Eliminating data artifacts: the use of adversarial data to more efficiently train ELECTRA-small entailment across datasets

Author Anonymized – Solo Project

Abstract

This study trains ELECTRA-small entailment models on full-sized and down-sampled Natural Language Inference (NLI) datasets and tests their performance on in and out of domain evaluation sets. Despite high in-domain accuracy, the base, non-adversarial models struggle to adapt to adversarial datasets without the inclusion of adversarial training data. We compare base and best performing models' accuracy on specific adversarial groupings and observe roughly equal performance boost across all the measured categories.

1 Introduction

Language processing systems, such as natural language inference (NLI) entailment classification, require large quantities of data to train effectively. Datasets designed for training and evaluation may contain artifacts representative of patterns that exist across the train and evaluation splits of prepared data. Resultant models may achieve high performance on in-domain evaluation despite overfitting to certain unintended artifactual patterns.

The use of adversarial datasets to stress-test a trained model can uncover common errors that do not exist in the training datasets (Jia and Liang, 2017; Wallace et al., 2019; Bartolo et al., 2020; Glockner et al., 2018; McCoy et al., 2019). Additional training on adversarial or contrast data is a common method to improve model performance across testing conditions (Lie et al., 2019; Zhou and Bansal, 2020; Morris et al., 2020). The diversification of training sets represents a balancing act both of sacrificing accuracy for one test set in favor of another as well as for processing time and data availability, where inflating the test set increases training time and provides diminishing returns on accuracy.

Nie at al. (2020) train three models (BERT, XLNet, and RoBERTa) on publicly available datasets: (1) the Stanford Natural Language Inference (SNLI; Bowman et al., 2015); (2) the Multi Natural Language Inference (MNLI; Adina et al., 2018); and (3) the Adversarial Natural Language Inference (ANLI; Nie et al., 2020). All three datasets are detailed in section 2. Nie et al. identify model improvement by combining training sets from all three sources and note that the adversarial data appears to more efficiently train the models.

This study trains an ELECTRA-small (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) model on the SNLI, MNLI, and ANLI datasets. We measure the performance from each possible combination of training sets across the evaluation splits for all three datasets. Additional training datasets are generated by replacing some non-adversarial data from SNLI and MNLI with data from the ANLI set to determine whether the model can achieve similar results despite a decrease in overall training data.

The base model (SNLI) and best model (SNLI + MNLI + ANLI) perform similarly on SNLI data (89.7% vs 89.6%) but show significant improvement over adversarial data (30.6% vs 52.6%). Down-sampled models show slightly worse performance than their full-sample counterparts.

Errors from the base and best performing models are compared and categorized to better understand model blind spots. We randomly sample the errors from each model and manually assign categories to explain each example's difficulty. The model makes mistakes within each adversarial category at

a rate proportional to that category's presence in the overall dataset. Similarly, the best model's improvement within those categories roughly matches the overall dataset composition.

2 Methodology

Entailment data consists of three components: a hypothesis (a statement we wish to verify), a premise (textual evidence for whether the hypothesis is true), and a label (the ultimate truth of whether the hypothesis is correct). The learning agent is tasked with assessing the entailment for each premise/hypothesis pair by assigning 0 (entailed), 1 (neutral), or 2 (contradiction). We compare the assigned label with the gold-label from the dataset to determine accuracy.

Each of our tested models features identical architecture: we use ELECTRA-small for its small size and documented high performance (Clark et al., 2020). Each model trained using default settings for the trainer from the HuggingFace API over three epochs with a batch size of 64 (Wolf et al., 2020).

We retrieve training and evaluation data from the following public datasets:

1. The Stanford Natural Language Inference dataset contains over 570,000 premise/hypothesis pairs hand-created by humans completing image captioning tasks.
2. The Multi Natural Language Inference dataset is a crowdsourced dataset of over 433,000 premise/hypothesis pairs modeled after the SNLI dataset, but with the intention of adding more genres of text.
3. The Adversarial Natural Language Inference dataset contains over 160,000 premise/hypothesis pairs specifically designed by humans to confuse AI agents. This data is broken into three categories: r1, r2, and r3.

In addition to the above standard datasets, we also create down-sampled versions of the SNLI and MNLI datasets such that their sizes match that of the ANLI training set, effectively replacing large portions of the non-adversarial data with adversarial data (as seen in Nie et al., 2020). The full list of training sets is shown in table 1 below.

Dataset ID	Description
S	SNLI training set
M	MNLI training set
ANLI	The combined ANLI r1 + r2 + r3 training sets
SM	Combined SNLI and MNLI training sets
SM ANLI	Combined SNLI, MNLI, and ANLI r1+r2+r3 training sets
S ANLI	Combined SNLI and ANLI r1+r2+r3 training sets
M ANLI	Combined MNLI and ANLI r1+r2+r3 training sets
S*	Randomly selected a portion of S equal to |ANLI|
S*ANLI	S* combined with ANLI
S*M*ANLI	S*ANLI combined with a randomly selected portion of M equal to |ANLI|

Table 1: training dataset descriptions

Once training is complete, each model is then assessed over the evaluation sets from SNLI, MNLI, and the ANLI r1, r2, and r3 datasets. We compare the raw percentage of examples with correct entailment that the various models achieve.

Individual errors made by the S and SM ANLI models on the SNLI and ANLI r1 evaluation set are sampled and manually annotated to better understand specific areas where the model makes errors, as well as which types of errors improve with the addition of adversarial data. Three subsets of 50 samples were randomly taken from the SNLI data: (1) from base model errors, (2) best model improvements, and (3) places where the best model performs worse than the base model. Similarly, two sets of 100 samples were randomly sampled from the erroneous data of the base model and from the examples fixed by the best model in the ANLI evaluation data.

We sort the erroneous data samples according to the categories described by Nie et al. (2020) in their review of the ANLI dataset:

- numerical reasoning – counting or comparing numbers;
- name/pronoun references – tracking subjects or understanding gendered names;
- standard inference – applying comparatives, negation, or cause and effect;
- lexical – identifying synonyms or antonyms;
- tricky inference – tracking complicated sentence restructuring;
- reasoning/facts – knowing a fact about the world or reasoning based on presented facts; and
- quality – the provided label is incorrect.

They also provide estimated percentages for the proportion of adversarial data represented by each category in the ANLI dataset. Note that one example can be assigned multiple category tags. We compare the manually annotated data output from our models' errors on the ANLI data with the percentage estimates from Nie et al. (2020) to assess which categories may be particularly difficult for the model.

3 Results

The percentage of evaluation examples correctly labelled by each model are shown in table 2. The base SNLI model achieves an 89.7% success rate over the SNLI test set, but below 31% for all adversarial test sets, and only 70.8% for the MNLI test set. No combination of models achieves over ~31% on adversarial data sets without the inclusion of specific adversarial data in training.

Trained dataset	**SNLI**	**MNLI**	**ANLI_r1**	**ANLI_r2**	**ANLI_r3**
Majority label	*33.9*	*35.4*	*33.4*	*33.4*	*33.5*
S	89.7	70.8	30.6	29.6	30.7
M	76.3	82.2	24.6	29.9	30.4
SM	89.7	81.7	25.2	29.0	30.7
ANLI	61.5	65.0	52.9	40.7	42.4
S ANLI	89.6	74.9	50.9	41.3	41.8
SM ANLI	89.6	82.1	52.6	42.4	39.0
S*	87.3	68.8	31.0	31.2	31.5
S* ANLI	87.5	73.7	51.5	40.7	41.5
S*M* ANLI	87.8	80.7	52.4	39.2	40.7

Table 2: model accuracies for different training data and evaluation data. S = SNLI; M = MNLI; * = down-sampled. All units are shown as percent correctly labelled out of total evaluation set size.

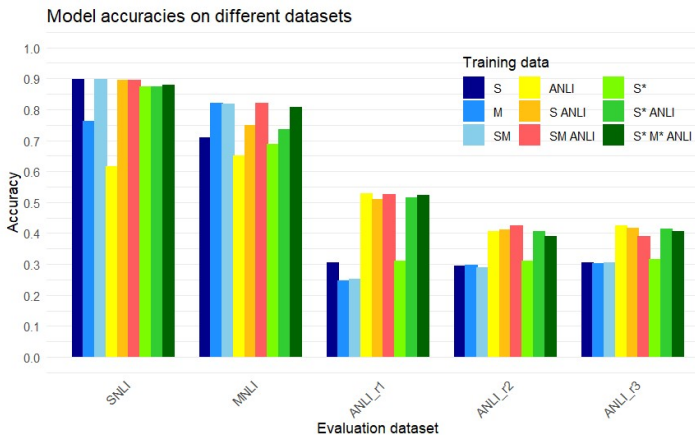

Figure 1: evaluation accuracy for each training dataset compared for each evaluation set.

The highest performing model on average uses the full training datasets from SNLI, MNLI, and ANLI r1, r2, and r3. It achieves 89.6% on SNLI, 82.1% on MNLI, and 52.6%, 42.4%, and 39% on the ANLI evaluation sets, respectively. Generally, the strongest boost towards accuracy on any evaluation set comes from the inclusion of that sets corresponding training partition.

Down-sampled training models reach similar accuracy levels but fall 0-2% short of their equivalent full-sample models; only for the r3 data does the down-sampled model outperform the full-dataset counterpart.

Since the base (S) and the best (SM ANLI) models achieve similar accuracy over the SNLI test set, we consider the changes in error type in SNLI data to see if the model truly did not change, or if it instead shifted which categories it identifies well. In total, just over 220 examples swapped from incorrect to correct (out of 9842 total examples or ~1377 correctly predicted examples), and vice versa, when comparing the base and the best models. We sampled 50 examples each from the errors made by the base model, from the 220 improved examples, and from the 225 worsened examples.

Figure 2 shows the results of this sampling below.

No singular category leaps out as driving the error rate in the base SNLI model, nor does one category dominate the improvements of the best model. Note that the SNLI data, unlike ANLI, was not constructed to confuse AI agents and, as such, the most common errors simply arise from an unrelated premise-hypothesis pairing; around 85% of the sampled errors involved a neutral label (either assigned or missed), which slightly overshoots the expected proportion based on majority-label statistics.

Since the best model showed more significant improvement over adversarial data than non-adversarial, we also consider areas of improvement in the ANLI dataset. We compare the categorical rates in our sampled data with the published summary by Nie et al. (2020) for the ANLI dataset, which allows us to determine whether our models disproportionately struggle with any one category of error relative to the overall composition of the dataset. We compare the full dataset estimate (n=500) with our sampled estimates of base model errors (n=100) and best model improvements (n=100). The categories reported by Nie et al. (pink) are compared with our sampled data (blue) in figure 3 below.

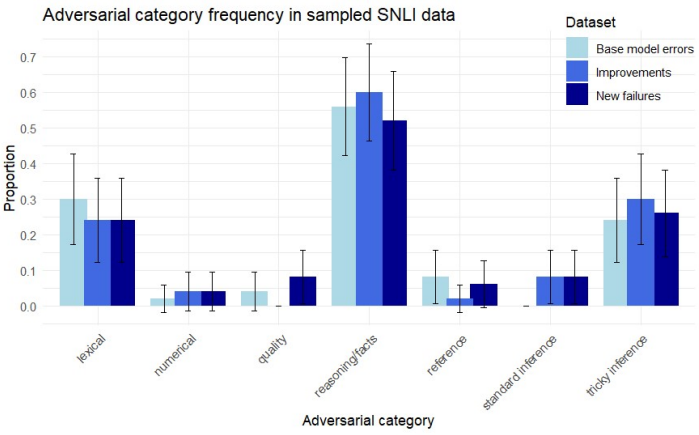

Figure 2: error categories sampled from SNLI data for erroneous and fixed examples. Units represent the proportion within each sample that fit into each category.

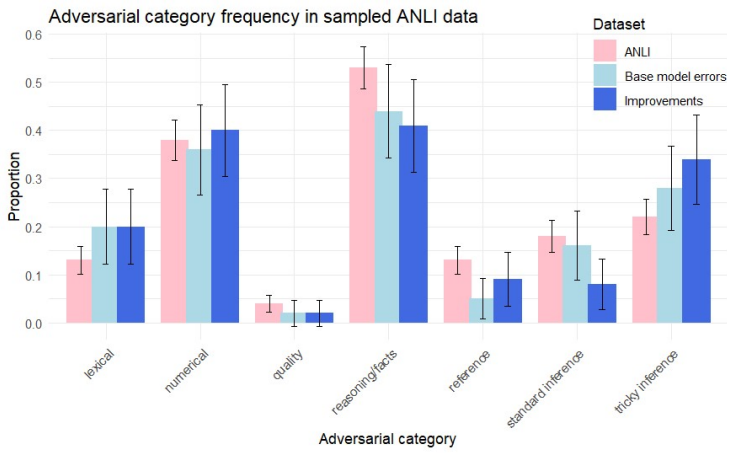

Figure 3: error categories reported in Nie et al. (2020) compared with sampled erroneous data in S and SM ANLI models

Most error categories fall within a reasonable range of the expected proportion in the ANLI dataset, which suggests that the model struggles roughly equally with the various types of adversarial data as they are presented in the dataset. The sampled errors include slightly fewer examples of reference confusion, but otherwise resemble the full ANLI dataset.

Similarly, the improvements (examples the model initially got wrong but fixed for the SM ANLI model) generally follow these same proportions, indicating that the model managed to learn across all categories and did not suffer from some kind of blind spot over one type of error. In other words, the categories roughly share the ~20% increase in overall model performance for the ANLI data.

4 Discussion

Despite the high accuracy it achieves on its own test set, the base SNLI model performs quite poorly on the adversarial dataset. Even against the MNLI data, which should be more similar, it drops 20% in accuracy, which indicates that it likely overfit to spurious trends within its training set. The fact that each evaluation set only reaches its peak accuracy when trained with its own training data indicates some presence of statistical artifacts that result in overfitting for all the datasets.

Only the ANLI trained dataset performs better on another evaluation set than its own. In part,

this may occur simply because significant work would be required to achieve a much higher accuracy on the adversarial data, whereas the non-adversarial data can be learned up to a certain accuracy with less specificity in training examples. Notably, however, the down-sampled SNLI data outperforms the ANLI data on the MNLI evaluation set, which indicates that adversarial data does not necessarily provide advantages over non-adversarial data in all cases.

When considering the role of data artifacts in affecting model performance, we must remember that these datasets are not truly independent. The SNLI dataset first established certain methodological conceptions for forming a dataset. The MNLI dataset later formed to intentionally expand the scope of the SNLI data (Adina et al., 2018). Finally, the ANLI dataset exists specifically to challenge the models created by the SNLI or MNLI datasets (or similar datasets; Nie et al., 2020). The inclusion of each example in the ANLI dataset is in part predicated on the inability of SNLI trained models to accurately label those examples.

We should consider these datasets' structures when we evaluate the individual errors the models made on the ANLI evaluation set. Individuals intentionally trying to repeatedly fool a language model created each example, which results in many of the same repeated tricks in the data. The categories described by Nie et al. (2020) were later assigned based on a sampling of the data. The proportion of challenge categories within the dataset may represent the relative difficulties the base model has on adversarial data, since only examples that challenged a model were included in the first place. Even though we sample our results to check for general overlap in categories, the assignment of categories also involves a subjective process that could easily differ between reviewers.

Not all examples within each category pose the same difficulty to the models, either. Upon qualitative review of the sampled data, the SNLI model appears to struggle to associate dates or times in any meaningful way, typically growing confused as soon as it needed to conduct chronological calculations. It also frequently mis-associates similar looking words (such as mall and hall) or fails to adapt to sentences with just one word changed. The base S model also particularly struggles to recognize an absence of supporting information. Though it improves in both areas with the addition of adversarial training data, it never reaches complete accuracy on either issue. In particular, the trained model continues to assume certain information is there due to the presence of related words (beach and water, for example). Although it improves somewhat on each of the categories defined by Nie et al., it never fully learns any one category. Moreover, the labels in the datasets do not always assign neutrality or contradiction in the same manner (likely due to human error or subjectivity), which would make mastery exceedingly difficult without overfitting to the dataset.

5 Conclusion

Despite attaining high accuracy on in-domain datasets, an ELECTRA-small model trained on exclusively SNLI data struggles to classify adversarial data. The addition of further datasets improves the models' performance on the evaluation splits for each additional dataset. The added data aids performance on external datasets but does not significantly affect accuracy on the SNLI data. Down-sampled non-adversarial data training sets produces models that fall slightly short of their full-sampled equivalents.

A comparison of adversarial categories in base datasets with errors made by the base S model and the best performing SM ANLI model does not identify any singular type of error category that dominates the model errors. The trained model appears to learn on the different types of adversarial data roughly equal to their presence in the adversarial dataset. Similarly, the accuracy turn-over on the SNLI data appear compositionally similar.

Further steps on this topic could include parameter searching to maximize the potential accuracy for each model. Similarly, the addition of more datasets might provide further perspective, since the SNLI, MNLI, and ANLI datasets were created with some ideological similarities (or intentional dissimilarities). One such option would be the NLI fever dataset

(Thorne et al., 2018). Alternatively, the error category analysis would benefit from a more robust approach to categorizing error categories such that subjective interpretations become less impactful. Modifying the evaluation set so that each adversarial category maintains equal presence could also elucidate the relative learning capacity of these models over different categories.

Acknowledgments

This paper was written following instructions provided under University of Texas at Austin's Natural Language Processing course, taught by Dr. Greg Durrett. Thank you to the full instructional staff of the course for providing knowledge and support that were instrumental in completing this project.

References

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

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

Clark, Kevin, Minh-Thank Luong, Quoc V. Le, and Chistopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv preprint (2020).*

Glockner, Max, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Jia, Robin and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Mccoy, Tom, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Morris, John X., Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *Conference on Epirical Methods in Natural Language Processing (2020).*

Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Nie, Yixin, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Ribeiro, Marco Tulio, 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, Association for Computational Linguistics.

Wallace, Eric, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint (2019)*.

Zhou, Xiang, and Mohit Bansal. 2020. Towards Robustifying NLI Models Against Lexical Dataset Biases. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Association for Computational Linguistics.