DERIVING EXPLAINABLE DISCRIMINATIVE ATTRIBUTES USING CONFUSION ABOUT COUNTERFACTUAL CLASS

Nakyeong Yang[1], *Taegwan Kang*[2], *Kyomin Jung*[1,2,*]

[1]Dept. of Artificial Intelligence, Seoul National University, Seoul, Korea
[2]Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea

ABSTRACT

Recently, Integrated Gradients-based (IG) methods have been commonly used to explain the decision process of deep neural networks (DNNs). However, they have only considered the information of the predicted class while neglecting the information of the rest classes. In this paper, we propose a novel counterfactual explanation method, Discriminative Gradients (DiscGrad) that derives explainable discriminative attributes by considering not only the predicted class but also the counterfactual classes. Specifically, we calculate the discriminative attributes by removing the attribute of the counterfactual classes, and this process makes it possible to derive only key discriminative attributes that contrast with other decisions. Also, we determine the weights for discriminative attributes using the degree of confusion about counterfactual classes. We evaluated our method by measuring how much logit decreases by perturbing important attributes. Experimental results on the widely used image and text datasets show that our proposed method outperforms the strong baseline, IG. In addition, we examine the relationship between class correlation and the performance of discriminative attribute to demonstrate the effectiveness of our method.

***Index Terms*—** explainable AI, counterfactual explanation, deep learning

1. INTRODUCTION

As deep neural networks (DNNs) are replacing humans in making decisions in various fields [1, 2, 3, 4], researches on explaining the decision process of DNNs have intensified. These researches have derived important features in the decision process by using gradient-based method [5, 6, 7, 8], activation map-based method [9, 10], and propagation-based method [11, 12, 13]. Among them, our work is focused on the gradient-based methods that use the gradient of the model's decision to its input as important features. In particular, Integrated Gradients-based (IG) methods have been commonly used because they do not require any modification to the original network and are simple to implement [6, 7, 14, 8].

*Corresponding Author.

However, these previous methods have only considered the information of the predicted class and disregarded the helpful counterfactual information in the rest classes. Specifically, IG methods have derived important attributes by using only the gradient information for the predicted class without considering any counterfactual classes. In order to solve these problems, it is necessary to derive the basis for the decision process using a counterfactual explanation. A counterfactual explanation describes a causal situation in the form: *"If X had not occurred, then Y would not have occurred"* [15]. In other words, the counterfactual explanation problem is to formulate the problem of finding X that will eventually lead to a change in the decision Y. This problem can be identified in various decision process cases in real life, such as *"The problem of classifying whether a person with specific personal information (eg, age, gender, property, debt, etc.) can borrow money from the bank"*. If a specific person is classified as non-loanable, where X is personal information, Y corresponds to a loanability. Here, it is possible to derive knowledge about how to change his loanability status from a *"not loanable state"* to a *"loanable state"* by modifying which personal information. By deriving discriminative attributes for counterfactual decisions, we can present better interpretability for human understanding.

In this paper, we propose a novel counterfactual explanation method Discriminative Gradients (DiscGrad), which is the method for deriving explainable discriminative attributes by considering not only the predicted class but also the counterfactual classes. Specifically, we use IG to derive the attributes for the predicted class and counterfactual class, and calculate the discriminative attribute by taking the difference between them. In addition, we derive the degree of confusion with the counterfactual class, and use it as a weight for each discriminative attribute. Depending on the degree of correlation between the predicted class and the counterfactual class, the discriminative attribute may be less meaningful. Therefore, we also use the degree of confusion as a weight to control the application of the total discriminative attribute. By removing counterfactual attribute and applying an appropriate weight to it, it is possible to derive only discriminative attribute that contrasts with other decisions. We evaluate our method on real-world image and text datasets. The experi-

mental results show that our approach definitely derives discriminative key attributes by showing superior performance than the existing method. In addition, we demonstrate the effectiveness of our method by visualizing discriminative attributes and identifying the relationship between class correlation and the performance of our method.

2. METHODOLOGY

2.1. Integrated Gradients

In this paper, we apply Integrated Gradients (IG) to extract the feature importance from the image and text data. For image, we use the pixel as a feature, and for text, the word is used as a feature. Formally, suppose we have a function $F : \mathbb{R}^d \rightarrow [0,1]^m$ that represents deep neural networks for multi-class classification, and let $x \in \mathbb{R}^d$ be the input feature and $x' \in \mathbb{R}^d$ be the baseline input. We use a black image as a baseline input for image processing and a zero vector for natural language processing. In detail, the contribution of the i-th feature in x to the prediction of c-th class using $F(x)$ is defined as follows:

$$IG_i^{(c)}(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F^{(c)}(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (1)$$

where $\partial F^{(c)}(x)/\partial x_i$ is the gradient of $F^{(c)}(x)$ along with the i-th feature. Since equation 1 is intractable for deep neural networks, we replace an integral part with the gradient addition formula:

$$IG_i^{(c)}(x) = \frac{1}{m} \times (x_i - x_i') \times \sum_{k}^{m} \frac{\partial F^{(c)}(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \quad (2)$$

where m is the number of steps in the Riemman approximation of the integral. By using equation 2, it is possible to calculate the contribution of each i-th feature to the decision with a specific class c.

2.2. Discriminative Attribute

Using the IG, it is possible to calculate the importance of each feature with a specific class c. Therefore, the importance of each feature to the predicted class y can be expressed as follows:

$$\begin{gathered} y = \underset{c}{\operatorname{argmax}}\, F^{(c)}(x) \\ A_i^{(y)}(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F^{(y)}(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \end{gathered} \quad (3)$$

In addition, the importance of each feature for the counterfactual class y' that does not correspond to the predicted class can be calculated as follows:

$$\begin{gathered} y' \in \{c | c \in C \cap c \neq y\} \\ A_i^{(y')}(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F^{(y')}(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \end{gathered} \quad (4)$$

Where C denotes the entire class set. We define the importance of each feature to the predicted class as the base attribute, and the importance of each feature to the counterfactual classes as the adversarial attribute. The base attribute and adversarial attribute calculated in the equation 3 and 4 are the results of deriving the importance of each feature for each class. Now, we can calculate the difference between them, and derive the discriminative attribute D as follows:

$$D_i^{(y,y')}(x) = max(A_i^{(y)}(x) - A_i^{(y')}(x), 0) \quad (5)$$

The discriminative attribute calculated through the equation 5 corresponds to the key attribute for the predicted class by removing the confusing factor between the predicted class and the counterfactual class. In the case of an attribute having a negative value in the adversarial attribute, it is the opposite attribute in predicting the class y', so it is not helpful at all in calculating the discriminative attribute for the class y. Therefore, in this work, an additional operation is performed by modifying the value of a negative attribute to 0 among adversarial attributes.

2.3. Weight Calculation using Class Confusion

The output of the deep neural network F contains clear information about which class a specific input x can be classified into. That is, if a high value of the output of $F^{(c)}(x)$ is derived for a class c that does not correspond to the predicted class, it can be interpreted as meaning that a strong confusion was detected between counterfactual class c and the predicted class y for the input x. In other words, model F has difficulty distinguishing between classes y and c. Therefore, we calculate the local confusion p between classes using the output of model F through the following formula:

$$\begin{gathered} z = [F^{(1)}(x), F^{(2)}(x), \cdots, F^{(C)}(x)] \\ p^{(y')}(x) = \frac{e^{z^{(y')}/t}}{\sum_{c \in C - \{y\}} e^{z^{(c)}/t}} \end{gathered} \quad (6)$$

We calculate the local confusion between the predicted class and the counterfactual classes through the Boltzmann distribution function because the range of the model output z is $\mathbb{R}$. Therefore, the total sum of the local confusion elements becomes 1 while limiting the range to $[0,1]$. The calculated local confusion can be used as a weight for each discriminative attribute by the following formula:

$$D_i^{(y)}(x) = \sum_{c \in C - \{y\}} p^{(c)}(x) \times D_i^{(y,c)}(x) \quad (7)$$

The discriminative attribute derived through the preceding process removes the confusing factor between the predicted class and other counterfactual classes. If there is no confusion between the predicted class and other classes for the input x, the discriminative attribute may not have a significant effect. To solve this problem, we define global confusion q, and the final attribute formula applying global confusion is as follows:

$$z' = [z^{(y)}, \sum_{c \in C - \{y\}} z^{(c)}]$$
$$q^{(y)}(x) = \frac{e^{(z'_1/t)} - e^{(z'_2/t)}}{e^{(z'_1/t)} + e^{(z'_2/t)}} \quad (8)$$
$$DiscGrad_i^{(y)} = q^{(y)}(x) \times A_i^{(y)}(x) + (1 - q^{(y)}(x)) \times D_i^{(y)}(x)$$

We define the weighted summed attribute value as DiscGrad, which stands for Discriminative Gradients. DiscGrad is an expression calculated in consideration of inter-class confusion with respect to input x. If inter-class confusion is low, the amount of reflected discriminative attributes may be small.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets We conduct experiments on three image datasets: FashionMNIST [16], CIFAR-10, CIFAR-100 [17] and four text datasets: AG's News, DBPedia, YahooAnswers [18], and 20 News Groups [19]. Also, we use pre-trained models for experiments. For image data, the Resnet-50 [20] is used, and for text data, the BERT [21] is used. The test accuracy for the image datasets is 93.7%, 94.4%, and 72.6%, respectively, and the test accuracy for the text datasets is 93.4%, 99.1%, 76.0%, and 81.8%, respectively[1].

Evaluation We evaluate the effectiveness of estimating word and pixel importance by the decrease of the golden class logit. More specifically, we measure the decrease of the golden class logit when perturbing a set of important features that are of top-most feature importance. The more classification performance degrades, the more important the feature is [14]. In this study, n features with high attribute values are derived through the existing method IG and our method. Then, after removing common features among the derived features, only the remaining features are masked, respectively, and how much the golden class logit of the classifier decreases is compared. We set n as 50 for image datasets, and 5 for text datasets.

3.2. Comparison with the Integrated Gradients

In this section, we show the results of evaluation by measuring golden class logit decrease when perturbing important features. Table 1 shows the performance results for the image datasets and Table 2 shows the performance results for the text datasets. First of all, we can observe that our method outperforms in both image and text. In the case of CIFAR-10, the performance difference between IG and DiscGrad is +0.0357. For CIFAR-100, the performance difference is +0.0581 and for FashionMNIST, the performance difference is +0.0024.

	CIFAR-10	CIFAR-100	FashionMNIST
IG (1)	-0.0469	-0.5569	-0.0748
DiscGrad (2)	-0.0826	-0.6150	-0.0772
(1)-(2)	+0.0357	+0.0581	+0.0024

Table 1. **The Performance Results for the Image Datasets.**

In the case of AG's News, the performance difference is +0.0207, and +0.0444 for 20 News Groups. Also, the performance difference is +0.0765 for DBPedia, and +0.0171 for YahooAnswers. These experimental results demonstrate that our method derives much more essential attributes than IG.

	AG's News	20 News groups	DBPedia	Yahoo Answers
IG (1)	-0.0155	-0.0408	+0.0408	-0.0080
DiscGrad (2)	-0.0362	-0.0853	-0.0358	-0.0250
(1)-(2)	+0.0207	+0.0444	+0.0765	+0.0171

Table 2. **The Performance Results for the Text Datasets.**

3.3. Visualization of the Discriminative attributes

To show the effect of the discriminative attribute, we visualize the results of discirminative attributes. Figure 1 shows the results of visualizing IG and our method. Figure 1-(a) shows the results of CIFAR-100 dataset, and Figure 1-(b) shows the results of AG's News dataset. As the result of the experiment, we can find that our method derived discriminative attributes better than IG. In particular, we can observe that the attribute of the confusing class, which had the highest logit value, was effectively removed.

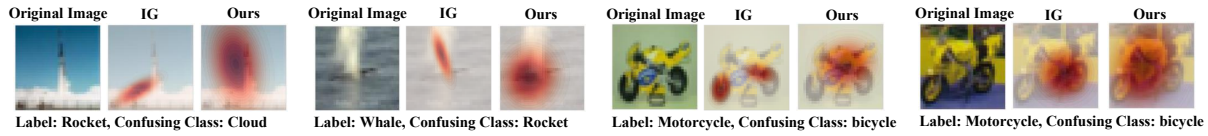
In Figure 1-(a), the first example shows that the attribute for the *"cloud"*, which is the counterfactual class, was well removed in deriving the discriminative attribute for the class *"Rocket"*. In the third example, because wheels are a common property of *"motorcycles"* and *"bicycles"*, our method highlights the body of the motorcycle rather than the wheels.

In Figure 1-(b), in the case of the first example, the relative importance of the words *"diplomatic"* and *"cyprus"* increased by removing the attribute of the counterfactual class *"Business"*. The second example shows that the importance of the words *"market"* and *"products"* has decreased while removing the attribute of *"Business"*, which is a counterfactual class.

3.4. Class Correlation and Discriminative Attribute

We examine the relationship between class correlation and the performance of discriminative attribute to demonstrate the effect of global confusion. We first derive a confusion matrix from the CIFAR-100 dataset. After that, we select 10 classes respectively to define two groups of class sets: high confusion class set and low confusion class set. Next, after calculating

[1]The accuracy of the models used in this study is slightly lower than that of state-of-the-arts, which is not a big problem because our method utilizes confusion between classes.



(a) CIFAR-100

Label : World **Confusing Class : Business**

(IG) EU sets date for Turkey talks , demands concession on Cyprus european union leaders offered to start membership talks with Turkey next Oct . 3 , as long as the turkish government ends its diplomatic standoff with historic rival Cyprus .

(Ours) EU sets date for Turkey talks , demands concession on Cyprus european union leaders offered to start membership talks with Turkey next Oct . 3 , as long as the turkish government ends its diplomatic standoff with historic rival Cyprus .

--

Label : Science/Technology **Confusing Class : Business**

(IG) Dell takes another cut at blade market the biggest danger to HP and IBM is a price war , said John Enck of Gartner . Blades are still premium - priced products from IBM and HP .

(Ours) Dell takes another cut at blade market the biggest danger to HP and IBM is a price war , said John Enck of Gartner . Blades are still premium - priced products from IBM and HP .

(b) AG's News

Fig. 1. **Visualization of the Discriminative Attributes for CIFAR-100 and AG's News datasets.** In this figure, we derived important features through our method and IG, and highlighted them in red. In Figure 1-(b), we marked the words with large differences between our method and IG in blue.

the discriminative attribute using these two class sets, we performed the same experiment as in Chapter 3.2. As the result, the high confusion class set showed +0.050, and the low confusion class set showed -0.004 for the difference between IG and discriminative attribute. This experimental result indicates that discriminative attribute is more effective as the correlation between classes in the dataset is higher. Therefore, it is reasonable to use the degree of confusion as a weight to control the application of the total discriminative attribute.

3.5. Parameter Search for Temperature of the Weights

The shape of the weight distribution can be changed by adjusting the temperature of the Boltzmann distribution when calculating the local confusion and global confusion. Figure 2 shows the result of calculating the difference in golden class logit reduction of IG and DiscGrad while controlling the temperature of local confusion and global confusion. Temperature 1 is a parameter of global confusion, and Temperature 2 is a parameter of local confusion. For example, CIFAR-100 shows the best performance at temperature 20.0 for global confusion. It indicates that DiscGrad gives a considerably large weight to discriminative attribute. Also, CIFAR-100 showed the best performance at temperature 1.0 for local confusion. This shows that the data of CIFAR-100 is usually confused with a small number of counterfactual classes. For image datasets, CIFAR-10 uses 0.1 and 2.0, and FashionMNIST uses 10.0 and 0.1 as temperatures for global and local confusion. In the case of text datasets, AG's News uses 0.5 and 0.1, DBPedia uses 0.1 and 1.0, Yahoo Answers uses 0.1 and 0.5, and 20 News Groups uses 0.1 and 2.0.

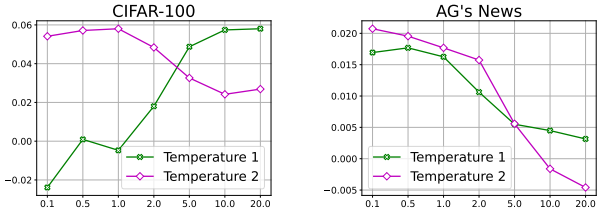


Fig. 2. **Temperature search results for Local Confusion and Global Confusion.** X-axis corresponds to temperature and Y-axis corresponds to the difference in golden logit reduction between IG and DiscGrad (IG-DiscGrad). The higher the difference, the better our method performs.

4. CONCLUSION

In this paper, we propose Discriminative Gradients (DiscGrad) which can derive discriminative attribute for the text and image datasets. We calculate the degree of confusion with the counterfactual classes, and use them as weights for each discriminative attribute. We demonstrate that DiscGrad outperforms the baseline approach for perturbation-based evaluation on widely-used datasets. In addition, we examine the relationship between class correlation and the performance of discriminative attribute to demonstrate the effectiveness of our method. Our method also has enormous value in that it does not require additional data and training for the interpretation model.

Acknowledgement

This work was supported by AIRS Company in Hyundai Motor Company & Kia Motors Corporation through HMC/KIA-SNU AI Consortium Fund.

5. REFERENCES

[1] Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, Bo Zhang, and Tat-Seng Chua, "Disease inference from health-related questions via sparse deep learning," *IEEE Transactions on knowledge and Data Engineering*, vol. 27, no. 8, pp. 2107–2119, 2015.

[2] Thai T Pham and Yuanyuan Shen, "A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform," *arXiv preprint arXiv:1706.02795*, 2017.

[3] Eunsuk Chong, Chulwoo Han, and Frank C Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Systems with Applications*, vol. 83, pp. 187–205, 2017.

[4] Jarmo Lundén and Visa Koivunen, "Deep learning for hrrp-based target recognition in multistatic radar systems," in *2016 IEEE Radar Conference (RadarConf)*. IEEE, 2016, pp. 1–6.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.

[7] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[8] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," *Nature Machine Intelligence*, pp. 1–12, 2021.

[9] 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*, 2016, pp. 2921–2929.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[11] 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*, vol. 10, no. 7, pp. e0130140, 2015.

[12] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks*. Springer, 2016, pp. 63–71.

[13] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3145–3153.

[14] Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael R Lyu, and Shuming Shi, "Towards understanding neural machine translation with word importance," *arXiv preprint arXiv:1909.00326*, 2019.

[15] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, pp. 841, 2017.

[16] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[17] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[18] Xiang Zhang, Junbo Zhao, and Yann LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, pp. 649–657, 2015.

[19] Ken Lang, "Newsweeder: Learning to filter netnews," in *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.