

# Uncertainty Quantification for Gradient-based Explanations in Neural Networks

Mihir Mulye

Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Germany  
 Heinrich-Heine University, Duesseldorf, Germany

mihir.mulye@hhu.de

Matias Valdenegro-Toro

Department of Artificial Intelligence  
 University of Groningen, Netherlands

m.a.valdenegro.toro@rug.nl

## Abstract

*Explanation methods help understand the reasons for a model's prediction. These methods are increasingly involved in model debugging, performance optimization, and gaining insights into the workings of a model. With such critical applications of these methods, it is imperative to measure the uncertainty associated with the explanations generated by these methods. In this paper, we propose a pipeline to ascertain the explanation uncertainty of neural networks by combining uncertainty estimation methods and explanation methods. We use this pipeline to produce explanation distributions for the CIFAR-10, FER+, and California Housing datasets. By computing the coefficient of variation of these distributions, we evaluate the confidence in the explanation and determine that the explanations generated using Guided Backpropagation have low uncertainty associated with them. Additionally, we compute modified pixel insertion/deletion metrics to evaluate the quality of the generated explanations.*

## 1. Introduction

With the increase in the application of neural networks in safety-critical cases such as autonomous driving, medical diagnosis, and quality control procedures, the number of methods developed to explain the network output has also increased. These explanation methods are now being coupled with the knowledge of human experts to understand the workings, debug and optimize the performance, and sometimes even extend the scope of application of these neural networks. As explanation methods gradually become indispensable, it is also critical to analyze the uncertainty in the output of these explanation methods themselves (see Figure 1) as it would help in justifying the trust being placed in these explanations.

In this paper, we attempt to solve this problem and investigate the possibility of combining the concepts of uncertainty estimation with the explainability of neural networks to quantify the confidence in the importance attributed to

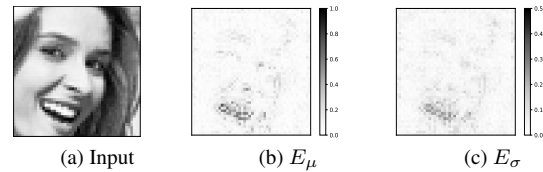


Figure 1. Visualization of explanation with uncertainty on an image from FER+ dataset using Guided Backpropagation as explanation method and Deep Ensembles as uncertainty estimation method. (a) Input Image, (b) Explanation Mean, and (c) Explanation Uncertainty.

the input features. To this end, we design and implement a pipeline that induces a distribution in the explanation method output of a neural network by utilizing the prediction distribution of the said neural network trained using uncertainty estimation methods. We verify the robustness of the aforementioned pipeline by testing it on two learning tasks namely, image classification and tabular regression. We select the CIFAR-10 [5] and FER+ [2] datasets for testing the pipeline on the task of image classification and choose the California Housing Dataset [9] for testing on the task of tabular regression.

By implementing our proposed approach, we attempt to answer the following research questions: RQ1: Could uncertainty estimation methods and explanation methods be combined to ascertain the explanation uncertainty of neural networks? RQ2: How can the explanation uncertainty be combined into a single representation of explanation uncertainty? RQ3: What are the possible approaches to evaluate explanation uncertainty? RQ4: Which combination of uncertainty estimation method and explanation method is the best to analyze the explanation uncertainty output for a neural network?

The contributions of this paper are: we propose and implement a pipeline that helps ascertain the uncertainty in the explanation method output of a neural network. To represent this, we compute the mean, the standard deviation, and the coefficient of variation of the generated explanation distribution. We propose the coefficient of variation as

a way to merge the explanation with its uncertainty into a single saliency explanation. Additionally, we modify the pixel insertion/deletion metric (originally defined for a single explanation instance) such that it can be used to quantify the quality of explanation distributions. Finally, we demonstrate the robustness of our proposed pipeline by testing it on two tasks namely, image classification (CIFAR-10 and FER+) and tabular regression (California Housing dataset).

## 2. Related Work

### 2.1. Uncertainty Estimation Methods

Classical neural networks produce outputs using fixed weights and parameters. The outputs of these networks are deterministic, in that, for multiple forward passes for a given input, the prediction of the network remains unchanged. However, several approaches have been proposed to train neural networks capable of generating outputs with uncertainty. These methods are chiefly divided into two categories: (i) Bayesian methods and their approximations and (ii) Non-Bayesian approaches.

Bayesian methods treat the weights of the neural network as probability distributions as opposed to singular values. However, as implementing the backpropagation algorithm over a parametric distribution is computationally intensive, approximations of Bayesian methods are used more extensively. Monte Carlo (MC) Dropout [4][14] belongs to this category. This method enables Dropout at inference time as well thereby inducing stochasticity in the model output. Another approach in this category is MC DropConnect [7][17] which is similar to MC-Dropout. However, in this case, weights are set to zero instead of activations during inference to produce stochasticity. Flipout [18] are also categorized as Bayesian Neural Networks. The weights are approximated with a Gaussian distribution  $q_\theta(W)$  and the loss that approximates Evidence Lower Bound (ELBO) is defined as:

$$L(\theta) = \text{KL}(\mathbb{P}(W), q_\theta(W)) - \mathbb{E}_{w \sim q_\theta(W)} [\log \mathbb{P}(y | x, w)] \quad (1)$$

where  $\text{KL}(\mathbb{P}(W), q_\theta(W))$  is the Kullback-Leibler divergence between the prior  $\mathbb{P}(W)$  and the approximate posterior weight distribution denoted by  $q_\theta(W)$ .  $-\log \mathbb{P}(y | x, w)$  is the negative log-likelihood that is calculated as an expectation by sampling weights from the  $q_\theta(W)$ . The negative log-likelihood takes the definition of mean squared error for regression and cross-entropy for classification.

Non-Bayesian methods, on the other hand, treat the weights of the network as deterministic. The uncertainty is introduced by training multiple copies of the same model architecture and using these trained model instances to predict an output distribution for a given input. A key aspect of this training paradigm is that each model instance is initial-

ized randomly which results in differently learned weights and subsequently a different output for the same given input. This output distribution can then be analyzed as output with uncertainty by computing its mean and standard deviation. Deep Ensembles [6] belong to this category.

As the models trained using uncertainty estimation generate multiple samples for the same given input, these samples need to be combined to estimate the model uncertainty. For the task of classification, the mean of the predicted probability vector is computed, and for the regression task, the mean ( $\mu(x)$ ) and the variance (and by extension standard deviation) ( $\sigma^2(x)$ ) for the samples is computed as:

$$\mu(x) = T^{-1} \sum f(x), \quad \sigma^2(x) = T^{-1} \sum (f(x) - \mu(x))^2 \quad (2)$$

where  $T$  is the number of samples (or models in the case of Ensembles). The uncertainty is then defined by the standard deviation  $\sigma(x)$ .

### 2.2. Saliency Explanation Methods

Explanation methods help in understanding why the neural network produces a particular output for a given input. In this section, we discuss several such approaches.

Class Activation Maps [21] uses the global average pooling layer in CNNs to produce heatmaps that highlight pixels relevant to the model for generating an output. A drawback of this approach is that it can only be applied to models that have a global average pooling layer in their architecture. Layer-wise Relevance Propagation [1] computes the relevance of input features by propagating the network prediction back to the input according to predefined local relevance propagation rules [8].

Integrated Gradients (IG)[15] produces explanations by integrating the gradient of output with respect to the input features along a linear path from a baseline image (usually a blank image with black pixels) to the input image. This is mathematically represented as:

$$\text{IG}_i(x, F) = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (3)$$

where  $x$  is the input,  $i$  is the feature dimension,  $x'$  is the baseline input, and  $\alpha$  is a factor of interpolation that determines the amount of feature perturbation, and  $F()$  is the function mapping the input to the output. Computing the integral in the Equation 3 is difficult therefore the solution is obtained using numerical approximation as:

$$\text{IG}_i(x, F) = (x_i - x'_i) \times m^{-1} \sum_{k=1}^m \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \quad (4)$$

where  $m$  denotes the number of steps in the numerical approximation of the integral.

Guided Backpropagation (GBP)[13] computes explanations by multiplying the gradients computed during backpropagation by the sign of corresponding activations from the forward pass. This amplifies the gradients corresponding to positive activation and suppresses gradients corresponding to negative activations. The mathematical representation is presented as:

$$f_i^{l+1} = \text{relu}(f_i^l) = \max(f_i^l, 0) \quad (5)$$

$$R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1} \quad (6)$$

where  $R$  is the reconstructed image obtained at layer  $l$ ,  $f_i^l$  is the positive forward pass activations and  $R_i^{l+1}$  stands for positive error signals.

Local Interpretable Model-agnostic Explanations (LIME) [11] proposes substituting a complex learning model with a simple surrogate model in a local neighborhood. This simplified model is easier to explain in that particular neighborhood. LIME involves sampling around the input to be explained, and then fitting a linear surrogate model that is explainable in that neighborhood. This approach can be applied to a variety of input modalities such as text, tabular data, and images.

### 2.3. Explanation Uncertainty in Neural Networks

With advances in both uncertainty estimation methods and explainability, researchers have attempted to combine these concepts and produce uncertainty in the explanation method output of neural networks. Slack et al. [12] utilize a Bayesian framework to analyze explanation uncertainty in LIME and KernelSHAP. Zhang et al. [20] aim to find sources of uncertainty in LIME by investigating randomness in sampling, and variation of explanation in differing proximities to data points.

Bykov et al. [3] combine Bayesian Neural Networks and Layer-wise Relevance Propagation to generate explanation uncertainty. Wickstrom et al [19] test the possibility of combining MC-Dropout with GBP to generate uncertainty in the explanation.

This work tests additional combinations of uncertainty estimation approaches and explanation methods systematically to produce explanation uncertainty. In the next sections, we discuss this in detail.

## 3. Explanation Uncertainty Pipeline

In this section, we discuss the proposed approach to compute the explanation uncertainty. To this end, we introduce the notation and the relevant underlying concepts. We also discuss the implementation of the pipeline along with providing the rationale for the selection of uncertainty estimation and explanation methods.

### 3.1. Explanation Uncertainty

Assume that a neural network  $f_\theta$  having parameters  $\theta$  is trained using uncertainty estimation methods such as MC-Dropout, Deep Ensemble, and Variational Inference among others. The inference for such a model involves multiple stochastic forward passes (MC-Dropout, Variation Inference), or multiple models (Deep Ensembles). This is mathematically formulated as:

$$\mu(x) = T^{-1} \sum_i f_\theta(x), \quad \sigma^2(x) = T^{-1} \sum_i (f_\theta(x) - \mu(x))^2 \quad (7)$$

where  $T$  denotes the number of stochastic forward passes or constituent ensemble models,  $\mu$  and  $\sigma^2$  denote the mean and the variance of the outcome of stochastic forward passes/ensemble models. The gradient-based saliency explanations can then generally be represented as:

$$E(x) = F\left(\hat{y}, \frac{\partial f_\theta(x)}{\partial x}\right) \quad (8)$$

where  $E(x)$  denotes the gradient-based saliency explanation,  $\hat{y}$  is the network prediction and  $x$  is the input.

To generate an explanation with uncertainty, we combine the previous two mathematical formulations as follows:

$$E_\mu(x) = T^{-1} \sum_i E(x) = T^{-1} \sum_i F\left(\hat{y}, \frac{\partial f_\theta(x)}{\partial x}\right) \quad (9)$$

$$E_\sigma(x) = T^{-1} \sum_i (F\left(\hat{y}, \frac{\partial f_\theta(x)}{\partial x}\right) - E_\mu(x))^2 \quad (10)$$

where  $E_\mu(x)$  denotes the explanation mean and  $E_\sigma(x)$  is the standard deviation of the explanations generated by outputs of multiple stochastic forward passes/models. If the explanations are drastically different, the standard deviation will be high. However, if the explanations are similar, the standard deviation will be low. Additionally, we compute the coefficient of variation, a metric that combines the standard deviation and the mean to concisely represent a distribution. It is defined as:

$$CV(x) = \frac{E_\sigma(x)}{E_\mu(x)} \quad (11)$$

where  $E_\sigma(x)$  and  $E_\mu(x)$  are the standard deviation and mean explanations respectively. The higher the coefficient of variation, the more the dissimilar the individual explanation heatmaps are. This metric can be helpful to quantify the confidence in the explanation of the network output for particular input features. Note that the explanation uncertainty is explicitly concerned with modeling uncertainty in the explanation and is unrelated to uncertainty explanation where the objective is to explain the uncertainty in a given prediction.

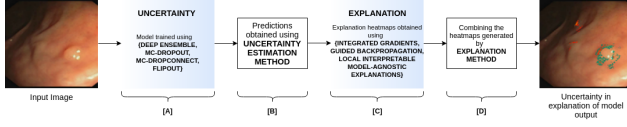


Figure 2. Overview of our proposed approach to explore the underlying uncertainty associated with the explanation of neural network output. The demonstrated input and output images have been taken from [19] for the sake of representation. The labels [A], [B], [C], and [D] depict the different stages of this pipeline and these labels are used as a reference for the discussions in the text.

### 3.2. Pipeline Implementation

Figure 2 provides a visual representation of our proposed pipelines. The various stages in this pipeline are denoted by [A], [B], [C], and [D]. In stage [A], we train a neural network using uncertainty estimation methods. This helps in generating an output distribution as visualized by [B]. Applying the explanation methods (denoted by [C]) to the samples of this output distribution then generates a distribution of explanation heatmaps as highlighted by [D].

It is by analyzing samples of this explanation distribution that we will quantify the confidence in the generated explanation. We compute the mean of the heatmap samples of the explanation distribution. The mean heatmap would be highly activated for those features that predominantly influenced the model outcome. The higher the activation of a particular feature in the mean heatmap, the higher the confidence in that feature impacting the model decision-making process. Next, we compute the standard deviation heatmap of the samples from the explanation distribution. The standard deviation provides insights into the amount of uncertainty in the confidence of the feature’s importance in the model decision-making. A higher standard deviation indicates higher uncertainty associated with the feature’s importance. Finally as discussed previously, we also compute the coefficient of variation of this explanation distribution. The higher the coefficient of variation is, the higher the uncertainty associated with the importance of that feature in the explanation.

### 3.3. Selection of Uncertainty and Explanation Methods

From the literature survey in previous sections, we identify several uncertainty estimation methods and explanation methods to test in this paper. Here, we provide the rationale for selecting these approaches:

**Uncertainty Estimation methods.** We identify a total of 4 uncertainty estimation approaches to test in our pipeline: *Deep Ensemble*, *MC-Dropout*, *MC-DropConnect*, *Flipout*.

**Explanation methods.** We select 3 explanation methods to test our proposed approach: *Guided Backpropagation*,

*Integrated Gradients*, *Local Interpretable Model-agnostic Explanations* (LIME).

These methods are selected by taking into consideration the complexity of their implementation, the amount of compute necessary to generate the output using these methods, and the availability of literature supporting their performance.

## 4. Experimental Setup

### 4.1. Image Classification Experiments

We test the proposed approach for the image classification task using the CIFAR-10 and the FER+ datasets. For this, we use the  $\{Deep\ Ensemble, MC-Dropout, MC-DropConnect, Flipout\}$  as uncertainty estimation methods and  $\{Guided\ Backpropagation, Integrated\ Gradients\}$  as the explanation methods. This provides us with a total of eight combinations of uncertainty estimation and explanation methods that we test. It should be noted that the LIME has not been used along with the image classification task as it is computationally expensive.

The neural network used is a variant of miniVGG discussed in [16]. The modified network has three sequential blocks each consisting of a *convolution* layer followed by *batch normalization* and a *max pooling* layer. The output of these blocks is supplied to a *flatten* layer and subsequently to two *dense* layers. The number of units in the final *dense* layer of this architecture has 10 and 8 units for CIFAR-10 and FER+ respectively as these are the total number of different classes available in these datasets.

After training the neural network with uncertainty estimation methods, we can generate multiple output predictions for a single input. Each of these predictions is a sample from the output distribution. As the next step, we apply the selected explanation methods to these output samples, resulting in an explanation heatmap per output sample. The size of each explanation heatmap is the same as our original input image. We visualize the mean, the standard deviation, and the coefficient of variation of these explanation distribution samples. As these quantities are also computed on a per-pixel basis, the heatmaps for these quantities will also be of the same size as the original input image.

**Pixel Deletion/Insertion.** We focus on quantifying the quality of the generated explanations using our proposed approach. To this end, we identify the *pixel deletion* and the *pixel insertion* metric proposed by [10] in their work. The objective is to compute the changes in the predicted class score by the subsequent addition/removal of pixels to/from an image. We plot this change in the class score against the number of pixels added/deleted and compute the area under this curve (AUC) which is the pixel insertion/deletion metric. The addition/removal of the pixels to/from the input image is governed by the explanation heatmap associated

with the image under consideration.

To compute the pixel deletion metric, we begin with the original input image and record the decrease in the class score as a function of number of pixels deleted from the original image. The order of pixel deletion is based on the importance of the pixel in the explanation heatmap of the input image. The most important pixels are deleted first and the class score is observed. This step is repeated till all the pixels in the input image are deleted. The pixel deletion metric is then defined as the AUC obtained by plotting the class score as a function of % of pixels deleted.

Conversely, to compute the pixel insertion metric, we begin with a noisy image and observe the increment in the class score as image pixels are gradually de-noised from this image based on their importance in the explanation heatmap. The process of adding pixels is similar to the one adopted while calculating the deletion metric. The most important pixels are used to de-noise the image first. This process is repeated till the entire image is de-noised and the corresponding class score is observed. The pixel insertion metric is then the AUC obtained by plotting the class score as a function of the % of pixels inserted.

It should be noted that a lower AUC for the pixel deletion corresponds to a good explanation as the removal of highly relevant pixels should diminish the class score significantly in the initial stages of the pixel removal process. Conversely, a higher AUC for the pixel insertion translates to a good explanation as the addition of highly important pixels early in the pixel insertion process should boost the class scores.

In this work, we modify the originally proposed pixel deletion/insertion metrics to suit our requirements. As opposed to computing the pixel deletion and pixel insertion directly on individual explanation heatmaps, we propose to compute these metrics on the mean and the standard deviation heatmap representation of the explanation distribution. This helps us to quantify the quality of these distributions. The insertion and removal of the pixels are governed by the importance of the pixel in the mean and the standard deviation heatmaps. Highly weighted pixels in the mean and the standard deviation heatmap are added/removed first. Additionally, we compute these metrics on batches of images belonging to the same class rather than computing on a per-image level. This provides a measure of the quality of explanations on a per-class basis.

**Prediction vs Ground Truth Neurons.** We propose modifying the computation of explanation uncertainty by varying the gradient computation step in our original approach. This modification is only tested with the FER+ dataset and is explained with the help of Figure 3.

In our approach, the explanation is generated by calculating the gradient of the network output with respect to the

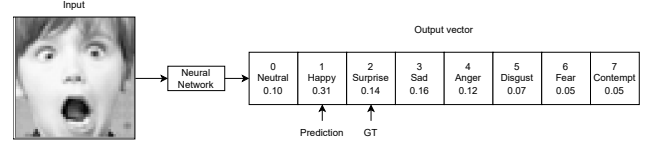


Figure 3. Illustration describing an alternative approach to compute gradients and generate explanations. The original approach generates explanations by computing the gradient of the network prediction logit with respect to the input image whereas, in the modified variant, we compute the gradient of the ground truth logit with respect to the input image.

input image. Mathematically, this is formulated as:

$$E(x) = \frac{\partial(\text{Output}_{\text{Pred}})}{\partial(\text{Input Image})}, E'(x) = \frac{\partial(\text{Output}_{\text{GT}})}{\partial(\text{Input Image})} \quad (12)$$

where  $\text{Output}_{\text{Pred}}$  is the same as network prediction from Figure 3. The explanation generated from this calculation helps us understand *why* the network predicted what it predicted.

However, now we modify the gradient computation step as can be seen in the Equation 12, where  $\text{Output}_{\text{GT}}$  is the activation of the unit corresponding to the ground truth (GT) as depicted in Figure 3. The explanation ( $E'(x)$ ) generated using this calculation helps to visualize the features/image regions that the network *should have* focused on for its prediction to match the ground truth. It is important to note that formulations listed in Equation 12 converge when the network prediction matches the ground truth for a given input.

In the case of the FER+ dataset, we observed that the network prediction often does not match the ground truth owing to the inherent nature of the dataset. Hence, we propose modifying the approach to compute the explanations. The process of generating the mean, the standard deviation heatmap, and the coefficient of variation heatmap is identical to our original approach.

## 4.2. Regression Experiments

We test our approach with tabular regression as it allows us to test its robustness. For this, we use the  $\{\text{Deep Ensemble, MC-Dropout, MC-DropConnect, Flipout}\}$  as uncertainty estimation methods and  $\{\text{Guided Backpropagation, Local Interpretable Model-agnostic Explanation}\}$  as the explanation methods. We select the LIME method as it can be applied to tabular datasets. Additionally, to the best of our knowledge, the GBP method has not been used to generate explanations for tabular data. We use a Multi-Layer Perceptron with 4 *dense* layers for this task.

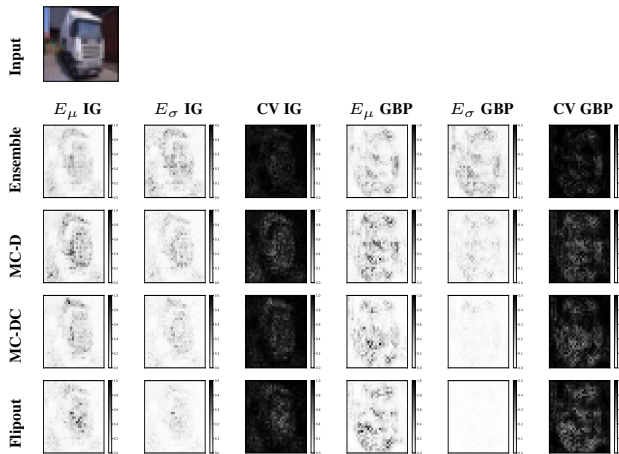


Figure 4. Visualization of the mean ( $E_\mu$ ), the standard deviation ( $E_\sigma$ ), and the coefficient of variation ( $CV$ ) heatmaps generated using the combinations of uncertainty estimation methods (Deep Ensemble, MC-Dropout, MC-DropConnect and Flipout) and explanation methods (Guided Backpropagation and Integrated Gradients). The uncertainty estimation methods are organized in individual rows and the explanation methods are aligned along the columns with the columns 1-3 representing the  $\mu$ ,  $\sigma$ , and the  $CV$  heatmaps for IG and the columns 4-6 representing the same for GBP.

## 5. Experimental Results

### 5.1. Image Classification Results

The explanation heatmaps for an image from the CIFAR-10 dataset are visualized in Figure 4. As previously described, the size of these heatmaps is the same as the size of the input image. In the subsequent discussions, the heatmaps will be referred to by their identifier namely the mean explanation heatmap as  $\mu$ , the standard deviation explanation heatmap as  $\sigma$ , and the coefficient of variation explanation heatmap as  $CV$ .

Figure 4 depicts the image of a *truck* taken from the CIFAR-10. We observe from the  $\mu$  heatmaps that the explanations generated by GBP are highly activated for relevant pixel regions as compared to those by IG. We also observe low activation in the  $\sigma$  heatmaps generated by the GBP as compared to IG. This implies that the explanations generated by the GBP have lower uncertainty as compared to those generated by IG.

Similar analysis can be conducted for images from the FER+ dataset. The explanation distribution obtained for an image taken from FER+ are presented in the first four rows of Figure 5. It can be observed in the  $\mu$  heatmaps that the GBP highlights the relevant facial features to a greater extent as compared to IG. This observation is supported by the  $\sigma$  heatmaps of both these methods as well. The  $\sigma$  and the  $CV$  heatmaps generated by the GBP have low activation

around the relevant pixel regions for all the uncertainty estimation methods. This implies that the explanations generated by the GBP have a higher agreement and are therefore highly confident. However, this is not the case for IG where the pixel activations in the  $\mu$  heatmaps are scattered. This indicates a low confidence in the generated explanation.

From these figures, we also observe that IG generates more noisier explanations as compared to GBP. This could be attributed to the accumulation of activations from multiple steps while calculating explanations using IG that result in spurious highlighting of less relevant regions.

**Pixel Deletion/Insertion.** The results for pixel deletion/insertion are shown in Figure 9 (in appendix) and average AUC are shown in Figure 1. The tables containing the AUCs for these plots are provided in the Supplementary Materials. It should be noted that a higher AUC for pixel insertion and a lower AUC for pixel deletion is desired to ensure good-quality explanations. The pixel insertion and deletion curves (see Supplementary Material) have been plotted for batches of images representing all the classes in both datasets (CIFAR-10 and FER+). The order of pixel insertion/deletion is determined using the mean and the standard deviation explanation heatmaps of the corresponding images from these datasets.

It is evident from these results that for CIFAR-10, the best explanations are generally generated for the *frog* class and the worst explanations are generated for the *dog* class. These observations are supported by high AUC for pixel insertion and low AUC for pixel deletion for these classes. Overall across uncertainty methods and datasets, Flipout seems to perform the best, while Ensembles seems to perform the worse, specially in deletion curve for CIFAR10, but the variation between area under insertion/deletion curve is small.

In the case of the FER+, the trends are not generalizable. This can be attributed to the fact that the FER+ is a difficult dataset to train a network on resulting in a relatively poor quality of model predictions and by extension, poor quality of explanations. This is also supported by the anomalous behavior of the pixel deletion plots for the case of Flipout. In these plots, the class score *increases* despite relevant pixels being deleted from the image.

### 5.2. Prediction vs Ground Truth Neurons

The presentation of the results is similar to that in Section 5.1 for the same input image. The results are shown in the last four rows of Figure 5.

We expect the output of both the variants of explanation computation to differ *only* when the network prediction *does not* match the ground truth for the samples obtained from forward passes/ensemble models.

Similar to the observations with original gradient computation, we find that in Figure 5 the explanations generated

|     |                | CIFAR10           |            |                     |            | FER+              |            |                     |            |
|-----|----------------|-------------------|------------|---------------------|------------|-------------------|------------|---------------------|------------|
|     |                | Insert $\uparrow$ |            | Delete $\downarrow$ |            | Insert $\uparrow$ |            | Delete $\downarrow$ |            |
|     |                | $E_\mu$           | $E_\sigma$ | $E_\mu$             | $E_\sigma$ | $E_\mu$           | $E_\sigma$ | $E_\mu$             | $E_\sigma$ |
| GBP | Ensemble       | 0.537             | 0.534      | 0.554               | 0.540      | 0.263             | 0.268      | 0.333               | 0.324      |
| GBP | MC-Dropout     | 0.551             | 0.541      | 0.320               | 0.323      | 0.354             | 0.347      | 0.293               | 0.281      |
| GBP | MC-DropConnect | 0.582             | 0.571      | 0.368               | 0.354      | 0.186             | 0.187      | 0.256               | 0.253      |
| GBP | Flipout        | 0.586             | 0.557      | 0.301               | 0.299      | 0.291             | 0.289      | 0.271               | 0.274      |
| IG  | Ensemble       | 0.547             | 0.542      | 0.606               | 0.571      | 0.304             | 0.304      | 0.364               | 0.361      |
| IG  | MC-Dropout     | 0.582             | 0.571      | 0.368               | 0.345      | 0.374             | 0.378      | 0.311               | 0.301      |
| IG  | MC-DropConnect | 0.585             | 0.562      | 0.310               | 0.310      | 0.271             | 0.266      | 0.239               | 0.244      |
| IG  | Flipout        | 0.598             | 0.589      | 0.334               | 0.337      | 0.328             | 0.341      | 0.308               | 0.323      |

Table 1. Area under the Insertion/Deletion curves, average across all classes, for CIFAR10 and FER+ datasets, computed separately for explanation mean  $E_\mu$  and explanation uncertainty  $E_\sigma$ . The up/down arrows indicate direction of best metric value.

by GBP have higher certainty in comparison to the explanations generated by IG. Additionally, we notice the modifications applied to the gradient computation step do not drastically alter the generated explanation. A slight variation in the activation of different pixel regions can be observed in the case of IG, but it is not as prominent in GBP.

A possible reason for no drastic changes in the explanation could be that the averaging effect of the multitude of explanation samples negates the impact of the modified gradient computation. It might be possible that the majority of the output prediction sample matches the ground truth thereby outweighing the effect of modified gradient computation associated with those incorrect output predictions. In such a case, where the model prediction increasingly matches the ground truth, the results of both the original gradient computation and the variation in gradient computation converge.

### 5.3. Regression Results

As the modality of input data changes for the tabular regression task, we change the depiction of the explanation as well. We visualize feature importance as bar charts for this task. The length of the bars shows the mean explanation and the error bars depict the standard deviation explanation. Figure 6 depicts the output obtained on a data point taken from the California Housing Dataset.

It can be observed that the mean explanations generated by LIME generally have a lower activation as compared to GBP. This indicates less agreement amongst the constituent explanations and in turn higher uncertainty in the explanation generated by LIME.

Furthermore, we find that for LIME, the value of the coefficient of variation is slightly higher for several features, especially in the case of Deep Ensemble and MC-DropConnect. This implies that these explanations are highly uncertain and should be used with caution. This can

be attributed to disagreement among the constituent explanations for that particular feature. The same is observed in the case of GBP for a couple of features, (to a smaller degree as compared to LIME), especially for the case of Deep Ensembles.

## 6. Conclusions and Future Work

In this section, we discuss the conclusions and answer the questions formulated in the previous section of this paper.

RQ1: We have demonstrated that additional combinations of uncertainty estimation and explanation methods can be implemented efficiently to generate explanation uncertainty by using our proposed approach. RQ2: We identified that the mean, the standard deviation, and the coefficient of variation can be used for concise analysis and representation of the resultant explanation distribution. The features having a high mean, low standard deviation, and subsequently a low coefficient of variation are generally more likely to have influenced the model output. RQ3: To analyze the quality of explanation distributions, we proposed and computed a modified version of the pixel insertion/deletion metric. We find that a higher value for pixel insertion and a lower value for pixel deletion are the characteristics of a good explanation. RQ4: From the analysis of the explanation representations, we deduce that any combination of uncertainty estimation with GBP (as an explanation method) can produce useful and highly confident explanation distributions. This can be supplemented by the observation that GBP tends to generate less noisy explanation heatmaps with higher mean and lower standard deviation for relevant features.

We expect that explanation uncertainty becomes a standard feature of saliency explanations as used by researchers and practitioners, same as it should be for uncertainty estimation methods, so a machine learning model can produce a prediction, with an uncertainty estimate, and an explanation, also with its uncertainty estimate, for a complete set of



Explanation for Predicted Class

Explanation for Ground Truth Class

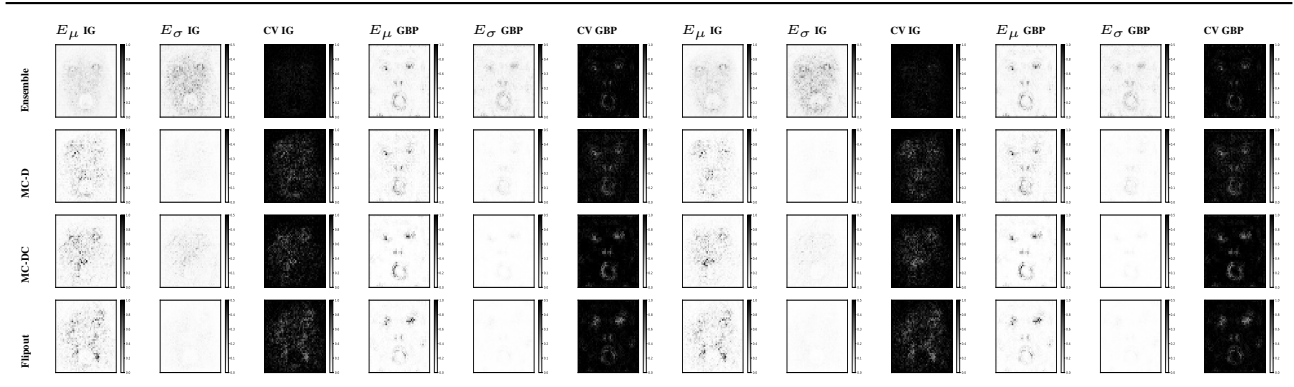


Figure 5. Visualization of the mean ( $E_\mu$ ), the standard deviation ( $E_\sigma$ ), and the coefficient of variation ( $CV$ ) heatmaps generated using the combinations of uncertainty estimation methods (Deep Ensemble, MC-Dropout, MC-DropConnect and Flipout) and explanation methods (Guided Backpropagation and Integrated Gradients). The uncertainty estimation methods are organized in individual rows and the explanation methods are aligned along the columns with the columns 1-3 representing the  $\mu$ ,  $\sigma$ , and the  $CV$  heatmaps for IG and the columns 4-6 representing the same for GBP.

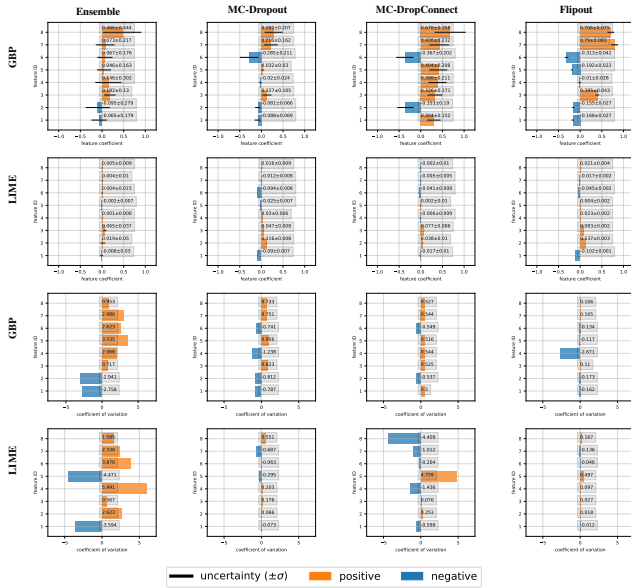


Figure 6. The explanations for the regression task with the mean and the standard deviation explanation (rows 1-2) and the coefficient of variation explanations (rows 3-4). The rows 1 and 3 represent the outputs of GBP and the rows 2 and 4 represent LIME. From left to right, the columns represent as follows: (i) Deep Ensemble (ii) MC-Dropout (iii) MC-DropConnect (iv) Flipout. The input features mapping in the plots is as follows: [1: longitude, 2: latitude, 3: housing median age, 4: total rooms, 5: total bedrooms, 6: population, 7: households, 8: median income].

information required for trustworthiness.

We hope that our proposed pipeline can be utilized to verify new developments in uncertainty estimation and explanation approaches. This work can be extended by incorporating a human-based evaluation of the explanations generated by algorithms. Having such feedback from users could provide valuable insights for improvements in the explanation methods.

## Broader Impact Statement

Both uncertainty estimation and saliency explanations have broad impact on society, as machine learning models are increasingly being used in real-world settings, where humans could be harmed, and uncertainty estimation produces better calibrated output uncertainties to detect when a model makes incorrect or ambiguous predictions. Similarly for saliency explanations, when explanation uncertainty is produced by using an uncertainty estimation method, this would allow human users to obtain additional information in order to decide if the explanation should be trusted or not.

In both cases, output uncertainty and saliency explanations can be misleading, and always additional validation is necessary when used by human users. This paper did not make a user study to evaluate the quality of explanations and uncertainty from a human perspective, we leave this for future work.

## References

- [1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10 (7):e0130140, 2015. 2
- [2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 1
- [3] Kirill Bykov, Marina M-C Höhne, Adelaida Creosteanu, Klaus-Robert Müller, Frederick Klauschen, Shinichi Nakajima, and Marius Kloft. Explaining bayesian neural networks. *arXiv preprint arXiv:2108.10346*, 2021. 3
- [4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2
- [5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1
- [6] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016. 2
- [7] Aryan Mobiny, Pengyu Yuan, Supratik K Moulik, Naveen Garg, Carol C Wu, and Hien Van Nguyen. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, 11(1):1–14, 2021. 2
- [8] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019. 2
- [9] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. 1
- [10] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 4
- [11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 3
- [12] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34:9391–9404, 2021. 3
- [13] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 3
- [14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 2
- [15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 2
- [16] Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516. IEEE, 2022. 4
- [17] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using drop-connect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013. 2
- [18] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018. 2
- [19] Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical image analysis*, 60:101619, 2020. 3, 4
- [20] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. ” why should you trust my explanation?” understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*, 2019. 3
- [21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2