REPROT: Explaining the Predictions of Complex Deep Learning Architectures for Object Detection Through Reducts of an Image

Leonardo Concepción[1,2], Marilyn Bello[3], Gonzalo Nápoles[4], Rafael Bello[1],
Pablo Mesejo[3], and Óscar Cordón[3]

[1] Department of Computer Science, Universidad Central de Las Villas, Cuba
[2] Faculty of Business Economics, Hasselt Universiteit, Belgium
[3] Universidad de Granada, Granada, Spain
[4] Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands

1 Introduction

Deep learning models, especially in fields like object detection, have achieved remarkable accuracy, yet they are criticized for their lack of transparency. This has led to a surge in research in Explainable AI (XAI), which seeks to make models interpretable [2, 4]. XAI methods can be categorized into model-specific techniques such as Grad-CAM [16], LRP [1] and DeepTaylor [10], and model-agnostic methods like LIME [14], SHAP [9], ANCHORS [15] and RISE [12]. Despite significant progress, challenges remain in explaining deep learning predictions effectively and efficiently across different architectures. Concerning the explanation of predictions of complex architectures for object detection, in this paper we summarize our research "Explaining the Predictions of Complex Deep Learning Architectures for Object Detection Through Reducts of an Image" [3].

2 Related Work

In contrast to model-specific post-hoc explanation approaches such as LRP and DeepTaylor, the model-agnostic explanation methods can be seamlessly applied to any machine learning model regardless of their architecture and internal processing. Notice that the particularities of complex neural architectures prevent these methods from being used directly, often requiring algorithmic adjustments. Examples of these networks include Yolo [18] and Mask R-CNN [6].

Ribeiro et al. introduced LIME, a model-agnostic method that explains predictions by perturbing input data and observing changes in output [14]. Other approaches, such as SHAP, provide local explanations using game theory [9]. However, these methods often struggle with scalability and generalization across different tasks and models. Recent advances, such as ANCHORS [15], offer improvements, but their explanations can be overly complex or difficult to interpret.

Rough Set Theory (RST), originally introduced by Pawlak [11], provides a foundation for reasoning with uncertainty and partial information. RST has been applied in various domains, including image analysis [8]. Our work in [3] extends the use of RST to explain deep learning models, focusing on image object detection.

3 Proposed Method: REPROT

REPROT (Reduct-based Explainability of Predictions using Prototypes) applies RST to generate explanations for object detection models. The core idea is to identify minimal subsets of image features, known as reducts, that are sufficient for detecting objects. Given an input image and a trained deep learning model (e.g., Inception [17], YOLOv5 [7] or Mask R-CNN [6, 13]), REPROT identifies the key features that contribute to the model's decision by applying RST-based reduction techniques.

Mathematically, we define a reduct $\mathcal{R}$ as a Region of Interest (ROI [5]) with the minimal subset of superpixels such that the object is detected. An image can exhibit more than one reduct, so the concept of multi-reduct is also defined. Finally, a prototype is an image built from a reduct. From this theoretical basis and for a given object detection task, REPROT computes reducts from superpixel-based segmentation of the image. By comparing the predictions of the deep learning model with these reducts, the method identifies the minimal set of superpixels necessary for the correct classification of an object. This provides an interpretable explanation of the model's decision-making process, highlighting the critical features within the prototype image.

4 Experiments and Results

We conducted experiments on standard object detection datasets, applying REPROT to explain the outputs of Inception, YOLOv5 and Mask R-CNN. The performance of REPROT was compared against state-of-the-art explainability methods, including LIME and ANCHORS. Our experiments show that REPROT provides more concise and interpretable explanations while maintaining fidelity to the model's original predictions. Furthermore, the confidence scores reported for the detected objects show our method alignment with the model's highest-confidence predictions, reinforcing the validity of the explanations.

5 Concluding Remarks

This paper presents REPROT, a novel method for explaining deep learning predictions using RST. To support our findings in the image processing domain, extending the definitions of information systems and reducts in RST is crucial. By focusing on minimal reducts of image features, REPROT provides concise and interpretable explanations across different object detection models. Our experiments demonstrate that REPROT outperforms other post-hoc explainability methods, offering a powerful tool for understanding complex AI models.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One **10**(7), e0130140 (2015)
2. Barredo, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020)
3. Bello, M., Nápoles, G., Concepción, L., Bello, R., Mesejo, P., Óscar Cordón: Reprot: Explaining the predictions of complex deep learning architectures for object detection through reducts of an image. Information Sciences **654** (2023). https://doi.org/10.1016/j.ins.2023.119851
4. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods for black box models. arXiv preprint arXiv:2102.13076 (2021)
5. Brinkmann, R.: The art and science of digital compositing: Techniques for visual effects, animation and motion graphics. Morgan Kaufmann (2008)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2961–2969 (2017)
7. Jocher, G., Nishimura, K., Mineeva, T., Vilariño, R.: yolov5. Code repository `https://github.com/ultralytics/yolov5` (2020)
8. Li, J., Ren, Y., Mei, C., Qian, Y., Yang, X.: A comparative study of multigranulation rough sets and concept lattices via rule acquisition. Knowledge-Based Systems **91**, 152–164 (2016)
9. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the International Conference on Neural Information Processing Systems. pp. 4768—-4777. ACM (2017)
10. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
11. Pawlak, Z.: Rough sets. International Journal of Computer & Information Sciences **11**(5), 341–356 (1982)
12. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the International Conference on Neural Information Processing Systems. vol. 1, pp. 91–99. ACM (2015)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. ACM (2016)
15. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE Conference on Computer Vision. pp. 618–626. IEEE (2017)

17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826. IEEE (2016)
18. Wang, C., Bochkovskiy, A., Liao, H.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696 (2022)