Enhancing the Value of Counterfactual Explanations for Deep Learning *

Yan Jia[1][0000−0002−5446−6565], John McDermid[1][0000−0003−4745−4272], and Ibrahim Habli[1][0000−0003−2736−8238]

University of York, York, UK
{yan.jia, john.mcdermid, ibrahim.habli} @york.ac.uk

Abstract. Counterfactual examples can be used to explain a specific clinical prediction from a deep learning model by identifying what kind of feature changes would produce a different result, i.e. flipping the prediction's classification. On-going research seeks to refine the metrics for discovering counterfactual examples, given a specific input to a deep learning model. Our work enhances this by using feature importance to reveal how much individual feature changes in the counterfactual example contribute to flipping the prediction's classification, compared with the original. Our approach does not depend on the specific metrics used for generating the counterfactual examples, so it is general. It can be used either to gain further insight when the counterfactual examples have already been generated or to influence the generation of the counterfactual examples. We illustrate this novel approach with a healthcare example.

Keywords: Explainability · Deep Learning · Counterfactual Examples.

1 Introduction

Clinical predictions based on Machine Learning (ML) are having an increasingly profound impact on the safety and quality of healthcare services [10], e.g. by recommending treatments. Our focus in this paper is on using explainability for ML-based systems to assist a clinician in achieving a desired healthcare goal.

Much work on explainability for ML-based models focuses on feature importance explanations, which score or rank the input features, conveying the relative importance of each input feature to the model output (or prediction) [7]. However, this does not help model users to understand what they should do in order to achieve a desired goal. More recently, Wachter et al [12] introduced counterfactual explainability which produces counterfactual examples that identify what changes in inputs to the ML model would be needed to reverse (or "flip") the ML model prediction. In this paper, we are interested in identifying changes in ML model inputs or patient conditions, that would enable a clinician to achieve a desired goal for a given patient.

The counterfactual examples should be close to the initial inputs to the model as smaller changes from the initial inputs are more likely to be achievable. Thus

* Supported by the Assuring Autonomy International Programme.

the approach should measure how far the predicted outcome of the counterfactual is from the desired outcome and the *distance* from the counterfactual to the initial input. When there are many features, searching for counterfactuals which combine changes in multiple inputs is computationally expensive, so it is necessary to find efficient solutions and to make simplifying assumptions, e.g. that the effects of small changes in inputs are additive in order to flip the prediction.

In some situations, it is desirable to provide a set of *diverse* counterfactuals, e.g. alternative changes in treatment, so that a user can choose which one to implement [5]. Our work builds on this idea and seeks to provide more insight into the different counterfactual examples. Specifically, our work enhances the value of counterfactual explanations for deep learning classifiers by revealing how much each input feature change in the counterfactual example contributes to flipping the decision. This novel combination of diverse counterfactual explanations and feature importance gives insight that enables users to choose which alternative to implement – thus making the ML models more actionable.

2 Background

Counterfactual explanations have been studied in philosophy and psychology and the work of Kahneman and Tversky in the 1970s and 1980s [4] presages many aspects of counterfactuals now addressed in ML. The introduction of counterfactual explanations for ML is more recent [12] but there is already some evidence that users prefer counterfactuals over feature importance methods [1].

Counterfactual explanations were formalised by Wachter et al [12]. Generally, given an input x, an ML classifier f, and a distance metric d, a counterfactual explanation x' which produces the desired output y can be generated by solving the optimisation problem:

$$x' = argmin\{yloss(f(x'), y) + d(x, x')\} \tag{1}$$

where *yloss* "pushes" the counterfactual x' towards a different classification than the initial input x, and the second term keeps the counterfactual x' close to the initial input x. There are four desirable properties for identifying good counterfactuals [7]. First, they should achieve the desired outcome as closely as possible, which is related to the first term in Equation 1. Second, the counterfactuals should be as close as possible to the original instance, which is related to the second term in Equation 1, i.e. the distance measure. Third, the counterfactuals should be *sparse*, i.e. an ideal counterfactual needs to change only a small number of features from the original instance. Fourth, it is desirable to have *diverse* counterfactuals. On-going research seeks to incorporate these properties in the loss function and optimisation methods. An overview of existing counterfactual explanation methods for ML is provided by Verma et al [11].

3 Method

Our method combines feature importance with counterfactuals. Specifically, it uses DeepLIFT (Deep Learning Important FeaTures) [9] to assign a contribution

score to each feature that changed in a counterfactual example. This can help users to understand how much individual feature changes in the counterfactual example contribute to flipping of the prediction's classification compared with the original instance. Where diverse counterfactual examples are available, the feature importance can help to choose between them.

DeepLIFT is an additive feature attribution method, developed specifically for use with deep neural networks (NNs). DeepLIFT compares the activation of each neuron for the input features of interest to its "reference activation" and attributes to each input a contribution score according to the difference. The "reference activation" is a user-defined reference input representing a background value. In order to enhance the value of counterfactual explanations, we assign a contribution score to each feature that changed in the counterfactual examples using DeepLIFT where the initial or original input features provides the "reference activation"[1].

We chose DeepLIFT because it compares the counterfactual examples to the initial instance and assigns the contribution scores according to the difference in the predictions. In addition, it considers both positive and negative contributions of features, hence identifying the sign of dependencies between the input features and the output. Further, the contribution score is generated by a single backwards pass through the NN so the scores can be generated efficiently.

If there are many features in the counterfactual example that have a very low contribution score, e.g. less than 1%, then that example might be discarded. This facilitates the identification of sparse counterfactual examples which is particularly important when choosing between diverse counterfactuals (see section-2).

4 Clinical Example

In Intensive Care Units (ICUs), mechanical ventilation is a common intervention that consumes a significant proportion of ICU resources [13]. It is of critical importance to determine the right time to wean the patient from mechanical support. However, assessing a patient's readiness for weaning is a complex clinical task and it is potentially beneficial to use ML to assist clinicians [6]. Our example uses Convolutional NN (CNN) based on the MIMIC-III data set [3] to predict readiness for weaning in the next hour. 25 patient features are included in the model as shown in Table 1. The predicted outcome is the probability of weaning readiness in the next hour with 0.5 as the threshold (0 means wean; 1 means continue). The CNN architecture and details of this example can be found in [2].

We illustrate our method with a patient's record at a particular time as the original instance to generate the counterfactual examples using DiCE (Diverse Counterfactual Explanation) [8]. DiCE can generate multiple diverse counterfactuals and works for any differentiable model. Thus it is widely applicable given the characteristics of commonly used deep learning methods. Four counterfactuals are shown in Table 1 along with the original instance, where "—" means the

[1] The contribution score for the features that didn't change in the counterfactual examples is zero, due to the way DeepLIFT works.

Table 1. Counterfactual examples for a given original instance with contribution scores (shown in blue and in parentheses)

Features	Original instance	Counterfactual Examples 1	2	3	4
Admit Type	Emergency	—	—	—	—
Ethnicity	White	—	—	—	—
Gender	Female	—	—	—	—
Age	78.2	—	—	—	—
Admission Weight	86.5	—	—	—	—
Heart Rate	119	—	—	—	—
Respiratory Rate	24	21.9 (≤ 0.01)	—	24.1 (≤ 0.01)	21.7 (≤ 0.01)
SpO2	98	—	—	96 (≤ 0.01)	—
Inspired O2 Fraction	100	—	—	—	—
PEEP set	10	1.1 (-0.23)	9.2 (≤ 0.01)	2 (-0.2)	5.1 (-0.12)
Mean Airway Pressure	14	—	15.2 (≤ 0.01)	—	14.8 (≤ 0.01)
Tidal Volume (observed)	541	—	540.1 (≤ 0.01)	541.9 (≤ 0.01)	541.9 (≤ 0.01)
PH (Arterial)	7.46	—	7.49 (≤ 0.01)	—	—
Respiratory Rate(Spont)	0	—	13.1 (-0.06)	—	—
Richmond-RAS Scale	-1	—	0 (-0.32)	—	2 (-0.37)
Peak Insp. Pressure	21	—	—	—	—
O2 Flow	5	—	7.3 (-0.01)	—	2.4 (0.02)
Plateau Pressure	19	—	—	—	—
Arterial O2 pressure	124	123.6 (≤ 0.01)	123.6 (≤ 0.01)	123.6 (≤ 0.01)	124.3 (≤ 0.01)
Arterial CO2 Pressure	33	—	—	—	—
Blood Pressure (systolic)	101	—	—	—	—
Blood Pressure (diastolic)	65	—	—	—	—
Blood Pressure (mean)	76	—	—	—	—
Spontaneous breathing trials	0	1 (-0.06)	1 (-0.06)	1 (-0.07)	—
Ventilator Mode	18	9 (-0.38)	1 (-0.44)	1 (-0.52)	—
Predicted outcome	0.93	0.27	0.04	0.14	0.46

feature in the counterfactuals is not changed from the original instance. In order to enhance the value of the counterfactual examples, each changed feature in the counterfactuals is assigned with a contribution score (shown in parentheses and in blue) to gain insight into how much it contributes to flipping the prediction. For example, in Example 1, the sum of contribution score from changing PEEP set, Spontaneous breathing trials, and Ventilator Mode is 0.67, which is the difference between the original prediction and the new prediction. The changes of Respiratory Rate and Arterial O2 pressure in the counterfactual Example 1 contribute less than 1% each, which is negligible.

The benefit of adding the contribution score in the counterfactual examples is twofold. First, it can help the user to choose which example to implement. In our four counterfactual examples, Example 1 is attractive as it avoids a lot of unnecessary changes which make little contribution by comparison with the others, especially Example 2. Also, it helps the users to prioritise the changes with high contribution scores. Second, it can also help to generate sparse counterfactuals through post filtering. For example, we can add constraints that if the contribution score in the counterfactuals is less than 1%, then the feature is left unchanged. In counterfactual Example 4, when the features Respiratory Rate and Arterial O2 pressure are kept the same as the original input, the new prediction score is the same as the counterfactual Example 4 to two decimal places. Thus, this will improve the sparsity of the counterfactual.

5 Conclusion

We have introduced a novel method to enhance the value of counterfactual explanations by revealing how much individual feature changes in the counterfactual example(s) contribute to flipping the prediction's classification. Our method uses DeepLIFT to generate contribution scores for the features in the counterfactual examples. We illustrated the method to show how it can help in choosing between diverse counterfactuals generated by DiCE, potentially enabling identification of sparse counterfactuals to implement, i.e. making the counterfactual more readily actionable. Although we have used a specific healthcare example and DiCE for producing the counterfactual examples to illustrate the method, we believe it is general as it does not depend on the specific metrics used for generating the counterfactual examples. Future work will include exploration of further examples and more extensive assessment of the method in a clinical setting.

References

1. Fernandez, C., Provost, F., Han, X.: Explaining data-driven decisions made by ai systems: The counterfactual approach. arXiv preprint arXiv:2001.07417 (2020)
2. Jia, Y., Kaul, C., Lawton, T., Murray-Smith, R., Habli, I.: Prediction of weaning from mechanical ventilation using convolutional neural networks. Artificial Intelligence in Medicine (2020 (Submitted))
3. Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
4. Kahneman, D., Tversky, A.: The simulation heuristic. Tech. rep., Stanford Univ Ca Dept Of Psychology (1981)
5. Kunaver, M., Požrl, T.: Diversity in recommender systems–a survey. Knowledge-Based Systems **123**, 154–162 (2017)
6. Kuo, H.J., Chiu, H.W., Lee, C.N., Chen, T.T., Chang, C.C., Bien, M.Y.: Improvement in the prediction of ventilator weaning outcomes by an artificial neural network in a medical icu. Respiratory care **60**(11), 1560–1569 (2015)
7. Molnar, C.: Interpretable Machine Learning. Lulu. com (2020)
8. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 607–617 (2020)
9. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning. pp. 3145–3153. PMLR (2017)
10. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine **25**(1), 44–56 (2019)
11. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596 (2020)
12. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
13. Wunsch, H., Wagner, J., Herlim, M., Chong, D., Kramer, A., Halpern, S.D.: ICU occupancy and mechanical ventilator use in the United States. Critical care medicine **41**(12) (2013)