Counterfactual Explanations for Misclassified Images: How Human and Machine Explanations Differ (Abstract Reprint)

Eoin Delaney[1,2,3], **Arjun Pakrashi**[1,3], **Derek Greene**[1,2,3], **Mark T. Keane**[1,2,3]

[1]School of Computer Science, University College Dublin, Belfield, Dublin, Ireland
[2]Insight Centre for Data Analytics, Belfield, Dublin, Ireland
[3]VistaMilk SFI Research Centre, Belfield, Dublin, Ireland

Abstract Reprint. This is an abstract reprint of a journal article by Delaney, Pakrashi, Greene, and Keane (2023).

Abstract

Counterfactual explanations have emerged as a popular solution for the eXplainable AI (XAI) problem of elucidating the predictions of black-box deep-learning systems because people easily understand them, they apply across different problem domains and seem to be legally compliant. Although over 100 counterfactual methods exist in the XAI literature, each claiming to generate plausible explanations akin to those preferred by people, few of these methods have actually been tested on users ($\tilde{\{}$7%). Even fewer studies adopt a user-centered perspective; for instance, asking people for their counterfactual explanations to determine their perspective on a "good explanation". This gap in the literature is addressed here using a novel methodology that (i) gathers human-generated counterfactual explanations for misclassified images, in two user studies and, then, (ii) compares these human-generated explanations to computationally-generated explanations for the same misclassifications. Results indicate that humans do not "minimally edit" images when generating counterfactual explanations. Instead, they make larger, "meaningful" edits that better approximate prototypes in the counterfactual class. An analysis based on "explanation goals" is proposed to account for this divergence between human and machine explanations. The implications of these proposals for future work are discussed.

References

Delaney, E.; Pakrashi, A.; Greene, D.; and Keane, M. T. 2023. Counterfactual explanations for misclassified images: How human and machine explanations differ. *Artificial Intelligence*, 324: 103995.