Maximum Entropy Baseline for Integrated Gradients

Hanxiao Tan  
AI Group  
TU Dortmund  
Dortmund, Germany  
hanxiao.tan@tu-dortmund.de

Abstract—Integrated Gradients (IG), one of the most popular explainability methods available, still remains ambiguous in the selection of baseline, which may seriously impair the credibility of the explanations. This study proposes a new uniform baseline, i.e., the Maximum Entropy Baseline, which is consistent with the “uninformative” property of baselines defined in IG. In addition, we propose an improved ablating evaluation approach incorporating the new baseline, where the information conservativeness is maintained. Furthermore, we explain the linear transformation invariance of IG baselines from an information perspective. Finally, extensive experiments indicate that IG with Maximum Entropy Baseline performs superiorly and the ablation tests derived from it are more plausible.

I. INTRODUCTION

Integrated Gradients (IG) [1] is one of the most widely used explainability methods at present, which illustrates the attribution of each pixel in the input x to the prediction of the model F. IG is a gradient-based approach that addresses the gradient saturation problem [2] of vanilla gradients by integrating them from a chosen baseline x’. IG is formulated as:

\[ IG_i = (x_i - x'_i) \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x} \, d\alpha \]  \hspace{1cm} (1)

The baseline is one of the most important parameters of IG, which strongly impacts the performance of the explanations. The existing studies, although specifically suggesting different baselines, are almost unanimous in their overall definition: missingness, i.e., the input that disables the model from capturing information. Currently the most prevalent baselines are: 1) Zeros: filling all input units with zeros. 2) Black (white) vectors: replace all units with the minimum (maximum) of the current input. 3) Random initializations. In addition, [3] complements several possible alternatives: 4) Max-distance pixels (Xdist): their study reveals that the explanations generated by IG tends to be more sensitive to pixels that differ significantly from the baseline, while ignoring those that are similar, even if they are located inside the object to be explained. Thus, they propose Xdist, i.e., replacing each pixel with the point that is farthest apart in chromaticity space. 5) Blurring: information is eliminated by blurring the input with kernel functions.

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B.
of the entropy.

- We quantitatively compare all possible baselines with the proposed evaluation metrics on different data sets.

The paper is structured as follows: Section II introduces the current studies on IG baselines and ablation tests. Sections III and IV elaborate the experiments related to the proposed baseline and ablation test, respectively. Section V presents the results of the experimental evaluation of the proposed methods.

II. RELATED WORK

Investigation of IG baseline. So far, there are no sufficient researches for IG baselines. Several studies [3], [6], [7] point out that the choice of baseline impacts dramatically on the explanation performance, with the appropriate baseline generating clearer attribution maps and the opposite distorting them. [6] also indicates that a number of specific baselines (e.g., zeros) do not satisfy linear transformation invariance and are hence unreliable. The proposer of IG [1] provides a definition of baselines, i.e., when the model prediction is neutral, and suggest that for most networks the zero baseline is applicable, while also specifying different applications for networks in different domains, e.g., the black background for image networks and the zero embedding vector for NLP tasks. As a complement, [7] proposes several additional baselines that are intuitively feasible, such as the Max-Distance and Blurred baselines. Although these baselines are either consistent with human intuition or address certain deficiencies of existing ones, there is no convincing argument that the proposed baselines are “uninformative”.

Ablation test. In LRP [4], the ablation test is first proposed as a validation for the explainability method, which has been followed or expanded in several studies relating to feature-wise attributions [5], [8]–[10]. However, [11] pointed out that the method suffers from the trouble that ablating a single feature might impair the feature correlation and data distribution, and they suggest that a new model should be retrained after ablating for accuracy validation (Remove & Retrain). Again, this also leads to debates that the explanations should be faithful to the data or the model [12], [13]. Leaving aside the controversy about feature correlation, we expose another potential risk of ablation experiments from the information perspective, i.e., non-conservation, and propose a corresponding solution.

III. MAX ENTROPY BASELINE

The ideal baseline is the input that “is neutral” [1] and “contains no information” [3]. However, measuring the amount of information embedded in the inputs is challenging since most models are black boxes and only final predictions are observable. On the other hand, we argue that the referent of “neutrality” is ambiguous. Existing referents for the baseline are 3 categories: uninformative for humans, data and models. Most baseline choices are for humans, such as zero, black and random baselines: it is intuitively assumed by humans that these values do not yield any information. Nevertheless, for the model to be explained, this may be problematic (e.g., a classifier that distinguishes between black and white images). Subsequently, those baselines for the dataset (e.g., Average of training data) also raise controversy: whether the explanation should be true to the data or the model [3], [12], [13]. So far, there is hardly any baseline for the model being proposed. To address the aforementioned flaws, we introduce the entropy of logits as a metric for quantifying information residual in the model. We denote the proposed baseline as $B_{X_{entn}}$, and is formulated as:

$$B_{X_{entn}} = \arg\max_x H(\text{Softmax}(f_i(x)))$$

where $f_i$ denotes the logits of the model and $H(\cdot)$ denotes the entropy function [14]:

$$H(A) = -\sum_{i=1}^{n} P(a_i)\log P(a_i)$$

Our newly proposed baseline has the following advantages:

- **Input information excluded.** Several baselines incorporate (or incompletely ablate) information from input instances to attain better visualizations. We argue that this violates the definition of baselines. For example, The Maximum Distance baseline (Xdist) [3], which calculates the maximum colorimetric distance of each pixel from the input, moderates the phenomenon of “attribution disappearance” in the explanations. Nevertheless, this baseline obviously contains extensive information about the input, such as object outlines (see figure 1).

- **Feature correlation Incorporated.** The plausibility of generating explanations based on the assumption of feature independence remains questionable. Owing to the strong correlation between features, learning a certain distribution based on the dataset is a more convincing solution [3]. Alternatively, the baseline we proposed is derived from the optimization of a trained model, which itself possesses the distribution of the dataset.

- **Linear transformation invariant.** [6] suggests that the generated explanation should remain constant when the input and the model undergo a uniform linear transformation (all pixels transformed with the same amplitude). A portion of the baselines fails this test (e.g., zero baseline), while several others (e.g., black baseline and ours) survive. In addition, we elucidate from an information perspective why all baselines fail the non-uniform linear transformation test in section III-C.

- **Computational simplicity.** Only a simple gradient ascent procedure is required once for each model to obtain $B_{X_{entn}}$, and is applicable to all predictions made by the model.

Figure 1 shows all currently available IG baselines and their generated explanations. Ahead of assessing the validity of the proposed baseline, we show our observations with respect to the correlation between the entropy of logits and the explanations of IG. In section III-A we show experimental observations in a tabular toy dataset, and in section III-B we...
extend the findings to the MNIST handwriting dataset. Note that for achievable computational complexity, in this section the baseline is treated as a uniform value ($B_{\text{Xentr}_w}$).

### A. Warm-up: tabular toy datasets

The lack of ground truth explanations is one of the obstacles to investigating baselines. Before experimenting on the real dataset, we simplified the problem by selectively creating features relevant to the labels to obtain ground truth attributions in advance. We artificially create tabular toy datasets whose ground truth explanations are available. An overview of the dataset is shown in figure 2. Our dataset involves 3 parameters, namely $f$, $k$, and $c$, representing the total number of features, the number of features correlated with the label and the total number of label categories, respectively. For each individual dataset, there is at least one feature correlated with the label ($k \geq 1$). Hence, we can obtain a unique ground truth explanation for this dataset, i.e., an attribution of 1 for the features relevant to the labels and 0 for the opposite.

Note that when $k > 1$, the labels must be computed jointly based on all relevant features, otherwise the model may learn only a part of them, leaving the ground truth explanations redundant. Each dataset consists of $10^4$ instances (most of the data are duplicates due to the limited combination of features), of which 80% are used as training data and the rest are for testing. A simple two-layer fully connected network is trained, which achieves 100% accuracy on both the training and test sets. We traverse the explanations of all the baselines in the valid data value range with 100-step IG. Owing to the high dimensionality, KL-loss struggles to reveal the distributional divergence, i.e., features related to the label should possess as large attributions as possible, otherwise should approximate to 0. It is notably that only the results of the instances where all features are 0 as the object being explained are shown, since according to the properties of the dataset, the ground truth explanations should be independent of the input instances, and we practically tested the input instances for all feature combinations and the results are consistent. On the other hand, we draw the entropy curve of the model over the valid range of data values. The prediction are fed into a Softmax function to assure that each logits neuron is positive, and then the entropy of this probability vector is calculated.

The results are reported in figure 3. Interestingly, the loss and entropy curves follow similar trends and reach extreme values at approximate baselines. These results are intuitively correct: the maximum predictive entropy implies the minimum information content of the input vector, which exactly coincides with the definition of the baseline in IG.

### B. Observation: MNIST handwriting dataset

To extend this conclusion to a more general scenario, a similar experiment is conducted on the MNIST handwritten dataset. However, the major challenge with real datasets compared to the previous one is that no prior knowledge is available about which features (pixels) in the instance are decisive for the labels, i.e., the ground truth explanations are not accessible. Existing studies treat those pixels located inside the object as ground truth [15], [16]. Nevertheless, no guarantee can be given that the model does not utilize any information from the background when inferring [17]. Our alternative approach is to observe the hybrid explanations generated by multiple explainers (average of explanations generated by Back-Propagation-Based, Taylor Decomposition, LRP and DeepLift series). Notwithstanding the inability to precisely restore the ground truth explanations, this approximated saliency map captures the trend of the attribution distribution given by the majority of explainers. Before training, all data are transformed to the value domain $[-0.42, 2.82]$ for higher accuracy. We train two different types of networks, the FC networks consisting of only fully connected layers and the CNN networks containing convolutional layers. Similarly, we search all the baselines in the (discretized) valid value range and produce the explanations by a 100-step IG. Owing to the high dimensionality, KL-loss struggles to reveal the distributional...
discrepancy, we therefore adopt Spearman’s rank correlation coefficient to measure the similarity between explanations. We randomly selected 100 instances from the test set for each model. The methodology of statistics is that we initialize a histogram containing \( m \) bins, each corresponds to a baseline within the (discretized) valid value range. For each instance, we try all \( m \) baselines and yield \( m \) explanations, we then compare each of them with the hybrid explanation. The baseline that generates the closest explanation to the hybrid one (with maximum Spearman’s coefficient) is counted in the corresponding bin. The curve of logits entropy is also plotted.

The results are demonstrated in figure 4 and the maximum of both recordings are annotated with a dashed line in the corresponding color. As can be observed from FC1 and CNN1, the baseline with the lowest loss almost coincides with the input of the maximum logits entropy. To exclude the possibility that the minimum loss baseline is around 0, two additional models are trained whose input of maximum logits entropy deviate far from the origin point (FC2 and CNN2). Although the minimum loss distributions are not perfectly concentrated at the maximum entropy values, significant following offsets can be observed. Additionally, we find that the reason why zero is feasible as a baseline alternative is that the global maximum of the entropy curve for a portion of neural networks lies approximately adjacent to the origin point. Nevertheless, the opposite also exists, for instance, the FC2 and CNN2 in our experiments. FC2 is derived from performing an identical linear transformation on the dataset and the bias of the first layer of the network for FC (the experiment is first proposed by [6] and is described in detail in the next section), while CNN2 is a convolutional neural network with more sophisticated architectures. The maximum values of their entropy curves deviate from the origin, at which point the zero baseline can no longer be considered as an approximation, while the baseline closest to the hybrid explanations can be observed to be shifting towards the positive direction of the X-axis, rather than remaining at the origin.

C. Unreliable? Linear shifts on baselines

Previous study point out that a fraction of the IG baselines fail the linear transformation test [6]. They shift the input instances and the parameters from the first layer of the model with the same linear offset, and subsequently observe whether the explanations generated by the explainer are consistent before and after the transformation. In the experiment, two alternative shifts are adopted as offsets, i.e., the uniform and non-uniform shifts, where the former is equally shifted at each pixel, while the latter is not subject to this restriction. They utilize black and zero baselines for IG, while only the former maintains the same explanation before and after the uniform shift. We reproduce the experiments and present them in figure 5 (the shift amplitudes, including all uniform shifts and the non-zero pixels of the cross-shaped shifts, are 0.5). We argue that the black baseline is adaptive and shifts with the input, whereas the zero baseline remains constant, which lacks fairness. Therefore, we transform the zero baseline with the same offset and observe the explanation invariance. The results are shown in the first row where the zero baseline maintains the consistent explanation after the transformation. The second row exhibits the results of the non-uniform shift, and interestingly, when we perform non-uniform shift for the zero baseline, the explanation is irreducible.

This can be explained by the entropy curves. Figure 6 plots the curves of the three models in parts of the valid value range. Suppose the entropy curve of the original model is \( H(x) \) (black curve), which turns into \( H_U(x) \) after the uniform linear transformation (red curve), and the transformation amplitude is \( A_s \). The correlation can be easily derived from the diagram:

\[
H_U(x) = H(x - A_s)
\] (4)

Equation 4 indicates that the information content of the two models is identical except for a phase difference of
Fig. 4. The entropy curves of the logits and density histograms of baselines which achieve the minimum Spearman loss with the hybrid explanations (HE). The red and blue dashed lines indicate the maximum values of the logits entropies and density boxes, respectively.

Fig. 5. Visualizations of the uniform and non-uniform linear transformations on the instances and baselines. The last three columns are the corresponding explanations. IG mask hold demotes the linear transformation is applied only to the instances, whose baselines remain non-transformed.

As. However, the model with cross-shaped transformation possesses a severely distorted entropy curve (blue curve). When the original model is explained with IG, the gradient is accumulated from an “uninformative” initiation to the destination to be explained, and an identical path can be found in the uniformly transformed model, with the starting and ending points translated by $A_s$ (e.g., bolded black and red segments), while the integrals are equivalent. In contrast, such a segment is absent in the curve of the model with the cross-shaped transformation. Consequently, linear invariance is only adequate for transformations that do not distort the model entropy curve.

IV. ABLATION-BASED EVALUATION METHODS

In this section, we elaborate on the flaw of the classical ablation-based evaluating metric for explainability methods, which monitors the prediction activations in IV-A, and propose a novel one, which targets the entropy as the surveillance IV-B.

A. Existing ablation test

The ablation approach and its variants are the most commonly used methods to evaluate explanations, which is comprehensively summarized in [3]. Let $x \in X$ denote an instance containing $n$ features $x_o = (x_o^1, x_o^2, ..., x_o^n)$ ($X$ is the set of all valid instances), the information quantity it contains before the ablation test is initiated as $I(x)$. After all $n$ features are ablated, the input now is noted as $x_\phi = (\phi^1, \phi^2, ..., \phi^n)$, where $\phi_s$ denotes the ablated substitutions, and the information quantity it holds is marked as $I(x_\phi)$, the ablation evaluation can be formulated as:

$$S_e \propto I(X_o) - I(X_\phi)$$

where $S_e$ indicates the score of the explainer and $X^p_\phi$ indicates the removal of $p$ positively attributed features.

Several studies [4], [5], [8] employed ablation approaches to validate the reliability of explainability methods, while differing slightly in the details. The major distinctions are:

- **Surveillance target.** Let $l$, $f$ represent the prediction label and the model to be explained. Frequently monitored objects are collectively defined as “prediction scores” and are divided into 1. the corresponding activations in logits: $f_l(x)$. 2. the corresponding cells in logits after taking
Softmax, which can be considered as the probability of the class: $\sigma(f_l(x))$. In these works, they consider the reduction of prediction scores as a result of “information removal” [5], the surveillance object is regarded as a measurement of “information quantity”, i.e., $I \approx \Delta f_l(x)$ or $\Delta \sigma(f_l(x))$.

- **Ablation destination.** The substitution value for ablation is also controversial. There are several competitors: zero, minimum value of the dataset, reversing the sign of the current pixel [4], blurring via Gaussian kernel [3] etc. The purpose of ablation is to conceal information about a specific pixel (or feature), the flipping destinations themselves should therefore carry no information, i.e., $I(x_\phi) \approx 0$.

However, we argue that such metrics suffer from non-conservation. According to the definition of the ablation test, the information quantity of the ablated input $x_\phi$ should satisfy:

$$\forall x \in X, I(x) \geq I(x_\phi)$$

(6)

Note that practically the information quantity is not directly available and the existing methods record the value of logits $f_l$ (or the softmax of the logits $\sigma(f_l)$) as substitutes. Thus, Equation 6 can be rewritten as:

$$\forall x \in X, f_l(x) \geq f_l(x_\phi) \text{ or } \sigma(f_l(x)) \geq \sigma(f_l(x_\phi))$$

(7)

Taking the CNN in Section III-B as an example, we employ a 1000-step gradient descent algorithm to minimize the value of the surveillance target and record the changing curve. Subsequently, we feed various $x_\phi$ into the model, derive the individual predicted values, and annotate them on the curve. Notably, the optimization process is performed in the valid range of the original dataset. According to figure 7, equation 7 is violated by the optimization results, and numerous inputs with lower monitoring target values can be identified in the valid range. If $f_l(x)$ is considered as the information quantity, $x_\phi$ suffers from extensive information redundancy (left plot). Compared to the former, $\sigma(f_l(x_\phi))$ mitigates this deficiency, while the information residual in $x_\phi$ is still visible (middle plot). Experiments show that the ablation test designed on the basis of inputs that are commonly understood by humans as “uninformative” may still contain information and may be problematic.

**B. Entropy-based ablation test**

To fulfill the conservativeness in Equation 6, we extend the definition in equation 2, that monitors the entropy of logits as the quantitative indicator of information, i.e.

$$I(x) \approx \frac{1}{H(\sigma(l))}$$

(8)

The increment of information diminishes the uncertainty of events, which can be expressed by the entropy. The introduction of entropy facilitates the understanding of “information quantity”. On the other hand, by redefining the unique ablated destination as $x_\phi = B_{X_{entr}}$, the conservativeness of the ablation assessment is assured

$$I(B_{X_{entr}}) \approx \frac{1}{H(\sigma(f_l(B_{X_{entr}})))}$$

(9)

Recalling equation 2, which yields

$$\forall x \in X, H(\sigma(f_l(x))) \leq H(\sigma(f_l(B_{X_{entr}})))$$

(10)

and refers to equation 8, equation 6 holds. To experimentally verify the conservativeness, we depict the entropy curve and mark out $B_{X_{entr}}$, as shown in figure 7 (right plot). Note that in this plot, the area above the entropy curve represents the information residual, which does not exist since $B_{X_{entr}}$ is the input that maximizes the entropy.

**V. Quantitative evaluations**

We choose two artificial and one real-world datasets for evaluation experiments: MNIST, CIFAR10 and Stanford Car Dataset. For MNIST, we train a fully connected network, which achieves 98.2% accuracy on the test set. For CIFAR10, we train a ResNet18 [18] network, whose test accuracy is 95.6%. For Stanford Car Dataset, we trained a ResNet152 with 92.2% accuracy. During evaluation, for MNIST we assess all 10,000 test data, while on the remaining two we select 1,000 examples from the dataset for evaluation.

As illustrated in Fig. 8, $X_{entr_u}$ and $X_{entr}$ are on par with or superior to other baselines in the same categories (i.e. $X_{entr_u}$ and $X_{entr}$ rank first, second and second amongst the Uniform and Non-Uniform baselines, respectively). Interestingly, the Max Distance baseline consistently performs worse than the remaining ones in the non-uniform baseline. There is extensive information about the input instances in the Max Distance baseline, including object boundaries, gray value information (see Figure 1), which validates our view that the more information about current instances is contained in the baseline, the less is integrated by IG and thus the generated explanation is weaker in terms of credibility.

Moreover, to exhibit the superiority of the improved maximum entropy ablation test, we compare the other ablation methods as well. We choose FC2 in Sec. III-B as the model for the reason that the maximum entropy baseline of FC2 deviates from the origin, enabling the results to be obvious at a glance. To exhibit the plausibility of the maximum entropy ablation, four different ablation tests are performed and shown in Fig. 9. The ablation tests with zero and the maximum value of the dataset as the destinations fail to reasonably evaluate the performance of explanations (randomly generated explanations outperform the IG with any baselines, which is counterintuitive). The reason that the minimum ablation also exhibits reasonable results is that, for FC2, the minimum value (1) is exactly approaching the maximum entropy baseline (see figure 4). Note that the evaluation experiments in this section are not absolutely precise. There are still numerous unaddressed issues in the ablation study, for example, perturbing a pixel corrupts the correlation between neighboring pixels. A potential future work is to enable ablation studies to take into account the maximum entropy theory and pixel-wise correlations simultaneously.
Fig. 7. The ablation experiments with the raw logits activations (left), softmax of logits activations (middle) and the entropy of logits as the surveillance target (right), respectively. The x-axis denotes the number of optimization steps and the y-axis represents the value of the corresponding monitored target. The shaded areas with colors represent “information residues”.

Fig. 8. Ablation tests on different IG baselines on MNIST, CIFAR10 and Stanford Car datasets. Bars from left to right: randomly generated saliency maps (comparison reference), zero, black, white, average of current input, max entropy with uniform values, max distance, average of training data, blurred, uniform, Gaussian noise and max entropy baselines, respectively. The red boxes represent baselines with uniform values on all pixels, while the blue boxes are free of this restriction. The bolded black bar in the box is the median, and the horizontal dashed line indicates the optimal median value of the baseline in the current category (uniform or non-uniform).

More importantly, such an “uninformative” baseline is not only desired in IG. Explainability methods such as KernelSHAP [19], Occlusion [20] and RISE [21] have all introduced the concept of information absence. Taking KernelSHAP as an example, the model substitutes part of the features with the baseline when sampling around the instance to be explained. Therefore, employing perturbation baselines with residual information may impair the performance of the surrogate model and thus diminish the credibility of the explanations (see Table I for the evaluation tests). Therefore, a well-accepted and effective baseline is pivotal not only for IG, but also for broader explainability research.

VI. CONCLUSION

This work identifies the conservation deficiencies of the existing IG baselines and ablation tests from an informational perspective, and proposes a new baseline and an enhanced ablating evaluation method based on the “missingness” necessitated by the explainability approaches. However, we acknowledge that existing ablation tests are still controversial in terms of, for example, feature correlation. In future work, we will direct our efforts to investigating more persuasive assessments for explainability methods.

REFERENCES

[1] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in Proceedings of the 34th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3319–3328. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html
[2] ——, “Gradients of counterfactuals,” 2016, arXiv preprint, arXiv:1611.02639.
[3] P. Sturmfels, S. Lundberg, and S.-I. Lee, “Visualizing the impact of feature attribution baselines,” Distill, vol. 5, no. 1, p. e22, 2020.
[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” PloS one, vol. 10, no. 7, p. e0130140, 2015.
[5] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, “Evaluating the visualization of what a deep neural network has learned,” IEEE transactions on neural networks and learning systems, vol. 28, no. 11, pp. 2660–2673, 2016.
[6] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, “The (un) reliability of saliency methods,” in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019, pp. 267–280.
[7] Z. Qi, S. Khorram, and F. Li, “Visualizing deep networks by optimizing with integrated gradients.” in CVPR Workshops, vol. 2, 2019.
[8] G. Montavon, “Gradient-based vs. propagation-based explanations: An axiomatic comparison,” in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019, pp. 253–265.
[9] T. Zheng, C. Chen, J. Yuan, B. Li, and K. Ren, “Pointcloud saliency maps,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1598–1606.
Fig. 9. Four different ablation tests. The ablation destinations are: zero (top left), the minimum (top right) and maximum (bottom left) values of the data set and the maximum entropy baseline (bottom right), respectively.

| Scores | Random | Uniform Baselines | Non-uniform Baselines |
|--------|--------|-------------------|-----------------------|
| Rdm.   | Zero   | Black | White | Avg. | X_{ent}\text{ru} | Xdist | TD. | Blur | UF. | Gaus. | X_{ent} |
| $\bar{S}$ | 0.157 | 0.582 | 0.571 | 0.208 | 0.585 | 0.581 | 0.318 | 0.570 | 0.534 | 0.408 | 0.163 | **0.593** |
| $\bar{\bar{S}}$ | 0.059 | 0.598 | 0.582 | 0.142 | 0.599 | 0.594 | 0.298 | 0.582 | 0.544 | 0.402 | 0.054 | **0.608** |
| $\sigma^2(S)$ | 0.044 | 0.074 | 0.079 | 0.049 | 0.076 | 0.075 | 0.059 | 0.073 | 0.070 | 0.068 | 0.054 | 0.075 |

**TABLE I**

EVALUATION EXPERIMENTS OF KERNELSHAP WITH VARIOUS BASELINES.

[10] H. Tan and H. Kotthaus, “Surrogate model-based explainability methods for point cloud mms,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2239–2248.

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

[12] M. Sundararajan and A. Najmi, “The many shapley values for model explanation,” in International conference on machine learning. PMLR, 2020, pp. 9269–9278.

[13] D. Janzing, L. Minorics, and P. Blöbaum, “Feature relevance quantification in explainable ai: A causal problem,” in International Conference on artificial intelligence and statistics. PMLR, 2020, pp. 2907–2916.

[14] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in In Workshop at International Conference on Learning Representations. Citeseer, 2014.

[16] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, “Top-down neural attention by excitation backprop,” *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[17] L. Arras, A. Osman, and W. Samek, “Ground truth evaluation of neural network explanations with clevr-xai,” *arXiv preprint arXiv:2003.07258*, 2020.

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

[19] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in neural information processing systems, vol. 30, 2017.

[20] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[21] V. Petsiuk, A. Das, and K. Saenko, “Rise: Randomized input sampling for explanation of black-box models,” *arXiv preprint arXiv:1806.07421*, 2018.