The Gaussian Discriminant Variational Autoencoder (GdVAE): A Self-Explainable Model with Counterfactual Explanations

Anselm Haselhoff, Kevin Trelenberg, Fabian Küppers, and Jonas Schneider

We integrate **counterfactual** and **prototype explanations** to analyze "what-if" scenarios and find classifier biases.

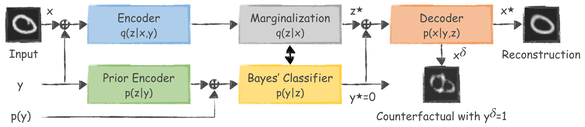
Relevance

Current visual counterfactual explanation (CF) methods require either post-hoc training or slow optimization during inference, often without guaranteed convergence. Our self-explainable model offers a fast, transparent analysis of the classifier's decision process and boundary, with user-specified confidence for CFs.

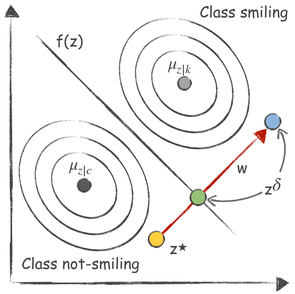
Desirable Properties of Explanations

- ***Realism:*** CFs should stem from the data manifold.
- ***Consistency:*** CFs should be conform with the desired classifier output $F(x^\delta) = \delta$ or $f(z^\delta) = \delta$.
- ***Proximity:*** CFs should minimally change the input.
- ***Transparency:*** Explanations (e.g., prototypes) are an intrinsic part of a white-box predictor.

Architecture



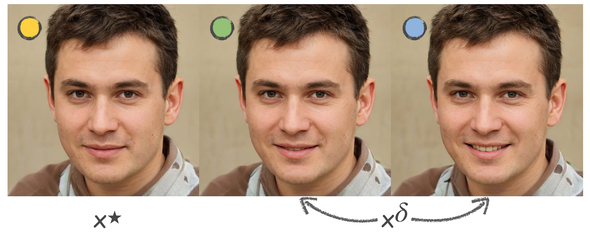
Counterfactual Generation in Latent Space



Prototypes — Global Explanations

Classifier decision is based on female prototypes!
Potential indicator of classifier or dataset bias?

Counterfactuals — Local Explanations



Fine-grained method to analyze the decision boundary.

FFHQ Counterfactual Examples

Results

Predictive Performance

Method	MNIST		CIFAR-10		CelebA - Gender	
	ACC% ↑	*MSE* ↓	*ACC%* ↑	*MSE* ↓	*ACC%* ↑	*MSE* ↓
Imp. Sampling	**99.0**	**1.04**	55.0	2.45	94.7	1.77
Ours	**99.0**	1.10	**65.1**	**1.71**	**96.7**	**0.91**
Black-box	99.3	1.12	69.0	1.45	96.7	0.82
ProtoVAE	**99.1**	1.51	76.6	2.69	96.6	1.32
Ours (data aug.)	98.7	**0.93**	**76.8**	**1.18**	**96.8**	**0.71**

Takeaway: GdVAE rivals black-box and state-of-the-art methods, especially on higher-dimensional images.

Quality of Counterfactual Explanations

Method		Consistency		Realism	Proximity	
	ρ_p ↑	*ACC%* ↑	*MSE* ↓	*FID* ↓	*MSE* ↓	
GANalyze	0.84	5.5	6.75	54.89	6.33	MNIST - 0/1
UDID	0.85	1.2	8.82	**38.89**	7.44	
ECINN	0.93	33.0	1.76	87.25	**3.47**	
EBPE	**0.97**	**44.6**	**0.50**	108.94	25.73	
C3LT	0.89	3.6	6.32	57.09	5.83	
Ours (L2)	0.95	42.9	0.95	91.22	4.58	
Ours (Maha.)	0.95	**44.6**	0.87	89.91	4.10	
GANalyze	0.78	15.2	5.42	147.43	13.47	CelebA - Smiling
UDID	0.86	15.8	4.22	178.23	13.73	
ECINN	0.72	21.3	5.68	95.35	1.16	
EBPE	**0.94**	**41.9**	**1.22**	191.67	1.54	
C3LT	0.90	11.8	3.94	101.46	3.97	
Ours (L2)	0.81	25.0	3.65	**85.52**	0.99	
Ours (Maha.)	0.82	25.7	3.51	85.56	**0.92**	

Takeaway: GdVAE consistently provides a better trade-off between metrics compared to post-hoc methods.

CelebA - Gender Bias of Smile Classifier

Hidden Attribute	*ACC%* ↑	*MSE* ↓
Male	89.9 ±1.37	0.92±0.03
Female	91.1±0.33	0.88±0.02

Takeaway: Smile classifier shows reduced performance and increased uncertainty in classifying males.

Project page
paper & code