Less Learn Shortcut: Analyzing and Mitigating Learning of Spurious Feature-Label Correlation

Yanrui Du1 *, Jing Yan2, Yan Chen2, Jing Liu2, Sendong Zhao1†, Qiaoqiao She2, Hua Wu2, Haifeng Wang2, Bing Qin1

1Harbin Institute of Technology, Harbin, China
2Baidu Inc., Beijing, China
{ yrdu, sdzhao, bqin} @ir.hit.edu.cn
{yanjing09, chenyan22, liujing46, sheqiaoqiao, wu_hua, wanghaifeng} @baidu.com

Abstract
Recent research has revealed that deep neural networks often take dataset biases as a shortcut to make decisions rather than understand tasks, leading to failures in real-world applications. In this study, we focus on the spurious correlation between word features and labels that models learn from the biased data distribution of training data. In particular, we define the word highly co-occurring with a specific label as biased word, and the example containing biased word as biased example. Our analysis shows that biased examples are easier for models to learn, while at the time of prediction, biased words make a significantly higher contribution to the models’ predictions, and models tend to assign predicted labels over-relying on the spurious correlation between words and labels. To mitigate models’ over-reliance on the shortcut (i.e. spurious correlation), we propose a training strategy Less-Learn-Shortcut (LLS): our strategy quantifies the biased degree of the biased examples and down-weights them accordingly. Experimental results on Question Matching, Natural Language Inference and Sentiment Analysis tasks show that LLS is a task-agnostic strategy and can improve the model performance on adversarial data while maintaining good performance on in-domain data.

1 Introduction
Pre-trained language models, e.g. BERT [Devlin et al., 2018], ERNIE [Sun et al., 2019] and RoBERTa [Liu et al., 2019b], have achieved great success on many NLP tasks. However, recent studies highlighted that pre-trained models tend to take dataset biases as a shortcut, rather than truly understand tasks [Schuster et al., 2019; Niven and Kao, 2019]. Models’ over-reliance on the shortcut results in their poor generalization ability and low robustness [Geirhos et al., 2020].

The phenomenon of shortcut learning has been widely studied in various NLP tasks. Many previous studies examine this phenomenon by constructing artificial adversarial examples, and employ adversarial data augmentation to enhance model robustness [Jia and Liang, 2017; Alzantot et al., 2018; Ren et al., 2019; Jin et al., 2020]. These studies reported high success rates on artificial adversarial examples, but it is uncertain if the models will perform well on real-world data distributions [Morris et al., 2020; Bender and Koller, 2020]. Additionally, recent work [Balkir et al., 2022] indicated that few studies have applied explainable methods to understand or investigate the impact of shortcut learning.

Previous works point out that shortcuts can be traced back to dataset biases [Lai et al., 2021; Gururangan et al., 2018; Kavumba et al., 2021; Du et al., 2021; McCoy et al., 2019; Liu et al., 2019a; Kavumba et al., 2021; Goyal et al., 2017; Ye and Kovashka, 2021; Dawkins, 2021]. For example, if “not” happens to be contradiction for most of the training data in Natural Language Inference (NLI) tasks, detecting “not” becomes a successful strategy for models’ prediction, thus leading to an unexpected performance on a shift distribution [Gururangan et al., 2018]. However, most studies are limited to analyzing task-specific shortcuts, which are prohibitive to be transferred to other tasks.

In this work, we analyze the correlations between simple features (e.g., words) and labels, which can be originated from the biased data distribution of any NLP task, to quantitatively investigate the shortcut learning behavior of NLP models. Existing work has argued that, for any NLP task, no single feature on its own should contain information about the labels, and any correlation between simple features and labels is spurious [Gardner et al., 2021]. Based on the above analysis, we propose a task-agnostic training strategy Less-Learn-Shortcut (LLS), which mitigates the shortcut behavior of models, thereby improving their performance on adversarial data.

To examine the spurious feature-label correlation, we first introduce two definitions: biased word, which is the word highly co-occurring with a specific label in a dataset, and biased example, which is the example containing at least one biased word. Then we quantitatively analyze the spurious feature-label correlations on the Question Matching (QM) task. Based on our analysis, we propose our training strategy LLS, with which biased training examples are down-weighted according to their biased degrees, and the mod-
els’ over-reliance on the biased words is penalized during fine-tuning. We conduct extensive experiments on QM, NLI and Sentiment Analysis (SA) tasks to evaluate our training strategy and compare it to other task-agnostic strategies such as Rewbias [Utama et al., 2020; Clark et al., 2019] and Forg. [Yaghoobzadeh et al., 2021]. Our experimental results demonstrate that LLS can improve the model performance on adversarial data while maintaining good performance on in-domain data, and can be easily transferred to different NLP tasks. Additionally, we explore the scenarios in which the above strategies are applicable.

In general, we have the following major findings and contributions:

• We reveal that biased examples (as defined in Sec. 2.2) are easier to be learned than other examples, and with an explainable method LIME [Ribeiro et al., 2016], we find that biased words make significantly higher contributions to models’ predictions than random words (see Sec. 3.1).

• We find that biased words will affect models’ predictions, and that models tend to assign labels highly correlated to the biased words (see Sec. 3.2).

• To mitigate the models’ over-reliance on the spurious correlation, we propose a training strategy Less-Learn-Shortcut (LLS). Experimental results show that LLS can improve the models’ performance on adversarial data while maintaining good performance on in-domain data. Furthermore, we compare LLS to existing strategies and reveal their respective applicable scenarios. (see Sec. 4).

2 Preliminary

In this section, we first introduce the QM datasets on which we analyze the spurious feature-label correlation, then we give the definitions of biased word and biased example. At last, we provide the settings of our experiments.

2.1 Datasets

We conduct our analysis on three datasets, LCQMC, DuQM and OPPO\(^1\), all of which are about QM task and collected from real-world applications. LCQMC [Liu et al., 2018] is a large-scale Chinese question matching corpus proposed by Harbin Institute of Technology in the general domain BaiduZhidao. DuQM [Zhu et al., 2021] is a fine-grained controlled adversarial dataset aimed to evaluate the robustness of QM models and generated based on the queries collected from Baidu Search Engine\(^2\). OPPO is collected from OPPO XiaoBu Dialogue application and can be downloaded on CCF Big Data & Computing Intelligence Contest. The data statistics are provided in Tab. 8 (in App. A).

2.2 Definitions

Here we provide the definitions we will use in this work. If we denote \(W\) as all words in the dataset, the set of examples containing a specific word \(w_i\) can be formalized as \(S(w_i)\), and the frequency of \(w_i\) can be formalized as \(f_{w_i}\). We define biased degree as \(d_{w_i}^{c_m}\) to measure the degree of word \(w_i\) co-occurring with category \(c_m\) (for QM task, \(c_m \in (0, 1)\)) and it can be denoted as

\[
d_{w_i}^{c_m} = \frac{|S(w_i, c_m)|}{|S(w_i)|} = \frac{|S(w_i, c_m)|}{f_{w_i}}
\]

where \(|S(w_i, c_m)|\) represents the number of examples with \(w_i\) and labeled with \(c_m\).

Biased word. A word highly correlated with a specific label in a dataset.\(^3\) To better discuss it, we define biased word as the word \(w_i\) with \(f_{w_i} \geq 3\) and \(d_{w_i}^{c_m} \geq 0.8\) for QM task in Sec. 2 and 3. It is worth mentioning that the biased words we analyze in this work are originated from the training set.

We further define biased word\(_0\) and biased word\(_1\) as the words highly correlated to category 0 and 1. As shown in Tab. 9 (in App. C), “便宜” (“handy”) occurs in 35 examples, 33 of which are with category 1, hence it is a biased word\(_1\). Tab. 1 shows that 27.24% (15864/58230) of words are biased words, and there are more biased word\(_0\) than biased word\(_1\) in LCQMC\(_{train}\).

Biased example. An example containing at least one biased word. As shown in Tab. 2, 41.15% of examples in LCQMC\(_{train}\) are biased examples, which are 25.97%, 32.25% and 24.98% in LCQMC\(_{test}\), DuQM and OPPO respectively. Since the biased words occur in almost half of the examples in LCQMC\(_{train}\), it is meaningful to study their effects on models. The examples without biased words are defined as unbiased example.

2.3 Experimental Setup

Models. We conduct our experiments on three popular publicly available pre-trained models, BERT-base\(^4\), ERNIE\(_{1.0}\)\(^5\) and RoBERTa-large\(^6\).

\(^1\)The datasets can be downloaded on https://luge.ai.
\(^2\)http://www.baidu.com.
\(^3\)Word is the smallest independent lexical item with its own objective or practical meaning. We use Lexical Analysis of Chinese [Jiao et al., 2018] (https://github.com/baidu/lac) for word segmentation in this work.
\(^4\)https://github.com/google-research/bert.
\(^5\)https://github.com/PaddlePaddle/ERNIE.
\(^6\)https://github.com/yumcui/Chinese-BERT-wwm.
Metrics. As most of the classification tasks, we use accuracy to evaluate the performance of models.

Training details. We use the integrated interface BertForSequenceClassification7 from huggingface for our experiment and use different learning rates for different pre-trained models. Specifically, for RoBERTa\textsubscript{large}, the learning rate is 5e-6. For BERT\textsubscript{base} and ERNIE1.0, the learning rate is 2e-5. The proportion of weight decay is 0.01 and the batch size is 64. We train two epochs for BERT\textsubscript{base} and ERNIE1.0, and train three epochs for RoBERTa\textsubscript{large}. Every 500 steps, we check the performance of models on LCQMC\textsubscript{dev} and choose the checkpoint with the highest accuracy as our main model, and report average results with three different seeds on LCQMC\textsubscript{test}, DuQM and OPPO.

3 Effect of Feature-Label Correlation

The dataset statistics in Sec. 2 show that 41.15\% of examples in LCQMC\textsubscript{train} contain biased words. It is a reasonable assumption that the spurious feature-label correlations would affect the models’ behavior and performance. To validate our assumption: 1) we conduct a behavior analysis of the model’s learning and deciding (See Sec. 3.1); 2) we discuss how the feature-label correlation affects the models’ performance by probing the relationship between the biased word and the predicted label (See Sec. 3.2). In Sec. 3.3 we discuss another type of shortcut word-overlap and argue that different shortcuts may interact together.

3.1 Feature-Label Correlation and Models’ Behavior

Models’ learning. To observe the models’ behavior during training, we separate LCQMC\textsubscript{train} into two subsets, biased examples and unbiased examples, and reorganize the train examples in 3 orders:

• bias-first: firstly biased examples, then unbiased examples;
• bias-last: firstly unbiased examples, then biased examples;
• random order: shuffle the examples randomly.

We finetune three models (BERT, ERNIE and RoBERTa) in above three orders and plot the training loss curves in Fig. 1.

Figure 1: Training loss curves of RoBERTa on LCQMC\textsubscript{train}, in which ● represents the time of finishing learning biased examples, and ▲ represents the time of finishing learning unbiased examples.

Models’ deciding. In this part, we provide a quantitative analysis of the spurious feature-label correlation’s impact on models’ deciding. If it is easier for a model to learn, will  

Figure 2: Probability of biased words and random words with the 1st, 2nd, 3rd, 4th contribution on three test sets. Bias-W and Rand-W represent biased words and random words respectively.

\( (a) \) Results on LCQMC\textsubscript{test}.
\( (b) \) Results on DuQM.
\( (c) \) Results on OPPO.
However, there are few quantitative analyses to discuss how the shortcut affects the models’ predictions exactly. In this section, we can conclude that the biased word makes significantly higher contributions than random words, which implies that models tend to pay more attention to biased words when deciding. With the analysis in this section, we can conclude that the biased word is a shortcut for the models and will affect the models’ behavior. It is therefore substantial to further analyze how it affects the models.

3.2 Feature-Label Correlation and Models’ Prediction

Existing works focus only on verifying the existence of shortcut [Bolukbasi et al., 2016; May et al., 2019; Ravfogel et al., 2020; Webster et al., 2020; Kaneko and Bollegala, 2021]. However, there are few quantitative analyses to discuss how the shortcut affects the models’ predictions exactly. In this part, we will focus on probing the relationship between the biased word and predicted label to analyze how the spurious feature-label correlations impact models. As the biased words are highly correlated to a specific label, it is a reasonable guess that the models tend to assign predicted labels highly correlated to the biased words.

Although the biased words tend to contribute more (discussed in Sec. 3.1), not all biased words make great contribution during predicting. To better analyze the impact of biased word on predicted label, we focus on the testing examples where biased word contributes the most, in which the effects of biased word would be more significant. For convenience, we define the examples in which the biased word makes the greatest contribution as focus-biased examples, and we present the statistics of biased examples and focus-biased examples in Tab. 10 and Tab. 11 (see App. D). To measure the tendency of models’ prediction, we define $T_{cm}$ as the tendency of model to predict of category $c_m$:

$$T_{cm} = \frac{|S_{pred}(c_m)|/|D|}{|S_{true}(c_m)|/|D|} = \frac{|S_{pred}(c_m)|}{|S_{true}(c_m)|} \quad c_m \in (0, 1)$$

where $|D|$ represents the number of observed examples, $|S_{true}(c_m)|$ and $|S_{pred}(c_m)|$ represent the number of examples with true label $c_m$ and predicted label as $c_m$, respectively. Specially, we observe the tendency of models’ prediction on “normal” biased examples and focus-biased examples, and denote them as $T_0$ and $T_{focus}$. The results are shown in Fig. 3. Fig. 3(a) to Fig. 3(c) show the influence of biased words on three test sets. On DuQM (Fig. 3(b)), it is obvious that $T_{focus}$ is higher than $T_0$ by averaged 7% with all three models, which implies that when biased word contributes the most, models have a high tendency to predict 0. The same result is shown on LCQMC$_{test}$ (Fig. 3(a)). However,
on OPPO (Fig. 3(c)), $T_0$ is slightly higher ($0.02\sim0.05$) than $T_1^{focus}$. We suppose that it is affected by the co-influencing of another shortcut and we provide an extensive experiment to discuss it in Sec. 3.3. Fig. 3(d) to Fig. 3(f) show the influence of biased word. As shown in Fig. 3(f), models tend to predict 1 when they concentrate on biased word on OPPO, that $T_1^{focus}$ is higher than $T_1$ by averaged $26\%$ with all three models. The comparison results on DuQM(Fig.3(e)) show the same tendency for all three models, that $T_1^{focus}$ is higher than $T_1$ by averaged $6\%$. On LCQMC(Fig.3(d)), $T_1^{focus}$ is almost close to $T_1$ with all three models.

Overall, we observe that when models pay more attention to biased words, they tend to assign labels over-relying on the biased words. Moreover, to explore why the tendency to 0 is not obvious on OPPO (Fig. 3(c)), we provide a further discussion about the influence of another shortcut word-overlap.

### 3.3 Word-Overlap: Another Shortcut for QM Models

In real-world scenarios, different shortcuts may interact together to affect the final prediction. Word overlap shortcut has been widely discussed in many MRC and NLI works [McCoy et al., 2019; Lai et al., 2021; Kaushik and Lipton, 2018]. For QM task, the models tend to predict 0 if a sentence pair has low word overlap, i.e., there are few common words between them, and vice versa. As the result of OPPO shown in Tab. 3, even if models focus on biased word, the tendency to 0 is not significant. We attribute the phenomenon to the word-overlap shortcut in the QM task. To eliminate the influence of word-overlap, we design an experiment on the examples in which the question pairs have high word-overlap. We use LevenshteinEdit distance to measure the overlapping degree. We report the models’ prediction tendency with short edit distance in Tab. 3. The results reflect that models have a higher tendency to predict 0 on focus-biased examples than “normal” biased examples, which implies that models tend to predict 0 if we try to eliminate the word-overlap shortcut.

Specifically, compared with “normal” biased examples, the average $T_0^{focus}$ of three models with edit distance 1 increases by 0.086, which is 0.040, 0.028, 0.027 and 0.022 for edit distance of 2, 3, 4 and 5.

Generally, we can deduce that models tend to assign labels relying on the feature-label correlation trick. By eliminating the influence of word-overlap, the models’ prediction tendency towards 0 becomes significant on OPPO. Besides the spurious correlations we study in this work, NLP models are also affected by many other shortcuts.

### 4 Less-Learn-Shortcut: A Training Strategy to Mitigate Models’ Over-Reliance on Feature-Label Correlation

In Sec. 3 we observe that the spurious feature-label correlation will affect models’ learning and deciding. To mitigate the models’ shortcut learning behavior, we propose a training strategy Less-Learn-Shortcut (LLS), with which all the biased training examples are penalized according to their biased degrees (in Sec. 4.1) during fine-tuning. Most of the existing strategies to mitigate shortcut learning include data augmentation [Jin et al., 2020; Alzantot et al., 2018] and adversarial training [Stacey et al., 2020], which are task-relevant. Our proposed method LLS is task-agnostic and can be easily transferred to different NLP tasks.

#### 4.1 Reweight Biased Examples

To mitigate the models’ over-reliance on the feature-label correlations, a straightforward idea is to down-weight the biased examples, so that the models are prevented from over-fitting the spurious correlations. In this section, we will introduce how we reweight the biased examples.

**Quantify the impact of correlation.** In Sec. 2, we have defined biased degree $d_{w_i}^{m}$ to measure the correlation between the word $w_i$ and the label $c_m$, which can quantify the impact of the correlation. The maximum biased degree of a word among all categories is denoted as $b_0^{w_i}$ (C represents all categories).

$$b_0^{w_i} = \max d_{w_i}^{m}, c_m \in C \quad (3)$$

Furthermore, some existing works show that the word frequency in the training data also influences the models’ prediction [Gu et al., 2020; Cui et al., 2016; Ott et al., 2018]. Considering both biased degree and word frequency, we formulate the impact of a biased word as

$$b_{w_i} = \max d_{w_i}^{m} + \alpha f_{w_i}, c_m \in C \quad (4)$$

where $f_{w_i}$ represents the frequency of words $w_i$ occurring in the training dataset, and $\alpha$ is a trade-off factor. Then the impact of a biased example can be formulated as the average impact of all biased words it contains:

$$b_{e} = \frac{1}{n} \sum_{i=1}^{n} b_{w_i} \quad (5)$$

| Model   | Dist. $\leq 1$ $T_0$ | Dist. $\leq 2$ $T_0^{focus}$ | Dist. $\leq 3$ $T_0^{focus}$ | Dist. $\leq 4$ $T_0^{focus}$ | Dist. $\leq 5$ $T_0^{focus}$ |
|---------|------------------|------------------|------------------|------------------|------------------|
| BERT    | 0.739            | 0.833            | 0.765            | 0.821            | 0.800            | 0.847            | 0.866            | 0.900            | 0.910            | 0.933            |
| ERNIE   | 0.761            | 0.857            | 0.779            | 0.876            | 0.841            | 0.847            | 0.894            | 0.905            | 0.946            | 0.970            |
| RoBERTa | 0.870            | 0.938            | 0.875            | 0.932            | 0.905            | 0.935            | 0.950            | 0.978            | 0.984            | 1.004            |
| $\overline{x}$    | 0.086            | 0.040            | 0.028            | 0.027            | 0.027            | 0.022            |
Detailed information about the experimental settings can be found in App.E. We present the average results of three different seeds and our performance improvements are statistically significant with a p-value of paired t-test less than 0.05.

Baseline. In addition to select Finetune as our baseline, we re-implement Rew\textsubscript{bias} strategies. The core idea of Rew\textsubscript{bias} is similar to our LLS strategy, which is down-weighting biased examples. Rew\textsubscript{bias} needs to additionally train a bias-only model to score biased examples. Forg. strategy uses examples forgotten by the model during training to do secondary training. The above two strategies are both task-agnostic and can be applied to any NLP task.

**QM task.** As shown in Tab. 4, for BERT and RoBERTa, our LLS strategy performs best on both the in-domain LCQMC\textsubscript{test} (87.86% and 88.46%) and adversarial DuQM (69.20% and 74.18%). For ERNIE, our LLS strategy performs best on in-domain LCQMC\textsubscript{test} (88.16%) and performs close to best on the adversarial DuQM (71.65%). Furthermore, we observe that although Rew\textsubscript{bias} and Forg. strategies improve the model performance on the adversarial DuQM, they only remain performance on LCQMC\textsubscript{test} and OPPO.
By contrast, our LLS strategy can improve the model performance on all three test sets.

To better investigate the contributions of different components of LLS, we compare LLS with two ablations: LLS$f$ only employs the biased degree to measure the impact of correlation, and does not consider the impact of word-overlap (Eq. 3, 5, and 7); LLS$s+f$ considers both biased degree and word frequency to measure the correlation, but also does not consider the impact of word-overlap (Eq. 4, 5, and 7). As shown in Tab. 4, LLS generally performs the best, indicating that considering word frequencies and excluding word-overlap has a positive effect.

**NLI task.** NLI task aims to determine the relationship between two sentences, whether a premise sentence entails a hypothesis sentence. It is normally formulated as a multi-class classification problem. In our experiments, we try two NLI datasets as the training sets, SNLI [Bowman et al., 2015] and MNLI [Williams et al., 2017]. Tab. 5 and 6 give the statistics of SNLI$_{train}$ and MNLI$_{train}$. Although only 2.96% of words in SNLI$_{train}$ are biased words, they occur in 4.84% of examples. Compared to SNLI$_{train}$, MNLI$_{train}$ is relatively unbiased and contains only 202 biased words (0.22%) and 993 biased examples (0.25%).

We first conduct our experiments on SNLI$_{train}$. We train models on SNLI$_{train}$ and evaluate them on the in-domain SNLI$_{test}$ and the adversarial HANS$_{test}$. SNLI is a dataset with three classes: entailment, neutral, and contradiction. HANS is a two-class dataset, entailment and non-entailment. As done in previous work [McCoy et al., 2019], to evaluate models on HANS$_{test}$, we convert neutral or contradiction labels to non-entailment. The experimental results are shown in Tab. 4. For BERT and ERNIE, Forg. strategy improves the model performance more significantly on the adversarial HANS$_{test}$. We present the statistics of forgotten examples (see App. F), and observe that for the large-scale SNLI$_{train}$, small models such as BERT and ERNIE are more likely to forget examples. Therefore, secondary training with forgotten examples can better help small models increase their robustness. In contrast, for the large RoBERTA model, Forg. strategy yields little and our LLS strategy performs better. Furthermore, compared to Finetune and Rew$_{bias}$ strategies, for all three models, our LLS strategy obtains a more significant benefit on the adversarial HANS$_{test}$ while maintaining good performance on the in-domain SNLI$_{test}$.

The results on MNLI$_{train}$ are shown in Tab. 7. Due to the fact that the MNLI$_{train}$ contains fewer biased examples, the effect of LLS is not significant. This suggests that LLS strategy is more effective for the training set with biased data distribution, helping models learn the spurious correlation less.

### Table 6: The statistics of biased examples in SNLI$_{train}$, MNLI$_{train}$ and Chnsenticorp$_{train}$. B-exp represents biased example.

| Dataset       | # Examples | # B-exp | % B-exp |
|---------------|------------|---------|---------|
| SNLI$_{train}$| 549,367    | 26,590  | 4.84%   |
| MNLI$_{train}$| 392,702    | 993     | 0.25%   |
| Chnsenticorp$_{train}$ | 9,600    | 9,151   | 95.11%  |

### Table 7: Performance (accuracy%) of BERT on MNLI$_{train}$.

| Model | Strategy | MNLI$_{test}$ | HANS$_{test}$ |
|-------|----------|--------------|---------------|
| BERT  | Finetune | 84.22        | 52.01         |
|       | LLS      | 84.36        | 52.40         |
|       | LLS$_{d}$ | 84.31        | 51.99         |
|       | LLS$_{d+f}$ | 84.49        | 52.24         |

**SA task.** SA task aims to determine whether a sentence has a positive or negative sentiment. In our experiment, we train models on Chnsenticorp$_{train}$ and evaluate them on the in-domain Chnsenticorp$_{test}$ and the adversarial SENTI$_{robust}$ [Wang et al., 2021]. As shown in Tab. 5 and Tab. 6, 14.05% of words in Chnsenticorp$_{train}$ are biased words, and they appear in 95.11% of the examples. Unlike NLI and QM tasks, SA task is a single-sentence classification task that is not affected by the word-overlap shortcut, thus we only report the results of LLS$_{d}$ and LLS$_{d+f}$ in Tab. 4. Compared to Rew$_{bias}$ and Forg. strategies, our LLS$_{d}$ and LLS$_{d+f}$ strategies obtain a more significant benefit on the adversarial SENTI$_{robust}$ and perform better on the in-domain Chnsenticorp$_{test}$. Additionally, it is worth noting that for the small-scale Chnsenticorp$_{train}$, models will not forget too many samples (see App. F) and Forg. strategy yields little.

In summary, our proposed LLS strategy can significantly improve the model performance on adversarial data while maintaining good performance on in-domain data. Our experiments show that existing strategies struggle to stably improve performance on in-domain data, making further research necessary. Furthermore, we reveal scenarios in which these strategies are applicable. Compared to the Rew$_{bias}$ strategy, LLS strategy demonstrates greater advantages on various tasks. However, LLS strategy is not applicable for the relatively unbiased dataset, such as MNLI$_{train}$. On the other hand, Forg. strategy shows its own advantages on SNLI$_{train}$. Specifically, when training a small model on a large-scale dataset, Forg. strategy is a good option to consider.

### 5 Conclusion

In this paper, we explore models’ shortcut learning behavior of spurious correlations between features and labels, and propose a training strategy LLS to mitigate the over-reliance of NLP models on the shortcut. Specifically, we observe that the models are prone to learn spurious correlations, and the biased words make significantly higher contributions to models’ predictions than random words. Moreover, we observe that the models tend to be misled by biased words to assign labels. To mitigate the over-reliance on biases, we propose a training strategy LLS to penalize the shortcut learning behavior of models. Experimental results show that LLS can improve the model performance on adversarial data while keeping good performance on in-domain data, and it is task-agnostic, which can be easily transferred to other tasks. In future research, we will explore how to better measure and formalize the shortcuts in the training data and generalize them as a class of problems.
**A Data Statistics**

Data statistics are presented in Tab. 8.

| Dataset | Word cnt. | Category | Total |
|---------|-----------|----------|-------|
|         | q1        | q2       | #0    | #1    | Total |
| L_{train} | 6.04      | 6.36     | 12.40 | 100,192 | 138,574 | 238,766 |
| L_{test}  | 5.51      | 5.61     | 11.12 | 6,250   | 6,250   | 12,500   |
| DuQM      | 4.66      | 4.80     | 9.46  | 7,318   | 2,803   | 10,121   |
| OPPO      | 4.82      | 4.71     | 9.53  | 7,160   | 2,840   | 10,000   |

Table 8: Data statistics. L_{train} denotes LCQMC training set, and L_{test} denotes LCQMC test set.

**B Training Loss**

![Training loss curve of BERT](a.png)

![Training loss curve of ERNIE](b.png)

Figure 4: Training loss curves of BERT and ERNIE on LCQMC_{train}, in which ● represents finishing learning biased examples, and ▲ represents finishing learning unbiased examples.

**C Examples of Bias-word**

Examples of bias-word_0 and bias-word_1 are given in Tab. 9.

**D Statistics of Bias-Example and Focus-Bias Examples**

The statistics of "normal" bias-examples and focus-bias examples are given in Tab. 10 and Tab. 11.

**E Experimental Settings**

Experimental settings are introduced Tab. 12.

**F Statistics of Forgotten Examples**

In Tab. 13, we record the number of examples forgotten by models on different tasks.
Acknowledgements

This work was supported by the National Key R&D Program of China [2021ZD0113302]; and the National Natural Science Foundation of China [62206079]; and Heilongjiang Provincial Natural Science Foundation of China [YQ2022F006].

References

[Alzantot et al., 2018] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhong Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2890–2896, Brussels, Belgium, 2018. Association for Computational Linguistics.

[Balkir et al., 2022] Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. Challenges in applying explainability methods to improve the fairness of nlp models. arXiv preprint arXiv:2206.03945, 2022.

[Bender and Koller, 2020] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, 2020.

[Bolukbasi et al., 2016] Tolga Bolukbasi, Kaï-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29:4349–4357, 2016.

[Bowman et al., 2015] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326, 2015.

[Clark et al., 2019] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683, 2019.

[Cui et al., 2016] Yiming Cui, Zhipei Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-attention neural networks for reading comprehension. arXiv preprint arXiv:1607.04423, 2016.

[Dawkins, 2021] Hillary Dawkins. Marked attribute bias in natural language inference. arXiv preprint arXiv:2109.14039, 2021.

[Devlin et al., 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[Du et al., 2021] Mengnan Du, Varun Manjunatha, R. Jain, Ruchi Deshpande, Franck Dernoncourt, Jixiang Gu, Tong Sun, and Xia Hu. Towards interpreting and mitigating shortcut learning behavior of nlu models. In NAACL, 2021.

[Gardner et al., 2021] Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah Smith. Competency problems: On finding and removing artifacts in language data. arXiv preprint arXiv:2104.08646, 2021.

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

[Goyal et al., 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6904–6913, 2017.

[Gu et al., 2020] Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. Token-level adaptive training for neural machine translation. arXiv preprint arXiv:2010.04380, 2020.

[Gururangan et al., 2018] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324, 2018.

[Jia and Liang, 2017] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328, 2017.

[Jiao et al., 2018] Zhenyu Jiao, Shuqi Sun, and Ke Sun. Chinese lexical analysis with deep bi-gru-crf network. arXiv preprint arXiv:1807.01882, 2018.

[Jin et al., 2020] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8018–8025. AAAI Press, 2020.

[Kaneko and Bollegala, 2021] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. arXiv preprint arXiv:2101.09523, 2021.

[Kaushik and Lipton, 2018] Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. arXiv preprint arXiv:1808.04926, 2018.

[Kavumba et al., 2021] Pride Kavumba, Benjamin Heinzerling, Ana Brassard, and Kentaro Inui. Learning to learn to be right for the right reasons. arXiv preprint arXiv:2104.11514, 2021.

[Lai et al., 2021] Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. Why machine reading comprehension models learn shortcuts? arXiv preprint arXiv:2106.01024, 2021.
[Liu et al., 2018] Xin Liu, Qingcai Chen, Chong Deng, Hua-jun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. LCQmc: A large-scale chinese question matching corpus. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1952–1962, 2018.

[Liu et al., 2019a] Nelson F Liu, Roy Schwartz, and Noah A Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. arXiv preprint arXiv:1904.02668, 2019.

[Liu et al., 2019b] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv, 2019.

[May et al., 2019] Chandler May, Alex Wang, Shikha Bordinia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561, 2019.

[McCoy et al., 2019] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007, 2019.

[Morris et al., 2020] John X Morris, Eli Litland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. Reevaluating adversarial examples in natural language. arXiv preprint arXiv:2004.14174, 2020.

[Niven and Kao, 2019] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. arXiv preprint arXiv:1907.07355, 2019.

[Ott et al., 2018] Myle Ott, Michael Auli, David Grangier, and Marc’Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In International Conference on Machine Learning, pages 3956–3965. PMLR, 2018.

[Ravfogel et al., 2020] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Winton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. arXiv preprint arXiv:2004.07667, 2020.

[Ren et al., 2019] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1085–1097, Florence, Italy, 2019. Association for Computational Linguistics.

[Ribeiro et al., 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

[Schuster et al., 2019] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. 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), pages 3419–3425, 2019.

[Stacey et al., 2020] Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training. arXiv preprint arXiv:2004.07790, 2020.

[Sun et al., 2019] Yu Sun, Shuhuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223, 2019.

[Utama et al., 2020] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Towards debiasing nlu models from unknown biases. arXiv preprint arXiv:2009.12303, 2020.

[Wang et al., 2021] Lijie Wang, Hao Liu, Shuyuan Peng, Hongxuan Tang, Xinyan Xiao, Ying Chen, Hua Wu, and Haifeng Wang. Dutrust: A sentiment analysis dataset for trustworthiness evaluation. arXiv preprint arXiv:2108.13140, 2021.

[Webster et al., 2020] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. arXiv preprint arXiv:2010.06032, 2020.

[Williams et al., 2017] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426, 2017.

[Yaghoobzadeh et al., 2021] Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, Timothy J Hazen, and Alessandro Sordoni. Increasing robustness to spurious correlations using forgettable examples. In EACL, 2021.

[Ye and Kovashka, 2021] Keren Ye and Adriana Kovashka. A case study of the shortcut effects in visual commonsense reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 3181–3189, 2021.

[Zhu et al., 2021] Hongyu Zhu, Yan Chen, Jing Yan, Jing Liu, Yu Hong, Ying Chen, Hua Wu, and Haifeng Wang. Duqm: A chinese dataset of linguistically perturbed natural questions for evaluating the robustness of question matching models. arXiv preprint arXiv:2112.08609, 2021.