Simple Interpretation-Based Data Augmentation Method for Mitigating Shortcut Learning

Zizhe CHEN
Chinese University of Hong Kong
1155173938@link.cuhk.edu.hk

Abstract

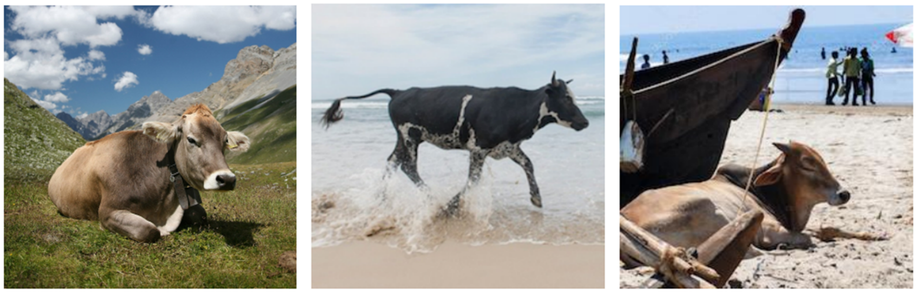
In spite of the remarkable success achieved by deep learning, it is important to acknowledge the presence of certain weaknesses, some of which have the potential to significantly impact performance. This report delves into one of the fundamental issues associated with deep learning, namely shortcut learning. Existing approaches for identifying and mitigating shortcut learning often necessitate prior knowledge of sensitive features, as well as additional training and data collection procedures, which can be both arduous and costly. Consequently, we propose the Interpretation-Based Data Augmentation Method (IBDAM) as an alternative approach to address shortcut learning in neural networks. IBDAM offers a simple and intuitive solution that does not require any prior knowledge or additional data collection. By employing IBDAM on meticulously constructed test examples that embody severe shortcut learning, we observe a notable enhancement in prediction accuracy ranging from 10% to 20%. Moreover, we extend the application of IBDAM to a chest X-Ray (CXR) dataset and implement the shorT method to evaluate its performance. Lastly, we conduct an analysis of the limitations inherent in IBDAM and outline potential avenues for future improvements.

1 Introduction

The advent of deep learning has precipitated the current surge in artificial intelligence, accompanied by a plethora of notable successes. However, beneath the veneer of achievement lie several weaknesses and limitations. This paper directs its focus toward one such weakness inherent in deep learning: shortcut learning. Shortcut learning pertains to the phenomenon wherein a neural network model attains high accuracy on independent and identically distributed (i.i.d) datasets by inadvertently learning irrelevant, sensitive features[3]. Figure 1 elucidates this concept through a tangible example in an object recognition task. In the initial image, a cow is depicted in a pasture, correctly identified by the neural network with a probability of 0.99. Conversely, in subsequent images featuring cows on a beach, the neural network fails to recognize the cows with comparable certainty. This failure stems from the network's reliance not on inherent cow features, but rather on contextual elements such as the presence of a pasture. This essence of shortcut learning is encapsulated in the schematic depicted in Figure 2. Here, the cow serves as a condition, the presence of which influences the image's content. The pasture and beach, on the other hand, represent attributes that influence both the image and the condition. Ideally, the prediction of the condition (i.e., the presence of a cow) should be solely contingent upon its intrinsic features within the image. However, the neural network's prediction is unduly influenced by extraneous attributes, thereby exemplifying the essence of shortcut learning.

Shortcut learning manifests across various domains within deep learning, including computer vision, natural language processing, and healthcare[3]. While models employing shortcut learning often achieve high accuracy on independent and identically distributed (i.i.d.) datasets, such as cows in pastures, their performance markedly deteriorates when tested on out-of-distribution (o.o.d.) datasets,

Preprint. Under review.



(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

[3]

Figure 1: A concrete example of shortcut learning in object recognization task.

such as cows on beaches[3]. Consequently, shortcut learning poses a significant threat to the overall efficacy of numerous learning tasks, including classification, contrastive learning, and few-shot learning[1,14,9]. It arises from dataset bias, and can itself be interpreted as a form of bias (for brevity, "bias" and "shortcut learning" are used interchangeably herein). Past research has introduced several effective methods for detecting and mitigating shortcut learning. Supervised approaches leverage prior knowledge of sensitive features associated with shortcut learning, necessitating additional training and data collection[3,1]. Alternatively, higher-level automated methods involve training additional adversary networks or implementing modifications to high-level features[14,11]. These methodologies are elaborated upon in Section 2.

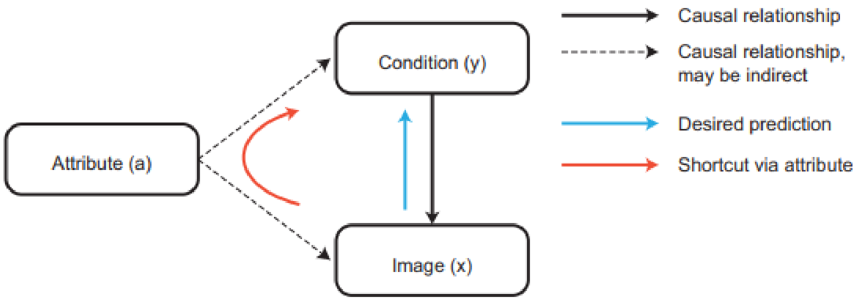
Presently, there exists a gap in research concerning the implementation and evaluation of simpler and more intuitive methodologies for mitigating shortcut learning automatically. In response, this project endeavors to introduce a straightforward, automated, and versatile approach to address shortcut learning. Given the recognized efficacy of data augmentation in enhancing the generalization of deep learning models[16], coupled with the insight provided by certain interpretation methods into neural network prediction mechanisms[15] [13], it becomes conceivable to synergize these fundamental concepts to impede the acquisition of shortcut features by neural networks. Accordingly, we propose an Interpretation-Based Data Augmentation Method (IBDAM) designed to automatically counteract shortcut learning phenomena. Remarkably, this method obviates the need for a priori knowledge of sensitive features, out-of-distribution data collection, or excessive computational overhead. Its simplicity and intuitive nature render it readily adaptable and modifiable for diverse tasks.

Contribution of this report:

1. Propose, describe, and implement the IBDAM method.
2. Design theoretical experiment with MNIST dataset and practical experiment with CIFAR10 dataset to verify the effectiveness of IBDAM.
3. Extend the IBDAM method to chest X-Ray dataset (CXR) and implement shorT in Pytorch to check the performance of IBDAM. Analysis the limitation and propose future improvement direction.

2 Related Work

Shortcut learning: A body of prior work has delved into the features that induce shortcut learning, as well as the detection and remediation thereof. By introducing various forms of noise and features to images within training datasets, one paper[10] demonstrates that unlike human vision, which exhibits a tendency to bias towards the overall shape of objects, neural networks do not exhibit such biases. Instead, they demonstrate a robust capability to learn and exploit any correlated features or noise within a given category. Building upon the insights of this research[10], I conducted further experiments to elucidate the behavior of neural networks. Addressing the issue of shortcut learning, this paper[3] proposed a method involving the construction of out-of-distribution datasets, which



1

Figure 2: A schematic of shortcut learning.

necessitates prior knowledge of sensitive features and data collection efforts. Similarly, another paper[1] introduced a multitask learning approach to quantify and regulate the impact of shortcut learning, albeit requiring prior knowledge of sensitive features and rigorous evaluation through model training. Given the implicit nature of features in images, manual analysis of every attribute proves prohibitively costly. Consequently, automated methods offer promise; for instance, one paper[11] trained an additional 'lens' network to effect subtle alterations to images, thereby eliminating shortcut features in self-supervised learning, while another paper[14] proposed implicit feature modifications to suppress shortcut solutions in contrastive learning. Reflecting upon these methodologies, it becomes apparent that the straightforward and intuitive concept underlying IBDAM remains unexplored in prior research. Thus, I propose the adoption of IBDAM and its application across various datasets to assess its efficacy.

Interpretation: Interpretation methods play a crucial role in generating insights from model predictions by producing heatmaps that highlight regions most pertinent to the prediction. Presently, a variety of interpretation methods are available. Randomized Input Sampling for Explanation (RISE) stands as a versatile interpretation approach applicable to any black-box model. RISE employs Monte Carlo sampling to mask a predetermined number of pixels and calculates the anticipated change in prediction value, thus deriving the importance of each pixel[13]. Despite its broad applicability, RISE entails significant computational overhead due to the necessity of multiple model inferences for a single interpretation, rendering it unsuitable for integration within IBDAM. Among contemporary interpretation methods, Grad-CAM enjoys prominence. Grad-CAM elucidates model predictions by backpropagating predictions to the neural network's last layer and computing the importance of each region via a weighted sum of channel activations[15]. Given that only partial backpropagation is requisite, Grad-CAM substantially reduces the computational burden of IBDAM. Additionally, numerous interpretation methods, such as Eigen-CAM and LayerCAM, which are based on Grad-CAM, exhibit varying performance across images and tasks[12 6]. In this report, we implement all these Grad-CAM based method with a open-source github repository[4].These methods will undergo application and evaluation within the framework of IBDAM.

Data augmentation: IBDAM employs data augmentation techniques to distort shortcut features, rendering them less readily learnable by neural networks. Given that positional shortcuts can effectively be addressed through random cropping of images[3], IBDAM concentrates on augmenting pixel values. Among the myriad augmentation methods available, Random Erasing and Cutout stand out as intuitive choices[19 2], exhibiting promising performance in experimental settings. Notably, the concept underlying IBDAM bears resemblance to interpretation-based cutout as proposed by the paper that proposed Cutout[2], albeit their exploration was confined to i.i.d. datasets without delving deeper into its implications. Similarly, one paper[18] advocate for Interpretability-Mask, a method distinct from IBDAM, which focuses on masking out background elements irrelevant to the object of interest. Furthermore, in an endeavor to regulate the extent of feature distortion, IBDAM incorporates experimentation with Gaussian Patch techniques[8]; however, empirical findings reveal that, under clean data conditions, Cutout outperforms Gaussian Patch.

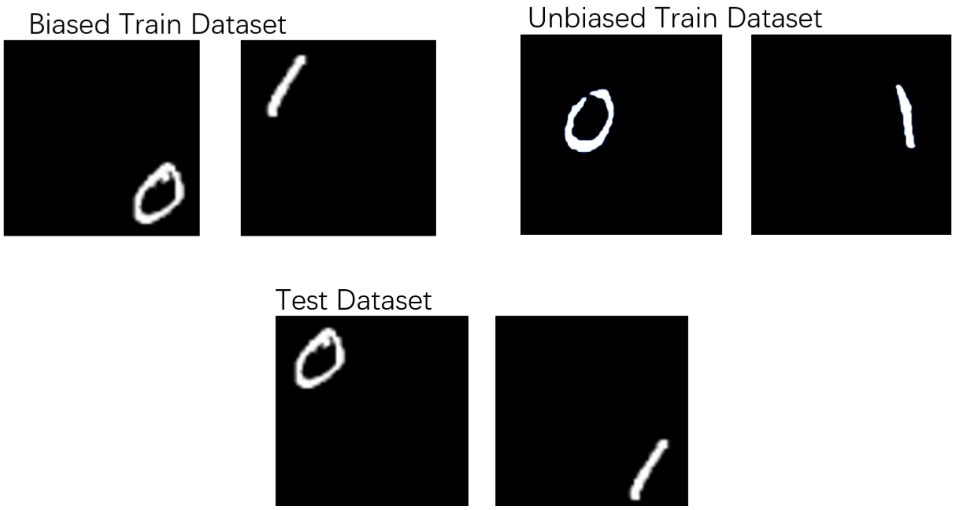


Figure 3: The biased train dataset.

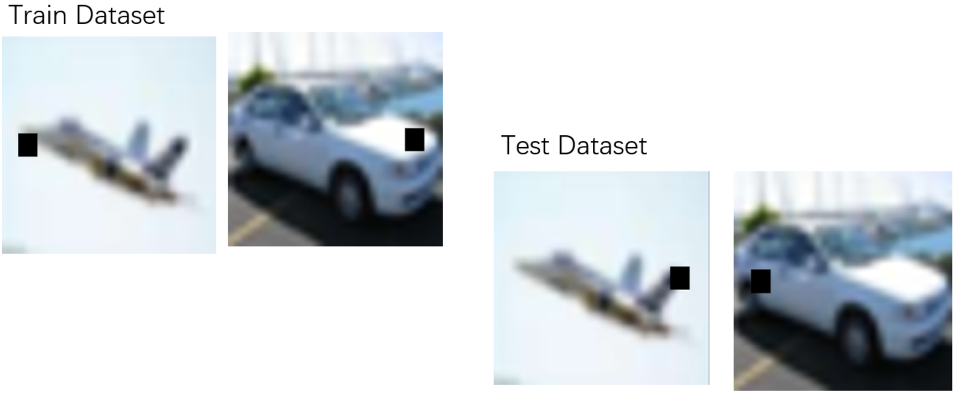


Figure 4: The ideal shortcut sub-dataset. The black patch on the images of train dataset is shortcut that manually added. In the test dataset, the location of black patch is opposite to the train dataset.

3 Data

There are three datasets, which are chosen according to different image types and difficulty, for the test, experiment, and evaluation of IBDAM. The first dataset is MNIST, which consists of 60000 handwritten numbers. It will be used to create 3 different sub-datasets shown in Figure 3. The first sub-dataset is the biased train dataset. It contains only two classes with "0" image as one class and "1" image as another class. For the "0" class, the number 0 is always in the bottom-right corner. For the "1" class, the number 1 is always in the top-left corner. The second sub-dataset is unbiased train dataset , in which there are also "0" class and "1" class, and the position of number 0 and 1 is randomly distributed in the image. The third sub-dataset is test dataset, in which the position of number 0 and 1 is opposite to the biased train dataset: The number 0 is always in the top-left corner and the number 1 is always in the bottom-right corner.

The second dataset is CIFAR-10. Since the train data and test data of CIFAR-10 are sampled from the i.i.d. dataset, it is used to preliminarily implement and experiment IBDAM in the i.i.d. dataset. In addition, to evaluate the performance of the IBDAM method, we also sample some images from the CIFAR-10 dataset and construct ideal shortcut sub-datasets, for example, as Figure 4 shows, containing only two classes of plane and car, one sub-dataset adds a black patch to the images and different patch position in different location of different classes is the shortcut for neural network. In the test dataset of this sub-dataset, the position of the black patch in opposite to the train dataset, so we can measure the degree of shortcut learning by evaluate the test loss and accuracy.

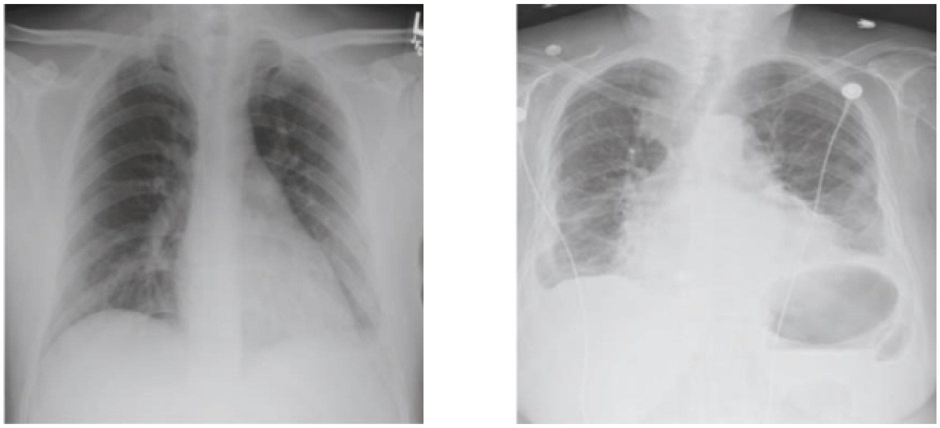


Figure 5: Two image samples from the NIH CXR dataset.

The third dataset that we use is the NIH Chest X-RAY(CXR) Dataset from the NIH Clinical Center (source: https://nihcc.app.box.com/v/ChestXray-NIHCC), which is comprised of 112120 X-ray images with disease labels from 30805 unique patients. One example of NIH CXR is Figure 2. Current research shows that the original NIH CXR dataset has bias on age, which can trigger shortcut learning, and more tricky shortcut learning on patient's age can be triggered by sample distribution with larger age bias from the dataset[1]. We will extend the IBDAM method to this dataset and use the shorT method[1] to evaluate its performance. Based on the evaluation result from NIH CXR dataset, we further discuss the limitation as well as future improvement of IBDAM method.

4 Approach

The conceptual framework of IBDAM is inherently intuitive, as illustrated in Figure 6. This methodology comprises three distinct steps. Firstly, training commences with basic data augmentation, a prerequisite for valid interpretation of model predictions. The model undergoes training on the dataset augmented solely through Horizontal Flip, Random Crop, and Normalization for 100 epochs or until convergence. Subsequently, this basic augmentation is consistently applied unless otherwise specified in subsequent steps. Upon completion of training, the model generates predictions utilized by Grad-CAM for interpretation. For each image within the training dataset (sans basic augmentation), Grad-CAM is employed to produce a corresponding heatmap devoid of image overlay (refer to Figure 7), which is then stored. The additional time cost incurred in this process equates to that of training for a single epoch. Furthermore, the storage space requirement is markedly reduced, ranging from 20 to 100 times smaller than the original dataset, owing to the predominantly monochromatic nature of the majority of heatmaps. Moreover, through proper encoding of Grad-CAM output, storage costs can be further minimized. Moving on to the second step, training resumes utilizing weights obtained from the initial training phase for an additional 75 epochs or until convergence. During this phase, both training data and corresponding heatmaps are employed. Varied intensity augmentation policies are applied based on distinct region highlights identified by Grad-CAM. Notably, among several augmentation policies explored, Cutout exhibits superior performance, and thus, only results employing Cutout will be presented in Section 5 for clarity. The additional time cost incurred in the third step is contingent upon image size and specific augmentation policies; for instance, with image dimensions of 224x224 and Erasing augmentation policy, the incremental time cost ranges from 5 to 10 seconds. It is imperative to highlight that, to ensure comprehensive learning of alternative features, heatmaps remain unchanged throughout step 3. Subsequent to step 2, additional augmentation ceases, and training continues for 25 epochs or until convergence in step 3. Detailed rationale for the inclusion of steps 2 and 3 in IBDAM is expounded upon in Section 5.

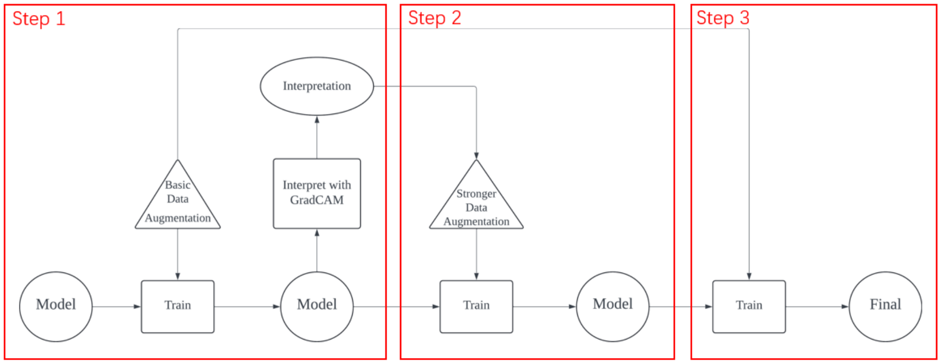


Figure 6: A schematic of the IBDAM method. The 'Model' always refers to the trained model from last step, and the arrow can be interpreted as 'apply to' relation.

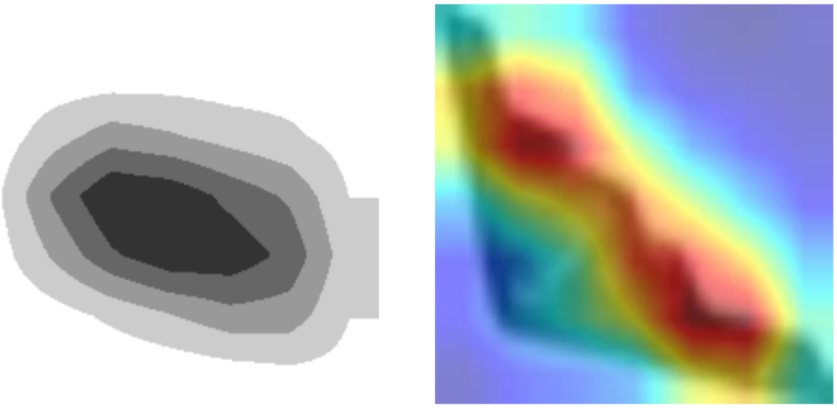


Figure 7: The pure heatmap compares with heatmap that overlaps with the image.

5 Experiments

5.1 Experiments on MNIST

In this section, we will show some important experiment result. All the model used in experiment is 18-layer Residual Net[5] implemented in Pytorch. We use SGD optimizer with default settings. The batch size is 64 and the training stop when the train loss and accuracy is converge. To ensure consistence between different dataset, images from all the dataset are resampled to size 224*224.

As we promise in Section 4, now, we show some experiments of the behavior of neural networks on the MNIST dataset to give intuition and reason of step 2 and step 3 of IBDAM. First experiment is Trigger experiment, the model is trained on the biased train dataset and test on the test dataset demonstrated in Figure 3 to see whether the shortcut learning of position can be triggered. We repeat this experiment for ten times. Each time, when the train loss converges to 0, with the test accuracy always approaches 0, the result is that the shortcut learning is always triggered.

Second is the Recover experiment. The model is trained with biased train dataset and then trained on the unbiased train dataset with random position of 1 and 0 in Figure 3. The goal is to see whether the model can recover from shortcut learning and converge to the unbiased dataset, which is decided by test accuracy on test dataset of Figure 3. With final test accuracy always approaches 100, the result is that model can recover from the shortcut learning.

Third is the Corrupt experiment. The model is trained with unbiased train dataset with random position of 1 and 0 and then trained with the biased train dataset. The goal is to see whether the model without shortcut learning will be corrupted by the dataset with obvious shortcut learning pattern,

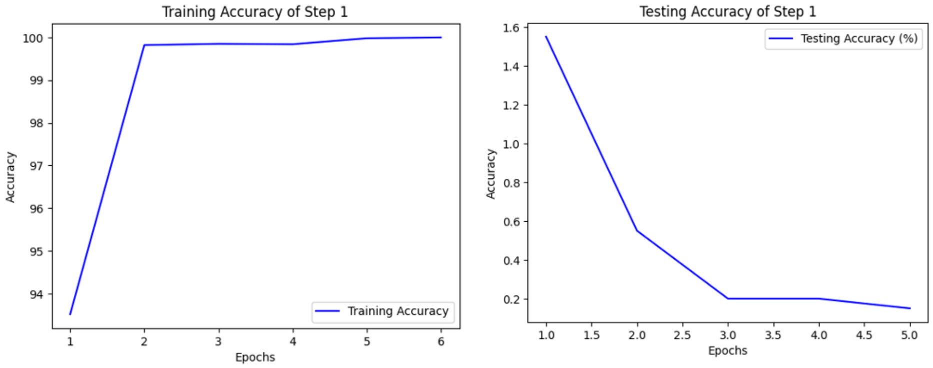


Figure 8: The plot of test accuracy and train accuracy in step 1.

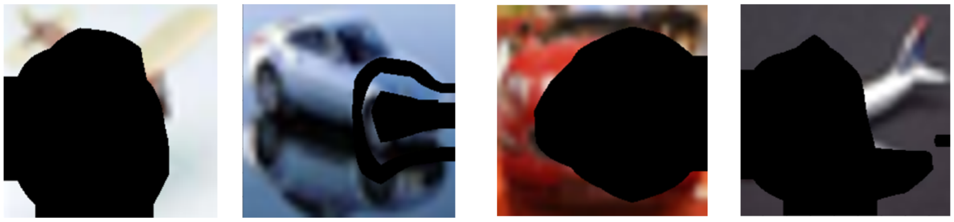


Figure 9: The images after being applied with stronger data augmentation in step 2.

which is decided by the test accuracy on test dataset. After being trained on biased train dataset for more than 50 epochs, the test accuracy is still over 90. Therefore, non-biased will not be easily corrupted by the shortcut learning pattern.

Although the above three experiments are simple, they reveal an important insight of the behavior of neural network while facing shortcut learning: biased model can recover from shortcut learning, but non-biased model will not be corrupted by shortcut learning pattern. We assume that neural network can roughly follow this rule in all other tasks. This is the intuition behind the step 2 and step 3 of IBDAM. Since biased model can recover from shortcut learning, in theory, in the step 3, where the extra data augmentation is removed, model can recover from any shortcut learning introduced by extra data augmentation of step 2. Since non-biased model cannot be corrupted by shortcut learning pattern, the shortcut learning corrected by step 2 cannot be reintroduced in step 3.

5.2 Experiments on Ideal Shortcut Dataset

Then, with the insight revealed, We will show the real pipeline of applying IBDAM through the following experiment on ideal shortcut sub-dataset sampled from CIFAR10 shown in Figure 4. In step 1, we train the model with this ideal shortcut sub-dataset. Figure 8 shows the plot of test accuracy and train accuracy during model training in step 1. When the train accuracy converge to 100%, the test accuracy approach to 0%, which means severe shortcut learning has happened during training.

In step 2, we apply stronger data augmentation based on the interpretation. For simplification, we apply Cutout as the stronger data augmentation. The images after being applied with stronger data augmentation is shown in Figure 9. Obviously, the feature that neural network depends on to perform shortcut learning, the black patch, is blocked and polluted by the stronger data augmentation. The training result of step 2 is shown in Figure 10. With the train accuracy converges to 95%, the test accuracy is around 50%. Although the test accuracy is even no better than the accuracy of random guess, at least compared with the 0% test accuracy in step 1, there is great improvement, which means that neural network can learn some desired feature after using stronger data augmentation method to block the shortcut learning label.

In step 3, we remove the stronger data augmentation and train the model with basic data augmentation until converge. Figure 11 show the plot of train accuracy and test accuracy. With train accuracy

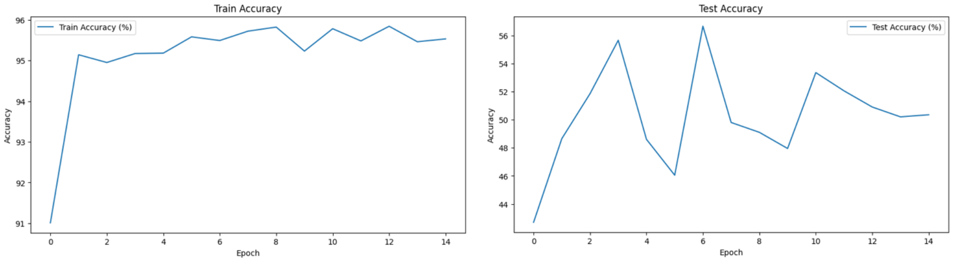


Figure 10: The plot of test accuracy and train accuracy in step 2.

Table 1: Effectiveness of Different Data Augmentation Method on Ideal Shortcut Dataset

Methods	Cutout	AutoAugmentation	Crop	IBDAM-gaussian	IBDAM-Cutout
Train Accuracy(%)	100	100	100	100	100
Test Accuracy(%)	0.75	0.5	0.85	**17.75**	**25**

converge to 100%, the test accuracy is around 25%. Although the accuracy is even worse than the accuracy of random guess, at least compared with the 0% accuracy in step 1 and compare with the result of applying other data augmentation techinques in Table 1, the accuracy is improved and the shortcut learning is mitigated. It is also important that the 25% test accuracy means neural network does not experience the catastrophic forgetting[10] and it still remembers some useful features learned in the step 2. I believe that with well design data augmentation methods, the performance of this IBDAM has great potential for improvement.

5.3 Experiments on Unbiased Dataset

Besides mitigating shortcut learning, IBDAM also has a good property that, if it is applied to dataset and model without bias and shortcut, it will not affect the accuracy of the model. We experiment this with a model trained on the original CIFAR10 dataset and the result is shown in the Figure 12. The first step is from epoch 0 to 100 and it is a normal training process. In step 2, the stronger data augmentation is applied and the performance of model inevitably decreases because large amount of features is blocked. When it comes to step 3, as soon as the stronger data augmentation is removed, the model retains the performance in step 1 and sometimes obtain better performance as Table 2 show.

5.4 Experiments on NIH CXR Dataset and Limitations

Furthermore, we extend the application of IBDAM to the NIH CXR dataset, which encompasses shortcut learning related to age attributes, aiming to assess its performance in a distinct dataset context. However, unlike the controlled environment of the Ideal Shortcut Dataset explored in Section 5.2, the shortcut learning associated with age within the NIH CXR dataset is implicit, necessitating quantitative metrics rather than human observation for assessment. To evaluate performance, we implement the shorT method in PyTorch due to unavailability of its source code from the authors,

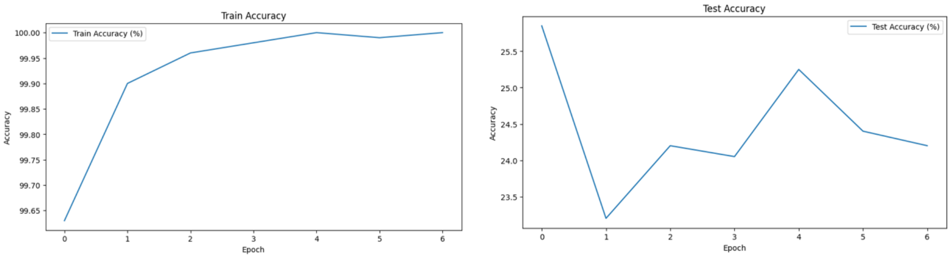


Figure 11: The plot of test accuracy and train accuracy in step 3.

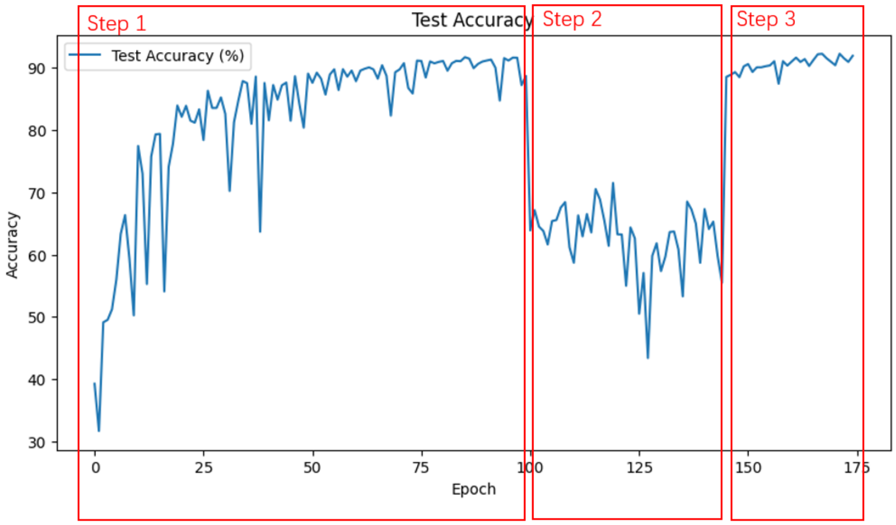


Figure 12: Train model with IBDAM on unbiased dataset CIFAR10.

Table 2: Effectiveness of Different Data Augmentation Method on Unbiased Dataset

Methods	Cutout	AutoAugmentation	Crop	IBDAM-gaussian	IBDAM-Cutout
Test Accuracy(%)	93.38	92.71	91.64	**92.22**	**93.63**

achieving results consistent with those reported in the original paper[1]. Following the framework of the shorT method, exemplified in Table 3, we employ the Mean Absolute Error (MAE) loss pertaining to age during transfer learning as a measure of age-related shortcut learning, as it reflects the degree to which the feature extractor encodes age information. As age-related shortcut learning manifests during classification tasks related to 'effusion' and 'no effusion'[1], we partition the dataset by randomly sampling 10,000 images from the 'No Findings' class and 2,000 images from the 'Effusion' class for both training and testing. Subsequently, employing identical settings to prior experiments, we train the model successfully until convergence. However, during interpretation, unexpected phenomena emerge, as depicted in Figure 13. Notably, in Figure 13(b), heatmap highlights focus on the area between the left and right lungs, devoid of significant age-related information, rendering the augmented data in Figure 13(c) ineffective in influencing age attributes. Similarly, Figure 13(d) highlights lung effusion features, yet fails to elucidate implicit age attributes. This phenomenon likely arises from Grad-CAM's inherent limitations as a prediction interpretation tool, wherein minor features contributing to prediction may be overshadowed by primary predictive features[15]. As anticipated, the application of IBDAM demonstrates no efficacy in mitigating age-related shortcut learning, as evidenced in Table 3. Moreover, experimentation with similar methods such as Eigen-CAM and LayerCAM[12 6] yield comparable interpretation outcomes under this scenario.

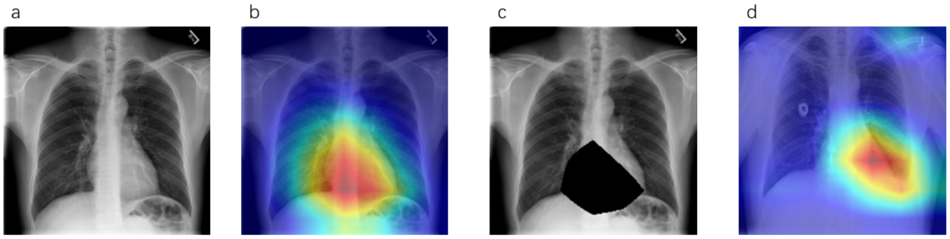


Figure 13: From left to right there are a, b, and c three images. The a image is a sample from train dataset with label "No Findings". The b image is the interpretation result on this image. The c image is the image after applying stronger data augmentation. The d image is the interpretation result on a sample from train dataset with label "Effusion".

Table 3: Mean Absolute Loss of Age Prediction after Transfer Learning on Age (-0.00x scale refers to the scale of age learning in shorT method)

Model	Original Model	-0.001 scale	-0.003 scale	-0.005 scale	IBDAM-Cutout
MAE	11.91	14.48	14.87	15.34	**12.03**

Upon identifying and comprehending the limitations of IBDAM, we promptly embarked on the development of alternative techniques, such as activation and gradient-guided dropout[7], capable of augmenting high-level feature dimensions. Our objective was to enhance performance on the NIH CXR dataset. While initial training endeavors were undertaken, resource constraints in terms of time and energy impeded the attainment of maturity and effectiveness for the new technique by the time of completing this report. However, we will include the results in the submitted code for your reference.

6 Conclusion

This report presents a comprehensive review of the shortcut learning problem, its implications on neural network performance, and existing methodologies aimed at its resolution. Building upon insights gleaned from prior research, we propose a significantly simpler and more intuitive interpretation-based data augmentation method for mitigating shortcut learning. The underlying assumptions and intuitive rationale of our proposed method, IBDAM, are elucidated through experimentation on the MNIST dataset, followed by performance evaluation on the CIFAR10 dataset. Furthermore, we extend the application of IBDAM to the NIH CXR dataset and conduct a thorough analysis of its limitations. Drawing from the findings of these experiments, we identify two primary avenues for future research. Firstly, the exploration of more tailored data augmentation techniques tailored specifically for IBDAM, and their evaluation across a broader spectrum of real-world shortcut datasets. Secondly, the development of methodologies capable of modifying high-level image features, thereby enhancing IBDAM's efficacy in handling implicit features.

References

[1] Brown Alexander, Tomasev Nenad, Freyberg Jan, Liu Yuan, Karthikesalingam Alan, and Schrouff Jessica. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature Communications*, 14(1):4314–4314, 2023. doi: https://doi.org/10.1038/s41467-023-39902-7.

[2] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[3] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. doi: 10.1038/s42256-020-00257-z.

[4] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[5] 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 (CVPR)*, June 2016.

[6] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[7] Rohit Keshari, Richa Singh, and Mayank Vatsa. Guided dropout. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4065–4072, 2019.

[8] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

[9] Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. Rectifying the shortcut learning of background for few-shot learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 13073–13085. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/6cfe0e6127fa25df2a0ef2ae1067d915-Paper.pdf`.

[10] Gaurav Malhotra and Jeffrey Bowers. What a difference a pixel makes: An empirical examination of features used by CNNs for categorisation, 2019. URL `https://openreview.net/forum?id=ByePUo05K7`.

[11] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6927–6937. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/minderer20a.html`.

[12] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

[13] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

[14] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4974–4986. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/27934a1f19d678a1377c257b9a780e80-Paper.pdf`.

[15] 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 (ICCV)*, Oct 2017.

[16] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), Jul 2019. doi: 10.1186/s40537-019-0197-0.

[17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[18] Hao Zhao, Jikai Wang, Zonghai Chen, Shiqi Lin, Peng Bao, and Meng Xu. Interpretability-mask: a label-preserving data augmentation scheme for better classification. *Signal, Image and Video Processing*, 17(6):2799–2808, 2023.

[19] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.