# Generating Natural Counterfactual Visual Explanations

**Wenqi Zhao**[1] , **Satoshi Oyama**[1,2] and **Masahito Kurihara**[1]
[1] Graduate School of Information Science and Technology, Hokkaido University
[2] Global Institution for Collaborative Research and Education, Hokkaido University
choubunnki@complex.ist.hokudai.ac.jp, {oyama, kurihara}@ist.hokudai.ac.jp

## Abstract

Counterfactual explanations help users to understand the behaviors of machine learning models by changing the inputs for the existing outputs. For an image classification task, an example counterfactual visual explanation explains: "for an example that belongs to class *A*, what changes do we need to make to the input so that the output is more inclined to class *B*." Our research considers changing the attribute description text of class *A* on the basis of the attributes of class *B* and generating counterfactual images on the basis of the modified text. We can use the prediction results of the model on counterfactual images to find the attributes that have the greatest effect when the model is predicting classes *A* and *B*. We applied our method to a fine-grained image classification dataset and used the generative adversarial network to generate natural counterfactual visual explanations. To evaluate these explanations, we used them to assist crowdsourcing workers in an image classification task. We found that, within a specific range, they improved classification accuracy.

## 1 Introduction

### 1.1 Background

The growing use of machine learning in various real-world applications has led to the interpretability of machine learning becoming an active research area [Doshi-Velez and Kim, 2017]. Deepening our understanding of machine learning model decision-making can help us create safer, more reliable products and reduce possible risks.

Most of the recent work on visual models has focused on finding the areas in the input image that has the greatest effect on the model decision-making. In our work, we explored counterfactual visual explanations from the perspective of counterfactual thinking. The feature that has the greatest effect on the model prediction result is what we currently believe is the essential counterfactual feature. We pass the counterfactual feature as a counterfactual explanation to the questioner. Figure 1 shows a counterfactual visual explanation between Nashville warbler and mourning warbler." A black chest patch is an attribute that the Nashville warbler does not have. If the bird in the picture has a black chest patch, the model's prediction result is more biased towards mourning warbler. So the black chest patch is a counterfactual explanation of why this bird is not a mourning warbler.

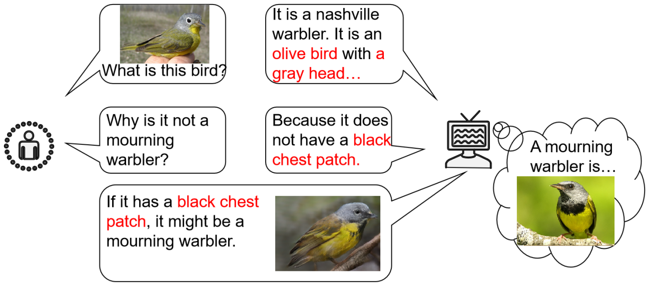
What is this bird?
It is a nashville warbler. It is an olive bird with a gray head…
Why is it not a mourning warbler?
Because it does not have a black chest patch.
A mourning warbler is…
If it has a black chest patch, it might be a mourning warbler.

Figure 1: Counterfactual visual explanation between Nashville warbler and mourning warbler.

### 1.2 Related Work

Many methods for creating a visual explanation of image classification have emerged in recent years. Explaining from a counterfactual perspective is an essential and relatively novel interpretable study.

Hendricks et al. [Hendricks *et al.*, 2018] trained a model for generating counterfactual explanations with the help of additional attribute annotations to explain why the model predicted results for class *I* rather than class *I'*. To solve a task similar to ours, Goyal et al. [Goyal *et al.*, 2019] proposed a technique for generating counterfactual visual explanations. For given images *i* and *i'* belonging to classes *I* and *I'*, respectively, the model prediction result is changed from class *I* to Class *I'* by identifying the spatial region in the two images and replacing the recognition region in *i* with the recognition region in *i'*. However, the quality of the generated results depends on whether the action postures of the subjects in the original image are consistent. If we use two images with a significant difference in attitude (such as an image of a bird flying from the front and an image of a bird resting on a branch), the result of region replacement may be unnatural.

## 2 Proposed Method

The goal of our research is to generate natural counterfactual visual explanations. To avoid the unnatural phenomenon when replacing the recognition area, we use predefined description texts and generative adversarial network to generate an image as a counterfactual visual explanation.

### 2.1 Predefined Description Texts

Our approach is to obtain the characteristics of birds through text descriptions and generate texts containing counterfactual features. If the features in the text description data are not detailed enough, the counterfactual text generation may be affected, so we make new text descriptions for each class of bird. We obtain the definition of each class of bird from the Internet and briefly specify its characteristic attributes as the *bill, head, breast, back, wings, tail, and feet*. The specific size of a bird is not an attribute that can be accurately obtained from an image, so it is not easy to express this attribute in the generated image. Therefore, size is not within the scope of our current research. Besides, we believe that designating the features of a bird to be dispersed into the features of a specified part may help us obtain the features of the missing element in the image from the user's input when putting into the application.

### 2.2 Generating Counterfactual Visual Explanations

We propose using a text-to-image generative adversarial network (GAN) model to generate the images. We look for counterfactual features belonging to class *B* that do not exist in class *A*. We use each counterfactual feature to replace the corresponding class *A* feature and output a counterfactual text. The counterfactual text contains the B-type features of one part and the A-type features of the remaining parts. Then we use a text-to-image GAN model and the counterfactual text to generate a counterfactual image. We use two GAN models (StackGAN [Zhang *et al.*, 2017] and AttnGAN [Xu *et al.*, 2018] to generate images. We use an image classification model to obtain the prediction probabilities of each counterfactual image for A and B.

We consider using the logarithm of probability for class A and class B as the classification results for counterfactual image *I*, as shown in equation 1, where $p(B)$ is the probability of to class B and $p(A)$ is the probability of class A. We choose the highest-scoring counterfactual image and combine its counterfactual feature with the prepared text to generate a counterfactual image explanation.

$$\log \frac{p\left(B\right)}{p\left(A\right)} \tag{1}$$

## 3 Evaluation

We recruited 50 random crowd workers and used an image classification task to evaluate our counterfactual visual explanations. The workers performed normal classification tasks based on the basic defined images of birds and actual images and do the same task after referring to the counterfactual visual explanations. The classification accuracy without the explanations was 0.81; with the explanations, it was 0.92. This demonstrates that our counterfactual images explanations can improve the accuracy of classification by crowd workers. Since there are many unstable factors in crowdsourcing experiments, we need to perform more crowdsourcing experiments to support our findings.

## 4 Conclusion

We have presented a method for generating counterfactual visual explanations and used a crowdsourcing task to evaluate their effectiveness for classification. The results demonstrated that the visual explanations improved classification accuracy to a certain extent. Our approach requires predefined attribute annotation text, and the generated model limits the quality of the generated counterfactual visual explanations. On the other hand, it enables us to focus on more comprehensive attribute features that cannot be obtained from a single image and generate more natural visual explanations.

In future work, we will focus on the generation of complex and unique shape features that are difficult to generate from generative adversarial models. In future work, we will focus on detailed feature generation in order to improve the quality of the counterfactual visual explanations.

## Acknowledgments

This work was partially supported by JSPS KAKENHI Grant JP18H03337, by the Telecommunications Advancement Foundation, and by the Global Station for Big Data and Cybersecurity, a project of the Global Institution for Collaborative Research and Education at Hokkaido University.

## References

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Goyal *et al.*, 2019] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*, 2019.

[Hendricks *et al.*, 2018] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

[Xu *et al.*, 2018] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[Zhang *et al.*, 2017] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.