Less is Better: Recovering Intended-Feature Subspace to Robustify NLU Models

Ting Wu
Fudan University
tingwu21@m.fudan.edu.cn

Tao Gui [*]
Fudan University
tgui@fudan.edu.cn

Abstract

Datasets with significant proportions of *bias* present threats for training a trustworthy model on NLU tasks. Despite yielding great progress, current debiasing methods impose excessive reliance on the knowledge of bias attributes. Definition of the attributes, however, is elusive and varies across different datasets. Furthermore, leveraging these attributes at input level to bias mitigation may leave a gap between intrinsic properties and the underlying decision rule. To narrow down this gap and liberate the supervision on bias, we suggest extending bias mitigation into feature space. Therefore, a novel model, **R**ecovering **I**ntended-Feature **S**ubspace with **K**nowledge-Free (RISK) is developed. Assuming that shortcut features caused by various biases are unintended for prediction, RISK views them as redundant features. When delving into a lower manifold to remove redundancies, RISK reveals that an extremely low-dimensional subspace with *intended features* can robustly represent the highly biased dataset. Empirical results demonstrate our model can consistently improve model generalization to out-of-distribution set, and achieves a new state-of-the-art performance [1].

1 Introduction

Pretrained language models have achieved remarkable performance on a wide range of natural language understanding (NLU) benchmarks (Devlin et al., 2019). However, when encountering more challenging test sets, they dramatically fail (McCoy et al., 2019). Studies indicate such a dilemma is mainly rooted in the model's reliance on specific *dataset biases* (Gururangan et al., 2018; Zhang et al., 2019a; Schuster et al., 2019), which correlate well with labels but not for the intended underlying task. For instance, on the natural language inference (NLI) task, models tend to use negation cues ("not", "no", etc.), for a *Contradiction* prediction, whereas a learner intended to learn the *underlying correlation* based on the context semantics.

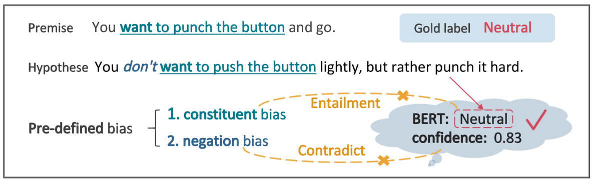

Figure 1: A toy example that illustrate bias in MNLI-matched dev set. BERT's prediction *Neutral* does not comply with the assumed decision rule (Entailment, Contradiction) caused by pre-defined bias.

To train a NLU model that captures the *underlying correlation* from biased datasets, current approaches focus on how to leverage kinds of supervision effectively. One of the most popular forms of such supervision is to explicitly construct a bias-only model under human annotations, e.g., a hypothesis-only model for NLI task, and factor it out from the main model through ensemble-based training (Clark et al., 2019; Utama et al., 2020a). Another empirical line of research shifts supervision from bias type annotations to weak model learners. They find models with limited capacity (Clark et al., 2020; Sanh et al., 2021) or training on limited dataset (Utama et al., 2020b) prone to extract shortcut patterns first, the observation of which can be utilized to mitigate dataset bias.

Despite the supervision on bias has shown effectiveness in bias mitigation, the fundamental questions remain unsolved. On one hand, acquirement of supervision on the bias either from human knowledge or model learning behaviours, is often a laborious and expensive cost. Moreover, considering the definition of bias attributes is elusive and varies across datasets, the external

[*] Corresponding author.

[1] Our code and data are available at https://github.com/CuteyThyme/RISK.git.

Proceedings of the 29th International Conference on Computational Linguistics, pages 1666–1676
October 12–17, 2022.

knowledge can not cover all types of biases in the dataset, leaving potential bias underexplored, e.g., bias beyond the definition or bias harder to learn. On the other hand, capturing bias only at the input examples is just like a black box, being oblivious to the intrinsic properties that drives model to make prediction. The toy example shown in Figure 1 reveals that predefined bias does not necessarily lead the model to learn the unintended decision rule (i.e., constituent bias triggers an Entailment prediction, negation bias triggers an Contradiction prediction). Hence, current debiasing methods inevitably fall short in above two limitations.

On account of the consensus that shortcut features induced by biased examples are detrimental for prediction, various kinds of biases can thus be equivalently viewed as redundancies. When delving into feature space, closer to the decision rule to remove these redundant features, supervision of the attributes from biased examples can be liberated as well. Therefore, we develop a novel model, **R**ecovering **I**ntended-Feature **S**ubspace with **K**nowledge-Free (RISK). Aimed with purifying redundancies from feature space, RISK reveals that for a highly biased dataset, a small subset of *informative* and *shared* features, i.e. intended ones, can give rise to a robust prediction. Concretely, RISK maps features into a lower manifold and learns an orthogonal projector spanned by *geometric median subspace* to recover the intended-feature subspace in an end-to-end manner.

Experimental results on three NLU tasks show RISK outperforms other methods by a large margin, indicating its potential to mitigate bias and the prerequisites of supervision on biased attributes can be liberated. Moreover, when transferring to more challenging out-of-distribution set, RISK can consistently improve the robustness of NLU models. To sum up, our contributions are three-fold as follows:

• We propose a novel *feature-based* debiasing model, termed as RISK. RISK is the initial attempt that free of the supervision on bias attributes.

• We reveal shortcut features as part of redundancy, and thus only leveraging the informative features shared across biased and bias-free examples can achieve the goal of bias mitigation.

• We conduct extensive experiments to validate the effectiveness of RISK in mitigating bias. Moreover, RISK exhibits great power to generalize to more challenging scenarios, showing its potential to robustify NLU models.

2 Bias Mitigation As Feature Redundancy

2.1 Problem Setup

Given training dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ including C classes, a NLU task requires the model to understand the semantic of input text x_i and then predict the target label y_i. Generally, the model is composed of a feature extractor $\mathcal{F}(\cdot) : \mathcal{F}(\mathbf{x}) \rightarrow \mathbf{z}$ and a linear classifier $g(\cdot) : g(\mathbf{z}) \rightarrow \hat{\mathbf{y}}$.

When the model is trained on a highly biased dataset, it will easily capture shortcut features in high-dimensional $\mathbf{z}$. Since the shortcut features are the unintended ones that induce predictions, we treat them as a kind of redundancy. Therefore, mitigating dataset bias can be subsumed under minimizing redundancy in feature space.

2.2 Feature Redundancy by Subspace Modeling

In statistical machine learning, feature subspace paves a path towards eliminating redundant features, as it sheds light on projecting high-dimensional feature onto one subspace, which can significantly capture its most significant information. A common formulation for subspace modeling is to find an orthogonal projection $\mathbf{P}$ of dimension d whose subspace can robustly represents the input features (Vaswani et al., 2018). Let $\mathbf{I}$ denote the identity matrix in the ambient space of the high-dimensional feature $\mathbf{z}$, and the least q-th power deviations formulation for $q > 0$ seeks $\mathbf{P}$ that minimizes:

$$\mathcal{L}(\mathbf{P}) = \sum_{i=1}^{N} \left\|(\mathbf{I} - \mathbf{P})z_i\right\|_2^q \tag{1}$$

Classically, taking $q = 2$ results in principal component analysis(PCA), which finds the orthogonal directions of maximum variance:

$$\mathcal{L}(\mathbf{P}) = \sum_{i=1}^{N} \left\|(\mathbf{I} - \mathbf{P})z_i\right\|_2^2$$

2.3 Geometric Median Subspace as solution.

However, even approximate minimization of Eq. 1 is nontrivial, since it has been shown to be NP hard for $1 \leq q < 2$, furthermore, $q < 1$ can result in a wealth of local minima. Literature have theoretically proven the preferable minimization

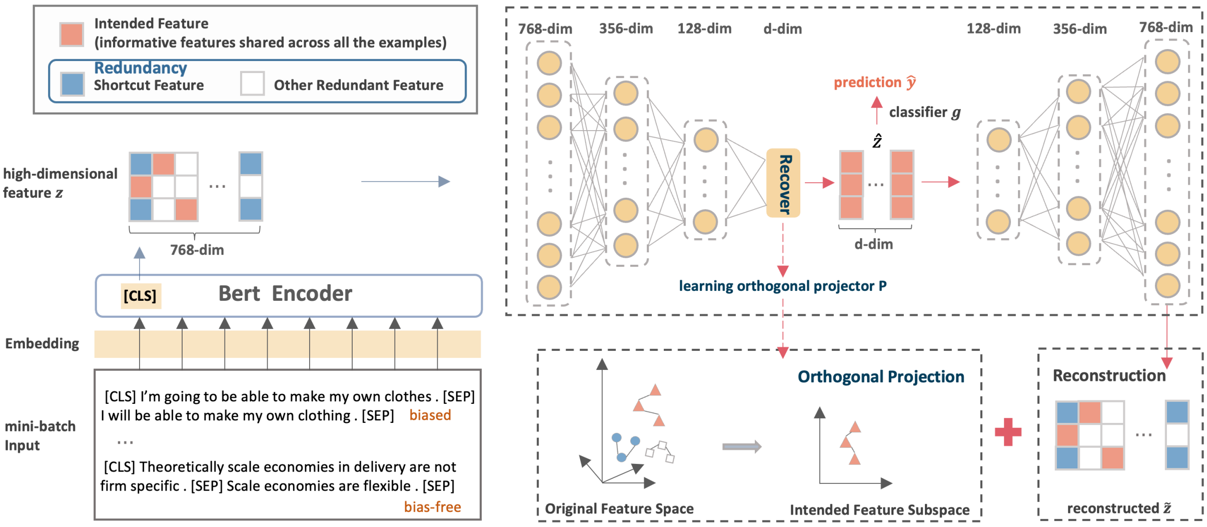
Intended Feature
(informative features shared across all the examples)
Redundancy
Shortcut Feature
Other Redundant Feature
high-dimensional feature z
768-dim
[CLS]
Bert Encoder
Embedding
mini-batch Input
[CLS] I'm going to be able to make my own clothes . [SEP] I will be able to make my own clothing . [SEP]
biased
[CLS] Theoretically scale economies in delivery are not firm specific . [SEP] Scale economies are flexible . [SEP]
bias-free
768-dim
356-dim
128-dim
d-dim
128-dim
356-dim
768-dim
prediction ŷ
classifier g
ẑ
Recover
d-dim
learning orthogonal projector P
Orthogonal Projection
Original Feature Space
Intended Feature Subspace
Reconstruction
reconstructed z̃

Figure 2: Model Architecture of RISK.

is $q = 1$ (Osborne and Watson, 1985; Nyquist, 1988), and thus equals to replace the least squares formulation in PCA with least absolute deviations as follows:

$$\mathcal{L}(\mathbf{P}) = \sum_{i=1}^{N} \left\| (\boldsymbol{I} - \boldsymbol{P}) z_i \right\|_2^1 \tag{2}$$

A nice interpretation of the minimizer of above equation is a *Geometric Median Subspace* (Fletcher et al., 2009), analogous geometric median in modeling centers of input features. Ideally, once we solve a orthogonal projection spanned by this geometric median space, we can achieve an robust estimation of all input features. Since the shortcut features are not shared by bias-free examples, they will be removed automatically as redundancy.

3 RISK: Feature-based Debiasing Without Supervision on Bias

Guided by the theoretical subspace modeling discussed above, in this section, we illustrate the detailed implementation of RISK. In practice, we adopt autoencoder as the main architecture. Leveraging autoencoder to map features into a lower manifold is the first stage of removing redundant features. We further add a simple but effective Recovery Layer within autoencoder to learn a orthogonal projection $\mathbf{P}$, leaving the shared informative features to perform final predictions.

3.1 Delving into Feature Space

We use BERT $\mathcal{F}_\theta$ to map each textual data point x_i into a high-dimensional feature space, that is, $\mathbf{z} = \mathcal{F}(\mathbf{x}, \theta)$. To be specific, $\mathbf{z}$ corresponds to [CLS] token embedding the last layer BERT outputs. It has been convinced that embeddings from pre-trained language models contain much redundency for down-stream NLU tasks (Dalvi et al., 2020). As for highly biased dataset, $\mathbf{z}$ will easily capture substantial shortcut ones. We thus categorize $\mathbf{z}$ into following two feature types:

Intended Features are the *informative* features *shared* across biased and bias-free examples.

Redundant Features include shortcut features that only correlate well with labels, and other redundant features (e.g., task-irrelevant, task-relevant but non-robust ones).

3.2 Autoencoder: To Be Informative Features

For the first stage of mitigating feature redundancy as a way of bias mitigation, we opt to employ an encoder $\mathscr{E}$ composed of a three-layer MLP to map $\mathbf{z}$ into a lower manifold.

Reconstruction loss. As shown in Figure 2, to be symmetric of the encoder $\mathscr{E}$ that project $\mathbf{z}$ into a lower manifold, we also train a decoder $\mathscr{D}$ that map $\hat{\mathbf{z}}$ into $\tilde{\mathbf{z}}$, i.e., the reconstruction representation of $\mathbf{z}$, formulating a bottleneck autoencoder as result. The reconstruction loss is thus defined as:

$$\mathcal{L}_{\text{recon}} = \sum_{i=1}^{N} \left\| z_i - \mathscr{D}(\hat{z}_i) \right\|_2^2 \tag{3}$$

The reconstruction term is used to ensure that a good reconstruction of the original feature can be obtained by using the learned low-dimensional subspace features. Notably, as we defined, *informative* is one of the key characteristics of the intended features. We can further prove that minimize the reconstruction error can serve as maximizing the lower bound of the mutual information between $\mathbf{z}$ and $\hat{\mathbf{z}}$. In general $\hat{\mathbf{z}}$ is not an exact reconstruction of $\mathbf{z}$, but rather in probabilistic terms as the mean of a distribution $p = (Z|\hat{Z} = \hat{\mathbf{z}})$, this yields an associated reconstruction error (Vincent et al., 2010) to be optimized:

$$\mathcal{L}_{\text{recon}} \propto -\text{log}p(\mathbf{z}|\hat{\mathbf{z}})$$

In conjuction with it, minimizing the reconstruction loss actually carry the following optimization:

$$\text{min}\mathbb{E}[\mathcal{L}_{\text{recon}}(\mathbf{z}, \hat{\mathbf{z}})] = \text{max}\mathbb{E}[log\mathbb{P}(\mathbf{z}|\hat{\mathbf{z}})] \quad (4)$$

Maximizing the expectation of the conditional probabilty $\mathbb{E}[log\mathbb{P}(\mathbf{z}|\hat{\mathbf{z}})]$ is equivalent to maximizing the mutual information between $\mathbf{z}$ and $\hat{\mathbf{z}}$ (Chen et al., 2022). This promises the subspace where $\hat{\mathbf{z}}$ lies in is informative and task-relevant for downstream task.

3.3 Recovery Layer: To Be Shared Features

In fact, only utilizing autoencoder can not promise the latent subspace as the intended one we defined before, since shortcut features dominated by biased examples also contain useful but not robust information for prediction. Therefore, going a step further to remove redundant features is needed.

Projection Loss. Leveraging the core idea of subspace modeling, *geometric median subspace* is a preferable minimum to solve the shared features in ideal. In this way, we can recast the problem into learning an orthogonal projector spanned by such median subspace. With the expansion of Eq. 2, the following projection loss function can be achieved:

$$\begin{aligned}\mathcal{L}_{\text{proj}}(\mathbf{A}) = \lambda_1 \sum_{i=1}^{N} \left\|z_i - \mathbf{A}^\top \mathbf{A} z_i\right\|_2^1 \\ + \lambda_2 \left\|\mathbf{A}\mathbf{A}^\top - \mathbf{I}_d\right\|_F^2\end{aligned} \quad (5)$$

we use $\mathbf{A}$ to denote the transformation that reduces feature dimension to d, and $\mathbf{A}^\top$ denotes the transpose of $\mathbf{A}$, $\mathbf{I}_d$ denotes the $d \times d$ identity matrix and $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm. Here λ is an hyperparameter represent the weight of the projection loss to the whole learning objective, for the simplicity, we let $\lambda_1 = \lambda_2$. We later show it associates with the dataset characteristics in Sec. 5.1.

It can be noted that the first term in the weighted sum of above loss function is close to Equation 2 as long as $\mathbf{A}\mathbf{A}^\top$ is close to an orthogonal projector. To enforce this requirement, we introduce the second term that imposes the nearness of $\mathbf{A}\mathbf{A}^\top$ to an orthogonal projection.

Practically, the transformation $\mathbf{A}$ is implemented as a linear MLP layer within the autoencoder $\mathscr{E}$, coined as the Recovery Layer. By applying the projection loss, the parameters of the trained Recovery Layer can approximate the minimal result of Eq. 2. In a sense, the Recovery Layer can be considered as bridging the connections between statistical machine learning and DNN.

3.4 Predictors Fitting in the Intended-Feature Subspace

Intuitively, as a robust model to defend against various distribution shift, it is expected to learn an optimal predictor g, which relies on only the intended features most relevant to current task to make predictions. So for the final step, we just fit a linear classifier in the recovered subspace:

$$g(\hat{\mathbf{z}}) = W^\top \hat{\mathbf{z}} + b$$

Along with minimizing the cross entropy between $g(\hat{\mathbf{z}})$ and $\mathbf{y}$, the final learning objective of RISK is summed into:

$$\mathcal{L}_{\text{RISK}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{proj}}$$

With the dual regularization of reconstruction loss and projection loss, we therefore promise the intended-feature subspace is de facto informative and shared.

4 Experiments

In this section, we provide comprehensive analysis on RISK through extensive experiments on three NLU tasks, and compare out-of-distribution as well as in-distribution accuracy of RISK with other debiasing methods to demonstrate its strength.

4.1 Tasks and Biased Datasets

We evaluate our approach on three NLU tasks: natural language inference (NLI), fact verification, and paraphrase identification.

Natural Language Inference aims to determine whether a premise sentence entails a hypothesis sentence. We use the MNLI dataset (Williams et al., 2018) for training, nevertheless, recent studies indicate that models trained on these NLI datasets tend to adopt shallow heuristics(e.g., lexical overlap, hypothesis-only) to predict (Gururangan et al., 2018; Poliak et al., 2018). Based on the findings, HANS(Heuristic Analysis for NLI Systems, McCoy et al. (2019)) is designed to contain many examples where the heuristics fail, and we condider it as the challenging set for evaluation.

Fact Verification requires models to validate a claim in the context of evidence. For this task, we use the training dataset provided by the FEVER challenge (Thorne et al., 2018). Studies show that models ignoring evidence can still achieve high accuracy on FEVER, accordingly, Fever-Symmetric dataset (Schuster et al., 2019) is used as the test sets for evaluation.

Paraphrase Identification is designed to identify whether a pair of sentences have the same thing. We train the models on QQP (Iyer and Csernai, 2017), a widely used dataset for the task. Similarly to MNLI, models trained on QQP are inclined to mark any sentence pairs with high word overlap as paraphrases despite clear clashes in meaning. As for the balance with respect to the lexical overlap heuristic in PAWS(Paraphrase Adversaries from Word Scrambling, Zhang et al. (2019b)) , we use it as our out-of-distribution set.

4.2 Baseline Methods

We compare RISK against seven debiasing models either with bias known or unknown. As for the bias-known models, supervision on bias is mainly the bias type, and for the bias-unknown models, supervision comes from a shallow model. For all these baseline methods, we adopt the BERT-base model (Devlin et al., 2019) as the main model.

Bias-known-prior Models. i). Reweighting (Clark et al., 2019) trains on a weighted version of the data to encourage the main model to focus on examples the bias-only model gets wrong. **ii). Product-of-Experts** (Clark et al., 2019) forces the main model to focus on learning from examples that are not predicted well by the bias-only model via logit ensembling. **iii). Learned-Mixin** (Clark et al., 2019) further improves this ensemble-based method by parameterizing the ensembling operation, allowing the main model to learn when to incorporate the output from the bias-only model for the ensembled prediction. **iv). Conf-reg** (Utama et al., 2020a) presents a novel *confidence regularization* method that encourage the main model to make predictions with lower confidence on examples that contained biased features.

Bias-known-free Models. For this line of research, models can bypass the need of hand-engineered bias-specific structures since a shallow model is utilized to identify biased examples automatically. **v). Self-debiasing** (Utama et al., 2020b) observe that BERT-base trained on a small subset of the training dataset can grasp the distribution of biased examples. **vi). Weak Learner** (Sanh et al., 2021) view models with limited capacity, i.e. Tiny-BERT (Turc et al., 2020), as the shallow one to obtain biased features. **vii). BERT+$\mathcal{F}_{\mathrm{BiLSTM}}$** (Yaghoobzadeh et al., 2021) employ example forgettting to find minority examples, and robustify the model by fine-tuning twice, first on the full training data and second on the minorities only.

4.3 Implementation Details

For each task, we utilize the training configurations that have been proven to work well in previous studies, that is, a learning rate of $5e^{-5}$ for MNLI and $2e^{-5}$ for FEVER and QQP, and choose AdamW as optimizer with a weight decay of 0.01. For fair comparison, we keep the same bias-only model for all the ensemble-based baselines. To tackle the high performance variance on test sets as observed by Clark et al. (2019), we run each experiment five times and report the mean accuracy scores.

As for the autoencoder, our multiple experiments reveal that when make sure the bottleneck architecture, the detailed dimension of each layer makes few differences. More implementation details such as λ, d selection can be found in Section 5.1.

4.4 Experimental Results

The extensive results of all the above mentioned methods are summarized in Table 1. The results on the original development and test sets of each task represent the in-distribution performance. Obviously, for all three tasks, RISK improves BERT-base by a large margin on the challenging test set. Moreover, it surpasses other baselines not only for the out-of-distribution test set, but also the in-distribution ones.

Out-of-distribution generalization and biases mitigation. The absence of explicit knowledge on bias attributes seemingly create a gap between

Model	MNLI			FEVER			QQP		
	ID	HANS	Δ	ID	Symm.	Δ	ID	PAWS	Δ
BERT-base	84.5	61.2	-	85.6	55.1	-	90.8	36.1	-
Reweighting	83.5	69.2	+8.0	84.6	61.7	+6.6	89.5	48.6	+12.5
Product-of-Experts	84.1	66.3	+5.1	82.3	62.0	+6.9	86.9	56.5	+20.4
Learned-Mixin	84.2	64.0	+2.8	83.3	60.4	+5.3	87.6	55.7	+19.6
Conf-reg	83.4	69.1	+7.9	86.4	60.5	+5.4	89.1	40.0	+3.9
Conf-reg$_{\mathbf{self-debias}}$ ♠	84.3	67.1	+5.9	87.6	60.2	+5.1	89.0	43.0	+6.9
Weak Learner	83.3	67.9	+6.7	85.3	58.5	+3.4	-	-	-
BERT+$\mathcal{F}_{\text{BiLSTM}}$	82.9	70.4	+9.2	86.5	61.7	+6.6	88.0	47.6	+11.5
RISK	**84.5**	**71.3**	**+10.1**	**88.3**	**63.9**	**+8.8**	**90.1**	**56.5**	**+20.4**
w/o Reconstruction Loss	84.2	69.2	+8.0	87.6	60.1	+5.0	90.5	50.6	+14.5
w/o Projection Loss	83.9	64.6	+3.4	86.5	57.7	+2.6	90.4	42.1	+6.0

Table 1: Model performance(accu.(%)) on in-distribution and corresponding challenge test set. ♠: Self-debiasing framework is implemented in conjunction with the bias-known-prior models, we select the version that achieves the best performance in the original paper, i.e., *Confidence Regularization* with annealing mechanism. "w/o Reconstruction Loss" represents RISK is trained without the regularization of reconstruction loss, and "w/o Projection Loss" represents RISK is trained without the regularization of projection loss.

the generalization ability of bias-known models and bias-unknown models. Though RISK furthur eliminate any supervision of specific bias signal, it still generalize well to the out-of-the distribution. To validate the effectiveness of RISK in mitigating bias, in Figure 3, we break down the results on HANS into three different heuristics that the dataset was built upon. The increase of the accuracy in comparison with BERT-base on the *non-entailment* category can reflect the degree to which this bais is removed. Although the overall accuracy of Conf-reg$_{\mathbf{self-debias}}$ on HANS is higher than that of Product-of-experts, as shown in Figure 3, it's debiasing capacity is actually the worst. However, RISK can do well in mitigating the three known biases, and is on par with Product-of-Experts, outperforming other baselines.

In-distribution performance retention. The mitigation of dataset bias often suffers from the trade-off between removing shortcut features and sacrificing in-distribution performance. Especially, on PAWS dataset, this trade-off becomes more pronounced. We can observe that previous methods all have a drop in in-distribution test set for MNLI and QQP, which can be attributable to their explicit omission of biased examples. In contrast, our method finds a balance point via intended-feature subspace, where the out-of-distribution performance is improved and the in-distribution is almost retained. For Fever, the in-distribution accuracy of RISK even increases compared to that of BERT-base.

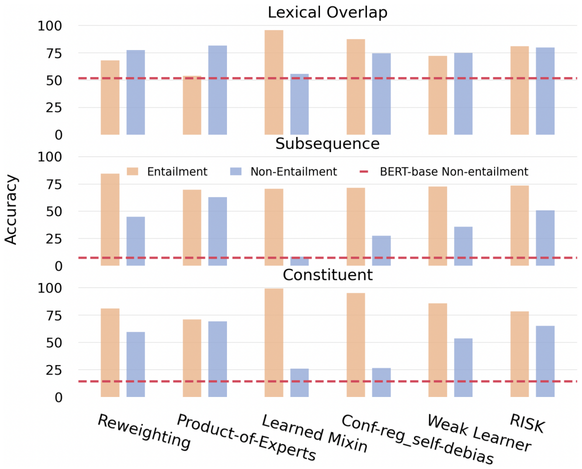

Figure 3: Performance of *RISK* and other baselines on the *entailment* and *non-entailment* categories for each heuristic(i.e., lexical overlap, subsequence and constituent) that HANS was designed to capture.

Ablation Studies. We assumed the reconstruction loss and projection loss are integral parts of RISK as they ensured the intended-feature subspace is *informative* and *shared*. To have an understanding of their impacts on the final performance respectively, we do the ablation studies, and results are shown in Table 1. Comparing the performance degradation, we can conclude that the projection loss plays a key role in helping mitigating dataset bias, and reconstruction loss can be viewed as a regularization that further bound the subspace to be more task-relevant to enhance the accuracy. As can be seen that faced with the removal of reconstruction loss or projection loss, in-distribution performances of the three tasks remain little affected.

5 Analysis and Discussion

In this section, we construct supplementary experiments to further analyze RISK's effectiveness. Free of supervision on bias, we reveal that RISK can deal with more challenging scenarios.

5.1 Hyper-parameter Exploration

To recover the intended feature, we introduce two hyperparameters, the weight λ of projection loss and the subspace dimension d. During the grid search for a fine-grained tuning, we find the values of this two hyperparameters have a close connection with intrinsic properties of dataset.

(1). λ reflects the hardness of challenging set. In the process of optimizing λ, we observe that for the three tasks, RISK achieves best out-of-distribution performance with different value of λ. For the sake of having a qualitative understanding on the three out-of-distribution test set, we compare the average of sentence length and constituency parse tree height of example in HANS, Fever-Symmetric and PAWS respectively.

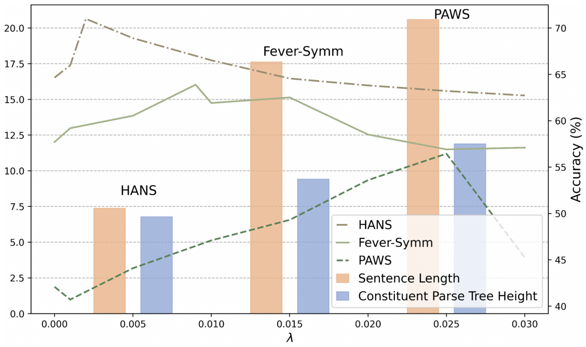

Figure 4: Bar chart represents average sentence length and constituent parse tree height of three out-of-distribution set. Line graph plots model performance with different λ.

As shown in Figure 4, we can observe that PAWS contains longer and syntactically more complex sentences. In contrast, HANS appears to be more easier for model to learn. Accordingly, easier HANS dataset requires a smaller weight of projection loss to obtain the best performance while PAWS requires a larger λ of 0.025. What's more, as for the harder patterns in PAWS for model to generalize, model performance on this task is more *sensitive* to the change of λ in a small range.

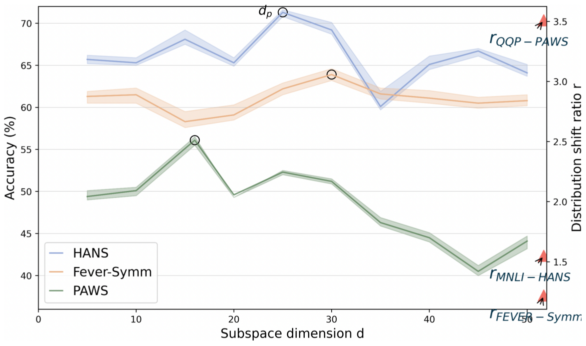

Figure 5: Model performance on HANS, Fever-Symm and PAWS with different subspace dimension d.

(2). d reflects the degree of distribution shift. To quantitatively describe the distribution shift, we propose *bias skewness* as an indicator of how biased a dataset is:

$$\text{bias skewness} = \frac{\#\ \text{biased examples}}{\#\ \text{bias} - \text{free examples}}$$

Thus, the ratio r of bias skewness between ID and OOD can mirror the distribution shift, the larger r, the greater distribution discrepancies. As shown in Figure 5, denote d_p as the optimal subspace dimension original training data set recovered to peak performance on the out-of-distribution set, and it turns out that $\text{PAWS}_{d_p} < \text{HANS}_{d_p} < \text{Fever-Symm}_{d_p}$. However, the ratio r reflects that $r_{\text{QQP-PAWS}} > r_{\text{MNLI-HANS}} > r_{\text{Fever-Symm}}$.

We can conclude that when faced with a larger distribution shift, the subspace dimension d on in-distribution training set should be smaller. In essence, d can be established a close connection with **intrinsic dimension** (Ansuini et al., 2019), i.e., the minimal number of parameters needed to represent a dataset. As our experiments reveal that a **16**-dimensional subspace with intended-features can represent the highly biased QQP training dataset well.

5.2 Transferability Analysis

We further examine the robustness of our approach along with other baselines by transferring to a more challenging scenario, training on MNLI but testing on Adversarial NLI. In our setting, Adversarial NLI contains not only human-crafted adversarial examples (Nie et al., 2020) but also those generated by textual adversarial attacks (TextFooler, Jin et al. (2020)). In general, models utilizing bias patterns that lack the ability to understand the underlying semantics are vulnerable to be attacked. Results are summarized in Table 2 as follows.

Model	R1	R2	R3	ANLI-m
BERT-base	0	28.9	28.8	33.0
Product-of-Experts	**25.2**	27.5	31.3	53.8
Learned-Mixin	23.6	28.0	30.9	54.9
Conf-reg$_{\text{self-debias}}$	21.8	27.4	31.0	48.5
RISK	25.1	**31.2**	**31.9**	**57.1**

Table 2: Model performance(accu.(%)) on adversarial MNLI. ANLI R1-R3 are challenging instances designed by human edition on input text. ANLI-m is adversarial MNLI-matched dataset generated by TextFooler based on blackbox BERT.

We can observe that vanilla BERT-base model trained on MNLI are vulnerable to those adversarial examples, especially ones generated by human edition, suggesting BERT relies overly on bias features to make predictions. On the other hand, either bias-known or bias-unknown models can more or less defend against these attacks. Compared to these baselines, RISK can consistently improve performance on all the adversarial test sets. This indicates the intended subspace has the power to robustify NLU models to various distribution shifts.

6 Related Work

We categorize the multiple lines of research devoted to mitigating dataset bias into three paradigms, in accordance with how the supervision is applied for bias mitigation.

6.1 Supervision from Bias Annotations

Concerns on robustness give rise to the discovery of a wide variety of biases in existing popular datasets, e.g., models make predictions only rely on the hypothesis in NLI datasets (Gururangan et al., 2018). Belinkov et al. (2019) utilize adversarial training to remove the known hypothesis-only features from model internal representations. Moreover, the understanding of specific dataset bias motivates the emergence of ensemble-based debiasing methods (Clark et al., 2019; He et al., 2019; Utama et al., 2020a) , which have shown promising improvements on the out-of-distribution performance. Generally, they view the known dataset biases as prior knowledge and design a simple bias-tailored model, namely the *bias-only model* and factor bias out of the *main model* through ensemble-based training. However, Xiong et al. (2021) theoretically prove that the inaccurate uncertainty estimations of the bias-only model can hurt the debiasing performance, and they propose to conduct calibration on the bias-only model.

6.2 Supervision from Model and Training

The excessive reliance on the assumption that specific types of biased features are known a-prior limits model's transferability. Correspondingly, this line of work seeks for the automatic identification of potentially biased examples, as their empirical results manifest that models with limited capacity (Clark et al., 2020; Sanh et al., 2021) or training on a fewer thousand examples (Utama et al., 2020b) exhibit different learning dynamics, and thus can be used to capture relatively shallow correlations.

Meanwhile, other observations have been made that a better use of minority examples(e.g., examples that are under-represented in the training distribution, or examples that are harder to learn) can play role in models' generalization as well. As Sagawa et al. (2020) point out, the fundamental reason leading to poor generalization lies in models' behaviour of *memorizing* the minority samples. Particularly, Tu et al. (2020) leverage the auxiliary tasks to help improve the generalization capability of pre-trained models on the minority groups. Yaghoobzadeh et al. (2021) propose to use *example forgetting* to find minority examples and make a second fine-tuning on those minorities.

6.3 Supervision from Augmentated Data

Data augmentation techniques have shown to be effective in regularizing models from overfitting to the training data(Novak et al., 2018). In this sense, when distribution shifts, the model will rely little on spurious correlations as a wider variety of predictive features are captured. This has attracted interest as a way to remove biases by explicitly modifying the dataset distribution(Min et al., 2020). Kaushik et al. (2020) and Srivastava et al. (2020) draw upon human-in-the-loop to augment existing training set with diverse and rich examples of potential unmeasured variables. Wang and Culotta (2021) further propose to automatically generate such counterfactual samples via a closet opposite matching strategy. Different from the augmentation of causal associations between features and classes, Wang et al. (2021) apply a cross-data analysis and knowledge-aware perturbations to identify spurious tokens on the stability of model predictions.

7 Conclusion

In this work, we shed light into feature subspace with the aim to create an underlying pathway — from the biased input examples to robust output prediction. Viewing shortcut features as redundancy, we construct a simple but effective Recovery Layer within the autoencoder structure for bias mitigation. Extensive experiments demonstrate the strengths of our model: better generalization, dataset-agnostic transferability and the robustness to more challenging scenarios. We believe this feature-based debiasing framework opens up new directions for establishing a trustworthy NLU model. Meanwhile, our concise motivation and implementation throw out a thought-provoking question, that is for model, for feature, sometimes less can be better.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially National Natural Science Foundation of China (No. 62076069, 61976056), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), and sponsored by CCF-Tencent Open Fund.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiahong Chen, Jing Wang, Weipeng Lin, Kuangen Zhang, and Clarence W de Silva. 2022. Preserving domain private representation via mutual information maximization. *arXiv preprint arXiv:2201.03102*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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)*, Hong Kong, China. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3031–3045, Online. Association for Computational Linguistics.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

P. Thomas Fletcher, Suresh Venkatasubramanian, and Sarang C. Joshi. 2009. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45:S143–S152.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Shankar Iyer and Nikhil Dandekarand Kornél Csernai. 2017. First quora dataset release: Question pairs. *data.quora.com*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

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

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. 2018. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*.

Hans Nyquist. 1988. Least orthogonal absolute deviations. *Computational Statistics & Data Analysis*, 6(4):361–367.

MR Osborne and GA Watson. 1985. An analysis of the total approximation problem in separable norms, and an algorithm for the total l_1 problem. *SIAM journal on scientific and statistical computing*, 6(2):410–424.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines for natural language inference. In *The Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, pages 8346–8356.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. 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, Hong Kong, China. Association for Computational Linguistics.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2020. Well-read students learn better: On the importance of pre-training compact models.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. In *ACL*, pages 8717–8729.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Namrata Vaswani, Thierry Bouwmans, Sajid Javed, and Praneeth Narayanamurthy. 2018. Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery. *IEEE Signal Processing Magazine*, 35(4):32–55.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Tianlu Wang, Diyi Yang, and Xuezhi Wang. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *ArXiv*, abs/2110.07736.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *AAAI*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ruibin Xiong, Yimeng Chen, Liang Pang, Xueqi Cheng, Zhi-Ming Ma, and Yanyan Lan. 2021. Uncertainty calibration for ensemble-based debiasing methods. *Advances in Neural Information Processing Systems*, 34.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

Guanhua Zhang, Bing Bai, Jian Liang, Kun Bai, Shiyu Chang, Mo Yu, Conghui Zhu, and Tiejun Zhao. 2019a. Selection bias explorations and debias methods for natural language sentence matching datasets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4418–4429, Florence, Italy. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.