Increasing Robustness to Spurious Correlations using Forgettable Examples

Yadollah Yaghoobzadeh[1] Soroush Mehri[2] Remi Tachet des Combes[2]
Timothy J. Hazen[1] Alessandro Sordoni[2]
[1]Microsoft Turing, Montréal
[2]Microsoft Research, Montréal
{yayaghoo,alsordon}@microsoft.com

Abstract

Neural NLP models tend to rely on spurious correlations between labels and input features to perform their tasks. Minority examples, *i.e.*, examples that contradict the spurious correlations present in the majority of data points, have been shown to increase the out-of-distribution generalization of pre-trained language models. In this paper, we first propose using example forgetting to find minority examples without prior knowledge of the spurious correlations present in the dataset. Forgettable examples are instances either learned and then forgotten during training or never learned. We empirically show how these examples are related to minorities in our training sets. Then, we introduce a new approach to robustify models by fine-tuning our models twice, first on the full training data and second on the minorities only. We obtain substantial improvements in out-of-distribution generalization when applying our approach to the MNLI, QQP, and FEVER datasets.

1 Introduction

Despite the impressive performance of current NLP models, these models often exploit spurious correlations: they tend to capture prediction correlations that hold for most examples but do not hold in general. For instance, in natural language inference (NLI) datasets, word-overlap between hypothesis and premise is highly correlated with the *entailment* label (McCoy et al., 2019; Zhang et al., 2019). Therefore, these models are brittle when tested on examples that cannot be solved by recurring to these correlations, limiting their application in real-world scenarios. *Out-of-distribution* or *challenging* sets are benchmarks carefully designed to break systems that rely on such correlations.

The paradigm of fine-tuning pre-trained language models (PLM) has pushed the state-of-the-art in a large variety of tasks involving natural language understanding (NLU) (Devlin et al., 2019; Wang et al., 2019). This is achieved by self-supervised learning from an enormous amount of text. PLMs also show increased robustness on challenging datasets (Hendrycks et al., 2019). This increase is attributed to an empirical finding that PLMs perform better on *minority examples* present in the training data (Tu et al., 2020). These minority examples violate the spurious correlations and therefore likely support the examples in challenging datasets.

Tu et al. (2020) find minority examples by manually dividing the training data into two groups, according to the known spurious correlations (e.g., word-overlap in NLI). They present an analysis of the robustness of PLMs and its connection to minority examples. In this work, we first introduce a systematic way to find minority examples that does not need prior knowledge of spurious correlations, a big limitation of the earlier work. We then present a simple approach that increases the robustness of PLMs further by tuning models more on these examples.

To identify the set of minority examples, we adopt *example forgetting* (Toneva et al., 2019). This statistic has been shown to relate to the hardness of examples, so we assume it is useful to find minorities in the training data. Based on the definition presented in Toneva et al. (2019), we consider an example *forgettable* if during training it is either properly classified at some point and misclassified later, or if it is never properly classified. This method is model- and task-agnostic. We show in our datasets that minority examples w.r.t to spurious correlations, such as word-overlap in NLI, are well represented in forgettable examples.

After finding minorities through forgettable examples, we propose a simple method to increase the robustness of PLMs further. We perform an additional fine-tuning on the minorities ex-

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3319–3332
April 19 - 23, 2021.

clusively, after fine-tuning on the whole training data. We find this strategy effective, as it increases robust accuracy, i.e., performance on out-of-distribution data, while minimally impacting performance on in-distribution examples. We evaluate our proposed methods in three tasks, including NLI (MNLI, Williams et al., 2017), paraphrase identification (QQP, Iyer et al., 2017) and fact verification (FEVER, Thorne et al., 2018). For each task, recent work has introduced out-of-distribution test sets targetting specific spurious correlations.

Our contributions are the following:

- We propose using forgettable examples as a new approach for finding minority examples from training data without prior knowledge of spurious correlations.
- We show how to exploit minority examples and increase the robustness of deep neural models. This method outperforms other baselines in three challenging datasets: HANS (McCoy et al., 2019), PAWS (Zhang et al., 2019) and FEVER-Symmetric (Schuster et al., 2019). Our method performs effectively when applied to both base and large versions of PLMs (e.g., $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$).
- We observe that finding minorities using a network shallower than the PLM is more effective to robustify it via fine-tuning.
- We show that training models only on forgettable examples leads to poor performance in our datasets, which contrasts with the vision results from Toneva et al. (2019). Our code is available at `github.com/sordonia/hans-forgetting`

2 Datasets

We consider three sentence pair classification tasks, namely natural language inference, paraphrase identification, and fact verification. In the following, we describe the datasets we choose for each task following an introduction of the task.

2.1 Natural Language Inference

The first task we consider is MNLI (Williams et al., 2017), a common natural language inference dataset containing more than 400,000 premise and hypothesis pairs annotated with textual entailment information (*neutral*, *entailment* or *contradiction*). Models trained on this dataset have been shown to capture spurious correlations, such as word-overlap between hypothesis and premise as a strong signal for the *entailment* label (Naik et al., 2018; McCoy et al., 2019). A series of diagnostic out-of-distribution test sets have been devised to test robustness against such heuristics, e.g., HANS.

HANS (McCoy et al., 2019, Heuristic Analysis for NLI Systems) is composed of both *entailment* and *contradiction* examples that have high word-overlap between hypothesis and premise (e.g. "*The president advised the doctor*" $\dashrightarrow$ "*The doctor advised the president*"). A model relying exclusively on the word-overlap feature would not have a higher than chance classification accuracy on HANS. As a matter of fact, BERT (Devlin et al., 2019) performance on this dataset is only slightly better than chance (McCoy et al., 2019). We consider HANS (size: 30k examples) and the MNLI matched dev (Williams et al., 2017) (size: 9815 examples) as our out- and in-distribution test sets for MNLI.

2.2 Paraphrase Identification

QQP (Iyer et al., 2017) is a widely used dataset for paraphrase identification containing over 400,000 pairs of questions annotated as either paraphrase or non-paraphrase. As a consequence of the dataset design, pairs with high lexical overlap have a high probability of being paraphrases. Similarly to MNLI, models trained on QQP are thus prone to learning lexical overlap as a highly informative feature and do not capture the common sense underlying paraphrasing. PAWS dataset is designed to test that.

PAWS (Zhang et al., 2019, Paraphrase Adversaries from Word Scrambling) is a question paraphrase dataset, well-balanced with respect to the lexical overlap heuristic. The accuracy of BERT is around 91.3% on QQP and only 32.2% on PAWS (Table 5). This makes it an interesting test-bed for our method. We use PAWS-QQP as our out-of-distribution set, which contains 677 questions pairs. Training examples from PAWS were never used to update our models. Following Zhang et al. (2019) and Utama et al. (2020), our QQP training and testing splits are based on Wang et al. (2017).

2.3 Fact Verification

The task of fact verification aims to verify a claim given an evidence. The labels are *support*, *refutes*, and *not enough information*. This task is defined as part of the Fact Extraction and Verifi-

cation (FEVER) challenge (Thorne et al., 2018). Schuster et al. (2019) show that models ignoring evidence can still achieve high accuracy on FEVER. They introduce an evaluation test set that challenges that bias. Following Utama et al. (2020), we use the **FEVER-Symmetric** datasets (Symm-v1 and Symm-v2 with 717 and 712 examples, respectively) for out-of-distribution evaluation[1].

3 Finding Minorities with Forgettables

We first define example forgetting and how to compute it. We then show that it can be used to find minority examples in the training data.

3.1 Forgettable examples

An example is forgotten if it goes from being correctly to incorrectly classified during training (each such occurrence is called a *forgetting event*). This happens due to the stochastic nature of gradient descent, in which gradient updates performed on certain examples can hurt performance on others. If an example is forgotten at least once or is never learned during training it is dubbed *forgettable*. Finding forgettable examples entails training the model on $\mathcal{D}$ and tracking the accuracy of each example at each presentation during training. The algorithm for computing forgettability is cheap (Toneva et al., 2019) and only requires storing the accuracy of each particular example at each epoch.

In Toneva et al. (2019), they extracted forgettable examples from a *shallower network* compared to their target model. This makes finding forgettables more efficient and also results in a more diverse set of examples, as the number of forgettable examples is usually higher for weaker models. Another factor is that the shallow models exhibit less memorization due to their fewer number of hyperparameters (Sagawa et al., 2020b) and therefore their forgettables are potentially more representative of the minorities.

We compute forgettable examples using two models with significantly lower capacity compared to PLMs. The first one is a "siamese" BoW classifier in which hypothesis and premise are independently encoded as a mean of word embeddings. This common model in NLP tasks has surprisingly good performance while relying only on the bag of lexical features. We also consider a siamese BiLSTM model. More details can be found in Appendix A. Finally, for comparison, we also experiment with the model used for HANS in SOTA baselines (Clark et al., 2019; Utama et al., 2020) (see also §4.2), as well as $\text{BERT}_{\text{BASE}}$ for in NLI.

[1] https://github.com/TalSchuster/FeverSymmetric

Dataset	Model	$\lvert\mathcal{F}\rvert$	Dev Acc.
MNLI (392,703)	BoW	63,390	64.0
	BiLSTM	46,740	69.6
	BERT	17,748	84.5
QQP (384,348)	BoW	71,116	81.1
	BiLSTM	76,634	84.3
	BERT	20,498	91.3
FEVER (242,911)	BoW	76,368	53.3
	BiLSTM	68,406	56.7
	BERT	21,066	84.4

Table 1: Number of "forgettable" examples along with the accuracy on the MNLI matched, QQP, and FEVER development set. BERT's forgettables are used only in MNLI experiments. The full training size is shown in parenthesis for each dataset.

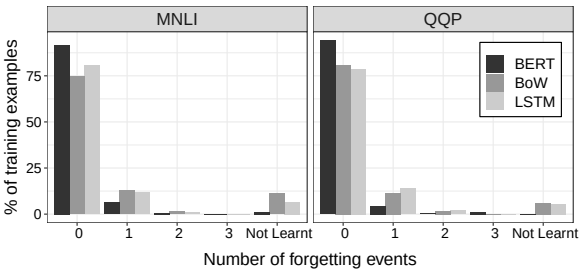

Figure 1: Forgetting events in MNLI and QQP training sets for the three models. A majority of examples are not forgotten during training.

We train the shallow models for five epochs and track forgetting statistics after each epoch. Table 1 shows the number of forgettable examples for BoW, BiLSTM and $\text{BERT}_{\text{BASE}}$ on the MNLI, QQP and FEVER training sets. The performance of the models on the dev set of MNLI is also included.

The distribution of forgetting events for each model can be found in Fig 1. We see that a majority of examples are not forgotten. Most of the forgettables are either those that are never learned or have only one forgetting event. In what follows, we denote the sets of forgettable examples from $\text{BERT}_{\text{BASE}}$, BiLSTM, and BoW as $\mathcal{F}_{\text{BERT}}$, $\mathcal{F}_{\text{BiLSTM}}$ and $\mathcal{F}_{\text{BoW}}$ respectively.

	MNLI		QQP		FEVER	
	P	$\neg P$	P	$\neg P$	P	$\neg P$
all	**.33**	.22	**.51**	.34	**.19**	.15
$\mathcal{F}_{\text{BoW}}$	.26	**.30**	.48	.49	.18	.17
$\mathcal{F}_{\text{BiLSTM}}$	.25	.28	.48	**.50**	.18	17
$\mathcal{F}_{\text{BERT}}$	.26	.26	.50	.50	**.19**	**.18**

Table 2: Average Jaccard index as a measure of word-overlap between two sentences grouped by P (positive) and $\neg P$ (non-positive).

	MNLI		QQP		FEVER	
	N	$\neg N$	N	$\neg N$	N	$\neg N$
all	**31.5**	10.4	**6.2**	4.0	**16.4**	1.2
$\mathcal{F}_{\text{BoW}}$	9.9	11.5	4.1	**4.6**	1.6	**3.1**
$\mathcal{F}_{\text{BiLSTM}}$	11.5	11.2	4.2	4.4	2.9	2.7
$\mathcal{F}_{\text{BERT}}$	14.2	**12.3**	4.2	4.4	6.3	2.5

Table 3: Percentage of examples containing one of the negative keywords in the hypothesis / second question / claim in MNLI / QQP / FEVER. We group examples by binary labels (N: negative and $\neg N$: non-negative) to show the distribution difference between forgettable and overall training examples.

3.2 Forgettable and minority examples

We focus on two important spurious correlations: word-overlap and contradiction-word. These correlations or biases are addressed in related work for MNLI and QQP (Tu et al., 2020; Zhou and Bansal, 2020). For convenience, we use "positive" for either entailment, supports or paraphrase, and "negative" for contradiction, refutes or non-paraphrase

High word-overlap between two sentences is spuriously correlated to the positive label in all three datasets. In Table 2, we show that on average, positive examples have higher word-overlap compared to non-positive ones. In other words, minorities w.r.t. word-overlap correspond to non-positive examples with high word-overlap and positive examples with low word-overlap. For MNLI and QQP, the distribution in $\mathcal{F}_{\text{BoW}}$ and $\mathcal{F}_{\text{BiLSTM}}$ exhibit an interesting behavior: on average, non-positive examples have higher word-overlap. $\mathcal{F}_{\text{BERT}}$ has the same average for both labels. For FEVER, the difference is also clear as the gap in word-overlap between positive and non-positive examples is lower for forgettables. The table allows to conclude that forgettable examples contain more minority examples than a random subset of the same size.

In Table 3, we perform a similar analysis for the presence of contradiction words in the second sentence, which is shown to correlate with negative class in MNLI (Naik et al., 2018; Zhou and Bansal, 2020) and FEVER (Schuster et al., 2019). We choose these contradiction words: {"not", "no", "doesn't", "don't", "never", "any"}, and analyze all three datasets. We observe here as well that forgettables contain more minority examples, as their percentage of examples with a contradiction word is lower for negative examples, which is the opposite than in the overall dataset (with the exception of $\mathcal{F}_{\text{BERT}}$ and FEVER).

4 Robustifying by Fine-Tuning on Minority Examples

Prior work shows that PLMs generalize to out-of-distribution data because they generalize better on minority examples from the training set (Tu et al., 2020). Here, we introduce a simple approach that exploits minority examples to increase robustness. In this approach, we fine-tune a PLM in two successive phases, first on the full training set, and then on the minority examples only. Our method does not need changes to the training objectives. An illustration is shown in Fig 2.

4.1 PLMs

We are interested in the robustness of large PLMs. In this work, we focus on two such models, BERT and XLNet, and experiment with both their base and large versions. $\text{BERT}_{\text{BASE}}$ being the model of choice in previous work (Clark et al., 2019; Zhang et al., 2019; Utama et al., 2020), it will serve as our default architecture. We adopt the Transformers library (Wolf et al., 2019). Our robust models are obtained by fine-tuning PLMs on the full training set for 3 epochs (using the default hyperparameters for each task) and then on the forgettable examples only, for 3 more epochs with a smaller learning rate. See B in Appendix for more details.

4.2 Baselines

Recently, multiple methods have been proposed to learn more robust models through mitigating biases (Clark et al., 2019; He et al., 2019; Mahabadi and Henderson, 2019; Utama et al., 2020). In these works, PLMs are fine-tuned on a re-weighted version of the source dataset, in which examples are weighted based on their hardness. Hardness is measured by training biased models using prior knowledge of the biases or spurious correlations, e.g., a

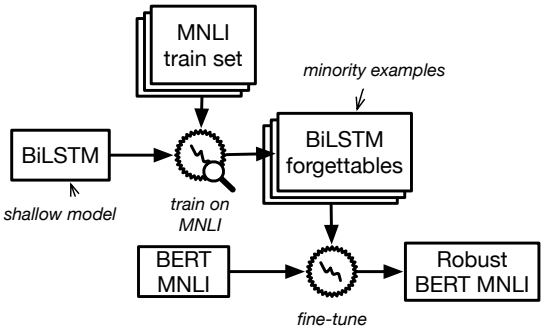

Figure 2: Our proposed framework to train more robust models for an example dataset (MNLI). We first detect minority examples through *forgettable* examples of a shallow model from the training set. We then fine-tune a PLM model (e.g. BERT) in two rounds: first on the full training set, and second on the forgettable subset exclusively. The final model is more robust.

linear model with word-overlap features for NLI. Compared to these works, our method does not need prior knowledge of spurious correlations and exploit the minority examples explicitly by further fine-tuning on them.

We consider the recent *confidence regularization* or "Reg-conf" technique from Utama et al. (2020), as our main baseline in all three tasks. This method is an improvement to the earlier related work in making more robust NLP models (He et al., 2019; Clark et al., 2019). Specifically, Reg-conf claims to maintain the in-distribution performance while improving out-of-distribution and is doing so without introducing new hyperparamters. For MNLI, we report the results of three other baselines of He et al. (2019), Clark et al. (2018) and Mahabadi and Henderson (2019).

Previous work generally design biased models assuming a priori knowledge of the specific dataset biases (with the exception of He et al. (2019), that use a BoW model for HANS). For HANS and PAWS, Clark et al. (2019) and Utama et al. (2020) employ a model with 7 input features, such as word-overlap between premise and hypothesis. To highlight the generality of our approach, we also add this biased model to the set of our shallow models for HANS and fine-tune on its forgettables ($\mathcal{F}_{\text{HANS}}$ with the size of around 200k). For FEVER-symmetric, Utama et al. (2020) consider an LSTM model that takes only the "claim" as input and ignores the "evidence". These baselines re-weight or confidence-regularize training examples using the biased models' performance.

Model	MNLI	HANS	Avg.
BERT	**84.4**$_{\pm 0.1}$	62.9$_{\pm 1.5}$	73.7$_{\pm 0.8}$
BERT+ $\mathcal{F}_{\text{BERT}}$	83.0$_{\pm 0.4}$	68.9$_{\pm 1.4}$	75.9$_{\pm 0.7}$
BERT+ $\mathcal{F}_{\text{BiLSTM}}$	82.9$_{\pm 0.4}$	70.4$_{\pm 0.9}$	76.7$_{\pm 0.5}$
BERT+ $\mathcal{F}_{\text{BoW}}$	83.1$_{\pm 0.3}$	**70.5**$_{\pm 0.7}$	**76.8**$_{\pm 0.4}$
BERT + Rand$_{63,390}$	84.3	63.6	73.9
BERT+ $\mathcal{F}_{\text{HANS}}$	83.9$_{\pm 0.4}$	69.5$_{\pm 0.9}$	76.7$_{\pm 0.5}$
Clark et al. (2019)			
Reweight	83.5	69.2	76.4
Learned Mixin	84.3	64.0	74.2
Mahabadi and Henderson (2019)			
Product of Experts	84.0	66.5	75.3
He et al. (2019)			
DRiFt-HYPO	84.3	67.1	75.7
Utama et al. (2020)			
Reg-conf$_{\text{hans}}$	84.3$_{\pm 0.1}$	69.1$_{\pm 1.2}$	76.7$_{\pm 0.6}$

Table 4: Results of our BERT models fine-tuned on different sources of forgettable examples. For each line, the accuracy on MNLI and HANS are shown, as well as their average.

4.3 Results

4.3.1 MNLI and HANS

In Table 4, we present the results of our models and four recent baselines. The first line reports the performance of BERT on MNLI and HANS. The following lines report the results obtained by fine-tuning BERT on the set of forgettable examples obtained using different shallow models. We also report the average performance between MNLI and HANS. The results confirm that tuning the model towards minority examples improves robustness with a slight drop in MNLI accuracy. Our best model is obtained by fine-tuning on $\mathcal{F}_{\text{BoW}}$, achieving a HANS mean accuracy of 70.5% (with a max of 71.3% over five seeds, which constitutes a +8.4% absolute improvement w.r.t to the initial BERT). To assess whether $\mathcal{F}_{\text{BoW}}$ is indeed responsible for the improvement, we also fine-tune BERT on the same number of randomly chosen examples (BERT + Rand$_{63,390}$), which leads to a negligible improvement.

Fine-tuning on $\mathcal{F}_{\text{BiLSTM}}$ is comparable to fine-tuning on $\mathcal{F}_{\text{BoW}}$, which demonstrates that both BoW and BiLSTM models learn similar spurious correlations. We also added results of fine-tuning BERT on its own forgettables for this task. Note

Model	QQP	PAWS	Avg.
BERT	**90.9**$_{\pm 0.4}$	34.5$_{\pm 1.5}$	62.7$_{\pm 0.6}$
BERT+ $\mathcal{F}_{\text{BOW}}$	89.0$_{\pm 0.9}$	**48.8**$_{\pm 5.2}$	**68.9**$_{\pm 2.2}$
BERT+ $\mathcal{F}_{\text{BILSTM}}$	88.0$_{\pm 0.8}$	47.6$_{\pm 4.1}$	67.8$_{\pm 1.7}$
Utama et al. (2020)			
BERT	91.0	34.3	62.6
Reg-conf$_{\text{hans}}$	89.1	39.8	64.5

Table 5: Results of BERT$_{\text{BASE}}$ trained on different sets of training examples. Accuracy (%) is reported on the QQP test set (size: 10k) and the PAWS dev set (size: 677), alongside their average.

that while it provides less improvement in robustness than on $\mathcal{F}_{\text{BILSTM}}$ or $\mathcal{F}_{\text{BOW}}$ [2], it does generate a significant 6.0% increase in performance. Finally, we also report fine-tuning results on $\mathcal{F}_{\text{HANS}}$, the biased model designed for HANS, and observe that it performs well with a smaller loss on MNLI and a smaller gain on HANS compared to $\mathcal{F}_{\text{BOW}}$ and $\mathcal{F}_{\text{BILSTM}}$.

Compared to other baselines, our approach achieves a comparable or better average accuracy of MNLI and HANS, despite its simplicity. In Fig. 3, we breakdown the results of our best performing model for the three different heuristics HANS was built upon. Our method does not suffer as much as other baselines in the entailment class, and still provides a significant improvement for non-entailment. (More analysis is presented in Appendix.)

4.3.2 QQP and PAWS

Here we report the results of our method applied to QQP and PAWS as out-of-distribution dataset. Results can be found in Table 5. We observe that our method improves out-of-distribution accuracy substantially. It is worth noting that the ground-truth labels in QQP contain noisy annotations (Iyer et al., 2017); a portion of performance loss on QQP could be attributed to that.

Our method outperforms Reg-conf$_{\text{hans}}$, while being simpler in terms of both the biased model and the training regime. We notice that Reg-conf$_{\text{hans}}$ also loses in-distribution performance[3].

[2]To eliminate the forgettables' size factor and focus on the type of model instead, we run an experiment where we sample from $\mathcal{F}_{\text{BOW}}$ the same numbers as $\mathcal{F}_{\text{BERT}}$. The result of our fine-tuning on that smaller $\mathcal{F}_{\text{BOW}}$ was still significantly better than $\mathcal{F}_{\text{BERT}}$.

[3]The authors report accuracy on each label individually and not the overall accuracy. We compute that based on their reported numbers.

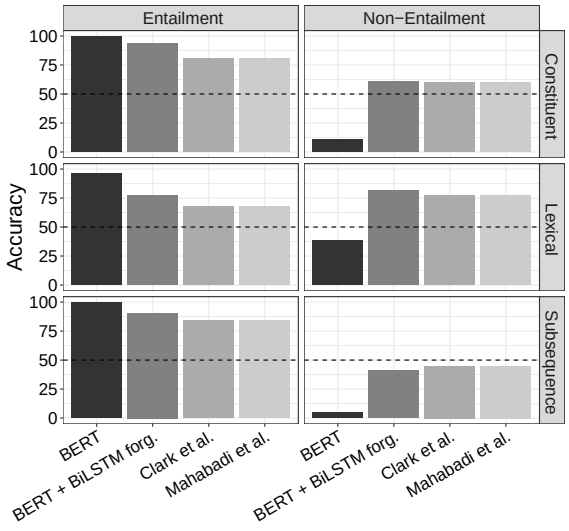

Figure 3: Performance of our BERT fine-tuned on the BiLSTM forgettables $\mathcal{F}_{\text{BILSTM}}$, and baselines on the "entailment" and "non-entailment" categories for each heuristic HANS was designed to capture.

Model	FEVER	Sym-v1	Sym-v2
BERT	86.1$_{\pm 0.3}$	57.7$_{\pm 1.3}$	64.7$_{\pm 1.1}$
BERT+ $\mathcal{F}_{\text{BOW}}$	**87.1**$_{\pm 0.2}$	61.0$_{\pm 1.4}$	**67.0**$_{\pm 1.5}$
BERT+ $\mathcal{F}_{\text{BILSTM}}$	86.5$_{\pm 0.4}$	**61.7**$_{\pm 1.2}$	66.6$_{\pm 1.0}$
Utama et al. (2020)			
BERT	85.8$_{\pm 0.1}$	57.9$_{\pm 1.1}$	64.4$_{\pm 0.6}$
Reweighting$_{\text{bigrams}}$	85.5$_{\pm 0.3}$	**61.7**$_{\pm 1.1}$	66.5$_{\pm 1.3}$
Reg-conf$_{\text{claim}}$	86.4$_{\pm 0.2}$	60.5$_{\pm 0.4}$	66.2$_{\pm 0.6}$

Table 6: Accuracy of different FEVER trained models on FEVER dev, and symmetric v1 and v2 datasets.

4.3.3 FEVER

In Table 6, we report the results of our method applied to the FEVER development and symmetric evaluation sets (see §2.3). Our approach again works well for both $\mathcal{F}_{\text{BOW}}$ and $\mathcal{F}_{\text{BILSTM}}$, but here we also gain on the original dev set when compared to the initial BERT$_{\text{BASE}}$ results. The gains of our method are larger than those of the Reg-conf$_{\text{claim}}$ baseline, which uses a biased model tailored to FEVER-symmetric.

4.4 Analysis

Final loss to detect minority examples An alternative way to find examples from the minority is to simply rank training examples based on their final loss value. In Fig 4, we compare that with our method based on forgettables. The two are

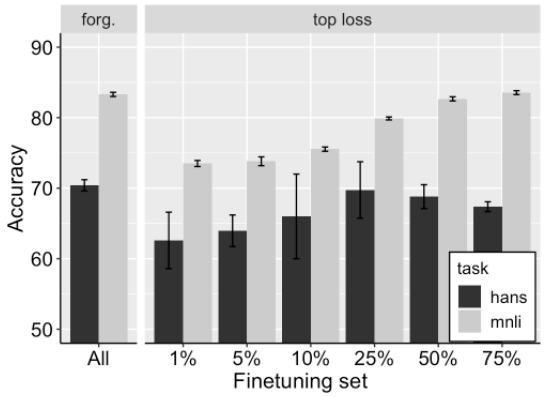

Figure 4: Accuracy on MNLI and HANS when the fine-tuning set is picked from examples in forgettables or in a range of percentages of examples with highest loss.

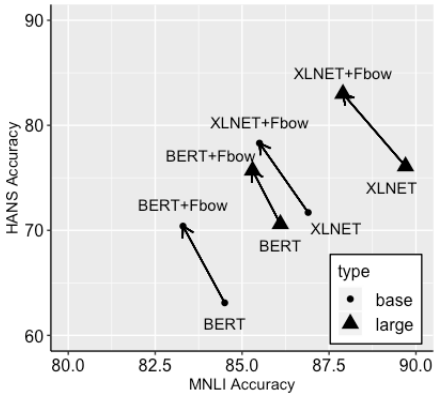

Figure 5: MNLI vs HANS accuracy for both base and large versions of BERT and XLNET.

Train examples	MNLI	HANS
$\mathcal{F}_{\text{BERT}}$ (17,748)	$49.6_{\pm 0.2}$	$37.9_{\pm 1.3}$
Random (17,748)	$74.7_{\pm 0.4}$	$50.8_{\pm 0.2}$
$\mathcal{F}_{\text{BiLSTM}}$ (46,740)	$66.7_{\pm 0.9}$	$54.0_{\pm 0.6}$
Random (46,740)	$78.8_{\pm 0.4}$	$51.3_{\pm 0.6}$
$\mathcal{F}_{\text{BoW}}$ (63,390)	$68.2_{\pm 0.8}$	$55.4_{\pm 1.4}$
Random (63,390)	$79.9_{\pm 0.4}$	$51.8_{\pm 0.2}$
All	$84.5_{\pm 0.1}$	$63.1_{\pm 1.2}$

Table 7: Results of $\text{BERT}_{\text{BASE}}$ models fine-tuned on the set of forgettable examples only.

Model	QQP	PAWS
$\text{BERT}_{\text{BASE}}$	90.9	34.5
$\text{BERT}_{\text{BASE}} + \mathcal{F}_{\text{BoW}}$	89.0	48.8
$\text{BERT}_{\text{LARGE}}$	**91.2**	36.0
$\text{BERT}_{\text{LARGE}} + \mathcal{F}_{\text{BoW}}$	88.3	54.4
$\text{XLNET}_{\text{BASE}}$	90.9	37.1
$\text{XLNET}_{\text{BASE}} + \mathcal{F}_{\text{BoW}}$	88.2	55.2
$\text{XLNET}_{\text{LARGE}}$	89.2	48.4
$\text{XLNET}_{\text{LARGE}} + \mathcal{F}_{\text{BoW}}$	87.8	**65.2**

Table 8: Average accuracy across random seeds on QQP and PAWS for BERT and XLNET base and large models, before and after fine-tuning on $\mathcal{F}_{\text{BoW}}$.

obviously related, as the examples that are never learned rank the highest w.r.t to the loss and are considered as forgettables. However, Fig 4 shows, for MNLI and HANS, that using forgettables produces better performance both in- and out-of-distribution. One additional issue with using the final loss to pick examples is the need to determine either a threshold value α on the loss (keep examples with a loss larger than α) or a number N of examples to retain. The optimal α or N might yield better performance but finding them implies using the out-of-distribution set.

Robustness of larger models We examine the performance of our method when applied to other PLMs and to larger networks by training BERT large and XLNET. Fig 5 shows the MNLI and HANS performance of those networks. Firstly, XLNet is noticeably more robust than BERT, compatible with its superior in-distribution performance (Yang et al., 2019). Secondly, we observe that the large versions generalize on HANS significantly better than their base counterparts (e.g., 76.1% vs 71.7% for XLNet, 70.6% vs 62.9% for BERT), confirming that larger models seem more robust. Lastly, $\text{XLNET}_{\text{LARGE}} + \mathcal{F}_{\text{BoW}}$ shows a +7% increase in performance, reaching 83.1% on HANS with a maximum score of 86.8% over three seeds. We also show Table 8 and Table 9) similar findings on QQP and FEVER. For instance, $\text{XLNET}_{\text{LARGE}} + \mathcal{F}_{\text{BoW}}$ achieves 65.2% on PAWS and 75.3% on FEVER-Sym-v1.

Training on forgettables only Toneva et al. (2019) showed that forgettable examples form the support of the training distribution. We follow their experimental setting and fine-tune BERT on the subset of forgettable examples only (i.e., without any fine-tuning on the whole dataset). Contrary to what was found in Toneva et al. (2019), we observe in Table 7 that the performance obtained by training only on forgettable examples is poor compared to random subsets of the same sizes; MNLI accuracy is only 37.9% for $\mathcal{F}_{\text{BERT}}$ compared to 74.7% for a random subset with the same size. The HANS accuracy is also poor. These results sug-

Model	FEVER	Symm-v1
$\text{BERT}_{\text{BASE}}$	86.1	57.7
$\text{BERT}_{\text{BASE}} + \mathcal{F}_{\text{BoW}}$	87.2	61.0
$\text{BERT}_{\text{LARGE}}$	86.9	59.7
$\text{BERT}_{\text{LARGE}} + \mathcal{F}_{\text{BoW}}$	86.5	67.8
$\text{XLNET}_{\text{BASE}}$	86.4	63.9
$\text{XLNET}_{\text{BASE}} + \mathcal{F}_{\text{BoW}}$	87.5	67.8
$\text{XLNET}_{\text{LARGE}}$	88.2	68.8
$\text{XLNET}_{\text{LARGE}} + \mathcal{F}_{\text{BoW}}$	**88.7**	**75.3**

Table 9: Average accuracy over seeds on FEVER and Fever-symm-v1 for BERT and XLNET base and large models, before and after fine-tuning on $\mathcal{F}_{\text{BoW}}$

gest an intrinsic difficulty in $\mathcal{F}_{\text{BERT}}$ that makes it hard for $\text{BERT}_{\text{BASE}}$ to generalize from it. However, as we showed previously, when starting from an already trained model, forgettables increase the out-of-distribution performance.

Calibration of models We look into the confidence of entailment when $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{BASE}} + \mathcal{F}_{\text{BoW}}$ trained on MNLI are applied to HANS. In Fig 6, we show that $\text{BERT}_{\text{BASE}}$ can discriminate HANS entailments from non-entailments but with a very large classification threshold. Fine-tuning on forgettables recalibrates the classification threshold on HANS and makes 0.5 as the optimum value.

Other diagnostic evaluations Fine-tuning on the forgettable examples of simple biased models improves robustness in the three challenging benchmarks HANS, FEVER-Symmetric and PAWS. We additionally evaluate the trained models listed in Table 4 on Stress tests (Naik et al., 2018), adversarial NLI (Nie et al., 2019) and MNLI-matched-hard (Gururangan et al., 2018). For these test sets, we do not observe improvements when evaluating the robust model using $\mathcal{F}_{\text{BoW}}$. We posit that specific biased models might be needed in some of these cases. As a validation, for MNLI-matched-hard, we design a BiLSTM model that only takes the hypothesis as input, and apply our method using the forgettables of that model to fine-tune $\text{BERT}_{\text{BASE}}$. We observe an increase in performance from 76.5% to 78.0% (averaged across five seeds). These results suggest that the forgettable examples of simple biased models like BoW or BiLSTM capture the more informative heuristics like word-overlap well. However, for less informative

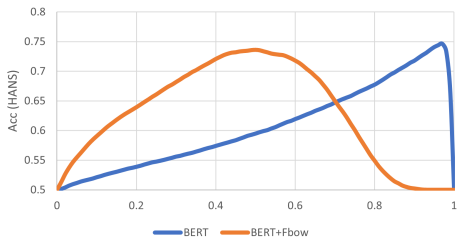

Figure 6: HANS accuracy vs classification threshold used to predict entailment/non-entailment. The base BERT model is overconfident in the entailment class while after fine-tuning on forgettables, we can improve model calibration.

heuristics like hypothesis-only features, a heuristic-designed biased model is a better choice since its forgettables likely violate the specific heuristic.

5 Related Work

A growing body of literature recently focused on **out-of-distribution generalization**, showing that it is far from being attained, even in seemingly simple cases (Geirhos et al., 2019; Jia and Liang, 2017; Dasgupta et al., 2018). In particular, and in contrast with what Mitchell et al. (2018) recommend, NLP models do not seem to "embody the symmetries that allow the same meaning be expressed within multiple grammatical structures". Supervised models seem to exhibit poor systematic generalization capabilities (Loula et al., 2018; Lake and Baroni, 2018; Baan et al., 2019; Hupkes et al., 2018) thus seemingly lacking *compositional* behavior (Montague, 1970). While this might seem at odds with the common belief that high-level semantic representations of the input data are formed (Bengio et al., 2009b), the reliance on highly predictive but brittle features is not confined to NLU tasks. It is also a perceived shortcoming of image classification models (Geirhos et al., 2019; Brendel and Bethge, 2019). To test systematically if machine learning models generalize beyond their training distribution, several challenging datasets have been introduced in NLP and other ML applications (Kalpathy-Cramer et al., 2015; Peng et al., 2019; Clark et al., 2019). Those test sets are made automatically from designed grammars (McCoy et al., 2019) and/or by human annotators (Zhang et al., 2019; Schuster et al., 2019).

Dataset re-sampling and weighting These techniques have been studied in order to solve class imbalance problem (Chawla et al., 2002) or co-

variate shift (Sugiyama et al., 2007), notably by importance weighted empirical risk minimization. In NLP, Clark et al. (2019); Mahabadi and Henderson (2019); He et al. (2019); Utama et al. (2020) give evidence of the effectiveness of re-weighting training examples to increase robustness. They generally assume *a priori* knowledge of the heuristics present in the dataset and up/down-weight examples concerning those heuristics. AFLITE (Sakaguchi et al., 2019) is an algorithmic method for bias removal in datasets without relying on prior knowledge about datasets. It filters out examples with a high average predictability score, relating them to points with biases or spurious correlations. Bras et al. (2020) adopt AFLITE to build more robust models. This method harms the in-distribution performance significantly. Comparatively, we aim to increase robustness while maintaining in-distribution performance, so we do not filter out easy examples in our approach. Our work also relates to distributionally robust optimization (Duchi and Namkoong, 2018; Hu et al., 2018) and the more recent group-DRO (Sagawa et al., 2020a), which does not assume access to target data and optimizes the worst-case performance under an unknown, bounded distribution shift. Recently, Swayamdipta et al. (2020) introduced a two-dimensional criterion to identify hard and easy examples. They consider both the confidence of an example (the average of its loss during training epochs) and the variability (those for which the loss has high variance) and show that ambiguous examples (high-variance and high-confidence) can enhance OOD accuracy. Easy examples (low-variance and high-confidence) can instead help model optimization. Although example forgetting is a coarser measure of variance of the loss, their results align with our findings: up-weighting hard/ambiguous examples enhance OOD generalization, but only training on those can harm optimization.

Curriculum Learning Dataset sampling is related to curriculum learning, where training proceeds along with a curriculum of samples with increasing difficulty (Bengio et al., 2009a). Kumar et al. (2010); Zhao and Zhang; Fan et al. (2017); Katharopoulos and Fleuret (2018); Kim and Choi (2018); Jiang et al. (2018) have shown the concept can be quite successful in a variety of areas. Our robustifying method is related to this concept. However, our models are first trained using i.i.d samples from the whole dataset and then fine-tuned on more difficult cases, i.e., the minorities.

Spurious correlations in NLU datasets like MNLI or FEVER are the subjects of many works. They include (i) the presence of specific words in the hypothesis or claim, for example, negation words like "not" are correlated with the contradiction label in entailment tasks (Naik et al., 2018; Gururangan et al., 2018), or bigrams like "did not" with the refute label, in fact, verification (Schuster et al., 2019); (ii) syntactic heuristics, like word-overlap between premise and hypothesis; and (iii) sentence length (Gururangan et al., 2018), and its correlation with labels. HANS (McCoy et al., 2019) and PAWS (Zhang et al., 2019) (which we evaluate on) generate plausible high word-overlap examples for both positive and negative classes. Glockner et al. (2018) build a new test example by simple lexical inference rules and show the brittleness of models on this out-of-distribution dataset. They also show that having supporting examples in training data is key to predict a test example correctly.

Feldman and Zhang (2020) show that when datasets are long-tailed, rare and atypical instances make up a significant fraction of the data distribution and **memorizing** them leads to better in-domain generalization. They find those rare and atypical examples using influence estimation. We instead study forgettable examples and their impact on out-of-distribution generalization. An interesting experiment would be to mine minority examples by influence estimation and compare with forgettable examples.

6 Conclusion

We introduced a novel approach, based on example forgetting, to extract minority examples and build more robust models systematically. Via example forgetting, we built a set of minority examples on which a pre-trained model is fine-tuned. We evaluated our method on large-scale models such as BERT and XLNet and showed a consistent improvement in robustness on three challenging test sets. We also showed that the larger versions obtain higher out-of-distribution performance than the base ones but still benefit from our method.

Acknowledgement

We thank Rabeeh Karimi Mahabadi, Yoshua Bengio and all the anonymous reviewers for their insightful comments.

References

Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes, and Elia Bruni. 2019. On the realization of compositionality in neural networks. *arXiv preprint arXiv:1906.01634*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009a. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Yoshua Bengio et al. 2009b. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*.

Wieland Brendel and Matthias Bethge. 2019. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

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)*.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1914–1925.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel Gershman, and Noah D. Goodman. 2018. Evaluating compositionality in sentence embeddings. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018*.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

John C. Duchi and Hongseok Namkoong. 2018. Learning models with uniform performance via distributionally robust optimization. *CoRR*, abs/1810.08750.

Yang Fan, Fei Tian, Tao Qin, Jiang Bian, and Tie-Yan Liu. 2017. Learning what data to learn. *arXiv preprint arXiv:1702.08635*.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

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)*.

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.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. 2019. Using pre-training can improve model robustness and uncertainty. 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 2712–2721. PMLR.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2034–2042.

Dieuwke Hupkes, Anand Singh, Kris Korrel, German Kruszewski, and Elia Bruni. 2018. Learning compositionally through attentive guidance. *arXiv preprint arXiv:1805.09657*.

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

R. Jia and P. Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*.

Jayashree Kalpathy-Cramer, Alba Garcia Seco de Herrera, Dina Demner-Fushman, Sameer K. Antani, Steven Bedrick, and Henning Mller. 2015. Evaluating performance of biomedical image retrieval systems - an overview of the medical image retrieval task at imageclef 2004-2013. *Comp. Med. Imag. and Graph.*, 39:55–61.

Angelos Katharopoulos and Franois Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80, pages 2530–2539.

Tae-Hoon Kim and Jonghyun Choi. 2018. Screenernet: Learning curriculum for neural networks. *CoRR*, abs/1801.00904.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-Paced Learning for Latent Variable Models. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 1–9.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888.

João Loula, Marco Baroni, and Brenden M. Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 108–114.

Rabeeh Karimi Mahabadi and James Henderson. 2019. Simple but effective techniques to reduce biases. *CoRR*, abs/1909.06321.

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*.

Jeff Mitchell, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Extrapolation in nlp. *arXiv preprint arXiv:1805.06648*.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

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*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1406–1415.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020a. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020b. An investigation of why overparameterization exacerbates spurious correlations. *arXiv preprint arXiv:2005.04345*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: an adversarial winograd schema challenge at scale. *CoRR*, abs/1907.10641.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÀžller. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005.

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.

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *CoRR*, abs/2007.06778.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 3261–3275.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37, pages 1–9.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8759–8771. Association for Computational Linguistics.

A Details of biased models (BoW and BiLSTM)

Both models are Siamese networks, with similar input representations and classification layers. For the input layer, we lower case and tokenize the inputs into words and initialize their representations with Glove, a 300-dimensional pretrained embedding (Pennington et al., 2014). For the classification task, from the premise and hypothesis vectors p and h, we build the concatenated vector $s = [p, h, |p - h|, p \odot h]$ and pass it to a 2-layer feedforward network. To compute p or h, the BoW model max-pools the bag of word embeddings, while the BiLSTM model max-pools the top-layer hidden states of a 2-layer bidirectional LSTM. The hidden size of the LSTMs is set to 200. Overall, BoW and BiLSTM contain 560K and 2M parameters, respectively.

B Hyperparameters and training time

We use a learning rate of 5e-5 for MNLI and QQP when training the PLMs on the full training and the learning rate of 1e-5 when fine-tuning on forgettables. For FEVER, we use 2e-5 and 5e-6 for the full training and the fine-tuning on forgettables, respectively.

With a 4x Tesla P100 GPU machine and batch-size 256 per GPU, one epoch of training on the full train set takes around 4-6 minutes for BOW and BiLSTM models in all of the three training tasks.

For $\text{BERT}_{\text{BASE}}$, with batch-size 32 per GPU, one epoch of training on the full train set takes around 30 / 20 / 30 minutes (per task). The maximum input length after tokenization is set to 128 in all the experiments.

B.1 Forgettables and word-overlap in MNLI

Model	Entailment			Non-Entailment		
	All	*High*	*Low*	*All*	*High*	*Low*
BERT	**84.0**	**89.9**	**76.0**	84.9	85.5	84.6
BERT + $\mathcal{F}_{\text{BOW}}$	80.2	85.1	73.4	**85.6**	86.9	**85.0**
BERT + $\mathcal{F}_{\text{BiLSTM}}$	79.9	85.2	72.4	**85.6**	**87.4**	84.8

Table 10: Fine-grained accuracy results of $\text{BERT}_{\text{BASE}}$ on the MNLI dev set split by word-overlap between hypothesis and premise.

In Table 10, we show the performance of our method on the MNLI dev set as a function of word-overlap, the main heuristic HANS was designed against. We split the evaluation set into High ($>$ mean) and Low ($<$ mean) word-overlap examples, where word-overlap is measured using the Jaccard Index between hypothesis and premise. We see in particular that entailment pairs with high word-overlap suffer from the fine-tuning on forgettables, while non-entailment improves (we observe a similar trend for QQP; see App. C). This supports the observations in 3.2 that the initial model relied on the spurious correlation of word-overlap and entailment to classify pairs and that by fine-tuning on forgettable examples, the performance on minorities increased.

C Forgettables and word-overlap in QQP

In Table 11, we show the performance of our method on the QQ evaluation set as a function of word-overlap, the main heuristic PAWS was designed against. We see in particular that paraphrase pairs with high word-overlap suffered from the fine-tuning, while non-paraphrase improved. This supports the intuition that the initial model relied on word-overlap to classify pairs as paraphrase, while forgettables help mitigate that phenomenon to some extent.

	Paraphrase			Non-Paraphrase		
Model	*All*	*High*	*Low*	*All*	*High*	*Low*
BERT	**90.0**	**90.8**	**88.9**	92.2	85.6	95.0
BERT + $\mathcal{F}_{\text{BiLSTM}}$	85.2	84.9	85.8	**93.0**	**87.3**	**95.4**
BERT + $\mathcal{F}_{\text{BoW}}$	87.3	87.2	87.4	92.6	86.4	95.2

Table 11: Fine-grained accuracy results of BERT on QQP development set before and after fine-tuning on forgettables. We split the evaluation set into High ($>$ mean) and Low ($<$ mean) word-overlap examples, where word-overlap is measured under the Jaccard Index between two sentences. Similar observations hold true in the case of MNLI.