4) IxG [56] and Grad-CAM [11, 53] seem to work better on CNN-based architectures [24, 58]. Methods developed for vision transformers, i.e., Rollout [1] and Chefer LRP [11], provide strong explanations, with Chefer LRP achieving the second highest explanation score.

Rollout

bagged networks (BagNet) [6] and contextualized attention maps (ProtoPNet) [12] are consistently better for VGG16 [58] than for ResNet-50 [24], indicating that different backbones vary in their explainability. We also observe that the relative ordering of different XAI methods changes across different backbones (e.g., RISE [44] and IxG [56]), showing the importance of evaluating future XAI methods on different backbones.

(4) IxG [56] and Grad-CAM [11, 53] seem to work better on CNN-based architectures [24, 58]. Methods developed for vision transformers, i.e., Rollout [1] and Chefer LRP [11], provide strong explanations, with Chefer LRP achieving the second highest explanation score.
cannot compute class-specific explanations, which could be disadvantageous in certain downstream tasks.

(5) Coarser methods like Grad-CAM [53] and RISE [44] tend to be more incomplete, which could be an artifact of poor localization [46] and our parts being fairly small.

(6) The B-cos network [9] outperforms competing methods on the same backbone, showing an improved explainability, which is in line with results from [9]. Surprisingly, we see a slight decrease in background independence.

(7) ProtoPNet [12] has poor correctness and contrastivity, which can be explained by related findings in that ProtoPNet similarities are not semantically reasonable [31] and by its poor localization (see Sec. 5.2 and Fig. 4).

(8) Especially correctness is a weakness of all examined methods, showing that, while existing approaches tend to accurately identify important areas, they lack the ability to reliably communicate the relative importance of each input area, or that the simple explanation of an attribution map might be too strong of a simplification of the actual model.

Finally, since XAI evaluation is multi-faceted and downstream task dependent, we doubt that a single score can be enough to properly compare methods. Hence, our work is less about ranking but more about automatically revealing the strengths and weaknesses of existing and future XAI methods in a systematic and theoretically grounded manner.

5.2. Custom evaluations

ProtoPNet. To quantitatively evaluate if prototypes used in [12] promote explainability (see Sec. 4.2), we conduct several sanity checks. First, we measure that out of 1000 examined ProtoPNet patches in the test image, only 510 touch the same part in the corresponding prototype patch, and thus, have a similarity that could be aligned with human notions of similarity. Further, we measure that 26.3% of the extracted test image patches do not touch a single bird part, indicating a poor localization of the explanation. Finally, 18.5% of the test image patches entail 4 or 5 different bird parts, which shows that the extracted patches are not local but global, and thus, do not promote the interpretability of the examined model. Qualitative results showing the described issues can be seen in Fig. 4. To summarize, many of the communicated prototypes of ProtoPNet [12] trained on our dataset are unlikely to be meaningful for humans. While a similar conclusion was drawn in a recent human study [31] confirming our findings, our proposed tools allow the same conclusion in a fully automatic manner.

Counterfactual visual explanations. To quantify if CVE [22] highlights meaningful swaps (see Sec. 4.2), we measure how many of the extracted swaps do not touch the same bird parts in the two images. When considering images that have the same viewpoint and differ only by one part (Fig. 5 – left), CVE with VGG16 [58] produces on average 1.65 swaps that do not touch the same part, and therefore, are semantically meaningless. When changing the viewpoint, CVE produces on average 1.3 swaps that do not touch the same part. When working with aligned viewpoints and birds that differ by multiple parts (4.07 parts on average), 2.75 such swaps are produced and 2.62 when also changing the viewpoint (Fig. 5 – right). Interestingly, the number of different parts has a stronger impact on the absolute number of unreasonable swaps, i.e., swaps that do not touch the same part in both images, than the viewpoint. Further, we observe that even in our controlled environment, CVE produces on average ~ 3 unreasonable swaps for images with on average only 4 different parts, indicating that many of the proposed swaps are semantically incorrect, and thus, do not promote coherent explanations.

6. Conclusion

In this work, we propose a novel approach to automatically analyzing XAI methods using a synthetic classification dataset that allows for full annotation and part interventions. Using this dataset, we propose an accompanying multi-dimensional analysis framework that faithfully assesses various important aspects of explainability and generalizes to different explanation types by using interface functions. With this easy-to-use tool, we analyzed 24 different setups to uncover various new insights and confirm findings from related work. This shows that, despite the synthetic setting, our findings appear to translate well to real data, and that our proposed tool is a practical and valuable asset for the analysis of future XAI methods. Finally, we showed how to develop tailored analyses to gain a better understanding of two specific XAI methods and discovered their weaknesses in an automatic and quantitative manner.
Acknowledgements. We thank Bernt Schiele, Moritz Böhle, and Sukrut Rao for their valuable feedback and helpful discussions. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 866008). The project has also been supported in part by the State of Hesse through the cluster projects “The Third Wave of Artificial Intelligence (3AI)” and “The Adaptive Mind (TAM)”.

References

[29] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick.


