Generating Counterfactual Images: Towards a C2C-VAE Approach

Ziwei Zhao*, David Leake, Xiaomeng Ye and David Crandall

Indiana University, Luddy School, Bloomington, Indiana, 47408 U.S.A.

Abstract

Generating semi-factual and counterfactual explanations from images requires methods for extracting and adjusting appropriate image features. This short paper presents initial research on a counterfactual generation method for images based on class-to-class variational autoencoders (C2C-VAEs). Initial experiments illustrate substantial speed increase in counterfactual generation while suggesting that the method achieves reasonable counterfactual quality compared to the state of the art. The paper closes by discussing tradeoffs of the approach.

Keywords

Class-to-class, Counterfactual, Explanation

1. Introduction

Counterfactual explanations contrastively explain the classification of a case with synthetic cases from other classes whose differences illuminate important factors. For example, a counterfactual explanation for being denied a loan might be "you would have received the loan had your salary been 5,000 euros higher." Counterfactual explanation has attracted great interest for its naturalness to people and potential compliance with the European General Data Protection Regulation. Keane et al. [1] identify over 100 current counterfactual explanation methods [1].

Much counterfactual explanation research addresses explanation of tabular data. For such data, features are clearly defined, facilitating adjusting them to generate explanations. However, for image data (e.g., to explain a tumor in an X-ray image), identification and modification of case features is challenging. Kenny and Keane [2] address this with a method, PIECE, that combines a CNN with a GAN to learn latent features, models their distribution, and modifies exceptional features to generate counterfactual and semi-factual images.

PIECE provides strong results but is computationally expensive; in addition, it is not applicable to one-shot learning settings. This short paper presents initial work on an approach to address these limitations, generating counterfactual explanations using a class-to-class variational autoencoder (C2C-VAE) [3], called CVC for **C**2C-**V**AE **C**ounterfactuals.

A class-to-class variational autoencoder (C2C-VAE) learns an embedding space representing

ICCBR XCBR'22: 4th Workshop on XCBR: Case-based Reasoning for the Explanation of Intelligent Systems at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

zz47@iu.edu (Z. Zhao); leake@iu.edu (D. Leake); xiaye@iu.edu (X. Ye); djcran@iu.edu (D. Crandall)

0000-0002-8666-34163 (D. Leake)

the difference patterns between two classes. Given a standard VAE that can extract a feature vector from a case, a C2C-VAE can embed the difference between two feature vectors into a difference embedding or reconstruct a feature difference from a difference embedding. Previously used to generate creative samples from limited data [3], a C2C-VAE can modify a source case s into a target case t, where the line connecting s and t in the embedding space possesses desirable characteristics for generating counterfactuals: (1) Most, if not all, of the line lies within the VAE embedding distribution (so cases on the line are valid), (2) The line follows a straightforward modification between the two classes (modifications are sparse), and (3) The line also allows perturbation of the source or the line itself (cases can be diverse). This paper describes CVC and initial results illustrating tradeoffs between CVC and PIECE, including a substantial speedup using CVC. The paper closes with some future directions. This paper focuses on the generation of counterfactuals, but semi-factuals could be similarly generated.

2. Background

Desiderata for Counterfactuals: For a given query s of class C_s, a counterfactual is a case cf of a different class C_t, and a semi-factual is a case sf of the class C_s. Keane and Smyth [4] propose three criteria for good counterfactuals: A good counterfactual cf avoids prolixity (it is minimally different from the query s), achieves sparsity (it differs from the query in the fewest features) and has plausibility (The counterfactual is realistic for the domain).

The PIECE Approach to Counterfactual Generation: Given a query Image s, a GAN generator G and a CNN classifier to be explained C, PIECE first performs GAN inversion, locating a latent vector z using gradient descent such that $G(z) = s$. PIECE then modifies the penultimate layer output of CNN $x = C(s)$ to x' by identifying exceptional features according to the weight vector of the last layer. The counterfactual output $G(z')$ is generated by optimizing z' such that the MSE loss between x' and $C(G(z'))$ is minimized.

The C2C-VAE Approach: A C2C-VAE [3] is a type of variational autoencoder that learns an embedding space of the difference pattern between features of two classes [3]. Because C2C-VAE works with case features, it requires a feature extractor f for domains in which pre-extracted features are not available (e.g. images) and, if new case generation is desired, a procedure f' to reconstruct a case from a feature vector. Both the feature extractor and case constructor can be implemented by the encoder and decoder of a standard variational autoencoder.

Given a pair of cases s and t, their features are $f(s)$ and $f(t)$, and the feature difference $f_\Delta(s,t) = f(s) - f(t)$. C2C-VAE encodes f_Δ using an encoder function g as $g(< f_\Delta, C_s, C_t >)$ and decodes this embedding using a decoder function g' as $f'_\Delta = g'(g(< f_\Delta, C_s, C_t >)$. With its encoder and decoder, C2C-VAE can sample a new feature difference embedding $g(< f_\Delta, C_s, C_t >)$ from a normal distribution and construct the corresponding feature difference $f_\Delta = g'(g(< f_\Delta, C_s, C_t >)$. C2C-VAE can also synthesize a case t of class C_t by adapting a source case s as $t = f'(f(s) - f_\Delta)$ (See Figure 1).

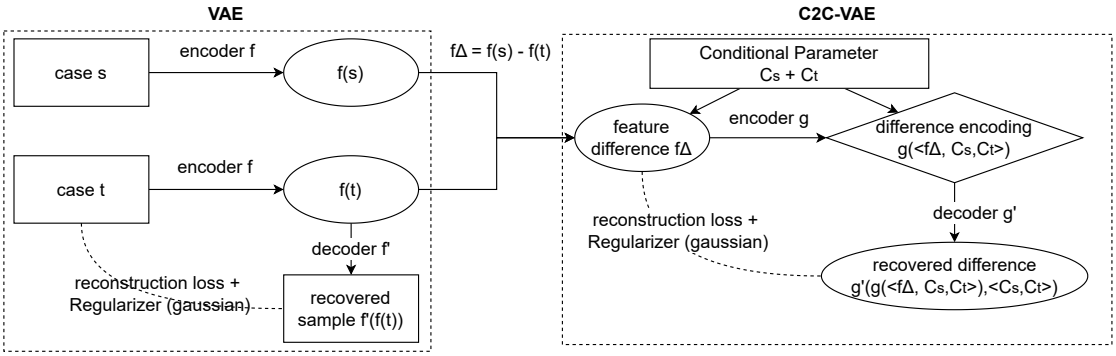


Figure 1: A VAE extracts (recovers) a feature vector from (to) a case. A C2C-VAE extracts (recovers) a difference embedding from (to) a feature difference [3].

(a) (b)

Figure 2: Comparison of average case generated by C2C-VAE (middle) and average case of the target class (right). C2C-VAE learns to change the query class while preserving query characteristics.

3. Using C2C-VAE to Generate Counterfactuals:

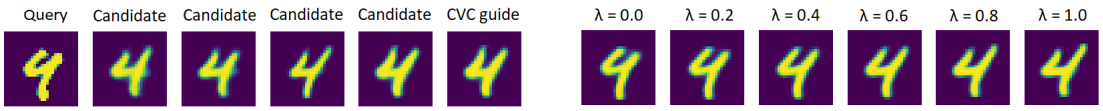
In previous work applying C2C-VAE to creative case generation [3], we noticed that the generated case t often preserves the visual characteristics of s. We hypothesized that this is due to the encoder-decoder pair of C2C-VAE, g and g', learning to recognize the feature differences in $f_\Delta(s,t) = f(s) - f(t)$ more related to class change C_s to C_t, therefore ignoring other less related features. This relates the sparsity needed for a good counterfactual, as shown in Figure 2.

Following the core design of C2C-VAE in Section 2, given a query s of the class C_s, C2C-VAE can be used to generate a guide t of another class C_t. A counterfactual cf of the class C_t can be found on the interpolation between $f(s)$ and $f(t)$ such that: cf is near the boundary of C_s and C_t (avoiding prolixity); cf is generated following an average difference $f_\Delta(s,t)$ pattern (sparsity, explained in the previous paragraph); And cf is interpolated from $f(s)$ and $f(t)$ within the embedding space of a standard VAE, thus conforming to the training data distribution (plausibility). Diversity of cf can be introduced by perturbing either s or $f_\Delta(s,t)$.

3.1. The CVC Algorithm to Generate Counterfactuals

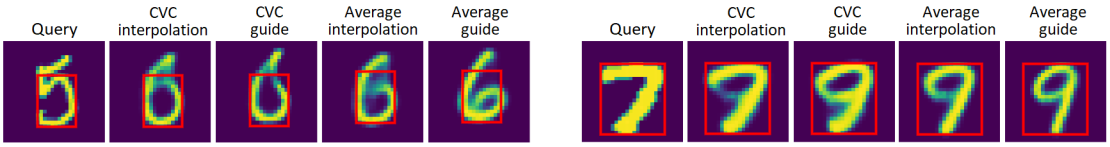
The CVC counterfactual generation algorithm is based on C2C-VAE and the Native Guide technique [5]. CVC has two steps: retrieval and adaptation.

Retrieval: CVC first randomly samples K random vectors $v_1 \dots v_K$ from a normal distribution with mean 0 and standard deviation σ (e.g. $\sigma = 1$) in the feature difference embedding



(a) Retrieval Process of CVC (b) Adaptation Process of CVC

Figure 3: CVC retrieval example with $K = 4$ and $\sigma = 0.5$, In (a), CVC uses C2C-VAE to generate 4 guide candidates from normal distribution $N(0, 0.5)$ and selects the guide case minimizing the difference between query and guide. In (b), CVC adaptation then synthesizes gradually changing semi-factual/counterfactual cases as λ increases.



(a) Preservation of bottom half of the query digit. (b) Preservation of width of the query digit.

Figure 4: Compared to interpolation with average case of the target class, interpolation with the guide case generated by CVC better preserves the visual characteristics of the query image that are less related to class change.

space of C2C-VAE. Given query s, its class C_s and a target class C_t, CVC uses C2C-VAE to generate K feature differences $f'_{\Delta 1} \dots f'_{\Delta K}$ by decoding $v_1 \dots v_K$ using equation 1.

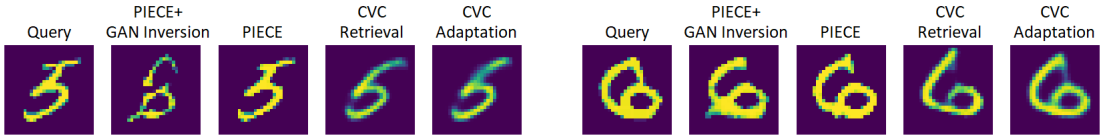
$$f'_{\Delta i} = g'(< v_i, C_s, C_t >) \tag{1}$$

CVC then generates K guide candidate cases $t_1, ..., t_K$ such that $t_i = f'(f(s) - f'_{\Delta i})$, and selects t_i as the guide case t that minimizes mean squared error $MSE(q, t_i)$ in pixel space. Figure 3a shows an example of the retrieval step. In contrast to the guide feature x' of PIECE, the CVC guide t can be directly visualized, therefore providing more explanability.

Adaptation: After selecting guide t and its corresponding VAE feature $f(t) = f(s) - f'_{\Delta i}$, CVC interpolates between $f(s)$ and $f(t)$ in the VAE's latent space to construct counterfactuals cf (or semi-factuals sf), as shown in equation 2.

$$cf = f'((1 - \lambda) * f(s) + \lambda * f(t)), 0 \leq \lambda \leq 1 \tag{2}$$

In equation 2, λ is a variable that determines the relative weight of interpolation between query s and guide t. A small λ value (<0.5) means the output is more similar to s, and a large λ value (>0.5) means the output is more similar to t. As shown in figure 3b, CVC is able to synthesize meaningful results for different λ values. The different values of λ allow CVC to find cf and sf that are closest to the decision boundary. For qualitative evaluation, we found that $\lambda = 0.5$ was suitable to visualize the difference between query, counterfactual and guide in our test domain.



(a) Counterfactual from class 3 to 5

(b) Counterfactual from class 8 to 6

Figure 5: Qualitative comparison of the counterfactual explanation generated by CVC and PIECE.

4. Evaluation

We performed an ablation study comparing adaptation results using (1) the CVC retrieved case and (2) the average case of the target class as guide. Figure 4 illustrates the observed trend that using the CVC-retrieved guide better preserves query characteristics.

We compared CVC to PIECE for the incorrect classifications test-set from Kenny and Keane [2]. We evaluated PIECE under two settings: (1) GAN inversion is accurate and pre-calculated before testing, and (2) GAN inversion is calculated at test time using gradient descent. For (1), we used the official implementation of PIECE provided by its authors. For (2), we implemented gradient descent following the equation in their paper. Because more effective methods of GAN inversion exist, processing time for an application using PIECE would be likely to fall between (1) and (2).

As observed by Kenny and Keane, it is difficult to quantitatively assess the counterfactual desiderata for image data. We measured both efficiency and proximity:

- **Inference time**: Time used to generate each image.
- **SSIM and PSNR**: We calculate Structural Similarity Index (SSIM) and Peak Signal to Noise Ratio (PSNR) between query and generated counterfactual image to measure their proximity.

Table 1 presents the results. Figure 5 illustrates qualitative results; full results are available online.[1] Figure 6 provides a heat map illustration of the differences between Figure 5 results.

Method	Inference time (seconds)	SSIM↑	PSNR↑
PIECE+(GAN inversion)	126.204	0.562	61.757dB
PIECE	25.715	**0.742**	64.067dB
CVC	**0.114**	0.735	**65.556**dB

Table 1
Efficiency and Proximity Results, the best results are highlighted in bold.

5. Conclusion

We proposed a novel counterfactual generation algorithm (CVC) that significantly reduces computational time comparing to the current state-of-the-art. Initial qualitative results suggest

[1]https://drive.google.com/drive/folders/1bZ_oy7eFt7LubmVSXmrOIIenO2yKwlW9?usp=sharing

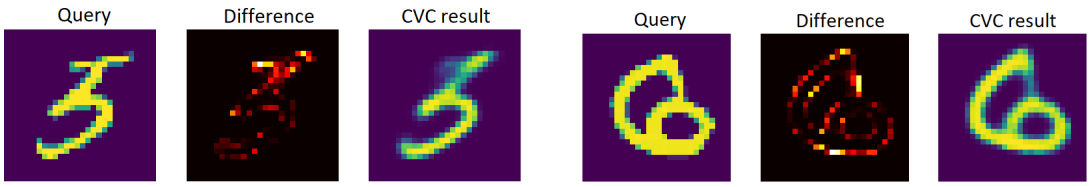


Figure 6: Difference between query and counterfactual image generated by CVC.

comparable quality counterfactuals. As a benefit inherited from C2C-VAE, CVC is applicable to one-shot learning settings (however this is not illustrated in this study). In addition, CVC does not require weight vectors of any layers of the CNN classifier to be explained, so is applicable to any black-box classifier.

This paper presents initial work. Additional evaluation is needed on other data sets and for image quality. The limited image reconstruction quality of the "vanilla" VAE structure we are currently using may prevent CVC from generating plausible results on higher resolution and more realistic images, so our future work will also involve testing other image synthesis models such as GANs. Another direction is to apply CVC to counterfactual generation for tabular data.

6. Acknowledgment

This work was funded by the Department of the Navy, Office of Naval Research (Award N00014-19-1-2655).

References

[1] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, ijcai.org, 2021, pp. 4466–4474.

[2] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2021, pp. 11575–11585.

[3] X. Ye, Z. Zhao, D. Leake, D. Crandall, Generation and evaluation of creative images from limited data: A classs-to-class VAE approach, in: Proceedings of the Thirteenth International Conference on Computational Creativity, 2022. https://computationalcreativity.net/iccc22/conference-program/.

[4] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI), in: Case-Based Reasoning Research and Development, ICCBR-21, Springer, Cham, 2020, pp. 163–178.

[5] E. Delaney, D. Greene, M. T. Keane, Instance-based counterfactual explanations for time series classification, in: Case-Based Reasoning Research and Development, ICCBR-21, Springer, 2021, pp. 32–47.