Counterfactual Training: Teaching Models Plausible and Actionable Explanations

Patrick Altmeyer[1][0000−0003−4726−8613] (✉), Arie van Deursen[1], and Cynthia C. S. Liem[1]

Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science {p.altmeyer}@tudelft.net

Abstract. Counterfactual Explanations (CE) have emerged as a popular method to explain the predictions made by opaque machine learning models in a post-hoc fashion. We propose a novel approach that leverages counterfactuals during the training phase of models.

Keywords: Counterfactual Explanations · Explainable AI

1 Introduction

2 Related Literature

2.1 Background on Counterfactual Explanations

[11, 5, 2]

2.2 Learning Representations

For example, joint-energy models

2.3 Generalization and Robustness

[9] generate counterfactual images for MNIST and ImageNet through independent mechanisms (IM): each IM learns class-conditional input distributions over a specific lower-dimensional, semantically meaningful factor, such as *texture*, *shape* and *background*. The demonstrate that using these generated counterfactuals during classifier training improves model robustness. Similarly, [1] argue that counterfactuals represent potentially useful training data in machine learning, especially in supervised settings where inputs may be reasonably mapped to multiple outputs. They, too, demonstrate the augmenting the training data of image classifiers can improve generalization.

[10] propose an approach using counterfactuals in training that does not rely on data augmentation: they argue that counterfactual pairs typically already exist in training datasets. Specifically, their approach relies on, firstly, identifying similar input samples with different annotations and, secondly, ensuring

that the gradient of the classifier aligns with the vector between pairs of counterfactual inputs using the cosine distance as a loss function (referred to as *gradient supervision*) (***this might be useful for our task as well***). In the natural language processing (NLP) domain, counterfactuals have similarly been used to improve models through data augmentation: [12], propose POLYJUICE, a general-purpose counterfactual generator for language models. They demonstrate empirically that augmenting training data through POLYJUICE counterfactuals improves robustness in a number of NLP tasks.

2.4 Link to Adversarial Training

[3] propose two definitional differences between Adversarial Examples (AE) and Counterfactual Explanations (CE): firstly, and more importantly according to the authors, the term AE implies missclassification, which is not the case for CE (***this might be a useful notion for use to distinguish between adversarials and explanations during training***); secondly, they argue that closeness plays a more critical role in the context of CE but confess that even counterfactuals that are not close might be relevant explanations. [7] show that CE and AE are equivalent under certain conditions and derive upper bounds on the distances between them.

2.5 Closely Related

[4] are the first to propose end-to-end training pipeline that includes counterfactual explanations as part of the training prodeduce. In particular, they propose a specific network architecture that includes a predictor and CE generator network (***akin a GAN?***), where the parameters of the CE generator network are learnable. Counterfactuals are generated during each training iteration and fed back to the predictor network (***here we are aligned***). In contrast, we impose no restrictions on the neural network architecture at all. (***to ensure the one-hot encoding of categorical features is maintained, they simple use softmax (might be interesting for CE.jl)***) Interestingly, the authors find that their approach is sensitive to the choice of the loss function: only MSE seems to lead to good performance. They also demonstrate theoretically, that the objective function is difficult to optimize due to divergent gradients and suffers from poor adversarial robustness. (***because partial gradients with respect to the classification loss component and the counterfactual validity component point in opposite directions***). To mitigate these issues, the authors use block-wise gradient descent: they first update with respect to classification loss and then use a second update with respect to the other loss components (***this might be useful for our task as well***). [8] propose a way to train models that are guaranteed to provide recourse for individuals with high probability. The approach builds on adversarial training (***here we are aligned***), where in this context adversarial examples are actively encouraged to exist, but only target attacks with respect to the positive class. The proposed method allows for imposing a set of actionable recourse ex-ante: for example, users can

impose mutability constraints for features (***here we are aligned***). (***To solve their objective function more efficiently, they use a first-order Taylor approximation to approximate the recourse loss component (might be applicable in our case)***)

[6] introduce Counterfactual Adversarial Training (CAT) with intention of improving generalization and robustness of language models. Specifically, they propose to proceed as follows: firstly, identify training samples that are subject to high predictive uncertainty (entropy); secondly, generate counterfactual explanations for those samples; and, finally, finetune the model on the augmented dataset that includes the generated counterfactuals.

3 Counterfactual Training

4 Experiments

4.1 Experimental Setup

4.2 Experimental Results

5 Discussion

6 Conclusion

Acknowledgments. A bold run-in heading in small font size at the end of the paper is used for general acknowledgments, for example: This study was funded by X (grant number Y).

Disclosure of Interests. It is now necessary to declare any competing interests or to specifically state that the authors have no competing interests. Please place the statement with a bold run-in heading in small font size beneath the (optional) acknowledgments, for example: The authors have no competing interests to declare that are relevant to the content of this article. Or: Author A has received research grants from Company W. Author B has received a speaker honorarium from Company X and owns stock in Company Y. Author C is a member of committee Z.

Bibliography

[1] Abbasnejad, E., Teney, D., Parvaneh, A., Shi, J., van den Hengel, A.: Counterfactual vision and language learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10041–10051 (2020). https://doi.org/10.1109/CVPR42600.2020.01006

[2] Altmeyer, P., Farmanbar, M., van Deursen, A., Liem, C.C.: Faithful model explanations through energy-constrained conformal counterfactuals. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10829–10837 (2024)

[3] Freiesleben, T.: The intriguing relation between counterfactual explanations and adversarial examples. Minds and Machines **32**(1), 77–109 (2022)

[4] Guo, H., Nguyen, T.H., Yadav, A.: Counternet: End-to-end training of prediction aware counterfactual explanations. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 577–589. KDD '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3580305.3599290, https://doi.org/10.1145/3580305.3599290

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

[6] Luu, H.L., Inoue, N.: Counterfactual adversarial training for improving robustness of pre-trained language models. In: Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation. pp. 881–888 (2023)

[7] Pawelczyk, M., Agarwal, C., Joshi, S., Upadhyay, S., Lakkaraju, H.: Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In: Camps-Valls, G., Ruiz, F.J.R., Valera, I. (eds.) Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 4574–4594. PMLR (28–30 Mar 2022), https://proceedings.mlr.press/v151/pawelczyk22a.html

[8] Ross, A., Lakkaraju, H., Bastani, O.: Learning models for actionable recourse. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2024)

[9] Sauer, A., Geiger, A.: Counterfactual generative networks (2021), https://arxiv.org/abs/2101.06046

[10] Teney, D., Abbasnedjad, E., van den Hengel, A.: Learning what makes a difference from counterfactual examples and gradient supervision. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 580–599. Springer (2020)

[11] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech. **31**, 841 (2017). https://doi.org/10.2139/ssrn.3063289

[12] Wu, T., Ribeiro, M.T., Heer, J., Weld, D.: Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 6707–6723. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.523, https://aclanthology.org/2021.acl-long.523