C^{2}L: Causally Contrastive Learning for Robust Text Classification

Seungtaek Choi[1*], **Myeongho Jeong**[1*], **Hojae Han**[2], **Seung-won Hwang**[2]

[1]Yonsei University, [2]Seoul National University
{hist0613, wag9611}@yonsei.ac.kr, {stovecat, seungwonh}@snu.ac.kr

Abstract

Despite the super-human accuracy of recent deep models in NLP tasks, their robustness is reportedly limited due to their reliance on spurious patterns. We thus aim to leverage contrastive learning and counterfactual augmentation for robustness. For augmentation, existing work either requires humans to add counterfactuals to the dataset or machines to automatically matches near-counterfactuals already in the dataset. Unlike existing augmentation is affected by spurious correlations, ours, by synthesizing "a set" of counterfactuals, and making a collective decision on the distribution of predictions on this set, can robustly supervise the causality of each term. Our empirical results show that our approach, by collective decisions, is less sensitive to task model bias of attribution-based synthesis, and thus achieves significant improvements, in diverse dimensions: 1) counterfactual robustness, 2) cross-domain generalization, and 3) generalization from scarce data.

Introduction

Deep learning models have been successful in natural language tasks, aiming to learn how to correlate input features to desired outputs. However, these models are reported to often rely on "spurious" features, hindering the robustness– An illustrative example from (Wang and Culotta 2020a) is, classifying the sentiment of "This Spielberg film was wonderful", to positive. Though the term *Spielberg* itself does not have a causal effect to predicting the review as positive, if the given dataset contains many positive reviews for his movies, it can correlate to positive. In contrast, another term *wonderful* has a causal effect. While both features have high predictive power in this particular dataset, using *Spielberg* for prediction does not generalize well to another set where his movie is not as favored.

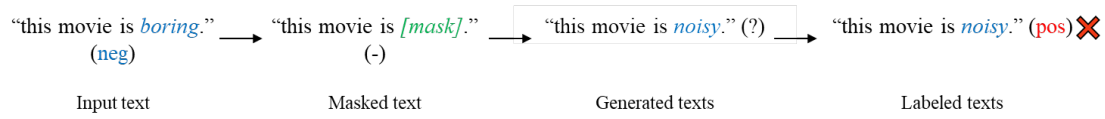
In this paper, we aim to reduce spurious correlations, to rely more on robust features, with more "causal" effects to predicting class labels and generalizing better to new data. For finding causal features in language tasks, existing efforts look for human annotations, namely counterfactual data augmentations (Kaushik, Hovy, and Lipton 2019), collecting the minimally-dissimilar yet differently labeled examples. By highlighting the different parts between original and counterfactual texts, the model could identify the causal correlations of the given task. In our example, the term *wonderful* is considered causal, as editing the term with *bad* generates its counterfactual pair of negative sentiment.

However, collecting such counterfactual pairs from human annotators is expensive and thus hard to scale up. Alternatively, there have been efforts to collect such counterfactual pairs automatically, to avoid high human-annotation costs: First, **matching-based** approaches (Wang and Culotta 2020b) find the most similar samples of opposite labels within the dataset. These will be effective when the given dataset contains counterfactual pairs, which may not be always practical. Second, **attribution-based** approaches (Moon et al. 2020; Liang et al. 2020; Han et al. 2021) no longer rely on the dataset coverage, but use attributions from the task model instead, such as attention or gradients, to decide "candidate" causal term as high-attributed features. This approach can be viewed as distilling causality from the task model itself, to reduce the dependence on the dataset coverage, though it may propagate the bias within the task model, e.g., *Spielberg* is highly-attributed.

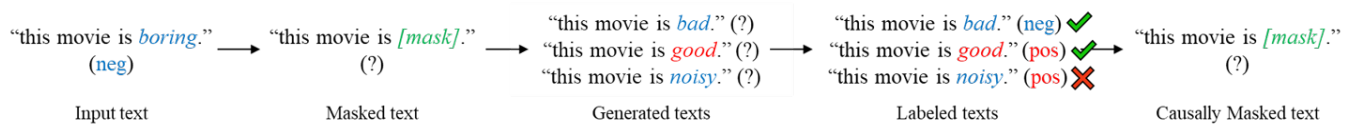
Our work is of combining the strength of the two approaches, by reducing the bias, from both dataset and task model. Specifically, we propose a novel pipeline for the unbiased causality identification, which consists of the following two steps: 1) candidate proposal and 2) candidate validation. The first step identifies "candidate" causal terms, by using attention or gradient as in (Moon et al. 2020; Liang et al. 2020). Then, the second step validates the causality of candidate terms, by estimating a causal "treatment" effect of the word, for classifying the given text into its class label, *i.e.*, "outcome", known as Individualized Treatment Effect (ITE) in causality literature (Shpitser and Pearl 2006; Shalit, Johansson, and Sontag 2017).

Our distinction is "robustifying" the second step, for which we contrast ITE validation of existing and our approaches as in Figure 1. Existing efforts (Wang and Culotta 2020b; Klein and Nabi 2020) first generate a single counterfactual example, then observe if the predicted label differs from the original label, which may be incorrect if the task is biased or noisy for the generated sentence as shown in Figure 1a. In contrast, we make a collective decision over multiple generations, as in Figure 1b, and consider decision "distributions" for making

[*]These authors contributed equally.



(a) Single decision



(b) Multiple decisions

Figure 1: Comparative illustration of (a) single and (b) multiple decisions for causality estimation. If we make a single decision, an incorrectly labeled (marked with X) counterfactual review, such as "this movie is *noisy*" with positive sentiment, would negatively affect the subsequent training process. On the other hand, making a collective decision from multiple counterfactual reviews is less sensitive to a few noisy assignments.

a validation. Though the task model still makes a mistake for "*noisy*", the overall decision checks if the predicted labels are diverse (or, close to uniform), which is much less sensitive to a few noisy/biased predictions.

Finally, in order for models to better understand the identified causal features, we propose a Causally Contrastive Learning (C^2L), which contrasts the original text with its causally related pairs. Specifically, we synthesize a counterfactual pair and factual pair, by masking the identified causal term and non-causal terms respectively. To illustrate, the example review with the term "*boring*" can be augmented into a counterfactual sentence, where the causal term is masked, *e.g.*, "the movie was `[MASK]`", such that its label cannot be determined. Similarly, a factual sentence is built by masking a non-causal term, such as "the" and "movie", still preserving the original semantics. That is, the model can learn that the term "*boring*" is causal to its label, by contrasting the original text with its counterfactual pair, as the masked token incurs label flips, while contrasting with the factual pair helps the model to learn being invariant to the non-causal features. With this causality-aware contrastive training, the robustness of the text classification model can be largely improved. Experimental results show that our method improves the robustness in various dimensions: 1) counterfactual examples, 2) cross-domain, and 3) data scarcity.

Methodology

This section discusses the overall framework of our proposed method, illustrated in Figure 2, which consists of three parts: (1) a base classification model, (2) a counterfactual sample synthesizing module, and (3) a contrastive learning (CL) objective.

Base Classification Model

The text classifier $f_\theta : x \rightarrow y$ maps an input x to the corresponding class among the N classes $y \in \{1, ..., N\}$. The input sequence x is a sequence of words $w_i \in \mathcal{V}$, *i.e.*, $x = [w_1, ..., w_T]$, where $\mathcal{V}$ is the vocabulary set and T is the length of the input. Here, the full corpus $\mathcal{D} = \{(x, y)\}$ is a collection of all inputs. The model parameters θ are trained to minimize the cross-entropy loss $\mathcal{L}_{task}$ between the predicted label $\hat{y}$ and the ground-truth label y:

$$\mathcal{L}_{task}(x_i, y_i; \theta) = -\sum_{j=1}^{N} y_i^j \log \hat{y}_i^j, \quad (1)$$

where j is the class index.

Unbiased Causality Identification

For more robust text classification, it is important to distinguish causal correlations from spurious correlations. Specifically, a causal feature is, in the context of text classification (Wang and Culotta 2020a,b), defined as follows: a word w is a **causal feature** in the input text x if, all else being equal, one would expect w to be a determining factor in assigning a label to x.

In this work, we present a novel pipeline, built on the existing label-indicative word identification methods, to efficiently narrow down the number of candidate words, followed by our **unbiased** causal word identification methods.

Step 1: Candidate Proposal The first step is selecting candidate tokens based on attribution scores. There are several ways to identify the causal features of the given text in our pipeline. For instance, (Moon et al. 2020) find causal tokens by collecting highly attended words of the trained classifier. However, the limitations of attention as explanation (i.e., an indicator of causal relationship) have been recently addressed (Jain and Wallace 2019; Grimsley, Mayfield, and Bursten 2020). As a more straightforward solution for capturing the attribution of each token, we can explore the tokens with the large gradients toward the gold label.

Formally, given an input x and its task label y, we compute the gradient magnitude g_i of i-th token w_i (specifically, the positional embedding $\mathbf{w}_i^p$ of token w_i) as follows:

$$g_i = \|\nabla_{\mathbf{w}_i^p} \mathcal{L}_{task}(x, y; \phi)\|^2, \quad (2)$$

where the magnitude is computed from the classifier f_ϕ. Note that the parameter ϕ is optimized only for the classification loss (Eq. 1).

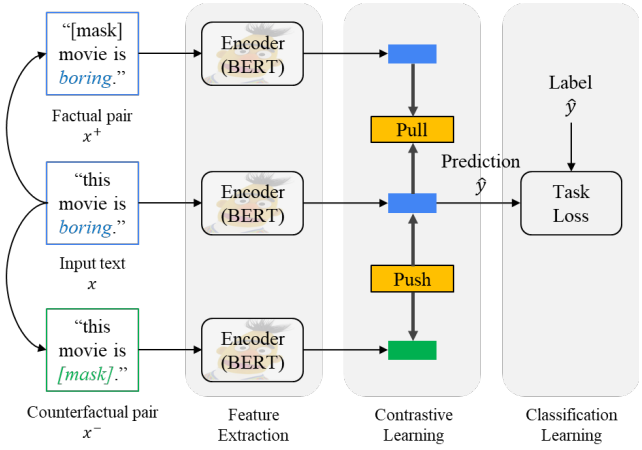


Figure 2: The overall framework of C^{2}L. With the causal triplets (x, x^+, x^-), we aim to map the representation of original text x close to factual pair x^+ but far apart from the counterfactual pair x^-.

The gradient-based score of token w is aggregated over all the training texts having the token w by:

$$s^{grad}(w) = \sum_{(x,y)\in\mathcal{D}} \frac{1}{n_{w,x}} \sum_{i\in\{1,...,T\}} \mathbb{I}(w_i = w) \cdot g_i, \quad (3)$$

where $\mathbb{I}$ is an indicator function, and $n_{w,x}$ is the number of word w in the input x.

Note that, as it is non-trivial to determine an absolute threshold for selecting the label-indicative tokens, we adopt an iterative process of validating the causality by testing the high-attributed tokens first. We terminate this process when we find a causal feature using the following validation step.

Step 2: Candidate Validation The second step is thus validating the candidate tokens (i.e., highly ranked according to $s^{grad}(w)$) based on its individualized treatment effect (ITE), which enables measuring how much the high-attributed tokens contribute to its label y. More specifically, one can test whether perturbing the high-attributed word leads to change its predicted label. For simplicity, here we denote the high-attributed word as w.

For this purpose, we leverage a pre-trained masked language model (LM), *i.e.*, BERT (Devlin et al. 2019), that is not fine-tuned to the dataset. Imagine that the LM provides unobserved counterfactual examples (or, words) in the train dataset, but possibly observed in the pre-training corpus. As the unobserved counterfactual examples are inherently unconstrained to dataset biases, there are more chances of being debiased from observed biases.

Our distinct intuition is that, if the masked text can be reconstructed into multiple examples and they are labeled as different classes, we can decide the masked term has a causal effect. As we make a collective decision on the distribution of multiple predictions, this decision is less sensitive to a few biased/noisy individual decisions. It thus provides more stable evidence to identify the masked word w as causal to the task label y.

Specifically, we feed the masked sentence $x_{w\rightarrow\texttt{[mask]}}$ into BERT, to generate counterfactual words $\hat{w}$ likely to occur at the position of w. However, as BERT is reportedly bad at finding the substitutes of the masked word, we adopt the dropout-based masking approach (Zhou et al. 2019), which partially masks the target word via dropout mechanism (Srivastava et al. 2014) to take the balanced consideration of the target word's semantics and contexts. Namely, we feed the original sentence x, where the token embedding of `[MASK]` token is added to the target word w as follows:

$$\mathbf{w}' = \lambda_d \mathbf{w} + (1 - \lambda_d)\mathbf{w}_{\texttt{[mask]}}, \quad (4)$$

where $\mathbf{w}_{\texttt{[mask]}}$ is the embedding of `[MASK]` token, and λ_d denotes a balancing coefficient for embedding dropout.

Then, the masked language model, *i.e.*, BERT, takes the partially-masked sentence as input, and samples top-k substitution candidates $\{\hat{w}_1, \hat{w}_2, ..., \hat{w}_k\}$ of target word w, from the likelihood distribution $P_{lm}(\hat{w}|x_{w\rightarrow\texttt{[mask]}})$. For the top-$k$ substitution candidates, we construct the k counterfactual sentences $\{x_{w\rightarrow\hat{w}_1}, x_{w\rightarrow\hat{w}_2}, ..., x_{w\rightarrow\hat{w}_k}\}$. We then collect the predicted labels of the sentences $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_k\}$ from the classifier f_ϕ. By testing whether the k labels are evenly distributed into the classes, we can decide the high-attributed token w as causal to its task label y. For example, if the generated $k = 4$ sentences are distributed in different classes, specifically two sentences at each class, we decide the token is causal[1]. If the token w cannot satisfy the condition, we move to the next mostly attributed token and repeat this second step until we find a causal feature[2]. We denote this method as multiple reconstruction (MR).

Contrastive Learning Objective

In this work, we leverage contrastive learning (Oord, Li, and Vinyals 2018; Hjelm et al. 2018) to better learn the causal structure of the classification task. For this purpose, we build the causal triplets (x, x^+, x^-) by utilizing the causal features obtained from above, as presented in Figure 2. First, the counterfactual pair x^- is built by masking out causal words, *i.e.*, replacing with a special token `[MASK]`, such that its label cannot be determined even with a similar syntactic structure. On the other hand, we mask one of the remaining words to generate a factual pair x^+ that is still recognized as the original label y, which helps to learn a model invariant to these features.

Formally, the contrastive objective aims to map the representation of x close to the representation of positive samples $\{x_j^+\}_{j=1}^J$, while the representation of negative samples $\{x_j^-\}_{j=1}^J$ far apart from x. To achieve this, we adopt the following margin-based ranking loss by (Zhang et al. 2020) for

[1]As explained here, we test $k = 4$ sentences for this purpose, which is empirically tuned. In our preliminary experiments, when we generate less than 4 sentences, this process suffers from incorrect predictions.

[2]In our experiments, augmentation increases BERT call to x15∼20 without back-propagation, translating to 5 additional hours in SST2 training, which is not significant as training is done offline.

model training:

$$\mathcal{L}_c(x;\theta) = \max(0, \Delta_m + \frac{1}{J}\sum_{j=1}^{J} s_\theta(x, x_j^+) - \frac{1}{J}\sum_{j=1}^{J} s_\theta(x, x_j^-)), \quad (5)$$

where J is the number of positive/negative pairs, Δ_m is a margin value which we set to 1 in this work, and $s_\theta(\cdot,\cdot)$ denotes the distance between the BERT representations.

For each causal triplets, minimizing the contrastive objective enables the model to learn the relationship between them and predict the right class from a more causal aspect. Finally, the parameters of the text classification network are trained to minimize the both loss terms together as follows:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda\mathcal{L}_c, \quad (6)$$

where λ is a balancing coefficient for the contrastive objective. Training the contrastive objective with our causal triplets (x, x^+, x^-), we name these models as **C**ausally **C**ontrastive **L**earning (C^{2}L) in the later experiments.

On Paired-input Task

Many important tasks such as Natural Language Inference (NLI) require to classify a pair of texts according to their logical relationship. Though the convention of BERT for such task is treating the input texts as one single segment, *i.e.*, "`[CLS]` text1 `[SEP]` text2 `[SEP]`", we observe that there can be interference between the texts during validation: For example, consider a masked hypothesis sentence "some men are `[MASK]` a sport", paired with a premise sentence "a soccer game with multiple males playing". If we feed the masked hypothesis together with the premise into BERT, the `[MASK]` token would be most likely replaced by the word "playing", which already appears in the premise sentence. The problem is that, it can deteriorate the quality of causality estimation, due to its low diversity (or, coverage) of the generated tokens.

To avoid such interference, we adopt separated-validation, which reconstructs each masked text independently by dropping the other text. Note that, such separation is only conducted when constructing k counterfactual sentences, while we collect the predictions on the concatenated texts.

One may ask similar interference can be observed across segments in long texts, *i.e.*, document-level input. We empirically compared with separated-validation to reduce interference, but differences were not statistically significant, despite computation overheads– We thus report separated-validations for long input only in Appendix.

Experiment Setup

Datasets

We first evaluate our method on the counterfactually-revised dataset (Kaushik, Hovy, and Lipton 2019), which was recently released to tackle the over-reliance problem of deep learning systems on *spurious* patterns in training data. To additionally validate the effectiveness of C^{2}L in cross-domain scenario, we use the following text classification datasets, which are widely used (Wang et al. 2018; Jain and Wallace 2019; Choi et al. 2020; Moon et al. 2020) and statistically diverse as well: Sentiment analysis experiments on IMDb (Maas et al. 2011), FineFood (McAuley and Leskovec 2013), and SST-2 (Socher et al. 2013) datasets. And, natural language inference experiments are conducted on MultiNLI (Williams, Nangia, and Bowman 2017) dataset.

Following the settings in (Kaushik, Hovy, and Lipton 2019; Moon et al. 2020), we use the official train and test splits if they exist, or we randomly divide the dataset with a 70:30 ratio, using them for train and test splits, respectively. To confirm convergence, we use 10% of the train set for validation purposes. The learning parameters were chosen by the best performance on the validation set. All the reported results are averaged over three trials.

Implementation Details

For experiments, we chose all the hyperparameters by the best performance on the validation set. For the BERT classifier, we train `bert-base-uncased` with a batch size of 16 for SST2, and 8 for IMDb/FOOD over 3 epochs, ensuring convergence. We used AdamW with a learning rate of $5e-5$ and the Linear scheduler with 50 warm-up steps.

For contrastive objective (Eq. 5), the balancing coefficient λ is tuned between [0.1, 1.0], and we observe giving larger λ is effective when fewer causal features are identified, such that setting the λ as 0.1, 0.7, and 1.0 performs well for SST-2, IMDb, and FineFood respectively. The number of positive/negative pairs J is set to 1 for the memory issue. The embedding dropout coefficient λ_d is tuned to 0.5.

Note that, for NLI tasks, we do not use the identified causal features (*i.e.*, $\lambda = 0$) for the "neutral" class samples, as the neutral class in itself means not belonging to any other classes, which is similar with our counterfactual pairs.

Comparison Methods

To closely observe how our proposed causal understanding contributes to the text classification task, we compare C^{2}L with the following baselines and ablations built on the standard BERT classifier.

SSMBA (Ng, Cho, and Ghassemi 2020): As a masking-based generative baseline, we implement SSMBA, a corrupt-and-reconstruct approach, that masks an arbitrary number of word positions and unmasks them using BERT. Following the original paper, we augment 5 samples for each sample with RoBERTa (Liu et al. 2019), and train the BERT-Base classifier on the augmented dataset with soft-label.

MASKER (Moon et al. 2020): Similar, but contrary to our approach, MASKER enforces BERT to make a prediction based solely on the surrounding contexts by masking out keywords. Following the training details written in the original paper, we reproduce the experiments with a validation set.

CL (pos/neg): As a basic implementation of contrastive learning, we collect positive (or, negative) contrastive pairs from a pool of the same (or, different) class samples, and give the contrastive objective (Eq. 5) with the pairs like C^{2}L. All the hyperparameters are the same with C^{2}L approaches.

Model	CF-IMDb		CF-NLI			
	Original	Revised	Original	RP	RH	RP & RH
BERT-Base	90.6±2.1	89.3±2.0	77.4±1.7	40.9±3.8	64.5±0.8	52.7±2.3
SSMBA	90.5±0.6	90.0±1.4	75.8±1.5	42.5±0.9	65.0±0.3	53.8±0.5
CL (pos/neg)	**91.3**±**0.4**	90.4±0.5	75.8±2.4	40.3±2.0	64.1±0.6	52.2±0.7
MCL (attn)	91.0±0.9	90.5±2.8	75.2±3.0	39.6±2.2	62.1±2.2	50.9±1.7
MCL (grad)	90.9±1.2	91.9±0.2	75.3±1.8	41.1±1.9	65.3±2.0	53.2±1.9
MCL (grad+HL)	89.0±0.3	90.9±0.7	76.7±1.6	41.5±2.0	64.6±1.7	53.0±1.8
MCL (grad+SL)	90.6±1.4	92.0±2.2	**78.3**±**1.1**	40.0±1.3	64.5±1.3	52.2±1.3
C^{2}L (-MR)	89.6±0.8	91.1±0.5	75.0±2.2	39.4±3.6	61.4±2.6	50.4±1.3
C^{2}L	**91.3**±**1.4**	**92.6**±**1.3**	76.2±1.7	**43.1**±**2.5**	**65.8**±**1.7**	**54.5**±**2.1**

Table 1: **Counterfactual Accuracy:** accuracy (%) on the counterfactually augmented IMDb and SNLI dataset (Kaushik, Hovy, and Lipton 2019). For CF-IMDb, all models are trained with the original 1.7k IMDb reviews, and evaluated on both original and counterfactually revised samples, which is the most difficult setting reported in the previous literature. For CF-NLI, we train all the models with the original 1.67k NLI samples, and report the following 3 types of counterfactual samples: revised premise (RP), revised hypothesis (RH), and combination of them (RP & RH).

MCL (grad): To confirm the effect of our candidate validation, we select the word w that has the highest candidate proposal score $s^{grad}(w)$ and mask the word. To distinguish it from the proposed causally contrastive learning, we call it masked contrastive learning (MCL).

MCL (attn): In MASKER (Moon et al. 2020), the attention values of the model are used for choosing the keywords, which are defined as the words with the highest attention values. For comparison with our gradient-based scoring, we leverage attention value as candidate proposal score without candidate validation step. Formally, let $\mathbf{a} = [a_1, ..., a_T]$ be attention values of the input embeddings, where a_i corresponds to the input word w_i. Then, the attention score of word w is aggregated over all the input texts having the word w by:

$$s^{attn}(w) = \sum_{(x,y)\in\mathcal{D}} \frac{1}{n_{w,x}} \sum_{i\in\{1,...,T\}} \mathbb{I}(w_i = w) \cdot \frac{a_i}{\|\mathbf{a}\|} \quad (7)$$

where $n_{w,x}$ is the number of word w in the input x, and $\|\cdot\|$ is l_2-norm. We mask the word w with the largest attention score $s^{attn}(w)$.

MCL (grad+HL): As a counterpart of our approach, we generate "single" counterfactual sentence for a word, and consider hard label-flip (HL) to decide the causality: When the label of the masked input is different from its original label, *i.e.*, $\arg\max y \neq \arg\max \hat{y}_{\hat{w}}$, we mask the word w. As discussed, the task model is biased to the given dataset, we categorize this as MCL.

MCL (grad+SL): Unlike MCL (grad+HL), erasing one may not necessarily flip the decision, but it can change the prediction significantly. Following that, we adopt Total Variance Distance (Jain and Wallace 2019) between two predictions: If the score $\text{TVD}(\hat{y}, \hat{y}_{\hat{w}}) = \frac{1}{2}\sum_{j=1}^{N} |\hat{y}^j - \hat{y}^j_{\hat{w}}|$ is larger than a threshold ϵ, we mask the word w, where the threshold is empirically tuned with the minimum value to flip a label in MCL (grad+HL).

C^{2}L (-MR): This is an ablation of our proposed C^{2}L regarding the multiple reconstruction method. While we "generate" multiple reconstructions using MLM, inspired by anchoring factual observation (Wang and Culotta 2020b), this "selects" from training samples, whose BERT similarity with the masked sample is over a threshold, empirically tuned to 0.95. When such a pair is found for w, we mask the word w.

Results and Discussion

We now proceed to empirically validate the effectiveness of C^{2}L. Based on the benchmark datasets, we address the following three key questions:

RQ1: Is C^{2}L robust for counterfactually-revised texts?
RQ2: Does C^{2}L adapt better to the new domain?
RQ3: Is C^{2}L robust for data-scarce cases?

RQ1: Robustness for Counterfactual Data

We first evaluate how C^{2}L contributes to the robustness of deep models on counterfactually revised datasets. The models are trained with the original datasets (1.7k IMDb reviews for sentiment analysis and 1.67k SNLI samples for natural language inference), and evaluated on both original and counterfactually revised samples. The results are presented in Table 1. In the table, we can observe the proposed approach, C^{2}L, outperforms all the baselines in the revised datasets, on which biases from the original training data cannot be relied. Specifically, C^{2}L improves the performance 3.3 points in revised IMDB and 1.8 points in revised NLI (RP & RH) from the BERT-Base. It demonstrates that, when the network better understands the causal correlations between input text and the task label, the network becomes more robust against spurious correlations as we claimed. It is noteworthy that

[3]Note that these results are different from those reported in the original paper, due to dataset settings.

Model	Sentiment						NLI	
	I→F	I→S	F→I	F→S	S→I	S→F	T→L	T→F
BERT-Base	87.9±2.2	86.9±0.2	82.8±1.8	74.4±1.5	88.1±0.8	81.3±3.2	80.8±0.7	79.4±0.1
SSMBA	88.6±0.1	87.1±0.7	83.7±0.7	74.7±1.0	88.0±0.2	80.5±0.3	80.6±0.4	79.7±0.4
MASKER* [3]	86.8±0.0	85.8±0.0	78.3±0.0	75.1±0.0	84.0±0.0	81.0±0.0	80.4±0.0	78.5±0.0
CL (pos/neg)	86.6±1.2	87.3±1.0	83.5±1.7	74.2±3.5	88.3±0.9	82.7±0.9	79.9±1.3	80.0±0.5
MCL (attn)	89.1±0.3	86.0±1.4	83.4±1.5	75.3±0.8	88.8±0.8	80.0±3.8	80.4±0.1	80.2±0.5
MCL (grad)	88.7±0.3	87.5±0.7	82.3±1.8	72.9±1.6	88.6±0.6	80.7±0.9	80.6±0.7	79.3±1.5
MCL (grad+HL)	88.5±0.8	86.9±1.0	83.8±0.8	75.0±0.8	89.2±0.7	82.4±4.1	81.0±0.7	80.1±0.3
MCL (grad+SL)	88.9±0.7	86.7±0.6	84.0±0.1	74.3±3.5	88.5±0.7	81.6±0.8	81.2±0.4	79.9±0.6
C^2L (-MR)	88.9±1.3	86.4±0.8	82.2±0.6	75.4±2.0	88.7±0.9	82.8±0.3	80.7±0.3	80.2±0.2
C^2L	**89.2**±0.5	**87.6**±0.5	**84.8**±0.9	**77.5**±0.2	**89.6**±0.4	**84.3**±1.3	**82.0**±0.7	**80.2**±0.6

Table 2: **Cross-Domain Accuracy:** accuracy (%) on the three sentiment analysis and multi-domain NLI datasets. * indicates that the results are reproduced by the original implementation. We denote each sentiment dataset as follows: IMDB (I), FineFood (F), and SST2 (S). For NLI, each domain is denoted as follows: Telephone (T), Letters (L), and FaceToFace (F). In-domain results are included in Appendix (Table 4).

ORIG: celebrity cameos do not automatically equal laughs.
GRAD: `[mask]` cameos do not automatically equal laughs.
OURS: celebrity cameos do `[mask]` automatically equal laughs.
ORIG: yet another sexual taboo into a really funny movie.
GRAD: yet another sexual `[mask]` into a really funny movie.
OURS: yet another sexual taboo into a really `[mask]` movie.
ORIG: has a rather unique approach to documentary.
GRAD: has a rather unique approach to `[mask]`.
OURS: has a rather `[mask]` approach to documentary.

Table 3: **Qualitative Analysis:** examples on the SST2 dataset. For comparison, we present an original input text (**ORIG**) with its masked pairs, where the word is selected by gradient method (**GRAD**) and our validation method (**OURS**) respectively.

C^2L adopts the self-supervision signals without requiring any additional human efforts (*e.g.*, revised training set).

Meanwhile, the results manifest that the baselines are less effective against the spurious correlations. For example, though MCL (grad+SL) achieves the best accuracy 78.3 in the original NLI test set, it fails to distinguish the causal correlations from spurious correlations, showing even lower performance than BERT-Base in the revised NLI test sets. This suggests that leveraging task models exploits the label-indicative features regardless of their spuriousness.

RQ2: Cross-Domain Generalization

The performance of neural networks can deteriorate under a domain shift between training and test data. Previous literature (Li et al. 2018; Moon et al. 2020) have shown that over-relying on the domain-specific keywords limits the generalization ability of networks, as the same keywords may not appear in another domain, for which we aim to remove such *spurious* features. We thus test whether the deep models can perform robustly from domain shifts.

Table 2 presents the classification accuracy for the cross-domain scenario, where each model is trained only on the source domain and evaluated on the target domain without further training. In the table, we can find C^2L is more robust, outperforming all the baselines in cross-domain settings. It demonstrates that contrasting causal triplets makes the network more robust against domain shifts.

Our study also confirms the following observations to our advantage: 1) the candidate validation step meaningfully contributes to performance gains, and 2) the validation is especially effective when multiple decisions are collected. More specifically, 1) MCL (grad+HL), MCL (grad+SL), C^2L (-MR) and C^2L show better performance than MCL (attn) or MCL (grad) in many cases. And 2) C^2L achieves the best performances in all the cross-domain settings, such as 84.8 at FineFood→IMDb and 84.3 at SST-2→FineFood with a significant performance gap, while in-domain accuracy still remains comparable[4]. This suggests that the proposed approach can effectively avoid spurious features, which cannot be generalized to new domains. In our extensive study, we additionally observe that identifying causal features makes the model better transfer to longer texts (Table 4 in Appendix).

For further analysis, we show qualitative examples in Table 3. We present the original text in SST-2 and its negative pairs masked by MCL (grad) and C^2L respectively. As discussed, we can observe that MCL (grad) gives high scores for the spurious words, such as the word "celebrity" or "taboo", where perturbing the word cannot change its label. Meanwhile, our approach identifies the causal words more accurately, such as "not" and "funny", where one can easily flip the label of the causally masked sentence by unmasking it with the words "always" and "boring".

[4]We report all the in-domain results in Appendix due to page limitation.

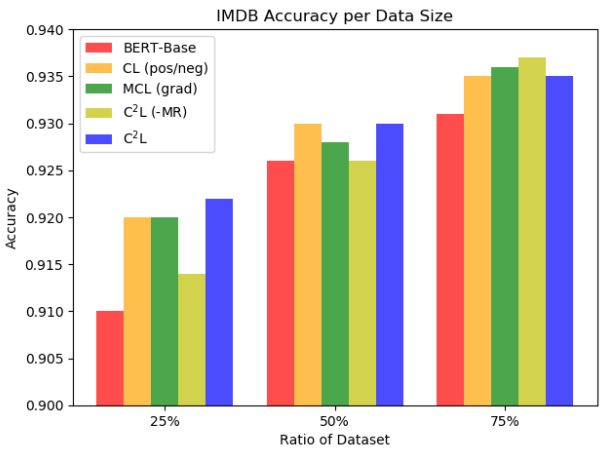


Figure 3: **Data Scarcity:** accuracy (%) on a varying number of training samples in the IMDb dataset.

RQ3: Data Scarcity

In many realistic applications, the training data is often too scarce. As a stress test for such data-scarce scenarios, we train each model with the IMDb dataset, but in different training set sizes (25%, 50%, and 75%) and evaluate in the official test set. The results are presented in Figure 3. From the results, we can find C^2L performs more robustly against data scarcity, outperforming the others when there are fewer training samples (25% and 50%). An interesting point is, the best performing model is MCL (grad) when there are enough data (75%). These trends indicate that MCL (grad) makes the best use of not only causal features but also spurious features, that can improve in-domain accuracy but cannot generalize to new data, as addressed in RQ1.

Among the results, it is noteworthy that C^2L (-MR) poorly works when there are not enough data, as it builds control groups of varying quality, which limits the quality of causality estimation in the data-scarce scenario. On the contrary, C^2L stably performs well regardless of the size of the given dataset, which demonstrates the advantages of leveraging language models.

Related work

Robust Text Classification

The goal of text classification is to assign labels to the given text. In spite of the recent advances for language understanding (Devlin et al. 2019; Liu et al. 2019), deep models are still challenged by spurious correlations, associating "free" with negative sentiment (Wang and Culotta 2020a), "gay" with toxicity (Wulczyn, Thain, and Dixon 2017), and "not" with contradiction (Gururangan et al. 2018). Against spurious correlations, recent work pursued additional human annotations, such as 1) human rationales (Zhang, Marshall, and Wallace 2016) and 2) counterfactually-augmented datasets (Kaushik, Hovy, and Lipton 2019; Khashabi, Khot, and Sabharwal 2020), for supervising neural attention (Zou et al. 2018; Bao et al. 2018; Choi et al. 2020), or model gradients (Liu and Avci 2019; Teney, Abbasnedjad, and Hengel 2020). However, due to annotation scarcity, the automatic annotation has been studied in the following directions: (Ng, Cho, and Ghassemi 2020; Wang and Culotta 2020b) generate the counterfactual sentences, (Garg and Ramakrishnan 2020) estimate token importance via counterfactual inference, and (Wang and Culotta 2020a; Klein and Nabi 2020) find a similar counterpart in the given dataset. Similar to our masking approach, (Moon et al. 2020; Liu, Chen, and Zhao) enforce the model to make a prediction based on the surrounding contexts by masking out some keywords (or, event mentions).

Our distinction: Our work shares the same goal of automating manual annotations, with the distinction of higher tolerance to noises and biases, by making an unbiased causal feature selection and avoiding noisy per-instance prediction, replaced by a group decision, represented by our proposed masked pairs.

Contrastive Learning

A recent advance for representation learning is contrastive objective (Oord, Li, and Vinyals 2018; Hjelm et al. 2018), which exploits the idea of learning by comparison to capture the subtle features of data. By recent studies, some crucial considerations for better contrastive learning are revealed, such as heavy data augmentation (Gontijo-Lopes et al. 2020; Tian et al. 2020), large sets of negatives (Chen et al. 2020b), and difficulty of negative pairs (Kalantidis et al. 2020). In this work, we focus on leveraging counterfactual samples, *i.e.*, minimally dissimilar pairs (Kaushik, Hovy, and Lipton 2019). Among the existing counterfactual contrastive learning approaches (Wang et al. 2020; Zhang et al. 2020), (Liang et al. 2020; Chen et al. 2020a) is closest to our work, by masking salient features, measured by task model, such as gradient-based explanation, *i.e.*, Grad-CAM (Selvaraju et al. 2017).

Our distinction: Unlike existing efforts using causal features for contrastive learning, to the best of our knowledge, our work is the first to study the debiasing of task models, without increasing annotation overheads on the human side.

Conclusion and Future Work

We studied the problem of counterfactual augmentation, with the distinction of collectively considering the counterfactual sentences, less biased by task models. Our empirical results validated that, with our causally contrastive objective, the robustness of text classification models can be significantly improved, showing the effectiveness of our approaches. We hope future research to explore generalization to other tasks (*e.g.*, question answering or conversation modeling) and languages (*e.g.*, low-resource languages).

Acknowledgments

This work was supported by SNU-NAVER Hyperscale AI Center and IITP granted by the Korea government (MSIT) [NO.2021-0-0268, AI Innovation Hub (SNU)] and SNU AI Graduate School Program [No.2021-0-01343]. Hwang is a corresponding author.

References

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving Machine Attention from Human Rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1903–1913.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020a. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10800–10809.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Choi, S.; Park, H.; Yeo, J.; and Hwang, S.-w. 2020. Less Is More: Attention Supervision with Counterfactuals for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6695–6704.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Garg, S.; and Ramakrishnan, G. 2020. BAE: BERT-based Adversarial Examples for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6174–6181.

Gontijo-Lopes, R.; Smullin, S. J.; Cubuk, E. D.; and Dyer, E. 2020. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*.

Grimsley, C.; Mayfield, E.; and Bursten, J. R. 2020. Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 1780–1790.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 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)*, 107–112.

Han, H.; Choi, S.; Jeong, M.; Park, J.-w.; and Hwang, S.-w. 2021. Counterfactual Generative Smoothing for Imbalanced Natural Language Classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3058–3062.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–3556.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive learning. *arXiv preprint arXiv:2010.01028*.

Kaushik, D.; Hovy, E.; and Lipton, Z. 2019. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

Khashabi, D.; Khot, T.; and Sabharwal, A. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 163–170.

Klein, T.; and Nabi, M. 2020. Contrastive Self-Supervised Learning for Commonsense Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7517–7523.

Li, Z.; Wei, Y.; Zhang, Y.; and Yang, Q. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liang, Z.; Jiang, W.; Hu, H.; and Zhu, J. 2020. Learning to Contrast the Counterfactual Samples for Robust Visual Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3285–3292.

Liu, F.; and Avci, B. 2019. Incorporating Priors with Feature Attribution on Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6274–6283.

Liu, J.; Chen, Y.; and Zhao, J. ???? Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *ACL*.

McAuley, J. J.; and Leskovec, J. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, 897–908.

Moon, S. J.; Mo, S.; Lee, K.; Lee, J.; and Shin, J. 2020. MASKER: Masked Keyword Regularization for Reliable Text Classification. *arXiv preprint arXiv:2012.09392*.

Ng, N.; Cho, K.; and Ghassemi, M. 2020. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *arXiv preprint arXiv:2009.10195*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085. PMLR.

Shpitser, I.; and Pearl, J. 2006. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 437–444.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Teney, D.; Abbasnedjad, E.; and Hengel, A. v. d. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355.

Wang, D.; Yang, Y.; Tao, C.; Kong, F.; Henao, R.; and Carin, L. 2020. Proactive Pseudo-Intervention: Causally Informed Contrastive Learning For Interpretable Vision Models. *arXiv preprint arXiv:2012.03369*.

Wang, Z.; and Culotta, A. 2020a. Identifying spurious correlations for robust text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 3431–3440.

Wang, Z.; and Culotta, A. 2020b. Robustness to Spurious Correlations in Text Classification via Automatically Generated Counterfactuals. *arXiv preprint arXiv:2012.10040*.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, 1391–1399.

Zhang, Y.; Marshall, I.; and Wallace, B. C. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 795–804.

Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. *Advances in Neural Information Processing Systems*, 33: 18123–18134.

Zhou, W.; Ge, T.; Xu, K.; Wei, F.; and Zhou, M. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3368–3373.

Zou, Y.; Gui, T.; Zhang, Q.; and Huang, X.-J. 2018. A lexicon-based supervised attention model for neural sentiment analysis. In *Proceedings of the 27th international conference on computational linguistics*, 868–877.