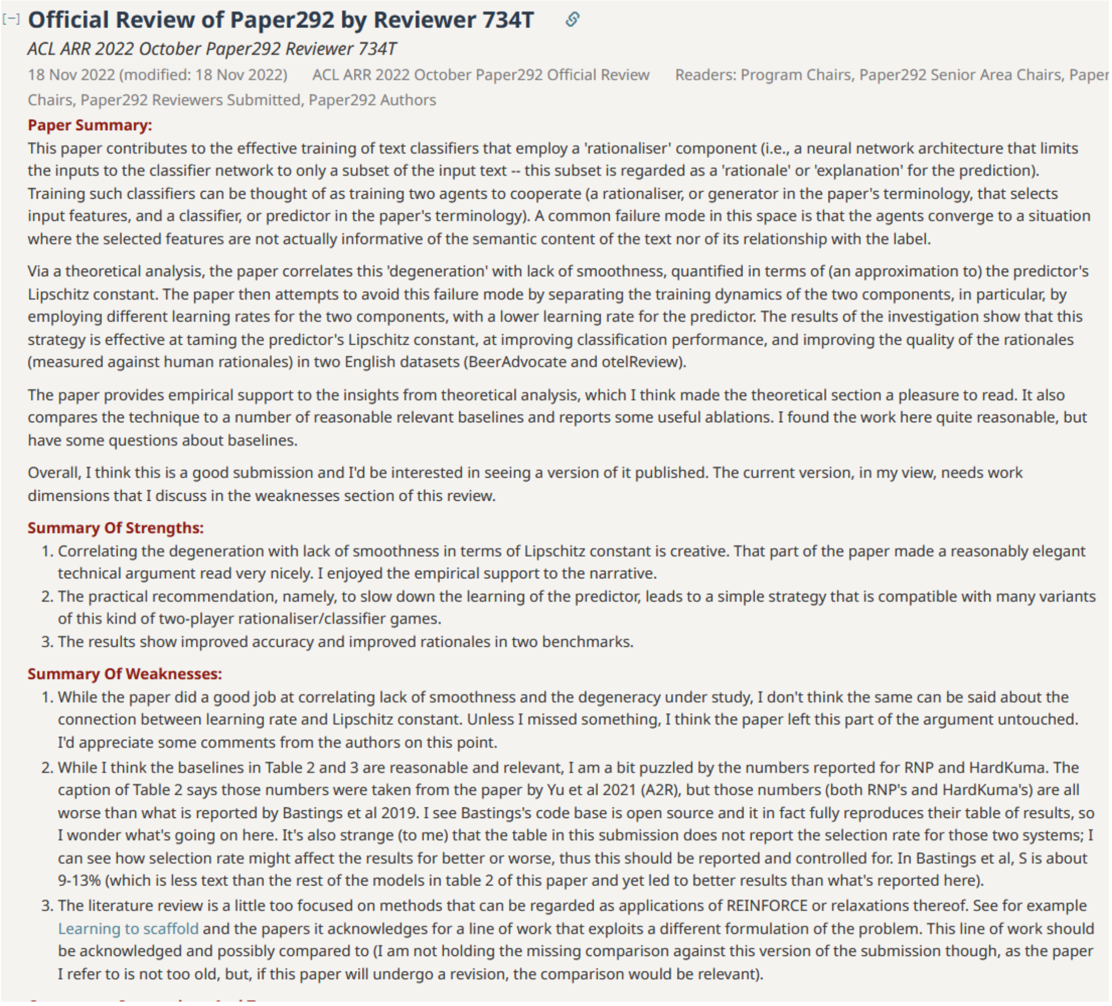


[−] **Official Review of Paper292 by Reviewer 734T**

ACL ARR 2022 October Paper292 Reviewer 734T

18 Nov 2022 (modified: 18 Nov 2022) ACL ARR 2022 October Paper292 Official Review Readers: Program Chairs, Paper292 Senior Area Chairs, Paper Chairs, Paper292 Reviewers Submitted, Paper292 Authors

Paper Summary:

This paper contributes to the effective training of text classifiers that employ a 'rationaliser' component (i.e., a neural network architecture that limits the inputs to the classifier network to only a subset of the input text -- this subset is regarded as a 'rationale' or 'explanation' for the prediction). Training such classifiers can be thought of as training two agents to cooperate (a rationaliser, or generator in the paper's terminology, that selects input features, and a classifier, or predictor in the paper's terminology). A common failure mode in this space is that the agents converge to a situation where the selected features are not actually informative of the semantic content of the text nor of its relationship with the label.

Via a theoretical analysis, the paper correlates this 'degeneration' with lack of smoothness, quantified in terms of (an approximation to) the predictor's Lipschitz constant. The paper then attempts to avoid this failure mode by separating the training dynamics of the two components, in particular, by employing different learning rates for the two components, with a lower learning rate for the predictor. The results of the investigation show that this strategy is effective at taming the predictor's Lipschitz constant, at improving classification performance, and improving the quality of the rationales (measured against human rationales) in two English datasets (BeerAdvocate and otelReview).

The paper provides empirical support to the insights from theoretical analysis, which I think made the theoretical section a pleasure to read. It also compares the technique to a number of reasonable relevant baselines and reports some useful ablations. I found the work here quite reasonable, but have some questions about baselines.

Overall, I think this is a good submission and I'd be interested in seeing a version of it published. The current version, in my view, needs work dimensions that I discuss in the weaknesses section of this review.

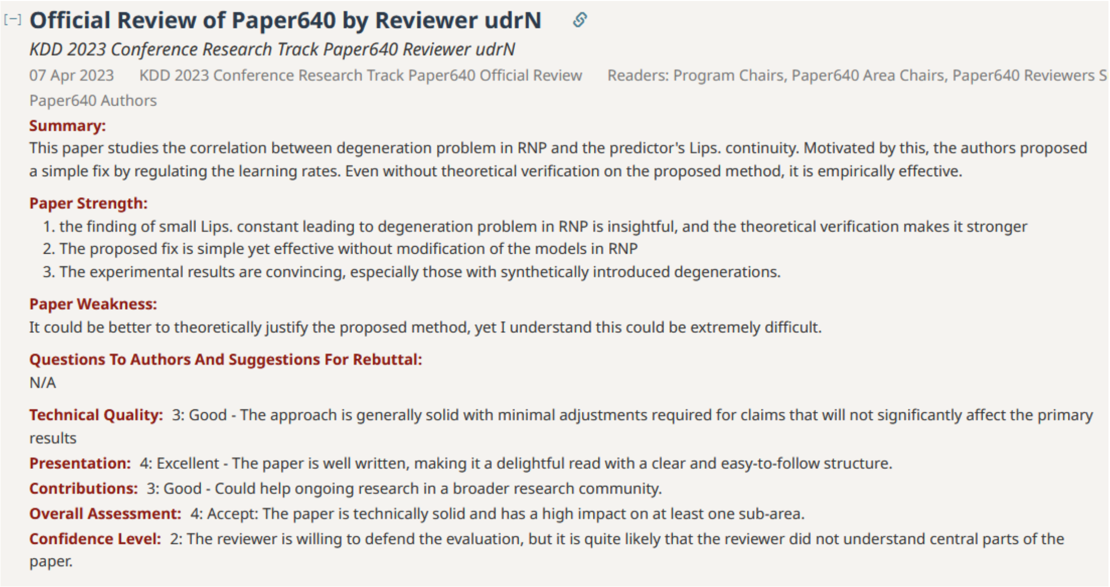
Summary Of Strengths:

1. Correlating the degeneration with lack of smoothness in terms of Lipschitz constant is creative. That part of the paper made a reasonably elegant technical argument read very nicely. I enjoyed the empirical support to the narrative.
2. The practical recommendation, namely, to slow down the learning of the predictor, leads to a simple strategy that is compatible with many variants of this kind of two-player rationaliser/classifier games.
3. The results show improved accuracy and improved rationales in two benchmarks.

Summary Of Weaknesses:

1. While the paper did a good job at correlating lack of smoothness and the degeneracy under study, I don't think the same can be said about the connection between learning rate and Lipschitz constant. Unless I missed something, I think the paper left this part of the argument untouched. I'd appreciate some comments from the authors on this point.
2. While I think the baselines in Table 2 and 3 are reasonable and relevant, I am a bit puzzled by the numbers reported for RNP and HardKuma. The caption of Table 2 says those numbers were taken from the paper by Yu et al 2021 (A2R), but those numbers (both RNP's and HardKuma's) are all worse than what is reported by Bastings et al 2019. I see Bastings's code base is open source and it in fact fully reproduces their table of results, so I wonder what's going on here. It's also strange (to me) that the table in this submission does not report the selection rate for those two systems; I can see how selection rate might affect the results for better or worse, thus this should be reported and controlled for. In Bastings et al, S is about 9-13% (which is less text than the rest of the models in table 2 of this paper and yet led to better results than what's reported here).
3. The literature review is a little too focused on methods that can be regarded as applications of REINFORCE or relaxations thereof. See for example Learning to scaffold and the papers it acknowledges for a line of work that exploits a different formulation of the problem. This line of work should be acknowledged and possibly compared to (I am not holding the missing comparison against this version of the submission though, as the paper I refer to is not too old, but, if this paper will undergo a revision, the comparison would be relevant).

We really appreciate the shortcomings raised by the anonymous reviewers regarding this paper and have corrected them in the KDD version. Specifically, we have completely rewritten the section on the relationship between Lipschitz continuity and learning rates. Words cannot express my gratitude.



Official Review of Paper640 by Reviewer udrN

KDD 2023 Conference Research Track Paper640 Reviewer udrN

07 Apr 2023 KDD 2023 Conference Research Track Paper640 Official Review Readers: Program Chairs, Paper640 Area Chairs, Paper640 Reviewers S
Paper640 Authors

Summary:
This paper studies the correlation between degeneration problem in RNP and the predictor's Lips. continuity. Motivated by this, the authors proposed a simple fix by regulating the learning rates. Even without theoretical verification on the proposed method, it is empirically effective.

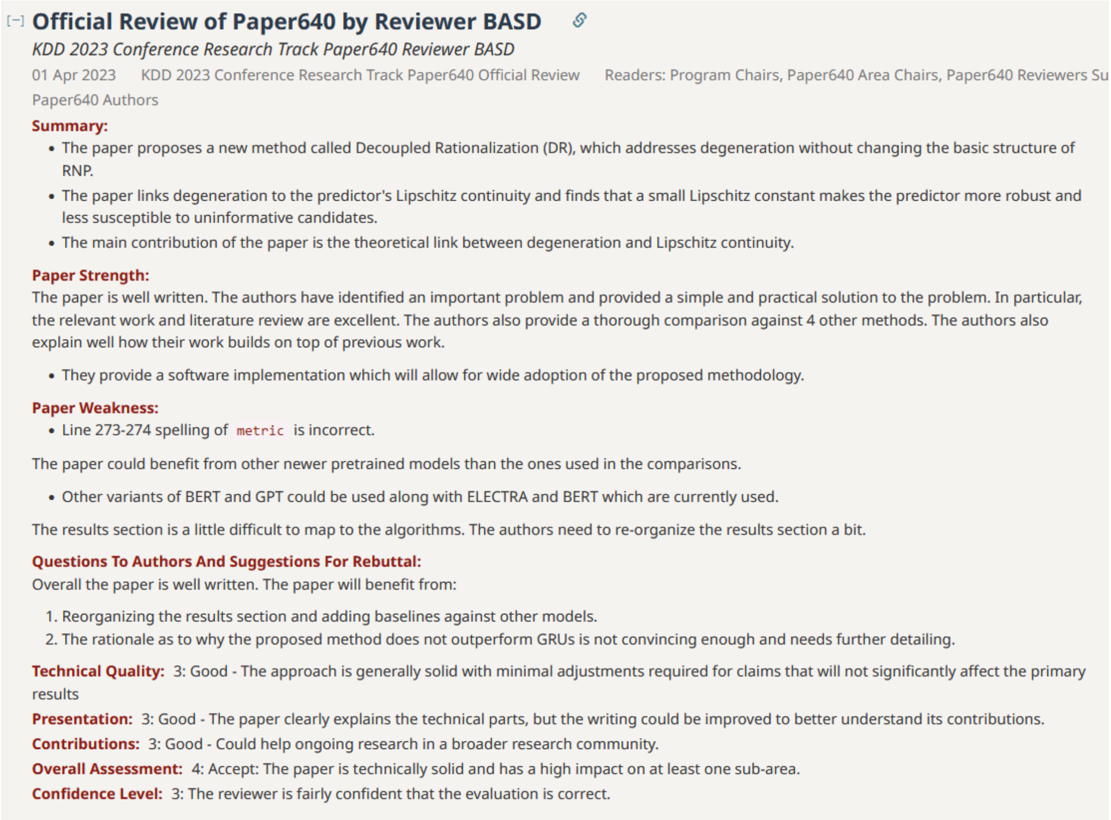
Paper Strength:

1. the finding of small Lips. constant leading to degeneration problem in RNP is insightful, and the theoretical verification makes it stronger
2. The proposed fix is simple yet effective without modification of the models in RNP
3. The experimental results are convincing, especially those with synthetically introduced degenerations.

Paper Weakness:
It could be better to theoretically justify the proposed method, yet I understand this could be extremely difficult.

Questions To Authors And Suggestions For Rebuttal:
N/A

Technical Quality: 3: Good - The approach is generally solid with minimal adjustments required for claims that will not significantly affect the primary results
Presentation: 4: Excellent - The paper is well written, making it a delightful read with a clear and easy-to-follow structure.
Contributions: 3: Good - Could help ongoing research in a broader research community.
Overall Assessment: 4: Accept: The paper is technically solid and has a high impact on at least one sub-area.
Confidence Level: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper.



Official Review of Paper640 by Reviewer BASD

KDD 2023 Conference Research Track Paper640 Reviewer BASD

01 Apr 2023 KDD 2023 Conference Research Track Paper640 Official Review Readers: Program Chairs, Paper640 Area Chairs, Paper640 Reviewers Su
Paper640 Authors

Summary:

- The paper proposes a new method called Decoupled Rationalization (DR), which addresses degeneration without changing the basic structure of RNP.
- The paper links degeneration to the predictor's Lipschitz continuity and finds that a small Lipschitz constant makes the predictor more robust and less susceptible to uninformative candidates.
- The main contribution of the paper is the theoretical link between degeneration and Lipschitz continuity.

Paper Strength:
The paper is well written. The authors have identified an important problem and provided a simple and practical solution to the problem. In particular, the relevant work and literature review are excellent. The authors also provide a thorough comparison against 4 other methods. The authors also explain well how their work builds on top of previous work.

- They provide a software implementation which will allow for wide adoption of the proposed methodology.

Paper Weakness:

- Line 273-274 spelling of `metric` is incorrect.

The paper could benefit from other newer pretrained models than the ones used in the comparisons.

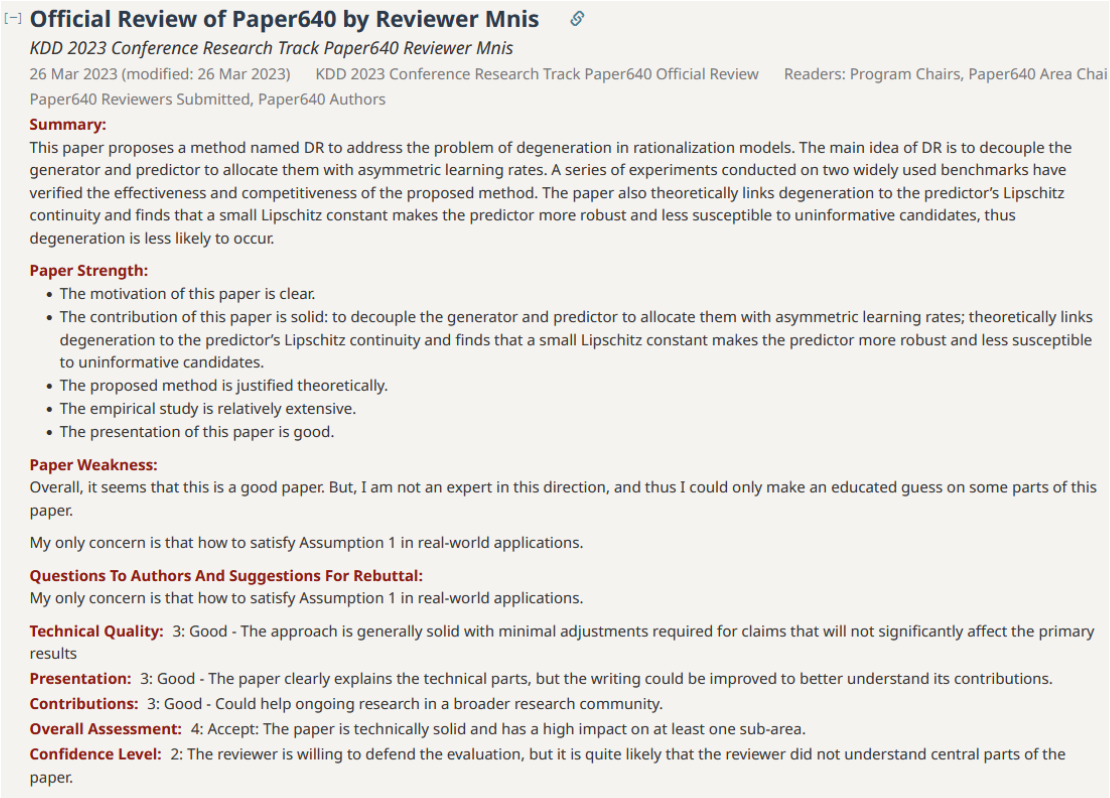
- Other variants of BERT and GPT could be used along with ELECTRA and BERT which are currently used.

The results section is a little difficult to map to the algorithms. The authors need to re-organize the results section a bit.

Questions To Authors And Suggestions For Rebuttal:
Overall the paper is well written. The paper will benefit from:

1. Reorganizing the results section and adding baselines against other models.
2. The rationale as to why the proposed method does not outperform GRUs is not convincing enough and needs further detailing.

Technical Quality: 3: Good - The approach is generally solid with minimal adjustments required for claims that will not significantly affect the primary results
Presentation: 3: Good - The paper clearly explains the technical parts, but the writing could be improved to better understand its contributions.
Contributions: 3: Good - Could help ongoing research in a broader research community.
Overall Assessment: 4: Accept: The paper is technically solid and has a high impact on at least one sub-area.
Confidence Level: 3: The reviewer is fairly confident that the evaluation is correct.



Official Review of Paper640 by Reviewer Mnis

KDD 2023 Conference Research Track Paper640 Reviewer Mnis

26 Mar 2023 (modified: 26 Mar 2023) KDD 2023 Conference Research Track Paper640 Official Review Readers: Program Chairs, Paper640 Area Chai
Paper640 Reviewers Submitted, Paper640 Authors

Summary:
This paper proposes a method named DR to address the problem of degeneration in rationalization models. The main idea of DR is to decouple the generator and predictor to allocate them with asymmetric learning rates. A series of experiments conducted on two widely used benchmarks have verified the effectiveness and competitiveness of the proposed method. The paper also theoretically links degeneration to the predictor’s Lipschitz continuity and finds that a small Lipschitz constant makes the predictor more robust and less susceptible to uninformative candidates, thus degeneration is less likely to occur.

Paper Strength:

- The motivation of this paper is clear.
- The contribution of this paper is solid: to decouple the generator and predictor to allocate them with asymmetric learning rates; theoretically links degeneration to the predictor’s Lipschitz continuity and finds that a small Lipschitz constant makes the predictor more robust and less susceptible to uninformative candidates.
- The proposed method is justified theoretically.
- The empirical study is relatively extensive.
- The presentation of this paper is good.

Paper Weakness:
Overall, it seems that this is a good paper. But, I am not an expert in this direction, and thus I could only make an educated guess on some parts of this paper.

My only concern is that how to satisfy Assumption 1 in real-world applications.

Questions To Authors And Suggestions For Rebuttal:
My only concern is that how to satisfy Assumption 1 in real-world applications.

Technical Quality: 3: Good - The approach is generally solid with minimal adjustments required for claims that will not significantly affect the primary results

Presentation: 3: Good - The paper clearly explains the technical parts, but the writing could be improved to better understand its contributions.

Contributions: 3: Good - Could help ongoing research in a broader research community.

Overall Assessment: 4: Accept: The paper is technically solid and has a high impact on at least one sub-area.

Confidence Level: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper.