Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 427–435
Marseille, 20-25 June 2022

Mitigating Dataset Artifacts in Natural Language Inference Through Automatic Contextual Data Augmentation and Learning Optimization

Michail Mersinias[1], Panagiotis Valvis[1]
Department of Computer Science, University of Texas at Austin
{mmersinias, pval}@utexas.edu

Abstract

In recent years, natural language inference has been an emerging research area. In this paper, we present a novel data augmentation technique and combine it with a unique learning procedure for that task. Our so-called *automatic contextual data augmentation (acda)* method manages to be fully automatic, non-trivially contextual, and computationally efficient at the same time. When compared to established data augmentation methods, it is substantially more computationally efficient and requires no manual annotation by a human expert as they usually do. In order to increase its efficiency, we combine *acda* with two learning optimization techniques: *contrastive learning* and a *hybrid loss* function. The former maximizes the benefit of the supervisory signal generated by *acda*, while the latter incentivises the model to learn the nuances of the decision boundary. Our combined approach is shown experimentally to provide an effective way for mitigating spurious data correlations within a dataset, called *dataset artifacts*, and as a result improves performance. Specifically, our experiments verify that *acda*-boosted pre-trained language models that employ our learning optimization techniques, consistently outperform the respective fine-tuned baseline pre-trained language models across both benchmark datasets and adversarial examples.

Keywords: natural language inference, data augmentation, learning optimization

1. Introduction

Inference has historically been a central topic in artificial intelligence, and recently so in the natural language domain. The so called *natural language inference (NLI)* task is to determine whether a natural language hypothesis *h* can justifiably be inferred from a natural language premise *p* (MacCartney and Manning, 2008). This is a challenging task due to the complex nature of natural language which entails informal reasoning, lexical semantic knowledge and structure as well as variability regarding linguistic expression.

In recent years, there has been a considerable amount of research in this particular area. The *Stanford Natural Language Inference (SNLI)* (Bowman et al., 2015) corpus and the *Multi-Genre Natural Language Inference (MNLI)* (Williams et al., 2017) corpus have provided robust benchmark datasets as they contain a large amount of annotated data. Furthermore, the advancement of pre-trained language models has contributed to the increased effectiveness of proposed solutions.

Although many of these solutions appear promising, as they report high accuracy on validation data, the task of natural language inference remains a work in progress. Recent research calls into question the learning which results from pre-trained language models in datasets such as *SNLI* and *MNLI* because it either predicts the right answer when it shouldn't, as it is the case with hypothesis-only baselines (Poliak et al., 2018), or predicts the wrong answer if minor modifications are made by utilizing *contrast sets* (Gardner et al., 2020), *checklist sets* (Ribeiro et al., 2020) or *adversarial attacks* (Jia and Liang, 2017). These observations all stem from the fact that a model may achieve high performance on a dataset by learning spurious correlations, which are called *dataset artifacts*, but it is then expected to fail in settings where these artifacts are not present.

In our work, we identify dataset artifacts and propose a data augmentation and learning optimization approach in order to achieve a higher and more robust performance than the respective fine-tuned pre-trained baseline language models from the *Hugginface Transformers* repository (Wolf et al., 2019). Specifically, our contributions can be summarized as follows. First, we propose *acda*, a novel data augmentation approach for the construction of adversarial examples to enrich the dataset for the purpose of enhancing the learning process. Compared to established data augmentation techniques such as *TextAttack* (Morris et al., 2020) and *Checklist* (Ribeiro et al., 2020), *acda* is substantially more computationally efficient and, moreover, fully-automatic as it requires no manual annotation by a human expert as these packages do. Furthermore, we propose a *hybrid loss* function which allows the *acda*-boosted models to learn the nuances of the decision boundary, thus providing results that are considerably more robust than those of the default NLL loss function of the fine-tuned baseline pre-trained language models, which is solely based on the maximum likelihood estimation (MLE) criterion. In addition, we make use of *contrastive learning* (Dua et al., 2021) which maximizes the benefit of the supervisory signal generated by *acda* and further increases performance. Finally, we perform a systematic comparison and demonstrate experimentally that the *acda*-boosted pre-trained language models which employ our learning optimization techniques, consistently outperform the respective fine-

[1]The authors contributed equally to this work.

tuned baseline pre-trained language models across both the *SNLI* and the *MNLI* datasets. In order to further demonstrate the effectiveness of our approach, we also provide a set of multiple hand-annotated adversarial examples where the *acda*-boosted models exhibit a considerably more robust behavior and performance than the fine-tuned baseline models.

2. Background and Related work

Textual entailment is the relationship between a natural language premise p and a natural language hypothesis h. It is positive when the truth of p requires the truth of h, that is, when a human annotator reading p would infer that h is most likely true. Likewise, it is negative when the truth of p contradicts the truth of h, that is, when a human annotator reading p would infer that h is most likely false. The absence of textual entailment is the lack of any relationship between p and h and in this case, the human annotator reading p would infer that the truth of p neither entails nor contradicts the truth of h. Thus, the goal of the natural language inference task is to determine whether h can justifiably be inferred from p. Specifically, based on the textual entailment relationship between p and h, there are three labels: *Entailment (ENT)* for positive textual entailment, *Neutral (NEU)* for the absence of textual entailment and *Contradiction (CON)* for negative textual entailment. Three examples from *SNLI* are presented in Table 1 below:

Premise	**Hypothesis**	**Label**
A soccer game with multiple males playing.	Some men are playing a sport.	Entailment (ENT)
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	Neutral (NEU)
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.	Contradiction (CON)

Table 1: Three examples (entailment, neutral, contradiction) from the SNLI dataset.

Since the publication of the *Stanford Natural Language Inference (SNLI)* (Bowman et al., 2015) and the *Multi-Genre Natural Language Inference (MNLI)* (Williams et al., 2017) datasets, there has been a considerable progress in the field of natural language inference due to the large amount of annotated data that these datasets provided. Numerous approaches based on recurrent neural networks, such as *LSTM*-based approaches which often utilize attention mechanisms, have produced decent results (Rocktäschel et al., 2015), (Chen et al., 2016), (Sha et al., 2016), (Munkhdalai and Yu, 2017), (Ghaeini et al., 2018). More recently, pre-trained language models have managed to provide an even higher performance on many tasks related to natural language, including natural language inference (Radford et al., 2018). Specifically, well pre-trained contextual language models such as *ELMo* (Peters et al., 1802) and *BERT* (Devlin et al., 2018) or *BERT*-based approaches (Zhang et al., 2020) are among those which achieve the highest performance for the *SNLI* and the *MNLI* datasets.

However, recent research shows that even though these pre-trained language models achieve high performance on benchmark datasets, they do so by learning spurious correlations, also called *dataset artifacts*. The models are then expected to fail in settings where these artifacts are not present, which may include real-world test sets of interest. The usage of contrast sets (Gardner et al., 2020), checklist sets (Ribeiro et al., 2020) or other adversarial sets (Jia and Liang, 2017), (Wallace et al., 2019), (Bartolo et al., 2020), (Glockner et al., 2018), (McCoy et al., 2019) makes performance plummet and thus highlights this issue.

In recent years, there has been a considerable effort in order to combat dataset artifacts in the natural language inference domain. Learning seems to be more robust when it focuses on hard subsets of data or data where the gold label distribution is ambiguous through dataset cartography (Swayamdipta et al., 2020) or other methods (Yaghoobzadeh et al., 2019), (Nie et al., 2020), (Meissner et al., 2021). Another approach is to train on sets of adversarial data such as challenge sets directly (Liu et al., 2019), (Zhou and Bansal, 2020) or adversarial sets generated by data augmentation (Ribeiro et al., 2020), (Morris et al., 2020). In our work, we propose our own novel method for creating adversarial sets through *automatic contextual data augmentation (acda)* which, when compared to the aforementioned data augmentation techniques, has the advantage of being substantially more computationally efficient and, at the same time, fully automatic as it requires no manual annotation by a human expert.

Finally, contrastive learning (Dua et al., 2021) is a learning optimization method which takes inspiration from contrastive estimation (Smith and Eisner, 2005) and extends the technique to supervised reading comprehension by carefully selecting appropriate neighbourhoods of related examples. In the original paper, it is used in the context of question answering and requires bundles of closely related question answering pairs which the authors call *instance bundles*. In our work, we show that the same technique can also be successfully used in natural language inference. In particular, our novel *acda* method displays great synergy with contrastive learning as it offers a natural way of creating multiple instance bundles of language inference examples that are both contextually closely related and of arbitrary size, which grows exponentially with the length of the hypothesis sentence. We also retain the authors' original technique of combining a Cross Entropy Loss with Maximum Likelihood Estimation (NLL Loss) through our proposed *hybrid loss*.

3. Analysis of Dataset Artifacts

The first task in solving the issue of dataset artifacts is to identify them. For this purpose, we conducted an exploratory analysis on the *SNLI* dataset and created our own set of hand-annotated adversarial examples. Note that these are examples that an original fine-tuned model classified correctly, but when the hypothesis is perturbed, even slightly, the prediction accuracy notably suffers. A subset is presented in Table 2 below:

	Premise	**Hypothesis**	**Label**	**Pred**	
1	Two women are embracing while holding to go packages.	One of the women is holding takeaway packages.	**ENT**	**CON**	✗
2	…	The packages contain food.	**ENT**	**CON**	✗
3	…	The women have bought food.	**ENT**	**CON**	✗
4	…	The women have bought lasagna.	**NEU**	**CON**	✗
5	A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.	A man is wearing black trousers.	**NEU**	**CON**	✗
6	…	His shirt features geometric designs.	**CON**	**ENT**	✗
7	A young boy in a field of flowers carrying a ball	He is carrying one ball.	**ENT**	**CON**	✗
8	…	Ball in field.	**ENT**	**CON**	✗
9	Two doctors perform surgery on patient.	The two doctors are performing brain surgery.	**NEU**	**CON**	✗
10	…	The patient is having heart surgery.	**NEU**	**CON**	✗
11	A white dog with long hair jumps to catch a red and green toy.	It is not a brown dog.	**ENT**	**CON**	✗
12	Kids are on a amusement ride.	Kids ride joyously an amusement ride.	**ENT**	**CON**	✗

Table 2: A sample of hand-annotated adversarial examples and the predictions of the highest performing fine-tuned baseline pre-trained language model (ELECTRA-Small).

By observing the 12 examples of Table 2, we can conclude that the highest performing fine-tuned baseline pre-trained language model, despite achieving a very high accuracy in the *SNLI* and the *MNLI* datasets, does not manage to classify any of our 12 hand-annotated adversarial examples correctly. This confirms the magnitude of impact dataset artifacts can have on performance. Specifically, we can make the following observations regarding dataset artifacts from the adversarial examples of Table 2.

First, the model's errors are mostly located around two particular classes, the neutral and the entailment classes. One of the potential artifacts at work here is a distance function between the premise and the hypothesis which the model learns instead of actual comprehension, and makes a prediction based on that distance artifact. Because the neutral class in particular cannot be adequately expressed by distance, or more accurately, the distance of hyponyms in embedding space can be very large and confuse the artifact's criterion, the result is that the model classifies these large distances as contradictions, which causes a substantial drop in performance.

Second, the model might perform well against trivial augmentations, such as introducing a negation in the form of adding a "not" word in the premise, but when adversarial examples use words which are further apart in embedding space from the premise words, results are much worse. Thus, the model clearly relies on learned artifacts instead of learned language comprehension. Apart from the distance function discussed above, another artifact is the set of words in the hypothesis that the model associates with a specific label regardless of context, only because it has observed those words accompanying that label multiple times during training. Recent research (Wallace et al., 2019) confirms our observation and discusses it in detail, providing examples such as "not" and "least" for the entailment class, "joyously" for the neutral class, and "nobody" and "never" for the contradiction class.

Specifically, in Table 2, we can observe, how the phrase "to go" is a synonym for words such as "takeaway" or "food", and yet the model produces an incorrect prediction for our adversarial examples 1, 2 and 3, which display a small and natural shift in language, the mere use of a synonym. The model also fails at example 4 where a more specific word is introduced, such as "lasagna", which is a hyponym of "food" and shifts the gold label to the neutral class, but the model perceives this as contradicting the premise. Furthermore, we can observe how in examples 9 and 10, even slight changes in context (specificity) cause the model to choose the contradiction class while the neutral class should have been appropriate. While this shows the effect of the distance function artifact, the most definitive example of the distance function artifact is likely to be example 6. We can observe how the model, by seeing the same phrase in both the premise and hypothesis, predicts en-

tailment and is unable to differentiate what the pattern is in reference to language. Reading comprehension would require that it can differentiate between "structure" and "shirt", and in such a case, the model would most likely make the correct prediction.

In conclusion, it is clear that while maintaining high performance on benchmark datasets remains an important indicator of performance, models should also be tested against adversarial examples, which are sometimes similar to real world sets, in order to ensure that their high performance is not a product of dataset artifacts. Thus, in our work, evaluation is carried out in regard to the prediction accuracy for both the benchmark datasets *(SNLI, MNLI)* and the hand-annotated adversarial set, a subset of which contains the 12 adversarial examples of Table 2 as they were presented and discussed above.

4. Our approach

We propose an approach which comprises three techniques towards mitigating dataset artifacts: a novel data augmentation procedure, contrastive learning, and a hybrid loss function. In what follows, we introduce our novel data augmentation technique, which we call *automatic contextual data augmentation (acda)*, and discuss its methodology as well as its benefits. We also present our learning optimization techniques of contrastive learning and hybrid loss, discuss their benefits and emphasize their synergy with *acda* in particular.

4.1. Data Augmentation

By referring to the adversarial examples as presented in Table 2, our observations naturally lead to an approach where contextual augmentation based on word groups could incentivise the model to learn the actual decision boundary instead of relying on dataset artifacts. That could happen if more hypotheses that are closely related with each other, such as our adversarial ones, were made available to the model, but which also included substantial contextual shifts, such as the ones the model fails at. In this case, there could be a benefit in performance. Moreover, we require that this augmentation procedure is fully automatic, i.e. it does not require a human expert to manually annotate each example, because otherwise the resources required would make the procedure infeasible. We devised such a data augmentation procedure that generates new examples which on one hand are non-trivial (as opposed, for example, to adding a "not" ahead of the hypothesis), while at the same time being robust in labelling the newly generated example correctly. To achieve non-trivial augmentation we employed *WordNet* (Miller, 1995) synsets and generated a new hypothesis, while leaving the premise as it is. This was done by replacing one word in the hypothesis with either a synonym, an antonym, a hyponym or a hypernym. In order to ensure that the labelling of the new example is sensible, we created and employed the set of rules shown in Table 3 below:

Old Label	Word Swap	New Label
ENT	Synonymn-Hypernym	**ENT**
ENT	Antonym	**CON**
ENT	Hyponym	**NEU**
NEU	Synonymn-Hypernym	**NEU**
NEU	Antonym	**UNK**
NEU	Hyponym	**UNK**
CON	Synonymn-Hypernym	**CON**
CON	Antonym	**UNK**
CON	Hyponym	**CON**

Table 3: Label generation rules for augmented examples using WordNet synsets.

Our data augmentation procedure scans the hypothesis sentence for nouns, and queries WordNet synsets for a replacement word. It then swaps each one of the nouns at a time and composes new examples using the labeling generation rules in Table 3. Observe that this procedure can be seen as replicating the generation of adverserial examples that caused the model performance to deteriorate. Therefore, the procedure yields a high number of new training examples from the most problematic areas of the decision boundary, which can now be used as part of training to incentivise the model against the reliance on artifacts.

The rules that result in the Unknown (UNK) label were not used as part of the augmentation. Because of the inherent ambiguity when replacing a word in these contexts, the supervisory signal can be corrupted and lead the model to learn nonsensical rules. Importantly, we note that the remaining rules are robust, but they are not infallible: there is still the possibility, however small, that a newly generated example gets an incorrect label assigned to it. However, this was deemed acceptable, because the inherent ambiguity in labelling any hypothesis is already only partially correct, even when done by human experts, as developed in detail in recent published research which shows that, indeed, numerous examples can be found in *SNLI* and other similar datasets where human experts disagree on which label to assign to a hypothesis (Dua et al., 2021). By keeping only the more robust rules for augmentation we ensure that the probability of generating a controversial example will be similar to the one induced by human experts, and will therefore not alter the undelying manifold of the dataset that the model is trying to learn.

The resulting augmentation benefits from being both fully automatic, as it does not require manual writing of new hypothesis or label annotation, while at the same time being non-trivial. For example, we can observe that by using Rule 1 in the hypothesis "A couple is playing with a dog outside", the word "dog" might be replaced by "animal" (a hypernym), which according to the rule will retain the *Entailment* label. This is logically correct, while at the same time produces an example where the swapped word can have a vector

of significant distance in embedding space, thus incentivising the model to discover the correct relations in the corpus and move away from the distance function artifact. As another example, we can consider swapping the same word with a hyponym such as "corgie". Because the original hypothesis label is *Entailment*, according to Rule 3, the new hypothesis "A couple are playing with a corgie outside" would get assigned the *Neutral* label, which is again logically correct and a valid datapoint for training a model.

We can further observe, that the number of possible augmented examples that can be generated grows exponentially with the length of the base hypothesis. This is because any noun in the hypothesis could be swapped by any word in its synset. In order to keep the training time bounded, our implementation enforces an upper bound of 10 augmented examples per hypothesis sentence. As anticipated, this approach leads to a 10 times larger dataset and training time also increases in a linear fashion. In our implementation, we use the `map()` method of the Huggingface `Trainer` class. This has the advantage of placing the augmented examples right below the original examples, as a result keeping the related examples together. This is very beneficial for our learning optimization techniques which we will discuss in the sections that follow.

4.2. Contrastive Learning

Having acquired a 10 times larger training set through *acda*, the question of taking maximum advantage of the training examples becomes pertinent. We decided to employ the recently published technique of contrastive learning (Dua et al., 2021) to further incentivise the model to learn the nuances of the decision boundary. According to the conducted research, one technique to achieve this is for the model to see *instance bundles* during training, that is, examples that are close together and belong to a specific area of the decision boundary in the same training batch. This approach has been used in unsupervised linguistic structure prediction (Smith and Eisner, 2005) and supervised reading comprehension (Dua et al., 2021).

Since *acda* places the augmented examples right after each original one, the dataset batches provided to the model in each iteration will consist of some number of original examples and their augmentations. This way, we manage to have a dataset consisting of multiple instance bundles and therefore, we gain the maximum benefit from contrastive learning. In our implementation, we disabled dataset shuffling in our `CustomTrainer` class by overloading the `_get_train_sampler()` method in the Huggingface `Trainer` class.

4.3. Hybrid Loss

Finally, as discussed above, the contrastive learning optimization technique re-focuses training in the localities of the current batch, but there lies the danger of the model learning to overfit these localities, while not being able to correctly classify examples that it has not seen and are further apart in decision space. In this scenario, the model is really learning many small multinomial classification problems, and misses out on larger scale rules in the classification manifold. In order to mitigate this, we decided to combine both the *Cross Entropy Loss (CE)* and the *NLL Loss*, which uses the *Maximum Likelihood Estimation (MLE)* criterion. We call this new combined loss function *Hybrid Loss* and define it as follows:

$$L(o,l) = \alpha \cdot \ L_{MLE}(o,l) + (1-\alpha) \cdot \ L_{CE}(o,l) \quad (1)$$

In the supervised setting, which includes our present natural language inference application, *MLE* (through the NLL Loss) is a much stronger training signal than *CE*. This is because *CE* does not provide a learning signal for the large space of alternative premises or hypotheses that are not in the neighbourhood of the current instance bundle. On the other hand, *CE* provides a much stronger signal for a small set of closely related and potentially confusing examples. Thus, the supervisory signal involves a smaller area of the decision boundary, as it will be made up of a small number of examples and their augmentations, all of which are close in decision space, as opposed to a larger number of examples all over the decision space. However, it will also be more complex in these locacilities, demanding a more fine-grained weight updating from the model and forcing it to learn the local properties of the decision boundary.

By combining both losses in a weighted average manner, we manage to retain the advantages of both loss functions. The *Cross Entropy Loss* ensures that part of the loss signal will be directly relevant to the shortcomings of the model in the localities of the decision boundary, enabling contrastive learning, while the *NLL Loss* will incentivise generalization in areas that the model has not seen, learning rules that can only be inferred by looking at unrelated examples. With this arrangement we ensure a balance between the large number of examples in a small area of the decision space, and a smaller number of examples all over that space. Intuitively, this can be thought as the *Hybrid Loss* using the NLL Loss to cause the largest modifications of the current decision boundary, affecting more of the decision space, and the Cross Entropy Loss to fine tune local areas according to the examples of each batch. In our implementation, we overloaded the `compute_loss()` method of the Huggingface `Trainer` class with our hybrid loss function as shown in Equation 1, with a value of 0.5 for the α parameter.

5. Experimental Evaluation

In this section, we experimentally evaluate our combined data augmentation and learning optimization approach on two benchmark datasets: *SNLI* and *MNLI*. Specifically, we utilize the Huggingface Transformers

Python package in order to train five models: four different BERT variants and the BERT-based ELECTRA-Small model. Then, we compare the fine-tuned baseline pre-trained language models with the respective *acda*-boosted pre-trained language models for both datasets. We select the best performing *acda*-boosted pre-trained language model and carry out an evaluation on our adversarial set in order to ensure that it successfully mitigates dataset artifacts. Finally, we present (Table 6) the outcome of our procedure on the same subset (Table 2) of our adversarial dataset that we used to demonstrate the influence of dataset artifacts.

5.1. Performance on Benchmark Datasets

We use *SNLI* and *MNLI* as our two benchmark datasets in order to present a comparison between fine-tuned baseline pre-trained language models from the Huggingface Transformers repository, and their respective *acda*-boosted pre-trained language models. Our goal is to show that our approach consistently improves performance regardless of model or dataset choice.

SNLI Dataset The first evaluation dataset is the *SNLI*, which is a collection of 570000 human-written English sentence pairs manually labeled for balanced classification (Bowman et al., 2015). We present the comparison between the fine-tuned baseline pre-trained language models and the respective *acda*-boosted pre-trained language models for *SNLI* in Table 4 below:

	Fine-tuned Baseline Model	**Acda-boosted Model**
BERT-Tiny	78.86	82.01
BERT-Mini	85.06	86.79
BERT-Small	87.27	87.90
BERT-Medium	88.92	89.01
ELECTRA-Small	89.02	89.82

Table 4: Comparison of fine-tuned baseline pre-trained language models and their respective *acda*-boosted pre-trained language models for the SNLI dataset.

Regarding the comparison results for *SNLI*, we can notice that *acda*-boosted pre-trained language models consistently outperform the respective fine-tuned baseline pre-trained language models. Specifically, we observe that models with a smaller architecture such as BERT-Tiny and BERT-Mini make the largest gains when they make use of *acda*, as their performance is increased by 3.15% and 1.73% respectively. The rest of the models display a performance increase between 0.1% and 0.8%, while the best performing model is the *acda*-boosted ELECTRA-Small with an accuracy of 89.82%. Therefore, we can conclude that our approach consistently increases performance across all models, particularly lightweight ones, for *SNLI*.

MNLI Dataset The second evaluation dataset is the *MNLI*, which is a crowd-sourced collection of 433000 sentence pairs annotated with textual entailment information. It is modeled on the *SNLI* corpus, but differs in that covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation (Williams et al., 2017). We present the comparison between the fine-tuned baseline pre-trained language models and the respective *acda*-boosted pre-trained language models for *MNLI* in Table 5 below:

	Fine-tuned Baseline Model	**Acda-boosted Model**
BERT-Tiny	65.24	69.06
BERT-Mini	72.54	75.22
BERT-Small	77.02	78.57
BERT-Medium	80.20	80.39
ELECTRA-Small	81.16	81.53

Table 5: Comparison of fine-tuned baseline pre-trained language models and their respective *acda*-boosted pre-trained language models for the MNLI dataset.

Regarding the comparison results for *MNLI*, we can notice that *acda*-boosted pre-trained language models consistently outperform the respective fine-tuned baseline pre-trained language models. Once again, we observe that models with a smaller architecture are the ones that receive the largest performance boost, even higher than the one observed for *SNLI*. Specifically, BERT-Tiny and BERT-Mini increase their performance by 3.82% and 2.68% respectively when they employ *acda*. The rest of the models display a variable performance increase between 0.19% and 1.55%, while the best performing model is the *acda*-boosted ELECTRA-Small with an accuracy of 81.53%. Therefore, we can reach the same conclusion as before, that is, our approach consistently increases performance across all models, particularly lightweight ones, for *MNLI*.

Computational Efficiency It is worth noting that we initially implemented our data augmentation rules for *acda*, as presented in Table 3, using the *TextAttack* package (Morris et al., 2020), as well as the *Checklist* package (Ribeiro et al., 2020). The result was a $\times 60$ increase in training time, while we also confirmed manually that they produced a smaller number of augmented examples in each iteration. According to the Huggingface training time estimator, this training procedure would take approximately 60 hours on Google Colab Pro for ELECTRA-Small. On the other hand, our own optimized implementation of *acda* only requires 9 hours of training for the same task, thus highlighting its computational efficiency.

5.2. Performance on Adversarial Examples

After showing that *acda*-boosted pre-trained language models provide a consistent improvement in performance for both the *SNLI* and the *MNLI* datasets when compared to the respective fine-tuned baseline pre-trained language models, we continue our evaluation by examining their behavior when facing adversarial examples. For this purpose, we make use of our hand-annotated adversarial set and specifically, the adversar-

ial examples of Table 2, which we discussed in Section 3. We can recall that the predictions of Table 2 are those of the best performing fine-tuned baseline pre-trained language model, ELECTRA-Small. Despite having a prediction accuracy of 81.16% and 88.92% for the *SNLI* and *MNLI* validation sets respectively, the model did not classify any of the 12 adversarial examples of Table 2 correctly. We present the same 12 adversarial examples with the predictions of the *acda*-boosted ELECTRA-Small in Table 6 below:

	Premise	**Hypothesis**	**Label**	**Pred**	
1	Two women are embracing while holding to go packages.	One of the women is holding takeaway packages.	**ENT**	**ENT**	✓
2	...	The packages contain food.	**ENT**	**ENT**	✓
3	...	The women have bought food.	**ENT**	**ENT**	✓
4	...	The women have bought lasagna.	**NEU**	**ENT**	✗
5	A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.	A man is wearing black trousers.	**NEU**	**NEU**	✓
6	...	His shirt features geometric designs.	**CON**	**ENT**	✗
7	A young boy in a field of flowers carrying a ball	He is carrying one ball.	**ENT**	**ENT**	✓
8	...	Ball in field.	**ENT**	**ENT**	✓
9	Two doctors perform surgery on patient.	The two doctors are performing brain surgery.	**ENT**	**NEU**	✗
10	...	The patient is having heart surgery.	**ENT**	**NEU**	✗
11	A white dog with long hair jumps to catch a red and green toy.	It is not a brown dog.	**ENT**	**ENT**	✓
12	Kids are on a amusement ride.	Kids ride joyously an amusement ride.	**ENT**	**ENT**	✓

Table 6: A sample of hand-annotated adversarial examples and the predictions of the highest performing *acda*-boosted pre-trained language model.

Comparing the fine-tuned baseline pre-trained language model results (Table 2) and the *acda*-boosted pre-trained language model results (Table 6), we notice a considerable improvement in prediction accuracy, and we can therefore conclude that the *acda*-boosted pre-trained language model exhibits a robust behavior against adversarial examples due to its resilience against dataset artifacts. Specifically, it manages to classify 8 out of the 12 adversarial examples correctly. We can attribute its success to the improved training procedure having moved the model further away from dataset artifacts and into greater reading comprehension. This is further proven by the fact that even when it comes to the adversarial examples which the *acda*-boosted pre-trained language model classifies incorrectly, we can manually confirm that, in the majority of the cases, the classification probability towards the gold label is significantly higher compared to the one produced by the respective fine-tuned baseline pre-trained language model.

6. Conclusions and Future Work

In this work we proposed a novel data augmentation technique, *acda*, discussed its advantages with respect to established data augmentation packages, and described how it can be naturally combined with a learning optimization method which utilizes contrastive learning and a hybrid loss function. We showed that the employment of this combined approach by pre-trained language models can lead to a consistent increase in performance, while requiring minimal computational cost regarding training time and resources. In particular, *acda*-boosted pre-trained language models consistently outperform the respective fine-tuned baseline pre-trained language models in benchmark datasets related to natural language inference. Furthermore, the *acda*-boosted pre-trained language models are also substantially more resilient to dataset artifacts and as a result display robust behavior and high performance against adversarial examples.

As a natural next step, we intend to further improve the data augmentation process by introducing more sophisticated rules. We believe that by expanding the rules in a structured manner, we can generate more closely related examples and improve performance metrics substantially. This will likely require a formal-logical treatment of the relationships between sentences when a word is swapped in a controlled manner. Similarly, coming up with a larger number of more complex rules, such as ones based on conditionals, is also promising as this would further increase the size of the training set in a meaningful way and, given the computational efficiency of our procedure, it would come at a minimal cost, as no computational cost is added on top of the training cost. Finally, we intend to create modified variants of *acda* in order to expand our methodology to other domains of interest within natural language processing, where reading comprehension is vital.

Acknowledgements

The authors wish to express their gratitude to Prof. Greg Durrett and his staff for their guidance and contributions.

7. Bibliographical References

Bartolo, M., Roberts, A., Welbl, J., Riedel, S., and Stenetorp, P. (2020). Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (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.

Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., and Inkpen, D. (2016). Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dua, D., Dasigi, P., Singh, S., and Gardner, M. (2021). Learning with instance bundles for reading comprehension. *arXiv preprint arXiv:2104.08735*.

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al. (2020). Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*.

Ghaeini, R., Hasan, S. A., Datla, V., Liu, J., Lee, K., Qadir, A., Ling, Y., Prakash, A., Fern, X. Z., and Farri, O. (2018). Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. *arXiv preprint arXiv:1802.05577*.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Liu, N. F., Schwartz, R., and Smith, N. A. (2019). Inoculation by fine-tuning: A method for analyzing challenge datasets. *arXiv preprint arXiv:1904.02668*.

MacCartney, B. and Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Meissner, J. M., Thumwanit, N., Sugawara, S., and Aizawa, A. (2021). Embracing ambiguity: Shifting the training target of nli models. *arXiv preprint arXiv:2106.03020*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., and Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

Munkhdalai, T. and Yu, H. (2017). Neural tree indexers for text understanding. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 11. NIH Public Access.

Nie, Y., Zhou, X., and Bansal, M. (2020). What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (1802). Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Sha, L., Chang, B., Sui, Z., and Li, S. (2016). Reading and thinking: Re-read lstm unit for textual entailment recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2870–2879.

Smith, N. A. and Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. (2020). Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019). Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yaghoobzadeh, Y., Mehri, S., Tachet, R., Hazen, T. J., and Sordoni, A. (2019). Increasing robustness to spurious correlations using forgettable examples. *arXiv preprint arXiv:1911.03861*.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., and Zhou, X. (2020). Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Zhou, X. and Bansal, M. (2020). Towards robustifying nli models against lexical dataset biases. *arXiv preprint arXiv:2005.04732*.