Utilizing Data Augmentation to Improve NLI Classification

Jay Upadhyaya
jau464
jayupa@utexas.edu

University of Texas at Austin

Abstract

Natural Language Understanding (NLU) models often struggle with linguistic phenomena such as negation and erroneous text, which poses a significant challenge in natural language inference (NLI). This paper will explore the weakness of an ELECTRA-small model initially trained on the Stanford Natural Language Inference (SNLI) dataset on such negations and noisy examples. To improve performance, adversarial data augmentation is utilized by injecting synthetically generated negated, misspelled, and grammatically erroneous examples into the training data to enhance the model's sensitivity. Preliminary results suggest that this approach yields moderate improvements in the model's ability to handle examples that display negation and noisy data along with enforcing the model to generalize results rather than relying on spurious connections between specific keywords.

1 Introduction

NLU models enable tasks such as question answering and natural language inference, but they tend to struggle with linguistic phenomena in real-world scenarios. More specifically, models occasionally fail for negation and noisy inputs, including misspellings, grammatical errors, and adversarial examples. These limit how well models can generalize in post-training contexts.

During the training phase, models sometimes learn surface-level patterns or spurious correlations within the training data. Therefore, it can be difficult to understand why NLU models make certain classification decisions, but one method to do so is by analyzing patterns in the training dataset and predictions in the evaluation set. Models that rely on such patterns in the training data may fail when encountering examples that do not follow such patterns, such as negations. Similarly, noisy data (characterized by syntactical and typographical errors) pose additional difficulties for models trained on cleaner datasets.

This paper investigates the performance of an ELECTRA-small model (Clark et. al, 2020) trained on the Stanford Natural Language Inference (SNLI) dataset, with a specific focus on its handling of negation and noisy examples. ELECTRA is incredibly effective at NLI tasks, and the initial portion of our study focuses on analyzing shortcomings in the model's capabilities. I first inspected the dataset to find patterns in the failed predictions on the evaluation set and found that it struggles with examples mentioned previously, such as hypotheses or premises that have negations, misspellings, or grammatical inaccuracies. By analyzing the shortcomings and addressing them through adversarial data augmentation, the goal is to improve the model's robustness . The contributions are as follows:

1. Identifying and quantifying the limitations of an ELECTRA-small model on negation and noisy data within the SNLI dataset.

2. Building upon an adversarial data augmentation approach (Wei & Zou, 2019), incorporating synthetically generated examples with negation, misspellings, and grammatical errors on the training of a model to make the model more resilient to such phenomena and increase the ability to generalize.

3. Demonstrating the efficacy of this approach through quantitative and qualitative evaluations, showing moderate improvements in loss and accuracy.

2 Approach

2.1 Analyzing Existing Limitations

Initially, I aimed to find which predictions ELECTRA-small failed on in the SNLI dataset. ELECTRA-small was trained on the unaugmented dataset and run on the evaluation set to generate predictions. Following this, I analyzed the types of errors found in the predictions. The model predicts 'entailment,' 'neutral,' or 'contradiction' based on the hypothesis and premise in the dataset, and Figure 1 below shows the confusion matrix of predictions made by the baseline model.

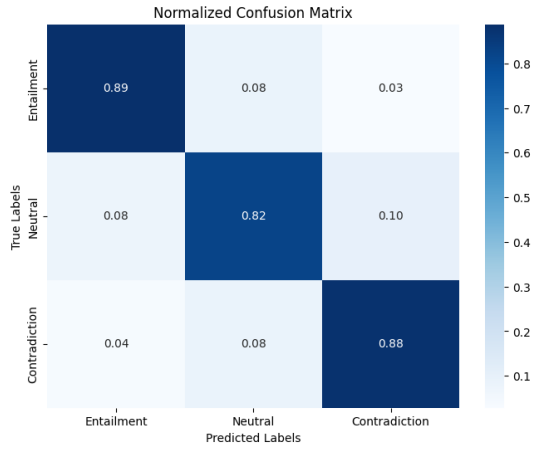

Figure 1: Predictions made by ELECTRA-small on the SNLI Dataset

Around 7% of the total predictions (∼17% of the total errors made) predict entailment when the hypothesis is a contradiction or vice versa. Following this, I analyzed the dataset and discovered examples where either negations are present or a negation could assist with a prediction. One such example where the model failed to predict was with the premise "A young boy with sandy blond-hair and white and black soccer uniform kicking for the goal while parents look on", and a hypothesis of "The boy's parents are not present." In this example, contradiction is the gold label, but the label predicted by ELECTRA is entailment.

Another subset of examples on which the model occasionally performs poorly are examples with typographical or grammatical errors. One such example is with the premise "A brown a dog and a black dog in the edge of the ocean with a wave under them boats are on the water in the background" and the hypothesis "The dogs are swimming among the boats". The model failed to predict the gold label (entailment) due to grammatical inconsistencies in the premise. In addition, 50 random examples were extracted from the SNLI dataset and modified to get a baseline percentage for each type of example.

2.2 Data Augmentation

To address these shortcomings, we extended the SNLI dataset by creating adversarial examples using programmatic rules. Misspellings were introduced by inserting random characters into randomly selected words in the premises and hypotheses. Grammatical errors were introduced similarly by swapping randomly selected adjacent words in the premises and hypotheses. Negated examples were formed by either removing or inserting negation keywords in sentences. The process of negation also introduced grammatical errors into the system (i.e. adding "not" after the first one negates the sentence but is not always syntactically correct). This process was done for each example in the SNLI dataset to maximize the model's exposure to such phenomena. By introducing randomness into the process, the model is less likely to create spurious correlations between specific words or phrases.

2.3 Experiments

To first analyze whether the model trained on the augmented data set was effective in handling these phenomena, its performance on the 50 randomly selected examples from 2.1 was an-

alyzed. Table 1 below compares the results of the ELECTRA-small model and the ELECTRA-Augmented models, showing that the fine-tuned model was more effective in handling the noisy and negated data. For all types of examples, the fine-tuned model was more accurate than the baseline ELECTRA-small model.

Table 1: ELECTRA-Small vs Augmented Model on Non-Standard Data

Example Type	Baseline	Fine-tuned
Original	0.94	0.96
Negated	0.50	0.66
Typographical Errors	0.82	0.86
Grammatical Errors	0.84	0.94

3 Results

The new model and the baseline model were then trained on different sample sizes. Following this, they were both run against an evaluation set to analyze each model's performance on a more exhaustive set of scenarios. Tables 2 and 3 below compare the accuracy and loss of the two models. Due to resource constraints, it was more efficient to compare both at lower training sample sizes, and the correlation allows us to extrapolate the results for larger datasets. The fine-tuned model, at the largest sample size of 110000 training samples, had an accuracy of ∼2.4% more than that of the baseline and a moderately lower loss.

Table 2: Accuracy for ELECTRA-Small vs Augmented Model on SNLI

Sample Size	Baseline	Fine-tuned
10000	0.751	0.759
20000	0.763	0.774
30000	0.772	0.785
40000	0.784	0.799
50000	0.791	0.810
110000	0.842	0.866

Table 3: Loss for ELECTRA-Small vs Augmented Model on SNLI

Sample Size	Baseline	Fine-tuned
10000	0.838	0.852
20000	0.825	0.824
30000	0.769	0.761
40000	0.753	0.749
50000	0.737	0.728
110000	0.626	0.558

The decrease in loss may be attributed to similar copies of examples being inserted into the databse; however, due to the random augmentation, the potential for over-fitting should be minimized. Further, the fine-tuned model correctly predicted the examples outlined in 2.1 that the baseline model initially failed.

Following the introduction of augmented data, the model is more capable in making predictions on noisy and non-standard data. Further, it handled negation cases far more effectively, as shown in the confusion matrix of the trained model in Figure 2. Compared to the baseline model, it made fewer entailment/contradiction errors, with only around ∼4% of the total predictions (or ∼12% of all errors) landing in those categories. Improvements extended to unseen examples, as showcased by the overall accuracy improvement, indicating enhanced generalization capabilities. The model still exhibits occasional reliance on patterns, indicating that this approach does not fully remove artifacts in training data.

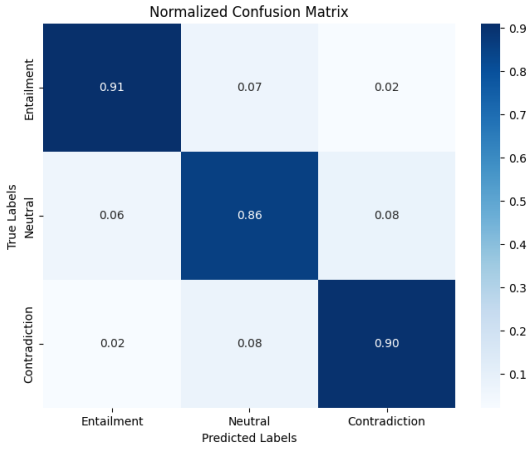

Figure 2: Predictions made by the fine-tuned model on the SNLI Dataset

4 Conclusion

Our model saw a ∼2.4% accuracy improvement upon the baseline ELECTRA-small model, which showcases that adversarial data augmentation can be an effective tool in improving model performance in real-world scenarios. However, it should be noted that this process is computationally expensive, as it has the potential to double or triple the size of the training dataset. Further steps could include injecting other types of data into the dataset to increase the robustness of the model, such as synonym or antonym replacement. Additionally, it would be worth exploring replacing examples in the dataset with perturbed examples rather than including copies to ensure overfitting is not an issue. In any case, the performance increase indicates that the data augmentation approach has the potential to break down dataset artifacts and should be considered to prevent models from learning spurious correlations from training data.

4.1 Acknowledgements

I'd like to acknowledge and thank Professor Durrett and the TAs for a fantastic semester. I have gained a lot of knowledge and interest in the field of NLP, and I'm excited to continue learning after the course.

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.

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).

Robin Jia and Percy Liang. 2017. *Adversarial examples for evaluating reading comprehension systems.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. https://arxiv.org/abs/1901.11196.*