Transferability of counterfactual examples across different models

Project Work - Responsible AI (by Bilal Zafar)

Frederik Hüttemann
Applied Computer Science
Ruhr-Universität Bochum
frederik.huettemann@rub.de

Abstract

An abstract should concisely (less than 300 words) motivate the problem, describe your aims, describe your contribution, and highlight your main finding(s). // In this project, we investigate the transferability of counterfactual examples across different models. Neural Networks are often used to make life critical decisions, such as in healthcare or criminal justice. They also find applications in less critical domains, such as image classification. However, these models are often criticized for being black boxes, which makes it difficult to understand their decisions. Counterfactual examples are a method to explain the decisions of a model by showing how the input would have to change in order to change the output.
It is often assumed that counterfactual examples are model-specific, so their the same counterfacutal example usually cannot be applied on different models. The goal of this project is to look deeper into this assumption and try to find out whether counterfactual examples can be transferred across different models. For this, different types of data and models are used. For the simple cases two numerical datasets are used. For the more complex case, the Imagenet dataset is converted into a binary classification task.
TBD: results and conclusion The results show that counterfactual examples can be transferred across different models, but the transferability depends on the type of data and the model used. In particular, counterfactual examples for numerical data can be transferred to at least some extent, while counterfactual examples for image data are not transferable at all. This suggests that the transferability of counterfactual examples is not a general property, but rather depends on the specific data and model used.

1 Introduction

The explains the problem, why it's difficult, interesting, or important, how and why current methods succeed/fail at the problem, and explains the key ideas of your approach and results. Though an introduction covers similar material as an abstract, the introduction gives more space for motivation, detail, references to existing work, and to capture the reader's interest.

2 Related Work

This section helps the reader understand the research context of your work by providing an overview of existing work in the area. **?**

3 Approach

In this project, an exploratory approach is taken. The goal is to find a limit on how far counterfactual examples can be transferred. This means that the project does not focus on a specific method, but rather on the general idea of counterfactual examples and their transferability.
Therefore, the first steps are to implement a method to generate counterfactual examples for different types of data.

4 Experiments

The experiments, which are described in this section, are designed to test the transferability of counterfactual examples across different models.

4.1 Data

The used data can be divided into two categories: numerical data and image data. For the numerical data, two datasets are used: the *cumpas* dataset (ProPublica (2016)) and the *ACS* dataset (Bureau (2018)).

4.2 Evaluation method

4.3 Experimental details

4.4 Results

Report the quantitative results that you have found. Use a table or plot to compare results and compare against baselines.

- If you're a default project team, you should **report the accuracy and Pearson correlation scores you obtained on the test leaderboard** in this section. You can also report dev set results if you'd like.
- Comment on your quantitative results. Are they what you expected? Better than you expected? Worse than you expected? Why do you think that is? What does that tell you about your approach?

5 Analysis

Your report should include *qualitative evaluation*. That is, try to understand your system (e.g., how it works, when it succeeds and when it fails) by inspecting key characteristics or outputs of your model.

6 Conclusion

Summarize the main findings of your project and what you have learned. Highlight your achievements, and note the primary limitations of your work. If you'd like, you can describe avenues for future work.

7 Ethics Statement

What are the ethical challenges and possible societal risks of your project, and what are mitigation strategies?

References

U.S. Census Bureau. 2018. American community survey (acs) dataset. Accessed: 2025-07-01.

ProPublica. 2016. Compas dataset. Accessed: 2025-07-01.

A Appendix (optional)

If you wish, you can include an appendix, which should be part of the main PDF, and does not count towards the 6-8 page limit. Appendices can be useful to supply extra details, examples, figures, results, visualizations, etc. that you couldn't fit into the main paper. However, your grader *does not* have to read your appendix, and you should assume that you will be graded based on the content of the main part of your paper only.