Is High Quality Data All You Need?

Swaroop Mishra
srmishr1@asu.edu

Anjana Arunkumar
aarunku5@asu.edu

Bhavdeep Sachdeva
bssachde@asu.edu

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University

Abstract

Even though deep neural models have achieved superhuman performance on many popular benchmarks, they have failed to generalize to OOD or adversarial datasets. Conventional approaches aimed at increasing robustness include developing increasingly large models and augmentation with large scale datasets. However, orthogonal to these trends, we hypothesize that a smaller, high quality dataset is what we need. Our hypothesis is based on the fact that deep neural networks are data driven models, and data is what leads/misleads models. In this work, we propose an empirical study that examines how to select a subset of and/or create high quality benchmark data, for a model to learn effectively. We seek to answer *if big datasets are truly needed to learn a task*, and whether a *smaller subset of high quality data* can replace big datasets. We investigate both *data pruning* and *data creation* paradigms to generate high quality datasets.

1 Introduction

Deep neural models such as EfficientNet-B7 [37], BERT [5] and RoBERTA [25] have achieved super-human performance on many popular benchmarks in various domains such as Imagenet [34], SNLI [2], and SQUAD [33]. However, their performance drops drastically on exposure to out of distribution (OOD) and adversarial datasets [13, 7, 14, 12]. Lots of resources and time are being invested in developing better models and architectures, such as transformer based approaches [42], that dominate leaderboards. *Since deep learning –a data driven approach– finds representation from data, shouldn't the focus be placed on creating 'better' datasets rather than developing increasingly complex models?*

Let us consider this through an analogy– a student (A) is asked to self-learn a concept by going through a question bank (Q_1), where there are 1000 solved questions. After self-learning, A is tested using 100 unsolved questions present at the end of the Q_1. While A achieves unprecedented performance (85/100), beating other students who are explicitly taught the concept, when tested on another 100 questions on the same topic from question bank (Q_2), A fails on 50 questions. Similarly, if A is interviewed by a teacher, A fails to answer 70 questions.

On analysis, we see that A has not truly learned the concept in Q_1; instead, A solves questions by relying on common question patterns seen in Q_1, and associating them with the provided answers. To fix this, suppose A is provided 1000 solved questions from Q_2. On testing, we find that A now correctly answers 90/100 unsolved questions from Q_2, but only 55/100 from Q_1, and 35/100 in the interview. Now, we provide A with 100 question banks in a similar manner, and find that A's performance on both Q_1 and Q_2 is 70/100, and is 40/100 for the interview. To improve interview performance, suppose that the interviewer prepares an additional question bank Q_i, then if A self-learns using both Q_1 and Q_i, then the scores for Q_1, Q_2, and the interview are 80, 45, and 80/100 respectively. However, if the interviewer changes, A again fails to correctly answer 70 questions.

Since the provision of additional question banks in different settings was not very effective, we

Pre-registration workshop NeurIPS (2020), Vancouver, Canada.

introduce a set of constraints in A's self-learning strategy that disallows A from picking up on questions patterns and answer associations. However, these constraints increase the time that A spends on self-learning, and the number of question banks required (also, in turn the money spent if question banks are rented on a time basis). We find that this improves A's accuracy in answering Q_1, Q_2, and interview questions to 95, 70, and 50/100 respectively.

Clearly, the above methods do not fully solve the problems in A's self-learning strategy. This leads us to question where the problem actually lies– *is it in the learning strategy or in the learning material?* Intuitively, improving A's learning strategy is conducive only if A is being provided high quality learning material without any scope for identifying question patterns and answer associations; this forces A to look beyond idiosyncrasies of the question banks that A is tested on.

How do we create high quality datasets for models to learn from? To define 'quality', we require a *quality index for machine learning*, similar to those used in the domains of power [1], air [15], food [9] and water [31]. Recently, based on a broad survey of AI literature, DQI [28] has been proposed as a data quality index for NLP; here, relevant text properties that lead to either spurious or inductive bias are identified and used to construct a formula that quantitatively evaluates benchmarks. DQI comprises of 133 terms; in this work we aim to study how some of these terms both individually and collectively can help models learn tasks in a few-shot setting. We aim to conduct two types of experiments: (i) dataset pruning on existing benchmarks using different DQI terms (individually and/or combined), and (ii) controlled crowdsourced creation of high quality datasets based on DQI.

2 Related Work:

Our progress in AI is evaluated by building and solving increasingly harder benchmarks. This in turn leads to the development of new models and architectures. This trend requires heavy resource investment, in terms of time, cost, hardware, etc. However, in this process, we must ask if we can *truly rely on our benchmarks*. A series of recent works have shown that models exploit spurious bias – unintended correlations between input and output [39, 3]– to solve tasks, instead of actually learning the task from underlying data features [10, 36, 32, 41, 18, 38, 21].

The mitigation of spurious bias has consequently become an increasingly prevalent track of research. Some of the most common methods to achieve this are dataset pruning [35, 22, 23, 44], residual learning [4, 11, 26], adversarial dataset creation [48, 30], and counterfactual data augmentation [19, 8].

Each of these methods focuses on a specific part of the data-model loop: (i) accepting/ rejecting a data sample created by a crowd-worker [30], (ii) retaining/ removing data with adversarial filtering [35, 22, 23], (iii) augmenting only counter factual data [19, 8], and/or (iv) including data only if it can fool the model [48, 30]; they are all commonly limited by binary evaluation, and can also introduce new kinds of bias, overfitting to a specific model or task [24].

Binary evaluation in particular, is extremely restrictive as it only allows inclusion or deletion of data, and further appends an overhead on human evaluators as there is uncertainty in class distinction. Some other limitations include: (i) resource wastage in the initial creation of 'biased' data, (ii) dataset creators are likely to repeat mistakes that lead to the making of biased data, as they do not learn what constitutes biased data, (iii) important aspects of bias– such as its dependency on a train-test split– are ignored, (iv) model training on each iteration increases time complexity, and (v) there is too much effort required on the part of crowdworkers/authors/experts without providing a suitable and/or illustrative feedback channel to educate data creators.

Using DQI to quantify benchmark quality can potentially address these issues; higher DQI implies lower bias and higher generalization.

3 Method

We intend to utilize DQI in two ways: (i) dataset pruning, and (ii) dataset creation. We start by addressing if ***(H1)** we can learn a task effectively with smarter sample selection(pruning).* However, pruning overlooks resource wastage in creating biased data. So, we investigate if ***(H2)** we can leverage our pruning approach to assist crowd workers in constructing a smaller, but higher quality dataset in the first place, such that pruning is no longer required.* Answering **H2** will justify the utility of our question–*Is high quality data all you need?*– and change the deep learning trend of creating big datasets.

3.1 Dataset Pruning:

Do we really need big datasets? Motivated by the process of human learning which relies on deep background knowledge about the world– we don't need access to hundreds of online materials to learn a topic, rather we intentionally avoid many noisy, distracting, and irrelevant materials– we probe this question. Considering that pre-training on large datasets has imparted linguistic knowledge to models like BERT [5] and RoBERTA [25], we realize that models no longer need to learn from scratch; instead, learning task-specific terminology (such as 'Entailment'/'Neutral'/'Contradiction'labels for Natural Language Inference) suffices, and might not necessitate the use of large datasets.

We therefore aim to find the high quality subset of benchmark data required to learn a task. Our approach is inspired by human tendency to: (i) estimate the presence of relevant materials from the total available material, (ii) remove redundant/irrelevant/known content from the initially selected material, and (iii) use background knowledge of the task, task priority, and time available for learning to heuristically sort and select relevant (i.e., high quality) content.

Algorithm: We mimic this material selection process in Algorithm 1. We use 2 modules for learning– (i) AFLite [3, 35], and (ii) DQI. AFLite is a recent technique for adversarial filtering of dataset biases using linear models, whereas DQI has a method to quantify quality of samples, with or without annotation.

Formalization: Let M be the model, full dataset D and pruned dataset S, and for each sample s, $E(s)$: evaluation score,$C(s)$: correct evaluation score, and $P(s)$: predictability score.

Algorithm 1: High Quality Sample Selection

```
Result: Input: Dataset $D$ and Models $M$:[Logistic Regression, SVM]; Hyper-Parameters $b$, $m$, $n$, $t$ and
        $tau$; Output: Pruned dataset $S$
a=0;
for $a < 100$ do
    Select $a$% random samples from $D$ and let $acc$ be IID accuracy of model $M$ at iteration $x$;
    if $acc(x) > acc(x-1)$ then
        a=a+b
    else
        a=a
    end
end
$D$= $a$% of samples from $D$;
Get $D$'s embeddings by finetuning RoBERTA on 10% of $D$ and discard this 10%;
$S = D$;
$E(s) = 0$ and $C(s) = 0$ for all $s$ in $S$ ;
while $\|S\| > n$ do
    forall $i \in m$ do
        Randomly select trainset of size t from S and let y =0;
        while y < 2 do
            Train M[y] on t and evaluate on rest of S i.e. V ;
            forall $s \in V$ do
                $E(s) = E(s) + 1$;
                if model prediction is correct then
                    $C(s) = C(s) + 1$
                end
            end
            y=y+1
        end
    end
    forall $s \in S$ do
        $P(s) = C(s)/E(s)$
    end
    Shortlist instances where $P(s) > tau$ ;
    Sort shortlisted instances based on DQI values and delete k lowest DQI instances
end
```

3.2 Dataset creation:

While dataset pruning verifies the hypothesis that *we don't need big datasets to learn a task when using pre-trained models*, other problems associated with pruning (Section 2), particularly resource and time wastage involved in creating biased data, remain unaddressed. We therefore plan to implement a DQI-in-the-loop approach as proposed in a recent work [28], to recreate datasets and answer our

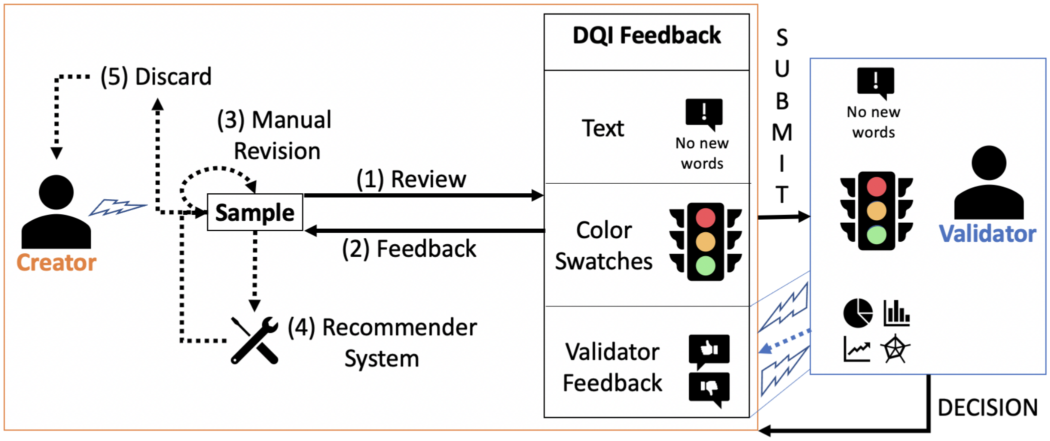

Figure 1: Data creation workflow. Creators have the choice of manually revising, fixing, or discarding samples. Validators may choose to provide feedback along with the decision to accept or reject samples. Dotted lines indicate steps where user choices are available.

second question– *Is a dataset created using DQI in a crowdsourcing setup equivalent to/better than one obtained via pruning?*

We propose a crowdsourcing workflow, as shown in Figure 1. This workflow will support data creators; newly created samples are evaluated by DQI, and feedback is given to the user about potential biases. Feedback can be shown to indicate if a particular aspect of the data created might lead to spurious bias– encouraging sample modification to decrease the presence of artifacts and increase generalization capability– using: (i) component-wise text feedback on specific sample characteristics, (ii) color-swatches, with easy-to-interpret traffic signal color coding (red, yellow, and green), and (iii) feedback from a data validator. We also intend to provide a recommender system based on DQI to further assist creators in converting red signals to green. We will experiment with these modules over varying granularities– i.e., feedback is provided at DQI component/sub-component/term levels. After potentially iterative cycles of feedback and revision, the sample will finally be submitted for bechmark inclusion.

Validators will evaluate submitted samples across different granularities in order to examine how each individual sample contributes to the overall quality of the current dataset state. This will utilize the first two feedback conditions presented to creators, along with visualizations, based on relevant text characteristics considered in each DQI component. The validators will then communicate sample decisions to the creators, with additional feedback (compulsory when the sample is rejected). This will enable continuous feedback about creator performance during creation, and also support data validators.

This framework aids in the creation of a new high quality dataset, and also enables opportunities for various novel applications (such as dataset refurbishment), inspiring the next generation of datasets and models.

4 Experimental Protocol

We perform an initial exploratory analysis and have promising results for dataset pruning. This is in line with our three step approach to mimic human learning (Section 2). We specifically address the third step– imparting background knowledge and heuristics in learning– and test 1 DQI term.

4.1 Exploratory Analysis:

In our preliminary experiments (Table 1), we utilize the first term of DQI C_1 (component 1) to prune SNLI [2] to $\sim 1-2\%$ of its original size (550K). When RoBERTA is trained with our pruned dataset, it achieves near-equal performance on the SNLI dev set, as well as competitive zero-shot generalization on: (i) NLI Diagnostics [43], (ii) Stress Tests [29], and (iii) Adversarial NLI [30]. This indicates that we might not need big datasets to learn a task.

4.2 Proposed Experiments:

We plan to further investigate high quality selection criterion by performing full scale pruning experiments on the SNLI[2] and MNLI [45], and SQUAD 1.1 [33] datasets. MNLI and SNLI will use the same OOD datasets as in Table 1. For SQUAD, we will use NewsQA[40], TriviaQA[16], SearchQA[6], HotpotQA[47], and Natural Questions[20] as OOD datasets, in line with a recent work [17]. We will also be recreating SNLI, MNLI, and SQUAD 1.1 using DQI-in-the-loop as demonstrated in in our workflow (Figure 1).

Pruning Experiments: We intend to prune based on additional terms in DQI [28]. DQI is calculated based on 7 components, 20 sub-components, and 133 terms. In order to short-list sub-components that we can reasonably expect to be useful, we will first conduct pruning on SNLI based on all 7 components, individually. We will compare IID accuracy for various pruned sizes (Table 1), and shortlist the two components that result in the highest IID accuracy of the pruned set. We will then prune with the terms of selected components (based on initial SNLI pruning), to varying sizes (similar to Table 1), to find out which DQI terms result in achieving higher IID accuracy, over all 3 datasets. We also plan to perform an ablation study of the DQI, AFLite and Coarse Action (Algorithm 1 lines 2-9) modules, by removing them from the algorithm, on the shortlisted components. In all these experiments our pruning happens purely based on IID test set accuracy. Zeroshot OOD evaluation is just done to ensure that the pruned dataset does not contribute mainly spurious bias.

In RoBERTA, we plan to change the learning rate from 1e-6 to 1e-5, and vary b as 100, 1000, 2000, and 5000. n is the target dataset size, and t is the training set size; we plan to vary both from 10 % of S (the pruned dataset) to 75% of S, in 15% increments. m will be varied from 8 to 124 in increments of either 16 or 32. Other hyperparameters will be fixed as per Hugging face transformers[46].

Expectations: In DQI, terms are synonymous with sub-components, except for C_2 and C_6, as these two components address quality at word, POS tag (adjective, adverb, noun, and verb), bigram, trigram, and sentence granularities; C_6 further calculates terms label-wise, which we will ignore for the purpose of pruning (-80 terms). If C_2 and/or C_6 are shortlisted based on component-wise pruning, they will contribute 8 and 40 terms respectively. In other cases, components will contribute 1-5 terms. We will prune all 3 datasets with the terms of selected components (based on initial SNLI pruning), to varying sizes, similar to Table 1. In recent work, word overlap [10, 3] and semantic textual similarity [27] have been dominant in producing spurious bias; we therefore expect to shortlist C_3 and C_5 in our component-wise experiments.

Previous work has found that the amount of artifacts in datasets is in the order: SNLI>SQUAD>MNLI [10, 3, 41, 38, 32, 28]. Accordingly, we expect the size of the equivalent pruned set (2% for SNLI) to be in reverse. Additionally, considering the human motivation for Algorithm 1, we expect our ablation experiments to affect performance in the order of DQI>AFLite>Coarse Action, with reverse order for effect on pruning time.

Creation Experiments: We will use our creation workflow (Section 3.2) in crowdsourcing, to recreate SNLI, MNLI, and SQUAD 1.1. For each respective dataset (without pruning), we will select the smallest pruned dataset size that results in IID accuracy within +/-5% of the original IID accuracy and create a similar size data using crowdsourcing setup. We will additionally perform ablation studies with subsets of creators, across the different quality feedback methods. We will be using the default hyperparameters mentioned in the DQI work [28].

Expectations: In the ablation studies, we expect number of samples, sample quality and IID/OOD performance to be affected in the following order: all feedback modes>recommender system>text feedback validator feedback>color-swatches. We expect the time involved to follow the reverse order.

Size (Random)	IID
5000	36.77
10000	77.45
15000	81.69

Size	IID	OOD ANLI			OOD NLI Diagnostics				OOD Stress Combined		
		R1	R2	R3	Knowl.	LS	Logic	PAS	Comp.	Distraction	Noise
550k	**89.64**	36.6	30.5	31.33	**57.64**	**62.23**	53.8	66.51	**51.63**	72.13	**79.52**
5000	87.47	32.6	31.8	28	50.35	61.14	48.37	**67.45**	35.29	65.72	73.97
10000	87.93	34.5	**33**	**31.67**	55.9	61.14	53.26	66.75	45.94	**74.88**	74.62
15000	88.95	**37.2**	28.3	29.17	56.6	56.79	**54.62**	65.8	45.94	70.66	77.71

Table 1: Left– Random Selection. Right– Pruned set results. Highlighted points: best performances.

References

[1] M. H. Bollen. Understanding power quality problems. In *Voltage sags and Interruptions*. IEEE press, 2000.

[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[3] R. L. Bras, S. Swayamdipta, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal, and Y. Choi. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*, 2020.

[4] C. Clark, M. Yatskar, and L. Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] M. Dunn, L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

[7] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[8] M. Gardner, Y. Artzi, V. Basmova, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, et al. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

[9] K. G. Grunert. Food quality and safety: consumer perception and demand. *European review of agricultural economics*, 32(3):369–391, 2005.

[10] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

[11] H. He, S. Zha, and H. Wang. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*, 2019.

[12] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

[13] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

[14] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*, 2019.

[15] A. P. Jones. Indoor air quality and health. *Atmospheric environment*, 33(28):4535–4564, 1999.

[16] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[17] A. Kamath, R. Jia, and P. Liang. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*, 2020.

[18] D. Kaushik and Z. C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*, 2018.

[19] D. Kaushik, E. Hovy, and Z. C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

[20] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[21] R. Le Bras, S. Swayamdipta, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal, and Y. Choi. Adversarial filters of dataset biases. *arXiv*, pages arXiv–2002, 2020.

[22] Y. Li and N. Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019.

[23] Y. Li, Y. Li, and N. Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[24] N. F. Liu, R. Schwartz, and N. A. Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. *arXiv preprint arXiv:1904.02668*, 2019.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[26] R. K. Mahabadi and J. Henderson. simple but effective techniques to reduce biases. *arXiv preprint arXiv:1909.06321*, 2019.

[27] S. Mishra, A. Arunkumar, C. Bryan, and C. Baral. Our evaluation metric needs an update to encourage generalization. *arXiv preprint arXiv:2007.06898*, 2020.

[28] S. Mishra, A. Arunkumar, B. Sachdeva, C. Bryan, and C. Baral. Dqi: Measuring data quality in nlp. *ArXiv*, abs/2005.00816, 2020.

[29] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*, 2018.

[30] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

[31] W. H. Organization. *Guidelines for drinking-water quality*. World Health Organization, 1993.

[32] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*, 2018.

[33] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[35] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

[36] R. Schwartz, M. Sap, I. Konstas, L. Zilles, Y. Choi, and N. A. Smith. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. *arXiv preprint arXiv:1702.01841*, 2017.

[37] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[38] S. Tan, Y. Shen, C.-w. Huang, and A. Courville. Investigating biases in textual entailment datasets. *arXiv preprint arXiv:1906.09635*, 2019.

[39] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[40] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016.

[41] M. Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. *arXiv preprint arXiv:1804.08117*, 2018.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[43] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[44] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[45] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[46] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[47] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[48] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.