Counterfactual Training: Teaching Models Plausible and Actionable Explanations

Patrick Altmeyer[1][0000−0003−4726−8613] (✉), Arie van Deursen[1], and Cynthia C. S. Liem[1]

Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science {p.altmeyer}@tudelft.net

Abstract. Counterfactual Explanations have emerged as a popular tool to explain predictions made by opaque machine learning models: they explain how factual inputs need to change in order for some fitted model to produce some desired output. Much existing research has focused on identifying explanations that are not only valid but also deemed plausible and desirable with respect to the underlying data and stakeholder requirements. Recent work has shown that under this premise, the task of learning plausible explanations is effectively reassigned from the model itself to the (post-hoc) counterfactual explainer. Building on that work, we propose a novel model objective that leverages counterfactuals during the training phase (ad-hoc) in order to minimize the divergence between learned representations and plausible explanations. Through extensive experiments, we demonstrate that our proposed methodology facilitates training models that inherently deliver plausible explanations while maintaining high predictive performance.

Keywords: Counterfactual Explanations · Explainable AI · Representation Learning

1 Introduction

Today's prominence of artificial intelligence (AI) has largely been driven by advances in **representation learning**: instead of relying on features and rules that are carefully hand-crafted by humans, modern machine learning (ML) models are tasked with learning these representations from scratch, guided by narrow objectives such as predictive accuracy [?]. Modern advances in computing have made it possible to provide such models with ever greater degrees of freedom to achieve that task, which has often led them to outperform traditionally more parsimonious models. Unfortunately, in doing so they also learn increasingly complex and highly sensitive representations that we can no longer easily interpret.

This trend towards complexity for the sake of performance has come under serious scrutiny in recent years. At the very cusp of the deep learning revolution, [?] showed that artificial neural networks (ANN) are sensitive to adversarial examples: counterfactuals of model inputs that yield vastly different model predictions despite being "imperceptible" in that they are semantically indifferent from their factual counterparts. Despite partially effective mitigation strategies such as **adversarial training** [?], truly robust deep learning (DL) remains unattainable even for models that are considered shallow by today's standards [?].

Part of the problem is that high degrees of freedom provide room for many solutions that are locally optimal with respect to narrow objectives [?][1]. Based purely on predictive performance, these solutions may seem to provide compelling explanations for the data, when in fact they are based on purely associative, semantically meaningless patterns. This poses two related challenges: firstly, it makes these models inherently opaque, since humans cannot simply interpret what type of explanation the complex learned representations correspond to; secondly, even if we could resolve the first challenge, it is not obvious how to mitigate models from learning representations that correspond to meaningless and implausible explanations.

The first challenge has attracted an abundance of research on **explainable AI** (XAI) which aims to develop tools to derive explanations from complex model representations. This can mitigate a scenario in which we deploy opaque models and blindly rely on their predictions. On countless occasions, this scenario has already occurred in practice and caused real harm to people who were affected adversely and often unfairly by automated decision-making systems (ADMS) involving opaque models [?]. Effective XAI tools can aide us in monitoring models and providing recourse to individuals to turn adverse outcomes (e.g. "loan application rejected") into positive ones ("application accepted"). [?] propose **counterfactual explanations** as an effective approach to achieve this: they explain how factual inputs need to change in order for some fitted model to produce some desired output, typically involving minimal perturbations.

To our surprise, the second challenge has not yet attracted any consolidated research effort. Specifically, there has been no concerted effort towards improving model **explainability**, which we define here as the degree to which learned representations correspond to explanations that are interpretable and deemed **plausible** by humans (see Definition **??**). Instead, the choice has typically been to improve the capacity of XAI tools to identify the subset explanations that are both plausible and valid for any given model, independent of whether the learned representations are also compatible with implausible explanations [?]. Fortunately, recent findings indicate that explainability can arise as byproduct of regularization techniques aimed at other objectives such as robustness, generalization and generative capacity [?].

Building on these findings, we introduce **counterfactual training**: a novel regularization technique geared explicitly towards aligning model representations with plausible explanations. Our contributions are as follows:

- We discuss existing related work on improving models and consolidate it through the lens of counterfactual explanations (Section **??**).

[1] For clarity: we follow standard ML convention in using "degrees of freedom" to refer to the number of parameters estimated from data.

- We present our proposed methodological framework that leverages faithful counterfactual explanations during the training phase of models to achieve the explainability objective (Section **??**).
- Through extensive experiments we demonstrate the counterfactual training improve model explainability while maintaining high predictive performance. We run ablation studies and grid searches to understand how the underlying model components and hyperparameters affect outcomes. (Section **??**).

Despite limitations of our approach discussed in Section **??**, we conclude that counterfactual training provides a practical framework for researchers and practitioners interested in making opaque models more trustworthy Section **??**. We also believe that this work serves as an opportunity for XAI researchers to reevaluate the premise of improving XAI tools without improving models.

2 Related Literature

To the best of our knowledge, our proposed framework for counterfactual training represents the first attempt to use counterfactual explanations during training to improve model explainability. In high-level terms, we define model explainability as the extent to which valid explanations derived for an opaque model are also deemed plausible with respect to the underlying data and stakeholder requirements. To make this more concrete, we follow [**?**] in tieing the concept of explainability to the quality of counterfactual explanations that we can generate for a given model. The authors show that counterfactual explanations—understood here as minimal input perturbations that yield some desired model prediction—are generally more meaningful if the underlying model is more robust to adversarial examples. We can make intuitive sense of this finding when looking at adversarial training (AT) through the lens of representation learning with high degrees of freedom: by inducing models to "unlearn" representations that are susceptible to worst-case counterfactuals (i.e. adversarial examples), AT effectively removes some implausible explanations from the solution space.

2.1 Adversarial Examples are Counterfactual Explanations

This interpretation of the link between explainability through counterfactuals on one side, and robustness to adversarial examples on the other, is backed by empirical evidence. [**?**] demonstrate that using counterfactual images during classifier training improves model robustness. Similarly, [**?**] argue that counterfactuals represent potentially useful training data in machine learning, especially in supervised settings where inputs may be reasonably mapped to multiple outputs. They, too, demonstrate the augmenting the training data of image classifiers can improve generalization. [**?**] propose an approach using counterfactuals in training that does not rely on data augmentation: they argue that counterfactual pairs typically already exist in training datasets. Specifically, their approach relies on, firstly, identifying similar input samples with different annotations and, secondly, ensuring that the gradient of the classifier aligns with the vector between pairs of counterfactual inputs using the cosine distance as a loss function. In the natural language processing (NLP) domain, counterfactuals have similarly been used to improve models through data augmentation: [?], propose *POLYJUICE*, a general-purpose counterfactual generator for language models. They demonstrate empirically that augmenting training data through *POLYJUICE* counterfactuals improves robustness in a number of NLP tasks. [?] introduce Counterfactual Adversarial Training (CAT), which also aims at improving generalization and robustness of language models. Specifically, they propose to proceed as follows: firstly, they identify training samples that are subject to high predictive uncertainty; secondly, they generate counterfactual explanations for those samples; and, finally, they fine-tune the given language model on the augmented dataset that includes the generated counterfactuals.

There have also been several attempts at formalizing the relationship between counterfactual explanations (CE) and adversarial examples (AE). Pointing to clear similarities in how CE and AE are generated, [?] makes the case for jointly studying the opaqueness and robustness problem in representation learning. Formally, AE can be seen as the subset of CE, for which misclassification is achieved [?]. Similarly, [?] show that CE and AE are equivalent under certain conditions and derive theoretical upper bounds on the distances between them.

Two recent works are closely related to ours in that they use counterfactuals during training with the explicit goal of affecting certain properties of post-hoc counterfactual explanations. Firstly, [?] propose a way to train models that are guaranteed to provide recourse for individuals to move from an adverse outcome to some positive target class with high probability. The approach proposed by [?] builds on adversarial training, where in this context susceptibility to targeted adversarial examples for the positive class is explicitly induced. The proposed method allows for imposing a set of actionability constraints ex-ante: for example, users can specify that certain features (e.g. *age*, *gender*, ...) are immutable. Secondly, [?] are the first to propose an end-to-end training pipeline that includes counterfactual explanations as part of the training procedure. In particular, they propose a specific network architecture that includes a predictor and CE generator network, where the parameters of the CE generator network are learnable. Counterfactuals are generated during each training iteration and fed back to the predictor network. In contrast to [?], we impose no restrictions on the neural network architecture at all.

2.2 Beyond Robustness

Improving the adversarial robustness of models is not the only path towards aligning representations with plausible explanations. In a work closely related to this one, [?] show that explainability can be improved through model averaging and refined model objectives. The authors propose a way to generate counterfactuals that are maximally **faithful** to the model in that they are consistent with what the model has learned about the underlying data. Formally, they rely on tools from energy-based modelling to minimize the divergence between the distribution of counterfactuals and the conditional posterior over inputs learned by the model. Their proposed counterfactual explainer, *ECCCo*, yields plausible explanations if and only if the underlying model has learned representations that align with them. They find that both deep ensembles [**?**] and joint energy-based models (JEMs) [**?**] tend to do well in this regard.

Once again it helps to look at these findings through the lens of representation learning with high degrees of freedom. Deep ensembles are approximate Bayesian model averages, which are most called for when models are underspecified by the available data [**?**]. Averaging across solutions mitigates the aforementioned risk of relying on a single locally optimal representations that corresponds to semantically meaningless explanations for the data. Previous work by [**?**] similarly found that generating plausible ("interpretable") counterfactual explanations is almost trivial for deep ensembles that have also undergone adversarial training. The case for JEMs is even clearer: they involve a hybrid objective that induces both high predictive performance and generative capacity [**?**]. This is closely related to the idea of aligning models with plausible explanations and has inspired our proposed counterfactual training objective, as we explain in Section **??**.

3 Counterfactual Training

Counterfactual training combines ideas from adversarial training, energy-based modelling and counterfactuals explanations with the explicit objective of aligning representations with plausible explanations that comply with user requirements. In the context of CE, plausibility has broadly been defined as the degree to which counterfactuals comply with the underlying data generating process [**? ? ?**]. Plausibility is a necessary but insufficient condition for using CE to provide algorithmic recourse (AR) to individuals affected by opaque models in practice. This is because for recourse recommendations to be **actionable**, they need to not only result in plausible counterfactuals but also be attainable. A plausible CE for a rejected 20-year-old loan applicant, for example, might reveal that their application would have been accepted, if only they were 20 years older. Ignoring all other features, this complies with the definition of plausibility if 40-year-old individuals are in fact more credit-worthy on average than young adults. But of course this CE does not qualify for providing actionable recourse to the applicant since *age* is not a mutable feature. For our intents and purposes, counterfactual training aims at improving model explainability by aligning models with counterfactuals that meet both desiderata, plausibility and actionability. Formally, we define explainability as follows:

Definition 1 (Model Explainability).

Let $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ *denote a supervised classification model that maps from the* D*-dimensional input space* $\mathcal{X}$ *to representations* $\phi(\mathbf{x};\theta)$ *and finally to the* K*-dimensional output space* $\mathcal{Y}$*. Assume that for any given input-output pair* $\{\mathbf{x},\mathbf{y}\}_i$ *there exists a counterfactual* $\mathbf{x}' = \mathbf{x} + \Delta : \mathbf{M}_\theta(\mathbf{x}') = \mathbf{y}^+ \neq \mathbf{y} = \mathbf{M}_\theta(\mathbf{x})$ *where* $\mathbf{y}^+$ *denotes some target output. We say that* $\mathbf{M}_\theta$ *is* ***explainable*** *to the extent that faithfully generated counterfactuals are plausible (i.e. consistent with the data) and actionable. Formally, we define these properties as follows:*

1. *(Plausibility)* $\int^{A} p(\mathbf{x}|\mathbf{y}^{+})d\mathbf{x} \rightarrow 1$ *where* A *is some small region around* $\mathbf{x}'$*.*
2. *(Actionability) Permutations* Δ *are subject to actionability constraints.*

We consider counterfactuals as faithful to the extent that they are consistent with what the model has learned about the input data. Let $p_\theta(\mathbf{x}|\mathbf{y}^{+})$ *denote the conditional posterior over inputs, then formally:*

3. *(Faithfulness)* $\int^{A} p_\theta(\mathbf{x}|\mathbf{y}^{+})d\mathbf{x} \rightarrow 1$ *where* A *is defined as above.*

The definitions of faithfulness and plausibility in Definition **??** are the same as in [?], with adapted notation. Actionability constraints in Definition **??** vary and depend on the context in which $\mathbf{M}_\theta$ is deployed. In this work, we focus on domain and mutability constraints for individual features x_d for $d = 1, ..., D$. We limit ourselves to classification tasks for reasons discussed in Section **??**.

3.1 Our Proposed Objective

To train models with high explainability as defined in Definition **??**, we propose the following objective,

$$\text{yloss}(\mathbf{M}_\theta(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}}\text{div}(\mathbf{x}, \mathbf{x}', y; \theta) + \lambda_{\text{adv}}\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}) \tag{1}$$

where yloss($\cdot$) denotes any conventional classification loss function (e.g. cross-entropy) that induces discriminative (predictive) performance. The two additional components in Equation **??** are explained in more detail below. For now, they can be sufficiently described as inducing explainability directly and indirectly by penalizing: 1) the contrastive divergence, div($\cdot$), between counterfactuals x' and observed samples x and, 2) the adversarial loss, advloss(.), with respect to counterfactuals. The tradeoff between the different components can be governed by adjusting the strengths of the penalties λ_{div} and λ_{adv}.

Directly Inducing Explainability through Contrastive Divergence [?] observe that any classifier can be re-interpreted as a joint energy-based model (JEM) that learns to discriminate output classes conditional on inputs and generate inputs. They show that JEMs can be trained to perform well at both tasks by directly maximizing the joint log-likelihood factorized as $\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x})$. The first factor can be optimized using conventional cross-entropy as in Equation **??**. To optimize $\log p_\theta(\mathbf{x})$ [?] minimize the contrastive divergence between samples drawn from $p_\theta(\mathbf{x})$ and training observations, i.e. samples from $p(\mathbf{x})$.

A key empirical finding in [?] was that JEMs tend to do well with respect to the plausibility objective in Definition **??**. If we consider samples drawn from $p_\theta(\mathbf{x})$ as counterfactuals, this is an expected finding, because the JEM objective effectively minimizes the divergence between the conditional posterior and $p(\mathbf{x}|\mathbf{y}^{+})$. To generate samples, [?] rely on Stochastic Gradient Langevin Dynamics (SGLD) using an uninformative prior for initialization. This is where we depart from their methodology: instead of generating samples through SGLD, we propose using counterfactual explainers to generate counterfactuals for observed training samples. Specifically, we have

$$\text{div}(\mathbf{x}, \mathbf{x}', y; \theta) = \mathcal{E}_\theta(\mathbf{x}, y) - \mathcal{E}_\theta(\mathbf{x}', y) \tag{2}$$

where $\mathcal{E}_\theta(\cdot)$ denotes the energy function. We generate samples $\mathbf{x}'$ by first randomly sampling the target class $y^+ \sim p(y)$ and then generating a counterfactual explanation for that target, similar to how conditional sampling is used to draw from $p_\theta(\mathbf{x})$ in [**?**]. In particular, we set $\mathcal{E}_\theta(\mathbf{x}, \mathbf{y}) = -\mathbf{M}_\theta(\mathbf{x})[y^+]$ where y^+ denotes the index of the target class.

Intuitively, the gradient of Equation **??** decreases the energy of observed training samples (positive samples) while at same time increasing the energy of counterfactuals (negative samples) [**?**]. As the generated counterfactuals get more plausible (Definition **??**) over the cause of training, these two opposing effects gradually balance each out [**?**].

The departure from SGLD allows us to tap into the vast repertoire of explainers that have been proposed in the literature to meet different desiderata. Typically, these methods facilitate the imposition of domain and mutability constraints, for example. In principle, any existing approach for generating counterfactual explanations is viable, so long as it does not violate the faithfulness condition. Like JEMs [**?**], counterfactual training can be considered as a form of contrastive representation learning.

Indirectly Inducing Explainability through Adversarial Robustness Based on our analysis in Section **??**, counterfactuals $\mathbf{x}'$ can be repurposed as additional training samples [**?**] or adversarial examples [**? ?**]. This leaves some flexibility with respect to the exact choice for advloss$(\cdot)$ in Equation **??**. An intuitive functional form to use, though likely not the only reasonable choice, is inspired by adversarial training:

$$\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}; \varepsilon) = \begin{cases} \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}) & \text{if } ||\varDelta||_\infty \leq \varepsilon \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Under this choice we treat the counterfactual $\mathbf{x}'$ as an adversarial example iff it is imperceptible, i.e. the magnitude of the perturbation of any individual feature is upper-bounded at ε.

3.2 Encoding Actionability Constraints

Many existing counterfactual explainers support domain and mutability constraints out-of-the-box. In fact, both types of constraints can be implemented for any counterfactual explainer that relies on gradient descent in the feature space for optimization [**?**]. In this context, domain constraints can be imposed by simply projecting counterfactuals back to the specified domain, if the previous gradient step resulted in updated feature values that were out-of-domain.

Mutability constraints can similarly be enforced by setting partial derivatives to zero to ensure that features are only mutated in the allowed direction, if at all.

Since actionability constraints are binding at test time, we should also impose them when generating $\mathbf{x}'$ during each training iteration to align model representations with user requirements. Through their effect on $\mathbf{x}'$, both types of constraints influence model outcomes through Equation **??**. Here it is crucial that we avoid penalizing implausibility that arises due to mutability constraints. For any mutability-constrained feature d this can be achieved by enforcing $\mathbf{x}[d] - \mathbf{x}'[d] := 0$ whenever perturbing $\mathbf{x}'[d]$ in the direction of $\mathbf{x}[d]$ would violate mutability constraints. Specifically, we set $\mathbf{x}[d] := \mathbf{x}'[d]$ if

1. Feature d is strictly immutable in practice.
2. We have $\mathbf{x}[d] > \mathbf{x}'[d]$ but feature d can only be decreased in practice.
3. We have $\mathbf{x}[d] < \mathbf{x}'[d]$ but feature d can only be increased in practice.

From a Bayesian perspective, setting $\mathbf{x}[d] := \mathbf{x}'[d]$ can be understood as assuming a point mass prior for $p(\mathbf{x})$ with respect to feature d. Intuitively, we think of this simply in terms ignoring implausibility costs with respect to immutable features, which effectively forces the model to instead seek plausibility with respect to the remaining features. This in turn results in lower overall sensitivity to immutable features, which we demonstrate empirically for different classifiers in Section **??**. Under certain conditions, this results holds theoretically[For the proof, see the supplementary appendix.]:

Theorem 1 (Protecting Immutable Features).
Let $f_\theta(\mathbf{x}) = \mathcal{S}(\mathbf{M}_\theta(\mathbf{x})) = \mathcal{S}(\Theta\mathbf{x})$ denote a linear classifier with softmax activation $\mathcal{S}$ (i.e. multinomial logistic regression*) where $y \in \{1, ..., K\} = \mathcal{K}$ and $\mathbf{x} \in \mathbb{R}^D$. If we assume multivariate Gaussian class densities with common diagonal covariance matrix $\Sigma_k = \Sigma$ for all $k \in \mathcal{K}$, then protecting an immutable feature from the contrastive divergence penalty (Equation* **??***) will result in lower classifier sensitivity to that feature relative to the remaining features, provided that at least one of those is mutable.*

It is worth highlighting that Theorem **??** assumes independence of features. This raises a valid concern about the effect of protecting immutable features in the presence of proxy features that remain unprotected. We discuss this limitation in Section **??**.

3.3 Illustration

To better convey the intuition underlying our proposed method, we illustrate different model outcomes in Example **??**.

Example 1 (Prediction of Consumer Credit Default).
Suppose we are interested in predicting the likelihood that loan applicants default on their credit. We have access to historical data on previous loan takers comprised of a binary outcome variable ($y \in \{1 = \text{default}, 2 = \text{no default}\}$) two

input features: 1) the subjects' *age*, which we define as immutable, and 2) the subjects' existing level of *debt*, which we define as mutable.

We have simulated this scenario using synthetic data with independent features and Gaussian class-conditional densities in Figure **??**. The four panels in Figure **??** show the outcomes for different training procedures using the same model architecture each time (a linear classifier). In each case, we show the linear decision boundary (green) and the training data colored according to their ground-truth label: orange points belong to the target class, $y^+ = 2$, blue points belong to the non-target class, $y^- = 1$. Stars indicate counterfactuals in the target class generated at test time using generic gradient descent for a fixed number of iterations.

In panel (a), we have trained our model conventionally, and we do not impose mutability constraints at test time. The generated counterfactuals are all valid, but not plausible: they are clearly distinguishable from the ground-truth data. In panel (b), we have trained our model with counterfactual training, once again not imposing mutability constraints at test time. We observe that the counterfactuals are clearly plausible, therefore meeting the first objective of Definition **??**.

In panel (c), we have used conventional training again, this time imposing the mutability constraint on *age* at test time. Counterfactuals are valid but involve some substantial reductions in *debt* for some individuals. By comparison, counterfactual paths are shorter on average in panel (d), where we have used counterfactual training and protected immutable features as described in Section **??**. The counterfactuals are also plausible with respect to the mutable feature. Thus, we consider the model in panel (d) as the most explainable according to Definition **??**.

`body_ecml_files/mediabag/../paper/figures/poc.pdf`

Fig. 1. Visual illustration of how counterfactual training improves explainability. See Example **??** for details.

4 Experiments

In this section, we present experiments that we have conducted in order to answer the following research questions:

Research Question 1 (Plausibility) *Does our proposed counterfactual training objective (Equation* **??***) induce models to learn plausible explanations?*

Research Question 2 (Actionability) *Does our proposed counterfactual training objective (Equation* **??***) yield more favorable algorithmic recourse outcomes in the presence of actionability constraints?*

Beyond this, we are also interested in understanding how robust our answers to RQ **??** and RQ **??** are:

Research Question 3 (Hyperparameters) *What are the effects of different hyperparameter choices with respect to Equation* **??***?*

4.1 Experimental Setup

4.2 Experimental Results

5 Discussion

1. Limited to classification models.
2. Proxy attributes of immutable features.

6 Conclusion

Acknowledgments. A bold run-in heading in small font size at the end of the paper is used for general acknowledgments, for example: This study was funded by X (grant number Y).

Disclosure of Interests. It is now necessary to declare any competing interests or to specifically state that the authors have no competing interests. Please place the statement with a bold run-in heading in small font size beneath the (optional) acknowledgments, for example: The authors have no competing interests to declare that are relevant to the content of this article. Or: Author A has received research grants from Company W. Author B has received a speaker honorarium from Company X and owns stock in Company Y. Author C is a member of committee Z.