ECCCos from the Black Box

Faithful Model Explanations through Energy-Based Conformal Counterfactuals

Patrick Altmeyer Mojtaba Farmanbar Arie van Deursen
Cynthia C. S. Liem

Delft University of Technology

2024-01-04

Open Questions

1. What makes a counterfactual **plausible**?

Open Questions

1. What makes a counterfactual **plausible**?
2. Why do we need plausibility?

Open Questions

1. What makes a counterfactual **plausible**?
2. Why do we need plausibility?
3. Is plausibility all we need?

Open Questions

1. What makes a counterfactual **plausible**?
2. Why do we need plausibility?
3. Is plausibility all we need?
4. What makes models more **explainable**?

Plausibility

There's no consensus on the exact definition of plausibility but we think about it as follows:

Definition (Plausible Counterfactuals)

Let $\mathcal{X}|\mathbf{y}^+ = p(\mathbf{x}|\mathbf{y}^+)$ denote the true conditional distribution of samples in the target class $\mathbf{y}^+$. Then for $\mathbf{x}'$ to be considered a plausible counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^+$.

Counter Example

- The counterfactual in Figure 1 is valid: it has crossed the decision boundary.
- But is it consistent with the data in the target class (blue)?

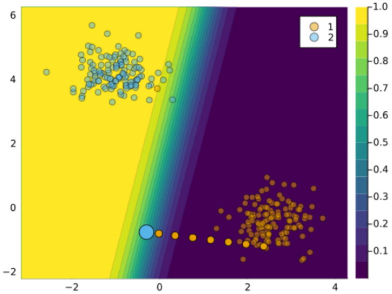


Figure 1: A valid but implausible counterfactual. Source: Altmeyer, Deursen, and Liem (2023)

Why Plausibility?

- Actionability: If a counterfactual is implausible, it is unlikely to be actionable.
- Fairness: If a counterfactual is implausible, it is unlikely to be fair.
- Robustness: If a counterfactual is implausible, it is unlikely to be robust.

But: Higher plausibility seems to require larger changes and hence increase costs to individuals.

Pick your Poison?

All of these counterfactuals are valid explanations for the model's prediction. Which one would you pick?

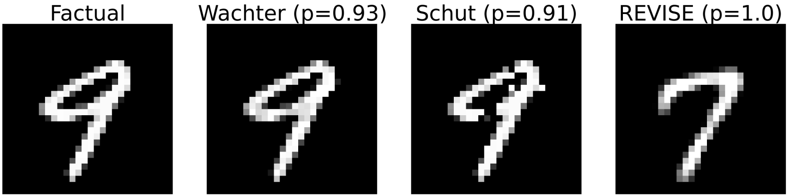
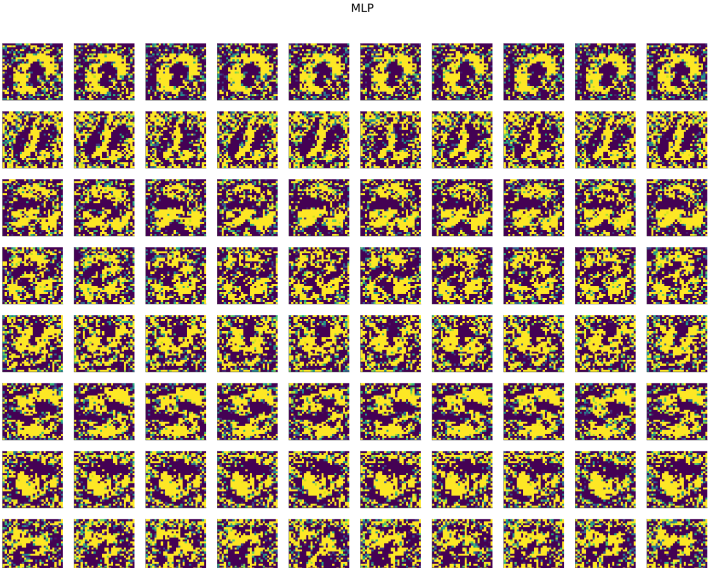


Figure 2: Turning a 9 into a 7: Counterfactual Examplanations for an Image Classifier.

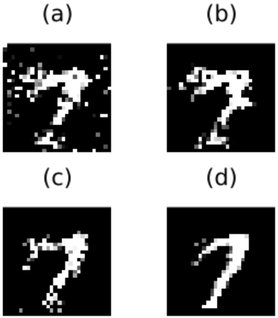
What do Models Learn?

These images are sampled from the posterior distribution learned by the model. Looks different, no?



Faithful Counterfactuals

We propose a way to generate counterfactuals that are as plausible as the underlying model permits (under review).



Definition (Faithful Counterfactuals)

Let $\mathcal{X}_\theta|\mathbf{y}^+ = p_\theta(\mathbf{x}|\mathbf{y}^+)$ denote the conditional distribution of $\mathbf{x}$ in the target class $\mathbf{y}^+$, where θ denotes the parameters of model M_θ. Then for $\mathbf{x}'$ to be considered a faithful counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^+$.

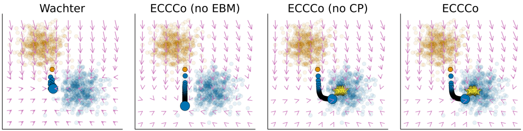


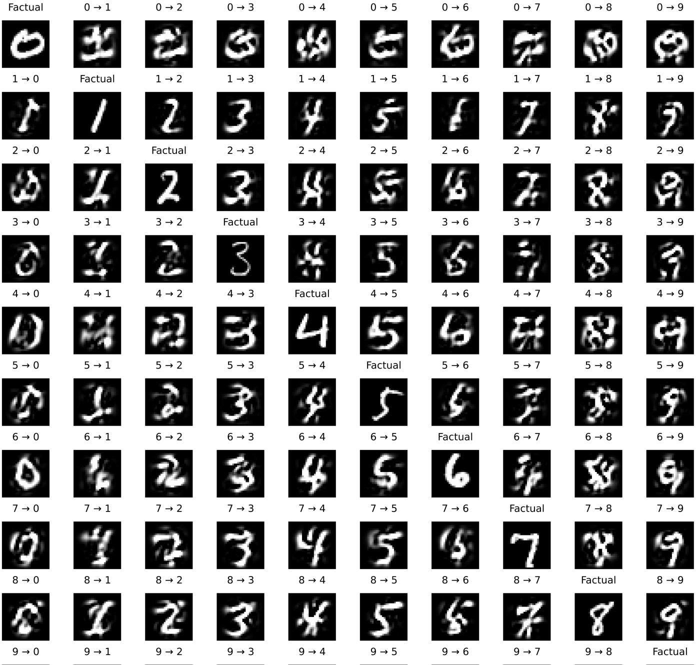
Figure 4: Gradient fields and counterfactual paths for different generators.

Improving Models

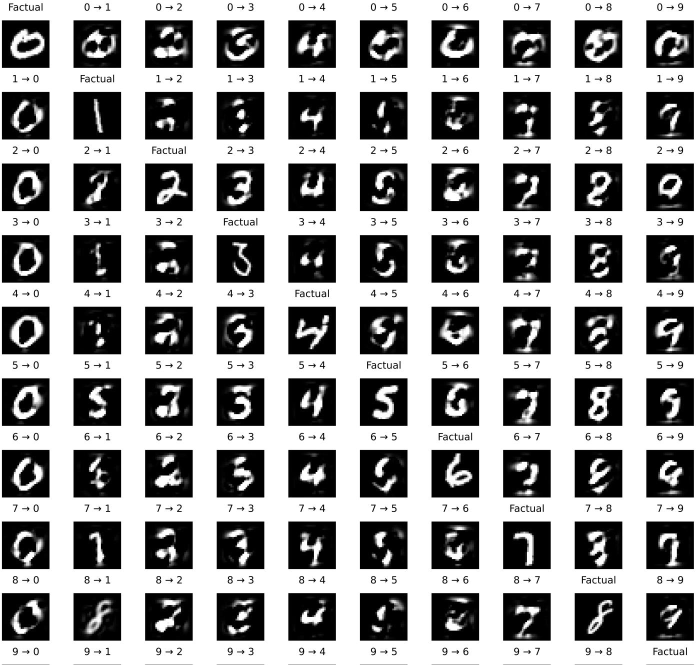
Now that we have a tool to faithfully explain models we may ask: **how** do models learn plausible explanations? Initial evidence:

1. Incorporating predictive uncertainty (e.g. ensembling).
2. Addressing robustness (e.g. adversarial training in Schut et al. (2021)).
3. Better model architectures.
4. Hybrid modelling (i.e. combining generative and discriminative models).

Example: Architecture



Example: JEM Ensemble



Questions?

With thanks to my co-authors Mojtaba Farmanbar, Arie van Deursen and Cynthia C. S. Liem. Slides powered by Quarto.

Figure 7: Takes you to my website.

Counterfactual Explanations

All the work presented today is powered by `CounterfactualExplanations.jl` .

There is also a corresponding paper, *Explaining Black-Box Models through Counterfactuals*, which has been published in JuliaCon Proceedings.

References

Altmeyer, Patrick, Arie van Deursen, and Cynthia Liem. 2023. "Explaining Black-Box Models Through Counterfactuals." *arXiv Preprint arXiv:2308.07198*.

Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2020. "Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One." In. https://openreview.net/forum?id=Hkxzx0NtDB.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11): 2278–2324.

Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. "Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties." In *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.