Supplemental materials

Contents:

On the Explainability of Convolutional Layers for Multi-Class Problems: Technical Appendix

Anonymous Authors[1]

[1]Anonymous Affiliation

Abstract

Technical appendix supplementing main paper, in particular providing pseudocode of the rule extraction algorithms and elaborating on reproducibility checklist questions where required.

This technical appendix comprises 2 parts:

- Pseudocode for ERIC rule extraction
- Elaboration on reproducibility checklist questions

Pseudocode for single-layer rule extraction with ERIC

Algorithm 1 extracts rules for each class separately using algorithm 2, adding them to a list of all rules each time.

Elaboration on reproducibility checklist question

Question 7.3 asks 'If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results'. We answered *partial* because instead of setting seeds we used multiple trials.

Algorithm 1: Tree_Extractor_Wrapper

Input: Matrix of quantised kernels $\boldsymbol{B}_{l^e}$, Matrix of output of M (one-hot encoding): $\boldsymbol{B}_{l^o}$
Parameter: Number of x-features n, pruning parameter ρ, atoms assigned to kernels in l^e: $\mathcal{L}_{l^e}$, atoms assigned to classes in l^o: $\mathcal{L}_{l^o}$
Output: A set of rules R

Let $R = \{\}$
for $\boldsymbol{b}_{l^o,j} \in \boldsymbol{B}_{l^o}$ **do**
  $R = R \cup Tree\_Extractor(\boldsymbol{B}_{l^e}, \boldsymbol{b}^s_{l^o,j}, \{\})$
end for
return R

Algorithm 2: Tree_Extractor (recursive)

Input: Matrix of quantised kernels $\boldsymbol{B}_{l^e}$, Output vector of M for class j $\boldsymbol{b}_{l^o,j}$, Literals in current path D
Parameter: Number of x-features n, pruning parameter ρ, atoms assigned to kernels in l^e: $\mathcal{L}_{l^e}$, atoms assigned to classes in l^o: $\mathcal{L}_{l^o}$
Output: A set of rules R

Let $R = \{\}$
Let $\hat{k} = \max_k(pearson\_correlation(\boldsymbol{b}_{l^e,k}, \boldsymbol{b}_{l^o,j}))$
for $s \in \{1, -1\}$ **do**
  Let $\boldsymbol{B}^s_{l^e}, \boldsymbol{b}^s_{l^o,j} = \{\boldsymbol{b}_{i,l^e}, b_{i,l^o,j} \mid b_{i,l^e,\hat{k}} = s\}$
  if $s = 1$ **then**
    $D^s = D \cup \mathcal{L}_{l^e,\hat{k}}$
  else
    $D^s = D \cup \neg\mathcal{L}_{l^e,\hat{k}}$
  end if
  if $|D^s| = n$ **or** $|\boldsymbol{B}^s_{l^e}|/|\boldsymbol{B}_{l^e}| < \rho$ **then**
    if $mode(\boldsymbol{b}^s_{l^o}) = 1$ **then**
      $C = \mathcal{L}_{l^o,j}$
    else
      $C = \neg\mathcal{L}_{l^o,j}$
    end if
    $R = R \cup \{\{D^s, C\}\}$
  else
    $R = R \cup Tree\_Extractor(\boldsymbol{B}_{l^e}, \boldsymbol{b}^s_{l^o,j}, D^s)$
  end if
end for
return R

View Reviews

Paper ID
3687

Paper Title
On the Explainability of Convolutional Layers for Multi-Class Problems

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).
Two methods, namely ERIC and SRAE, providing layer-wise explanations for the behaviour of convolutional layers of CNNs are compared, and found to provide similar results in terms of fidelity. The differences of the explanations the method provide are discussed.

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)
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).
- Well written paper, a nice read
- Analysing the explanations of layers of CNNs is very relevant for AAAI community
- Multi-class context

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).
- Novelty is limited
- Contribution, while valuable, remains a bit shallow for AAAI
- Reproducibility details are provided, but no actual code

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. Page 2, left column top: should it be "not P_3" in the rule?
2. After Eq (4), what is \rho? It does not appear in equations?
3. Do you intend to provide the code for your experimental evaluation to allow for reproducibility?

13. {Detailed Feedback for the Authors} Please provide other detailed, constructive, feedback to the authors.
In this paper ERIC and SRAE methods, providing layer-wise explanations for the behaviour of convolutional layers of CNNs are compared, and found to provide similar results in terms of fidelity. The differences of the explanations the method provide are discussed.

The paper provides an experimental study of differences of previously published algorithms. The related work is discussed adequately and the experimental evaluation seems well done. While I find the topic of the paper interesting and relevant for AAAI audience, the novelty seems a bit weak for AAAI.

It would be easier to compare accuracy vs. fidelity results (Figure 3) if the same scale was used for x- and y-axis

Style issues: use capital letters when referring to equations, figures etc. That is, "Figure 1" (not "figure 1"), for equations, use \eqref instead of \ref. It would be preferable to have equations (1)-(4) placed where they are needed (within text) and not afterwards.

Minor details:
- page 2, left column top: should it be "not P_3" in the rule
- after Eq (4), what is \rho? It does not appear in equations.

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
Borderline reject: Technically solid paper where reasons to reject, e.g., lack of novelty, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

20. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

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 examines two methods for explaining the classifications of a given layer of a neural network using logic programs. The methods are applied to the last convolutional layer of a VGG16 performing various classification tasks and evaluated in terms of accuracy and fidelity. The main contribution/novelty of this work seems to be the extension of these methods to multi-class classification problems.

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

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.?
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).
The experiments are well designed and the evaluation of the two explainers (in terms of fidelity and accuracy) seems thorough. The paper is well written. It's difficult for me to judge how novel/impactful this work is as I don't work in this particular subfield of model explanability.

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).
It's unclear to me how useful these methods would be as "explanations" for a human end user or another machine learning system, as this is not tested. The explanations do not seem like they could be easily interpreted by humans -- although the logic statements are clear, it's not clear what the "atoms" of these statements represent (they do not clearly correspond to objects, features, etc. in the image, although the authors speculate that some of them can be interpreted as objects).

12. {Questions for the Authors} Please provide questions that you would like the authors to answer during the author feedback period. Please number them.
Why are vertical flips included in the data augmentation for the Places dataset? This seems like an unnatural variation to include for scenes, which unlike objects, are generally only photographed in a canonical upright orientation.

13. {Detailed Feedback for the Authors} Please provide other detailed, constructive, feedback to the authors.
If these explainers are to be used in applications, it would be valuable to include some experiments showing that the explanations can be interpreted (by users, or whatever system is likely to use these explanations in a downstream task).

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
Borderline reject: Technically solid paper where reasons to reject, e.g., lack of novelty, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

20. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Responses to Reviewer Questions

Responses to comments by reviewers are provided where appropriate:

Reviewer 1

12. {Questions for the Authors}
1. Page 2, left column top: should it be "not P_3" in the rule? – **This is now fixed**
2. After Eq (4), what is \rho? It does not appear in equations? – **\rho was defined before the equations and is not related to them. We have reworded the relevant sentence for clarity**
3. Do you intend to provide the code for your experimental evaluation to allow for reproducibility? **Unfortunately we are unable to provide the code due to company policy on intellectual property**

13. {Detailed Feedback for the Authors}
…
It would be easier to compare accuracy vs. fidelity results (Figure 3) if the same scale was used for x- and y-axis **We agree, but space limitations made this difficult.**

Style issues: use capital letters when referring to equations, figures etc. That is, "Figure 1" (not "figure 1"), for equations, use \eqref instead of \ref. **We have adjusted the text accordingly**

It would be preferable to have equations (1)-(4) placed where they are needed (within text) and not afterwards. **Although I would normally agree, this is less efficient in terms of space and caused our paper to run over the limit**

Reviewer 2

12. {Questions for the Authors}
Why are vertical flips included in the data augmentation for the Places dataset? This seems like an unnatural variation to include for scenes, which unlike objects, are generally only photographed in a canonical upright orientation. **True, but we chose this augmentation method because it yielded higher accuracy in the test set (which was not augmented), and we wanted the original CNN to have as high accuracy as possibly in order to more readily yield higher fidelity from the extracted programs. As we argue in the paper, we generally observe higher fidelity in CNNs with higher accuracy.**

13. {Detailed Feedback for the Authors}
If these explainers are to be used in applications, it would be valuable to include some experiments showing that the explanations can be interpreted (by users, or whatever system is likely to use these explanations in a downstream task).
True, but:
1) This is beyond the scope of the paper since as we state in the opening paragraphs our focus is on the accuracy of explainable models – interpretable models must nonetheless still be accurate.
2) There is no agreed metric for interpretability in the literature.
3) Interpretability of the methods we compare were originally observed using different metrics, detailed in the original literature in both cases.