Mitigating Dataset Artifacts Through Fine Tuning and Ensemble-Based Training

Carlos Barbosa and Diego Sariol
The University of Texas at Austin

1 Abstract

This study explores the power and generalizability of the ELECTRA-small model in the context of Natural Language Inference (NLI). We evaluate ELECTRA's performance on the Stanford Natural Language Inference (SNLI) and Adversarial NLI (ANLI) datasets, as well as an available dataset comprising of unique and contrastive examples. Our analysis shows insights into the model's ability to handle typical examples alongside its sensitivity to minimal yet critical changes in input that alter the semantic relationship between the sentence pairs. The results highlight both the strengths and limitations of ELECTRA-small in understanding nuanced language variations; therefore, contributing to ongoing discussions about the effectiveness and reliability of language models in the task of NLI.

2 Introduction

Natural Language Inference (NLI) is a critical task in the field of natural language processing, where the objective is to determine the relationship between two sentences: typically, a premise and a hypothesis. This task plays a pivotal role in understanding and interpreting human language, with applications stretching from text summarization all the way to question and answering. The introduction of transformer-based models like ELECTRA has revolutionized NLI by introducing architectures that power-up self-attention mechanisms to better capture contextual relationships in text. Unlike traditional models that process text in sequence, transformers can process all words, allowing a more complete and nuanced understanding of the presented language. However, despite their sophistication, the robustness and generalizability of these models, particularly in handling diverse and nuanced linguistic scenarios, remain an open question. This paper aims to investigate the abilities of the ELECTRA-small model in deciphering complex language constructs, focusing on its performance in scenarios involving subtle linguistic variations.

Our approach requires a comprehensive evaluation of the ELECTRA-small model on established benchmarks such as the Stanford Natural Language Inference (SNLI) dataset, the Adversarial NLI (ANLI) dataset, and the “Martin Nguyen” Contrast dataset. An additional model is created using ensemble-based learning where a baseline model is established and trained on a difficult dataset, which allows our ELECTRA model to train on the residual (Clark et al., 2019). This allows to expand the limits of the model's capabilities; we extend our investigation to include unique and contrastive examples within these datasets and even introduce a publicly available dataset created specifically to introduce lexical differences (Glockner et al., 2018). These examples are precisely crafted to introduce minimal yet semantically significant alterations, posing a challenge to the model's ability to accurately interpret nuanced linguistic signs. Contrastive examples, a key element of our evaluation strategy, involve modifications to original examples that change their meaning or logical inference using additional information or small lexical changes (Gardner et al., 2020). We created our own self-annotated validation set of 50 examples as well as utilizing a publicly available contrast set composing of over 33 thousand training examples. This technique helps in measuring the model's ability to discern fine-grained semantic differences, a crucial aspect of language understanding.

Furthermore, we investigate unique examples that diverge from standard training paradigms. These examples are designed to test the model's ability to generalize and maintain

consistent performance in the face of atypical or adversarial conditions, further examining its understanding of complex language constructs. The importance of this study lies in its detailed examination of the current capabilities and limitations of NLI models, particularly those based on transformers. By analyzing the model's performance across a diverse corpus of examples, from standard to adversarial crafted, we aim to uncover the details of language understanding and the extent to which these four models can generalize beyond typical training scenarios. Our findings, whether the model accuracies increase or not, are expected to provide valuable insights for developing more robust, reliable, and nuanced NLI systems, paving the way for future advancements in natural language processing.

3 Contrastive Analysis

Analyzing the ELECTRA natural language inference model, the performance was examined on the SNLI validation set and a self-annotated contrastive dataset. These datasets contain Premise Hypothesis pairs labeled as entailment (0), neutral (1), or contraction (2). The goal is to evaluate how well ELECTRA-small can make predictions when faced with these complex textual inferences.

The confusion matrices for the two different evaluation sets show that ELECTRA-small displays a tendency to play it safe by predicting "neutral" rather than committing to entailment or contradiction judgments. For the original SNLI validation set, the model can achieve 89.2% accuracy which sets a baseline for the second half of this paper. Out of 1054 examples, it incorrectly labeled entailment relationships as neutral 238 times, and contradictions as neutral 231 times. Additionally, direct confusions between entailment and contradictions occurred 141 times, indicating issues with disambiguating on these classes. **Table 1** shows the confusion matrix of label classification of the NLI model trained solely on SNLI dataset and **Table 2** follows the same principle to illustrate the labeling output for the same model on the contrastive dataset.

		Gold Label		
		0	1	2
Predicted	0	0	229	77
	1	238	0	231
	2	64	215	0

Table 1 - NLI SNLI on SNLI

		Gold Label		
		0	1	2
Predicted	0	0	4	2
	1	3	0	9
	2	5	6	0

Table 2 - NLI SNLI on Contrast set

When evaluating the same NLI model on the self-annotated contrastive validation set, it is only able to achieve 42% accuracy. This is less than half of the total examples being predicted correctly, out of a total of 50 examples. This contrastive dataset was created by selecting unique Premise and Hypothesis pairs from the original SNLI validation set of 1054 examples. To properly create a contrastive dataset, the introduction of negation, contradictions, and even small lexical differences are made to essentially “fool” the model and its ability to make accurate predictions on examples it may have already seen.

By examining some of ELECTRA's errors, we can better understand where it fails in making legitimate logical inferences:

- In an SNLI example, the premise states, "A child is playing in the park" while the hypothesis claims "The child is at school." This contradictory pair was incorrectly predicted as neutral.
- In another SNLI case, the model failed to recognize the entailment between "A musician is playing the guitar" and the more general statement "The musician is playing an instrument."

Similar issues appeared in the contrastive dataset:

- Unable to reconcile the contradiction between “A man in the middle east with a corn-on-the-cob cart selling corn.” and “An old middle eastern man is buying corn-on-the-cob from his brother's cart.” ELECTRA guessed neutral.

- Given the premise of "Under a blue sky with white clouds, a child reaches up to touch the propeller of a plane standing parked on a field of grass." and hypothesis of "A child is reaching to touch the hull of a plane.", the model should have identified the obvious difference between the two parts of the airplane. However, it predicted entailment.

In summary, while ELECTRA shows promising language inference capabilities, it struggles with semantically complex examples that require deeper comprehension. Further training focused on contrasting examples and targeted linguistic phenomena to enhance the model's inference capacities. Building a richer understanding of the relationships between textual concepts will help move toward more human-like language mastery. The following next two sections will discuss the pipeline process for constructing and training three new models to further evaluate results in accuracy performance on the contrastive datasets and the original SNLI validation set.

4 "Breaking NLI" / Contrast Fine Tuning

The (Glockner et al., 2018) research study set out to disrupt NLI predictions and thus make it much more difficult for the model to accurately predict a correct label by introducing small lexical differences. Unlike the ANLI dataset, the examples differ from SNLI by "at most one word." To properly fine-tune the already pre-trained model, training resumption can be performed with the new dataset. This unique and new set of challenging examples for the model will allow for it to adapt to a new dataset with its pre-existing training and weights from the SNLI training set. The limitations and flaws of the model become more apparent as the difficulty of the subtle lexical differences increase as this can completely alter the Premise Hypothesis pair.

When analyzing the evaluation results of this fine-tuned model, similar mistakes are made once again, which reveals that this specifically designed dataset created to break NLI predictions, does decrease the original SNLI validation set accuracy to 75.2% seen in **Table 3** and decreases the contrast set accuracy to 42% seen in **Table 4**. Analyzing the evaluation results for the second fine-tuned (FT) model on the "Martin Nguyen" dataset can be seen in **Table 5** and **Table 6**. This newly fine-tuned model increased the contrast evaluation set to 50% accuracy and decreased the original SNLI validation set accuracy to 88.1%. By examining some of the fine-tuned ELECTRA's errors, we can better understand where it fails and where it progresses.

		Gold Label		
		0	1	2
Predicted	0	0	137	23
	1	170	0	21
	2	688	1404	0

Table 3 - NLI on Fine Tuned SNLI

		Gold Label		
		0	1	2
Predicted	0	0	2	1
	1	2	0	2
	2	4	15	0

Table 4 - NLI on Fine Tuned - Contrast set

By examining some of ELECTRA's errors, we can better understand where it fails in making legitimate logical inferences:

- In a SNLI set example, the Premise reads "A woman prepares ingredients for a bowl of soup." The Hypothesis states that "A soup bowl prepares a woman." The correct gold label should be entailment; however, the model is unable to understand the lexical difference of the sentence pairs and predicts entailment.
- In a contrast set example, the Premise reads "A group of young people with instruments are on stage." The Hypothesis states that "People are performing as a band on stage for a crowd." The correct gold label should be entailment; however, the model is unable to understand the semantic context of the sentence pairs and predicts neutral.

Predicted \ Gold Label	0	1	2
0	0	175	65
1	320	0	388
2	74	151	0

Table 5 - NLI on Fine Tuned Nguyen on SNLI

Predicted \ Gold Label	0	1	2
0	0	3	4
1	5	0	6
2	0	7	0

Table 6 - NLI on FT Nugeyen on Contrast

5 Ensemble-Based Training

Another interesting approach to consider when dealing with dataset artifacts is to introduce ensemble-based training. The concept involves constructing a baseline model which is comparatively simpler to the ELECTRA-small transformer, training it on a dataset known to have difficult challenges for ELECTRA, and then resume residual training with the more advanced model. This process begins by constructing and initializing a baseline Logistic Regression model and the ANLI dataset. All three rounds of this dataset will be combined, and this model will train, update its weights, and make its own predictions. Once this stage of the pipeline has been completed, all that remains is the model's vectorizer and the saved model itself. Going back to the previous section and reusing the SNLI dataset means that this significantly "weaker" model can be utilized to essentially prune the SNLI training set to remove all Premise and Hypothesis pairs that were predicted correctly. This leaves a filtered dataset behind which ELECTRA-small can resume training on.

Analyzing the evaluation results will be done once again on both the original SNLI validation set and our self-annotated contrast set to understand the influence of ensemble-based training as it decreases both the SNLI validation and contrast test set accuracies to 82.2% and 40% respectively which can be seen in **Table 7** and **Table 8**.

Predicted \ Gold Label	0	1	2
0	0	448	222
1	326	0	291
2	184	496	0

Table 7 - NLI Ensemble SNLI

Predicted \ Gold Label	0	1	2
0	0	5	5
1	4	0	6
2	1	9	0

Table 8 - NLI Ensemble SNLI - Constrast

This decrease in both validation and contrast sets can better be understood below:

- In a SNLI set example, the Premise reads “Two small children are gathering water from a large container.” The Hypothesis states that “Two kids are collecting water from a stream.” The correct gold label should be contradiction; however, the model predicts neutral which reveals it does not understand the difference in item context.
- In a contrast set example, the Premise reads "An Indian woman is washing and cleaning dirty laundry at a lake and in the background is a kid who appears to have jumped into the lake." The Hypothesis states that “An Indian woman is doing her clean laundry in a lake.” The correct gold label should be contradiction; however, the model is unable to understand the single lexical difference and predicts entailment.

6 Detailed Comparison

On NLI, the F1 score is a critical indicator of a model’s efficacy, a balanced measure of precision and recall. Our investigation contrasts four distinct approaches: the standard NLI applied on SNLI dataset, an ensemble method utilizing multi NLI models on SNLI, a fine-tuned NLI model of SNLI and a fine-tuned model on the “Martin Nguyen” contrast dataset.

The standard NLI on SNLI performed with an F1 score of 0.893, highlighting its proficiency in

standard scenarios. The ensemble method, when applied to SNLI dataset, results in an F1 score of 0.821, which, while still lower than the standard NLI on SNLI, reflects a robust performance. However, the ensemble approach when was evaluated on contrastive dataset, its F1 score dropped to 0.401. This decrease highlights the challenges that ensemble models face in adapting to contrastive scenarios that deviate from the patterns learned from the original SNLI dataset.

To contrast, the fine-tuned NLI model present a different result. On the SNLI dataset, its F1 score was 0.748, which is a decrease from the standard NLI model but suggests an expected trade-off to the model being optimized for a broader set of challenges. Notably, when evaluated on the contrast dataset, the fine-tuned model scored an F1 of 0.422, outstanding the ensemble method in this more complex context. On "Martin Nguyen" contrast dataset, the evaluation on contrast yields a F1 score of 0.504. This is notably higher than the other model's performance on the same evaluation method, suggesting that the "Martin Nguyen" approach improved adaptation to contrastive scenarios. In the other side, on the SNLI dataset, the Nguyen achieved a F1 score of 0.882, just slightly under the standard NLI model, but still showing a significant proficiency.

The contrast dataset scores particularly show the adaptive strengths and weakness on the tested models. While the standard NLI model's F1 score on the contrastive dataset was 0.463, it demonstrates that even a model that performs well on a standard dataset can struggle with more complex examples. Similarly, the Nguyen dataset's performance across different contexts highlights is flexibility and robustness in handling varied NLI challenges. Figure 1 illustrates the overall F1 score's performance of each dataset and their evaluation method.

7 Conclusion

Through the creation of four differently initialized and trained models, the influence of a model's pre-training can be better understood. The first model was solely trained on the SNLI training set to set a baseline for the following three models to be created. The second model utilized the previously created model; however, additional training was performed on all three rounds of the ANLI training set to further fine-tune the pre-existing and pre-trained SNLI model. Similarly, the third model was the original SNLI trained model, fine-tuned on the "Martin Nguyen" HuggingFace contrast dataset. The fourth and final model was constructed using ensemble-based training. This meant that a simple Logistic Regression model was instantiated, initialized, and trained on the ANLI dataset once again. This model was then used to filter the SNLI training set to prune out examples that the Logistic Regression model predicted correctly. The remaining examples could be utilized to train a new SNLI model on the residual examples.

The error types discovered from the in-depth analysis and comparison of the four models reveals ELECTRA-small's inability to make accurate predictions when faced with both difficult semantic and contextual relationships between the Premise and Hypothesis pairs. These underlying issues with the model's ability to make logical inferences whether it be negation and or contextual relation, reveal the further influence of additional training through fine-tuning or ensemble-based training. Future improvements that can aid in a better understanding of ELECTRA's Natural Language Inference could be expanded to the creation of new contrastive and adversarial datasets used in the pre-training phase. This would allow for the model to be exposed to example pairs that it most frequently predicted incorrectly.

References

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China, November. Association for Computational Linguistics.

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).

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah Smith. 2021. Competency problems: On finding and removing artifacts in language data. arXiv preprint arXiv:2104.08646.

Martin Nguyen. Contrast NLI Dat set. https://huggingface.co/datasets/martn-nguyen/contrast_nli

Max Glockner, 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 6 Computational Linguistics (Volume 2: Short Papers), pages 650–655, Melbourne, Australia, July. Association for Computational Linguistics.

Samuel R. Bowman, 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, September. Association for Computational Linguistics

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. arXiv preprint arXiv:1910.14599v2.

Appendix

Accuracy table

Accuracy		Evaluation Method	
		SNLI	Contrast
Dataset	NLI Ensemble SNLI	0.822	0.401
	NLI on SNLI	0.893	0.463
	NLI on Fine Tuned SNLI	0.752	0.422
	Nguyen	0.881	0.504

F1 Score table

F1 Score		Evaluation Method	
		SNLI	Contrast
Dataset	NLI Ensemble SNLI	0.821	0.401
	NLI on SNLI	0.893	0.463
	NLI on Fine Tuned SNLI	0.748	0.422
	Nguyen	0.882	0.504

F1 Score Comparison