IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Improve SQuAD fine-tuned ELECTRA model on Adversarial QA

Gunjan Srivastava

Department of Computer science & Data Science, University of Austin, Texas

Abstract: To evaluate the performance of model, our primary approach has been to measure accuracy of model on validation dataset. Benchmark datasets used to train and validate NLP models may often overestimate the performance. These observations all stem from the fact that a model may achieve high performance on a dataset by learning spurious correlations, also called dataset artifacts. The model is then expected to fail in settings where these artifacts are not present, which may include real-world testbeds of interest. Our aim in this paper is to evaluate pre-trained QA model fine-tuned on SQuAD benchmark dataset and rigorously test it against challenging dataset; identify different types of issues in the model; and further define an approach to improve the model performance on one of the specified issues.

Index Terms: Adversarial QA, LLM, Google ELECTRA, SQuAD

1. INTRODUCTION

Question Answering task (QA), or the ability to read the comprehension and answer questions on it can be a challenging task for machine. It requires them to understand local and global context, how humans interact, understand grammar, complex sentences, relationships, paraphrase, ambiguity in comprehensions etc. Consider question "What is the colour of the table?" posed in the small comprehension in Figure 1.

Context: There is an oval brown table in the room.

Question: What is the color of the table?
Answer: Brown

Figure 1: Example - Context, Question, answer

To answer this question, one must understand shape, size, colour and grammar to answer the colour is brown.

Our goal of the NLP model is generalization. It should be able to answer question with high accuracy given any comprehension and question type. The standard approach for evaluation is using training-validation split to estimate the accuracy of the model. While the performance on held out data is a useful indicator, held-out dataset is often not comprehensive and may suffer from **selection bias**. Further, analysing performance on a single aggregate statistic, it becomes difficult to know where the model is failing.

To build the NLP model for Question answering (QA) task, we picked pre-trained model ELECTRA-small (Clark et al., 2020) and fine- tuned it on SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al.,2016). SQuAD is a crowdsourced training data with Comprehension, questions, and Answer span of size 100K.

To measure model's performance – different evaluation approaches have been proposed to measure the robustness of the QA model, such as contrast sets (Gardner et al., 2020), adversarial challenge sets (Jia and Liang, 2017), model ablations (Kaushik and Lipton, 2018), and checklist sets (Ribeiro et al., 2020). We use **checklist** approach to synthetically build challenging datasets which helps us diagnose the models and uncover weakness on multiple contexts and types of questions. We will discuss detailed methodology of model diagnostics in section 3.
https://github.com/marcotcr/checklist

We picked an issue 'model robustness' or Adversarial scenarios where model failed to answer correctly when an irrelevant sentence gets added to comprehension, at the start/end or comprehension is challenging for model to understand the ambiguity and complex relationship between sentences in the paraphrase. We call this as adversarial challenging datasets.

2. FINE-TUNE ELECTRA ON SQUAD

To study the Machines effectiveness on question answering task (QA), we initialized our model as ELECTRA-small (Clark et al., 2020). ELECTRA is a small and a compute effective model as compared to BERT-small. It has better performance as compared to its peer model, as its task is defined over all inputs instead of just masked inputs, like in case of BERT. It uses generator-discriminator architecture to achieve this result. To fine-tune the ELECTRA model on QA tasks, we use SQuAD 100k questions for Machine comprehension task proposed by Rajpurkar et al., 2016. Through SQuAD, our model can learn to answer question on Dates, numbers, Person, Locality, Other Entity, Adjectives, Verb, Other. Figure 2 Summarizes the answer types in SQuAD data set.

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Figure 2. Answer types in SQuAD benchmark dataset credits: Rajpurkar et al.2016

To fine-tune the ELECTRA model on QA tasks, we use SQuAD 100k questions for Machine comprehension task proposed by Rajpurkar et al., 2016. Through SQuAD, our model can learn to answer question on Dates, numbers, Person, Locality, Other Entity, Adjectives, Verb, Other. Figure 2 Summarizes the answer types in SQuAD dataset.

To evaluate performance of model, we used Exact-Match(EM) and F1-Score both on Validation as well as our synthetically generated checklist dataset. Our SQuAD fine-tuned can achieve EM- **78.25 and F1-score of 86.04.**

3. CHECKLIST approach to measure model robustness

SQuAD fine-tuned model achieves a good result with EM score of 78.25 and F1-score of 86.04. However, has our model generalized well for all Machine comprehension examples and scenarios? Is our model consistent? Are the questions which we need to evaluate to understand model weakness before fixing them. To evaluate model performance, we will use checklist approach as described in Ribeiro et al 2020 to build test dataset for various scenarios.

SQuAD fine-tuned model achieves a good result with EM score of 78.25 and F1-score of 86.04. However, has our model generalized well for all Machine comprehension examples and scenarios? Is our model consistent? Are the questions which we need to evaluate to understand model weakness before fixing them. To evaluate model performance, we will use checklist approach as described in Ribeiro et al 2020 to build test dataset for various scenarios.

3.1 What is Checklist

Ribeiro et al 2020, proposed an approach to build CheckLists, a comprehensive behavioural testing of NLP models. It helps us define test cases to evaluate capability of our NLP model like, Vocabulary, Negation, Robustness, Named-Entity-recognition, POS etc.

It further defines an approach to evaluate capability on aspects like **MFT** (minimum functionality test, similar to unit test to check basic functionality of the model), **INV**(invariance test, or perturbations to inputs which should not change the output of the model), **DIR**(Directional expectations test – adding new sentence to the input should not change the output of the model).

3.2 Statistics of test dataset

We used GitHub repository(https://github.com/marcotcr/checklist) of Ribeiro et al 2020 to build the test cases and a challenging dataset. Tried integrating directly using the repository, however due to python incompatibility was forced to build synthetic dataset manually.

Manually created synthetic test dataset of various test cases around **Coreference** (he/she), **Negation** (Either or both Context and question has negation), **Robustness** (Adding adversarial sentence to context, introducing typos in the questions), **Taxonomy** (Distinguish Animal vs Vehicle, Nationality vs Jobs, more/less keywords in the context, match properties like color, shape etc.)

Temporal (Change in one person role with time, understanding who comes first/last, before/after), **Vocab** (Comparison great/less, nice/less nice open/less open, quite open). Statistics of the dataset is presented in Figure 3 and table 1.

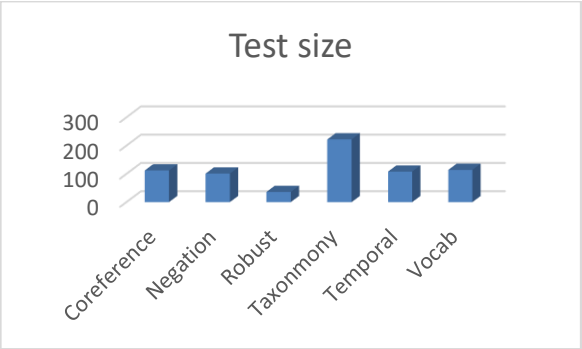

Figure 3: types of test cases and size of test data.

Table 1: Summary of Test cases

Capability	Test type	Description	No of records
Negation	MFT	Context /Question has negation	101
Robust	DIR	add irrelevant sentence to C	18
	INV	Swap adjacent characters in Q (typo)	18
Taxonmony	MFT	Animal vs Vehicle	88
	MFT	match properties to categories	134
Temporal	MFT	change in one person only	72
	MFT	Understanding before/after, last/first	36
Vocab	MFT	Comparison	114
Coreference	MFT	he/she	112
Grand Total			693

3.3 Performance of SQuAD fine-tuned model on Challenging dataset

To identify model weakness on various scenarios, we evaluated performance of SQuAD fine-tuned model on CheckLists challenging dataset (section 3.2 of this paper). Figure. 4 Presents comparison of Exact match score of our SQuAD fine-tuned model on Challenging test cases and Figure 5 Displays models F1-score on challenging datasets.

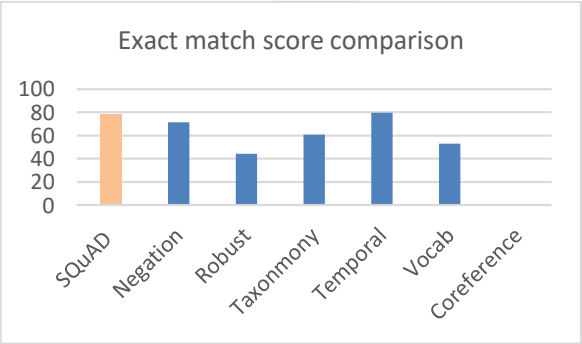

Figure 4: Originally our SQuAD fine-tuned model achieved EM score of **78.25**, when evaluated on challenging dataset, its performance (EM metrics) degraded across capability test and lowest in Adversarial Question (44) and Conference (0).

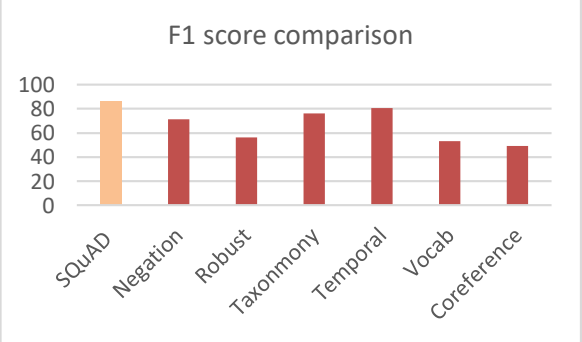

Figure 5: Originally our SQuAD fine-tuned model achieved F1-score score of **86.04**, when evaluated on challenging dataset, its performance (F1-score) degraded across capability test and lowest in Adversarial(robust) Question (**56**) , Vocab(**53**) and Conference (**49**).

We observed performance issues of our SQuAD fine-tune Electra model **across capability test** (section 3.2) created using checklist approach.

We found poor performance on challenging Adversarial QA tasks, where distractor was appended to the comprehension or complex question where model did not learn relationship among sentences in the complex passages and instead learnt superficial cues to answer, which resulted in incorrect prediction. We will pick model performance on Adversarial QA task as an area of improvement for in this paper.

4. Training model on Adversarial QA

4.1 Background

SQuAD1.1 (Rajpurkar et al., 2016) has been an established benchmark dataset for building model for question answer tasks. Data collection and annotation for NLP model development have been following crowdsourcing approach which have led to breakthrough; however, it does have some drawbacks like crowdsourcing by non-experts,
distant supervision, human error, and quality assurance.

This leads to dataset biases and model blind-spots where model learns spurious correlations and fails when evaluated on fine-tuned task and challenging questions and comprehensions.
To improve performance of our model on adversarial QA, we will use **model-in-the-loop** approach (Bartolo et al 2020) and use dataset generated with the help of BIDAF, BERT, RoBERTa model to build stronger datasets for the adversarial tasks.

4.2 What is model-in-the loop approach?

Used SQuaD1.1 dataset as input to model-in-the loop approach where similar annotation protocol was used, approach investigate the annotation setup where an annotator must compose a question for which the model predicts the wrong
answer. As a result, only the samples where model prediction was **wrong** were retained in the dataset.

Annotation strategy was applied to three different models, BIDAF, BERT, RoBERTa with each model being stronger compared to the prior. Models can, to a considerable degree, learn to generalise to more challenging questions, based
on training sets collected with both stronger and weaker models in the loop. Figure. 6 depicts the visual for model-in-the loop annotation strategy.

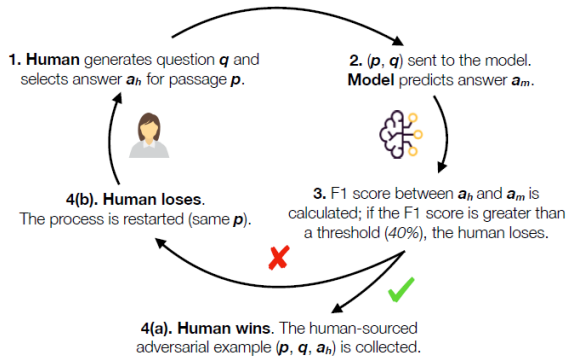

Figure 6: Overview of the annotation process to collect adversarial written questions from humans using a model in the loop. (Reference: Bartolo, Roberto et al 2020)

Dataset collected with model-the-loop approach is part of work done in "**Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension**" (Bartolo, Roberts et al 2020).

We used hugging face dataset repository for of "adversarial_qa". Table 2 provides general details on the number of passages and question-answer pairs used in the different dataset splits.

Table 2: Number of passages and question-answer pairs for each data resource. (Reference: Bartolo et al 2020)

	#Passages			*#QAs*		
Resource	*Train*	*Dev*	*Test*	*Train*	*Dev*	*Test*
$\mathcal{D}_{\text{SQuAD}}$	18,891	971	1,096	87,599	5,278	5,292
$\mathcal{D}_{\text{BiDAF}}$	2,523	278	277	10,000	1,000	1,000
$\mathcal{D}_{\text{BERT}}$	2,444	283	292	10,000	1,000	1,000
$\mathcal{D}_{\text{RoBERTa}}$	2,552	341	333	10,000	1,000	1,000

4.3 Experiment

We defined following hypothesis for the experiment:

a. How does our model when trained on each of these separate adversarial datasets perform on validation?
b. Does training in weaker adversarial dataset, i.e., dataset generated using weaker model help us generalize on stronger datasets?
c. How would our adversarial trained model perform on non-adversarial tasks?
d. Can generalization be achieved for both adversarial and non-adversarial task when trained on combined dataset SQuad+DIBAF|BERT|RoBERTa(Adversarial dataset)

A potential problem with the focus on challenging questions is that they might be very distinct from one another, leading to difficulties in generalization.

4.4 Observation

We trained our ELECTRA-small model on all possible combination of SQuAD + Adversarial dataset. When trained individually on each dataset SQuAD, BIDAF, BERT, RoBERTa – We observe following (statistics in Table 3):

i. Model trained on SQuAD generalizes better on BIDAF adversarial dataset. However, performance of SQuaD trained model degrades on stronger adversarial dataset BERT and RoBERTa.
ii. Performance of Adversarial trained model does not generalize well on SQuaD (Non-adversarial) dataset.

Further, we trained model **on SQuaD combined with different adversarial created datasets**, we observed following (statistics in Table 4):

i. Performance on challenging dataset improves when model was trained on SQuAD + one of Adversarial dataset (BIDAF, BERT, RoBERTa). Performance of combined dataset training improves as compared to model trained on just SQuaD dataset.

   For ex – EM score on BERT improves from 21.47 to 32.15 when trained on SQuAD + BERT. Similar observation can be made for BIDAF and RoBERTa.
ii. Performance of Model when trained on SQuAD + Adversarial dataset (BIDAF, BERT, RoBERTa) improves drastically as compared to model trained on just Adversarial dataset. For example – Performance on RoBERTa dataset improves from EM score of **14.70** to Similar observation can be seen for BIDAF and BERT.
iii. Model trained with **SQuAD + BIDAF** show good result in both **Adversarial** dataset (BIDAF, BERT, RoBERTa) and **Non- Adversarial** (SQuAD). SQuAD +BIDAF achieves EM score of ~32 across Adversarial dataset, just little behind performance of SQuAD+RoBERTA EM ~33. It's performance on Non-Adversarial (SQuAD) is comparable to model just trained on SQuAD. Achieve EM 72 and F1-score 82 (against original SQuAD fine-tuned model EM 78 F1-score 86).

Table 3. shows the results of our model when trained on each dataset separately and performance (EM and F1-score) across the datasets for each of instance of trained model.

Model	Trained On	Evaluation (Validation) Dataset							
		SQuaD		BIDAF		BERT		RoBERTa	
		EM	F1	EM	F1	EM	F1	EM	F1
ELECTRA-small	SQuaD	78.25	86.04	28.60	40.87	12.70	21.47	13.50	22.95
	BIDAF	43.63	56.86	15.30	26.28	11.40	21.32	12.90	21.15
	BERT	40.12	53.36	14.60	26.22	14.50	24.78	14.50	24.79
	RoBERTa	37.79	51.31	14.60	27.12	12.90	23.50	14.70	24.82

Table 4. shows the results of our model when trained on combined datasets complexity and its performance on Non-adversarial (SQuAD) and Adversarial dataset as the datasets gets stronger.

Model	Trained On	Evaluation (Validation) Dataset							
		SQuaD		BIDAF		BERT		RoBERTa	
		EM	F1	EM	F1	EM	F1	EM	F1
ELECTRA-small	SQuaD	78.25	86.04	28.60	40.87	12.70	21.47	13.50	22.95
	SQuaD+BIDAF	71.60	80.55	33.30	45.61	22.20	31.32	22.60	31.74
	SQuaD+BERT	69.35	78.11	33.30	45.30	22.80	32.15	22.20	31.65
	SQuaD+RoBERTa	66.80	76.40	30.40	42.54	21.90	31.76	24.10	33.23

5. Conclusion

We evaluated performance of SQuAD fine-tuned ELECTRA-small model on multiple capability tests using Checklist approach (Ribeiro et al 2020) and identified areas where model could not infer correctly. Example: Negation, Taxonomy, Co- reference, Vocab, Adversarial QA tasks etc. We saw model performed worst in case Adversarial QA tasks.

To improve the model on Adversarial QA tasks, we investigated RC annotation paradigm which define a model-in-the loop ("Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension", Bartolo et al 2020) approach for collecting dataset for Adversarial learning. Applying this approach with progressively stronger models in the loop (BIDAF, BERT, RoBERTa), we produced three separate adversarial datasets.

Training and fine-tuning our ELECTRA-small model on above dataset, we evaluated multiple hypothesis to check model generalization on adversarial task with progressively stronger datasets and on non-adversarial tasks. We found model when trained on SQuAD + BIDAF can improve the model for significantly from EM score of ~12 to ~32 (refer to Table 4) on all the three Adversarial dataset generated using BIDAF, BERT, RoBERTa and also perform well on SQuAD dataset EM 72 and F1-score 81 (non-adversarial task). Model-in-loop approach show a direction where model when trained on Adversarial dataset using Bartolo et al 2020 approach can achieve generalization across Adversarial and Non-Adversarial tasks.

6. Related works

Liu et al 2019 described "Inoculation by Fine- Tuning: A Method for Analysing Challenge Datasets" to evaluate performance of SQuaD trained model on adversarial tasks and presented approach to create challenging QA datasets for model assessment when trained on challenge datasets and it result on both challenge and original test.

References

[Rajpurkar et al.2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

[Clark et al.2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR).

[Ribeiro et al.2020] 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, July. Association for Computational Linguistics.

[Liu et al.2019] Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. In Proceedings of the 2019 Conference of the North American Chapter of the Association for7 Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2171–2179, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[Jia and Liang 2017] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proc. of EMNLP.

[Gardner et al.2020] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna 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.

[Kaushik and Lipton2018] Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5010–5015, Brussels, Belgium, October- November. Association for Computational Linguistics.

[Bartolo et al.2020] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. Transactions of the Association for Computational Linguistics, 8:662–678.