Constructing Contrast Sets Based on Adversarially Generated Examples

Charles Wang
thewangclass@gmail.com

Abstract

Machine learning models have made significant progress on the task of Natural Language Inference. However, many studies have shown these models do not actually learn the language semantics as intended. Rather, the datasets the models are trained and evaluated on allow them to achieve high accuracy scores with simple decision rules, enabling them to perform well on the test set without a deep understanding of language semantics. Using a model-in-the-loop approach, we constructed new datasets consisting of mislabeled examples. We used these mislabeled examples to construct new contrast sets consisting of self-authored examples. We then trained our model on combinations of the newly created datasets with SNLI and MultiNLI. We found that our newly trained models all show an improvement when evaluated on data they have not seen during training, leading us to believe this technique will help models generalize to outside datasets.

1 Introduction

Natural language inference (NLI) is a task focused on identifying whether a premise sentence entails, contradicts, or is neutral with a hypothesis sentence (Bowman et al., 2015). This is an example of a three-way classification problem. Machine learning based approaches to this task require enormous amounts of data to train increasingly data-hungry models. One approach to meet this demand is crowd-sourcing the data creation process.

Two popular crowd-sourced datasets are the Stanford Natural Language Inference (SNLI) and the Multi-Genre Natural Language Inference (MultiNLI, referred to as MNLI) corpus (Bowman et al., 2015; Williams et al., 2018). These two datasets added over 900K examples for machine learning models to train on, leading to rapid breakthroughs and classification performance competitive to that of humans. However, numerous problems with these datasets have been discovered, showing that models trained on these datasets exploit dataset biases in making their predictions and do not generalize to even simple examples outside of the dataset itself. A meaningful NLI model should have to rely on both the premise and hypothesis in making its prediction, but a classifier with access to only the hypothesis sentence outscored many early machine learning models (Poliak et al., 2018). This hypothesis-only trained model could score over 50%, well over the baseline of majority-class, without ever knowing the premise. Other research conducted at the same time showed sentence length, the presence of negation, and other annotation artifacts could be used as strong predictors of the label (Gururangan et al., 2018).

Almost every dataset contains annotation artifacts and dataset biases, resulting in numerous attempts to address these issues. Datasets have been constructed consisting of adversarial examples, where a small change has been made to the hypothesis but the premise-hypothesis pairing still retains the same gold label (Glockner et al., 2018). Other researchers have used a model-in-the-loop approach to construct datasets consisting of examples models get wrong and including them in the training set (Bartolo et al., 2020). Some studies exposed models to data from a challenge dataset as part of the training process to see if it improved the model performance during evaluation in a process known as inoculation (Liu et al., 2019). Contrast sets, manually perturbed test instances that result in a change in the gold label, were created as evaluation benchmarks to better examine whether models learned the desired lexical semantics or merely used simple decision rules for their clas-

sifications (Gardner et al., 2020).

Using a model-in-the-loop approach, we construct new datasets consisting of mislabeled examples. We use these mislabeled examples to construct new contrast sets consisting of self-authored examples. We then train our model on combinations of the newly created datasets with SNLI and MultiNLI. We find that our newly trained models all show a slight improvement on data they have not seen during training.

We propose involving a trained model in the process of constructing new training examples. We train a BERT based model on the standard training split of a dataset and then evaluate the model on the standard test split, which produces examples the model correctly and incorrectly classified. We utilize the incorrectly predicted examples as an adversarial set and construct a contrast set out of these examples. Then we train our model on this new data and evaluate on the standard test split. Our results show little statistical improvement in accuracy scores. However, we believe there is still further to explore in this area based on other things discovered during our experimental process.

2 Experiment Setup

2.1 Model

For our experiments, we use ELECTRA-small model. ELECTRA-small is a machine learning model with the same architecture as state-of-the-art BERT but with a smaller number of parameters, allowing it to run with less compute power in a reasonable amount of time. 81.3 is an expected result for ELECTRA on MNLI, or 77.4 on GLUE (Clark et al., 2020)

2.2 Datasets

We focus on the SNLI and MNLI datasets for our experiments. There is already substantial literature about the biases that exist in these datasets. Discussion has already been given to some of the more general mistakes. Poliak showed NLI datasets contain statistical irregularities, known as annotation artifacts, that allow a hypothesis-only classifier to score very high on a premise-blind classification test. Their work showed many words were highly-correlated with each label, making the presence of such a word in a premise-hypothesis pair a likely annotation artifact machine learning models could pick up as part of their training process (Poliak et al., 2018). Likewise, Gururangan simultaneously discovered the correlation between words and certain labels, and their team revealed an additional correlation between hypothesis sentence length and the gold label. For SNLI, neutral hypotheses tended to be long and entailed hypotheses were generally shorter. Their work also found entailment hypotheses tend to contain gender-neutral references, neutral hypotheses tend to have purpose clauses, and contradiction hypotheses tend to have negation. Using their premise-oblivious model, they partitioned the SNLI and MNLI test sets into two subsets - those their premise-oblivious model accurately classified were labeled Easy and those it misclassified were labeled Hard. (Gururangan et al., 2018)

For our experiment, we initially trained our ELECTRA-small model on SNLI, MNLI, and then on both SNLI+MNLI train set. We then evaluated each of these three models on the SNLI, MultiNLI Matched (M), and MultiNLI Mismatched (Mm) evaluation sets. We also evaluated our trained models on Gururangan's SNLI Hard (S-H), MultiNLI Matched Hard (M-H), and MultiNLI Mismatched Hard (Mm-H) test sets. Table 1 shows the accuracy scores of each model on each test set.

2.3 Exploration

Our initial results seem comparable to those of other researchers. The SNLI trained model initially scores above 80% on the SNLI test set, only to fall to below 70% for the SNLI-hard set. We see similar accuracies and drops when evaluated on the SNLI test set for the models trained on MNLI and SNLI+MNLI, confirming prior work (Gururangan et al., 2018). The same observations can be seen when evaluated on MultiNLI Matched and the MultiNLI Mismatched sets.

All of our trained models perform better than the majority-class baseline of about 33%. Additionally, the accuracy scores are above Gururangan's fastText hypothesis-only classifer of 67.0 on SNLI, 53.9 on MutliNLI Matched, and 52.3 on MultiNLI Mismatched (Gururangan et al., 2018). The scores are also above Poliak's hypothesis-only classifier (Poliak et al., 2018).

2.4 Experiment Focus

Our attention turns to understanding the makeup of the examples our models incorrectly labeled. Table 2 shows all of the misclassified premise-

Training Dataset	Evaluation Dataset Accuracy %								
	$\mathcal{D}_S$	$\mathcal{D}_{S-H}$	Δ	$\mathcal{D}_M$	$\mathcal{D}_{M-H}$	Δ	$\mathcal{D}_{Mm}$	$\mathcal{D}_{Mm-H}$	Δ
SNLI	89.6	78.5	-12.4	68.5	58.7	-14.3	68.9	57.4	-16.7
MNLI	76.6	66.2	-13.6	81.4	73.6	-9.6	82.2	74.0	-10.0
S/MNLI	89.5	78.0	-12.9	81.8	74.2	-9.3	81.5	72.9	-10.6

Table 1: ELECTRA-small model trained on various datasets evaluated on standard test splits. $\mathcal{D}_S$ and $\mathcal{D}_{S-H}$ are SNLI and SNLI-H. $\mathcal{D}_M$ and $\mathcal{D}_{M-H}$ are MNLI Matched and Hard. $\mathcal{D}_{Mm}$ and $\mathcal{D}_{Mm-H}$ are MNLI Mismatched and Hard. SNLI shows the greatest % decrease between regular evaluation and hard evaluation test sets.

hypothesis pairs across all the test sets with the information publicly available[1].

From this data, we can see the majority of the mislabeled examples were from mislabeling either contradiction/entailment as neutral, or vice-versa. This suggests that without neutral, our model correctly predicts entailment as entailment and contradiction as contradiction with fairly high accuracy. We hypothesize many of the wrongly classified entailment as contradiction and contradiction as entailment examples express one or more of the annotation artifacts that are known to be highly correlated with a certain class.

2.5 Creating Adversarial and Contrast Sets

We decided to focus our attention on premise-hypothesis pairs all three of the trained models misclassified. In addition to these examples, we also focused on the examples only the SNLI+MNLI trained model misclassified (while the other two models correctly classified). We found the model trained on SNLI+MNLI made the least number of unique misclassifications. Finally, we chose to focus any further experiments on the SNLI+MNLI trained model due to namely time, compute power, and storage capacity considerations.

We decided to focus only on the MNLI Matched errors. One reason is MNLI is a newer dataset than SNLI and was created partially to address some of the issues found in SNLI (Williams et al., 2018). Another reason is we can maintain train/test splits when evaluating on MNLI Mismatched, so genres in MNLI Mismatched would not be seen in any of the training data for MNLI Matched.

We conducted some data analyses and found the union of misclassifications for the models trained on SNLI, MNLI, and SNLI+MNLI. We looked at only the MNLI Matched examples. We broke these misclassifications down by genre, and further by type of misclassifications. Table 3 shows the results.

As seen in the table, we see most of the misclassifications are between neutral and entailment, or neutral and contradiction, rather than between entailment and contradiction. We also see that overall, the genres of Government and Telephone have the least mislabeled while Fiction and Slate have the most mislabeled. Of particular note is how relatively few misclassifications of contradiction as entailment there are for the Government genre of $\mathcal{D}_{sm}$.

From these premise-hypothesis pairs, we created four sets of 100 examples each to augment our SNLI and MNLI training sets. We randomly chose 20 premise-hypothesis pairs from $\mathcal{D}_{sm}$ for each genre. 10 of these would come from entailment misclassified as contradiction and the other 10 would come from contradiction misclassified as entailment. This results in 100 premise-hypothesis pairs. If these categories did not provide enough examples, we would randomly select from entailment/contradiction incorrectly labeled as neutral. If there still were not enough examples, we would randomly select from neutral incorrectly labeled as entailment or contradiction. For example, for the Telephone genre in $\mathcal{D}_{sm}$, there is only 1 contradiction example misclassified as entailment. Since there is not enough examples, We look first to contradiction misclassified as neutral, or $2 \rightarrow 1$. We grab 8 more examples there to make a total of 9. We then go to neutral misclassified as contradiction for the last remaining example, to make a total of 10. We followed the same procedure for selecting premise-hypothesis pairs from $\mathcal{D}_{union}$. One thing to note is the lack of entailment samples labeled as contradiction and contradiction labeled as entailment in $\mathcal{D}_{sm}$. This could be due to the other trained models making similar classification errors on the same premise-hypothesis pairs.

[1]MNLI Hard test sets were found on Kaggle without gold labels. Request to the author via email for gold labels went unanswered.

Trained On	Gold → Prediction	**Evaluated On**			
		$\mathcal{D}_S$	$\mathcal{D}_{S-H}$	$\mathcal{D}_M$	$\mathcal{D}_{Mm}$
SNLI	0 → 1	292	171	656	600
	0 → 2	49	37	273	255
	1 → 0	207	175	595	671
	1 → 2	199	146	529	475
	2 → 0	65	34	459	448
	2 → 1	214	137	582	610
MNLI	0 → 1	326	169	457	412
	0 → 2	281	110	197	172
	1 → 0	392	261	391	346
	1 → 2	601	261	391	346
	2 → 0	169	90	170	169
	2 → 1	531	236	328	371
S/MNLI	0 → 1	270	181	444	432
	0 → 2	54	39	184	150
	1 → 0	214	187	284	312
	1 → 2	218	140	380	361
	2 → 0	55	27	149	169
	2 → 1	224	142	349	396

Table 2: Evaluation of trained models on the original test sets. The premise-hypothesis pairs the model incorrectly classified, broken down by gold label and predicted label. Of particular note is how relatively small the number of entailment/contradiction misclassifications are when compared to neutral.

Gold → Prediction	$\mathcal{D}_{sm}$					$\mathcal{D}_{union}$				
	Fiction	Gov	Slate	Telephone	Travel	Fiction	Gov	Slate	Telephone	Travel
0 → 1	8	11	15	11	15	46	32	53	50	32
0 → 2	3	7	8	1	6	18	9	23	12	16
1 → 0	7	7	9	10	5	32	19	41	33	22
1 → 2	14	10	14	6	10	28	26	51	33	37
2 → 0	3	0	2	1	1	12	10	32	12	30
2 → 1	9	8	12	8	9	32	26	46	37	34

Table 3: Number of premise-hypothesis pairs incorrectly classified, broken down by genre and type of misclassification. $\mathcal{D}_{sm}$ represents the examples only the SNLI+MNLI trained model missed. $\mathcal{D}_{union}$ represents the examples all three of the models missed.

We used these two sets of 100 examples each as adversarial inputs for fine-tuning the model; if the model trained on this new challenge set performed well, then the weakness probably lies with the original dataset. We chose 100 examples due to prior work showing as few as 100 examples is effective (Liu et al., 2019) time constraints, the number of misclassified contradiction as entailment examples in $\mathcal{D}_{sm}$, and time constraints.

We then used these adversarial sets to craft contrast sets. For each premise-hypothesis pair, we created a new hypothesis sentence with a label the original hypothesis was wrongly classified as. For example, if the premise is *"Visit at sundown or out of season to get the full flavor of the setting."*, the hypothesis is *"The setting truly comes alive with fewer people during sundown or out of tourist season."*, and it was misclassified as contradiction, then we would write a hypothesis that actually has a gold label of contradiction such as *"The best time to savor the full flavor of the setting is during the peak of tourist season when there are lots of people."*

These two contrast sets, one for $\mathcal{D}_{sm}$ and one for $\mathcal{D}_{union}$, result in 4 sets of 100 examples each for 400 total examples. We then combined these with the SNLI and MNLI training sets in the following combinations for training:

1. $\mathcal{D}_{union,a}$, $\mathcal{D}_{sm,a}$
2. $\mathcal{D}_{union,a}$, $\mathcal{D}_{union,c}$
3. $\mathcal{D}_{sm,a}$, $\mathcal{D}_{sm,c}$
4. $\mathcal{D}_{sm,c}$, $\mathcal{D}_{union,c}$
5. $\mathcal{D}_{union,a}$, $\mathcal{D}_{sm,a}$, $\mathcal{D}_{sm,c}$, $\mathcal{D}_{union,c}$

where $_a$ stands for adversarial and $_c$ stands for contrast.

We trained our model on SNLI+MNLI augmented with each of these five datasets and then evaluated the models on SNLI, SNLI Hard, MNLI Matched, MNLI Matched Hard, MNLI Mismatched, and MNLI Mismatched Hard.

3 Results

Table 4 shows the results of the experiment. Once again, we see the typical drop in evaluation accuracy from the original to the hard version of the test set. The % drops seem to hover around the same value as the original accuracy decrease. Another interesting observation is how while there is some performance degradation on $\mathcal{D}_S$, every model performed better on $\mathcal{D}_{S-H}$. This could be because the model is learning lexical semantics necessary to classify $\mathcal{D}_{S-H}$ properly from the new examples, as opposed to relying on heuristics learned solely from the original training set. Ironically, this could cause it to perform worse on the original that was easily classified with a hypothesis-only classifier that relies on annotation artifacts. Something else that immediately stands out is how, even though the new training data came from the examples missed on the MNLI Matched dataset, all the models trained on various combinations of the new training data performed better than the original on MNLI Mismatched. Adding only 400 training examples to the already existing over 800K examples from SNLI+MNLI already leads to a minimum of 0.1% improvement on a evaluation test set that was never seen by the model before. This strengthens the claim that the additional examples help the model perform well on other datasets it has never seen before.

4 Discussion

We reflect on our results and discuss avenues for future work to be performed.

4.1 Closer Look at Misclassified Examples

Many of the misclassified examples followed patterns discovered in earlier work. The majority of the entailment incorrectly labeled as contradiction had some form of negation or a lack of activity in the hypothesis. This corroborates with earlier research of possible crowd-sourced annotator strategies at producing these contradiction hypotheses (Gururangan et al., 2018; Poliak et al., 2018). We also find many examples mislabeled as neutral with modifiers, superlatives, or assigning a purpose to the premise. Many of the misclassified as entailment premise-hypothesis pairings also had much shorter hypotheses lengths.

We believe many of the mislabeled examples could also benefit from applying real-world knowledge. For example, prompt/pairID 129601/n was in $\mathcal{D}_{sm}$ due to being misclassified as neutral when the gold label was contradiction. However, what is the concept of *forever*? Depending on the context, forever could very well be, or seem like to the speaker, two years. Another example that deals with time is prompt/pairID 4082/e. What is the length of a term? In the United States, a typical term may be 2, 4, or 6 years. And yet the gold label here is entailment for Helms turning 81 by the time his fifth term ends.

Premise: Took forever
Hypothesis: Lasted two years
Gold Label: Contradiction

Bowman and Williams both briefly discuss the problems surrounding annotator decisions about the coreference between entities and events across the premise-hypothesis pairs, notably that it can lead to different annotators assigning different labels (Bowman et al., 2015; Williams et al., 2018). Bowman attempted to address coreference via prompts to the annotator, but this led to annotation artifacts (Gururangan et al., 2018). Williams used prompts to address the coreference issue as well, and it remains to be seen whether his attempts also lead to exploitable artifacts in the datasets.

Although our work in creating the contrast sets involved mostly drawing from entailment and contradiction mislabeled examples, we believe future work should look deeply into these neutral-labeled examples when building contrast sets to better delineate label boundaries.

Training Dataset	Evaluation Dataset Accuracy %								
	$\mathcal{D}_S$	$\mathcal{D}_{S-H}$	Δ	$\mathcal{D}_M$	$\mathcal{D}_{M-H}$	Δ	$\mathcal{D}_{Mm}$	$\mathcal{D}_{Mm-H}$	Δ
S/MNLI	89.5	78.0	-12.9	81.8	74.2	-9.3	81.5	72.9	-10.6
1	89.5	79.5	-11.2	83.0	74.8	-9.9	82.1	73.8	-10.1
Δ	0.0	+1.9		+1.5	+0.8		+0.7	+1.2	
2	88.8	78.1	-12.0	82.6	73.5	-11.0	82.1	73.4	-10.6
Δ	-0.8	+0.1		+1.0	-0.9		+0.7	+0.7	
3	89.2	78.5	-12.0	82.1	74.8	-8.9	82.0	73.6	-10.2
Δ	-0.3	+0.6		+0.4	+0.8		+0.6	+1.0	
4	89.1	78.8	-11.6	81.6	74.0	-9.3	82.2	73.5	-10.6
Δ	-0.4	+1.0		-0.2	-0.3		+0.9	+0.8	
5	89.0	78.2	-12.1	82.9	74.4	-10.3	82.0	73.0	-11.0
Δ	-0.6	+0.3		+1.3	+0.3		+0.6	+0.1	

Table 4: Training on combined datasets and evaluated on various evaluation sets. S/MNLI stands for SNLI and MNLI, -H stands for Hard, -M stands for Matched, and -Mm stands for Mismatched. 1, 2, 3, 4, and 5 are SNLI+MNLI datasets augmented with the corresponding $\mathcal{D}_{union,a}$, $\mathcal{D}_{sm,a}$, $\mathcal{D}_{sm,c}$, and $\mathcal{D}_{union,c}$ mentioned previously in Creating Adversarial and Contrast Sets.

4.2 Annotator Disagreement

While we were conducting our experiment, we began to wonder whether the level of annotator disagreement could be correlated with certain labels just as certain words were highly predictive of the label. When producing our contrast sets, we noticed the difficulty in producing examples for anything involving a neutral classification without relying on heuristics other researchers have found. Additionally, we were hesistant on several of the assigned gold labels. Taking a look at the SNLI dataset, we see some alarming statistics: 6.8% of gold labels are not the original author's labels, and only 58.3% of labels have unanimous agreement. For the MNLI dataset, we see similar statistics: 5.6% of gold labels are not the original author's labels, and only 58.2% of labels have unanimous agreement. We believe these inter-annotator labels should be looked at further to see if any correlation with annotator agreement and gold label exists.

5 Conclusion

We investigated constructing contrast sets that requires a model in the loop to misclassify the standard test split of a dataset. Using the same model trained on three different datasets, we found the premise-hypothesis pairs each trained model incorrectly labeled. We specifically looked at the examples our model trained on the most data missed, as well as the union of all missed examples. We randomly sampled 100 examples and used these to construct an additional 100 contrast examples. We found training our model on the original dataset and these new adversarially generated and contrast examples leads to a minor improvement on the original test set. We find that our newly trained models all show an appreciable improvement given the difference in magnitude of additional data introduced compared to the total training examples. These results lead us to believe this technique will help models generalize to outside datasets.

References

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.

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. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh 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. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

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 Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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 for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, 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.