CounteRGAN: Generating Counterfactuals for Real-Time Recourse and Interpretability using Residual GANs (Supplementary material)

Daniel Nemirovsky[1] **Nicolas Thiebaut**[2] **Ye Xu**[3] **Abhishek Gupta**[3]

[1] Amazon, Seattle, WA, U.S.A., nemird@amazon.com
[2] Hired, New York, NY, U.S.A., nicolas.thiebaut@hired.com
[3] Meta, Menlo Park, CA, U.S.A., {yexu, abigupta}@fb.com

1 ADDITIONAL EXPERIMENT: PIMA INDIANS DIABETES DATASET

Following the experiments in Wachter et al. [2017b], we utilize the Pima Indians Diabetes dataset (Smith et al. [1988]). It is composed of low dimensional tabular data and helps to validate the CounteRGAN's versatility and its applicability to diverse use cases. The dataset contains 8 features describing the relevant characteristics of patients useful for predicting diabetes. The target label is positive if the patient has diabetes (268 examples) and negative otherwise (500 examples). We use stratified (label balanced) sampling with 80% of the dataset being assigned to the train set and the remaining 20% for the test set. The classifier is the same as the neural network architecture used in Wachter et al. [2017b] and achieves an accuracy of 74.68% on the test set.

	White-box classifier				Black-box classifier		
	RGD	CSGP	GAN	CounteRGAN	RGD	CSGP	CounteRGAN
↑ Prediction gain	0.15 ± 0.01	0.13 ± 0.02	0.15 ± 0.03	**0.33 ± 0.04**	**0.17 ± 0.00**	0.13 ± 0.00	**0.16 ± 0.02**
↓ Realism	2.20 ± 0.24	2.03 ± 0.11	3.33 ± 0.11	**1.79 ± 0.11**	2.22 ± 0.01	**1.98 ± 0.01**	2.13 ± 0.12
↓ Actionability	1.64 ± 0.20	**1.14 ± 0.19**	9.46 ± 0.53	6.91 ± 0.43	1.75 ± 0.02	**1.29 ± 0.02**	2.97 ± 0.12
↓ Latency (ms)	1,195.91 ± 5.65	3,211.67 ± 11.65	1.68 ± 0.06	**1.51 ± 0.03**	2,525.99 ± 1.23	15,921 ± 23.66	**1.82 ± 0.12**
↓ Batch latency (s)	204.58	483.88	0.26	**0.23**	453.45	2,228.23	**0.32**

Table 1: Diabetes test data results (mean and 95% confidence interval). The arrows indicate whether larger ↑ or lower ↓ values are better, and the best results are in bold. The realism metric typically ranges from 1.84 (mean reconstruction error on the test set) to 2.44 (reconstruction error on random Gaussian noise). Computations are performed using the entire test set (154 samples).

For this experiment we introduce the important concept of *mutable* and *immutable features*. For most practical applications of counterfactual search, certain features may be hard or impossible to change and can be considered immutable. Though features typically vary in their degree of mutability, for the purposes of this experiment we consider features as either mutable or immutable. For the Pima Indians Diabetes dataset, we consider *Pregnancies*, *Age*, and *Diabetes Pedigree Function* features to be immutable. We use *Glucose*, *Insulin*, *Body Mass Index*, *Tricept Skin Fold Thickness*, and *Blood Pressure* as mutable features. In practice, we apply counterfactual search with no modifications, then simply cancel the perturbations applied to immutable features.

Table 1 summarizes our findings for this experiment. On this dataset, all methods appear equally capable of improving classifier prediction gain. The CounteRGAN generates more realistic instances, and the CSGP outputs the sparsest counterfactuals. Even on this low-dimensional dataset, the CounteRGAN is able to meet or exceed the evaluation metrics of counterfactuals produced by existing methods while heavily outperforming them in terms of latency. This includes >1,000x to >2,000x improvements for individual counterfactuals on white-box and black-box models respectively and from 3 to 4 orders of magnitude for batch generation of all counterfactuals.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

The evaluation results validate that the proposed CounteRGAN method is capable of overcoming the main limitations of existing methods, namely the lack of realism and high latency. It also provides similar or better prediction gain and actionability on high dimensional images and a low-dimensional tabular dataset. The impressive latency improvements are pivotal with regard to real-time applicability and scalability. This is due to the generator only needing a forward-pass through the neural network as opposed to performing a new counterfactual search for every data point, as required by existing methods.

2 PROOF OF THEOREM 1

Theorem 1. *If the discriminator is systematically allowed to reach its optimum, and the generator has sufficient capacity, then the minimax optimization of the value function*

$$\mathcal{V}_{\mathrm{CounteRGAN-wt}}(D,G) = \frac{\sum_i C_t(x_i)\log D(x_i)}{\sum_i C_t(x_i)} + \frac{1}{N}\sum_i \log\left(1 - D(x_i + G(x_i))\right), \tag{1}$$

converges to the Nash equilibrium. The full generator's output distribution p_{g_+} converges to a distribution p_{C_t} defined by

$$p_{C_t}(x) = \mathcal{N}_t\ C_t(x)\ p_{\mathrm{data}}(x), \tag{2}$$

where N_t is a normalization constant.[1]

Proof. We first introduce the full generator output function $G_+(x) = x + G(x)$, and note that the value function defined by equation 1 can be written as

$$\mathcal{V}_{\mathrm{CounteRGAN-bb}}(D,G) = \mathbb{E}_{x\sim p_{C_t}} \log D(x) + \mathbb{E}_{x\sim p_{g_+}} \log\left(1 - D(x)\right), \tag{3}$$

since the first term on the r.h.s. of Equation 1 is a weighted sampling estimate of $\mathbb{E}_{x\sim p_{C_t}} \log D(x)$, and for the second term, the equality $\mathbb{E}_{x\sim p_{g_+}} \log\left(1 - D(x)\right) = \mathbb{E}_{x\sim p_{\mathrm{data}}} \log\left(1 - D(G_+(x))\right)$ is a consequence of the Radon–Nikodym theorem.

From the expression of the value function in equation 3, Proposition 1 of Goodfellow et al. [2014a] implies that for any generator G the optimal discriminator is

$$D^*(x) = \frac{p_{C_t}(x)}{p_{g_+}(x) + p_{C_t}(x)}. \tag{4}$$

The value function for an ideal discriminator thus reads:

$$\mathcal{V}^*(G) = \mathcal{V}(D^*,G) = \mathbb{E}_{x\sim p_{C_t}} \log \frac{p_{C_t}(x)}{p_{g_+}(x) + p_{C_t}(x)} + \mathbb{E}_{x\sim p_{g_+}} \log \frac{p_{g_+}(x)}{p_{g_+}(x) + p_{C_t}(x)}. \tag{5}$$

To find the distribution $p^*_{g_+}$ that minimizes $\mathcal{V}^*$ under the probability normalization constraint, $\int p_{g_+}(x)\mathrm{d}x = 1$, we introduce a Lagrange multiplier μ. We then compute the functional derivative of $\mathcal{V}^*$ with respect to p_{g_+} using the shortened notation for $p = p_{C_t}(x)$ and $q = p_{g_+}(x)$ in the following equation

$$\frac{\delta \mathcal{V}^*}{\delta q} = \frac{\partial}{\partial q}\left[p\log\left(\frac{p}{p+q}\right) + q\log\left(\frac{q}{p+q}\right) + \mu q\right] = \log\left(\frac{q}{p+q}\right) + \mu. \tag{6}$$

The optimum of $\mathcal{V}^*$ is attained for

$$\frac{\delta V}{\delta p^*_{g_+}}(x) = 0 \quad \Longleftrightarrow \quad p^*_{g_+}(x) = \frac{p_{C_t}(x)}{\exp(\mu) - 1}, \tag{7}$$

from which the normalization constraint leads to

$$\int \frac{p_{C_t}(x)}{\exp(\mu) - 1}\mathrm{d}x = 1 \quad \Longleftrightarrow \quad \exp(\mu) = 2, \tag{8}$$

[1] Explicitly, $\mathcal{N}_t = \left(\int C_t(x)\ p_{\mathrm{data}}(x)\mathrm{d}x\right)^{-1}$ but it doesn't need to be computed for our purpose.

such that

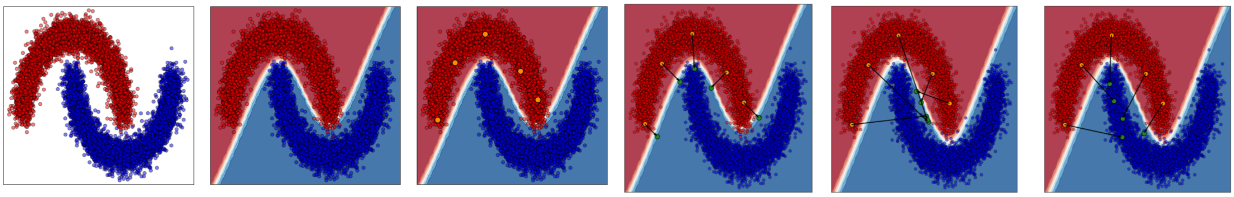
$$p^*_{g_+}(x) = p_{C_t}(x) \tag{9}$$

for all x. Hence $\mathcal{V}^*$ has a unique optimum[2] that is reached when

$$p^*_{g_+} = p_{C_t}. \tag{10}$$

The fact that p_{g_+} converges to the optimum when using the alternating gradient updates follows from Proposition 2 in Goodfellow et al. [2014a]. □

3 SYNTHETIC DATASET EXAMPLE



(a) Original distribution of data points. **(b)** Decision boundary of trained classifier. **(c)** Data points for counterfactuals search. **(d)** Regularized gradient descent (RGD). **(e)** Standard GAN. **(f)** CounterGAN.

Figure 1: Comparing how three different counterfactual search techniques are able to achieve their objectives while producing significantly different counterfactuals on a synthetic and binary class dataset.

Figure 1 provides an example of counterfactual search using a synthetic dataset meant to illustrate the challenges faced by counterfactual generation methods. The data points shown in (a) can be interpreted as the known populations from two different societies (red/blue). An ML classifier has been trained to predict the type of society a person belongs to based on their weight (x-axis) and height (y-axis). The solid white line in (b) represents the classifier's decision boundary such that all predictions for points falling within the red shaded region are classified as persons belonging to the red society and vice-versa. The five selected orange points in (c) represent persons from the red society we seek to provide counterfactuals for. These counterfactuals should provide meaningful recourse regarding how to turn themselves into realistic looking persons of the blue society, as predicted by the classifier. The counterfactuals generated by an existing method (d) produce the correct classification result (blue) but the suggested changes would mean that the transformed individuals would not look like the rest of the known populace of the blue society (lack of realism). Using a standard GAN, the counterfactuals always result in the same or similar looking persons of the blue society. While these results are more realistic than those obtained with the previous method, the suggested changes may be harder to apply to some original persons than others (i.e., lower sparsity) and hence less actionable. The proposed CounteRGAN method (f) results in counterfactuals that are of the desired classification (blue) and are most realistic and actionable than those obtained with previous methods. Red society members seeking to imperceptibly infiltrate the blue society would benefit the most from the meaningful recourse provided by this method.

4 CODE

The corresponding code to reproduce all the results and methods will be available by the date of publication.

References

Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38, April 2018.

[2] The optimum is a minimum here since $\mathcal{V}^*$ is a convex functional of p_{g_+}, as can be seen from the form of the second functional derivative $\frac{\delta^2 V}{(\delta p^*_{g_+})^2}(x) = \frac{p_{C_t}(x)}{p_{g_+}(x)(p_{g_+}(x)+p_{C_t}(x))}$, which is always positive.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. January 2017.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2018. doi: 10.1109/wacv.2018.00097. URL `http://dx.doi.org/10.1109/WACV.2018.00097`.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016.

Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. June 2015.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 592–603. Curran Associates, Inc., 2018.

Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*. Citeseer, 2012.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. December 2014b.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. June 2016.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. March 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. November 2016.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems, 2019.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, January 2020.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2280–2288. Curran Associates, Inc., 2016.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. December 2017.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

David K Lewis. Causation. *J. Philos.*, 70(17):556–567, 1973.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.*, 19(6):1236–1246, November 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. November 2014.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 607–617, New York, NY, USA, January 2020. Association for Computing Machinery.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 2642–2651. JMLR.org, August 2017.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. *Proceedings of The Web Conference 2020*, Apr 2020. doi: 10.1145/3366423.3380087. URL `http://dx.doi.org/10.1145/3366423.3380087`.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. September 2019.

ProPublica. Compas, 2017. URL `https://github.com/propublica/compas-analysis/`.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. November 2015.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. February 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7 (4):233–242, December 2017.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL `http://dx.doi.org/10.1007/s11263-019-01228-7`.

Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. December 2016.

Jack W Smith, J E Everhart, W C Dickson, W C Knowler, and R S Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261, November 1988.

Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. October 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. March 2017.

M Tavakolian, C G Bermudez Cruces, and A Hadid. Learning to detect genuine versus posed pain from facial expressions using residual generative adversarial networks. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–8, May 2019.

N Tollenaar and P G M van der Heijden. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 176(2):565–584, 2013.

Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. July 2019.

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated Decision-Making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, June 2017a.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. November 2017b.

Jifeng Wang, Xiang Li, Le Hui, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. December 2017.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. December 2016.

Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. RIS-GAN: Explore residual and illumination with generative adversarial networks for shadow removal. November 2019.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image translation using Cycle-Consistent adversarial networks, 2017a.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal Image-to-Image translation. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 465–476. Curran Associates, Inc., 2017b.