Adversarial Filters of Dataset Biases

Ronan Le Bras [1] **Swabha Swayamdipta** [1] **Chandra Bhagavatula** [1] **Rowan Zellers** [1 2] **Matthew E. Peters** [1]
Ashish Sabharwal [1] **Yejin Choi** [1 2]

Abstract

Large neural models have demonstrated human-level performance on language and vision benchmarks, while their performance degrades considerably on adversarial or out-of-distribution samples. This raises the question of whether these models have learned to solve a *dataset* rather than the underlying *task* by overfitting to *spurious dataset biases*. We investigate one recently proposed approach, AFLITE, which adversarially filters such dataset biases, as a means to mitigate the prevalent overestimation of machine performance. We provide a theoretical understanding for AFLITE, by situating it in the generalized framework for optimum bias reduction. We present extensive supporting evidence that AFLITE is broadly applicable for reduction of measurable dataset biases, and that models trained on the filtered datasets yield better generalization to out-of-distribution tasks. Finally, filtering results in a large drop in model performance (e.g., from 92% to 62% for SNLI), while human performance still remains high. Our work thus shows that such filtered datasets can pose new research challenges for robust generalization by serving as upgraded benchmarks.

1. Introduction

Large-scale neural networks have achieved superhuman performance across many popular AI benchmarks, for tasks as diverse as image recognition (ImageNet; Russakovsky et al., 2015), natural language inference (SNLI; Bowman et al., 2015), and question answering (SQuAD; Rajpurkar et al., 2016). However, the performance of such neural models degrades considerably when tested on out-of-distribution or adversarial samples, otherwise known as data "in the wild" (Eykholt et al., 2018; Jia & Liang, 2017). This phenomenon indicates that high performance of the strongest AI models is often confined to specific *datasets*, implicitly making a closed-world assumption. In contrast, true learning of a *task* necessitates generalization, or an open-world assumption. A major impediment to generalization is the presence of spurious *biases* – unintended correlations between input and output – in existing datasets (Torralba & Efros, 2011). Such biases or *artifacts*[3] are often introduced during data collection (Fouhey et al., 2018) or during human annotation (Rudinger et al., 2017; Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; Geva et al., 2019). Not only do dataset biases inevitably bias the models trained on them, but they have also been shown to significantly inflate model performance, leading to an overestimation of the true capabilities of current AI systems (Sakaguchi et al., 2020; Hendrycks et al., 2019).

Many recent studies have investigated task or dataset specific biases, including language bias in Visual Question Answering (Goyal et al., 2017), texture bias in ImageNet (Geirhos et al., 2018), and hypothesis-only reliance in Natural Language Inference (Gururangan et al., 2018). These studies have yielded domain-specific algorithms to address the found biases. However, the vast majority of these studies follow a *top-down* framework where the bias reduction algorithms are essentially guided by researchers' intuitions and domain insights on particular types of spurious biases. While promising, such approaches are fundamentally limited by what the algorithm designers can *manually* recognize and enumerate as unwanted biases.

Our work investigates AFLITE, an alternative *bottom-up* approach to *algorithmic* bias reduction. AFLITE[4] was recently proposed by Sakaguchi et al. (2020)—albeit very succinctly—to systematically discover and filter *any* dataset artifact in crowdsourced commonsense problems. AFLITE employs a model-based approach with the goal of removing spurious artifacts in data beyond what humans can intuitively recognize, but those which are exploited by powerful models. Figure 1 illustrates how AFLITE reduces dataset biases in the ImageNet dataset for object classification.

[1]Allen Institute for Artificial Intelligence [2]Paul G. Allen School of Computer Science, University of Washington. Correspondence to: Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula <{ronanlb,swabhas,chandrab}@allenai.org>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[3]We will henceforth use *biases* and *artifacts* interchangeably.
[4]Stands for Lightweight Adversarial Filtering.

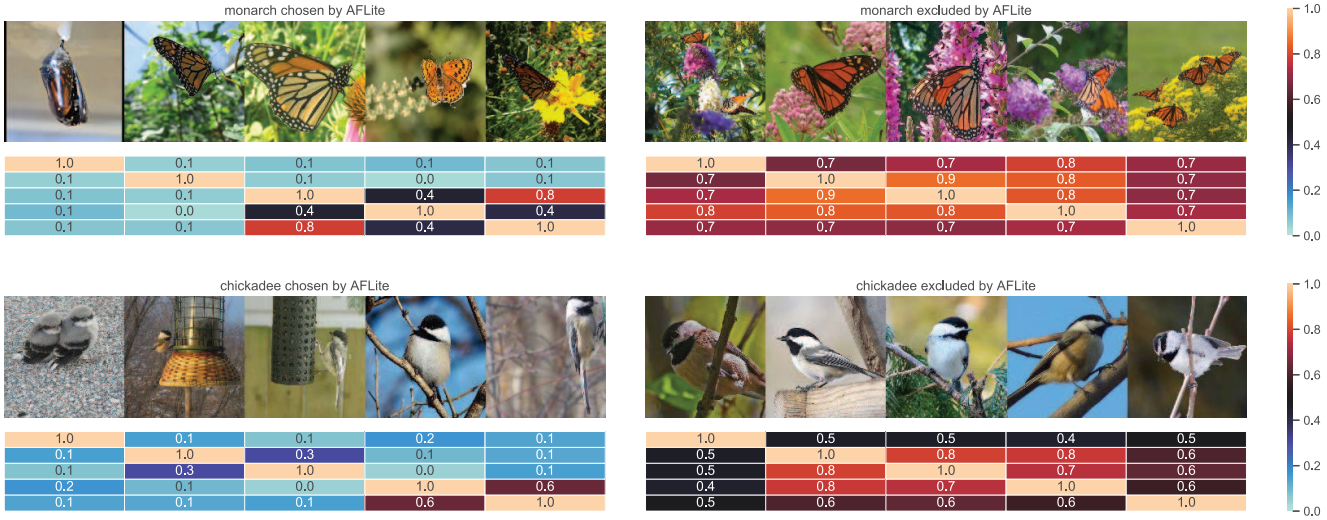

Figure 1. Example images of the Monarch Butterfly and Chickadee from ImageNet. On the left are images in each category which were retained (filtered in) by AFLITE, and on the right, the ones which were removed (filtered out). The heatmap shows pairwise cosine similarity between EfficientNet-B7 features (Tan & Le, 2019). The retained images (left) show significantly greater diversity – such as the cocoon of a butterfly, or the non-canonical chickadee poses – also reflected by the cosine similarity values. This diversity suggests that the AFLITE-filtered examples present a more accurate benchmark for the task of image classification.

This paper presents the first theoretical understanding and comprehensive empirical investigations into AFLITE. More concretely, we make the following four novel contributions.

First, we situate AFLITE in a theoretical framework for optimal bias reduction, and demonstrate that AFLITE provides a practical approximation of AFOPT, the ideal but computationally intractable bias reduction method under this framework (§2).

Second, we present an extensive suite of experiments that were lacking in the work of Sakaguchi et al. (2020), to validate whether AFLITE truly removes spurious biases in data as originally assumed. Our baselines and thorough analyses use both synthetic (thus easier to control) datasets (§3) as well as real datasets. The latter span benchmarks across NLP (§4) and vision (§5) tasks: the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets for natural language inference, QNLI (Wang et al., 2018a) for question answering, and the ImageNet dataset (Russakovsky et al., 2015) for object recognition.

Third, we demonstrate that models trained on AFLITE-filtered data generalize substantially better to out-of-domain samples, compared to models that are trained on the original biased datasets (§4, §5). These findings indicate that spurious biases in datasets make benchmarks artificially easier, as models learn to overly rely on these biases instead of learning more transferable features, thereby hurting out-of-domain generalization.

Finally, we show that AFLITE-filtering makes widely used AI benchmarks considerably more challenging. We consistently observe a significant drop in the in-domain performance even for state-of-the-art models on all benchmarks, even though human performance still remains high; this suggests that currently reported performance on benchmarks might be inflated. For instance, the best model on SNLI-AFLITE achieves only 63% accuracy, a 30% drop compared to its accuracy on the original SNLI. These findings are especially surprising since AFLITE maintains an identical train-test distribution, while retaining a sizable training set.

In summary, AFLITE-filtered datasets can serve as upgraded benchmarks, posing new research challenges for robust generalization.

2. AFLITE

Large datasets run the risk of prioritizing performance on the data-rich *head* of the distribution, where examples are plentiful, and discounting the *tail*. AFLITE seeks to minimize the ability of a model to exploit biases in the head of the distribution, while preserving the inherent complexity of the *tail*. In this section, we provide a formal framework for studying such bias reduction techniques, revealing that AFLITE can be viewed as a practical approximation of a desirable but computationally intractable optimum bias reduction objective.

Formalization Let Φ be any feature representation defined over a dataset $\mathcal{D} = (X, Y)$. AFLITE seeks a subset $S \subset \mathcal{D}, |S| \geq n$ that is maximally resilient to the features uncovered by Φ. In other words, for any identically-distributed train-test split of S, learning how to best exploit the features Φ on the training instances should not help models generalize to the held-out test set.

Let $\mathcal{M}$ denote a family of classification models (e.g., logistic regression, support vector machine classifier, or a particular neural architecture) that can be trained on subsets S of $\mathcal{D} = (X, Y)$ using features $\Phi(X)$. We define the *representation bias of* Φ *in* S *w.r.t* $\mathcal{M}$, denoted $\mathcal{R}(\Phi, S, \mathcal{M})$, as the best possible out-of-sample classification accuracy achievable by models in $\mathcal{M}$ when predicting labels Y using features $\Phi(X)$. Given a target minimum reduced dataset size n, the goal is to find a subset $S \subset \mathcal{D}$ of size at least n that minimizes this representation bias in S w.r.t. $\mathcal{M}$:

$$\underset{S \subset D,\, |S| \geq n}{\operatorname{argmin}} \; \mathcal{R}(\Phi, S, \mathcal{M}) \tag{1}$$

Eq. (1) corresponds to *optimum bias reduction*, referred to as AFOPT. We formulate $\mathcal{R}(\Phi, S, \mathcal{M})$ as the expected classification accuracy resulting from the following process. Let $q : 2^S \rightarrow [0, 1]$ be a probability distribution over subsets $T = (X^T, Y^T)$ of S. The process is to randomly choose T with probability $q(T)$, train a classifier $M_T \in \mathcal{M}$ on $S \setminus T$, and evaluate its classification accuracy $f_{M_T}\big(\Phi(X^T), Y^T\big)$ on T. The resulting accuracy on T itself is a random variable, since the training set $S \setminus T$ is randomly sampled. We define the expected value of this classification accuracy to be the representation bias:

$$\mathcal{R}(\Phi, S, \mathcal{M}) \triangleq \mathbb{E}_{T \sim q}\left[f_{M_T}\big(\Phi(X^T), Y^T\big)\right] \tag{2}$$

The expectation in Eq. (2), however, involves a summation over exponentially many choices of T even to compute the representation bias for a *single* S. This makes optimizing Eq. (1), which involves a search over S, highly intractable. To circumvent this challenge, we refactor $\mathcal{R}(\Phi, S, \mathcal{M})$ as a sum over instances $i \in S$ of the *aggregate* contribution of i to the representation bias across all T. Importantly, this summation has only $|S|$ terms, allowing more efficient computation. We call this the *predictability score* $p(i)$ for i: on average, how reliably can label y_i be predicted using features $\Phi(x_i)$ when a model from $\mathcal{M}$ is trained on a randomly chosen training set $S \setminus T$ not containing i. Instances with high predictability scores are undesirable as their feature representation can be exploited to confidently correctly predict such instances.

With some abuse of notation, for each $i \in S$, we denote $q(i) \triangleq \sum_{T \ni i} q(T)$ the marginal probability of choosing a subset T that contains i. The ratio $\frac{q(T)}{q(i)}$ is then the probability of T conditioned on it containing i. Let $f_{M_T}\big(\Phi(x_i), y_i\big)$ be the classification accuracy of M_T on i. Then the expectation in Eq. (2) can be written in terms of $p(i)$ as follows:

$$\begin{aligned}
&\sum_{T \subset S} q(T) \cdot \frac{1}{|T|} \sum_{i \in T} f_{M_T}\big(\Phi(x_i), y_i\big) \\
&= \sum_{T \subset S} \sum_{i \in T} q(T) \cdot \frac{f_{M_T}\big(\Phi(x_i), y_i\big)}{|T|} \\
&= \sum_{i \in S} \sum_{\substack{T \subset S \\ T \ni i}} q(T) \cdot \frac{f_{M_T}\big(\Phi(x_i), y_i\big)}{|T|} \\
&= \sum_{i \in S} q(i) \sum_{\substack{T \subset S \\ T \ni i}} \frac{q(T)}{q(i)} \, \frac{f_{M_T}\big(\Phi(x_i), y_i\big)}{|T|} \\
&= \sum_{i \in S} q(i)\, \mathbb{E}_{T \subset S,\, T \ni i}\left[\frac{f_{M_T}\big(\Phi(x_i), y_i\big)}{|T|}\right] \\
&= \sum_{i \in S} p(i)
\end{aligned}$$

where $p(i)$ is the predictability score of i defined as:

$$p(i) \;\triangleq\; q(i) \cdot \mathbb{E}_{T \subset S,\, T \ni i}\left[\frac{f_{M_T}\big(\Phi(x_i), y_i\big)}{|T|}\right] \tag{3}$$

While this refactoring works for any probability distribution q with non-zero support on all instances, for simplicity of exposition, we assume q to be the uniform distribution over all $T \subset S$ of a fixed size. This makes both $|T|$ and $q(i)$ fixed constants; in particular, $q(i) = \binom{|S|-1}{|T|-1} / \binom{|S|}{|T|} = \frac{|T|}{|S|}$. This yields a simplified predictability score $\tilde{p}(i)$ and a factored reformulation of the representation bias from Eq. (2):

$$\tilde{p}(i) \triangleq \frac{1}{|S|}\, \mathbb{E}_{T \subset S,\, T \ni i}\left[f_{M_T}\big(\Phi(x_i), y_i\big)\right] \tag{4}$$

$$\mathcal{R}(\Phi, S, \mathcal{M}) = \sum_{i \in S} \tilde{p}(i) \tag{5}$$

Although this refactoring reduces the exponential summation underlying the expectation in Eq. (2) to a linear sum, solving Eq. (1) for optimum bias reduction (AFOPT) remains challenging due to the exponentially many choices of S. However, the refactoring does enable computationally efficient heuristic approximations that start with $S = \mathcal{D}$ and iteratively filter out from S the most predictable instances i, as identified by the (simplified) predictability scores $\tilde{p}(i)$ computed over the current candidate for S. AFLITE adopts a greedy slicing approach. Namely, it identifies the instances with the k highest predictability scores, removes all of them from S, and repeats the process up to $\lfloor \frac{|\mathcal{D}|-n}{k} \rfloor$ times. This can be viewed as a scalable and practical approximation of (intractable) AFOPT for optimum bias reduction. In Appendix §A.1, we compare three such heuristic approaches. In all cases, we use a fixed training set size $|S \setminus T| = t < n$. Further, since a larger filtered set is generally desirable, we terminate the filtering process early (i.e., while $|S| > n$) if the predictability score for every i falls below a pre-specified early stopping threshold $\tau \in [0, 1]$.

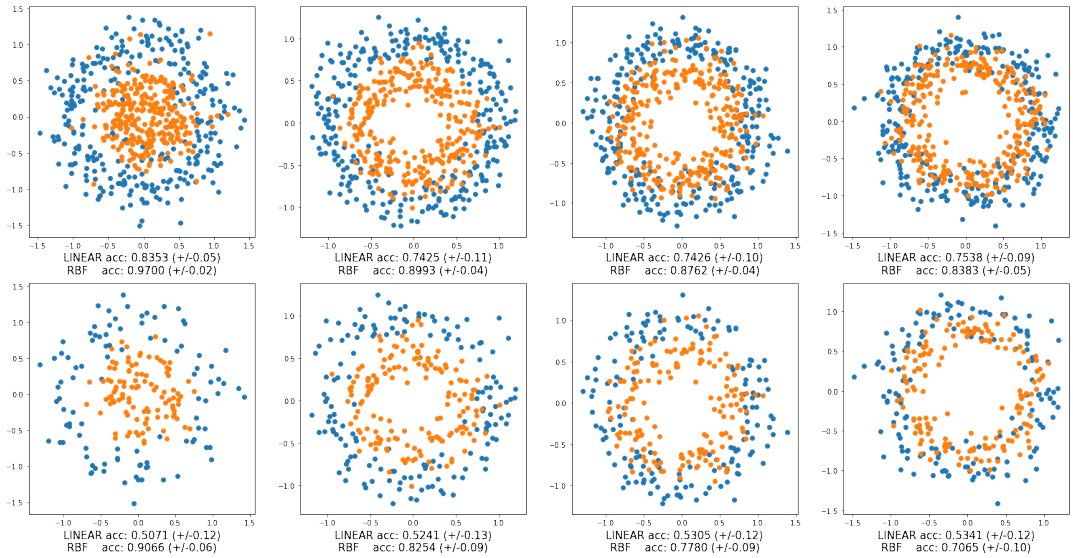

Figure 2. Four sample biased datasets as input to AFLITE (top). Blue and orange indicate two different classes. Only the original two dimensions are shown, not the bias features. For the leftmost dataset with the highest separation, we flip some labels at random, so even an RBF kernel cannot achieve perfect performance. AFLITE makes the data more challenging for the models (bottom).

Algorithm 1 AFLITE

Input: dataset $D = (X, Y)$, pre-computed representation $\Phi(X)$, model family $\mathcal{M}$, target dataset size n, number of random partitions m, training set size $t < n$, slice size $k \leq n$, early-stopping threshold τ
Output: reduced dataset S
$S = D$
while $|S| > n$ **do**
  `// Filtering phase`
  forall $i \in S$ **do**
    Initialize multiset of out-of-sample predictions $E(i) = \emptyset$
  for *iteration* $j : 1..m$ **do**
    Randomly partition S into $(T_j, S \setminus T_j)$ s.t. $|S \setminus T_j| = t$
    Train a classifier $\mathcal{L} \in \mathcal{M}$ on $\{(\Phi(x), y) \mid (x, y) \in S \setminus T_j\}$ ($\mathcal{L}$ is typically a linear classifier)
    forall $i = (x, y) \in T_j$ **do**
      Add the prediction $\mathcal{L}(\Phi(x))$ to $E(i)$
  forall $i = (x, y) \in S$ **do**
    Compute the predictability score $\tilde{p}(i) = |\{\hat{y} \in E(i) \text{ s.t. } \hat{y} = y\}| \,/\, |E(i)|$
  Select up to k instances S' in S with the highest predictability scores subject to $\tilde{p}(i) \geq \tau$
  $S = S \setminus S'$
  if $|S'| < k$ **then**
    break
return S

Implementation Algorithm 1 provides an implementation of AFLITE. The algorithm takes as input a dataset $D = (X, Y)$, a representation $\Phi(X)$ we are interested in minimizing the bias in, a model family $\mathcal{M}$ (e.g., linear classifiers), a target dataset size n, size m of the support of the expectation in Eq. (4), training set size t for the classifiers, size k of each slice, and an early-stopping filtering threshold τ. Importantly, for efficiency, $\Phi(X)$ is provided to AFLITE in the form of *pre-computed* embeddings for all of X. In practice, to obtain $\Phi(X)$, we train a first "warm-up" model on a small fraction of the data based on the learning curve in low-data regime, and do not reuse this data for the rest of our experiments. Moreover, this fraction corresponds to the training size t for AFLITE and it remains unchanged across iterations. We follow the iterative filtering approach, starting with $S = D$ and iteratively removing some instances with the highest predictability scores using the greedy slicing strategy. Slice size k and number of partitions m are determined by the available computation budget.

At each filtering phase, we train models (linear classifiers) on m different random partitions of the data, and collect their predictions on their corresponding test set. For each instance i, we compute its *predictability score* as the ratio of the number of times its label y_i is predicted correctly, over the total number of predictions for it. We rank the instances according to their predictability score and use the greedy slicing strategy of removing the top-k instances whose score is not less than the early-stopping threshold τ. We repeat this process until fewer than k instances pass the τ threshold in a filtering phase or fewer than n instances remain. Appendix §A.5 provides details of hyperparameters used across different experimental settings, to be discussed in the following sections.

3. Synthetic Data Experiments

We present experiments under a synthetic setting, to evaluate whether AFLITE successfully removes examples with

Table 1. Zero-shot SNLI accuracy on three out-of-distribution evaluation tasks, comparing *RoBERTa*-large models trained on the original SNLI data (D, size 550k), AFLITE-filtered data ($D(\phi_{RoBERTa})$, size 182k), and on a random subset with the same size as the filtered data (D_{182k}). The reported accuracy is averaged across 5 random seeds, and the subscript denotes standard deviation. On the HANS dataset, all models are evaluated on the non-entailment cases of the three syntactic heuristics (*Lexical overlap*, *Subsequence*, and *Constituent*). The NLI-Diagnostics dataset is broken down into the instances requiring world and commonsense knowledge (*Knowl.*), logical reasoning (*Logic*), predicate-argument structures (*PAS*), or lexical semantics (*LxS.*). Stress tests for NLI are further categorized into *Competence*, *Distraction* and *Noise* tests.

	HANS			NLI-Diagnostics				Stress		
	Lex.	*Subseq.*	*Constit.*	*Knowl.*	*Logic*	*PAS*	LxS.	*Comp.*	*Distr.*	*Noise*
D	$88.4_{2.2}$	$28.2_{3.4}$	$21.7_{7.1}$	$51.8_{1.6}$	$57.8_{1.7}$	$\mathbf{72.6}_{1.3}$	$65.7_{1.9}$	$77.9_{2.5}$	$\mathbf{73.5}_{2.9}$	$\mathbf{79.8}_{0.8}$
D_{182k}	$56.6_{14.7}$	$19.6_{5.6}$	$13.8_{2.9}$	$\mathbf{56.4}_{0.8}$	$53.9_{1.5}$	$71.2_{1.1}$	$65.6_{1.7}$	$68.4_{3.0}$	$73.0_{3.0}$	$78.6_{0.4}$
$D(\phi_{RoBERTa})$	$\mathbf{94.1}_{3.5}$	$\mathbf{46.3}_{6.0}$	$\mathbf{38.5}_{15.2}$	$53.9_{1.6}$	$\mathbf{58.7}_{1.2}$	$69.9_{0.9}$	$\mathbf{66.5}_{1.7}$	$\mathbf{79.1}_{1.0}$	$72.0_{1.8}$	$79.5_{0.4}$

Table 2. SNLI accuracy on Adversarial NLI using *RoBERTa*-large models pre-trained on the original SNLI data (D, size 550k) and on AFLITE-filtered data ($D(\phi_{RoBERTa})$, size 182k). Both models were finetuned on the in-distribution training data for each round (*Rd1*, *Rd2*, and *Rd3*).

	Adversarial-NLI		
	Rd1	*Rd2*	*Rd3*
D	58.5	48.3	50.1
$D(\phi_{RoBERTa})$	**65.1**	**49.1**	**52.8**

spurious correlations from a dataset. We synthesize a dataset comprising two-dimensional data, arranged in concentric circles, at four different levels of separation, as shown in Figure 2. The label (color) indicates the circular region the data point is situated in. As is evident, a linear function is inadequate for separating the two classes; it requires a more complex non-linear model such as a support vector machine (SVM) with a radial basis function (RBF) kernel.

To simulate spurious correlations in the data, we add class-specific artificially constructed features (biases) sampled from two different Gaussian distributions. These features are only added to 75% of the data in each class, while for the rest of the data, we insert random (noise) features. The bias features make the task solvable through a linear function. Furthermore, for the first dataset, with the largest separation, we flipped the labels of some biased samples, making the data slightly adversarial even to the RBF. Both models can clearly leverage the biases, and demonstrate improved performance over a baseline without biases.[5]

Once we apply AFLITE, as expected, the number of biased samples is reduced considerably, making the task hard once again for the linear model, but still solvable for the non-linear one. The filtered dataset is shown in the bottom half of Fig. 2, and the captions indicate the performance of a linear and an SVM model (detailed results for each are provided in Appendix §A.3, for better visibility). Under each separation level, our results show that AFLITE indeed removes examples with spurious correlations from a dataset. Moreover, AFLITE removes most of the flipped examples in the first dataset.

[5]We use standard implementations from scikit-learn: `https://scikit-learn.org/stable/`.

4. NLP Experiments

As our first real-world data evaluation for AFLITE, we consider out-of-domain and in-domain generalization for a variety of language datasets. The primary task we consider is natural language inference (NLI) on the Stanford NLI dataset (Bowman et al., 2015, SNLI). Each instance in the NLI task consists of a premise-hypothesis sentence pair, the task involves predicting whether the hypothesis either *entails*, *contradicts* or is *neutral* to the premise.

Experimental Setup We use feature representations from *RoBERTa*-large, $\phi_{RoBERTa}$ (Liu et al., 2019b), a large-scale pretrained masked language model. This is extracted from the final layer before the output layer, trained on a random 10% sample (warm-up) of the original training set. The resultant filtered NLI dataset, $D(\phi_{RoBERTa})$, is compared to the original dataset D as well as a randomly subsampled dataset D_{182k}, with the same sample size as $D(\phi_{RoBERTa})$, amounting to only a third of the full data D. The same *RoBERTa*-large architecture is used to train the three NLI models.

4.1. Out-of-distribution Generalization

As motivated in Section §1, large-scale architectures often learn to solve datasets rather than the underlying task by overfitting on unintended correlations between input and output in the data. However, this reliance might be hurtful for generalization to out-of-distribution examples, since they may not contain the *same* biases. We evaluate AFLITE for

Table 3. Dev accuracy (%) on the original SNLI dataset D and the datasets obtained through different AFLITE-filtering and other baselines. D_{92k} indicates a randomly subsampled train dataset of the same size as $D(\phi_{RoBERTa})$. Δ indicates the difference in performance (or size, last row) between the full model and the model trained on $D(\phi_{RoBERTa})$.

	Train Data					
Model	D	D_{92k}	$D(\Phi_{ESIM+GLoVe})$	$D(\phi_{BERT})$	$D(\phi_{RoBERTa})$	Δ
ESIM+ELMo (Peters et al., 2018)	88.7	86.0	61.5	54.2	51.9	-36.8
BERT (Devlin et al., 2019)	91.3	87.6	74.7	61.8	57.0	-34.3
RoBERTa (Liu et al., 2019b)	92.6	88.3	78.9	71.4	62.6	-30.0
Max-PPMI	54.5	52.0	41.1	41.5	41.9	-12.6
BERT -HypOnly	71.5	70.1	52.3	46.4	48.4	-23.1
RoBERTa -HypOnly	72.0	70.4	53.6	49.5	48.5	-23.5
Human performance	88.1	88.1	82.3	80.3	77.8	-10.3
Training set size	550k	92k	138k	109k	92k	-458k

this criterion on the NLI task.

Gururangan et al. (2018), among others, showed the existence of certain annotation artifacts (lexical associations etc.) in SNLI which make the task considerably easier for most current methods. This spurred the development of several out-of-distribution test sets which carefully control for the presence of said artifacts. We evaluate on four such out-of-distribution datasets: HANS (McCoy et al., 2019b), NLI Diagnostics (Wang et al., 2018a), Stress tests (Naik et al., 2018) and Adversarial NLI (Nie et al., 2019) (c.f. Appendix §A.4 for details). Given that these benchmarks are collected independently of the original SNLI task, the biases from SNLI are less likely to carry over.[6]

Table 1 shows results on three out of four diagnostic datasets (HANS, NLI-Diagnostics and Stress), where we perform a zero-shot evaluation of the models. Models trained on SNLI-AFLITE consistently exceed or match the performance of the full model on the benchmarks above, up to standard deviation. To control for the size, we compare to a baseline trained on a random subsample of the same size (D_{182k}). AFLITE models report higher generalization performance suggesting that the filtered samples are more informative than a random subset. In particular, AFLITE substantially outperforms challenging examples on the HANS benchmark, which targets models purely relying on lexical and syntactic cues. Table 2 shows results on the Adversarial NLI benchmark, which allows for evaluation of transfer capabilities, by finetuning models on each of the three training datasets (*Rd1*, *Rd2* and *Rd3*). A *RoBERTa*-large model trained on SNLI-AFLITE surpasses the performance in all three settings.

[6]However, these benchmarks might contain their own biases (Liu et al., 2019a).

4.2. In-distribution Benchmark Re-estimation

AFLITE additionally provides a more accurate estimation of the benchmark performance on several tasks. Here we simply lower the AFLITE early-stopping threshold, τ in order to filter most biased examples from the data, resulting in a stricter benchmark with 92k train samples.

SNLI In addition to *RoBERTa*-large, we consider here pre-computed embeddings from *BERT*-large (Devlin et al., 2019), and *GloVe* (Pennington et al., 2014), resulting in three different feature representations for SNLI: ϕ_{BERT}, $\phi_{RoBERTa}$ from *RoBERTa*-large (Liu et al., 2019b), and $\Phi_{ESIM+GLoVe}$ which uses the ESIM model (Chen et al., 2016) with *GloVe* embeddings. Table 3 shows the results for SNLI. In all cases, applying AFLITE substantially reduces overall model accuracy, with typical drops of 15-35% depending on the models used for learning the feature representations and those used for evaluation of the filtered dataset. In general, performance is lowest when using the strongest model (*RoBERTa*) for learning feature representations. Results also highlight the ability of weaker adversaries to produce datasets that are still challenging for much stronger models with a drop of 13.7% for *RoBERTa* using $\Phi_{ESIM+GLoVe}$ as feature representation.

To control for the reduction in dataset size by filtering, we randomly subsample D, creating D_{92k} whose size is approximately equal to that of $D(\phi_{RoBERTa})$. All models achieve nearly the same performance as their performance on the full dataset – even when trained on just one-fifth the original data. This result further highlights that current benchmark datasets contain significant redundancy within its instances.

We also include two other baselines, which target known dataset artifacts in NLI. The first baseline uses Point-wise Mutual Information (PMI) between words in a given instance and the target label as its only feature. Hence it captures the extent to which datasets exhibit word-association

Table 4. Dev accuracy (%) on the original (D) and AFLITE-filtered ($D(\phi_{RoBERTa})$) MultiNLI-matched and QNLI datasets. The *-PartialInput* baselines show models trained on only *Hypotheses* for MultiNLI instances and only *Answers* for QNLI. Δ indicates the difference in accuracy of the full model and the filtered model.

		Train Data		
Task	Model	D	$D(\phi_{RoBERTa})$	Δ
QNLI	*BERT*	86.6	55.8	-30.8
	RoBERTa	90.3	66.2	-24.1
	BERT-PartialInput	59.7	43.2	-16.5
	RoBERTa-PartialInput	60.3	44.4	-15.9
Multi-NLI	*BERT*	92.0	63.5	-28.5
	RoBERTa	93.7	77.7	-16.0
	BERT-PartialInput	62.6	56.6	-6.0
	RoBERTa-PartialInput	63.9	59.4	-4.5

biases, one particular class of spurious correlations. While this baseline is relatively weaker than other models, its performance still reduces by nearly 13% on the $D(\phi_{RoBERTa})$ dataset. The second baseline trains only the hypothesis of an NLI instance (*-HypOnly*). Such partial input baselines (Gururangan et al., 2018) capture reliance on lexical cues only in the hypothesis, instead of learning a semantic relationship between the hypothesis and premise. This reduces performance by almost 24% before and after filtering with *RoBERTa*. AFLITE, which is agnostic to any particular known bias in the data, results in a drop of about 30% on the same dataset, indicating that it might be capturing a larger class of spurious biases than either of the above baselines.

Finally, to demonstrate the value of the iterative, ensemble-based AFLITE algorithm, we compare with a baseline where using a single model, we filter out the most predictable examples in a single iteration — a non-iterative, single-model version of AFLITE. A *RoBERTa*-large model trained on this subset (of the same size as $D(\phi_{RoBERTa})$) achieves a dev accuracy of 72.1%. Compared to the performance of *RoBERTa* on $D(\phi_{RoBERTa})$ (62.6%, see Table 3), it makes this baseline a sensible yet less effective approach. In particular, this illustrates the need for an iterative procedure involving models trained on multiple partitions of the remaining data in each iteration.

MultiNLI and QNLI We evaluate the performance of another large-scale NLI dataset multi-genre NLI (Williams et al., 2018, MultiNLI), and the QNLI dataset (Wang et al., 2018a) which is a sentence-pair classification version of the SQuAD (Rajpurkar et al., 2016) question answering task.[7] Results before and after AFLITE are reported in Table 4. Since *RoBERTa* resulted in the largest drops in performance across the board in SNLI, we only experiment with *RoBERTa* as adversary for MultiNLI and QNLI. While *RoBERTa* achieves over 90% on both original datasets, its performance drops to 66.2% for MultiNLI and to 77.7% for QNLI on the filtered datasets. Similarly, partial input baseline performance also decreases substantially on both dataset compared to their performance on the original dataset. Overall, our experiments indicate that AFLITE consistently results in reduced accuracy on the filtered datasets across multiple language benchmark datasets, even after controlling for the size of the training set.

Table 3 shows that human performance on SNLI-AFLITE is lower than that on the full SNLI.[8] This indicates that the filtered dataset is somewhat harder even for humans, though to a much lesser degree than any model. Indeed, removal of examples with spurious correlations could inadvertently lead to removal of genuinely easy examples; this might be a limitation of a model-based bias reduction approach such as AFLITE (see Appendix §A.8 for a qualitative analysis). Future directions for bias reduction techniques might involve additionally accounting for unaltered human performance before and after dataset reduction.

5. Vision Experiments

We evaluate AFLITE on image classification through ImageNet (ILSVRC2012) classification. On ImageNet, we use the state-of-the-art EfficientNet-B7 model (Tan & Le, 2019) as our core feature extractor $\Phi_{\text{EN-B7}}$. The EfficientNet model is learned from scratch on a fixed 20% sample of the ImageNet training set, using RandAugment data augmentation (Cubuk et al., 2019). We then use the 2560-dimensional features extracted by EfficientNet-B7 as the underlying representation for AFLITE to use to filter the remaining dataset, and stop when data size is 40% of ImageNet.

Adversarial Image Classification In Table 5, we report performance of image classification models on ImageNet-A, a dataset with out-of-distribution images (Hendrycks et al., 2019). As shown, all EfficientNet models struggle on this task, even when trained on the entire ImageNet.[9]

[7] QNLI is stylized as an NLI classification task, where the task is to determine whether or not a sentence contains the answer to a question.

[8] Measured based on five annotator labels provided in the original SNLI validation data.

[9] Notably, there is a large difference in the degree of out-of-distribution generalization performance for NLP and vision tasks. NLP tasks benefit from the availability of pretrained representations from large language models, such as *RoBERTa*. In vision, however, while (pre)training on ImageNet alone is often sufficient to learn competitive features, such strong pretrained representations are not available. Moreover, ImageNet has many classes and a skewed distribution of data (Vodrahalli et al., 2018). Hence, it is considerably harder to find a smaller subset of data which generalizes well to adversarial challenge sets, such as ImageNet-A.

Table 5. Top-1 accuracy on ImageNet-A (Hendrycks et al., 2019), an adversarial evaluation set for image classification. The most powerful model EfficientNet-B7 improves by 2% on out-of-distribution ImageNet-A images when trained on AFLITE-filtered data $D(\Phi_{\text{EN-B7}})$.

	Model	
Train Data	EfficientNet-B5	EfficientNet-B7
D	16.5	20.6
$D_{40\%}$	5.9	8.5
$D(\Phi_{\text{EN-B7}})$	7.2	10.4

However, we find that training on AFLITE-filtered data leads to models with greater generalization, in comparison to training on a randomly sampled ImageNet of the same size, leading to up to 2% improvement in performance.

In-distribution Image Classification In Table 6, we present ImageNet accuracy across the EfficientNet and ResNet (He et al., 2016) model families before and after filtering with AFLITE. For evaluation, the Imagenet-AFLITE filtered validation set is much harder than the standard validation set (also see Figure 1). While the top performer after filtering is still EfficientNet-B7, its top-1 accuracy drops from 84.4% to 63.5%. A model trained on a randomly filtered subsample of the same size though suffers much less, most likely due to reduction in training data.

Overall, these results suggest that image classification – even within a subset of the closed world of ImageNet – is far from solved. These results echo other findings that suggest that common biases that naturally occur in web-scale image data, such as towards canonical poses (Alcorn et al., 2019) or towards texture rather than shape (Geirhos et al., 2018), are problems for ImageNet-trained classifiers.

6. Related Work

Adversarial Filtering AFLITE is related to Zellers et al. (2018)'s adversarial filtering (AF) algorithm, yet distinct in two key ways: it is (i) much more broadly applicable (by not requiring over generation of data instances), and (ii) considerably more lightweight (by not requiring re-training a model at each iteration of AF). Variants of this AF approach have recently been used to create other datasets such as HellaSwag (Zellers et al., 2019) and Abductive NLI (Bhagavatula et al., 2019) by iteratively perturbing dataset instances until a target model cannot fit the resulting dataset. While effective, these approaches run into three main pitfalls. First, dataset curators need to explicitly devise a strategy of collecting or generating perturbations of a given instance. Second, the approach runs the risk of distributional bias where a discriminator can learn to distinguish between machine generated instances and human-generated ones. Finally it requires re-training a model at each iteration, which is computationally expensive especially when using a large model such as *BERT* as the adversary. In contrast, AFLITE focuses on addressing dataset biases from existing datasets instead of adversarially perturbing instances. AFLITE was earlier proposed by Sakaguchi et al. (2020) to create the Winogrande dataset. This paper presents more thorough experiments, theoretical justification and results from generalizing the proposed approach to multiple popular NLP and Vision datasets.

Table 6. Results on ImageNet, in Top-1 accuracy (%). We consider training on the 40% challenging instances, as filtered by AFLITE ($D(\Phi_{\text{EN-B7}})$), and compare this to a random 40% subsample of ImageNet ($D_{40\%}$). We report results on the ImageNet validation set before and after filtering with AFLITE. Δ indicates the difference in accuracy when trained on the full data and the filtered data. Notably, evaluating on AFLITE-filtered ImageNet is much harder—resulting in a drop of nearly 21 percentage points in accuracy for the strongest model.

	Train Data			
Model	D	$D_{40\%}$	$D(\Phi_{\text{EN-B7}})$	Δ
EfficientNet-B0	76.3	69.6	50.2	-26.1
EfficientNet-B3	81.7	75.1	57.3	-24.4
EfficientNet-B5	83.7	78.6	62.2	-21.5
EfficientNet-B7	84.4	78.8	63.5	-20.9
ResNet-34	78.4	65.9	46.9	-31.5
ResNet-50	79.2	68.9	50.1	-29.1
ResNet-101	80.1	70.1	52.2	-27.9
ResNet-152	80.6	71.0	53.3	-27.3

Data Selection for Debiased Representations Li & Vasconcelos (2019) recently proposed REPAIR, a method to remove representation bias by dataset resampling. The motivation in REPAIR is to learn a probability distribution over the dataset that favors instances that are hard for a given representation. In contrast to AFLITE, the implementation of REPAIR relies on in-training classification loss as opposed to out-of-sample generalization accuracy. RESOUND (Li et al., 2018) quantifies the representation biases of datasets, and uses them to assemble a new K-class dataset with smaller biases by sampling an existing C-class dataset ($C > K$). Dataset distillation (Wang et al., 2018b) optimizes for a different objective function compared to AFLITE: it aims to synthesize a small number of instances to approximate the model trained on the original data. Dasgupta et al. (2018) introduce an NLI dataset that cannot be solved using only word-level knowledge and requires some compositionality. The authors show that debiasing training corpora and augmenting them with minimal contrasting examples makes models more suited to learn the compositional structure of language. Finally, Sagawa et al. (2020)

analyze the tension between over-parameterization and using all the data available. It advocates for subsampling the majority groups as opposed to upweighting minority groups in order to achieve low worst-group error. This is in line with the filtering approach that AFLITE adapts, as well as the out-of-distribution and robustness results we observe.

Learning Objectives for Debiasing Another line of related work focuses on removing bias in data representations via the design of learning objectives for debiasing. Arjovsky et al. (2019) propose Invariant Risk Minimization as an objective that promotes learning representations of the data that are stable across environments. Instead of learning optimal classifiers, AFLITE aims to remove instances that exhibit artifacts in a dataset. Belinkov et al. (2019) propose an adversarial removal technique that encourages models to learn representations free of hypothesis-only biases. He et al. (2019) propose DRiFt, a debiasing algorithm that first learns a biased model using only known biased features and then trains a debiased model that fits the residuals of the biased model. Similarly, Clark et al. (2019) propose learning a naive classifier using only bias features, to be used in an ensemble along with other classifiers containing more general features. Each of the previous approaches target only *known* NLI biases, based on prior knowledge; we show AFLITE is capable of removing even those examples which exhibit previously *unknown* spurious biases. Finally, Elazar & Goldberg (2018) show that adversarial training effectively mitigate demographic information leakage, but fail to remove it completely when dealing with text data.

7. Conclusion

We present a deep-dive into AFLITE – an iterative greedy algorithm that adversarially filters out spurious biases from data for accurate benchmark estimation. We provide a theoretical framework supporting AFLITE, and show its effectiveness in bias reduction on synthetic and real datasets, providing extensive analyses. We apply AFLITE to four datasets, including widely used benchmarks such as SNLI and ImageNet. On out-of-distribution and adversarial test sets designed for such benchmarks, we show that models trained on the AFLITE-filtered subsets achieve better performance, indicating higher generalization abilities. Moreover, we show that the strongest performance on the resulting filtered datasets drops significantly (by 30 points for SNLI and 20 points for ImageNet). We hope that dataset creators will employ AFLITE to identify unknown dataset artifacts before releasing new challenge datasets for more reliable estimates of task progress on future AI benchmarks. All datasets and code for this work will be made public.

Acknowledgments

We would like to thank Noah A. Smith, Nicholas Lourie, Ana Marasovic̀ and Daniel Khashabi for insightful discussions about this work as well as the anonymous reviewers for their valuable feedback. This research was supported in part by NSF (IIS-1524371), the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256082, DARPA CwC through ARO (W911NF15-1- 0543), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI. Computations on `beaker.org` were supported in part by credits from Google Cloud.

References

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *CVPR*, 2019.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2019. URL `https://arxiv.org/abs/1907.02893`. ArXiv:1907.02893.

Balog, M., Tripuraneni, N., Ghahramani, Z., and Weller, A. Lost relatives of the Gumbel trick. In *ICML*, 2017.

Belinkov, Y., Poliak, A., Shieber, S. M., Durme, B. V., and Rush, A. M. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *ACL*, 2019.

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., tau Yih, S. W., and Choi, Y. Abductive commonsense reasoning. In *ICLR*, 2019. URL `https://arxiv.org/abs/1908.05739`.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015. URL `https://www.aclweb.org/anthology/D15-1075`.

Chen, Q., Zhu, X.-D., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. Enhanced LSTM for natural language inference. In *ACL*, 2016. URL `https://www.aclweb.org/anthology/P17-1152`.

Clark, C., Yatskar, M., and Zettlemoyer, L. 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)*, pp.

4069–4082, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1418. URL https://www.aclweb.org/anthology/D19-1418.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., and Goodman, N. D. Evaluating compositionality in sentence embeddings. *CogSci*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Elazar, Y. and Goldberg, Y. Adversarial removal of demographic attributes from text data. In *EMNLP*, 2018.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. X. Robust physical-world attacks on deep learning models. In *CVPR*, 2018.

Fouhey, D. F., Kuo, W.-c., Efros, A. A., and Malik, J. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2018.

Geva, M., Goldberg, Y., and Berant, J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *EMNLP*, 2019. URL https://www.aclweb.org/anthology/D19-1107.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Gumbel, E. J. and Lieblein, J. Statistical theory of extreme values and some practical applications: A series of lectures. In *Applied Mathematics Series*, volume 33. National Bureau of Standards, USA, 1954.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. Annotation artifacts in natural language inference data. In *NAACL*, 2018. URL https://www.aclweb.org/anthology/N18-2017/.

He, H., Zha, S., and Wang, H. Unlearn dataset bias in natural language inference by fitting the residual. *ArXiv*, 2019. URL https://arxiv.org/abs/1908.10763.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2016.

Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, 2017. URL https://www.aclweb.org/anthology/D17-1215.

Kim, C., Sabharwal, A., and Ermon, S. Exact sampling with integer linear programs and random perturbations. In *AAAI*, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980. arXiv:1412.6980.

Kool, W., van Hoof, H., and Welling, M. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *ICML*, 2019.

Li, Y., Li, Y., and Vasconcelos, N. RESOUND: Towards action recognition without representation bias. In *ECCV*, 2018.

Li, Y. C. and Vasconcelos, N. REPAIR: Removing representation bias by dataset resampling. In *CVPR*, 2019.

Liu, N. F., Schwartz, R., and Smith, N. A. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *NAACL*, 2019a. doi: 10.18653/v1/N19-1225. URL https://www.aclweb.org/anthology/N19-1225.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M. S., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. S., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach, 2019b. URL https://arxiv.org/abs/1907.11692. ArXiv:1907.11692.

Maddison, C. J., Tarlow, D., and Minka, T. A* sampling. In *NeurIPS*, 2014.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2016.

McCoy, R. T., Min, J., and Linzen, T. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *ArXiv*, abs/1911.02969, 2019a.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*, 2019b. doi: 10.18653/v1/P19-1334. URL https://www.aclweb.org/anthology/P19-1334.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., and Neubig, G. Stress test evaluation for natural language inference. In *ICCL*, 2018. URL https://www.aclweb.org/anthology/C18-1198.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial NLI: A new benchmark for natural language understanding, 2019. URL https://arxiv.org/abs/1910.14599. arXiv:1910.14599.

Pennington, J., Socher, R., and Manning, C. D. GloVe: Global vectors for word representation. In *EMNLP*, 2014. URL https://www.aclweb.org/anthology/D14-1162.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. S. Deep contextualized word representations. In *NAACL*, 2018. URL https://www.aclweb.org/anthology/N18-1202.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. Hypothesis only baselines in natural language inference. In **SEM*, 2018. URL https://www.aclweb.org/anthology/S18-2023.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2016. URL https://www.aclweb.org/anthology/D16-1264.

Rudinger, R., May, C., and Durme, B. V. Social bias in elicited natural language inferences. In *EthNLP@EACL*, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. doi: 10.1007/s11263-015-0816-y.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. *ArXiv*, abs/2005.04345, 2020.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. WINOGRANDE: An adversarial winograd schema challenge at scale. In *AAAI*, 2020. URL https://arxiv.org/abs/1907.10641.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. *CVPR*, 2011.

Tsuchiya, M. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *LREC*, 2018.

Vieira, T. Gumbel-max trick and weighted reservoir sampling, 2014. URL https://bit.ly/310I39S.

Vodrahalli, K., Li, K., and Malik, J. Are all training examples created equal? an empirical study. *arXiv preprint arXiv:1811.12569*, 2018.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2018a. URL https://arxiv.org/abs/1804.07461.

Wang, T., Zhu, J., Torralba, A., and Efros, A. A. Dataset distillation, 2018b. URL http://arxiv.org/abs/1811.10959. arXiv:1811.10959.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, 2018. doi: 10.18653/v1/N18-1101. URL https://www.aclweb.org/anthology/N18-1101.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*, 2018. URL https://www.aclweb.org/anthology/D18-1009.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In *ACL*, 2019. URL https://www.aclweb.org/anthology/P19-1472.

A. Appendix

A.1. Filtering Heuristics

We present three heuristic approaches that approximate the optimum bias reduction problem (AFOPT): **(A)** A simple *greedy approach* starts with the full set $S = \mathcal{D}$, identifies an $i \in S$ that maximizes $\tilde{p}(i)$, removes it from S, and repeats up to $|\mathcal{D}| - n$ times. **(B)** A *greedy slicing approach* identifies the instances with the k highest predictability scores, removes all of them from S, and repeats the process up to $\lfloor \frac{|\mathcal{D}|-n}{k} \rfloor$ times. **(C)** A *slice sampling approach*, instead of greedily choosing the top k instances, randomly samples k instances with probabilities proportional to their predictability scores (cf. Appendix §A.2 for more details).

All three strategies could be further improved by considering not only the predictability score of the top-k instances but also (via retraining without these instances) how their removal would influence the predictability scores of other instances in the next step. We found our computationally lighter approaches to work well even without the additional overhead of such look-ahead. AFLITE implements the greedy slicing approach, and can thus be viewed as a scalable and practical approximation of (intractable) AFOPT for optimum bias reduction. We leave the empirical investigation into other proposed strategies for future work.

A.2. Slice Sampling Details

As discussed in Appendix §A.1 (**C**), the *slice sampling approach* can be efficiently implemented using what is known as the Gumbel method or Gumbel trick (Gumbel & Lieblein, 1954; Maddison et al., 2014), which uses random perturbations to turn sampling into a simpler problem of optimization. This has recently found success in several probabilistic inference applications (Kim et al., 2016; Jang et al., 2016; Maddison et al., 2016; Balog et al., 2017; Kool et al., 2019). Starting with the log-predictability scores $\log \tilde{p}(i)$ for various i, the idea is to perturb them by adding an independent random noise γ_i drawn from the standard Gumbel distribution. Interestingly, the maximizer i^* of $\gamma_i + \log \tilde{p}(i)$ turns out to be an exact sample drawn from the (unnormalized) distribution defined by $\tilde{p}$. Note that i^* is a random variable since the γ_i are drawn at random. This result can be generalized (Vieira, 2014) for slice sampling: the k highest values of Gumbel-perturbed log-predictability scores correspond to sampling, without replacement, k items from the probability distribution defined by $\tilde{p}$. The Gumbel method is typically applied to exponentially large combinatorial spaces, where it is challenging to scale up. In our setting, however, the overhead is minimal since the cost of drawing a random γ_i is negligible compared to computing $\tilde{p}(i)$.

Table 7. Mean *Dev* accuracy (%) on two models trained on four synthetic datasets before (D) and after ($D(\Phi)$) AFLITE. Standard deviation across 10 runs with randomly chosen seeds is provided as a subscript. The datasets, also shown in Fig. 2 differ in the degree of separation between the two classes. Both models (SVM with an RBF kernel & linear classifier with logisitic regression) perform well on the original synthetic dataset, before filtering. The linear classifier performs well on the data, because it contains spurious artifacts, making the task artificially easier for it. However, after AFLITE, the linear model, relying mostly on the spurious features, clearly underperforms.

Class Separation	Model	D	$D(\Phi)$
0.8	SVM-RBF	97.0_{02}	90.7_{06}
	Logistic Reg.	83.5_{05}	50.7_{12}
0.7	SVM-RBF	89.9_{04}	82.5_{09}
	Logistic Reg.	74.3_{11}	52.4_{13}
0.6	SVM-RBF	87.6_{04}	77.8_{09}
	Logistic Reg.	74.3_{10}	53.1_{12}
0.4	SVM-RBF	83.8_{05}	70.7_{10}
	Logistic Reg.	75.4_{09}	53.4_{12}

A.3. Results on Synthetic Data Experiments

As discussed in Section §3, Figure 2 shows the effect of AFLITE on four synthetic datasets containing data arranged in concentric circles at four degrees of class separation. For greater visibility, we have provided the accuracies of the SVM with RBF kernel and logistic regression in Table 7.

In summary, a stronger model such as the SVM is more robust to the presence of artifacts than a simple linear classifier. Thus, the implications for real datasets is to move towards models designed for reasoning about a specific task, hence avoiding a dependence on spurious artifacts.

A.4. NLI Out-of-distribution Benchmarks

We describe the four out-of-distribution evaluation benchmarks for NLI from Section §4.1 below:

- HANS (McCoy et al., 2019b) contains evaluation examples designed to avoid common structural heuristics (such as word overlap) which could be used by models to correctly predict NLI inputs, without true inferential reasoning.
- NLI Diagnostics (Wang et al., 2018a) is a set of handcrafted examples designed to demonstrate model performance on several fine-grained semantic categories, such as logical reasoning and commonsense knowledge.
- Stress tests for NLI (Naik et al., 2018) are a collection of tests targeting the weaknesses of strong NLI models, to check if these are robust to semantics (competence),

Table 8. Hyperparameters for the AFLITE algorithm, used for in-distribution benchmark estimation on different datasets. m denotes the size of the support of the expectation in Eq. (4), t is the training set size for the linear classifiers, k is the size of each slice, and τ is an early-stopping filtering threshold. For ImageNet, we set $n = 640K$ and hence do not need to control for τ. In every other setting, we set τ as above, and hence do not need to control for n. Detailed definitions for each hyperparameter is provided in Section §2.

	Synthetic	SNLI	MultiNLI	QNLI	ImageNet
m	128	64	64	64	64
t	100	50K	40K	10K	32.7K
k	1	10K	10K	2K	33.6K
τ	0.75	0.75	0.75	0.75	-

irrelevance (distraction) and typos (noise).

- Adversarial NLI (Nie et al., 2019) consists of premises collected from Wikipedia and other news corpora, and human generated hypotheses, arranged at different tiers of the challenge they present to a model, using a human and model in-the-loop procedure.

Recent work (McCoy et al., 2019a) has observed large variance on out-of-distribution test sets with random seeds. Hence, we report the mean and variance across 5 random seeds in all settings in Table 1. Since Adversarial NLI involves finetuning the model, and not just reporting on a different test set, we skip this step in Table 2.

A.5. Hyperparameters for AFLITE

Table 8 shows hyperparameters used to run AFLITE to obtain filtered subsets for in-distribution benchmark estimation on different datasets. Target dataset size, n and the early stop filtering threshold τ are interdependent, as the predictability score threshold determines what examples to keep, which in turn influences the desired size of the dataset, n. For ImageNet, we set $n = 640K$ and do not control for τ. We use much larger values for t and k for ImageNet than in all NLP experiments, where the use of powerful language representations (such as *RoBERTa*) allows us to get reasonable performance even with smaller training sets; ImageNet does not offer any such benefits arising from pretrained representations.

For all out-of-distribution NLP experiments, we explicitly control for the size of n, as discussed in the corresponding sections in the paper. In these cases, we typically end up using slightly larger n, allowing for the final models to get more exposure to task data which is, to a degree, helpful for out-of-distribution generalization. In ImageNet, we use the same hyperparameters in both sets of experiments. In particular, we explicitly set $n = 182K$ for SNLI, and $n = 640K$ for ImageNet AFLITE-filtering for the out-of-distribution generalization experiments.

A.6. Hyperparameters for NLP experiments

For all NLP experiments, our implementation is based on the GLUE (Wang et al., 2018a) experiments in the Transformers repository (Wolf et al., 2019) from Huggingface.[10] We used the Adam optimizer (Kingma & Ba, 2014) for every training set up, with a learning rate of 1e-5, and an epsilon value of 1e-8. We trained for 3 epochs for all *NLI tasks, maintaining a batch size of 92. All above hyperparameters were selected using a grid search; we kept other hyperparameters unaltered from the original HuggingFace repository. Each experiment was performed on a single Quadro RTX 8000 GPU.

A.7. Hyperparameters for ImageNet

We trained our ImageNet models using v3-512 TPU pods. For EfficientNet (Tan & Le, 2019), we used RandAugment data augmentation (Cubuk et al., 2019) with 2 layers, and a magnitude of 28, for all model sizes. We trained our models using a batch size of 4096, a learning rate of 0.128, and kept other hyperparameters the same as in (Tan & Le, 2019). We trained for 350 epochs for all dataset sizes - so when training on 20% or 40% of ImageNet (or a smaller dataset), we scaled the number of optimization steps accordingly. For ResNet (He et al., 2016), we used a learning rate of 0.1, a batch size of 8192, and trained for 90 epochs.

A.8. Qualitative Analysis of SNLI

Table 9 shows some examples removed and retained by AFLITE on the NLI dataset.

[10] https://github.com/huggingface/transformers

Table 9. Examples from SNLI, removed (top) and retained (bottom) by AFLITE. As is evident, the retained instances are slightly more challenging and capture more nuanced semantics in contrast to the removed instances. Removed instances also exhibit larger word overlap, and many other artifacts found in Gururangan et al. (2018). Two examples per label are shown, the AFLITE-filtered dataset contains many more `neutral` examples, as opposed to those labeled as `contradiction`.

REMOVED BY AFLITE		
Premise	**Hypothesis**	**Label**
A woman, in a green shirt, preparing to run on a treadmill.	A woman is preparing to sleep on a treadmill.	`contradiction`
The dog is catching a treat.	The cat is not catching a treat.	`contradiction`
Three young men are watching a tennis match on a large screen outdoors.	Three young men watching a tennis match on a screen outdoors, because their brother is playing.	`neutral`
A girl dressed in a pink shirt, jeans, and flip-flops sitting down playing with a lollipop machine.	A funny person in a shirt.	`neutral`
A man in a green apron smiles behind a food stand.	A man smiles.	`entailment`
A little girl with a hat sits between a woman's feet in the sand in front of a pair of colorful tents.	The girl is wearing a hat.	`entailment`
RETAINED BY AFLITE		
Premise	**Hypothesis**	**Label**
People are throwing tomatoes at each other.	The people are having a food fight.	`entailment`
A man poses for a photo in front of a Chinese building by jumping.	The man is prepared for his photo.	`entailment`
An older gentleman speaking at a podium.	A man giving a speech	`neutral`
A man poses for a photo in front of a Chinese building by jumping.	The man has experience in taking photos.	`neutral`
People are waiting in line by a food vendor.	People sit and wait for their orders at a nice sit down restaurant.	`contradiction`
Number 13 kicks a soccer ball towards the goal during children's soccer game.	A player passing the ball in a soccer game.	`contradiction`