NoiseGrad — Enhancing Explanations by Introducing Stochasticity to Model Weights

Kirill Bykov*,[1, 2] **, Anna Hedström***,[1, 2] **, Shinichi Nakajima**[1, 3] **, Marina M.-C. Höhne**[1, 2]

[1] ML Group, TU Berlin, Germany
[2] Understandable Machine Intelligence Lab
[3] RIKEN AIP, Tokyo, Japan
kirill.bykov@campus.tu-berlin.de, anna.hedstroem@tu-berlin.de, nakajima@tu-berlin.de, marina.hoehne@tu-berlin.de

Abstract

Many efforts have been made for revealing the decision-making process of black-box learning machines such as deep neural networks, resulting in useful local and global explanation methods. For local explanation, stochasticity is known to help: a simple method, called *SmoothGrad*, has improved the visual quality of gradient-based attribution by adding noise to the input space and averaging the explanations of the noisy inputs. In this paper, we extend this idea and propose *NoiseGrad* that enhances both local and global explanation methods. Specifically, NoiseGrad introduces stochasticity in the weight parameter space, such that the decision boundary is perturbed. NoiseGrad is expected to enhance the local explanation, similarly to SmoothGrad, due to the dual relationship between the input perturbation and the decision boundary perturbation. We evaluate NoiseGrad and its fusion with SmoothGrad — *FusionGrad* — qualitatively and quantitatively with several evaluation criteria, and show that our novel approach significantly outperforms the baseline methods. Both NoiseGrad and FusionGrad are method-agnostic and as handy as SmoothGrad using a simple heuristic for the choice of the hyperparameter setting without the need of fine-tuning.

Introduction

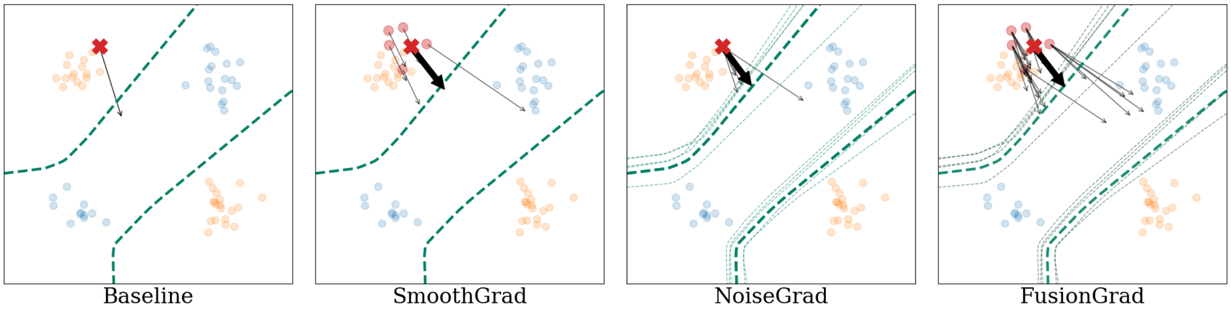
The ubiquitous usage of Deep Neural Networks (DNNs), fueled by their ability to generalize and learn complex non-linear functions, has presented both researchers and practitioners with the problem of non-interpretability and opaqueness of Machine Learning (ML) models. This lack of transparency, coupled with the widespread use of these highly complex models in practice, represents a risk and a major challenge for the responsible usage of artificial intelligence, especially in security-critical areas, e.g. the medical field. In response to this, the field of eXplainable AI (XAI) has emerged intending to make the predictions of complex algorithms comprehensible for humans.

One possible dichotomy of post-hoc explanation methods can be carried out on the basis of whether these methods refer to the global or local properties of a learning machine. The local level XAI aims to explain a model decision of an *individual* input (Guidotti et al. 2018), for which various methods, such as Layer-wise Relevance Propagation (LRP) (Bach et al. 2015), GradCAM (Selvaraju et al. 2019), Occlusion (Zeiler and Fergus 2014), MFI (Vidovic et al. 2016), Integrated Gradient (Sundararajan, Taly, and Yan 2017), have proven effective in explaining DNNs. In contrast, global explanations aim to illustrate the decision process as a whole, without the connection to individual data samples. Recently, methods belonging to the Activation-Maximization (Erhan et al. 2009) family of methods have become widely popular, such as DeepDream (Mordvintsev, Olah, and Tyka 2015), GAN-generated explanations (Nguyen et al. 2016) and Feature Visualization (Olah, Mordvintsev, and Schubert 2017).

For local explanation, gradient-based methods are most popular due to their simplicity, however, they tend to suffer from the gradient shattering effect, which often results in noisy explanation maps (Samek et al. 2021). As a remedy, Smilkov et. al. proposed a simple method, called SmoothGrad (Smilkov et al. 2017), where stochasticity is introduced to the input. Specifically, it adds Gaussian Noise to the input features n times, computes the n corresponding explanations, and takes the average over the n explanations. SmoothGrad is applicable to any local explanation method and has been practically proven to reduce the visual noise in the explanation map.

The mechanism behind SmoothGrad's enhancement of explanations is not yet well understood. One could argue that SmoothGrad averages out the shattering effect. However, SmoothGrad performs best when the added noise level is around 10%–20% of the signal level, which not only smooths out peaky derivatives but is large enough to cross the decision boundary. From this fact, we hypothesize that SmoothGrad perturbs the test sample in order to get a signal from the steepest part of the decision boundary. This motivated us to explore another way of using stochasticity: instead of adding noise to the input, our proposed method — NoiseGrad (NG) — draws samples from the network weights from a *tempered* Bayes posterior (Wenzel et al. 2020), such that the decision boundaries of some models are close to the test sample, which results in more precise explanations.

Our hypothesis leads to a natural and easy way of hyperparameter choice: the noise level added to the weights (which corresponds to the temperature of the tempered Bayes posterior) is chosen such that the relative perfor-



Baseline

SmoothGrad

NoiseGrad

FusionGrad

Figure 1: Illustration of the differences in explanation behavior between Baseline (gradient-based explanation), SmoothGrad, NoiseGrad, and FusionGrad for a toy experiment. Given training samples of two classes (orange and blue dots), a 3-layer MLP was trained for binary classification, where the learned decision boundary is shown by the green dashed line. The gradient explanations for a fixed test sample (red point) are shown as black arrows and the mean explanation as a bold black arrow. (a) For the baseline method, the explanation is the gradient itself. (b) SmoothGrad enhances the explanation by sampling points in the neighborhood (small red dots), and averaging their explanations (bold black arrow). (c) NoiseGrad enhances the explanation by averaging over perturbed models, indicated by multiple decision boundaries (thin green dashed lines). (d) FusionGrad combines SmoothGrad and NoiseGrad by incorporating both stochasticities in the input space and the model space.

mance drop is around 5%. In addition, we approximate the tempered Bayes posterior by multiplicative noise applied to the network weights — in the same spirit as MC dropout (Gal and Ghahramani 2016). Thus, our proposed method NoiseGrad can be implemented as easily as SmoothGrad with an automatic hyperparameter choice and is applicable to any model architecture and explanation method. Our experiments empirically support our hypothesis and show quantitatively and qualitatively that NoiseGrad outperforms SmoothGrad and combining NoiseGrad with SmoothGrad, which we refer to as FusionGrad, further boosts the performance. An overview of our proposed methods is shown on a toy experiment in Figure 1.

Another advantage of NoiseGrad over SmoothGrad is that it is straightforwardly applicable to global explanations as well. For example, we can replace the objective function for activation maximization with its average over the model samples, which is expected to stabilize the image representing the features captured by neurons. Our experiments demonstrate that NoiseGrad improves global explanations in terms of human interpretability and vividness of illustrated abstractions.

Our main contributions include:

- We propose a novel method, *NoiseGrad*, that improves local and global explanation methods by introducing stochasticity to the model parameters.
- The performance gain by NoiseGrad and its fusion with SmoothGrad, *FusionGrad*, for local explanations, is shown qualitatively and quantitatively using different evaluation criteria.
- We observe that NoiseGrad is further capable of enhancing global explanation methods.

Background

Let $f(\cdot;\hat{W}):\mathbb{R}^d \to \mathbb{R}^k$ be a neural network with learned weights $\hat{W} \subset \mathbb{R}^S$ that maps a vector $\boldsymbol{x} \in \mathbb{R}^d$ from the input domain to a vector $y \in \mathbb{R}^k$ in the output domain. In general, attribution methods could be viewed as an operator $E\left(\boldsymbol{x}, f(\cdot,\hat{W})\right)$ that attributes relevances to the features of the input $\boldsymbol{x}$ with respect to the model function $f(\cdot,W)$. More in-depth discussion about the different explanation methods used can be found in the Appendix.

Enhancing local explanations by adding noise to the inputs A recently proposed popular method, called SmoothGrad (SG), seeks to alleviate noise and visual diffusion of saliency maps by introducing stochasticity to the inputs (Smilkov et al. 2017). SmoothGrad adds Gaussian noise to the input and takes the average over N instances of noise:

$$E_{SG}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} E\left(\boldsymbol{x}+\xi_i, f(\cdot,\hat{W})\right), \quad \xi_i \sim \mathcal{N}(\mathbf{0},\sigma_{\text{SG}}^2\mathbf{I}) \tag{1}$$

where $\mathcal{N}(\mu,\boldsymbol{\Sigma})$ is the Normal distribution with mean μ and covariance $\boldsymbol{\Sigma}$ and $\mathbf{I}$ the identity matrix. The authors of the original paper state that SG allows smoothening the gradient landscape, thus providing better explanations. Later SmoothGrad has also proven to be more robust against adversarial attacks (Dombrowski et al. 2019).

Enhancing explanations by approximate Bayesian learning From a statistical perspective, training DNNs with the most commonly used loss functions and regularizers, such as categorical cross-entropy for classification and MSE for regression, can be seen as performing *maximum a-posteriori (MAP) learning*. Hence, the resulting weights can be thought of as a point estimate for a posterior mode in the parameter space, capturing no uncertainty information. Recently, Bykov et al. (2021) showed that incorporating information about posterior distribution can enhance local explanations for DNNs. Intuitively, in contrast to the MAP learning, where point estimates of weights represent one deterministic decision-making strategy, a posterior distribution represents an infinite ensemble of models, which employ different strategies towards the prediction. By aggregating the variability of the decision-making processes of networks, we can obtain a broader outlook on the features that were used for the prediction, and thus deeper insights into the models' behavior (Grinwald et al. 2022).

Since exact Bayesian Learning is intractable for DNNs, a plethora of approximation methods have been proposed, e.g., Laplace Approximation (Ritter, Botev, and Barber 2018), Variational Inference (Graves 2011; Osawa et al. 2019), MC dropout (Gal and Ghahramani 2016), Variational Dropout (Kingma, Salimans, and Welling 2015; Molchanov, Ashukha, and Vetrov 2017), MCMC sampling (Wenzel et al. 2020). Since most of the approximation methods require full retraining of the network or evaluation of the second-order statistics, which are computationally expensive, we use a cruder approximation with multiplicative noise to draw model samples for NoiseGrad.

Method

As mentioned previously, the mechanism why SmoothGrad improves explanations has not been well understood. In empirical experiments we found that SmoothGrad with a recommended 10%–20% noise level is large enough to cross the decision boundary, resulting in a significant classification accuracy drop. This finding implies that SmoothGrad does not only smooth the peaky derivative but also collects signals from the steepest part of the likelihood, i.e., decision boundary, by perturbing the input sample with large noise.

Motivated by this observation, we propose another way of introducing stochasticity – instead of perturbing the input, we perturb the model itself. More specifically, we propose a new method — NoiseGrad — which draws network weight samples from a *tempered* Bayes posterior (Wenzel et al. 2020), i.e., the Bayes posterior with a temperature higher than 1. The temperature should be so high that the decision boundaries of some model samples are close to the test sample, which reinforces the signals for explanations.

Local explanation with NoiseGrad Mathematically, we define the local explanation with NoiseGrad (NG) as follows

$$E_{\mathrm{NG}}(\boldsymbol{x}) = \tfrac{1}{M}\textstyle\sum_{i=1}^{M} E\left(\boldsymbol{x}, f(\cdot, \mathcal{W}_i)\right), \qquad (2)$$

where $\{\mathcal{W}_i\}, i \in [1, M]$ are samples drawn from a tempered Bayes posterior. Since approximate Bayesian learning is computationally expensive, we approximate the posterior with multiplicative Gaussian noise – in the spirit of MC dropout (Gal and Ghahramani 2016): $\mathcal{W}_i = \hat{W} \cdot \eta_i$, with $\eta_i \sim \mathcal{N}(\mathbf{1}, \sigma^2_{\mathrm{NG}}\mathbf{I})$, where $\mathbf{I}$ refers to an identity matrix. By averaging over a sufficiently large number of samples M, we expect NG to smooth the signal and also to collect amplified signals from models whose decision boundary is close to the test sample. This and the smoothing capabilities of the NG method can be observed in Figure 2, where the gradients are plotted for each grid-point on the toy dataset used before in Figure 1 for both Baseline and NG. From the results shown, we can observe that NoiseGrad in fact smoothens out the gradient.

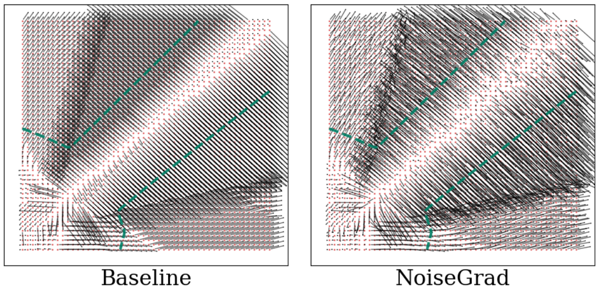


Figure 2: Illustration of the impact of NoiseGrad on a gradient flow map. For the same problem as in Figure 1, for each grid-point the gradient is computed (left). On the right we can observe the effect of NoiseGrad — it smoothens the gradients by perturbing the decision boundary.

FusionGrad We also propose FusionGrad (FG), a combination of NoiseGrad and SmoothGrad, to incorporate both stochasticities in the input space and the model space

$$E_{\mathrm{FG}}(\boldsymbol{x}) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \tfrac{1}{M}\sum_{j=1}^{M} E\left(\boldsymbol{x} + \xi_j, f(\cdot, \mathcal{W}_i)\right) \qquad (3)$$

where $\xi_j \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathrm{SG}}\mathbf{I})$, N the number of noisy inputs, and M the number of model samples. We show in our experiments that FG further boosts the performance of NG, providing the best qualitative and quantitative performances.

Global explanation with NoiseGrad Unlike SmoothGrad, NoiseGrad can be used to enhance global explanation methods. An important class of global explanation methods is activation maximization (AM) (Erhan et al. 2009), which synthetically creates an input that maximizes a given function $g(x)$. Usually, the activation of a particular neuron is maximized, and thus the generated input can embody the main concepts and abstractions that the DNN is looking for. Many variants with different types of regularization have emerged (Nguyen et al. 2016; Olah, Mordvintsev, and Schubert 2017) — regularization is necessary because otherwise, AM might synthesize adversarial inputs, which would not convey visual information to the investigator.

NoiseGrad enhancement of any AM technique could be performed as follows: given the target function $g(x, \hat{W})$ for a model $f(\cdot; \hat{W})$, we sample M models with the NG procedure, and maximize the average function over the number of perturbed models:

$$\arg\max_{x \in \mathcal{C}} \tfrac{1}{M}\textstyle\sum_{i=1}^{M} g(x, \mathcal{W}_i), \qquad (4)$$

where $\mathcal{C}$ is a regularized input domain, specific to a particular implementation of an AM method.

Heuristic for hyperparameter setting One of the reasons for the popularity of SmoothGrad is that it does not require hyperparameter tuning: it works well if the number N of noisy samples is sufficiently large, and the input noise level

σ_{SG} is set to a value in the recommended range, 10%–20%, compared to the signal level.

A major question is if one can set the noise level σ_{NG} for NoiseGrad in a similar way that does not require finetuning. We put forward a simple hypothesis: since we need signals from models whose decision boundaries are close to the test sample, we might choose the noise level σ_{NG} such that we observe a certain performance drop. For the classification setting with balanced classes, we can use accuracy as a performance measure: from experimental results (discussed more in-depth in the Appendix) we recommend to set the relative accuracy drop $\mathrm{AD}(\sigma) = 1 - (\mathrm{ACC}(\sigma) - \mathrm{ACC}(\infty))/(\mathrm{ACC}(0) - \mathrm{ACC}(\infty))$ to around 5%, where $\mathrm{ACC}(\sigma)$ denotes the classification accuracy at the noise level σ. Note that $\mathrm{ACC}(0)$ and $\mathrm{ACC}(\infty)$ correspond to the original accuracy and the chance level, respectively. This rule of thumb can be used for various model architectures with different scales, as shown in the next section.

As a heuristic for FusionGrad, we recommend to half both σ_{SG} and σ_{NG} (as found by their respective heuristics) to equal the contribution from the input perturbation and the model perturbation. Further, we empirically found that 10 samples are sufficient for both methods. With those heuristics, NoiseGrad and FusionGrad can be used as effortlessly as SmoothGrad. A detailed discussion on the relationship between the explanation quality (localization criteria) and accuracy drop can be found in the Appendix.

Experiments On Local Explanations

In this section, we explain datasets and evaluation metrics used for evaluating our proposed methods for local *attribution quality*.

Datasets To measure the goodness of an explanation, one typically needs to resort to proxies for evaluation since no ground-truth for explanations exists. Similar to Arras, Osman, and Samek (2020) and Yang and Kim (2019); Romijnders (2017), we therefore design a controlled setting for which the ground-truth segmentation labels are simulated. For this purpose, we construct a semi-natural dataset CMNIST (customized-MNIST), where each MNIST digit (LeCun, Cortes, and Burges 2010) is displayed on a randomly selected CIFAR background (Krizhevsky, Hinton et al. 2009). To ensure that the explainable evidence for a class lies in the vicinity of the object itself, rather than in its contextual surrounding, we uniformly distribute CIFAR backgrounds for each MNIST digit class as we construct the CMNIST dataset. Ground-truth segmentation labels for the explanations are formed by creating different variations of segmentation masks around the object of interest such as a squared box around the object or the pixels of the object itself. Moreover, to understand the real impact of SOTA, we use the PASCAL VOC 2012 object recognition dataset (Everingham et al. 2010) and ILSVRC-15 dataset (Russakovsky et al. 2015) for evaluation, where object segmentation masks in the forms of bounding boxes are available. Further details on training- and test splits, preprocessing steps and other relevant dataset statistics can be found in the Appendix.

In an explainability context, the question naturally arises whether object localization masks can be used as ground-truth labels for explanations of natural datasets in which the independence of the models from the background cannot be guaranteed. We, therefore, report quantitative metrics only on the controlled semi-natural dataset but report qualitative results on the natural dataset as well.

Evaluation metrics While the debate of what properties an attribution-based explanation ought to fulfill continues, several works (Montavon, Samek, and Müller 2018; Alvarez Melis and Jaakkola 2018; Carvalho, Pereira, and Cardoso 2019) suggest that in order to produce human-meaningful explanations one metric alone is not sufficient. To broaden the view of what it means to provide a good explanation, we evaluate the explanation-enhancing methods using four well-studied properties — *localization* (Zhang et al. 2018; Kohlbrenner et al. 2020; Theiner, Müller-Budack, and Ewerth 2021; Arras, Osman, and Samek 2020), *faithfulness* (Bach et al. 2015; Samek et al. 2016; Bhatt, Weller, and Moura 2020; Nguyen and Martínez 2020; Rieger and Hansen 2020), *robustness* (Alvarez Melis and Jaakkola 2018; Montavon, Samek, and Müller 2018; Yeh et al. 2019), and *sparseness* (Nguyen and Martínez 2020; Chalasani et al. 2020; Bhatt, Weller, and Moura 2020). While there exists several empirical interpretations, or operationalizations, for each of these qualities, we selected one metric per category. We adopted *Relevance Rank Accuracy* (Arras, Osman, and Samek 2020) to express localization of the attributions, *Faithfulness correlation* (Bhatt, Weller, and Moura 2020) to capture attribution faithfulness, applied *max-Sensitivity* (Yeh et al. 2019) to express attribution robustness and *Gini index* (Chalasani et al. 2020) to assess the sparsity of the attributions. All evaluation measures are clearly motivated, defined, and discussed in the Appendix.

Explanation methods NoiseGrad is *method-agnostic*, which means, that it can be applied in conjunction with *any* explanation method. However, in these experiments, we focus on a popular category of post-hoc gradient-based attribution methods and use *Saliency* (SA) (Mørch et al. 1995) as the base explanation method in the experiments similar as in (Smilkov et al. 2017). Since the majority of model-aware local explanation methods make use of the model gradients, we argue that a potential explanation improvement on SA with our proposed method may also be transferred to an improvement of a related gradient-based explanation method. As the comparative baseline explanation method (Baseline), we employ the Saliency explanation, which adds no noise to either the weights nor the input. We report results on additional explanation methods in the Appendix.

Model architectures Explanations were produced for networks of different architectural compositions such as ResNet (He et al. 2016), VGG (Simonyan and Zisserman 2014) and LeNet (LeCun et al. 1998). All networks were trained for image classification tasks so that they showcased a comparable test accuracy to a minimum of 86%, 92% and 86% classification accuracy for CMNIST, PASCAL VOC 2012, and ILSVRC-15 datasets respectively. For more details on the model architectures, optimization configurations,

Method	Localization (↑)	Faithfulness (↑)	Robustness (↓)	Sparseness (↑)
Baseline	0.7315 ± 0.0505	0.3413 ± 0.1549	0.0763 ± 0.0265	$\mathbf{0.6272 \pm 0.0475}$
SG	0.8263 ± 0.0483	0.3465 ± 0.1601	0.0590 ± 0.0235	0.5310 ± 0.0635
NG	0.8349 ± 0.0367	$\mathbf{0.3635 \pm 0.1536}$	0.0224 ± 0.0080	0.5794 ± 0.0533
FG	$\mathbf{0.8435 \pm 0.0358}$	$\mathbf{0.3697 \pm 0.1465}$	$\mathbf{0.0153 \pm 0.0058}$	0.5721 ± 0.0532

Table 1: Comparison of attribution quality where the noise levels are set by the heuristic. ↑ and ↓ indicate the larger is the better and the smaller is the better, respectively. The values of the best method and the methods that are not significantly outperformed by the best method, according to the Wilcoxon signed-rank test for $p = 0.05$, are bold-faced.

Method	LeNet	VGG11	VGG16	RN9	RN18	RN50
Baseline	0.922 ± 0.033	0.961 ± 0.017	0.967 ± 0.015	0.926 ± 0.026	0.913 ± 0.04	0.912 ± 0.035
SG	0.940 ± 0.048	0.971 ± 0.025	0.963 ± 0.034	0.975 ± 0.015	0.962 ± 0.026	0.967 ± 0.022
NG	0.949 ± 0.029	0.982 ± 0.011	$\mathbf{0.985 \pm 0.011}$	0.978 ± 0.012	0.963 ± 0.023	$\mathbf{0.973 \pm 0.017}$
FG	$\mathbf{0.961 \pm 0.025}$	$\mathbf{0.984 \pm 0.011}$	0.982 ± 0.013	$\mathbf{0.982 \pm 0.011}$	$\mathbf{0.975 \pm 0.016}$	0.969 ± 0.019

Table 2: Attribution ranking scores (AUC) for different architectures with noise levels set by our proposed heuristic. We can observe that either NG or FG outperforms Baseline and SG. For the sake of space, we refer ResNets as RNs.

and training results, we refer to the Appendix.

Results

In the following, we present our experimental results. The findings can be summarized as follows: (i) both NG and FG offer an advantage over SG measured with several metrics of attribution quality and (ii) as a heuristic, choosing the hyperparameters for NG and FG according to a classification performance drop of 5% typically result in explanations with a high attribution quality.

Quantitative evaluation We start by examining the performance of the methods considering the four aforementioned attribution quality criteria applied to the absolute values of their respective explanations. The results are summarized in Table 1, where the methods (Baseline, SG, NG, FG) are stated in the first column and the respective values for localization, faithfulness robustness, and sparseness in columns 2-5. The scores were computed and averaged over 256 randomly chosen test samples from the CMNIST dataset, using a ResNet-9 classifier and the Saliency as the base attribution method. The Quantus library was employed for XAI evaluation[1] (Hedström et al. 2022).

The noise level for SG, NG, and FG is set by the heuristic, which was described in the method section. We conducted the same experiment with additional base attribution methods and datasets and found similar tendencies, which are reported in the Appendix. In the Appendix, we also investigated how the different noise levels for FusionGrad (σ_{NG} and σ_{SG}) influence the ranking of attributions, as well as performed model parameter randomization sanity checks (Adebayo et al. 2018).

From Table 1, we can observe a significant attribution quality boost by our proposed methods, NG, and FG in comparison to the baselines, Baseline and SG. For each of the examined quality criteria, the values range between $[0, 1]$. For localization, faithfulness, and sparseness higher values are better and for robustness lower values are better. The combination of SmooothGrad and NoiseGrad, i.e., FusionGrad is *significantly* better than either method alone. In summary, we conclude that NG outperforms SG on all four criteria and Baseline on the most criteria except Sparseness, and FG further boosts the performance. In general, any perturbation naturally degrades the sparseness, and therefore Baseline gives the best sparseness score. Note that NG and FG both improve the other criteria with less degradation of sparseness in comparison to SG. It is important to also emphasize that evaluation of explanation methods should always be viewed holistically — i.e., while Baseline may be the most sparse explanation, it would *not* be the overall preferred option since it is the least faithful, localized, and robust explanation of them all. Further evaluation results on the ILSVRC-15 dataset can be found in the Appendix.

Heuristic applied to different architectures In Table 2, we present ranking AUC scores for different model architectures trained on CMNIST, using the recommended heuristic to set the noise level.

We can observe that appropriate noise levels are chosen with the proposed heuristic — where NG and FG significantly outperform Baseline and SG.

Qualitative evaluation Figure 3 shows attribution maps for an image from the PASCAL VOC 2012 dataset for Baseline, SG, NG, and FG for two attribution methods, Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017) and GradientSHAP (GradSHAP) (Lundberg and Lee 2017). Compared to Baseline and SmoothGrad, NG and FG demonstrate improved localized attribution with improved vividness. Semantic meaningful features such as the nose and the eyes of the dog are highlighted by NG but not by SG, indicating that our methods can find additional attributional evidence for a class that SG or Baseline explanation does not. Furthermore,

[1] Code can be found at https://github.com/understandable-machine-intelligence-lab/quantus

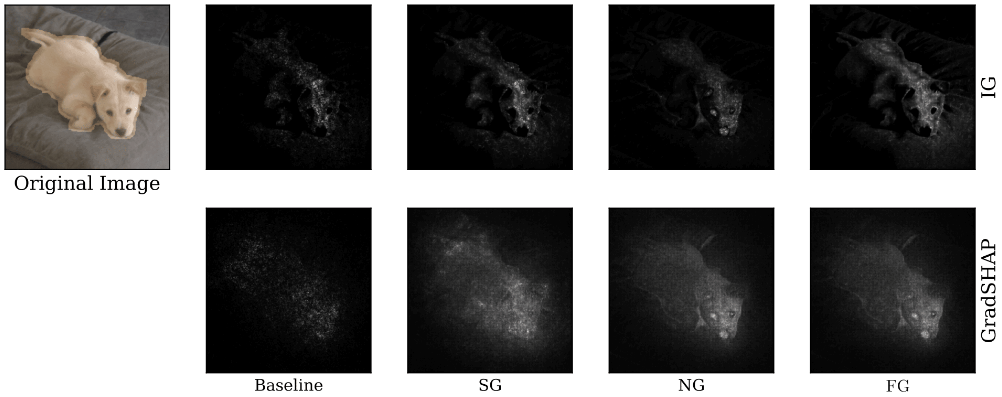


Figure 3: Attribution maps by Baseline, SG, NG, and FG for two base explanation methods Integrated Gradients (IG) and GradientSHAP (GradSHAP), for an image from the PASCAL VOC 2012 dataset. We can observe that both NG and FG improve the sharpness of the attributions compared to Baseline and SG. Moreover, NG highlights semantic features of the dog, such as the nose and the eyes, which are not visible for Baseline and SG.

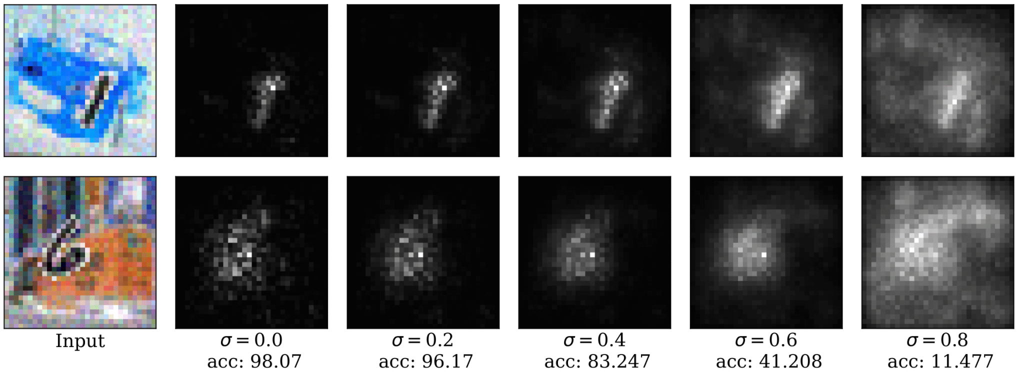


Figure 4: Illustration of NG-enhanced Saliency explanations for the CMNIST dataset: we observe an improvement of the localization ability of the explanation when increasing the hyperparameter σ until $\sigma \leq 0.4$ — afterward, if the noise amplitude becomes too large, the models lose their predicting ability, which results in noisy attribution maps.

as we enumerated several test samples to find representative qualitative characteristics that distinguish the different approaches, we could conclude that attributions of NG and FG are typically more crisp and concise compared to Baseline and SG explanations.

Figure 4 shows the noise level dependence of the NG attribution map with Saliency as the base attribution. Visually, the attribution seems to improve with a noise level between $\sigma_{\text{NG}} \in [0.2, 0.4]$, which is chosen by our heuristic as well. More examples are given in the Appendix.

Experiments On Global Explanations

Finally, we apply NoiseGrad to enhance the global explanations generated by Feature Visualisation [2] (Olah, Mordvintsev, and Schubert 2017). We applied FV to the output neurons for different classes of a ResNet-18 network pre-trained on ImageNet dataset with and without NoiseGrad.

Figure 5 indicates the feature visualization images by the Baseline AM (top row) and by AM using NoiseGrad (bottom row). We can observe that the visualized abstractions with NoiseGrad are more vivid and more human-understandable, implying that the NG-enhanced global explanation can convey improved recognizability of underlying high-level concepts. More examples and experiments can be found in the Appendix.

Conclusion

In this work, we demonstrated that the use of stochasticity in the parameter space of deep neural networks can enhance techniques for eXplainable AI (XAI).

Our proposed NoiseGrad draws samples from the approximated tempered Bayes posterior, such that the decision boundary of some model samples is close to the test sam-

[2] For generating global explanations following library was used https://github.com/Mayukhdeb/torch-dreams

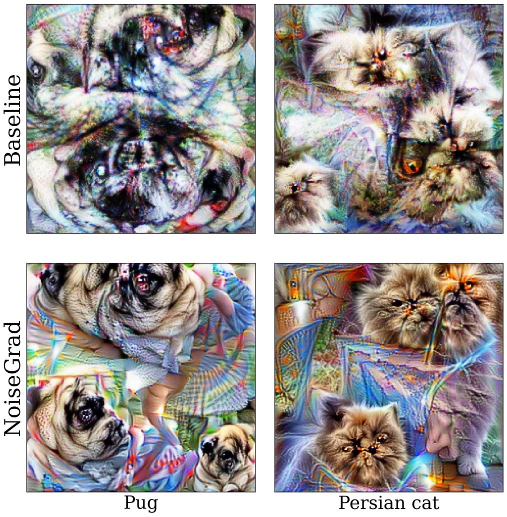


Figure 5: Global explanation by activation maximization (AM) without (top row) and with (bottom) NoiseGrad, applied to ResNet-18 network pre-trained on ImageNet dataset. For each column, the output neuron for the specified class is explained.

ple, effectively amplifying the gradient signals. In our experiments on local explanations, we have shown the advantages of NoiseGrad and its fusion with the existing SmoothGrad method qualitatively and quantitatively on several evaluation criteria. A notable advantage of NoiseGrad over SmoothGrad is that it can also enhance global explanation methods by smoothing the objective for activation maximization (AM), leading to enhanced human-interpretable concepts learned by the model. We believe that our idea of introducing stochasticity in the parameter space facilitates the development of practical and reliable XAI for real-world applications.

Limitations Since the number of parameters in a DNN is usually larger than the number of features in the input data, NoiseGrad is more computationally expensive than SmoothGrad (more discussion in the Appendix).

In addition, explanation evaluation is still an unsolved problem in XAI research and each evaluation technique comes with individual drawbacks. Further research is needed to establish a sufficient set of quantitative evaluation metrics beyond the four criteria used in this paper.

Future work To broaden the applicability of our proposed methods, we are interested to investigate the performance of NG and FG on other tasks than image classification such as time-series prediction or NLP. We also want to further explore how NG and FG explanations change when alternative ways of adding noise to the weights of a neural network are employed e.g., by adding different levels of noise to different layers or individual neurons.

Acknowledgements

This work was partly funded by the German Ministry for Education and Research through the third-party funding project Explaining 4.0 (ref. 01IS20055) and as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref 01IS18037A).

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 9505–9515.

Alvarez Melis, D.; and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Arras, L.; Osman, A.; and Samek, W. 2020. Ground Truth Evaluation of Neural Network Explanations with CLEVR-XAI.

Bach, S.; Binder, A.; on, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7).

Bhatt, U.; Weller, A.; and Moura, J. M. 2020. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*.

Bykov, K.; Höhne, M. M.-C.; Creosteanu, A.; Müller, K.-R.; Klauschen, F.; Nakajima, S.; and Kloft, M. 2021. Explaining Bayesian Neural Networks. *arXiv preprint arXiv:2108.10346*.

Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8): 832.

Chalasani, P.; Chen, J.; Chowdhury, A. R.; Wu, X.; and Jha, S. 2020. Concise Explanations of Neural Networks using Adversarial Training. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1383–1391. PMLR.

Dombrowski, A.-K.; Alber, M.; Anders, C.; Ackermann, M.; Müller, K.-R.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, 13567–13578.

Erhan, D.; Bengio, Y.; Courville, A.; and Vincent, P. 2009. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3): 1.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of ICML*.

Graves, A. 2011. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24.

Grinwald, D.; Bykov, K.; Nakajima, S.; and Höhne, M. M.-C. 2022. Visualizing the diversity of representations learned by Bayesian neural networks. *arXiv preprint arXiv:2201.10859*.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5): 1–42.

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

Hedström, A.; Weber, L.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; and Höhne, M. M. C. 2022. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations.

Kingma, D. P.; Salimans, T.; and Welling, M. 2015. Variational Dropout and the Local Reparameterization Trick. In *Advances in NIPS*.

Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; and Lapuschkin, S. 2020. Towards Best Practice in Explaining Neural Network Decisions with LRP. arXiv:1910.09840.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Molchanov, D.; Ashukha, A.; and Vetrov, D. 2017. Variational Dropout Sparsifies Deep Neural Networks. In *Proceedings of ICML*.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73: 1–15.

Mørch, N. J. S.; Kjems, U.; Hansen, L. K.; Svarer, C.; Law, I.; Lautrup, B.; Strother, S. C.; and Rehm, K. 1995. Visualization of neural networks using saliency maps. In *Proceedings of International Conference on Neural Networks (ICNN'95), Perth, WA, Australia, November 27 - December 1, 1995*, 2085–2090. IEEE.

Mordvintsev, A.; Olah, C.; and Tyka, M. 2015. Inceptionism: Going Deeper into Neural Networks. https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Accessed: 2022-04-25.

Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; and Clune, J. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*, 3387–3395.

Nguyen, A.; and Martínez, M. R. 2020. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*.

Olah, C.; Mordvintsev, A.; and Schubert, L. 2017. Feature visualization. *Distill*, 2(11): e7.

Osawa, K.; Swaroop, S.; Jain, A.; Eschenhagen, R.; Turner, R. E.; Yokota, R.; and Khan, M. E. 2019. Practical Deep Learning with Bayesian Principles. In *Advances in NeurIPS*.

Rieger, L.; and Hansen, L. K. 2020. A simple defense against adversarial attacks on heatmap explanations. *arXiv preprint arXiv:2007.06381*.

Ritter, H.; Botev, A.; and Barber, D. 2018. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning.

Romijnders, R. 2017. Simple Semantic Segmentation. https://github.com/RobRomijnders/segm. Accessed: 2022-04-25.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Samek, W.; Binder, A.; on, G.; Lapuschkin, S.; and Müller, K.-R. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11): 2660–2673.

Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; and Müller, K.-R. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3): 247–278.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2): 336–359.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328. PMLR.

Theiner, J.; Müller-Budack, E.; and Ewerth, R. 2021. Interpretable Semantic Photo Geolocalization. arXiv:2104.14995.

Vidovic, M. M.-C.; Görnitz, N.; Müller, K.-R.; and Kloft, M. 2016. Feature importance measure for non-linear learning algorithms. *arXiv preprint arXiv:1611.07567*.

Wenzel, F.; Roth, K.; Veeling, B. S.; Swiatkowski, J.; Tran, L.; Mandt, S.; Snoek, J.; Salimans, T.; Jenatton, R.; and Nowozin, S. 2020. How Good is the Bayes Posterior in Deep Neural Networks Really? *arXiv:2002.02405*.

Yang, M.; and Kim, B. 2019. Benchmarking Attribution Methods with Relative Feature Importance. *CoRR*, abs/1907.09701.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A. S.; Inouye, D. I.; and Ravikumar, P. 2019. On the (in) fidelity and sensitivity for explanations. *arXiv preprint arXiv:1901.09392*.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zhang, J.; Bargal, S. A.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10): 1084–1102.