Bridging Explanations and Logics: Opportunities for Multimodal Language Models

Nicolas Sebastian Schuler[1], Vincenzo Scotti[1[0000−0002−8765−604X]], Matteo Camilli[2[0000−0003−2491−5267]], and Raffaela Mirandola[1[0000−0003−3154−2438]]

[1] KASTEL,
Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131, Karlsruhe, Germany
nicolas.schuler@student.kit.edu,
vincenzo.scotti@kit.edu, raffaela.mirandola@kit.edu
[2] DEIB, Politecnico di Milano, Via Golgi 42, 20133, Milano (MI), Italy
matteo.camilli@polimi.it

Abstract. As subsymbolic Artificial Intelligence (AI) systems have become increasingly integrated into decision support tools, there is a consequent need for transparent and interpretable models. In this sense, eXplainable AI (XAI) techniques offer insights into model behavior; however, they often lack the formal rigor –typical of symbolic AI– required for causal interpretation and verification. This paper presents a framework designed to bridge the gap between subsymbolic explanations and symbolic reasoning through the application of Multimodal Language Models (MLMs). By combining the output of XAI methods with symbolic knowledge bases encoded in logic programming languages, we enable abductive reasoning to yield causal interpretations of the explanations produced over a prediction. In this instance, MLMs serve as intersymbolic translators, converting visual or textual explanations into structured logical assertions that can be processed by inference engines or subsequent verification purposes. With this approach, we aim not only to enhance the interpretability of existing AI systems, but also to promote the use of sound reasoning as a means to increase the trustworthiness of the entire AI-based system. In this paper, we outline our proposed methodology, illustrating it with a running example, and discuss both future directions (including prompt optimization and ontology generation) and challenges (like hallucinations and scalability). Our work aims at contributing to the emerging field of intersymbolic AI, which calls for the integration of symbolic and subsymbolic paradigms in the pursuit of trustworthy AI.

Keywords: XAI · MLM · Abduction.

1 Introduction

The pervasive presence of Artificial Intelligence (AI) that has emerged in the last years, given the impressive results introduced by *subsymbolic* AI techniques, has pushed for the advancement of *explainability* and eXplainable Artificial Intelligence (XAI) [8]. This area of AI deals with the production of human-understandable

justifications for model predictions. Techniques like LIME [6], SHAP [3], or Grad-CAM [7] have quickly become standards for attributing the output predictions of a model to specific input features, offering insights into what aspects of the input most influenced a decision.

Explanations in XAI are typically categorized into (i) *attributive explanations*, which highlight most likely relevant input features, and (ii) *counterfactual explanations*, which describe how minimal changes to the input could alter the output. More recent research directions include *causal and mechanistic interpretations*, which aim to extract the internal pathways and dependencies that lead to a specific prediction from a model.

When it comes to causality, given the limitations of subsymbolic AI, more rigorous and formal techniques like those rooted in *symbolic* AI represent an invaluable tool. Formal reasoning methods, particularly those grounded in logic, offer a sound framework for interpreting and validating causal claims. Among the classical forms of inference, which are (i) *deduction*, (ii) *induction*, and (iii) *abduction*, abduction may turn out to be fundamental in the context of XAI as it seeks the most plausible causes for observed effects [2, 4, 1]. Thus, abduction can be use to bridge between a model's output and a structured, causal interpretation of its explanation.

However, abductive reasoning requires a priori, explicit encoding, using a logical formalism, of all the structured knowledge to reason upon. Moreover, this information is often absent from the raw outputs provided by subsymbolic models. This introduces two key challenges: (i) translating the outputs of XAI techniques into logic-compatible formats, and (ii) incorporating relevant world knowledge to support plausible abductive reasoning, while accounting for the inherent uncertainty of data-driven models.

We propose to exploit Multimodal Language Models (MLMs) to address these challenges. These models can combine several input modalities, have a vast –yet possibly faulty– built-in knowledge of various topics, and they can generate structured outputs. These features make MLMs a suitable bridge between subsymbolic explanations and symbolic reasoning for explainability. In fact, we argue that MLMs can work as *intersymbolic translators*, parsing all the necessary inputs and generating symbolic representations compliant with the reasoning engine.

The remainder of the paper is organized as follows. Section 2 presents our envisioned methodology. Section 3 discusses ongoing and future research directions. Finally, Section 4 concludes the paper.

2 Methodology

Figure 1 illustrates a high-level schema of our approach. As shown in the schema, we use a common image classification task as a running example throughout the section. Specifically, let us consider a Neural Network (NN) trained to distinguish between images of cats and dogs. The outcome of the NN classifier for a given input image is a probability distribution from which we can identify the selected label as the most probable class (e.g., cat). To understand this decision, we may employ an explainability technique (e.g., Grad-CAM) to produce a heatmap highlighting

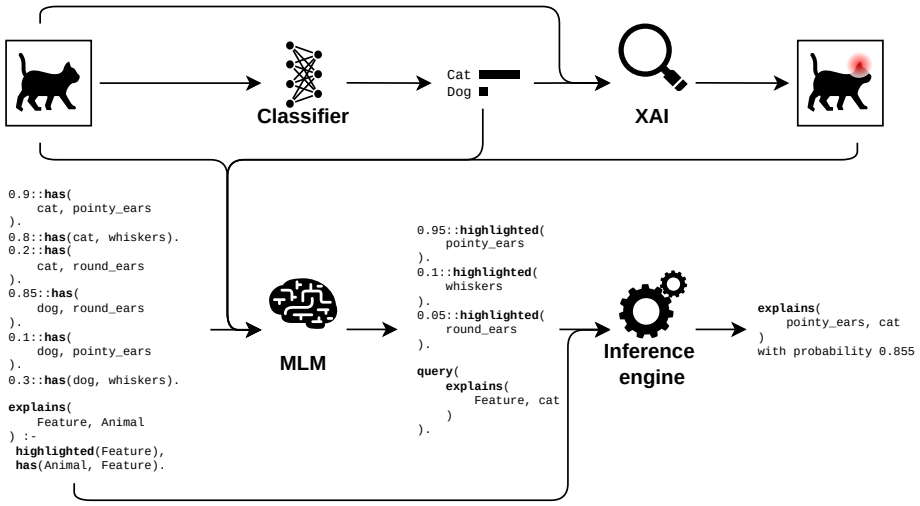

Fig. 1: Explainability abduction pipeline with probabilistic reasoning

the regions of the image that most influenced the predicted label. As an example, the heatmap could prominently highlight the area of the image that includes the animal's ears. While the attributive explanation provides a visual cue, it lacks semantic grounding, as it does not explain why the highlighted pixels (corresponding to the ears) are relevant to the classification.

Our approach addresses this gap by introducing a symbolic knowledge base, encoded in a logic programming language such as *Prolog* (or its probabilistic extension *ProbLog* [5]). This knowledge base contains ontological relationships and domain-specific facts, like the `has(cat, pointy_ears)` and `has(cat, whiskers)` expressions Figure 1.

Given the highlighted region (the ears) and the predicted label (cat), we can now apply abductive reasoning to infer the most plausible explanation: the model predicted cat because the image features pointy ears, which are characteristic of cats. This abductive step formalizes the causal interpretation of the explanation, grounding the visual attribution in symbolic knowledge.

The critical step in this process is the translation from the output of the XAI model explanation (e.g., a heatmap or a saliency map) into a symbolic representation that the reasoning engine can utilize. This is where MLMs can serve as a bridge between low-level model outputs and high-level symbolic reasoning. Indeed, MLMs can interpret visual inputs (e.g., heatmaps), extract relevant features (e.g., "ears are highlighted"), and generate structured logical assertions (e.g., `has(cat, pointy_ears)`) that are compatible with the symbolic knowledge base.

Moreover, MLMs can serve as mediators between the subsymbolic and symbolic layers by performing the following tasks: (i) mapping visual features to semantic concepts using contextual knowledge, (ii) generating logic-compatible hypotheses

for abductive reasoning, or (iii) handling uncertainty by interfacing with probabilistic logic systems like ProbLog. This MLM-based bridging mechanism enables an intersymbolic reasoning pipeline: the NN provides a prediction, the XAI model offers a localized explanation, the MLM translates this into symbolic form, and the logic engine performs abductive inference to yield a causal, interpretable explanation. Moreover, all of this can be done while preserving the probabilistic nature of the original model, enabling formal reasoning over its outputs.

3 Further developments

Our approach opens up multiple research opportunities. One immediate direction is the systematic exploration of prompting strategies and pipeline configurations. Despite their capabilities, MLM respond differently depending on how they are used on a specific task: for example, larger models may work well in an end-to-end fashion, while smaller ones may respond equally well if the task is broken down into smaller steps. These aspects directly affect the quality and structure of the generated symbolic representations, and at the same time, indirectly affect the abductive reasoning process.

Beyond prompt engineering, a more advanced evolution of this proposed framework could enable the exploitation of built-in world knowledge in MLMs to automatically construct or augment the symbolic ontology used for reasoning. In fact, instead of relying solely on manually curated knowledge bases, MLMs could be tasked with generating logical assertions such as taxonomies, part-whole relationships, or typical features of objects. For instance, starting simply from the original classification task, an MLM can generate facts like `has(cat, whiskers)` or `has(cat, pointy_ears)`, which can then be validated and integrated into the reasoning engine for subsequent verification.

While this approach can be effective in reducing the manual effort and domain expertise required to build symbolic knowledge bases, it also introduces risks. The main risks are those related to hallucinations, which are the generation of content that is neither factual nor faithful to the current context. In the context of formal reasoning, these errors can lead to unsound or misleading conclusions. Therefore, it will be crucial to investigate the reliability and limits of MLM-generated knowledge, and to develop mechanisms for validation, correction, and uncertainty quantification.

Moreover, the computational cost of this intersymbolic pipeline, especially when involving large-scale MLMs and probabilistic reasoning engines, raises practical concerns. Efficient implementations, caching strategies, and selective reasoning mechanisms will be essential to make this approach scalable and applicable to real-world scenarios.

4 Conclusion

While the combination of explainability, abduction, and MLMs offers interesting perspectives for intersymbolic AI, it also presents several open challenges. Addressing these will require interdisciplinary collaboration across machine learning, formal

methods, and knowledge representation communities. We believe that this line of research would be beneficial not only to XAI per se, but also to other disciplines, like medicine or law, which require AI models to be explainable for decision-making.

Acknowledgments. This work was supported by funding from the pilot program Core Informatics at KIT (KiKIT) of the Helmholtz Association (HGF) and supported by the German Research Foundation (DFG) - SFB 1608 - 501798263 and KASTEL Security Research Labs, Karlsruhe.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hoffman, R.R., Miller, T., Clancey, W.J.: Psychology and AI at a Crossroads: How Might Complex Systems Explain Themselves? The American Journal of Psychology **135**(4), 365–378 (Dec 2022). https://doi.org/10.5406/19398298.135.4.01
2. Josephson, J.R., Josephson, S.G.: Abductive Inference: Computation, Philosophy, Technology. Cambridge University Press, Cambridge (Oct 2009)
3. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 4765–4774 (2017), https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
4. Miller, T.: Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023. pp. 333–342. ACM (2023). https://doi.org/10.1145/3593013.3594001, https://doi.org/10.1145/3593013.3594001
5. Raedt, L.D., Kimmig, A., Toivonen, H.: Problog: A probabilistic prolog and its application in link discovery. In: Veloso, M.M. (ed.) IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. pp. 2462–2467 (2007), http://ijcai.org/Proceedings/07/Papers/396.pdf
6. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144. ACM (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
7. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128**(2), 336–359 (2020). https://doi.org/10.1007/S11263-019-01228-7, https://doi.org/10.1007/s11263-019-01228-7
8. Speith, T.: A review of taxonomies of explainable artificial intelligence (XAI) methods. In: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022. pp. 2239–2250. ACM (2022). https://doi.org/10.1145/3531146.3534639, https://doi.org/10.1145/3531146.3534639