From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Master's Theses, 2022-2026 Computer Science, 2023

2024

Augmenting Concept-Labeled Datasets with Counterfactuals for Interpretability

Sinha, Prachi

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01b2774008t

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Augmenting Concept-Labeled Datasets with Counterfactuals for Interpretability

Prachi Sinha

A master's thesis
presented to the faculty
of princeton university
in candidacy for the degree
of master of science of engineering

recommended for acceptance by
the department of
computer science

May 2024



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Abstract

Concept-based interpretability methods use a pre-defined set of human-understandable concepts to explain a model's output. These explanations are learned by "probing" the model using a concept-labeled dataset and learning how concepts can be linearly combined to predict outputs. However, there are several limitations to current methods. Explanations are highly dependent on the probe dataset, and they may reflect correlations between concepts and outputs, rather than causation. To address these issues, we augment a probe dataset with counterfactual images in which particular semantic concepts have been removed. Using such counterfactual images, we attempt to improve the accuracy of explanations for a scene classification model, and more directly learn the causal relationships between concepts and outputs. We find that explanation accuracy improves in certain cases, but is generally still limited by the probe dataset's alignment with the model, and that counterfactual images do highlight difference in concept importance in the presence and absence of other strongly co-occurring concepts.

Acknowledgements

I would like to thank my advisor, Vikram Ramaswamy, for their continuous support, encouragement, and guidance throughout the process of this thesis. I want to thank Olga Russakovsky and the Visual AI group for being so welcoming and creating an environment that was both kind and inspired me to explore new academic spaces and push my boundaries. Specifically from the Visual AI group, I would like to thank Xinran Liang for being a great collaborator, and Ye Zhu for her insights and taking the time to discuss the best directions to take this project.

I would like to thank my lovely friends at Princeton who have also gone through this journey with me. I am so excited to see you succeed and continue doing so in the future.

I would like to thank my family for their unconditional love and support, for giving me pep talks, dropping off care packages, and always asking about my progress and encouraging me. I could not have made it this far with you.

Contents

Abstract iii

Acknowledgements iv

List of Tables vii

List of Figures viii

1 Introduction **1**

2 Counterfactual Generation **3**

2.1 Approach 3

2.2 Results 4

2.3 Discussion and Future Work 4

3 Explanations **6**

3.1 Approach 6

3.2 Experimental Setup 7

3.3 Results 8

3.4 Discussion and Future Work 8

4 Importance Scores **12**

4.1 Approach 12

4.2 Experimental Setup 13

4.3 Results 13

4.4 Discussion and Future Work 16

5 Conclusion **17**

List of Tables

2.1 The top-20 biased concepts and counterparts as described in Singh et al. [12]. 5

3.1 The accuracy of the Original and Augmented Explanation model across 30 representative scenes. 10

3.2 This table displays the top 10 positive concepts (and their coefficients) which contribute to the prediction of scenes "bus station" and "nursing home" across the Original and Augmented Explanation models. In this example, we see that the biased concept "handbag" drops from the top-10 concepts when moving from the original to augmented explanation model for bus station. In contrast, we see that the rank of hair drier, which is also strongly biased with person, increases when predicting nursing home in the augmented explanation model. 11

4.1 The average importance score and standard deviation for wine glass for scene restaurant. 16

List of Figures

2.1 Two examples of removal-based counterfactual generation using Large Mask Inpainting. On the left, the clock face is removed, and on the right we have a bus being removed. 5

4.1 Suppose (painting, lamp) is a strongly biased pair. For an image containing painting and lamp, we generate counterfactuals in which:
Edit A: "painting" has been removed
Edit B: "lamp" has been removed
Edit C: both "painting" and "lamp" have been removed. 14

4.2 Using the differences in class output score, we get two importance scores for painting, one in the presence of lamp, and one in the absence of lamp. 14

4.3 The importance of keyboard for scene home-office in the presence and absence of mouse. 15

4.4 The importance of wine glass for scene restaurant in the presence and absence of person. 15

Chapter 1

Introduction

We have seen an exponential growth in the applications and capabilities of deep neural networks in computer vision. However, many fields have been slow to adopt these models because in high-risk and highly-regulated industries such as finance and healthcare it is necessary to be able to understand why decisions are made in a human-interpretable manner [4]. Many methods of interpretability have been proposed, including concept-based interpretability. Concept-based methods score a predefined set of semantic concepts to explain a model's decision. These scores are generated by evaluating the model of interest using a probe dataset and correlating the model predictions with the concepts labeled in the probe's images. For example, we would expect that for scene classification, if we are given an image of a bedroom, the concepts "bed" and "nightstand" would have a high, positive score, and "grass" would have a negative score. However, recent work has shown that there are still many limitations to these methods.

Ramaswamy et al. [12] find that the choice of dataset used to probe the model has a significant impact on the explanations, and that in many cases, the semantic concepts used for the explanations are harder for the model to learn than the actual outcomes. This suggests that the concept-based explanations may not be true to the actual decision-making process, and may reflect correlations between concepts and outputs, rather than causation.

Prior work has been done to address this issue of causation vs correlation; however, they have largely focused on tying local regions of the input to the model's output. Ruth et al. [5] use extremal perturbation analysis and feature inversion to elucidate important features for classification. Goyal et al.[7] present a method to identify regions of the image, which when substituted with a similar region from a distractor image, change the classification output.

In this work we use counterfactual images generated through semantic edits of concept-labeled images. We propose two uses of counterfactual image generation to improve explanation models and more directly learn the casual effect of concepts.

First, probe datasets may contain concept biases that negatively influence how explanation models learn the correlations between specific concepts and model outcomes. We can reduce some of these biases by augmenting the probe dataset with counterfactual images that disentangle strongly biased concepts.

Second, we can use counterfactual images to directly learn how the presence or absence of a concept influences the model output.

Chapter 2

Counterfactual Generation

2.1 Approach

In order to reduce bias between concepts in a probe dataset, we want to augment the probe dataset with counterfactual images that disentangle co-occurring concepts. Singh et al. examine the Common Objects in Context (COCO) dataset and identify twenty pairs of concepts which are strongly biased [14]. From here on, given a pair of concepts where $concept_1$ is highly biased with $concept_2$, we will refer to $concept_1$ as a "biased concept" and $concept_2$ as its "counterpart."

To disentangle these biased concepts from their counterparts, we look at images where biased concepts and their counterparts are both present in the image, and generate a counterfactual image that removes the counterpart.

For each of the biased concepts, we use COCO annotations to generate masks that cover every instance of the concept in the image of interest. We then use Large Mask Inpainting (LaMa) [15] to generate images where the masked area is removed.

2.2 Results

The set of top biased concepts as described in Singh et al. [14] are listed in Table 2.1, and examples of counterfactual images generated using LaMa are in 2.1.

2.3 Discussion and Future Work

We did not explicitly evaluate the generated images, and as can be seen in Figure 2.1, the quality of the results can vary greatly depending on the size of the area to be removed, the contextual background, and more. We considered applying existing quantitative metrics such as Fréchet Inception Distance (FID) [9], which measures the similarity between two datasets; however, because we are generating counterfactual images and thus expect the distribution of the generated images to shift from the prior distribution, FID score may not work well in this case. Other methods such as comparing CLIP directional similarity [6] also are limited by the accuracy of a secondary model in recognizing concepts which exist in an image.

Lacking quality of generated images may have contributed to downstream effects on the accuracy of the augmented explainability model and non-linear importance scores. In the future, we may use human evaluation to filter at a high level which removals are more likely to result in quality counterfactuals before introducing them into the following pipeline.

Biased Concept	Counterpart
baseball glove	person
skateboard	person
tennis	person
hair	towel
toaster	ceiling
snowboard	person
remote	person
sports ball	person
clock	building
skis	person
keyboard	mouse
microwave	oven
spoon	bowl
potted plant	vase
bus	road
car	road
apple	fruit
handbag	person
wine glass	person
cup	table

Table 2.1: The top-20 biased concepts and counterparts as described in Singh et al. [12].



Figure 2.1: Two examples of removal-based counterfactual generation using Large Mask Inpainting. On the left, the clock face is removed, and on the right we have a bus being removed.

Chapter 3

Explanations

3.1 Approach

As described in Ramaswamy et al. [13], the classification model we are trying to explain can be decomposed into two functions, $F = g \circ f$. Function f maps the image to a feature space, $f : I \rightarrow \mathbb{R}^N$, and function g maps the features to an class output space, $g : \mathbb{R}^N \rightarrow \mathbb{R}^D$. In order to explain the classification model using human-understandable concepts, we further decompose g into two functions, $g : h_{exp} \circ h_{conc}$. Function h_{conc} maps the feature space to a human-understandable concept space, $h_{conc} : \mathbb{R}^N \rightarrow \mathbb{R}^C$, and function h_{exp} maps the concepts to the class output space, $h_{exp} : \mathbb{R}^C \rightarrow \mathbb{R}^D$. From here, we will refer to F as the "Classification" model, h_{conc} as the "Concept-Recognition" model, and h_{exp} as the "Explanation" model.

In order to evaluate the effect of introducing counterfactual images into the training set of the Explanation model, we train two versions of the Explanation model: "Original", which uses the probe dataset images as-is, and "Augmented" which introduces the counterfactual images into its training set. In both cases, for consistency, we use the Concept-Recognition model to annotate the concepts present in the images for training and testing the Explanation model.

3.2 Experimental Setup

Our experimental set up follows that which is described in Ramaswamy et al. [13] [12].

Our Classification model of interest is a RESNET-18 [8] model trained on the Places365 dataset [16]. The probe dataset we are using is Common Objects in Context (COCO) [11], with "thing" annotations from the original release and additional "stuff" annotations from Caesar et al. [3].

We want the internal representations of the Concept Recognition model to align as closely as possible with the Classification model. The Concept Recognition model is built by replacing the final layer of the Classification model with a new linear layer, and training the model to learn the annotated concepts in each image by optimizing a Binary Cross-Entropy Loss.

The Explanation model is a simple baseline model that is built using sklearn's Logistic Regression. The model input is a vector representing the concept-recognition score for the 171 possible concepts in an image, and the output is the top-predicted scene for that image. For training we use L1 loss, liblinear solver, and tune the regularization strength on the validation set. We evaluate the explanation model's accuracy on a set of 30 classes in which the biased concepts are well-represented. With this baseline explanation model, we are evaluating a "least-faithful" explanation, as described in Ramaswamy et al. [13]. The accuracy is measured by comparing the Explanation model's output to the top-scoring class from the Classification model's output.

When training the Original Explanation model, to ensure the distribution of images is similar across the training sets for Original and Augmented, we duplicate images in the Original Explanation model training set, such that the number of duplicates for an image matches the number of counterfactuals of that image included in the training set of the Augmented Explanation model.

3.3 Results

We find that the accuracy of the Augmented Explanation Model does not significantly improve when compared to the Original Model, and in some cases actually drops as can be seen in Table 3.1. However, in some cases we do see changes in the coefficients for the top-contributing concepts used in the explanation for a scene as can be seen in Table 3.2. In this example, we see that the biased concept "handbag" drops from the top-10 concepts when moving from the Original to Augmented model explanation for bus station. Because handbag is heavily biased with person, it is possible that the coefficient for handbag drops in the augmented explanation because it is more specifically the presence of handbag with person that contributes to predicting bus station. In contrast, we see that the coefficient for hair drier increases when predicting nursing home. Hair drier is also strongly biased with person. In this case, it is possible that we see an increase in coefficient because there are instances of hair drier without person in nursing home that the model is still able to categorize accurately and better attribute to hair drier alone.

3.4 Discussion and Future Work

We find that augmenting the probe set used to train the explanation model has little impact on the Explanation model's accuracy, though it does have some impact on the coefficient scores assigned to each concept. There are several reasons why we may have seen little improvement or even a drop in accuracy with the augmented model. As described in Ramaswamy et al. [12], the explanation model generally does better when the probe dataset is similar to the original dataset the Classification model was trained on. We attempted to neutralize biases in the probe dataset such that it may better align with the original dataset; however, by introducing counterfactual images, we may have in fact pushed the distribution of the probe dataset further away from aligning with the original dataset. Additionally, as mentioned in Section **??**, it is possible that the quality of the images being used for training

the Explanation model may be low which could have a negative impact on the model's ability to map concepts to predictions and learn accurate coefficients.

Going forward, I plan to try augmenting the Explanation model's training set with an alternative type of counterfactual which may be of better quality and closer to the distribution of original images in the dataset. These counterfactuals are generated by replacing biased concept counterparts with contextually similar concepts (rather than simply removing them).

Scene	Orig Accuracy	Aug Accuracy	Aug - Orig Accuracy
dorm_room	0.5724	0.5855	0.0131
nursing_home	0.552	0.504	-0.048
ski_slope	0.9176	0.9275	0.0099
ice_skating_rink/outdoor	0.4505	0.4595	0.009
pizzeria	0.9028	0.9127	0.0099
bakery/shop	0.7674	0.7752	0.0078
street	0.739	0.7287	-0.0103
bus_station/indoor	0.954	0.954	0
kitchen	0.8516	0.8281	-0.0235
athletic_field/outdoor	0.5	0.5172	0.0172
baseball_field	0.8914	0.8914	0
park	0.6598	0.6649	0.0051
crosswalk	0.5923	0.5494	-0.0429
soccer_field	0.7227	0.7545	0.0318
stadium/baseball	0.6574	0.7222	0.0648
florist_shop/indoor	0.8489	0.8273	-0.0216
living_room	0.6607	0.6875	0.0268
home_office	0.756	0.7143	-0.0417
office	0.3981	0.3796	-0.0185
restaurant	0.4865	0.5045	0.018
tower	0.7845	0.7931	0.0086
synagogue/outdoor	0.6061	0.5758	-0.0303
utility_room	0.5632	0.6092	0.046
snowfield	0.1923	0.2308	0.0385
market/outdoor	0.5688	0.5413	-0.0275
ball_pit	0.5733	0.5867	0.0134
bathroom	0.957	0.9385	-0.0185
shower	0.3824	0.3824	0
storage_room	0.4783	0.4348	-0.0435

Table 3.1: The accuracy of the Original and Augmented Explanation model across 30 representative scenes.

	bus_station/indoor		nursing_home	
Rank	**Original**	**Augmented**	**Original**	**Augmented**
1	(wall-panel: 13.2948)	(wall-panel: 13.4147)	(ceiling-tile: 14.3204)	(ceiling-tile: 15.3409)
2	(structural-other: 11.1545)	(bus: 10.2322)	(ground-other: 12.8346)	(ground-other: 10.8276)
3	(bus: 10.6564)	(floor-stone: 5.7802)	(clothes: 8.1273)	(microwave: 7.8644)
4	(floor-stone: 6.0967)	(metal: 4.7318)	(microwave: 6.0204)	(clothes: 7.7512)
5	(floor-tile: 5.7267)	(plant-other: 4.3282)	(scissors: 5.4430)	**(hair drier: 5.8492)**
6	(wall-concrete: 4.8241)	(suitcase: 3.8360)	(paper: 4.6416)	(plastic: 4.5325)
7	(plant-other: 4.2617)	(wall-concrete: 3.6809)	(clouds: 4.3295)	(cupboard: 4.4444)
8	**(handbag: 3.6117)**	(fog: 3.5436)	(laptop: 4.3238)	(house: 4.3089)
9	(metal: 3.5407)	(wall-brick: 3.0728)	**(hair drier: 4.3230)**	(paper: 4.1313)
10	(suitcase: 2.8238)	(ceiling-other: 2.8574)	(plastic: 4.1641)	(clouds: 4.0474)

Table 3.2: This table displays the top 10 positive concepts (and their coefficients) which contribute to the prediction of scenes "bus station" and "nursing home" across the Original and Augmented Explanation models. In this example, we see that the biased concept "handbag" drops from the top-10 concepts when moving from the original to augmented explanation model for bus station. In contrast, we see that the rank of hair drier, which is also strongly biased with person, increases when predicting nursing home in the augmented explanation model.

Chapter 4

Importance Scores

4.1 Approach

Many concept-based interpretability methods rely on using a linear combination of concepts to explain model outputs. [1] [2][10]. However, as we observed when trying to improve the Explanation model, it can be difficult to fully disentangle concepts and in some cases, if the Classification model has learned some strong correlations between concepts, then attempting to learn an accurate coefficient for each concept may not work well. We hypothesize that the assumption of linearity between concepts may hold for pairs of concepts which are unbiased, but for strongly biased concepts, the importance of a concept by itself it different than when it occurs in an image with its counterpart.

We approach this problem using counterfactual generation. For a given image that contains a biased concept and its counterpart, we generate three counterfactual images (1) Remove the biased concept, (2) Remove the counterpart, and (3) Remove the biased concept and counterpart. An example is illustrated in Figure 4.1.

We want to see how the output class score changes with the removal of concepts to directly learn the importance of those concepts (refer to Figure 4.2). Specifically, we are interested in whether the importance of a biased concepts is influenced by the presence or

absence its counterpart.

4.2 Experimental Setup

For each image that contains a strongly biased pair, we generate the three described edits. Taking the class output which is predicted on the original image ($class_i$), we evaluate the Classification model on each of the edits, and find the score for $class_i$. We map the change in score to the change in that concept. We do not assume perfect removal in the edits, and instead apply a threshold such that if the score for a concept (predicted by the Concept-Recognition model) scales down by at least 0.5, then we consider the concept removed. If the concept score does not scale down by 0.5, then we discard that particular set of images as the quality of removal may not be sufficient for calculating importance score.

In the cases where the concept is considered removed, we assign the difference in class score to be the importance of the concept which has been edited.

4.3 Results

We find that for some scenes, biased concepts exhibit different importance scores depending on whether or not their counterpart is present (Figure 4.4), but in other scenes, biased concepts have consistent importance scores (Figure 4.3). We also note; however, that the calculated importance scores (taken as the mean of importance scores for that concept across a given scene) tend to have a high variation. In the example of wine-glass given in Figure 4.4, Table 4.1, the average importance scores are approximately within one standard deviation of one another.

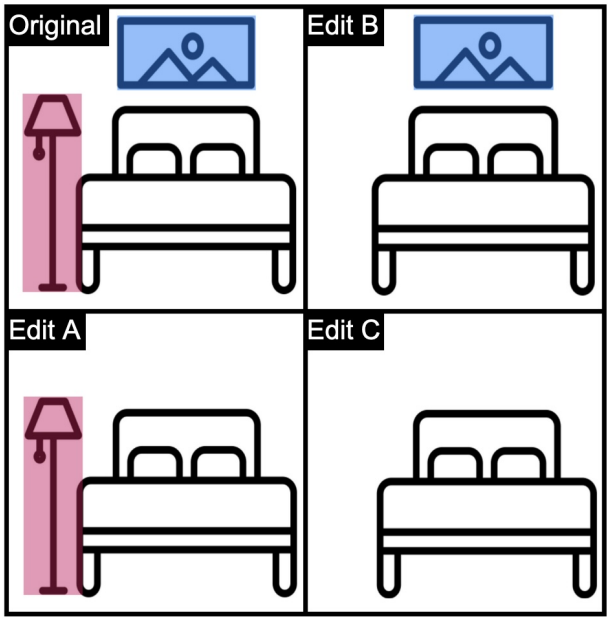


Figure 4.1: Suppose (painting, lamp) is a strongly biased pair. For an image containing painting and lamp, we generate counterfactuals in which:
Edit A: "painting" has been removed
Edit B: "lamp" has been removed
Edit C: both "painting" and "lamp" have been removed.

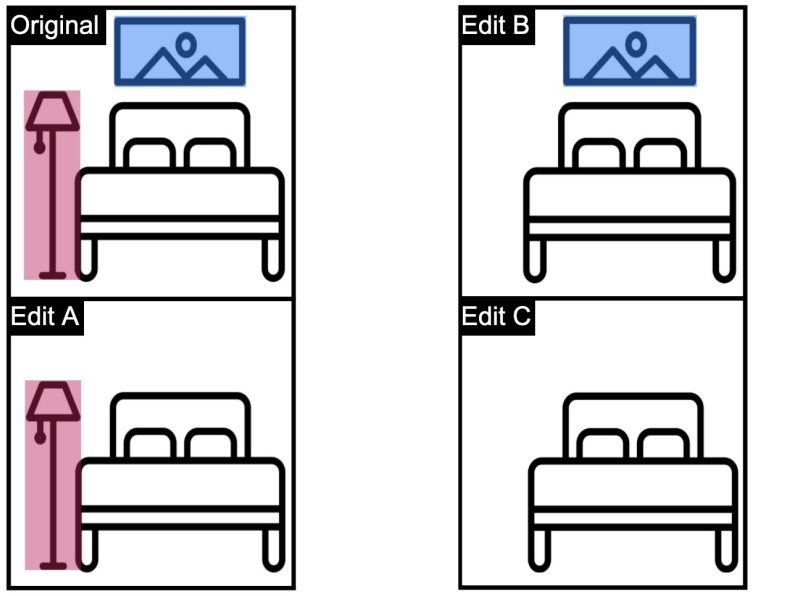


Figure 4.2: Using the differences in class output score, we get two importance scores for painting, one in the presence of lamp, and one in the absence of lamp.

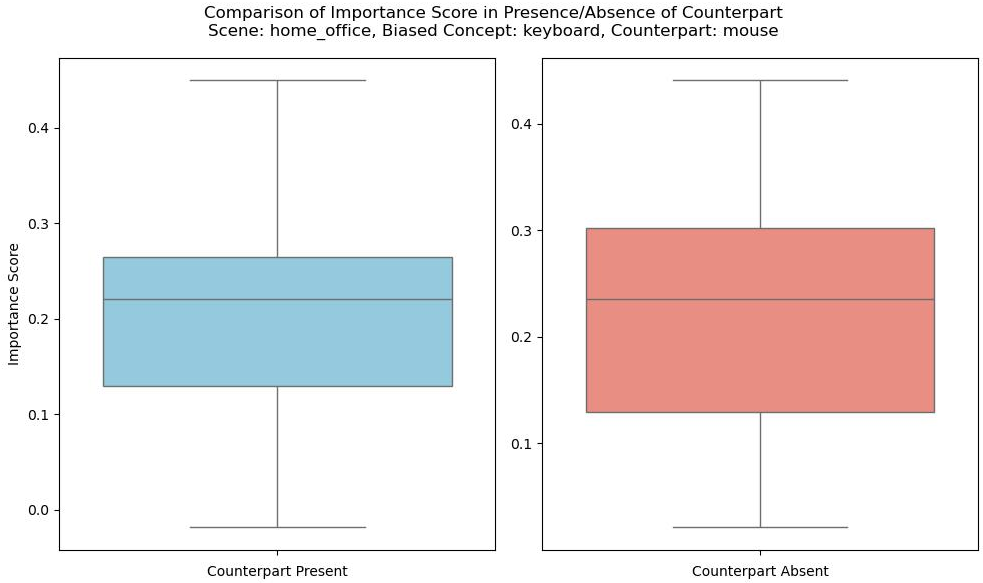


Figure 4.3: The importance of keyboard for scene home-office in the presence and absence of mouse.

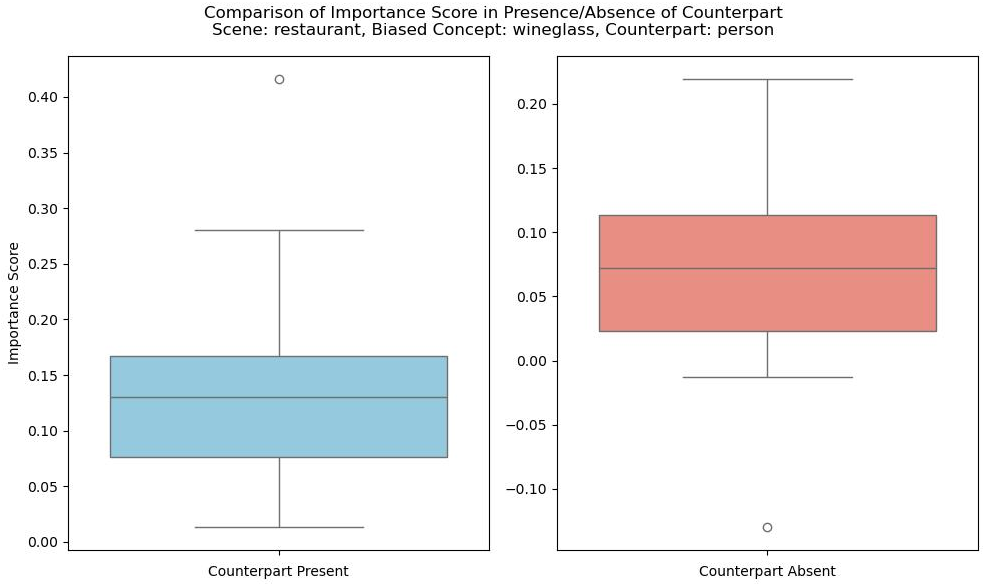


Figure 4.4: The importance of wine glass for scene restaurant in the presence and absence of person.

Wine-glass Importance for Restaurant		
	Importance Score	Standard Deviation
Person Present	0.1455	0.0976
Person Absent	0.071	0.0844

Table 4.1: The average importance score and standard deviation for wine glass for scene restaurant.

4.4 Discussion and Future Work

Using counterfactual generation we are able to more directly learn the casual relationships between concepts and model outcomes; however, methods to effectively quantify the effects of concepts on model outcomes as coefficients to be linearly combined still requires further work.

Chapter 5

Conclusion

Prior work has highlighted several limitations in existing concept-based interpretability methods. We attempt to address the issues of misalignment of probe datasets used for explanations, and learning correlations as opposed to causality using counterfactual generation. We have encountered many further limitations in the course of this work, from the capabilities of generative AI applied to counterfactual images, to the issues of non-linearity in concept-based explanations. However, we have also highlighted several further avenues for research which I look forward to investigating!

Bibliography

[1] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[3] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[4] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[5] R. Fong, M. Patrick, and A. Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[6] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[7] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[10] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[12] V. V. Ramaswamy, S. S. Kim, R. Fong, and O. Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10932–10941, 2023.

[13] V. V. Ramaswamy, S. S. Kim, R. Fong, and O. Russakovsky. Ufo: A unified method for controlling understandability and faithfulness objectives in concept-based explanations for cnns. *arXiv preprint arXiv:2303.15632*, 2023.

[14] K. K. Singh, D. Mahajan, K. Grauman, Y. J. Lee, M. Feiszli, and D. Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11070–11078, 2020.

[15] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[16] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.