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 tool to explain predictions made by opaque machine learning models: they explain how factual inputs need to change in order for some fitted model to produce some desired output. Much existing research has focused on identifying explanations that are not only valid but also deemed desirable with respect to the underlying data and stakeholder requirements. Recent work has shown that under this premise, the task of learning desirable explanations is effectively reassigned from the model itself to the (post-hoc) counterfactual explainer. Building on that work, we propose a novel model objective that leverages counterfactuals during the training phase (ad-hoc) in order to minimize the divergence between learned representations and desirable explanations. Through extensive experiments, we demonstrate that our proposed methodology facilitates training models that inherently deliver desirable explanations while maintaining high predictive performance.

Keywords: Counterfactual Explanations · Explainable AI · Representation Learning

1 Introduction

Today's prominence of artificial intelligence (AI) has largely been driven by advances in **representation learning**: instead of relying on features and rules that are carefully hand-crafted by humans, modern AIs are tasked with learning these representations from scratch, guided by narrow objectives such as predictive accuracy [4]. Modern advances in computing have made it possible to provide such AIs with ever greater degrees of freedom to achieve that task, which has often led them to outperform traditionally more parsimonious models. Unfortunately, in doing so they also learn increasingly complex and highly sensitive representations that we can no longer easily interpret.

This trend towards complexity for the sake of performance has come under serious scrutiny in recent years. At the very cusp of the deep learning revolution, [5] showed that artificial neural networks (ANN) are sensitive to adversarial examples (AE): counterfactuals of model inputs that yield vastly different model predictions despite being semantically indifferent from their factual counterparts.

Despite partially effective mitigation strategies such as **adversarial training**, truly robust deep learning (DL) remains unattainable even for models that are considered shallow by today's standards [8].

Part of the problem is that high degrees of freedom provide room for many solutions that are locally optimal with respect to narrow objectives [17]. Based purely on predictive performance, these solutions may seem to provide compelling explanations for the data, when in fact they are based on purely associative, semantically meaningless patterns. This poses two related challenges: firstly, it makes these models inherently opaque, since humans cannot simply interpret what type of explanation the complex learned representations correspond to; secondly, even if we could resolve the first challenge, it is not obvious how to mitigate models from learning representations that correspond to meaningless and undesirable explanations.

The first challenge has attracted an abundance of research on **explainable AI** (XAI) which aims to develop tools to derive explanations from complex model representations. This can mitigate a scenario in which we deploy opaque models and blindly rely on their predictions. On countless occasions, this scenario has already occurred in practice and caused real harm to people who were affected adversely and often unfairly by automated decision-making systems involving opaque models [10]. Effective XAI tools can aide us in monitoring models and providing affected individuals with recourse [16].

To our surprise, the second challenge has not yet attracted any consolidated research effort. Specifically, there has been no concerted effort towards improving model **explainability**, which we define here as the degree to which learned representations correspond to explanations that are deemed desirable by humans. Instead, the choice has typically been to improve the capacity of XAI tools to identify the subset explanations that are both desirable and valid for any given model, independent of whether the learned representations are also compatible with undesirable explanations [2]. Fortunately, recent findings indicate that explainability can arise as byproduct of regularization techniques aimed at other objectives such as robustness, generalization and generative capacity [14].

Building on these findings, we introduce **counterfactual training**: a novel regularization technique geared explicitly towards aligning model representations with desirable explanations. Our contributions are as follows:

- We discuss existing related work on improving models and consolidate it through the lens of counterfactual explanations (Section 2).
- We present our proposed methodological framework that leverages faithful counterfactual explanations during the training phase of models to achieve the explainability objective (Section 3).
- Through extensive experiments we demonstrate the counterfactual training improve model explainability while maintaining high predictive performance. We run ablation studies and grid searches to understand how the underlying model components and hyperparameters affect outcomes. (Section 4).

Despite limitations of our approach discussed in Section 5, we conclude that counterfactual training provides a practical framework for researchers and practi-

tioners interested in making opaque models more trustworthy Section 6. We also believe that this work serves as an opportunity for XAI researchers to reevaluate the premise of improving XAI tools without improving models.

2 Related Literature

2.1 Background on Counterfactual Explanations

[16, 7, 2]

2.2 Learning Representations

For example, joint-energy models

2.3 Generalization and Robustness

[13] 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*. They 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.

[15] 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: [18], 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

From this perspective, adversarial training induces models to "unlearn" representations that are susceptible to the semantically most meaningless explanations—adversarial examples.

[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. [11] show that CE and AE are equivalent under certain conditions and derive upper bounds on the distances between them.

2.5 Closely Related

[6] 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***). [12] 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)***)

[9] 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] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)

[5] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014)

[6] 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

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

[8] Kolter, Z.: Keynote Addresses: SaTML 2023 . In: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). pp. xvi–xvi. IEEE Computer Society, Los Alamitos, CA, USA (Feb 2023). https://doi.org/10.1109/SaTML54575.2023.00009, https://doi.ieeecomputersociety.org/10.1109/SaTML54575.2023.00009

[9] 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)

[10] O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown (2016)

[11] 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

[12] Ross, A., Lakkaraju, H., Bastani, O.: Learning models for actionable recourse. In: Proceedings of the 35th International Conference on Neural In-

formation Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2024)
[13] Sauer, A., Geiger, A.: Counterfactual generative networks (2021), https://arxiv.org/abs/2101.06046
[14] Schut, L., Key, O., Mc Grath, R., Costabello, L., Sacaleanu, B., Gal, Y., et al.: Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In: International Conference on Artificial Intelligence and Statistics. pp. 1756–1764. PMLR (2021)
[15] 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)
[16] 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
[17] Wilson, A.G.: The case for bayesian deep learning (2020)
[18] 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