How to Think About Benchmarking Neurosymbolic AI?

Johanna Ott[3,*], Arthur Ledaguenel[1,2,*], Céline Hudelot[1] and Mattis Hartwig[3,4,*]

[1]MICS, CentraleSupélec, Université Paris-Saclay, Paris, France
[2]IRT SystemX, Paris-Saclay, France
[3]German Research Centre for Artificial Intelligence (DFKI), Lübeck, Germany
[4]singularIT GmbH, Leipzig, Germany

Abstract

Neurosymbolic artificial intelligence is a growing field of research aiming at combining neural networks with symbolic systems, including their respective learning and reasoning capabilities. This hybridization can take many shapes which adds to the fragmentation of the field and makes it difficult to compare the existing approaches. If some efforts have been made in the community to define archetypical means of hybridization, many elements are still missing to establish principled comparisons. Amongst those missing elements are formal and broadly accepted definitions of neurosymbolic tasks and their corresponding benchmarks. In this paper, we start from the specific task of multi-label classification with the integration of propositional background knowledge to illustrate how such a benchmarking framework could look like. Based on the benchmarking of one granular task we zoom out and discuss important elements and characteristics of building a full benchmarking suite for more than just one task.

1. Introduction

Neurosymbolic artificial intelligence (AI) is a trending research topic [1]. In general, neurosymbolic AI focuses on bringing together concepts from the logic-focused symbolic world and the neural or connectionist's world [2, 3, 4, 5].

The potential of the field is based on the "best of both worlds" perspective, i.e., that by combining neural and symbolic, the respective strengths are maintained while the weaknesses are minimized. Thus, the objectives are extensive and include, amongst others, improved performance [6, 7, 8, 9], explainability [10, 6, 11, 12, 13, 14] and generalization [15, 10, 11, 12, 13, 9, 16].

Contrasting its promise of generalization, the field of neurosymbolic AI exhibits a progress-hampering level of fragmentation, e.g. in the evaluation and the architectural landscape. There have been several attempts to structure the architectural approaches in the neurosymbolic AI field [17, 2, 18, 19, 20, 21]. In this paper, we focus on the fragmented evaluation landscape, i.e. the tasks, datasets and metrics used to evaluate neurosymbolic systems. Although not the focus of this paper, we believe that further work on a clear, unified architectural taxonomy is needed

NeSy2023: 17th International Workshop on Neural-Symbolic Learning and Reasoning, Siena, Italy
*Corresponding author. These authors contributed equally.
b00782280@essec.edu (J. Ott); arthur.ledaguenel@irt-systemx.fr (A. Ledaguenel); mattis.hartwig@dfki.de (M. Hartwig)

and that the current ambiguity about the separation of the neural, symbolic, and neurosymbolic worlds adds to the fragmented evaluation landscape.

Previous researchers have highlighted the fragmentation problem and emphasized the need for a more systematic approach to evaluating neurosymbolic AI [22, 19, 1]. Although efforts have been made to tackle this issue [23, 24, 25, 26, 27, 28], they have primarily remained at a narrow and specific level, i.e. they propose specific tasks and benchmarks, including datasets and evaluation metrics. Only a few exceptions, such as the panel discussion "The future of (neuro-symbolic) AI" at the IBM Neuro-Symbolic AI Workshop 2022 [29] and the presentation by Madhyastha and subsequent open discussion at NeSy2022 conference[22], have addressed the neurosymbolic benchmark fragmentation issue on a level beyond a specific benchmark.

In this position paper, we seek to complement the prior work tackling the fragmented neurosymbolic benchmark landscape by facing the challenge on a higher-level, focusing on the question of how to think about benchmarking neurosymbolic AI. We give an example for the setup of a specific benchmark on the task of multi-label classification with symbolic background knowledge. We include the thought process of coming up with a formal definition of the task, a suitable dataset, and a selection of metrics. Additionally, we discuss the implications for adding further benchmarks using our proposed thought process and thus contribute to a more principled benchmarking landscape for neurosymbolic AI.

2. Benchmarking neurosymbolic systems on a specific task

A neurosymbolic benchmark can be designed to answer two main questions: "What performance level can neurosymbolic systems reach on a given task?" and "How does hybridization of neural and symbolic components help on a given task?". The first question takes an outside view focusing on observable behavior while the second question takes an inside view focusing on the design of agents. The inside and the outside view are two well-known and deeply grounded perspectives in AI research [30]. We agree with Russell that in general artificial intelligence should be measured taking an outside view. However, answering the second question with the inside view can give further insights on how to design AI agents by understanding how and when to use neurosymbolic architectures. Additionally it might help directing the research efforts of the neurosymbolic community because advancement in task performance can be better linked to the architectural setup of the agent.

Hence, in this section, we describe the challenges of benchmarking the task of multi-label classification with symbolic background knowledge so that the two questions (internal and external) can be answered. We cover the formal definition of the task, the underlying dataset and the metrics. Although a task is not *per se* neurosymbolic, our chosen task covers elements that are linked to a neural (image classification) and to a symbolic (background knowledge) domain. This setup makes relatively straight-forward to use agents with a neurosymbolic architecture, and is suitable for a neurosymbolic benchmark that answers both questions.

2.1. Task formalism

Setting a formal definition of the task is a necessary preliminary step to compare neurosymbolic systems in a principled way. To be practical, the formalism also has to be comprehensive enough

to incorporate diverse datasets (in terms of modality and background knowledge structure) and avoid a fragmentation of the field into multiple narrower tasks definitions.

Multi-label classification with background knowledge is mapping inputs $x \in \mathbb{R}^d$ to binary labels $\mathbf{y} \in \{0,1\}^k$ such that these labels satisfy some background knowledge. This background knowledge is expressed as a propositional formula α using symbols from the signature $\mathcal{S} := \{Y_j\}_{1 \leq j \leq k}$ and logical connectors $\{\neg, \wedge, \vee\}$ with their standard semantics. For lighter notations, we identify a label $\mathbf{y} \in \{0,1\}^k$ with the propositional valuation mapping each Y_j to y_j. Therefore, we note $\mathbf{y} \vDash \alpha$ if the corresponding valuation models α. A dataset for that task is $\mathscr{D} := (x^i, \mathbf{y}^i)_{1 \leq i \leq n}$ with $x^i \in \mathbb{R}^d$, $\mathbf{y}^i \in \{0,1\}^k$ such that all labels in the dataset satisfy the background knowledge, i.e. $\forall 1 \leq i \leq n, \mathbf{y}^i \vDash \alpha$.

This formalism encompasses standard classification tasks like independent binary classification (where $\alpha = \top$ since every combination is valid) and multi-category classification (where $\alpha = (\bigvee_{1 \leq j \leq k} Y_j) \wedge (\bigwedge_{1 \leq j < l \leq k} (\neg Y_j \vee \neg Y_l))$ enforces that one and only one atom is true at a time).

Since we formally introduced our task we need to discuss the dataset and the metrics to complete our benchmark.

2.2. Datasets

Building an appropriate dataset for multi-label classification with background knowledge poses a substantial challenge. It must contain large amounts of data amenable to neural processing and whose labels present some significant structure expressible in the language of propositional logic. Efforts to build such datasets were often led by researchers trying to measure the performance of their neurosymbolic system, meaning that different systems are rarely evaluated on the same datasets and that datasets are often custom built to fit the capacity of a given system.

We observed three patterns in how datasets were created: **symbolic** datasets where a symbolic reasoning task is turned into a learning task (e.g. finding the shortest path in a weighted graph [31]), **compositional** datasets where instances are tuples of a base sub-symbolic classification dataset constrained to respect a given structure (e.g. the MNIST SUDOKU dataset [26]) and **hierarchical** datasets where classes of a sub-symbolic classification dataset are chosen in a hierarchy of concepts (e.g. classes in ImageNet [32] are chosen amongst synsets of the WordNet hierarchy [33]).

To turn this collection of datasets into an efficient benchmark for multi-label classification with background knowledge, further aspects need to be considered. On a fundamental level, we observe an inverse relation between the complexity of the sub-symbolic features and the complexity of the symbolic structure of the dataset, which means that the zone of complex sub-symbolic features and complex symbolic structure is not well covered by existing datasets. [25] is a dataset of traffic videos (complex sub-symbolic features) where labels satisfy a rich set of constraints (complex symbolic structure). It constitutes a first step to cover that void and more efforts should be invested in that direction. On the practical side, we need to set up a standard on how to represent, store and operate neurosymbolic datasets and their corresponding background knowledge, to allow rapid testing of any system on any dataset.

2.3. Metrics

To evaluate neurosymbolic systems inside our benchmark we use a combination of performance metrics (the outside view) and control metrics (the inside view). Examples of standard performance metrics are cross-entropy loss, individual accuracy, f1-score, collective accuracy, top-k accuracy. Likewise, for standard control metrics we have number of trainable parameters, number of hyper-parameters, number of FLOPS.

Besides, new control or performance metrics specific to neurosymbolic tasks might be beneficial. One example for such a performance metric is semantic consistency which tracks how many predictions of a given system match the constraints expressed by the background knowledge (see [34] or [25] for instance).

To settle on a limited set of metrics for the multi-label classification with background knowledge task (which also can be used for other tasks), we suggest to use collective accuracy and semantic consistency as performance metrics and network size (number of trainable parameters) as a control metric. The semantic consistency metric helps us understand how much the system integrates background knowledge. Collective accuracy is a very demanding metric that is robust to imbalanced datasets: we generally observe a strong correlation between collective accuracy and f1-score for instance. Eventually, network size is a good first order approximation for model capacity.

3. Broaden the focus on a collection of tasks

To extend the thoughts on the specific task from the previous section to cover more of the neurosymbolic AI field, a natural next step is to focus on transferring the approach on more tasks. We draw confidence in the transferability of the proposed thought process from the observation that benchmarks cited in the preceding sections have already incorporated some of our suggestions (e.g. implementing control metrics). Furthermore, existing benchmarks may benefit from our thought process to improve their comparability. For instance, in visual reasoning, CLEVR [35] and CLEVRER [36] benchmarks do not provide a formal definition of the task, which makes comparisons between systems and with other datasets hard to establish. Moreover, both underlying datasets lack sub-symbolic complexity compared to classic computer vision datasets: the community could greatly benefit from filling that void.

Expanding the focus from a specific task to a collection of tasks, i.e., creating a benchmarking suite, raises another critical question: Which tasks should be included? The diversity of tasks has been identified as a key consideration for a benchmarking suite by the discussion panel in [29]. Potential tasks should have ranging difficulties for both the neural and the symbolic architecture. Also similar to the Glue [37] or GlueCon [38] benchmarking suites, several different capabilities and skills should be needed to solve the tasks.

4. Conclusion

This position paper contributes an example thought process for designing neurosymbolic AI benchmarks. Of course a single position paper cannot fully solve all questions around building

a unified benchmarking system, but, in contrast to other papers in the field so far, we refrained from marketing an individual dataset and focused more on the questions around the design phase of a benchmark. We also discussed the implications for broadening the approach to multiple tasks which will be a valuable starting point for future benchmarking discussions and designs. Next steps could include to validate our approach on more tasks and add further thoughts to the discussion around important characteristics of a more holistic benchmarking suite.

References

[1] K. Hamilton, A. Nayak, B. Bozic, L. Longo, Is neuro-symbolic ai meeting its promise in natural language processing? a structured review, ArXiv abs/2202.12205 (2022).

[2] M. K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neuro-symbolic artificial intelligence: Current trends, 2021. URL: https://arxiv.org/abs/2105.05330. doi:10.48550/ARXIV.2105.05330.

[3] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, L. Zhou, Neuro-symbolic approaches in artificial intelligence, National Science Review 9 (2022). URL: https://doi.org/10.1093/nsr/nwac035. doi:10.1093/nsr/nwac035. arXiv:https://academic.oup.com/nsr/article-pdf/9/6/nwac035/43952954/nwac035_supplemental_file.pdf, nwac035.

[4] Z. Susskind, B. Arden, L. K. John, P. Stockton, E. B. John, Neuro-symbolic ai: An emerging class of ai workloads and their characterization, 2021. URL: https://arxiv.org/abs/2109.06133. doi:10.48550/ARXIV.2109.06133.

[5] P. Hitzler, M. K. Sarker, T. R. Besold, A. D. Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K. U. Kühnberger, L. C. Lamb, P. M. H. V. Lima, L. D. Penning, G. Pinkas, H. Poon, G. Zaverucha, Neural-Symbolic Learning and Reasoning: A Survey and Interpretation, volume 342, 2022. doi:10.3233/FAIA210348.

[6] D. Lyu, F. Yang, B. Liu, S. Gustafson, Sdrl: Interpretable and data-efficient deep reinforcement learning leveraging symbolic planning, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 2970–2977. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4153. doi:10.1609/aaai.v33i01.33012970.

[7] D. Demeter, D. Downey, Just add functions: A neural-symbolic language model, in: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 2020. doi:10.1609/aaai.v34i05.6264.

[8] F. Yang, D. Lyu, B. Liu, S. Gustafson, Peorl: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making, in: IJCAI International Joint Conference on Artificial Intelligence, volume 2018-July, 2018. doi:10.24963/ijcai.2018/675.

[9] H. Jiang, S. Gurajada, Q. Lu, S. Neelam, L. Popa, P. Sen, Y. Li, A. G. Gray, Lnn-el: A neuro-symbolic approach to short-text entity linking, in: Annual Meeting of the Association for Computational Linguistics, 2021.

[10] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=rJgMlhRctm.

[11] Y. Feng, X. Yang, X. Zhu, M. A. Greenspan, Neuro-symbolic natural logic with introspective

revision for natural language inference, Transactions of the Association for Computational Linguistics 10 (2022) 240–256.

[12] K. Zheng, K.-Q. Zhou, J. Gu, Y. Fan, J. Wang, Z. xiao Li, X. He, X. E. Wang, Jarvis: A neuro-symbolic commonsense reasoning framework for conversational embodied agents, ArXiv abs/2208.13266 (2022).

[13] Y. Liang, J. Tenenbaum, T. A. Le, S. N, Drawing out of distribution with neuro-symbolic generative models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 15244–15254. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/6248a3b8279a39b3668a8a7c0e29164d-Paper-Conference.pdf.

[14] B. Finzel, A. Saranti, A. Angerschmid, D. Tafler, B. Pfeifer, A. Holzinger, Generating explanations for conceptual validation of graph neural networks: An investigation of symbolic predicates learned on relevance-ranked sub-graphs, KI - Künstliche Intelligenz 36 (2022) 271–285. doi:10.1007/s13218-022-00781-7.

[15] M. B. Ganapini, M. Campbell, F. Fabiano, L. Horesh, J. Lenchner, A. Loreggia, N. Mattei, F. Rossi, B. Srivastava, K. B. Venable, Combining fast and slow thinking for human-like and efficient decisions in constrained environments, in: International Workshop on Neural-Symbolic Learning and Reasoning, 2022.

[16] X. Chen, C. Liang, A. W. Yu, D. Song, D. Zhou, Compositional generalization via neural-symbolic stack machines, in: Advances in Neural Information Processing Systems, volume 2020-December, 2020.

[17] S. Bader, P. Hitzler, Dimensions of neural-symbolic integration - a structured survey, 2005. URL: https://arxiv.org/abs/cs/0511042. doi:10.48550/ARXIV.CS/0511042.

[18] H. A. Kautz, The third ai summer: Aaai robert s. engelmore memorial lecture, AI Mag. 43 (2022) 93–104.

[19] A. d'Avila Garcez, L. C. Lamb, Neurosymbolic ai: the 3rd wave, Artificial Intelligence Review (2023). doi:10.1007/s10462-023-10448-w.

[20] F. V. Harmelen, A. ten Teije, A boxology of design patterns for hybrid learning and reasoning systems, Journal of Web Engineering 18 (2019) 97–124.

[21] L. d. Raedt, S. Dumančić, R. Manhaeve, G. Marra, From statistical relational to neuro-symbolic artificial intelligence, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4943–4950. URL: https://doi.org/10.24963/ijcai.2020/688. doi:10.24963/ijcai.2020/688, survey track.

[22] P. Madhyastha, Towards a benchmark suite for neural-symbolic approaches for learning and reasoning, 2022. URL: https://ijclr22.doc.ic.ac.uk/program/index.html, 16th International Workshop on Neural-Symbolic Learning and Reasoning.

[23] Ö. Yılmaz, A. S. d'Avila Garcez, D. L. Silver, A proposal for common dataset in neural-symbolic reasoning studies, in: NeSy@HLAI, 2016.

[24] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, R. B. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1988–1997.

[25] E. Giunchiglia, M. C. Stoian, S. Khan, F. Cuzzolin, T. Lukasiewicz, Road-r: the autonomous driving dataset with logical requirements, Machine Learning (2023). doi:10.

1007/s10994-023-06322-z.
[26] E. Augustine, C. Pryor, C. Dickens, J. Pujara, W. Y. Wang, L. Getoor, Visual sudoku puzzle classification: A suite of collective neuro-symbolic tasks, in: International Workshop on Neural-Symbolic Learning and Reasoning, 2022.
[27] A. D. Lindström, S. S. Abraham, Clevr-math: A dataset for compositional language, visual and mathematical reasoning, volume 3212, 2022.
[28] C. Cornelio, V. Thost, Synthetic Datasets and Evaluation Tools for Inductive Neural Reasoning, in: N. Katzouris, A. Artikis (Eds.), Inductive Logic Programming, Springer International Publishing, Cham, 2022, pp. 57–77.
[29] F. Rossi, H. Kautz, G. Marcus, L. Lamb, L. Kaelbling, Closing, 2022. URL: https://video.ibm.com/recorded/131288165, iBM Neuro-Symbolic AI Workshops.
[30] S. J. Russell, Artificial intelligence a modern approach, Pearson Education, Inc., 2010.
[31] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. V. D. Broeck, A semantic loss function for deep learning with symbolic knowledge, volume 12, International Machine Learning Society (IMLS), 2018, pp. 8752–8760.
[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.
[33] G. A. Miller, Wordnet, Communications of the ACM 38 (1995) 39–41. URL: https://dl.acm.org/doi/10.1145/219717.219748. doi:10.1145/219717.219748.
[34] K. Ahmed, S. Teso, K.-W. Chang, G. Van den Broeck, A. Vergari, Semantic probabilistic layers for neuro-symbolic learning, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 29944–29959. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c182ec594f38926b7fcb827635b9a8f4-Paper-Conference.pdf.
[35] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2901–2910.
[36] K. Yi*, C. Gan*, Y. Li, P. Kohli, J. Wu, A. Torralba, J. B. Tenenbaum, Clevrer: Collision events for video representation and reasoning, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=HkxYzANYDB.
[37] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. URL: https://aclanthology.org/W18-5446. doi:10.18653/v1/W18-5446.
[38] H. R. Faghihi, A. Nafar, C. Zheng, R. Mirzaee, Y. Zhang, A. Uszok, A. Wan, T. Premsri, D. Roth, P. Kordjamshidi, Gluecons: A generic benchmark for learning under constraints, ArXiv abs/2302.10914 (2023).