Clarity: an improved gradient method for producing quality visual counterfactual explanations

Claire Theobald

Université de Lorraine, CNRS, LORIA, F-54000 Nancy France
claire.theobald@loria.fr

21-04-2023

Explaniable AI

- Machine learning models such as deep neural networks have become more and more complex over the past years.
- Their increase in performance has come with a tradeoff in explainability.
- Complex models are hard to explain in a form comprehensible by humans.

Counterfactual explanations

- In this talk, I will focus on **counterfactual explanations**.
- Given a classifier C, an input X and its predicted class $y = C(X)$, a counterfactual explanation X' of X for a *target class* $y' \neq y$ is an input as close as possible to X but of predicted class y'.
- Example: "You were unable to sell your apartment for 150k€ because the surface is too low. If the apartment had a higher surface, then the apartment would have sold for 150k€."

Counterfactual explanations

- Counterfactual explanations must be **realistic**: they must be likely under the data distribution $p(X)$. They also need to be **unambiguous**; i.e. they must clearly represent the target class y'.
- I will only focus on *visual counterfactual explanations*, meaning counterfactual explanations on images.
- Quantifying realism of a counterfactual can sometimes be simple, but other times it can be way trickier, especially on images. We will only judge realism by a systematic visual inspection of counterfactual images.

Counterfactual explanations in image space

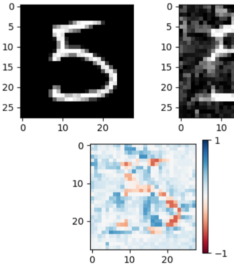
Formally, the counterfactual X' of X with target class y' is generated by a gradient descent using the following objective function [1]:

$$\mathcal{L}_{CE}(X') = L(C(X'), y') + \lambda\, d(X, X')$$
$$X' \leftarrow X' - \eta\, \nabla \mathcal{L}_{CE}(X')$$

Counterfactual explanations in image space

Formally, the counterfactual X' of X with target class y' is generated by a gradient descent using the following objective function [1]:

$$\mathcal{L}_{CE}(X') = L(C(X'), y') + \lambda\, d(X, X')$$
$$X' \leftarrow X' - \eta \nabla \mathcal{L}_{CE}(X')$$



Counterfactual explanation $5 \rightarrow 1$

Counterfactual explanations in image space

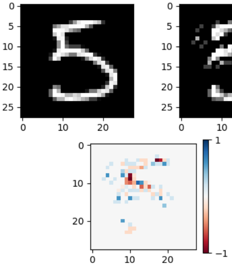
- Schut et al. [2] proposed a sparse modification of the input image by modifying one pixel at at time, based on the Jacobian Saliency Map Attack (JSMA) [3].
- They also use an ensemble of models to take into account epistemic uncertainty and indirectly minimize it.

$$\mathcal{L}_{Schut}(X') = \frac{1}{M} \sum_{m=1}^{M} L(C_m(X'), y'). \tag{1}$$

Counterfactual explanations in image space

- Schut et al. [2] proposed a sparse modification of the input image by modifying one pixel at at time, based on the Jacobian Saliency Map Attack (JSMA) [3].
- They also use an ensemble of models to take into account epistemic uncertainty and indirectly minimize it.

$$\mathcal{L}_{Schut}(X') = \frac{1}{M} \sum_{m=1}^{M} L(C_m(X'), y'). \tag{2}$$

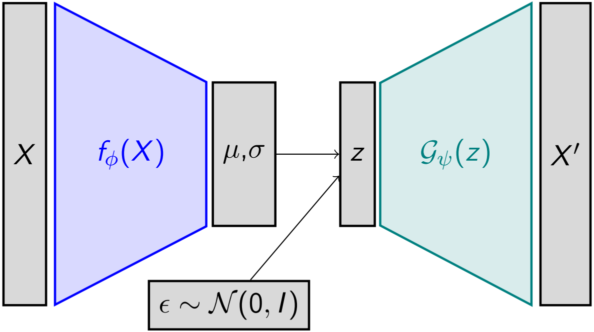


Counterfactual explanations in latent space

- Image space based counterfactual algorithms fail to produce realistic counterfactuals.
- This is because the image space is a highly dimensional space with lots of sparsity and low level information, while images are generally processed at a higher level with global features.
- A solution is to use a latent space generated by a *Variational autoencoder* (VAE).

Counterfactual explanations in latent space

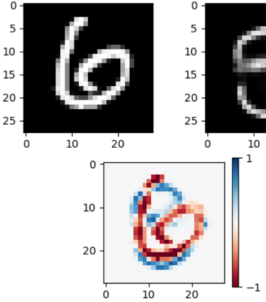
- Variational autoencoder: maps the input X to a latent Gaussian distribution $q_\phi(z|X) = \mathcal{N}(\mu, \sigma I)$, then maps the latent space back to the image space: $X' = \mathcal{G}_\psi(z)$.



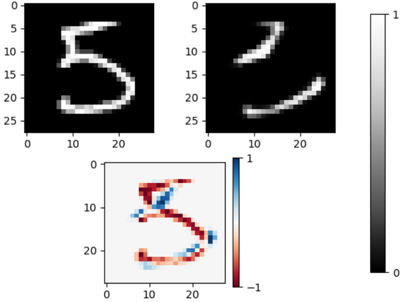
Counterfactual explanations in latent space

- *REVISE* [4] is an algorithm that generates a counterfactual using the latent space of a VAE.
- Let $q_\theta(z|X)$ be the normal variational posterior which samples a latent variable $z \in \mathcal{Z} \subset \mathbb{R}^d$ and $\mathcal{G}_\psi$ the decoder of the VAE.

$$\mathcal{L}_{Revise}(z') = L(C(\mathcal{G}_\psi(z')), y') + \lambda\, d(X, \mathcal{G}_\psi(z'))$$
$$z' \leftarrow z' - \eta \nabla \mathcal{L}_{Revise}(z')$$
$$X' \leftarrow \mathcal{G}_\psi(z')$$



(a) $6 \rightarrow 3$ (REVISE)



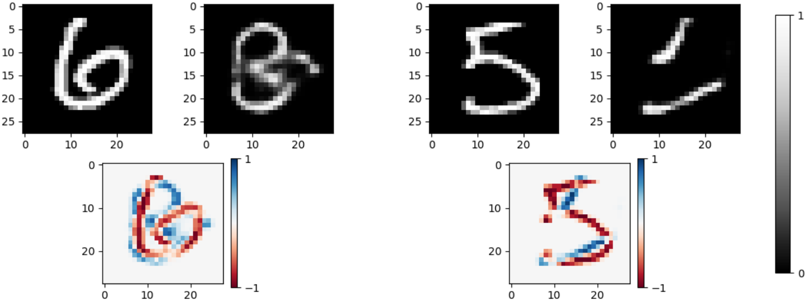
(b) $5 \rightarrow 1$ (REVISE)

Still not satisfactory...

Counterfactual explanations in latent space

We checked whether a modified version of REVISE, called REVISE-ENSEMBLE, which uses an ensemble of classifiers in order to take into account epistemic uncertainty, can yield more realistic results.

$$\mathcal{L}_{Revise-e}(z') = \frac{1}{M} \sum_{m=1}^{M} L(C_m(\mathcal{G}_\psi(z')), y') + \lambda\, d(X, \mathcal{G}_\psi(z'))$$



(c) $6 \rightarrow 3$ (REVISE-E)

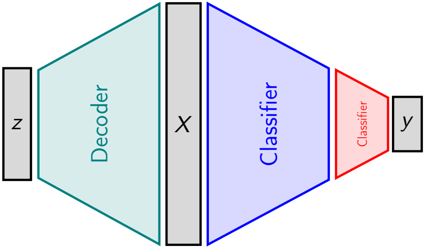
(d) $5 \rightarrow 1$ (REVISE-E)

Using the latent space as a basis for classification models

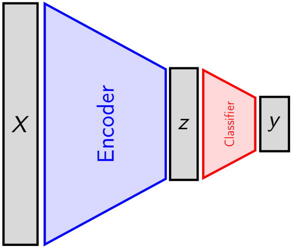
- Instead of using pre-trained classifiers on the image space, we propose to use a classifier trained directly on the latent space.

$$\mathcal{L}_{latent}(z') = L(C(z'), y') + \lambda\, d(z, z')$$

Using the latent space as a basis for classification models

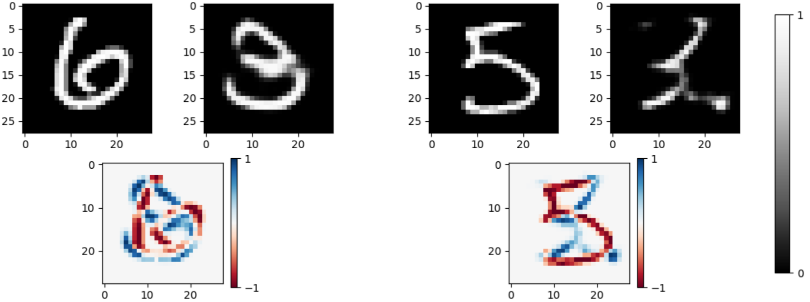


(e) REVISE architecture



(f) Latent space classifier

Difference of architecture between REVISE and using a latent space classifier. Blocks of same color share the same architecture.



(a) $6 \rightarrow 3$ (b) $5 \rightarrow 1$

Counterfactual explanations using a latent space classifier

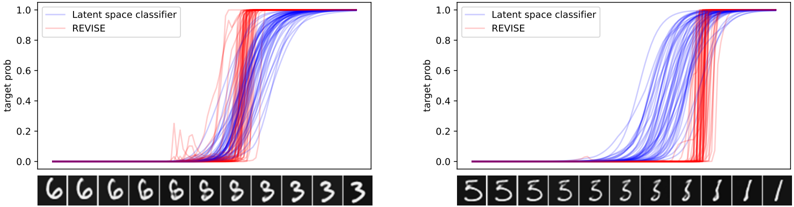
Using the latent space as a basis for classification models

- Why are those results more realistic ?
- We interpolate two images of two different classes into the latent space given the equation

$$z(t) = (1-t)\,z_1 + t\,z_2,\ t \in [0,1],$$

given two images X_1 and X_2 with respective classes y_1 and y_2. Then, we plot the target probability $t \mapsto P(Y = y_2 \,|\, z(t))$.

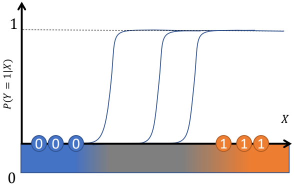
Using the latent space as a basis for classification models



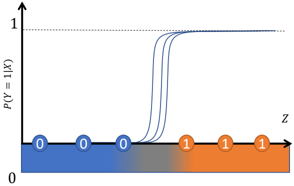
(a) $6 \rightarrow 3$ (b) $5 \rightarrow 1$

Probability of the target class with respect to the interpolation in the latent space. In red: REVISE. In blue: latent space classifier.

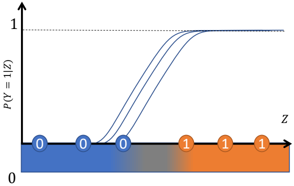
Using the latent space as a basis for classification models



(a) Classifier trained and viewed from image space.

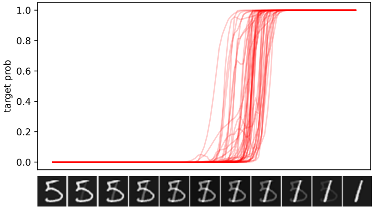


(b) Classifier trained in image space, viewed from latent space.

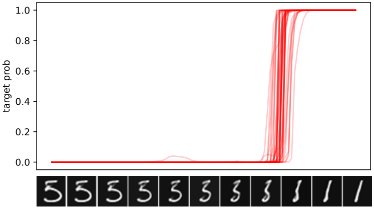


(c) Classifier trained and viewed from latent space.

Using the latent space as a basis for classification models



(d) Image space interpolation



(e) Latent space interpolation

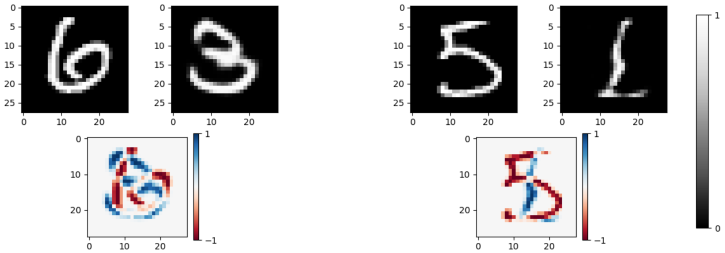
Probabilities of the target class from the ensemble of image space classifiers, with respect to the interpolation in the image and latent space respectively, from 5 to 1.

Clarity: an explainable Bayesian latent space classifier

- Just using a single classifier is not enough, as there is still a lot of variance between the possible models.
- To take into account epistemic uncertainty, we use an ensemble of classifiers $(C_m)_{m=1}^{M}$

$$\mathcal{L}_{Clarity}(z') = \frac{1}{M}\sum_{m=1}^{M} L(C_m(z'), y') + \lambda\, d(z, z')$$

Clarity: an explainable Bayesian latent space classifier

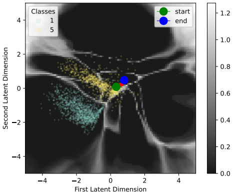


(a) $6 \to 3$ (b) $5 \to 1$

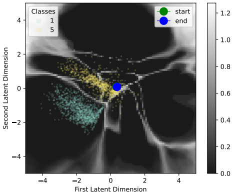
Counterfactual explanations using *Clarity*

Clarity: an explainable Bayesian latent space classifier

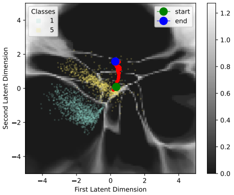
Counterfactual trajectory in a 2D latent space.



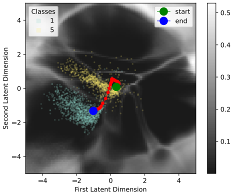
(a) Gradient descent



(b) Schut

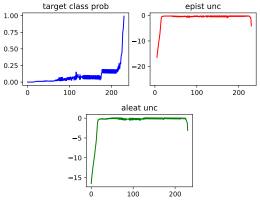


(c) REVISE-E

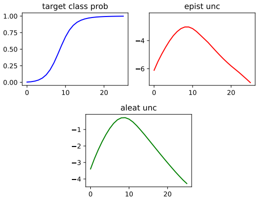


(d) Clarity

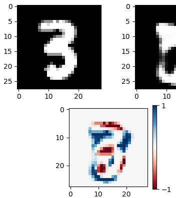
Uncertainty consistency and realism between Clarity and REVISE-E.



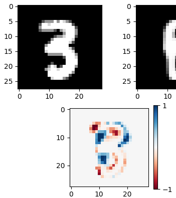
(e) REVISE-E



(f) Clarity



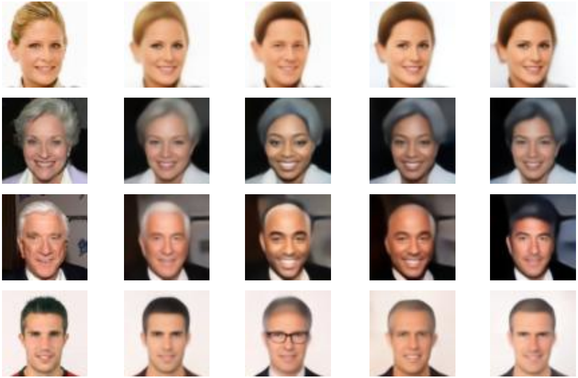
(g) 3 to 8 (REVISE-E)



(h) 3 to 8 (Clarity)

Results on the CelebA dataset.

Original	VAE	R-E (image)	R-E (latent)	Clarity



Conclusion

- There is something to be learned and gained from using classifiers that are explainable by design and can produce realistic explanations.
- This work aims to give insights on the benefits of leveraging the structure of semantic latent space for realistic explanations.
- Our classifiers rely on a latent space of a generative model, which can be further improved.
- Further research can be done in improving the realism of images by linking uncertainty estimates and realism.

References

[1] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law and Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[2] L. Schut, O. Key, R. McGrath, *et al.*, "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties," in *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, A. Banerjee and K. Fukumizu, Eds., ser. Proceedings of Machine Learning Research, vol. 130, PMLR, 2021, pp. 1756–1764. [Online]. Available: http://proceedings.mlr.press/v130/schut21a.html.

References

[3] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, IEEE, 2016, pp. 372–387. DOI: 10.1109/EuroSP.2016.36. [Online]. Available: https://doi.org/10.1109/EuroSP.2016.36.

[4] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *CoRR*, vol. abs/1907.09615, 2019. arXiv: 1907.09615. [Online]. Available: http://arxiv.org/abs/1907.09615.