DSSV 2023
Data Science, Statistics & Visualisation

ECDA 2023
European Conference on Data Analysis

Multi-Modal Counterfactual Explanations for Image Classification

Camille Dams[1], James Hinns[1], and David Martens[1]

[1]University of Antwerp, Department of Engineering Management, Prinsstraat 13, Antwerp 2000, Belgium

In this paper, we present a novel model-agnostic method of generating multi-modal counterfactual explanations for image classification models.

Our approach leverages a pre-trained image segmentation model to identify and segment meaningful components within an input image. Each segment is systematically blurred to observe if it leads to a change in the predicted class. If a class change is detected, we consider the corresponding segment as a counterfactual instance. To provide an interpretable explanation, we generate a textual sentence describing the model's decision at this instance, utilising the labels returned by the segmentation model, as well as the counterfactual image itself. To increase the coverage of explanations, we combine detected segments and check for counterfactuals, optimising with meta-heuristics.

By incorporating semantic and multi-modal elements, our method enables a more meaningful and context-aware representation of counterfactuals than previous approaches that relied on edge-based segmentation [1]. Semantically based segmentation models provide a closer fit to how a human may describe a counterfactual, based on whole objects, rather than edges. Offering text labels for the segments that lead to counterfactuals allows the findings of many local explanations to be represented in frequency tables, which display trends across the class. As well as segment name, and the predicted class prior and post blurring, the frequency tables also include the positions and surrounding objects to counterfactual segments.

These tables provide a lossless form of aggregation between explanations, not possible with previous methods, allowing users to explore trends in their models predictions and identify unwanted behaviour. As interpretability is simply the degree to which a human can understand the cause of a decision, in order to increase interpretability, explanations should be tailored to specific users. Based on the parallels of this problem to many found in data visualisation, we propose an approach that follows Shneiderman's mantra for the creation of visualisation systems [2]:

- **Overview First:** Presenting a table aggregating local explanation results by textual descriptions of their properties.
- **Zoom and Filter:** Allows aggregation of explanations in different ways, for example, by segment name, prior and post classification, and position. Results can then be filtered after aggregation, by any of the recorded properties.
- **Details on Demand:** Providing counterfactual images along with textual descriptions for a specific data instance.

We show via experimentation how our method performs against other model-agnostic counterfactual methods, and how changing the segmentation model can affect results. We also demonstrate the different criterion for which we may optimise segment combinations, and how this affects the counterfactuals generated. Finally, we show how the trends from our aggregation of explanations can lead to insights to the model as a whole.

Keywords — XAI, Machine Learning, Counterfactual Explanations, Multi-Modal

[1] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. de Oliveira, and D. Martens, "Explainable image classification with evidence counterfactual," Pattern Analysis and Applications , vol. 25, no. 2, pp. 315–335, 2022.

[2] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in Proceedings 1996 IEEE symposium on visual languages , pp. 336–343, IEEE, 1996.