Counterfactual Explanations for Inferred Medical Treatments

Xinmiao Lin

Abstract

Deep learning models gain popularity in many domains with their universal matching capabilities but their decision making process is usually opaque. In personalized medicare where models decide on a treatment plan and doctors cannot understand the reasons behind this decision, it is unlikely that the doctors will trust the models. Counterfactual explanations help the models to tell explain the effects of an alternative treatment plan proposed by physicians. In this work, we develop a counterfactual inference model for medical treatments in an interactive framework. Our model consisted of one prediction model and one explanation model. The prediction model predicts the treatment plan for a patient, and the explanation model provides counterfactual explanations for that decision. The physicians could query the model with counterfactual treatment plans and the model will return the predicted risks and outcomes. Our work delivers good experimental results and shows that models can earn trust from the physicians.

1. Introduction

Deep learning is widely applied in many domains, such as handwritten zip code recognition [5], images classification [3], DeepFake [7], etc. In medicare fields where the deep learning models work closely with physicians to identify breast cancer [2] or to predict drug effects [8], the physicians are unlikely to trust the models if the models lack explanations for their decisions. Explainable machine learning models is an active research area that targets to explain the decisions made by deep learning models. Some applications include Visual Question Answering (VQA) [9], machine teaching [10], Explainable-AI [1].

A common causation explanation is using counterfactual, for example, in a scene of car accident, the inspectors may wonder "what if the car drove at 55 mph instead of 70 mph?" This question is discriminative to that specific situation only, and we could never get an answer of it unless we can peek through the parallel universe. Although the generation of counterfactual explanations is intrinsically hard, they are useful in medicare when the physicians give a treatment plan or identify a tumor from a CT-scan. Because one can alway ask questions such as "what if you give Anastrozole instead of Docetaxel to treat breast cancer? [1]"

In this work, we develop a VQA model targeting the treatment plans for breast cancer patients. The VQA model consisted of two submodels. The first model is the prediction model that takes as input the patient information and gives a treatment plan including the medicine and doses. The second model is the VQA model that provides counterfactual explanations. When the doctor is in doubt of the model's decisions, the doctor can ask the model "why did you choose medicine A, instead of B?" The model will then give answers such as : "with medicine A, the patient is likely to have effects C and D." Our work is inspired from this work [4] that uses two models to provide counterfactual explanations in videos.

We test our model in the Breast-Cancer dataset and evaluate the results with oncologists specialized in breast cancer in a single-blind experiment setting.

2. Method and Experiments

2.1. Datasets

2.1.1 Breast-Cancer

The Breast-Cancer dataset contains 100,000 anonymous patient background information, the breast cancer description and the treatment given. Each patient has a profile with a list of categorical values, for example, age, gender, cancer type (IDC, DCIS, ILC, LCIS, etc.), and so on. In addition, each patient is treated in one of the following ways: surgery, chemotherapy, hormonal therapy, biological therapy or radiation therapy [2]. The effects of the treatment are also documented in full details, such as the stage, size of the tumor, etc.

2.2. Prediction Model

The VQA model consisted of two models: prediction model and explanation model. The prediction model is a ResNet-50 [3] on the Breast-Cancer dataset with features that combine the patient profile information and the treatment received, and predicts the effects that the patient may have.

[1] https://www.cancer.gov/about-cancer/treatment/drugs/breast

[2] https://www.cdc.gov/cancer/breast/basic$_i nfo/treatment.htm$

In the evaluation phase, the prediction model takes as the input a certain patient profile and returns the best treatment plan with their effects.

2.3. Explanation Model

The explanation model is a LSTM model [6] trained to recognize counterfactual questions, e.g., "what if the patient X is treated with chemotherapy instead of hormonal therapy?", and parse the questions into a tuple of datapoint $(X, treatment)$ where X contains all the information of the patient X. Then, the tuple is given to the prediction model to get a treatment plan and a list of the predicted effects. The effects Y are then rephrased into a sentence in the format, "if the patient X is treated with $treatment$, then the patient will have effects Y." The counterfactual statement is then returned to the doctor for the evaluation of its validity.

3. Evaluation and Results

3.1. Evaluation

It is generally hard to evaluate the validity of treatment plans, and even so their counterfactual explanations. In our experiments, we invite 20 oncologists to evaluate the validity of the treatment plans. We also invite the oncologists to write treatment plans for a small subset of test data.

We conduct 2 set of experiments. The first type of experiments are on the validation of the treatment plans given by the prediction model and the other oncologists. The validation is single-blinded which means that the oncologists do not know whether the treatment plan is given by an oncologist or the model. During the evaluation of the treatment plans, the oncologists ask counterfactual questions which proposes an alternative treatment. These counterfactual questions are submitted to the evaluation model and also given to the other oncologists for counterfactual explanations. The second type of experiments evaluate the counterfactual explanations. The model predicted counterfactual explanations are mixed with the ones given by the oncologists, then they are asked to evaluate the validity of the explanations.

3.2. Results

We present the results of the first experiment in table 1. Each physician has been asked to give treatment plans for 20 patients, and these 20 patients belong to a test set of 140 patients. We see that more oncologists agree unanimously on the treatment plans given by the prediction model than by the oncologists themselves.

Each oncologist asks counterfactual questions for 140 patients. Out of these 140 questions, 20 will be answered by each oncologist, and 120 will be answered by the explanation model. In table 2, we present the results of the satisfaction of the oncologists with respect to the counterfactual explanations. The results show that more oncologists agree unanimously on the counterfactual explanations given by the explanation model than by themselves. This is quite normal because the model has knowledge of 100,000 patients and the different treatment plans, while the majority of the doctors will not be able to see 100,000 patients in their entire lifetime. Another factor could be attributed to the limited time to give treatment plans while in almost all situations in real life, multiple oncologists work together to give a treatment plan and they are allowed more time.

	Agree	Neutral	Disagree
physician-treatment	64%	15%	21%
model-treatment	72%	10%	18%

Table 1: The table of the first experiment. The first column denotes the type of treatment plans given by the oncologists or the prediction model. The metrics measure the treatment plans that all the oncologists agree. If there is at least one oncologist that disagrees or is neutral, then the treatment plan does not count toward the agree metric.

	Agree	Neutral	Disagree
physician-explanation	59%	30%	11%
model-explanation	75%	20%	5%

Table 2: The table of the second experiment. The first column denotes the type of counterfactual explanations given by the oncologists or the explanation model. The metrics measure the satisfaction of counterfactual explanations that all the oncologists agree. For example, if there is at least one oncologist that disagrees or is neutral, then the explanation does not count toward the agree metric.

4. Conclusion

In order for the deep learning models to be widely applied in healthcare, the models need to be explainable. In this work, we propose to use counterfactual explanations to justify treatment plans predicted by the model for breast cancer patients. Our model is evaluated by oncologists and shows promising results. We believe that counterfactual explanations are essential for learning and our work provides some beneficial background for future work to apply deep models in medicare.

References

[1] D. Doran, S. Schulz, and T. R. Besold. What does explainable AI really mean? A new conceptualization of perspectives. *CoRR*, abs/1710.00794, 2017.

[2] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber. Using deep learning to enhance cancer diagnosis and classification. 06 2013.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[4] A. Kanehira, K. Takemoto, S. Inayoshi, and T. Harada. Multimodal explanations by predicting counterfactuality in videos. *CoRR*, abs/1812.01263, 2018.

[5] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[6] Y. Liu, C. Sun, L. Lin, and X. Wang. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090, 2016.

[7] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019.

[8] N. Tatonetti, P. Ye, R. Daneshjou, and R. Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4:125ra31 – 125ra31, 2012.

[9] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. *CoRR*, abs/1810.02338, 2018.

[10] X. Zhu, A. Singla, S. Zilles, and A. N. Rafferty. An overview of machine teaching. *CoRR*, abs/1801.05927, 2018.