Cover Letter

Title: Data Augmentation for More Robust Natural Language Reasoning

Tracking number: 11716

Name of the conference: AAAI2022

Main reasons for rejection:

1. Our description of "Mutation" maybe not clear enough that make Reviewer #1 and Review #2 confused.

2. The neural architecture used to predict the correct answer is not discussed in the last version.

Changes to address the reviewers' comments

1. We revised Figure 3 and added more description for "Mutation" in Section 2.2.

2. We added the description for the neural architecture used to predict the correct answer in Section 3.1.

3. We didn't consider "Adv" and "Syn" in this version, because they generate a new wrong choice that is semantically similar to the original wrong choice, thereby making them not suitable for testing short-circuiting. The numbers in Table 4 may vary for this change.

Reviews

Reviewer #1

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).

they propose biologically inspired operations (crossover and mutation) that can effectively be used to augment training data to teach the existing models to be more robust in their tasks. They also generate some "stress test cases" generated by different operators such as Neg, NER, PR etc. Results show that augmented models(trained on the augmented data) become more robust against both difficult cases and original test data.

2. Novelty How novel are the concepts, problems addressed, or methods introduced in the paper?

Fair: The paper contributes some new ideas.

3. Soundness Is the paper technically sound?

Good: The paper appears to be technically sound, but I have not carefully checked the details.

4. Impact How do you rate the likely impact of the paper on the AI research community?

Fair: The paper is likely to have moderate impact within a subfield of AI.

5. Clarity Is the paper well-organized and clearly written?

Good: The paper is well organized but the presentation could be improved.

6. Evaluation If applicable, are the main claims well supported by experiments?

Good: The experimental evaluation is adequate, and the results convincingly support the main claims.

7. Resources If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)

Fair: The shared resources are likely to be moderately useful to other AI researchers.

8. Reproducibility Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)

Good: key resources (e.g., proofs, code, data) are available and key details (e.g., proofs, experimental setup) are sufficiently well-described for competent researchers to confidently reproduce the main results.

9. Ethical Considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?

Good: The paper adequately addresses most, but not all, of the applicable ethical considerations.

10. Reasons to Accept Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

Experimental results show that augmented models become more robust against both difficult cases and original test data, beating back-translation, which is a recent strong baseline.

11. Reasons to Reject Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

The mutation idea in AI is not super novel, but I'm not sure whether it has been explored in NLP data augmentation field, if not, then it still has some novelty.

12. Questions for the Authors Please provide questions that you would like the authors to answer during the author feedback period. Please number them.

Can you tell me whether this idea has been explored in NLP data augmentation field? as far as I know, it has been explored in RL. Can you extend this idea to more settings like multiple choice machine reading comprehension datasets, such as RACE?

13. Detailed Feedback for the Authors Please provide other detailed, constructive, feedback to the authors. no

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category

Weak Accept: Technically solid, moderate to high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

Reviewer #2

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).

This paper proposes two data augmentation methods to make the models more robust. The experiment results show the robustness improvements.

2. Novelty How novel are the concepts, problems addressed, or methods introduced in the paper?

Fair: The paper contributes some new ideas.

3. Soundness Is the paper technically sound?

Fair: The paper has minor, easily fixable, technical flaws that do not impact the validity of the main results.

4. Impact How do you rate the likely impact of the paper on the AI research community?

Good: The paper is likely to have high impact within a subfield of AI OR moderate impact across more than one subfield of AI.

5. Clarity Is the paper well-organized and clearly written?

Good: The paper is well organized but the presentation could be improved.

6. Evaluation If applicable, are the main claims well supported by experiments?

Good: The experimental evaluation is adequate, and the results convincingly support the main claims.

7. Resources If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)

Fair: The shared resources are likely to be moderately useful to other AI researchers.

8. Reproducibility Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)

Fair: key resources (e.g., proofs, code, data) are unavailable but key details (e.g., proof sketches, experimental setup) are sufficiently well-described for an expert to confidently reproduce the main results.

9. Ethical Considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?

Not Applicable: The paper does not have any ethical considerations to address.

10. Reasons to Accept Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

> 1. Two data augmentation methods (Crossover and Mutation) are proposed in this paper. The methods are simple. However, the methods work well according to their experiment results.
>
> 2. Clear analysis has been done. Attention maps are shown.

11. Reasons to Reject Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

> 1. The second proposed data augmentation method ”Mutation” seems a little unconvincing. Are the meanings the same after the swaping two consecutive words in the right or wrong choice? It seems unclear to me. In Figure 3, the wrong is the swaped right choice. It makes more sense.
>
> 2. Stress Test Cases are used for evaluating the data augmentation method. However, the technique can also be used for data augmentation on the training data. This could be more efficient. This lead a question why still proposed the two techniques.

12. Questions for the Authors Please provide questions that you would like the authors to answer during the author feedback period. Please number them.

> 1. See Below.
>
> 2. In this papaer, four different datasets are used. However, the dataset size is relatively small. RELOR is the largest dataset. However, the proposed method does not work better than baseline and back-translation. Could you explain this?

13. Detailed Feedback for the Authors Please provide other detailed, constructive, feedback to the authors.

> 1. Figure 3 seems unclear. In the figure, the wrong choice is the swaped right choice. However, the paragraph description under the figure is different.

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category

Weak Accept: Technically solid, moderate to high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

Reviewer #3

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).

This paper discusses learning models for answering Multiple Choices Questions. Based on the assumption that models may not learn the underlying logical connections between the question and the answers, the authors investigate solutions to mitigate this issue. They illustrate this with the very limited connections that a language model makes between each possible answer and the corresponding premise in the MCQ. The proposition is with the training phase, for which a data augmentation scheme is proposed. Especially, two genetically inspired operators are proposed, which allow to generate new samples aiming at challenging the model to not rely much on spurious text features. This idea is tested on four popular MCQ datasets, as well as a stress test cases which are instances that may be harder for the models to tackle. Ultimately, the objective is to provide a more robust learning approach to solve MCQ as well as better answer logical-based questions in MCQs.

2. Novelty How novel are the concepts, problems addressed, or methods introduced in the paper?

Good: The paper makes non-trivial advances over the current state-of-the-art.

3. Soundness Is the paper technically sound?

Good: The paper appears to be technically sound, but I have not carefully checked the details.

4. Impact How do you rate the likely impact of the paper on the AI research community?

Good: The paper is likely to have high impact within a subfield of AI OR moderate impact across more than one subfield of AI.

5. Clarity Is the paper well-organized and clearly written?

Excellent: The paper is well-organized and clearly written.

6. Evaluation If applicable, are the main claims well supported by experiments?

Fair: The experimental evaluation is weak: important baselines are missing, or the results do not adequately support the main claims.

7. Resources If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)

Good: The shared resources are likely to be very useful to other AI researchers.

8. Reproducibility Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)

Good: key resources (e.g., proofs, code, data) are available and key details (e.g., proofs, experimental setup) are sufficiently well-described for competent researchers to confidently reproduce the main results.

9. Ethical Considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?

Not Applicable: The paper does not have any ethical considerations to address.

10. Reasons to Accept Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

- A very simple idea

- Evaluation is made on i.d.d. test set but also on "stress test cases" to test models' robustness

- Evaluation is not specific to a single language model – three were tested and each provide the same consistent results.

- The proposed data augmentation scheme makes the MCQ models more robust baselines as illustrated by stress tests. 11. Reasons to Reject Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

- The baseline seems stronger for i.d.d MCQs

- The evaluation average metric used for analysis is questionable

- The prediction mechanism for MCQ is not clear from the paper only (may not be self-consistent)

12. Questions for the Authors Please provide questions that you would like the authors to answer during the author feedback period. Please number them.

A. About encoding and data augmentation

A1. In Figure 1, is there an impact on the attention map after fine-tuning the language model on the MCQ corpus unsupervisedly using masked language modeling ?

A2. In MCQs, questions and answers are rather single-sentences. Why not using dedicated single-sentence representation approaches like sentence BERT and the likes ? Wouldn't a better (hopefully) encoder influence results or even attention maps ?

A3. In table 1, about the adverb operator: how is the adverb selected ? Is there a limited pool of adverbs that is being used ? Corollary: cannot this operation affect less attention if the adverb pool is small ?

B. On evaluation and results

B1 This may be popular in MCQ papers, but the neural architecture used to predict the correct answer is not discussed. How the models infer the proposed answer? This may be a trivial question, but there are several ways to do that, and the one used here is not provided making the paper rather not self-consistent.

B2. In C+M data augmentation scheme, there is an augmentation from crossover and a augmentation from mutations. Unless if I am incorrect, this approach therefore benefits from twice as much training samples than round trip back translation, crossover or mutation data augmentation, which make the comparison unfair in my opinion for that one. What happens if BT, C or M generate two times augmented data ? (back translation can be another language for another augmented sample, crossover just requires two times of MCQ questions pairs, while doubling for mutations only imply to double the sampling of MCQ questions to augment).

B3. In Table 4, the average of the 4 datasets is not a weighted average. With the different in the number of testing samples in each datasets (Table 2), it is difficult to make use of the last two column of the table although those two columns lead the analysis. For instance, if the average is weighted as I think it makes sense if we want an average performance number, then the baseline BT+B is the best in average. Or do not compute the average and provide an analysis given each dataset specificity.

Other remarks:

O1. The baseline seems stronger when the dataset contains more training and testing samples (eg. BT+B on RELOR), which tends to make the approach irrelevant as dataset grows. In practice, the dataset may be small (for annotation costs), so a few shots study could make sense for your approach.

O2. In the introduction, the paragraph starting at ”In contrast, the proposed (...) increase in the original test data” breaks the reading flow and since no mutation or crossover was discussed previously, it is not understandable. I add that, until that paragraph, the introduction is crystal clear to me.

O3. Please double check the list of references. Some cited papers from arxiv are published. For instance, the UDA paper was published at NeurIPS20.

13. Detailed Feedback for the Authors Please provide other detailed, constructive, feedback to the authors. I enjoyed reading the paper as the motivations are clearly illustrated and the proposed ideas are simple yet they are useful for MCQs models.

The crossover and mutation ideas are very practical and the proposition to include stress test suites to test models robustness is very interesting and should inspire other tasks to do so as well. I mostly regret that the main table of results may be biased (for the two last columns and for the BT+C+M rows) and that the paper does not detail some aspects of how the predictions are made or the impact of standard operations on their results, like unsupervised fine-tuning or using sentence-level encoders. My questions above list my concerns, and I would be happy to revise based on feedback from the authors.

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category

Weak Accept: Technically solid, moderate to high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

Reviewer #4

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).

To learn logical reasoning ability and improve the robustness of the nli model, this paper proposes a data augmentation method. They adopt various well-designed data augmentation methods such as crossover, mutation, and back-translation. Experiments show that the proposed method outperforms the method wo data augmentation.

2. Novelty How novel are the concepts, problems addressed, or methods introduced in the paper?

Fair: The paper contributes some new ideas.

3. Soundness Is the paper technically sound?

Good: The paper appears to be technically sound, but I have not carefully checked the details.

4. Impact How do you rate the likely impact of the paper on the AI research community?

Fair: The paper is likely to have moderate impact within a subfield of AI.

5. Clarity Is the paper well-organized and clearly written?

Good: The paper is well organized but the presentation could be improved.

6. Evaluation If applicable, are the main claims well supported by experiments?

Not applicable: The paper does not present an experimental evaluation (the main focus of the paper is theoretical).

7. Resources If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper’s reproducibility checklist)

Not applicable: For instance, the primary contributions of the paper are theoretical.

8. Reproducibility Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper’s reproducibility checklist.)

Fair: key resources (e.g., proofs, code, data) are unavailable but key details (e.g., proof sketches, experimental setup) are sufficiently well-described for an expert to confidently reproduce the main results.

9. Ethical Considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?

Not Applicable: The paper does not have any ethical considerations to address.

10. Reasons to Accept Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

1. It is a nice motivation to force the model to learn logical reasoning ability.

2. This paper conducts detailed experiments based on four datasets.

11. Reasons to Reject Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).

1. The main weakness is that this paper do not propose new methods. They adopt previous well-designed data augmentation methods such as crossover, mutation, and back-translation.

2. This is no clues in this paper show that the model has learned logical reasoning ability after the data augmentation.

3. It would be nice if the authors can give more detailed analysis to show this augmentation method can improve the robustness of the model.

12. Questions for the Authors Please provide questions that you would like the authors to answer during the author feedback period. Please number them.

n\a

13. Detailed Feedback for the Authors Please provide other detailed, constructive, feedback to the authors.

n\a

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category

Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility, incompletely addressed ethical considerations.

Data Augmentation for More Robust Natural Language Reasoning

Abstract

Biases in training data may lead to the fragility in neural models that makes choices in natural language reasoning questions without referring to the context or premises. To encourage models to learn more from the logical connections between premises and choices, we propose two biologically inspired operations that can generate new training data that "forces" the model to look at the premises. They can augment any type of multiple choice reasoning dataset, and can be applied to any supervised learning models. Results show that models trained with the augmented data become more robust against both stress test cases and original test data, beating the strong back-translation baseline.

1 Introduction

Multiple-choice questions (MCQs) are a widely used format in Natural Language understanding tasks. For example, causal reasoning task [Gordon *et al.*, 2012], story ending prediction task [Mostafazadeh *et al.*, 2017], argument reasoning comprehension task [Habernal *et al.*, 2017], and reading comprehension task [Yu *et al.*, 2020] are mostly in the form of MCQs which are made up of a premise and two or more choices. Below is an example question taken from the COPA [Gordon *et al.*, 2012] dataset, which tests commonsense causal reasoning.

Example 1 *An MCQ from COPA:*

Premise: *The man hurt his back.*
Choice 1: *He stayed in bed for several days.* ✓
Choice 2: *He went to see a psychiatrist.* ✗

One of the primary goals of training MCQs models is generalization. Mostly, models are trained on training data and tested with the validation-test split standard paradigm. While accuracy on held-out data is a useful indicator, held-out datasets are often not comprehensive and contain the same biases as the training data [McCoy *et al.*, 2019]. Furthermore, this single aggregate statistic is difficult to figure out the robustness and the reason for choosing correctly. There has been speculation [Sharma *et al.*, 2018; Srinivasan *et al.*, 2018; Zellers *et al.*, 2018] that many models did not really "understand" the semantical and logical connection between the premise and the choices, but do well only due to spurious statistical features in the choices. It may make the models fragile.

From the speculation above, we observe some problems through testing with white-box [Vig, 2019] and black-box [Ribeiro *et al.*, 2020] methods. For white-box test, previous work uses the attention map to explain the good performance of advanced neural models on such natural language (NL) reasoning problems. We plot the attention map between the words in the full question from the final encoder layer of the model. We show such a plot of Example 1 in Figure 1. The diagram clearly shows that there's virtually no connection between the first choice and the premise (emphasized with the red box) when the model is processing the full question, while the attention between the words within the first choice remains the same when the model processes only the choices without the premise. We call the phenomenon "short circuit" in natural language reasoning in this paper.

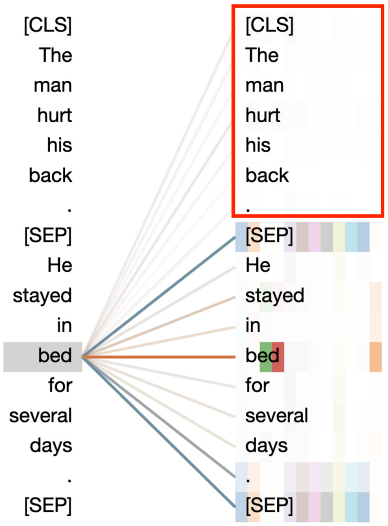


Figure 1: Attention map showing that BERT short-circuits on a COPA question.

Due to the limited interpretability of the model, there are two kinds of black-box tests. One is called "ending-only tests" in some literature [Sharma *et al.*, 2018; Bras *et al.*, 2020], which we refer to as "choice-only test" here since our focus is on multiple-choice questions. For example, BERT, when fine-tuned on the COPA data, can answer the question in Example 1 correctly. When we remove the premise from the same question and feed it to the same BERT model, it still gets the correct answer (Choice 1). This result from the "choice-only" test seems to suggest that the model can make the correct prediction without even looking at the premise of the question. The other is stress test [Jurafsky *et al.*, 2020], which guides users in what to test, by providing a list of linguistic capabilities, like taxonomy and negation. It breaks down potential capability failures into specific behaviors. We generate many stress test cases for MCQs and observed that many models are fragile. Through testing, we basically confirm that the model only pays attention to the characteristics of choices instead of the relationship between the premise and choices.

One straightforward way to improve model robustness is to generate more training examples by the types of stress test that the model is struggling with. However, many of the stress tests come with certain constraints on the choice construction, which limits the number of cases that can be generated automatically, and consequently their ability to serve as a general data augmentation method. In contrast, the proposed *crossover* and *mutation* are easily scalable and applicable for generating abundant data to encourage models to pay more attention to the relationship rather than choices only.

To that end, we applied crossover, mutation and back-translation [Xie *et al.*, 2019] to augment BERT, XLNet [Yang *et al.*, 2019] and RoBERTa [Liu *et al.*, 2019c] on ROC [Mostafazadeh *et al.*, 2017], COPA, ARCT [Habernal *et al.*, 2017] and RECLOR [Yu *et al.*, 2020] and saw up to 27% increase in accuracy on the stress tests and 10% increase in the original test data.

This paper makes two main contributions:

1. We propose to use both crossover and mutation operations to augment training data that teaches the models to consider premises in questions. Our experiments confirm the validity of this approach and show substantial improvement to model robustness, not only on the stress tests but also on the original test data.
2. We experimentally explain the behavior in three fine-tuned strong NL reasoning models and the effectiveness of our augmentation method on weakening the impact of short circuits.

2 Approach

We first present the preliminaries on stress test operators that are used to create stress test cases for measuring model robustness. Then we modify some of these methods to create training data to enhance the robustness of the models.

2.1 Preliminaries

The typical approach for detecting and evaluating model fragility is to construct out-of-distribution stress test in addition to the original in-distribution test set and observe model performance on these stress test.

In this paper, we consider the operators for constructing stress test listed in Table 1. Some of the operators were mentioned in previous literature, others are proposed here (marked with *). Each operator creates stress test instance corresponding to a specific MCQ by keeping the right choice and generating a new **wrong** choice. The new wrong choice can be generated from two sources depending on the operator: a) right choice of the original MCQ; b) wrong choice of the original MCQ.

Among all operators listed in Table 1, *Neg+*, *Neg-*, *NER*, *PR*, *PI*, and *Voice* operate on the original right choice and generate a new wrong choice that is linguistically similar to the right choice. The resultant stress test cases are not only able to evaluate the general model robustness but also more likely to detect whether models are exploiting spurious features in choices alone to make predictions rather than considering the relationship between premise and choices. The remaining two operators *Adv* and *Syn* generate a new wrong choice that is semantically similar to the original wrong choice, thereby making them not suitable for testing short-circuiting.

Oper.	Description and Example
Neg+	Add negation (r→w) Input: *They called the police to come to my house.* ✓ Output: *They **didn't** called the police to come to my house.* ✗
Neg-	Remove negation (r→w) Input: *Ben **never** starts working out.* ✓ Output: *Ben starts working out.* ✗
NER	Randomly replace person names (r→w) Input: *A big wave knocked **Mary** down .* ✓ Output: *A big wave knocked **Kia** down .* ✗
PR*	Switch pronoun by gender or quantity (r→w) Input: ***She** had a great time .*✓ Output: ***He** had a great time .* ✗
PI*	Instantiate pronoun by randome person (r→w) Input: ***They** gave Tom a new latte with less ice .* ✓ Output: ***Nathanael** gave Tom a new latte with less ice .* ✗
Voice	Swap subject and object (r→w) Input: ***Kara** asked **the neighbors** not to litter in their yard .* ✓ Output: ***the neighbors** asked **Kara** not to litter in their yard .* ✗
Adv	Add adverbs for emphasis (w→w) Input: *The ocean was a calm as a bathtub .* ✗ Output: ***In fact** the ocean was a calm as a bathtub .* ✗
Syn	Replace adj/adv with synonym (w→w) Input: *Dawn felt **happy** about getting away with it .* ✗ Output: *Dawn felt **glad** about getting away with it .* ✗

Table 1: Stress test operators considered in this paper. First line in each cell describes the operation, the remaining lines in the cell give examples of how the operators work using input and output. r→w indicates the operator turns a right choice into a wrong choice, while w→w indicates the operator turns a wrong choice into another wrong choice.

2.2 Improving Model Robustness by Data Augmentation

If a model is shown to be fragile by the stress tests, its performance may decline, especially when applied to out-of-distribution test data. To make models more robust, one natural thought is to generate more data to encourage models to focus on the relation between the premise and choices. Intuitively, all the operations that can generate proxy tests can be used to create more training data. But in reality, most of these operations are not scalable and cannot generate enough data for training, e.g., *NER* requires the presence of named entities in the choices. Meanwhile, these stress test operators are fine-grained operations on specific features, and it's hard to enumerate all the possible features for training, not to mention combinations of them.

To circumvent the above issues with existing operators, we propose two genetically inspired operators, namely *crossover* and *mutation*, which serve as highly scalable and generally applicable techniques for data augmentation. These two operators are not only simple, but also not limited to fine-grained specific features.

Crossover is visually illustrated in Figure 2. It operates across two different questions. For each MCQ in the dataset that is correctly answered by the model, we substitute its original wrong choice with the right choice from another randomly sampled MCQ. The substituted choice in the new question is guaranteed to be wrong. For *crossover*, we only consider swapping the right choices between two questions rather than wrong choices. The two choices were originally true answers in their respective questions, and presumably carry spurious features if the model was short-circuiting. *Crossover* tries to balance data from a broad perspective. One choice labeled with "True" will also be labeled "False" in our generated new cases, like the green choice in A is the right choice, but wrong for the new question A'. Hence to tell if one choice is better than the other, the model is encouraged to consider the premise.

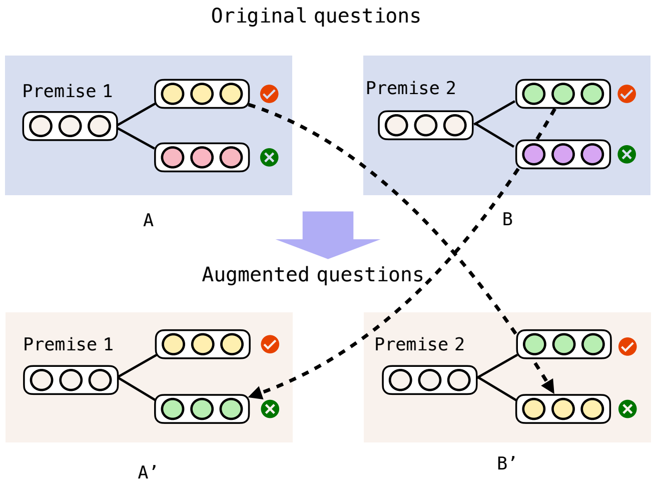


Figure 2: The Crossover Operation: the true choice of both questions are used to replace the false choices of these questions to create two new questions.

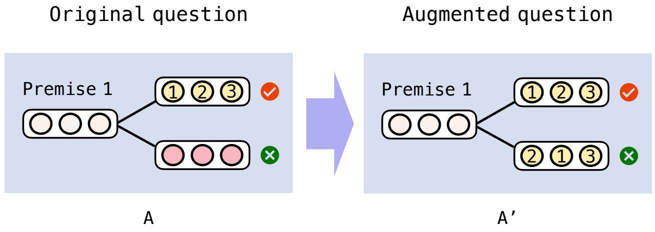


Figure 3: The Mutation Operation: the true choice of a question is used to replace the false choices of this question to create new questions.

Mutation is visually illustrated in Figure 3. *Mutation* operation swaps two consecutive words either in the right choice or wrong choice of the original MCQ, each with 50% probability. Intuitively, mutation has the potential to be effective at improving model robustness: it not only encourages the model to look into the premise due to its two very similar choices (same set of tokens), but also makes the model more sensitive to find differences in word orders and enhances the model's prior grammatical knowledge.

3 Experiments

To verify the effectiveness of *crossover* and *mutation*, We evaluate it on four popular reasoning tasks which consist of multiple-choice natural language understanding (NLU) questions. Three transformer-based models, BERT, XLNet, and RoBERTa are employed as the testbed for our experiments. In this section, we first show the experimental setup. Second, we evaluate models with different augmentation methods on original test data and stress tests. Third, we explore the reason for the fragility of raw models and the strengths of enhanced models with augmented data by choice-only test. Finally, we use a case study to further explain the reason for model improvement.

3.1 Experimental Setup

In this section, we will show our setup for datasets, models and test operators.

Datasets

We experiment on 4 datasets from four different tasks:

ROC is a story ending prediction dataset. The task is to identify the correct ending of a four-sentence story premise from two alternative choices.

COPA is a causal reasoning dataset, an example is previously shown in Section 1. Given a premise, COPA requires choosing the more plausible, causally related choice.

ARCT is an argument reasoning comprehension dataset. There may exist an alternative warrant choice in which the reason is connected to the claim.

RECLOR is a reading comprehension dataset that requires logical reasoning.

Examples and statistics of these datasets are shown in Table 2.

Models

We mainly investigate three popular classifiers based on pre-trained language models. There are several available versions of pre-trained models differing in the number of layers and parameters. We choose to use the base version of each model. We train and test all the models on a server: a GeForce GTX 1080 Ti GPU with 11G RAM and Intel(R) Xeon(R) CPU E5-2630 with 128G of RAM.

BERT (BT) is a popular attention model, which applies the bidirectional training of Transformer.

XLNet (XL) is trained with Permutation Language Modeling and without NSP.

RoBERTa (RB) is an improved pre-training procedure of BERT.

Besides the original models (marked as w/o), we also train these three models with four competing data augmentation methods: back-translation [Xie *et al.*, 2019] (B), crossover (C), mutation (M) and crossover + mutation (C+M). We start by implementing and comparing with back-translation method as our baseline which is popularly used in NLU tasks. While there exists promising data augmentation methods [Qu *et al.*, 2020; Chen *et al.*, 2021] that are based on dynamic perturbation of hidden states, back-translation is by far the most effective data augmentation method that operates on the input level. The expanded data volume for each augmentation

Dataset	Premise	Choices	Training size	Test size
COPA	I pushed the door.	The door opened. ✓ The door locked. ✗	500	500
ROC	Sarah was home alone. She wanted to stay busy. She turned on the TV. She found a reality show to watch.	Sarah then happily watched the show. ✓ Sarah could not find anything to watch. ✗	1871	1871
ARCT	**Reason**: Milk isn't a gateway drug even though most people drink it as children. **Claim**: Marijuana is not a gateway drug.	**Warrant 1**: Milk is similar to marijuana. ✓ **Warrant 2**: Milk is not marijuana. ✗	1210	444
RECLOR	**Context**:In a business...to financial prosperity. **Question**:The reasoning in the argument is flawed because the argument	A: ignores the fact that in... the family 's prosperity.✓ B: presumes, without... the family's prosperity. ✗ C: ignores the fact... even if they pay high wages. ✗ D: presumes, without providing...can succeed. ✗	4638	500

Table 2: Examples for all 4 datasets considered in this paper.

method is consistent with the original data volume. The expanded data volume is equal to the original data volume and the size of new train dataset has doubled.

Stress Test Cases

Following previous research [Jurafsky *et al.*, 2020], we will test the effectiveness of different data augmentation methods by looking at the robustness of models against different stress tests. We create these stress test cases using the operators introduced in Table 1. Different operations generate different number of cases as shown in Table 3. The standard test data accuracy and stress test accuracy are utilized for all data augmentation methods to ensure a fair comparison and thorough evaluation for both effectiveness and robustness.

Stress	ROC	COPA	ARCT	RECLOR
Neg+	1,797	492	297	375
Neg-	94	2	152	119
NER	362	0	5	0
PR	1,073	328	71	72
PI	861	219	56	91
Voice	1,014	246	174	263
Adv	1,850	496	444	500
Syn	653	25	303	289
Total	11,446	2,808	2,390	2,709

Table 3: Number of stress test cases [1]generated by different operators for the four datasets.

Generating Augmented Data

We first apply the proposed two operators *crossover* (C) and *mutation* (M) as well as their combination *crossover+mutation* (C+M) to generate additional training data. For each MCQ in the original training set, we follow the description in Section 2.2 to generate one additional MCQ using C, M, or C+M.

For back-translation, we also generate one additional MCQ for each MCQ in the original training set by conducting a round-trip English-to-French and French-to-English translation over each wrong choice. The translation model we utilized is the mBART model based on HuggingFace [Wolf *et al.*, 2020] library. The quantity of augmented data generated by each method is the same as the original training data of each dataset, hence allowing a controlled comparison.

3.2 Evaluation

In this subsection, we explore models with different data augmentation methods, back-translation, *crossover* and *mutation*, from overview and detail perspectives. Overview means we analyze the results from the original test accuracy result and overall stress test accuracy result. Detailed perspective means we will show the stress test accuracy results separately with different operators.

Overview results

Our overall comparison experiments results are shown in Table 4. The numbers in the "Original" columns denote the percentage of cases in the original test set that is correctly predicted by the models. For example, ROC has 1871 test cases (Table 2). The scores in the "Stress" columns are the percentage of stress test cases correctly predicted by the models.

In Table 4, we can find that vanilla BERT, XLNet, and RoBERTa models on different datasets are mostly not robust on stress test. Compared with testing on original test data, the accuracy rate has dropped by 12.12% for XLNet model trained with ROC. Similarly, all three models perform much worse than before on COPA (7.47%), RECLOR (18.3%) and ARCT (14.86%) datasets. It indicates that the original models are fragile and can be easily confused by small perturbations. Furthermore, there are two possible sources for model fragility, model structure and spurious features in training data. Since the model is black-box and hard to interpret, we explore the source from the data. If the model can get better performance on stress test with a data augmentation method, it suggests that the source for model fragility is from data rather than model structure. Meanwhile, it indicates that this data augmentation method is effective for enhancing model robustness.

We can also find that model performances with these data augmentation methods and the vanilla transformer-based models have achieved similar performance (±2.2) mostly

[1] The number denotes the number of questions which can be transfered to a new stress test case with a certain operation.

Model	ROC		COPA		ARCT		RELOR		Average of 4 Datasets	
	Original	**Stress**	**Original**	**Stress**	**Original**	**Stress**	**Original**	**Stress**	**Original**	**Stress**
BT(w/o)	86.58	80.53	62.00	55.53	63.96	57.19	45.60	33.24	64.53	56.62
BT+B	86.75	82.50	68.60	68.03	68.47	55.39	48.60	35.02	68.11	60.24
BT+C	87.07	83.20	72.80	78.84	68.92	63.78	47.00	47.63	**68.95**	68.04
BT+M	86.48	86.06	70.40	76.32	67.79	63.45	46.80	42.77	67.87	66.34
BT+C+M	86.75	85.67	72.40	77.27	67.57	67.38	43.60	47.92	67.58	**69.56**
XL(w/o)	90.81	78.69	61.40	53.93	75.45	60.59	56.00	37.70	70.92	57.73
XL+B	90.43	82.01	63.20	63.44	79.05	63.25	57.0	42.66	**72.42**	62.84
XL+C	89.47	86.01	67.80	74.22	74.55	65.58	54.40	50.59	71.56	69.10
XL+M	90.17	88.44	62.20	69.75	74.10	71.17	53.60	49.96	70.02	69.83
XL+C+M	90.22	90.76	67.20	81.64	77.03	73.37	54.2	51.11	72.16	**74.22**
RB(w/o)	92.73	81.23	76.40	74.56	78.83	65.84	51.00	31.82	74.59	68.16
RB+B	92.46	76.70	77.00	79.31	81.31	65.91	51.00	34.97	**75.44**	64.22
RB+C	91.18	88.59	79.00	82.52	77.93	65.31	50.40	51.00	74.63	71.86
RB+M	92.62	88.47	72.60	82.85	77.03	74.90	52.00	51.02	73.56	74.31
RB+C+M	91.88	90.76	74.00	86.73	75.00	73.17	48.40	50.04	72.32	**75.16**

Table 4: Overview test on 4 models with or without(w/o) data augmentation. All numbers are percentages (%). +B = augmented with back-translation, +C = augmented with crossover, +M = augmented with mutation.

from the average original test column, demonstrating that leveraging diverse changes to choices won't harm the effectiveness of models in most cases. Consistent with other people's research [Chen *et al.*, 2021], back-translation can help to improve the accuracy of the model on the original tests slightly.

Besides, among all the augmentation strategies, we observe that back-translation only takes slightly improvement (no more than 5%) on stress test for all models in 4 tasks. Moreover, great gains can be obtained by integrating *crossover* (+C), *mutation* (+M) and their combination (+C+M) from the aggregate results in the "Stress" columns. Especially, the augmented models with combination (+C+M) method surpass original models greatly. The stress test result for XLNet on COPA has 27.71% improvement. The performance gap between the original test and stress test becomes smallar. *Crossover* and *mutation* also perform consistently better than back-translation. It indicates that these two methods are effective for improving the robustness and generalization of the models. They can further complement each other.

Detailed Results

We have described the performance of models with different augmentation methods in the aggregated stress test above. Moreover, we explore models with detailed stress tests. Concretely, different aspects of stress test data are utilized for testing. The corresponding results are presented in Figure 4. We observe that the vanilla model with purple and back-translation with green exhibit worse across different aspects than other lines (especially the vanilla model). The models trained with data augmented by *crossover* and *mutation* (the red line) are mostly more robust than others. It is consistent with our overview results in Table 4. Please refer to the Appendix for complete numbers of results.

Furthermore, in Figure 4, we can also find that the accuracy performance points for "Syn" and "Adv" are concentrated but scattered on other operator aspects. As described in Section 2.2, all stress tests (except "Syn" and "Adv) can evaluate whether a model considers the premise by using similar choices. Thus it indicates that our data augmentation method improves the model robustness by encouraging models to pay attention to the relation between premise and choices. To further explore model ability on avoiding short circuits, we also use "choice-only" tests and white-box attention map observation in the next sections.

3.3 Choice-only Test

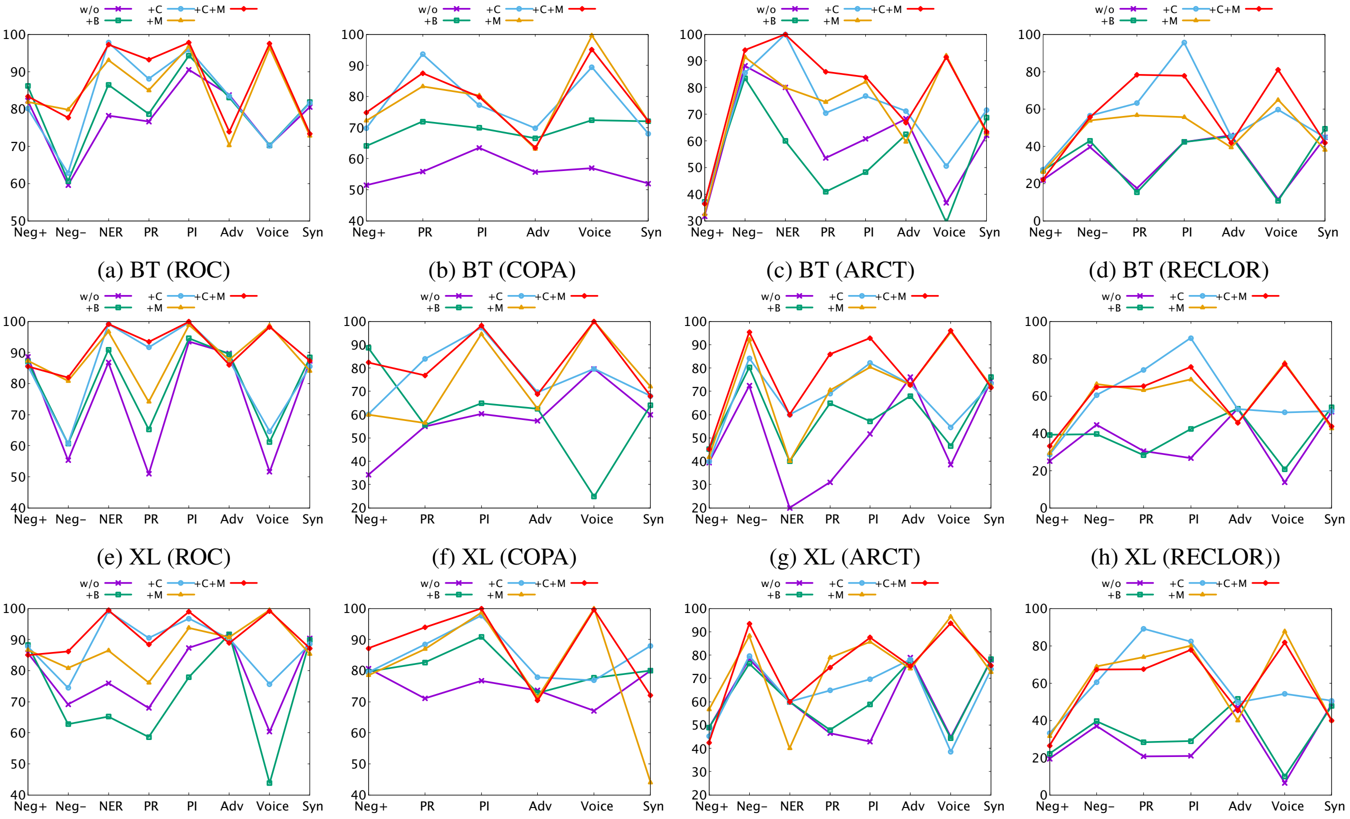
In choice-only test, we only feed choices for a model without premise which is replaced by an empty string. In this way, models have to make selection without the premise, not to mention the relationship between premise and choices. We expect the model selection results to be consistent with random selection Because if a model can easily get the "right" result which is labeled for the relationship between premise and choices in choice-only test, it may indicate this model have the opportunity to cheat in evaluation procedure and become fragile. Another possibility that is also not ruled out is that even if the model can tell the result with only choices, it still chooses to look at the premise context. Although high scores do not necessarily correspond to models not looking forward, low scores necessarily mean that models cannot conclude that solely relying on choices.

In Figure 5, we can find that the choice-only model accuracy rate has dropped with *crossover* and *mutation* augmentation methods (red line is the lowest among all). Some rates are similar to random selection, like BT+C on ROC (51.2%) and RB+C+M on ARCT (54.8%), which indicates that models are less likely to cheat anymore. In other words, models are more likely to consider premise. The results on the choice-only tests provide another perspective for us to confirm that models augmented with crossover and mutation can reduce short circuits and thus model fragility.

3.4 Case Study

To glean more insight on how our data augmentation methods help enhance model robustness on natural language reasoning, we perform case study by analyzing the change of attention patterns.

We present an example from COPA to describe the decision made on attention map. The story example is shown in Table 2. In this example, the word "pushed" in the premise has great relevance with the word "opened" in the right choice



(a) BT (ROC)

(b) BT (COPA)

(c) BT (ARCT)

(d) BT (RECLOR)

(e) XL (ROC)

(f) XL (COPA)

(g) XL (ARCT)

(h) XL (RECLOR))

(i) RB (ROC)

(j) RB (COPA)

(k) RB (ARCT)

(l) RB (RECLOR)

Figure 4: Detailed stress test with different aspects on 4 different tasks. The x-axis in the figures indicates different stress test aspects and the y-axis indicates model accuracy in percentage.

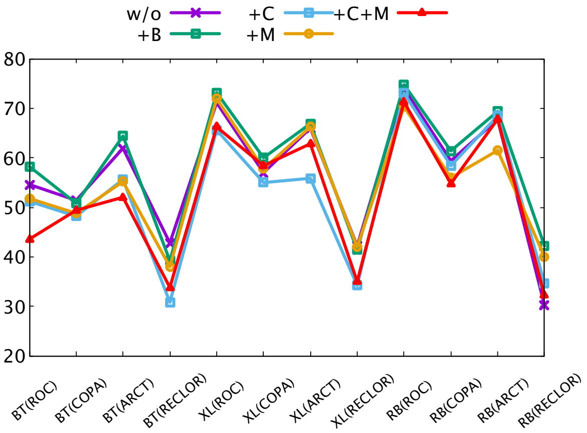


Figure 5: The choice-only results of different data augmentation methods with 3 models on 4 tasks.

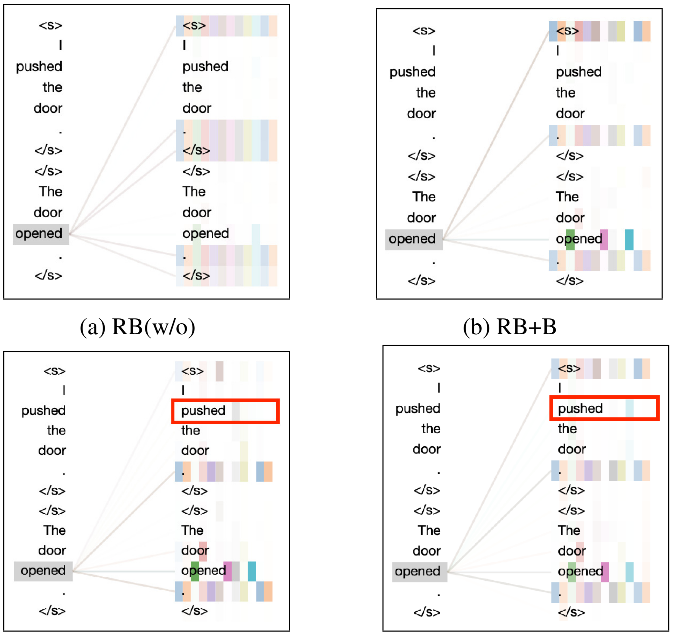
from human knowledge. The relationship between these two words is the key to answer this question. We explore different models with the augmentation method with attention map to visualize whether these two words have a relation or not. The attention map is visualized through an off-the-shelf tool [Vig, 2019].

In Figure 6, RoBERTa trained on the original training set fails to pick up the relation between “pushed” and “opened”. After training with *crossover* data augmentation, the model learns to build contextual reasoning by attending to relevant concept in the premise. Similar trends also exist for the combination of *crossover* and *mutation* operation in Figure 6d. These observations empirically demonstrate the effectiveness of our methods to encouraging the model to pay attention to premise in terms of improving model robustness. However, back-translation in Figure 6b seems not to improve model by enhance the relation between premise and choices.

4 Related Work

Our work is related to three lines of research: spurious feature analysis, data augmentation, and model probing.

Spurious Feature Analysis. Prior studies [Sharma *et al.*, 2018; Srinivasan *et al.*, 2018; Zellers *et al.*, 2018] have discovered that some NLP models are able to achieve surprisingly good accuracy on natural language understanding tasks in multiple-choice question form even without looking at the context. Such phenomenon is identified via so called “hypothesis-only” test, as referred to in some literatures. Concurrent work [Sanchez *et al.*, 2018] further showed that models sometimes bear insensitivity to certain slight but semantically significant perturbations in the hy-



(a) RB(w/o)

(b) RB+B

(c) RB+C

(d) RB+C+M

Figure 6: Attention map on a COPA example for models.

pothesis, leading to suspicions that the high hypothesis-only performance stems from statistical correlations between spurious cues in the hypothesis and the label. Such spurious cues can be categorized into lexicalized [Naik *et al.*, 2018; Sanchez *et al.*, 2018; McCoy *et al.*, 2019] and unlexicalized [Bowman *et al.*, 2015]: the former mainly contains n-gram and cross-ngram spans that are indicative of certain labels, while the latter involves word overlap, sentence length and BLUE score between the premise and the hypothesis. Instead of unearthing the specific cues in the dataset, our work directly diagnoses whether models are simply exploiting the short circuit in hypothesis alone and mitigates such reasoning behavior accordingly.

Data Augmentation. Data augmentation is a widely used technique to enhance the robustness of neural network models both in computer vision [Perez and Wang, 2017] and NLP [Belinkov and Bisk, 2018; Minervini and Riedel, 2018; Yanaka *et al.*, 2019]. In many cases, augmentation with one kind of example improves accuracy on that particular case, but does not generalize to other cases, suggesting that models overfit to the augmentation set [Iyyer *et al.*, 2018; Liu *et al.*, 2019b]. In particular, McCoy *et al.* found that augmentation with HANS examples may generalize to a different word overlap challenge set, but only for examples similar in length to HANS examples. We reduce the hypothesis-only short circuit inference behavior of NLP models via several simple yet scalable augmentation methods aiming at teaching models to reason over relations between context and hypothesis.

Model Probing. Ever since the emergence of large pre-trained language models, much works have focused on the analysis of their inner workings. As a result, a considerable amount of linguistic properties are shown to be encoded in the contextualized representations and attention heads [Goldberg, 2019; Clark *et al.*, 2019; Liu *et al.*, 2019a; Tenney *et al.*, 2019]. In contrast, we are concerned with model's higher level reasoning capability, in particular the short circuiting behavior, as reflected in downstream performance through diagnostic stress test.

5 Conclusion

We observe that models can select correctly without premise and pay little attention to premise on attention map. Inspired by a speculation that models can short circuit the premises on MCQs and become fragile, we propose two data augmentation methods *crossover* and *mutation*. Our expriment results show that, while the proposed methods do not always improve results on the original datasets, they significantly and consistently increase the accuracy on stress test. They improve the model robustness and generalization capability. We also analyze the reason for this improvement with detailed stress test, choice-only test and case study. We conclude that our data augmentation methods can encourge models to pay more attention to the premise of questions.

References

[Belinkov and Bisk, 2018] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[Bowman *et al.*, 2015] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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 2015. Association for Computational Linguistics.

[Bras *et al.*, 2020] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR, 2020.

[Chen *et al.*, 2021] Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. HiddenCut: Simple data augmentation for natural language understanding with better generalizability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4380–4390, Online, August 2021. Association for Computational Linguistics.

[Clark *et al.*, 2019] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of* *the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[Goldberg, 2019] Yoav Goldberg. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287, 2019.

[Gordon *et al.*, 2012] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[Habernal *et al.*, 2017] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task. *CoRR*, abs/1708.01425, 2017.

[Iyyer *et al.*, 2018] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Jurafsky *et al.*, 2020] Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors. *ACL 2020*, 2020.

[Liu *et al.*, 2019a] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Liu *et al.*, 2019b] Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 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 2019. Association for Computational Linguistics.

[Liu *et al.*, 2019c] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[McCoy *et al.*, 2019] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[Minervini and Riedel, 2018] Pasquale Minervini and Sebastian Riedel. Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[Mostafazadeh *et al.*, 2017] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain, April 2017. Association for Computational Linguistics.

[Naik *et al.*, 2018] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Perez and Wang, 2017] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017.

[Qu *et al.*, 2020] Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding, 2020.

[Ribeiro *et al.*, 2020] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 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 2020. Association for Computational Linguistics.

[Sanchez *et al.*, 2018] Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Sharma *et al.*, 2018] Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Srinivasan *et al.*, 2018] Siddarth Srinivasan, Richa Arora, and Mark Riedl. A simple and effective approach to the story cloze test. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 92–96, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Tenney *et al.*, 2019] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[Vig, 2019] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[Xie *et al.*, 2019] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation. *CoRR*, abs/1904.12848, 2019.

[Yanaka *et al.*, 2019] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019.

[Yu *et al.*, 2020] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[Zellers *et al.*, 2018] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.