Do We Need to Create Big Datasets to Learn a Task?

Swaroop Mishra[*] **Bhavdeep Sachdeva**[*]
Department of Computer Science, Arizona State University
{srmishr1, bssachde}@asu.edu

Abstract

Deep Learning research has been largely accelerated by the development of huge datasets such as Imagenet. The general trend has been to create big datasets to make a deep neural network learn. A huge amount of resources is being spent in creating these big datasets, developing models, training them, and iterating this process to dominate leaderboards. We argue that the trend of creating bigger datasets needs to be revised by better leveraging the power of pre-trained language models. Since the language models have already been pretrained with huge amount of data and have basic linguistic knowledge, there is no need to create big datasets to learn a task. Instead, we need to create a dataset that is sufficient for the model to learn various task-specific terminologies, such as 'Entailment', 'Neutral', and 'Contradiction' for NLI. As evidence, we show that RoBERTA is able to achieve near-equal performance on $\sim 2\%$ data of SNLI. We also observe competitive zero-shot generalization on several OOD datasets. In this paper, we propose a baseline algorithm to find the optimal dataset for learning a task.

1 Introduction

Large scale datasets such as Imagenet (Russakovsky et al., 2015) in Vision, and SQUAD (Rajpurkar et al., 2016) and SNLI (Bowman et al., 2015) in NLP have accelerated our progress in deep learning. The general trend has been to create large scale datasets for various tasks such as Abductive NLI (Bhagavatula et al., 2019), DROP (Dua et al., 2019), and SWAG (Zellers et al., 2018). The process of creating big datasets involves heavy investment in resources, that further increases when models are developed in response to these datasets, and trained to top leaderboards. This makes deep learning research and development inaccessible to communities where resources are scarce. Additionally, the heavy computation involved in training models adversely affects the environment on a broader scale (Schwartz et al., 2019). This leads us to the question: *Do we always need to create big datasets?*

We probe this question with motivation from the process by which we learn a certain topic/task. *Even though we have access to hundreds of materials available online, we do not need to go through all of them in order to learn the specific topic. In fact, we intentionally avoid certain materials which are noisy, distracting, or irrelevant to the topic.* Humans have deep background knowledge about the world which makes this possible. With the recent developments in language modelling, pre-training on huge datasets have imparted linguistic knowledge to models like BERT (Devlin et al., 2018) and RoBERTA (Liu et al., 2019). With this knowledge, models need not learn everything from scratch; instead they should just learn task specific terminologies such as 'Entailment', 'Neutral', and 'Contradiction' for NLI, which might not necessitate the use of big datasets.

There are certain other factors that recommend against creating big datasets. A growing number of recent works (Poliak et al., 2018; Geva et al., 2019; Kaushik and Lipton, 2018; Schwartz et al., 2017; Mishra et al., 2020; Bras et al., 2020) have exposed the presence of spurious bias in many popular benchmarks. Spurious bias represents unintended correlations between input and output (e.g.: the word 'not' is most often associated with the label 'contradiction'(Gururangan et al., 2018)). Spurious bias makes a task easy for models, allowing them to exploit instead of learning generalizable features like humans. Models finetuned on these benchmarks fail to generalize in Out of Distribution (OOD) and Adversarial settings. Since the sources of these spurious biases: data collection,

[*] equal contribution

Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pages 169–173
Online, November 20, 2020. ©2020 Association for Computational Linguistics

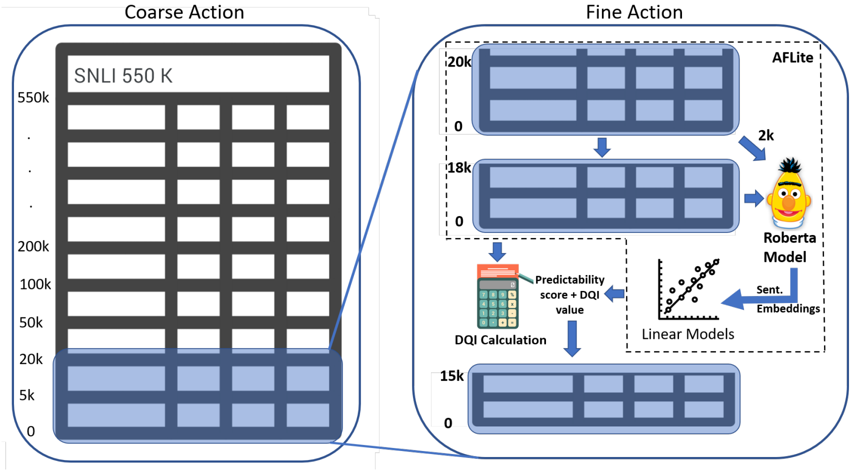

Figure 1: Proposed baseline approach to select the optimal data necessary to learn a task.

and crowdsourcing are hard to control, carefully selecting a smaller and optimal dataset may be a viable alternative.

We propose a baseline in this paper to find the optimal dataset for learning a task. Our approach is inspired by the human tendency to first make a rough estimate of the presence of relevant materials by glancing at various parts of an entire set of available materials. After selecting a slice of existing materials in the first phase, they remove redundant/easy/already known/possible distracting content from the slice. Finally, they use heuristics based on their background knowledge about the task to sort based on relevance and select the most optimal content based on the priority of the task and availability of time to learn it. We utilize several recently proposed modules in our baseline.

We prune SNLI (Bowman et al., 2015) to $\sim 2\%$ of its original size using our baseline. Our results show that RoBERTA on training with this pruned set achieves near-equal performance on the SNLI dev set and competitive zero-shot generalization on three OOD datasets (i) NLI Diagnostics (Wang et al., 2018), (ii) Stress Tests (Naik et al., 2018), (iii) Adversarial NLI (Nie et al., 2019). Our analysis shows that big datasets not only prevent generalization, but also impact IID testset performance. Interestingly, we find the annotation of those data to be correct and not noisy. This indicates that, certain data samples might be distracting a model by acting against the inductive bias created by rest of the dataset. Our finding opens up the possible existence of such distractors in real datasets, encouraging NLP community to explore the optimal selection of certain samples in a dataset instead of trying to dominate a leaderboard with the entire dataset.

2 Proposed Algorithm

We mimic the relevant material selection process in humans to propose algorithm for selecting the optimal dataset necessary to learn a task, as illustrated in Figure 1. We use robotics terminology (Rauch et al., 2019) to explain the stages of learning (i) coarse action (ii) fine action. Algorithm 1 details our approach. We briefly explain each stage below.

Formalization: Let D represent the entire dataset, s represent samples, M be the model, S be the pruned set, $E(s)$, $C(s)$ and $P(s)$ be the evaluation score, correct evaluation score and predictability score of each sample s respectively. In this preliminary work, we just explore the first term of DQI_{c1}. Expanding this to other terms will be the immediate future work.

Coarse Action: We start with a random subset of $a\%$ of dataset (D), train model (M) on it and calculate accuracy on the IID testset. We iteratively append a random subset of $b\%$ of data from the rest of D, train M on the combined data and calculate accuracy on the testset. We continue adding $b\%$ of data until the testset accuracy stops increasing. L1-L8 of algorithm 1 explains coarse action.

Fine Action: We use two key modules (i) AFLite (Bras et al., 2020; Sakaguchi et al., 2019) and (ii) DQI (Mishra et al., 2020) for fine action on the data selected after coarse action. AFLite is a recent technique for adversarial filtering of dataset biases, whereas DQI has a method to quantify quality of samples with or without annotation.

AFLite: In our setup, AFLite randomly selects 10% of data (selected after coarse action) for fine tuning on M, and then discards them. It randomly partition the data into train and test set, and does it in parallel several times. It trains linear models (logistic regression and SVM) with the train data, and evaluate on the test data. It combines parallel sessions by calculating predictability score ($P(s)$) of every data as the number of time it has been correctly predicted ($C(s)$) divided by the number of times it has been evaluated ($E(s)$). It then shortlists samples for which predictability score is greater than a threshold (tau).

DQI: DQI stands for Data Quality Index. It is a compilation of various linguistic parameters related to dataset biases. It has seven components –(i) Vocabulary, (ii) Inter-Sample N-gram Frequency and Relation, (iii) Inter-Sample STS (Semantic Textual Similarity), (iv) Intra-Sample Word Similarity, (v) Intra-Sample STS, (vi) N-Gram Frequency per Label, (vii) Inter-Split STS – that cover various possible inter/intra-sample interactions (a subset of which leads to biases) in an NLP dataset. DQI has a total of 20 subcomponents and 133 terms. Higher DQI is meant to indicate lower existence of spurious bias and higher generalizable features.

Leveraging AFLite and DQI in Fine Action: We use DQI in the pruning step of AFLite; instead of sorting samples based on the predictability score, we sort them based on the DQI values. L9-L34 and L34-36 of algorithm 1 explain the usage of AFLite and DQI respectively in fine action.

Size	Performance on IID test set
5000	36.77
10000	77.45
15000	81.69
20000	**84.69**
25000	80.96

Table 1: Coarse Action results on SNLI dataset

3 Results

Hyperparameters: We use $a = 5000$, $b = 5000$ and use other hyperparameters from AFLite (Bras et al., 2020) and DQI (Mishra et al., 2020) papers.

Analysis: Table 1 shows that IID testset accuracy decreases after 20k, so we stop there and proceed for fine action with 20k data. With fine action, we prune 20k data further to the size of 5k-15k, as shown in Table 2. Our results in Table 2 shows that the pruned datasets achieves near equal performance on IID testset and competitive performance on various sections of three OOD datasets. Since we have included just the first term of DQI_c1, we perform ablation study of that specific term. Our results in Table 3 shows that the first term of DQI_{c1} helps in improving performance on most of the cases. Interestingly we observe that, 20k data has lower IID testset accuracy than 5k, 8k, 10k, 12k and 15k datasets, as shown in Table 2.

Algorithm 1: Optimal Sample Selection

```
   Result: Input: Dataset $D$, Hyper-Parameters $a$, $b$, $m$,
      $n$, $t$ and $tau$ and Output: Pruned dataset $S$
1  for $a < 100$ do
2      Randomly Select $a\%$ of samples from $D$ and
        Check IID testset accuracy of model $M$;
3      if IID accuracy is increasing then
4          a=a+b
5      else
6          break
7      end
8  end
9  $D$= $a\%$ of samples from $D$;
10 Fine tune RoBERTA on 10 % of $D$ and get the
    embeddings of rest of the dataset $D$. Discard 10 %
    of $D$ used in Training.;
11 $S = D$;
12 $E(s) = 0$ and $C(s) = 0$ for all $s$ in $S$ ;
13 while $\|S\| > n$ do
14     forall $i \in m$ do
15         Randomly select trainset of size t from $S$ ;
16         Train Logistic Regression on t and evaluate
            on rest of $S$ i.e. $V$ ;
17         forall $s \in V$ do
18             $E(s) = E(s) + 1$;
19             if model prediction is correct then
20                 $C(s) = C(s) + 1$
21             end
22         end
23         Train SVM on t and evaluate on $V$ ;
24         forall $s \in V$ do
25             $E(s) = E(s) + 1$;
26             if model prediction is correct then
27                 $C(s) = C(s) + 1$
28             end
29         end
30     end
31     forall $s \in S$ do
32         $P(s) = C(s)/E(s)$
33     end
34     Shortlist instances for which $P(s) > tau$ ;
35     Sort shortlisted instances based on DQI values
        and delete $k$ instances with lowest DQIs
36 end
37
```

4 Discussion

We perform a preliminary analysis on the 15k samples retained using our algorithm and observe that the 15k retained data contains 4939, 5058 and 4983

Size	IID Test	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**
20k	84.69	33.1	32.2	30.42	39.93	51.9	39.95	63.21	33.79	57.77	61.84
5k	87.47	32.6	31.8	28	50.35	61.14	48.37	**67.45**	35.29	65.72	73.97
8k	87.54	34.7	31.5	28.92	51.74	55.98	51.63	65.57	40.21	68.99	75.08
10k	87.93	34.5	**33**	**31.67**	55.9	61.14	53.26	66.75	45.94	**74.88**	74.62
12k	88.56	32.6	32.7	30.67	49.31	57.61	50.82	66.27	39.03	67.84	73.67
15k	88.95	**37.2**	28.3	29.17	56.6	56.79	**54.62**	65.8	45.94	70.66	77.71

Table 2: Fine Action on the selected subset of SNLI post coarse action: Highlighted points have best performances. First row represent the original SNLI dataset of size 550k, second row represents the dataset of size 20k selected after coarse action. Last five rows (5k-15k) represent the dataset retained after pruning in fine action.

Size	IID Test	OOD ANLI			OOD NLI Diagnostics				OOD Stress Combined		
		R1	R2	R3	Knowl.	LS	Logic	PAS	Comp.	Distraction	Noise
5k	**86.76**	34.4	**29.8**	**27.75**	**50.04**	**56.52**	**47.01**	**65.33**	38.1	66.14	**72.01**
8k	**87.08**	**33.1**	**31.2**	**28.42**	**46.18**	56.52	**47.01**	**65.57**	**37.93**	**68.91**	**71.85**
10k	88.39	**33.5**	**31.6**	**30.42**	**53.47**	**60.05**	**47.28**	**66.27**	**40.05**	**67.02**	**73.26**
12k	**88.38**	33.9	**31.1**	**29.17**	51.04	**57.42**	**50.27**	66.98	**38.81**	69.42	76.32
15k	**88.92**	**35.4**	33.9	**28.5**	**49.31**	57.88	**51.9**	67.22	50.24	**70.45**	**75.07**

Table 3: Ablation Study for DQI_{c1}: Highlighted points show sections of various datasets where addition of DQI has resulted in higher performance

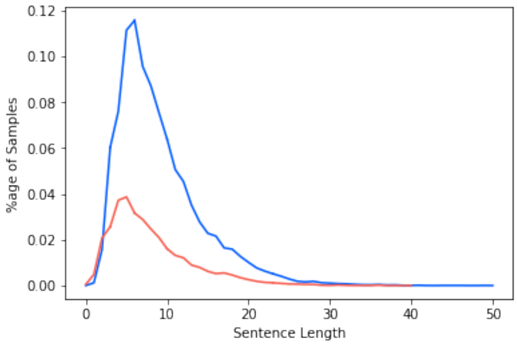

Figure 2: Sentence length vs. percentage of samples for the dataset retained (15k, Blue) and removed (5k, Red) after fine action

samples of contradiction, entailment and neutral respectively. This is similar to the distribution of the original SNLI dataset which has around 183k samples for each class. However, figure 2 illustrates that sentence length of retained and removed samples follow different distributions.

5 Conclusion

We propose a baseline approach to find the optimal set of samples required to learn a task. Our approach mimics humans in identifying relevant materials for learning a task. In the first stage, our algorithm finds a rough estimate as part of the coarse action. The second stage leverages two recently proposed modules AFLite (for adversarial filtering of dataset biases) and DQI (for quantifying the quality of data) to perform fine action. We show the efficacy of our baseline by pruning SNLI to 2% of its original size. Our results show that RoBERTA on training with this pruned set achieves near-equal performance on the SNLIdev set and competitive zero-shot generalization on three OOD datasets. Our analysis shows that big datasets not only prevent generalization, but also impact IID performance. Our findings about distracting samples will encourage community to look for the possible existence of such distractors in real datasets and subsequently explore the optimal selection of samples in a dataset instead of trying to dominate a leaderboard with the entire dataset. Studying the effect of our algorithm on model training time, memory footprint, model interpretation research and better understanding of how deep learning models work in general are some of the potential future directions to explore.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

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

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020. Dqi: Measuring data quality in nlp. *ArXiv*, abs/2005.00816.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

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

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Christian Rauch, Vladimir Ivan, Timothy Hospedales, Jamie Shotton, and Maurice Fallon. 2019. Learning-driven coarse-to-fine articulated robot tracking. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6604–6610. IEEE.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2019. Green ai. *arXiv preprint arXiv:1907.10597*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. *arXiv preprint arXiv:1702.01841*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.