ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and
Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Clarity: a Deep Ensemble for Visual Counterfactual Explanations

Claire Theobald[1], Frédéric Pennerath[2], Brieuc Conan-Guez[2]
Miguel Couceiro[1], and Amedeo Napoli[1]

1- Université de Lorraine, CNRS, LORIA, F-54000 Nancy France

2- Université de Lorraine, CentraleSupélec, CNRS, LORIA, F-57000 Metz France

Abstract. Counterfactual visual explanations are aimed at identifying changes in an image that will modify the prediction of a classifier. Unlike adversarial images, counterfactuals are required to be realistic. For this reason generative models such as variational autoencoders (VAE) have been used to restrain the search of counterfactuals on the data manifold. However such gradient-based approaches remain limited even when they deal with simple datasets such as MNIST. Conjecturing that these limitations result from a plateau effect which makes the gradient noisy and less informative, we improve the gradient estimation by training an ensemble of classifiers directly in the latent space of VAEs. Several experiments show that the resulting method called Clarity delivers counterfactual images of high-quality, competitive with the state-of-the-art.

1 Introduction

Deep Neural Networks are powerful predictive tools while in many application areas, they are also expected to deliver interpretable results that can be well understood and accepted by human agents. One technique increasing interpretability is based on the generation of *counterfactuals* [1, 2]: given a classifier C, an input X and its predicted class $y = \arg\max(C(X))$ with $C(X)$ the output distribution of the classifier, a *counterfactual* X' of X for a *target class* $y' \neq y$ is an input as close as possible to X but of predicted class $y' = \arg\max(C(X'))$. Moreover we add the constraint that X' must be a realistic/plausible example of class y' according to human perception, which distinguishes a counterfactual from an *adversarial attack*. A counterfactual thus expresses the minimal modifications that must be made to X so that it is interpreted by the classifier and by human beings as belonging to the target class y' instead of y. While counterfactuals are often presented in the context of low dimensional tabular data, the present work falls within the realm of *visual counterfactual explanations*, where inputs X are images. Generating realistic counterfactual explanations in a high dimensional space such as an image space, is challenging. Indeed, a fundamental issue is that we are trying to produce a realistic image representing a high-level concept, *e.g.*, a real-world object, by modifying low-level data (pixels).

In this paper we focus on *gradient-based methods* for generating *visual counterfactuals*. Under this term, we understand methods such as [2, 3, 4] based on gradient descent to find the counterfactual image X', typically by searching an image that substantially maximizes the probability $P(Y = y'|X')$ of the target class y', starting from X. In this context of gradient-based methods, we propose a new approach based on techniques enabling the design of image counterfactuals of high quality while other gradient-based methods fail to do so, even on an image classification problem as simple as MNIST. These techniques help to construct models *explainable by design*, which is highly desirable in application where explainability is needed and expected.

2 Method Clarity

Our approach, called *Clarity*, consists in computing the gradient from an *ensemble* of classifiers *directly trained in the latent space of a VAE.* We show that only the combined contributions of the VAE and of an ensemble model enable a stable computation of the gradient (i.e., reduce its variance), ensure the convergence of the gradient descent and finally result in quality visual counterfactuals.

Similarly to [3], we restrict the space in which an explanation can be generated to the low dimensional latent space $\mathcal{Z}$ of a VAE. Let $q_\theta(z|X)$ be the normal variational posterior which samples a latent variable $z \in \mathcal{Z} \subset \mathbb{R}^d$ from an image $X \in \mathcal{X} \subset \mathbb{R}^k$, with $d \ll k$, thanks to an encoder network parameterized by θ. Let $z \mapsto \mathcal{G}_\psi(z)$ be the generative function, implemented as a decoder network parameterized by ψ, that decodes the latent variable z into the image X, i.e. $\mathcal{G}_\psi(z) \approx X$. Clarity seeks for an optimal latent representation z' of a counterfactual, and then uses the decoder to produce the explanation X'. Moreover, Clarity relies on an ensemble model such as the one used in [4] to improve the counterfactual quality. The idea is to implicitly minimize the epistemic uncertainty associated with the produced counterfactual to ensure a high degree of realism. We denote $(C_m^{latent})_{m=1}^M$ the ensemble of classifiers.

Importantly, our method differs from others in that these classifiers C_m^{latent} are trained from the latent space $\mathcal{Z}$ rather than the image space $\mathcal{X}$. This design improves algorithm convergence but prevents Clarity from being applied to a classifier whose input space is $\mathcal{X}$. Clarity seeks to minimize the objective function defined in Equation 1. In order to obtain a deterministic algorithm (see Algorithm 1), the mean of the variational posterior $q_\theta(z|X)$ is chosen as the starting point of the optimization.

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

where L denotes the cross-entropy loss, d^{latent} is the L1 distance computed in the latent space $\mathcal{Z}$ and z is the encoding of X. The first term of $\mathcal{L}_{Clarity}$ forces the counterfactual to be a realistic element of the target class y' while the second term penalizes the counterfactual from going too far from the original image X. The hyperparameter λ balances these two objectives. The counterfactual is finally obtained by decoding the minimizer z' of $\mathcal{L}_{Clarity}$, *i.e.*, $X' = \mathcal{G}_\psi(z')$.

```
**Algorithm 1** Clarity
**Input:** Original image $X$, output probability $p_m(y' \mid z')$ of classifier $C_m^{latent}$ for target class $y'$, target probability $\gamma$, maximum number of iterations $N$, encoder $q_\theta(z|X) = \mathcal{N}(\mu, \Sigma)$ and decoder $\mathcal{G}_\psi(z)$
**Output:** counterfactual image $X'$
$\mu, \Sigma \leftarrow q_\theta(z|X)$
$z \leftarrow \mu,\ z' \leftarrow z,\ i \leftarrow 0$
**while** $\frac{1}{M}\sum_{m=1}^{M} p_m(y' \mid z') \leq \gamma$ *and* $i \leq N$ **do**
    $S(z', y') = \nabla_{z'} \left( \frac{1}{M} \sum_{m=1}^{M} L(C_m^{latent}(z'), y') + \lambda\, d^{latent}(z, z') \right)$
    $z' \leftarrow optimizer(z', S(z', y'))$
    $i \leftarrow i + 1$
**end**
**return** $X' = \mathcal{G}_\psi(z')$
```

3 Related methods

In 2017, Watcher and al. [2] first propose to generate counterfactuals by minimizing the objective function $\mathcal{L}_{Watcher}(X') = L(C(X'), y') + \lambda\, d(X, X')$. In this first approach, there is no latent space, and therefore the unique classifier C is defined directly in the input space $\mathcal{X}$. This first method is well adapted to tabular data, but fails to converge for high-dimensional data such as images.

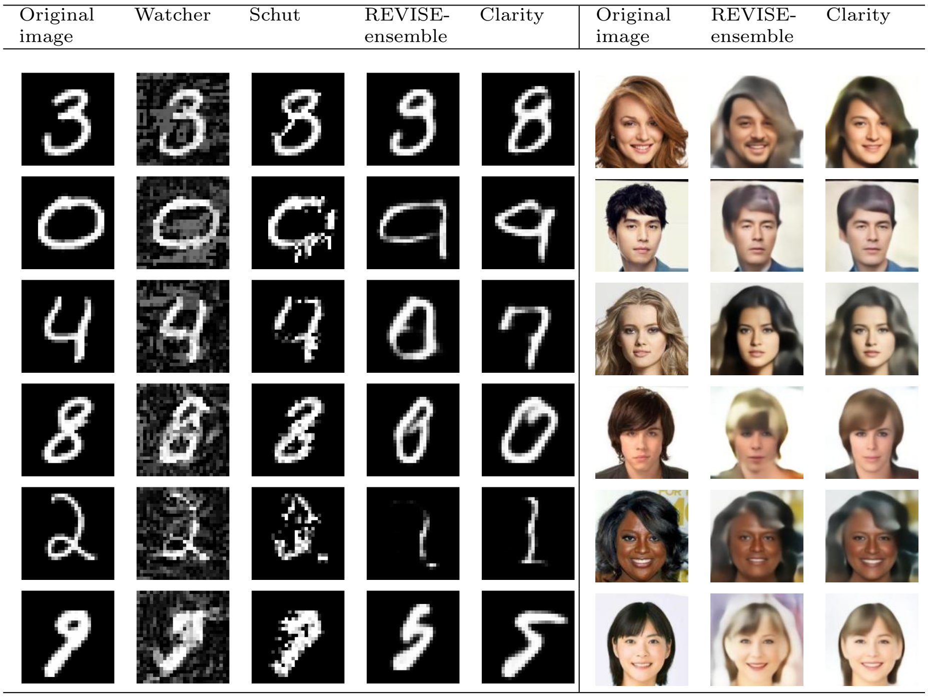
In 2019, Joshi and al. [3] propose the REVISE method, which seeks for the explanation in the latent space of a VAE by optimizing the objective function $\mathcal{L}_{REVISE}(z') = L(C(\mathcal{G}_\psi(z')), y') + \lambda\, d(X, \mathcal{G}_\psi(z'))$. Unlike Clarity, the classifier C of REVISE is defined in the image space $\mathcal{X}$, allowing REVISE to be applied to any classifier. However, as shown in the experiments, this choice has a negative impact on the quality and convergence speed of REVISE. For a fair comparison with Clarity, an ensemble version, called REVISE-ensemble, is also introduced: $\mathcal{L}_{REVISE-e}(z') = \frac{1}{M}\sum_{m=1}^{M} L(C_m(\mathcal{G}_\psi(z')), y') + \lambda\, d(X, \mathcal{G}_\psi(z'))$.

Finally, in 2021, Schut and al. [4] propose to minimize the epistemic uncertainty associated to the counterfactual example. The epistemic uncertainty can be interpreted as a measure of realism and can be estimated thanks to an ensemble model. Schut and al. show that minimizing an objective function composed of a cross entropy term and of an uncertainty term amounts to minimize a simpler objective function: $\mathcal{L}_{Schut}(X') = \frac{1}{M}\sum_{m=1}^{M} L(C_m(X'), y')$. The proximity of the counterfactual to the original image doesn't appear in the objective function, and is implicitly obtained by sparse updates of X'.

4 Experiments

In this section, we compare Clarity and the three other methods on two reference datasets: MNIST (digit prediction) and CelebA (hair color prediction). Due to space constraints, Clarity is only compared with the best performing method on CelebA: REVISE-ensemble. For all methods, the best hyperparameter λ is

Table 1: MNIST dataset on the left: Digit counterfactual explanation. From top to bottom, the explanations are: $3 \rightarrow 8, 0 \rightarrow 9, 4 \rightarrow 7, 8 \rightarrow 0, 2 \rightarrow 1$, $9 \rightarrow 5$. CelebA dataset on the right: Hair color counterfactual explanation. From top to bottom, the explanations are: Brown → Black, Black → Grey, Blond → Black, Brown → Blond, Black → Grey, Black → Blond.


Original image
Watcher
Schut
REVISE-ensemble
Clarity
Original image
REVISE-ensemble
Clarity

chosen by visual inspection. The target probability $P(Y = y'|X')$ is set to 0.99. The dimension of the latent space is 16 for MNIST and 512 for CelebA. Other parameters depend on the dataset. For each method, we train CNN classifiers with adversarial training to improve the quality of the counterfactuals.

In column 2 of Table 1, we observe that Watcher's method is not well suited to image data as it fails to produce examples belonging to the target class. For Schut's method (column 3), counterfactuals are still very noisy. As observed in the original paper [4], this algorithm does not converge for every pair of starting and target classes. Schut's method is best suited to deal with tabular data.

For the MNIST dataset, REVISE-ensemble's counterfactuals are generally better and less noisy thanks to the regularization induced by the latent space (Table 1, column 4). The explanation even preserves handwriting features such as the thickness of the line. But there are still some cases where the counterfactual examples are unrealistic ($8 \rightarrow 0$), are ambiguous ($0 \rightarrow 9$) or do not

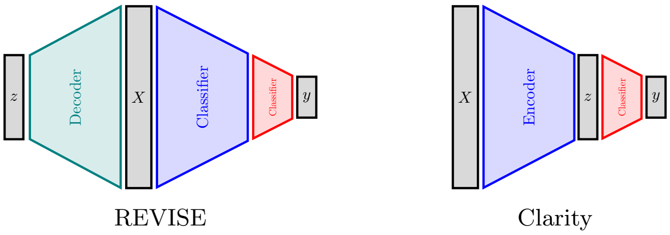


Fig. 1: REVISE's classifier in blue has the same architecture as Clarity's encoder.

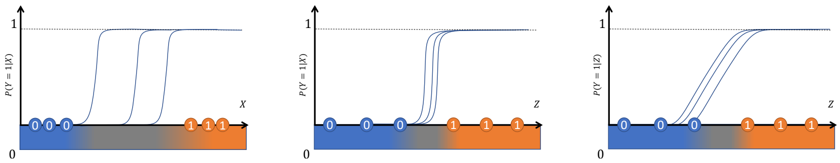
belong to the target class (3 → 8). It should be noticed that REVISE-ensemble only marginally improves the realism of counterfactuals compared to REVISE. (experiments with REVISE are not presented in this paper due to space constraints). This phenomenon is a consequence of the plateau effect described in Section 5. For MNIST, Clarity proposes counterfactuals of higher quality without degrading classifier performance: 98% accuracy on the test set for Clarity's classifier, 99% for a classifier trained in the image space.

For the CelebA dataset, the goal of the explanations is to change the hair color without modifying anything else. On Table 1, we can see that REVISE-ensemble distorts faces more often than Clarity (change gender, add makeup). Clarity is more conservative, so the counterfactual is closer to the original image.

5 Discussion

In this section we justify Clarity's rationale. Figure 1 emphasizes the differences in the architectures of REVISE and Clarity: REVISE uses a classifier in which both blue and red parts are trained for the classification task, whereas only the red part after the encoder is trained in Clarity. Ultimately, the architectures of REVISE's classifier and Clarity's "Encoder + Classifier" are identical. Clarity's architecture offers two advantages over REVISE: Firstly Clarity is an order of magnitude faster than REVISE. Indeed, at each iteration of the optimizer, the decoder (green) and the classifier (blue and red) are evaluated for REVISE, while Clarity only evaluates the classifier (red).

Secondly, Clarity's gradient is less noisy and more informative, which allows a better convergence. This can be explained by considering for instance a binary classification problem, and $X(t)$ an example evolving regularly between classes 0 and 1 with respect to $t \in [0, 1]$. For a classifier trained in the image space (Watcher's and Schut's cases), the target probability function $t \mapsto P(Y = 1 \,|\, X(t))$ is very steep in the low density zone (inter-class transition: grey zone in Figure 2a). This steep variation results in the appearance of a plateau corresponding to probabilities close to 1. In this zone, the algorithm stops prematurely, preventing it from converging properly. For the cases of REVISE and REVISE-ensemble, even if the classifier is viewed from the latent space (C is composed with $\mathcal{G}_\psi$), the plateau effect is still present in the compressed inter-class transition zone of the latent space. Therefore the same



Classifier trained and viewed from image space.

Classifier trained in image space, viewed from latent space.

Classifier trained and viewed from latent space.

Fig. 2: Probability of the target class $Y = 1$ for three different classifiers. The x-axis represents either the image space or latent space. Blue and orange circles represent classes 0 and 1. The grey zone represents the area between classes with out-of-distribution images. This zone is compressed in the latent space.

convergence problem appears (Figure 2b). For Clarity, on the other hand, the target probability function shows smoother variations, as classifiers are trained in the latent space (Figure 2c). As a result, Clarity demonstrates better convergence properties (less local minima), and benefits even more from the ensemble model it is based on (lower gradient variance).

6 Conclusion

In this paper we propose a method based on generative models and ensembles of classifiers, that improves the quality of counterfactual visual explanations. This method contributes to make models more interpretable by design, and capable of generating realistic and unambiguous counterfactual images with minimal changes. Moreover, such counterfactual images are an order of magnitude faster to compute. At a more fundamental level, we give insights why applying a gradient-based method to a classifier trained in the latent space of a VAE is more likely to produce counterfactual images of high quality compared to a classifier trained in the image space. We think that the Clarity approach proposes new and effective directions in the search of counterfactual explanations.

References

[1] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 2022.

[2] S. Watcher, B. D. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31:2018, 2017.

[3] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. 2019.

[4] L. Schut, O. Key, R. McGrath, L. Costabello, B. Sacaleanu, M. Corcoran, and Y. Gal. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1756–1764. PMLR.