Uninformative Input Features and Counterfactual Invariance: Two Perspectives on Spurious Correlations in Natural Language

Anonymous ACL submission

Abstract

The natural language processing community has become increasingly interested in spurious correlations, and in methods for identifying and eliminating them. Gardner et al. (2021) argue that due to the compositional nature of language, *all* correlations between labels and individual input features are spurious. This paper analyzes this proposal in the context of a toy example, demonstrating three distinct conditions that can give rise to feature-label correlations through a simple PCFG. Linking the toy example to a structured causal model shows that (1) feature-label correlations can arise even when the label is invariant to interventions on the feature, and (2) feature-label correlations may be absent even when the label *is* sensitive to interventions on the feature. Because input features will be individually correlated with labels except in very rare circumstances, mitigation and stress tests should focus on those correlations that are counterfactually invariant under plausible causal models.

1 Introduction

Spurious correlations have increasingly preoccupied researchers in machine learning (Geirhos et al., 2020) and related fields, including natural language processing (Gururangan et al., 2018; McCoy et al., 2019, *inter alia*). However, the notion is frequently used without a formal definition. Gardner et al. (2021) propose a definition in terms of conditional probabilities: a feature X_i is spuriously correlated with the label Y unless $P(Y \mid X_i)$ is uniform. The definition can be generalized from uniformity to independence ($X_i \perp\!\!\!\perp Y$) without affecting the claims of the paper. They go on to argue that "in a language understanding problem, . . . *all* simple correlations between input features and output labels are spurious" (emphasis in the original). The property that individual input features should be independent of labels — which we will call *marginally uninformative input features (UIF)* — is treated as an assumption about the nature of language processing and also as a desideratum that datasets should satisfy: if the label can be predicted from input features alone, then the dataset is too easy.[1]

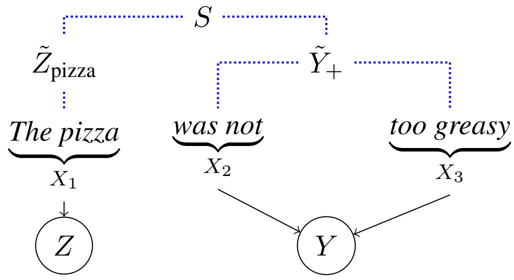

Figure 1: An instance from the toy model. The upper part of the figure corresponds to f_X, the function that generates the text via a PCFG (see fig. 2): nodes correspond to non-terminals in the grammar and edges represent context-free derivations. The lower part of the figure corresponds to the causal model of the sentiment Y and target Z. Here nodes correspond to variables and edges correspond to causal relationships.

The principle of UIF is based on the insight that linguistic context can invert the semantics of any subspan of a text (via, e.g., syntactic negation or discourse relations). Furthermore, the frequency of negation and other forms of semantic inversion may vary across datasets and deployment settings. A predictor that relies on, e.g., negation being rare, cannot be said to have truly achieved competence in the language processing task, and may perform poorly in domains in which these high-level distributional properties shift.

An especially provocative assertion of Gardner et al. is that all correlations between labels and individual input features have the same status. In the sentence *the pizza was amazing*, sup-

[1] To formalize the UIF assumption, it is necessary to clarify which features are "input features": bytes, phonemes, wordpieces, words, phrases, or sentences? The selection of input features is a property of the model and not the dataset; one could use character-level features for natural language inference or sentence-level features for sentiment analysis.

pose that both *pizza* and *amazing* are correlated with positive sentiment because the reviewers like pizza. There is an intuitive difference between these two correlations, because the modified sentence *the **movie** was amazing* should have the same label as the original, while *the pizza was **greasy*** should not. This intuition can be formalized using the framework of causality, which has generally treated spurious correlations as those that arise without a direct causal explanation (Simon, 1954). Given a causal model of the data generating process, we can compute the *interventional* distribution $P(Y \mid \mathrm{do}(X_1 := x_1), X_2, X_3)$, which corresponds to the distribution over Y in a data generating process in which the value of X_1 is surgically set to x_1 (Pearl, 1995; Peters et al., 2017; Feder et al., 2021).[2] When such interventions do not affect Y for any given example, we say that Y and X_1 are *counterfactually invariant* (Veitch et al., 2021). Violations of UIF are particularly troubling when they are accompanied by counterfactual invariance, because non-causal correlations often do not transfer to other domains (Schölkopf et al., 2012; Bühlmann, 2020).

This paper uses a toy example to relate the UIF property to (1) the production probabilities in probabilistic context-free grammars (PCFGs), and (2) counterfactual invariance in structured causal models. The connection to PCFGs provides additional motivation for the UIF criterion from the perspective of domain generalization, while clarifying the scenarios that can give rise to violations of UIF, which Gardner et al. attribute too narrowly to "bias and priming effects" in annotators. The connection to counterfactual invariance highlights the ways in which these concepts do and do not align. Efforts to remove artifacts from the training and evaluation of NLP systems will be most productive when focused at the intersection of these two views of spurious correlations: violations of UIF for input features to which the label is counterfactually invariant according to a causal model of the data.

2 Toy Example

Consider a simplified targeted sentiment analysis task (Mitchell et al., 2013), in which the sentiment is Y, the target is Z, and the sentences are all of the form (X_1, X_2, X_3), with X_1 specifying a target noun phrase, X_2 a copula-like expression, and X_3 a predicative adjectival phrase. For example, $Y = \textsc{Pos}$, $Z = \textsc{Pizza}$, $X_1 =$ *the pizza*, $X_2 =$ *turned out to be*, $X_3 =$ *crispy and delicious*. We will treat this data as generated from the causal model shown in fig. 2. This causal model can be summarized by two assertions: (1) the target Z is a direct effect of only the span X_1; (2) the sentiment label Y is a direct effect of only the spans X_2 and X_3. The function f_X can represent any generative model of text: an n-gram model, a grammar-based formalism, a deep autoregressive network, etc.

$$
\begin{aligned}
U :=& N_U && (1)\\
(X_1, X_2, X_3) :=& f_X(U, N_X) && (2)\\
Z :=& f_Z(X_1, N_Z) && (3)\\
Y :=& f_Y(X_2, X_3, N_Y). && (4)
\end{aligned}
$$

Figure 2: Causal model for the toy example shown in fig. 1. N_U, N_X, N_Y, N_Z indicate independent noise variables, and f_X, f_Y, f_Z indicate deterministic functions that map from causes to effects (for more details on the notation, see Peters et al., 2017).

Aside on the direction of causation. We treat the text as the cause of the labels, rather than the converse. This distinction is somewhat vexed (Schölkopf et al., 2012; Jin et al., 2021). In some cases the direction of causation is clear from the task (e.g., table-to-text generation, summarization, and translation), but often the problem could be framed in either direction: perhaps the writer had the label in mind when producing the text, and thus the text is an effect of the label; or perhaps it is better to think of the annotator, who must read the text to arrive at the label, regardless of the writer's intentions. When the labels cause the text, the notion of counterfactual invariance can be restated in terms of the invariance of text features to perturbations on labels, e.g. $P(X_1 \mid \mathrm{do}(Y := y), Z)$. As the toy example is meant to serve only an expository purpose, we leave elaboration of the relationship of UIF to such models for future work.

2.1 Counterfactual invariance $\not\Rightarrow$ UIF

The causal model implies several counterfactual invariance properties: intervention on X_1 will not affect Y, nor will intervention on X_2 or X_3 affect Z. This is because X_1 blocks the influence of X_2 and X_3 on Z, and vice versa for Y. Conversely, (X_3, Y) are not counterfactually invariant in gen-

[2] Space does not permit a discussion of the distinction between interventions and counterfactuals (see Pearl, 2009).

eral because X_3 is an ancestor of Y in the causal graph, and similarly for (X_2, Y) and (X_1, Z).

Counterfactual invariance does not imply that the associated input features are marginally uninformative of the label. Consider a classical spurious correlation in which pizza tends to receive positive sentiment and sushi receives negative sentiment. This correlation is produced when f_X encodes a PCFG with the top-level production:

$$\begin{aligned} S \rightarrow \quad & \tilde{Z}_{\text{pizza}}\ \tilde{Y}_{+} && (1+\alpha)/4 \\ & \tilde{Z}_{\text{sushi}}\ \tilde{Y}_{-} && (1+\alpha)/4 \\ & \tilde{Z}_{\text{pizza}}\ \tilde{Y}_{-} && (1-\alpha)/4 \\ & \tilde{Z}_{\text{sushi}}\ \tilde{Y}_{+} && (1-\alpha)/4, \end{aligned} \tag{5}$$

with the right column indicating the probability of each rule expansion and $\alpha \in [-1, 1]$.[3] The nonterminal symbols $\tilde{Z}_{\text{pizza}}, \tilde{Z}_{\text{sushi}}, \tilde{Y}_{+}, \tilde{Y}_{-}$ are intentionally chosen to correspond to the labels Z and Y. Subsequent rules in the grammar can then be designed to ensure that $\tilde{Z}_{\text{pizza}}$ usually produces values of X_1 that make $Z = \text{PIZZA}$ likely, and analogously for the other non-terminals and associated labels. The unification of PCFGs and structured causal models is shown in fig. 1.

When $\alpha \neq 0$, there may be an association between X_1 and (X_2, X_3). As a result, there exist (x_1, x_1') such that,

$$\begin{aligned} & P(Y|X_1 = x_1) \\ & \quad = \sum_{X_2, X_3} P(Y \mid X_2, X_3) P(X_2, X_3 \mid X_1 = x_1) \\ & \quad \neq \sum_{X_2, X_3} P(Y \mid X_2, X_3) P(X_2, X_3 \mid X_1 = x_1') \\ & \quad = P(Y|X_1 = x_1'), \end{aligned} \tag{6}$$

creating a violation of UIF. The same argument can be applied to $P(Z \mid X_2)$ and $P(Z \mid X_3)$. UIF is also violated in $P(Z \mid X_1)$, $P(Y \mid X_2)$, and $P(Y \mid X_3)$, but for a different reason: these distributions are conditioned on the direct causal parents of the labels in f_Y and f_Z. Manipulation of the data distribution to ensure that $\alpha = 0$ (deconfounding $\tilde{Y}$ and $\tilde{Z}$) can remove only the violations of UIF induced by f_X, but not those induced by the direct causal relationships encoded in f_Y and f_Z.

[3]The stochasticity of the grammar is encoded in the deterministic function f_X through the noise variable N_X. Let $N_X \sim \text{Uniform}(0, 1)$, and choose the first rule expansion of S when $N_X < (1+\alpha)/4$, the second rule expansion when $(1+\alpha)/4 \leq N_X < (1+\alpha)/2$, and so on.

Discussion. The example shows how violations to UIF can emerge via confounding, creating classical spurious correlations in the sense of Simon (1954): informativeness despite counterfactual invariance. Such correlations are unlikely to be robust because it is not difficult to imagine a domain in which the sign of α changes, impairing the performance of predictors that have learned the spurious correlation. In contrast, feature-label correlations that arise directly from the causal model, such as (Z, X_1), are only damaging under more extreme forms of concept shift, in which the meanings of the features themselves change.[4]

2.2 UIF $\not\Rightarrow$ Counterfactual Invariance

Violations of counterfactual invariance can occur even when UIF is satisfied. To show this, we supply two more productions for the grammar:

$$\begin{aligned} \tilde{Y}_{+} \rightarrow \quad & \text{COP}_{+}\ \text{ADJP}_{+} && \beta_{+} \\ & \text{COP}_{-}\ \text{ADJP}_{-} && 1 - \beta_{+} \end{aligned} \tag{7}$$

$$\begin{aligned} \tilde{Y}_{-} \rightarrow \quad & \text{COP}_{+}\ \text{ADJP}_{-} && \beta_{-} \\ & \text{COP}_{-}\ \text{ADJP}_{+} && 1 - \beta_{-} \end{aligned} \tag{8}$$

Here the non-terminal COP_{+} produces a "positive" copula in X_2 (*is*, *was*, *is universally agreed to be*), COP_{-} produces a negated copula in X_2 (*isn't*, *wasn't*, *was the furthest possible thing from*), ADJP_{+} produces positive-sentiment adjectival phrases in X_3 (*great*, *delicious*), and ADJP_{-} produces negative-sentiment adjectival phrases in X_3 (*disappointing*, *totally unappetizing*). There are two special cases of interest:

- When $\beta_{+} = \beta_{-}$, the probability of using a negated copula is independent of Y, so X_2 satisfies UIF with regard to Y, while X_3 generally does not.
- When $\beta_{+} = 1 - \beta_{-}$, the use of negation is balanced to make the distribution over sentiment terms independent of Y, so X_3 satisfies UIF with Y, while X_2 generally does not.

[4]This basic intuition is sometimes formalized as the *principle of sparse mechanism shift*, which states that complex causal systems are usually composed of smaller independent parts, and that domain shifts typically affect only a few components (Schölkopf et al., 2021). A related principle arises in the context of natural language: distributional frequencies are more likely to change across domains than categorical facts about language. Biber (1991), for example, makes this argument explicitly in the analysis of register. In our model, the implication is that the probabilistic rule expansions in f_X are more likely to change than the basic properties of the lexicon, which govern which terminal symbols can be emitted by each non-terminal.

Combining these cases, both X_2 and X_3 satisfy UIF with Y when $\beta_+ = \beta_- = \frac{1}{2}$, meaning that negated and non-negated copula are equally likely and are independent of Y.

Discussion. UIF is violated not only by confounding, as discussed in the previous section, but also in mild settings that do not meet any reasonable definition of bias: unless $\beta_+ = \beta_- = 1/2$ then at least one of X_2 and X_3 is marginally informative of Y. Furthermore, UIF has no impact on the counterfactual invariance of X_2 and X_3 on Y. Neither is counterfactually invariant even when the generative model is parametrized to make UIF hold for all input features (see also Pearl, 2009, page 185). This is because the overall sentiment can be directly affected by adding or removing negation and by flipping the polarity of the sentiment-carrying adjective.

3 Conclusions

In the toy example, violations of UIF arise from three distinct phenomena: confounding between the sentiment and the target ($\alpha \neq 0$, leading to $X_1 \not\perp Y$); confounding between the sentiment and the use of negation ($\beta_+ \neq \beta_-$, leading to $X_2 \not\perp Y$); and lack of a perfect balance in the probability of negation between positive- and negative-sentiment examples ($\beta_+ \neq 1 - \beta_-$, leading to $X_3 \not\perp Y$.) The conditions required to satisfy UIF are thus progressively less plausible as we move from X_1 to X_3, and full UIF is achieved only in the perfectly balanced case of $\alpha = 0, \beta_+ = \beta_- = \frac{1}{2}$. The number of such constraints will increase with the size of the grammar, making UIF vanishingly rare in more general settings. Note that this general conclusion follows from the PCFG analysis, and can be derived without reference to causality.

The toy example also demonstrates the disconnect between the UIF view of spurious correlations and the causal view: counterfactual invariance does not imply UIF because X_1 can be marginally informative of Y even when X_1 and Y are counterfactually invariant (these are the artifacts that we want to remove); UIF does not imply counterfactual invariance because both X_2 and X_3 can be uninformative of Y even when Y is sensitive to interventions on both features. From a theoretical perspective, it is unsurprising that these two views diverge, because UIF is a purely observational criterion while counterfactual invariance requires an explicit causal model. Indeed, this relationship is discussed in depth by Pearl (2009, §6.3), albeit outside the context of language. The two perspectives can be seen as complementary, in that violation of UIF is a necessary but insufficient condition for a spurious correlation in the causal sense.

It is of course possible to quibble with the causal model presented here, and in real applications it is likely impractical to construct full causal models of language. How then can we use causal insights to go beyond sensitivity analysis to design better benchmarks and more robust language understanding systems? In some cases it is possible to elaborate partial causal models of a task, with associated invariance properties: for example, the sentiment of a movie review should be invariant to (though not independent of) the identities of the actors in the movie. Several existing approaches can be viewed as instantiations of partial causal models: for example, data augmentation, causally-motivated regularizers, stress tests, and "worst-subgroup" performance metrics (and associated robust optimizers) can be seen as enforcing or testing task-specific invariance properties that provide robustness against known distributional shifts (e.g., Lu et al., 2020; Ribeiro et al., 2020; Kaushik et al., 2021; Koh et al., 2021; Veitch et al., 2021). Such approaches generally require domain knowledge about the linguistic and causal properties of the task at hand — or to put it more positively, they make it possible for such domain knowledge to be brought to bear.

A final observation, pertaining to both UIF and counterfactual invariance, is the parallel treatment of X_2 (the copula) and X_3 (the adjectival phrase). From a lexical semantic perspective, only X_3 is directly associated with the sentiment, while X_2 plays a functional role by potentially reversing X_3. It may therefore seem undesirable to learn a correlation between X_2 and Y, and preferable to attach that relationship exclusively to X_3. Yet neither UIF nor counterfactual invariance is capable of making such a distinction. While it is possible to enforce uninformativeness on X_2 heuristically, e.g. by sampling or augmenting the data to ensure $\beta_+ = \beta_-$, those same heuristics could be applied to enforce uninformativeness on X_3 by making $\beta_+ = 1 - \beta_-$. Singling out X_2 requires additional justification. Such a principle might be found in the multitask setting, in which we prefer feature-label informativeness to be sparse, with each feature directly informing only a few labels.

References

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Peter Bühlmann. 2020. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2021. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Herbert A Simon. 1954. Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, 49(267):467–479.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34.