Adversarial Dataset Evaluation Against Varied NLI Trained Models

Individual (not group) project submission
{student name omitted here for purposes of anonymous peer grading}

Abstract

Evaluation of natural language inference models are constantly strained with adversarial and contrast datasets in an effort to improve learning on language and to continue progress in the work to develop systems that can truly understand language. Methods for training models on adversarial sets sometimes reply on semantics that are not necessarily natural language but intentionally misleading. Insights may be gained by increasing the training dataset size to naturally minimize the skewing such examples may have on the model. Insights presented here discuss accuracy on the base SNLI dataset vs a merged SNLI/MNLI dataset. Each model is then fine-tuned with adversarial input and each model is re-on adversarial datasets and original inputs to contrast performance improvement.

1. Introduction

Classic testing of model performance relies typically on splitting a dataset into a training set and evaluation set. Model are trained on the training set while the example in the evaluation set are held out so that evaluation can be done on new examples unseen by the model. The concept is sound in many areas utilizing unsupervised learning but there are noted issues where the model can learn on attributes that are non-semantic in nature which find patterns that are more predictive to the output but failing to ultimately assuage the meaning of the broader discourse context. (Rimell et al., 2009, Paperno et al., 2016).

In this paper, we evaluate multiple adversarial datasets for natural language processing by fine-tuning model training on the SNLI and combined SNLI/MNLI datasets. The models are fine-tuned using the ANLI dataset and the Contradictory My Dear Watson dataset.

Watson Dataset Example	
Premise: "The streets are crammed with vendors selling shrine offerings of sweets, curds, and coconut, as well as garlands and holy images."	
Hypothesis: "Vendors have lined the streets with torches and fires.",	
Gold Label:	2, Contradiction
SNLI Base Model Predicted Label:	1, Neutral
After Fine-tuning SNLI Model with Watson Dataset:	2, Contradiction
MNLI-SNLI Combined Base Model:	2, Contradiction
SNLI fine-tuned with Reddit Sarcasm data:	1, Neutral

Figure 1: Example from the Watson dataset. The base model classifies the premise incorrectly but adversarial fine tuning improves performance. Training on a larger dataset (SNLI/MNLI Combined) also improves predictive performance

Machine learning has many applications if models can be trained to discern language meaning for applications in fact-checking statements, evaluating the probability of an article being fake-news, or analyzing text to provide summary paragraphs, auto-captioning or providing language meaning models for purposes of language translation or sarcasm detection.

The goal of this project is to evaluate model performance through either increasing the number of examples the model is trained on, fine-tuning models with a corpus that matches the intent, and evaluating various models with adversarial sets to determine the best performing general use model. Each mode, base models and fine-tuned models are evaluated by predicting whether a given hypothesis is related to the premise statement by

contradiction (label ==2), *entailment (label == 0)* or *neutral (label == 0)* if neither of those is true.

First, a base model is trained with a standard dataset using ELECTRA: a BERT-like model that is pre-trained as a discriminator in a set-up resembling a generative adversarial network (GAN). It is used as a base for faster model training times.

The ELECTRA model is first trained on the SNLI dataset. The SNLI dataset comes from a crowdsourced task involving Flicker image captions without the image. Human workers were tasked with developing neutral statements, contradictory statements, and entailment responses. This dataset has received criticism for its noted bias on protected characteristics such as gender and race as well as human patterns used to create contradictory statements (i.e. overuse of the word "not" biases the model select statements with 'not' and contradictory) (He et al, 2019). An alternative NLI model is trained on a combination of SNLI and MNLI datasets for contrasting performance from the base SNLI and all datasets are used to fine-tune each model to determine performance enhancements.

Each of the model weights (SNLI and SNLI/MNLI trained models) are then fine-tuned with additional data sources: CNLI (Watson), ANLI, and a maverick Reddit-Sarcasm dataset. The intent of the Watson data is to broaden the content of the examples to hopefully limit some of the generative biases and to make the model more robust as it encounters adversarial datasets. The model is alternatively fine tuned using the ANLI data set (Nie et al 2019), a human-and-model-in-the-loop model generated dataset to examine if the adversarial training made the model more robust. The final dataset used for fine-tuning comes from a contradictory Reddit-sourced dataset that is comprised of comments that were marked as sarcastic. The intent of the Reddit dataset is to see if any of the models are able to perform the rather complex task of identifying sarcasm.

The two base trained models (SNLI and SNLI/MNLI) and the fine-tuned models using the Watson, ANLI, and Reddit datasets are each evaluated on accuracy and loss using withheld evaluation data sourced from each of the data sources.

2. The NLI Task and Models

2.1 Task Overview

Natural Language Inference (NLI) is also known as Recognizing Textual Entailment (RTE) is a task of determining whether the given "hypothesis" and "premise" can be logically true (entailment), logically untrue (contradiction) or are undetermined (neutral) to each other. For example, let us consider premise "The streets are crammed with vendors selling shrine offerings of sweets, curds, and coconut, as well as garlands and holy images." From Figure 1 and its hypothesis as "Vendors have lined the streets with torches and fires.". The task of NLI model is to predict whether the two sentences are either entailment, contradiction, or neutral. In this case, it is a contradiction. The vendors are selling torches and fires. The hypothesis contradicts this premise.

Natural language inference models can be used to automate tasks when trained on task specific datasets but are not reliable in a general use sense. They can be used, for example, to automate auThis project uses the ELECTRA [Clark et al 2020] which is based on the BERT architecture [Devlin et al 2019] with an improved model training schema. BERT takes a long time computationally to arrive at an answer due to the masking used on the generative model. Predicted tokens must be predicted using only 15% of the tokens available. ELECTRA improves this by training a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not instead of training a model that predicts the original identities of the corrupted tokens. This method of pre-training is more efficient because the task is defined over all input tokens rather than just the small masked out subset. Gains in training time and efficiency are particularly strong for small models, which is alignment with this project.

2.2 SNLI Dataset

The SNLI dataset [Bowman et al 2015] corpus is comprised of a collection of 570k sentence pairs with labels of [0, 1, or 2] which correspond to [*entailment, neutral,* and *contradiction*]. The examples are generated from mechanical turk micro-task human workers who both generate and label the sentence pairs. Much of the corpus is based on Flickr image captions with the image removed. The hypothesis sentences are then constructed using human input. This has led to

some criticism on the frequency of using words like 'not' to construct *contradiction* examples as well as gender and racial biasing [He et al, 2019]. The transformer readily picks up these unintended biases and word frequency usage and has trouble generating outside of the dataset.

2.3 Multi-NLI Dataset

The Multi-NLI dataset was concatenated with the SNLI dataset to construct an alternate base trained model. The Multi-NLI dataset [MNLI 2017] is comprised of 433K sentence pairs, modeled like SNLI. MultiNLI expands the scope and diversity of the data the model is trained on using content from ten different genres (Face-to-face, Telephone, 9/11, Travel, Letters, Oxford University Press, Slate, Verbatim, Goverment and Fiction) of written and spoken English data. Once concatenated with the SNLI dataset, this model is trained on over 4.8M sentence pairs with the idea that the model will be more robust when presented with adversarial and contrast sets.

2.4 Watson CNLI Dataset

The Contradictory My Dear Watson [Watson 2020] dataset provides an alternate to the base trained model using a smaller dataset that is also contracted using examples from diverse genres. This dataset challenges the models by presenting sentence pairs from 15 different languages: Arabic, Bulgarian, Chinese, German, Greek, English, Spanish, French, Hindi, Russian, Swahili, Thai, Turkish, Urdu, and Vietnamese. Performance is compared on the various models after fine-tuning using the Watson dataset to compare it to the performance of the much larger Multi-NLI/SNLI concatenated dataset. The Watson dataset is used to fine tune both the SNLI and the SNLI/Multi-NLI base models.

2.5 Adversarial NLI Dataset

The Adversarial NLI (ANLI) dataset [Nie et all 2019] is a human model-in-the-loop, an iteratively constructed dataset that contains texts that are intentionally selected to be difficult tasks for standard NLI models to handle. Adversarial training can enhance robustness of the model. Generally as with added model robustness, models generally have more difficulty with language generalization. Fine-tuning pre-trained models using the Adversarial NLI dataset has been shown to improve generalization even on models that have been been pre-trained with a large number text corpora. The ANLI dataset is normally comprised of three rounds, each progressively more difficult. For this project all three rounds were combined into a singular ANLI challenge set to determine general performance and to assess model generalization. This also makes ANLI a good benchmark to assess the progress of NLI models. The models are all assessed using the ANLI held out test set and also fine-tuned on the ANLI training set and then re-evaluated.

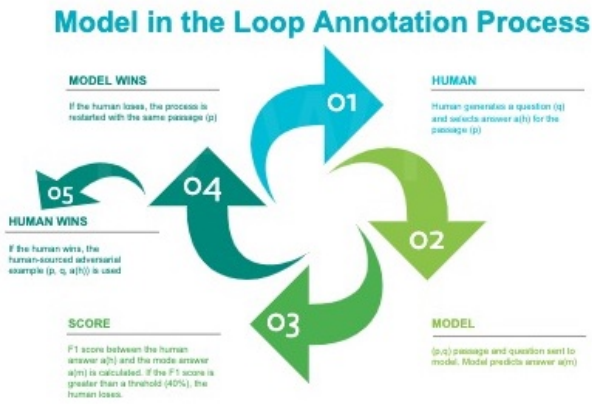

Figure 2: Human-Model-in-the-Loop Annotation Process

Reddit Sarcasm Dataset

The Reddit Sarcasm dataset [Reddit 2017] is used as an out-of -domain dataset to challenge the pre-trained models. The sentence pairs are constructed from Reddit topics and comments. User comments that are self-labeled with an \s are fairly reliably well-labeled sarcastic replies. The Reddit dataset also challenges the models semantically as it may include shortcut text slang phrases such as irl (in real life). Each sentence pair is labeled as either **0**, *Entailment* or **2**, *Contradiction* for the sarcastic comments. The models are then exercised on fine-tuned Reddit models as well as evaluated on a held-out Reddit test to determine generalized performance on the following subtasks:

- Predicting sarcasm while training on comments that may present sentiment heavy words, racism, internet slang, and forum specific phrases.
- Corpus with unusual linguistic features such as caps, italics, or elongated words that infer contradictory meaning. e.g., "Yeahhh, I'm sure THAT is the right answer", a sarcastic response meaning the opposite or "Excuse me" vs "Excuuuuse me!", which has the opposite meaning.

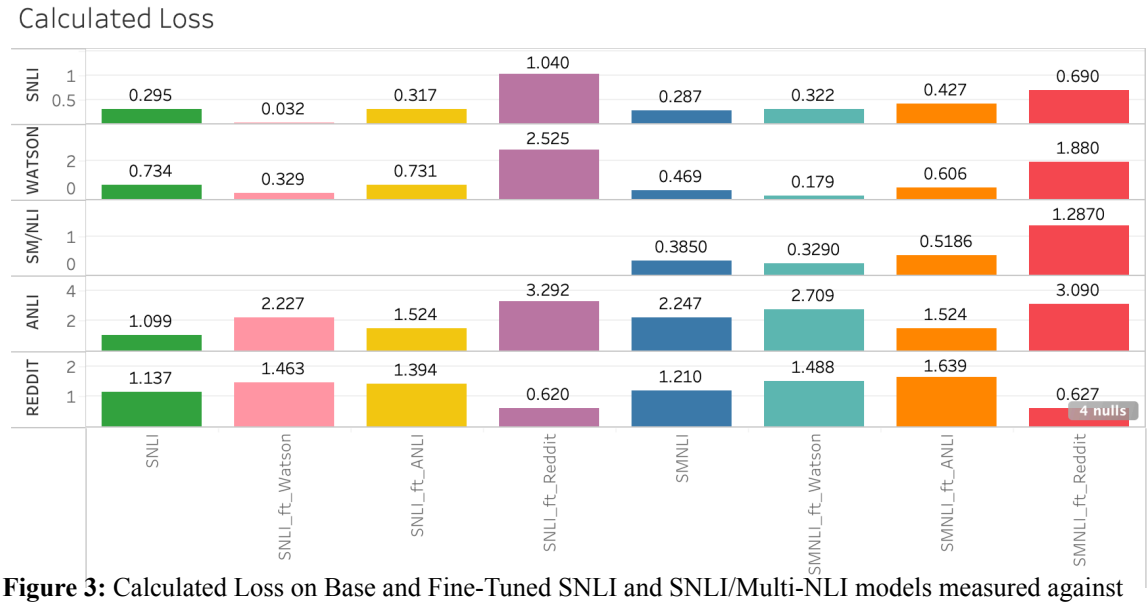

Figure 3: Calculated Loss on Base and Fine-Tuned SNLI and SNLI/Multi-NLI models measured against various challenge datasets.

• Semantic evaluation on topics that people tend to react to sarcastically (i.e. politics)

2.6 Implementation

Two base models are constructed: SNLI and Multi-NLI/SNLI concatenated. The models are then evaluated on test sets Watson, ANLI and Reddit. Each model is then fine tuned on training sets from Watson or ANLI or Reddit and then evaluated on each of the held out test sets from each dataset to determine loss and model prediction accuracy.

3 Results

3.1 The Base SNLI Trained Model

The SNLI trained model performed well on the SNLI test set but had greater ambiguity on the purity of its decision (increased loss) on the Watson, ANLI and Reddit datasets with the Reddit dataset causing the most loss and accuracy issues (See Figures 3 and 4). The ANLI dataset, as expected, provided quite a challenge on the SNLI base model.

3.2 Fine Tuned SNLI Datasets

The SNLI base model was fine tuned on the Watson dataset (SNLI_ft_Watson), the ANLI dataset (SNLI_ft_ANLI) and the Reddit dataset (SNLI_ft_Reddit) (See figure 3). The fine-tuned models performed best on test sets from the similar semantic corpora used in fine-tuning. The mode that performed the best in generalization is the NLI model fine tuned on the ANLI dataset.

The base SNLI model was improved significantly after fine tuning likely due to to overcoming or correcting some drawbacks of the SNLI datatset. Because the SNLI dataset is constructed use image captions, those caption (premies) are short as their nature is photo descriptions and don't contain temporal reasoning, beliefs, or modalities that are introduced in the fine-tuning datasets from Watson, ANLI and Reddit. In addition, because the original SMLI premises are short, the resulting hypothesis also tend to be short so the benchmark is not challenging enough where models can easily achieve human accuracy levels. Adding any additional corpus to fine-tune the model resulted in performance improvements in loss (Figure 3) and model accuracy (Figure 4).

Fine-tuned SNLI models achieve task-specific accuracy improvements of 10-20% by fine-tuning these models with specific data. This means that these models do much better using inputs from sources that are specific to the task at hand and still lack the ability to do general language semantic inference.

3.2 Base Trained Multi-MLI/SNLI Model

The model trained on the concatenated Mul-NLI and SNLI datasets provide to be more robust than the model trained on SNLI alone. The model seemed to be more sure of its decision on the ANLI and Reddit datasets, as expected. One note

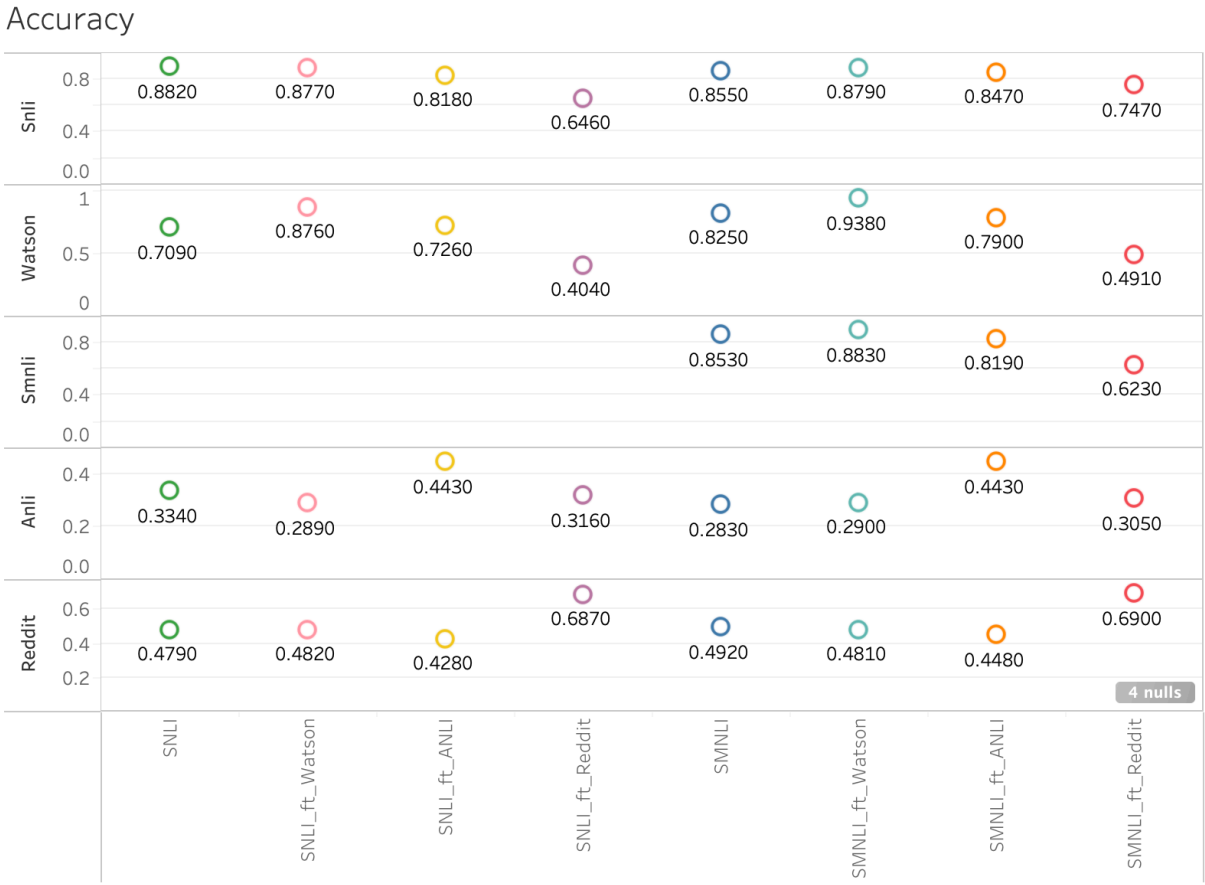

Figure 4: Calculated Accuracy on Base and Fine-Tuned SNLI and SNLI/Multi-NLI models measured against various challenge datasets.

of interest was increased loss on the ANLI dataset compared to the SNLI base model. It's possible that the increased diversity of test led to higher than expected performance loss for adversarial attack.

MNLI overcomes the drawbacks of short hypotheses and premise statements in the SNLI by drawing from published and spoken works over 10 genres including face to face spoken transcriptions, government speeches, letters and press reports, letters, the public report on the 9/11 terrorist attack, popular culture texts from Slate Magazine, telephone transcriptions from two-sided telephone conversations, five non-fictions works on textiles and and child development , travel guides, and several fiction works written between 1912 and 2010. Due to the wide variety and sourcing of these works and their sources, the model works to generalize better than the SNLI base model, as expected.

3.3 Fine Tuned Multi-NLI/SNLI Models

The Multi-NLI/SNLI model was fine tuned using the Watson, ANLI, and Reddit datasets. Performance improvements for task-specific fine-tuning was seen in line with the fine-tuned SNLI models but the Multi-NLI/SNLI model seemed to carry a baseline of lower overall loss indicating that the model was more confident in its classification decision. There is likely less bias introduced from this pre-trained base than what was seen on the SNLI base model. The Watson fine-tunes SNLI/Multi-NLI model saw the highest accuracy on the Watson dataset of over 93% (Figure 4). The lowest performing model was the out-of-domain Reddit Sarcasm dataset but once fine-tuned using the Sarcasm inputs, the resulting fine-tuned model seemed fairly adept at identifying sarcastic commentary with accuracy of 69% with just the fine-tuning and no transformer parameter or performance tuning accounting for the accuracy increase or loss reduction.

4 Conclusion

Pre-trained models provide several advantages to construct low-resource domain specific tasks on new datasets. By utilizing these pre-trained models and fine-tuning them with domain or task specific inputs enabled rapid low-cost performance improvements. These fine-tuned

models are then able to tackle a broad set of NLP tasks from identification of classification outside of the trained language (Watson) to adversarial attack (ANLI) and an attempt at using fine-tuned models to be utilized as binary classifiers for tasks as diverse as sarcasm detection (Reddit).

We were able to verify that low-resources solutions exist by utilizing pre-trained models and fine-tuning them for specific domain tasks but these models faced challenges when asked to generalize to an out-of-domain test set. The datasets based on human operators tasks with dataset specific construction tasks show the highest bias in language. We show here that increasing the training data can overcome some of those embedded biases for contradictor language. The best promise came from the Multi-NLI /SNLI dataset which seemed to help eliminate some human biases but retained racial and gender biases due to the texts used to construct the corpus. Depending on the task, further research will be needed to eliminate human biases in constructing future NLI dataset. However, when training on domain specific entries, it may be an improvement to detection of out-of-bounds discourse as is found on internet forums or comments that may include such biases and models may be required to identify those semantics without sanitizing the data to remove these biases. The results discussed here show that by increasing the amount of training data and fine-tuning the pre-trained models to domain specific tasks, low-resource models can be constructed that show remarkable performance gains. Additional performance gains are likely by fine-tuning the underlying transformer structure and model parameters.

References

[Bowman et al 2015] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large anno- tated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

[Clark et al.2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre- training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR).

[Devlin et al 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguis- tics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Com- putational Linguistics.

[Nie et all 2019] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. arXiv preprint arXiv: 1910.14599

[He wt al 2019] He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 132–142, Hong Kong, China. Association for Computational Linguistics.

[Paperno et al 2016] D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Association for Computational Linguistics (ACL).

[Reddit 2017] Mikhail Khodak and Nikunj Saunshi and Kiran Vodrahalli, A Large Self-Annotated Corpus for Sarcasm, 2017, https://arxiv.org/abs/ 1704.05579}

[Rimell et al 2009] L. Rimell, S. Clark, and M. Steedman. 2009. Un- bounded dependency recovery for parser evaluation. In Empirical Methods in Natural Language Process- ing (EMNLP).

[Seonhoon et al 2019] Kim, Seonhoon, Inho Kang, and Nojun Kwak. "Semantic sentence matching with densely-connected recurrent and co-attentive information." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[Watson 2020] Amy Jang, Ana Sofia Uzsoy, Phil Culliton, Contradictory, My Dear Watson, Kaggle, 2020, https://kaggle.com/competitions/ contradictory-my-dear-watson

[MNLI 2017] Williams, Adina and Nangia, Nikita and Bowman, Samuel R., A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference}, 2017, https://arxiv.org/abs/ 1704.05426