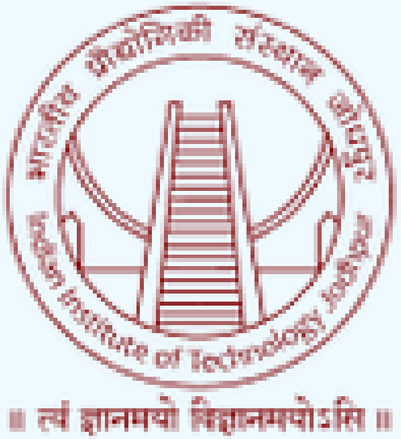


IIT Jodhpur

Dependable AI

Generative Counterfactual Introspection For Explainable Deep Learning

Renu Sankhla (B21AI028)

Contents

Abstract

In this paper, we propose an introspection technique for deep neural networks utilizing a generative model. This technique facilitates salient editing of input images for model interpretation, enabling intervention to alter predictions. We showcase its effectiveness by answering counterfactual inquiries, demonstrating meaningful changes to input images. Our approach unveils intriguing classifier properties on MNIST and CelebA datasets.

Introduction

Addressing challenges in understanding complex neural network decision boundaries, our work introduces a generative counterfactual introspection framework.

Existing methods predominantly focus on correlation between inputs and outputs, lacking discriminative and counterfactual insights crucial for explaining deep neural networks.

Leveraging powerful generative models and latent feature editing mechanisms, we pioneer a framework for generating actionable counterfactual explanations.

Our approach goes beyond existing prototype and criticism-based methods by providing inherently interpretable counterfactual visual explanations.

By exploring decision boundaries and relationships between classes, we offer actionable feedback and generate counterfactual **prototypes and criticisms,** essential for understanding model behavior.

Literature Review

1. Various model introspection methods, including CNN interpretation techniques like GradCAM and localized model approximations, have emerged to enhance interpretability of predictions.

2. Interpretation approaches such as LIME and decision process modeling via partition trees aim to attribute prediction results directly into the input domain.

3. With a growing emphasis on causal understanding, interpretation methods focusing on counterfactual reasoning have been proposed, addressing issues like fairness evaluation and patch-based editing for prediction changes.

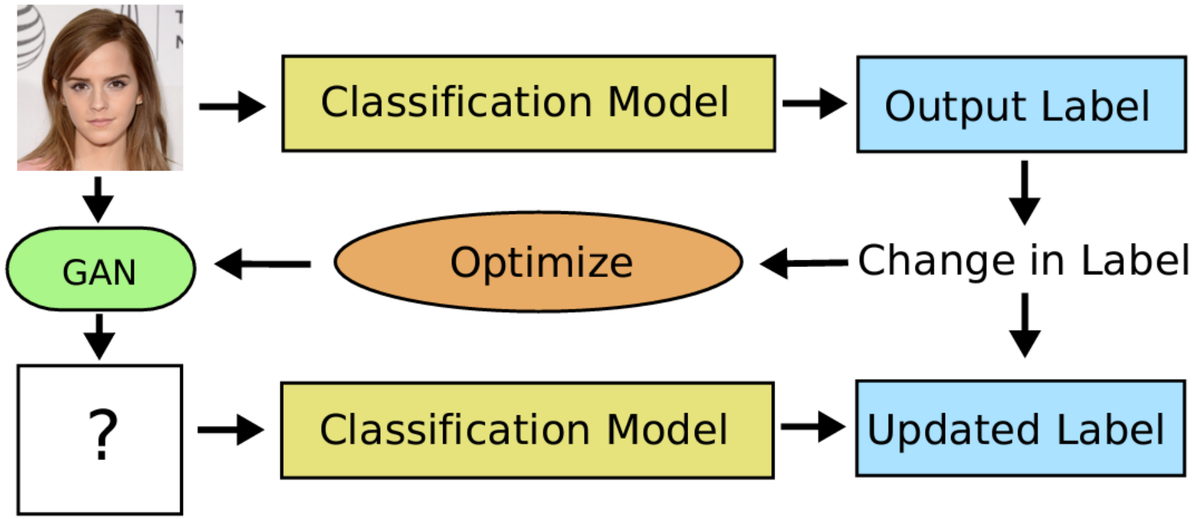


Figure 1: The illustration of the generative counterfactual introspection concept.

4. Text-based explanations and examining the relationship between the trained model and training dataset also contribute to understanding model behavior.

5. Concerns about the safety of deep neural networks, particularly regarding adversarial samples, have spurred specialized optimization approaches and the utilization of generative adversarial networks to ensure meaningful input modifications.

Methods

1. Minimal Change Counter factual Example Generation
2. Approximate Solution

Minimal Change Counter factual Example Generation

Objective : Our aim is to generate minimal change counterfactual examples for a given query image I such that altering key attributes leads the network to either change its decision or increase confidence in the query class.

Utilizing Generative Adversarial Networks (GANs) : We leverage GANs to transform latent factors into synthetic samples resembling training data, enabling the manipulation of salient attributes in an unsupervised manner.

Generative Editing Models : We denote generative editing models as $G(I;A)$ or $G(I;Lo)$, depending on whether actionable attributes are known or unknown, respectively.

Formulation of Problem : We formulate the minimal change counterfactual explanation generation problem, aiming to minimize the difference between the original image and its manipulated version while achieving the desired target attribute.

Loss Function : Our approach involves an alternating loss function to promote solutions that maximize class confidence, ensuring meaningful alterations while generating prototypes.

$$\begin{aligned} \min_{A'} \quad & \|I - I(A')\|_p \\ \text{s.t.} \quad & c' = C(I(A')) \\ & I(A') = G(I; A') \end{aligned}$$

Approximate Solution

Challenges with Non-linear Formulations : Due to the non-linear and non-convex nature of deep neural network formulations, finding closed-form solutions becomes impractical.

Relaxed Optimization Problem : We propose a relaxed version of the optimization problem, allowing for efficient solutions using gradient descent algorithms.

$$\min_{A'} \quad \lambda \cdot loss_{C,c'}(I(A')) + \|I - I(A')\|_p$$

Objective Function : Our approach involves minimizing a combination of cross-entropy loss and the difference between the original and manipulated images.

Efficient Optimization: The differentiability of both the classifier C and generator G enables us to compute gradients using back-propagation, facilitating gradient descent for solving the minimal change counterfactual example generation problem.

I**terative Refinement** : To achieve explanations with minimum change, one can iteratively solve the optimization problem, continuously updating parameters such as λ using bisection search or similar methods for one-dimensional optimization.

Experiment

1 MNIST dataset

It is a widely used dataset in machine learning, containing 60,000 training and 10,000 test images of handwritten digits ranging from 0 to 9.

2 CelebA dataset

It is a large-scale face attributes dataset with over 200,000 celebrity images, each annotated with explicit attributes such as age, gender, and presence of accessories like glasses or hats.

MNIST Dataset

Problem Description: Classifying hand-written digits into one of ten classes (0 to 9).

Classifier and Image Generator: Utilizing a pretrained classifier achieving 99.10% accuracy on the test set and a DCGAN architecture with a 10D latent space as the image generator.

Optimization Method : Employing the proposed optimization approach to update the latent vector L0 to generate meaningful modifications of the image answering the counterfactual query.

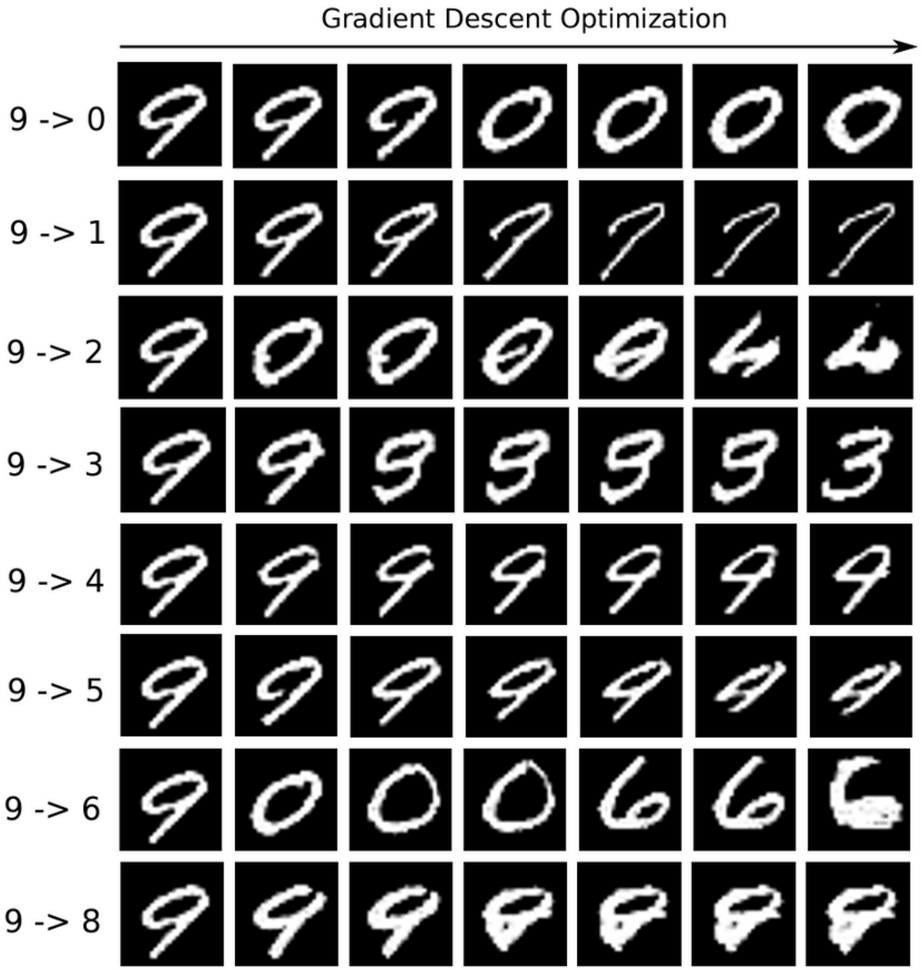


Figure 2: Finding criticism of the digit 9 class.

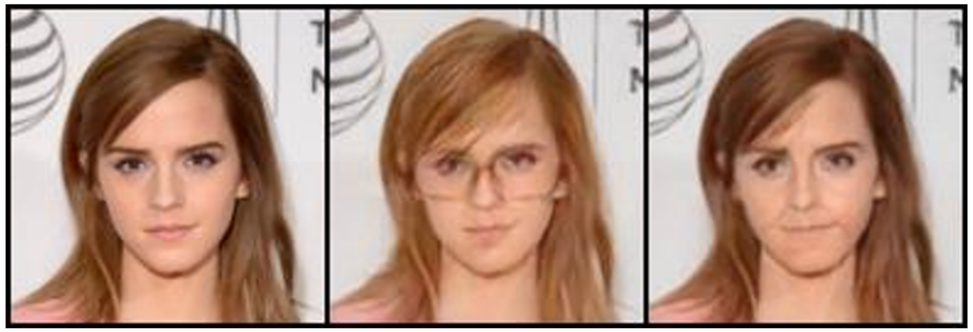
Results Visualization: : We see that meaningful changes to the image of digit 9 are depicted to alter its prediction, showcasing the optimization path and the resulting images altering the classified label.

Insights and Observations: The optimization explores the manifold of possible meaningful images, revealing insights into classifier decision boundaries and the preference regarding similarity between digits.

Figure 4: The effect of regularization on the optimization path for finding criticism of 9 in the direction of 7.

Regularization for Interpretability : Incorporating a regularization term ensures minimal and consistent modifications to the images, facilitating smooth optimization

CelebA Dataset



(a) Original (b) Optimized (c) "Old" Target

Problem Description: The classifier is trained to classify celebrity face images into "young" or "old" categories, achieving an average accuracy of 90.89% on the CelebA testing set.

Generative Editing Method: AttGAN is employed as the generative editing method, allowing for modifications based on additional attributes such as hair color, glasses, bangs, and baldness.

Results Visualization: We see that different editing schemes for input images, showcasing modifications driven by classifier preference and explicit attribute changes.

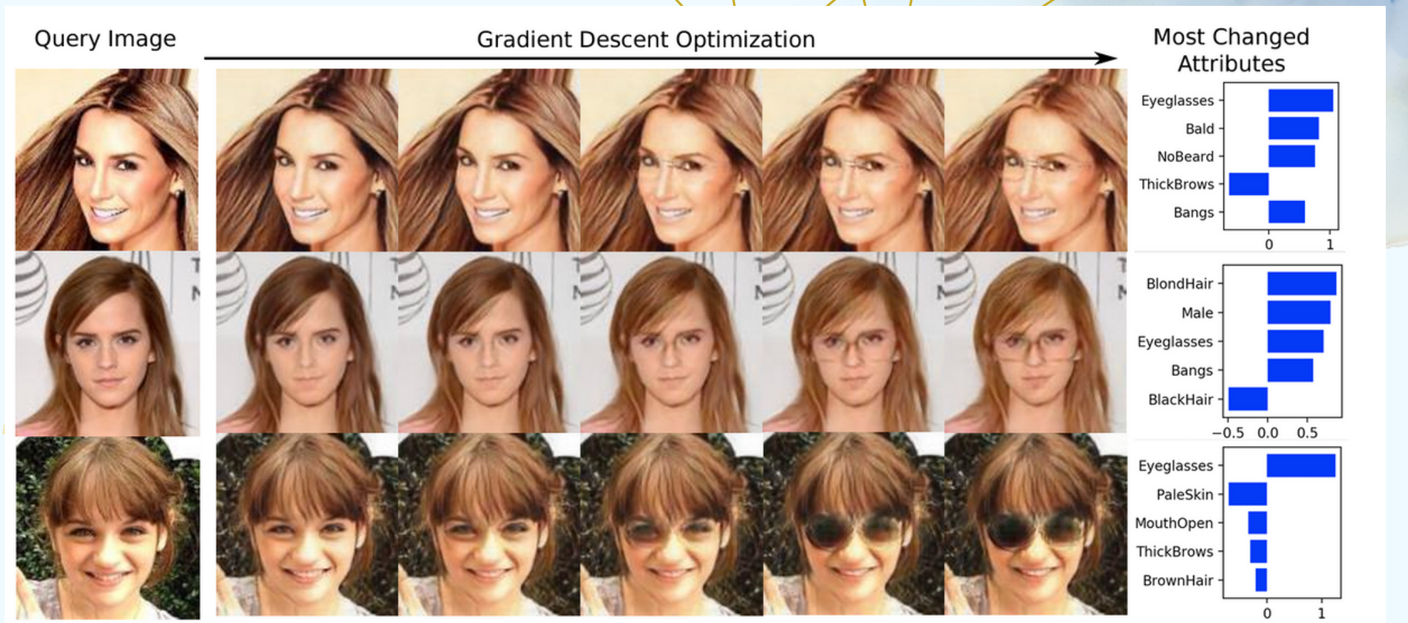
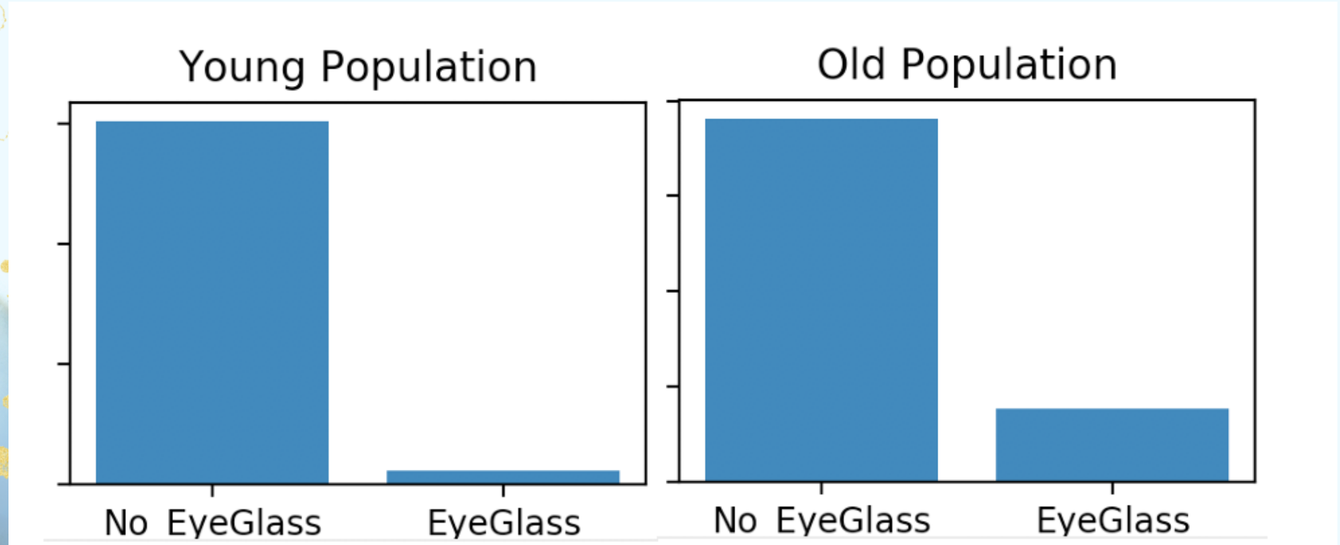


Figure 6: Illustrate attributes changes (beside the young/old attribute) that will make the images appear older for the given classifier. The left most column shows the original image. The right most column shows the top five most changed attributes and their relatively changes.



Attribute Manipulation and Optimization : we explore attribute changes, apart from the "young/old" attribute, that influence the perceived age of the images, revealing interesting patterns such as the presence of eyeglasses.

Understanding Classifier Behavior: By investigating prototype and criticism examples for male celebrity faces labeled as "old", we gain insights into the classifier's behavior and potential biases in the training data.

The potential bias in the CelebA dataset. The percentage of people having eye glass is much higher in the population labeled as “old”.

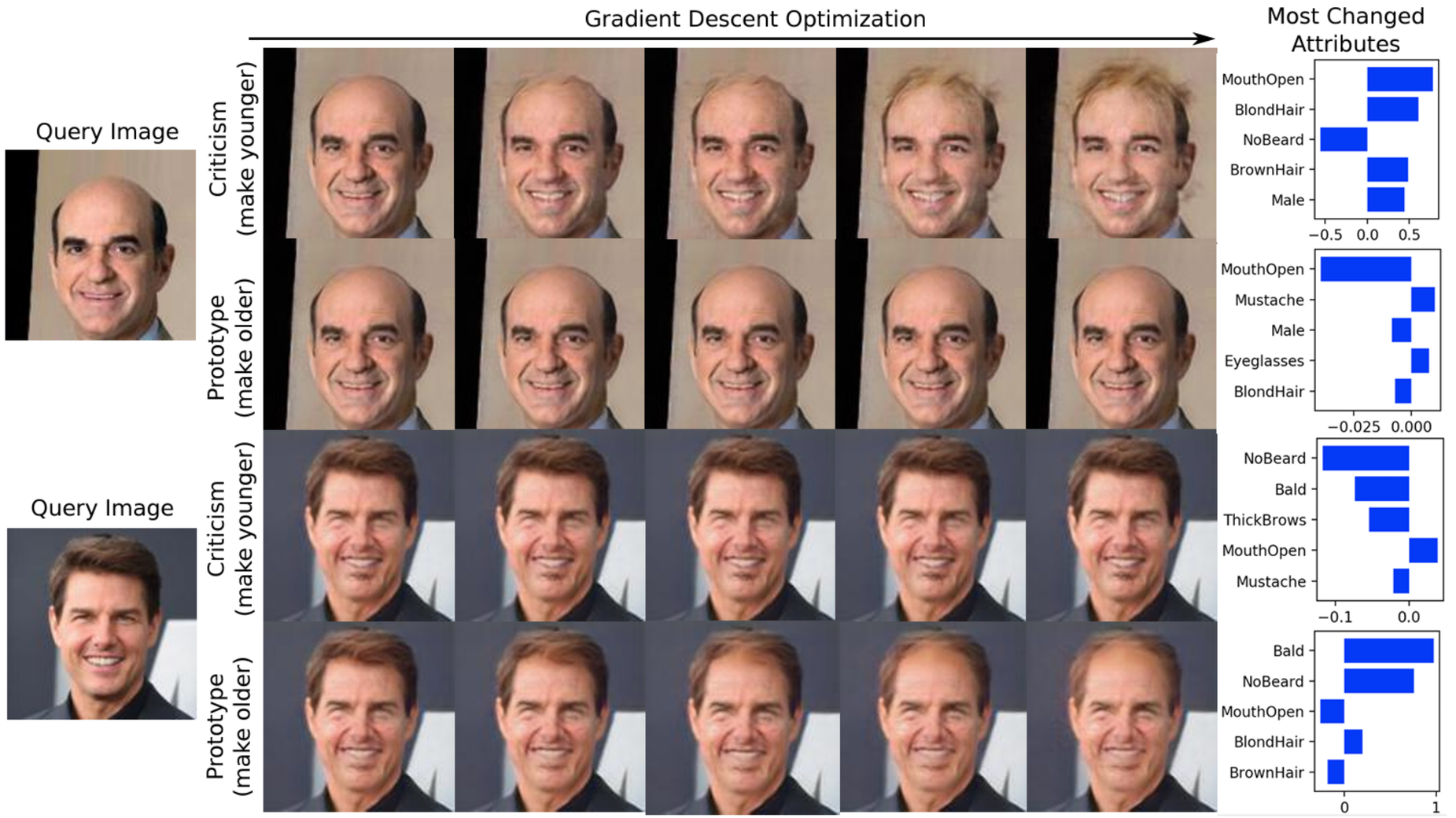


Figure 8: Prototype and criticism for the images with ground truth label “old”. The left most column shows the original image. The right most column shows the top five most changed attributes and their relatively changes.

Actionable Insights : The top five most changed attributes provide actionable insights for achieving desired changes, highlighting the importance of hair features in discriminating age groups

Conclusion

In conclusion, our research showcases the efficacy of generative counterfactual introspection for explainable deep learning. By exploring the MNIST and CelebA datasets, we elucidate classifier decision boundaries and uncover potential biases. Through attribute manipulation and optimization, we gain valuable insights into classifier behavior, aiding in the interpretation of model predictions. Our findings highlight the importance of understanding the impact of attribute changes on model predictions for enhanced interpretability and fairness. Overall, our approach offers actionable insights for improving model transparency and addressing challenges in deep learning interpretability.

THANK YOU