SVEA: A Small-scale Benchmark for Validating the Usability of Post-hoc Explainable AI Solutions in Image and Signal Recognition

Sam Sattarzadeh, Mahesh Sudhakar, Konstantinos N. Plataniotis
Department of Electrical & Computer Engineering, University of Toronto
sam.sattarzadeh@mail.utoronto.ca

Abstract

Novel solutions in the area of Explainable AI (XAI) have made a significant breakthrough in increasing the trust of end-users in Machine Learning (ML) models. However, validating the performance of these solutions remains a challenging task. In this work, we focus on evaluating the methods that attribute a model's decision to their input features. The prior metrics on this topic fail to consider multiple properties that a usable explainability solution should satisfy. Also, conducting experiments to assess the concreteness of the explanations provided by these solutions in large-scale datasets consumes excessive time and resources. To overcome these shortcomings, we propose the Small-scale Visual Explanation Analysis (SVEA) benchmark, which comprises the recent minimal MNIST-1D dataset. Our proposed benchmarking tool aids the practitioners and researchers to perform experiments on the Explainable AI methods without the need to access expensive computational devices. Furthermore, we offer a framework to evaluate various characteristics of the state-of-the-art XAI methods and include several widely used interpretability solutions in the SVEA benchmark to perform a thorough analysis of their completeness and understandability. The results obtained from our proposed evaluation metric suggest that specific approaches lack the ability to transfer the chosen model's understanding to a second interpretable model by the explanations generated. The users can replicate our experiments within few minutes before working extensively on other larger datasets, thereby saving a lot of experimental time and effort.

1. Introduction

As one of the considering directions in Trustworthy AI [16], Explainable AI has become a highly demanding field in a variety of real-world applications, such as healthcare, autonomous driving, criminal justice, and finance [15, 4, 24]. The solutions proposed in this area lead a large group of developers, theorists, ethicists, and stakeholders to achieve a more concrete understanding of the decision mechanism of Machine Learning (ML) models and increase their trust in the decisions made by these models [22]. However, evaluating the performance of explanation methods remains a significant challenge in both industrial and academic research [11].

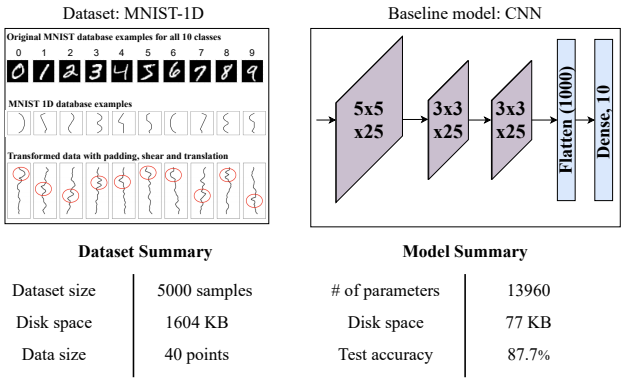

Figure 1. Key features of the dataset and model employed in our proposed benchmark.

A concrete explainable AI method must accompany both the target model and end-users as an interface between computer-based predictors and humans. In general, explanations are expected to be *faithful* from the model's perspective. They should correctly depict the exact behavior of the target model in a particular scope. From the users' side, explanations should be *understandable* and must provide them with transparent and interpretable insights regarding the model's decision-making procedure [4]. If an explainable AI method's output satisfies 'faithfulness' and 'understandability', it can be further extended for functional purposes such as model understanding, model debugging, and detecting dataset biases.

The expected properties of explanation algorithms are organized more accurately in the "Explainability fact sheet" proposed in [31]. According to this fact sheet, each explainability solution should satisfy several properties to be con-

sidered "usable" from the users' point of view. Developing a framework to assess these properties simultaneously is an ambitious task, especially in large-scale applications such as high-resolution image processing. The need for excessive resources such as ample memory space and faster GPUs, the training time, and providing feature-level annotations or human feedback are common shortcomings among most of these experiments that hinder the students and researchers from taking advantage of these valuable frameworks.

To circumvent these limitations, we introduce a small-scale framework to benchmark the attribution methods that are functional in ML-based image and signal processing applications. Attribution methods are a specific subgroup of *post-hoc* and *local* explainability methods (according to the terminologies defined in [31]) that take a trained target model and image (or signal data) as inputs and output a heatmap with the same shape as the input that highlights the features that are the most important towards the model's prediction.

Our proposed framework inherits the MNIST-1D dataset [10], a low-dimensional analogous of the renowned MNIST dataset [14]. This dataset was initially developed to enable the researchers to study deep learning cases that mainly focus on data recognition on a much small scale. Unlike these objectives, we utilize the MNIST-1D dataset to inspect the faithfulness and understandability of the post-hoc interpretability solutions in small scales. Carrying experiments on such a scale aids in promoting future research in this field as a whole. Our contributions in this work can be summarized as follows:

- We present a low-memory and low-compute benchmark to compare the performance of state-of-the-art attribution methods in various aspects. Our benchmarking tool allows researchers and practitioners to explore solutions for model interpretability without dealing with the burdensome problems in large-scale environments.
- We propose a validation experiment that strictly measures the usability of attribution methods without the need for complicated feature-level annotations or human resources.

2. Related Works

As stated in [7], the approaches to validate the explanation algorithms are classified into three types: application level, human level, and function level.

2.1. Application Level Validation

The explanations are evaluated on real tasks at this level by comparing with the explanations provided by a domain expert. In terms of attribution methods, the visual explanations are compared with ground truth annotations such as bounding boxes or segmentation masks. These metrics are also termed as *ground truth-based* metrics [25].

For instance, evaluation metrics such as Pointing Game (PG) [35] and its expanded version, Energy-based Pointing Game (EBPG) [33, 19], quantify the fraction of energy in a set of explanations that are located inside their corresponding ground truth labels. Moreover, the Bounding box metric [26] as an adaptive analogous for mean Intersection over Union (mIoU), calculates the portion of the most highlighted features captured by the annotation masks. In a more novel experiment designated in [34], attribution methods are evaluated by being applied on two different models trained on a crafted dataset containing foreground and a background class.

2.2. Human Level Validation

Human-level validation experiments evaluate the understandability and satisfaction of the explanations by including people in the loop. This type of evaluation, which relies on getting direct feedback from the users engaged with the model, can be performed by asking the users to rate the explanations generated by explainability methods or utilizing the explanations to perform specific tasks [7, 13, 21]. To carry human-level validation experiments, some prior works [24, 27] created interfaces that enable the individuals to compare multiple explanations in various aspects, such as class discrimination, visual clarity, and trustworthiness.

2.3. Function Level Validation

This type of validation primarily evaluates the explanations' correctness by measuring the correlation between the model's behavior and the provided explanations. When it comes to evaluating attribution methods, function-level validation can be conducted in several ways. For instance, pairs of metrics such as "Drop and Increase rate" and "Insertion and Deletion" validate the faithfulness of explanations by observing the model's output when it is fed only with the input features indicated as important by the explanations [18, 5, 6, 33, 9]. The Remove and Retrain (ROAR) metric runs by retraining the target model from scratch using only the features that scored the highest by an attribution method [12]. The sensitivity-n experiment operates by statistical computation of the covariance of the explanations and the model's predictions by applying random perturbations in the input domain [3]. Instead of pre-defined ground truth labels, these experiments consider the model's prediction to the given input as an evaluation baseline. Hence, they are termed as *model truth-based* metrics. Furthermore, another series of function-level experiments focus on evaluating the explanation algorithms' sensitivity to the model's specifications and parameters [2].

Compared with the two former validation types, a significant advantage of function-level experiments is that they

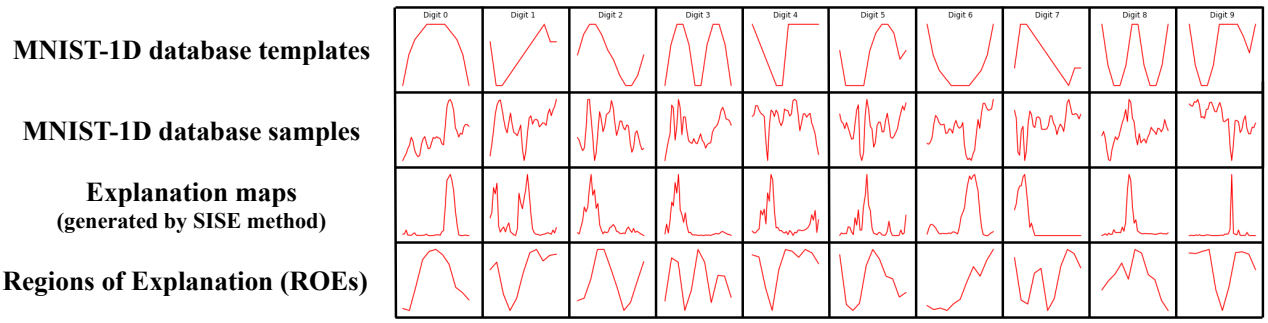

Figure 2. Samples from the MNIST-1D data [10] and their corresponding Regions of Explanation (ROEs) extracted using SISE method [25].

do not require providing extravagant information such as annotations and user feedback to operate. Hence, they are compatible with any ML-based model regardless of their application. Our proposed validation framework can be categorized into function-level experiments. Unlike the prior experiments in the same category, our benchmark takes account for a broader range of properties to guarantee the usability of attribution methods [31].

3. Small-scale Visual Explanation Analysis (SVEA) Benchmark

The Small-scale Visual Explanation Analysis (SVEA) benchmark is designed to perform the function-level validation tasks quickly and efficiently reproducible. To achieve our defined goal, we utilize MNIST-1D [10], a synthetic dataset containing low-dimensional encoded data. When a non-interpretable model trained on this dataset is provided to a well-performing attribution method, it can decode the data in a representative manner by discovering the model's understanding of the underlying data. We also designate an evaluation framework that works by asking an interpretable model to replicate the target model's task by employing the baseline model's explanation.

Our proposed benchmark includes four main components: 1) The training and test data in the MNIST-1D dataset, 2) A trained baseline predictive model, 3) An interpretable linear classifier, and 4) A set of state-of-the-art visual explanation algorithms. The significant advantage of our proposed benchmark is that it allows simulations to run without demanding an unreasonable amount of time, memory, and external acceleration devices like GPUs.

3.1. The MNIST-1D Dataset

As stated above, this minimal analogous for the MNIST dataset was initially introduced to be applied in case studies such as predicting lottery tickets, observing deep double descent, and meta-learning [10]. However, this dataset has not yet been employed for analyzing model interpretation techniques in small scales. In this work, we utilize this dataset to considerably decrease the computational overhead of our validation experiments on novel attribution methods.

The MNIST-1D dataset is functional in real-world digit classification. Though the training samples in this dataset are 20 times smaller than MNIST, this dataset distinguishes the critical machine learning models more broadly in terms of test accuracy. In terms of structure, the main difference between these datasets is that instead of handwriting images representing the digits 0-9, the samples in the MNIST-1D dataset are formed based on ten one-dimensional template patterns, as shown in Fig. 2 which resemble the original digits. Each MNIST-1D sample is created as follows:

1. Select the template for each class label. (Each template is a one-dimensional pattern consisting of 12 points.)
2. Pad the template by randomly adding 36-60 points.
3. Apply random transformations such as translation, 1-D shearing, and Gaussian noise addition.
4. Finally, scale and downsample the pattern to 40 points.

This procedure implies that in each MNIST-1D sample, at least 70% of the data points do not represent the actual template related to the sample's correct label. Moreover, since the samples are affected by random translation, the spatial information of the patterns is not a reliable evidence in the classification procedure. Hence, translation-variant models such as linear classifiers fail to achieve a high classification accuracy in the MNIST-1D dataset. Fig. 3 illustrates this massive gap between a logistic regressor and a Convolutional Neural Network (CNN).

3.2. Region of Explanation (ROE)

To address the drawbacks of models with spatial inductive biases, a simple but novel idea is to train them using the visual explanations reached from a more complicated and translation-invariant model instead of the original data. This idea connects linear signal classification and Explainable AI. The intuition behind this idea is to replace the noisy data

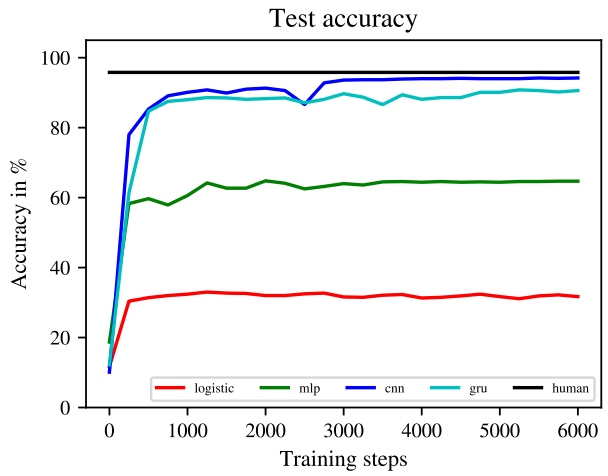

Figure 3. The test accuracy for different classification models trained on the MNIST-1D dataset, as reported in [10].

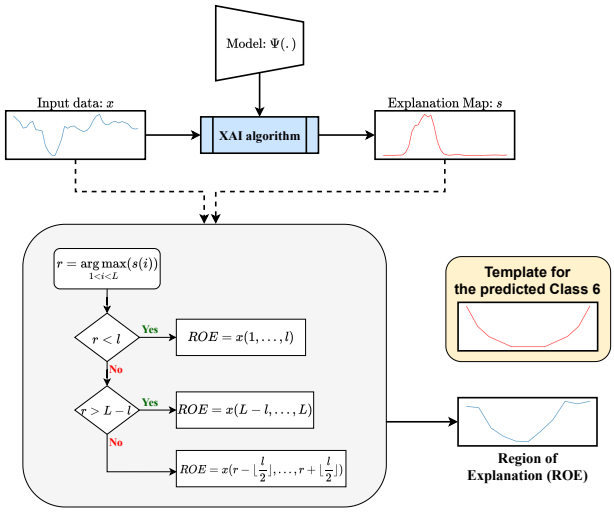

Figure 4. The schematic diagram to extract the Region of Explanation (ROE) for a given sample input. Each visual explanation algorithm outputs a set of connected data points with dimension $l = 12$ as the ROE associated with the input data given to the baseline model.

with their corresponding common features (CF) to avoid manipulating the linear models in their learning process. As defined in [34], a common feature is a set of points with some semantic meaning that commonly appear in all examples of one class. In terms of the MNIST-1D dataset, common features are a set of connected data points representing a template related to a specific digit. Hence, a usable explanation method is expected to extract the common feature from each given sample with a well-trained model on this dataset.

Given a set of a model, input data, and attribution method, we define the term Region of Explanation (ROE) as a set of connected points in the input domain highlighted by the attribution method as the most important in the model's prediction. Taking this definition into account, the ROEs determined by an ideal attribution method should point out the common features, which are the defined digit templates, in our case. Thus, extracting ROEs using a well-performed attribution method helps to learn an interpretable classifier more accurately while eliminating the effect of spatial variation among the dataset.

3.3. Baseline Model

We trained an extremely shallow CNN with the same architecture used in [10] as the baseline, shown in Fig. 1. This selection allows us to evaluate the attribution methods that are specialized to be applied only to CNNs (e.g., Grad-CAM, Grad-CAM++, SISE [27, 5, 25]), as well as the popular model-agnostic methods (e.g., Integrated Gradients, RISE [32, 18]). Considering that the baseline model contains only 13,960 trainable parameters, training this model from scratch is not a time-consuming process, even if done without a GPU. For our use case, we trained this network for 200 epochs using Stochastic Gradient Descent (SGD) optimizer and achieved a test accuracy of 87.7%. More details regarding the performance of the baseline CNN are provided in the supplementary material.

Denoting the baseline model as $\Psi : \mathbb{R}^L \rightarrow \mathbb{R}^{|C|}$ and the input as $x \in \mathbb{R}^L$, an attribution method generates the explanation for the model's top prediction indicated as $s \in \mathbb{R}^L$, where L is the size of the data (40 in our case), and $C = \{0, 1, ..., 9\}$ is the set of output classes (digits). In particular, we define the ROE for this pair as a window of size $l = 12$, centered by the data point with the highest importance score in s, based on the condition,

$$l < \underset{1 \leq i \leq L}{\arg\max}(s(i)) < L - l \tag{1}$$

If this condition is not satisfied, it implies that the highest scored point is adjacent to the signal's origin or end. Then, the window matches the first or last l points. Fig. 4 shows the procedure to extract ROE for a given input.

3.4. ROE Understandability Test

The usability of attribution methods can be validated by employing the ROEs they determine to learn and evaluate a translation-variant model. Unlike the ground truth-based metrics that measure the explanations' understandability by matching them with human-crafted annotations, our proposed test evaluates the ability of the explanations to transfer the baseline model's understanding to a second model whose functionality is easy to interpret. Also, similar to the ROAR experiment [12], our test accounts for validating the "completeness" of the explanations generated by attribution methods and assess whether the explanations can be generalized and framed into a specific context.

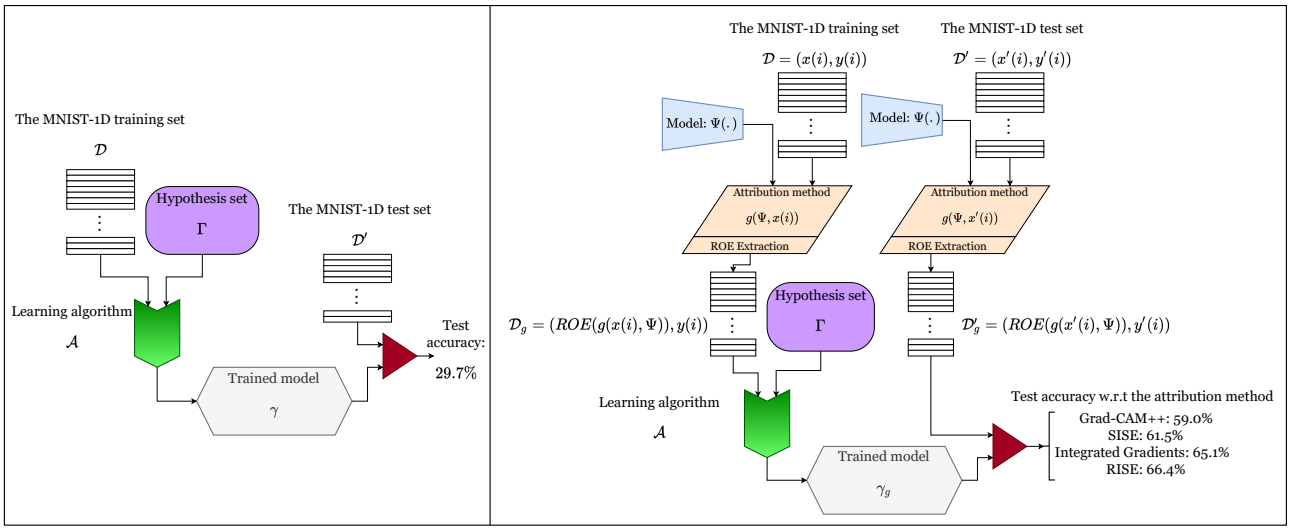

Figure 5. The schematic of the Region of Explanation (ROE) Understandability Test. Linear classifiers are not able to reach a high test accuracy, when trained on the original MNIST-1D training data (left subfigure). In the ROE understandability test, the original MNIST-1D data are replaced with the ROEs obtained by applying a visual explainabilty algorithm to the baseline model. The increase in the test accuracy of the linear classifier indicates the understandability of the employed visual explainability algorithm (the right subfigure).

We coin our proposed test as *ROE Understandability Test*. To conduct this test, we choose a 10-class linear classifier as the hypothesis set, which is denoted by $\Gamma : \mathbb{R}^L \rightarrow C$. The applied Support Vector Machine (SVM) [20] learning algorithm is defined as $\mathcal{A}$. This function receives the set of hypotheses Γ and a training set as input and returns a trained function $\gamma \in \Gamma$. Representing the training set of the MNIST-1D dataset as $\mathcal{D} = \{(x(i), y(i)) | i \in \{1, ..., N\}\}$ where $N = 4000$ is the number of training samples and $x(i)$ and $y(i)$ indicate the i-th training data and label respectively, the trained classifier is formulated as,

$$\gamma = \mathcal{A}(\Gamma, \mathcal{D}) \tag{2}$$

On the other hand, we denote the test set of the MNIST-1D as $\mathcal{D}' = \{(x'(i), y'(i)) | i \in \{1, ..., N'\}\}$ where $N' = 1000$ is the number of test samples and $x'(i)$ and $y'(i)$ represent the i-th test data and label, the test accuracy of the trained SVM is calculated as,

$$z(\gamma, \mathcal{D}') = \mathbb{E}_{i \in \{1,...,N'\}}[\gamma(x'(i)) \neq y'(i)] \tag{3}$$

In our benchmark, we replace the original training and test data with the ROEs derived by attribution methods when applied to the baseline model $\Psi(.)$. Each selected attribution method denoted as the function $g : (\Psi(.), \mathbb{R}^L) \rightarrow \mathbb{R}^L$, extracts the ROE from the input data as instructed in the previous subsections. For each input x, this region of explanation is notated as $ROE(g(x, \Psi))$. Hence, an attribution method g reformats the datasets $\mathcal{D}$ and $\mathcal{D}'$ as follows:

$$\mathcal{D}_g = \{(ROE(g(x(i), \Psi)), y(i)) | i \in \{1, ..., N\}\} \tag{4}$$

$$\mathcal{D'}_g = \{(ROE(g(x'(i), \Psi)), y'(i)) | i \in \{1, ..., N'\}\} \tag{5}$$

Considering these reformatted datasets, the linear classifier model trained with the training set $\mathcal{D}_g$ is formulated as:

$$\gamma_g : \mathbb{R}^L \rightarrow C = \mathcal{A}(\Gamma, \mathcal{D}_g) \tag{6}$$

The model's test accuracy R_g is evaluated by replacing the original test set with only the ROEs for the test data. The achieved accuracy rate is considered a metric to quantify the extent to which the attribution method g can transfer the baseline model's understanding to a human-understandable model.

$$R_g = \mathbb{E}_{i \in \{1,...,N'\}}[\gamma(ROE(g(x'(i), \Psi)) \neq y'(i)] \tag{7}$$

The test accuracy reached from the Eqn. (7) is the output of the ROE test when it is applied on the attribution method g. The higher this value is, the more soundness, completeness, contextfulness, and actionability (refer Sec. 5 for definitions) is offered by the method g.

The detailed methodology of the proposed ROE test is depicted in Fig. 5. Using any attribution method g that offers sound and complete explanations, the expected result is that pre-processing the training and test data by extracting the ROEs defined by the attribution method improves the linear model's classification accuracy $(z(\gamma_g, \mathcal{D'}_g) > z(\gamma, \mathcal{D}'))$, given that the baseline model is well-trained. The achieved higher test accuracy R_g indicates the better ability of the attribution method g in forming concrete, sound, and understandable explanations and its usability in real-world tasks.

4. Empirical Results

With the components introduced in the previous section, we implement the SVEA framework for the MNIST-

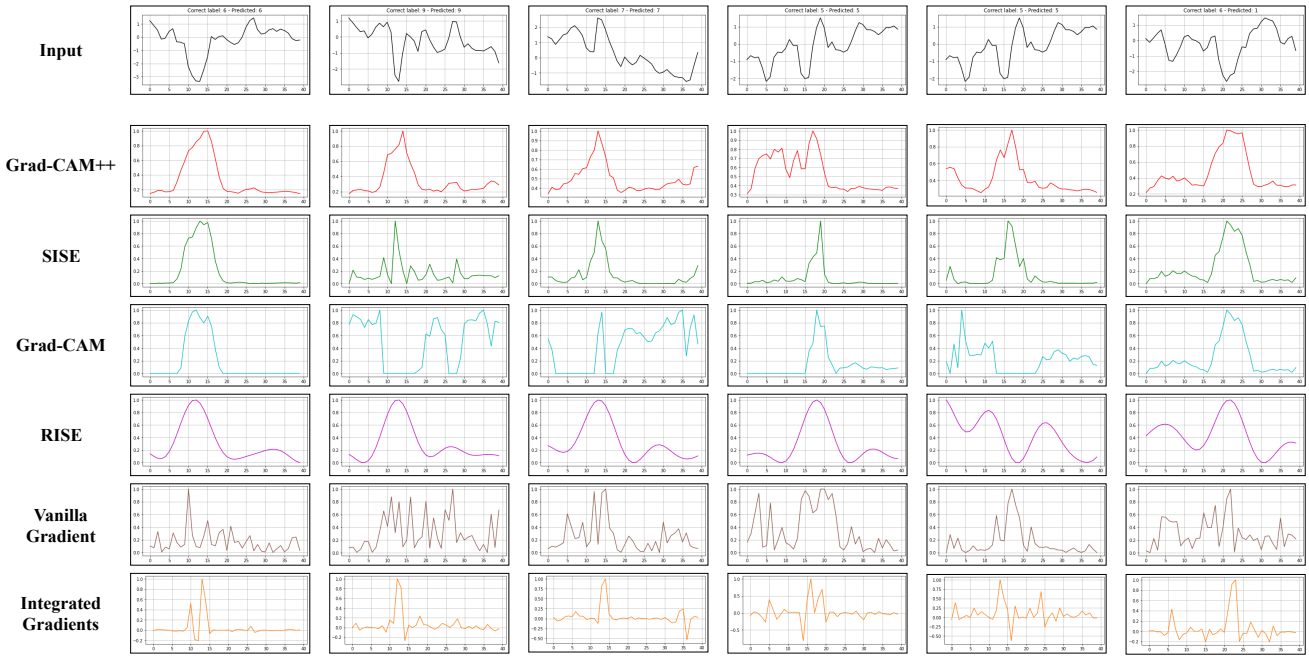

Figure 6. The qualitative results of six different attribution methods on multiple samples. The first four columns show the results for the data correctly classified by the baseline model, while the two latter columns depict the explanations for the data incorrectly classified by the baseline model.

1D dataset[1] along with various attribution methods (however, we have converted the source code to TensorFlow 2.x framework [1]). We have selected the following visual explanation methods to evaluate in our benchmark.

- **Backpropagation-based methods:** operate by backpropagating the signals from the model's output to its input or hidden features, e.g., Vanilla Gradient (VG) [29, 8], Integrated Gradients (IG) [32], SmoothGrad (SG) [30].
- **Perturbation-based methods:** run by feeding the target model with the perturbed copies of the input. e.g., Randomized Input Sampling for Explanation (RISE) [18] and Semantic Input Sampling for Explanation (SISE)[2] [25].
- **CAM-based methods:** are specialized for CNNs and aim to visualize the high-level features extracted by the convolutional units of these networks in a specific layer. e.g., Grad-CAM [27], Grad-CAM++ [5], Score-CAM [33], XGrad-CAM [9].

For each attribution method, a summarized fact sheet according to [31] is provided in the supplementary material that includes their methodology and other notable implementation details in our proposed benchmark.

[1] https://github.com/greydanus/mnist1d

[2] As far as the baseline model consists of only one convolutional block, we calculate SISE explanation maps by applying their framework only on the last convolutional layer.

4.1. Qualitative Analysis

Fig. 6 represent the explanations generated by six different attribution methods in few samples. Additional qualitative results for more samples and methods are included in our supplementary material. In summary, these results suggest that 1) The methods such as Grad-CAM++ and SISE [5, 25] that take account for the presence of smaller patterns detected by the baseline model show a more extraordinary ability in highlighting the related regions confidently, 2) In some cases, the "gradient saturation" problem which is addressed in prior works like [28] hurdles the ability of the methods such as Grad-CAM and Vanilla Gradient that highly rely on signal backpropagation [27, 29] to concretely estimate the importance of the input features, 3) The mentioned limitation in Grad-CAM and Vanilla Gradient are circumvented to a satisfying extent in some methods such as Integrated Gradients and RISE, by employing unique ideas such as perturbation-based analysis and calculating the path integral of input gradients, 4) However, in some cases, Integrated Gradients and RISE fail to explain the reason that the model is unable to make a correct prediction.

4.2. Quantitative Analysis

In addition to the ROE Understandability Test, we calculate a pair of *model truth-based* metrics named "Drop%" and "Increase%". It was initially introduced to compare the performance between Grad-CAM and Grad-CAM++ [5] and then expanded in later works [6, 33, 9]. This metric

Metric	Vanilla Gradient	XGrad-CAM	Score-CAM	Grad-CAM	Grad-CAM++	SISE	Integrated Gradient	RISE
ROE Test	32.7	39.3	42.1	47.6	59.0	61.5	65.1	66.4
Drop%	29.54	24.69	27.03	26.18	17.81	12.19	9.85	7.64
Increase%	29.6	37.7	28.9	36.4	33.9	38.5	40.4	41.0

Table 1. Results of the quantitative metrics applied on the state-of-the-art visual explanation methods. For Drop%, the lower is the better. For Increase% and our proposed metric (ROE Understandability Test), the higher is the better. All values are reported in percentage.

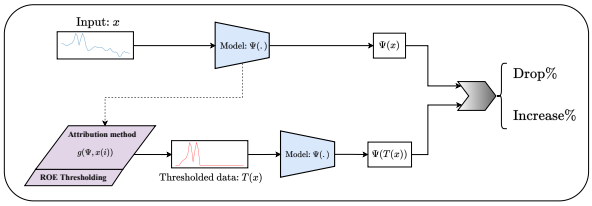

Figure 7. The framework for calculating Drop% and Increase% on individual data in our proposed benchmark.

assesses the faithfulness of attribution methods by probing the model's behavior when fed only with the features highlighted by the attribution method.

The intuition for "Drop%" is that when the essential features for the model's prediction are retained while the other features are masked, the model's confidence score should not drop significantly. In the same manner, the intuition for "Increase%" is that in some cases, by perturbing the unimportant features, the model's confidence in a prediction may increase. Unlike feature perturbation techniques in prior works, we perturb the features that do not fall into a region of explanation (ROE) as determined by the attribution method to be evaluated. The detailed methodology of calculating these metrics for a given attribution method is shown in Fig. 7.

To conduct the ROE Understandability Test, we employed an SVM learning algorithm to train 10 linear classifiers in a 'one-vs-rest' manner. We set the learning algorithm $\mathcal{A}$ to minimize the square of a hinge loss by applying 10,000 steps of the gradient descent optimization method. Since the learning algorithm configurations are fixed in the training procedure, all attribution methods are evaluated in this framework fairly. As reported in Fig. 5 and Table 1, in case of the usage of original MNIST-1D data to train the linear classifiers, the achieved test accuracy is too low (29.7%). However, processing the data by extracting the ROEs from the baseline model results in a remarkable improvement in training the linear classifiers. This improvement is expected since the well-performed attribution methods help us discard the spatial information that manipulates the linear classifiers while retaining the semantic patterns.

Moreover, the suggestions obtained through qualitative evaluation are verified by the results presented in Table 1. For instance, "Vanilla Gradient" that reaches the lowest scores in the ROE test also generates visually unclear and confusing explanations, especially for few samples whose template-related patterns are highly influenced by the applied transformations. The attribution methods that score features' importance in a size-invariant manner reach ROE test accuracy higher than 50%. Also, RISE and Integrated Gradients, two model-agnostic methods that address the "gradient saturation" problem in backpropagation-based methods, reach the top ranks both in ROE Test and in terms of Drop/Increase rates.

5. Discussion

According to the properties defined in [31], the highest scores achieved in our quantitative evaluation framework verifies the following properties in visual explainability algorithms:

- **Soundness:** Attribution methods should be able to represent how the baseline model discriminates the patterns related to different classes. The higher test accuracy achieved by the linear model depicts more solidity to represent the discrimination by the attribution method.
- **Completeness:** ROE Understandability Test measures how the explanations are generalized across the dataset applied to the baseline model. Unlike metrics such as Drop/Increase rates that score the faithfulness of the explanations separately, the ROE test evaluates all generated explanations across the target dataset simultaneously and in a unified interpretable framework.
- **Contextfulness:** The ROEs that the linear classifier fails to predict correctly indicate the explanations that lack understandability. Observing the linear model's predictions can help the users estimate unreliable explanations generated by the evaluated attribution method.
- **Actionability:** Through the ROEs given to the linear classifier, an actionable model can determine the importance of the input features. Since the classifiers are trained using the ROEs corresponding to the training data, the classifier's weight parameters corresponding to the correct class may imitate the guideline for the end-users in the visual explanations.

Below we discuss the advantages and limitations of our proposed benchmark and evaluation framework and its functionality for the individuals interested in researching Explainable AI.

5.1. Computational Complexity

In large-scale applications, applying evaluation metrics such as Remove and Retrain (ROAR) [12] that operate by training a network with a reformatted dataset require an enormous amount of GPU time. However, in our low-compute benchmark, the cost of performing this type of validation decreases remarkably.

For conducting the ROE understandability test, the computational cost of evaluating each attribution method is equivalent to 1) applying the attribution method to the entire data in the MNIST-1D dataset, 2) saving a dataset containing the extracted ROEs with a size of 30% of the MNIST-1D dataset, 3) training a classifier with 130 trainable parameters by the SVM algorithm. Using a CPU with a disk space that is extremely larger than the overall size of the SVEA benchmark components, training a multi-class linear classifier with 10,000 iterations takes only *17.9 seconds*, on average. Thus, performing this validation can not be considered a time-consuming or resource-exhausting process.

5.2. Transfer Ranking

Compared with Drop and Increase rates, the ROE Understandability Test ranks attribution methods with a slight variance. This slight variation in the standings is acceptable since the aspects based on which the attribution methods are evaluated differ (to some extent) between the ROE test and the prior validation frameworks. The differences between the rankings provided in Table 1 is because the Drop and Increase metrics quantify the faithfulness that the explanations had provided, while the scores assigned in the ROE Understandability Test are sensitive to a broader range of properties that attribution methods should satisfy.

5.3. The Limitations of the SVEA Benchmark

Despite that our proposed benchmark is functional for measuring the concreteness and correctness of the visual explanation algorithms in a small-scale experimental environment, some methods' performance may fluctuate when applications with larger scales are included. Of course, interpreting deeper machine learning models with complicated structures and millions of trainable parameters is a more challenging task for all attribution methods compared to the baseline model in the SVEA benchmark.

For instance, when CAM-based methods are applied to deep CNNs with several convolutional blocks, they would generate blurry and noisy explanation maps since they run by visualizing the deepest convolutional layer of the CNN's feature extractor. Later works such as [23, 25, 17] attempt to circumvent this issue by aggregating the information obtained by visualizing multiple layers of the CNN. Also, the RISE methods suffer from the same problem while dealing with high-dimension inputs. In this method, generating explanations using random perturbation masks distributes the energy in explanation maps across the whole input domain.

Besides, in image processing applications where numerous high-level features may be formed from the interaction between image pixels, backpropagation-based methods like Vanilla Gradient and Integrated Gradient usually produce extremely sparse explanation maps. Though the method Integrated Gradients shows outperforming soundness and completeness in our small-scale benchmark, empirical results in prior works suggest that this method fails to analyze abstract features detected by the target model in large-scale applications [33, 25].

Another shortcoming of the SVEA benchmark is that this framework cannot measure the complexity of attribution methods when applied in large-scale tasks. For example, the RISE method ranked first in our leaderboard, operates extremely slow in image recognition tasks, as this method works by feeding the target model with numerous masked copies of the input image [18]. The SISE method decreases this computational overhead substantially by eliminating the need for employing random masks and replacing them with smaller sets of attribution masks [25]. This simplification is not apparent to measure in a computationally inexpensive setup.

6. Conclusion

In this work, we employed the MNIST-1D dataset to create SVEA, a low-memory and minimalist benchmark for evaluating the visual explanations generated by state-of-the-art attribution methods in small scales, before transforming the empirical results to large scales, such as image processing experiments. The SVEA benchmark eliminates the need for conducting exhaustive experiments to perform high-level quantitative evaluations. We also proposed the ROE Understandability Test, a function-level validation metric that compares an attribution method's usability from numerous aspects. Our experiments' empirical results show a high correlation between our proposed metric and the prior evaluation frameworks. We believe that our proposed benchmark and evaluation metric becomes a stepping stone for future research in the field of Explainable AI.

References

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82 – 115, 2020.

[5] A. Chattopadhay, A. Sarkar, P. Howlader, and V. 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, 2018.

[6] S. Desai and H. G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 972–980, 2020.

[7] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[8] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009. Also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.

[9] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

[10] Sam Greydanus. Scaling down deep learning, 2020.

[11] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[12] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9737–9748. Curran Associates, Inc., 2019.

[13] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1675–1684, New York, NY, USA, 2016. Association for Computing Machinery.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, Dec. 1989.

[15] Zachary C. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, Sept. 2018.

[16] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Anil K. Jain, and Jiliang Tang. Trustworthy ai: A computational perspective, 2021.

[17] Fanman Meng, Kaixu Huang, Hongliang Li, and Qingbo Wu. Class activation map generation by representative class selection and multi-layer feature fusion. *arXiv preprint arXiv:1901.07683*, 2019.

[18] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[19] Vipin Pillai and Hamed Pirsiavash. Explainable models with consistent interpretations. *UMBC Student Collection*, 2021.

[20] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[21] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability, 2021.

[22] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable ai, 2018.

[23] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020.

[24] 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*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[25] Sam Sattarzadeh, Mahesh Sudhakar, Anthony Lem, Shervin Mehryar, K. N. Plataniotis, Jongseong Jang, Hyunwoo Kim, Yeonjeong Jeong, Sangmin Lee, and Kyunghoon Bae. Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation, 2020.

[26] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

[27] 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.

[28] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[29] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

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

[31] Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 56–67, New York, NY, USA, 2020. Association for Computing Machinery.

[32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[33] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[34] Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance, 2019.

[35] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Stan Sclaroff, and Adel Bargal, Sarah. Top-down neural attention by excitation backprop, 2016.