

CLEVR-XAI: A Benchmark Dataset for the Ground Truth Evaluation of Neural Network Explanations

Leila Arras*, Ahmed Osman*, Wojciech Samek

Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany

Abstract

The rise of deep learning in today’s applications entailed an increasing need in explaining the model’s decisions beyond prediction performances in order to foster trust and accountability. Recently, the field of explainable AI (XAI) has developed methods that provide such explanations for already trained neural networks. In computer vision tasks such explanations, termed *heatmaps*, visualize the contributions of individual pixels to the prediction. So far XAI methods along with their heatmaps were mainly validated qualitatively via human-based assessment, or evaluated through auxiliary proxy tasks such as pixel perturbation, weak object localization or randomization tests. Due to the lack of an objective and commonly accepted quality measure for heatmaps, it was debatable which XAI method performs best and whether explanations can be trusted at all. In the present work, we tackle the problem by proposing a ground truth based evaluation framework for XAI methods based on the CLEVR visual question answering task. Our framework provides a (1) selective, (2) controlled and (3) realistic testbed for the evaluation of neural network explanations. We compare ten different explanation methods, resulting in new insights about the quality and properties of XAI methods, sometimes contradicting with conclusions from previous comparative studies. The CLEVR-XAI dataset and the benchmarking code can be found at <https://github.com/ahmedmagdiosman/clevr-xai>.

*Equal contribution

Email addresses: leila.arras@hhi.fraunhofer.de (Leila Arras), ahmed@osman.ai (Ahmed Osman), wojciech.samek@hhi.fraunhofer.de (Wojciech Samek)

Keywords: explainable AI, Evaluation, Benchmark, Convolutional Neural Network, Visual Question Answering, Computer Vision, Relation Network

1. Introduction

With the renaissance of neural networks in the last decade, the domains of application of deep learning have been continuously increasing. Indeed these models were shown to reach excellent performance on various large-scale prediction tasks, e.g., in computer vision [1], language understanding [2] or medical diagnosis [3]. At the same time, concerns were raised to whether such high performance is based on genuinely solving a given problem, or if it may partly rely on exploiting spurious correlations found in the data [4, 5, 6, 7]. The field of explainable AI (XAI) [8, 9] has recently developed various techniques to uncover the decision making process of the model and help unmasking such Clever Hans¹ predictors [7]. From an end-user perspective it might also be desirable or even legally required [10, 11], especially for safety-critical systems, to accompany a model's decision with an *explanation* in order to trace it back to the decisive parts of the input. Beyond that, XAI bears also the potential to help improve model performance and efficiency [12], or to enable new data-driven scientific discoveries [13, 14, 15].

In the vision domain, the explanation can take the form of a *heatmap*, where each pixel in an input image gets assigned a *relevance* value or score, indicating its relative contribution to the final decision. Methods providing such heatmaps in a deterministic and unambiguous way on an already trained neural network (so-called *post-hoc* explanation methods) include, amongst others, Class Saliency Map [16], Grad-CAM [17], Gradient \times Input [18], Integrated Gradients [19], Layer-wise Relevance Propagation [20], Excitation Backpropagation [21], Guided Backpropagation [22]. Another set of explanation methods

¹Clever Hans was a horse that was supposed to be able to perform simple calculations, but actually arrived at the correct solution by utilizing a spurious correlation, unintentionally reflected in the body language of the enquirer.

25 include some aspects of randomness in the heatmap computation: they require
generating additional perturbed training data samples or involve solving an *ad-*
hoc optimization problem to provide a single heatmap [23, 24, 25, 26, 27]. In
this work, we will mainly focus on the first set of methods², because they exploit
exactly the same amount of information as is available at prediction time, i.e.,
30 a single input data point and the trained neural network’s parameters. Typi-
cally such explanation methods implement a custom backward pass through the
network, or are based on gradient integration.

In previous works, visual explanations were mostly validated on real-world
image classification tasks, either through pixel perturbation analyses [28], or
35 by using the pixels’ relevances as an object detection signal [29]. While these
evaluations might be justified in a scenario where no ground truth explana-
tion is available for a given task, they could also potentially create a mismatch
between the explanation’s primary goal (explain the model’s *current* decision,
which might be also based on the image’s context or on dataset biases [4]), and
40 the evaluation criterion (track the *change* in model prediction when perturbing
pixels, or *localize* an object’s bounding box).

In the present work, we propose instead to evaluate explanations *directly*
against ground truth object coordinates using a restricted setup of synthetic,
albeit realistically rendered, images of 3D geometric shapes. To the best of
45 our knowledge, this is the first ground truth based and realistic testbed for
the evaluation of neural network explanations proposed in the literature. To
this end, we leverage a synthetic visual question answering (VQA) task, namely
the CLEVR³ diagnostic dataset [30], which was initially proposed to diagnose
the visual reasoning abilities of VQA models. We argue that VQA is a more
50 spatially grounded and selective setup for evaluating visual explanations than

²The only exception to this is what we additionally consider the SmoothGrad and VarGrad
methods, since they were often used in previous comparative studies.

³The acronym CLEVR stands for Compositional Language and Elementary Visual Rea-
soning.

single-object image classification, since it has more variability: the number of objects, their location and size vary across images, while in image classification the object of interest would often lie in the middle of the image and occupy a great portion of the image. Further, in VQA, not every object present in the image is relevant to the prediction. Indeed the explanation is *modulated* by the question which de-facto selects the important objects for a given prediction (i.e., the heatmap is not solely dependent on the image’s content), while in standard image classification tasks every image has a single relevant object which shall be recognized to make the final classification decision.

More precisely, our contributions to the evaluation of XAI methods can be summarized as follows:

- We create a new large-scale visual question answering dataset, denoted as CLEVR-XAI, made of questions and the corresponding pixel-level ground truth masks, which can be used as a new benchmark for evaluating visual explanations. Our dataset comprises 39,761 simple questions (CLEVR-XAI-simple) and 100,000 complex questions (CLEVR-XAI-complex).
- We propose two novel quantitative metrics suitable to evaluate visual explanations w.r.t. these ground truth masks: relevance mass accuracy and relevance rank accuracy. We expect our new metrics, together with our dataset⁴, will foster the evaluation of XAI methods on a well-defined common ground, which is crucially needed in the XAI research field [10].
- We systematically evaluate and compare ten different explanation methods on our new benchmark and metrics, using a trained Relation Network model [31] (a state-of-the-art model on the CLEVR task), leading to new findings regarding the quality and properties of XAI methods, which are sometimes contradicting with the conclusions drawn from previous comparative studies in vision XAI.

⁴Our dataset and code will be made publicly available upon paper acceptance at <https://github.com/ahmedmagdiosman/clevr-xai>.

Our work is divided as follows. Section 2 motivates our approach in relation to previous works. Section 3 introduces our new benchmark dataset, CLEVR-
80 XAI as well as our novel quantitative metrics for heatmap evaluation: relevance mass accuracy and relevance rank accuracy. This Section also details the relevance pooling techniques we considered in our work. Section 4 defines the XAI methods and their variants we tested in our empirical comparative study of visual explanations. In Section 5 and 6, we present our experiments and discuss
85 our results. Finally Section 7 concludes our work.

2. Previous Work on Evaluating Explanations

A widely used approach to evaluate explanations in the vision domain was initially introduced as pixel-flipping or region-perturbation analysis [20, 28] (variants of this analysis were also proposed in subsequent works [32, 33]). It
90 consists in repeatedly altering a sequence of pixels (or boxes of pixels) in an input image, accordingly to their *relevance* ordering, i.e., their importance for the prediction, and measure the effect of this perturbation on the model’s prediction. The higher the effect, measured for example in terms of prediction performance drop, the more accurate was the relevance. Additionally, we notice that
95 such evaluation ideally encompasses two baseline perturbation schemes: one random perturbation (corresponding to uninformative relevance), and a brute force search pixel-perturbation scheme which optimizes the measured performance metric (the latter representing the best case ordering of the relevance, which can also be seen as a valid explanation and is related to occlusion-based
100 relevance [34]). One potential issue with the pixel-perturbation evaluation is that the model might receive as input images that lie outside the actual training data distribution, i.e., out of data manifold samples. This can potentially lead to artifacts and unreliable model predictions. In contrast, our evaluation approach is based on unmodified input images, from the same distribution as
105 during model training.

Another perturbation-based evaluation technique consists in performing ran-

domizations of the model weights or training data [35] (in the latter case the model is re-trained on the same images but with randomly permuted class labels), and then measure the similarity between the original and the randomized explanations, in order to verify that the explanations are indeed sensitive to the model’s parameters and the data generating process. While this type of analyses can serve as an auxiliary sanity check for explanations, it doesn’t constitute a direct assessment of the explanation’s quality for a given real prediction. In a similar manner, further work proposes to measure the cosine similarity of explanations along the model’s lower layers, when initializing the relevance at a given higher layer randomly, in order to test if backward-propagation based explanations converge to a single direction throughout the propagation process, which would render these explanations effectively independent of higher layers [36]. Recent work from [37] also modifies the training or test data, or the model weights, and then compares the similarity of the resulting modified explanations with the original ones or some ground truth masks. Lastly, authors of [38] propose the Remove and Retrain evaluation method. Here a new training dataset is created by replacing the most important pixels in each image by an uninformative value (mean value per channel), accordingly to the pixels’ relevance ranking. The model is then re-trained on that modified data, and the prediction accuracy of the re-trained model is compared to a random pixel importance baseline, the higher the decrease w.r.t. that baseline the more accurate the explanation. However, a drawback of all previously mentioned perturbation-based evaluations is that they are based on input data, or on a model, that are different from the model of interest which is actually being deployed and explained at test time. In other words, they introduce a discrepancy between the original model’s training/test set configurations and the XAI evaluation setup, which can induce misleading or inaccurate results when comparing explanation methods.

A further commonly used approach for evaluating explanations in the visual domain is to use the pixel relevances for weakly supervised object localization (or segmentation) in real-world image classification tasks, e.g., by applying a

threshold on the relevances, and then computing the Intersection over Union (IoU), or related metrics (such as the per-pixel accuracy), w.r.t. object bound-
140 ing box annotations (or segmentation masks) as a measure of relevance accuracy [16]. A closely related evaluation consists in verifying that the single pixel with the highest relevance lies inside the object corresponding to the model’s predicted class, which is denoted as the pointing game accuracy [29]. As noticed in previous work a naive baseline pointing to the center of the image
145 already performs well on the latter type of evaluation [39]. More generally such localization-based evaluations could also be misleading, since they assume the model’s classification decision is based solely on the object itself, and not on its context or background, which on real-world image datasets can not be ensured. Indeed it was previously shown that state-of-the-art image classifiers fail to recognize objects in an unexpected location (e.g., a cow on a beach, rather than
150 on grassland), or conversely they identify nonexistent objects solely based on the image’s context (a lush hilly landscape is mistakenly recognized as a “herd of sheep”) [4, 40]. Moreover, the IoU type of metric favors a relevance distribution that fully covers the object’s surface in the image, while a trained classifier
155 might as well rely on parts of this area to make its decision. The two metrics we will introduce in Section 3 do not have this limitation and only assume the “major” part of the relevance lies on the object(s) of interest (in terms of mass or rank), but the relevances’ spread does not need to fully match the object boundaries.

160 Besides, most previous works evaluating XAI methods in computer vision consider as a model a convolutional neural network (CNN) which was trained for image classification on the ImageNet dataset [41]. In this task the goal is to classify images into one object category. However, as previously mentioned, such type of classification decision using real-world images can be based on various confounding features such as the object’s texture, the image’s background,
165 dataset biases or image generation artifacts [4, 7]. Therefore for this task evaluations based on object localization are not reliable. In contrast, we employ a *synthetic* generated dataset of 3D rendered shapes. This allows us to ensure

that the objects’ background does not contain any information: it is made of
170 the same uniformly colored plane surface for all images. In other words we can
assume the relevant information for the considered prediction task is on the
objects. Furthermore, we argue that VQA is a more *selective* task than image
classification for evaluating explanations. Indeed, for image classification the
model acts as a feature detector: it will process all relevant parts of the image
175 independently of the predicted class, up to the final linear layer that will in-
tegrate these features and assign a prediction score to each class. Hence, the
image processing pipeline is greatly identical for all classes (up to the output
layer): all important features in the image will be detected. Therefore it is
not surprising that the heatmaps for different classes may look very similar.
180 In contrast, in VQA, the hidden layer representations are modulated by the
question, and thus different questions on the same image will induce different
heatmaps. This makes VQA an ideal task to evaluate if an explanation can
select different objects in the image depending on the question, reflecting the
neural network’s internal processing of both the image and the specific ques-
185 tion. Lastly, the VQA model is naturally trained to deal with several objects
present in the input image at the same time, whereas on single-object image
classification tasks there is always one object to recognize and classify. Hence
the property of “class-discriminateness”, which is often stated as a desirable
property for an explanation method in the XAI literature (e.g., [42, 43, 39]),
190 meaning the ability to focus on the right object when *multiple* objects are present
in the image, might be a slightly overstated expectation in the general case of
a CNN trained on standard ImageNet single-object classification (if the CNN
was instead trained for multi-label classification, this behavior might however
be justified).

195 More particularly, our VQA-based evaluation benchmark is composed of
10,000 images, each containing 3 to 10 objects with different attributes (8 col-
ors, 2 materials, 2 sizes and 3 shapes) placed at random locations in the image.
Objects are also often situated near to one another, or occluded by one another:
this forces the model to focus on the objects themselves, and largely reduces

200 object detection strategies relying on the “negative space” around objects, e.g.,
to recognize the objects’ shape. This allows us to generate pixel-level ground
truth masks that depend on the target objects of each question (thus our an-
notations are more fine-grained than the bounding boxes used in weak object
localization). For each image we generate approx. 4 simple questions and 10
205 complex questions (more details on the ground truth generation for each type
of question will be given in Section 3).

In recent years such synthetic datasets were more commonly employed for the
fine-grained analysis of the concept learning and reasoning capabilities of neural
network models, in particular of VQA models [44, 45, 30, 46, 47, 48, 49, 50].
210 In contrast, we use a synthetic VQA task not for model evaluation, but for
evaluating the explanations of the model’s decisions. Further, the CLEVR task
was also used for evaluating the quality of attention heatmaps generated by
internal modules of a highly customized model [51]. Instead our aim is to
evaluate the input images’ explanations obtained through an XAI method, using
215 a general neural network architecture.

Closely related to our work is the evaluation approach taken in [52], which is
based on an image classification dataset of synthetic flowers with known discrim-
inative features (these are mainly composed of the petal’s color or the stem’s
color), and where the evaluation consists in calculating the IoU of the heatmaps
220 w.r.t. ground truth masks of the corresponding flower components. In their
considered task, all flowers have a unique size and each image contains a single
flower to classify. Hence our VQA setup has more variability. Relatedly, a recent
work proposed another synthetic dataset based on binary image classification
of abstract animals [53]. In this task the main discriminating feature is the
225 animal’s arm position, though the data generation process allows for the intro-
duction of other confounding features (such as the background’s color or the
animal’s position) in order to create controlled biases (this approach is similar,
e.g., to the Decoy MNIST dataset introduced in [54]). The intended purpose
is to serve as a basis for the evaluation of explanations (in particular human-
230 based), although thus far no comparative study in XAI was conducted using

this dataset.

Finally, other works relied on the human-based visual assessment of the explanation’s quality [19, 43, 22, 24, 23, 26, 37, 55]. While these studies give complementary insights about the usefulness of explanations from an end-user
235 perspective, and may contribute to an intuitive understanding of the properties of explanations, they can not replace an objective and systematic quantitative evaluation of XAI in computer vision. Further, when designing such human-based evaluations, one should be aware that a plausible explanation for a human may not necessarily reflect or be consistent with the model’s underlying decision
240 process (which we want to uncover through XAI) [56, 57], thus both must be taken apart carefully as far as possible.

Table 1 recapitulates the most commonly used approaches to quantitatively evaluate explanations in the vision domain, and contrasts their main properties with our proposed new approach.

245 In the next Section we will introduce our CLEVR-XAI benchmark dataset as well as our novel quantitative metrics for evaluating explanations in this testbed. Furthermore, we define the different relevance pooling techniques we considered in our work.

Table 1: Evaluations of visual explanations (aka heatmaps) most commonly used in the XAI literature. Through the task we denote the model’s training objective (and corresponding dataset). If the evaluation and model training are based on *synthetic* datasets, this enables a tighter control over biases and artefacts (which can’t be entirely controlled for in real-world collected images, or in human-generated annotations). We denote as *selective* a learning task where each image contains several objects of potential interest, but which are not all relevant for the prediction, and thus require the model to distinguish and select the relevant one(s). Lastly, we differentiate if the evaluation is based on the *same* model and *same* data as during model training, as opposed to evaluations that rely either on a modified model (due to re-training or parameter randomization), or on modified inputs, and that induce a discrepancy between the training/test and XAI evaluation configurations.

Evaluation of XAI	Task (dataset)	Properties				Proposed/used in
		synthetic	selective	same model	same data	
Pixel Perturbation		✗	✗	✓	✗	[20, 28, 58, 24, 25, 27]
Data Randomization Test		✗	✗	✗	✗	[35, 37]
Model Randomization Test	object classification	✗	✗	✗	✓	[35, 59, 36, 37]
Remove and Retrain	(ImageNet etc.)	✗	✗	✗	✗	[38]
Object Localization/Segmentation		✗	✗	✓	✓	[16, 29, 60, 42, 25]
Pointing Game		✗	✗	✓	✓	[29, 42, 25]
Ours: CLEVR-XAI	VQA (CLEVR)	✓	✓	✓	✓	

3. Ground Truth Evaluation of Heatmaps

250 3.1. A new Benchmark for Visual Explanations: CLEVR-XAI

Our new benchmark for evaluating explanations in the visual domain was built upon the original CLEVR dataset generator [30], we call it CLEVR-XAI.

CLEVR is a synthetic VQA task that was designed to diagnose the reasoning abilities of VQA models by avoiding the biases present in real-world
255 human-annotated datasets (such as [61, 62, 63]), and allowing full control on the data generation pipeline. The CLEVR dataset is comprised of 3 splits of training/validation/test sets with 70,000/15,000/15,000 images, and resp. 699,989/149,991/149,988 questions (i.e., there are approx. 10 questions per image), and the prediction problem is framed as a classification task with 28 possible
260 answers. CLEVR images contain 3D objects rendered under various lighting directions and positioned on a grey plane surface, each object having 4 types of attributes (color, material, size, and shape). The image generation process encompasses the creation of a *scene graph* which contains all necessary information to describe the visual scene such as object locations, attributes, and
265 inter-object relations. Further, the question generation is done via a *functional program* and various program templates. The functional program is made of a sequence of *basic functions* such as querying and comparing attributes, counting objects, and checking the existence of a certain object. Once the functional program is built, it can be applied on a scene graph to yield a ground truth
270 answer. Further, the program can be used to identify relevant objects along the question’s processing pipeline.

In our evaluation framework, a model is first trained on the original CLEVR training set. And subsequently, the explanations obtained on this trained model with various XAI methods are evaluated on the separate CLEVR-XAI evaluation
275 set, which is made of standard CLEVR-type questions that are augmented with the information from our generated ground truth masks (since those masks are not available in the original CLEVR dataset). Our CLEVR-XAI evaluation

set comprises approx. 140k questions⁵, divided into two types:

- 39,761 simple questions (CLEVR-XAI-simple): This subset contains only
280 simple queries about an object’s attributes with no inter-object relations.
Hence these questions have only a *single* target object that can be used
for the ground truth mask generation.
- 100,000 complex questions (CLEVR-XAI-complex): This subset contains
285 *all* types of questions as in the original CLEVR task [30], i.e., they include
spatial relations between objects as well as attribute comparison between
objects and object counting. Therefore *several* objects in the image can
be relevant for such questions and act as ground truths.

Both subsets share the same set of 10,000 images (i.e., for each image, there
are approx. 4 simple questions and 10 complex questions).

290 In the following we briefly describe the image generation process, which is
common to both subsets. Then, for each subset we will detail the question
generation process as well as the ground truth masks generation.

Image Generation. To generate the CLEVR-XAI images, we use the same
pipeline as in the CLEVR data generator [30]. Additionally, we use the scene
295 graph to create a segmentation mask of each object present in the scene (we
achieved this in practice by rendering a secondary image where the light sources
are deactivated and by assigning to each object a unique color). These segmen-
tation masks will serve as a basis for generating the questions’ ground truth
masks. In total we generated 10,000 images, where each image contains 3 to 10
300 objects with random attributes and locations.

⁵For comparison, the ImageNet image classification tasks contain 100k to 150k test images,
and the ImageNet object detection task contains 40k test images [41]. Hence our CLEVR-XAI
dataset can be considered as an evaluation at large scale.

3.2. CLEVR-XAI-simple

Question Generation. We generate CLEVR-XAI-simple questions in the same fashion as in the CLEVR data generator [30], except that for this subset we consider only specific question types where the answer can be inferred by examining a single object in the image. Our goal for this subset is indeed to consider the most restricted setup possible in order to remove any ambiguity in the ground truth mask generation. For instance, for the question “*What material is the tiny cyan sphere?*”, the relevant object of the question must be the only tiny cyan sphere in the scene. Hence we use only simple questions that query an object’s attribute, the corresponding question types are: `query_shape`, `query_color`, `query_size`, `query_material`. For each question type, we randomly sample a question template, and generate this way 4 questions per image. Finally ill-posed questions are discarded (as was done in [30]). This leaves us with a total of 39,761 questions.

Ground Truth Masks. Using the object segmentation masks created during the image generation step and each question’s functional program, we can automatically identify the single target object of each question. With this information, we generate two types of ground truth (GT) masks for evaluating explanations:

- *GT Single Object:* The first type of mask we generate is based solely on the target object’s pixels, which are marked as True, while remaining objects’ pixels, as well as the background, are set to False.
- *GT All Objects:* The second ground truth mask we generate is less discriminative and encompasses all objects’ pixels in the scene, those pixels are set to True, while only the scene’s background is set to False. This latter mask will allow us to perform a weak sanity check on the explanation to verify whether the relevance is indeed assigned to objects and not to the background, since in our synthetic task, by construction, the background shall be uninformative.

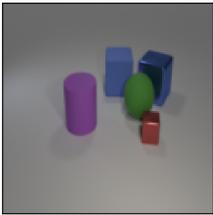
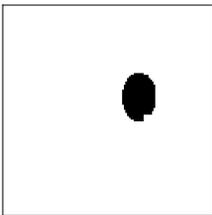
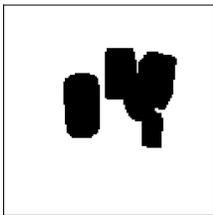
Image	Question/Answer	Program
	<p>What is the sphere made of?</p> <p><i>rubber</i></p>	<pre>[scene, filter_shape, unique, query_material]</pre>
Ground Truth	GT Single Object	GT All Objects
		

Figure 1: Example CLEVR-XAI-simple data point. We create two ground truths: one containing the single target object of the question, and one containing all objects.

Figure 1 provides an example CLEVR-XAI-simple question with corresponding ground truth masks.

3.3. CLEVR-XAI-complex

Question Generation. Unlike CLEVR-XAI-simple, CLEVR-XAI-complex questions are generated exactly like in the original CLEVR dataset [30]. We create in this way 10 questions per image, i.e., in total 100,000 questions. The resulting question distribution is approx. the same as in the original CLEVR dataset, i.e., grouped by question family we obtain: 36% *Query Attribute*, 18% *Compare Attribute*, 9% *Compare Integer* (these are questions that compare the sizes of objects' sets), 24% *Count* and 13% *Exist* questions. Hence, the CLEVR-XAI-complex subset of questions comprises far more complex questions than CLEVR-XAI-simple. This includes spatial relationships between objects, attribute comparison as well as logical operators (**and**, **or**) in the question formulation. The corresponding functional programs of the questions have a length between 2 and 24 (which means questions are made out of a sequence of up to

24 CLEVR basic functions). This drastically complicates the process of finding
345 relevant objects in the scene for building the corresponding ground truth masks.
Since there is no unique way how to define the relevant objects for such complex
questions, we create 3 different ground truth masks per question, each based on
different assumptions on how to process the question’s functional program.

As introduced earlier, the functional program of each question is made of
350 a sequence of basic functions (this sequence can further be chain- or tree-
structured). In this sequence each function takes as input the output of the
previous function. The first function in the program is always the scene (it
returns the entire set of objects present in the image), and the last function of
the program returns the true answer to the VQA question. Input and output of
355 each function can be attribute values, integer or boolean, but most importantly
they can also be sets of objects (for a detailed list of all basic functions, their
input and output, we refer the reader to the Supplementary Material in [30]).
We build our ground truths on the sets of objects returned by the intermediate
functions in the functional program.

360 *Ground Truth Masks.*

- *GT Unique:* This is the most discriminative ground truth mask we create.
It is based on the union of the output of all `unique` functions occurring
in the program. The function `unique` is used in the program when one
object has to be uniquely identified among a set of objects present in the
365 scene to answer the question. For instance, in the question “*How many
objects are left from the grey cube?*”, there must be a unique grey cube
in the image such that the question is not ill-posed. Such type of unique
objects are selected to build our *GT Unique* masks (note that a complex
question can have several unique objects, for instance when the question
370 contains a logical operator `and/or` in the program).
- *GT Unique First-non-empty:* To create this ground truth we iterate through
the functional program in reverse order and include to the ground truth

the first non-empty set of objects we find which is returned by any intermediate function in the program (we exclude the `scene` function which is the first function in the program since it is not discriminative and just returns all the objects of the scene). If the program is tree-structured (this can for instance occur when the function contains a logical operator), we aggregate the first non-empty sets from each of the two tree branches. Additionally, we incorporate to this ground truth all `unique` function outputs. In other words, the *GT Unique First-non-empty* set of objects is a superset of the *GT Unique*.

- *GT Union*: This is the least discriminative ground truth mask we create that is still related to the question. It simply includes the union of all sets of objects that are returned by every function present in the program (excluding the first function, which is the `scene` function and just returns all the objects in the scene). Per construction the *GT Union* is a superset of the *GT Unique First-non-empty*.
- *GT All Objects*: Like in CLEVR-XAI-simple, this mask contains all objects of the scene, and is independent of the question.

Figure 2 provides an example CLEVR-XAI-complex question with corresponding ground truth masks. The given question is chain-structured (for an example with a tree-structured question see Appendix A). For this question the output of the `unique` function is the small brown sphere, it serves to build the *GT Unique*. The *GT Unique First-non-empty* further includes the first non-empty set of objects returned by the functions processed in reversed order of the program, in this case the output of the function `same_size`: this corresponds to all other small objects present in the scene (the grey cube, the purple cylinder and the cyan cube). The *GT Union* further contains the output of all intermediate functions in the program, for the given question this means the objects returned by the function `filter_material` (which is a superset of the output of the second filter function `filter_shape`), i.e., all shiny objects in the scene (thus it additionally encompasses the large purple cylinder). Finally the *GT All*

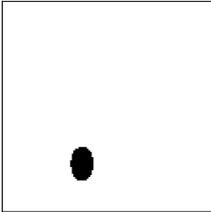
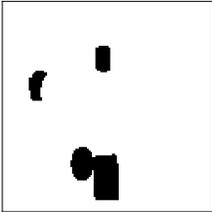
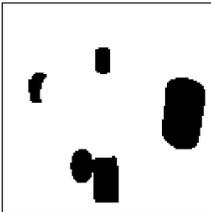
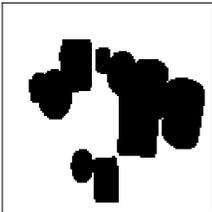
Image	Question/Answer	Program
	<p>Is there any other thing that has the same size as the shiny sphere?</p> <p><i>yes</i></p>	<pre>[scene, filter_material, filter_shape, unique, same_size, exist]</pre>
Ground Truth	GT Unique	GT Unique First-non-empty
		
	GT Union	GT All Objects
		

Figure 2: Example CLEVR-XAI-complex data point. The functional program is used to determine which objects in the scene are considered as ground truths.

Objects simply contains all objects. The example question further illustrates the fact that objects can be situated next to one another, and partially occluded.

405 In the Appendix B we provide further statistics on the CLEVR-XAI dataset (number of objects and number of pixels per ground truth mask).

3.4. Relevance Pooling

For visual explanations, evaluation w.r.t. ground truth masks requires the heatmap to be a 2D positive-valued image with a single channel. Indeed, we
410 care about the spatial location of the relevances rather than their channel (color) values. For most XAI methods, the original heatmap mirrors the shape of the

model’s input (i.e., 3 channel values for RGB images). Thus, there are a number of ways to *pool* the multiple channel values down to a single-channel heatmap. Since there is no consensus in the literature on how to perform this pooling step, 415 in our work we employ ten different pooling techniques. In the experimental Section 5, we will report the results with the pooling technique that led the highest performance for each XAI method and evaluation metric.

The ten pooling techniques we utilize are the following:

$$\text{sum,pos:} \quad R_{pool} = \max(0, \sum_{i=1}^C R_i)$$

$$\text{sum,abs:} \quad R_{pool} = \left| \sum_{i=1}^C R_i \right|$$

$$\text{l1-norm:} \quad R_{pool} = \sum_{i=1}^C |R_i|$$

$$\text{max-norm:} \quad R_{pool} = \max(|R_1|, |R_2|, \dots, |R_C|)$$

$$\text{l2-norm:} \quad R_{pool} = \sqrt{\sum_{i=1}^C R_i^2}$$

$$\text{l2-norm-sq:} \quad R_{pool} = \sum_{i=1}^C R_i^2$$

$$\text{pos,sum:} \quad R_{pool} = \sum_{i=1}^C \max(0, R_i)$$

$$\text{pos,max-norm:} \quad R_{pool} = \max(\max(0, R_1), \max(0, R_2), \dots, \max(0, R_C))$$

$$\text{pos,l2-norm:} \quad R_{pool} = \sqrt{\sum_{i=1}^C (\max(0, R_i))^2}$$

$$\text{pos,l2-norm-sq:} \quad R_{pool} = \sum_{i=1}^C (\max(0, R_i))^2$$

where R_{pool} is the pooled relevance at the current pixel, i is the channel index which starts from 1 to the number of channels C ($C = 3$ for RGB images) and 420 R_i is the relevance value at channel i . Note that the last 4 pooling techniques are just a variant of the first 6 pooling techniques where a ReLU preprocessing is applied to the relevance.

3.5. Evaluation Metrics: Relevance Mass Accuracy and Relevance Rank Accuracy

425 Given the already pooled heatmap, we propose two metrics to evaluate the accuracy of an explanation. These metrics only assume the “major” part of the relevance should lie inside the ground truth mask, either in terms of relevance mass or in terms of relevance ranking (hence our metrics do not require the heatmap to be binarized and to closely match the ground truth surface, as is
430 the case for example for metrics typically used for evaluations based on weak object localization). Both metrics deliver a number in the range $[0, 1]$, the higher this value the more accurate the relevance heatmap.

Relevance Mass Accuracy. The relevance mass accuracy is computed as the ratio of the sum of the relevance values lying within the ground truth mask over the sum of all relevance values over the entire image. In other words, it measures how much “mass” does the explanation method give to pixels within the ground truth. It can be written as:

$$\text{Mass Accuracy} = \frac{R_{within}}{R_{total}}, \quad (1)$$

$$\text{with } R_{within} = \sum_{\substack{k=1 \\ \text{s.t. } p_k \in GT}}^{|GT|} R_{p_k} \quad \text{and} \quad R_{total} = \sum_{k=1}^N R_{p_k} \quad (2)$$

where R_{p_k} is the relevance value at pixel p_k , GT is the set of pixel locations that lie within the ground truth mask, $|GT|$ is the number of pixels in this mask
435 and N is the total number of pixels in the image.

Relevance Rank Accuracy. The relevance rank accuracy measures how much of the high intensity relevances lie within the ground truth. It is calculated through the following steps. Let K be the size of the ground truth mask. Get the K highest relevance values. Then count how many of these values lie within the ground truth pixel locations, and divide by the size of the ground truth. Informally, this can be written as:

$$P_{top\ K} = \{p_1, p_2, \dots, p_K \mid R_{p_1} > R_{p_2} > \dots > R_{p_K}\} \quad (3)$$

where $P_{top\ K}$ is the set of pixels with relevance values $R_{p_1}, R_{p_2}, \dots, R_{p_K}$ sorted in decreasing order until the K -th pixel.

Then, the rank accuracy is computed as:

$$\text{Rank Accuracy} = \frac{|P_{top\ K} \cap GT|}{|GT|} \quad (4)$$

4. Explanation Methods

In this work we consider ten different explanation methods and evaluate them
440 against our CLEVR-XAI benchmark. In this Section we will briefly describe
these methods and detail the specific variants (including hyperparameters) for
each method we tested in our experiments. Note that in a practical XAI use-
case one doesn't have access to an evaluation metric to test an explanation's
quality. But since in this work we are interested in thoroughly comparing a
445 set of XAI methods, we allow for tuning⁶ those methods against our CLEVR-
XAI benchmark. More broadly, we believe that an "ideal" XAI method shall
not contain any free parameter and ought to be applicable in a straightforward
manner without requiring any tuning.

In our evaluation we mainly focus on *post-hoc* and deterministic XAI meth-
450 ods (as opposed to sampling-based or optimization-based methods), i.e., we em-
ploy methods that, given an input and a trained neural network, always deliver
the same explanation (the only exception to this is that we additionally con-
sider the sampling-based SmoothGrad and VarGrad methods, since they were
often used in previous works). Our XAI comparative study includes the same
455 set of explanation methods as were previously compared in [59, 35, 38], resp.
most of the methods that were compared in [32, 36, 37]. It additionally contains

⁶By tuning we mean that we consider alternative variants (resp. hyperparameters) choices for each XAI method, and ultimately report the results for the best performing variant w.r.t. our CLEVR-XAI benchmark. This is different to how tuning is usually done in machine learning, where a validation set is used for tuning the model's hyperparameters and the results are reported on a separate test set.

methods which were omitted in some of these prior works, such as Layer-wise Relevance Propagation (LRP) [20] or Grad-CAM/Guided Grad-CAM [42].

We now briefly review how each of these methods computes the pixels' 460 relevance values. To this end let us introduce some notations. We refer to $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ as the input image (with C channels, height H and width W), and to $x_i \in \mathbb{R}$ as a single input variable (i.e., x_i corresponds to a specific pixel location in horizontal and vertical direction and to a specific channel). Further we denote by $f_c(\cdot)$ the model's prediction function for some target class c (i.e., 465 the pre-softmax prediction score for class c). In our work we use as the target class the model's predicted answer.

Since our considered model is a VQA classifier, the prediction function $f_c(\cdot)$ takes as input both the image \mathbf{x} and the question \mathbf{q} . However, in order to simplify notation, we omit the dependence on the question, as in this work 470 we are only interested in the explanations obtained on the image-side. An explanation method delivers for each single input variable x_i a scalar *relevance* value $R_i \in \mathbb{R}$. These relevances are subsequently pooled along the channel axis (as was described in Section 3.4) to obtain a positive-valued heatmap of size $H \times W$, the latter finally serves as a basis for our quantitative evaluation using 475 the metrics from Section 3.5 and the CLEVR-XAI ground truth masks.

4.1. Gradient and Gradient×Input

One method to obtain the relevance R_i is based on the partial derivative of the model's prediction function w.r.t. the input variable [16], we denote it as Gradient:

$$R_i = \frac{\partial f_c}{\partial x_i}(\mathbf{x}) \quad (5)$$

A similar method further multiplies the partial derivative with the input variable's value [18], we denote it as Gradient×Input:

$$R_i = \frac{\partial f_c}{\partial x_i}(\mathbf{x}) \cdot x_i \quad (6)$$

The partial derivatives can be obtained in one gradient backward pass through the network. For both methods, we additionally consider a variant which is

based on the *squared* partial derivative [64, 65], instead of the simple partial
 480 derivative (i.e., $(\frac{\partial f_c}{\partial x_i}(\mathbf{x}))^2$, resp. $(\frac{\partial f_c}{\partial x_i}(\mathbf{x}))^2 \cdot x_i$, for Eq. 5 and 6).

4.2. SmoothGrad and VarGrad

Two other methods we consider are based on the mean or variance of the
 previous gradient-based relevance, when taking as input noisy versions of the
 input image. Defining by $g_i^{[k]}$ the gradient-based relevance of a noisy version k
 of the input image for variable x_i :

$$g_i^{[k]} = \frac{\partial f_c}{\partial x_i}(\mathbf{x} + (x_{\max} - x_{\min}) \cdot \mathcal{N}^{[k]}(0, \sigma^2)) \quad (7)$$

where x_{\max} and x_{\min} represent the extremal values of the original image’s single
 input variables, and $\mathcal{N}(0, \sigma^2)$ is the sampled additive gaussian noise with
 standard deviation σ (each sample being indexed by k).

Then, the SmoothGrad relevance [43] is computed as:

$$R_i = \frac{1}{n} \sum_{k=1}^n g_i^{[k]} \quad (8)$$

485 where n is the number of random samples.

We additionally consider three other variants of the SmoothGrad method,
 by taking into account a multiplication with the input variable’s value and/or
 squaring the gradient-based relevance (i.e., in Eq. 8, we use the following defini-
 tions of $\tilde{g}_i^{[k]}$ in place of $g_i^{[k]}$: $\tilde{g}_i^{[k]} = g_i^{[k]} \cdot x_i$, $\tilde{g}_i^{[k]} = (g_i^{[k]})^2$ and $\tilde{g}_i^{[k]} = (g_i^{[k]})^2 \cdot x_i$).
 490 Such squared variants of the gradient-based relevance were, e.g., used in [38].

The VarGrad method [59, 66] is similar to SmoothGrad, just that instead of
 a mean a variance is computed:

$$R_i = \frac{1}{n} \sum_{k=1}^n (g_i^{[k]} - (\frac{1}{n} \sum_{k=1}^n g_i^{[k]}))^2 \quad (9)$$

Both SmoothGrad and VarGrad have two hyperparameters: the gaussian
 noise’s standard deviation σ and the number of samples n . We tried the
 following values in our experiments: $\sigma = [0.02, 0.05, 0.10, 0.20, 0.30]$ and $n =$
 $[20, 50, 100, 300]$. (The SmoothGrad authors recommend to use σ between 0.10
 495 and 0.20 and $n = 50$, according to their qualitative experiments [43].)

4.3. Deconvnet and Guided Backpropagation

Deconvnet and Guided Backpropagation are two methods that modify the standard gradient backward pass through the neural network’s ReLU layers.

Given a hidden layer neuron which is ReLU activated in the forward pass and defined as $x_j^{l+1} = \text{ReLU}(x_j^l) = \max(0, x_j^l)$ where j stands for the neuron’s index and l for the layer’s index. Then in the standard gradient backward pass, given R_j^{l+1} the gradient at the output of the ReLU layer, the gradient at the input is computed as:

$$R_j^l = \mathbb{1}_{x_j^l > 0} \cdot R_j^{l+1} \quad (10)$$

Instead in Deconvnet [34] the gradient at the input of the ReLU layer is computed as:

$$R_j^l = \text{ReLU}(R_j^{l+1}) \quad (11)$$

And in Guided Backpropagation [22] it is computed as:

$$R_j^l = \text{ReLU}(\mathbb{1}_{x_j^l > 0} \cdot R_j^{l+1}) \quad (12)$$

The latter method was also re-invented in [60] under the name of DeSaliNet.

500 For both methods we tried two variants: applying the modified ReLU gradient backpropagation only to ReLU layers within convolutional layers, or applying it to all ReLUs present in the network.

4.4. Grad-CAM and Guided Grad-CAM

Grad-CAM [42, 17] is a method derived from the standard gradient-based 505 relevance of the last convolutional layer in the network.

Following the authors’ notation, denoting by A^k the features maps of dimension $h \times w$ at the last convolutional layer (k being the feature map’s index, and K the total number of feature maps). First the feature map’s gradient-based relevance is global-average-pooled to obtain a feature map’s importance weight α_k , i.e.:

$$\alpha_k = \frac{1}{h \times w} \sum_{i=1}^h \sum_{j=1}^w \frac{\partial f_c}{\partial A_{ij}^k}(\mathbf{x}) \quad (13)$$

Then the feature maps are linearly combined and passed through a ReLU processing to finally obtain the Grad-CAM heatmap H of dimension $h \times w$:

$$H = \text{ReLU}(\sum_{k=1}^K \alpha_k \cdot A^k) \quad (14)$$

To obtain a heatmap of the same dimension as the input image, the Grad-CAM heatmap is upsampled through bilinear interpolation. Note that the resulting heatmap is already 2D and positive (thus no need to apply any further pooling technique from Section 3.4 when using this explanation method). We further observe that Grad-CAM can be seen as a variation of Gradient \times Input, which is computed in the last convolutional layer, and where the gradient is first averaged in spatial location.

In our considered model (see Section 5.1), the last convolutional layer has the following structure: `conv` \rightarrow `relu` \rightarrow `batchnorm`. In our experiments we tried using each one of these layers for the Grad-CAM computation (the Grad-CAM authors used models with the following structure: `conv` \rightarrow `relu` \rightarrow `maxpool`, and implemented Grad-CAM on the `relu` layer). Further we found that for some data points the resulting Grad-CAM heatmaps are zero-valued, i.e., uninformative, we discard these data points from our analysis.

Additionally the authors of [42, 17] proposed the Guided Grad-CAM explanation method: it is simply as element-wise multiplication of the Grad-CAM and Guided Backpropagation input image relevances.

4.5. Integrated Gradients (IG)

Another method we consider is Integrated Gradients (IG) [19]. The relevances R_i of the input variables x_i are based on approximating the following integral through a Riemann sum:

$$R_i = (x_i - x'_i) \cdot \int_{\alpha=0}^1 \frac{\partial f_c(\mathbf{x}' + \alpha \cdot (\mathbf{x} - \mathbf{x}'))}{\partial x_i} d\alpha, \quad (15)$$

where \mathbf{x}' is a *baseline* image that needs to be chosen when applying the method. Ideally it should be an image with near-zero prediction score and containing no signal. The IG authors used a zero-valued image, i.e., a black image, for that

purpose [19]. In our experiments we also used a zero-valued baseline. Further we experimented with two other baselines: the mean image, as well as the mean channel values over the flattened images (which we computed over the CLEVR training set). Since the CLEVR images contain a grey background, this resulted in two grey-valued baselines.

Another choice to be made when employing the IG method is the number of steps used for integral approximation. We tried the values [300, 1000, 3000, 10000, 30000] for each image separately until we reached a sufficient precision for the integral approximation, and we used the Riemann sum with midpoint rule for numerical integration. Indeed one can compute the precision of the integral approximation by exploiting the IG’s completeness property:

$$\sum_i R_i = f_c(\mathbf{x}) - f_c(\mathbf{x}') \quad (16)$$

In our experiments we ensure that the resulting relative error is lower than 0.01 and discard from our analysis data points with higher errors (the original authors suggest a value of 0.05).

4.6. Layer-wise Relevance Propagation (LRP)

Another method we consider is Layer-wise Relevance Propagation (LRP). It was initially introduced in [20], and later justified via Deep Taylor Decomposition in [67, 68]. It consists in redistributing the model’s prediction score $f_c(\mathbf{x})$ via a custom backward pass through the network that follows a local conservation principle. During this backward pass each neuron in the network gets assigned its own *relevance* value, up to the input layer neurons (i.e., the input variables).

In practice, through ReLU layers the relevance is backpropagated as the identity, while weighted linear connections (such as convolutional and fully-connected layers) serve to redistribute the relevance in *proportion* to neuron contributions from the forward pass. More precisely, given a linear layer with the forward pass equation $z_j = \sum_i z_i \cdot w_{ij} + b_j$, then two main LRP rules can

be used⁷ to compute the neuron’s relevance R_i , given the relevances of the connected higher-layer neurons R_j . One is the ϵ -rule:

$$R_i = \sum_j \frac{z_i \cdot w_{ij}}{z_j + \epsilon \cdot \text{sign}(z_j)} \cdot R_j, \quad (17)$$

where ϵ is a stabilizer (we use $\epsilon = 0.001$). The other is the α_1, β_0 -rule:

$$R_i = \sum_j \frac{(z_i \cdot w_{ij})^+}{\sum_k (z_k \cdot w_{kj})^+ + (b_j)^+} \cdot R_j, \quad (18)$$

where $(\cdot)^+$ denotes the $\max(0, \cdot)$ operation. Note that when using the latter LRP propagation rule in all layers of the network LRP is equivalent to Excitation
 545 Backpropagation [29, 21].

We experimented with different LRP rules configurations: either using each rule uniformly in all layers of the network, or a combination of both rules. In the latter case we employed the ϵ -rule in the classifier layers of the network (i.e., the last 3 layers of the network, see model description in Section 5.1), or
 550 else in all fully-connected layers (i.e., including the Relation Network module of the network), applying the α_1, β_0 -rule in the remaining layers. Additionally, we considered using a specific LRP rule for the input layer (namely the box-rule, the w^2 -rule and the flat-rule, for a definition of these rules we refer to [69]), as well as subsuming consecutive convolutional/fully-connected and batchnorm
 555 layers into a single linear layer before applying LRP (according to a recommendation in [69]), otherwise we used for batchnorm layers the same rule as in the convolutional layers.

5. Experiments

5.1. Model

560 The model we consider for the evaluation of XAI methods is a Relation Network (RN) model [31].

⁷More recently the LRP γ -rule was introduced [69]. We did not experiment with this rule since it contains an hyperparameter γ which, in a general use-case, one doesn’t know how to choose/tune when there is no evaluation criterion available for the explanation.

There are two reasons why we choose this model: i) it is a state-of-the-art model on the CLEVR dataset [30], ii) it is a simple architecture made of very common layers (convolutional layers, batchnorm layers, fully-connected
565 layers, element-wise summation, ReLU activations, plus an LSTM for processing the question). The latter is very important to not obfuscate the evaluation of explanation methods by the complexity of the neural network model, and first establish the suitability of XAI methods w.r.t. ground truths on widely used and standard neural network components.

570 More precisely, the RN model comprises a 4-layer CNN to extract feature maps from the image. Then, the pixels in the last convolutional feature maps are pair-wise concatenated with the question representation obtained from the last hidden state of an LSTM. These representations are passed through a 4-layer MLP of fully-connected layers, summed-up in place, and finally fed to a
575 3-layer MLP of fully-connected layers for classification.

Since the original authors [31] did not release their code, we re-implement their model and train it from scratch on the original CLEVR training set [30], using the validation set for hyperparameter tuning. As an image preprocessing step, we rescaled the image values to the range $[0, 1]$ (we did not center the data
580 since the RN model contains batchnorm layers), further our input images have 3 channels (RGB). For more details on the model specification and training we refer to the Appendix C. Note also that since the RN model takes input images of size 128×128 , while the original CLEVR images have size 320×480 , we need to resize our Ground Truth masks to this former size too, see the Appendix D
585 for more details on this step.

Our trained RN model reaches 93.3% accuracy on the CLEVR test set (the original authors report 95.5% accuracy).

On the CLEVR-XAI-simple questions it reaches 98.2% accuracy, and on CLEVR-XAI-complex 93.3%. In the latter case the performance differs per
590 question type: for questions involving counting (`count`, `equal_integer`, `greater_than`, `less_than`) the accuracy is lower, between 72.2% and 93.0%, while on the remaining question types (`equal_color`, `equal_material`, `equal_shape`, `equal_size`,

query_color, query_material, query_shape, query_size, exist) the accuracy is higher, between 94.2% and 98.3% (which is consistent with results from[31]).

595 We now move to the evaluation of XAI methods on the CLEVR-XAI-simple and the CLEVR-XAI-complex questions, respectively.

5.2. Simple Questions: CLEVR-XAI-simple

Tuning XAI Methods. In our evaluation paradigm we start by tuning each XAI method, for each type of relevance accuracy metric separately, using the *GT*
600 *Single Object*. By tuning we mean selecting for each XAI method the variant/hyperparameter choice, as well as the pooling technique, that lead to the highest relevance accuracy. The resulting method’s best variant and pooling technique will further be applied on the CLEVR-XAI-complex questions and other ground truth masks to compare XAI methods.

605 We employ the *GT Single Object* for tuning since it’s the most restricted and controlled setup among all questions and ground truth masks we created. Indeed here we consider only the case of simple queries about an object’s attribute. And, in this minimal configuration, we might reasonably assume that a good explanation method shall focus its relevance on the single target object of the
610 question which must have triggered the model’s response.

For this tuning step, as well as for the remaining experiments, we only use data points which are correctly classified by the model in order to avoid noise in the evaluation stemming from false predictions, and we consider all pooling techniques introduced in Section 3.4 (i.e., we allow each XAI method to rely on
615 a different pooling).

The resulting relevance accuracies are reported in Table 2 and 3, left column (note that we list the XAI methods in the same order for all tables in this work: this order generally reflects the relative quality of the XAI methods across all our experiments). A detailed listing of the corresponding XAI methods
620 variants/hyperparameters and pooling techniques can be found in Appendix E.

A nice finding of this tuning step is that, for every XAI method, tuning with either *mass* accuracy or *rank* accuracy leads to the same best vari-

ant/hyperparameter choice for each method⁸. Also the relative ordering of the XAI methods remains largely consistent across the two accuracy metrics. This
625 highlights the fact that both metrics are well-suited to measure the quality of an explanation method. Concerning the most adequate pooling technique per accuracy metric, we find the l2-norm-sq and pos,l2-norm-sq poolings to be the most appropriate poolings for the mass accuracy metric, and the max-norm pooling stood out as the appropriate pooling across all XAI methods for the
630 rank accuracy metric.

In the following we briefly review for each XAI method which of its variant led the best relevance accuracy, and in the next subsection, we will compare the relative performance of XAI methods.

For LRP applying the α_1, β_0 -rule in every hidden layer and the box-rule in
635 the input layer led to best results. Note that this setup is conform to the configuration brought forward by the Deep Taylor Decomposition framework [67, 68]. Although more recently a composite application of the LRP rules was advocated in the particular case of CNNs for image classification for discriminativity reasons (according to [69, 70]), in our VQA setup this composite application of the
640 α_1, β_0 -rule and ϵ -rule, where the latter rule is used on the classifier layers of the network (i.e., the last three layers) in addition to using the box-rule in the input layer, either led to slightly inferior results or results on-par with the best setup mentioned earlier, while applying the ϵ -rule in all fully-connected layers and the α_1, β_0 -rule in all convolutional layers consistently lead to worse results. More-
645 over, using the ϵ -rule uniformly in all layers performed very poorly, although this rule was shown to perform well in other contexts, such as CNNs for text classification or recurrent neural networks for sequential data [71, 72, 73, 74], which is probably due to the specific nature and spatial structure of neural network

⁸Except for some gradient-based relevance techniques where using the unsquared gradient sometimes lead to better results for the rank accuracy metric, and for the Grad-CAM method where the used layer for the relevance computation differs per accuracy metric. See Appendix E for more details.

models in computer vision compared to other data domains. Lastly, we didn't
650 find that combining the batchnorm layers with their nearest convolutional/fully-
connected layer before applying LRP led to an improvement in the relevance's
accuracy.

Further, we note that simply applying the LRP α_1, β_0 -rule in all layers of
the network, including the input layer, already performs very well: this is the
655 result we report under Excitation Backprop [29] in our tables. This particular
LRP setup was also shown to perform very well in the context of neural network
pruning for image classification CNNs [12].

For IG the baseline image with mean channel values led to the best results,
nearly followed by the mean image baseline. However, using a zero-valued base-
660 line (as was done by the original authors in [19]) leads to very poor results: the
resulting mass and rank accuracy are only 0.27 in this case. This highlights the
crucial importance of choosing an appropriate baseline for the IG method (for
a general discussion on the impact of the baseline choice we refer to [75]).

Guided Backprop performed around 0.10 accuracy better when applying the
665 modified backpropagation rule to all ReLU layers in the network, rather than
only to ReLUs within convolutional layers. For Deconvnet that's the oppo-
site: applying the modification only within the CNN part of the network led to
approx. 0.15 better accuracy (in the other case the accuracy is near zero).

Grad-CAM alone performs generally very poorly (and a bit worse when using
670 the `conv` layer for its computation, rather than the `relu` or `batchnorm` layer).
When augmented with Guided Backprop the Grad-CAM performance is highly
boosted: Guided Grad-CAM reached its best performance when Grad-CAM
was computed on the `relu` layer, and the Guided Backprop's backpropagation
rule was applied to all ReLUs in the network.

675 SmoothGrad and VarGrad both reached their highest performance when the
number of samples was set to 300 and the noise level to 0.05. If we had taken
instead no more than 100 samples and a noise level of at least 0.10 (as suggested
by the original authors in [43]), then these methods would have lost at least 0.10
mass accuracy and 0.08 rank accuracy, rendering them not much better than

680 the simpler Gradient and Gradient×Input methods. This illustrates the high sensitivity of the sampling-based XAI methods to the specific hyperparameter choices.

Comparing XAI Methods. Now when comparing the XAI methods to one another using Table 2 and 3, left column, as well as the results in subsequent
685 tables (which will be more detailed in the remainder of this work) we find that Grad-CAM and Deconvnet are by far the worst performing methods. Guided Grad-CAM performs relatively well, but this is likely mainly due to the benefit added by its Guided Backprop component. That Grad-CAM performs even worse than Gradient×Input is not really surprising, since the Grad-CAM’s internal computation resembles an “averaged Gradient” times Input in the CNN’s
690 last layer feature maps: the gradient information is diluted through the global average pooling operation, leading to loss of precision in spatial location.

Gradient×Input performs slightly worse than Gradient. In general we expect the former to be very sensitive to the specific image preprocessing: if we had
695 instead center our image data to have a mean of zero (in our case we merely rescaled the image values to the range $[0,1]$), then the images’ background would have a near zero value, and consequently the Gradient×Input relevance would have been near zero in these image regions too, which would have probably boosted the Gradient×Input’s performance.

700 SmoothGrad and VarGrad are a bit better than their unsmoothed counterparts Gradient and Gradient×Input, however this is due to hyperparameter finetuning, and in a practical application it would be unclear how to choose/tune those hyperparameters. Further we observe that VarGrad is generally no better than SmoothGrad.

705 Two methods that perform remarkably well in our evaluation, although their computation is very efficient (a single backward pass through the network) are LRP and Guided Backprop, though LRP generally outperforms the latter.

IG also performs well, however it’s computation is more expensive (multiple backward passes are required), and in general, it remains an open question how

710 to choose a good baseline image for this XAI method. Note also that if we
compare the IG and LRP results on the most selective setups we considered,
i.e., the *GT Single Object* on CLEVR-XAI-simple and the *GT Unique* and *GT*
Unique First-non-empty on CLEVR-XAI-complex, then the LRP method is at
least 0.11 better in mass accuracy, and at least 0.08 better in rank accuracy,
715 than IG.

Now if we look at the variances of the accuracy results, and if we exclude from
this analysis the Deconvnet and the Grad-CAM methods (since these methods
anyway perform poorly), then the two methods with the lowest variance are
LRP and Guided Backprop, while the gradient-based methods show the highest
720 variances. IG, SmoothGrad, VarGrad, Gradient and Gradient \times Input show es-
pecially high variance in terms of relevance mass accuracy, while SmoothGrad
has particularly high variance on the relevance rank accuracy. The only excep-
tion to this behavior for the IG method is when we look at the *GT All Objects*
results in Table 6-7 and 8-9: here LRP and IG both have the lowest variance,
725 and also perform on-par in terms of relevance accuracy mean and median.

More generally, we observe that for the mass metric the variances are higher
than for the rank metric. This is to be expected since for the rank metric
only the pixels' relevance *ordering* matters, and if a few outliers exist in the
relevance heatmap with very high values which are located outside the ground
730 truth mask, this metric will be less affected by those noisy values. In this
way the rank accuracy would still reflect the global ordering pattern of the
relevance distribution. On the contrary the mass accuracy will be more sensitive
to represent the position of a few pixels with very high relevance values. In this
sense both accuracy metrics are complementary.

735 Additionally to the relevance accuracy results presented in this Section, in
the Appendix F we provide some further qualitative results, namely example
heatmaps for a few data points from our CLEVR-XAI dataset.

Impact of Model Confidence and Question Difficulty. We explored whether the
relevance accuracy of an XAI method depends on the model's prediction confi-

740 dence, and the difficulty of the considered VQA question.

In the first case we selected data points for which the model’s predicted softmax probability is very high (higher than 0.99999), this gives us the results in Table 2 and 3, middle column.

745 For the second analysis we retained only questions for which the *GT Single Object* mask encompasses at least 1000 pixels. This excluded objects of the category *small* (in the CLEVR task objects have two possible size attributes: large or small). With this restriction we further seek to exclude *large* objects that may be occluded, which would render the corresponding question more difficult to answer (the CLEVR scenes indeed often contain occluded objects).
750 Our results for this setup are reported in Table 2 and 3, right column.

We observe that in both cases the relevance accuracy generally tends to increase (i.e., moving from the left to the middle column, or from the left to right column increases the accuracy’s mean and median), while the accuracy’s standard deviation decreases. This demonstrates that the relevance accuracy
755 is not a static value that only depends on the XAI method’s quality, but also reflects the model’s confidence in the prediction or the difficulty of a given question.

Validating our Evaluation Approach. So far in our evaluation we have assumed that the “major” part of the relevance shall lie within an object for a VQA
760 question about that object. This assumption may potentially be questioned or challenged. Or at least, conversely, there is one type of questions where the model’s predicted answer can only be *found* on the object itself: these are queries about an object’s color. If the expected role of an XAI method is to highlight the image region that triggered the model’s answer, then we can safely
765 assume the relevance to focus on the *GT Single Object* mask for such type of questions.

To analyze this particular use case we conducted an experiment where this time we tune the XAI methods on the *GT Single Object* using correctly predicted data points for which the question type is restricted to be `query_color`. This

770 gives us the results in Table 4 and 5 (which are to be compared with the left
column of Table 2 and 3).

We find that this restricted setup generally doesn't affect the best variant/hyperparameter choice and pooling technique for each XAI method⁹, and also doesn't modify much the absolute accuracy results (e.g., for the three best
775 performing XAI methods, the accuracy increase is no more than 0.02).

This experiment demonstrates that the VQA model mainly focuses on the target object of the question, no matter which type of question is being asked, and validates our prior assumption that the target objects' of the questions can be utilized as ground truth masks to evaluate XAI methods.

780 *Weak Sanity Check.* Now that we have evaluated and compared the XAI methods on the *GT Single Object*, we report their performance on the *GT All Objects* masks. We recall that the latter masks are not selective and independent of the question: they simply contain all the objects in the scene. By using the latter GT we can perform a weak sanity check on the XAI methods: how much relevance lies on the objects at all vs. the background. Given that our VQA task
785 is controlled, the background shall be uninformative to the prediction and consequently very little relevance shall be assigned to it. Our results are reported in Table 6 and 7.

For this experiment, IG, Excitation Backprop and LRP perform best: these
790 methods assign more than 90% of the relevance's mass on objects, and more than 72% of the most relevant pixels lie within objects, on average. Further, for these methods, the standard deviation is more than halved compared to the *GT Single Object* (Table 2 and 3, left column). This indicates that for these

⁹The only noticeable difference we observe is for the LRP method: this time the composite application of the α_1, β_0 -rule and ϵ -rule (where the latter is applied to the classifier layers of the network, i.e., the last three layers) performs *on-par* with a uniform application of the α_1, β_0 -rule on all hidden layers, in both cases using the box-rule for the input layer. While when considering all question types the composite configuration performed slightly worse than the uniform application of the α_1, β_0 -rule.

Table 2: Relevance *mass* accuracy on **CLEVR-XAI-simple**, using the single target object of the question as Ground Truth, i.e., *GT Single Object*. We consider 3 sets of questions: all correctly predicted, correctly predicted with high confidence (softmax proba > 0.99999), and correctly predicted with large target object (greater than 1000 pixels).

Relevance Mass Accuracy	correctly predicted		correctly predicted		correctly predicted	
	all		& proba>0.99999		& nb pixels>1000	
	mean (std)	median	mean (std)	median	mean (std)	median
LRP [20]	0.85 (0.17)	0.91	0.90 (0.09)	0.93	0.91 (0.09)	0.94
Excitation Backprop [29]	0.80 (0.20)	0.87	0.85 (0.14)	0.90	0.88 (0.13)	0.92
IG [19]	0.71 (0.27)	0.81	0.75 (0.25)	0.85	0.80 (0.24)	0.90
Guided Backprop [22]	0.58 (0.20)	0.62	0.63 (0.16)	0.66	0.76 (0.13)	0.78
Guided Grad-CAM [42]	0.58 (0.24)	0.63	0.64 (0.21)	0.68	0.83 (0.13)	0.86
SmoothGrad [43]	0.60 (0.33)	0.69	0.61 (0.33)	0.72	0.80 (0.25)	0.91
VarGrad [59]	0.58 (0.34)	0.68	0.60 (0.34)	0.71	0.80 (0.25)	0.91
Gradient [16]	0.49 (0.35)	0.49	0.51 (0.34)	0.53	0.82 (0.25)	0.93
Gradient×Input [18]	0.43 (0.34)	0.37	0.44 (0.34)	0.39	0.77 (0.27)	0.89
Deconvnet [34]	0.18 (0.16)	0.13	0.18 (0.16)	0.13	0.42 (0.21)	0.41
Grad-CAM [42]	0.09 (0.10)	0.05	0.10 (0.10)	0.07	0.29 (0.10)	0.28

Table 3: Relevance *rank* accuracy on **CLEVR-XAI-simple**, using the single target object of the question as Ground Truth, i.e., *GT Single Object*. We consider 3 sets of questions: all correctly predicted, correctly predicted with high confidence (softmax proba > 0.99999), and correctly predicted with large target object (greater than 1000 pixels).

Relevance Rank Accuracy	correctly predicted		correctly predicted		correctly predicted	
	all		& proba>0.99999		& nb pixels>1000	
	mean (std)	median	mean (std)	median	mean (std)	median
LRP [20]	0.72 (0.15)	0.75	0.76 (0.10)	0.79	0.73 (0.11)	0.75
Excitation Backprop [29]	0.69 (0.18)	0.74	0.74 (0.13)	0.78	0.73 (0.13)	0.76
IG [19]	0.53 (0.18)	0.55	0.56 (0.17)	0.59	0.65 (0.13)	0.67
Guided Backprop [22]	0.53 (0.14)	0.55	0.57 (0.11)	0.58	0.63 (0.09)	0.64
Guided Grad-CAM [42]	0.52 (0.17)	0.55	0.56 (0.14)	0.59	0.68 (0.10)	0.69
SmoothGrad [43]	0.49 (0.19)	0.53	0.51 (0.18)	0.56	0.58 (0.15)	0.61
VarGrad [59]	0.46 (0.24)	0.51	0.47 (0.24)	0.53	0.64 (0.16)	0.67
Gradient [16]	0.34 (0.18)	0.34	0.35 (0.18)	0.36	0.55 (0.12)	0.56
Gradient×Input [18]	0.31 (0.17)	0.30	0.31 (0.17)	0.31	0.52 (0.12)	0.53
Deconvnet [34]	0.21 (0.15)	0.18	0.21 (0.15)	0.18	0.37 (0.15)	0.38
Grad-CAM [42]	0.17 (0.23)	0.01	0.19 (0.24)	0.06	0.55 (0.20)	0.58

Table 4: Relevance *mass* accuracy on **CLEVR-XAI-simple**, using the single target object of the question as Ground Truth, i.e., *GT Single Object*. Here we consider only questions that are correctly predicted and query an object’s color.

Relevance Mass Accuracy	correctly predicted & query color	
	mean (std)	median
LRP [20]	0.87 (0.15)	0.92
Excitation Backprop [29]	0.82 (0.18)	0.89
IG [19]	0.70 (0.28)	0.79
Guided Backprop [22]	0.64 (0.16)	0.67
Guided Grad-CAM [42]	0.60 (0.24)	0.66
SmoothGrad [43]	0.55 (0.35)	0.64
VarGrad [59]	0.54 (0.35)	0.63
Gradient [16]	0.49 (0.35)	0.48
Gradient×Input [18]	0.42 (0.35)	0.35
Deconvnet [34]	0.24 (0.18)	0.20
Grad-CAM [42]	0.09 (0.10)	0.05

Table 5: Relevance *rank* accuracy on **CLEVR-XAI-simple**, using the single target object of the question as Ground Truth, i.e., *GT Single Object*. Here we consider only questions that are correctly predicted and query an object’s color.

Relevance Rank Accuracy	correctly predicted & query color	
	mean (std)	median
LRP [20]	0.73 (0.13)	0.76
Excitation Backprop [29]	0.71 (0.16)	0.76
IG [19]	0.54 (0.18)	0.55
Guided Backprop [22]	0.56 (0.11)	0.58
Guided Grad-CAM [42]	0.53 (0.18)	0.57
SmoothGrad [43]	0.49 (0.20)	0.53
VarGrad [59]	0.43 (0.24)	0.48
Gradient [16]	0.37 (0.18)	0.37
Gradient×Input [18]	0.33 (0.18)	0.33
Deconvnet [34]	0.26 (0.16)	0.25
Grad-CAM [42]	0.16 (0.23)	0.00

Table 6: Relevance *mass* accuracy on **CLEVR-XAI-simple**, using all objects in the image as Ground Truth, i.e., *GT All Objects*. We consider all correctly predicted questions.

Relevance Mass Accuracy	all correctly predicted	
	GT All Objects	
	mean (std)	median
LRP [20]	0.96 (0.03)	0.97
Excitation Backprop [29]	0.90 (0.10)	0.94
IG [19]	0.97 (0.03)	0.98
Guided Backprop [22]	0.72 (0.12)	0.74
Guided Grad-CAM [42]	0.73 (0.14)	0.74
SmoothGrad [43]	0.73 (0.25)	0.81
VarGrad [59]	0.72 (0.26)	0.80
Gradient [16]	0.66 (0.29)	0.73
Gradient×Input [18]	0.56 (0.31)	0.58
Deconvnet [34]	0.29 (0.17)	0.26
Grad-CAM [42]	0.32 (0.11)	0.33

Table 7: Relevance *rank* accuracy on **CLEVR-XAI-simple**, using all objects in the image as Ground Truth, i.e., *GT All Objects*. We consider all correctly predicted questions.

Relevance Rank Accuracy	all correctly predicted	
	GT All Objects	
	mean (std)	median
LRP [20]	0.74 (0.08)	0.75
Excitation Backprop [29]	0.76 (0.08)	0.77
IG [19]	0.72 (0.07)	0.72
Guided Backprop [22]	0.48 (0.09)	0.48
Guided Grad-CAM [42]	0.49 (0.10)	0.49
SmoothGrad [43]	0.34 (0.12)	0.32
VarGrad [59]	0.29 (0.15)	0.26
Gradient [16]	0.36 (0.09)	0.36
Gradient×Input [18]	0.33 (0.09)	0.33
Deconvnet [34]	0.23 (0.09)	0.22
Grad-CAM [42]	0.39 (0.15)	0.40

methods, when the relevance is not located on the target object of the question,
795 then it is mainly situated on other objects and not on the background. There is
also a large gap between these methods and the remaining XAI methods: other
methods present a mean mass accuracy of at most 0.73 and mean rank accuracy
of at most 0.49. Finally, overall, the relative ordering of the XAI methods w.r.t.
the accuracies on this weaker sanity check remains largely unchanged.

800 5.3. Complex Questions: CLEVR-XAI-complex

Using for each XAI method the best performing variant as was previously
tuned on the simple questions *GT Single Object*, we report in this Section their
accuracy on the different Ground Truths masks we generated for complex ques-
tions. Further, we consider only questions that are correctly predicted and that
805 do not involve counting (since the VQA model was shown to perform less good
in terms of prediction performance on these types of questions, see 5.1) in order
to avoid noise in the evaluation of the XAI methods. Additionally we discard
questions of the type `exist`, where the true answer is `no`, since for these ques-
tions obviously no object in the scene matches the ground truth answer, and
810 hence such questions can be ambiguous to use for the XAI evaluation. Our
results are reported in Table 8 and 9.

Impact of Ground Truth Selectivity. We observe that the more selective the
Ground Truth, i.e., the less objects it contains, the lower the relevance accu-
racy. However, interestingly, even for the most selective Ground Truth, i.e. *GT*
815 *Unique*, the relevance accuracy is still high in absolute value, and only less than
0.08 accuracy lower than when using the *GT Single Object* on simple ques-
tions, for the XAI methods LRP, IG and Guided Backprop. This highlights
the fact that even on complex questions these XAI methods are able to reveal
the unique objects in the scene that are important for a given VQA question.
820 Further, the relative ordering of the XAI methods across the different complex
questions' Ground Truths remains largely unchanged: this illustrates the fact
that our CLEVR-XAI-complex Ground Truths are all appropriate to evaluate
and compare XAI methods.

Table 8: Relevance *mass* accuracy on **CLEVR-XAI-complex**, using different types of Ground Truths. We consider all correctly predicted questions that do not involve counting.

Relevance Mass Accuracy	correctly predicted & no counting							
	GT Unique		GT Unique First-non-empty		GT Union		GT All Objects	
	mean (std)	median	mean (std)	median	mean (std)	median	mean (std)	median
LRP [20]	0.82 (0.19)	0.90	0.84 (0.17)	0.90	0.91 (0.12)	0.95	0.96 (0.03)	0.97
Excitation Backprop [29]	0.78 (0.21)	0.85	0.80 (0.19)	0.86	0.86 (0.15)	0.91	0.91 (0.08)	0.94
IG [19]	0.67 (0.29)	0.77	0.70 (0.27)	0.79	0.87 (0.19)	0.95	0.97 (0.03)	0.98
Guided Backprop [22]	0.65 (0.17)	0.68	0.65 (0.16)	0.69	0.71 (0.14)	0.74	0.75 (0.10)	0.76
Guided Grad-CAM [42]	0.64 (0.20)	0.69	0.65 (0.20)	0.69	0.71 (0.16)	0.74	0.76 (0.11)	0.77
SmoothGrad [43]	0.49 (0.34)	0.52	0.51 (0.33)	0.55	0.61 (0.30)	0.69	0.65 (0.28)	0.73
VarGrad [59]	0.48 (0.34)	0.51	0.50 (0.33)	0.54	0.60 (0.30)	0.68	0.64 (0.29)	0.72
Gradient [16]	0.45 (0.34)	0.43	0.46 (0.33)	0.45	0.56 (0.31)	0.60	0.62 (0.29)	0.67
Gradient×Input [18]	0.39 (0.33)	0.31	0.40 (0.32)	0.33	0.48 (0.31)	0.46	0.53 (0.30)	0.53
Deconvnet [34]	0.17 (0.14)	0.13	0.18 (0.14)	0.14	0.24 (0.16)	0.20	0.27 (0.16)	0.24
Grad-CAM [42]	0.15 (0.12)	0.12	0.16 (0.12)	0.13	0.27 (0.14)	0.27	0.34 (0.11)	0.35

Table 9: Relevance *rank* accuracy on **CLEVR-XAI-complex**, using different types of Ground Truths. We consider all correctly predicted questions that do not involve counting.

Relevance Rank Accuracy	correctly predicted & no counting							
	GT Unique		GT Unique First-non-empty		GT Union		GT All Objects	
	mean (std)	median	mean (std)	median	mean (std)	median	mean (std)	median
LRP [20]	0.64 (0.16)	0.66	0.64 (0.15)	0.66	0.70 (0.13)	0.73	0.78 (0.07)	0.78
Excitation Backprop [29]	0.62 (0.17)	0.65	0.63 (0.16)	0.65	0.70 (0.14)	0.73	0.78 (0.07)	0.79
IG [19]	0.49 (0.19)	0.51	0.51 (0.17)	0.53	0.65 (0.14)	0.69	0.74 (0.06)	0.75
Guided Backprop [22]	0.50 (0.13)	0.52	0.50 (0.12)	0.51	0.51 (0.11)	0.52	0.54 (0.08)	0.55
Guided Grad-CAM [42]	0.50 (0.15)	0.51	0.49 (0.14)	0.51	0.51 (0.12)	0.53	0.55 (0.09)	0.56
SmoothGrad [43]	0.33 (0.17)	0.32	0.34 (0.17)	0.33	0.35 (0.14)	0.35	0.35 (0.12)	0.35
VarGrad [59]	0.31 (0.20)	0.30	0.31 (0.20)	0.30	0.32 (0.17)	0.31	0.31 (0.15)	0.30
Gradient [16]	0.28 (0.14)	0.28	0.29 (0.14)	0.29	0.34 (0.12)	0.34	0.38 (0.09)	0.38
Gradient×Input [18]	0.25 (0.14)	0.25	0.26 (0.13)	0.25	0.30 (0.11)	0.31	0.34 (0.09)	0.34
Deconvnet [34]	0.17 (0.12)	0.15	0.17 (0.11)	0.16	0.21 (0.10)	0.20	0.23 (0.10)	0.22
Grad-CAM [42]	0.22 (0.22)	0.18	0.23 (0.22)	0.20	0.34 (0.20)	0.36	0.42 (0.16)	0.43

Comparison with Simple Questions. We note that the relative ordering of the
825 XAI methods on complex questions is also largely consistent with their ordering
on the simple questions. This supports the fact that both CLEVR-XAI-simple
and CLEVR-XAI-complex subsets of questions are pertinent, and can deliver
fine-grained insights on which objects in the scene the XAI methods are focusing
on when explaining a given prediction, e.g., in order to answer questions such
830 as: How much relevance is assigned to unique objects in the scene (through
using *GT Unique*, resp. *GT Single Object* for simple questions)? And how
much relevance is assigned to all objects involved in the VQA question vs. all
objects in the scene (with *GT Union* vs. *GT All Objects*)?

6. Discussion

835 *Main findings.* In a nutshell, our findings in terms of the comparison of the XAI
methods w.r.t. their relevance accuracy can be summarized as follow: LRP,
IG and Guided Backprop are among the best performing methods (Guided
Grad-CAM also belongs to this category but mainly due to its Guided Back-
prop component), Grad-CAM and Deconvnet perform very poorly, while the
840 gradient-based methods SmoothGrad, VarGrad, Gradient and Gradient \times Input
perform on an intermediate level. For a more detailed analysis on the properties
of these methods taking into account the variance of their results, their sensitiv-
ity w.r.t. hyperparameter choices, and their computational cost, we refer to the
Section 5.2 of this work. Furthermore, we would like to notice that our compar-
845 ative study already includes an advantage for the IG and Grad-CAM methods.
Indeed for the former we ignore data points for which we didn't achieve a pre-
cise enough numerical integration (with our considered number of integration
steps), and for the latter we ignore zero-valued uninformative heatmaps, while
for all other XAI methods were considered all data points of our dataset.

850 *Comparison with previous works.* Through our extensive and careful experi-
ments with the CLEVR-XAI benchmark, we could not confirm the conclusions

drawn from previous evaluation works. To the best of our knowledge, our findings even contradict with most of the claims provided by all large comparative studies we are aware of.

855 For instance, the authors of [35] perform randomization tests on the model parameters and training data to evaluate and compare XAI methods. They find that Guided Backprop and Guided Grad-CAM fail to pass their sanity checks, and according to the authors this implies that these methods are *“incapable of supporting tasks that require explanations that are faithful to the model or the*
860 *data generating process”* [35]. At the same time, they find that Gradient and Grad-CAM successfully pass their tests. In our controlled VQA framework, we find exactly the opposite: Grad-CAM performs very poorly, gradient-based relevances also do not perform well, while Guided Backprop and Guided Grad-CAM are among the top performing XAI methods. We suspect the tested objectives
865 from [35] to be too disconnected from the actual prediction problem and presumably prone to perturbation-based artefacts, and thus to be inadequate to allow for a fine-grained relative comparison of the quality of XAI methods. Indeed in their evaluation they track the change in explanation under perturbation of the model (either through parameter randomization or re-training on random-
870 ized labels), while we keep a clean unmodified configuration based on the same model and data as during model training. Hence we believe that our evaluation setup is more controlled and reliable, which can explain the contradiction in the results. Further, as we pointed out in Section 2, we argue that our considered VQA task is more selective, while the image classification tasks used in [35] tend
875 to produce heatmaps that are often similar for different classes and thus to some degree are insensitive to higher layers of the network since these models act as feature detectors over a single object present in the image. In other words, the model was not trained to focus on one (or more) object(s) and ignore the others, as is the case in our multi-object VQA setup.

880 Another work that is based on the Remove And Retrain paradigm [38] concluded that Gradient, IG, and Guided Backprop are *“worse or on par with a random assignment of [input feature] importance”*. This again contradicts

our results: in our study, IG and Guided Backprop are among the best performing methods, and thus far from representing random feature importance assignments. We suspect again the evaluation protocol of this work to be too disconnected from the actual prediction task and the original model, to serve as a precise evaluation of XAI methods to explain a given prediction.

Lastly, in [36] the authors find that several modified backpropagation-based XAI methods, including LRP and Guided Backprop, which in our experiments are among the top performing methods, are “*independent of the parameters of later layers*” in the model due to a convergence of the explanation to a rank-1 matrix, and hence do “*not explain the decision of deep neural networks faithfully*”, while methods relying on the gradient such as IG, SmoothGrad and Grad-CAM are not prone to this convergence issue. In our VQA setup we instead find empirically a high connection between the relevance heatmaps and the target objects of each question, which even increases with the model’s confidence (i.e., the output of the network). Maybe the phenomenon described in [36] becomes predominant in the asymptotic case of a neural network with a very high number of layers, and might be relevant for further improving XAI methods. However, in our practical use case of a neural network with 12 layers, we could not confirm that this behavior alters the quality of modified backpropagation-based explanations in comparison to gradient-based ones.

Limitations. We are aware that XAI evaluations w.r.t. Ground Truths should always been carried out with care. In many cases there probably exists no *perfect* Ground Truth. In our study for example, the ground truth masks do not take into account the objects’ shadows, while in principle the VQA model could use them for answering certain questions (such as queries about the objects’ shapes). However, for our benchmark we preferred relying on a dataset with realistic looking illumination conditions and shadows, rather than using a more artificial task (such as the sort-of-CLEVR toy task from [31] for example) to keep the evaluation grounded in a practical and natural visual scene environment. Further, we believe the specific dataset we created, which is based on

the CLEVR task [30], is particularly well-suited for XAI evaluation in computer vision due to its complexity and variability: objects are often occluded and at different locations. This forces the model to focus on the objects themselves and not on the background to detect their shapes (this would be less the case in a toy task with no overlapping object boundaries, where the model could develop a strategy that always includes a portion of the “negative space” around objects for shape recognition). Besides, we validated our approach through an experiment taking into account only questions about an object’s color in Section 5.2.

Finally, although we acknowledge that our evaluation setup is not perfect, we firmly believe it is more reliable and trustworthy than previously conducted evaluation tests in XAI.

Outlook. For future evaluations based on our CLEVR-XAI benchmark, it remains an open question to whether the relevance accuracy upper bound of 1.0 can be reached by an XAI method at all (that was not tested in this study), when using one of the selective Ground Truth masks we generated (on the unselective *GT All Objects* this upper bound is already nearly reached by the IG and LRP methods for the mass accuracy). Indeed one possible reason for having an accuracy lower than 1.0 stems the classifier’s uncertainty (we showed that the higher the model’s confidence, the higher the relevance accuracy) on some more difficult to predict data points from our dataset. For such data points the VQA model could be hesitating or confused, and consequently the relevance gets more diffused across multiple objects in the scene, rather than concentrated on the target objects of the question. Nevertheless, we believe this does not invalidate our evaluation approach, since the model’s underlying uncertainty is the same for all XAI methods, and thus our benchmark dataset is still suited to compare the *relative* performance of XAI methods.

In a similar manner, the relatively high standard deviations we observe in our experiments shall not be interpreted as a lack of stability of our results: these are not induced by randomness in the evaluation process, but are mainly due to the difficulty of some of the questions (we showed that the higher the

number of pixels of the ground truth masks, the lower the standard deviation).

Further, in terms of directions for improving XAI methods, it would be interesting to explore whether the strengths of LRP and Guided Backprop could be combined to build an even stronger XAI method, or to investigate possible theoretical connections between those methods. Indeed both methods are efficient (they require a single backward pass), further they do not necessitate to choose/tune a numerical hyperparameter (like this is the case, e.g., in SmoothGrad), or to determine a baseline image (like this is the case for IG). IG also performs well, but is far more expensive to compute. Moreover, methods such as SmoothGrad and IG rely on perturbed input images for the averaging, resp. the integration, of gradients. Such perturbed data points may lie outside the actual training data’s manifold, bearing the risk to lead to artefacts and unreliable model behavior.

Lastly, for future work on evaluating XAI our benchmark could be tested on other optimization-based or sampling-based XAI methods (in this work we mainly tested deterministic XAI methods). Also it would interesting to train a neural network model that is conform to the Deep Taylor Decomposition theoretical framework [67], i.e., a Relation Network model [31] which is constrained to have negative biases and positive-valued prediction scores, to check if this increases the performance of the LRP method (in the present work we used a standard Relation Network model trained with no particular constraints).

7. Conclusion

In the present work we have proposed a novel VQA-based evaluation paradigm for computer vision XAI. Our benchmark dataset (CLEVR-XAI) with corresponding Ground Truth annotations and metrics can be used to analyze and improve current XAI methods for neural networks.

Compared to previous evaluation approaches, which were mainly based on CNN object classifiers, our framework has the advantage of being altogether selective, controlled and realistic, and to further rely on the same model and

data as during model training. Hereby we hope to pave the way for a more reliable and trustworthy assessment of XAI methods in computer vision.

Among the methods we tested in our comparative study, we found Layer-
975 wise Relevance Propagation, Integrated Gradients and Guided Backpropagation to be the most accurate XAI methods, while Deconvnet and Grad-CAM were the least accurate ones to back up the model's predictions.

References

- [1] C. Lu, X. Tang, Surpassing human-level face verification performance on
980 LFW with GaussianFace, in: AAAI Conference on Artificial Intelligence, 2015, pp. 3811–3819.
- [2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep
bidirectional transformers for language understanding, in: Proceedings of
the North American Chapter of the Association for Computational Lin-
985 guistics: Human Language Technologies (NAACL), 2019, pp. 4171–4186.
- [3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural
networks, *Nature* 542 (2017) 115–118.
- [4] R. Geirhos, J. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge,
990 F. A. Wichmann, Shortcut Learning in Deep Neural Networks, *Nature
Machine Intelligence* 2 (2020) 665–673.
- [5] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, A. Rohrbach, Women
also snowboard: Overcoming bias in captioning models, in: Proceedings of
the European Conference on Computer Vision (ECCV), 2018, pp. 793–811.
- 995 [6] V. Manjunatha, N. Saini, L. S. Davis, Explicit bias discovery in visual
question answering models, in: Proceedings of the Conference on Computer
Vision and Pattern Recognition (CVPR), 2019, pp. 9554–9563.

- [7] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, *Nature Communications* 10 (2019) 1096.
- [8] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Vol. 11700 of *Lecture Notes in Computer Science*, Springer, 2019.
- [9] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, *Proceedings of the IEEE* 109 (3) (2021) 247–278.
- [10] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, *Information Fusion* 58 (2020) 82–115.
- [11] EU-GDPR, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), *Official Journal of the European Union* L 119 59 (2016) 1–88.
- [12] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, W. Samek, Pruning by explaining: A novel criterion for deep neural network pruning, *Pattern Recognition* 115 (2021) 107899.
- [13] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, *Nature Communications* 8 (2017) 13890.
- [14] F. Horst, S. Lapuschkin, W. Samek, K.-R. Müller, W. I. Schöllhorn, Explaining the unique nature of individual gait patterns with deep learning, *Scientific Reports* 9 (2019) 2391.

- [15] M. Widrich, B. Schäfl, M. Pavlović, H. Ramsauer, L. Gruber, M. Holzleitner, J. Brandstetter, G. K. Sandve, V. Greiff, S. Hochreiter, G. Klambauer, Modern Hopfield Networks and Attention for Immune Repertoire Classification, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- 1030
- [16] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, in: *International Conference on Learning Representations (ICLR)*, 2014.
- [17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, *International Journal of Computer Vision* 128 (2) (2020) 336–359.
- 1035
- [18] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Interpretable deep learning by propagating activation differences, *arXiv preprint arXiv:1605.01713* (2016).
- 1040
- [19] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.
- [20] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLoS ONE* 10 (7) (2015) e0130140.
- 1045
- [21] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top-down neural attention by excitation backprop, *International Journal of Computer Vision* 126 (10) (2018) 1084–1102.
- [22] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: *International Conference on Learning Representations (ICLR)*, 2015.
- 1050

- [23] M. T. Ribeiro, S. Singh, C. Guestrin, “Why Should I Trust You?”: Explaining the Predictions of Any Classifier, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.
- [24] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NIPS), 2017, pp. 4765–4774.
- [25] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the International Conference on Computer Vision (ICCV), 2017.
- [26] L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction difference analysis, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.
- [27] J. Chen, L. Song, M. Wainwright, M. Jordan, Learning to Explain: An Information-Theoretic Perspective on Model Interpretation, in: Proceedings of the International Conference on Machine Learning (ICML), 2018, pp. 883–892.
- [28] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a Deep Neural Network has learned, IEEE Transactions on Neural Networks and Learning Systems 28 (11) (2017) 2660–2673.
- [29] J. Zhang, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top-down neural attention by excitation backprop, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 543–559.
- [30] J. Johnson, B. Hariharan, L. v. d. Maaten, L. Fei-Fei, C. L. Zitnick, R. Girshick, CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1988–1997.

- [31] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 4967–4976.
- 1085 [32] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for Deep Neural Networks, in: *International Conference on Learning Representations (ICLR)*, 2018.
- [33] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, M. Botvinick, On the importance of single directions for generalization, in: *International Conference on Learning Representations (ICLR)*, 2018.
- 1090 [34] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 818–833.
- [35] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity Checks for Saliency Maps, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 9505–9515.
- 1095 [36] L. Sixt, M. Granz, T. Landgraf, When Explanations Lie: Why Many Modified BP Attributions Fail, in: *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 9046–9057.
- 1100 [37] J. Adebayo, M. Muelly, I. Liccardi, B. Kim, Debugging tests for model explanations, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [38] S. Hooker, D. Erhan, P. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 9737–9748.
- 1105 [39] J. Gu, Y. Yang, V. Tresp, Understanding individual decisions of cnns via contrastive backpropagation, in: *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2019, pp. 119–134.

- [40] S. Beery, G. Van Horn, P. Perona, Recognition in Terra Incognita, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.
- [41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.
- [42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the International Conference on Computer Vision (ICCV), 2017.
- [43] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, in: Proceedings of the International Conference on Machine Learning Workshop on Visualization for Deep Learning, 2017.
- [44] A. Kuhnle, A. Copestake, Shapeworld - a new test methodology for multi-modal language understanding, arXiv:1704.04517 (2017).
- [45] A. Kuhnle, Evaluating visually grounded language capabilities using microworlds, Tech. Rep. UCAM-CL-TR-942, University of Cambridge (2020).
- [46] J. Kim, M. Ricci, T. Serre, Not-So-CLEVR: learning same–different relations strains feedforward neural networks, Interface Focus 8 (20180011) (2018).
- [47] D. Bahdanau, H. de Vries, T. J. O’Donnell, S. Murty, P. Beaudoin, Y. Bengio, A. Courville, CLOSURE: Assessing Systematic Generalization of CLEVR Models, arXiv:1912.05783 (2019).
- [48] D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, A. Courville, Systematic Generalization: What Is Required and Can It

- Be Learned?, in: International Conference on Learning Representations (ICLR), 2019.
- [49] D. Barrett, F. Hill, A. Santoro, A. Morcos, T. Lillicrap, Measuring abstract reasoning in neural networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2018, pp. 511–520.
1140
- [50] A. Holzinger, A. Saranti, H. Mueller, Kandinskypatterns – an experimental exploration environment for pattern analysis and machine intelligence, arXiv:2103.00519 (2021).
- [51] D. Mascharka, P. Tran, R. Soklaski, A. Majumdar, Transparency by design: Closing the gap between performance and interpretability in visual reasoning, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2018, p. 4942–4950.
1145
- [52] J. Oramas, K. Wang, T. Tuytelaars, Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks, in: International Conference on Learning Representations (ICLR), 2019.
1150
- [53] M. Schuessler, P. Weiß, L. Sixt, Two4two: Evaluating interpretable machine learning - a synthetic dataset for controlled experiments, in: International Conference on Learning Representations Workshop on Responsible AI, 2021.
- [54] A. S. Ross, M. C. Hughes, F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 2662–2670.
1155
- [55] A. Holzinger, A. Carrington, H. Müller, Measuring the Quality of Explanations: The System Causability Scale (SCS), Künstliche Intelligenz 34 (2020) 193–198.
1160
- [56] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the

- 1165 58th Annual Meeting of the Association for Computational Linguistics
(ACL), 2020, pp. 4198–4205.
- [57] A. Holzinger, M. Kickmeier-Rust, H. Müller, Kandinsky patterns as IQ-test for machine learning, in: Machine Learning and Knowledge Extraction. CD-MAKE 2019, Vol. 11713 of Lecture Notes in Computer Science, Springer, 2019, pp. 1–14.
- 1170 [58] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 3145–3153.
- [59] J. Adebayo, J. Gilmer, I. J. Goodfellow, B. Kim, Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values, in:
1175 International Conference on Learning Representations (ICLR), 2018.
- [60] A. Mahendran, A. Vedaldi, Salient deconvolutional networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 120–135.
- [61] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh,
1180 VQA: Visual Question Answering, in: International Conference on Computer Vision (ICCV), 2015, pp. 2425–2433.
- [62] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: Proceedings of the Conference on Computer Vision and
1185 Pattern Recognition (CVPR), 2017, pp. 6325–6334.
- [63] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don’t just assume; look and answer: Overcoming priors for visual question answering, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4971–4980.

- 1190 [64] Y. Dimopoulos, P. Bourret, S. Lek, Use of some sensitivity criteria for
choosing networks with good generalization ability, *Neural Processing Let-*
ters 2 (6) (1995) 1–4.
- [65] M. Gevrey, I. Dimopoulos, S. Lek, Review and comparison of methods
to study the contribution of variables in artificial neural network models,
1195 *Ecological Modelling* 160 (3) (2003) 249–264.
- [66] J. Seo, J. Choe, J. Koo, S. Jeon, B. Kim, T. Jeon, Noise-adding methods
of saliency map as series of higher order partial derivative, in: *Proceedings*
of the ICML Workshop on Human Interpretability in Machine Learning,
2018.
- 1200 [67] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Ex-
plaining nonlinear classification decisions with deep Taylor decomposition,
Pattern Recognition 65 (2017) 211–222.
- [68] G. Montavon, W. Samek, K.-R. Müller, Methods for Interpreting and Un-
derstanding Deep Neural Networks, *Digital Signal Processing* 73 (2018)
1205 1–15.
- [69] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-
wise relevance propagation: An overview, in: *Explainable AI: Interpreting,*
Explaining and Visualizing Deep Learning, Vol. 11700 of *Lecture Notes in*
Computer Science, Springer, 2019, pp. 193–209.
- 1210 [70] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, S. La-
puschkin, Towards best practice in explaining neural network decisions
with LRP, in: *IEEE International Joint Conference on Neural Networks*
(IJCNN), 2020, pp. 1–7.
- [71] L. Arras, F. Horn, G. Montavon, K.-R. Müller, W. Samek, "What is rel-
1215 evant in a text document?": An interpretable machine learning approach,
PLoS ONE 12 (7) (2017) e0181142.

- [72] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis, in: Proceedings of the EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA), 2017, pp. 159–168.
- [73] N. Poerner, B. Roth, H. Schütze, Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018, pp. 340–350.
- [74] L. Arras, J. Arjona-Medina, M. Widrich, G. Montavon, M. Gillhofer, K.-R. Müller, S. Hochreiter, W. Samek, Explaining and Interpreting LSTMs, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Vol. 11700 of Lecture Notes in Computer Science, 2019, pp. 211–238.
- [75] P. Sturmfels, S. Lundberg, S.-I. Lee, Visualizing the impact of feature attribution baselines, Distill (2020). doi:10.23915/distill.00022.

Appendix A. Additional CLEVR-XAI-complex example

Figure A.3 provides a CLEVR-XAI-complex example data point for a question with a tree-structured functional program. Note that the program does not contain any `unique` function, therefore the *GT Unique* mask is undefined for this question. The *GT Unique First-non-empty* mask includes the first non-empty set of objects returned by the functions in the program processed in reversed order, i.e., the output of the `union` function, which corresponds to the big purple cube. The *GT Union* further contains the output of all intermediate functions in the program, in particular this corresponds to the output of the first filter function in each program branch (`filter_size` in `branch1` and `filter_color` in `branch2`), i.e., all large objects and all green objects (thus the two large cubes and the small green cylinder). The *GT All Objects* simply contains all the objects in the scene.

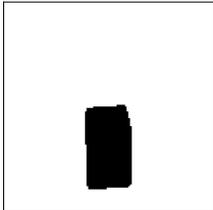
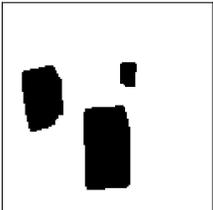
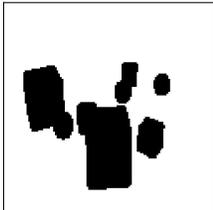
Image	Question/Answer	Program
	<p>What number of objects are large purple blocks or green metallic cubes?</p> <p><i>one</i></p>	<pre>branch1 = [scene, filter_size, filter_color, filter_shape] branch2 = [scene, filter_color, filter_material, filter_shape] program = [(branch1,branch2), union, count]</pre>
Ground Truth	GT Unique	GT Unique First-non-empty
	undefined	
	GT Union	GT All Objects
		

Figure A.3: Example CLEVR-XAI-complex data point. The functional program is used to determine which objects in the scene are considered as ground truths.

Note that in general, as illustrated with the previous example, for CLEVR-
1245 XAI-complex questions some ground truth masks can be undefined or empty. This is why in Table B.10 for three ground truths there are less than 100,000 data points.

Appendix B. CLEVR-XAI dataset statistics

In Table B.10 we provide statistics for the CLEVR-XAI benchmark dataset.
1250 The program length is the number of basic functions present in the question's functional program and is an indicator of the question's complexity. For the

Table B.10: CLEVR-XAI dataset statistics

CLEVR-XAI-simple (39,761 data points)

program length: 4 min, 6 max, 5.1 mean, 0.7 std

Ground Truth	number of questions	number of pixels				number of objects			
		min	max	mean	std	min	max	mean	std
<i>GT Single Object</i>	39,761	40	2040	428	321	1	1	1	0
- small target object	20,393	40	548	189	75	1	1	1	0
- large target object	19,368	125	2040	679	286	1	1	1	0
<i>GT All Objects</i>	39,761	347	6789	2654	1071	3	10	6.5	2.3

CLEVR-XAI-complex (100,000 data points)

program length: 2 min, 24 max, 11.1 mean, 4.3 std

Ground Truth	number of questions	number of pixels				number of objects			
		min	max	mean	std	min	max	mean	std
<i>GT Unique</i>	89,873	43	4320	837	572	1	4	2.0	0.9
<i>GT Unique First-non-empty</i>	99,786	43	5908	1042	676	1	10	2.5	1.1
<i>GT Union</i>	99,786	45	6789	1958	1116	1	10	4.8	2.3
<i>GT All Objects</i>	100,000	347	6789	2650	1072	3	10	6.5	2.3

CLEVR-XAI-simple’s *GT Single Object* we additionally distinguish between two cases depending on the size attribute of the target object of the question: note the high variability in the number of pixels within each size category, indicating
 1255 that objects can be partially occluded and located at different places (in the front or in the back of the scene’s plane surface).

Lastly the number of pixels per ground truth we report were calculated on images of size 128×128 (since the neural network we use in our experiments takes as input images of size 128×128 , for more details on the ground truth mask
 1260 resizing see Appendix D). In our dataset release we also provide the possibility to use ground truths of size 320×480 (the original size of the CLEVR dataset images [30]) or any other chosen image size.

Appendix C. Neural Network used for the XAI Evaluation

Here we provide more details on the Relation Network (RN) model architecture and training [31], which we used for the empirical evaluation of XAI methods.

The CNN part of the network is made of 4 layers, each with the following structure: `conv` \rightarrow `relu` \rightarrow `batchnorm`. Each convolutional layer has 24 kernels of size 3×3 , stride 2, and no padding.

The LSTM part of the network is a unidirectional LSTM with word embeddings of size 32, and a hidden layer of size 128.

The Relation Network part of the model is made of 4 fully-connected layers of size 256, each followed with ReLU activation, and a final element-wise summation layer.

The classifier part of the network contains 3 fully-connected layers, where the first two layers have size 256 and are followed each by ReLU activation. Additionally, the second layer uses dropout ($p = 0.5$). The output layer has size 28 (the number of possible answers for the CLEVR VQA task [30]).

For preprocessing the questions, we removed punctuation and applied lowercasing, this leaves us with a vocabulary of size 80.

For preprocessing the images, we rescaled the pixel values to the range $[0, 1]$, and resized the images to the size 128×128 (the original CLEVR images have size 320×480).

Training was done with the Adam optimizer, using a batch size of 64, an initial learning rate of $2.5e-4$, clipping the gradient norm to 5.0, l2-norm regularization of $4e-5$, and decreasing the learning rate by a factor of 0.95 once the validation accuracy does not improve within 10 epochs. Training was done for a maximum of 1200 epochs.

During training, we also applied data augmentation (random cropping and random rotation of the images), as described in the original publication [31].

Appendix D. Resizing of Ground Truth Masks

Note that since the RN model we use in our experiments takes as input images of size 128×128 , the ground truth masks used for the XAI evaluation also have to be resized to this size (initially these masks have the same size as the original CLEVR dataset’s images [30], i.e. 320×480). To this end we proceeded in the following way: we resized the masks using the same operation as was used for resizing the standard input images, starting with masks having a value of 1.0 on the ground truth pixels, and 0.0 elsewhere. Then, after resizing the masks, we set all pixels having a non-zero value to True, and the remaining pixels to False. This way we ensure that the resulting masks also include the objects’ borders (which are slightly blurred and dilated due to the resizing operation).

Appendix E. XAI methods hyperparameters

In table E.11 we recapitulate the hyperparameters and pooling techniques we used for each XAI method, after tuning them for each relevance accuracy metric, as was described in Section 5.2.

Table E.11: XAI methods hyperparameters

Relevance Mass Accuracy		
Method	Hyperparameters	Pooling
LRP [20]	α_1, β_0 -rule in the hidden layers, box-rule in the input layer	l2-norm-sq
Excitation Backprop [29]	α_1, β_0 -rule in all layers	l2-norm-sq
IG [19]	mean channel values as baseline image, nb of integration steps \in [300, 1000, 3000, 10000, 30000] such that relative error < 0.01	pos,l2-norm-sq
Guided Backprop [22]	modified gradient backward pass for all ReLU layers in the network	l2-norm-sq
Guided Grad-CAM [42]	element-wise multiplication of Guided Backprop with modified gradient backward pass for all ReLU layers in the network & Grad-CAM applied to the ReLU layer output in the last convolutional layer of the network	l2-norm-sq
SmoothGrad [43]	nb of samples 300, noise level 0.05, gradient squared	l2-norm-sq
VarGrad [59]	nb of samples 300, noise level 0.05	l2-norm-sq
Gradient [16]	gradient squared	l2-norm-sq
Gradient \times Input [18]	gradient squared	l2-norm-sq
Deconvnet [34]	modified gradient backward pass only for ReLU layers within the convolutional neural network part of the network	pos,l2-norm-sq
Grad-CAM [42]	Grad-CAM applied to the batchnorm layer output in the last convolutional layer of the network	none
Relevance Rank Accuracy		
Method	Hyperparameters	Pooling
LRP [20]	α_1, β_0 -rule in the hidden layers, box-rule in the input layer	max-norm
Excitation Backprop [29]	α_1, β_0 -rule in all layers	max-norm
IG [19]	mean channel values as baseline image, nb of integration steps \in [300, 1000, 3000, 10000, 30000] such that relative error < 0.01	max-norm
Guided Backprop [22]	modified gradient backward pass for all ReLU layers in the network	max-norm
Guided Grad-CAM [42]	element-wise multiplication of Guided Backprop with modified gradient backward pass for all ReLU layers in the network & Grad-CAM applied to the ReLU layer output in the last convolutional layer of the network	max-norm
SmoothGrad [43]	nb of samples 300, noise level 0.05, gradient unsquared	max-norm
VarGrad [59]	nb of samples 300, noise level 0.05	max-norm
Gradient [16]	gradient unsquared	max-norm
Gradient \times Input [18]	gradient squared	max-norm
Deconvnet [34]	modified gradient backward pass only for ReLU layers within the convolutional neural network part of the network	max-norm
Grad-CAM [42]	Grad-CAM applied to the ReLU layer output in the last convolutional layer of the network	none

Appendix F. Heatmaps

In Figures F.12-F.20 we provide example heatmaps for CLEVR-XAI questions. In all cases we use as the target class for the XAI method the neural network’s *predicted* class.

1310 We selected the example data points automatically in the following way: we conducted a search over the CLEVR-XAI-simple and CLEVR-XAI-complex subsets, resp., and retrieved the points with the highest predicted softmax probabilities. For Fig. F.12- F.14 we considered all correctly predicted data points from CLEVR-XAI-simple. For Fig. F.15- F.17 we considered all falsely predicted
1315 data points from CLEVR-XAI-simple that query an object’s color. And for Fig. F.18- F.20 we considered all correctly predicted data points from CLEVR-XAI-complex.

For the heatmap visualization we depict the raw heatmap (after pooling the relevance along the channel axis according to the pooling technique that was
1320 tuned towards the relevance *mass* accuracy, as was described in Section 5.2) and color code the relevance value in red (normalizing the color to the maximum value per heatmap), using Matplotlib’s `seismic` colormap. Additionally we visualize the original image overlaid with the heatmap. For the latter visualization we first apply gaussian blur to the pooled heatmap with standard deviation
1325 0.02 times the image dimension, and subsequently color code the heatmap with Matplotlib’s `gist_ncar` colormap (normalized to the extremal relevance values per heatmap).

Especially Fig. F.15 and F.17 illustrate why falsely predicted data points should not be taken into account when evaluating XAI methods: indeed for
1330 these questions the model focused on the wrong object in the image to produce the answer. For Fig. F.15 it obviously focused on the small brown cylinder rather than on the small yellow ball. For Fig. F.17 it appears to have focused on the large blue cube rather than on the small purple cube. On Fig. F.16, however, it seems the neural network was able to detect the right object (the small red
1335 ball), but apparently it mistakenly interpreted its color as being brown.

Table F.12: Heatmaps for a correctly predicted CLEVR-XAI-simple question (raw heatmap and heatmap overlayed with original image), and corresponding relevance *mass* accuracy.

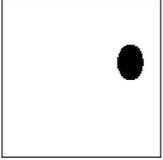
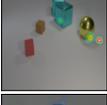
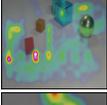
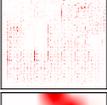
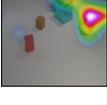
What is the material of the large ball? <i>metal</i>			GT Single Object
LRP [20]			0.98
Excitation Backprop [29]			0.97
IG [19]			0.80
Guided Backprop [22]			0.58
Guided Grad-CAM [42]			0.72
SmoothGrad [43]			0.03
VarGrad [59]			0.02
Gradient [16]			0.25
Gradient×Input [18]			0.17
Deconvnet [34]			0.04
Grad-CAM [42]			0.19

Table F.13: Heatmaps for a correctly predicted CLEVR-XAI-simple question (raw heatmap and heatmap overlayed with original image), and corresponding relevance *mass* accuracy.

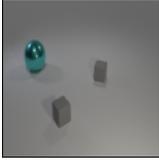
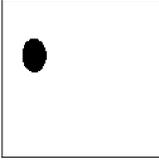
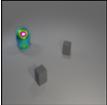
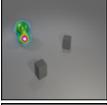
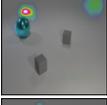
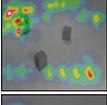
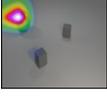
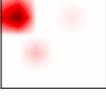
What is the large cyan ball made of? <i>metal</i>			GT Single Object
LRP [20]			0.97
Excitation Backprop [29]			0.97
IG [19]			0.96
Guided Backprop [22]			0.79
Guided Grad-CAM [42]			0.85
SmoothGrad [43]			0.14
VarGrad [59]			0.09
Gradient [16]			0.10
Gradient×Input [18]			0.07
Deconvnet [34]			0.08
Grad-CAM [42]			0.27

Table F.14: Heatmaps for a correctly predicted CLEVR-XAI-simple question (raw heatmap and heatmap overlayed with original image), and corresponding relevance *mass* accuracy.

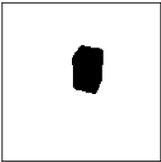
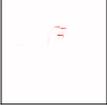
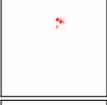
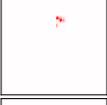
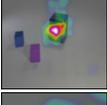
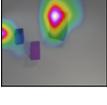
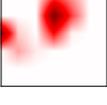
What is the material of the large block? <i>metal</i>			GT Single Object
LRP [20]			0.94
Excitation Backprop [29]			0.82
IG [19]			0.99
Guided Backprop [22]			0.80
Guided Grad-CAM [42]			0.82
SmoothGrad [43]			0.75
VarGrad [59]			0.75
Gradient [16]			0.85
Gradient×Input [18]			0.67
Deconvnet [34]			0.43
Grad-CAM [42]			0.28

Table F.15: Heatmaps for a falsely predicted CLEVR-XAI-simple question (raw heatmap and heatmap overlaid with original image), and corresponding relevance *mass* accuracy.

What is the color of the rubber ball? true: <i>yellow</i> predicted: <i>brown</i>			GT Single Object
LRP [20]			0.0
Excitation Backprop [29]			0.0
IG [19]			0.0
Guided Backprop [22]			0.0
Guided Grad-CAM [42]			0.0
SmoothGrad [43]			0.0
VarGrad [59]			0.0
Gradient [16]			0.0
Gradient×Input [18]			0.0
Deconvnet [34]			0.0
Grad-CAM [42]			0.0

Table F.16: Heatmaps for a falsely predicted CLEVR-XAI-simple question (raw heatmap and heatmap overlaid with original image), and corresponding relevance *mass* accuracy.

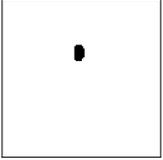
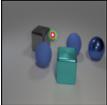
What is the color of the small shiny thing? true: <i>red</i> predicted: <i>brown</i>			GT Single Object
LRP [20]			0.88
Excitation Backprop [29]			0.83
IG [19]			0.48
Guided Backprop [22]			0.44
Guided Grad-CAM [42]			0.64
SmoothGrad [43]			0.05
VarGrad [59]			0.02
Gradient [16]			0.04
Gradient×Input [18]			0.06
Deconvnet [34]			0.07
Grad-CAM [42]			0.10

Table F.17: Heatmaps for a falsely predicted CLEVR-XAI-simple question (raw heatmap and heatmap overlaid with original image), and corresponding relevance *mass* accuracy.

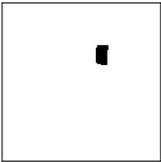
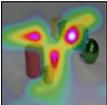
What color is the small shiny block? true: <i>purple</i> predicted: <i>blue</i>			GT Single Object
LRP [20]			0.37
Excitation Backprop [29]			0.40
IG [19]			0.58
Guided Backprop [22]			0.28
Guided Grad-CAM [42]			0.52
SmoothGrad [43]			0.74
VarGrad [59]			0.72
Gradient [16]			0.94
Gradient×Input [18]			0.94
Deconvnet [34]			0.30
Grad-CAM [42]			0.03

Table F.18: Heatmaps for a correctly predicted CLEVR-XAI-complex question (raw heatmap and heatmap overlayed with original image), and relevance *mass* accuracy.

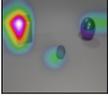
What material is the large object that is left of the big purple metallic ball? <i>metal</i>			GT Unique First-non-empty
LRP [20]			0.97
Excitation Backprop [29]			0.91
IG [19]			0.99
Guided Backprop [22]			0.88
Guided Grad-CAM [42]			0.90
SmoothGrad [43]			0.42
VarGrad [59]			0.40
Gradient [16]			0.63
Gradient×Input [18]			0.42
Deconvnet [34]			0.04
Grad-CAM [42]			0.38

Table F.19: Heatmaps for a correctly predicted CLEVR-XAI-complex question (raw heatmap and heatmap overlayed with original image), and relevance *mass* accuracy.

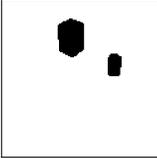
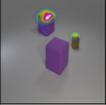
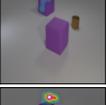
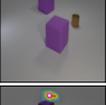
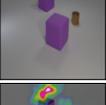
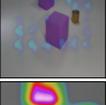
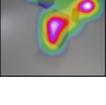
There is a large purple cube that is behind the brown shiny cylinder; what material is it? <i>metal</i>			GT Unique First-non-empty
LRP [20]			0.92
Excitation Backprop [29]			0.83
IG [19]			0.98
Guided Backprop [22]			0.81
Guided Grad-CAM [42]			0.83
SmoothGrad [43]			0.76
VarGrad [59]			0.74
Gradient [16]			0.14
Gradient×Input [18]			0.08
Deconvnet [34]			0.07
Grad-CAM [42]			0.13

Table F.20: Heatmaps for a correctly predicted CLEVR-XAI-complex question (raw heatmap and heatmap overlayed with original image), and relevance *mass* accuracy.

There is a big ball on the right side of the blue metal object; what is it made of? <i>rubber</i>			GT Unique First-non-empty
LRP [20]			0.72
Excitation Backprop [29]			0.66
IG [19]			0.46
Guided Backprop [22]			0.74
Guided Grad-CAM [42]			0.81
SmoothGrad [43]			0.55
VarGrad [59]			0.53
Gradient [16]			0.93
Gradient×Input [18]			0.83
Deconvnet [34]			0.13
Grad-CAM [42]			0.14