Handling Negation Bias in Natural Language Inference

Grant Haislip
University of Texas at Austin
granthaislip@utexas.edu

Abstract

My goal is to use adversarial data and repeated training on negation data to improve the natural language inference of the ELECTRA-small model. I initially train and evaluate the model on training and test data from SNLI. I then use filtered and contrast data to expose the faults of the model and attempt to fix them with random adversarial data and repeated training on negation phrases with a non-contradiction label. I found that both techniques were able to decrease the bias in negation words and slightly improve overall accuracy. Negation words used in entailment and neutral relations were less likely to be assumed as a contradiction.

1 Introduction

When training a model on sample data, the model can work well against the engineered test data but fail to perform well in the real world. This occurs due to the train dataset having systematic gaps, like annotation artifacts, that stray the model away from its intended behavior. One of these issues is negation words causing bias to assume contradiction (Gardener et al., 2020). The table below shows an example of the model incorrectly predicting these cases. The bias arises from crafted training data not properly reflecting negation phrasing outside of a contradiction context.

Sentence	Pred	Actual
Group of guys sitting in a circle	Cont.	Neut.
There are no females around them		

Table 1: Example of incorrect predicted relation made by ELECTRA-small. Prediction of contradiction instead of neutral

Adversarial data is synthesized data used to train a model so that it captures a specific malfunction in expected behavior. This can be used to lightly train a model for handling unexpected situations (Belinkov et al., 2019). Data enhancement by repeated training pinpoints small portions of samples that do not follow typical behavior. It then repeatedly trains the records to reduce bias against them (Zhou et al., 2020).

By using adversarial data, I hope to reduce bias in negation words and phrases. This will be done by converting negation data with a contradiction label to a different label at low probability. By training on the newly changed data, the model will better behave such that negation words have less bias in cases of entailment and neutrality, meaning that the model should not overfit negation words to be contradictions. With repeated training, I hope to better balance out entailment and neutral data with negation words to those of contradiction. Ideally, leveraging the repeated data will cause the model to better understand these phrases that do not amount to a contradiction.

2 Approach

2.1 Model

The ELECTRA-small model will be used for this experiment. The method created by Google Research is used to train transformer networks for learning self-supervised language representation. The small variant uses less layers and parameters, and a smaller hidden size so training can reasonably be done on a single GPU.

2.2 Data Sources

The data to be used is the Stanford Natural Language Inference (SNLI) dataset. This includes both a training and test set. Each record represents a given sentence, the hypothesis of the sentence, and their relationship. The relationship can be entailment, contradiction, or neutral. Along with this, a handcrafted set of contrast data will be created to better expose faults in the model. This includes an entailment and contradiction pair that irregularly use negation wording. Table 2 below gives an example pair of the created contrast data.

Sentence	Label
A statue that is not at a museum that no one seems to be looking at	Entail.
There is a statue that not many people seem to be interested in	
A statue at a museum that people are not forgetting to see	Cont.
There is a statue that not many people seem to be interested in	

Table 2: A pair of contrast data for an entailment and contradiction relation.

2.3 AdversD: Adversarial Data

For the first approach, which will be called AdversD, training data with presence of negation words and a contradiction label will have its label swapped to entailment. The probability of this occurring will be set somewhat low so that negation word weights are reasonably balanced yet capture possible negation phrases not properly represented in the dataset. An example of an adversarial change can be seen in Table 3.

Sentence	Label
A woman and a child holding on to the railing while on trolley	Contr.
The people are not holding onto anything	
Data now edited	
A woman and a child holding on to the railing while on trolley	Entail.
The people are not holding onto anything	

Table 3: AdversD example with previous and then edited data record from contradiction to entailment

2.4 RepT: Repeated Training

With the second approach, which will be called RepT, training data with a non-contradiction label that has negation words in either the original sentence or the hypothesis will be repeatedly trained a fixed number of times. This will help reduce bias by providing a more balanced training set for negation words in contexts that do not amount to a contradiction.

Sentence	Label
New sport is being played to show appreciation to the kids who can not walk	Entail.
People are playing a sport in honor of crippled people	
New sport is being played to show appreciation to the kids who can not walk	Entail.
People are playing a sport in honor of crippled people	
New sport is being played to show appreciation to the kids who can not walk	Entail.
People are playing a sport in honor of crippled people	

Table 4: RepT example of entailment with the original data and repetition constant of 2

2.5 Hyperparameters

Using the two techniques above, the repeated training constant and probability of adversarial data were finetuned to seek out the best results. Training epochs was fixed at 3 and training batch size was fixed at 30.

2.4 Experiments

I first trained the base model with the default SNLI dataset to obtain a control to compare with the added techniques. Afterwards, the model was analyzed with the default evaluation dataset, a filtering of this dataset of records with negation wording and a non-contradiction label, and lastly a contrast dataset.

After implementing AdversD and RepT, numerous experiments with a variety of different hyperparameter values were done to determine an

optimal configuration for the model with the new changes. The negation words used for two techniques were "no", "not", "nor", "neither", "never", "none", "nobody", "nothing", "nowhere", "false", "hardly", "barely", "isn't", "don't", "doesn't", "won't", "can't", "couldn't", "wouldn't", "shoudn't", "wasn't", and "scarcely."

These experiments were measured by comparing the accuracy score between the models trained using the base dataset, the filtered dataset, and the handcrafted contrast dataset. Specifically, the models were first trained with an adversarial rate of 1% and a repeated data constant of 1. The hyperparameters of the two experiments were slowly increased until negation word weights caused lower accuracy than the control. The results presented include the highest accuracy scores for AdversD, RepT, and both AdversD and RepT used together.

3 Results

To first create a control group, the default train data of SNLI was used against the ELECTRA-small model using a training batch size of 30 and a train epoch size of 3. The results below show its accuracy against the default test data of SNLI, the filtered data, and the handcrafted contrast data.

Data	Accuracy (%)
SNLI Test Data	89.687
Filtered Data	78.378
Contrast Data: Contradiction	15.000
Contrast Data: Entailment	90.000

Table 5: Results for the control model

As expected, the ELECTRA-small model performed worse than average on data with negation wording and a non-contradiction label. Combining the filtered and entailment datasets exposed an accuracy decrease of over 10%. Contrary to expectations, the model performed much worse with the irregular contradiction data. The model would mostly guess entailment or neutral unless contradiction phrasing was explicit. Exact wording and adjacent negation placement were needed for contradiction phrasing to be understood.

Following this, AdversD was trained with varying probabilities. It was found that a 10% probability that created 981 adversarial data entries resulted in the highest performing model.

Data	Accuracy (%)
SNLI Test Data	89.859
Filtered Data	81.081
Contrast Data: Contradiction	27.500
Contrast Data: Entailment	90.000

Sentence	Control Pred	AdversD Pred
Trying very hard not to blend any of the yellow paint into the white	Cont.	Neut.
Someone is painting a house		
A woman in a white hijab decides not to dig into the ground	Entail.	Cont.
The lady digs into the ground		

Table 6 and 7: Accuracy results and fixed test cases for the AdversD model with adversarial probability of 10%

The adversarial data changes caused model to have improved accuracy over the control. Both contradiction phrasing in the contrast data and non-contradiction phrasing in the filtered data that use negation wording were better understood in less explicit cases. Shown above, the phrase “very hard not to” did not have enough bias to assume contradiction, and the phrase “decides not to” was better learned given its context. Further increase of the adversarial probability led to incorrect behavior in contradiction phrases the control model was able to handle.

Next, RepT was trained with different values for the repetition constant. A constant value of 2 was found to be the best performing hyperparameter; however, RepT only had a slightly better accuracy to that of the control. Increasing the constant beyond 2 caused significantly lower performance.

Data	Accuracy (%)
SNLI Test Data	89.799
Filtered Data	79.730

Contrast Data: Contradiction	20.000
Contrast Data: Entailment	90.000

Table 8: Results for the RepT model with repetition constant of 2

Lastly, both AdversD and RepT were used alongside each other in training. The hyperparameter values with highest accuracy were AdversD probability of 8% and RepT repetition constant of 2. Combined, they did outperform the control model, yet AdversD alone still had slightly higher accuracy results against the contradiction contrast data and the overall SNLI test data.

Data	Accuracy (%)
SNLI Test Data	89.850
Filtered Data	81.081
Contrast Data: Contradiction	20.000
Contrast Data: Entailment	90.000

Table 9: Results for the AdversD+RepT model with an adversarial probability of 10% and a repetition constant of 2

4 Conclusions

The two techniques and their combination resulted with better performance from reducing contradiction bias in negation words. The highest performing approach was AdversD with an adversarial probability of 10%, leading to only a 0.17% increase in overall accuracy to that of the control, but filtering that data shows a 2.7% increase in entailment and neutral relations that contain negation words. Using adversarial data by occasionally flipping the label of a phrase with a negation word helped to balance the weight of negation words so they do not always assume contraction when appearing in either the statement or hypothesis sentences.

However, the ELECTRA-small model still underperforms with irregular negation phrasing. This was especially the case for contradiction relations with irregular negation wording. It is likely that this malfunction arises from training and test data that do not reflect many sentence structures outside of clear, straightforward entailment and contradiction. Other than the techniques used in this paper, developing more thorough data to train with can lead to the model performing better in these cases. A future experiment for this could be to do more research on possible negation phrasing, define their structures, and add new data to SNLI that appropriately reflects them. What comes from that addition could be an exhaustive, more resilient model that achieves improved behavior in real world use.

Acknowledgments

Thank you Dr. Durrett and the TAs for all the lessons you've taught me and help you've given me in this course.

References

Belinkov, Y., Poliak, A., Shieber, S.M., Durme, B.V., & Rush, A.M. 2019. On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference. *ArXiv*, arXiv:1907.04389.

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N.F., Mulcaire, P., Ning, Q., Singh, S., Smith, N.A., Subramanian, S., Tsarfaty, R., Wallace, E., Zhang, A.Q., & Zhou, B. 2020. Evaluating NLP Models via Contrast Sets. *ArXiv*, arXiv:2004.02709.

Zhou, Xiang and Mohit Bansal. 2020. Towards Robustifying NLI Models Against Lexical Dataset Biases. *ArXiv* , arXiv:2005.04732.