Submission Summary

Conference Name
European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2025

Track Name
Research

Paper ID
908

Paper Title
Counterfactual Training: Teaching Models Plausible and Actionable Explanations

Abstract
We propose a novel training regime termed counterfactual training that leverages counterfactual explanations to increase the explanatory capacity of models. Counterfactual explanations have emerged as a popular post-hoc explanation method for opaque machine learning models: they inform how factual inputs would need to change in order for a model to produce some desired output. To be useful in real-word decision-making systems, counterfactuals should be (1) plausible with respect to the underlying data and (2) actionable with respect to the user-defined mutability constraints. Much existing research has therefore focused on developing post-hoc methods to generate counterfactuals that meet these desiderata. In this work, we instead hold models directly accountable for the desired end goal: counterfactual training employs counterfactuals ad-hoc during the training phase to minimize the divergence between learned representations and plausible, actionable explanations. We demonstrate empirically and theoretically that our proposed method facilitates training models that deliver inherently desirable explanations while promoting robustness and preserving high predictive performance.

Created
3/7/2025, 9:07:38 AM

Last Modified
3/15/2025, 12:32:26 PM

Authors
Patrick Altmeyer (Delft University of Technology) <p.altmeyer@tudelft.nl>
Aleksander Buszydlik (Delft University of Technology) <A.J.Buszydlik@student.tudelft.nl>
Arie van Deursen (Delft University of Technology) <Arie.vanDeursen@tudelft.nl>
Cynthia C. S. Liem (Delft University of Technology) <C.C.S.Liem@tudelft.nl>

Primary Subject Area
Responsible ML & DM -> Interpretability and Explainability

Secondary Subject Areas
Deep Learning -> Representation Learning

Responsible ML & DM -> Robustness and Uncertainty

Domain Conflicts

tudelft.nl; ing.com; jetbrains.com

Submission Files

main.pdf (747.6 Kb, 3/15/2025, 12:22:50 PM)

Supplementary Files

supp.pdf (8.7 Mb, 3/15/2025, 12:29:40 PM)

Submission Questions Response

1. Keywords

Please provide a list of keywords separated by commas (maximum 6):

Counterfactual Training, Counterfactual Explanations, Algorithmic Recourse, Explainable AI, Representation Learning

2. Student Paper

Is the first author a student?

Yes

3. Potential Reviewers

Please provide the email address of at least one author who holds a PhD, has a strong background in ML or DM, and is not already a member of the Program Committee to serve as a potential reviewer.

C.C.S.Liem@tudelft.nl

4. Ethical Considerations

Does this paper address ethical concerns related to AI and machine learning?

Yes

5. Ethical Considerations

Does the research involve human subjects or sensitive data?

No

6. Ethical Considerations

If you answered 'Yes' to the previous question, briefly mention how you tackled that.

[Not Answered]

7. Authors Agreement

By submitting the paper, the authors agree to the following terms:
** One of the authors commits to reviewing, as outlined in the Call for Papers (CfP): https://ecmlpkdd.org/2025/submissions-research-track/*
** Adhere to the ethical guidelines for authors and reviewers.*
** Register and present the paper in person at the conference if it is accepted.*
** Display the paper title and author names on the conference website if the paper is accepted.*

Agreement accepted