Sparse Visual Counterfactual Explanations in Image Space

Valentyn Boreiko Maximilian Augustin Francesco Croce Philipp Berens Matthias Hein

University of Tübingen

Abstract

Visual counterfactual explanations (VCEs) are an important tool to understand decisions of image classifiers as they show under which changes of the image the decision of the classifier would change. Their generation in image space is challenging and requires robust models due to the problem of adversarial examples. Existing techniques to generate VCEs in image space suffer from spurious changes in the background. Our novel perturbation model for VCEs together with its efficient optimization via our novel Auto-Frank-Wolfe scheme yields sparse VCEs which lead to subtle changes specific for the target class. Moreover, we show that VCEs can be used to detect undesired behavior of ImageNet classifiers due to spurious features in the dataset.

1. Introduction

Counterfactual explanations (CEs) introduced in [13] is a form of instance-specific explanations. Current approaches to generate CEs for classifier decisions try to answer the question: "What is the minimal change δ of the input x so that the perturbed input $x + \delta$ is classified as the desired target class with sufficiently high confidence and is realistic?". From the developer's perspective, CEs are interesting for debugging as they allow to detect spurious features which the classifier has picked up. In [12], with an extensive overview on the related literature, five criteria for CEs are formulated: i) **validity:** the changed input $x + \delta$ should have the desired target class, ii) **actionability:** the change δ should be possible to be realized by the human, iii) **sparsity:** the change δ should be sparse so that the change is interpretable for humans, iv) **realism:** the changed input $x + \delta$ should lie close to the data manifold, v) **causality:** CEs should maintain causal relations between features. Note that generating CEs for images, visual counterfactual explanations (VCEs), is very similar to that of generating adversarial examples [9, 11].

We make the following contributions: i) we show that the l_2-metric used for the generation of VCEs in [9, 11] leads to changes all over the image (see Fig. 1) which are unrelated to the object, especially for ImageNet models; ii) we propose sparse VCEs based on the l_p-metric for $p = 1.5$. Since an efficient projection onto $l_{1.5}$-balls is not available, we develop a novel Auto-Frank-Wolfe (AFW) optimization scheme with an adaptive step-size for the generation of $l_{1.5}$-VCEs. iii) we illustrate that VCEs can detect spurious features in ImageNet classifiers, e.g., a spurious feature "watermark" in the class granny smith (Sec. 4), showing their usefulness as a "debugging tool" for ML classifiers.

2. Visual Counterfactual Explanations (VCEs)

We assume in the paper that the classifier, $f : \mathbb{R}^d \to \mathbb{R}^K$, outputs for every input x a probability distribution $\hat{p}_f(y|x)$ ($y \in \{1, \dots, K\}$) over the classes. The l_p-distance on $\mathbb{R}^d$ is

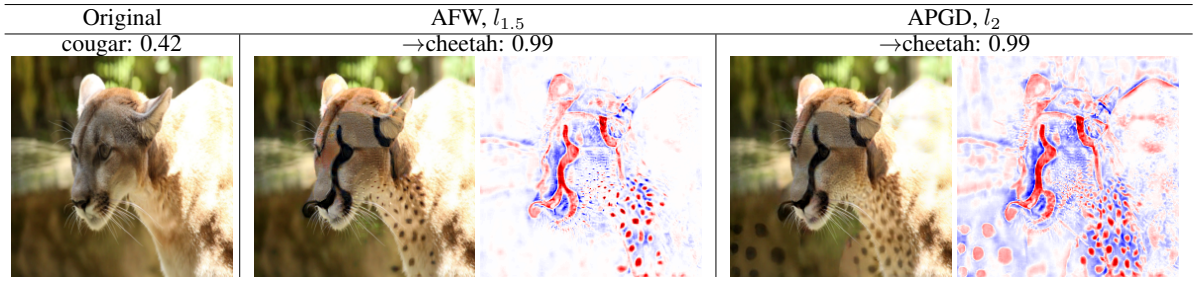

Figure 1. VCEs together with difference maps for the change "cougar ⟶ cheetah" for an adversarially robust ImageNet model [5, 7]. Our novel $l_{1.5}$-VCEs yield more sparse changes which are mainly focused on the object compared to the previously considered l_2-VCEs [9, 11].

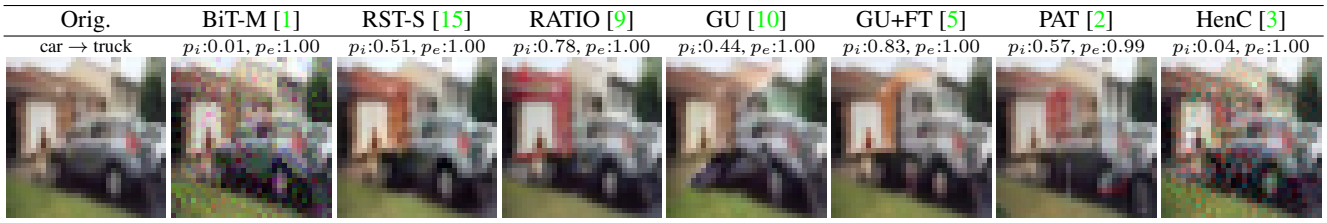

Figure 2. **CIFAR10:** l_2-VCEs of radius $\epsilon = 2.5$ of different classifiers for the change "car → truck" We denote by p_i resp. p_e the confidence in the target class for the original image and the generated VCE. All generated VCEs are valid as high confidence in the target class is achieved but only adversarially robust models, see Tab. 1, show class-specific changes.

defined as: $\|x-y\|_p = \left(\sum_{i=1}^d |x_i - y_i|^p\right)^{\frac{1}{p}}$. We generate VCEs by solving

$$\underset{x\in[0,1]^d \cap B(x_0,\epsilon)}{\arg\max} \quad \log \hat{p}_f(k|x), \tag{1}$$

where $B(x_0,\epsilon) = \{x \in \mathbb{R}^d \,|\, d(x,x_0) \leq \epsilon\}$. The constraint $x \in [0,1]^d$ is necessary to generate valid images. The choice of the distance metric is crucial for the quality of the VCEs (see Sec. 2.2). The parameter ϵ can be interpreted as "perturbation budget" with respect to the chosen metric. We solve it with adversarial attacks, that is either APGD [4] or our novel Frank-Wolfe based method AFW (details below). For both we use 5 random restarts of 75 iterations. For the quantitative evaluation of the image quality of VCEs we use FID scores by generating $10,000$ VCEs from the test set for the in-distribution (ID) evaluation, where the target class is the second most likely class according to an ensemble of all classifiers (see Fig. 2). An evaluation using FID scores on the ID test set only is in our setting problematic, as methods with no (or minimal) change would get the best FID score. Thus, we also generate $10,000$ VCEs from an out-of-distribution (OD) dataset (the first 10k of 80M Tiny Images respectively ImagetNet-A and ImageNet-O) where the target label corresponds to the decision of the ensemble. In our experience from the qualitative inspection of the images, the average of FID scores on ID and OD images reflects best the realism and quality of the VCEs.

2.1. What kind of (adversarial) robustness is required for VCEs?

While previous work [9, 11] has shown that l_2-adversarially robust models lead to realistic VCEs, there has been no study so far about what kind of adversarial robustness in terms of the employed threat model is needed nor if more robust models also have better generative properties. For this purpose, we qualitatively compare different CIFAR-10 classifiers in Fig. 2 and quantitatively in Tab. 1. GU+FT stands for fine-tuning the GU model [5] to get multiple-norm-robust. From Fig. 2 one observes that the two non-robust models BiT-M and HenC do not produce any meaningful counterfactuals. Surprisingly, the RST-s model has some adversarial robustness but its l_2-VCEs do minimal changes to the image, with little class-specific features of the target class. Thus the FID score for the ID is low, but the FID score of the OD is high. Moreover, the PAT-model, trained for robustness with respect to a perceptual distance produces VCEs that show strong artefacts. The

Table 1. **CIFAR-10:** Evaluation of (robust) classifiers for standard accuracy, l_1-, $l_{1.5}$- and l_2-robust accuracy (RA) evaluated at $\epsilon_1 = 12$, $\epsilon_{1.5} = 1.5$, and $\epsilon_2 = 0.5$ respectively (first 1k test points). Further, FID scores for l_1-, $l_{1.5}$-, and l_2-VCEs at $\epsilon_1 = 20$, $\epsilon_{1.5} = 6$, $\epsilon_2 = 2.5$ for in-and out-of-distribution inputs and their average are shown. For all classifiers except RATIO $l_{1.5}$-VCEs attain the best average FID score.

	BiT-M	RST-s	RATIO	GU	GU+FT	PAT	HenC
Acc.	97.4	87.9	94.0	94.7	90.8	82.4	95.8
l_1-RA	0.0	36.5	34.3	33.4	58.0	32.9	0.0
$l_{1.5}$-RA	0.0	70.4	75.4	76.8	76.7	59.2	0.3
l_2-RA	0.0	71.4	79.9	81.7	79.2	62.4	0.1
FID scores for l_1-VCE							
ID	25.1	26.0	24.4	31.1	10.2	29.1	22.7
OD	79.5	72.6	57.8	71.4	52.7	72.2	79.5
Avg.	52.3	49.8	41.1	51.3	31.5	50.6	51.1
FID scores for $l_{1.5}$-VCE							
ID	12.2	8.5	11.7	12.3	9.2	14.4	18.8
OD	62.7	51.6	30.4	52.5	43.4	51.6	62.4
Avg.	**42.5**	**30.1**	19.5	**32.4**	**26.3**	**33.0**	**40.6**
FID scores for l_2-VCE							
ID	55.4	10.3	12.2	15.8	11.9	18.8	37.9
OD	83.9	50.7	26.0	53.9	41.2	49.0	67.2
Avg.	69.7	30.5	**19.1**	34.9	26.7	33.9	52.6

Table 2. **ImageNet:** Accuracy and $l_{1.5}$-, l_2-robust accuracy (RA) at $\epsilon_{1.5} = 12.5, \epsilon_2 = 2$ for the l_2-adv. robust model from Madry [8] and Madry [8]+FT, and FID scores for l_1, $l_{1.5}$- and l_2-VCEs, at $\epsilon_1 = 400$, $\epsilon_{1.5} = 50$, $\epsilon_2 = 12$, generated on in(ID)- and /out-distribution(OD) images and their average. The best FID score are achieved for $l_{1.5}$-VCEs for the Madry [8]+FT model.

	Accuracies			**FID scores (ID/OD/AVG)**		
	Acc.	l_2-RA	$l_{1.5}$-RA	l_1-VCE	$l_{1.5}$-VCE	l_2-VCE
Madry [8]	57.9	45.7	37.4	13.6/41.6/27.6	**8.4**/24.3/16.4	**8.4/22.8/15.6**
[8] +FT	57.5	44.6	40.1	9.6/35.7/22.6	**6.9/22.6/14.8**	7.9/23.1/15.5

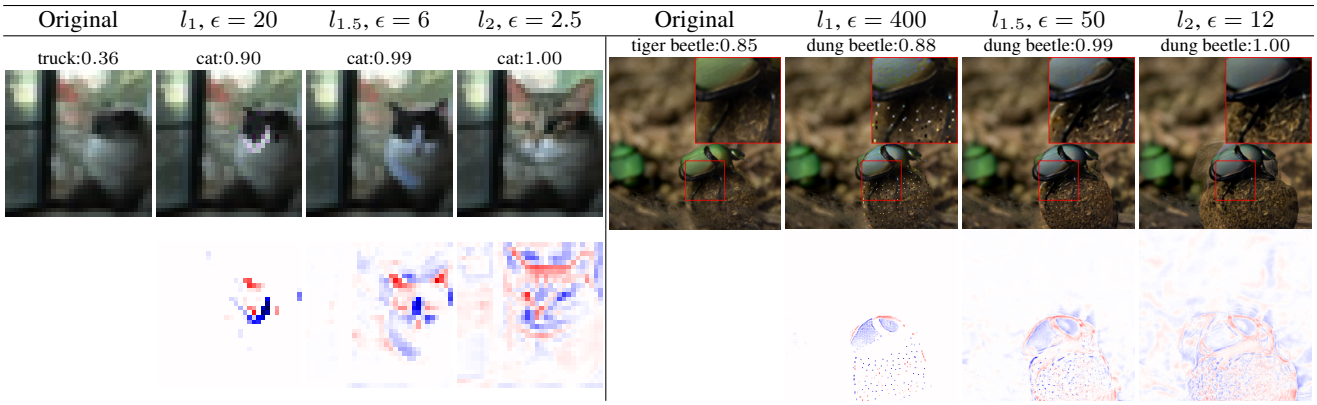

Figure 3. **CIFAR10 (left) and ImageNet (right):** l_p-VCEs into correct class (original images are misclassified) for the multiple-norm adversarially robust models GU+FT (CIFAR10) and Madry [8]+FT (ImageNet). l_1-VCEs are too sparse and introduce artefacts and l_2-VCEs change the background. Our $l_{1.5}$-VCEs are sparse and object-related (see difference maps in the second row).

best VCEs are generated by RATIO, GU and GU+FT, which also have the highest l_2-adversarial robustness. Among them, RATIO and GU+FT produce the most visually realistic VCEs and also have the best FID scores for in- and out-distribution. In particular, the multiple-norm finetuning of the GU model seems to significantly boost the generative properties, both for l_2-VCEs and the $l_{1.5}$-VCEs (see Sec. 2.2).

2.2. Sparse VCEs via the $l_{1.5}$-metric

As a compromise between l_1 (too sparse) and l_2 (non-sparse), we propose to use the $l_{1.5}$-metric for the perturbation model in Eq. (1). Figs. 1 and 3 show that the changes of $l_{1.5}$-VCEs are sparse and localized on the object. The FID scores of the $l_{1.5}$-VCEs (generated with $\epsilon = 6$ for CIFAR-10 and $\epsilon = 50$ for ImageNet) of all classifiers for CIFAR-10 can be found in Tab. 1, where again the RATIO and GU+FT model work best. Both images and FID scores indicate that $l_{1.5}$-VCEs have higher **realism** and **sparsity** than l_1- and l_2-VCEs. The ID FID score of the non-robust BiT-M model is surprisingly good, but this is an "artefact" of the optimization reaching maximum confidence in the interior of the $l_{1.5}$-ball resulting in rather small changes. The OD FID reveals that the quality of the generated VCEs is, as expected, low. This shows that the quantitative evaluation of VCEs using FID scores has to be done with great care. The results for ImageNet for the l_2-robust model of Madry [8] and Madry [8]+FT are in Tab. 2.

3. Auto-Frank-Wolfe for l_p-VCEs

For deep models, the optimization problem for l_p-VCEs

$$\max_{x \in B_p(x_0,\epsilon) \cap [0,1]^d} \log \hat{p}(y|x), \tag{2}$$

is non-convex and related to targeted adversarial attacks, for which AutoPGD (APGD) [4] has been shown to be very effective. APGD requires projections onto l_p-balls which are available either in closed form for l_2 and l_∞ or can be computed efficiently for l_1. In other cases, there is no such projection available. Thus, in order to generate l_p-VCEs for $p > 1$, we propose an adaptive version of the Frank-Wolfe (FW) algorithm [6], named Auto-Frank-Wolfe (AFW). FW has the advantage that it is projection-free and thus allows to use arbitrary l_p norm balls for $p > 1$ or their intersection with $[0,1]^d$ which is required for l_p-VCEs. At each iteration k, FW maximizes the first-order Taylor expansion at the iterate x^k of the objective in the feasible set, i.e.

$$s^k = \operatorname*{arg\,max}_{s \in B_p(x_0,\epsilon) \cap [0,1]^d} \left\langle s, \nabla_{x^k} \log \hat{p}(y|x^k) \right\rangle, \tag{3}$$

and the next iterate is the convex combination

$$x^{k+1} = (1 - \gamma^k) x^k + \gamma^k s^k. \tag{4}$$

The choice of the learning rate $\gamma^k \in (0,1)$ is crucial: in the context of adversarial attacks, [6] use a fixed value γ_0 for every k, while [14] decrease it as $\frac{\gamma_0}{\gamma_0 + k}$. In both cases the schedule is agnostic of the total budget of iterations, and γ_0 needs to be tuned, while we choose γ^k adaptively following [4]. Also, [6, 14] do not consider the image domain constraints $[0,1]^d$ but rather solve Eq. (3) for l_p-ball constraints only (which has a closed form solution) and clip it to $[0,1]^d$. This is suboptimal, especially when p is close to 1. The following proposition shows that it is possible to solve Eq. (3) efficiently in the intersection $B_p(x_0,\epsilon) \cap [0,1]^d$ for $p > 1$.

Proposition 3.1 *Let $w \in \mathbb{R}^d$, $x \in [0,1]^d$, $\epsilon > 0$ and $p > 1$. The solution δ^* of the optimization problem*

$$\operatorname*{arg\,max}_{\delta \in \mathbb{R}^d} \langle w, \delta \rangle \quad \textit{s.th.} \ \|\delta\|_p \leq \epsilon, \ x + \delta \in [0,1]^d \tag{5}$$

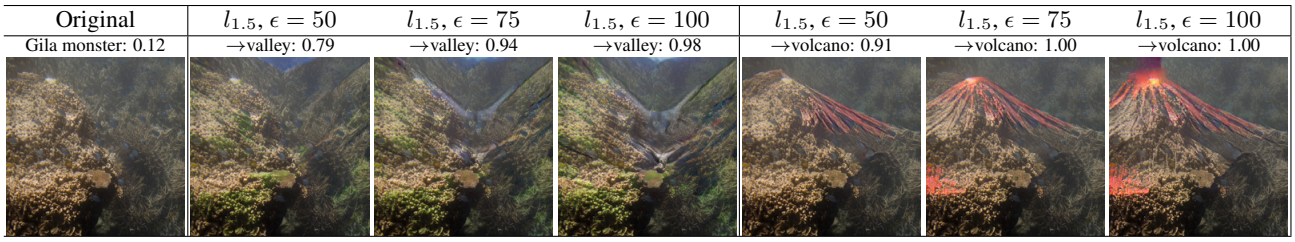

Figure 4. $l_{1.5}$-VCEs for Madry [8]+FT with varying radii for a misclassified image of class "coral reef" for the target classes: "coral reef", "cliff", "valley" and "volcano" (same wordnet category "geological formation").

is given, with the convention $\operatorname{sign} 0 = 0$*, by*

$$\delta_i^* = \min \left\{ \gamma_i, \left(\frac{|w_i|}{p\mu^*} \right)^{\frac{1}{p-1}} \right\} \operatorname{sign} w_i, \quad i = 1, \ldots, d,$$

where $\gamma_i = \max\{-x_i \operatorname{sign} w_i, (1 - x_i) \operatorname{sign} w_i\}$ *and* $\mu^* > 0$*, and can be computed in* $O(d \log d)$ *time.*

4. Finding spurious features with $l_{1.5}$-VCEs

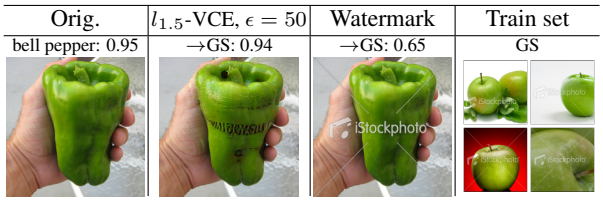

Figure 5. The $l_{1.5}$-VCE with target class "granny smith" (GS) for Madry [8]+FT shows that the model has associated a spurious "text" feature with this class. This is likely due to "iStockphoto" watermarked images in its training set (right). Adding the watermark changes the decision of the classifier to GS.

Failure A, Watermark text as spurious feature for "granny smith": We detected this failure when creating VCEs for the target class "granny smith". We consistently observed text-like features on the generated $l_{1.5}$-VCEs which are not related to this class. In Fig. 5 we illustrate the $l_{1.5}$-VCE for an image from the class "bell pepper".

Failure B, Cages as spurious feature for "white shark": The next failure was detected using $l_{1.5}$-VCEs for the shark classes where frequently grid-like structures appear - but only for VCEs with target class "white shark". A typical situation is shown in Fig. 6, where the original image is from class "coral reef".

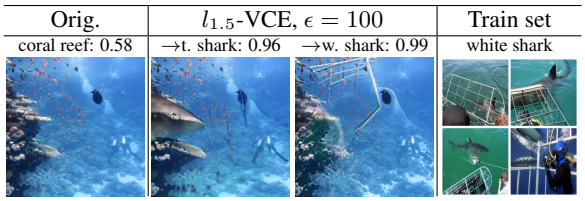

Figure 6. The $l_{1.5}$-VCE for an image from class coral reef with target "tiger shark" shows a tiger shark, but with target "white shark" grid-like structures as spurious feature. The training set of white shark (right) contains many images with cages.

References

[1] A. Kolesnikov et al. Big transfer (bit): General visual representation learning. In *ECCV*, 2020.

[2] C. Laidlaw et al. Perceptual adversarial robustness: Defense against unseen threat models. In *ICLR*, 2021.

[3] D. Hendrycks et al. AugMix: A simple data processing method to improve robustness and uncertainty. *ICLR*, 2020.

[4] F. Croce et al. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

[5] F. Croce et al. Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model. *arXiv preprint arXiv:2105.12508*, 2021.

[6] J. Chen et al. A Frank-Wolfe framework for efficient and effective adversarial attacks. In *AAAI*, 2019.

[7] L. Engstrom et al. Adversarial robustness as a prior for learned representations, 2019.

[8] L. Engstrom et al. Robustness (python library), 2019.

[9] M. Augustin et al. Adversarial robustness on in- and out-distribution improves explainability. In *ECCV*, 2020.

[10] S. Gowal et al. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593v2*, 2020.

[11] S. Santurkar et al. Image synthesis with a single (robust) classifier. In *NeurIPS*, 2019.

[12] S. Verma et al. Counterfactual explanations for machine learning: A review. *arXiv preprint, arXiv:2010.10596*.

[13] S. Wachter et al. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *HJLT*, 2018.

[14] T. Tsiligkaridis et al. Understanding frank-wolfe adversarial training. *arXiv preprint arXiv:2012.12368*, 2020.

[15] Y. Carmon et al. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019.