A reproduction and expansion of "Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension"

Student @ The University of Texas
Natural Language Processing

Abstract

This paper is an investigation into the work of (Bartolo et al 2020), which explored the benefits of using human-generated adversarial training examples fed into a model for further development as a "model-in-the-loop". We will take this work and try to re-use it in a cost-effective iterative tuning method leveraging pre-trained models based on the SQuAD and SQuAD v2 datasets. We show that this technique resulted in a significant reduction in performance and in some cases, total loss of critical predictive outputs. However, the datasets from the original paper are still proven worthy when they are concatenated to the SQuAD v2 dataset for fun fine-tuning and analysis.

1 Introduction

In the 2020 paper "Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension" (Bartolo et al 2020), the hypothesis and supporting discussion focus on the use of trained humans to create a base set of adversarial entries and then the use of models to create unique adversarial datasets relative to each model type.

The original paper also discusses the potential for adversarial datasets to become outdated as model technologies advance. So, part of this investigation is to check if the original datasets are still helpful in improving the performance of these models.

In the first part of this paper, we will explore the performance of the Stanford dataset SQuAD (Rajpukar et al 2016) and the advanced version that introduces a set of adversarial data on top of the original data SQuAD v2 (Rajpukar et al 2016) when trained with modern models.

In section two, we will deviate from the original paper to see if we can improve the performance results in section one. The differentiation to the original work will be in re-training models fine-tuned on SQuAD and SQuAD v2 with the paper's proposed adversarial data, whereas the original work trained on the proposed datasets or combined the original SQuAD with the new data into one extensive dataset for the training.

2 Comparison of Electra Models with SQuAD and SQuAD v2

To establish a starting point for the investigation in section 3, we will create fine-tuned models built from the HuggingFace (Wolf et all 2020) library for both base models and data.

The target models for this section will be the "google/electra-small-discriminator" 14M parameter model (ES) and the "google/electra-base-discriminator" 110M parameter model (EB) (Clark et al 2020).

EB is expected to produce better results than ES due to the significant changes to the hidden size in the model neural network architecture. However, this much larger model comes at the cost of nearly 4x the training time on the same dataset and only achieved a modest decrease in loss, as shown in Figure 1.

The SQuAD dataset contains 87.6K training samples and 10.6K validation samples with paragraphs of text, questions, and acceptable (ground truth) answers.

SQuAD v2 is an extension of SQuAD to include over 50K adversarial (unanswerable) questions, where the entry is related to an entry in the original SQuAD data. The goal is to strengthen the model by forcing it to understand when it should say, "I don't know." The final size of this dataset is 130K training rows and 11.9K validation rows.

2.1 Setup

All model training was done with a standard set of hyperparameters, fixed epochs, and a set random-seed to control variability in these results.

The models will be fine-tuned using both datasets, creating four new models. We will abbreviate the names of the resulting models as shown in Table 1.

	Models	
Datasets	electra-small	electra-base
SQuAD	ES-S	EB-S
SQuAD v2	ES-S2	EB-S2

Table 1: Notation for fine-tuned models

The resulting training from the four models shows an extensive range in performance as measured by loss. The ES models converge to similar loss, which makes sense because the data related to non-adversarial learning is the same between the different datasets. However, the EB models do not converge to similar loss within the limit of 3 epochs. This indicates that the adversarial entries impact the performance, and the model training needs to be extended by a few epochs.

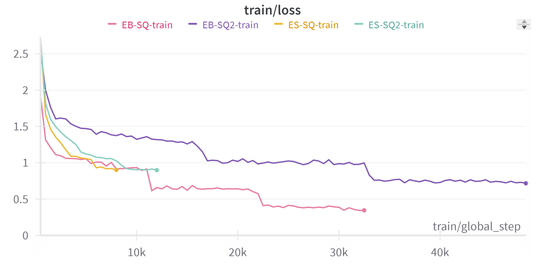


Figure 1: Training loss on the four fine-tuned models

2.2 Experimentation

We will run all four newly created foundation models against the two datasets, resulting in eight evaluations as indicated in Table 2.

2.3 Metric Math for Q&A in NLP

Standard confusion matrix math for performance differs from how QA models compute similar metrics. The primary metrics used are F1 and Exact Match.

F1 is defined as the harmonic mean of precision and recall and takes the standard form of

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{1}$$

The difference is in the calculations of Precision and Recall.

Precision (P) is the sum of the number of times a match is found between the prediction and an answer for all entries divided by the number of predictions in the evaluation.

$$P = \sum_{i=1}^{n} \frac{correct_i}{predictions_i} \tag{2}$$

Recall (R) is calculated as the sum for all entries of the number of times a match is found between the prediction and an answer divided by the count of answers in the evaluation.

$$R = \sum_{i=1}^{n} \frac{correct_i}{answers_i} \tag{3}$$

The new metric, exact match (EM), is the count of times at least one answer is matched by a prediction divided by the total number of evaluations.

$$EM = \frac{matches_i}{evaluations_i} \tag{4}$$

2.4 Performance Results

The first comparison is how each model has performed when tested against the SQuAD data (Fig 4). This evaluation has an apparent performance loss for the model trained on the SQuAD v2 dataset. This indicates that the introduction of the adversarial components is having an impact. In the F1 metric, it is difficult to say if this loss is significant. However, since EM is a measure of exact similarity between prediction and answer, it appears that the models trained on SQuAD v2 are weaker in their ability to answer the questions correctly.

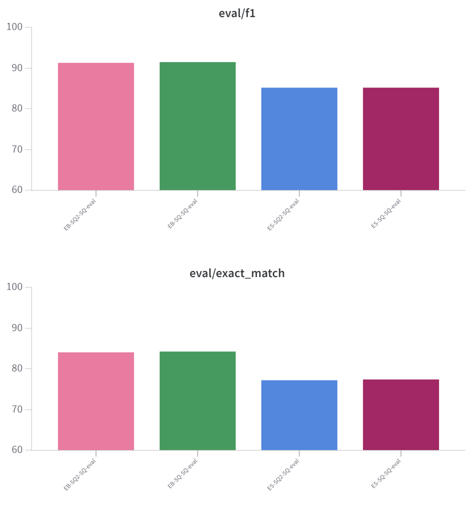


Figure 2: F1 and EM on SQuAD

The 2nd comparison (Fig 5) now examines how these models perform when challenged with V2

	Models			
Datasets	ES-S	ES-S2	EB-SQ	EB-SQ2
SQuAD (SQ)	ES-SQ-SQ	ES-SQ2-SQ	EB-SQ-SQ	EB-SQ2-SQ
SQuAD v2 (SQ2)	ES-SQ-SQ2	ES-SQ2-SQ2	EB-SQ-SQ2	EB-SQ2-SQ2

Table 2: Notation for evaluations based on fine-tuned models

evaluation data, which will include both v1 evaluation questions and adversarial questions. The resulting f1 scores now heavily favor the models trained on V2 data. This makes sense because the evaluation in V2 now includes a new measurement where the answers to adversarial questions should be “I don’t know,” represented as a prediction of Nothing or 0. The models trained on the V1 data will not have examples of this result and will try to make a prediction when there should not be one.

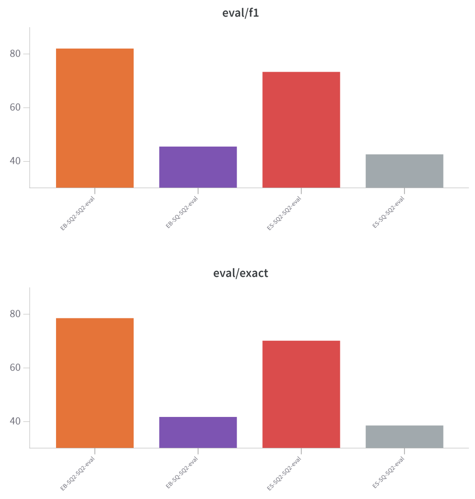


Figure 3: f1 and EM on SQuAD v2

If we explore deeper into the results through grouping and analysis with the Captum python package (Kokhlikyan et al 2020), the evaluation when using the SQuAD v2 dataset includes a breakdown of the metrics on the questions that have an answer, and when they do not (the adversarial entries). We can compare the models side-by-side, focusing solely on the V1 data 4. In both f1 and EM metrics, the models trained on V1 data perform significantly better than those trained on V2.

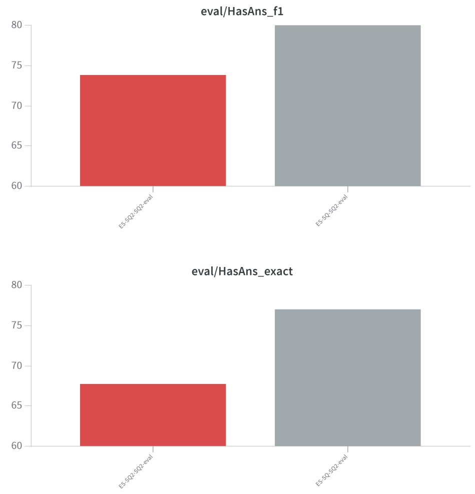


Figure 4: f1 and EM for "Has Answer" on SQuAD v2

2.5 Examples and discussion of errors introduced by Squad V2

Using the results shown in the Fig 6 charts, we can focus this section on the items being successfully predicted with the models trained on V1 data and not successfully predicted with the models trained on V2 data. This will set the stage for determining if iterative training using the data from the original work will improve these problem areas.

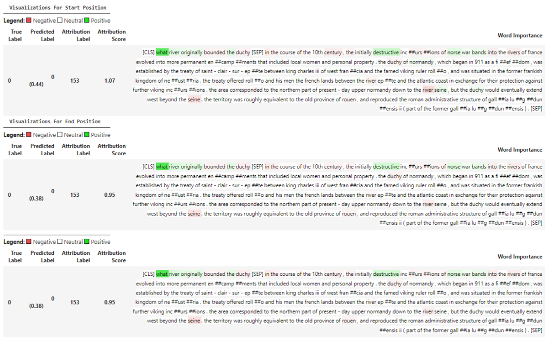
2.5.1 E1 – Predictions that are empty in the Squad V2 results

The biggest bucket of S2 errors is when there is no prediction. Since we are examining only the results for items with an answer, introducing the adversarial entities has weakened the model’s predictive ability. The top three observation on grouped error types is described below, with the contribution from each group indicated in Table 3

Example:

- Answers:
  - ’Seine’
  - ’Epte’
  - ’Seine’
- Prediction: [CLS]

We can see from this analysis that some confusion in the model caused the very low ranks of some critical words in the answers. The resulting



Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.44)	153	1.07	[CLS] what river originally bounded the duchy [SEP] in the course of the 10th century , the initially destructive inc ##urs ##ions of norse war bands into the rivers of france evolved into more permanent en ##camp ##ments that included local women and personal property . the duchy of normandy , which began in 911 as a fi ##ef ##dom , was established by the treaty of saint - clair - sur - ep ##te between king charles iii of west fran ##cia and the famed viking ruler roll ##o , and was situated in the former frankish kingdom of ne ##ust ##ria . the treaty offered roll ##o and his men the french lands between the river ep ##te and the atlantic coast in exchange for their protection against further viking inc ##urs ##ions . the area corresponded to the northern part of present - day upper normandy down to the river seine , but the duchy would eventually extend west beyond the seine . the territory was roughly equivalent to the old province of rouen , and reproduced the roman administrative structure of gall ##ia lu ##g ##dun ##ensis ii (part of the former gall ##ia lu ##g ##dun ##ensis) . [SEP]

Visualizations For End Position

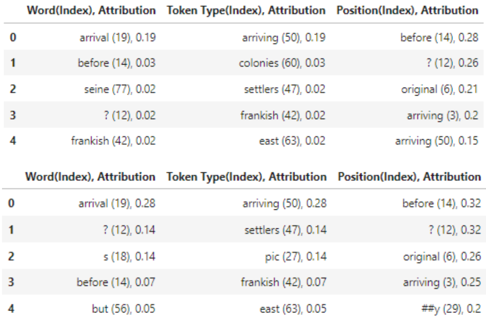
Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.38)	153	0.95	[CLS] what river originally bounded the duchy [SEP] in the course of the 10th century , the initially destructive inc ##urs ##ions of norse war bands into the rivers of france evolved into more permanent en ##camp ##ments that included local women and personal property . the duchy of normandy , which began in 911 as a fi ##ef ##dom , was established by the treaty of saint - clair - sur - ep ##te between king charles iii of west fran ##cia and the famed viking ruler roll ##o , and was situated in the former frankish kingdom of ne ##ust ##ria . the treaty offered roll ##o and his men the french lands between the river ep ##te and the atlantic coast in exchange for their protection against further viking inc ##urs ##ions . the area corresponded to the northern part of present - day upper normandy down to the river seine , but the duchy would eventually extend west beyond the seine . the territory was roughly equivalent to the old province of rouen , and reproduced the roman administrative structure of gall ##ia lu ##g ##dun ##ensis ii (part of the former gall ##ia lu ##g ##dun ##ensis) . [SEP]

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.38)	153	0.95	[CLS] what river originally bounded the duchy [SEP] in the course of the 10th century , the initially destructive inc ##urs ##ions of norse war bands into the rivers of france evolved into more permanent en ##camp ##ments that included local women and personal property . the duchy of normandy , which began in 911 as a fi ##ef ##dom , was established by the treaty of saint - clair - sur - ep ##te between king charles iii of west fran ##cia and the famed viking ruler roll ##o , and was situated in the former frankish kingdom of ne ##ust ##ria . the treaty offered roll ##o and his men the french lands between the river ep ##te and the atlantic coast in exchange for their protection against further viking inc ##urs ##ions . the area corresponded to the northern part of present - day upper normandy down to the river seine , but the duchy would eventually extend west beyond the seine . the territory was roughly equivalent to the old province of rouen , and reproduced the roman administrative structure of gall ##ia lu ##g ##dun ##ensis ii (part of the former gall ##ia lu ##g ##dun ##ensis) . [SEP]

Figure 5: E1 Visualization of word attributions



	Word(Index), Attribution	Token Type(Index), Attribution	Position(Index), Attribution
0	arrival (19), 0.19	arriving (50), 0.19	before (14), 0.28
1	before (14), 0.03	colonies (60), 0.03	? (12), 0.26
2	seine (77), 0.02	settlers (47), 0.02	original (6), 0.21
3	? (12), 0.02	frankish (42), 0.02	arriving (3), 0.2
4	frankish (42), 0.02	east (63), 0.02	arriving (50), 0.15

	Word(Index), Attribution	Token Type(Index), Attribution	Position(Index), Attribution
0	arrival (19), 0.28	arriving (50), 0.28	before (14), 0.32
1	? (12), 0.14	settlers (47), 0.14	? (12), 0.32
2	s (18), 0.14	pic (27), 0.14	original (6), 0.26
3	before (14), 0.07	frankish (42), 0.07	arriving (3), 0.25
4	but (56), 0.05	east (63), 0.05	##y (29), 0.2

Figure 6: E1 Word attribution for start and end tokens

prediction associates the text with the classification token [CLS] common in BERT-style models. When introducing the new adversarial examples, it is expected that the target words can move up the attribution list and help remove these errant predictions of [CLS] and None.

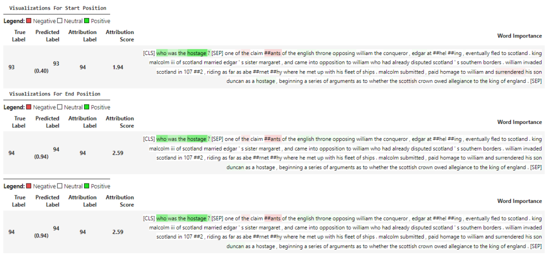
2.5.2 E2 – An answer is in the prediction

In this case, the prediction contains extra words or information that is not expected to be part of the answer. The prediction is almost correct and, in some cases, might be an acceptable version of an answer. However, it is not an "exact match".

Example:

- Answers:
  - 'Duncan'
  - 'Duncan'
  - 'Duncan'
- Prediction: son Duncan

In this example of error type 2, we can see that the target single-word answer is high in the positional index for both the start and end tokens. However, it is not marked high in the word importance plots. For this specific example, the word "son" is an appositive noun and could be strongly associated with the noun Duncan, resulting in the incorrect prediction. In this case, it is possible that adversarial entries can help separate such noun-noun compounds and create a cleaner person-associated prediction.



Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
93	93 (0.40)	94	1.94	[CLS] who was the hostage ? [SEP] one of the claim ##ants of the english throne opposing william the conqueror , edgar at ##hel ##ing , eventually fled to scotland . king malcolm iii of scotland married edgar ' s sister margaret , and came into opposition to william who had already disputed scotland ' s southern borders . william invaded scotland in 107 ##2 , riding as far as abe ##rnet ##hy where he met up with his fleet of ships . malcolm submitted , paid homage to william and surrendered his son duncan as a hostage , beginning a series of arguments as to whether the scottish crown owed allegiance to the king of england . [SEP]

Visualizations For End Position

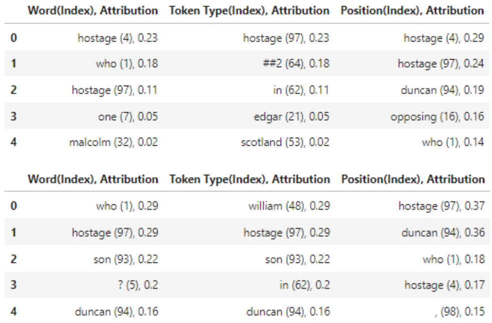
Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
94	94 (0.94)	94	2.59	[CLS] who was the hostage ? [SEP] one of the claim ##ants of the english throne opposing william the conqueror , edgar at ##hel ##ing , eventually fled to scotland . king malcolm iii of scotland married edgar ' s sister margaret , and came into opposition to william who had already disputed scotland ' s southern borders . william invaded scotland in 107 ##2 , riding as far as abe ##rnet ##hy where he met up with his fleet of ships . malcolm submitted , paid homage to william and surrendered his son duncan as a hostage , beginning a series of arguments as to whether the scottish crown owed allegiance to the king of england . [SEP]

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
94	94 (0.94)	94	2.59	[CLS] who was the hostage ? [SEP] one of the claim ##ants of the english throne opposing william the conqueror , edgar at ##hel ##ing , eventually fled to scotland . king malcolm iii of scotland married edgar ' s sister margaret , and came into opposition to william who had already disputed scotland ' s southern borders . william invaded scotland in 107 ##2 , riding as far as abe ##rnet ##hy where he met up with his fleet of ships . malcolm submitted , paid homage to william and surrendered his son duncan as a hostage , beginning a series of arguments as to whether the scottish crown owed allegiance to the king of england . [SEP]

Figure 7: E2 Visualization of word attributions



	Word(Index), Attribution	Token Type(Index), Attribution	Position(Index), Attribution
0	hostage (4), 0.23	hostage (97), 0.23	hostage (4), 0.29
1	who (1), 0.18	##2 (64), 0.18	hostage (97), 0.24
2	hostage (97), 0.11	in (62), 0.11	duncan (94), 0.19
3	one (7), 0.05	edgar (21), 0.05	opposing (16), 0.16
4	malcolm (32), 0.02	scotland (53), 0.02	who (1), 0.14

	Word(Index), Attribution	Token Type(Index), Attribution	Position(Index), Attribution
0	who (1), 0.29	william (48), 0.29	hostage (97), 0.37
1	hostage (97), 0.29	hostage (97), 0.29	duncan (94), 0.36
2	son (93), 0.22	son (93), 0.22	who (1), 0.18
3	? (5), 0.2	in (62), 0.2	hostage (4), 0.17
4	duncan (94), 0.16	duncan (94), 0.16	, (98), 0.15

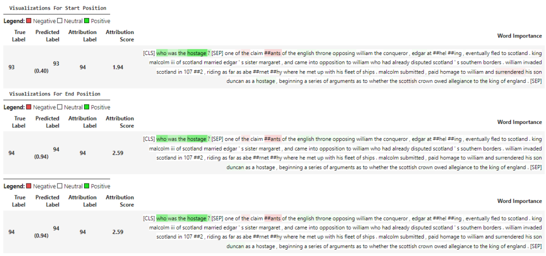
Figure 8: E2 Word attribution for start and end tokens

2.5.3 E3 – The prediction is in an answer

Similar to the prior error mode, in this case, the prediction has insufficient information to make it an exact match to an answer. However, the prediction can be found in one of the answers.

Example:

- Answers:
  - 'set of triples'
  - 'triple'
  - 'the set of triples (a, b, c) such that the relation a × b = c holds'
- Prediction: triples



Visualizations For Start Position

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
93	93 (0.40)	94	1.94	[CLS] who was the hostage ? [SEP] one of the claim ##ants of the english throne opposing william the conqueror , edgar at ##hel ##ing , eventually fled to scotland . king malcolm iii of scotland married edgar ' s sister margaret , and came into opposition to william who had already disputed scotland ' s southern borders . william invaded scotland in 107 ##2 , riding as far as abe ##rnet ##hy where he met up with his fleet of ships . malcolm submitted , paid homage to william and surrendered his son duncan as a hostage , beginning a series of arguments as to whether the scottish crown owed allegiance to the king of england . [SEP]

Visualizations For End Position

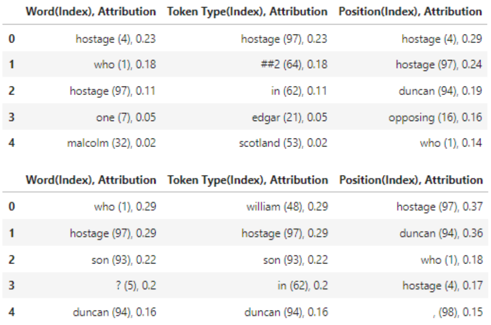
Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
94	94 (0.94)	94	2.59	[CLS] who was the hostage ? [SEP] one of the claim ##ants of the english throne opposing william the conqueror , edgar at ##hel ##ing , eventually fled to scotland . king malcolm iii of scotland married edgar ' s sister margaret , and came into opposition to william who had already disputed scotland ' s southern borders . william invaded scotland in 107 ##2 , riding as far as abe ##rnet ##hy where he met up with his fleet of ships . malcolm submitted , paid homage to william and surrendered his son duncan as a hostage , beginning a series of arguments as to whether the scottish crown owed allegiance to the king of england . [SEP]

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
94	94 (0.94)	94	2.59	[CLS] who was the hostage ? [SEP] one of the claim ##ants of the english throne opposing william the conqueror , edgar at ##hel ##ing , eventually fled to scotland . king malcolm iii of scotland married edgar ' s sister margaret , and came into opposition to william who had already disputed scotland ' s southern borders . william invaded scotland in 107 ##2 , riding as far as abe ##rnet ##hy where he met up with his fleet of ships . malcolm submitted , paid homage to william and surrendered his son duncan as a hostage , beginning a series of arguments as to whether the scottish crown owed allegiance to the king of england . [SEP]

Figure 9: E3 Visualization of word attributions



	Word(Index), Attribution	Token Type(Index), Attribution	Position(Index), Attribution
0	hostage (4), 0.23	hostage (97), 0.23	hostage (4), 0.29
1	who (1), 0.18	##2 (64), 0.18	hostage (97), 0.24
2	hostage (97), 0.11	in (62), 0.11	duncan (94), 0.19
3	one (7), 0.05	edgar (21), 0.05	opposing (16), 0.16
4	malcolm (32), 0.02	scotland (53), 0.02	who (1), 0.14

	Word(Index), Attribution	Token Type(Index), Attribution	Position(Index), Attribution
0	who (1), 0.29	william (48), 0.29	hostage (97), 0.37
1	hostage (97), 0.29	hostage (97), 0.29	duncan (94), 0.36
2	son (93), 0.22	son (93), 0.22	who (1), 0.18
3	? (5), 0.2	in (62), 0.2	hostage (4), 0.17
4	duncan (94), 0.16	duncan (94), 0.16	, (98), 0.15

Figure 10: E3 Word attribution for start and end tokens

This failure type appears to be true "near misses." In this example, the word importance shows that "triple" appears many times more than "triples." And the word "triples" is not in the start or end attribution lists. However, there is an apparent problem in the data where hash marks separate one of the answers from being considered with a "set of triples" showing up as a "set of triple ##s."

This could be addressed with better pre-filtering data to remove punctuation from the source and restore better word forms. For this exercise, we must see if the human-created adversarial components might reduce this mistake and lead to a positive prediction.

2.6 Contribution by error type

Error type	% of errors	% of population
E1 - None	28.8%	4.6%
E2 - Excess	4.3%	0.7%
E3 - Sparse	3.1%	0.5%
Totals	36.2%	5.8%

Table 3: Contribution % based on error type

3 Iterative Training Using D(RoBERTa)

One base dataset $adversarialQA$ and three tuned datasets are in the adversarial_qa repository (Hugging Face Inc 2021).

3.1 Setup

To simplify the experiment and align (partially) the source of this “model-in-the-loop” dataset with the foundations of the Electra models, we will use only one of the three available datasets from the original paper, D(RoBERTa). We will focus exclusively on Electra-small in the future since that is the basis for groups identified in Section 2.

The Electra-small fine-tuned model, ES-S2, will be used as the source model in a new fine-tuning training session using D(RoBERTa). This new model, ES-S2-DR, will then be evaluated against ES-S and ES-S2 to produce new evaluations for comparison and investigation.

3.2 Did we improve on the prior error types?

In all three error types, the iterative tuning technique resulted in a significant loss of performance. The impact against all errors is shown in Table 5.

3.2.1 E1 – Predictions that are empty in the Squad V2 results

The new models produced no entries with "None" as the prediction. This was surprising since we expected the original fine-tuning results to stay functional and only learn from introducing the new data.

3.2.2 E2 – An answer is in the prediction

There was an overall increase in the number of items that fit into this category, jumping from 0.7% of the population to 2.0% as shown in Table 5

3.2.3 E3 – The prediction is in an answer

Similarly, the incident rate for items where the prediction is actually inside one of the ground truths has gone up from 0.5% of the population to 1.4%.

Error type	% of errors	% of population
E1 - None	0.0%	0.0%
E2 - Excess	11.1%	2.0%
E3 - Sparse	7.7%	1.4%
Totals	18.8%	3.4%

Table 4: Contribution % after iterative training

3.3 Conclusion

This didn’t work. The analysis of each error type suggests that there is too much overlap between the adversarial examples in SQuAD V2 and the adversarial examples created in the original work. The iterative training is possibly re-learning against these new examples and losing predictive power gained through the updated SQuAD V2 data.

4 One Last Attempt - Reproduce (sort of) the Original Paper

We are making one last investigation to see what will happen if we combine SQuAD V2 with the D(RoBERTa) dataset from the original paper. The original paper included performance results for training on a combination of SQuAD and D(RoBERTa), so this effort will advance against the original experiments.

4.1 Setup

This training will be done on the ES-S2 fine-tuned model as ES-S2+DR.

We will compare the results against the ES-S2 and ES-S2-DR models to see if the increase in errors in any of the three failure types is at least restored to the performance found in the fine-tuned models from Section 2.

4.2 Results

The initial measurement of f1 and EM looks promising, as shown in Figure 11. The unified data has fixed the problem of missing None predictions, resulting in equivalent full performance on the Square v2 evaluation. When looking specifically at the items with an answer similar to the analysis in Section 3, we can see that performance has slightly improved.

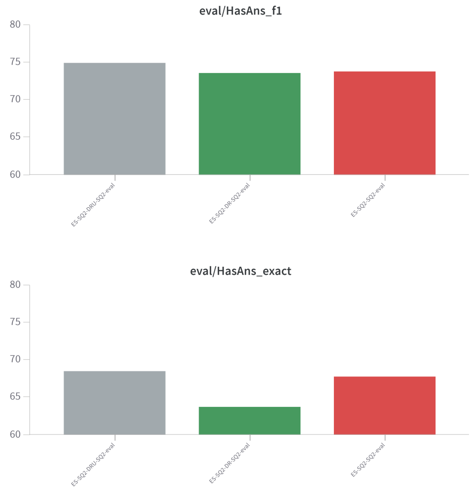


Figure 11: F1 and EM for "Has Answer" items after training with unified SQuAD v2 and D(RoBERTa)

Error type	% of errors	% of population
E1 - None	6.1%	1.0%
E2 - Excess	4.7%	0.5%
E3 - Sparse	2.1%	0.3%
Totals	12.9%	1.8%

Table 5: Contribution % after iterative training

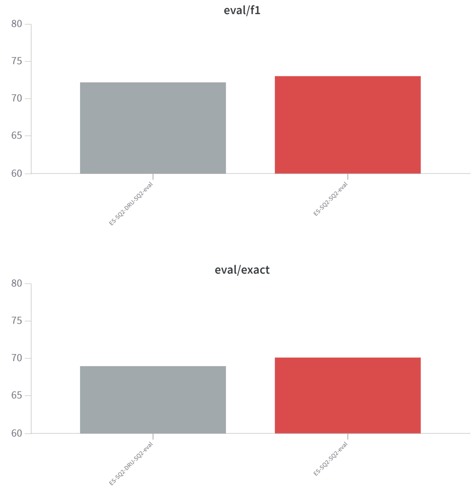


Figure 12: F1 and EM after training with unified SQuAD v2 and D(RoBERTa)

5 Conclusion

The initial purpose of this investigation was to determine if it was possible to use the models proposed in "Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension" as iterative learning agents on top of more modern models and datasets.

The essential idea is to reduce the overall computational cost for model advancement by focusing on small subsets of critical pieces of data. In this case, possibly due to collisions between the new data and the datasets from the paper, the result was not supportive of the hypothesis. For the SQuAD v2 dataset, simply introducing the D(RoBERTa) dataset for incremental training on a fine-tuned model is not a viable solution.

However, as an extension to the original plan, it appears that the paper's original work to concatenate their datasets with other data is still viable. The paper used SQuAD v1 data concatenated with D(RoBERTa) with results that outperformed their basic models. We moved that idea forward by combining SQuAD v2 with D(RoBERTa) and demonstrated an improvement in performance against SQuAD v2 alone.

References

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

[2] Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*. ICLR, 2020. https://openreview.net/pdf?id=r1xMH1BtvB

[3] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush. *Transformers: State-of-the-Art Natural Language Processing*, 2020. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[4] Hugging Face Inc.. *Huggingface Datasets*. GitHub repository, 2021. https://github.com/huggingface/datasets

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. arXiv e-prints, 2016. https://arxiv.org/abs/1606.05250

[6] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, Orion Reblitz-Richardson. *Captum: A unified and generic model interpretability library for PyTorch*. arXiv preprint, 2020. https://arxiv.org/abs/2009.07896