Do Joint Energy-Based Models Produce More Plausible Counterfactual Explanations?

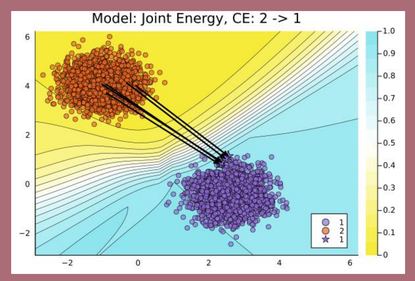
Giacomo Pezzali [g.pezzali@student.tudelft.nl]
Supervisors: Cynthia C. S. Liem, Patrick Altmeyer
EEMCS, Delft University of Technology, The Netherlands

1. Introduction

- **Counterfactual Explanations** (CE)[1]: given an input classified by a model in an undesired class, find how that input should change to instead be classified in the target class. A good CE is:
  - **Plausible**: the proposed change makes the new input indistinguishable from the real-world examples of the target class.
  - **Faithful**: the proposed change is representative of what the model has learned from its training.
- **Explainable Model**: a model whose faithful CEs are also plausible.
- **Energy-Constraint Conformal Counterfactual** (ECCCo)[1]: a technique to generate CE with a specific focus on faithfulness.
- **Joint Energy-based Models** (JEM)[2]: an alternative way of training certain classifier architectures to behave both as classifiers (high accuracy) and generative (low generative loss) models.
- **Implausibility**: metric to evaluate CEs. A model whose ECCCo-generated CEs have low implausibility is considered explainable.

$$\text{impl}(x', \mathcal{X}_{c_T}) = \frac{1}{|\mathcal{X}_{c_T}|} \sum_{x \in \mathcal{X}_{c_T}} \text{dist}(x, x')$$

where x' is the counterfactual explanation, $\mathcal{X}_{c_T}$ is the subset of points from the training set originally labelled as the target class c_T and $\text{dist}(\cdot, \cdot)$ is the euclidean distance.



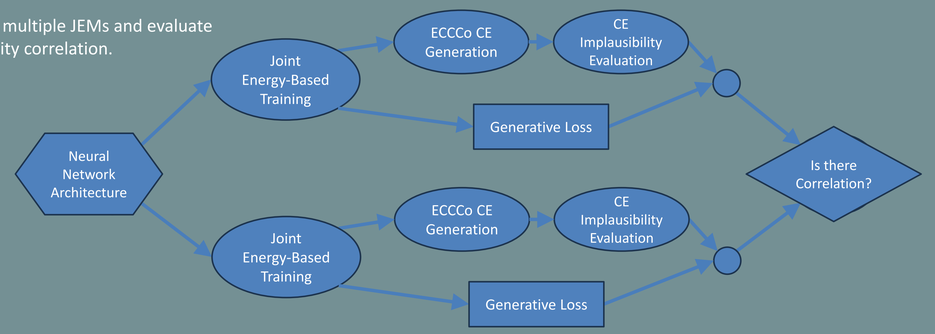
Counterfactuals generated from a joint energy model

2. Research Questions

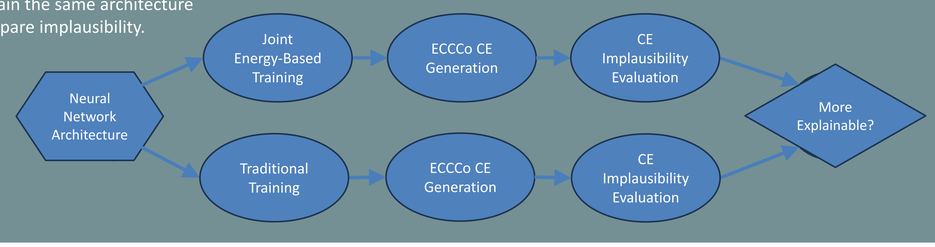
1. When training a JEM, does its generative loss affect its explainability?
2. Given the same architecture, does the JEM training improve explainability of the model vs. classical training?

3. Methodology

Intra-model Experiment: train multiple JEMs and evaluate generative loss and implausibility correlation.



Training-based Experiment: train the same architecture classically and as JEM and compare implausibility.



4. Results

Intra-model Experiment: correlation between generative loss and implausibility. 1 would indicate perfect linearity between generative capabilities and explainability of a model. Computed on multiple datasets and neural network architectures.

Dataset	Correlation
Circles	0.0117
Linearly Separable	0.2056
Overlapping	0.1506
MNIST [Altmeyer]	-0.2536
MNIST [Le Cun]	**0.5142**
California Housing	-0.2460
German Credit	**-0.6261**
German Credit [Zhao]	0.0078

Training-based Experiment: comparison between classical model implausibility and JEM implausibility. * markings indicate significant improvement in explainability using JEM, † markings indicate significant loss of explainability.

Training	Circles		Linearly Separable		Overlapping	
	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)
Classical	**0.38 ± 0.25**	**0.33 ± 0.30**	4.24 ± 0.51	2.37 ± 0.03	**0.91 ± 0.20**	**0.85 ± 0.26**
Joint-Energy	3.16 ± 0.56 ††	1.91 ± 1.87 ††	**3.17 ± 0.78** **	**2.13 ± 0.28** **	1.56 ± 0.86 ††	0.91 ± 0.20

Training	MNIST [Altmeyer]		MNIST [Le Cun]	
	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)
Classical	18.80 ± 1.83	**16.50 ± 1.66**	18.02 ± 1.47	**17.07 ± 2.06**
Joint-Energy	**18.79 ± 1.83**	19.17 ± 2.49 †	**17.90 ± 1.51**	17.53 ± 1.82

Training	California Housing		German Credit		German Credit [Zhao]	
	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)
Classical	**1.62 ± 0.31**	1.62 ± 0.31	**4.75 ± 0.21**	**4.74 ± 0.43**	4.94 ± 0.07	4.90 ± 0.07
Joint-Energy	1.89 ± 0.10	**1.43 ± 0.43**	4.94 ± 0.05	4.75 ± 0.18	**4.81 ± 0.04** *	**4.72 ± 0.07** **

5. Conclusion & Limitations

- No relevant influence was found of the generative capability on the explainability of a joint-energy model.
- The use of JEM training produced both more explainable models and less explainable models than classical training.
- Various experimental paths are still open:
  1. Comparing JEMs varying the importance of the generative training objective in training.
  2. Work to counterbalance the training instability of JEMs to more clearly compare them and classically trained models.

References

[1] P. Altmeyer, M. Farmanbar, A. van Deursen, and C. Liem, "Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals," English, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 10, pp. 10 829–10 837, 2024, ISSN: 2159-5399. DOI: 10.1609/aaai.v38i10.28956.
[2] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," in International Conference on Learning Representations, 2020. [Online]. Available: https://openreview.net/forum?id=Hkxzx0NtDB.