Understanding the Reason for Misclassification by Generating Counterfactual Images

Muneaki Suzuki[1] Yoshitaka Kameya[1,*] Takuro Kutsuna[2] Naoki Mitsumoto[2]

Abstract

Explainable AI (XAI) methods contribute to understanding the behavior of deep neural networks (DNNs), and have attracted interest recently. For example, in image classification tasks, attribution maps have been used to indicate the pixels of an input image that are important to the output decision. Oftentimes, however, it is difficult to understand the reason for misclassification only from a single attribution map. In this paper, in order to enhance the information related to the reason for misclassification, we propose to generate several counterfactual images using generative adversarial networks (GANs). We empirically show that these counterfactual images and their attribution maps improve the interpretability of misclassified images. Furthermore, we additionally propose to generate transitional images by gradually changing the configurations of a GAN in order to understand clearly which part of the misclassified image cause the misclassification.

1 Introduction

Deep neural networks (DNNs) have demonstrated high accuracy in several visual recognition tasks, and are expected to be applied to more critical situations (e.g., medical diagnosis and automatic driving). Despite the remarkable performance, their decisions are difficult to understand, which often restricts the application field of DNNs. Many studies on explainable AI (XAI) [1] have attempted to mitigate this problem. In image classification tasks, attribution map methods have been used to confirm whether a classifier network focuses on the correct region in the input image or not. Typically, attribution maps are created by back-propagating a relevance score from the output layer to the input layer, and highlight the important input pixels in warm colors or the unimportant pixels in cold colors. In particular, relative attributing propagation (RAP) [7] creates attribution maps which clearly separate the important and unimportant regions. Oftentimes, however, a single attribution map is not sufficient in exploring the reason for misclassification. For example, the image in Figure 1 (left) is the one misclassified into "standard poodle," and the correct class is "brown bear." Although its attribution map strongly highlights the creature's nose, we cannot immediately be sure that this is the reason for misclassification.

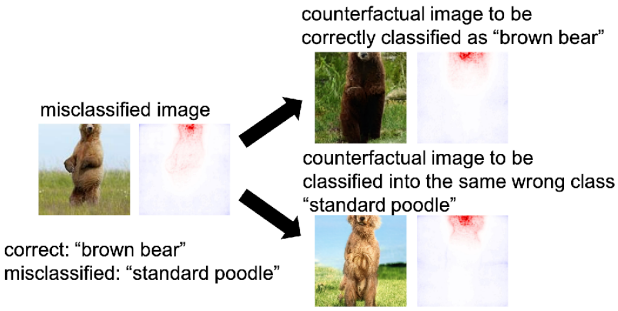


Figure 1: An example of a misclassified image (left) and the generated counterfactual images (right).

In this paper, we propose to generate counterfactual images similar to the misclassified image using generative adversarial networks (GANs). We use BigGAN [3] which generates high resolution and realistic images, and optimize the input noise vector of BigGAN so that the generated counterfactual images have high-level features which have high similarity to those of the misclassified image. Then, we compare the generated images that will be correctly classified into the original input image's ground truth class ("brown bear" in Figure 1), and the ones that will be classified into the same wrong class ("standard poodle"). In addition, we create attribution maps for the generated counterfactual images. By this comparison, we would be able to understand the reason for the misclassification more clearly. Figure 1 (right) illustrates two exemplar counterfactual images generated by the proposed method and their attribution maps. By comparing these materials, one may consider that the thinness of necks can be a discriminating factor between "standard poodle" and "brown bear," and the misclassification might have been caused by the thinness of the creature's neck.

In this paper, we further propose to generate transitional counterfactual images from the ones with higher confidence for the correct class to those for the wrong class by gradually changing the configurations of BigGAN. These transitional images would make clearer the discriminating factors between correctly-classified images and misclassified images, and thus give more clues to the reason for misclassification.

[1]Dept. of InformationEngineering, Meijo University
[2]DENSO CORPORATION
*ykameya@meijo-u.ac.jp

This paper is outlined as follows. First, Section 2 describes the background and the related work of the paper. Section 3 presents the proposed method. Section 4 reports the experimental results, and Section 5 concludes the paper.

2 Related Work

Attribution map methods create a heatmap which highlights important regions typically by backpropagating a relevance score depending on each neuron's contribution. So far, dozens of attribution map methods have been proposed, e.g., layer-wise relevance propagation (LRP) [2] and relative attributing propagation (RAP) [7]. RAP is an improvement of LRP which does not cancel out positive and negative relevance scores, and creates attribution maps which clearly separate important regions and unimportant regions. In this work, we use attribution maps created by RAP.

In generating counterfactual images, we use GANs. By simultaneously training the generator which generates fake images from the noise vector and the discriminator which discriminates between real images and fake images, we make the generator generate realistic fake images. BigGAN [3] is known to generate high resolution images. The generator of BigGAN requires three input configurations: the noise vector, the class vector and the diversity/quality threshold. The class vector is a k-dimensional vector whose elements take on [0, 1] and sum up to unity, where k is the number of classes. By specifying this, we give weights to the classes of images to be generated. Typically we use a one-hot vector for generating images of a particular class. Furthermore, we can generate morphed images when a couple of classes are specified at once. The diversity/quality threshold ranges from 0.0 to 1.0, and balances between the diversity and the quality of the images to be generated. If this threshold is 1.0, generated images will be of high diversity but of low quality.

Instance-based explanation methods present images relevant to the test image under study. The method using the influence function [6] searches for the images which have the positive or negative effects by evaluating the loss of the test image with respect to training images. In this paper, we focus on misclassified images, and generate images whose high-level features are similar to those of the test image. Thanks to using a GAN, many explanatory instances would be available even if such instances are few in the training dataset.

Counterfactual explanation methods [4, 5] modify the test image under study to increase the confidence for a particular class. Some methods identify the region in the test image to be replaced with uninformative pixels or another image's fragment in order to change the classifier network's decision. In this paper, we generate a series of transitional counterfactual images from the ones to be correctly classified to those to be classified into the wrong class.

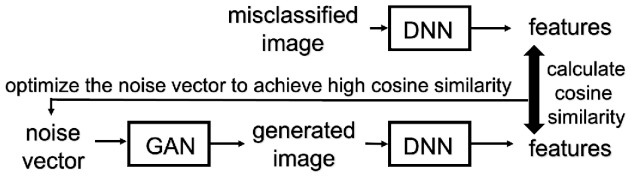


Figure 2: The outline of generating single counterfactual images.

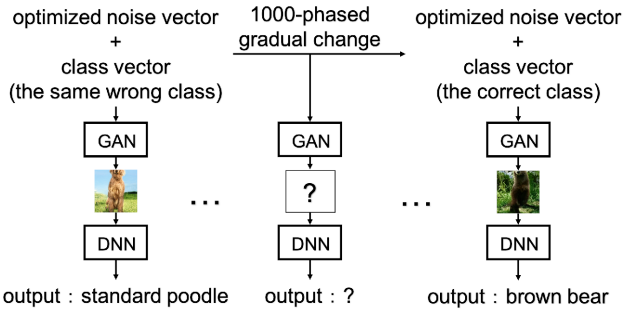


Figure 3: The outline of generating transitional images.

3 Proposed Method

Figure 2 outlines the proposed method. Here, we optimize the noise vector z, which is an input of BigGAN, in order to generate counterfactual images having high-level features similar to those of the misclassified image. We conduct gradient descent for solving the optimization problem in Eq. 1, where $\phi(x)$ refers to the high-level features of each image x extracted by the discriminator. That is, we adjust the noise vector z in order to enable the generator G to generate images whose high-level features has a high cosine similarity to those of the misclassified image x^*.

$$\hat{z} = \underset{z}{\operatorname{argmax}} \cos(\phi(x^*),\ \phi(G(z))) \tag{1}$$

Besides, we also propose to generate transitional counterfactual images as additional clues. Figure 3 outlines the procedure. To generate such images, by linear interpolation, we gradually change the input configurations (the noise vector and the class vector) of BigGAN over 1000 phases. We choose the optimized noise vector which generates a counterfactual image with high confidence for the misclassified class as the starting point, and that for the correct class as the goal. The high-level features of these two end-point images must have high cosine similarity to those of the misclassified image.

Finally, for all the generated images, we conduct classification, and create attribution maps using RAP, in a standard manner.

4 Experiments

We chose RMSprop as the gradient descent method, since it worked most stably in the preliminary experiments. We updated the noise vector 2000 times. High-level features were extracted by VGG-19 truncated fully connected layers and the shape of the extracted high-level features was (7, 7, 512). We used BigGAN which has three input configurations: the noise vector, the class vector, and the diversity/quality threshold. In the experiments, we initialized the noise vector randomly in 32 ways, and the class vector in three ways: a one-hot vector representing the correct class, a one-hot vector representing the wrong class, and a vector between them. The last one is actually the output of the softmax function. The diversity/quality threshold was fixed at 1.0 in order to generate high diversity images. Combining 32 noise vectors and three class vectors, 96 counterfactual images were finally generated for each misclassified image. We input all the generated counterfactual images into VGG-19 to perform classification, and obtained their attribution maps.

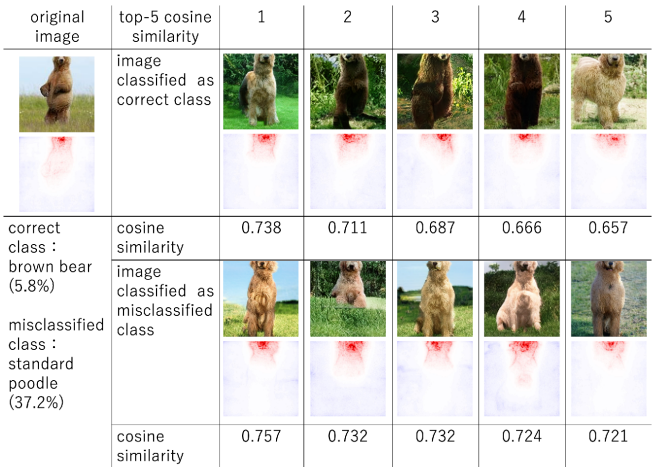
4.1 Comparative Counterfactual Images

Figure 4 shows a part of the experimental results. The original input images we pick up here are the misclassified ones. Upper-row images are the generated counterfactual images with the highest cosine similarity which were classified into the correct class, and lower-row images are those classified into the same wrong class. Generating a counterfactual image took approximately 83 seconds on average.

In Figure 4a, the original image was classified into "standard poodle," but the correct class is "brown bear." All generated images look similar to the original image, but their attribution maps for each class are slightly different. The attribution maps of the images classified into "brown bear" primarily highlight the face, but those of "standard poodle" highlight not only the face but the bottom of the neck and detect the neck edges. At the moment, one may find that the creature in the original image has a thin neck like the images classified into "standard poodle," whose attribution maps highlight the face and the bottom of the neck. Furthermore, he/she may say more generally that a brown bear having a thin neck like the original image tends to be misclassified into "standard poodle."

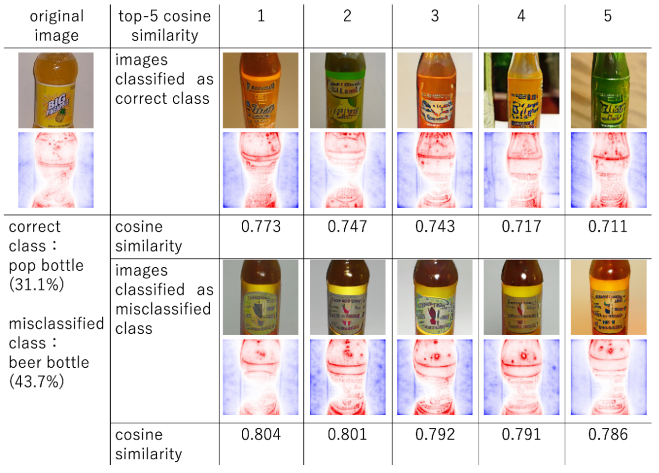
In Figure 4b, the original image was classified into "gibbon," but the correct class is "brown bear." The images classified into "brown bear" have big ears above their faces, and their attribution maps highlight the top of the head. The images classified into "gibbon" have a round face, and their attribution maps primarily highlight the face. In the original image, the ear is hidden by some leaves, and the attribution map primarily highlights the face.

In Figure 4c, the original image was classified into



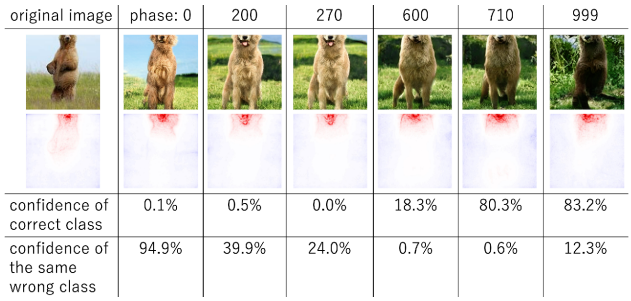
(a) "brown bear" classified into "standard poodle"

(b) "brown bear" classified into "gibbon"



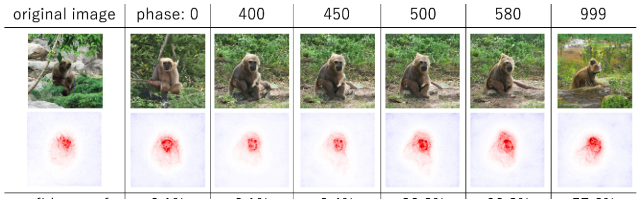
(c) "pop bottle" classified into "beer bottle"

Figure 4: Counterfactual images having high-level features similar to those of the misclassified image.



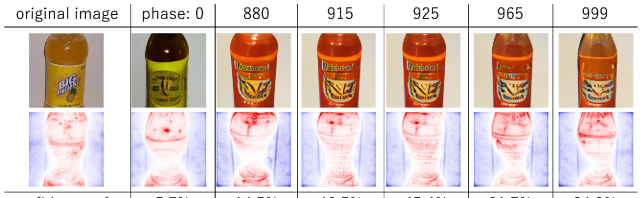
original image	phase: 0	200	270	600	710	999
confidence of correct class	0.1%	0.5%	0.0%	18.3%	80.3%	83.2%
confidence of the same wrong class	94.9%	39.9%	24.0%	0.7%	0.6%	12.3%

(a) "brown bear" classified into "standard poodle"



original image	phase: 0	400	450	500	580	999
confidence of correct class	0.1%	0.1%	0.4%	30.2%	99.2%	57.6%
confidence of the same wrong class	86.1%	38.9%	28.7%	1.7%	0.0%	1.6%

(b) "brown bear" classified into "gibbon"



original image	phase: 0	880	915	925	965	999
confidence of correct class	5.7%	14.5%	40.5%	45.4%	81.7%	84.9%
confidence of the same wrong class	91.2%	4.9%	30.1%	24.1%	17.0%	13.5%

(c) "pop bottle" classified into "beer bottle"

Figure 5: Results of generating transitional images.

"beer bottle," but the correct class is "pop bottle." In the images classified into "pop bottle," the bottles are upwardly tapered. In the images classified into "beer bottle," the upper side of the bottles is rounded and a light is reflected. The attribution maps for each class highlight the shapes of the top of bottles, and the ones in the "beer bottle" class focus on a light reflection. One may guess that the original image was classified into "beer bottle" since the bottle has a rounded shape and a light reflection.

4.2 Transitional Counterfactual Images

Figures 5 and 6 show the generated transitional images. The original misclassified images are the ones also shown in Figure 4. Figure 6 shows the changes of the confidence for each class. The blue solid line indicates the confidence for the correct class and the orange dotted line indicates that for the misclassified

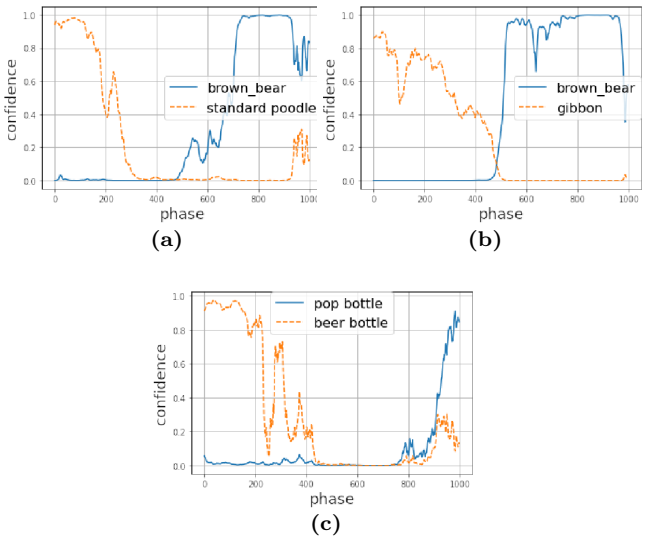


Figure 6: Change in confidence for each class in Figure 5.

class. Figure 5 shows typical transitional images. The 0th image is a counterfactual image having high confidence for the misclassified class, and the 999th image has high confidence for the correct class.

In Figure 5a, as the neck gets thinner like the 200th to 270th images, the confidence for "standard poodle" decreases, and as the neck gets thicker like the 600th to 710th images, the confidence for "brown bear" increases. Their attribution maps turn not to highlight the neck gradually. In Figure 5b, as something like a ear start to appear like the 500th image, the confidence for "gibbon" increases and their attribution maps highlight the ear gradually. In Figure 5c, as the upper side of the bottle becomes thinner and the light reflection fades like the 880th to 999th images, the confidence for "pop bottle" increases and the confidence for "beer bottle" decreases. These results would support the observations in Section 4.1.

5 Conclusion and Future Work

In this paper, we proposed to generate counterfactual images similar to misclassified images, and empirically showed that these images would help us to guess the reason for misclassification. In addition, we proposed to generate transitional images and confirm what differences in the transitional images makes the changes in the confidence for each class. In future, we would like to conduct quantitative or large-scale subjective evaluation of the proposed methods, and improve the predictive performance of the classifier network exploiting generated counterfactual images. We also plan to apply the proposed method to other datasets where we need to build GANs from scratch.

References

[1] Adadi, A. et al.: Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access, Vol. 6, 2018.

[2] Bach, S. et al.: On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation, PLOS ONE, Vol. 10, No. 7, 2015.

[3] Brock, A. et al.: Large Scale GAN Training for High Fidelity Natural Image Synthesis, Proc. of ICLR-19, 2019.

[4] Chang, C.-H. et al.: Explaning Image Classifiers by Counterfactual Generation, Proc. of ICLR-19, 2019.

[5] Goyal, Y. et al.: Counterfactual Visual Explanations, Proc. of ICML-19, 2019.

[6] Koh, P.-W. et al.: Understanding Black-box Predictions via Influence Functions, Proc. of ICML-17, 2017.

[7] Nam, W.-J. et al.: Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks, Proc. of AAAI-20, 2020.