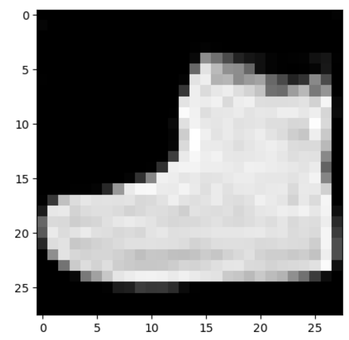
PIECE: On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Leaning



Yu Li, Yifan Wang

Poster Presentations in context of Interpretable Machine Learning

1 TL;DR

- PIECE was used to generate counterfactual and semi-factual images, which were applied to the MNIST and Fashion-MNIST datasets. The images generated by PIECE are more comprehensible compared to images generated by other algorithms.

2 Motivation & Problem Setting

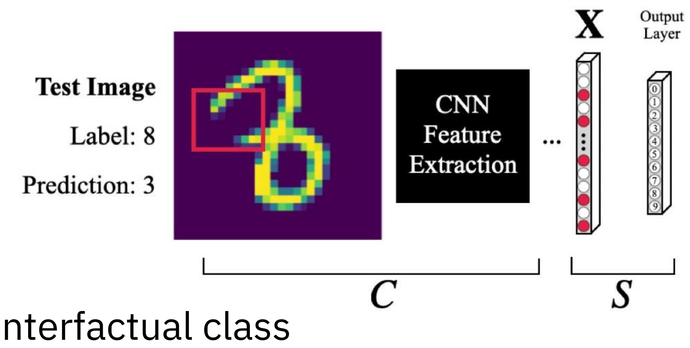
Motivation

- **PIECE for generating Semi- / Counterfactual** [Kenny, Keabe et al. 2020]
- **Counterfactual by C-Min-Edit** [Wachter, Mittelstadt, and Russel. 2017]

Problem Setting

Counterfactual explanations in iML have gained popularity, but the less-explored semi-factuals, similar to human reasoning, have been neglected. The method PIECE (Exceptionality-based Contrastive Explanations) generate plausible counterfactuals and semi- factuals for black-box CNN classifiers in computer vision.

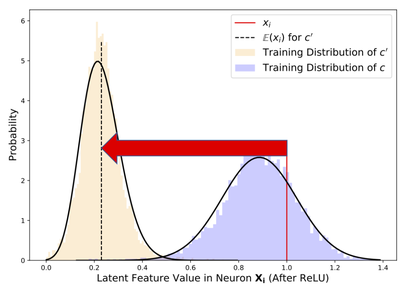
3 Approach



Identify the Counterfactual class

$$z = \arg\min_{z_0} ||C(G(z_0)) - C(I)||_2^2 + ||G(z_0) - I||_2^2 \quad (1)$$

$$\arg\max_{z} || S\left(C\left(G(z)\right)\right) - Y_c||_2^2 \quad (2)$$



Approach in PIECE Method

1. Identifying exceptional features

a. define the hurdle models:

$$p(x_i) = (1 - \theta_i)\delta_{(x_i)(0)} + \theta_i f_i(x_i),\ s.t.\ x_i \geq 0 \quad (3)$$

b. x_i is considered as an exceptional feature x_e when:

x_i does not activate:

$$x_i = 0 \mid p(1 - \theta_i) < \alpha \quad (4)$$

$$x_i > 0 \mid p(\theta_i) < \alpha \quad (5)$$

x_i does activate:

$$\theta_i F_i(x_i) < \alpha \mid x_i > 0 \quad (6)$$

$$(1 - \theta_i) + \theta_i F_i(x_i) > 1 - \alpha \mid x_i > 0 \quad (7)$$

2. Modify the Exceptional to the Expected

Algorithm 1: Modify exceptional features in x to produce x'

Input: x: The latent features of the test image I
Input: w: The weight vector connecting $\mathbf{X}$ to c'
Input: $\{x_e\}_{e=1}^n \in x$: The exceptional features (ordered lowest to highest probability)

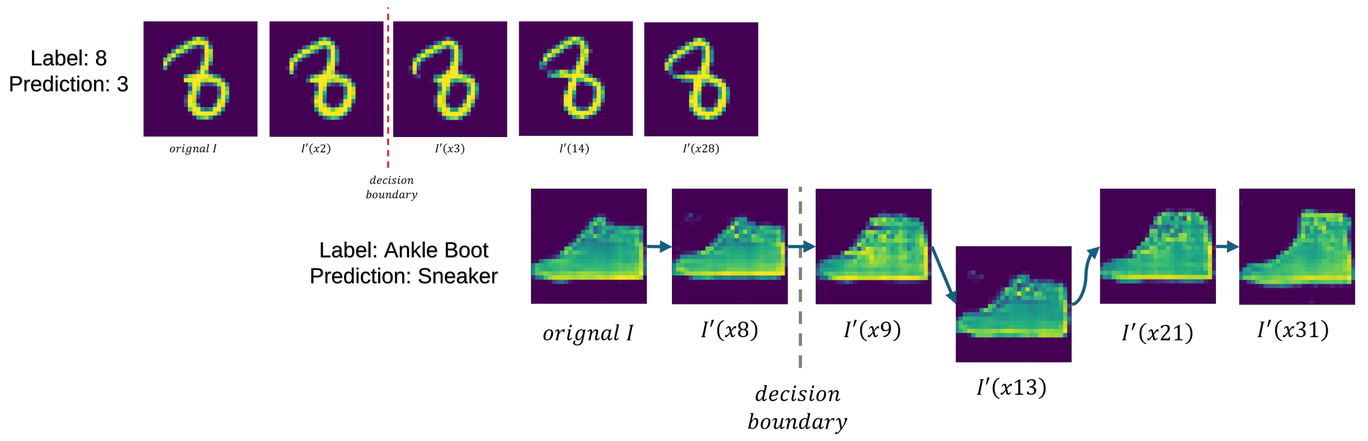
```
foreach x_e in {x_e}_{e=1}^n ∈ x do
    if w_e > 0 and x_e discovered with Eq. (4), Eq. (5), or Eq. (6) then
        x_e ← E[X_e]        // Using PDF modelled for c' in Eq. (3)
    else if w_e < 0 and x_e discovered with Eq. (5) or Eq. (7) then
        x_e ← E[X_e]        // Using PDF modelled for c' in Eq. (3)
return x (now modified to be x')
```

3. Visualizing the Explanation with GAN

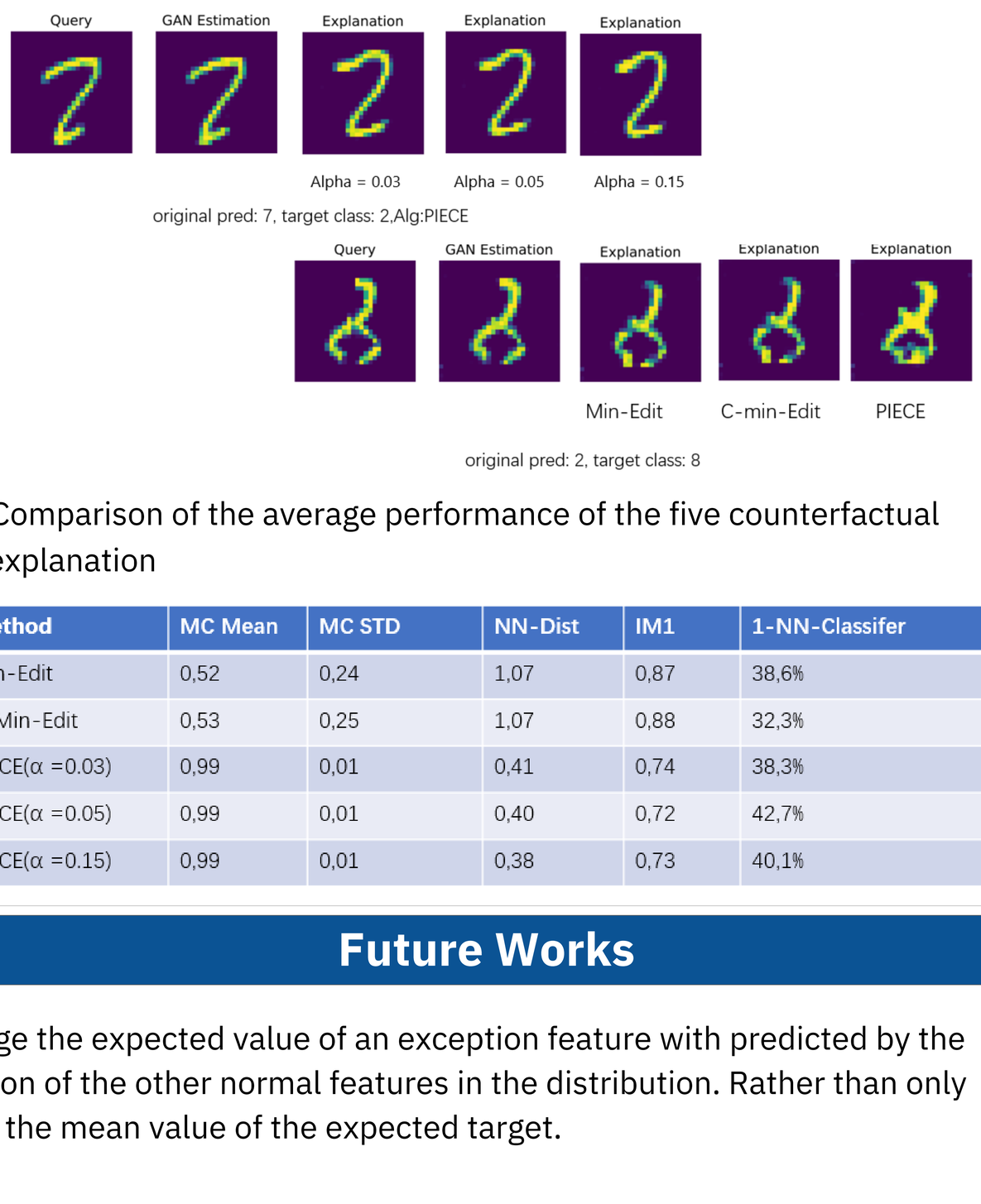
$$z' = \arg\min_{z} || C\left(G(z)\right) - x'||_2^2 \quad (8)$$

4 Key Insights

- Example of generating semi-/ counterfactual images on MNIST Digits and Fashion



- Using different algorithms and probability density function threshold α to generate images



- Comparison of the average performance of the five counterfactual explanation

Method	MC Mean	MC STD	NN-Dist	IM1	1-NN-Classifer
Min-Edit	0,52	0,24	1,07	0,87	38,6%
C-Min-Edit	0,53	0,25	1,07	0,88	32,3%
PIECE(α =0.03)	0,99	0,01	0,41	0,74	38,3%
PIECE(α =0.05)	0,99	0,01	0,40	0,72	42,7%
PIECE(α =0.15)	0,99	0,01	0,38	0,73	40,1%

5 Future Works

Change the expected value of an exception feature with predicted by the position of the other normal features in the distribution. Rather than only using the mean value of the expected target.