Mitigating Shortcut Learning with InterpoLated Learning

Michalis Korakakis[1,3] **Andreas Vlachos**[1] **Adrian Weller**[2,3]
[1]Department of Computer Science and Technology, University of Cambridge
[2]Department of Engineering, University of Cambridge
[3]The Alan Turing Institute
{mk2008,av308,aw665}@cam.ac.uk

Abstract

Empirical risk minimization (ERM) incentivizes models to exploit shortcuts, i.e., spurious correlations between input attributes and labels that are prevalent in the majority of the training data but unrelated to the task at hand. This reliance hinders generalization on minority examples, where such correlations do not hold. Existing shortcut mitigation approaches are model-specific, difficult to tune, computationally expensive, and fail to improve learned representations. To address these issues, we propose InterpoLated Learning (InterpoLL) which interpolates the representations of majority examples to include features from intra-class minority examples with shortcut-mitigating patterns. This weakens shortcut influence, enabling models to acquire features predictive across both minority and majority examples. Experimental results on multiple natural language understanding tasks demonstrate that InterpoLL improves minority generalization over both ERM and state-of-the-art shortcut mitigation methods, without compromising accuracy on majority examples. Notably, these gains persist across encoder, encoder-decoder, and decoder-only architectures, demonstrating the method's broad applicability.

1 Introduction

Models trained via empirical risk minimization (ERM) are prone to using shortcuts (Geirhos et al., 2020), i.e., superficial patterns between input features and labels that occur in the majority of training data due to idiosyncrasies in dataset collection and annotation, yet are irrelevant to the task at hand. For example, in MNLI (Williams et al., 2018), most "entailment" examples exhibit high word overlap between the premise and the hypothesis (McCoy et al., 2019). As a result, ERM-trained models achieve high in-distribution (ID) accuracy by relying on this shortcut, but fail on minority examples where the heuristic breaks down, such as instances where high word overlap corresponds to "contradiction." Reliance on shortcuts impairs out-of-distribution (OOD) generalization, particularly under distribution shifts, where minority examples are underrepresented in training but more prevalent at test time (Yang et al., 2023b). Shortcut learning is widespread across NLP tasks, including text classification (Park et al., 2018; Dixon et al., 2018; Borkan et al., 2019) and question answering (Jia and Liang, 2017; Shinoda et al., 2023; Mikula et al., 2024). This highlights a fundamental issue: ERM-trained models often solve the dataset rather than the underlying task.

Typical shortcut mitigation methods include synthetically augmenting minority examples (Zhao et al., 2018; Min et al., 2020; Lee et al., 2021; Wu et al., 2022, *inter alia*), and learning an example weight distribution to up-weight minority examples (Utama et al., 2020b; Sanh et al., 2021; Liu et al., 2021; Korakakis and Vlachos, 2023). However, these techniques primarily improve the classification layer operating on top of model representations (Izmailov et al., 2022). In fact, they do not learn representations different from those of ERM and can even reinforce shortcut features within the model (Mendelson and Belinkov, 2021). Additionally, many shortcut mitigation methods rely on auxiliary models, increasing computational overhead and hyper-parameter tuning complexity (Pezeshki et al., 2024). The introduction of auxiliaries can further destabilize training when their learning dynamics differ from those of the primary learner model (Amirkhani and Pilehvar, 2021).

To address these limitations, we propose InterpoLated Learning (InterpoLL), an interpolation-based technique that mitigates shortcut learning and improves generalization on minority examples. In contrast to existing methods that rely on group annotations, which are impractical to obtain in real-world settings, InterpoLL operates without requiring prior knowledge of minority or majority group

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9191–9206

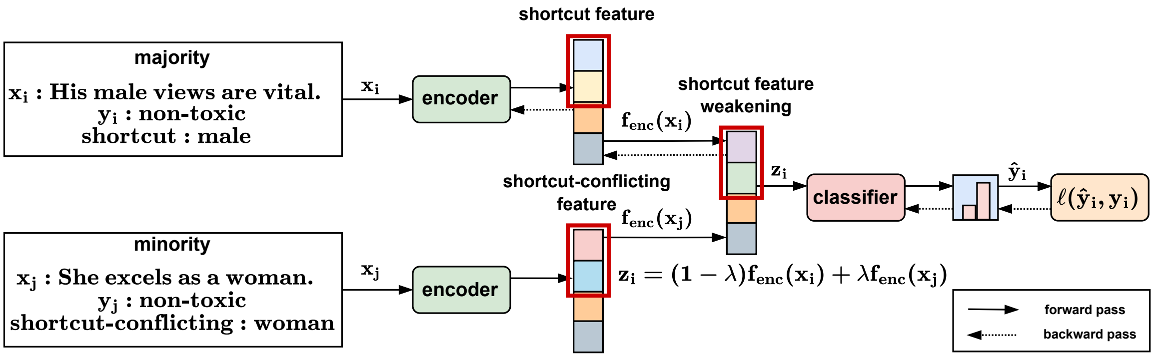

Figure 1: Illustration of InterpoLated Learning (InterpoLL) for a majority example x_i. The term "male" in x_i is over-represented in the "non-toxic" class, creating a shortcut that models exploit. Conversely, the minority example x_j, also labeled "non-toxic," contains the under-represented term "female," which counters this shortcut. To mitigate shortcut reliance, during the forward pass for x_i, we interpolate between the encoder representations of x_i and x_j as $z_i = (1-\lambda)f_{\text{enc}}(x_i) + \lambda f_{\text{enc}}(x_j)$. This interpolation reduces the influence of shortcut features while preserving the label of x_i, as x_j shares the same ground-truth label. The interpolated representation z_i is then used in the backward pass for x_i to compute the loss and update model parameters.

labels. InterpoLL interpolates the representations of majority examples with those of intra-class minority examples that exhibit shortcut-mitigating features (Figure 1). This discourages the model from relying on shortcuts for prediction, and shifts learning toward features that are predictive across all examples.

We evaluate InterpoLL on six natural language understanding datasets. Results show that InterpoLL achieves substantial improvements in minority generalization over ERM and state-of-the-art methods. Notably, it outperforms approaches that rely on prior knowledge of minority examples in the training data. InterpoLL also enhances domain generalization without access to domain-specific data. These improvements are consistent across encoder, encoder-decoder, and decoder-only architectures, as well as different model scales.

Detailed analyses demonstrate that, compared to existing shortcut mitigation methods, InterpoLL: (1) integrates with multiple techniques for identifying minority and majority examples; (2) reduces the presence of shortcut features in model representations; (3) demonstrates greater robustness to noise in the training data; (4) balances training dynamics between minority and majority examples; and (5) maintains a runtime comparable to ERM.

2 Method

2.1 Preliminaries

Problem Setting We consider a supervised learning task where $x \in \mathcal{X}$ is the input and $y \in \mathcal{Y}$ is the ground-truth label. Following Sagawa et al. (2020), we further assume that the training dataset contains distinct groups of examples within some classes. Some of these groups are well-represented and strongly associated with labels, e.g., high word overlap and "entailment" in natural language inference (NLI) datasets (McCoy et al., 2019), while others are not, e.g., negation words in the hypothesis and "entailment" (Gururangan et al., 2018). We refer to the examples belonging to the well-represented groups associated with a particular class as majority examples g_{maj} of said class, and the rest as minority g_{min}.[1]

We want to train a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ. The model f_θ consists of an encoder $f_{\text{enc}} : \mathcal{X} \to \mathcal{Z}$ mapping the input x_i to the representation z_i, and a classifier $f_{\text{cls}} : \mathcal{Z} \to \mathcal{Y}$ which takes the representation z_i as input and maps it to a softmax probability over $\mathcal{Y}$. The model f_θ is typically optimized to minimize the average training loss through ERM. Given a loss function $\ell(f_\theta(x_i), y_i)$, e.g., cross-entropy loss, ERM minimizes the following objective:

$$J_{\text{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)\,. \tag{1}$$

Training with ERM causes the model to em-

[1] Note that some examples can be majority for a particular class, and other examples exhibiting the same patterns can be minority for another, e.g., in NLI, examples containing negation words in the hypothesis are majority for the "contradiction" class, but minority for the "entailment" class.

Algorithm 1 InterpoLated Learning (InterpoLL)

Input: Training dataset $\mathcal{D}$, auxiliary model f_ϕ, learner model $f_\theta = \{f_{\text{enc}}, f_{\text{cls}}\}$

1: $g_{\text{min}}, g_{\text{maj}} = \textsc{InferMinMaj}(f_\phi, \mathcal{D})$ ▷ classify examples into minority and majority
2: Sample $B_1 \sim \mathcal{D}$
3: $B_2 = \emptyset$
4: **for** $(x_i, y_i) \in B_1$ **do**
5:     **if** $(x_i, y_i) \in g_{\text{maj}}$ **then**
6:         Sample $(x_j, y_j) \sim g_{\text{min}}$ s.t. $y_i = y_j$ ▷ sample intra-class minority example
7:         Sample $\lambda \sim \text{Uniform}(0, 0.5)$ ▷ sample interpolation ratio
8:         $z_i \leftarrow (1-\lambda) f_{\text{enc}}(x_i) + \lambda f_{\text{enc}}(x_j)$ ▷ add minority features to majority example
9:     **else**
10:         $z_i \leftarrow f_{\text{enc}}(x_i)$ ▷ use f_{enc} for minority examples
11:     $B_2 \leftarrow B_2 \cup (x_i, z_i, y_i)$
12: Update θ using B_2

phasize shortcut features prevalent in majority examples (Puli et al., 2023). However, at test time, minority examples with features contradicting the shortcuts may become more prevalent (Koh et al., 2021). As a result, an ERM-trained model generalizes poorly.

Inferring Minority/Majority Examples We assume no prior knowledge about the minority and majority examples in the training dataset. To infer them, we use an under-parameterized ERM-trained auxiliary model f_ϕ, as under-parameterization exacerbates model reliance on shortcut features within majority examples (Sanh et al., 2021; Liu et al., 2021). Formally, we categorize examples in the training dataset $\mathcal{D}$ as minority if they are misclassified by f_ϕ, i.e., $(x_i, y_i) \in g_{\text{min}}$ iff $f_\phi(x_i) \neq y_i$, and as majority if they are correctly classified by f_ϕ, i.e., $(x_i, y_i) \in g_{\text{maj}}$ iff $f_\phi(x_i) = y_i$.

2.2 InterpoLated Learning

Here, we address the limitation of the ERM training objective, which fails to produce models that perform consistently across all examples in a dataset, by introducing InterpoLated Learning (InterpoLL). Our goal is to enable the learner model f_θ to focus on features that are predictive across all examples. To achieve this, we interpolate the representations of majority examples with those from minority examples within the same class, whose features contradict the shortcuts. This process weakens the influence of shortcuts, preventing the model from relying on them for predictions. The overall procedure of InterpoLL is detailed in Algorithm 1.

Concretely, for each majority example $(x_i, y_i) \in g_{\text{maj}}$ in the current mini-batch, we randomly select a minority example (x_j, y_j) from the identified set of minority examples g_{min}, ensuring that $y_i = y_j$ (line 6). Then we interpolate their representations (line 8):

$$z_i = (1-\lambda) f_{\text{enc}}(x_i) + \lambda f_{\text{enc}}(x_j)\,, \qquad (2)$$

where the interpolation ratio λ is sampled from $\text{Uniform}(0, 0.5)$. This process diminishes the shortcut features within x_i by incorporating shortcut-mitigating representations from x_j. Sampling λ from $\text{Uniform}(0, 0.5)$ ensures that the representation of x_i is minimally altered, allowing the model to fit majority examples without exploiting shortcuts (line 7). The label of the interpolated majority example x_i remains unchanged, as both x_i and x_j belong to the same class.

Then, we train the learner f_θ as in ERM (Equation 1) with the following modification: during the forward pass for a majority example $(x_i, y_i) \in g_{\text{maj}}$, instead of using the representation provided by f_{enc}, we use z_i as obtained in Equation 2. The backward pass, however, proceeds as normal, i.e., z_i is used to compute the loss and gradients that update all the parameters of the learner model f_θ. Conversely, for minority examples, we do not alter their representations via Equation 2; thus, both the forward and backward passes use the representations provided by f_{enc} (line 10).

Intuition InterpoLL can be viewed as a form of implicit data augmentation, as it repeatedly adds representations from intra-class minority examples to majority examples. This process increases the prevalence of minority example features, leading to a more uniform distribution of examples.

Method	MNLI			FEVER			QQP			Avg		
	ID	OOD	Stress	ID	OOD	Stress	ID	OOD	Stress	ID	OOD	Stress
ERM	84.9±0.3	62.4±1.3	62.9±0.8	88.4±0.4	55.9±1.6	62.8±2.5	90.2±0.5	33.8±0.6	63.5±1.8	87.8	50.7	63.1
Prior Minority/Majority Example Information Required												
GroupDRO	84.3±0.4	72.5±0.6	63.1±1.5	87.5±0.4	64.1±0.5	64.8±1.7	89.5±0.3	52.9±0.8	66.2±1.2	87.1	63.2	64.7
PoE	83.8±0.7	69.7±1.5	61.9±0.8	86.5±0.6	61.9±1.8	63.2±2.1	89.1±0.6	51.3±1.3	64.1±1.5	86.5	61.0	63.1
Conf-reg	84.7±0.6	71.9±1.8	62.5±1.3	87.1±0.4	62.4±1.9	63.9±0.8	89.8±0.4	52.7±1.5	65.3±1.6	87.2	62.3	63.9
JTT	84.1±0.4	71.8±0.5	62.8±0.9	87.1±0.3	62.3±1.1	64.2±0.9	89.3±0.5	52.5±0.6	65.5±1.4	86.8	62.2	64.2
DFR	84.4±0.5	72.3±0.9	63.1±1.1	87.2±0.3	62.7±0.7	64.6±1.2	89.4±0.4	53.5±1.3	65.9±0.6	87.0	62.8	64.5
No Prior Minority/Majority Example Information Required												
Minimax	83.5±0.4	72.4±0.6	62.6±1.4	85.3±0.5	62.6±0.5	64.5±0.8	87.9±0.7	53.8±1.6	65.8±1.4	85.6	62.9	64.3
Weak-learn	83.8±0.8	71.8±1.3	62.3±0.6	85.6±0.3	62.1±1.1	64.1±1.3	88.3±0.3	53.2±1.1	65.7±1.7	85.9	62.4	64.0
CNC	83.1±1.4	72.4±1.1	62.7±2.1	84.7±0.9	62.8±1.5	63.7±2.5	88.9±0.8	53.5±1.2	64.6±2.1	85.6	62.9	63.7
RNF	83.6±0.8	69.9±0.6	62.2±1.7	85.1±0.8	61.4±1.4	64.2±2.4	87.6±0.5	52.2±0.7	64.2±1.8	85.4	61.2	63.5
InterpoLL	84.6±0.5	**<u>75.6</u>**±0.7	**<u>66.5</u>**±0.8	87.8±0.6	**<u>68.7</u>**±0.8	**<u>69.8</u>**±1.1	89.8±0.5	**<u>56.9</u>**±1.0	**<u>70.5</u>**±0.6	87.4	**67.1**↑3.9	**68.9**↑4.2

Table 1: Accuracies on the MNLI, FEVER, and QQP datasets. We report results on in-distribution development sets (ID), out-of-distribution test sets (OOD), and stress test sets (Stress). All numbers are averaged over five runs (with different random seeds), and standard deviations are reported. The best-performing results are bolded, while underlining indicates statistically significant improvements over the ERM-trained baseline (t-test, $p < 0.05$). Values in blue denote improvements over the next best result.

3 Experimental Setup

Datasets We conduct experiments on six English datasets covering NLI and text classification. For NLI, we use MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), and QQP (Chen et al., 2017), and evaluate minority generalization using OOD and stress test sets (Naik et al., 2018). For text classification, we use FDCL18 (Founta et al., 2018), CivilComments-WILDS, and Amazon-WILDS (Koh et al., 2021), and evaluate minority generalization based on minority-group accuracy on the test set.

Comparisons We compare InterpoLL against shortcut mitigation methods that require prior knowledge of minority and majority examples: **GroupDRO** (Sagawa et al., 2019), **PoE** (Karimi Mahabadi et al., 2020), **Conf-reg** (Utama et al., 2020a), **JTT** (Liu et al., 2021), and **DFR** (Kirichenko et al., 2023). We also evaluate InterpoLL against shortcut mitigation methods that do not rely on such knowledge: **Weak-learn** (Sanh et al., 2021), **Minimax** (Korakakis and Vlachos, 2023), **CNC** (Zhang et al., 2022), and **RNF** (Du et al., 2021b).

Implementation Details We use BERT-base (Devlin et al., 2019) from HuggingFace (Wolf et al., 2020) as the learner model, and TinyBERT (Jiao et al., 2020) as the auxiliary model in InterpoLL for our main experiments. We extract sentence embeddings from the CLS token of the final layer to perform interpolations. Each experiment is conducted five times with different random seeds, and results are reported as mean accuracy with standard deviations.

4 Main Results

Natural Language Inference Based on the results in Table 1, we observe that although all shortcut mitigation methods improve minority generalization compared to ERM, no single previously proposed method consistently outperforms the others. Among methods that require prior knowledge of minority and majority examples, GroupDRO achieves the best OOD and stress test accuracy. Methods that do not rely on such prior information generally perform worse, except for InterpoLL, which achieves the highest OOD and stress test accuracy overall. Compared to GroupDRO, the next best method, InterpoLL improves OOD accuracy by 3.9% and stress test accuracy by 4.2%.

Text Classification Table 2 presents the average and minority test set accuracies for FDCL18 and CivilComments-WILDS, and the 10th percentile accuracy for Amazon-WILDS, following Koh et al. (2021). Similar to the results from NLI, the effectiveness of previous approaches varies across these datasets. However, InterpoLL consistently outperforms all methods, achieving the highest minority accuracy, 4.2% higher than the next best approach, GroupDRO.

Method	FDCL18 Average	FDCL18 Minority	CivilComments-WILDS Average	CivilComments-WILDS Minority	Amazon-WILDS 10th Percentile	Avg Average	Avg Minority
ERM	81.3±0.2	35.6±2.5	85.2±0.3	63.5±0.9	53.8±0.8	83.2	49.6
Prior Minority/Majority Example Information Required							
GroupDRO	76.2±1.5	57.3±3.4	82.1±0.4	69.5±0.9	53.3±0.0	79.2	63.4
PoE	69.5±2.3	53.1±2.7	83.8±0.7	69.6±0.8	52.9±0.8	76.6	61.4
Conf-reg	76.3±1.2	55.9±1.9	84.2±0.5	67.2±1.2	53.4±0.8	80.2	61.6
JTT	76.1±1.4	55.6±1.7	83.9±0.4	69.9±0.8	52.9±0.8	80.0	62.8
DFR	76.3±1.2	56.9±1.4	83.7±0.6	69.8±0.7	53.4±0.8	80.0	63.4
No Prior Minority/Majority Example Information Required							
Minimax	75.9±1.1	56.8±1.8	83.2±0.4	68.4±0.8	52.9±0.0	79.6	62.6
Weak-learn	71.6±1.3	56.5±2.0	83.1±0.5	67.9±0.7	52.4±0.0	77.4	62.2
CNC	75.6±0.7	56.7±1.3	83.5±0.9	70.1±1.4	53.4±0.0	79.6	63.4
RNF	72.3±0.8	54.9±1.6	83.8±0.4	67.1±0.9	52.9±0.8	78.0	61.0
InterpoLL	78.8±0.6	<u>**61.2**</u>±0.7	84.7±0.7	<u>**73.9**</u>±0.3	<u>**55.2**</u>±0.0	81.8	**67.6**↑4.2

Table 2: Model performance on the FDCL18, CivilComments-WILDS, and Amazon-WILDS datasets. For FDCL18 and CivilComments-WILDS, we report average and minority test accuracy, and for Amazon-WILDS, we report 10th Percentile accuracy (Koh et al., 2021). All numbers are averaged over five runs (with different random seeds), and standard deviations are reported. The best-performing results are bolded, while underlining indicates statistically significant improvements over the ERM-trained baseline (t-test, $p < 0.05$). The value in blue indicates improvement over the next best result. Amazon-WILDS is excluded from the calculations in the last two columns.

5 Domain Generalization

Experimental Setup We focus on domain generalization, where a model is applied to out-of-domain settings without access to any additional domain data. We hypothesize that by reducing the model's reliance on shortcut features for predictions, InterpoLL facilitates the acquisition of task-relevant features, thereby enhancing domain generalization. To this end, we conduct experiments using GLUE-X (Yang et al., 2023a). We restrict our comparison to methods that do not rely on prior assumptions about shortcuts, as GLUE-X does not provide information on minority or majority examples. Detailed dataset statistics are provided in Table 11 in the Appendix.

Method	SST-2	MNLI	QNLI	RTE	MRPC	QQP	Avg
ERM	87.2	68.7	77.1	57.9	58.4	66.5	69.3
Minimax	88.1	70.2	77.3	57.7	58.8	67.5	69.9
Weak-learn	87.5	68.9	77.5	57.8	58.6	66.7	69.5
RNF	87.8	69.8	77.4	56.2	58.4	66.8	69.4
InterpoLL	**90.1**	**71.5**	**81.2**	**60.6**	**60.9**	**69.8**	**72.4**↑2.5

Table 3: Domain generalization results on GLUE-X. The value in blue indicates an improvement over the next best result.

Results Table 3 shows domain generalization results. We observe that Minimax, Weak-learner, and RNF offer only minor improvements over ERM, with average gains of 0.6%, 0.2%, and 0.1%, respectively. In contrast, InterpoLL achieves average improvements of 3.1% over ERM, 2.5% over Minimax, 2.9% over Weak-learner, and 3.0% over RNF.

6 Analysis

Model	Method	HANS	PAWS	Sym.	RTE	MRPC	Avg
BERT-l	ERM	73.8	42.9	65.2	64.9	61.7	61.7
	InterpoLL	**78.2**	**59.4**	**69.1**	**68.6**	**63.4**	**67.7**
RoBERTa-l	ERM	74.7	45.4	68.1	74.4	64.6	65.4
	InterpoLL	**80.1**	**60.5**	**74.5**	**77.3**	**66.9**	**71.9**
XLNet-l	ERM	75.3	47.6	69.4	75.1	64.3	66.4
	InterpoLL	**81.6**	**59.7**	**78.2**	**78.4**	**67.5**	**73.1**
ELECTRA-l	ERM	76.2	49.3	70.5	76.3	68.4	68.1
	InterpoLL	**81.8**	**61.6**	**78.4**	**77.9**	**70.3**	**74.0**
T5-l	ERM	77.8	50.1	71.8	79.1	70.8	69.9
	InterpoLL	**82.9**	**63.9**	**78.9**	**81.5**	**72.7**	**76.0**
T5-3B	ERM	78.9	51.2	72.4	79.9	71.6	70.8
	InterpoLL	**84.5**	**65.1**	**79.7**	**82.8**	**74.2**	**77.3**
GPT2-m	ERM	61.2	44.9	64.7	65.8	56.1	58.5
	InterpoLL	**67.8**	**54.3**	**69.2**	**68.7**	**59.5**	**63.9**
GPT2-l	ERM	70.8	47.5	66.8	74.9	61.4	64.3
	InterpoLL	**77.4**	**59.6**	**72.1**	**77.8**	**63.3**	**70.0**

Table 4: Performance for large models trained with ERM and InterpoLL. The experimental setup for MNLI, FEVER, and QQP follows Section 3, while the setup for RTE and MRPC follows Section 5. The best results are bolded.

Comparison Across Model Sizes and Architectures We investigate whether the generalization improvements achieved by InterpoLL extend to larger models and different architectures. Specifi-

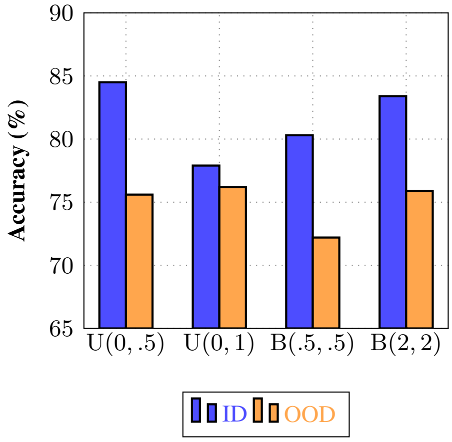

Figure 2: In-distribution (ID) accuracy on MNLI and out-of-distribution (OOD) generalization on HANS for different interpolation ratios λ used in InterpoLL: uniform distributions $\mathrm{U}(0, 0.5)$ and $\mathrm{U}(0, 1)$; beta distributions $\mathrm{B}(0.5, 0.5)$ and $\mathrm{B}(2, 2)$.

cally, we evaluate its performance on large-scale encoder models, including BERT-large (Devlin et al., 2019) (≈340M params), RoBERTa-large (Liu et al., 2019) (≈355M params), ELECTRA-large (Clark et al., 2020) (≈340M params), and XLNet-large (Yang et al., 2019) (≈340M params). We also evaluate encoder-decoder models, such as T5-large (Raffel et al., 2020) (≈770M params) and T5-3B (Raffel et al., 2020) (≈3B params), as well as decoder-only models, including GPT2-medium (Radford et al., 2019) (≈350M params) and GPT2-large (Radford et al., 2019) (≈750M params). For T5 models, interpolations are applied in the final layers of both the encoder and decoder, while for GPT2 models, interpolations are applied in the final layer of the decoder. The results in Table 4 show that InterpoLL improves average OOD performance by 6.0 points for BERT-large, 6.5 points for RoBERTa-large, 6.7 points for XLNet-large, 5.9 points for ELECTRA-large, 6.1 points for T5-large, 6.5 points for T5-3B, 5.4 points for GPT2-medium, and 5.7 points for GPT2-large.

Impact of Interpolation Ratio We investigate the impact of the interpolation ratio λ on InterpoLL when it is sampled from four different distributions: (1) $\mathrm{Uniform}(0, 0.5)$, (2) $\mathrm{Uniform}(0, 1)$, (3) U-shaped $\mathrm{Beta}(0.5, 0.5)$, and (4) bell-shaped $\mathrm{Beta}(2, 2)$. Figure 2 presents results on MNLI. We observe that sampling λ from $\mathrm{Uniform}(0, 0.5)$ achieves better performance by sufficiently weakening the shortcuts, while still enabling the model to learn majority features. Other distributions cause a drastic drop in accuracy either by incorporating a higher proportion of minority features into majority examples, impairing the model's ability to fit them, or by failing to sufficiently weaken the shortcuts.

Method	MNLI		CivilComments	
	ID	HANS	Average	Minority
ERM	84.9	62.4	85.2	63.5
no auxiliary	84.5	75.3	84.9	73.5
BERT-tiny	84.6	**75.6**	84.7	**73.9**
BERT-base	84.1	75.2	84.5	73.6
BERT-large	84.5	74.8	84.1	73.2
under-trained	84.3	75.1	84.6	73.8
regularized	84.4	74.9	84.4	73.7

Table 5: Accuracies on MNLI and CivilComments-WILDS using different auxiliary models for InterpoLL.

Impact of Auxiliary To evaluate how auxiliary model choice affects InterpoLL's effectiveness, we experiment with BERT models of varying sizes. We also evaluate two additional variants: a BERT-base model trained with ERM for 3 epochs (under-trained) (Utama et al., 2020b), and a BERT-base model trained with weight decay set to 1 (regularized) (Liu et al., 2021). Lastly, we assess a variant of InterpoLL that operates without an auxiliary model, where the learner itself identifies minority and majority examples (no auxiliary) (Korakakis and Vlachos, 2023). In this setup, the learner is trained for 2 epochs, and its misclassifications are used as a proxy for identifying minority and majority examples. Results from Table 5 show that all auxiliary variants achieve comparable generalization. Notably, the no-auxiliary InterpoLL variant still outperforms the other shortcut mitigation methods on HANS and CivilComments-WILDS, as shown in Tables 1 and 2.

InterpoLL Variants To better understand the effects of different components of InterpoLL, we conduct experiments with two variants: (1) inter-class InterpoLL (inter) interpolates the representations of majority examples with those of inter-class minority examples, in contrast to standard InterpoLL, which involves intra-class interpolations, (2) inverse InterpoLL (inv) reverses the interpolation direction, i.e., interpolating the representations of minority examples with those of majority examples, thus weakening the features of the former. Figure 3 presents results for models trained on MNLI and evaluated on HANS. We find that intra-class interpolations outperform inter-class ones, as the latter may introduce irrelevant features into major-

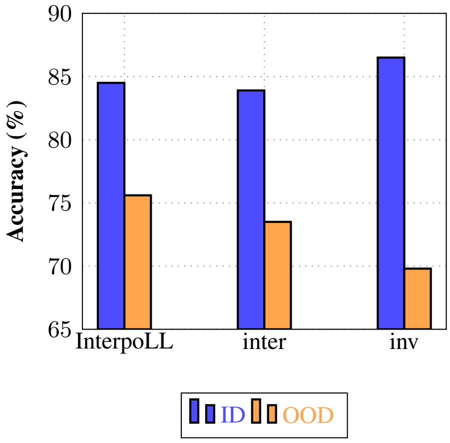

Figure 3: In-distribution (ID) accuracy on MNLI and out-of-distribution (OOD) generalization on HANS for two InterpoLL variants: inter, which performs inter-class interpolation between minority and majority examples; and inv, an inverse variant that weakens minority representations by interpolating them with those of majority examples.

ity examples. Similarly, reversing the interpolation direction improves accuracy on majority examples but reduces it on minority examples.

Method	**Overlap↓**	**Subsequence↓**	**Negation↓**
ERM	2.9±0.3	2.5±0.3	2.7±0.3
GroupDRO	3.1±0.2	2.9±0.5	2.9±0.3
PoE	3.9±0.2	4.1±0.4	3.0±0.3
Conf-reg	4.5±0.3	4.2±0.3	3.2±0.6
JTT	3.9±0.2	4.1±0.4	3.0±0.3
DFR	3.0±0.4	2.5±0.5	2.8±0.5
Minimax	3.5±0.2	3.2±0.2	3.2±0.3
Weak-learn	3.4±0.3	3.3±0.5	3.2±0.5
CNC	2.8±0.6	2.6±0.4	2.6±0.5
RNF	3.3±0.5	3.8±0.5	3.2±0.3
InterpoLL	**2.7**±0.3	**2.5**±0.3	**2.5**±0.2

Table 6: Probing results for Overlap, Subsequence, and Negation shortcut categories on MNLI. Higher values indicate greater extractability of shortcut features from model representations.

Shortcut Extractability We evaluate the extractability of shortcut features from model representations using minimum description length probing (Voita and Titov, 2020; Mendelson and Belinkov, 2021; Lyu et al., 2023). We report compression, where higher values suggest increased shortcut extractability from the model representations. Table 6 shows the probing results on MNLI. We observe that prior methods increase the extractability of shortcuts from model representations, as indicated by high compression values.

Method	**MNLI**		**CivilComments**	
	ID	HANS	Average	Minority
ERM	41.8	38.1	44.3	40.1
GroupDRO	35.9	34.7	39.5	37.7
RNF	38.3	37.6	37.6	35.2
JTT	36.9	33.4	38.9	36.4
Minimax	35.1	34.3	37.9	37.3
InterpoLL	**64.8**	**63.9**	**62.8**	**61.3**

Table 7: Model performance on MNLI and CivilComments-WILDS with 5% synthetic label noise in the training data.

InterpoLL exhibits the lowest compression values, while achieving the highest minority generalization performance (Table 1). Table 14 in the Appendix presents detailed results.

Robustness to Noise To assess InterpoLL's robustness under noisy conditions, we introduce 5% synthetic label noise into the training data and evaluate performance on MNLI and CivilComments-WILDS. The results in Table 7 indicate that InterpoLL exhibits substantially greater resilience to noise compared to ERM and other shortcut mitigation methods.

Method	**Time (hr)**
ERM	4
DFR	4
PoE	6
Weak-learn	6
Conf-reg	7
RNF	5
JTT	8
CNC	39
Minimax	5
InterpoLL	5

Table 8: Training runtime (in hours) on MNLI.

Training Runtime We assess the efficiency of shortcut mitigation methods by measuring training time. As shown in Table 8, InterpoLL introduces only a minor computational overhead compared to ERM, while most other methods approximately double ERM's training time. Notably, InterpoLL is substantially faster than CNC, which also targets the reduction of shortcut features in model representations.

Model Training Dynamics We analyze model training dynamics to assess InterpoLL's ability to balance training across minority and majority examples, thereby counteracting the tendency of

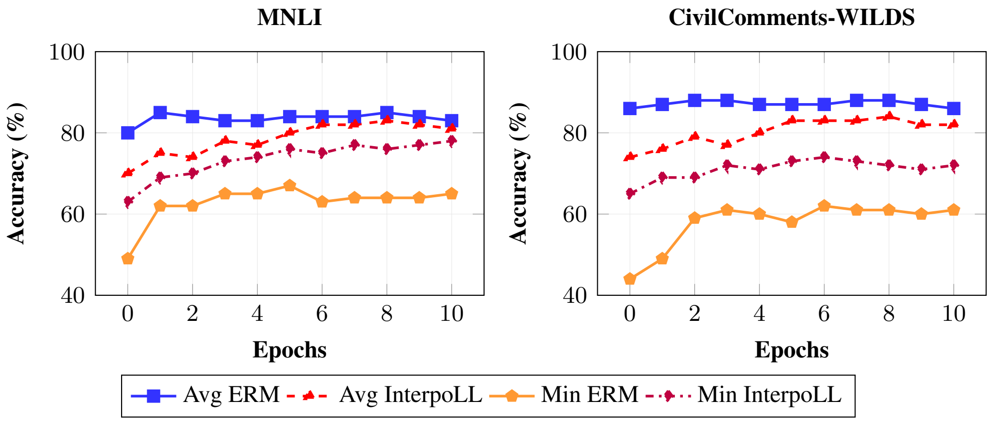

Figure 4: Training set accuracy over epochs for minority and majority examples on MNLI and CivilComments-WILDS, using ERM and InterpoLL.

Dataset	BERT-Tiny	BERT-Large	Under-trained	Regularised
MNLI	95.1%	93.9%	94.6%	94.1%
FEVER	97.1%	95.4%	96.3%	96.1%
QQP	94.9%	93.7%	93.5%	92.7%
CivilComments-WILDS	96.8%	94.8%	95.9%	95.9%

Table 9: Recall of minority examples across different datasets and auxiliary model variants.

ERM-trained models to focus on fitting shortcut features within majority examples first. Figure 4 shows the training set accuracy over epochs on the MNLI and CivilComments-WILDS datasets. We see that while an ERM-trained model eventually achieves high accuracy for minority examples, this only occurs towards the end of training. InterpoLL demonstrates more balanced training throughout, mitigating the model's initial preference for fitting majority examples.

Effectiveness of Minority Example Identification To verify the reliability of different auxiliary models in inferring minority examples, we compute recall, defined as the proportion of ground-truth minority examples correctly identified. We conduct experiments on the MNLI, FEVER, QQP, and CivilComments-WILDS datasets. Table 9 indicates that all auxiliary models achieve high recall.

Weakening Shortcut Features vs Augmenting Minority Examples We compare InterpoLL with mixup (Zhang et al., 2018) and LISA (Yao et al., 2022), two interpolation-based data augmentation techniques. Mixup performs interpolations between random pairs of examples, whereas LISA interpolates examples that either have the same label but come from different example groups, or have different labels but the same example group. Conversely, InterpoLL targets the weakening of shortcut features prevalent in majority examples by interpolating their representations with those of intra-class minority examples exhibiting shortcut-mitigating features. To ensure that majority representations are not substantially altered, the interpolation ratio λ is sampled from a $\mathrm{Uniform}(0, 0.5)$ distribution, rather than from $\mathrm{Uniform}(0, 1)$.

Table 10 presents the experimental results on the MNLI and CivilComments-WILDS datasets. On MNLI, InterpoLL achieves a 6.3% improvement in terms of minority accuracy compared to mixup, and a 3.4% improvement compared to LISA. Similarly, on CivilComments-WILDS, InterpoLL outperforms mixup and LISA by 7.0% and 3.2%, respectively. These results underscore InterpoLL's effectiveness in improving minority generalization over interpolation-based data augmentation methods.

7 Related Work

Mixup InterpoLL is inspired by mixup (Zhang et al., 2018), a data augmentation method that generates synthetic examples by interpolating random pairs of examples and their labels. Recent mixup extensions apply interpolations in representation space (Verma et al., 2019), explore alternative interpolation strategies (Guo et al., 2019b; Mai et al., 2022; Yin et al., 2021; Collins et al., 2023), and integrate mixup with regularization

Method	MNLI ID	MNLI HANS	CivilComments-WILDS Avg	CivilComments-WILDS Minority	Amazon-WILDS 10th %	GLUE-X
ERM	84.9±0.3	62.4±1.3	85.2±0.3	63.5±0.9	53.8±0.0	69.3
mixup	84.7±0.3	69.3±1.4	84.9±0.2	66.9±1.1	52.4±0.0	69.4
LISA	84.5±0.5	72.2±0.8	84.4±0.1	70.7±0.4	53.9±0.0	–
InterpoLL (ours)	84.6±0.5	**75.6**±0.7	84.7±0.7	**73.9**±0.3	**55.2**±0.0	**72.4**

Table 10: Model performance on MNLI, CivilComments-WILDS, Amazon-WILDS, and GLUE-X (averaged over five runs). InterpoLL and ERM results for MNLI and CivilComments-WILDS are repeated from Tables 1 and 2. LISA requires group annotations and is therefore not applicable to GLUE-X.

techniques (Jeong et al., 2022; Kong et al., 2022). In NLP, Guo et al. (2019a) use mixup on word and sentence embeddings, whereas, Yoon et al. (2021) propose a mixup variant that conducts interpolations on the input text. Several other works leverage mixup for model calibration (Zhang and Vaidya, 2021; Park and Caragea, 2022b,a). Zhang et al. (2020) perform mixup to augment the unlabeled pool in active learning, while Korakakis et al. (2024) use it to identify informative and diverse unlabeled instances for annotation.

Shortcut Mitigation in NLP Existing approaches for mitigating shortcuts can be categorized into two main groups, model- and data-centric (Du et al., 2022). Model-centric approaches typically assume that examples with shortcuts are learned more quickly than minority ones due to ERM training. To this end, a popular approach is ensemble an auxiliary model that relies on prior shortcut knowledge for predictions with a learner model during training, such that the latter is discouraged from using shortcuts (He et al., 2019; Clark et al., 2019; Stacey et al., 2020; Xiong et al., 2021; Ghaddar et al., 2021). More recently, Honda et al. (2024) extended this setup in a post-hoc manner by combining the predictions of multiple models under the assumption that each model exploits different features for prediction. Other model-centric methods include contrastive learning (Lyu et al., 2023), knowledge distillation (Du et al., 2021a), forgettable examples (Yaghoobzadeh et al., 2021), adversarial training (Belinkov et al., 2019), and pruning (Meissner et al., 2022). Data-centric methods aim to reduce the presence of shortcuts by modifying the training data itself, typically through data augmentation (Wang and Culotta, 2021; Stacey and Rei, 2024; Cheng and Amiri, 2024) or example filtering (Bras et al., 2020; Swayamdipta et al., 2020).

8 Conclusion

In this work, we introduced InterpoLL, an interpolation-based technique designed to mitigate shortcut learning and improve minority generalization without compromising accuracy on majority examples. The effectiveness of InterpoLL is demonstrated across various datasets, covering a range of natural language understanding tasks. Our analyses show that InterpoLL enhances domain generalization, remains effective across different model architectures and scales, and reduces shortcut features in model representations.

Limitations

Similar to other shortcut mitigation methods, InterpoLL can result in reduced ID accuracy. Moreover, while our probing experiments demonstrate that InterpoLL reduces shortcut-feature extractability, our analysis does not examine whether this suppression inadvertently impacts other aspects of the learned representations. Exploring such potential trade-offs is an important direction for future work. Finally, our experiments focus on specific English-language natural language understanding tasks, leaving InterpoLL's broader applicability to other tasks and languages unexplored.

Acknowledgments

Michalis Korakakis is supported by the Cambridge Commonwealth, European and International Trust, the ESRC Doctoral Training Partnership, and the Alan Turing Institute. Andreas Vlachos is supported by the ERC grant AVeriTeC (GA 865958). Adrian Weller acknowledges support from a Turing AI Fellowship under grant EP/V025279/1.

References

Hossein Amirkhani and Mohammad Taher Pilehvar. 2021. Don't discard all the biased instances: Investigating a core assumption in dataset bias mitigation techniques. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4720–4728, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 491–500. ACM.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.

Jiali Cheng and Hadi Amiri. 2024. FairFlow: Mitigating dataset biases through undecided learning for natural language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975, Miami, Florida, USA. Association for Computational Linguistics.

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)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Katherine M. Collins, Umang Bhatt, Weiyang Liu, Vihari Piratla, Ilia Sucholutsky, Bradley C. Love, and Adrian Weller. 2023. Human-in-the-loop mixup. In *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216 of *Proceedings of Machine Learning Research*, pages 454–464. PMLR.

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 67–73. ACM.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2022. Shortcut learning of large language models in natural language understanding: A survey. *CoRR*, abs/2208.11857.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021a. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Hassan Awadallah, and Xia Ben Hu. 2021b. Fairness via representation neutralization. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12091–12103.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

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

Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. End-to-end self-debiasing framework for robust NLU training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929, Online. Association for Computational Linguistics.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019a. Augmenting data with mixup for sentence classification: An empirical study. *CoRR*, abs/1905.08941.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019b. Mixup as locally linear out-of-manifold regularization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3714–3722. AAAI Press.

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

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.

Ukyo Honda, Tatsushi Oka, Peinan Zhang, and Masato Mita. 2024. Not eliminate but aggregate: Post-hoc control over mixture-of-experts to address shortcut shifts in natural language understanding. *Transactions of the Association for Computational Linguistics*, 12:1268–1289.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. 2022. On feature learning in the presence of spurious correlations. In *NeurIPS*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2022. Augmenting document representations for dense retrieval with interpolation and perturbation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 442–452, Dublin, Ireland. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Fanshuang Kong, Richong Zhang, Xiaohui Guo, Samuel Mensah, and Yongyi Mao. 2022. DropMix: A textual data augmentation combining dropout with mixup. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 890–899, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michalis Korakakis and Andreas Vlachos. 2023. Improving the robustness of NLI models with minimax training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14339, Toronto, Canada. Association for Computational Linguistics.

Michalis Korakakis, Andreas Vlachos, and Adrian Weller. 2024. ALVIN: Active learning via INTerpolation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22715–22728, Miami, Florida, USA. Association for Computational Linguistics.

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheon-Eum Park, and Kyomin Jung. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3181–3185. ACM.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yougang Lyu, Piji Li, Yechang Yang, Maarten de Rijke, Pengjie Ren, Yukun Zhao, Dawei Yin, and Zhaochun Ren. 2023. Feature-level debiased natural language understanding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13353–13361. AAAI Press.

Zhijun Mai, Guosheng Hu, Dexiong Chen, Fumin Shen, and Heng Tao Shen. 2022. Metamixup: Learning adaptive interpolation policy of mixup with metalearning. *IEEE Trans. Neural Networks Learn. Syst.*, 33(7):3050–3064.

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.

Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in NLU. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7607–7613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lukáš Mikula, Michal Štefánik, Marek Petrovič, and Petr Sojka. 2024. Think twice: Measuring the efficiency of eliminating prediction shortcuts of question answering models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2179–2193, St. Julian's, Malta. 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.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Seo Yeon Park and Cornelia Caragea. 2022a. A data cartography based MixUp for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4244–4250, Seattle, United States. Association for Computational Linguistics.

Seo Yeon Park and Cornelia Caragea. 2022b. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. 2024. Discovering environments with XRM. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Aahlad Manas Puli, Lily H. Zhang, Yoav Wald, and Rajesh Ranganath. 2023. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. *CoRR*, abs/2308.12553.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2023. Which shortcut solution do question answering models prefer to learn? In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13564–13572. AAAI Press.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

Joe Stacey and Marek Rei. 2024. Distilling robustness into natural language inference models with domain-targeted augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2239–2258, Bangkok, Thailand. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

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.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

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.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14024–14031. AAAI Press.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 2660–2676, Dublin, Ireland. 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. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13657–13669.

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.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023a. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750, Toronto, Canada. Association for Computational Linguistics.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. 2023b. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39584–39622. PMLR.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. 2022. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 25407–25437. PMLR.

Wenpeng Yin, Huan Wang, Jin Qu, and Caiming Xiong. 2021. BatchMixup: Improving training by interpolating hidden states of the entire mini-batch. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4908–4912, Online. Association for Computational Linguistics.

Soyoung Yoon, Gyuwan Kim, and Kyumin Park. 2021. SSMix: Saliency-based span mixup for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3225–3234, Online. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Michael Zhang, Nimit Sharad Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. 2022. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26484–26516. PMLR.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020. SeqMix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579, Online. Association for Computational Linguistics.

Wancong Zhang and Ieshan Vaidya. 2021. Mixup training leads to reduced overfitting and improved calibration for the transformer architecture. *CoRR*, abs/2102.11402.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Dataset Details

Task	ID	OOD
Sentiment Analysis	SST-2	IMDB Yelp Amazon Flipkart
Natural Language Inference	MNLI-m	MNLI-mm SNLI SICK
Question Answering NLI	QNLI	NewsQA
Textual Entailment	RTE	SciTail HANS
Paraphrase	MRPC	QQP Twitter
	QQP	MRPC Twitter

Table 11: Dataset statistics for out-of-domain generalization experiments on GLUE-X.

B Additional Analysis

Method	MNLI ID	HANS
ERM	58.9	54.3
InterpoLL	65.3	71.9

Table 12: Performance comparison on MNLI and HANS when the dominant shortcuts are removed from the training data.

B.1 Hard-to-learn Shortcuts

To evaluate InterpoLL in settings where no known easy-to-learn shortcuts are present, we removed all MNLI training examples containing: (i) high word overlap between premise and hypothesis with the "entailment" label, and (ii) negation in the hypothesis with the "contradiction" label. As shown in Table 12, InterpoLL maintains its strong performance, outperforming ERM in both ID and OOD generalisation.

B.2 Synthetic Shortcut

We modified the MNLI training data by appending a prefix in the hypothesis containing the ground-truth label with probability p, or a random label

Method	p = 0.2	p = 0.4	p = 0.6	p = 0.8
ERM	80.1	75.2	60.9	55.9
InterpoLL	80.3	75.1	73.5	72.9

Table 13: Performance comparison between ERM and InterpoLL for different values of p for the synthetic shortcut.

otherwise. The modified test set always includes a random label in the prefix. We hypothesize that when the shortcut prevalence is low ($p < 0.5$), the shortcut becomes harder to exploit as it no longer forms a dominant pattern. The results on the modified MNLI test set in Table 13 confirm that InterpoLL remains effective even when shortcuts are less likely to be learned early on.

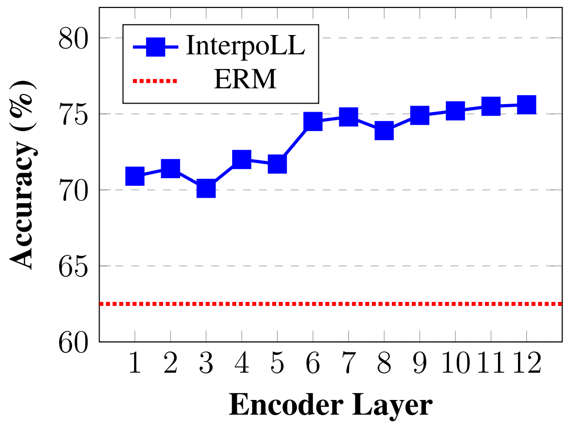

Figure 5: Model performance on HANS for different layers used in InterpoLL for the interpolations between minority and majority examples for a BERT-base model.

B.3 Impact of Interpolation Layer

Figure 5 illustrates the effect of using different encoder layers for conducting interpolations. Models are trained on MNLI and evaluated on HANS. We observe that the most effective layers for interpolations are situated in the upper regions of the model, where higher-level features are typically processed (Jawahar et al., 2019). Notably, InterpoLL outperforms ERM in minority generalization, regardless of the encoder layer used for the interpolations.

Method	Overlap Compr. ↓	Overlap Acc. ↓	Subsequence Compr. ↓	Subsequence Acc. ↓	Negation Compr. ↓	Negation Acc. ↓
ERM	2.9±0.3	86.9±0.7	2.5±0.3	87.8±1.4	2.7±0.3	95.1±0.5
GroupDRO	3.1±0.2	87.1±0.6	2.9±0.5	88.5±1.3	2.9±0.3	96.1±0.6
PoE	3.9±0.2	93.2±1.4	4.1±0.4	94.3±2.1	3.0±0.3	96.3±0.7
Conf-reg	4.5±0.3	94.1±0.5	4.2±0.3	94.8±1.9	3.2±0.6	96.8±0.5
JTT	3.9±0.2	93.2±1.4	4.1±0.4	94.3±2.1	3.0±0.3	96.3±0.7
DFR	3.0±0.4	87.1±0.7	2.5±0.5	87.8±1.5	2.8±0.5	95.3±0.4
Minimax	3.5±0.2	89.1±0.8	3.2±0.2	89.3±1.9	3.2±0.3	96.5±0.6
Weak-learn	3.4±0.3	88.9±0.5	3.3±0.5	89.9±2.1	3.2±0.5	96.9±0.7
CNC	2.8±0.6	86.8±0.7	2.6±0.4	87.7±1.3	2.6±0.5	94.9±0.6
RNF	3.3±0.5	87.3±1.8	3.8±0.5	92.4±1.2	3.2±0.3	96.5±2.2
InterpoLL	**2.7**±0.3	**86.7**±0.4	**2.5**±0.3	**87.6**±1.5	**2.5**±0.2	**94.8**±0.7

Table 14: Probing results for Overlap, Subsequence, and Negation shortcut categories on the MNLI dataset. Higher values in both compression (Compr.) and accuracy (Acc.) metrics indicate greater extractability of shortcut features from the model's representations.