Leveraging Neurosymbolic AI for Slice Discovery

Michele Collevati, Thomas Eiter, and Nelson Higuera
Institute of Logic and Computation
Technische Universität Wien
Favoritenstraße 9–11, 1040 Vienna, Austria
{michele.collevati, thomas.eiter, nelson.ruiz}@tuwien.ac.at

Abstract

Remarkable recent developments in deep neural networks have significantly contributed to advancing the state-of-the-art in the Computer Vision (CV) field. However, several studies have shown that CV models often make systematic errors on important subsets of data called slices, which are groups of data that share a set of attributes. It would be beneficial to solve the *slice discovery problem*, which consists of detecting semantically meaningful slices on which the model performs poorly, using the Neurosymbolic (NeSy) AI approach. The main advantage of this approach is that it exploits the strengths of both subsymbolic and symbolic AI to extract human-readable logical rules that describe underperforming slices, and to improve the explainability of CV models thanks to a modular architecture.

1 Introduction

Computer Vision is a field of AI that enables computer systems to extract valuable information from digital images and videos. Following the remarkable recent developments of deep neural networks, significant achievements have been made in advancing state-of-the-art performance in various tasks [14, 22, 10], among which it is crucial to mention safety-critical applications, such as autonomous driving [24].

However, empirical studies, e.g. [21], show that CV models struggle to generalize to new data slightly different from those on which they were initially trained and tested. A related problem is the presence of important subsets of data, called slices, for which deep learning models often make systematic errors [9]. A *slice* is defined as a group of data sharing a set of attributes. For instance, regarding the task of identifying collapsed lungs in chest X-rays, it was observed in [20] that CV models base predictions on the presence of chest drains, which is a device used in treatment. Consequently, such models often make prediction errors on crucial slices where chest drains are absent, posing a significant risk of life-threatening false negatives.

Accurately detecting underperforming slices, called *rare* slices, allows one to carefully analyze such prediction errors and subsequently improve the model. However, identifying rare slices is a complex task, especially for high-dimensional data, e.g. images, where slices are very difficult to spot and extract; furthermore, it is non-trivial to understand what makes slices rare. In view of this, the *slice discovery problem* [9] has been described as mining unstructured input data for semantically meaningful slices on which the model performs poorly. *We propose to tackle the slice discovery problem via the NeSy AI approach [11], as we believe that combining the strengths of Deep Learning (DL) and recent Knowledge Representation and Reasoning (KRR) methods will be key aspects to obtain a satisfactory solution.*

Indeed, the previous issues highlight a need for a deeper understanding of principles underlying CV models and their limitations. Addressing this problem is an urgent need for both science and industry because of the growing societal impact of such models, especially in safety-critical applications. In

The First Austrian Symposium on AI, Robotics, and Vision (AIROV24).

particular, we need to understand the following: When do these models work? When do they fail? And why?

2 Neurosymbolic AI for Slice Discovery

Several studies [20, 23, 1] have shown that machine learning models often make systematic errors on critical data slices. Consequently, recent research [9, 7, 19] has proposed automated *Slice Discovery Methods (SDMs)* to identify semantically meaningful slices in which the model exhibits prediction errors. An optimal SDM should automatically detect data slices containing coherent instances that closely correspond to a concept understandable by humans and in which the model underperforms.

The NeSy AI approach for SDM is valuable because it allows for the exploitation of the advantages of both worlds of subsymbolic and symbolic AI. For solving the slice discovery problem, we aim to leverage artificial neural networks to capture structured representations of images in the form of *scene graphs*, which were proposed in [13]. Such graphs capture detailed semantic knowledge of an image in terms of objects present and their relationships, as found e.g. by a vision component. Scene graphs play a key role in our proposed approach because they provide, augmented with possible domain knowledge, the semantic base for obtaining a semantic description of rare slices. To this end, rule extraction methods are utilized in order to characterize rare slices via a set of logical rules. The rules extracted should capture the relationships between the objects in the scene, which to the best of our knowledge is not possible with the previous SDMs proposed in the literature. Furthermore, critical advantages of logic-based rule extraction methods over other forms of machine learning include:

1. *Generalization:* Logic-based systems can generalize well from small amounts of data, making it possible to learn complex knowledge without the need for large datasets.

2. *Existing knowledge:* Logic-based systems can start from an existing knowledge base rather than learn everything from scratch.

3. *Interpretability:* Since the hypothesis is expressed using logical rules, it is compact, human-readable, and can be used to generate explanations.

In this way, we intend to provide explainability for the CV model via symbolic representations, which constitute high-level abstract knowledge well suited for reasoning.

A primary advantage of the NeSy AI approach is the modularity of its architecture, which allows one to solve different subproblems with the most appropriate tools of subsymbolic and symbolic AI in an integrated manner. In the architecture we are going to propose, this integration occurs through a loop that starts with the construction of scene graphs and then moves on to learning from them, followed by reasoning and searching for explanations at the symbolic level, and finally ending in the improvement of the CV model under examination. The main challenges that arise from our approach are as follows:

1. Derive suitable scene graphs from the images to identify slices the model performs poorly on. By "suitable" we mean getting correct information but with an appropriate level of detail, including relevant relationships. Since there can be many features in a scene, the difficulty lies in capturing those relevant to the description of rare slices.

2. Translate scene graphs into logical representations suitable to be processed by rule extraction methods. The challenge is to find a logical encoding of the scene graphs that adequately characterizes the images while remaining tractable by rule extraction systems.

3. Specify the space of all rules that a learning task may learn to efficiently obtain rules that accurately describe rare slices. Complex knowledge may be represented through rules with variables and expressive features such as negation as failure to model defaults and exceptions, choice rules to model alternatives, hard constraints to exclude scenarios, and weak (soft) constraints to model preferences. The problem, in this case, consists of defining a hypothesis space of tractable size without losing rules that potentially characterize the slices.

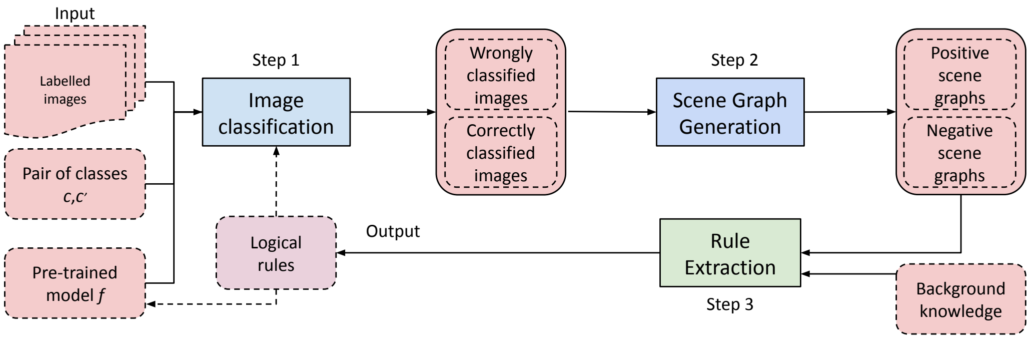

Figure 1: Overview of the proposed neurosymbolic SDM architecture.

3 Proposal for an SDM Architecture

In order to realize a neurosymbolic SDM approach, we envisage an architecture and workflow as shown in Figure 1. Consider an image classification setting, i.e. the task of assigning a class c from a set C of classes to an entire image. Let $D = \{(i_k, c_k)\}_{k=1}^n$ be a labelled dataset having n samples, where i_k is an image in the set I of images and c_k is a class in C, and $f : I \rightarrow C$ be a pre-trained model. As input, our SDM takes f, D, and a fixed pair of classes c, c' with $c \neq c'$. As output, it returns a set of logical rules describing the commonalities among the images that the model f misclassified into class c' instead of c.

Step 1: *Image classification.*
Let c be a fixed class from C. We classify each labelled image (i, c) by f and divide the images I into the following disjoint sets, where c' is from $C \setminus \{c\}$:

$$E^+_{c,c'} = \{i \in I \mid f(i) = c', (i, c) \in D\}, \text{ and}$$
$$E^-_c = \{i \in I \mid f(i) = c, (i, c) \in D\};$$

that is, we split I into the images misclassified into class c' from those correctly classified into c.

Step 2: *Scene Graph Generation.*
The *Scene Graph Generation (SGG) problem* consists of generating a scene graph from an input image. More precisely, a scene graph is a directed graph with three node types: object, attribute, and relation. Objects in an image are located by bounding boxes, and each can have several attributes, e.g. `color = blue` or `state = standing`. Furthermore, there can be different types of relations between paired objects, such as actions like `walking` or spatial relations like `behind`. Scene graphs provide a powerful semantic representation of images we intend to exploit to understand visual data and spot their misclassification patterns. To this end, SGG methods, which are mainly based on deep neural networks (see [16, 3] for an overview), are used for generating the scene graph for each image contained in the sets $E^+_{c,c'}$ and E^-_c. As a result, we obtain the sets $E^+_\mathcal{G}$ and $E^-_\mathcal{G}$ of the positive and negative scene graphs corresponding to the images that were wrongly and correctly classified by f, respectively.

Step 3: *Rule Extraction.*
We specify the *rule extraction problem* instance by translating each scene graph into a logical representation (i.e., facts and possibly further information) that will constitute with further domain knowledge the background knowledge describing the semantic information about the objects in the image. Then, the background knowledge is fed into a rule extraction system along with the positive and negative examples. Notably, the positive examples are the input images for which the model made an incorrect prediction, i.e., the images in the sets $E_{c,c'}$, as we look for an explanation of why the model fails. For rule extraction, *Inductive Logic Programming (ILP)* [5, 6] is a candidate method to obtain relevant logical rules from the background knowledge B that distinguish wrongly from correctly classified images. Given positive and negative examples and possibly further background information, an ILP system aims to find a hypothesis H, i.e. a set of rules, which entails all the positive examples and none of the negative ones. The extracted rules can be used to mitigate the

influence of rare slices on model performance, leading to a more robust model. This improvement can be achieved by augmenting the training data with additional samples.

The architecture produces logical rules as output, which capture the conditions an image must satisfy so that it is hard for the model to correctly classify it. These rules can then be used to enhance the training process of the models or in the image classification by checking the classified images against the rules to check their difficulty. Furthermore, the rules can be used to generate specific data to the rare slices. This can be easily achieved in the case of synthetic datasets, which usually are accompanied by a dataset generator which can be edited to follow the rules, or even in the case of real images, where generative models have been used in practice to generate volumes of data based in specific prompts.

Lastly, an advantage of obtaining logical rules is that they are easy to understand for humans and are inherently explainable, as these are obtained using formal methods. The rules allow us to create explanations for the neural network's behaviour and can also be used to create *contrastive explanations*. The latter are based on logical abduction and are used to know which changes in the input are needed to produce a change in the output in a desired direction. Furthermore, contrastive explanations can be readily implemented in logical programming languages [8]. This technique can then be used in combination with data generation methods to test how neural networks react to changes in the data distribution and whether we can automatically mine logical rules to mend their deficiencies.

4 Conclusion

We propose using the NeSy AI approach to address the slice discovery problem. We believe SDMs can benefit from such an approach because it allows one to exploit the potential of DL and KRR methods simultaneously. In particular, we presented an SDM architecture that leverages SGG as a subsymbolic component to obtain a structured representation of images. Scene graphs, which are a kind of knowledge graphs, play a key role here because they are the input to the subsequent symbolic module, which consists of rule extraction methods fundamental to discovering and describing rare slices through a set of logical rules. In this way, compact and human-readable logical rules can be obtained that improve the interpretability and explainability of the CV model under examination. Our modular NeSy AI approach thus results in a loop integrating different components that collaborate with each other, potentially providing causal and contrastive explanations.

We are conducting preliminary experiments using Super-CLEVR [17] as the synthetic image dataset, YOLOv5 [12] as the object detection model, and ILASP [15] as the rule extraction method. The latter is a system for inductively learning answer set programs from examples. Answer Set Programming (ASP) [2, 18] is a popular declarative problem solving method rooted in logic programming and nonmonotonic reasoning that offers high expressiveness and is supported by a suite of efficient solvers. Initial results suggest that our approach is promising. As with symbolic approaches in general, a trade-off between scalability versus expressiveness of the logical rules considered for scenes of increasing complexity will prevail, whose study and optimization in the SDM context pose challenging research tasks.

Acknowledgments

The project leading to this research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101034440.

References

[1] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digit. Medicine*, 2, 2019.

[2] Gerhard Brewka, Thomas Eiter, and Miroslaw Truszczynski. Answer set programming at a glance. *Commun. ACM*, 54(12):92–103, 2011.

[3] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A Comprehensive Survey of Scene Graphs: Generation and Application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):1–26, 2023.

[4] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Slice Finder: Automated Data Slicing for Model Validation. In *Proc. ICDE 2019*, pp. 1550–1553. IEEE, 2019.

[5] Andrew Cropper and Sebastijan Dumancic. Inductive Logic Programming At 30: A New Introduction. *J. Artif. Intell. Res.*, 74:765–850, 2022.

[6] Andrew Cropper, Sebastijan Dumancic, Richard Evans, and Stephen H. Muggleton. Inductive logic programming at 30. *Mach. Learn.*, 111(1):147–172, 2022.

[7] Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. The Spotlight: A General Method for Discovering Systematic Errors in Deep Learning Models. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1962–1981. ACM, 2022.

[8] Thomas Eiter, Tobias Geibinger, Nelson Higuera, and Johannes Oetsch. A Logic-based Approach to Contrastive Explainability for Neurosymbolic Visual Question Answering. In *Proc. IJCAI 2023*, pagespp. 3668–3676. ijcai.org, 2023.

[9] Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering Systematic Errors with Cross Modal Embeddings. In *Proc. ICLR 2022*, OpenReview.net, 2022.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 Proc. CVPR 2016*, pp. 770–778. IEEE Computer Society, 2016.

[11] Pascal Hitzler and Md. Kamruzzaman Sarker, eds. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, vol. 342 *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021.

[12] Glenn Jocher. YOLOv5 by Ultralytics, May 2020.

[13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image Retrieval using Scene Graphs. In *Proc. CVPR 2015*, pp. 3668–3678. IEEE Computer Society, 2015.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, 60(6):84–90, 2017.

[15] Mark Law, Alessandra Russo, and Krysia Broda. The ILASP system for learning Answer Set Programs. `www.ilasp.com`, 2015.

[16] Hongsheng Li, Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Xia Zhao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey. *Neurocomputing*, 566:127052, 2024.

[17] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L. Yuille. Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning. In *Proc. CVPR 2023*, pp. 14963–14973. IEEE, 2023.

[18] Vladimir Lifschitz. *Answer Set Programming*. Springer, 2019.

[19] Jan Hendrik Metzen, Robin Hutmacher, N. Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of Systematic Errors of Image Classifiers on Rare Subgroups. *CoRR*, abs/2303.05072, 2023.

[20] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In Marzyeh Ghassemi, ed, *Proc. ACM CHIL 2020*, pp. 151–159. ACM, 2020.

[21] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, eds, *ICML 2019*, vol. 97 *Proc. Machine Learning Research*, pp. 5389–5400. PMLR, 2019.

[22] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Yoshua Bengio and Yann LeCun, eds, *Proc. ICLR 2015*, 2015.

[23] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019.

[24] Mengshi Zhang, Yuqun Zhang, Lingming Zhang, Cong Liu, and Sarfraz Khurshid. DeepRoad: GAN-Based Metamorphic Testing and Input Validation Framework for Autonomous Driving Systems. In Marianne Huchard et al., eds, *Proc. ASE 2018*, pp. 132–142. ACM, 2018.