Iterative Counterfactual Data Augmentation

Mitchell Plyler
Min Chi

Department of Computer Science, North Carolina State University
mlplyler@ncsu.edu, mchi@ncsu.edu

Abstract

Counterfactual data augmentation (CDA) is a method for controlling information or biases in training datasets by generating a complementary dataset with typically opposing biases. Prior work often either relies on *hand-crafted rules* or algorithmic CDA methods which can leave unwanted information in the augmented dataset. In this work, we show **iterative** CDA (ICDA) with initial, high-noise interventions can converge to a state with significantly lower noise. Our ICDA procedure produces a dataset where one target signal in the training dataset maintains high mutual information with a corresponding label and the information of spurious signals are reduced. We show training on the augmented datasets produces rationales on documents that better align with human annotation. Our experiments include six human produced datasets and two large-language model generated datasets.

Code — https://github.com/mlplyler/ICDA

Introduction

Counterfactual data augmentation (CDA) is a method that can reduce targeted biases in training datasets, and ideally, reduces those biases in models trained on those datasets (Lu et al. 2020). During CDA, counterfactuals are generated with roughly speaking the opposite bias of some original dataset. The original and counterfactual samples are concatenated into an augmented dataset where ideally their unwanted biases are balanced and canceled. In the literature, CDA can target specific biases through hand-crafted rules (Lu et al. 2020), or general, unwanted biases with human annotators (Kaushik, Hovy, and Lipton 2020). Both of these methods are costly either in expert knowledge or in annotator labor. Alternatively, cheaper, model-driven interventions can be noisy, both in where the interventions are made on source documents and how the interventions are made (Plyler, Green, and Chi 2021). In this work, we leverage rationale networks (Lei, Barzilay, and Jaakkola 2016) to decide where to make counterfactual interventions (Plyler, Green, and Chi 2021).

Given a sample, rationale networks seek a *subset* of the input with which to make a decision. Typically a portion of the network, the rationale selector, extracts some text, and another portion of the network, a classifier, makes a decision using the extracted text. The network learns to select subsets of text that make better predictions than other subsets of text (Lei, Barzilay, and Jaakkola 2016). Prior work (Chen et al. 2018) has shown that these networks are seeking the subset of text that maximizes the mutual information between the selected text and the prediction task. Often, this maximum mutual information (MMI) signal aligns with human reasoning and the ideal rationale network will find a policy that mimics the behaviour of human experts (Lei, Barzilay, and Jaakkola 2016). With an aligned rationale model, we can use CDA to maintain a training dataset's information that aligns with human reasoning and we can reduce information that does not align (Plyler, Green, and Chi 2021). This makes rationale networks a potentially ideal candidate for selecting where to make counterfactual interventions on documents.

Unfortunately, a core challenge for rationale networks are co-varying or spurious signals that cause the rationale network to converge to a policy where, in some cases, the spurious signals are used to make the prediction (Chang et al. 2019). Plyler, Green, and Chi showed that CDA with rationale derived interventions can help lower the mutual information between spurious text and a target label, and help the rationale network get closer to the optimal, MMI policy. Plyler, Green, and Chi argued the benefits of CDA are dependent on the error rate of the rationale selector used to make the interventions, and empirically, they showed a second rationale model, trained on the augmented dataset, had a lower rationale error rate. If the second rationale model, the one trained on the augmented dataset, has a lower error rate than the initial model, we should see more benefits from CDA using that second rationale model instead of the initial. In fact, it stands to reason we can iteratively apply CDA with a new rationale model that is improving with each iteration.

This work builds on the ideas of rationalization (Lei, Barzilay, and Jaakkola 2016), counterfactual data augmentation (Lu et al. 2020), their combination (Plyler, Green, and Chi 2021), and fixed-point processes to show the potential benefits of applying counterfactual data augmentation **iteratively**. Starting from an initial, noisy rationale model, we create a counterfactual dataset, train a new rationale network on the augmented data, and use that new network to create counterfactuals for the next iteration. We present information theoretic analysis, in a simplified setting, showing this iterative CDA (ICDA) algorithm forms a fixed-point process that should converge to a rationale network that better aligns with the maximum mutual information signal. Empirically, we show that our iterative process produces rationale models that align closely with human annotations. We perform experiments on six real, or human generated, datasets in the RateBeer and TripAdvisor settings. We also show that ICDA fits into the modern paradigm of generating a dataset using a large language model and training a light-weight classifier on that generated dataset. Across all eight experiments, ICDA outperforms the baselines.

Method

Problem Definition

Consider the typical supervised learning problem with inputs X and output labels Y. We seek to train some model, F, that maps $Y \leftarrow F(X)$. A variation on the typical supervised problem is the rationale learning problem. A rationale model uses some subset of the input, $X_M \subseteq X$, to make prediction, $Y \leftarrow F(X_M)$. The rationale model is tasked with *learning which subset of input to use* to make the **same** prediction as using X. Typically, there is one signal in the input which would be ideal for the rationale model to select. Following (Plyler, Green, and Chi 2021), we will call this subset X_1 and the label that corresponds with this subset Y_1. Often, X_1 is the subset of text that a human annotator would select as the *rationale* for making the label prediction. Therefore, the quality of the rationale model can be assessed by measuring the agreement or alignment between the human annotated rationales X_1 and the selected rationales X_M. A successful rationale model therefore learns to select the subset X_1. $X_M \leftarrow X_1$. **It is important to point out that this problem definition does not include the human annotations, ground-truth X_1, in the training dataset and they are only available in the test set for evaluation.**

In this work, we assume there are multiple signals or aspects in the input and some of these signals are correlated with the target label Y_1. More specifically, X_1 refers to the desired signals in the dataset and X_2 represents another subset of input text that are undesired or spurious signals. Often, we analyze the two aspect case where X_1 is desired and X_2 is considered spurious but correlated with the label. The general case would be for N signals or subsets in the dataset which may or may not be disjoint. $X \leftarrow \{X_1, X_2, ..., X_N\}$. A typical multi-aspect example are hotel reviews were aspects within the reviews could refer to the hotel's location, cleanliness, service, etc.

Background: Noisy CDA

Chen et al. showed these rationale networks are seeking the subset or signal in the training dataset that maximizes the mutual information with the labels under some constraints G. These constraints are typically over the size or contiguity of the rationales. This is referred to as the maximum mutual information (MMI) criteria (Chen et al. 2018).

$$\max_G I(X_M; Y) \quad \text{subject to} \quad M \sim G(X) \tag{1}$$

In our problem definition, the rationale network will ideally learn $X_M \leftarrow X_1$. For the typical rationale network to be applicable, we are assuming that X_1 is the most informative signal in the input. We also assume there is high mutual information between the spurious signal and the target label, $I(X_2, Y_1)$. Typically, the dimensionality of the problem precludes an exhaustive search over subsets, so the rationale network seeks this MMI solution through Monte Carlo (Lei, Barzilay, and Jaakkola 2016). The Monte Carlo approximation, and the high correlation or mutual information between X_2 and Y_1 means that our rationale network sometimes makes the mistake of selecting X_2 instead of X_1. The central hypothesis of (Plyler, Green, and Chi 2021) is that reducing the mutual information between X_2 and Y_1, $I(X_2, Y_1)$ will help the rationale network in identifying the correct relationship $X_M \leftarrow X_1$. Plyler, Green, and Chi showed that with perfect CDA, the spurious mutual information, $I(X_2, Y_1)$ would be eliminated in the augmented dataset. Perfect CDA would require perfect knowledge of X_1 and would make the process self-redundant. They reasoned that lowering mutual information between the spurious signal and the label more than the lower the mutual information between the target label and the target text would help the rationale network. They first defined the change in mutual information $\Delta I^a_{X,Y}$ from the original dataset (X, Y) to the augmented dataset (X^a, Y^a).

$$\Delta I^a_{X_i,Y_j} = I(X_i, Y_j) - I(X^a_i, Y^a_j) \tag{2}$$

They then defined conditions when CDA will be successful:

$$\Delta I^a_{X_2,Y_1} - \Delta I^a_{X_1,Y_1} > 0 \tag{3}$$

We adopt the same definition of success in this work. In their work, they proposed a three stage approach: train an initial rationale selector, generate a counterfactual dataset, and finally train a second rationale selector on the augmented dataset. They reasoned that if Equation 3 was satisfied, the second rationale selector should be better at identifying the target signal than the first. They analyzed an error model where the first or initial selector made the mistake of selecting and modifying the spurious signal X_2 instead of the original signal X_1 at an error rate α. In the augmented dataset, they derived conditional probabilities dependent on the initial conditional probabilities in the original dataset and the error rate of the initial selector, α.

Plyler, Green, and Chi argued that the benefits of CDA are dependent on the error rate of the rationale selector and showed that there is a minimum error rate, α, necessary for CDA to be beneficial. When the target signal is more informative than the spurious signal, $I(X_1, Y_1) > I(X_2, Y_1)$, the minimum error to see CDA benefits is actually pretty small. In fact, the initial selectors from Plyler, Green, and Chi imply CDA should not be beneficial.

Helpful Counterfactual Generation Errors

In this section, we will demonstrate how a suboptimal counterfactual generation process can, in fact, be advantageous for CDA by increasing the minimum error budget, thereby enhancing the benefits derived from CDA. Plyler, Green, and

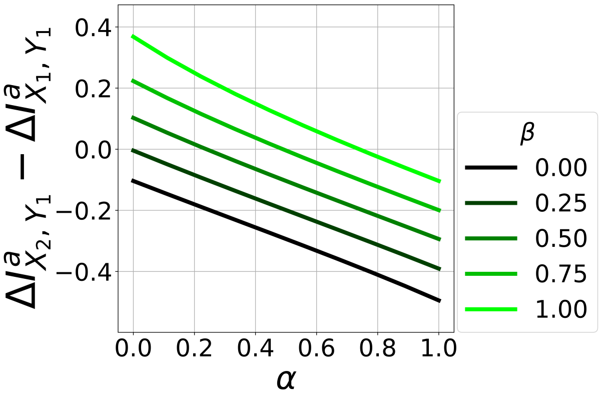


Figure 1: Increasing β helps CDA.

Chi's error or noise analysis was of course an approximation and it focused solely on the worst-case scenario: CDA perfectly flips X_2 when generating the counterfactual based on an incorrect rationale. In this section, we step back from the worst-case scenario by first considering the situation where X_2 is not perfectly flipped when it is modified during counterfactual creation.

Consider the example document

This beer smells great. It tastes fantastic.

If we consider Y_1 our desired label corresponding to the smell aspect, then the subset of text that we should consider for this document would be x_1 = *This beer smells great.* and the subset that we label spurious x_2 = *It tastes fantastic.* Consider the α noise from Plyler, Green, and Chi, we incorrectly select the spurious subset X_2 and we change its content to match the flipped label. This error would produce the counterfactual document:

This beer smells great. It tastes terrible.

Now, consider the scenario where we select the incorrect portion of text x_2 but we still flip its sentiment correctly according to X_1 to match the flipped label.

It tastes fantastic → *It smells terrible.*

The new counterfactual document would read:

This beer smells great. It smells terrible.

This document is obviously nonsensical, and it adds noise to our augmented dataset, but we will show this error is not as harmful as previous analysis where CDA flipped the taste sentiment.

Our augmented dataset now has two documents:

$< positive >$ *This beer smells great. It tastes fantastic.*

$< negative >$ *This beer smells great. It smells terrible.*

In half the augmented examples, we have the original conditional distributions: $P(Y_1|X_1)$ and $P(Y_1|X_2)$ just from including the original, unaltered sample. In the other half of the augmented examples, we have: $P(Y_1|X_1) = P(Y_1)$ and $P(Y_1|X_2) = P(Y_1)$ in the counterfactual samples. Lets say this happens only when we select the incorrect text which happens with probability α, and when we flip the sentiment of the selected text correctly which we say happens with probability β.

When Plyler, Green, and Chi defined conditional probabilities in the augmented dataset, $P(Y_1^a|X_1^a)$ and $P(Y_1^a|X_2^a$, they had an α portion of the documents that defined the error case and a $(1 - \alpha)$ portion of the documents that defined the non-error case. Here we refine the analysis and define the error case for X_1 as

$$P(Y_1^a|X_1^a)_{\text{error}} = \alpha\Big((1-\beta)P(Y_1) + \beta\Big(\frac{1}{2}P(Y_1) + \frac{1}{2}P(Y_1|X_1)\Big)\Big) \quad (4)$$

This shows for the α portion of documents, where we grab the correct text, with probability $(1 - \beta)$, we change the taste text as before and we have $P(Y_1)$. With probability β, we change the taste text to have negative smell sentiment, so half of the augmented dataset has no mutual information with the label $P(Y_1)$ and half maintain the information with the label $P(Y_1|X_1)$. Of course, our analysis should acknowledge the possibility that the rationale is correct, X_1 is modified, but the counterfactual is in correct, $1 - X_2$ is inserted. Therefore, we will say that the correct $1-X_1$ is inserted with probability β and the incorrect text is inserted with probability $1 - \beta$. We assume that the inserted X matches the counterfactual label $1-Y_1$. We can add back in the correct portion of the augmented dataset, $(1 - \alpha)P(Y_1|X_1)$ and simplify to find the relation. We can repeat the analysis for $P(Y_1|X_2)$ to find the augmented conditional probabilities:

$$\begin{aligned} P(Y_1^a|X_1^a) &= \Big(\frac{\alpha}{2} + \frac{1}{2} - \frac{\beta}{2}\Big)P(Y_1) \\ &\quad + \Big(-\frac{\alpha}{2} + \frac{1}{2} + \frac{\beta}{2}\Big)P(Y_1|X_1), \\ P(Y_1^a|X_2^a) &= \Big(-\frac{\alpha}{2} + \frac{1}{2} + \frac{\beta}{2}\Big)P(Y_1) \\ &\quad + \Big(\frac{\alpha}{2} + \frac{1}{2} - \frac{\beta}{2}\Big)P(Y_1|X_2). \end{aligned} \quad (5)$$

To analyze the problem, we approximate using binary variables (Plyler, Green, and Chi 2021): $p(Y_1|X_1) = .95$, $p(X_1) = p(X_2) = p(Y_1) = \frac{1}{2}$. Figure 1 shows this β error affects the success criteria of CDA. Remember the intersection of the x-axis shows the break even point of CDA and everything on the positive y-axis is a benefit while everything below the x-axis is a detriment. As our β error increases, we are actually increasing our error budget for seeing benefits from CDA. This suggests a degenerate counterfactual generator that only injects $X_1|1 - Y_1$ regardless of the context of the document, X_2, can actually be beneficial from an information theoretic view. We use this insight in section to strategically select counterfactual examples that increase our β rate.

Algorithm 1: Iterative CDA Procedure

```
Require: $D$ is a dataset with documents $X$ and labels $Y$.
  $D' \leftarrow D$
  while not converged do
    $S \leftarrow train\_selector(D')$
    $D^c \leftarrow infer\_counterfactuals(D, S)$
    $D^a \leftarrow concatenate(D, D^c)$
    $D' \leftarrow D^a$
  end while
```

Iterative Counterfactual Data Augmentation

Plyler, Green, and Chi argued the benefits of CDA are proportional to the error rate, α, of the initial rationale selector $S^{k=0}$. If training on counterfactually augmented data yields a lower error rationale selector, why not use that new rationale selector for another round of CDA? This question motivates our iterative approach where with each CDA iteration we train a better rationale selector and we lower the error rate of our counterfactual interventions.

Algorithm 1 outlines the ICDA method. We initialize our training dataset D' with some unaugmented, source dataset D. We train a rationale selector S^k on D' using both the original dataset D and the selector S^k to infer a set of counterfactuals D^c for the kth iteration. The augmented dataset D^a is the concatenation of the original dataset D and the counterfactual dataset D^c. The augmented dataset D^a becomes our training dataset D' for the next iteration. In this section, we will show that the error rate α of the selector S decreases with each iteration.

In our ICDA procedure, the error rate of the $k+1$th iteration's selector S^{k+1}, α^{k+1}, is dependent on the error rate of S^k, α^k:

$$\alpha^{k+1} = \psi(\alpha^k) \tag{6}$$

Where ψ is our iterative operator and consists of the functions $infer\_counterfactuals(D, S^k)$ and $train\_selector(D')$ from Algorithm 1. To illustrate the convergence of process6 and Algorithm 1, we will revisit the simplified scenario involving binary random variables. In this binary variable context, we will define our rationale scheme as follows:

Definition 1. *Given random variables X_1, X_2, and Y_1 along with n observations of these variables, our MMI rationalization scheme selects the variable in X that maximizes the agreement with expected error α.*

First notice the error rate of the rationale selector in scheme 1 is dependent on the difference in mutual information of the spurious signal and the target signal, δ. We will call this relation operator R:

$$\delta := I(X_1, Y_1) - I(X_2, Y_1) \tag{7}$$

$$\alpha^{k+1} = R(\delta, n) \tag{8}$$

An increase in δ would increase the gap between $P(Y_1|X_1)$ and $P(Y_1|X2)$ and therefore decreases the rate of incorrectly choosing X_2 which is the expected error α^{k+1}. Also note the strength of the rationale selector is dependent on the number of observations n. With enough observations, the rationale selector correctly identify the subset of X with the highest mutual information. For any positive δ, $\alpha^{k+1} \to 0$ as $n \to \infty$.

In this work, we derive the difference in mutual information, δ^a, in the augmented dataset using the augmented conditional probabilities in equations 5. Notice these conditional probabilities are dependent on the initial conditional and marginal probabilities of X_1, X_2, and Y_1 in the training dataset. We define these as θ. The conditional probabilities in the augmented dataset are also dependent on our helpful error β, and the error rate of the selector α^k. Going forward, we will assume $\beta = 1 - \alpha^k$ because of the counterfactual generation process in Section Implementation where we randomly sample from a set of candidate rationales which were produced with an error α^k. The properties of our augmented dataset are defined by our counterfactual generation process, equations 5, and the initial conditions in the training dataset. We will call the transform from the initial dataset to the augmented dataset operator J. J is shown in Figure 2 for our initial conditions and varying α^k.

$$\delta^a = J(\alpha^k, \theta) \tag{9}$$

Now, from Equation 8, we have:

$$\alpha^{k+1} = R(\delta^a) = R(J(\alpha^k, \theta)) \tag{10}$$

Now, with our definitions of operators R and J, we can show convergence of Algorithm 1.

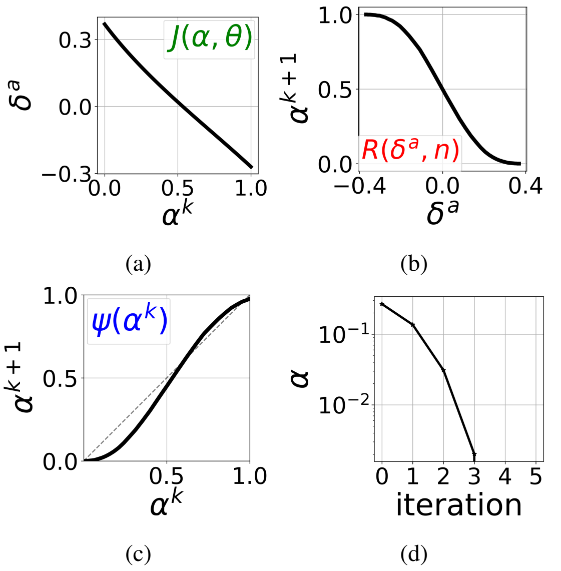
Theorem 1. *Process 6 converges to expected error $\alpha^{k+1} = R(\delta^a = I(X_1, Y_1), n)$ under Algorithm 1 for rationale scheme 1, an n such that $R^{-1} < J$ for some $\alpha \in [R(\delta^a = I(X_1, Y_1), n), \alpha_T)$, and an initial $\alpha^{k=0}$ such that $R(\delta^a = I(X_1, Y_1), n) <= \alpha^{k+1} < \alpha_T$.*

Our first condition on the rationale selector is driven by a sufficiently large n such that there is some region of α^k such that $\psi(\alpha^k)$ lies below the line $\alpha^k = \alpha^{k+1}$. This is the region where α iterations are decreasing. Our second condition on the error rate of the initial selector $\alpha^{k=0}$ ensures that we start in this decreasing region. We also know that $\psi(\alpha^k)$ is monotonic because it is a composition of two monotonic functions.

Finally, notice that our final error is lower bounded by $R(\delta = I(X_1|Y_1), n)$ which is the expected error when there is no mutual information $I(X_2|Y_1)$. This is how well the rationale selector is expected to perform when there is no spurious mutual information in the dataset. With these conditions, we can see that our iteration 6 is monotonically decreasing and lower bounded and therefore converges to $\alpha^{k+1} = R(I(X_1|Y_1), n)$ (Bibby 1974).

Theorem 1 guarantees convergence for an initial α in the region such that $\psi(\alpha^k)$ is less than the line $\alpha^k = \alpha^{k+1}$. The α^k where $\psi(\alpha^k)$ crosses that line is our α_T. Above α_T, our iteration is actually increasing toward another fixed point at $R(\delta^a(-I(X_2, Y_1), n)$ which is the augmented dataset where there is no mutual information $I(X_1, Y_1)$.

To demonstrate an example in the binary setting, we will assume $p(Y_1|X_1) = .9$, $p(Y_1|X_2) = .85$, $p(X_1) =$



(a)

(b)

(c)

(d)

Figure 2: (a) Definition of operator J. (b) MC Definition of operator R for $P(Y_1|X_1) = .95$. (c) Definition of fixed-point process 6 from (a) and (b). (d) shows convergence in simulation for an initial point $\alpha^{k=0} = .27$ determined by J and initial conditions θ.

$p(X_2) = p(Y_1) = \frac{1}{2}$, and our selector has 35 observations. Notice X_1 is our MMI solution. Figure 2 shows example relations for operator J, R, the map from α_k to α_{k+1}, and finally the α convergence through iterations. Note, for these initial conditions, algorithm 1 converges **quadratically**. Relation R is computed through simulation. See Appendix Additional Notes on ICDA Convergence for more information on Theorem 1 and examples under more initial conditions.

Implementation

Algorithm 1 outlines our iterative approach to counterfactual data augmentation. In this section, we will detail our implementation and make ICDA work in practice. First, we are iterating on the MMI implementation of the rationale scheme, so the first or zeroth iteration of ICDA is vanilla MMI rationalization. Subsequent iterations are the MMI scheme but on different datasets, specifically new counterfactually augmented datasets with each iteration.

In this work, we focus on rationalization over sentences. This scheme was actually used in the seminal work (Chen et al. 2018), but most other works have focused on rationalization over tokens (Lei, Barzilay, and Jaakkola 2016) (Yu et al. 2019) (Chang et al. 2019). We use the hierarchical transformer network (Pappagari et al. 2019) shown in Figure 3 in Appendix Implementation. A token level transformer encodes each sentence into a representation, and a sentence level transformer operates over the sentence representation. We apply the rationale criteria over the sentence representation. Sparsity is a desiderata of rationales (Lei, Barzilay, and Jaakkola 2016), so one can choose any of the sparsity constraints at this point. Our experiments use one sentence as the rationale, but this scheme does generalize to rationales over multiple sentences. After the rationale sentence is selected, the sentence representation produced by the token level transformer is left unmasked for the selected rationale, and we mask vectors of sentences not selected in the rationale. We then apply a pooling operation, max-out in our case, and perform classification using the selected sentence rationale representations.

Plyler, Green, and Chi generated counterfactual documents following a process that depended on the context in unmodified portion of the document. Their method relied on training generative models that worked with the initial rationale selector. To generalize to our iterative scheme, we could naively train new generative models for each CDA iteration that worked with the the rationale selector at each iteration, but this would obviously be computationally expensive. Alternatively, we follow a methodology that is more closely related to (Zeng et al. 2020) where we simply shuffle rationales between documents to generate the counterfactual samples. This means our counterfactual generation process no longer explicitly respects the context of the rest of the document, X_2.

Specifically, for a given rationale selector S, and dataset D, we generate a set of rationales A. We divide this set by the class label of the source documents, so for a binary classification problem, we have A_0 and A_1. To generate a counterfactual, we sample from the set with the flipped source label. Our generation process becomes:

$$Y_1^c \leftarrow 1 - Y_1; \quad X_1^c \leftarrow A|(1 - Y_1) \tag{11}$$

We sample new counterfactual documents during each epoch of training on the augmented datasets. We should also point out that this sampling approach is appropriate and works because we are rationalizing at the sentence level. In the beer review context, swapping one smell sentence for another smell sentence tends to produce coherent documents. If we were to rationalize at the token level, naively swapping rationales would most likely produce incoherent samples, and one would need to consider a generative approach like that used by (Plyler, Green, and Chi 2021) or one using more modern LLMs (Li et al. 2024).

We limit our augmented dataset size to be of equal size to the original dataset. We include original samples in the augmented dataset where the classifier had the lowest prediction error. Motivated by the analysis on helpful β errors, we know during counterfactual generation, it is helpful to insert a rationale of the correct aspect even if the rationale on the original document was incorrect. We therefore limit our rationale set for counterfactual generation A to be sourced from only documents where we made a correct prediction, and we take the 10% rationales, per class, where the model was most confident in correct prediction. We found that this limited rationale set helped the dev set classification loss converge in some cases.

During the iteration in Algorithm 1, $train\_selector$ is a key operation that produces a rationale selector S. In our implementation, we train three rationale selectors by varying the random seed across the three runs. During the zeroth iteration, we select the rationale model with the best loss on a dev set. For future iterations, on counterfactual data, we do the three runs over the random seeds and another run that is initialized by the rationale selector from the previous run and fine-tuned on the counterfactually augmented dataset. For the first counterfactual iteration, we select the model with the lowest loss on the augmented dev dataset. We want the iterative process to converge to a stable set of rationales. To encourage this, we select the model with the lowest loss that also has a decreasing rate of change in the rationale set from the previous iteration. If A^k defines the rationales produced during iteration k, the change in the rationale set is defined as $\Delta A = \frac{|A^{k+1} - A^K|}{|A^{k+1}|}$. After the first counterfactual iteration. We are seeking models that have a decreasing change in the rationale set: $\Delta A^{k+1} < \Delta A^k$. We define convergence of the iterative procedure, or the stopping criteria, as a selection of the rationale model initialized by the model from the previous iteration that is not improved by training on the augmented dataset. A more detailed version of our iterative algorithm is shown in Appendix Details.

Experiments

In this work, we adopt two common benchmark datasets for the rationale problem: RateBeer (McAuley, Leskovec, and Jurafsky 2012) and Tripadvisor (Wang, Lu, and Zhai 2010). These are two multi-aspect datasets where reviews describe beers and hotels respectively, but most reviews cover multiple aspects of the product.

For the RateBeer dataset, reviews typically describe the appearance, smell, taste, palatability, and overall rating of the beer. McAuley, Leskovec, and Jurafsky annotated about 1,000 reviews assigning each sentence in the review to an aspect label(s). These annotations have been used to evaluate the rationale alignment in (Lei, Barzilay, and Jaakkola 2016) and many works thereafter. The annotations are strictly in the test split of the data and are not used for hyper-parameter tuning. The key metric in rationale works, and this work, is the precision of the machine generated rationale relative to the ground-truth human annotated rationales. This measures the rate the machine generated rationales are also selected by human annotators. We follow the common practice of evaluating our methods on the appearance, smell, and palatability aspects (Lei, Barzilay, and Jaakkola 2016) (Chang et al. 2019).

The TripAdvisor dataset is much like the RateBeer dataset, but instead of smell, taste, etc., the reviews describe aspects like the location, service, and cleanliness of the hotel. The dataset was originally curated by (Wang, Lu, and Zhai 2010) and human-labeled rationales were collected by (Bao et al. 2018). Again, these rationales are in the test set only. Note there is a separate dataset for each aspect in the beer and hotel datasets.

We also evaluated ICDA on two LLM-generated datasets. Note, a popular paradigm in machine learning now is to generate a synthetic dataset using an LLM and training a classifier on the generated dataset. This saves an engineer the effort of curating a dataset traditionally, and it allows the engineer to deploy a much small, cheaper classifier as compared the LLM. We show that ICDA can fit into this paradigm with two example tasks: blue-tooth headphone reviews ”Connectivity” and restaurant reviews ”Restaurant”. For the ”Connectivity” dataset, we prompt the LLM in such a way that the connectivity aspect of the headphones are most informative of the label. For the ”Restaurant” dataset, we did not prescribe a desired rationale, but we found that most models converged to a ”food” aspect. Exact details for generating the datasets, and statistics relating to all datasets, are in Appendix Data.

We compare our iterative approach to three baselines from the literature. Note, these baselines have been reimplemented so that they align with our studied rationale criteria for fair comparison. The maximum mutual information (MMI) method from (Lei, Barzilay, and Jaakkola 2016) (Chen et al. 2018) is our first key baseline. This can be considered the base approach for the CDA and ICDA approaches. We also implement the complement control method from (Yu et al. 2019). Note that this reimplementation is the same as MMI+minimax and we did not use the introspective model. We are iterating on a version of counterfactual data augmentation from (Plyler, Green, and Chi 2021), so we of course include one CDA iteration as a baseline. Specifically, during our ICDA runs, we take the iteration-0 runs as the MMI implementation, the iteration-1 runs as the CDA implementation, and the last iteration as the ICDA run. This keeps the implementation details consistent between the methods and we can fully see the affect of the iterations.

Results

Table 2 shows the results on the LLM generated datasets. Our method, ICDA, showed an improvement over all baselines. The ICDA results here are further iterations on the MMI and CDA results. Based on the theory in Section Iterative Counterfactual Data Augmentation, we should expect CDA to show an improvement over MMI and ICDA to show another improvement over CDA. This was true for the Connectivity dataset, but on average, it wasn’t true for the Restaurant dataset. The convergence plots are shown in Appendix Additional Results and show that our experiments converged to the ”Food” aspect on the restaurant dataset $\frac{2}{3}$ times and the drop in average rationale precision for CDA is a product of that $\frac{1}{3}$ experiment converging to another aspect. Remember, for this restaurant dataset, we did not necessary prescribe a ”desired” aspect apriori.

Table 1 shows the rationale precision results on the human generated and annotated test datasets. Again, our ICDA method outperformed the baselines on datasets when averaged over the three runs of the experiment. To see how each run converged, see the charts in Appendix Convergence Plots.

Related Work

The concept of rationales was created in (Lei, Barzilay, and Jaakkola 2016) and its relation to mutual information was shown by (Chen et al. 2018). There have been a variety of follow-up works tackling different issues faced by these networks. Yu et al. addressed the problem of rationale degen-

	Appearance	Smell	Palatability	Location	Service	Cleanliness
COMP	56.4 ± 11.3	50.8 ± 4.0	30.9 ± 5.9	60.3 ± 6.9	64.7 ± 4.0	50.5 ± 7.2
MMI	47.5 ± 21.3	56.0 ± 10.8	23.1 ± 3.9	71.0 ± 4.1	66.8 ± 4.3	55.8 ± 2.7
CDA	62.9 ± 23.1	77.8 ± 11.9	24.7 ± 5.4	84.7 ± 2.2	70.0 ± 5.4	53.9 ± 15.7
ICDA	**66.6** ± 17.5	**93.6** ± 2.8	**37.6** ± 21.0	**88.0** ± .8	**74.0** ± 4.5	**56.0** ± 17.1

Table 1: Rationale Precision on the test set on real data. The appearance, smell, and palatability datasets are sourced from the beer and hotel datasets. Mean and standard deviation are reported.

	Connectivity	Restaurant
COMP	56.9 ± 4.9	41.3 ± 7.5
MMI	50.3 ± 8.9	54.6 ± 5.0
CDA	57.9 ± 7.6	51.5 ± 11.7
ICDA	**63.2** ± 5.2	**57.4** ± 17.2

Table 2: Rationale Precision on the test set on LLM generated data. Mean and standard deviation over are reported.

eration through the idea of complement control: minimizing the amount of information left in the complement of the rationale. Chang et al. leveraged the idea of counterfactuals for rationales, but did these counterfactuals were different selections over a single input document. Chang et al. used a variety of training dataset environments and invariant learning to find a rationale policy that generalizes across the domains. Our work builds most directly on Plyler, Green, and Chi which uses generative models to create a counterfactual dataset. Liu et al. and Zhang et al. have both taken a causal prospective on the rationalization problem where (Liu et al. 2023) builds counterfactuals during training by perturbing the rationales and Zhang et al. evaluated differences in predictions using the whole document versus the rationales. We selected MMI (Lei, Barzilay, and Jaakkola 2016) (Chen et al. 2018) and CDA (Plyler, Green, and Chi 2021) as methods to re-implement as baselines because our work is most directly built on these methods. We also re-implemented (Yu et al. 2019) because it directly generalizes to the rationalization over sentences setting while many methods are tied to the token-rationalization scheme. Its also important to point out that our data augmentation method is offline from model training and the $train\_rationale$ procedure in Algorithm 1 can generally be replaced with any rationalization strategy.

Prior work in the causal modeling community focused on identifying causal signals and helping models ignore spurious signals. Sun et al. 2021 shows that identifying such causal signals helps models become shift-invariant. Veitch et al. 2021 leverages causal inference and proposes the idea of counterfactual invariance as a model requirement and training strategy for avoiding spurious correlations.

Counterfactual data augmentation was introduced as a method for controlling gender bias in datasets (Lu et al. 2020). They hand-crafted a strategy for changing gender carrying terms in the datasets in such a way that the augmented dataset would have less gender bias. This idea has been built upon in the literature. Zeng et al. created counterfactuals by swapping named entities in a training dataset, and we should note our method of shuffling rationales between documents is similar counterfactual generation process. Kaushik, Hovy, and Lipton used human annotators to generate counterfactual datasets and showed that this can help downstream models generalize out-of-domain and Deng et al. adapted this strategy to active learning. Li et al. explored prompting LLMs for counterfactual generation.

Impact, Limitations, and Conclusions

While the datasets studied were relatively benign, one can imagine this as a methodology for controlling more serious spurious signals in training datasets. To that end, we should also acknowledge that this method relies on generating counterfactuals without human intervention or annotation. For errant cases, ICDA could produce counterfactual documents that are not factual, and a user should be aware of the dataset they are using and how the algorithm could manipulate that data.

Our theory in the binary section assumed the rationale model at iteration k will have the expected error from R and is not inherently random. The NLP rationale learning problem is inherently stochastic. Lei, Barzilay, and Jaakkola framed this a reinforcement learning problem where the selector is our agent and the hard selection over tokens are the available actions. It is often the case that our agent converges to a point where the rationales are low quality. We therefore selected the best MMI model, according to dev set prediction loss, from three runs with different random seeds as a our initial rationale selector for future iterations. Tables 1 and 2 show the MMI results vary. This noise is not factored into the theoretical framing for the binary case.

The ICDA approach is more computationally expensive than the baselines because it is running the baselines multiple times. Figure 2 suggests ICDA should converge quadratically. Across all experiments, Figure 4 in the Appendix shows ICDA converged in five or less iterations. We always trained a model to convergence on each augmented dataset. It might be possible to save computation, and allow ICDA to see better counterfactuals earlier, by not training to convergence during the earlier iterations.

This work presents an iterative approach to counterfactual data augmentation. We show how a process that starts with noisy interventions can self correct and converge to a process with less noise. Specifically, we showed how initial rationale models that aligned relatively poorly with human annotations could be iteratively improved through a CDA scheme that does not rely on human annotation, domain specific knowledge, or generative models. Our iterative approach outperformed the baselines for both human sourced datasets as well as LLM generated datasets.

Acknowledgments

This research was supported by the NSF Grants: Integrated Data-driven Technologies for Individualized Instruction in STEM Learning Environments (1726550), CAREER: Improving Adaptive Decision Making in Interactive Learning Environments (1651909), and Generalizing Data-Driven Technologies to Improve Individualized STEM Instruction by Intelligent Tutors (2013502).

References

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving Machine Attention from Human Rationales. *arXiv:1808.09367 [cs]*. ArXiv: 1808.09367.

Bibby, J. 1974. Axiomatisations of the average and a further generalisation of monotonic sequences. *Glasgow Mathematical Journal*, 15(1): 63–65.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. S. 2019. A Game Theoretic Approach to Class-wise Selective Rationalization. *arXiv:1910.12853 [cs, stat]*. ArXiv: 1910.12853.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. S. 2020. Invariant Rationalization. *arXiv:2003.09772 [cs, stat]*. ArXiv: 2003.09772.

Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. *arXiv:1802.07814 [cs, stat]*. ArXiv: 1802.07814.

Deng, X.; Wang, W.; Feng, F.; Zhang, H.; He, X.; and Liao, Y. 2023. Counterfactual Active Learning for Out-of-Distribution Generalization. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11362–11377. Toronto, Canada: Association for Computational Linguistics.

Kaushik, D.; Hovy, E.; and Lipton, Z. C. 2020. LEARNING THE DIFFERENCE THAT MAKES A DIFFER- ENCE WITH COUNTERFACTUALLY-AUGMENTED DATA. 17.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. Austin, Texas: Association for Computational Linguistics.

Li, Y.; Xu, M.; Miao, X.; Zhou, S.; and Qian, T. 2024. Prompting Large Language Models for Counterfactual Generation: An Empirical Study. ArXiv:2305.14791 [cs].

Liu, W.; Wang, J.; Wang, H.; Li, R.; Deng, Z.; Zhang, Y.; and Qiu, Y. 2023. D-Separation for Causal Self-Explanation.

Lu, K.; Mardziel, P.; Wu, F.; Amancharla, P.; and Datta, A. 2020. Gender Bias in Neural Natural Language Processing. In Nigam, V.; Ban Kirigin, T.; Talcott, C.; Guttman, J.; Kuznetsov, S.; Thau Loo, B.; and Okada, M., eds., *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, Lecture Notes in Computer Science, 189–202. Cham: Springer International Publishing. ISBN 978-3-030-62077-6.

McAuley, J.; Leskovec, J.; and Jurafsky, D. 2012. Learning Attitudes and Attributes from Multi-Aspect Reviews. *arXiv:1210.3926 [cs]*. ArXiv: 1210.3926.

Pappagari, R.; Żelasko, P.; Villalba, J.; Carmiel, Y.; and Dehak, N. 2019. Hierarchical Transformers for Long Document Classification. ArXiv:1910.10781 [cs, stat].

Plyler, M.; Green, M.; and Chi, M. 2021. Making a (Counterfactual) Difference One Rationale at a Time. In *Advances in Neural Information Processing Systems*, volume 34, 28701–28713. Curran Associates, Inc.

Wang, H.; Lu, Y.; and Zhai, C. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, 783–792. New York, NY, USA: Association for Computing Machinery. ISBN 9781450300551.

Yu, M.; Chang, S.; Zhang, Y.; and Jaakkola, T. 2019. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4092–4101. Hong Kong, China: Association for Computational Linguistics.

Zeng, X.; Li, Y.; Zhai, Y.; and Zhang, Y. 2020. Counterfactual Generator: A Weakly-Supervised Method for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7270–7280. Online: Association for Computational Linguistics.

Zhang, W.; Wu, T.; Wang, Y.; Cai, Y.; and Cai, H. 2023. Towards Trustworthy Explanation: On Causal Rationalization. In *Proceedings of the 40th International Conference on Machine Learning*, 41715–41736. PMLR. ISSN: 2640-3498.