View Reviews

Paper ID
11716

Paper Title
Genetically Inspired Data Augmentation for Multiple-Choice Natural Language Reasoning

Track Name
Main Track

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 w/o 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.