Analyzing and Mitigating Dataset Artifacts in NLI

Eric Yeung
Masters in Data Science
University of Texas in Austin
Austin, Tx 77030
ericyeung281@utexas.edu

Abstract

In order to analyze the effects of data set artifacts and the potential methods of mitigating such artifacts, the ELECTRA-small model was used in order to analyze the results of the SQuAD data set in NLI (natural language inference). Models were initially trained on the full SNLI data set, and the model's accuracy was testing on SNLI test sets as well as contrasting test sets (MNLI) and antagonistic test set (ANLI). These accuracy metrics were then compared to a new model created using both the SNLI and ANLI datasets for training with the intent of mitigating the data set artifacts by increasing the scope of data introduced to the model. The result showed indicated incremental increase in accuracy an all data sets, but the effect of the new model in combating artifacts is inconclusive.

1 Introduction

This report details the analysis of the ELECTRA-small ([2] model in NLI. The model was initially trained using data sets from SQuAD ([4]) and tested with various data sets. The purpose was to analyze the accuracy of the model to classify sentence pairs as either contradiction, neutral, entailment. The results of the model's inference would be analyzed to determine if any data set artifacts exist and how strongly they affect the model's ability to accurately make inferences on different types of sentence pairs. Once these artifacts were analyzed and categorized, a new model would be trained using not only the SNLI data set from SQuAD, but also ANLI in an attempt to remove priority on data set artifacts. In order to accurately make comparisons between the base model and the new mixed-model, the models will be tested using various different data sets to show any changes in accuracy or significant changes in the class of examples that the model categorizes incorrectly.

1.1 Approach

The initial ELECTRA-small model is trained using the testing set from the SNLI dataset. This model is then tested using the SNLI, ANLI and MNLI datasets to test the model accuracy against multiple types of data. The SNLI data set provides a baseline accuracy on the model with which we will be comparing with the adjusted model. The ANLI and MNLI data sets will also be analyzed for accuracy, but the focus of the analysis will be on the types of sentences pairing that the model has difficulty solving. After collecting the initial model data, a new model will be trained using both the training set from SNLI as well as the r1_training set from the ANLI data set. The objective of training with the ANLI test set is to introduce significantly different data from the original training set. By using a deliberately generated antagonistic training set, the hope is that the new model will have a stronger understanding of relationships between words [1]. The new mixed model will be analyzed using the same testing sets from the original model and the same metrics will be analyzed. These metrics being the accuracy on each data set as well as analysis on which type of sentence pair examples the model appears to get incorrectly most of the time. We will compare the two model

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

results and determine whether conclusions can be drawn regarding the effects of data artifacts on the model's performance.

2 Results

2.1 Base model analysis

The base ELECTRA-small model trained from only SNLI training sets resulted in the follow accuracy for each of the corresponding testing sets in as shown in the table below. From all mistakes made for each data set, the percentage of mistakes made for each prediction:label pairing is shown in table 2 with 0,1 and 2 representing entailment, neutral, and contradiction classifications respectively.

Table 1: Accuracy of base model

SNLI	**MNLI (GLUE)**	**ANLI (r1_testing)**
%	%	%
0.89250	0.69628	0.24500

Table 2: % of mistake pairing to total number of mistakes for given test set

Label Pair	**SNLI**	**MNLI (GLUE)**	**ANLI (r1_testing)**
0:1	0.201	0.193	0.113
0:2	0.065	0.162	0.166
1:0	0.225	0.184	0.180
1:2	0.222	0.181	0.156
2:0	0.071	0.080	0.186
2:1	0.216	0.170	0.199

Looking into each dataset individually, model had the most difficulty handling sentence pairs in the SNLI data set that involved neutral sentiment in either the premise or hypothesis. As shown in table 2, the percentage of mistakes made for each prediction:label pair indicates the model's difficulty handling neutral sentiment. This may potentially hint at data set artifacts as the model could be seeing similar, but not identical, strings in the premise and hypothesis and making a decision without deeply understanding the context.

For example,one of the pairs the model incorrectly labeled is as follows.
Premise: "Two men on bicycles competing in a race."
Hypothesis: "Men are riding bicycles on the street."
With the similar contexts of "men" and "bicycles" the model may have incorrectly assigned entailment without understanding that a race does not inherently imply riding on a street.

Premise: "Men stand on steps leading out of a body of water"
Hypothesis: "Men are standing by a docked boat." This example indicates potential artifact between words that would often appear in the same context.
In this case, the model appears to process "water" and "boat" and lean towards entailment without referring to the context of the sentence itself.

More applicably, it seems that among the SNLI data set, a commonality among mislabels is the introduction of new context or information.
Premise:"A man wearing sunglasses is sitting on the steps outside, reading a magazine."
Hypothesis: "The man is reading a spoon with the words "HELP ME" on it."
For this example, the model incorrectly guessed neutral when the label was contradiction potentially because "reading" in the model has been seen in context with many other object which causes the model to lean towards neutral or because the introduction of new context of "HELP ME" or "spoon" confused the model.

The final noticeable class of error made by the model involves words that likely have very low representation in the model.
Premise:"The two farmers are working on a piece of John Deere equipment."
Hypothesis: "2 Men are making a pool together"

In this example, "John Deere" is likely a string that has very low representation in the data set and the model is likely unable to provide strong relationships between it and other words or phrases.

The MNLI test set presents syntactical confusion to many of the sentence pairs the model has difficulty dealing with.
Premise:"Oh, what a fool I feel! "
Hypothesis:"I am beyond proud."
The altering of normal sentence structure or the use of multiple negatives appear to be a common trend of mistake within the MNLI data set for the base model. As one would expect, the introduction of multiple negatives cause the model to lose track of meaning within the context of a sentence. This is highlighted by the predictions of the model. Out of the 2981 pairs that the model labeled incorrectly, the model only predicted "contradiction" 25.7% of the time. This is a different behavior from the SNLI test set where most of the mistakes were concentrated among pairs involving neutral sentiment. The model failing to predict contradiction potentially indicates that the model is able to identify some relationship between contexts between the premise and hypothesis, but is not strong enough to parse through the complex syntax or noise that is associate with spoken speech examples.

Because the MNLI dataset samples from novels and speech, the use of idiom and colloquialism also affect the model's accuracy. While trained on the SNLI data set which is a much more carefully curated set of sentence pairs, the model has difficulty assigning meaning and context to phrases such as "off you go" and Just suddenly popped up." In addition to slang, pulling data from human speech results in filler words and noise to be introduced to the premise which the model has not been trained to account for. For example, "it's actually there well Iraq has had uh designs on that place since nineteen twenty two so you know it wasn't like something that just suddenly popped up" The introduction of "well", "uh", and "um" can break up the syntactical structure of a sentence and harm the model's accuracy.

The ANLI test set presents the most difficult challenge for the model as it is an adversarial human-and-model-in-the-loop generated data set. The base model was only able to correctly label 24.5% of the pairs within the test set. Interestingly, the model leaned heavily towards prediction contradiction and predicted entailment significantly less. Among the 705 incorrectly labeled pairs, 27.9% were labeled entailment, and 38.4% were labeled as contradiction. Analyzing the mislabeled pairs, the mistaken labels can be categorized as mistakes of context or mistakes of syntax.

For mistakes of context, the hypothesis will often propose a statement that requires additional knowledge beyond the scope of the premise. For example, in the sentence pair below, the model must identify that London is a location that exists within Britain which exists within Europe in order to correctly understand the context between the premise and hypothesis. Additionally, the ANLI data set involves instances in which the model must understand concepts of ranking and ordering. In this example specifically, the model must be able to understand what it means to be the first store as opposed to being any store opened in Europe.
Premise:"Ernest Jones is a British jeweller and watchmaker. Established in 1949, its first store was opened in Oxford Street, London. Ernest Jones specialises in diamonds and watches, stocking brands such as Gucci and Emporio Armani. Ernest Jones is part of the Signet Jewelers group."
Hypothesis:"The first Ernest Jones store was opened on the continent of Europe."

For mistakes of syntax, the ANLI data set seeks to confuse the model, by altering the syntactic structure between premise and hypothesis to reduce the effect of data set artifacts on prediction as well as introducing superfluous information in an attempt to confuse the model's ability to detect the context required to compare the premise and hypothesis. The following example, presents "Macintosh" as three distinct but related entities within the premise and hypothesis: as a operating systems, a series of personal computers, and a business. Additionally, the premise includes excessive information that is unrelated to the hypothesis that would hinder the base model's ability to correctly predict context given the original training data.
Premise:"The family of Macintosh operating systems developed by Apple Inc. includes the graphical user interface-based operating systems it has designed for use with its Macintosh series of personal computers since 1984, as well as the related system software it once created for compatible third-party systems."
Hypothesis:"Macintosh is a business that owns Apple Inc."

2.2 Revised Model: Mixed Model

The revised model is an ELECTRA-small model trained using both the SNLI training set and the r1_training set from the ANLI data set. This model was created with the intention of reducing data artifacts and improving model accuracy by introducing antagonistic data to the training and giving training data that intention of introducing new contextual relationships that are not present in the SNLI data set [3]. The 16946 pairs in r1_training set are appended to the existing 550152 pair and used to train the mixed model. The mixed model's performance on the SNLI test set, MNLI test set, and r1_testing set are shown in table 3.

Table 3: Accuracy of models

Model	**SNLI** %	**MNLI (GLUE)** %	**ANLI (r1_testing)** %
Mixed Model	0.89565	0.69893	0.40400
Base Model	0.89250	0.69628	0.24500

The mixed model accuracy improved significantly on the ANLI test set and marginally improved on the SNLI and MNLI test sets. This result is unsurprising given that the relationships learned by the model through the ANLI test set involved the learning of complex relationships between words in a context, such as order, or rankings.

Looking once again through the sentence pairs the model predicted incorrectly, the mixed model seems to have increase weight towards guessing entailment across the entire test set. Among the incorrect labels, it is the only category in which the number of incorrect guesses increased, and entailment labels account for the largest increase in the model's accuracy. The MNLI test set yields similar results with an increase in entailment predictions across the data set. Surprisingly, the number of incorrect predictions increased in the case where the model predicted contradiction but the true label was entailment.

The ANLI results are similar to the other two test sets, with heavy predictive weight towards entailment when the model is unable to predict correctly. With the largest increase in accuracy, the mixed model was able to more accurately predict contradictory statements within the test set. This change in the model's behavior is noted in Table 4.

Table 4: Prediction\label % for each model

Label Pair	**Mixed Model**	**Base Model**
0:1	0.113	0.163
0:2	0.166	0.280
1:0	0.180	0.153
1:2	0.156	0.122
2:0	0.186	0.151
2:1	0.199	0.131

3 Analysis

The results of the mixed model indicates that the ANLI training set was able to create generalizations for obscure context or relationships between words. Across all three data sets, the model increased its predictions for entailment implying that the model was able to determine some kind of relationship between the premise and hypothesis across all types of data. Given the generated nature of the ANLI data set, the mixed model was unable to improve its performance on the MNLI test set. On the contrary, the mixed model increased the number of incorrect predictions in the case of predicting contradiction on pairs with a entailment true label. The spoken nature of the MNLI data set is an artifact which the mixed model is unable to accurately handle. With the introduction of colloquial phrases and slang in the MNLI data set and the presence of fill words within the premises, the mixed model is able to make sparse connections between the premise\hypothesis pairs, but does not have sufficient training to make accurate prediction. This is supported by the significant drop in "neutral" predictions made by the model. From the base model to mixed model, the amount of incorrect "neutral" predictions decreased by 2%.

The addition of the ANLI data was unable to increase the model's accuracy in parsing out context between similar words. We know this, because this type of ambiguous context is present in both the SNLI dataset and the ANLI data set, as seen in the examples in the base model discussion. Because there is no significant increase in accuracy in the SNLI test set, we can say that the mixed model is able to better recognize the existence of relationships between words, but is no better at parsing the context of that relationship between the premise and hypothesis. This low level understanding of relationship without context is what allows the model to guess contradiction and entailment more frequently, but unable to guess with increased accuracy.

With the understanding that there is no significant increase in the model's ability to understand context, there is one conclusion for the mixed model's accuracy on the ANLI test set. The introduction of the ANLI test set, has allowed the mixed model to better understand the concept of sequences and ordering. Of the examples in the base model incorrectly predicted, the mixed model was able to accurately predict examples involving a relationship of time or ranking. By introducing ANLI into the training set, the mixed model was able to learn this type of data artifact.

To better improve this mixed model, The simplest solution would be to add additional data to cover all of the data artifacts unaccounted for by the model. The colloquialism in the MNLI data set for example. A larger data set would also reduce out-of-dataset words which would significantly increase the model's ability to determine context. To prevent the effect of filler words or noise on the model's prediction, a part of speech tagger could be implemented to identify filler words and tangential phrases, and have the model be trained to be less biased towards those phrases.

References

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

[2] Quoc V. Le Kevin Clark, Minh-Thang Luong and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *Proceddings of the International Conference on Learning Representations(ICLR)*, 2020.

[3] Roy Schwartz Nelson F. Liu and Noah A. Smith. 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, 2019. Association for Computational Linguistics.

[4] Konstantin Lopyrev Pranav Rajpurkar, Jian Zhang and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceddings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics.