UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Software Engineering Curriculum

Mykyta Skliarov

A Comparative Evaluation of Explainability Techniques for Image Data

Master's Thesis (30 ECTS)

Supervisor: Radwa ElShawi, PhD

Tartu 2024

A Comparative Evaluation of Explainability Techniques for Image Data

Abstract:
The use of machine learning has increased dramatically in the last decade across many domains, especially in computer vision, where high-performing convolutional deep neural networks have reached and even surpassed human performance in many areas. To answer increasing needs in transparency for these black-box models the community of eXplainable AI have produced various techniques to explain their predictions. A popular way to do so for image data is via saliency maps. At the same time, objectively evaluating the quality of these techniques is not an easy task, due to the multifaceted nature of interpretability. In this work we perform a thorough comparative evaluation of six popular saliency map explainability techniques, namely LIME, SHAP, GradCAM, GradCAM++, IntGrad and SmoothGrad, using five quantitative function-grounded metrics present in literature, specifically fidelity, stability, identity, separability and time, on three commonly used benchmarking datasets, and three well-known model architectures, to determine pros and cons of each of the techniques. Though we find that no single technique dominates in all metrics, the obtained results show that IntGrad and SmoothGrad performed well on our fidelity and stability tests, with SHAP also achieving high results in fidelity. All techniques but LIME and SmoothGrad score highly on identity metric, and all but LIME - on separability, while GradCAM and GradCAM++ were by far the fastest. We also note the caveats we identified in the metrics, suggesting that more work is needed to gain a full picture of the quality of the different XAI techniques.

Keywords:
eXplainable Artificial Intelligence, Interpretable Machine Learning, Metrics, Evaluation

CERCS: P176 - Artificial intelligence

Pildiandmete selgitamismeetodite võrdlev hindamine

Lühikokkuvõte: Masinõppe kasutamine on viimasel kümnendil paljudes valdkondades märkimisväärselt kasvanud, eriti arvutinägemise valdkonnas, kus suure jõudlusega konvolutsioonilised süva-neuronivõrgud on saavutanud ja isegi ületanud inimese jõudluse paljudes valdkondades. Et vastata nende black-box mudelite läbipaistvuse kasvavale vajadusele, on eXplainable AI kogukond loonud erinevaid tehnikaid nende prognooside selgitamiseks. Populaarne viis seda teha pildiandmete puhul on kasutada silmapaistvuse kaardid. Samas ei ole nende meetodite kvaliteedi objektiivne hindamine tõlgendatavuse mitmetahulisuse tõttu lihtne ülesanne. Käesolevas töös teostame kuue populaarse silmapaistvuse kaardi selgitamise tehnika, nimelt LIME, SHAP, GradCAM, GradCAM++, IntGrad ja SmoothGrad, põhjaliku võrdleva hindamise, kasutades kirjanduses esinevaid viit kvantitatiivset funktsioonipõhist mõõdikut, täpsemalt truudust, stabiilsust, identiteeti, eraldatavust ja aega, kolme üldkasutatava võrdlusandmestiku ja kolme tuntud mudelite arhitektuuri põhjal, et määrata kindlaks iga tehnika plussid ja miinused. Kuigi me leiame, et ükski tehnika ei domineeri kõigis mõõdikutes, näitavad saadud tulemused, et IntGrad ja SmoothGrad esinesid meie usaldusväärsuse ja stabiilsuse testides hästi, kusjuures SHAP saavutas ka kõrgeid tulemusi usaldusväärsuse osas. Kõik meetodid peale LIME ja SmoothGrad said kõrgeid tulemusi identiteedi mõõtkavas ja kõik peale LIME - eraldatavuse mõõtkavas, samas kui GradCAM ja GradCAM++ olid kaugelt kõige kiiremad. Samuti märgime, et meetrikate puhul on ettevaatusabinõusid, mis viitavad sellele, et erinevate XAI-tehnikate kvaliteedist täieliku pildi saamiseks on vaja teha rohkem tööd.

Võtmesõnad:
Seletatav tehisintellekt, tõlgendatav masinõpe, mõõdikud, hindamine

CERCS: P176 - Tehisintellekt

Contents

1 Introduction

Usage of artificial intelligence (AI) and machine learning (ML) models has increased dramatically in recent years, especially in computer vision (CV). With the success of deep learning, AI has approached human performance and even outperformed humans in some CV challenges (e.g. ImageNet image classification challenge [RDS+15]). ML models have been adopted in various domains including healthcare [NAL+21], criminal justice [FBL16, Cho16], education [BH22], autonomous driving [YLCT20] etc.

Many of the deployed models are black-box models, meaning that they are either too complicated for a human to comprehend (e.g. deep neural networks or DNNs), or proprietary models. Since their inner workings are opaque, such models can have low transparency, which poses risks in security, fairness, ethics, and limits trust in them, despite their impressive performance. The risks that such models pose to society are recognised not only by the ML community, but also by policy makers. In 2018 the general data protection regulation (GDPR) was imposed by the European Parliament forcing industries to "explain" any decision taken when automated decision making takes place: "a right of explanation for all individuals to obtain meaningful explanations of the logic involved" [EC]. In 2019, the High-Level Expert Group on AI presented the ethics guidelines for trustworthy AI [Hig19]. While legal experts may hold differing opinions on these clauses, there is a consensus that the implementation of such a principle is urgently needed, presenting a significant and open scientific challenge.

To address these challenges the field of eXplainable Artificial Intelligence (XAI) has emerged. It aims to shift focus from predictive accuracy as the sole metric of quality of ML models to the aspects of interpretability and explainability. As the field grows, so does the number of different techniques to help interpret and explain the predictions of ML models. They can be used by businesses as part of the decision making process in human-in-the-loop systems, or by researches during model debugging process. These techniques are varied and span different model types, data types and explanation types. For image data an intuitive way to provide an explanation of a prediction is naturally through visualisation. The most popular branch of visualisation techniques are the ones that generate saliency maps (SM), which show the importance of different parts of the image towards the final output of the model. Some of the ways saliency maps can be generated is by exploiting backpropagation process of neural networks, as done in Integrated Gradients [STY17] technique; by exposing internal state of the model, such as in GradCAM [SCD+17] technique; or by perturbing the input image and measuring the change in the output of the model, as done in Local Interpretable Model-agnostic Explanations (LIME) [RSG16] technique. An advantage of saliency maps over other visualisation techniques is that the explanations produced by them can be directly correlated to and overlaid onto the feature space of the input image, which is intuitive and simple to understand for a human user. At the same time, thy have drawn criticism for being unreliable, misleading or insufficient in some cases [AGM+18, Rud19].

As the number of XAI techniques expands and their adoption increases, so grows the need to evaluate them in order to allow researchers and businesses to understand which techniques are more or less appropriate for their use cases. This is made difficult by the fact that there is no unified agreed-upon definition of interpretability in the XAI community [Rud19]. While it is considered to be multi-faceted and domain-specific notion, preventing us from measuring it directly, existing research has identified a number of properties that make a XAI technique useful, as well as various ways to measure them [NTP$^+$23]. These properties can be categorised by what aspect of quality they represent: the quality of the content of the explanation (e.g. correctness w.r.t. the black-box mode), the presentation aspect of the explanation (e.g. the size of the explanation), or the way the user interacts with the explanation (e.g. how much does the user's prior knowledge matters). Another way to divide evaluation criteria is into function-grounded metrics, human-grounded metrics and application-grounded metrics [DVK17]. The latter two usually involve evaluating the quality of the explanations through humans (non-experts or experts respectively) by collecting their answers through questionnaires or interviews and performing quantitative analysis afterwards. These are important, especially for qualitative evaluation, and, since some desirable properties of XAI techniques can be considered subjective (e.g. those related to presentation or human-explanation interaction), such experiments may be the only way to measure them. However, they are labour-intensive and as such are difficult to scale up, standardize and impossible to fully automate. The former, functionally-grounded metrics, do not require humans. Instead they evaluate some property of a XAI technique by measuring its performance on some proxy task. It can be challenging to find a good proxy task, but since these metrics don't involve humans, they may arguably be more objective. Additionally, since they can be automated, they can be more appropriate for benchmarking, non-critical domains and when speed of development matters.

Even though a multitude of quantitative evaluation methods for measuring different desirable properties of XAI explanations were developed already, there is no agreed-upon unified set of metrics [NTP$^+$23]. In the absence of consensus on this topic, there has been a lack of overall quantitative comparative evaluations of popular and recent XAI techniques for image data, especially comparing performance on different underlying black-box ML models. We aim to close this research gap by first identifying a set of function-grounded metrics that can be used to quantitatively evaluate the quality of such techniques from different angles, then performing rigorous experiments to compare different saliency map techniques using the selected metrics across different datasets and models.

Our main contribution is as follows: we conduct a detailed comparative evaluation of six recent and popular local interpretability techniques for image data that generate explanations in form of saliency maps, namely LIME, SHAP, GradCAM, GradCAM++, IntGrad and SmoothGrad, using five quantitative function-grounded metrics present in literature, specifically *fidelity*, *stability*, *identity*, *separability*, and *time*, with each of the metrics measuring different aspect of quality of the explanation, on three commonly used general-purpose datasets (CIFAR10, SVHN and Imagenette), using models of three well-known state-of-the-art deep convolutional neural network architectures (VGG16BN, Resnet50, Densenet121) for each dataset. We analyse the experimental results and identify strengths and weaknesses of each explainability technique. To ensure transparency and repeatability we provide access to our source code and the complete results of our experiments in [Skl24].

The rest of this Thesis is organised as follows. In Section 2 we establish common terminology, recall existing taxonomy of interpretability techniques, as well as the different properties of the techniques and the possible ways to measure them. We finish the section with an overview of some other recent comparative evaluations of XAI techniques. Section 3 provides an overview of the XAI techniques that were considered in this Thesis. In Section 4 we describe the evaluation metrics that were used in this study. Section 5 describes the details of our experimental setup in terms of the datasets, the models used and the meta-parameters of the techniques. The experimental results are reported in Section 6 before we conclude in Section 7.

2 Background and Related Work

In this section we first aim to establish a baseline of understanding about the terminology used throughout this Thesis. We then recall the established taxonomy of existing XAI techniques. Next, we review the different desirable properties of explanations produced by the techniques and some of the ways they can be measured in practice. Lastly, we look at which some of the recent comparative studies for interpretability techniques.

2.1 Definitions and Terminology

Interpretability is a domain-specific notion that lacks an all-purpose definition [Rud19]. In a large part, this is because different contexts may require different types of explanations and no definition may capture all the possible contexts. In this work we will adopt the definition of interpretability as the "degree to which a human can understand the cause of a decision made by an ML model" [Mil19]. While at a glance interpretability may appear a binary property, recent research argues that it should be considered a multifaceted one, which can be broken down into a number of smaller characteristics [NTP+23] which we explore in Section 2.3. *Explainability* is a related term, which is often used interchangeably, or sometimes distinguished from interpretability, although there is no consensus what the distinction actually is [BDMZM21]. In this paper, for simplicity, we will use them interchangeably.

An *explanation* can be defined as a presentation of (aspects of) the reasoning, functioning and/or behavior (depending on the type of the explanation) of a machine learning model in human-understandable terms [NTP+23]. In this definition *reasoning* refers to the process on how a model came to a particular decision, *functioning* refers to the (internal) workings of the model, and *behavior* refers to how the model globally operates (without analysing the internal workings). Inclusive *or* indicates that a single explanation can fulfil multiple objectives at once.

2.2 Taxonomy of Explainability Techniques

Here we reference the taxonomy of XAI techniques present in literature such as [BGG+23], identifying main distinctions by which different techniques can be divided.

The first distinction we can make is between inherently interpretable models and post-hoc explanation techniques.

- *Inherently interpretable* models are a class of models which provide a decision in a form where the reason for that decision is directly accessible from the model structure. These models are usually constrained in model form, e.g. by feature sparsity, model simulatability and others. This class usually refers to "simpler" models, such as decision trees, linear models, and rules-lists, but also includes highly sophisticated models, such as neural networks with interpretability mechanisms (e.g. attention mechanism) baked in during training process.

- *Post-hoc* explanation aim to provide explanations for black-box model after it has been trained, and without changing the model. This may, for example, involve building a separate post-hoc model to explain the output of the original black-box model.

Post-hoc techniques can be further subdivided into global and local techniques.

- *Global* techniques aim at explaining the overall logic of a model.

- *Local* techniques explain the decision of the model in a specific instance.

Another division of the post-hoc methods splits them into model-agnostic and model-specific.

- *Model-Agnostic* techniques can be used to provide explanation for any type of black-box model. This way the explanation is decoupled from the inner structure of the ML model, allowing us to change the underlying model independently of the interpretation technique.

- *Model-Specific* techniques can be used to provide explanation only for a specific type of black-box models, most commonly Neural Networks.

Furthermore, it is possible to divide post-hoc explainability techniques by type of explanation provided. Different data types may require different types of explanation. For image data the subdivision is as follows.

- *Saliency Maps* techniques provide explanation in form of a map which highlights the importance of each input pixel.

- *Concept Attribution* techniques show the importance of a "concept" to the model output. For example, the importance of the presence of stripes to a prediction of a zebra. The concept can be provided by a user or derived automatically from the input.

- *Prototype* techniques provide a series of images that exemplify the predicted class.

- *Counterfactual* techniques provide a set of examples that are similar to the input image, but predicted as a different class; or showcase the minimum modifications required for the input to be predicted as a different class.

The focus of this Thesis - saliency map techniques - can be even further subdivided by the mechanism the importance of the pixels is computed. [GM23].

- *Perturbation-based* techniques calculate the importance of a pixel by measuring the output change when perturbing this the input. Usually these methods partition the input image into super-pixels and measure their importance, instead of doing this for each pixel.

- *Activation-based* techniques are based on weighting the feature maps produced at the last layer of the CNN to explain the predicted class. These methods generate saliency maps with the resolution of the feature maps at the final layer, which are then upscaled to the resolution of the input, resulting in very coarse maps.

- *Backpropagation-based* techniques consist of backpropagating information from the output of the model to the input image to yield a high-resolution saliency map, which indicates the impact of each pixel on the decision.

The detailed description of each method used in this Thesis is provided in Section 3.

2.3 Evaluation of Explainability Techniques

Since the emergence of XAI, the number of explainability techniques has been increasing steadily. Naturally, to answer the demand for *quality* explainability techniques, a multitude of ways to evaluate them has been developed, and more and more papers that introduce new techniques include evaluations of their methods [NTP$^+$23].

To systematise evaluation methods, [DVK17] proposes three major categories:

- *Function-grounded* metrics evaluate the interpretability technique via some proxy tasks. These metrics do not require humans and as such can be automated. Defining a good proxy task can be a challenge and depends on the context and the requirements of the specific explanation task. One example authors provide for this category is evaluating a technique w.r.t another one, already proven to be of high quality in human-based studies.

- *Application-grounded* metrics rely on human experts to validate the explanation, depending on the context of the task. These are usually employed in critical domains. An obvious example is employing doctor experts to validate explanations produced in medical setting.

- *Human-grounded* evaluation methods evaluate explanation through non-expert humans. These are most appropriate to evaluate general qualities of the explanation.

Another way to classify evaluation methods is by looking at which desirable traits of an explanation they measure and how. Authors of [NTP$^+$23] surveyed 361 papers that introduce, apply or evaluate one or more interpretability methods and structured the commonly evaluated properties into twelve distinct categories. These properties, the authors refer to as Co-12, cover different aspects of explanation quality and can be quantitatively evaluated independently from each other. The paper summarizes the methods which have been used to functionally evaluate each of them, although, naturally, some properties are easier to evaluate, or generalize evaluation for, than others.

The following categories are related to the qualities of content of the explanation and the functional qualities of the explanation. This is the most commonly evaluated category with several evaluation approaches existing for each property.

- *Correctness* (also fidelity [RŠB18], truthfulness [DVK17] or faithfulness [LSL$^+$20]) describes how well the explanation approximates the prediction of the black box model. High correctness is always desired, otherwise the interpretability technique becomes counter-productive to explaining the behavior of the underlying black-box model and becomes misleading instead. The most popular way to evaluate this property is by single or incremental deletion, which involves deleting or obscuring input deemed important by the XAI technique and measuring correlation in output of the original black-box model and the explanations. A variation of this is incremental insertion where features are incrementally added. Another popular way - is by using a synthetic dataset that follows a particular reasoning and checking if the produced explanations follow the same reasoning.

- *Completeness* addresses the extent to which the explanation explains the predictive model. High completeness is desired, but should be balanced with compactness and correctness as to not overwhelm the user. Output-completeness, i.e. whether the explanation holds enough data to explain output of the model, can also be measured by single or incremental deletion.

- *Consistency* (also identity in [Hon18]). Describes how deterministic and invariant to implementation the explanation method is. The idea is to confirm that the identical inputs have identical explanations, regardless of the underlying black-box model. One way to do this is by measuring the difference in explanation of the same images using the same technique on the same or on different black-box models.

- *Continuity* (also stability in [Hon18, AMJ18]) describes how continuous and generalizable the explanation function is. Small variations in the input should not drastically change the explanation. In essence, similar inputs should have similar explanations. A popular way to evaluate this in practice is to measure the difference in explanation after slightly perturbing original input, for instance adding slight noise.

- *Contrastivity* denotes whether the explanation is discriminative toward the other targets or events, or in other words, whether it answers the question of “why this prediction was made instead of another prediction?" [CPC19]. This generally refers to a class of explainability techniques called counterfactual explanations. This is also related to the separability axiom introduced by Honnager in [Hon18], which denotes that dissimilar instances should have dissimilar explanations. A straightforward way to measure this in case of saliency maps is to compare explanations for different targets or logits: different targets should have different explanations.

- *Covariate complexity* describes how complex the (interactions of) features in the explanation are. Features should be comprehensible [CPC19], while non-complex interactions between features are desired [DVK17]. For image data this can be measured by calculating overlap between human-interpretable concepts (which can be manually labeled) and the features in the explanation.

The next three categories are related to presentation aspects of the explanation. These are much less frequently evaluated and can be difficult to evaluate objectively in some cases.

- *Compactness* (selectivity in [CPC19]). Describes the size of the explanation. Explanations should be sparse, short and not redundant to avoid presenting an explanation that is too big to understand [CPC19]. This is commonly evaluated by measuring explanation size or sparsity.
- *Composition* considers the presentation format and organization of the explanation [CPC19]. This property is about "how" something is explained as opposed to "what" is explained. Some formats are generally more comprehensible than others. This quality is difficult to objectively evaluate in most cases, but for text data the authors identified the measure of perceptual realism, i.e. how believable a generated explanation is compared to real text samples.
- *Confidence* describes whether the probability information is present in the explanation as well as its accuracy. This measure can include both the presence of the confidence of the original model's prediction [CPC19], as well as the confidence of the explainability technique itself [Mil19]. Presence of this information is usually a design choice for a XAI researcher. It is rarely explicitly evaluated, but the process of measuring it is comparable to single incremental deletion.

These final three categories are related to the way the user interacts with the explanation. These are most often evaluated with user studies, but there are some quantitative methods for evaluating these as well.

- *Context* addresses the degree to which the user's level of knowledge and expertise should be taken into the account. It is desirable that the explanations are relevant to the user [Mil19] and this property indicates whether the type of stakeholder is taken into account. One method to evaluate this quality involves designating some features as “untrustworthy” based on domain knowledge and measuring changes in explanations after removing those features.

- *Coherence* describes the degree to which the explanation is consistent with the user’s prior knowledge, beliefs and general consensus [Mil19]. This is related to plausibility and it is important not to conflate it with correctness, but evaluate separately. This is one of the most commonly measured properties. For image data, “location coherence” is often assessed by comparing a heatmap or localization explanation to ground-truth object bounding boxes, segmentation masks etc.

- *Controllability* indicates whether the user can interact with the explanation, correct or otherwise control it, which can also be interpreted as the social aspect of the explanations [Mil19]. The presence of user-tunable (hyper-)parameters is usually a design choice (and not a common one) and is not evaluated often. As an example authors present a paper that measured accuracy of textual explanations after iterative user feedback.

According to Nauta et al. [NTP+23], 90% of papers that introduce a new XAI method perform quantitative evaluation by three or fewer properties. Coherence and Output-completeness are evaluated most often, followed by Correctness, Compactness and Covariate complexity.

2.4 Related Studies

In this section we examine some recent comparative studies of explainability techniques for image data to see which techniques are being evaluated by which metrics, as well as which models and datasets are used. For each metric evaluated we will reference it with a property mentioned in Section 2.3.

Bodria et al. [BGG+23] identified and employed a number of metrics to evaluate various interpretability techniques on image, tabular and text data.

- Fidelity (or correctness) is noted as one of the most popular metrics in literature for all types of data and techniques. For image data authors evaluate fidelity using incremental deletion and insertion methods and measuring area under curve (AUC).

- Another popular metric is stability, which measures the consistency of explanations for similar records. Authors reference Lipschitz constant [AMJ18] that can be used to evaluate stability.

- Accuracy, precision and recall, as well as running time are also measured. These are commonly used for evaluating quality of both interpretability techniques and other ML models.

Specifically, authors evaluate fidelity for nine saliency map techniques, including LIME, IntGrad, SmoothGrad, GradCAM and GradCAM++, across 100 images from each of the MNIST, CIFAR10 and Imagenet datasets, using custom CNN models for MNIST and CIFAR10 and VGG16 model for Imagenet. Stability, while mentioned, is evaluated only for text and tabular data, not images.

Li et al. [LSL^{+}20] selected five metrics to evaluate and compare different saliency map methods:

- Faithfulness (or correctness) was evaluated by incremental insertion and measuring iAUC - area under insertion curve.

- Localization (as a measure of coherence) was evaluated via Pointing Game proposed in [ZBL^{+}18] on a dataset with known ground truth, by counting the ratio of data samples in which the pixel with the highest relevance score lies in the bounding box. This metric is quite loose, since counting a single pixel as a hit can only be accurate if the bounding box is small enough compared to the image size. To amend this authors propose to instead measure the proportion of the overlap between the salient area and the ground truth to the salient area.

- False-positives, as another aspect of correctness, were measured by performing checks against a synthetic dataset with known ground truth developed in [YK19]. Two metrics were calculated as proposed in the original paper: Model Contrast Score (MCS), Input Dependence Rate (IDR).

- Class Sensitivity (as a measure of contrastivity) check was performed by measuring the difference between explanation for highest and lowest confidence classes in the dataset. The idea here is that a good explanation method would have dissimilar explanation for different classes.

- Stability (continuity) was measured by perturbing input and measuring the difference in the output of the explanation technique.

Authors use these five metrics to evaluate seven SM techniques, including IntGrad and GradCAM, on images from MS COCO and VOC datasets, using VGG16 and Resnet50 black-box models.

Huber et al. [HLA22] focused on evaluating correctness of different saliency map techniques using the following metrics:

- Sanity checks for saliency maps, as first proposed in [AGM^{+}18] involving model parameter randomization check – randomly perturbing the internals of the predictive model and checking whether the explanation changes.

- Incremental insertion method, measuring area under insertion curve was also used.

The experiments were performed using a deep reinforcement learning black-box model and datasets from four Atari games.

Cerekci et al. [CAD$^+$24] and Saporta et al. [SGA$^+$22] also employed Pointing Game [ZBL$^+$18] method to compare coherence of several XAI techniques on medical datasets (mammograms and chest X-Rays respectively) with expert-marked ground truth. In the latter the authors also present a modification of this method, measuring the overlap between the saliency method segmentation and the expert segmentation, along with the less strict hit rate evaluation which was proposed in the original paper.

To conclude, while there have already been a number of comparative studies evaluating the quality of explainability techniques, they primarily focus on fidelity, like in [BGG$^+$23, HLA22] or coherence, like in [CAD$^+$24, SGA$^+$22]. However, we believe that this alone is insufficient to measure the overall quality of the technique, since, as mentioned in Section 2.3, it is a multifaceted notion. Additionally, the practice of standardizing the measurements on different model architectures across different datasets, i.e. using each model type on each dataset and vice versa is not widely applied, with only [LSL$^+$20] performing evaluations in this way. To avoid overlaps with the latter study, we differentiate our work by selecting different datasets and different techniques for our experiments, and use different methods to calculate the metrics.

3 Reference Explainability Techniques

In this section we provide a short overview of the post-hoc explainability techniques for image data considered in this Thesis. Examples of the explanations produced using each metric are presented on Figure 2

3.1 LIME

Local Interpretable Model-agnostic Explanations (LIME) [RSG16] is a local model-agnostic technique that exploits local surrogate models and can be used to produce saliency maps when used on image data, but can also be used on tabular and text data, which is a rare quality. It is a perturbation-based technique, which in case of image data first segments the input image into superpixels, then perturbs the image by replacing different superpixels with solid, usually neutral color. The perturbed variations are then fed to the black-box model and a linear surrogate model is learned on top, which is weighted by the distance of the perturbed variants to the original image. One of the biggest drawbacks of this technique is the instability of the explanations it produces. Since the technique samples perturbed instances around the original image randomly, it may produce different explanations for the same image. Another significant drawback is computational power required to produce an explanation, though this can be configured, albeit with probable loss of fidelity.

When used with images the segmentation algorithm is crucial to the quality of the resulting explanation. For small resolution images a badly-configured segmenter may result in superpixels the size of the entire image which results in meaningless explanations.

3.2 SHAP

SHapley Additive exPlanations [LL17] is a local-agnostic explanation method which has both model-agnostic and model-specific variations and, like LIME, can be used on images, text and tabular data. It aims to explain the prediction of an instance by measuring the contribution of each feature towards the prediction. The method is based on the concept of Shapley Values from coalitional game theory. SHAP treats features (or groups of features, e.g. superpixels) of the input as players in a coalition acting together to change the outcome (the prediction) and tells us how to fairly distribute the "payout" (the impact towards the prediction) between the players. The Shapley value explanation can be represented an additive feature attribution method, a linear model. Formally SHAP specifies explanation as:

$$g(z') = \phi_0 + \sum_{M}^{j=1} \phi_j z'_j$$

where g is the explanation model, $z' \in \{0,1\}^M$ is the simplified features, M is the maximum coalition size and $\phi_j \in R$ is the feature attribution for a feature j, the Shapley values. To compute such explanation we simulate all combinations of some of the features playing (being present) and some not (being absent). One can notice here similarity to LIME, but wheres LIME assumes linearity in local model behaviour (which is not guaranteed), SHAP guarantees a fair distribution of the "payout" among the features. This in theory should result in more trustworthy and stable explanations. Like LIME, SHAP is also quite computationally intensive.

SHAP can produce a number of models that differ in how they approximate the computation of Shapley values. In our experiments we opted to use `PartitionExplainer`, which is a model-agnostic estimator. It computes Shapley values recursively based on a hierarchy of features that defines feature coalitions. Unlike the `KernelExplainer` the paper originally came out with that has an exponential runtime, `PartitionExplainer` has a quadratic runtime, which makes it more suitable for use on image data.

3.3 IntGrad

Integrted Gradients [STY17] is a local model-specific, data-agnostic explainability technique. It utilises gradients of a black-box model and as such can only be applied to differentiable models. Given an input I and a baseline I', IntGrad constructs a straight path from I' to I and computes gradients of the points along the path. In case of images, the points are taken by overlapping I on I' and gradually modifying the opacity of I. Finally, Integrated Gradients are obtained by cumulating the gradients of these points. Given a baseline I', the explanation for an image I for model f is formally defined as:

$$g(f, I, I') = (I - I') \int_{a=0}^{a=1} \frac{\partial f(I' + \alpha(I - I'))}{\partial I} d\alpha$$

The choice of the baseline can drastically affect the resulting explanations. A good baseline should be neutral, representing an absence of an object of any class. Typically a constant black or white image is used, though in other cases it can be beneficial to use a different kind of baseline that better represents a status quo in the image. Using a black baseline image, can lower the importance of dark pixels in the source image, while using white baseline can lower the importance of light pixels.

3.4 SmoothGrad

SmoothGrad [STK$^+$17] is a local model-specific, data-agnostic technique. It is not a standalone technique, but can be used with any Gradient-based explainability method.The saliency maps produced by Gradient-based techniques such as IntGrad tend to be quite noisy, especially for pixel-wise explanations. This is due to high fluctuations in the derivatives at small scale. SmoothGrad overcomes this by adding Gaussian noise to the input image and averaging the output of the underlying Gradient-based technique. Formally, given a saliency method $f(x)$ which produces a saliency map s, its smoothed version $\hat{f}$ can be expressed as:

$$\hat{f} = \frac{1}{n}\sum_{1}^{n} f(x + \mathcal{N}(0, \sigma^2))$$

where n is the number of samples, and $\mathcal{N}(0, \sigma^2)$ is the Gaussian noise.

The amount of noise that has to added depends on the dataset and model architecture, but the authors suggest setting noise level at 10-20%.

3.5 GradCAM

Gradient-weighted Class Activation Map [SCD+17] is a model-specific local explanation technique for image data. GradCAM uses the class-specific gradient information flowing into the target convolutional layer of a CNN model (as opposed to flowing back to input as in Gradient methods) to assign saliency values to each neuron for a particular decision and produce a coarse localization map of the important regions in the image. An overview of the method is presented on Figure 1.

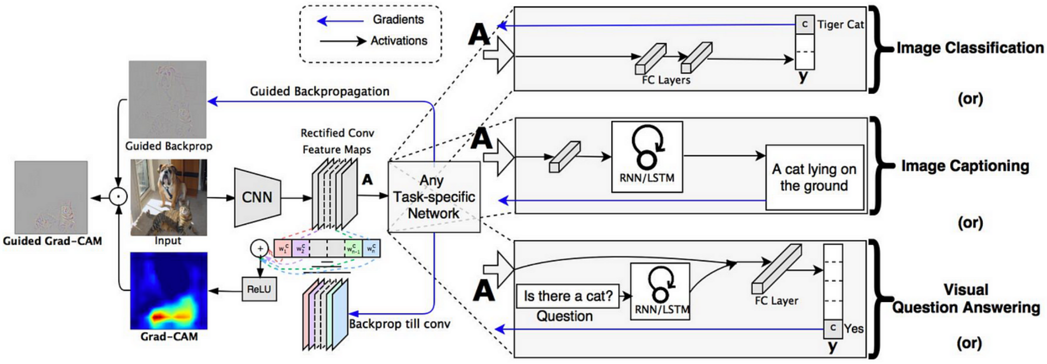

Figure 1. Grad-CAM Overview: Given an image and class of interest, the image is forward-propagated through the CNN network and the activation maps for the layers of interest are obtained. The gradients are set to zero for all classes except for the desired class, then backpropagated to the rectified convolutional feature maps of interest, which are combined to compute the coarse Grad-CAM heatmap. [SCD+17]

First, for each feature map A^k in the last convolutional layer a gradient with respect to a predicted class c, y^c is calculated, which can formally be expressed as:

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where weight a_k^c is the importance of feature map A^k for calss c. Then a weighted combination of the resulting tensors is performed, followed by ReLU (Rectified Linear Unit activation functio) operation which produces the final heatmap:

$$g(I, f)_c = ReLU(\sum_k a_k^c A^k)$$

For this technique and its derivatives the choice of the target convolutional layer can significantly change the result. Typically the final convolutional layer is chosen, since it contain the most relevant high-level information, but in some cases choosing a higher layer, or even averaging across several layers may yield better results.

3.6 GradCAM++

GradCAM++ [CSHB18] is an extension of GradCAM solving some of the issues present in that technique. Since GradCAM aggregates the feature map gradients all with the same weight, the highlighted areas may not capture the whole object and if multiple instances of the same object are present, the method does not highlight them all. GradCAM++ overcomes these shortcomings by explicitly assigning different weights α_c^k to gradients of different activation maps when aggregating them:

$$a_k^c = \sum_i \sum_j w_{ij}^{kc} ReLU(\frac{\partial y^c}{\partial A_{ij}^k})$$

Additionally, ReLU is applied to each gradient before aggregating, the idea being that for visualisation purposes the positive gradients are preferred to highlight the importance of features, rather than negative ones that suppress the neuron activation.

4 Quantitative Metrics for Assessing Explanations

Here we provide an overview of the qualitative metrics and methods that were used in this Thesis to evaluate quality of the saliency map explanations generated by the XAI techniques described in Section 3. All metrics we selected are agnostic to the evaluated saliency map technique, the dataset and the black-box model. They are also function-grounded, meaning they can be automated and do not require human involvement, including for labeling ground truth.

4.1 Fidelity

Fidelity, or correctness, of the explanation refers to how well the explanation approximates the behaviour of the underlying black-box model, i.e. whether the features (pixels or superpixels) highlighted by an explanation method are truly important to the ML model. The most popular way to evaluate this is by modifying the input and measuring changes in the output of the black-box.

In our experiments we opted to use the *deletion* method, first proposed in [PDS18], also referred to in other studies as area under deletion curve, *dAUC*. In this approach the pixels are incrementally deleted from the original image in the order of decreasing importance as indicated in the explanation. After each deletion the modified image is fed to the black-box and the probability of the originally predicted class is recorded. Once the whole image has been masked the scores are normalised and plotted as a function of the proportion of the pixels removed. The *dAUC* is defined as the area under curve (AUC) of this graph. The intuition of this method is that removing the features important to the decision would result in early and drastic decrease in the prediction score, minimising the AUC.

The biggest advantage of this method is that it requires no additional ground truth labeling. This means that it can be easily deployed and automated. One disadvantage of this method is the high number of inferences required, which can potentially result a lot of computation. For example, for a 224 x 224 resolution image 50176 inferences would be needed. To alleviate this, in our experiments we have performed the deletions of pixels in batches equal to the highest dimension of the image. The second disadvantage of this method is the possibility of data shifts, since removing pixels from the images could drag the input out of the learned distribution of the black-box model. Further research is required to tackle this issue.

4.2 Stability

Stability, or continuity, refers to how generalizable the explanation function is, i.e. whether similar input images have similar explanations. To measure this quality we utilised the Lipschitz constant as proposed in [AMJ18]:

$$L_x = max \frac{\| e_x - e_{x'} \|}{\| x - x' \|}, \forall x' \in \mathcal{N}_x$$

where x is the explained instance, e_x the explanation and $\mathcal{N}_x$ is a neighborhood of instances x' similar to x. In other words, we measure the maximum distance between the explanations for similar images in a neighborhood, and the explanation for the original image, bounded by the distance between the similar images and the original image. In our experiments to construct the neighborhood of similar images, we considered five images from the evaluated dataset with closest Euclidian distance to the original image.

4.3 Identity

Identity, or consistency, describes how deterministic the interpretability technique is. One simple way to measure this is by checking whether the technique generates the same explanations for the same input [Hon18]. In our experiments we calculate this by generating saliency maps twice for every image in the dataset and checking whether the distance between two explanations for any particular image is zero. More deterministic XAI techniques should produce higher proportion of identical explanations, ideally a 100%. Since this is a binary metric, it only makes sense to evaluate as a mean across many images.

4.4 Separability

Separability, or contrastivity, aims to evaluate how discriminative an explanation towards other targets that are not the same. To measure this in our experiments, we utilise the method proposed in [Hon18]: comparing the explanation of each image in a dataset of unique inputs to explanations produced for all other inputs from the same dataset. More discriminative techniques should produce higher percentage of distinct explanations. This metric is also binary and should be averaged across a large number of images.

4.5 Time

Finally we evaluated the the average time it took to produce explanation for each of the XAI techniques as a measure of computational complexity. Some techniques (especially perturbation-based) require a lot of compute, which may become a bottleneck in deployment.

5 Experimental Setup

In this section we provide details about the datasets, the ML models and the XAI techniques we used in this experimental study. All experiments were done in Google Colaboratory [Bis19] using standard Python runtime environment with T4 GPU hardware accelerator. Our code along with the results in full is available online [Skl24].

5.1 Datasets

We performed our experiments using three popular, RGB, general-purpose, image classification datasets in this study:

- CIFAR10 (Canadian Institute for Advanced Research) [Kri09], a subset of Tiny Images dataset [TFF08], composed of 32x32 images belonging to 10 different classes, such as "airplane", "cat" or "truck".
- Imagenette [HG20], a subset of ImageNet dataset [DDS+09], with images of 10 easily classified classes, e.g. "tench", "gas pump", "parachute". Imagenette contains images of varying resolution which we resize to 224x224.
- SVHN (Street View House Numbers) [NWC+11], a digit classification benchmark dataset containing 32x32 images of printed digits (from 0 to 9) cropped from pictures of house number plates.

For each dataset a total of 1000 images were evaluated, 100 images of each class. All datasets are available for download through `PyTorch` machine learning library [PGM+19].

5.2 Models

For the underlying black-boxes to be explained we used three types of well-known, deep convolutional neural network models for each dataset:

- VGG16BN [SZ14] is a CNN composed of 16 weighted layers: 13 convolutional layers and 3 fully-connected layers. The BN variation features batch-normalisation layers after each convolutional layer. This network is well-known for its simplicity and efficiency in image classification tasks. It is a relatively large network, having about 138 million parameters.
- ResNet50 [HZRS16] in a network composed of 50 layers, including 16 residual blocks, with each block being composed of several convolutional layers with residual connections. It's main innovation is the use of skip connections to mitigate the vanishing gradient problem, allowing the training of very deep neural networks. ResNet50 has approximately 25.6 million parameters.

- DenseNet121 [HLVDMW17] is a CNN characterised by its dense block structure, with each of its layers being connected to every other layer in a feedforward fashion. Additionally, to help reduce the number of parameters, it uses bottleneck layers. This achieves the goal without reducing the number of features learned by the network and DenseNet121 has around 7.5 million parameters.

Pretrained models for CIFAR10 and SVHN datasets were sourced from the `detectors` package [Dad23], while for Imagenette we used `PyTorch` models pretrained on ImageNet. Models for CIFAR10 have 94% accuracy, for SVHN - 96%, and ImageNet-trained models we used for Imagenette were 82-88% accurate on that dataset.

5.3 Explainability Techniques

Implementations for explainability techniques were sourced from public repositories: LIME from the official repository referenced in [RSG16], SHAP from the official repository referenced in [LL17], GradCAM and GradCAM++ from [Gc21], IntGrad and SmoothGrad from [KMM+20].

All techniques were used with default settings, the settings recommended in the original paper, or the official documentation for the implementation. For LIME the number of samples was kept at 1000, the occlusion colour was kept at "the mean colour of the input image", and no custom segmenter was provided. For PartitionExplainer in SHAP the `blur` type masker was used. For GradCAM and GradCAM++ we used the following layers as target layers: for VGG16BN model - the second-to-last convolutional layer (using the final layer didn't generate any explanation for smaller images), for Resnet50 and Densenet121 - the respective final convolutional layers. For IntGrad and SmoothGrad we used constant black image as baseline. For SmoothGrad we set the amount of noise to be added to 20%, and the number of evaluations with noise was kept to the default five. For all techniques the explanation was generated for the highest scoring class. Lastly, for all explanations the negative contributions were not considered and instead set to zero or disabled, and the remaining importance values were scaled to between zero to one to ensure comparability between different explanations.

6 Quantitative Comparison and Results

In this section we present the results of our experiments comparing XAI techniques referenced in Section 3 using metrics from Section 4 on datasets and pretrained black-box models described in Section 5. We first go over the results for each of the metrics, presenting a summary with aggregated scores at the end. For all metric lower score is better.

In Figure 2 we visualise some examples of typical saliency maps produced by each of the techniques on sample images from the evaluated datasets.

6.1 Fidelity Results

Following the definition of the dAUC metric in Section 4.1 proposed in [PDS18] we performed experiments on the pretrained models to evaluate their fidelity w.r.t. the black-box model. The results, averaged across 1000 images from each dataset are presented in Table 1.

Table 1. Fidelity results using dAUC metric on images from CIFAR10, SVHN and Imagenette datasets with VGG16BN, ResNet50 and Densenet121 pretrained models. The best scores are highlighted in bold.

		CIFAR10	SVHN	Imagenette
VGG16BN	LIME	0.432	0.489	0.113
	SHAP	0.248	**0.207**	0.067
	GradCAM	0.503	0.470	0.084
	GradCAM++	0.504	0.472	0.091
	IntGrad	**0.195**	0.276	**0.054**
	SmoothGrad	0.204	0.288	0.055
Resnet50	LIME	0.520	0.493	0.130
	SHAP	0.320	0.249	0.087
	GradCAM	0.327	**0.242**	0.129
	GradCAM++	0.334	0.243	0.131
	IntGrad	0.227	0.290	**0.069**
	SmoothGrad	**0.223**	0.287	0.072
Densenet121	LIME	0.374	0.385	0.116
	SHAP	0.227	**0.183**	**0.082**
	GradCAM	0.269	0.204	0.114
	GradCAM++	0.272	0.203	0.116
	IntGrad	**0.168**	0.194	0.086
	SmoothGrad	**0.168**	0.189	0.090

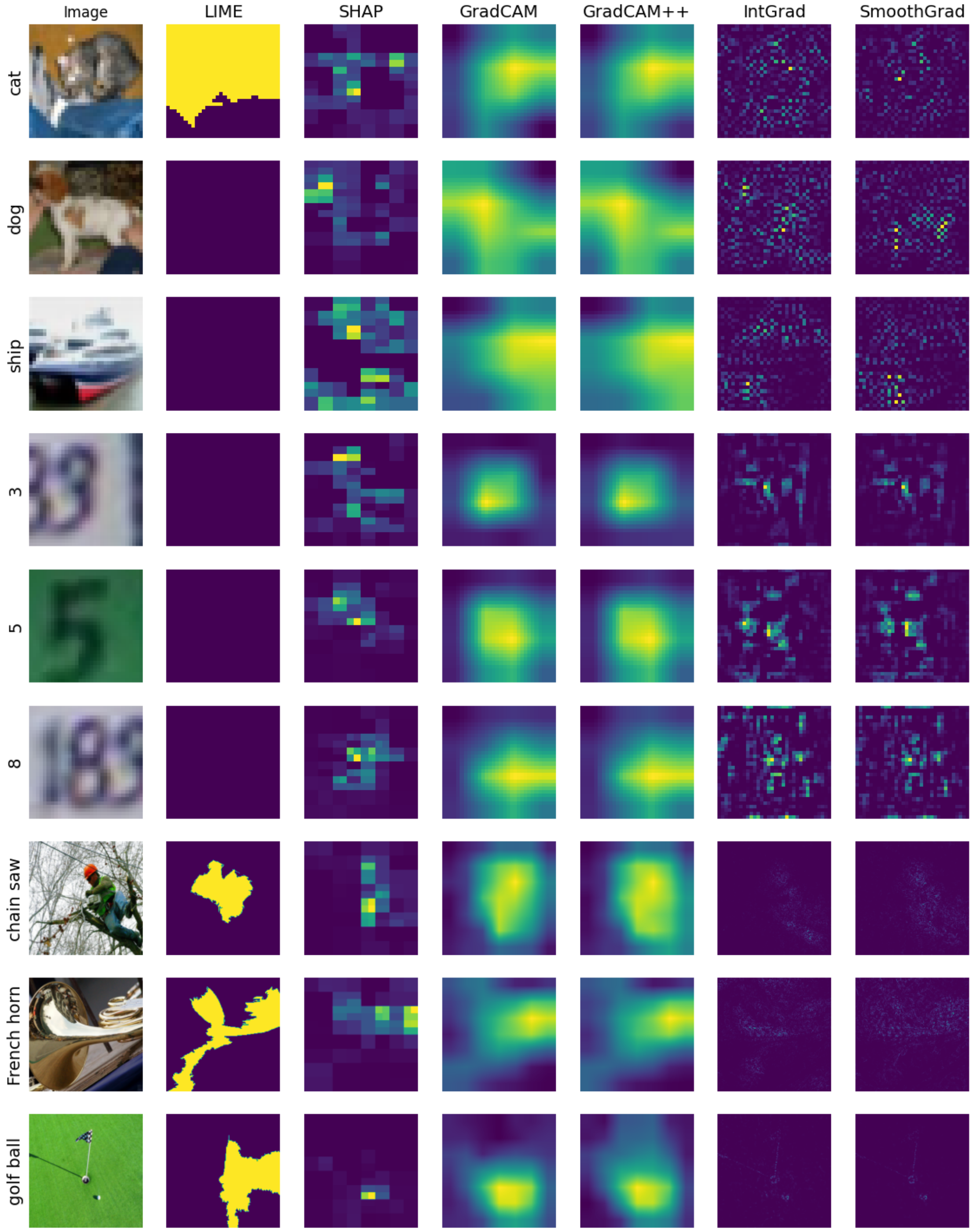

Figure 2. Examples of explanations produced by techniques described in Section 3 on the evaluated datasets described in Section 5.1. The black-box model used is Resnet50. The first column is the original image from the dataset labeled with class as predicted by the model.

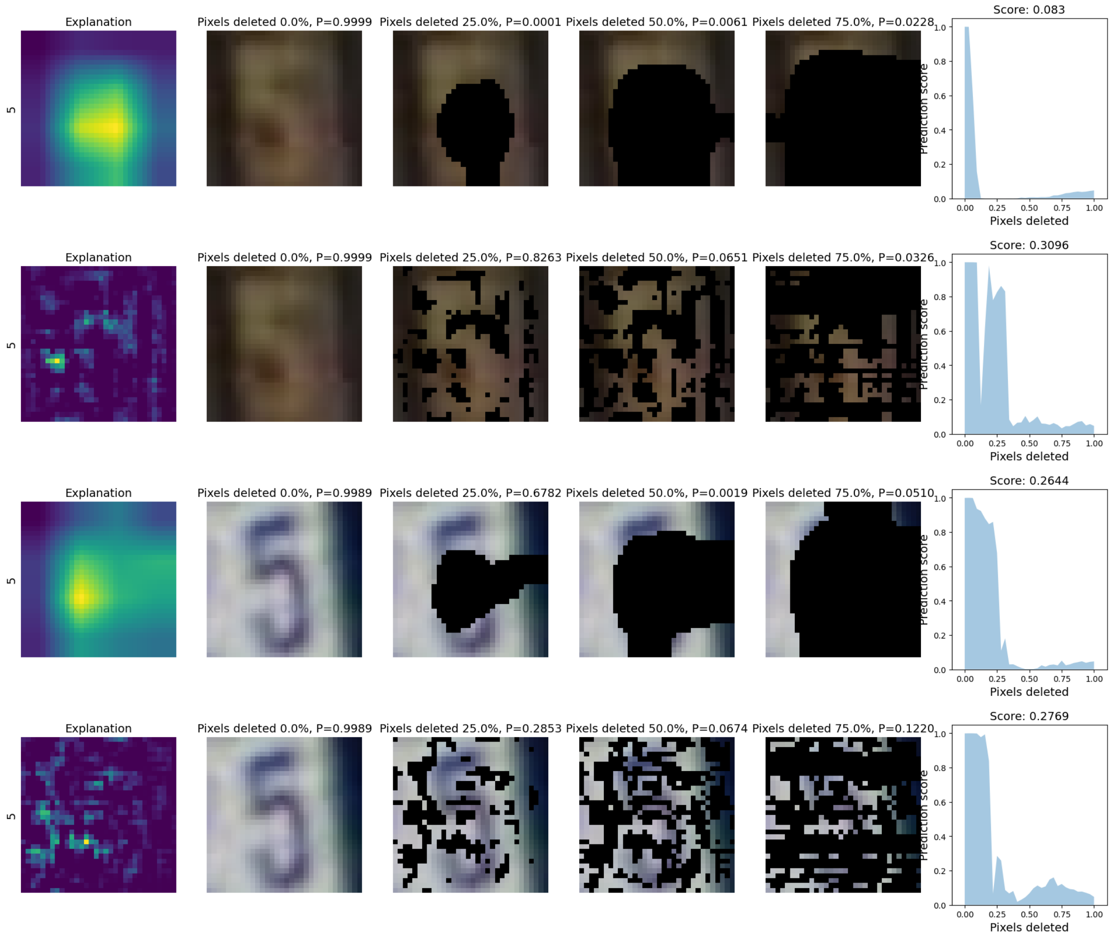

Figure 3. Examples of the deletion process for GradCAM and IntGrad on SVHN for white digits on dark background vs dark digits on white backgound. First column contains the explanation and the image label predicted by the model. Columns 2-5 show intermediate states of the deletion process. The final column shows the AUC chart and the final score for the explanation. The black-box model is Resnet50.

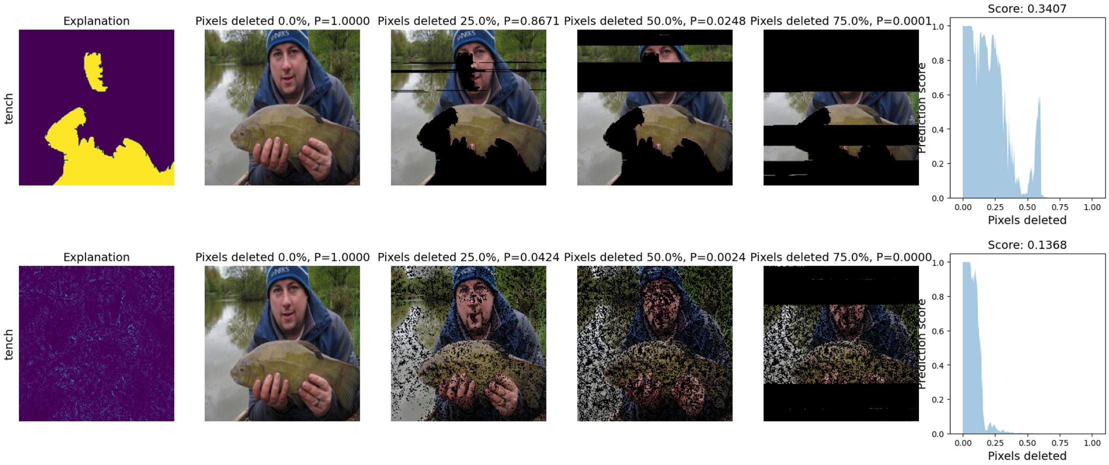

Figure 4. Examples of the deletion process for LIME and SmoothGrad with Resnet50 on Imagenet. First column contains the explanation and the image label predicted by the model. Columns 2-5 show intermediate states of the deletion process. The final column shows the AUC chart and the final score for the explanation.

From Table 1 we can see that IntGrad outperformed the other methods in most cases with SmoothGrad offering comparable performance. While the saliency maps they produced were noisy and sparse, removing the pixels they highlight resulted the sharpest drop in dAUC across all the models in CIFAR and most models in Imagenette, while on SVHN the two were overtaken by other methods. SHAP scored best on SVHN when using VGG16BN or Densenet121 and on Imagenette with Densenet121, but also performed well on the other two datasets, placing just behind SmoothGrad and IntGrad. GradCAM and GradCAM++ placed second-to-last in most experiments, with GradCAM scoring usually slightly higher. They performed particularly poorly with VGG16BN on CIFAR and SVHN, where their saliency maps highlighted sections of the image far from the actual object. This could be because VGG16BN is a more shallow model compared to the Resnet50 and Densenet121 and its final layers and not as discriminative. Either way, it highlights the dependence on the model structure present in these methods. Finally, LIME ranked worst in most cases. The explanations it produced were very coarse, highlighting large portions of the image. In particular, on CIFAR and SVHN, due to to poor performance of the default segmenter, it often produced explanation of constant value, which caused the deletion process to delete pixels essentially at random. However, on Imagenette, where the segmenter worked as intended, it performed much better, scoring close to the two activation-based techniques on Resnet50 and Densenet121.

Although this metric is a popular method to measure fidelity, in practice we can notice some problems. Namely, the choice of the occlusion colour is an important one and does affect the results. In our experiments we used solid black, but replacing dark colours with solid black may not drastically affect model confidence. This is especially noticeable on SVHN dataset, which contains images of digits on contrasting background. If the explanation focuses on a dark digit on white background - replacing it with black colour did not affect the model prediction that much. That is, unless the saliency map is very coarse, which explains relatively high scores for GradCAM and GradCAM++ (assuming they are at least somewhat localised on the object, like on Resnet50 and Densenet121). On the other hand, IntGrad and SmoothGrad often highlighted the pixels around the digit. In case of dark background, replacing them with black colour also did not reduce model confidence as sharply. Some examples of the phenomenon are provided on Figure 3.

Additionally, we can note how moving from lower resolution images of CIFAR and SVHN to higher resolution images of Imagenette resulted in sharp drop in all dAUC scores. This could be because larger images contain more data for model to consider and that makes it easier to reduce accuracy, or induce an out-of-distribution situation. Accuracy of the model and the number of classes the model was trained on could also play a role here. Additional examples for Imagenette are provided on Figure 4

6.2 Stability Results

Following Section 4.2, to evaluate the stability of XAI techniques we used the Lipschitz constant proposed in [AMJ18]. Specifically, we measured the maximum discrepancy between the explanation for the original image and the explanations for the five closest different images from the evaluated dataset, bounded by the difference between the explained images. The averaged results across 1000 images from each of the three datasets using each of the three pretrained models are presented in Table 2.

From the results presented in Table 2 we can see a similar distribution to the fidelity metric. SmoothGrad and IntGrad score best in all cases but one. This can be explained by the sparsity of the saliency maps they produce. Since they highlight select pixels, and not large parts of an image like other techniques, the distances between the explanations are smaller, even if the explanations highlight completely different parts of the image. SHAP scores third behind the tho gradient-based techniques on CIFAR and Imagenette, as well as SVHN with VGG16BN. GradCAM and GradCAM++ again score worst when using VGG16BN on smaller images, but perform only slightly worse than SHAP in other cases, even outperforming it on SVHN with Resnet50 and Densenet121. Between the two, GradCAM++ consistently shows slight improvement in stability over GradCAM. LIME shows the lowest stability among the evaluated techniques on Imagenette, and non-VGG16BN CIFAR10, but scores surprisingly well on SVHN. Upon closer inspection this is because, due to the poor performance of the segmenter. Over 80% of the explanations produced for SVHN ended up being of constant value, so in a lot of cases the distance between explanations for all of the closest images equaled zero.

Looking at the results we can note that this evaluation method seems to be heavily

Table 2. Stability results for XAI techniques using the Lipschitz constant. The best scores are highlighted in bold.

		CIFAR10	SVHN	Imagenette
VGG16BN	LIME	1.341	1.048	1.445
	SHAP	0.797	1.264	0.680
	GradCAM	1.568	3.031	0.781
	GradCAM++	1.561	3.012	0.762
	IntGrad	0.450	1.061	0.172
	SmoothGrad	**0.447**	**1.054**	**0.164**
Resnet50	LIME	1.366	1.821	1.477
	SHAP	0.817	1.291	0.668
	GradCAM	0.952	1.170	0.901
	GradCAM++	0.872	1.077	0.884
	IntGrad	**0.469**	0.993	**0.159**
	SmoothGrad	0.484	**0.959**	0.160
Densenet121	LIME	1.626	**0.461**	1.463
	SHAP	0.892	1.346	0.692
	GradCAM	0.997	1.231	0.885
	GradCAM++	0.940	1.168	0.866
	IntGrad	0.513	1.015	0.184
	SmoothGrad	**0.491**	0.988	**0.163**

dependent on the sparsity of explanation the XAI technique produces. Techniques like IntGrad and SmoothGrad that highlight fewer pixels, or highlight them with more gradient will score higher than techniques like LIME that produce binary saliency maps highlighting large parts of the image, even though they may highlight roughly the same area of the image containing the object of interest. Lastly, since in this approach we generate the neighbourhood of similar images from the set of images that are evaluated, it may not be reliable for small-scale evaluations.

6.3 Identity Results

As suggested in [Hon18] that we referenced in Section 4.3 we measure the consistency of the model by generating explanations twice for each image and comparing them to each other. We do this for a 1000 images in each dataset and the percentage of non-identical explanations for each technique, dataset and model is provided in Table 3.

From the results in Table 3 we see that SHAP, GradCAM, GradCAM++ and IntGrad all achieve perfect score on this metric, meaning they are completely deterministic and always generate the same explanation for the same model when using the same underly-

Table 3. Identity results for XAI techniques as percentage of non-identical explanations for the same image. The best scores are highlighted in bold.

		CIFAR10	SVHN	Imagenette
VGG16BN	LIME	0.2	0.1	52.1
	SHAP	**0.0**	**0.0**	**0.0**
	GradCAM	**0.0**	**0.0**	**0.0**
	GradCAM++	**0.0**	**0.0**	**0.0**
	IntGrad	**0.0**	**0.0**	**0.0**
	SmoothGrad	100.0	100.0	100.0
Resnet50	LIME	0.9	0.4	55.1
	SHAP	**0.0**	**0.0**	**0.0**
	GradCAM	**0.0**	**0.0**	**0.0**
	GradCAM++	**0.0**	**0.0**	**0.0**
	IntGrad	**0.0**	**0.0**	**0.0**
	SmoothGrad	100.0	100.0	100.0
Densenet121	LIME	0.3	0.2	54.5
	SHAP	**0.0**	**0.0**	**0.0**
	GradCAM	**0.0**	**0.0**	**0.0**
	GradCAM++	**0.0**	**0.0**	**0.0**
	IntGrad	**0.0**	**0.0**	**0.0**
	SmoothGrad	100.0	100.0	100.0

ing black-box. SmoothGrad fails this metric in all cases, meaning no two explanations it generates are the same. This makes sense, considering it works by averaging explanations for variations of the original image with random noise added. LIME generates non-identical explanations around 50% of the time on Imagenette, and under 10% for smaller images. The latter is, again, because the default segmenter generates very large superpixels, and, as the Imagenette results show, with more appropriately configured segmenter producing smaller superpixels we can expect this number to be higher. The inconsistency of the technique, however, is expected, since it uses random sampling to generate explanations.

Since this metric was initially proposed in the context of tabular data, we feel like it may not be very descriptive for image data where there are usually a lot more features. A difference in the value of even one pixel would cause the image to count as non-identical, while being probably non-perceptible for a human looking at the explanation. At the same time the difference between the non-identical explanation may vary significantly. For example, as shown on Figure 5, non-identical SmoothGrad explanations look largely the same, while non-identical LIME explanations can at times highlight very different parts of the image. Some possible ideas to improve this measure could be to count explanations

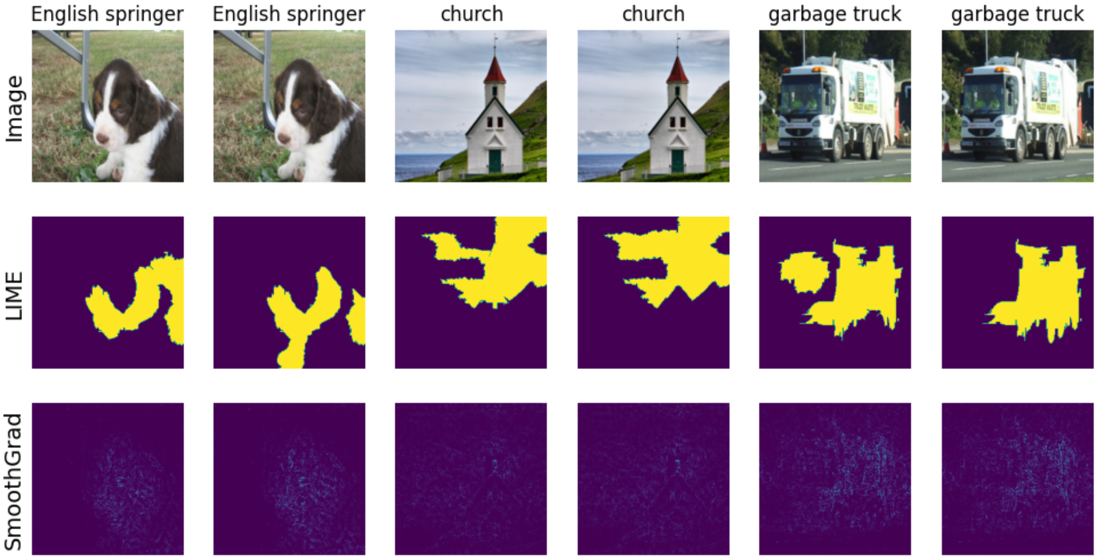

Figure 5. Example of instabilities in the explanations produced by LIME and SmoothGrad with Resnet50 on Imagenette. The first row is the original image labeled as classified by the black-box model, the second and third row are examples of non-duplicate predictions generated by LIME and SmoothGrad respectively.

as non-identical only above a certain threshold in the number of non-identical saliency values, or some measure of distance; or measure distance between explanations for identical images instead.

6.4 Separability Results

Applying the second metric from [Hon18], separability, we calculate the percentage of images with the same explanation as another non-identical image from the evaluated dataset. The results are presented in Table 4.

Table 4. Separability results for XAI techniques as percentage of images explained identically to another image. The best scores are highlighted in bold.

		CIFAR10	SVHN	Imagenette
VGG16BN	LIME	62.6	81.5	**0.0**
	SHAP	**0.0**	**0.0**	**0.0**
	GradCAM	**0.0**	**0.0**	**0.0**
	GradCAM++	**0.0**	**0.0**	**0.0**
	IntGrad	**0.0**	**0.0**	**0.0**
	SmoothGrad	**0.0**	**0.0**	**0.0**
Resnet50	LIME	64.2	78.0	**0.0**
	SHAP	**0.0**	**0.0**	**0.0**
	GradCAM	**0.0**	**0.0**	**0.0**
	GradCAM++	**0.0**	**0.0**	**0.0**
	IntGrad	**0.0**	**0.0**	**0.0**
	SmoothGrad	**0.0**	**0.0**	**0.0**
Densenet121	LIME	68.8	85.6	**0.0**
	SHAP	**0.0**	**0.0**	**0.0**
	GradCAM	**0.0**	**0.0**	**0.0**
	GradCAM++	**0.0**	**0.0**	**0.0**
	IntGrad	**0.0**	**0.0**	**0.0**
	SmoothGrad	**0.0**	**0.0**	**0.0**

From the results in Table 3 we can see that all evaluated techniques but LIME satisfy this metric on all models and datasets, while LIME satisfies it on Imagenette, but not on other datasets. On CIFAR and SVHN LIME produced over 60% and over 75% respectively of identical explanations for non-identical images. This is of course due to the default segmenter generating very large superpixels when partitioning the images, which resulted in the majority of the explanations being constant value maps.

Even though the majority of techniques have passed this metric, we believe for image data this metric can serve as a simple sanity check for pertrubation-based approaches like LIME which rely on segmentation algorithms to generate superpixels. A non-zero score may indicate that the segmenter needs additional configuration.

6.5 Time Results

The final metric we measured was time needed to produce a single explanation, as a measure of computational complexity. Average values across 2000 images for each evaluated XAI technique, dataset and model are presented in Table 5.

Table 5. Time results for XAI techniques in seconds. The best scores are highlighted in bold.

		CIFAR10	SVHN	Imagenette
VGG16BN	LIME	0.830	0.805	10.053
	SHAP	0.441	0.451	4.247
	GradCAM	**0.007**	0.008	**0.020**
	GradCAM++	**0.007**	**0.007**	0.024
	IntGrad	0.025	0.025	0.607
	SmoothGrad	0.124	0.124	3.032
Resnet50	LIME	1.722	1.969	6.912
	SHAP	0.902	1.006	2.612
	GradCAM	**0.021**	**0.020**	0.026
	GradCAM++	0.024	**0.020**	**0.024**
	IntGrad	0.135	0.134	0.365
	SmoothGrad	0.689	0.678	1.838
Densenet121	LIME	3.687	3.639	7.700
	SHAP	2.258	2.474	3.125
	GradCAM	**0.054**	**0.058**	0.061
	GradCAM++	0.056	0.059	**0.055**
	IntGrad	0.075	0.075	0.335
	SmoothGrad	0.309	0.309	1.712

As show in the results in Table 5, the metric largely stays consistent across models and dataset and is directly proportional to the number of black-box model inferences needed to produce an explanation. Activation-based techniques GradCAM and GradCAM++ only require one forward pass and performed the best. They are followed by IntGrad which required 50 inferences in our experiments. SmoothGrad multiplies that number by the number of noisy images in uses, in our case 5, requiring 250 inferences in total. SHAP requires 500 evaluations by default, while LIME requires 1000. While these values are configurable, lowering them is likely to degrade the performance and affect other metrics and should be approached with care.

6.6 Results Summary

To summarise our experimental comparison we provide aggregated results for fidelity, stability, identity, separability and time metrics on all explainability techniques, averaged across the three datasets and the three black-box models in Table 6.

Table 6. Results summary averaged across all models and datasets. The best scores are highlighted in bold.

	Fidelity	Stability	Identity	Separability	Time
LIME	0.339	1.339	14.0	49.0	4.146
SHAP	0.186	0.938	**0.0**	**0.0**	1.946
GradCAM	0.260	1.280	**0.0**	**0.0**	**0.031**
GradCAM++	0.263	1.238	**0.0**	**0.0**	**0.031**
IntGrad	**0.173**	0.557	**0.0**	**0.0**	0.197
SmoothGrad	0.175	**0.545**	100.0	**0.0**	0.979

Though LIME clearly achieves lowest score in all metrics, no other technique excels in all five metrics. IntGrad, SmoothGrad and SHAP show comparable performance in fidelity with IntGrad pushing slightly ahead of the other two and SHAP lagging slightly behind. In terms of stability, IntGrad and SmoothGrad display significantly better performance than the rest of the evaluated techniques, with SmoothGrad taking the edge. All techniques but LIME and SmoothGrad achieved perfect score in identity, and all but LIME - in separability. As for time, GradCAM and GradCAM++ are by far the fastest.

To summarise, we found that no XAI technique could achieve best results in all of the tested metrics, meaning that the choice of a technique remains a difficult one, requiring careful and detailed examination from multiple angles according to the desired qualities.

7 Conclusion and Future Work

In conclusion, over the course of this study we first identified five quantitative function-grounded metrics that can be used to measure different aspects of quality of saliency map XAI techniques for image data: fidelity, stability, identity, separability and time. We then conducted a thorough comparative evaluation of six popular SM interpretability techniques with regards to those metrics across three general-purpose image classification datasets and three well-known models. The results show that no single technique technique could serve as a silver bullet when it comes to explaining black-box model predictions on image data. In the end of the day, choosing the right technique for the use case at hand is a matter of prioritising aspects of quality that matter the most in the particular situation. That said, IntGrad performed quite well on all metrics.

As we can see from the obtained results, the dataset and the underlying black-box model may affect the performance drastically, particularly for model-specific techniques, so they should be an important factor in that selecting the right saliency map technique for the job. Another important factor are the meta-parameters of the XAI technique and more experiments are needed to evaluate that, most notably for LIME.

Lastly, we noted that some of the metrics seem to have certain caveats, meaning they probably do not present the full picture. Measuring fidelity using dAUC, we can see that the choice of the occlusion colour affects the result, and that the scores differ significantly for larger and smaller images. When measuring stability using the Lipschitz constant, we can see that it is seemingly biased towards sparser explanations, though this is not necessarily a bad thing. The definition of identity may be too strict for images where the number of features is usually very large, creating an opening for misinterpretation, while a broader metric could be more descriptive. Though we used one method per property of XAI technique in our evaluations, ideally more metrics should be used to achieve more precise and robust results.

References

[AGM$^+$18] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[AMJ18] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[BDMZM21] Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, and Cecilia Zanni-Merk. Towards a terminology for a fully contextualized xai. *Procedia Computer Science*, 192:241–250, 2021.

[BGG$^+$23] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023.

[BH22] Ryan S Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41, 2022.

[Bis19] Ekaba Bisong. *Google Colaboratory*, pages 59–64. Apress, Berkeley, CA, 2019.

[CAD$^+$24] Esma Cerekci, Deniz Alis, Nurper Denizoglu, Ozden Camurdan, Mustafa Ege Seker, Caner Ozer, Muhammed Yusuf Hansu, Toygar Tanyel, Ilkay Oksuz, and Ercan Karaarslan. Quantitative evaluation of saliency-based explainable artificial intelligence (xai) methods in deep learning-based mammogram analysis. *European Journal of Radiology*, 173:111356, 2024.

[Cho16] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. big data 5, 2 (2017), 153–163, 2016.

[CPC19] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[CSHB18] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter* *conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[Dad23] Eduardo Dadalto. Detectors: a python library for generalized out-of-distribution detection. `https://github.com/edadaltocg/detectors/`, 5 2023.

[DDS$^+$09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DVK17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[EC] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

[FBL16] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[Gc21] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[GM23] Tristan Gomez and Harold Mouchère. Computing and evaluating saliency maps for image classification: a tutorial. *Journal of Electronic Imaging*, 32(2):020801–020801, 2023.

[HG20] Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.

[Hig19] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Report, European Commission, Brussels, April 2019.

[HLA22] Tobias Huber, Benedikt Limmer, and Elisabeth André. Benchmarking perturbation-based saliency maps for explaining atari agents. *Frontiers in Artificial Intelligence*, 5:903875, 2022.

[HLVDMW17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[Hon18] Milo Honegger. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054*, 2018.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[KMM$^+$20] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

[Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[LSL$^+$20] Xiao-Hui Li, Yuhan Shi, Haoyang Li, Wei Bai, Yuanwei Song, Caleb Chen Cao, and Lei Chen. Quantitative evaluations on saliency methods: An experimental study. *arXiv preprint arXiv:2012.15616*, 2020.

[Mil19] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[NAL$^+$21] Oded Nov, Yindalon Aphinyanaphongs, Yvonne W Lui, Devin Mann, Maurizio Porfiri, Mark Riedl, John-Ross Rizzo, and Batia Wiesenfeld. The transformation of patient-clinician relationships with ai-based medical advice. *Communications of the ACM*, 64(3):46–48, 2021.

[NTP$^+$23] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[NWC$^+$11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and* *unsupervised feature learning*, volume 2011, page 7. Granada, Spain, 2011.

[PDS18] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[RŠB18] Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 159–175, 2018.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[SCD+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[SGA+22] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

[Skl24] Mykyta Skliarov. A comparative evaluation of explainability techniques for image data. `https://github.com/DataSystemsGroupUT/explainability-comparison-for-images/`, 5 2024.

[STK$^+$17] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[TFF08] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[YK19] Mengjiao Yang and Been Kim. Bim: Towards quantitative evaluation of interpretability methods with ground truth. *arXiv preprint arXiv:1907.09701*, 2019.

[YLCT20] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

[ZBL$^+$18] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Mykyta Skliarov**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   A Comparative Evaluation of Explainability Techniques for Image Data,
   *(*title of thesis)

   supervised by Radwa ElShawi, PhD.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Mykyta Skliarov
14/05/2024