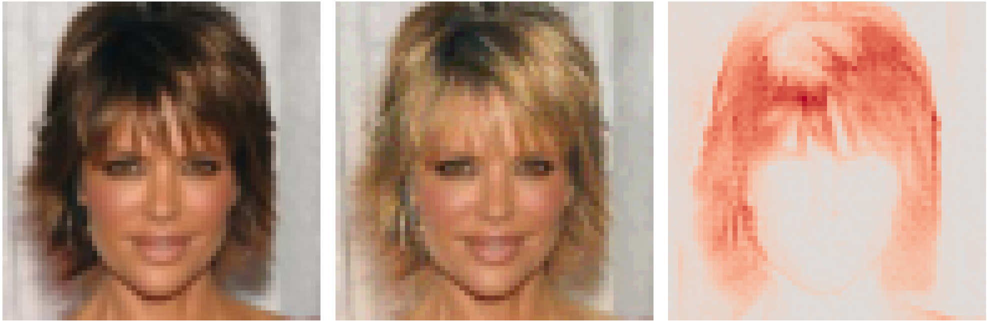
Diffeomorphic Explanations with Normalizing Flows

Ann-Kathrin Dombrowski*, Jan E. Gerken*, Pan Kessel

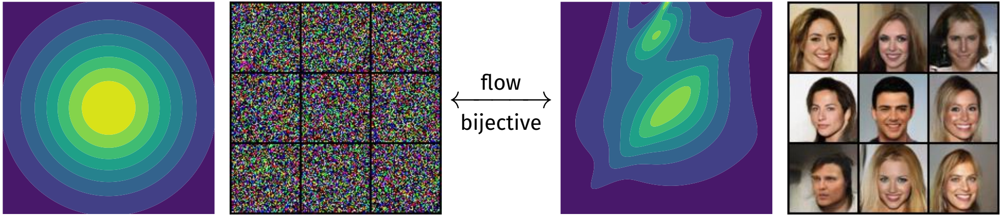
*equal contribution



original x counterfactual x' heatmap δx

Outline

Normalizing Flows

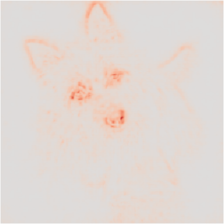


base distribution | samples from base distribution | learned distribution | samples from learned distribution

Explanations for classifiers

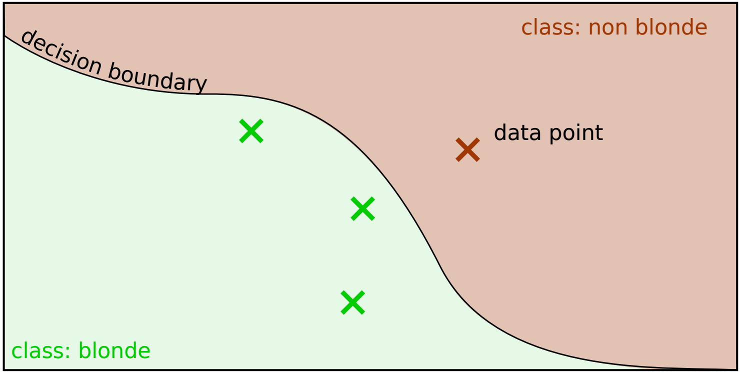
image x

classifier $\xrightarrow{f(x)}$ prediction: dog $\xrightarrow[\frac{\partial f(x)}{\partial x}]{\text{explanation}}$

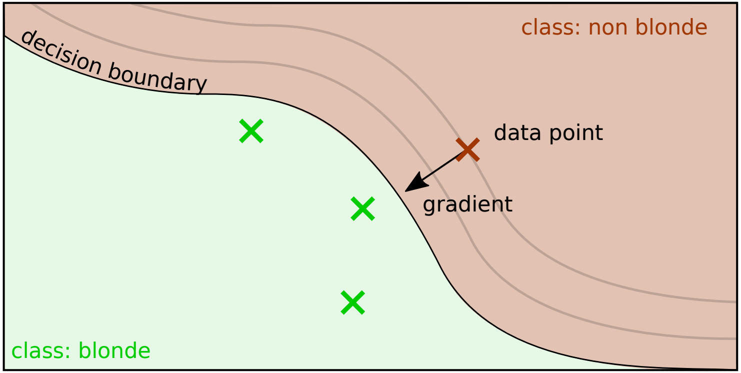


explanation

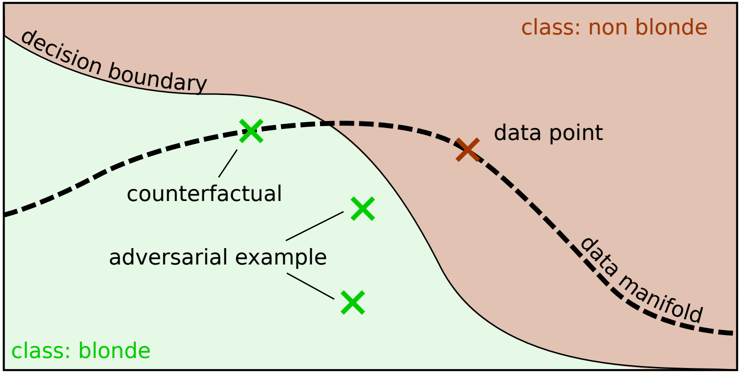
Finding counterfactuals



Gradient ascent

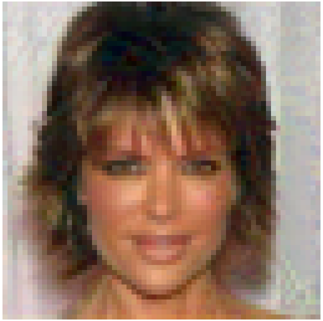


Finding counterfactuals *on* the data manifold

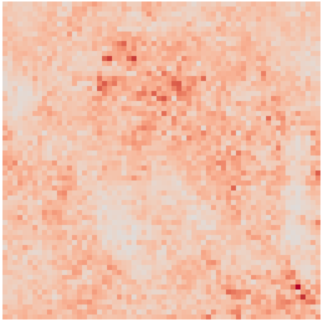


Adversarial examples lie *off* the data manifold

original x
not blonde $(p \approx 0.99)$



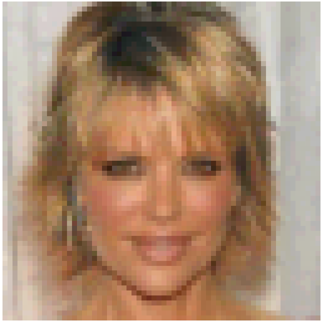
adversarial example x'
blonde $(p \approx 0.99)$



heatmap δx

Finding counterfactuals *on* the datamanifold

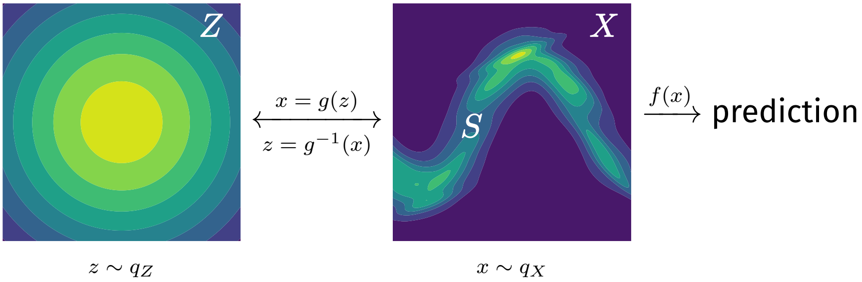
original x
not blonde ($p \approx 0.99$)



counterfactual x'
blonde ($p \approx 0.99$)

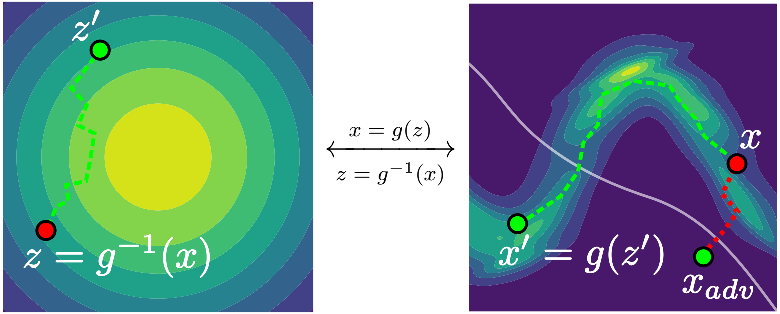
heatmap δx

Method



- choose sample x
- find representation in base space $z = g^{-1}(x)$
- update z with gradient $\frac{\partial f_k(g(z))}{\partial z}$ until target class has desired probability

Intuition



gradient ascent in Z gradient ascent in X

Gradient ascent in base space

Gradient ascent in X for class k of the classifier f with learning rate λ:

$$x^{(t+1)} = x^{(t)} + \lambda \frac{\partial f_k}{\partial x}(x^{(t)})$$

Theorem

Gradient ascent in the base space Z is given by

$$x^{(t+1)} = x^{(t)} + \lambda\, \boldsymbol{\gamma^{-1}(x^{(t)})} \frac{\partial f_k}{\partial x}(x^{(t)}) + \mathcal{O}(\lambda^2)$$

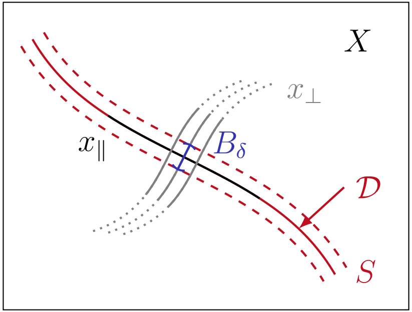
where $\gamma^{-1}(x) = \left(\frac{\partial g}{\partial z}\frac{\partial g}{\partial z}^T\right)(g^{-1}(x))$ is the inverse of the induced metric on X from Z under the flow g.

Coordinates on X

Approximate the data manifold by

$$S = \mathcal{D} \times B_{\delta_1} \times \cdots \times B_{\delta_{N-n}}$$

and use Gaussian normal coordinates



The induced metric in normal coordinates

Theorem

In Gaussian normal coordinates, γ^{-1} is given by

$$\gamma^{-1} = \begin{pmatrix} \gamma_{\mathcal{D}}^{-1} & & & \\ & \gamma_{B_{\delta_1}}^{-1} & & \\ & & \ddots & \\ & & & \gamma_{B_{\delta_{N-n}}}^{-1} \end{pmatrix}$$

and, for well-trained flows, $\gamma_{B_{\delta_i}}^{-1} \to 0$ for $\delta_i \to 0$.

$\Rightarrow$ *For gradient ascent in Z, the learning rate in $x_\perp$ directions is scaled by a vanishing factor. Therefore, we stay on the data manifold.*

Sketch of proof

- For well-trained flows g, almost all probability mass is concentrated in $S = \mathcal{D} \times B_{\delta_1} \times \cdots \times B_{\delta_{N-n}}$

$$1 - \epsilon < \int_S q_X(x)\,\mathrm{d}x = \int_S \sqrt{\det|\gamma|}\, q_Z(g^{-1}(x))\,\mathrm{d}x$$

- When $\delta_i \to 0$, the integration domain shrinks to zero, but the value of the integral is bounded from below

- Hence, the metric $\gamma_{B_{\delta_i}}$ has to diverge, i.e. $\gamma_{B_{\delta_i}}^{-1} \to 0$.

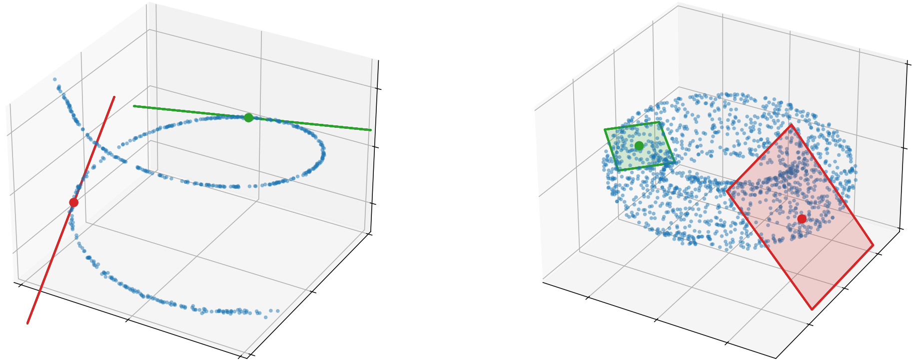
Tangent space from induced metric

- Perform singular value decomposition of the Jacobian $\frac{\partial g}{\partial z} = U\,\Sigma\,V$
- Rewrite the inverse induced metric as

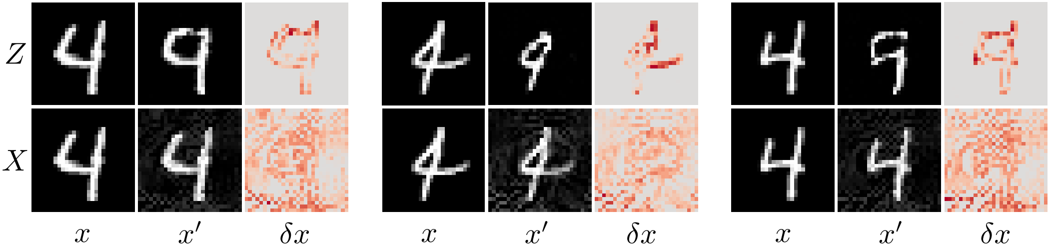
$$\gamma^{-1} = \frac{\partial g}{\partial z}\frac{\partial g}{\partial z}^{T} = U\,\Sigma^2\,U^T$$

- For n dimensional data manifold: n large singular values
- Corresponding left-singular vectors span the tangent space of the data manifold

Tangent space from induced metric

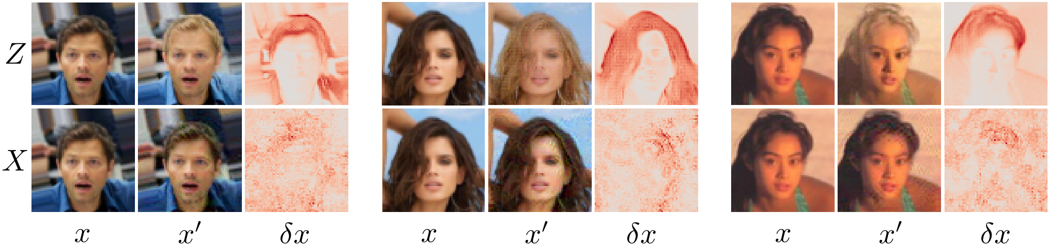


Experiments with MNIST



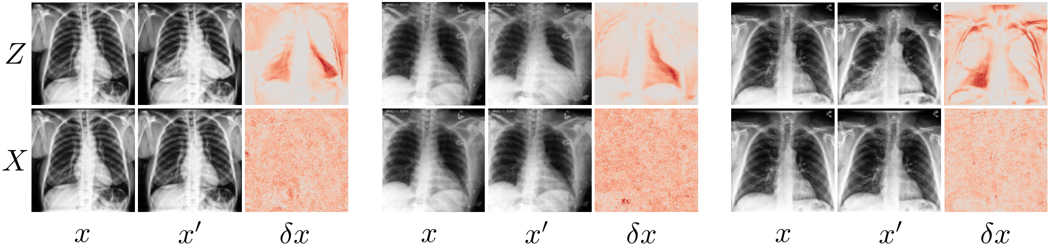
- task: $4 \to 9$
- Flow: RealNVP
- Classifier: CNN with 10 classes (test acc: 99%)

Experiments with CelebA



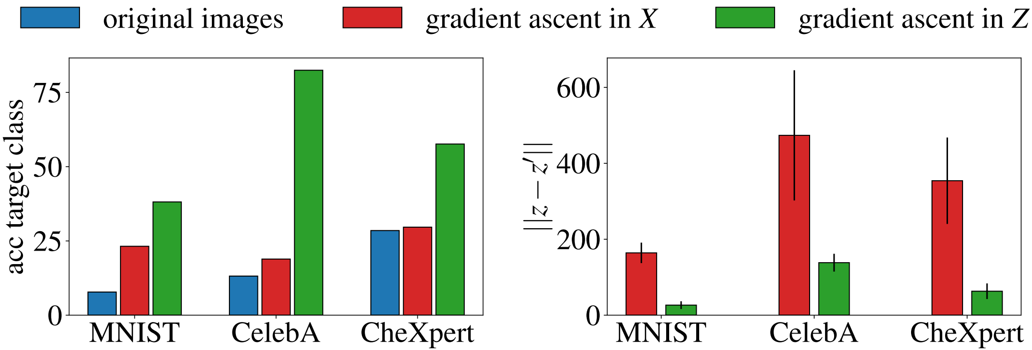
- task: *not blonde* $\rightarrow$ *blonde*
- Flow: Glow
- Classifier: Binary CNN trained on *blonde/not blonde* attribute (test acc: 94%)

Experiments with CheXpert



- task: *healthy* $\rightarrow$ *cardiomegaly*
- Flow: Glow
- Classifier: Binary CNN trained on *cardiomegaly/healthy* attribute (test acc: 86%)

Quantitative Evaluation



Conclusion

Summary

- Counterfactual examples provide explanations for classifiers
- Gradient ascent in the input space of the classifier leads off the data manifold
- Gradient ascent in the base space of the flow leads to counterfactuals on the data manifold
- Derived theoretically and shown experimentally

Open questions

- Dependence on flow architecture
- Robustness of counterfactuals against manipulations