Semantic Attribution For Explainable Uncertainty Quantification

Hanjing Wang[1], Shiqiang Wang[2], and Qiang Ji[3]

[1] Rensselaer Polytechnic Institute, Troy, NY, USA
wangh36@rpi.edu (The Corresponding Author)
[2] IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA
wangshiq@us.ibm.com
[3] Rensselaer Polytechnic Institute, Troy, NY, USA
jiq@rpi.edu

Abstract. Bayesian deep learning, with an emphasis on uncertainty quantification, is receiving growing interest in building reliable models. Nonetheless, interpreting and explaining the origins and reasons for uncertainty presents a significant challenge. In this paper, we present semantic uncertainty attribution as a tool for pinpointing the primary factors contributing to uncertainty. This approach allows us to explain why a particular image carries high uncertainty, thereby making our models more interpretable. Specifically, we utilize the variational autoencoder to disentangle different semantic factors within the latent space and link the uncertainty to corresponding semantic factors for an explanation. The proposed techniques can also enhance explainable out-of-distribution (OOD) detection. We can not only identify OOD samples via their uncertainty, but also provide reasoning rooted in a semantic concept.

Keywords: Bayesian Deep Learning · Uncertainty Attribution.

1 Introduction

While conventional deep learning has made remarkable strides in various domains, it is not without its shortcomings. One notable limitation is the inability of these models to effectively quantify the uncertainties associated with their predictions. This can lead to overconfidence in unfamiliar territories, making the models ill-equipped to identify attacks stemming from data perturbations and out-of-distribution inputs.

Predictive uncertainty can be categorized into two distinct types: epistemic uncertainty and aleatoric uncertainty. Epistemic uncertainty arises due to the model's limited understanding of the input, often stemming from a lack of sufficient training data. Aleatoric uncertainty represents the inherent randomness or noise present within the data itself.

Bayesian deep learning (BDL) models present a well-founded framework for estimating the two types of uncertainties. In contrast to conventional point-estimated models, BDL models emphasize the construction of the posterior distribution of model parameters. By generating predictions from a diverse set of models obtained through sampling the parameter posterior, BDL enables the systematic quantification of predictive uncertainties, providing a more comprehensive understanding of the model's performance and confidence in various situations.

While present BDL methods concentrate on enhancing the accuracy and efficiency of uncertainty quantification (UQ), these approaches are often treated as "black boxes" with limited explainability. Uncertainty attribution (UA) is an essential aspect of UQ that focuses on understanding and explaining the sources and causes of uncertainty within a predictive model. This process offers valuable insights into the model's behavior and allows for enhanced interpretability, trustworthiness, and decision-making in BDL models.

The majority of recently suggested UA methods offer a localized explanation for images that possess high uncertainty. These techniques, often known as local uncertainty attribution strategies, seek to localize the predicted uncertainty by generating an uncertainty map of the input data. This map helps identify the most problematic regions that contribute significantly to prediction uncertainty. By evaluating the contribution of each pixel or data point to the uncertainty, the transparency of BDL models can be substantially improved.

However, local uncertainty attribution falls short in certain scenarios. For instance, when image uncertainty stems from low resolution or random noise, it's impossible to isolate it to a specific region, as these imperfections pervade the whole image. Hence, we propose "semantic uncertainty attribution" for uncertainty reasoning, aiming at identifying the primary factors responsible for input data imperfections that contribute to predictive uncertainty. This method is especially useful in detecting uncertainty sources when data disturbances affect an entire image. Data imperfection mainly arises from data perturbation, indicating noise levels, and data anomaly, reflecting input deviation from the training data distribution. Aleatoric uncertainty basically gauges input perturbations, while epistemic uncertainty measures input anomalies. Nonetheless, image data imperfections can stem from noise, resolution, lighting, object positioning, camera parameters, etc. A deeper understanding of uncertainty sources is desirable, encompassing various types of input perturbations and input anomalies.

In brief, we introduce semantic uncertainty attribution as a method for rationalizing the origins of uncertainty within semantic concepts. Initially, we identify and disentangle the pertinent task-specific factors using a variational autoencoder. Following this, we associate the estimated uncertainty derived from the BDL models with the latent semantic factors to enhance interpretability.

2 Related Work

Classification Attribution Previous attribution methods have been predominantly developed for classification attribution (CA) using deterministic neural networks to determine the contribution of image pixels to the classification score.

Existing CA methods can be broadly categorized into two groups: gradient-based methods and perturbation-based methods. Gradient-based methods [21, 27, 20, 23, 25, 8, 19, 22] leverage gradient information as input attribution, providing insights into the relationship between input features and model predictions. For example, [21] employed the raw gradient to compute the importance of features. To smooth out these raw gradients, [20] multiplied the gradients with the inputs. [22] combined the gradients from several noisy inputs, while [23] accumulated the gradients through a path integral from a reference input to the target input. Perturbation-based methods [17, 15, 6, 3, 5, 26] offer an alternative approach to attributing the contributions of different features by modifying the input and observing the subsequent changes in the model's output.

Local Uncertainty Attribution As shown by [24], CA methods may not be reliable when applied directly to localize uncertainty for identifying problematic regions in the input data. However, various CA methods can be adapted for local uncertainty attribution. Gradient-based CA methods can be extended for uncertainty attribution by adjusting their focus from the model output to the uncertainty. In this paper, we extend gradient [21], Input-G [20], SmoothGrad [22], and IG [23] for UA. Likewise, perturbation-based methods can also be applied to uncertainty localization by observing the changes in uncertainty corresponding to input alterations. In recent years, several methods have been specifically designed for UA. For instance, CLUE [1] and its variants [11, 12] focus on creating an improved image with minimal uncertainty by modifying the uncertain input using a generative model. The attribution map is generated by measuring the difference between the original input and the modified input. To further enhance pixel-wise attributions, [14] combined the CLUE method with the path integral technique. To relax the assumptions of the generative model, [24] proposed UA-Backprop for attributing the input within a single backward pass.

Semantic Uncertainty Attribution Studies for identifying key factors causing high uncertainty in the input data are limited. This complex field necessitates a deep understanding of the data generation process to isolate intertwined factors during learning. [16] attempted to identify uncertainty sources in image object classification by building a variational autoencoder (VAE) with disentangled latent representations. They developed a classification model to predict labels directly from these disentangled representations, computing each factor's attribution score from its optimized uncertainty reduction. However, this field is still under-researched. The applicability of unsupervised disentangled learning for uncertainty attribution remains uncertain. The framework's dual use of features for reconstruction and classification may impair the latter. It also requires solving computation-intensive multi-optimization problems, highlighting a need for more efficient methods. Evaluations of uncertainty attribution methods must also be carefully designed due to the absent ground truth reasoning.

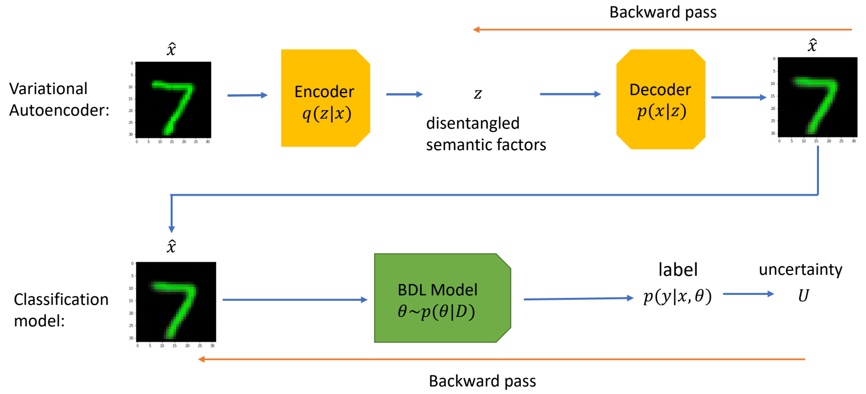

Fig. 1. The overall process of the semantic uncertainty attribution. The blue lines represent the forward propagation and the orange lines represent the backward propagation.

3 Proposed Methods

3.1 Overall Framework

In this study, we present a unique framework using a pre-trained VAE to interpret uncertainty produced by a BDL model. This requires two pre-trained models: a disentangled VAE and a BDL model for classification. The VAE encodes inputs into a disentangled latent space, where each latent factor symbolizes a specific semantic factor. We utilize a Bayesian classification model since it can explain both aleatoric and epistemic uncertainties. Differing from [16], the VAE and the classification model are trained independently to preserve their performance. The attribution process is reserved only for images with high uncertainty.

During the uncertainty reasoning process, our initial step is to link the pre-trained VAE with the classification model for a combined attribution effort. To accomplish this, we feed the image into the VAE for reconstruction. Subsequently, the reconstructed image serves as the input for the classification model to estimate uncertainty. Notably, this approach allows the estimated uncertainty to be a function of the disentangled latent factors, thereby enabling a backward computation of each semantic factor's contribution to the uncertainty. An overall framework is shown in Figure 1 and more details are elaborated in the subsequent sections.

3.2 Disentangled VAE

To identify the essential factors, we propose leveraging recent advancements in disentangled representation learning [9, 7, 2, 13, 18]. We aim to construct a variational autoencoder with an encoder $q(z|x)$ and a decoder $p(x|z)$ to learn interpretable latent representations $z = [z_1, z_2, \cdots, z_K]^T$ that correspond to perturbation factors. The learning of these latent representations varies, which depends on the level of supervision provided. In the initial phase, our assumption is that there is no supervision for the semantic factors and we only have access to the classification label for each image.

To learn the interpretable factors, we employ the $\beta-$VAE [7]. The loss function for the $\beta-$VAE of a single input x is presented as follows:

$$L(x) = -\mathrm{E}_{z \sim q(z|x)}[\log p(x|z)] + \beta D_{KL}(q(z|x)||p(z)) \tag{1}$$

where D_{KL} represents the Kullback-Leibler (KL) divergence and $p(z) = \mathcal{N}(0, I)$ is designated as a standard Gaussian distribution, which acts as the prior distribution for z. The hyperparameter β, which is greater than 1, is utilized to create a balance between the reconstruction performance and the capacity for disentanglement. In the case of a standard VAE, β is equal to 1. However, it has been demonstrated that a larger β value can enhance the performance of disentanglement. It's important to underscore that, in the absence of supervision, we cannot guarantee the disentanglement of all relevant factors. However, when data are methodically generated from diverse factors, we can strive to disentangle these factors to the best of our ability using those unsupervised methods. Besides $\beta-$VAE, other methods can also be employed such as [2, 13, 18]. The performance of attribution is reliant on the disentangled factors, and we cannot assign attribution to a factor that the VAE has not recognized. Within the framework of unsupervised disentangled representation learning, we can discern the meaning of each factor by visualizing an image trajectory.

In our forthcoming research, we plan to disentangle the factors of interest by leveraging various degrees of supervision. When strong supervision is accessible, which implies we have knowledge of these factors' values, we can develop a model that maps the inputs to their corresponding factors. In cases where weak supervision is provided, such as inputs being grouped by similar semantic factors except for one distinct factor, without knowing the specific values of the factors, we can use weakly-supervised disentangled representation learning methods like the one proposed by [9]. In most cases, self-supervision combined with standard data augmentation techniques proves sufficient for disentangling numerous factors essential for classification such as resolution, illumination, rotation, color, and random Gaussian noise.

3.3 Classification Model

In this study, we utilize the deep ensemble method [10] for classification [4], enabling calculating both aleatoric and epistemic uncertainties. Basically, N models parameterized by $\{\theta_s\}_{s=1}^S$ are trained independently with different initializations, donated as $p(y|x, \theta_i)$. In the context of classification tasks, entropy can be

[4] There is debate over whether deep ensemble is a Bayesian method. We believe it is since each ensemble component can serve as a mode of $p(\theta|\mathcal{D})$.

used to assess the uncertainty in class predictions [4]:

$$\underbrace{\mathcal{H}\left[\mathrm{p}(y|x,\mathcal{D})\right]}_{\text{Total Uncertainty}} = \underbrace{\mathcal{I}\left[y,\theta|x,\mathcal{D}\right]}_{\text{Epistemic Uncertainty}} + \underbrace{\mathbb{E}_{\mathrm{p}(\theta|\mathcal{D})}\left[\mathcal{H}[\mathrm{p}(y|x,\theta)]\right]}_{\text{Aleatoric Uncertainty}} \tag{2}$$

where $\mathcal{H}$ and $\mathcal{I}$ represent the entropy and mutual information, respectively. $\mathcal{D}$ is the training data and $p(\theta|\mathcal{D})$ is the posterior distribution of parameters. More specifically,

$$\begin{aligned} &\mathcal{H}\left[p(y|x,\mathcal{D})\right] = \mathcal{H}\left[E_{p(\theta|\mathcal{D})}[p(y|x,\theta)]\right] \approx \mathcal{H}\left[\frac{1}{S}\sum_{s=1}^{S} p(y|x,\theta^s)\right]. \\ &\mathbb{E}_{p(\theta|\mathcal{D})}\left[\mathcal{H}[p(y|x,\theta)]\right] \approx \frac{1}{S}\sum_{s=1}^{S}\mathcal{H}(p(y|x,\theta^s)). \end{aligned} \tag{3}$$

3.4 Semantic Uncertainty Attribution

Forward Propagation With a pre-trained VAE and the classification model at hand, the process initiates with forward propagation, where high-uncertain images are first fed into the VAE. The VAE's encoder produces the mean and covariance matrix of the latent representation z. The mean from the distribution of z is then used as the input for the decoder to generate a reconstructed image. Following this, the reconstructed image is inputted into the BDL model for estimating uncertainty. It's important to highlight that the proposed method is applicable to any form of uncertainty as defined in Eq. (2).

Backward Propagation Rather than resorting to a simplistic attribution method, we exploit the current advancements in explainable AI techniques for effective and precise UA. We incorporate some gradient-based CA techniques and adapt them for UA purposes. Our findings indicate that several methods perform quite satisfactorily, eliminating the need to formulate a specific algorithm exclusively for semantic UA. For the backpropagation step, we test several approaches, restricting our focus only to gradient-based methods to ensure optimal efficiency.

SemanticUA-G: In this approach, we utilize the absolute values of the raw gradients from uncertainty U to the latent representation z as the attribution scores. These scores are represented by Eq. (4):

$$A_G(z) = \left|\frac{\partial U}{\partial z}\right|. \tag{4}$$

In this section, the latent variable z refers to the mean value of the distribution generated by the VAE encoder. The vector $A(z) = [A(z_1), A(z_2), \cdots, A(z_N)]^T$ matches the size of z. Each element, $A(z_i)$, represents the contribution of the ith factor to the overall uncertainty.

SemanticUA-InputG: To account for the potential noise in the raw gradients, an alternative approach is to utilize the InputG method [20]:

$$A_{InputG}(z) = \left| z \odot \frac{\partial U}{\partial z} \right|. \tag{5}$$

where $\odot$ represents the element-wise product.

SemanticUA-SG: SmoothGrad [22] aims to reduce the impact of noisy gradients by aggregating attributions from multiple noisy inputs. Let's denote T as the number of noisy images created by adding Gaussian noise. These noisy images are fed into the encoder, resulting in T noisy latent representations denoted as $\{z^{(t)}\}_{t=1}^{T}$. The computation of $z^{(t)}$ involves applying the function $E_z[q(z|x + \epsilon^{(t)}))]$, where $\epsilon^{(t)} \sim \mathcal{N}(0, \sigma^2 I)$ is a random noise sampled from a Gaussian distribution with a mean of 0 and covariance matrix $\sigma^2 I$. Here, σ represents a hyperparameter and I denotes the identity matrix. Finally, the attribution score is generated using Eq. (6):

$$A_{SG}(z) = \frac{1}{T} \sum_{t=1}^{T} A_G(z^{(t)}). \tag{6}$$

SemanticUA-IG: We can utilize the integrated gradient (IG) [23] method for uncertainty attribution. The adapted version of this method for UA establishes a path integral from a reference latent representation z_0 to z, accumulating the uncertainty gradients U with respect to the latent representations on the path from z_0 to z, as demonstrated in Eq. (7):

$$A_{IG}(z) = (z - z_0) \odot \int_0^1 \frac{\partial U(z_0 + \alpha(z - z_0))}{\partial z} d\alpha. \tag{7}$$

Given that the reference input z_0 bears no uncertainty, the property of completeness is fulfilled, ensuring that the sum of the attribution scores equates to the uncertainty itself:

$$U = \sum_i A_{IG}(z_i). \tag{8}$$

Since IG requires a reference input z_0, we randomly choose z_0 by the encoding of an image from the training data with low uncertainty.

4 Experiment

Dataset We employ the colored MNIST dataset for image classification and the disentanglement of semantic factors. As part of our future work, we plan to extend our analysis to include more datasets.

Disentangled VAE To achieve the disentanglement of semantic factors, we employ β-VAE, an unsupervised approach. We can visualize the meaning of each factor through an image trajectory. The implementation code, utilizing the default model architecture and hyperparameters, can be found at the following URL: `https://github.com/YannDubs/disentangling-vae`.

Bayesian Deep Learning Model For approximating the posterior distribution of parameters and performing uncertainty quantification, we utilize the deep ensemble method [10]. This approach involves training five ensemble models independently, each initialized differently. We employ the following architecture: Conv2D-Relu-Dropout-Conv2D-Relu-MaxPool2D-Dropout-Dense-Relu-Dropout-Dense-Softmax. Each convolutional layer consists of 32 convolution filters with a kernel size of 4×4. Additionally, we utilize a max-pooling layer with a kernel size of 2×2, two dense layers with 128 units, and a dropout probability of 0.25. The batch size is set to 128, and the maximum epoch is 50. We employ the SGD optimizer with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0005.

Baselines We introduce several attribution methods, namely SemanticUA-G, SemanticUA-InputG, SemanticUA-SG, and SemanticUA-IG, by employing different techniques. For SemanticUA-IG, we approximate the integration in Eq. (7) by sample average, where 200 samples of α are linearly generated between 0 and 1. For SemanticUA-SG, the number of noisy images, denoted as T, is equal to 200. The added noise is sampled from a Gaussian distribution with 0 mean and 0.1 standard deviation. Due to the relatively limited exploration in this area, we conduct a comparison of our method exclusively with SourceUA [16].

4.1 Explainable OOD Detection

In this experiment, our objective is to detect synthetically generated out-of-distribution samples and provide explanations for why they differ from the training data. The experiment is structured in a way that we have prior knowledge of the underlying reasons for problematic images. We then assess the capability of our approaches to identify these reasons through the utilization of the backpropagation step.

Experiment Setting For training the disentangled VAE, we employ the colored MNIST dataset, which allows us to disentangle a specific factor representing the digit's color. Figure 2 illustrates that z_5 corresponds to the color factor of the digit. In this experiment, we train the classification model exclusively on red images. Consequently, when encountering images with green or blue colors, the BDL classification model should exhibit large uncertainty. Thus, we can conclude that the primary factor contributing to the observed predictive uncertainty is the color itself. Subsequently, we apply our semantic UA method to identify the reasons behind the high uncertainty observed in images with different colors. We utilize total uncertainty for this evaluation.

Evaluation Metric We design the accuracy (ACC) by the percentage of the number of successes in the detection of the color factor that contributes most to the uncertainty. Let $\{z^m\}_{m=1}^{M}$ denote the encodings of the testing images. The

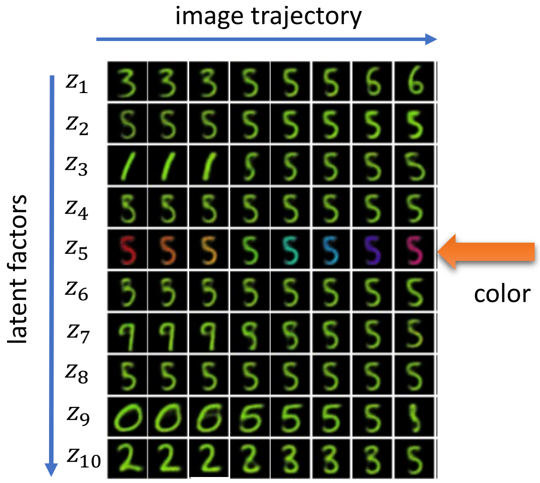

Fig. 2. The image trajectory for the disentangled VAE. Each row represents the image trajectory for a particular semantic factor. The reconstructed images are displayed by smoothly transitioning the values of z_k from small to large while keeping the other latent factors fixed.

detection accuracy can be calculated in the following equation:

$$ACC = \frac{1}{M}\sum_{m=1}^{M}\delta(\arg\max_{k=1:K} A(z_k^m) = 5) \tag{9}$$

where δ is a delta function that returns 1 only when $\arg\max_{k=1:K} A(z_k^m) = 5$ and otherwise, it will return 0. Our objective is not only to ensure that the "color" latent factor is the primary source of uncertainty but also to ensure that it is the sole source of uncertainty. This is because, during the generation of problematic images, we solely modify the color information. As a result, we anticipate that the color factor can account for nearly 100% of the uncertainty. The explanation rate for factor z_k can be defined as follows:

$$ER(k) = \frac{1}{M}\sum_{m=1}^{M}\frac{\exp(A(z_k^m))}{\sum_{j=1}^{K}\exp(A(z_j^m))}. \tag{10}$$

In our experiment, our expectation is that $ER(5)$, the explanation rate for the color factor, should be maximized for images with different colors in comparison to the training data.

Experiment Results The results presented in Table 1 demonstrate the substantial improvement achieved by our semantic uncertainty attribution framework compared to SourceUA. This improvement can primarily be attributed to the fact that SourceUA directly utilizes disentangled latent representations

Table 1: ACC, ER(5), and empirical runtime for semantic UA evaluation on colored MNIST dataset. The last column displays the time required to attribute a single input. In the case of green and blue images, where the color factor primarily drives uncertainty, we anticipate a higher value for ACC and ER(5). Conversely, for red images, we expect lower values for ACC and ER(5) because color is not the primary source of predictive uncertainty in this context.

Method	Green		Blue		Red		UA Time
	ACC ↑	ER(5) ↑	ACC ↑	ER(5) ↑	ACC ↓	ER(5) ↓	
SemanticUA-G	0.733	0.220	0.523	0.139	0.013	**0.095**	0.04s
SemanticUA-InputG	0.450	0.137	0.400	0.122	0.107	0.096	0.04s
SemanticUA-SG	0.623	0.257	0.467	0.118	**0.000**	0.096	0.65s
SemanticUA-IG	**1.000**	**0.503**	**0.987**	**0.448**	**0.000**	0.100	0.66s
SourceUA	0.050	0.101	0.057	0.102	0.027	0.101	28.44s

from the VAE as input to the classifier, which can negatively impact the accuracy of uncertainty quantification in the classification model. Among the various methods within our framework, SemanticUA-IG stands out with the highest performance. It achieves nearly 100% accuracy in identifying the underlying causes of uncertainty when provided with ground truth information.

5 Conclusion

In conclusion, this paper has explored the application of semantic uncertainty attribution in the context of Bayesian deep learning, with a particular focus on uncertainty quantification and interpretation. By leveraging the variational autoencoder, we successfully disentangled various semantic factors within the latent space, enabling us to attribute uncertainty to specific factors. This approach significantly enhances the interpretability of our models by providing explanations for high uncertainty in individual images. Moreover, we have demonstrated the utility of these techniques in improving OOD detection, where not only can we identify OOD samples based on their uncertainty, but we can also offer explanations grounded in semantic concepts.

In our future work, we plan to expand our disentanglement efforts to encompass additional classification-relevant factors such as image resolution, illumination, and random noise. We will explore various disentangled representation learning methods to effectively disentangle and attribute these factors to different types of uncertainty, thereby enhancing our understanding of the distinctions between aleatoric and epistemic uncertainty. Furthermore, we aim to leverage the estimated semantic uncertainty attribution to improve model performance, taking a step beyond mere explanation toward the development of actionable Bayesian deep learning techniques. By incorporating these advancements, we can construct models that not only provide explanations but also actively utilize uncertainty information to drive improved performance and decision-making.

References

1. Antorán, J., Bhatt, U., Adel, T., Weller, A., Hernández-Lobato, J.M.: Getting a clue: A method for explaining uncertainty estimates. arXiv preprint arXiv:2006.06848 (2020)
2. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. Advances in neural information processing systems **31** (2018)
3. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. Advances in neural information processing systems **30** (2017)
4. Depeweg, S., Hernandez-Lobato, J.M., Doshi-Velez, F., Udluft, S.: Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In: International Conference on Machine Learning. pp. 1184–1193. PMLR (2018)
5. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2950–2958 (2019)
6. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE international conference on computer vision. pp. 3429–3437 (2017)
7. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework (2016)
8. Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., Bolukbasi, T.: Guided integrated gradients: An adaptive path method for removing noise. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5050–5058 (2021)
9. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.: Deep convolutional inverse graphics network. Advances in neural information processing systems **28** (2015)
10. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles (2016), `http://arxiv.org/abs/1612.01474`
11. Ley, D., Bhatt, U., Weller, A.: {\delta}-clue: Diverse sets of explanations for uncertainty estimates. arXiv preprint arXiv:2104.06323 (2021)
12. Ley, D., Bhatt, U., Weller, A.: Diverse, global and amortised counterfactual explanations for uncertainty estimates. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 7390–7398 (2022)
13. Margonis, V., Davvetas, A., Klampanos, I.A.: Wela-vae: Learning alternative disentangled representations using weak labels. arXiv preprint arXiv:2008.09879 (2020)
14. Perez, I., Skalski, P., Barns-Graham, A., Wong, J., Sutton, D.: Attribution of predictive uncertainties in classification models. In: The 38th Conference on Uncertainty in Artificial Intelligence (2022)
15. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)
16. Rey, L.A.P., İşler, B., Holenderski, M., Jarnikov, D.: Identifying the sources of uncertainty in object classification
17. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

18. Sarhan, M.H., Eslami, A., Navab, N., Albarqouni, S.: Learning interpretable disentangled representations using adversarial vaes. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 37–44. Springer (2019)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
20. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
21. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
22. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
23. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
24. Wang, H., Joshi, D., Wang, S., Ji, Q.: Gradient-based uncertainty attribution for explainable bayesian deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12044–12053 (2023)
25. Xu, S., Venugopalan, S., Sundararajan, M.: Attribution in scale and space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9680–9689 (2020)
26. Yang, Q., Zhu, X., Fwu, J.K., Ye, Y., You, G., Zhu, Y.: Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1376–1383. IEEE (2021)
27. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)