View Reviews

Paper ID

2349

Paper Title

Symbolic Knowledge-Extraction Evaluation Metrics: The FiRe Score

Track Name

Main Track

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).

The authors design a metric, FiRe score, to evaluate the effectiveness of symbolic knowledge extraction. The score considers accuracy and readability trade-off. The authors provide some math analysis on the score design.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.

The authors discuss a relatively novel question in this paper: how to balance the fidelity-readability trade-off in symbolic knowledge extraction. However, I cannot see too much novelty for the score design.

The paper needs to improve readability. It would be better to immediately introduce the key concepts when mentioning them, like decompositional and pedagogical in the first page. The logic between paragraphs also takes much time to understand. Besides, some properties about FiRe score has better way to demonstrate, and it does not need 4 sub-figures (about half page).

The experiment and results are not sufficient. We can see the extraction with better readability and predictive performance leads to lowest FiRe score, but not on the contrary. Besides, the authors do not discuss the situation that both readability and predictive performance are not improved at the same time.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.

1. Why there is a "0.05" in the FiRe score formula? Could you explain all the magic numbers mentioned in this paper?

2. If readability and predictive performance are not improved at the same time, is the FiRe score still effective?

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?

Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?

Fair: The paper has minor technical flaws. For example, the proof of a theorem has some fixable errors or the experimental evaluation is weak.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?

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

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?

Poor: The paper is unclear and very hard to understand.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily

reproducible? (It may help to consult the paper's reproducibility checklist.)checklist.)
Good: key resources (e.g., proofs, code, data) are available and sufficient details (e.g., proofs, experimental setup) are described such that an expert should be able to reproduce the main results.

9. {Evaluation: 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 of some use to other AI researchers.

10. {Evaluation: 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.?
Fair: The paper addresses some applicable ethical considerations but fails to address some important ones.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Borderline reject: Technically solid paper where reasons to reject, e.g., poor novelty, outweigh reasons to accept, e.g. good quality. Please use sparingly.

13. (CONFIDENCE) How confident are you in your evaluation?
Somewhat confident, but there's a chance I missed some aspects. I did not carefully check some of the details, e.g., novelty, proof of a theorem, experimental design, or statistical validity of conclusions.

14. (EXPERTISE) How well does this paper align with your expertise?
Mostly Knowledgeable: This paper has little overlap with my current work. My past work was focused on related topics and I am knowledgeable or somewhat knowledgeable about most of the topics covered by the paper.

16. I acknowledge that I have read the author's rebuttal (if applicable) and made changes to my review as needed.
Agreement accepted

Reviewer #7

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).
The paper provides an evaluation metric for symbolic knowledge extraction named FiRe as a metric to evaluate and compare different knowledge extractors, which defines an evaluation metric for symbolic knowledge extractors. In performing the evaluation, two main components are considered: fidelity and readability of the extracted knowledge.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.
Strengths
1. This paper proposes an important study that attempts to develop a widely accepted, well-founded, and reasonable definition and measurement of the assessment criteria of interpretable AI through different metrics.
2. When constructing the evaluation function, some key evaluation criteria were considered.

Weaknesses
1. The choice of many parameters is not explained in detail in the text, which is difficult to follow this formula.
2. Only the small Iris dataset was used for the training test, rather than training on the larger dataset. Generally, more knowledge will be extracted with larger datasets, and this part needs further explanation.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
1. It is mentioned in this paper that readability is related to the form of rule representation, the readability of individual atoms that constitute knowledge, and the form of rules. But the formula uses only the number of rules as a readability indicator. Why are the other parameters discarded, and how would they affect the overall assessment?
2. As described in the formula, only the number of rules is considered a readability indicator. Does the same number

of different kinds of parameters yield the same ψ value (e.g., choosing rule 1,2, or rule 3,4) and does each rule have the same expressiveness? And how would this affect the final evaluation?
3. For different classification tasks (regression and classification tasks), different prediction losses p seem to be used, how does this affect the generic evaluation criteria?

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?
Good: The paper appears to be technically sound. The proofs, if applicable, appear to be correct, but I have not carefully checked the details. The experimental evaluation, if applicable, is adequate, and the results convincingly support the main claims.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?
Fair: The paper is likely to have modest impact within a subfield of AI.

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?
Good: The paper is well organized but the presentation has minor details that could be improved.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)checklist.)
Fair: key resources (e.g., proofs, code, data) are unavailable and/or some key details (e.g., proof sketches, experimental setup) are unavailable which make it difficult to reproduce the main results.

9. {Evaluation: 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 of some use to other AI researchers.

10. {Evaluation: 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.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Borderline reject: Technically solid paper where reasons to reject, e.g., poor novelty, outweigh reasons to accept, e.g. good quality. Please use sparingly.

13. (CONFIDENCE) How confident are you in your evaluation?
Somewhat confident, but there's a chance I missed some aspects. I did not carefully check some of the details, e.g., novelty, proof of a theorem, experimental design, or statistical validity of conclusions.

14. (EXPERTISE) How well does this paper align with your expertise?
Mostly Knowledgeable: This paper has little overlap with my current work. My past work was focused on related topics and I am knowledgeable or somewhat knowledgeable about most of the topics covered by the paper.

Reviewer #8

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).
The study of interpretability evaluation metric is an important step in XAI. Previous papers focus more on the readability of predictors. This paper proposes a novel metric combining model predictiveness and interpretability. The authors analyze some mathematical properties and conduct experiments to show the interpretability of the metric. This novel metric is parameterized and the authors discuss the hyperparmaeter tuning as well.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.
Strengths:

1. The study of interpretability evaluation metric is an important step in XAI.
2. The authors propose a novel metric with analytical properties.

Weaknesses:
1. Writing should be improved. e.g. "Existing techniques use to require tuning of hyper-parameters."
2. The proposal of this metric form is not well motivated. Why these three variables? Why multiplicative form?
3. The advantages of the metric is not well demonstrated. What could we know in terms of interpretability with this new metric compared to previous metric?

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.

It would be much helpful if the authors could provide further evidence and justifications on the weaknesses mentioned above.

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?

Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?

Fair: The paper has minor technical flaws. For example, the proof of a theorem has some fixable errors or the experimental evaluation is weak.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?

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

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?

Good: The paper is well organized but the presentation has minor details that could be improved.

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

Fair: key resources (e.g., proofs, code, data) are unavailable and/or some key details (e.g., proof sketches, experimental setup) are unavailable which make it difficult to reproduce the main results.

9. {Evaluation: 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 of some use to other AI researchers.

10. {Evaluation: 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.?

Fair: The paper addresses some applicable ethical considerations but fails to address some important ones.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.

Reject: For instance, a paper with poor quality, inadequate reproducibility, incompletely addressed ethical considerations.

13. (CONFIDENCE) How confident are you in your evaluation?

Very confident. I have checked all points of the paper carefully. I am certain I did not miss any aspects that could otherwise have impacted my evaluation.

14. (EXPERTISE) How well does this paper align with your expertise?

Very Knowledgeable: This paper significantly overlaps with my current work and I am very knowledgeable about most of the topics covered by the paper.

16. I acknowledge that I have read the author's rebuttal (if applicable) and made changes to my review as needed.

Agreement accepted

Reviewer #9

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).

The authors introduce the FiRe score metric in this paper to evaluate the effectiveness of a symbolic knowledge-extraction technique while also considering the readability of the extracted knowledge. It can be used to assist users in selecting the appropriate extraction process for a certain fidelity/readability trade-off, stated as a parameter. To put it more accurately, it is a compact score combining both a readability assessment and a predictive performance evaluation.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.

This paper is difficult to follow. It is not clearly stated the contribution and significance. The author state they proposed the FiRe which can be used to evaluate and compare SKE algorithms, but not clearly showing what is the advantages.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.

What are the questions or conclusions the experiments want to show? Is it only want to show the effectiveness of the newly proposed methods? What is the advantage of FiRe?

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?

Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?

Fair: The paper has minor technical flaws. For example, the proof of a theorem has some fixable errors or the experimental evaluation is weak.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?

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

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?

Poor: The paper is unclear and very hard to understand.

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

Fair: key resources (e.g., proofs, code, data) are unavailable and/or some key details (e.g., proof sketches, experimental setup) are unavailable which make it difficult to reproduce the main results.

9. {Evaluation: 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 of some use to other AI researchers.

10. {Evaluation: 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.?

Fair: The paper addresses some applicable ethical considerations but fails to address some important ones.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.

Reject: For instance, a paper with poor quality, inadequate reproducibility, incompletely addressed ethical considerations.

13. (CONFIDENCE) How confident are you in your evaluation?

Somewhat confident, but there's a chance I missed some aspects. I did not carefully check some of the details, e.g., novelty, proof of a theorem, experimental design, or statistical validity of conclusions.

14. (EXPERTISE) How well does this paper align with your expertise?

Knowledgeable: This paper has some overlap with my current work. My recent work was focused on closely related topics and I am knowledgeable about most of the topics covered by the paper.

16. I acknowledge that I have read the author's rebuttal (if applicable) and made changes to my review as needed.

Agreement accepted