European Journal of Epidemiology (2022) 37:563–568
https://doi.org/10.1007/s10654-022-00892-3

ESSAY

A structural characterization of shortcut features for prediction

David Bellamy[1,2] · Miguel A. Hernán[1,2,3] · Andrew Beam[1,2,4]

Received: 15 June 2022 / Accepted: 19 June 2022 / Published online: 6 July 2022

Abstract
With the rising use of machine learning for healthcare applications, practitioners are increasingly confronted with the limitations of prediction models that are trained in one setting but meant to be deployed in several others. One recently identified limitation is so-called *shortcut learning*, whereby a model learns to associate features with the prediction target that do not maintain their relationship across settings. Famously, the watermark on chest x-rays has been demonstrated to be an instance of a shortcut feature. In this viewpoint, we attempt to give a structural characterization of shortcut features in terms of causal DAGs. This is the first attempt at defining shortcut features in terms of their causal relationship with a model's prediction target.

Keywords Causal inference · Prediction models · Machine learning

Introduction

Data analyses can be carried out for 3 tasks: description, prediction, and counterfactual prediction; the latter is used for causal inference [1]. Each of these tasks requires the combination of data with different types of external or *expert knowledge*. For example, when performing a causal inference task, a core piece of expert knowledge is the specification of the set of variables to adjust for confounding [2]. Domain experts may select those adjustment variables after representing their knowledge about the underlying causal structure using causal directed acyclic graphs (DAGs) [3].

In contrast, the role of expert knowledge about the causal structure is less clear for prediction tasks. The goal of prediction is to learn a mapping from a set of input features to the prediction target. For example, the pixels from a chest X-ray (the features) may be used to predict the presence of a disease (the target). Recently, prediction models based on machine learning have demonstrated impressive results on a broad set of problems that have historically been challenging prediction tasks [4–10]. The success in machine learning has largely been driven by *deep learning* [11] techniques, which are extremely flexible models that impose very few assumptions and limited amounts of explicit expert knowledge [8]. The canonical deep learning approach is to include as many features as possible and to learn, solely from data, which combinations and transformations of these variables result in the most accurate prediction model [12]. Thus, it seems that the role of expert knowledge has an increasingly small role to play in prediction modeling. However, recent work has brought into question the wisdom of the purely data-driven paradigm in prediction modeling. In this viewpoint, we describe the role of expert knowledge in prediction using causal DAGs.

An example of a shortcut feature

Suppose that we train a deep learning model to predict whether a patient has COVID-19 (the prediction target) using chest X-rays (the input features) from a healthcare system of 4 hospitals. The goal is that the model uses pathophysiological markers such as lung opacity, like a human physician, to create a clinical decision support tool that can

✉ Andrew Beam
andrew_beam@hms.harvard.edu

[1] CAUSALab, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[2] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[3] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[4] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

be deployed in the 4 hospitals as well as in additional hospitals that did not provide training data.

However, the model might find a shortcut to predict disease using the training data but without using pathophysiological markers. Because some hospitals are more likely to treat patients with COVID-19, the X-rays from those hospitals are more likely to correspond to individuals with COVID-19 than the X-rays from other hospitals. Therefore, like in a recent study, a deep learning model may associate the presence of a hospital-specific *watermark* in the top-right corner of each X-ray with a greater probability of COVID-19 diagnosis [13]. The watermark in the chest X-ray is an example of a predictor that has been referred to as a *shortcut feature* [13, 14] in the machine learning literature. While the presence of a watermark for a specific hospital does shift the relative probability of COVID-19, the predictive power of the watermark is of limited use when the deep learning model is applied to X-rays from new hospitals with different watermarks or no watermark at all.

The characterization and avoidance of shortcut features is an active area of research [15–28]. Here we argue that expert knowledge is required to define what a shortcut feature is and to distinguish between shortcut features and other features.

Three types of input features in prediction models

The causal DAG in Fig. 1A depicts the prediction target (COVID-19 diagnosis) and the input features (lung opacity and watermark from the X rays). The arrow from COVID-19 to lung opacity represents the causal effect of the infection that is mediated by changes in the lung, while the arrow from COVID-19 to watermark represents the causal effect of infection that is mediated by admitting hospital, and the absence of an arrow from COVID-19 to number of ribs represents the lack of effect of infection on the bones. We say that a path in a causal DAG is *open* when it only includes two types of nodes: colliders (or their descendants) that are conditioned on, or non-colliders that are not [29, 30]. Two variables connected by an open path are expected to be statistically associated.

Using causal DAGs, we can consider 3 types of input features based on the paths that link them to the prediction target:

(a) Non-shortcut features are connected with the prediction target exclusively through open paths that are conserved across the training and deployment settings, e.g., lung opacity.
(b) Shortcut features are connected with the prediction target through one or more open paths that are not conserved across the training and deployment settings, e.g., watermarks in the X-rays from the original hospitals (which do not provide information on COVID-19 diagnosis in other hospitals).
(c) Irrelevant features are not connected with the prediction target through any open paths in the training data, e.g., number of ribs.

Our classification is qualitative (because we do not provide a quantitative definition of "conserved path") and informal, but it is sufficient to extract some conclusions about the role of expert knowledge for prediction tasks and a basis for the development of more formal approaches.

The role of expert knowledge

Expert knowledge is not required to conclude that irrelevant features (c) are not helpful to improve the prediction model. Though they may be statistically associated with the prediction target by chance in finite samples, irrelevant features (c) will be automatically discarded by the deep learning model given enough training data. In contrast, both non-shortcut (a) and shortcut (b) features will be incorporated into the predictive model, but the only way to distinguish

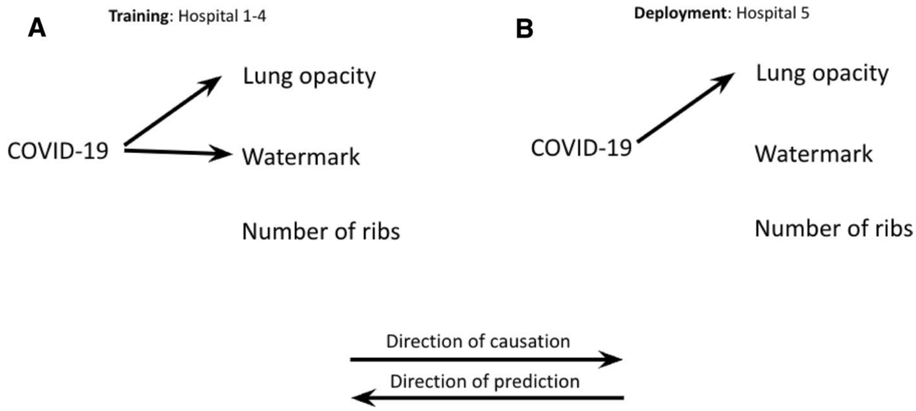

Fig. 1 Causal DAG underlying the prediction of COVID-19 status from chest X-rays across two different contexts. The training context (left) includes 4 hospitals with different COVID-19 prevalence and watermarking patterns. The deployment context (right) includes a fifth hospital with no watermark on the X-ray

between non-shortcut and shortcut features is the use of causal knowledge.

The classification of a feature as a shortcut feature depends on the set of deployment settings under consideration. Therefore, building a reliable prediction model requires that we restrict the model's access to shortcut features relative to the expected deployment contexts. In the chest X-ray example, lung opacity is a non-shortcut feature (a) because we expect it to be affected by COVID-19, regardless of which hospital performed the X-ray. In contrast, the watermark is a shortcut feature (b) because it lies on an open path to the prediction target in the hospitals used to train the model, but this path no longer exists in other hospitals in which the model is deployed, since the deployment hospital uses a different or no watermark.

This failure of predictive models to distinguish features (a) and (b) reflects the model's ignorance about the underlying causal structures in the training and deployment contexts. Therefore, model developers need to use expert knowledge to characterize the relevant set of deployment contexts, identify the shortcut features, and exclude the shortcut features from the data before the model is trained. In our example, model developers could accomplish this by cropping out all watermarks from the X-rays. Some shortcut features may be less obvious to experts or may be difficult to define concisely. For example, it has been shown that scanners from different manufacturers encode information in chest X-rays in slightly different ways [31]. This results in a nebulous scanner "signature", invisible to the human eye, that can result in shortcut learning if different patient populations are scanned using machines from different companies [31].

Causal structures for shortcut features

In general, features (a) and (b) do not have to be causally affected by the prediction target. Instead, the open path between a shortcut feature and the outcome may have a variety of causal structures (see Fig. 2), where a subset of the open path differs between the training and deployment contexts. In the most extreme case, a subset of the arrows and nodes are present in one setting, but absent in the other. Figure 2 illustrates some possible causal structures for shortcut features.

The left column shows the same causal structure that we examined in Fig. 1. The prediction target precedes the features temporally thus the direction of causation and prediction are opposite, also known as anti-causal prediction [32]. The central column shows a shortcut feature that shares a common cause with the prediction target in the training data but not in the deployment setting. The right column shows a shortcut feature that affects the prediction target in the training setting but not in the deployment setting. Figure 2 is not meant to be an exhaustive characterization of causal structures that permit shortcut learning, as we can imagine other structures with paths containing colliders between the shortcut feature and the prediction target that change between training and deployment.

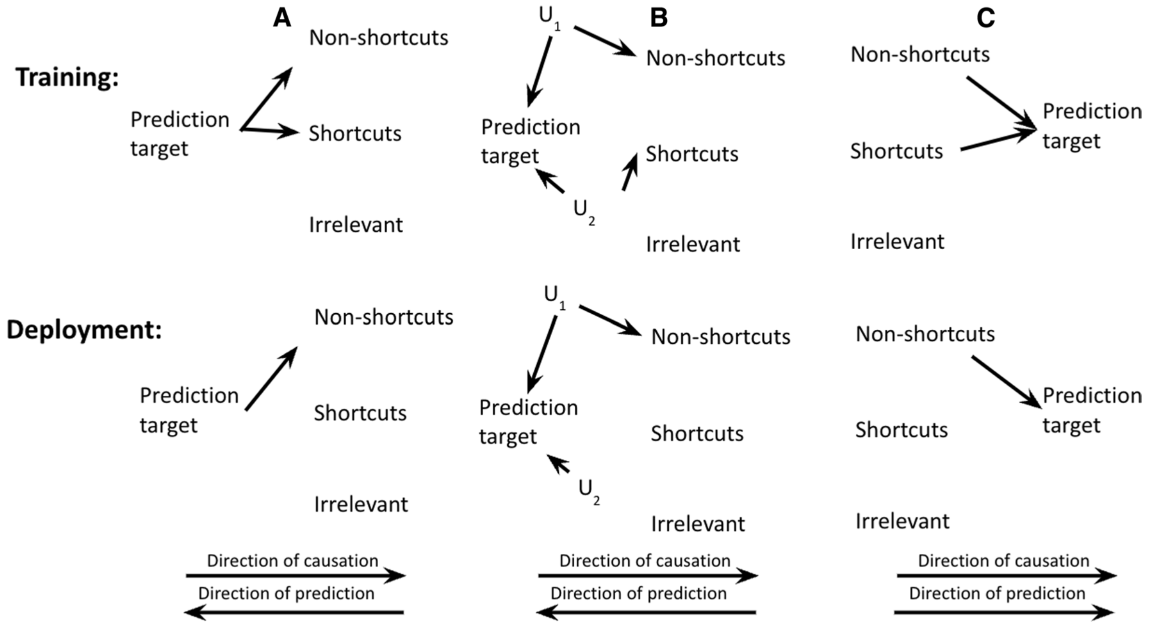

Fig. 2 Causal DAGs in the training (top row) and deployment (bottom row) context. Panel (**A**) depicts the same causal structure as Fig. 1, panels (**B**) and (**C**) depict alternative causal structures for shortcut features

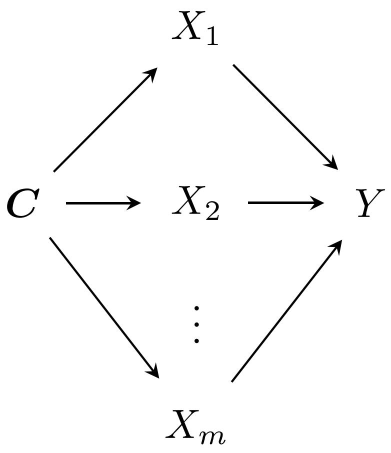

Fig. 3 Causal DAG from Wang et al. [25]. The features X_1, …, X_m cause the prediction target Y and share a common cause C

Other views on shortcut features

Shortcut features have attracted a great deal of attention from research groups in the fields of statistics, machine learning, and causal inference. We offer a brief overview of related efforts and provide a contrast with the characterization we provide in this work.

Arjovsky et al. [15] introduced *invariant risk minimization* (IRM) for data collected across a set of pre-specified settings: the model is trained to minimize the maximum error across settings (i.e., a minimax objective) and, by leveraging the pre-specification of deployment settings, the IRM framework is expected to discard shortcut features even if the causal structure of shortcut features is not explicitly characterized.

Other work has defined shortcuts in the *representations* learned by deep learning models rather than in the features themselves. For example, Wang and Jordan [25] consider a setting in which the input features cause the prediction target Y and there is an unobserved common cause C of all features (Fig. 3). Under the representation learning framework, models attempt to learn a mapping from the input feature space to a lower-dimensional representation of each sample. In general, we would like this representation to be highly predictive of the outcome and free of shortcuts, which the authors operationalize by defining a measure of a representation's *sufficiency* for the outcome. However, their notion of sufficiency does not apply to individual features but instead applies to the black box representation of the inputs. While this framework may enable the development of methods that optimize for representations with high sufficiency, it does not advance our reasoning about individual features, which is often the most useful level for reasoning about shortcuts in epidemiology.

In the field of statistical transportability, previous work has been concerned with estimating the conditional distribution of Y given X in a target domain using the distribution learned in the training domain and limited data from the target domain [33–39]. Correa and Bareinboim [26] developed an approach for solving this task that relies on factorizing the target conditional distribution according to its underlying causal DAG using a novel formalism known as a c*-factors. It has yet to be shown if the conditions required by this approach from this work can be stated in terms of graphical conditions that can be easily read off a DAG without requiring c*-factors. However, shortcut features per our definition would violate these transportability conditions. We believe that a direct structural characterization of shortcuts at the level of individual features (as opposed to at the level of c*-factors) is the most useful for practical discussion and enables intuitive reasoning about the surrounding issues.

Conclusion

Shortcut features pose a significant challenge to the safe and reliable deployment of prediction models. Models that learn shortcut features are unreliable and have the potential to cause catastrophic errors if used in clinical decision-making. An explicitly causal characterization of shortcut features, as we proposed here, facilitates the incorporation of expert knowledge into prediction models and may guide future work on remedies to the problem.

Acknowledgements The authors would like to thank Dr. Issa Dahabreh and other members of the CAUSALab at Harvard for their helpful feedback on this manuscript.

Author contributions All authors contributed to the conception of the ideas in this work and participated in the drafting, editing, and approval of the final manuscript.

Funding Dr. Beam received funding from the National Institutes of Health (K01 HL141771). Dr. Hernan received funding from the National Institutes of Health (P50 MH115846). Dr. Beam has equity and receives consulting fees from Generate Biomedicines, Inc. The authors have not disclosed any funding.

Data availability No data was used to produce the results in the manuscript.

Code availability No code was used to produce the results in the manuscript.

Declarations

Competing interests The authors have no other relevant competing interests to declare.

Consent to participate This manuscript did not involve any human participants.

Consent to publish This manuscript did not involve any human participants.

Ethical approval This study is a viewpoint and did not require ethics approval.

References

1. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. Chance. 2019;32:42–9.
2. Hernán MA, Robins JM. Causal inference: what if. Boca Raton: Chapman & Hall/CRC; 2020.
3. Pearl J, Glymour M, Jewell NP. Causal inference in statistics: a primer. Wiley; 2016.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. NIPS. 2012. p. 4.
5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
6. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;304:649–56.
7. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1711.05225
8. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319:1317–8.
9. Schmaltz A, Beam AL. Sharpening the resolution on data matters: a brief roadmap for understanding deep learning for medical data. Spine J. 2020. https://doi.org/10.1016/j.spinee.2020.08.012.
10. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2:719–31.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
12. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35:1798–828.
13. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv. 2020. https://doi.org/10.1101/2020.09.13.20193565.
14. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Machine Intell. 2020;2:665–73.
15. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. arXiv [stat.ML]. 2019. Available: http://arxiv.org/abs/1907.02893
16. Cheng PW, Lu H. 5 Causal invariance as an essential constraint for creating representation of the world: generalizing the invariance of causal power. The Oxford handbook of causal reasoning. 2017;65.
17. Creager E, Jacobsen J-H, Zemel R. Environment Inference for Invariant Learning. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. PMLR; 18--24 2021; 2189–2200.
18. Lu C, Wu Y, Hernández-Lobato JM, Schölkopf B. Invariant causal representation learning. 2020. Available: https://openreview.net/pdf?id=K4wkUp5xNK
19. Lu C, Wu Y, Hernández-Lobato JM, Schölkopf B. Nonlinear invariant risk minimization: a causal approach. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2102.12353
20. Moraffah R, Shu K, Raglin A, Liu H. Deep causal representation learning for unsupervised domain adaptation. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1910.12417
21. Moyer D, Gao S, Brekelmans R, Galstyan A, Ver Steeg G. Invariant representations without adversarial training. Adv Neural Inf Process Syst. 2018;31. Available: https://proceedings.neurips.cc/paper/2018/hash/415185ea244ea2b2bedeb0449b926802-Abstract.html
22. Puli A, Zhang LH, Oermann EK, Ranganath R. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2107.00520
23. Veitch V, D'Amour A, Yadlowsky S, Eisenstein J. Counterfactual invariance to spurious correlations: why and how to pass stress tests. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2106.00545
24. Kilbertus N, Parascandolo G, Schölkopf B. Generalization in anti-causal learning. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1812.00524
25. Wang Y, Jordan MI. Desiderata for Representation Learning: A Causal Perspective. arXiv [stat.ML]. 2021. Available: http://arxiv.org/abs/2109.03795
26. Correa JD, Bareinboim E. From Statistical Transportability to Estimating the Effect of Stochastic Interventions. IJCAI. 2019; 1661–1667.
27. Paul MJ. Feature selection as causal inference: experiments with text classification. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics; 2017. pp. 163–172.
28. Zhao H, Combes RTD, Zhang K, Gordon G. On learning invariant representations for domain adaptation. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. PMLR; 09—15, 2019;7523–7532.
29. Hernan MA, Robins JM. Causal inference causal inference: what if. Boca Raton, FL, USA: CRC Press; 2018.
30. Pearl J. Causality. Cambridge University Press; 2009.
31. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2019;2:31.
32. Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. The MIT Press; 2017.
33. Pearl J, Bareinboim E. Transportability of causal and statistical relations: a formal approach. Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewPaper/3769
34. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset shift in machine learning. MIT Press; 2008.
35. Zhang K, Schölkopf B, Muandet K, Wang Z. Domain adaptation under target and conditional shift. In: Dasgupta S, McAllester D, editors. Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA: PMLR; 17—19, 2013; 819–827.
36. Zhang K, Gong M, Schoelkopf B. Multi-source domain adaptation: a causal view. Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewPaper/10052
37. Magliacane S, van Ommen T, Claassen T, Bongers S, Versteeg P, Mooij JM. Domain adaptation by using causal inference to predict invariant conditional distributions. Adv Neural Inf Process Syst. 2018;31. Available: https://proceedings.neurips.cc/

paper/2018/hash/39e98420b5e98bfbdc8a619bef7b8f61-Abstract.html

38. Rojas-Carulla M, Schölkopf B, Turner R, Peters J. Invariant models for causal transfer learning. J Mach Learn Res. 2018;19:1309–42.
39. Tian J, Pearl J. A general identification condition for causal effects. eScholarship, University of California; 2002.