Natural Language Processing - Final Project Report

Anonymous Group

Abstract

In Natural Language Processing (NLP) there are often trained models that in practice appear to perform very well when using test datasets, but when challenged with real-world examples it becomes evident that these trained models are not actually valid. In this paper, we will explore one of these pre-trained models on a Natural Language Inference (NLI) dataset. We will perform various analyses including Checklist sets and Contrast sets to determine if the pre-trained model actually has a high performance. Then we will use Adversarial datasets in an attempt to improve the model's performance and actually learn while also having a high evaluation rate.

1 Introduction

In this paper, we will examine the performance of the Electra-small model (Clark et al., 2020). This small model is computationally more efficient than a large model. The model is also similar to BERT Models, but its training methods are better and more developed. We will then train this model on a Natural Language Inference (NLI) dataset. We chose to train on the Stanford NLI dataset (Bowman et al., 2015) we will refer to this dataset as SNLI.

The SNLI dataset is made up of image captions from a Flickr30K corpus (Young et al. 2014). Stanford used crowd-sourcing in which they asked the public to read these image captions, and write three alternate captions:

1. A caption that is **definitely true** based on the original caption – **Entailment**
2. A caption that is **might be true** based on the original caption – **Neutral**
3. A caption that is **definitely not true** based on the original caption – **Contradiction**

In the SNLI dataset, the original image captions are referred to as the **Premise** and the alternate captions are the **Hypotheses**. We are training a model to predict the **Relation** between the Hypothesis and the Premise. The 3 relations are labeled as the following:

- **0 = Entailment (Definitely True)**
- **1 = Neutral (Might be True)**
- **2 = Contradiction (Definitely Not True)**

Every entry in the SNLI dataset consists of a Premise, a Hypothesis, and a relationship Label. The dataset is split into 3:

1. Train set consisting of 550,152 entries
2. Validation set consisting of 10,000 entries

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: SNLI Dataset Example

Premise	Hypothesis	Label
"A couple walk hand in hand down a street."	"A couple is walking together."	0 (entailment)
"A couple walk hand in hand down a street."	"The couple is married."	1 (neutral)
"A couple walk hand in hand down a street."	"A couple is sitting on a bench."	2 (contradiction)

Table 2: Electra-Small Model Labels vs. Checklist Labels

Labels	Electra-Small Model	Checklist
Positive (Entailment)	0	2
Neutral	1	1
Negative (Contradiction)	2	0

3. Test set also consisting of 10,000 entries

An example entry from the SNLI dataset is in Table 1: SNLI Dataset Example, from Hugging Face (Angeli 2015). We trained our model, the Electra-small model, on the SNLI train dataset for three epochs which resulted in a **train loss of 0.41313**. We then ran an evaluation on our trained model which resulted in an **evaluation accuracy score of 0.89748 and an evaluation loss of 0.37624**.

This is a fairly high accuracy of almost 90%, but has our trained model actually learned, or is the high accuracy due to some other correlation that it learned instead? We are going to use a few analysis methods to determine this and see if we can fix the model to perform better.

2 Analysis

There are a number of reasons we need to perform analysis on our model to understand if the model is learning or if the model has found a workaround allowing it to perform well. Since the SNLI data is from crowd-sourcing, one of our main concerns is that it might be subject to bias. To test to see if the model is actually learning we are going to run some analysis. In our analysis of the model, we decided to perform two different methods of changing data: using checklist sets (Ribeiro et al., 2020) and using contrast sets (Gardner et al., 2020).

2.1 Checklist

Checklist sets are a method of behavioral testing for NLP Models from the Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (Ribeiro et al., 2020). Checklist uses various linguistic capabilities that it then makes unit tests from in order to evaluate the NLP model to see if it has learned different linguistic capabilities. These tests then provide an output to see what behaviors the model has failed to learn. For our model, we will use the Vocabulary + POS (Part of Speech) capability which determines if the model has a sufficient vocabulary, and tests to see if the model can manage words having different parts of speech. We will use this Vocabulary + POS capability in order to test the model's understanding of sentiment. This test will check if the model can appropriately determine the sentiment of a word: positive, negative, or neutral sentiment.

Checklist has three separate test types used to evaluate the different capabilities. These tests are Minimum Functionality tests (MFT), Invariance tests (INV), and Directional Expectation tests (DIR). In this paper, we will focus on the MFT tests. The Minimum Functionality test (MFT) checks behavior and capability using a small sample of easy examples and labels. The MFT model basically creates unit tests that are very useful for determining when the model is not actually learning. It can determine where the model is not mastering the capability at hand and is instead using some shortcut to receive high accuracy and perform well.

In the Checklists test, we are testing our base model on the data from the "Twitter US Airline Sentiment" dataset that we are given in the source code from the Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (Ribeiro et al., 2020). We also must adjust our model labels after each prediction in order to run the checklist tests because the Checklist labels are 2= Positive, 1 = Neutral, and 0 = Negative which is the reverse of the Electra-small model. Table 2: Electra-Small Model Labels vs. Checklist Labels displays these label changes.

Table 3: Minimum Functionality Test Results for Vocabulary + POS Capability

Test Name	Cases Failed \Total Test Cases	Failure Rate %
Single Positive Words	3\34	8.8%
Single Negative Words	35\35	100%
Single Neutral Words	12\13	92.3%
Sentiment-laden Words in Context	278\500	55.6%
Neutral Words in Context	500\500	100%

When testing Vocabulary + POS capabilities using MFT tests we will run tests for Single Positive Words, Single Negative Words, Single Neutral Words, Sentiment-laden Words in Context, and Neutral Words in Context. This is an example from Beyond Accuracy: Behavioral Testing of NLP Models with CheckList of how these test cases are made (Ribeiro et al., 2020).

- Start with the Template "I {NEGATION} {POS-VERB} the {THING}."
  - {NEGATION} = {didn't, can't say I, ...}
  - {POS-VERB} = {love, like, ...}
  - {THING} = {food, flight, service, ...}
- Then produces test cases similar to these: "I didn't love the food." and "I can't say I love the service."

We will run the tests for single positive words, single negative words, single neutral words, Sentiment-laden words in context, and neutral words in context then determine the failure rate for each test to see where our model is failing to learn. The results from these Minimum Functionality tests are in Table 3: Minimum Functionality Test Results for Vocabulary + POS Capability. The list below shows examples from the tests where our model fails for the Vocabulary + POS Capabilities.

- Single Positive Words
  - great
  - welcomed
  - value
- Single Negative Words
  - bad
  - frustrating
  - lousy
- Single Neutral Words
  - Italian
  - Australian
  - see
- Sentiment-laden Words in Context
  - That is a poor pilot.
  - This is an awful company.
  - This was an awful seat.
- Neutral Words in Context
  - That is a commercial seat.
  - That was a commercial customer service.
  - The flight was Israeli.

When analyzing the results from Table 3 and looking at the failed examples it is very evident our model performs very poorly with Negative and Neutral examples. This shows that our model has not actually learned and is finding some other way to perform well on evaluation to produce a high accuracy score. In the Fixing-it section, we will work on trying to fix the issues our model is failing on to improve our NLP model's performance.

2.2 Contrast Datasets

The second analysis we are doing is changing the data using contrast datasets. In Contrast datasets, we manually change the dataset in a small but meaningful way that can change the label of the hypothesis. By doing so we generate the examples that create the boundary line for classification between different labels. Figure 1 is an example of contrast sets from the paper Evaluating Models' Local Decision Boundaries via Contrast Sets on Textual Perturbations used in Contrast Datasets (Gardner et al., 2020).

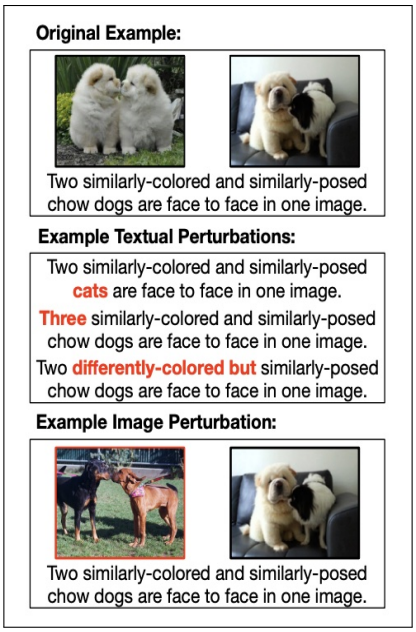

Figure 1: Contrast datasets example

We made these contrast datasets by going through some examples from our training and testing data and created 50 examples of contrast data manually. We tweaked the data a bit to change the label. Below is an example of this:

1.
   - Premise: "A man, woman, and child enjoying themselves on a beach."
   - Old Hypothesis: "A family of three is at the beach" (entailment)
   - New Hypothesis: "The family will go to the beach tomorrow" (neutral)
2.
   - Premise: "Two women who just had lunch hugging and saying goodbye."
   - Old Hypothesis: "There are two women in this picture" (entailment)
   - New Hypothesis: "There are three women in this picture" (contradiction)

In addition to this, we created many labels with simple negation (not) in them that did not belong to contradiction to test the efficiency of our model when dealing with negations. Below is an example of this:

1.
   - Premise: "A woman with a green headscarf, blue shirt, and a very big grin."
   - Hypothesis: "The woman is not sad."
   - Label: Entailment
2.
   - Premise: "A man playing an electric guitar on stage."
   - Hypothesis: "The man is not performing for charity."
   - Label: Neutral

After implementing these changes we ran our base model that is trained on the SNLI dataset. The performance significantly decreased on the contrast dataset and the **accuracy fell from 89% to 56%.** We inspected to see where the model was failing and discovered that our model predicted contradiction in most of the negation cases. Below are two examples of our model predicting poorly on negation tests:

1.
   - Premise: "This church choir sings to the masses as they sing joyous songs from the book at a church."
   - Hypothesis: "A choir not singing at a book store"
   - Label: Entailment

Table 4: Contrast Sets Result Table

	Entailment	**Neutral**	**Contradiction**
Total True labels	20	17	13
Correctly predicted	14 (70%)	12 (70%)	2 (15%)
Incorrect predicted	6 (30%)	5 (30%)	11 (85%)

- Predicted label: Contradiction

2.
   - Premise: "A few people in a restaurant setting, one of them is drinking orange juice."
   - Hypothesis: "The people are not drinking wine"
   - Label: Neutral
   - Predicted label: Contradiction

Overall we had 28 hypotheses with negations and 7 incorrect prediction labels. All of the incorrect predictions were indeed contradictions.

Our model was also unable to distinguish the slight changes in our data and therefore predicted the tweaked data incorrectly. Below are examples of the wrong predictions made by the model for the tweaked data we made:

1.
   - Premise: "Two women who just had lunch hugging and saying goodbye."
   - Old Hypothesis: "There are two women in this picture" (Entailment)
   - New Hypothesis: "There are three women in this picture" (Contradiction)
   - Predicted label: Entailment
2.
   - Premise: "Two older men are talking."
   - Old Hypothesis: "Two people are having a conversation." (Entailment)
   - New Hypothesis: "Three people are having a conversation" (Contradiction)
   - Predicted label: Entailment
3.
   - Premise: "A big brown dog swims towards the camera."
   - Old Hypothesis: "A dog swims towards the camera." (Entailment)
   - New Hypothesis: "A dog swims towards the underwater camera." (Neutral)
   - Predicted label: Entailment

From the evidence in Table 4, it is clear that our model needs more training towards examples with contradictions as it did really badly on these. This is one thing we will work on fixing in the next section.

3 Fixing It

3.1 Method

In order to fix our model we will use an adversarial dataset to train the model (Liu et al., 2019; Zhou and Bansal, 2020; Morris et al., 2020). In recent years, large-scale benchmark models like SNLI and SQuAD are being tested on multiple tasks to track their performances. Researchers have seen that instead of the models understanding the meaning in a general way like humans do they are actually following statistical patterns in order to classify the data. This has made it easy for the researchers to create examples that are used as tests to show where the model lacks in performance when compared to a real human being.

In order to fix our model we will use the dataset from the paper Adversarial NLI: A New Benchmark for Natural Language Understanding and the associated code (Nie et al., 2020). We will start by taking our base model, the one trained on the SNLI dataset, and training it on the adversarial dataset. Then we will test our new model in two ways: 1. By testing it on the base SNLI test dataset and 2. testing the model on the same contrast dataset from part 2.2 Contrast datasets.

Table 5: New Model Tests Evaluation Results

Tests	Loss	Accuracy
Contrast Dataset	1.6086	0.6800
SNLI Test Dataset	0.7353	0.8086

Table 6: Results Before and After Fixing the Model

Models	SNLI Test Dataset	Contrast Test Dataset
SNLI Base Model	89%	56%
Adversarial Dataset + SNLI Model	80%	68%

3.2 Results

In both testing methods, we combined three adversarial datasets and the SNLI dataset for training. In one we tested the new model on the contrast dataset and in the second we tested the model on the SNLI dataset. The results of the two tests are in Table 5: New Model Tests Evaluation Results.

As you can see from Table 5, both tests on the new model **1. On the contrast dataset with an evaluation accuracy of 68%** and **2. On the SNLI test dataset with an evaluation accuracy of 81%,** performed fairly well (we are going to use 60% accuracy as a benchmark for a good evaluation accuracy).

In Table 6: Results Before and After Fixing the Model, we have compared the results of the model on the two tests before and after training on the three additional adversarial datasets. Our updated model did comparatively better when tested on the contrast dataset, but a little lower than the base performance on the SNLI test dataset. We expect that this lower percentage on the SNLI test dataset is due to the model actually having to learn and not using a cheat or workaround to have high performance.

3.3 New Models Performance on Contrast Data

When examining our new model's performance on the contrast dataset test compared to that of our base model we see the accuracy increase from 56% to 68% by training on adversarial datasets. Let's dive into the analysis of the predictions generated by our model.

Our base model was unable to break the connection between negation with contradiction, however, our final model was able to unlearn this pattern and perform better with negation examples. In Table 7, Negation Examples in Old and New Model, see that not only the number of wrong predictions were less but the wrong predictions as contradictions were almost 50% rather than 100% in our base model. This is a great improvement in our new model, showing that our final model was able to handle the small changes made in our original dataset and predict them correctly, unlike our base model. The following list shows the examples from the Analysis section(2.2) but now includes how our final model performed.

1.
   - Premise: "Two women who just had lunch hugging and saying goodbye."
   - Old Hypothesis: "There are two women in this picture" (Entailment)
   - New Hypothesis: "There are three women in this picture" (Contradiction)
   - Gold Label: Contradiction
   - Predicted label By Base Model: Entailment

Table 7: Negation Examples in Old and New Model

Models	Total Negation Exs.	Wrongly Predicted	Contradiction Predicted
SNLI Base Model	28	7	7
Adversarial Dataset + SNLI Model	28	6	3

Table 8: Old vs. New Model Incorrect Predictions

	Entailment	**Neutral**	**Contradiction**
Total True Labels	20	17	13
Old Model Incorrect Predictions	6 (30%)	5 (30%)	11 (85%)
New Model Incorrect Predictions	7 (35%)	2 (12%)	8 (60%)

- Predicted label By Final Model: Contradiction

2.
   - Premise: "Two older men are talking."
   - Old Hypothesis: "Two people are having a conversation." (Entailment)
   - New Hypothesis: "Three people are having a conversation" (Contradiction)
   - Gold Label: Contradiction
   - Predicted label By Base Model: Entailment
   - Predicted label By Final Model: Contradiction
3.
   - Premise: "A big brown dog swims towards the camera."
   - Old Hypothesis: "A dog swims towards the camera." (Entailment)
   - New Hypothesis: "A dog swims towards the underwater camera." (Neutral)
   - Gold Label: Neutral
   - Predicted label By Base Model: Entailment
   - Predicted label By Final Model: Neutral

From the above examples and Table 8: Old vs. New Model Incorrect Predictions, it is very evident that our Final model did significantly better in Neutral and Contradictory cases which is where both our Checklist tests and Contrast tests failed before; however, the performance decreased in the case of entailment.

A possible next step is to research how to improve the model further on entailment-focused data that can make our predictions more accurate.

4 Conclusion

We began by training our Electra-small model on the Stanford NLI dataset and evaluated this NLP model. The model had high accuracy, but we needed to check if the accuracy was overestimating the NLP model's performance. We began by doing rigorous testing on our model using two methods: Checklist and Contrast data. These analysis methods allowed us to explore the annotation artifacts and biases in the SNLI dataset, which affected the trained model's overall performance. After further analysis, we were able to find the systematic gaps that are typically present in datasets, especially Crowd Sourcing datasets like this one, and tried to close these gaps by further fine-tuning our model in hopes of improvement.

In order to fix these issues, we used adversarial data to train our state-of-the-art model and find the loopholes where our data was not performing well. We trained on the adversarial examples that can fool a model into misclassifying the data and overcome the vulnerabilities that were present in our base Electra-small model trained on the SNLI dataset.

After we trained on these new adversarial datasets and the SNLI datasets we evaluated our new model and compared its performance to our base model. Then we discussed the areas where our final model outperformed the base SNLI model and showed specific example outputs from our results.

While we rigorously tested and improved our model, there are still thousands of new test examples across a wide variety of datasets that in the future we could run to improve the model further. Over the last few years, there have been very successful model advancements in terms of evaluation methodologies in Natural Language Processing and we hope to see the trend of improvements continue in the future.

References

[Angeli 2015] Angeli, G. 2015. Snli · datasets at hugging face. snli · Datasets at Hugging Face. Retrieved November 30, 2022, from https://huggingface.co/datasets/snli

[Bowman et al. 2015] 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.

[Clark et al. 2020] 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).

[Eight 2019] Eight, F. 2019, October 16. Twitter us airline sentiment. Kaggle. Retrieved November 30, 2022, from https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

[Gardner et al. 2020] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna 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.

[Masolo 2021] Masolo, C. 2021, August 4. Sentiment analysis on US Twitter Airline Dataset - 1 of 2. Medium. Retrieved November 30, 2022, from https://towardsdatascience.com/sentiment-analysis-on-us-twitter-airline-dataset-1-of-2-2417f204b971

[Morris et al. 2020] John X. Morris, 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.

[Nelson et al. 2019] 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, June. Association for Computational Linguistics.

[Nie et al. 2020] Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. 2020, July 10. Adversarial NLI: A New Benchmark for Natural Language Understanding. Retrieved December 5, 2022, from http://aclanthology.lst.uni-saarland.de/2020.acl-main.441.pdf

[Phan 2022] Phan, C. 2022, May 24. NLP: Explore Data Artifacts in SNLI dataset with checklist and electra. Medium. Retrieved November 30, 2022, from https://towardsdatascience.com/nlp-explore-data-artifacts-in-snli-dataset-with-checklist-and-electra-ebbdd1b83cd0

[Potts 2020] Potts, C. 2020, April 29. Natural Language Inference. CS224U: Natural Language Understanding - Spring 2020. Retrieved December 5, 2022, from https://stanford.edu/class/cs224u/2020/

[Potts 2020] Potts, C. 2020. Natural language Inference: Dataset artifacts and adversarial testing. CS224u: Natural language Understanding. Retrieved December 6, 2022, from https://web.stanford.edu/class/cs224u/slides/cs224u-nli-part3-handout.pdf

[Riberio et al. 2020] Marco Tulio Ribeiro, 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, Online, July. Association for Computational Linguistics.

[Young et al. 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2:67–78.

[Zhou and Bansal 2020] Xiang Zhou 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, Online, July. Association for Computational Linguistics.