View Reviews

Paper ID

8250

Paper Title

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

Reviewer #3

Questions

1. Summarize the contributions made in the paper with your own words

This paper proposes a self-rationalization method, RExC, which combines extractive rationalization and knowledge generation to produce natural language explanations. Experimental results show that this method achieves SOTA for a variety of natural language and vision-language tasks. A big takeaway is that the natural language explanations of self-rationalization methods can be improved by also including extractive rationalizations and vice-versa; another takeaway is that generating knowledge snippets can significantly aid explanation quality.

2. Novelty, relevance, significance

High significance, as the method clearly demonstrates significant improvements in the quality of not only natural language explanations but also extractive rationales. That the method combines existing modules -- e.g. HardKuma selection, knowledge grounding -- does not detract from the novelty, but rather makes the method more intuitive to understand.

3. Soundness

The paper is sound.

4. Quality of writing/presentation

The paper is generally well-written, but there should be more formal description of RExC -- see "Summary".

5. Literature

Literature seemed sound.

6. Basis of review (how much of the paper did you read)?

I read the full paper and appendix.

7. Summary

Update after reading the author response: The author response clarified the questions I had about the paper. I thought it was a strong paper, and now I think it is even stronger.

==============================

This paper proposes a self-rationalization method, RExC, for classification tasks in natural language and vision-language. Given an input, the model is trained to both predict the classification label and to generate a natural language explanation for its prediction. There are two components of the proposed model that distinguish it from prior work: incorporating extractive rationales into the prediction (i.e. selecting the words in the input that best explain the model's prediction before generating a natural language explanation) and including a knowledge selection module (where the extractive rationale is fed into a generative knowledge module to produce supporting facts that guide the final natural language explanation). After selecting an extractive rationale and generating knowledge snippets from the rationale, the module produces a natural language explanation and then predicts the class label. The individual components of RExC are initialized with pretrained transformer modules (such as BART/COMET).

The experimental results show that RExC consistently provides natural language explanations that are most similar to those of humans (and most interpretable by human evaluators), without sacrificing final task performance accuracy (and sometimes improving it). Additional experiments show that it's not only the natural language explanations that are similar to those of humans; the extractive rationales are more plausible than models that don't train on NLEs or use knowledge snippets, indicating the importance of incorporating these components into self-explanation models. Additional experiments show the faithfulness of the natural language explanations and extractive rationales.

This is a strong paper. The method is intuitive and the empirical applications demonstrate that RExC is an important contribution. Specifically, the fact that including a) extractive rationales and b) knowledge modules as part of NLE methods is a significant advancement. I would argue for the acceptance of this paper into ICML.

My main issue with the paper is the writing: although most of the paper is well-written, I thought the explanation of RExC was unclear. Although there is a helpful figure, each component of RExC is described in English without any math or exposition-- for example, there is no provided intuition for the HardKuma distribution (which selects rationales and knowledge snippets), and without any formal math, I cannot be certain about the specific structure they use for latent selection. I was also confused by the inputs for each module in Figure 2: it seems that after the knowledge selection step, only the knowledge snippets are passed into NLE generation and task prediction. Is this correct? Looking at the example in Figure 1a, it seems impossible to predict "[person2] is guarding [person3]" from just the background knowledge. In other words, I assume that the input is being used somewhere in these steps, but without any formal math in the descriptions we cannot assess how.

Another thing I would've liked to see is an ablation with the same model used for RExC but dropping only the ER selector. This ablation is mentioned in the experiments section ("w/o ER"), but the experiments table does not include the results. One of the interesting results for me in this paper is that the ER selection is so important for generating high-quality NLEs, but we don't actually see the ablation where everything is kept the same but only the ER is dropped (we see an ablation without knowledge selection or ER, and the results significantly degrade, but there is no ablation without only ER).

Small questions/notes:
- How large is RExC, and how does the model size compare to the other models?
- From Figure 3, it seems that the RExC NLEs are more detailed than the SOTA NLEs. Is this a general pattern for RExC? And if so, is length correlated with NLE quality?
- It's hard to assess how good the comprehensiveness and sufficiency of the ERs/knowledge snippets are in Table 4 without some kind of baseline (it could be a very simple baseline, just to give us something to compare against).

Overall, I think this is a strong paper, that could be even stronger with more details about the model. I would argue for its acceptance into ICML.

8. Miscellaneous minor issues

It looks like there's a footnote in Table 1 ("1") that is missing.

10. [R] Phase 1 recommendation. Should the paper progress to phase 2?

Yes

Reviewer #4

Questions

1. Summarize the contributions made in the paper with your own words

The paper considers the problem of producing explanations for the predictions of a learned model. Explanations of two types are considered: attributive explanations that identify the features in the data that support the prediction, and natural language explanations that explain the prediction based on commonsense knowledge. The proposed solution outperforms state-of-the-art approaches in an empirical evaluation.

2. Novelty, relevance, significance
The main novelty seems to be that both forms of explanations are supported at the same time, and are trained jointly, which presumably is also the reason for the improved empirical performance.

3. Soundness
The paper is an empirical one, and the methodology followed seems sound.

4. Quality of writing/presentation
The paper is mostly easy to follow, although somewhat dense at points.

5. Literature
The paper seems to be well-positioned with respect to relevant literature.

6. Basis of review (how much of the paper did you read)?
I have read the entire paper, but I did not consult the appendix or other sources.

7. Summary
A new architecture for producing explanations through attribution and natural language explanations that empirically outperforms SOTA.

10. [R] Phase 1 recommendation. Should the paper progress to phase 2?
Yes

Reviewer #5

Questions

1. Summarize the contributions made in the paper with your own words
This paper presents a white-box method for knowlege-grounding for sel-frationalizing models. The proposed model, RExC, answers quesions of textual entainment tasks (Natural Language Inference) and visual Q/A tasks by extracting relevant rationales (for questions) and related background knowledge from external sources. I believe that the paper combine the benefits of the end-to-end training and explainable rationale. Empricial performance demonstrate the proposed model performs better compared to existing baseline models.

2. Novelty, relevance, significance
It is an interesting work on knowledge-grounding for self-rationalizing models.
One issue is that the internal procedure in Figure 2 is not fully explainable since knowledge snippets are not decoded into natural language. Authors said that the final hidden representations (vectors) are used. If it is the case, the vector representations may not be directly (one-to-one) mapped to the natural language as shown in Figure 1. Thus, this part should be clarified.
Also, it is not clear where the empirical performance comes. That is, utilizing existing SOTA methods as internal modules could improve the performance in NLI and VQA.

the outcome looks fully explainables in extractive rationales and knowledge grounding in motivation.

3. Soundness
The proposed methods are rigorous in that the used components are state-of-the-art in NLP.

4. Quality of writing/presentation
There are some prepresentation issues in writing. However, the issues are not critical.

5. Literature

The paper includes important work in black-box knowledge-grounding models and core components that authors used.

6. Basis of review (how much of the paper did you read)?

I carefully read the method (Section 2) of the paper. However, I look through quickly the experimental parts.

7. Summary

The paper presents an interesting self-rationalzing model for knowledge-grounding. It would be good to see that the proposed methods outperform existing SOTA models in several tasks.
However, it would be good to check whether the internal vector representations are clearly explainable (for human readable self-rationalization). Also, it is not clear the sources (reasons) of the empirical performance.

After the rebuttal ---------------

Authors answered my questions well.