Provably Better Explanations with Optimized Aggregation of Feature Attributions

Thomas Decker [1 2] **Ananta R. Bhattarai** [2 3] **Jindong Gu** [4] **Volker Tresp** [1 5] **Florian Buettner** [2 6 7]

Abstract

Using feature attributions for post-hoc explanations is a common practice to understand and verify the predictions of opaque machine learning models. Despite the numerous techniques available, individual methods often produce inconsistent and unstable results, putting their overall reliability into question. In this work, we aim to systematically improve the quality of feature attributions by combining multiple explanations across distinct methods or their variations. For this purpose, we propose a novel approach to derive optimal convex combinations of feature attributions that yield provable improvements of desired quality criteria such as robustness or faithfulness to the model behavior. Through extensive experiments involving various model architectures and popular feature attribution techniques, we demonstrate that our combination strategy consistently outperforms individual methods and existing baselines.

1. Introduction

The practice of quantifying the influence of individual features through attribution methods has been established as a popular paradigm to enhance the transparency of complex machine learning models. These approaches typically produce heatmaps highlighting individual input features, such as pixels or image regions, relevant to a specific model prediction (see Figure 1, left). However, while a multitude of techniques has been developed for this purpose, concerns and doubts regarding the reliability of individual methods persist (Adebayo et al., 2018; 2020; Zhou et al., 2022). For instance, empirical evidence has revealed that single methods often exhibit unreasonable sensitivity to input perturbations (Kindermans et al., 2019; Alvarez-Melis & Jaakkola, 2018; Dombrowski et al., 2019; Ghorbani et al., 2019; Lin et al., 2023) and critically depend on the concrete hyperparameter choice (Bansal et al., 2020; Sturmfels et al., 2020; Pahde et al., 2023). This lack of explanation robustness not only causes unstable attribution results but can even be exploited for malicious manipulations (Baniecki & Biecek, 2023). On top of that, some methods may fail to identify relevant features (Hooker et al., 2019; Zhou et al., 2022) and different techniques frequently disagree substantially when explaining the same prediction (Krishna et al., 2022; Neely et al., 2021). These findings contest the actual fidelity of a single attribution result for the purpose model explainability. In addition to these observations, there is a growing body of theoretical work that highlights the limitations of individual attribution methods (Nie et al., 2018; Sixt et al., 2020; Kumar et al., 2021; Bilodeau et al., 2024; Fokkema et al., 2023). More specifically, in (Han et al., 2022), the authors establish a "no-free lunch" theorem for model explanations, which implies that a single attribution method cannot universally approximate the behavior of any model faithfully.

Nevertheless, each feature attribution method derives importance based on different mechanisms and each can be associated with individual benefits and shortcomings. As a consequence, the question arises of how to best combine them to attain better explainability of opaque predictions.

In this work, we explore the capabilities of convex combinations across different attribution results to improve the overall reliability of explanations. Guided by established quality criteria for feature attributions (Nauta et al., 2023), we propose an effective strategy to derive convex weightings such that the corresponding aggregation of different outcomes yields significant improvements in robustness and faithfulness. This is underpinned by a theoretical analysis showing that the improvements in relevant quality metrics are provable and even close to optimal with high probability. Our specific contributions are the following:

- We introduce an innovative approach for combining the results generated by various feature attribution methods or different variants of the same method.

[1]LMU Munich [2]Siemens AG [3]Technical University of Munich [4]University of Oxford [5]Munich Center for Machine Learning (MCML) [6]Goethe University Frankfurt [7]German Cancer Research Center (DKFZ). Correspondence to: Thomas Decker <thomas.decker@siemens.com>, Florian Buettner <florian.buettner@dkfz.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

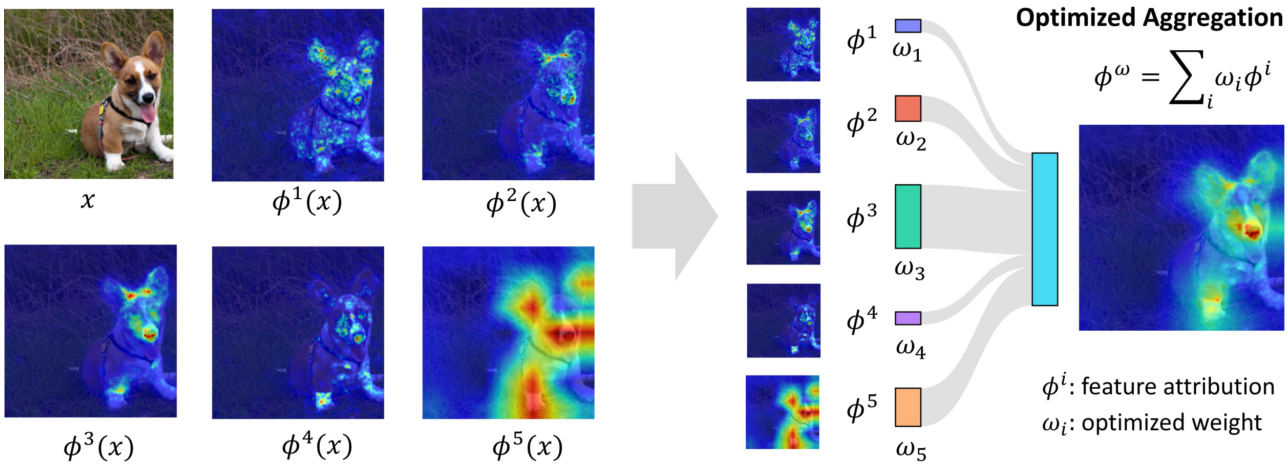

Figure 1. **Disagreement across attribution methods (left):** Different feature attribution methods $(\phi^1, \ldots, \phi^5)$ provide distinct perspectives about which particular features of an input x are important for an opaque model prediction $f(x)$. Oftentimes they tend to disagree causing ambiguity about which inputs truly matter. **Our Optimized Aggregation approach (right):** We study how to combine all individual attribution results fruitfully to attain better explanations. We propose a novel aggregation approach to retrieve optimal convex weights ω_i such that the aggregated feature attribution $\phi^\omega = \sum_i \omega_i \phi^i$ is provably more robust and more faithful to the underlying model.

- We show that our method can be effectively employed to optimize explanations according to commonly used measures of quality, including robustness and faithfulness to the model's behavior based on a unifying framework.
- We conduct a rigorous theoretical analysis establishing provable improvements of explanation quality and corresponding optimality bounds for our approach.
- We manifest these findings through a series of experiments involving popular feature attribution techniques and model architectures, consistently outperforming existing baselines and individual methods.

2. Problem Setup

Our goal is to enhance the reliability of explanations by developing effective strategies for combining diverse feature attribution results. To illustrate, let's consider explaining a prediction, denoted as $f(x)$, of a classification model $f : \mathbb{R}^d \to \mathbb{R}$, for an input instance $x \in \mathbb{R}^d$. A feature attribution method $\phi : \mathbb{R}^d \to [0,1]^d$, explains the prediction $f(x)$ by associating to each separate input x_i a normalized importance score $\phi_i(x)$. Suppose we have access to k different attribution methods, denoted by $\phi^1, \ldots, \phi^k$, each offering a distinct perspective. Further, let $\omega_1, \ldots, \omega_k$ be scalar weights such that $\sum_i \omega_i = 1$, and each weight is non-negative ($\omega_i \geq 0$). To aggregate distinct attribution outcomes we consider the weighted sum $\sum_i \omega_i \phi^i(x)$ yielding a novel explanation that combines individual insights. Our objective is to determine prediction-specific weights ω_i in a manner that provably improves desired quality metrics. This ultimately leads to more reliable and better explanations via aggregation (see Figure 1, right).

3. Background and Related Work

3.1. Measuring attribution quality

Evaluating the fidelity of explanations is a challenging endeavor due to missing knowledge about an objective ground truth. However, several quantitative metrics have been proposed to assess different aspects concerning the quality of feature attribution results (Nauta et al., 2023; Hedström et al., 2023a). In our study, we focus on two prominent categories of explanation quality:

Robustness Many explanation methods exhibit instabilities under small input perturbations (Alvarez-Melis & Jaakkola, 2018) leading to significantly different feature attribution results for almost identical inputs. While this not only casts doubts regarding the explanatory integrity of the considered technique, it might further be exploited to manipulate explanations intentionally (Baniecki & Biecek, 2023). A popular metric to quantitatively measure attribution robustness is Max-Sensitivity (Yeh et al., 2019):

$$\text{SENS}_{MAX} : \max_{\|\varepsilon\| \leq \delta} \quad \|\phi(x) - \phi(x+\varepsilon)\|$$

This quantity is typically estimated using a Monte Carlo approach by sampling a fixed number of small perturbations, evaluating the explanations, and storing the maximal distortion. An alternative metric for attribution robustness is Average-Sensitivity (Bhatt et al., 2021):

$$\text{SENS}_{AVG} : \mathbb{E}_\varepsilon \left[\|\phi(x) - \phi(x+\varepsilon)\|\right]$$

where $\varepsilon \sim \mathbb{P}_\varepsilon$ is a small random input perturbation, typically either Gaussian or uniformly distributed with mean zero.

Faithfulness The goal of faithfulness metrics is to measure how aligned an attribution result is with the actual model behavior in the sense that perturbing important features should also alter the model prediction accordingly. While different mathematical formulations have been proposed, a prominent choice is Infidelity (Yeh et al., 2019):

$$\text{INFD} : \mathbb{E}_I \left[(I^T \phi(x) - (f(x) - f(x - I)))^2 \right]$$

Here, $I \in \mathbb{R}^d$ describes a probabilistic perturbation such as replacing random parts of x with a fixed baseline value or Gaussian noise (Yeh et al., 2019). Similarly, (Bhatt et al., 2021) proposed to quantify faithfulness to the model's behavior using a correlation measure:

$$\text{FCOR} : \text{corr}_I \left(I^T \phi(x), (f(x) - f(x - I)) \right)$$

Thus, Faithfulness Correlation (FCOR) measures how correlated the attribution scores are with prediction changes under corresponding input modifications.

Other metrics Beyond robustness and faithfulness, additional dimensions of explanation quality have also been investigated in the literature. Alignment metrics (Arras et al., 2022; Decker et al., 2023) measure to which extent an explanation matches a desirable ground truth derived from domain knowledge and randomization-based sanity checks (Adebayo et al., 2018; Hedström et al., 2023b) ensure a sufficient dependence of the attribution result on the examined model. Moreover, Complexity metrics (Bhatt et al., 2021; Chalasani et al., 2020) quantify how comprehensible a model explanation is given the premise that sparser attributions are more informative to humans due to reduced cognitive load. Please refer to (Nauta et al., 2023) for a more comprehensive overview of available metrics.

3.2. Aggregating explanations

The idea of aggregating multiple feature attribution results **within** the same method is already anchored in popular explainability techniques. SmoothGrad (Smilkov et al., 2017) and UniformGrad (Wang et al., 2020) combine gradients in the proximity of the input and VarGrad (Adebayo et al., 2018) uses the variance of gradients within a neighborhood to derive feature importance. Similarly, Integrated Gradients (Sundararajan et al., 2017) and GradSHAP (Erion et al., 2021) aggregate gradients along a specific path towards predetermined baseline values. While such techniques combine gradients following input perturbation, NoiseGrad (Bykov et al., 2022) averages gradients under model parameter modifications to form a final explanation. In (Rebuffi et al., 2020) the authors analyze how the combination of attribution results obtained from different layers can improve the final explanation. The authors of (Bhatt et al., 2021) propose to enhance explanations by combining the Shapley Values of an instance with the ones obtained from its nearest neighbors in the training dataset.

On the other hand, the idea of aggregating attribution results **across** distinct methods has received considerably less attention. In (Rieger & Hansen, 2019), the authors propose two basic ways to combine distinct explanations which are defined as follows. $\text{AGG}_{Mean} = \frac{1}{k} \sum_{i=1}^{k} \phi^i(x)$ simply averages different attribution outcomes and AGG_{Var} incorporates also feature-wise variability to downgrade the importance of features where methods tend to disagree on:

$$\text{AGG}_{Var} = \frac{1}{k} \sum_{i=1}^{k} \frac{\phi^i(x)}{\sigma(\phi^1, \ldots, \phi^k) + \epsilon}$$

where $\sigma(\phi^1, \ldots, \phi^k) \in \mathbb{R}^d$ describes the feature-wise standard deviation across the different attribution results and ϵ is a small constant promoting numerical stability.

Nevertheless, a theoretically grounded strategy of how to best combine different attribution results for desired improvements is still missing and we aim to address this gap in the remainder of this paper.

4. Optimizing Explanations with Aggregation

Generalized L2 metrics for explanations In this section, we introduce a general class of quality metrics for explanation methods that can efficiently be improved via cross-method combination as shown later.

Definition 4.1. Let $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}$ a be a quality metric for feature attribution results. Then, $\mathcal{Q}$ belongs to the class of generalized $L2$ metrics if there exist suitable random variables $\gamma_1 \in \mathbb{R}^{g \times d}$ and $\gamma_2 \in \mathbb{R}^g$ such that :

$$\mathcal{Q}(\phi(x)) = \mathbb{E}_{\gamma_1, \gamma_2} [\|\gamma_1 \phi(x) - \gamma_2\|_2^2]$$

Conceptually, any such metric evaluates the quality of an attribution result $\phi(x)$ using the following intuitive principle. First, a linear query γ_1 is applied to extract certain information from the attribution results. Second, the obtained information content is compared to a desired query outcome γ_2 using the squared Euclidean distance. Thus, a smaller value of $\mathcal{Q}$ implies a better attribution result according to the considered criteria. Evaluating such metrics can simply be performed by estimating the expectation with a finite set of metric evaluation samples denoted by $\{(\gamma_1^{(j)}, \gamma_2^{(j)})\}_{i=1}^m$. Note that common quality metrics introduced above are generalized $L2$ metrics. For example, Average-Sensitivity can be recovered in the following way: Let $\mathcal{I}_d$ be the d-dimensional identity matrix, then setting $\gamma_1 = \mathcal{I}_d$ and $\gamma_2 = \phi(x + \varepsilon)$ with $\varepsilon \sim \mathbb{P}_\varepsilon$ results in SENS_{AVG} with respect to the squared Euclidean norm. Similarly, Infidelity can be obtained by choosing $\gamma_1 = I^T$ and $\gamma_2 = f(x) - f(x - I)$.

We further show in Appendix B how other categories of quality metrics can be expressed within this framework.
In conclusion, many established quality criteria for model explanations can be assessed based on a corresponding $L2$ formulation. Next, we show that enhancing such metrics through convex combinations leads to a well-posed optimization problem.

Deriving optimal weights Remember that our goal is to combine multiple attribution results to provably improve desired quality criteria. Suppose k different attribution outcomes $\phi^1(x) \dots \phi^k(x)$ for which we seek optimal convex weight factors $\boldsymbol{\omega} = (\omega_1, \dots, \omega_k)$. First, note that evaluating an aggregated attribution result $\phi^\omega = \sum_i \omega_i \phi^i$ via a generalized $L2$ metric $\mathcal{Q}$ reads:

$$\mathcal{Q}(\phi^\omega) = \mathbb{E}\left[\|\gamma_1 \phi^\omega - \gamma_2\|_2^2\right] = \mathbb{E}\left[\|(\gamma_1 \Phi)\omega - \gamma_2\|_2^2\right]$$

where $\Phi \in \mathbb{R}^{d \times k}$ describes the matrix of stacked individual attribution results $\Phi = (\phi^1, \dots, \phi^k)$. Therefore, optimizing for convex weights ω reduces to solving:

$$\min_\omega \;\; \mathbb{E}\left[\|(\gamma_1 \Phi)\omega - \gamma_2\|_2^2\right] \quad \text{s.t.} \quad \omega_i \geq 0, \;\; \sum_{i=1}^k \omega_i = 1$$

To ease the notation, we denote the set of feasible aggregation weights by $\Omega = \{\omega \in \mathbb{R}^k : \omega_i \geq 0, \; \sum_{i=1}^k \omega_i = 1\}$ and define $\gamma_{2,k:} \in \mathbb{R}^{g \times k}$ as matrix storing k copies of γ_2 in its columns. By setting $\Gamma := (\gamma_1 \Phi - \gamma_{2,k:}) \in \mathbb{R}^{g \times k}$, it holds within the set of feasible weights Ω that:

$$\min_{\omega \in \Omega} \;\; \mathcal{Q}(\phi^\omega) \;\; \Leftrightarrow \;\; \min_{\omega \in \Omega} \;\; \omega^T \mathbb{E}\left[\Gamma^T \Gamma\right] \omega$$

Hence, searching for the best way to aggregate different attribution outcomes ends up in a constrained quadratic program with convex constraints. This exhibits a global optimum and can efficiently be solved using corresponding numerical solvers (Boyd & Vandenberghe, 2004). On top of that, this observation also enables us to optimize multiple generalized $L2$ metrics simultaneously as quadratic forms are additive. Suppose we seek to improve q independent metrics $\mathcal{Q}_1, \dots \mathcal{Q}_q$ with associated parameters Γ_q as defined above. Then, for any scalers $\lambda_1, \dots \lambda_q$ we have:

$$\sum_{j=1}^q \lambda_j \mathcal{Q}_j(\phi^\omega) = \omega^T \mathbb{E}\left[\sum_{j=1}^q \lambda_j \Gamma_j^T \Gamma_j\right] \omega$$

This implies that searching for convex weights that directly improve multiple metrics prioritized by λ_i can also be expressed as a single constrained quadratic program and thus efficiently be solved.
In addition to its numerical appeal, optimizing explanations via aggregation in this way also comes with theoretical benefits in the form of provable improvement guarantees and probabilistic optimality bounds.

Provable improvement through aggregation The following theorem allows us to precisely quantify the gain in explanation quality induced via convex aggregation.

Theorem 4.2. *Let $\phi^\omega = \sum_i \omega_i \phi^i$ be the aggregated explanation, then the quality metric of ϕ^ω is always at least as good as the weighted metrics of the individual attributions:*

$$\mathcal{Q}(\phi^\omega) = \sum_i \omega_i \mathcal{Q}(\phi^i) - \mathbb{E}_{\gamma_1}\left[\sum_i \omega_i \|\gamma_1(\phi_i - \phi^\omega)\|_2^2\right]$$

Note that this result can be related to the error ambiguity decomposition for ensemble learning introduced in (Krogh & Vedelsby, 1994) and we conduct the proof in Appendix A. The achievable gain via aggregation $\mathbb{E}_{\gamma_1}\left[\sum_i \omega_i \|\gamma_1(\phi_i - \phi^\omega)\|_2^2\right] \geq 0$ depends on how diverse the different explanations behave under queries γ_1 compared to the aggregated one. Moreover, its non-negativity ensures that the quality of the aggregated explanation is at least as good as the equivalently weighted individual attribution qualities since lower values of $\mathcal{Q}$ imply improvements.

Generalization bounds for estimated weights Obtaining optimal weights usually requires approximating the objective based on a limited set of metric evaluation samples $\{(\gamma_1^{(j)}, \gamma_2^{(j)})\}_{i=1}^m$. Hence, the resulting estimate $\hat{\omega} = \arg\min_{\omega \in \Omega} \frac{1}{m} \sum_{j=1}^m \|\gamma_1^{(j)} \phi^\omega - \gamma_2^{(j)}\|_2^2$ may deviate from the ideal combination weights as it might not generalize well to unseen metric evaluations. As a consequence, it would be desirable to ensure that the quality improvement with estimated aggregation weights is close enough to the best possible strategy concerning the entire quality metric $\mathcal{Q}$. The following theorem establishes a corresponding result.

Theorem 4.3. *Consider a generalized $L2$ metric denoted by $\mathcal{Q}$ with $\max_{\gamma_1} \|\gamma_1\|_1 \leq c_1$ and $\max_{\gamma_1, \gamma_2} \|\gamma_1 \phi^i - \gamma_2\|_2^2 \leq c_2$. Additionally, let Ω represent the set of feasible weights ω and $\phi^\omega = \sum_{i=1}^k \omega_i \phi^i$ denote an aggregated feature attribution result. Suppose $\hat{\omega}$ is an estimate of aggregation weights obtained from m metric evaluation samples given by:*

$$\hat{\omega} = \arg\min_{\omega \in \Omega} \;\; \frac{1}{m} \sum_{j=1}^m \;\; \|\gamma_1^{(j)} \phi^\omega - \gamma_2^{(j)}\|_2^2$$

Then there exist a constant $C(c_1, c_2) > 0$ depending on c_1 and c_2 such that with probability of at least $(1 - \delta)$:

$$\mathcal{Q}(\phi^{\hat{\omega}}) - \min_{\omega \in \Omega} \mathcal{Q}(\phi^\omega) \leq C \sqrt{\frac{4 \log(16k/\delta)}{m}}$$

To prove this statement, we develop appropriate bounds on the Rademacher complexity of vector-valued functions based on a concentration result from (Maurer, 2016) and the specific properties of generalized $L2$ metrics over convex combinations of normalized feature attribution results. The

Table 1. SENS_{AVG} (S_{AVG}) and SENS_{MAX} (S_{MAX}) results for gradient-based attribution methods and different aggregation strategies across several model architectures. Our approach AGG_{robust} consistently outperforms all other techniques followed by AGG_{opt} as second best.

Feature Attribution	VGG16		AlexNet		ResNet18		MobileNetV2		MLPMixer	
	$S_{AVG} \downarrow$	$S_{MAX} \downarrow$	$S_{AVG} \downarrow$	$S_{MAX} \downarrow$	$S_{AVG} \downarrow$	$S_{MAX} \downarrow$	$S_{AVG} \downarrow$	$S_{MAX} \downarrow$	$S_{AVG} \downarrow$	$S_{MAX} \downarrow$
Saliency	0.994	1.214	0.800	0.942	0.964	1.163	1.022	1.265	1.200	1.591
Guided BP	0.515	0.650	0.430	0.512	0.483	0.611	0.832	1.060	-	-
DeepLift	0.893	1.090	0.791	0.932	0.857	0.999	0.894	1.075	0.961	1.165
IntGrad	0.888	1.065	0.724	0.854	0.838	0.991	0.910	1.084	0.941	1.141
InputxGrad	0.988	1.214	0.807	0.957	0.956	1.157	1.029	1.292	1.107	1,419
SmoothGrad	0.784	0.913	0.622	0.719	0.779	0.898	0.858	0.992	0.643	0.739
VarGrad	0.599	0.949	0.571	0.914	0.553	0.829	0.747	1.183	0.554	0.910
AGG_{Mean}	0.596	0.734	0.480	0.583	0.529	0.644	0.586	0.724	0.663	0.853
AGG_{Var}	0.582	0.700	0.476	0.574	0.518	0.618	0.568	0.686	0.631	0.788
AGG_{faith} **(ours)**	0.644	0.833	0.535	0.679	0.471	0.578	0.792	1.036	0.696	0.892
AGG_{opt} **(ours)**	0.456	0.584	0.364	0.449	0.427	0.538	0.536	0.701	0.483	0.642
AGG_{robust} **(ours)**	**0.424**	**0.543**	**0.349**	**0.426**	**0.410**	**0.513**	**0.505**	**0.654**	**0.473**	**0.634**

full derivation and the precise expression for the constant C are given in Appendix A. Intuitively, theorem 4.3 guarantees that the maximum potential deviation of our aggregation approach from the optimal improvement can be bounded with high probability. Moreover, the worst-case performance gap diminishes with order $\mathcal{O}(1/\sqrt{m})$ for increasing number of metric evaluation samples m.

Optimal aggregation for desired improvements Based on the generalized framework above we propose different aggregation strategies to intentionally enhance specific properties of feature attribution results. To explicitly enhance explanation robustness we obtain combination weights ω^{robust} by optimizing Average-Sensitivity as related $L2$ metric.

$$\text{AGG}_{robust}: \quad \omega^{robust} = \arg\min_{\omega \in \Omega} \text{SENS}_{AVG}(\phi^{\omega})$$

Equivalently, to optimize for faithfulness we can compute ω^{faith} by considering Infidelity as underlying objective:

$$\text{AGG}_{faith}: \quad \omega^{faith} = \arg\min_{\omega \in \Omega} \text{INFD}(\phi^{\omega})$$

As a default strategy to increase explanation quality via aggregation, we further propose improving both metrics simultaneously. We coin this approach AGG_{opt} optimizing for better feature attributions more generically:

$$\text{AGG}_{opt}: \quad \omega^{opt} = \arg\min_{\omega \in \Omega} \text{INFD}(\phi^{\omega}) + \text{SENS}_{AVG}(\phi^{\omega})$$

5. Experiments

We conducted a multifaceted empirical evaluation to investigate the capacities of our proposed aggregation strategies to intentionally enhance desired properties of explanations. All our aggregation strategies are optimized using only a small amount of metric evaluation samples to approximate the underlying metric ($m_{agg} = 50$). We explicitly test how well the improvements generalize to a larger sample of novel metric evaluations ($m_{eval} = 200$) and if they transfer to alternative quality measures. The findings presented in this section are based on the ImageNet `ILSVRC2012` dataset and concrete implementation details are documented in Appendix C. Accompanying source code is released at `https://github.com/thomdeck/aggopt`.

5.1. Quantitative evaluation of quality improvements

Increasing robustness via AGG_{robust} and AGG_{opt} We examine to which extent our aggregation approach can mitigate typically encountered instabilities of gradient-based explanations on convolutional models. For this purpose, we consider seven corresponding attribution techniques as well as four different ways of combining them including the two simple baselines AGG_{Mean} and AGG_{Var} and our proposed strategies AGG_{robust} and AGG_{opt}. All resulting explanations are computed for 500 random samples from ImageNet across five popular computer vision models and we evaluated their robustness based on the metrics SENS_{AVG} and SENS_{MAX}. The corresponding results in Table 1 indicate that our approach AGG_{robust} consistently outperforms all individual attribution methods as well as all other aggregations followed by AGG_{opt}, which is almost always second best. Remember that AGG_{robust} directly optimizes for SENS_{AVG} using a small number of metric evaluations. Thus, the generalization performance for this metric, now evaluated with a higher number of unseen metric evaluation samples, is in line with our theoretical framework. On top of that, the additional superiority in terms of the alternative metric SENS_{MAX} demonstrates that our approach also improves attribution robustness in general.

Table 2. INFD and FCOR results for different attribution methods and aggregation strategies across several model architectures. Our approach AGG_{faith} consistently outperforms all other techniques and AGG_{opt} is either second best or comparable.

Feature Attribution	DenseNet121		ResNet18		MobileNetV2		DeiT		SwinT	
	INFD ↓	FCOR ↑	INFD ↓	FCOR ↑	INFD ↓	FCOR ↑	INFD ↓	FCOR ↑	INFD ↓	FCOR ↑
GradSHAP	2.846	0.303	3.196	0.369	0.509	0.261	0.512	0.120	0.446	0.094
IntGrad	2.913	0.263	3.371	0.312	0.522	0.223	0.507	0.122	0.444	0.094
InputxGrad	3.097	0.205	3.587	0.259	0.538	0.192	0.516	0.113	0.446	0.079
SmoothGrad	2.604	0.388	2.916	0.444	0.494	0.296	0.392	0.297	0.367	0.208
GradCAM	2.646	0.388	2.922	0.459	0.478	0.319	0.385	0.311	0.373	0.227
GradCAM++	2.687	0.376	2.988	0.438	0.484	0.306	0.487	0.213	0.394	0.187
EigenCAM	3.044	0.251	3.381	0.347	0.538	0.231	0.568	0.058	0.439	0.107
AGG_{Mean}	2.661	0.370	2.928	0.444	0.479	0.293	0.449	0.237	0.377	0.213
AGG_{Var}	2.675	0.368	2.945	0.442	0.481	0.294	0.447	0.238	0.381	0.212
AGG_{robust} **(ours)**	2.678	0.339	2.956	0.415	0.483	0.282	0.407	0.288	0.366	0.224
AGG_{opt} **(ours)**	2.514	0.380	2.729	0.458	0.467	0.304	0.339	0.368	0.341	0.265
AGG_{faith} **(ours)**	**2.390**	**0.406**	**2.595**	**0.481**	**0.443**	**0.325**	**0.335**	**0.372**	**0.335**	**0.275**

Increasing faithfulness via AGG_{faith} and AGG_{opt} Similar to the robustness experiments, we evaluate the metrics Infidelity INFD and Faithfulness correlation FCOR to validate the capabilities of our approach for improving attribution fidelity. For both metrics, features are perturbed by replacing randomly selected pixels with the corresponding values of a blurred image version. More details regarding the precise design of these metrics and the underlying perturbations are specified in Appendix C1. We again considered five popular model architectures including two transformer-based ones and selected seven applicable attribution methods. Table 2 summarizes the results and shows that our dedicated aggregation approach AGG_{faith} performs best in every scenario, followed again by AGG_{opt} being either second best or comparable. The consistent superiority for INFD again supports our theory that optimizing aggregation using only a small amount of evaluation samples is enough to attain sustainable quality enhancement generalizing to novel out-of-sample metric evaluations unseen during optimization. Interestingly, the additional improvement concerning FCOR seems to be particularly strong for the two transformer-based models.

Improving additional quality metrics To further substantiate the benefits of our proposed methods, we also investigated how well the improvements generalize to additional quality metrics that express alternative notions of robustness and faithfulness. Stability metrics (Agarwal et al., 2022a) offer a complementary approach to evaluating the robustness of explanations by quantifying the sensitivity of attribution results relative to changes in various quantities of interest. More specifically, Relative Input Stability (RIS) assesses how explanations vary relative to input changes, Relative Representation Stability (RRS) examines variations relative to changes in the model's internal representations, and Relative Output Stability (ROS) evaluates sensitivity relative to changes in output prediction probabilities. The corresponding results in Table 3 over 500 samples on a ResNet18 show that our dedicated approach AGG_{robust} also significantly improves all stability metrics. During the experiments above, we used blurring as base perturbation when optimizing aggregation for faithfulness and when evaluating Infidelity and Faithfulness correlation. To check how well the improvements generalize to slightly altering notions of faithfulness, we computed variations of these metrics based on alternative corruptions (Hedström et al., 2023a) such as pixel value replacement with zeros (INFD_0 and FCOR_0) and the image mean ($\text{INFD}_{\bar{x}}$ and $\text{FCOR}_{\bar{x}}$). Even though we explicitly kept blurring as perturbation during weight optimization for AGG_{opt} and AGG_{faith} the resulting explanations still perform best when evaluated with different corruptions. Finally, we computed the Remove and Debias (ROAD) metric (Rong et al., 2022) that assesses the fidelity of explanations by removing the top features identified by an attribution method and estimating the subsequent decrease in prediction confidence. In Table 3 we report the outcomes of ablating the most relevant $p = 10, 20, 30$ percent of pixels in an image (MoRF_p) while additional percentiles are deferred to Appendix D1. All results consistently indicate that also this metric can be improved by relying on one of our proposed methods.

Overall, the results of all experiments manifest that our aggregation techniques achieve desired and generalizing improvements in explanation quality in line with our theoretical analysis. They also suggest that the aggregation strategy AGG_{opt}, which optimizes for faithfulness and robustness simultaneously, is an effective default approach to attain better explainability via aggregation when both criteria matter. All these findings are confirmed on four

Table 3. Additional quality metric results on a ResNet18 for different attribution methods and aggregation strategies. Across all evaluations, one of our proposed approaches performs best, and AGG_{opt} is consistently at least second best or comparable.

Feature Attribution	Stability ↓			Infidelity ↓		Faith. Corr. ↑		ROAD ↓		
	RIS	RRS	ROS	$INFD_0$	$INFD_{\bar{x}}$	$FCOR_0$	$FCOR_{\bar{x}}$	$MoRF_{10}$	$MoRF_{20}$	$MoRF_{30}$
Deeplift	8.10	5.26	8.92	3.59	3.64	0.33	0.32	-1.10	-2.02	-3.02
VarGrad	7.22	4.73	7.86	3.18	3.12	0.42	0.42	-2.37	-4.69	-6.79
GuidedBP	3.61	2.31	3.91	3.21	3.10	0.44	0.45	-3.09	-4.86	-6.31
IntGrad	7.69	4.88	8.28	3.14	3.12	0.43	0.43	-0.85	-1.79	-2.82
SmoothGrad	8.18	5.43	9.32	3.61	3.61	0.33	0.32	-1.57	-2.72	-3.84
InputxGrad	9.47	6.08	10.28	3.84	3.90	0.27	0.27	-0.63	-1.34	-2.24
Saliency	9.13	5.89	9.99	3.50	3.52	0.34	0.36	-0.58	-1.24	-2.04
AGG_{Mean}	5.33	3.41	5.75	3.16	3.23	0.43	0.41	-2.21	-3.98	-5.68
AGG_{Var}	5.19	3.33	5.61	3.18	3.22	0.42	0.42	-2.21	-3.98	-5.69
AGG_{faith} **(ours)**	5.73	3.70	6.25	**2.82**	**2.78**	0.47	**0.48**	-2.66	-4.50	-6.12
AGG_{opt} **(ours)**	3.62	2.33	3.96	2.83	2.80	**0.49**	0.47	-3.30	-5.41	-7.14
AGG_{robust} **(ours)**	**3.27**	**2.09**	**3.55**	2.97	2.93	0.46	0.46	**-3.36**	**-5.48**	**-7.17**

additional datasets in Appendix D2. Furthermore, we provide supplementary ablation studies in Appendix D4, which imply beneficial effects resulting from an increasing number of combined explanations and greater method diversity.

5.2. Understanding how optimized aggregation helps

In Figure 2 we display seven concrete examples with corresponding individual and aggregated attribution results to gain further insights into how optimized aggregation succeeds in improving explanations. Notice that our generic aggregation approach AGG_{opt} enhances feature attributions essentially via two mechanisms. Particularly in the first two images, all considered methods highlight intuitively relevant but diverging regions causing ambiguity about which pixels truly matter. AGG_{opt} improves the explanations by combining all perspectives to complement each other, which also leads to a visually more convincing explanation. In the last two images, some individual methods seem to fail by producing rather deteriorated results. For such instances AGG_{opt} performs automatic method selections intrinsically and aggregates only valid attribution outcomes to form an enhanced explanation that is more representative of the underlying model. Consulting the distribution of aggregation weights retrieved during both experiments in section 5.1 also reveals that the optimal weighting is highly model-dependent and even exhibits strong variability across samples. In Figure 4 we present corresponding boxplots for the weights of AGG_{opt} obtained for the two considered sets of attribution techniques. For all methods, the allocated weights during the experiments vary substantially among samples covering oftentimes even the entire possible range between 0 and 1. Moreover, the distribution of allocated weights does not transfer across models as for different architectures other methods are most favored. This provides further evidence that a single attribution method seems unable to explain every prediction for all model architectures faithfully and supports our approach to rather aggregate them in an optimizing manner.

5.3. Enhancing individual methods via aggregation

Many feature attribution methods rely on several hyperparameters and their concrete choice can greatly impact the resulting explanation (Bansal et al., 2020). A popular example is LIME (Ribeiro et al., 2016), which derives feature importance by fitting a linear surrogate model to approximate the model behavior in the vicinity of an input. When applied to image data, LIME typically computes attributions at the level of superpixels and incorporates an $L1$ regularization to enforce a certain level of sparsity via LASSO (Ribeiro et al., 2016; Garreau & Mardaoui, 2021). However, the requirement of fixing the regularization strength in advance might result in inferior explanations in cases where the number of important features does not match the enforced level of sparsity. To evaluate if optimized aggregation can effectively mediate this critical hyperparameter choice, we conducted the following experiment. We aggregate six different versions of LIME covering two different superpixel algorithms (SLIC (Achanta et al., 2012) and squared patches), each exhibiting either no, low, or high level of sparsity regularization. Using the available bounding box information for ImageNet, we distinguish between images where the object to be classified is particularly small ($< 20\%$ of the total picture) or rather larger ($> 60\%$). We randomly selected 200 images per object size and Figure 5 displays the average weights allocated by AGG_{opt} to different variants of LIME grouped by sparsity regularization. For both considered

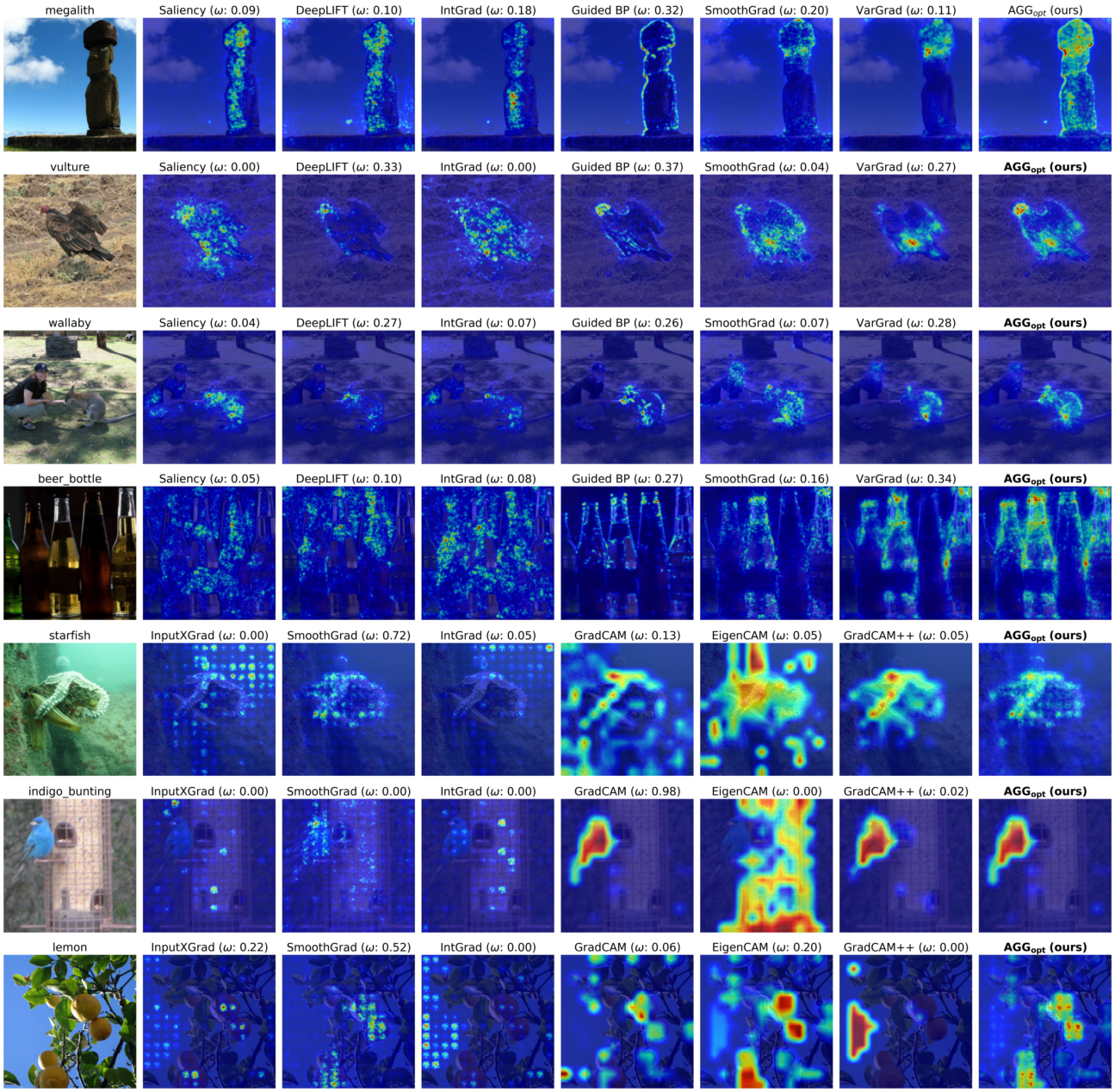

Figure 2. Individual outcomes of different feature attribution methods as well as our approach AGG_{opt} **(right column)** for seven images based on VGG16 (row 1-5) and DeiT (row 6-8). In addition to the quantitative improvements established in section 5.1 for robustness and faithfulness, our aggregation strategy also produces visually more intuitive and convincing explanations. It succeeds in enhancing the attribution results by combining several valid perspectives to complement each other (e.g rows 1 and 2) and by automatically discarding seemingly deteriorated explanations (e.g. rows 7 and 8).

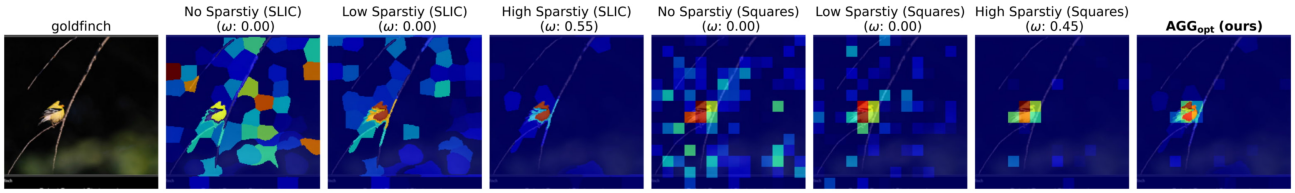

Figure 3. Individual attribution results of different LIME variants varying by superpixel structure and sparsity regularization on VGG16. The object to be classified is rather small and AGG_{opt} automatically combines only the sparsest explanations to enhance the explanation.

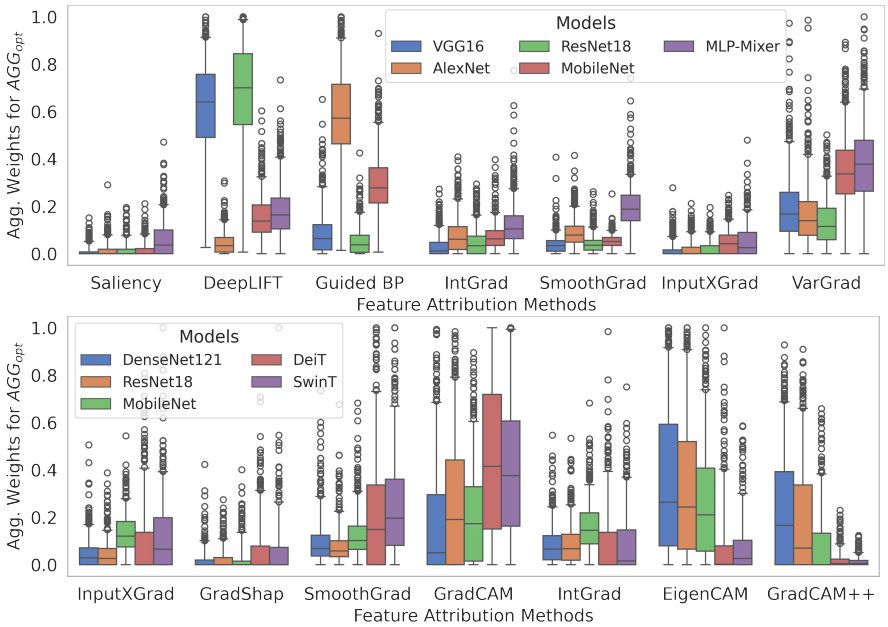

Figure 4. Boxplots of aggregation weights obtained by AGG_{opt} for the two considered sets of attribution methods during the evaluations in section 5.1 for robustness (top) and faithfulness (bottom) based on 500 samples. For each method, the allocated weight differs substantially among samples as most distributions cover almost the entire range between 0 and 1. There is also high variability across models indicating that a single method alone is unable to provide a reliable explanation for every prediction consistently.

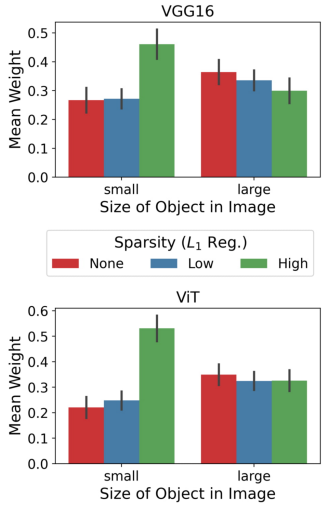

Figure 5. Average aggregation weights obtained by AGG_{opt} while optimizing the results from different versions of LIME on VGG16 (top) and ViT (bottom) including 95% confidence intervals as error bars. For smaller objects, significantly more weight is put on higher sparsity regularization.

models, significantly more weight is put on the explanations resulting from higher sparsity regularization when the object to be classified is small compared to larger ones. This is also exemplified in Figure 3, where the prediction for an image containing a rather small object is explained. AGG_{opt} optimizes the results of all considered LIME variants by aggregating only the two sparsest attributions matching the location of interest.

This demonstrates that our proposed aggregation approach also boosts the performance of individual attribution techniques when combining different versions of the same method.

6. Discussion and Conclusion

In this work, we provided the first theoretically grounded approach to optimally leverage distinct feature attribution results for improving explanations of opaque models. A downside of our technique is the higher inference time especially compared to relying on a single method only. However, since the main purpose of explainability techniques is to reliably increase the transparency of particularly critical decisions we argue that the added computational costs are minor (see Appendix D3) and well justified for the sake of provably better results. Another limiting aspect is the reliance on existing feature attribution methods and their validity as any uniform weakness might also compromise the aggregation. Hence, we recommend considering a sufficiently diverse set of individual techniques and we provide corresponding ablation studies in Appendix D4.

A natural extension of our work is to consider more sophisticated strategies beyond convex weighting to perform aggregation, such as voting algorithms or other ensemble methods. Furthermore, we specifically focused on combining fairly comparable feature attribution techniques. Future work could also explore how to best incorporate supplementary insights derived from concept-based (Hitzler & Sarker, 2022), optimization-based (Dabkowski & Gal, 2017; Fong et al., 2019; Jethani et al., 2021) or counterfactual explanations (Guidotti, 2022) to even further enhance explainability with aggregation.

Acknowledgements

We acknowledge the support from the German Federal Ministry for Economic Affairs and Climate Action (BMWK) via grant agreement 19I21039A.

Impact Statement

This paper presents work that aims at enhancing the transparency of predictions made by opaque machine learning

models. Enabling more reliable explanations can be associated with a variety of societal benefits by promoting trust, accountability, and compliance with regulatory and ethical standards. Hence, any research dedicated to explainable machine learning has the potential to contribute to more responsible development and deployment of machine learning-based technology.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. *Advances in Neural Information Processing Systems*, 33:700–712, 2020.

Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., and Lakkaraju, H. Rethinking stability for attribution-based explanations. *arXiv preprint arXiv:2203.06877*, 2022a.

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. OpenXAI: Towards a transparent evaluation of model explanations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022b.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Arras, L., Osman, A., and Samek, W. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. ISSN 1566-2535.

Baniecki, H. and Biecek, P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *arXiv preprint arXiv:2306.06123*, 2023.

Bansal, N., Agarwal, C., and Nguyen, A. Sam: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 8673–8683, 2020.

Bhatt, U., Weller, A., and Moura, J. M. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3016–3022, 2021.

Bilodeau, B., Jaques, N., Koh, P. W., and Kim, B. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Bykov, K., Hedström, A., Nakajima, S., and Höhne, M. M.-C. Noisegrad—enhancing explanations by introducing stochasticity to model weights. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6132–6140, 2022.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Chalasani, P., Chen, J., Chowdhury, A. R., Wu, X., and Jha, S. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pp. 1383–1391. PMLR, 2020.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Dabkowski, P. and Gal, Y. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

Decker, T., Lebacher, M., and Tresp, V. Does your model think like an engineer? explainable ai for bearing fault detection with deep learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Dicker, L. Sparsity and the truncated l^2-norm. In *Artificial Intelligence and Statistics*, pp. 159–166. PMLR, 2014.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for

image recognition at scale. In *International Conference on Learning Representations*, 2021.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021.

Fokkema, H., de Heide, R., and van Erven, T. Attribution-based explanations that provide recourse cannot be robust. *Journal of Machine Learning Research*, 24(360):1–37, 2023.

Fong, R., Patrick, M., and Vedaldi, A. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2950–2958, 2019.

Garreau, D. and Mardaoui, D. What does lime really see in images? In *International conference on machine learning*, pp. 3620–3629. PMLR, 2021.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3681–3688, 2019.

Gildenblat, J. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

Han, T., Srinivas, S., and Lakkaraju, H. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems*, 35:5256–5268, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., and Höhne, M. M.-C. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34): 1–11, 2023a.

Hedström, A., Weber, L., Lapuschkin, S., and Höhne, M. Sanity checks revisited: An exploration to repair the model parameter randomisation test. In *XAI in Action: Past, Present, and Future Applications*, 2023b.

Hitzler, P. and Sarker, M. Human-centered concept explanations for neural networks. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342(337):2, 2022.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jethani, N., Sudarshan, M., Aphinyanaphongs, Y., and Ranganath, R. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1459–1467. PMLR, 2021.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Krogh, A. and Vedelsby, J. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems*, 34:26598–26608, 2021.

Lin, C., Covert, I., and Lee, S.-I. On the robustness of removal-based feature attributions. *arXiv preprint arXiv:2306.07462*, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision

transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016 Proceedings 27*, pp. 3–17. Springer, 2016.

Muhammad, M. B. and Yeasin, M. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Neely, M., Schouten, S. F., Bleeker, M. J., and Lucic, A. Order in the court: Explainable ai methods prone to disagreement. *arXiv preprint arXiv:2105.03287*, 2021.

Nie, W., Zhang, Y., and Patel, A. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International conference on machine learning*, pp. 3809–3818. PMLR, 2018.

Pahde, F., Yolcu, G. U., Binder, A., Samek, W., and Lapuschkin, S. Optimizing explanations by network canonization and hyperparameter search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3819–3828, June 2023.

Rebuffi, S.-A., Fong, R., Ji, X., and Vedaldi, A. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8839–8848, 2020.

Ribeiro, M. T., Singh, S., and Guestrin, C. " 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*, pp. 1135–1144, 2016.

Rieger, L. and Hansen, L. K. Aggregating explanation methods for stable and robust explainability. *arXiv preprint arXiv:1903.00519*, 2019.

Rong, Y., Leemann, T., Borisov, V., Kasneci, G., and Kasneci, E. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, pp. 18770–18795. PMLR, 2022.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Sixt, L., Granz, M., and Landgraf, T. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pp. 9046–9057. PMLR, 2020.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Sturmfels, P., Lundberg, S., and Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Wang, Z., Wang, H., Ramkumar, S., Mardziel, P., Fredrikson, M., and Datta, A. Smoothed geometry for robust attribution. *Advances in neural information processing systems*, 33:13623–13634, 2020.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9623–9633, 2022.

A. Theoretical Proofs

In this section, we conduct the proofs of the theoretical results presented in the main paper.

Probable Improvements via Aggregation

Theorem A.1. *Let $\phi^\omega = \sum_i \omega_i \phi^i$ be the aggregated explanation then the quality of ϕ^ω is better than the the weighted quality of the individual attribution results:*

$$\mathcal{Q}(\phi^\omega) = \sum_i \omega_i \mathcal{Q}(\phi^i) - \mathbb{E}_{\gamma_1}\left[\sum_i \omega_i \|\gamma_1(\phi_i - \phi^\omega)\|_2^2\right]$$

Proof. The proof is similar to a related result established in (Krogh & Vedelsby, 1994):

$$\begin{aligned}
\sum_{i=1}^k \omega_i \mathcal{Q}(\phi^i) - \mathcal{Q}(\phi^\omega) &= \sum_{i=1}^k \omega_i \mathbb{E}\left[\|\gamma_1\phi^i - \gamma_2\|_2^2\right] - \mathbb{E}\left[\|\gamma_1\phi^\omega - \gamma_2\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^k \omega_i \|\gamma_1\phi^i - \gamma_2\|_2^2 - \|\gamma_1\phi^\omega - \gamma_2\|_2^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^k \omega_i \left({\phi^i}^T \gamma_1^T \gamma_1 \phi^i - 2\gamma_2^T \gamma_1 \phi^i + \gamma_2^T \gamma_2\right) \; - \; {\phi^\omega}^T \gamma_1^T \gamma_1 \phi^\omega - 2\gamma_2^T \gamma_1 \phi^\omega + \gamma_2^T \gamma_2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^k \omega_i \left({\phi^i}^T \gamma_1^T \gamma_1 \phi^i\right) \; - \; {\phi^\omega}^T \gamma_1^T \gamma_1 \phi^\omega\right] = \mathbb{E}\left[\sum_{i=1}^k \omega_i \left(\gamma_1(\phi^i - \phi^\omega)\right)^T \left(\gamma_1(\phi^i - \phi^\omega)\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^k \omega_i \|\gamma_1(\phi_i - \phi^\omega)\|_2^2\right]
\end{aligned}$$

□

Generalization Bound The proof of the generalization bound leverages the following result from (Shalev-Shwartz & Ben-David, 2014), where we slightly adapted the notation to better match our setup:

Theorem A.2 (26.5.3 in (Shalev-Shwartz & Ben-David, 2014))**.** *Let $(\mathcal{X} \times \mathcal{Y})$ be a probability space and $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a bounded loss function with $\ell(x, y) \leq L$. Let $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \ell(x_i, y_i)$ be an empirical estimator for the minimum of $\mathcal{L}(f) = \mathbb{E}\left[\ell(f(X), Y)\right]$. Also,for Rademacher variables $\varepsilon_i \in \{-1, 1\}$ the empirical Rademacher Complexity of a function set $\mathcal{F}$ is defined by:*

$$\hat{\mathcal{R}}_m(\mathcal{F}) = \frac{1}{m}\mathbb{E}_\varepsilon\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \varepsilon_i f(x_i)\right]$$

Then, it holds with a probability of at least $1 - \delta$:

$$\mathcal{L}(\hat{f}) - \min_{f \in \mathcal{F}} \mathcal{L}(f) \leq 2\hat{\mathcal{R}}_m(\ell \circ \mathcal{F}) + 5L\sqrt{\frac{2\ln(8/\delta)}{m}}$$

On top of that, we need the following Lemmas:

Lemma A.3. *Suppose that for all $x, y \in S \subset \mathbb{R}^g$ we have $\|x - y\|_2 \leq c$ for a constant c. Then, the squared Euclidean distance $l(x, y) = \|x - y\|_2^2$ is Lipschitz continuous in its first argument with Lipschitz constant $L = 2c$.*

Proof. The statement follows from quadratic factorization and the reverse triangular inequality.
For all $x, x', y \in S$ it holds:

$$\begin{aligned}
|l(x, y) - l(x', y)| &= |\|x - y\|_2^2 - \|x' - y\|_2^2| \\
&= |(\|x - y\|_2 + \|x' - y\|_2)\,(\|x - y\|_2 - \|x' - y\|_2)\,| \\
&\leq |(\|x - y\|_2 + \|x' - y\|_2)|\;\|x - y - x' + y\|_2 \\
&\leq 2c\;\|x - x'\|_2
\end{aligned}$$

□

Lemma A.4. *For $i = 1, \dots, m$, let $x_i \in \mathbb{R}^k$ and $\varepsilon_i \in \{-1, 1\}$ be Rademacher variables, so $\mathbb{P}(\varepsilon = \pm 1) = 1/2$. Then:*

$$\mathbb{E}_\varepsilon \left[\|\sum_{i=1}^m \varepsilon_i x_i\|_\infty \right] \leq \sqrt{m} \max_{i=1,\dots,m} \|x_i\|_\infty \sqrt{2\ln(2k)}$$

Proof. The proof leverages Massart's lemma and is for instance conducted within the proof of Lemma 26.11 in (Shalev-Shwartz & Ben-David, 2014) □

Now we are equipped to proof Theorem 4.3:

Theorem A.5. *Let $\mathcal{Q}$ be a generalized $L2$ metric with $\max_{\gamma_1} \|\gamma_1\|_1 \leq c_1$ and let $\Phi = (\phi^1, \dots, \phi^k)$ be the matrix of stacked attribution outcomes to be aggregated into $\phi^\omega = \sum_{i=1}^k \omega_i \phi^i$. Suppose that $\max_{\gamma_1,\gamma_2} \|\gamma_1 \phi^i - \gamma_2\|_2^2 \leq c_2$ as well as $\|\phi^i\|_\infty \leq 1$ for all $i = 1, \dots, k$. Also let Ω be the set of feasible weights ω and let $\hat{\omega}$ be an aggregation weight estimate obtained from m metric evaluations given by*

$$\hat{\omega} = \arg\min_{\omega \in \Omega} \frac{1}{m} \sum_{j=1}^m \|\gamma_1^{(j)} \phi^\omega - \gamma_2^{(j)}\|_2^2$$

Then there exist a constant $C(c_1, c_2) > 0$ depending on c_1 and c_2 such that with probability of at least $(1 - \delta)$:

$$\mathcal{Q}(\phi^{\hat{\omega}}) - \min_{\omega \in \Omega} \mathcal{Q}(\phi^\omega) \leq C \sqrt{\frac{\ln(16k/\delta)}{m}}$$

Proof. The theorem can be interpreted as an extension and adaptation of Theorem 26.15 in (Shalev-Shwartz & Ben-David, 2014) to the specifics of our setup. We develop appropriate bounds on the Rademacher complexity of vector-valued functions based on a concentration result from (Maurer, 2016) and the specific properties of generalized $L2$ metrics over convex combinations of normalized feature attribution results.
Let $\mathcal{F} = \{f : \mathbb{R}^{g \times k} \to \mathbb{R}^g,\ f(A) = A\omega \mid \omega \in \Omega\}$, then with Theorem A2 above we immediately get

$$\mathcal{Q}(\phi^{\hat{\omega}}) - \min_{\omega \in \Omega} \mathcal{Q}(\phi^\omega) \leq 2\hat{\mathcal{R}}_m(\ell \circ \mathcal{F}) + 5c_2 \sqrt{\frac{2\ln(8/\delta)}{m}} \tag{1}$$

where l is the squared Euclidean distance. To ease the notation in the following, define $A^{(i)} := \gamma_1^{(i)} \Phi$ and the j-th row of $A^{(i)}$ as $A_{j:}^{(i)} \in \mathbb{R}^k$. Using the assumption that $\max_{\gamma_1,\gamma_2} \|\gamma_1 \phi^i - \gamma_2\|_2^2 \leq c_2$, we know from Lemma A.3 that l is Lipschitz continuous with constant $L = 2\sqrt{c_2}$. Therefore, we are able to leverage a corresponding result from (Maurer, 2016) $(\star)$ to upper-bound the empirical Rademacher complexity of the vector-valued function set. More precisely, it holds:

$$\begin{aligned}
m\hat{\mathcal{R}}_m(\ell \circ \mathcal{F}) &= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \varepsilon_i \|f(A^{(i)}) - \gamma_2^{(i)}\|_2^2 \right] \overset{(\star)}{\leq} \sqrt{2} L \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sum_{j=1}^g \varepsilon_{i,j} f_j(A^{(i)}) \right] \\
&= 2\sqrt{2c_2} \mathbb{E}_\varepsilon \left[\sup_{\omega \in \Omega} \sum_{i=1}^m \sum_{j=1}^g \varepsilon_{i,j} (A^{(i)}\omega)_j \right] = 2\sqrt{2c_2} \mathbb{E}_\varepsilon \left[\sup_{\omega \in \Omega} \left\langle \sum_{i=1}^m \sum_{j=1}^g \varepsilon_{i,j} A_{j:}^{(i)}, \omega \right\rangle \right] \\
&\leq 2\sqrt{2c_2} \mathbb{E}_\varepsilon \left[\sup_{\omega \in \Omega} \|\sum_{i=1}^m \sum_{j=1}^g \varepsilon_{i,j} A_{j:}^{(i)}\|_\infty \|\omega\|_1 \right] = 2\sqrt{2c_2} \mathbb{E}_\varepsilon \left[\|\sum_{i=1}^m \sum_{j=1}^g \varepsilon_{i,j} A_{j:}^{(i)}\|_\infty \right]
\end{aligned}$$

where ε_i as well as $\varepsilon_{i,j}$ are Rademacher variables and the last two steps follow from the Hölder inequality as well as the constraints on ω. Next, notice that the term $\sum_{i=1}^m \sum_{j=1}^g \varepsilon_{i,j} A_{j:}^{(i)}$ sums over all row of A^i across all samples. Hence we can reindex the term as a sum over all consecutive rows in the sample denoted by $a^l \in \mathbb{R}^k$ with $l = 1, \dots, gm$.
Applying Lemma A.4 yields:

$$\mathbb{E}_\varepsilon \left[\|\sum_{i=1}^m \sum_{j=1}^g \varepsilon_{i,j} A_{j:}^{(i)}\|_\infty \right] = \mathbb{E}_\varepsilon \left[\|\sum_{l=1}^{gm} \varepsilon_l a^l\|_\infty \right] \leq \sqrt{gm} \max_{l=1,\dots,gm} \|a^l\|_\infty \sqrt{2\ln(2k)}$$

Notice that bounding $\max_{l=1,\dots,gm}\|a^l\|_\infty$ requires to bound the maximal entry of $\gamma_1\Phi$ that could be encountered while computing the metric. Using the assumed constraints on γ_1 and Φ we obtain:

$$\max_{l=1,\dots,gm}\|a^l\|_\infty \leq \max_{i,j}|(\gamma_1\Phi)_{i,j}| \leq \max_{\gamma_1}\|\gamma_1\|_1 \max_j\|\phi^j\|_\infty \leq c_1$$

Therefore, we finally have an upper bound on the empirical Rademacher complexity given by:

$$\hat{\mathcal{R}}_m(\ell\circ\mathcal{F}) \leq 2c_1\sqrt{2gc_2}\sqrt{\frac{2\ln(2k)}{m}} \tag{2}$$

Combining (1) and (2) and setting $C := \max\{4c_1\sqrt{2gc_2}, 5c_2\}$ gives:

$$\mathcal{Q}(\phi^{\hat{\omega}}) - \min_{\omega\in\Omega}\mathcal{Q}(\phi^\omega) \leq 4c_1\sqrt{2gc_2}\sqrt{\frac{2\ln(2k)}{m}} + 5c_2\sqrt{\frac{2\ln(8/\delta)}{m}} \leq C\sqrt{\frac{4\ln(16k/\delta)}{m}}$$

where the last step utilizes the fact that $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$ to merge the two square roots. □

B. Additional Generalized L2 Metrics for other Dimensions of Explanation Quality

In the main paper, we showed that popular metrics for feature attribution such as Infidelity and Average-Sensitivity are generalized $L2$ metrics. Below we also show that metrics regarding other quality criteria can be expressed as such.

Alignment Metrics Alignment metrics, also referred to as localization metrics, measure to which extent an attribution result corresponds to an expected explanation grounded in domain knowledge. For image classification models, such metrics typically quantify how well important image regions overlap with the actual location of the classified object in the image. A simple way to achieve this is to define $\gamma_2^\star$ as desired attribution results and measure alignment via the squared Euclidean distance: $\mathcal{Q}(\phi(x)) = \|\phi(x)-\gamma_2^\star\|_2^2$. Another possibility that closely resembles the logic of localization metrics for computer vision models is to measure if important features lie within a region of interest. Let $\mathcal{I}\subset\{1,\dots d\}$ be an index indicating the position of an object to be detected by a model. Then $\mathcal{Q}(\phi(x)) = \|\phi(x)-\phi_{\mathcal{I}}(x)\|_2^2$ is a generalized L2 metric that captures how much attribution mass is allocated to the region of interest.

Randomization-based sanity checks Randomization-based sanity checks have been developed to verify that an attribution result is not abstract and does indeed depend sufficiently on the model of interest. Typically they asses whether feature attributions change if certain parameters of the model are randomized. If an attribution result is invariant to parameter randomization it might not be reliably explain the examined model. To express this via a generalized $L2$ metric, suppose we are interested in explaining the prediction of a model f_θ with parameters θ. Let $\phi_\theta(x)$ be a feature attribution result obtained from the original model f_θ. Further, let $f_{\tilde{\theta}}$ denote the corresponding model where all parameters or a specific subset is randomized based on $\tilde{\theta}\sim\mathbb{P}_{\tilde{\theta}}$. Then, the variability of $\phi_\theta(x)$ under parameter randomization can be computed via $\mathcal{Q}(\phi_\theta(x)) = -\mathbb{E}_{\tilde{\theta}}\|\phi_\theta(x)-\phi_{\tilde{\theta}}(x)\|_2^2$. Note that we incorporate a negative sign to indicate that invariant attribution results correspond to lower quality.

Complexity To express complexity measures for feature attributions one can use the truncated $L2$ norm $\|\cdot\|_{2,t}$ as a sparsity measure (Dicker, 2014). This implies that $\mathcal{Q}(\phi(x)) = \|\min\{\phi(x),t\}\|_2^2$ where $\min\{\cdot,\cdot\}$ denotes the elementwise minimum operator and t a predefined noise threshold. Note that improving this metric $\mathcal{Q}$ requires pushing more entries of $\phi(x)$ below the threshold t which also promotes sparsity and reduces complexity. To translate this metric to the generic formulation proposed in Definition 4.1. one needs to set $\gamma_1\in\mathbb{R}^{d\times d}$ and $\gamma_2\in\mathbb{R}^d$ like this:

$$(\gamma_1)_{i,j} = \begin{cases} i=j: 1 \text{ if } |\phi_i|<t \text{ else } 0 \\ i\neq j: 0 \end{cases} \qquad (\gamma_2)_i = \begin{cases} -t & \text{if } |\phi_i|>t \\ 0 & \text{else} \end{cases}$$

C. Experimental Details

C.1. Metric Details

Robustness Throughout all experiments Average-Sensitivity (SENS_{AVG}) and and Max-Sensitivity (SENS_{MAX}) are computed using uniformly distributed corruptions $\varepsilon \sim \mathcal{U}[-0.1, 0.1]$:

$$\text{SENS}_{AVG} : \mathbb{E}_{\varepsilon}\left[\|\phi(x) - \phi(x+\varepsilon)\|\right]_2^2 \qquad \text{and} \qquad \text{SENS}_{MAX} : \max_{\varepsilon} \|\phi(x) - \phi(x+\varepsilon)\|_2^2$$

To optimize the aggregation weights for AGG_{robust} and AGG_{opt} the expectation is estimated using only m_{agg} samples for ε. During the evaluation in section 5.1, the metrics are computed using $m_{eval} = 200$ unseen samples to explicitly check for generalization. During evaluation both metrics are computed using the implementation provided by `Quantus` (Hedström et al., 2023a).

Faithfulness To compute the Infidelity metric (INFD) we rely on original design principles proposed by the authors (Yeh et al., 2019). In particular, we utilized binary perturbations $I \in \{0,1\}^d$ that randomly select an image area of 20% such that $I^T\phi$ equals the sum of attribution scores allocated to the selected region. From this quantity, we subtract the prediction change caused by replacing the selected image area with the corresponding values of a blurred image version x_b. This can be formalized using a map $h : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ with $h(x, x_b, I)_i = (x_b)_i$ if $I_i = 1$ and $h(x, x_b, I)_i = x_i$ else. We also incorporated the normalization utilized by the authors in their implementation (Yeh et al., 2019). For the Faithfulness correlation metric (FCOR) we use the same kind of perturbation and use the Pearson Correlation as correlation measure corr as proposed in (Bhatt et al., 2021). This results in:

$$\text{INFD} : \mathbb{E}_I\left[(I^T\phi(x) - (f(x) - f(h(x, x_b, I)))^2\right] \qquad \text{and} \qquad \text{FCOR} : \text{corr}_I\left(I^T\phi(x), f(x) - f(h(x, x_b, I))\right)$$

To optimize the aggregation weights for AGG_{faith} and AGG_{opt} the expectation and correlation is estimated using only m_{agg} samples of I. During the evaluation in section 5.1, the metrics are again computed using $m_{eval} = 200$ fresh samples to explicitly check for generalization.

Stability All stability metrics have been computed based on their implementation in `OpenXAI` (Agarwal et al., 2022b). For Relative Representation Stability (RRS) we used the activation of the final layer before the classification happens as underlying representation to compute the metric.

ROAD To compute the Remove and Debias metric (Rong et al., 2022) we leveraged the implementation provided by the `pytorch-gradcam` library (Gildenblat, 2021). Therefore, MoRF_p (Most relevant first) corresponds to the average decrease in confidence for the correct class of an image resulting from removing the p percent of the most important pixels as indicated by an attribution result. Note, that feature removal is performed using noisy linear imputation which has been demonstrated to produce consistent results matching the outcomes of retaining-based metrics such as Remove and Retrain (Hooker et al., 2019).

C.2. Feature Attribution Methods and Aggregations

Individual Methods During the experiments, we evaluated in total twelve different feature attribution techniques. The methods Saliency (Simonyan et al., 2013), InputxGrad (Shrikumar et al., 2016), Guided Backpropagation(Springenberg et al., 2014), DeepLift (Shrikumar et al., 2017), Integrated Gradients (Sundararajan et al., 2017), GradSHAP (Lundberg & Lee, 2017), SmoothGrad (Smilkov et al., 2017), VarGrad (Adebayo et al., 2018), Shapley Values (Castro et al., 2009), LIME (Ribeiro et al., 2016) and Feature Ablation are computed using the corresponding implementation provided by `Captum` (Kokhlikyan et al., 2020). For the methods GradCAM (Selvaraju et al., 2017), EigenCAM (Muhammad & Yeasin, 2020) and GradCAM++ (Chattopadhay et al., 2018) we utilized the `pytorch-gradcam` (Gildenblat, 2021) library. All attribution results are normalized to lie within the range $[0, 1]$ by taking the absolute value and rescaling them based on the maximum to ensure comparability.
Note that we excluded Guided Backpropagation when evaluating the MLPMixer architecture since the method was originally designed only for networks with the ReLU activation function.
For the experiments in section 5.3 we utilized Lime with three different LASSO regularization parameters λ. More precisely high sparsity regularization corresponds to $\lambda_{high} = 0.1$, medium to $\lambda_{medium} = 0.01$ and no regularization uses an ordinary least square regression approach to estimate the Lime coefficients. For the SLIC variant we provided a feature mask using the SLIC algorithm (Achanta et al., 2012) partitioning an image into approximately 100 superpixels.

Optimized Aggregation The aggregation weights for our combination approaches are optimized by estimating the underlying $L2$ metric using m metric evaluation samples yielding $\widehat{\text{SENS}}_{AVG}$ and $\widehat{\text{INFD}}$. In particular, we have:

$$\widehat{\text{SENS}}_{AVG}(\phi^\omega) = \frac{1}{m}\sum_{j=1}^{m}\|\phi^\omega(x) - \phi^\omega(x+\varepsilon^{(j)})\|_2^2$$

$$\widehat{\text{INFD}}(\phi^\omega) = \frac{1}{m}\sum_{j=1}^{m}\|\big(I^{(j)}\big)^T\phi^\omega(x) - \big(f(x) - f(h(x, x_b, I^{(j)}))\big)\|_2^2$$

and the weights are computed by solving:

$$\text{AGG}_{robust}: \quad \omega^{robust} = \arg\min_{\omega\in\Omega} \widehat{\text{SENS}}_{AVG}(\phi^\omega)$$

$$\text{AGG}_{faith}: \quad \omega^{faith} = \arg\min_{\omega\in\Omega} \widehat{\text{INFD}}(\phi^\omega)$$

$$\text{AGG}_{opt}: \quad \omega^{opt} = \arg\min_{\omega\in\Omega} \widehat{\text{INFD}}(\phi^\omega) + \widehat{\text{SENS}}_{AVG}(\phi^\omega)$$

All objectives are reformulated as constrained quadratic programs using the logic described in section 4 of the main paper and optimized using the default solver provided by `cvxpy` (Diamond & Boyd, 2016). For AGG_{opt} we additionally normalized both metrics using the Frobenius norm of the respective parameter matrix $\|\Gamma^T\Gamma\|_F$ to ensure comparability between the two considered metrics.

C.3. Model Details

We downloaded all convolutional models,including VGG in from (Simonyan et al., 2013), AlexNet (Krizhevsky et al., 2012), ResNet18 (He et al., 2016), MobileNetV2 (Sandler et al., 2018) and DenseNet121 (Huang et al., 2017), from `torchvison` with pre-trained weights. All transformer-based models are downloaded using the `timm` library (Wightman, 2019). More precisely, we utilized the following model variants:

DeiT (Touvron et al., 2021): `deit_tiny_patch16_224.fb_in1k`
ViT (Dosovitskiy et al., 2021): `vit_tiny_patch16_224.augreg_in21k_ft_in1k`
SwinT (Liu et al., 2021): `swin_tiny_patch4_window7_224.ms_in1k`
MLPMixer (Tolstikhin et al., 2021): `mixer_b16_224.goog_in21k_ft_in1k`

D. Additional Experiments and Results

D.1. Extended results for ROAD

Table 4. Remove and Debiase (ROAD) metric results on a Resnet18 where MoRF_p evaluates the average decrease in confidence caused by removing the top p percent of the most relevant pixels as indicated by the explanation method.

Method	MoRF_{10}	MoRF_{20}	MoRF_{30}	MoRF_{40}	MoRF_{50}	MoRF_{60}	MoRF_{70}	MoRF_{80}	MoRF_{90}	**Average** ↓
Deeplift	-1.10	-2.02	-3.02	-4.15	-5.42	-6.85	-8.56	-10.57	-13.03	-6.30
VarGrad	-2.37	-4.69	-6.79	-8.45	-10.05	-11.35	-12.57	-13.63	-14.67	-9.84
GuidedBP	-3.09	-4.86	-6.31	-7.62	-8.94	-10.17	-11.35	-12.59	-14.01	-9.77
IntGrad	-0.85	-1.79	-2.82	-3.96	-5.31	-6.86	-8.67	-10.70	-13.13	-6.23
SmoothGrad	-1.57	-2.72	-3.84	-4.94	-6.09	-7.35	-8.74	-10.38	-12.73	-6.82
InputxGrad	-0.63	-1.34	-2.24	-3.37	-4.63	-6.18	-8.05	-10.23	-12.87	-5.72
Saliency	-0.58	-1.24	-2.04	-3.10	-4.25	-5.65	-7.31	-9.33	-12.13	-5.18
AGG_{Mean}	-2.21	-3.98	-5.68	-7.33	-8.82	-10.27	-11.60	-12.92	-14.31	-9.57
AGG_{Var}	-2.21	-3.98	-5.68	-7.33	-8.82	-10.28	-11.60	-12.93	-14.31	-9.57
AGG_{faith}	-2.66	-4.50	-6.12	-7.49	-8.79	-9.99	-11.28	-12.57	-14.05	-9.50
AGG_{opt}	-3.30	-5.41	-7.14	-8.62	-9.96	-11.25	-12.45	-13.51	-14.58	-10.25
AGG_{robust}	**-3.36**	**-5.48**	**-7.17**	**-8.78**	**-10.15**	**-11.40**	**-12.65**	**-13.70**	**-14.68**	**-10.37**

D.2. Results on other Datasets

To substantiate the findings in the main paper, we repeated the experiments in section 5.1 on four additional datasets, namely CIFAR10 as well as three medical image datasets BloodMNIST, DermaMNIST and PathMNIST (Yang et al., 2023). Tables 5 and 6 summarize the corresponding results for the considered faithfulness and robustness metrics based on 500 images evaluated with a pre-trained ResNet18 model.

Table 5. INFD and FCOR results for different attribution methods and aggregation strategies for a ResNet18 model. Our approach AGG_{faith} consistently outperforms all other techniques and AGG_{opt} is either second best or comparable.

Feature Attribution	CIFAR10		BloodMNIST		DermaMNIST		PathMNIST	
	INFD ↓	FCOR ↑	INFD ↓	FCOR ↑	INFD ↓	FCOR ↑	INFD ↓	FCOR ↑
Saliency	4.129	0.159	11.60	0.393	0.324	0.321	5.394	0.152
DeepLift	3.928	0.252	15.12	0.278	0.354	0.233	5.294	0.159
IntGrad	4.016	0.229	12.97	0.290	0.321	0.309	5.145	0.190
InputxGrad	4.326	0.133	13.56	0.206	0.322	0.320	5.466	0.118
SmoothGrad	3.736	0.313	13.04	0.326	0.360	0.276	5.395	0.127
VarGrad	3.607	0.319	11.84	0.379	0.380	0.145	6.106	0.103
AGG_{Mean}	3.802	0.290	12.71	0.390	0.332	0.362	5.153	0.183
AGG_{Var}	3.817	0.290	12.73	0.391	0.335	0.361	5.132	0.186
AGG_{opt} **(ours)**	3.538	0.343	11.80	0.414	0.322	0.370	5.135	0.187
AGG_{faith} **(ours)**	**3.342**	**0.378**	**10.66**	**0.465**	**0.281**	**0.433**	**4.850**	**0.286**

Table 6. SENS_{AVG} (S_{AVG}) and SENS_{MAX} (S_{MAX}) results for gradient-based attribution methods and different aggregation strategies for a ResNet18 model. Our approach AGG_{robust} consistently outperforms all other techniques followed by AGG_{opt} as second best.

Feature Attribution	CIFAR10		BloodMNIST		DermaMNIST		PathMNIST	
	S_{AVG} ↓	S_{MAX} ↓	S_{AVG} ↓	S_{MAX} ↓	S_{AVG} ↓	S_{MAX} ↓	S_{AVG} ↓	S_{MAX} ↓
Saliency	0.916	1.143	0.696	0.882	0.787	0.993	0.942	1.108
DeepLift	0.805	1.016	0.514	0.644	0.679	0.936	0.723	0.818
IntGrad	0.820	1.029	0.481	0.622	0.673	0.861	0.833	0.955
InputxGrad	0.910	1.152	0.708	0.893	0.795	1.000	0.932	1.076
SmoothGrad	0.818	0.978	0.549	0.683	0.566	0.694	0.828	1.000
VarGrad	0.617	0.953	0.449	0.677	0.390	0.606	0.583	0.961
AGG_{Mean}	0.553	0.699	0.384	0.526	0.475	0.615	0.537	0.655
AGG_{Var}	0.549	0.689	0.386	0.526	0.474	0.609	0.536	0.653
AGG_{opt} **(ours)**	0.492	0.657	0.343	0.485	0.411	0.567	0.457	**0.619**
AGG_{robust} **(ours)**	**0.491**	**0.650**	**0.339**	**0.476**	**0.389**	**0.547**	**0.439**	0.627

D.3. Computation times of different aggregation strategies

In Table 7 we report the time required to retrieve optimal aggregation weights across 7 explainers for different models evaluated on an NVIDIA RTX A5000 GPU and averaged over 100 samples with corresponding standard deviations:

Table 7. Inference times to perform weight optimization based on seven explanation methods for AGG_{faith} and AGG_{robust} as average over 100 samples with corresponding standard deviation.

Time (s)	VGG16	ResNet18	MobileNetV2	DenseNet121	DeiT	SwinT
AGG_{faith}	0.79 ±0.06	0.75 ±0.06	0.77 ±0.06	1.49 ±0.51	1.33 ±0.92	0.83 ±0.09
AGG_{robust}	22.57 ±0.13	7.54 ±0.25	13.06 ±0.35	37.22 ±4.01	17.96 ±1.78	33.34 ±1.17
AGG_{opt}	23.39 ±0.14	8.31 ±0.23	13.91 ±0.27	38.52 ±3.28	19.39 ±4.59	34.47 ±1.13

We believe the additional computational cost imposed by our aggregation technique is minor compared to the strong improvements in explanation metrics.

D.4. Ablation Studies regarding number and diversity of combined explanations

Varying the number of methods to be aggregated We anticipate that our method will benefit from an increasing number of considered attributions by automatically down-weighting disadvantageous explanations. This behaviour is also exemplified in the last two rows of Figure 2 where deteriorated results received zero weight. To further investigate this, we performed a dedicated experiment in which we increased the number of feature attribution methods incrementally from 2 to 7 on a ResNet18 over 100 samples. The results in Table 8 and 9 show that the metrics do indeed get better for robustness and faithfulness, but the improvements seem to saturate at a certain point. The ordered set of explainers that were used for this experiment is: DeepLift, VarGrad, GuidedBackprop, SmoothGrad, IntGrad, InputxGrad, Saliency.

Table 8. Robustness metrics for AGG_{robust} combining an increasing number of explanation methods

AGG_{robust}	2	3	4	5	6	7
$S_{AVG} \downarrow$	0.52	0.43	0.42	0.41	0.41	0.41
$S_{MAX} \downarrow$	0.68	0.54	0.52	0.52	0.51	0.50

Table 9. Faithfulness metrics for AGG_{faith} combining an increasing number explanation methods

AGG_{faith}	2	3	4	5	6	7
INFD $\downarrow$	2.69	2.46	2.44	2.43	2.43	2.43
FCOR $\uparrow$	0.46	0.49	0.50	0.51	0.51	0.50

Combining different types of explanation techniques Concerning the diversity of explainers to use, we argue that our approach can be applied to fruitfully combine all methods that output fairly comparable explanations. This includes gradient-based and perturbation-based ones and we expect that also a higher diversity will be advantageous. To further investigate this, we evaluated our method again on 100 samples with a ResNet18 using 3 gradient-based methods, 3 perturbation-based methods, and a combination of all 6. The results in Tables 10 and 11 indicate that including both types does also benefit our aggregation approach. For instance, the perturbation-based methods seem to be significantly more robust than the gradient-based ones but by combining them, we can even further improve their robustness. We can also enhance faithfulness this way. The gradient-based methods used in this experiment are DeepLift, SmoothGrad, InputxGrad and the perturbation-based methods used in this experiment are Lime, Shapley Values and Feature Ablation.
All methods have been computed based on their default implementation in `Captum` (Kokhlikyan et al., 2020) while we additionally used a feature mask of 16x16 patches for all perturbation-based methods.

Table 10. Robustness metrics for AGG_{robust} combining different types of explanation methods

AGG_{robust}	Grad.-based	Pert.-based	Both
$S_{AVG} \downarrow$	0.65	0.49	**0.46**
$S_{MAX} \downarrow$	0.77	0.61	**0.57**

Table 11. Faithfulness metrics for AGG_{faith} combining different types of explanation methods

AGG_{faith}	Grad.-based	Pert.-based	Both
INFD $\downarrow$	2.82	2.92	**2.76**
FCOR $\uparrow$	0.46	0.44	**0.47**