Generalized but not Robust? Comparing the Effects of Data Modification Methods on Out-of-Domain Generalization and Adversarial Robustness

Tejas Gokhale[*] **Swaroop Mishra**[*] **Man Luo**[*]
Bhavdeep Singh Sachdeva **Chitta Baral**
Arizona State University
{tgokhale, srmishr1, mluo26, bssachde, chitta}@asu.edu

Abstract

Data modification, either via additional training datasets, data augmentation, debiasing, and dataset filtering, has been proposed as an effective solution for generalizing to out-of-domain (OOD) inputs, in both natural language processing and computer vision literature. However, the effect of data modification on adversarial robustness remains unclear. In this work, we conduct a comprehensive study of common data modification strategies and evaluate not only their in-domain and OOD performance, but also their adversarial robustness (AR). We also present results on a two-dimensional synthetic dataset to visualize the effect of each method on the training distribution. This work serves as an empirical study towards understanding the relationship between generalizing to unseen domains and defending against adversarial perturbations. Our findings suggest that more data (either via additional datasets or data augmentation) benefits both OOD accuracy and AR. However, data filtering (previously shown to improve OOD accuracy on natural language inference) hurts OOD accuracy on other tasks such as question answering and image classification. We provide insights from our experiments to inform future work in this direction.

1 Introduction

Deep neural networks have emerged as a widely popular architectural choice for modeling tasks in multiple domains such as (but not limited to) computer vision (Yuille and Liu, 2021), natural language processing (Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017), and audio (Hannun et al., 2014). While these models are highly capable of learning from training data, recent studies show that they are quite prone to failure on new test sets or under distribution shift (Taori et al., 2020), natural corruptions (Hendrycks and Dietterich, 2019), adversarial attacks (Goodfellow et al., 2015), spurious correlations (Beery et al., 2018), and many other types of "unseen" changes that may be encountered after training. This shortcoming stems from the *i.i.d.* assumption in statistical machine learning which guarantees good performance only on test samples that are drawn from an underlying distribution that is identical to the training dataset. For instance, digit recognition models trained on the black-and-white MNIST training images are almost perfect ($> 99\%$ accuracy) on the corresponding test set, yet their performance on colored digits and real-world digits from street number plates is less than 75%. Similarly, state-of-the-art NLP models have been shown to fail when negation is introduced in the input (Kassner and Schütze, 2020). These findings pose a significant challenge to the practical adoption of these models and their reliability in the real-world.

To test model performance beyond the traditional notion of in-domain (ID) generalization, two prominent ideas have emerged: out-of-domain (OOD generalization) *a.k.a.* domain generalization[1], and adversarial robustness. The OOD generalization objective expects a model which is trained on distribution $\mathcal{D}$ to perform reliably on unseen distributions $\mathcal{D}_e, e \in \{1, \dots, n\}$, that differ from $\mathcal{D}$. For a trained classifier f^*, OOD accuracy on previously unseen distribution $\mathcal{D}_e$ is defined as:

$$\text{acc}^e_{\text{OOD}} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_e} [\mathbb{I}(f^*(\mathbf{x}) = \mathbf{y})] \qquad (1)$$

To define adversarial robustness, consider an input $\mathbf{x}$ and a true label $\mathbf{y}$. For a classifier loss function ℓ, a loss-maximizing perturbation δ^* within Δ_ϵ (an ϵ-bounded neighborhood of $\mathbf{x}$) is defined as:

$$\delta^*_{\mathbf{x}} = \max_{\delta\in\Delta_\epsilon} \ \ell(f^*(\mathbf{x}+\delta), \mathbf{y}). \qquad (2)$$

The second idea is that of adversarial robustness. Recent work on adversarial examples has revealed

[*] Equal Contribution

[1] In this paper we use these two terms interchangeably.

Findings of the Association for Computational Linguistics: ACL 2022, pages 2705 - 2718
May 22-27, 2022 ©2022 Association for Computational Linguistics

the vulnerability of deep neural networks against small perturbations of the original data. Adversarial robustness in such under this setting is defined as the accuracy of the classifier on adversarial samples $\mathbf{x} + \delta_{\mathbf{x}}$, where the perturbation lies within an ℓ_p norm bound: $||\delta_{\mathbf{x}}||_p < \epsilon$.

$$\text{acc}_{rob} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \mathbb{I}(f^*(\mathbf{x} + \delta_{\mathbf{x}}) = \mathbf{y}). \quad (3)$$

In the context of text classification, the norm-bound can also be in the form of small character-level or word-level perturbations such as swapping, inserting, or deleting characters or words. In essence, adversarial robustness measures the invariance of the classifier to small perturbations of the input.

Various methods have been developed that either improve OOD generalization or improve adversarial robustness. Notable among these are techniques that modify the distribution of the training dataset. In this paper, we focus on three major data modification techniques – the use of additional datasets (also known as multi-source training), data augmentation, and data filtering; in addition we also consider model-based debiasing techniques which do not alter the data distribution explicitly. We study the performance of these methods on three representative tasks – natural language inference (NLI), extractive question answering (QA), and image classification (IC).

Our first aim in this paper is to understand whether the increase or decrease in OOD generalization by each method over the naive baseline (standard training on the source dataset) is consistent across tasks. To further conduct fine-grained analysis, we also analyze the effect of these methods on in-domain (ID) accuracy on the test set for each task, since in the ideal case improvement in OOD performance should not come at the cost of in-domain accuracy.

Recent work seeks to understand the relationships between in-domain and out-of-domain performance: for instance, Miller et al. (2021) empirically show that ID and OOD performance are strongly correlated, Raghunathan et al. (2020); Yang et al. (2020) show a trade-off between robustness and accuracy for adversarially trained models. However it is not clear how methods *designed for OOD generalization* affect robustness. This is largely because work on domain generalization reports only IID and OOD metrics, and work on robustness reports only ID and robustness metrics. Our second aim is to understand the effect of these generalization methods on adversarial robustness.

In addition to our experiments on NLP and vision tasks, we also provide an experiment on a synthetic binary classification dataset where points lie in a 2-dimensional feature space and are separated by concentric circles into class labels. This setting allows us to visualize the effect of data modification techniques on the training distribution and the resulting performance.

Our findings can be summarized as follows:

- More data benefits OOD generalization,
- Data filtering hurts OOD generalization, and
- Data filtering significantly hurts adversarial robustness on all benchmarks.

These findings and our additional analysis raise new questions for robustness and domain generalization research. Significant among these are the importance of both diversity and number of training samples for inductive bias and generalization guarantees, the problems associated with data filtering in terms of robustness, and the importance of a comprehensive set of evaluation metrics that could be adopted for future work.

2 Categorization of Domain Generalization Methods

In this section, we provide a categorization of methods that are typically used as baselines for domain generalization. We briefly explain the method and provide relevant related work in which these ideas are used as methods for domain generalization. Throughout this paper, we will refer to the original training distribution as the *"source"* and the out-of-distribution datasets as the *"targets"*.

Single-Source Training (SS) refers to the "vanilla" baseline which is trained only on the source dataset, without any dataset modification. SS utilizes no other information apart from the single source dataset $\mathcal{D}$ and updates parameters θ of classifier f to minimize the risk on the source using approaches such as ERM (Vapnik and Chervonenkis, 1991).

$$\underset{\theta}{\text{minimize}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \ell(f(\mathbf{x};\theta), \mathbf{y}). \quad (4)$$

Multi-Source Training (MS). This method is identical to SS except that additional training datasets $\mathcal{D}'$ are used for risk minimization.

$$\underset{\theta}{\text{minimize}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}\cup\mathcal{D}'} \ell(f(\mathbf{x};\theta), \mathbf{y}). \quad (5)$$

Usually $\mathcal{D}'$ are designed for the same task as $\mathcal{D}$ but may have different styles, characteristics, or sources of collection. For instance, while both SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) are datasets for natural language inference with identical class labels, SNLI was collected from image captions, while MNLI was collected from Open American National Corpus[2].

Gulrajani and Lopez-Paz (2020) provide an extensive comparitive study of models trained for multi-source domain generalization for image classification and surprisingly find that if multiple source domains are available, ERM is empirically the best approach as compared to specially designed DG methods such as meta-learning (Li et al., 2018a), learning domain-invariant features (Ganin et al., 2016), invariant risk minimization (Arjovsky et al., 2019), etc. These findings have also been observed on text classification experiments in (Koh et al., 2021). Hendrycks et al. (2020a) show that pre-training transformer architectures on diverse data leads to higher OOD accuracies on multiple tasks such as semantic textual similarity, sentiment classification, reading comprehension and natural language inference.

Data Augmentation (`DA`). When additional training distributions are not directly available, transformations of samples in $\mathcal{D}$ using pre-defined augmentation functions can be used to create $\mathcal{D}'$ and train the model. Such data augmentation functions are typically derived from existing knowledge about the invariance of the task w.r.t. certain transformations. For instance, for image classification, addition of small noise, small translations, scaling, etc. are common data augmentation functions, since they do not change the true label for the image. Similarly, for text inputs, synonyms of words are commonly used since they do not change the semantics of the sentence. NLP data augmentation techniques include UDA (Xie et al., 2020), EDA (Wei and Zou, 2019), and back-translation for question answering (Longpre et al., 2019).

Data Filtering (`DF`). Dataset filtering has been previously explored for quality control, such as, removing noise and artifacts to curate and improve publicly sourced datasets. However, there has been recent interest in considering `DF` as a method for bias reduction and generalization. This idea can be traced back to work by Zellers et al. (2018, 2019), that proposed `DF` as an algorithmic method to avoid annotation artifacts and spurious correlations during dataset construction. AFLite (Bras et al., 2020) extended this idea to a generic filtering methodology that can work without any pre-defined rules or strategies. Instead, AFLite operates by utilizing several weak learners (such as support-vector machines) trained over small subsets to identify samples that are easy to classify. It is argued that such samples are more likely to carry biases, and as such, could be removed. AFLite suggests that reduction of a dataset to even 10% of the original size can boost OOD accuracy on NLI. In the vision domain, similar ideas have been proposed concurrently, including REPAIR (Li and Vasconcelos, 2019) and RESOUND (Li et al., 2018b), in which instead of completely removing samples, biased samples are assigned smaller weights. However these methods require a prior knowledge of the bias variable. Liu et al. (2021) have recently proposed a simple approach which upweights samples which have higher loss – this is shown to improve worst-group accuracy without having access to the bias variable.

Model De-biasing (`DB`). Methods under this category do not directly alter the training dataset, but instead resort to changes in the modeling technique – these changes can be in terms of the optimization function, regularization, additional auxiliary costs, etc. The main idea in `DB` is to utilize known biases (or identify unknown biases) in the data distribution, model these biases in the training pipeline, and use this knowledge to train robust classifiers (Clark et al., 2019; Wu et al., 2020; Bhargava et al., 2021). In the image classification literature, there is growing consensus on enforcing a consistency on different views (or augmentations) of an image in order to achieve debiasing (Hendrycks et al., 2020c; Xu et al., 2020; Chai et al., 2021; Nam et al., 2021). Unlike `DF`, model de-biasing does not directly alter the training distribution, but instead allows the model to learn which biases to ignore.

3 Toy Example: Concentric Circles

We begin with a simple two-dimensional example to illustrate our experimental setting and to show how each method affects the distribution of the training set. Consider the set of points shown in Figure 1 where the points belong to two class labels (either `0` or `1`) and are seen to lie on concentric circles. Points with label `0` are closer to the origin,

[2] `https://www.anc.org/`

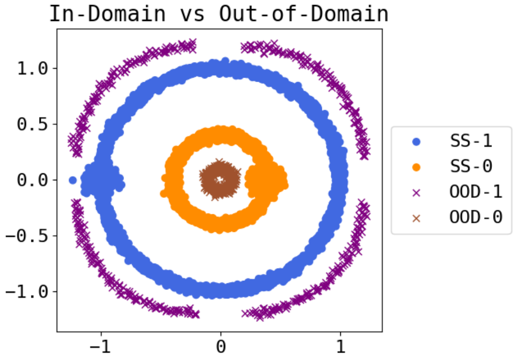

Figure 1: Our toy experimental setting consists of points in $\mathbb{R}^2$ belonging to two classes (0/1). This illustration shows the discrepancy between the source dataset (SS) and the out-of-domain dataset (OOD).

while points with label 1 are closer to a distance of 1 from the origin. Our aim is to start with the single source dataset and train the model to generalize on the out-of-domain (OOD) dataset. An important thing to note here is that the source dataset contains a subset of points with label 0 (orange) clustered around $(0.4, 0.0)$ and a subset with label 1 clustered around $(-1, 0.0)$. This implies that class-0 is biased towards $x > 0$, while class-1 is biased towards $x < 0$. In total, our SS dataset consists of 10000 samples, of which 20% are biased.

We apply three data modifications: additional source (MS), gaussian data augmentation (DA) $\sim\mathcal{N}(0, 0.1)$, and data filtering (AFLite) which reduces the dataset size to 10%. Note that we do not show model debiasing (DB) here, since it does not alter the data distribution. Figure 2 shows the effect on the data distribution. The most striking is the effect of DF which removes all samples previously in the biased clusters near $(0.4, 0.0)$ and $(-1.0, 0.0)$.

Equipped with these resulting datasets, we train a linear SGD classifier with log-loss and evaluate the robustness of each model in terms of in-domain and OOD accuracies. We also evaluate adversarial robustness by using standard PGD attacks. Results are shown in the textboxes in Figure 2. It can be seen that data filtering significantly hurts both OOD generalization and robustness. This finding motivates our experiments to understand the effect of each method for NLP and vision tasks.

4 Experiments

In this section, we present three tasks and their corresponding experimental setup, evaluation protocol and our findings. A summary of methods belong to each category is provided in Table 1 and the abbreviations SS, MS, DA, DB, DF are used henceforth.

4.1 Natural Language Inference (NLI)

NLI is the task of determining whether a *hypothesis* is true (entailment), false (contradiction), or undetermined (neutral) given a *premise*.

Methods. We use RoBERTa as the backbone model for each method and SNLI (Bowman et al., 2015) as our source training corpus. A model trained with expected risk minimization (ERM) on SNLI alone, forms our single-source (SS) baseline. A model trained with a combination of SNLI and MNLI (Williams et al., 2018) forms our multi-source (MS) baseline. We apply EDA (Wei and Zou, 2019) to augment our training dataset with 100% of additional data to train a DA model. The LMH debiasing method from Clark et al. (2019) represents our DB model. For data filtering, we use AFlite (Bras et al., 2020) to filter out 90% of the SNLI training data, and use the remaining 10% data to train our DF model – this setting is based on the experiments from (Bras et al., 2020).

Evaluation Protocol. We report accuracy on the SNLI test set (IID), and to evaluate generalization, we report accuracy on NLI diagnostics (Wang et al., 2018), Stress test evaluation (Naik et al., 2018a) and HANS (McCoy et al., 2019a). We use two metrics for evaluating robustness:

- *model-based robustness* uses BAE adversarial attack (Garg and Ramakrishnan, 2020), implemented using TextAttack (Morris et al., 2020), and reports robustness as number of queries (sequential perturbations) needed to fool the model.
- *model-free robustness* uses six pre-defined operations to transform SNLI test inputs into adversarial examples. These six methods are: CLARE (Li et al., 2021a), character-swap (Pruthi et al., 2019), Checklist (Ribeiro et al., 2020), EDA (Wei and Zou, 2019), counter-fitted embeddings (Emb) (Alzantot et al., 2018a).

Results. Table 2 shows the performance of each method in terms of in-domain and out-of-domain accuracy. We observe that four methods all improve the generalization performance on average but decrease the in-domain performance. Especially, DF method is the best in terms of OOD accuracy, but is the worst in terms of in-domain performance. We also see a trend that four methods improve the generalization in all sets of NLI-

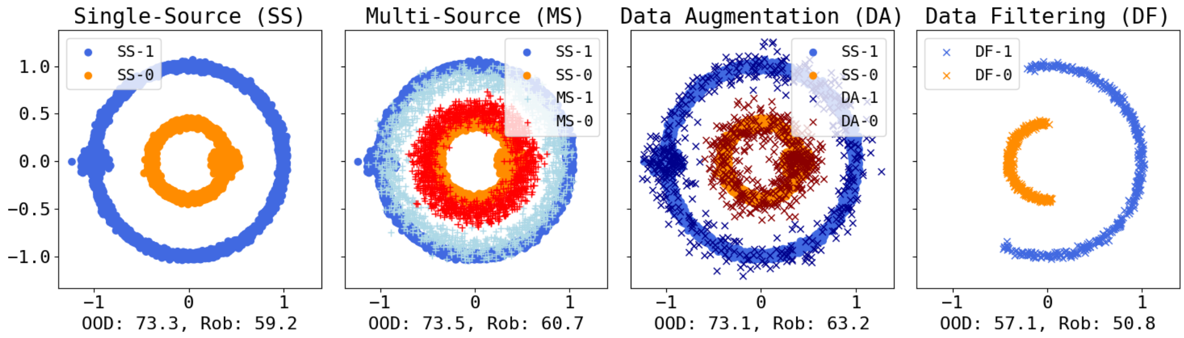

Figure 2: This figure illustrates the effect of data modification techniques on the training distribution. The leftmost figure shows the training distribution in the single-source setting. The introduction of a second dataset or Data-augmentation (done using small perturbations of source samples with Gaussian noise) makes the distribution more diverse in the multi-source (MS) and data augmentation (DA) setting respectively. On the other hand, data filtering, in order to remove spurious correlations from the dataset, removes points from certain sectors of the distribution. The effect of each strategy on OOD generalization and robustness is shown below each plot.

Method Category	**Tasks**		
	Natural Language Inference	Question Answering	Image Classification
SS (Single-Source ERM)	SNLI	NQ (Kwiatkowski et al., 2019)	MNIST
MS (Multi-Source ERM)	SNLI + MNLI	NQ + SQuAD+NQA+HQA+SQA+TQA	MNIST + USPS
DA (Data Augmentation)	EDA (Wei and Zou, 2019)	QG (Chan and Fan, 2019)	M-ADA (Qiao et al., 2020)
DB (Model De-biasing)	LMH (Clark et al., 2019)	Mb-CR(Wu et al., 2020)	RandConv (Xu et al., 2020)
DF (Data Filtering)	AFLite (Bras et al., 2020)	AFLite (adapted for QA)	AFLite

Table 1: List of method categories and specific methods that we use under each task setting in nour experiments. Details for each can be found in Section 4 for the corresponding task.

Diagnostics and HANS, while all four methods do not show improvement on generalization on Distraction and Noise sets of Stress dataset.

Table 3 shows the robustness evaluation. We see that except for DF, all methods improve the robustness under both model-based and model-free evaluation. MS improves the robustness in all transformations except for EDA. DA achieves the best robustness by model-based evaluation but is not consistent in terms of different transformations of model-free evaluation. DB improves the robustness in terms of every transformation and achieves the best robustness in terms of average of model-free evaluation. DF significantly hampers the model-free robustness with a drop in all transformations.

4.2 Question Answering (QA)

We focus on extractive QA. Given a passage (or "context") and a question, the task is to extract the answer span from the passage.

Methods. We use BERT (Devlin et al., 2019) as the backbone model for each method. We use MRQA (Fisch et al., 2019) which is a collection of 12 publicly available multi-domain QA datasets – with Natural Questions (NQ) (Kwiatkowski et al., 2019) as the source dataset. SQuAD, NewsQA, HotpotQA, SearchQA, and TriviaQA are used as additional datasets for multi-source training. Similar to NLI, we use EDA for DA by applying EDA on the question. We apply the augmentation to all samples in the training set and combine them with the original set to train a DA model. For model debiasing (DB), we use Mb-CR approach (Wu et al., 2020), where a teacher and bias models are trained *a* priori, and are used for debiasing.

We modify AFLite for our QA task of span prediction, since AFLite was originally designed for classification tasks. To do so, we first randomly divide the training set into 10 subsets (or folds) $S_{1:10}$. For $k \in \{1, \ldots, 10\}$, we pick S_k as the held-out test set, and train models on the rest, and obtain 10 such models. At test time, models are used for predicting an answer by only looking at the context (without access to the question) – this allows us to identify strong spurious correlations in the dataset. Based on the predictions, samples are sorted on the basis of their F1 score. A higher F1 score implies that the model is more likely to answer the question

Method	In-Domain Acc. (%)	OOD Acc. (%)										
		NLI-Diagnostics				Stress Test			HANS			Avg
		Kno.	Lex.	Log.	PAS	Comp.	Distr.	Noise	Lex.	Subs.	Consti.	
SS	**89.6**	51.8	65.7	57.8	72.6	77.9	73.5	79.8	88.4	28.2	21.7	61.74
MS	87.8	52.1	66.8	57.8	72.8	79.6	72.4	79.2	92.0	33.6	26.7	63.30
DA	87.2	52.1	66.0	58.1	72.6	79.6	71.8	79.2	92.8	32.8	26.4	63.14
DB	81.8	52.4	66.0	58.4	72.8	79.3	71.8	79.5	92.2	33.8	27.5	63.37
DF	62.6	53.9	66.5	58.7	68.9	79.1	72.0	79.5	94.1	46.3	38.5	**65.75**

Table 2: NLI Result: In-domain (IID) accuracy and out-of-domain generalization (OOD) on the NLI benchmark using SNLI as source dataset. [3] *See Table 1 for method abbreviations.*

Method	Model Based #Num Queries	Model Free Accuracy (%)						
		CharSwap	EasyData	Embedding	WordNet	CheckList	CLARE	Avg
SS	53.56	81.3	72.0	81.9	77.0	89.4	76.3	79.65
MS	54.44	81.5	71.6	82.0	78.2	89.2	77.5	80.00
DA	**55.06**	77.7	74.1	80.7	80.2	86.6	80.5	79.97
DB	54.82	81.5	72.4	82.3	78.0	89.2	77.0	**80.07**
DF	51.13	65.2	56.8	66.2	62.5	72.3	62.5	64.25

Table 3: NLI Result: Comparison of robustness in terms of model-based evaluation (number of queries needed to fool the model) and model-free (accuracy on adversarial transformations). [2] *See Table 1 for method abbreviations.*

without even knowing the question. We retain 10% samples with the lowest F1 scores – these represent the task since the model is not likely to predict the correct answer without knowing the question.

Evaluation Protocol. We report exact-match (EM) accuracy for MRQA. To evaluate the generalization performance, we use six OOD development sets from MRQA: DROP, RACE, BioASQ, TextbookQA, RelationExtraction, and DuoRC. For robustness, we use the "Morphues" attack (Tan et al., 2020) on the question as the model-based evaluation, the attack method is similar to NLI. Model-free methods are the same as NLI.

Results. Table 4 shows the performance of each method in terms of in-domain and out-of-domain accuracy. We observe that two methods, MS and DB, improve the generalization performance on each out-of-domain dataset and also improve the in-domain performance. The improvement of MS is larger than DB. DA improves on some out-of-domain datasets but not all, and it also improves the in-domain performance. DF dramatically reduces both out-of-domain and in-domain datasets.

Table 5 shows that except for DF, all methods improve over SS for both model-based and model-free robustness evaluation. MS, DA, and DB improve the robustness in all transformations of model-free evaluation as well as the model-based evaluation, where MS achieves the best performance in model-based and model-free evaluation. DF significantly hampers the model-free robustness with drop in all transformations, meanwhile, the model-based robustness also drops.

4.3 Image Classification

We conduct our experiments on the standard domain generalization benchmark "Digits", which is a collection of handwritten digit classification datasets belonging to 10 classes (digits 0–9). Following standard practice(Volpi et al., 2018), we train models on 10000 images from MNIST (LeCun et al., 1998) as the source, and use SVHN (Netzer et al., 2011), SYN and MNIST-M (Ganin and Lempitsky, 2015) as the OOD datasets.

Methods. We use DigitNet (Volpi et al., 2018) as our backbone image classifier architecture. Our SS baseline uses MNIST for training; MS uses MNIST and USPS (Denker et al., 1988). For data augmentation we rely on M-ADA (Qiao et al., 2020) which is a perturbation-based min-max algorithm to create augmented data. Our debiasing method is RandConv (Xu et al., 2020) which utilizes a random convolutional layer to generate novel views of each input image, and a KL-divergence based loss function that encourages the classifier to predict consistent predictions for each version of the image. This leads to the model being debiased on spurious features like background, texture, or color of digits. We use AFLite as our DF method.

Method	In-Domain EM. (%)	OOD EM. (%)						
		DROP	RACE	BioASQ	TBQA	R.E.	DuoRC	**Avg**
SS	63.76	20.09	19.29	33.91	28.61	62.82	32.71	32.91
MS	65.07	26.88	27.45	45.01	40.52	72.86	43.44	**42.69**
DA	63.84	19.23	19.73	32.31	28.54	61.97	32.31	32.35
DB	64.58	20.83	19.73	34.64	31.20	63.64	35.98	34.34
DF	49.56	9.25	11.72	20.94	19.63	45.28	21.45	21.38

Table 4: QA Result: Source (IID) accuracy and domain generalization (OOD) on the Question Answering benchmark with NaturalQuestions as source dataset. EM: Exact-Match. *See Table 1 for method abbreviations.*

Method	Model Based #Queries	Model Free EM. (%)						
		CharSwap	EasyData	Embedding	WordNet	CheckList	CLARE	**Avg**
SS	19.55	60.29	52.17	61.21	58.41	63.22	61.92	59.54
MS	21.97	62.22	52.65	63.22	59.84	64.42	63.55	**60.98**
DA	21.91	60.88	54.52	62.02	59.82	63.42	62.36	60.5
DB	20.40	61.62	53.16	62.35	59.32	64.03	63.01	60.58
DF	19.19	47.97	42.48	48.55	47.19	49.34	48.72	47.38

Table 5: QA Result: Comparison of robustness in terms of model-based evaluation (number of queries needed to fool the model) and model-free (accuracy on adversarial transformations). [2] *See Table 1 for method abbreviations.*

Method	In-Domain Acc. (%)	OOD Acc. (%)			
		MNIST-M	SVHN	SYNTH	Avg
SS	98.40	58.09	33.85	45.94	45.96
MS	98.54	59.79	33.87	48.42	47.36
DA	99.30	67.94	42.55	48.95	53.15
DB	98.86	87.67	54.95	63.37	68.66
DF	95.27	51.04	22.07	27.83	33.65

Table 6: Source (in-domain) accuracy and domain generalization (OOD accuracy) on the Digits benchmark with MNIST-10k as source dataset.[2]

Evaluation Protocol. We report IID accuracy on the MNIST test set and generalization as the accuracy on our OOD datasets. For evaluating adversarial robustness we use Foolbox (Rauber et al., 2017) and use 10 attack methods (both ℓ_2 and ℓ_∞ versions of FGSM, PGD, BIM, AUN, and DeepFool). Robustness is calculated as the accuracy for 20 values of ϵ between $[0, 2]$, and is plotted as robustness curves for visualization, along with the average values for area under the curve (AUC).

Results. Table 6 shows the performance of each method in terms of in-domain and OOD accuracy. MS, DA and DB, improve the generalization performance on each OOD dataset and also improve the in-domain performance, where DB displays best generalization capacity. DF dramatically reduces the OOD performance with significant reduction across all datasets; the in-domain accuracy also decreases. Figure 3 shows robustness (accuracy) and area under the curve (AUC) for each plot. It can be observed that DF is worse than SS for all 10 attack variants. We observe that DA and DB are better than SS, and the drop for DF is the largest.

5 Analysis

Based on the results of three tasks, we have the following observations about the performance of each method compared to the SS baseline:

- MS increases OOD accuracy on all three tasks and robustness on two tasks (NLI and QA).
- DA increases OOD on two tasks (NLI and IC) and robustness on all three tasks.
- DB increases OOD on three tasks and robustness on two tasks (NLI and QA).
- DF decreases OOD on two tasks (QA and IC) and robustness on all three tasks.

Decrease in NLI in-domain accuracy is seen for all methods, even though these lead to increase in OOD accuracy. This suggests that the training dataset (SNLI) has a large shift w.r.t. OOD datasets.

More data implies more OOD generalization: While this trend is observed for both MS and DA, there is one anomaly – DA for the QA task leads to marginal decrease compared to SS (a difference of 0.56%). This finding is aligned with Longpre et al. (2019), who report no significant effect of data augmentation (back translation) on OOD performance for question answering. This points to the need for improving data augmentation techniques in QA.

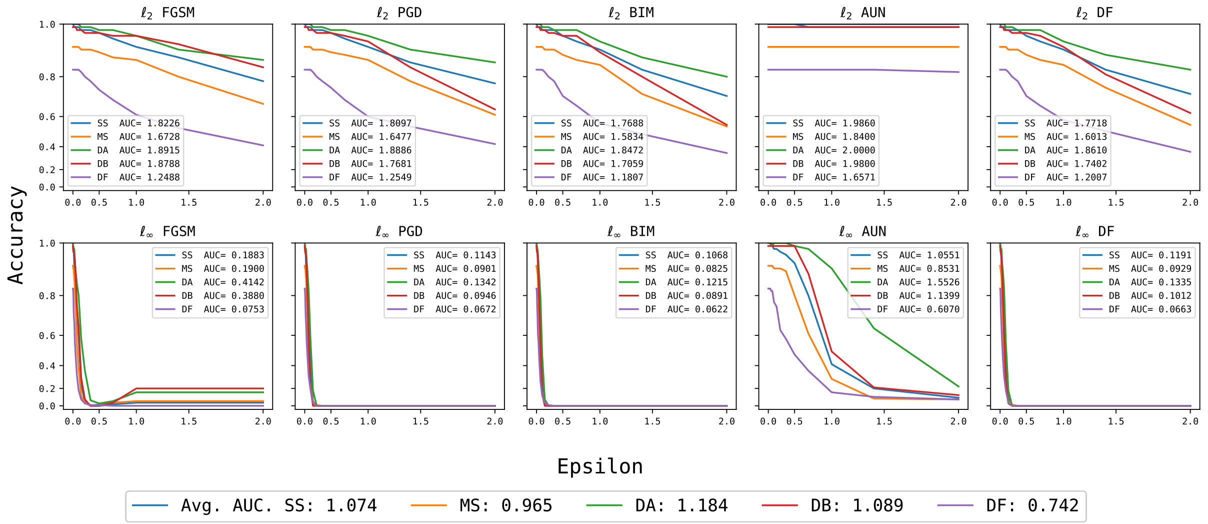

Figure 3: Evaluation of adversarial robustness (using 10 attack methods) for MNIST10k.

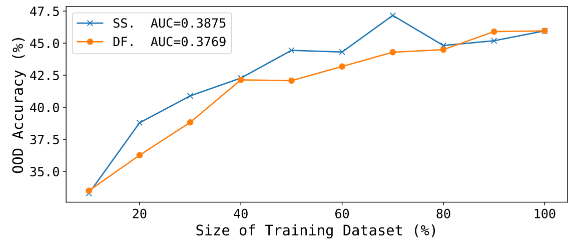

Figure 4: Comparison between SS and DF models trained with different percentages of MNIST10k.

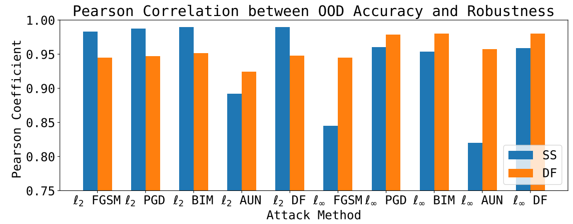

Figure 5: Pearson Correlation between OOD accuracy and robustness for SS and DF models on MNIST10k.

On the other hand, the performance drop due to DF is significantly large for QA (11.53%).

Decrease in MNIST robustness: For MNIST, the DA method (M-ADA (Qiao et al., 2020)) is the best in terms of robustness and also improves OOD accuracy. M-ADA is an "adversarial data augmentation" method, i.e., it uses a min-max objective to find loss-maximizing perturbations and uses these perturbations as augmented data. It is therefore intuitive that such a method would do well on the adversarial robustness metric (although robustness evaluation was not reported by Qiao et al. (2020)).

Marginal Improvement on Robustness: From the results, it is easy to see that the improvement on OOD is more noticeable than robustness, for example, MS improves OOD performance by $\sim 10\%$, but improves only by $\sim 1\%$ under model-free evaluation. While this observation is reasonable since each method is designed to improve the generalization, new methods that improve both generalization and robustness should be encouraged.

5.1 Correlation between Adversarial Robustness and OOD Generalization

Our experiments reveal the alarming finding that across the board, DF reduces adversarial robustness. To investigate further, we conduct an analysis on the Digits benchmark and compare SS and DF when trained with equal amounts of data ($\{10\%, 20\%, \ldots, 100\%\}$). Note that for SS the data are sampled randomly, while for DF the data are obtained via AFLite data filtering. Results are shown in Figure 4. It can be observed that the OOD accuracy increases as the size of the dataset increases, and is greater for SS than DF. To understand how an increase in OOD accuracy affects robustness, we also compute the robustness values at each size of training data, and compute the Pearson correlation coefficient for each attack method – positive correlation implies that as OOD accuracy increases, robustness also increases. Figure 5 shows clear evidence in favor of positive correlation; interestingly, SS has higher correlation for ℓ_2 attacks, while DF is higher for ℓ_∞ attacks. The evidence is clear: OOD generalization increases with the size of the dataset and adversarial robustness is positively correlated with OOD generalization.

Our experiments show that the size of the training set directly affects both robustness and generalization. While removing 90% data increased OOD accuracy in NLI, the effect was the exact opposite for QA and MNIST. The key idea in domain generalization is that the test distributions are unknown and little information about them is available apart from the fact that there is no task shift. Without this prior knowledge, deciding whether (or how much) to filter a dataset is a challenging task.

6 Related Work

In Section 2 we have provided relevant work that falls into one of our five modeling categories. Here, we discuss additional literature on robustness and generalization and new efforts towards dataset creation, benchmarks, and evaluation.

Generalization Benchmarks. Hendrycks et al. (2020b) have constructed a robustness benchmark for multiple language understanding tasks by splitting training sets from existing benchmarks according to topics, styles, and vocabulary; this has been subsequently used to study robustness of model rankings (Mishra and Arunkumar, 2021). Benchmarks have also been constructed to study dataset artifacts and generalization capabilities of models (Mishra et al., 2020a,b; Mishra and Sachdeva, 2020). MRQA (Fisch et al., 2019) is a benchmark for evaluating domain generalization of question answering (reading comprehensive) models. MRQA contains 6 datasets each for training, development, and evaluation. For image classification, many benchmarks have been proposed to evaluate domain generalization, such as PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), Digits (Volpi et al., 2018), and WILDS (Koh et al., 2021) which is a compendium of domain generalization bechmarks for tasks such as image classification, text sentiment and toxicity prediction.

Corruption Robustness. Hendrycks and Dietterich (2019) introduced ImageNet-C and CIFAR-C to test robustness along corruptions such as weather, noise, blur, and digital artifacts, and ImageNet-P which tests robustness against small tilts and changes in brightness. MNIST-C was introduced by Mu and Gilmer (2019) for similar corruptions of handwritten digit images.

Adversarial and Contrastive Sets. Generation of adversarial examples (Jia and Liang, 2017; Ribeiro et al., 2018; Iyyer et al., 2018; Alzantot et al., 2018b) and approaches to defend against word substitution (Jia et al., 2019) have been explored. Contrastive examples have been introduced as a means for evaluation, for example, manually crafted contrast sets for textual entailment (Gardner et al., 2020) or template-based (McCoy et al., 2019b; Glockner et al., 2018; Naik et al., 2018b). Model-in-the-loop dataset creation methods have also been proposed for various NLP tasks (Nie et al., 2020; Arunkumar et al., 2020; Kiela et al., 2021) and visual question answering (Sheng et al., 2021; Li et al., 2021b).

7 Discussion

Recently, Miller et al. (2021) have empirically shown linear trends between in-distribution and out-of-distribution performance on multiple image classification tasks, across various model architectures, hyper-parameters, training set size, and duration of training. They also show that there are certain settings of domain shift under which the linear trend does not hold. Our work empirically shows that while data filtering may benefit OOD generalization on the NLI benchmark, this does not hold for other tasks such as image classification and question answering. This suggests that data filtering may benefit generalization in certain types of domain shift, but not on others. Concurrently, Yi et al. (2021) have theoretically shown that models robust to input perturbations generalize well on OOD distribution within a Wasserstein radius around the training distribution. Our empirical observations in this paper in both vision and language domains, agree with the theory of Yi et al. (2021).

In this work, we conduct a comprehensive study of methods which are designed for OOD generalization on three tasks: NLI, QA, and IC. We evaluate each method on in-domain, OOD, and adversarial robustness. [4] Our findings suggest that more data typically benefits both OOD and robustness. Data filtering hurts OOD accuracy on two out of three tasks, and also hurts robustness on all three tasks. In context of our findings and work by Miller et al. (2021); Yi et al. (2021), we recommend that methods designed either for robustness or generalization should be evaluated on multiple aspects and not on the single metric that they are optimized for.

[4] Code for our experiments will be released at `https://github.com/tejas-gokhale/gen-vs-rob`.

Acknowledgements

This work was funded in part by DARPA SAIL-ON program (W911NF2020006) and DARPA CHESS program (FA875019C0003). The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

Broader Impact

One underlying assumption behind using large datasets for training (or pre-training) vision and language models is that larger datasets increase the likelihood of obtaining a diverse set of samples to reduce overfitting. However, recent studies (Bender et al., 2021; Stanovsky et al., 2019) serve as cautionary tales when employing uncurated internet data to train large language models, and discuss how large data does not necessarily imply that models will learn the dievrse distribution. At the same time, the inverse (small data aids diversity) is also not true (as shown by this paper) and comes with its own problems – for instance, Figure 2 shows that dataset filtering can lead to much larger changes in the data distribution beyond notions of proportionality and fairness. As such, the decision on how many and what samples to remove can also introduce its own set of biases. Data curation is a challenging problem and needs further task-specific study since the concepts of bias and fairness often depend on the task definition and specifications of ideal outcomes. Insights from this paper could help researchers and practitioners in choosing appropriate approaches for improving generalization and robustness.

References

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

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

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Anjana Arunkumar, Swaroop Mishra, Bhavdeep Sachdeva, Chitta Baral, and Chris Bryan. 2020. Real-time visual feedback for educative benchmark creation: A human-and-metric-in-the-loop workflow.

Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. 2021. Ensembling with deep generative views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14997–15007.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

JS Denker, WR Gardner, HP Graf, D Henderson, RE Howard, W Hubbard, LD Jackel, HS Baird, and I Guyon. 1988. Neural network recognizer for hand-written zip code digits. In *Proceedings of the 1st International Conference on Neural Information Processing Systems*, pages 323–331.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. In *International Conference on Learning Representations*.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Dan Hendrycks and Thomas G. Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020a. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020b. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020c. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages

2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and William B Dolan. 2021a. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021b. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *International Conference on Computer Vision (ICCV)*.

Yi Li and Nuno Vasconcelos. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9572–9581.

Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018b. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 220–227, Hong Kong, China. Association for Computational Linguistics.

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

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

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR.

Swaroop Mishra and Anjana Arunkumar. 2021. How robust are model rankings: A leaderboard customization approach for equitable evaluation. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 13561–13569.

Swaroop Mishra, Anjana Arunkumar, Chris Bryan, and Chitta Baral. 2020a. Our evaluation metric needs an update to encourage generalization. *arXiv preprint arXiv:2007.06898*.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020b. Dqi: A guide to benchmark evaluation. *arXiv preprint arXiv:2008.03964*.

Swaroop Mishra and Bhavdeep Singh Sachdeva. 2020. Do we need to create big datasets to learn a task? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 169–173.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Norman Mu and Justin Gilmer. 2019. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*.

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

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

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. 2021. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591.

Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12553–12562. IEEE.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. 2020. Understanding and mitigating the tradeoff between robustness and accuracy. *Proceedings of Machine Learning Research*.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. 2017. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *arXiv preprint arXiv:2106.02280*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin'time! combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599.

Vladimir N Vapnik and A Chervonenkis. 1991. The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recognition and Image Analysis*, 1(3):284–305.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5339–5349.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

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

Mingzhu Wu, Nafise Sadat Moosavi, Andreas Rücklé, and Iryna Gurevych. 2020. Improving qa generalization by concurrent modeling of multiple biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 839–853.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. 2020. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. 2020. A closer look at accuracy vs. robustness. In *NeurIPS*.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. 2021. Improved ood generalization via adversarial training and pretraing. In *International Conference on Machine Learning*, pages 11987–11997. PMLR.

Alan L Yuille and Chenxi Liu. 2021. Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, 129(3):781–802.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.