# Elaboration for Neurosymbolic Computation (Work in Progress)

Thomas Eiter and Nelson Higuera

Knowledge Based Systems Group, TU Wien, Vienna, Austria.

**Abstract.** Neurosymbolic Computation aims to unify the two main branches in AI, namely, neural networks and logic. Answer Set Programming (ASP) is a good candidate for the logic part as it offers a declarative and expressive language. We consider a Visual Question Answering (VQA) problem, where we want to answer a question using visual input, for which neural and neurosymbolic approaches achieved good results. Our interest is here with elaboration of the questions, when new predicates, of increasing sophistication, become available. To this end, we experimentally compare a neural-based approach against a neurosymbolic one over the CLEVR dataset. The results show that the latter approach is more robust to the new questions, achieving high accuracy, and also provides the benefit of producing explainable answers. On the downside, the neurosymbolic approach requires that the semantics of the questions respective predicates have to be manually coded. Preliminary work is being done to relieve this by using Inductive Logic Programming to learn the semantics of the predicates.

**Introduction.** Visual Question Answering (VQA) is the field of problems where we want to answer questions, posed in natural language, against an image or video [1]. To solve this, different capabilities such as language parsing, object recognition, and reasoning are needed. Different approaches to VQA exist, among them neural-based approaches, with end-to-end processing of questions by neural networks [2], as well as neurosymbolic approaches [3] that combine neural and logic modules in a system [4, 5]. CLEVR [6] is a synthetic dataset for VQA created to diminish the bias introduced by human made questions. In CLEVR, questions are about geometric objects concerning their color, shape, size, and matter, as well as position and spatial relationship. State of the art performers on this dataset are, in the neural-based field, the Memory, Attention, and Composition (MAC) system [7], and in the neurosymbolic field, NS-VQA [8]. Both systems achieve very high accuracy of 98.9% and 99.8%, respectively. These frameworks also share architectural similarities, as they both combine a CNN [9] and LSTM [10].

**Experiments.** We are interested in the elaboration of the CLEVR scenario, when new predicates and questions are added, and want to see the effects on the two approaches, tested on the representative systems. To this end, we introduce 20 new questions templates separated into three groups, namely, the Between,

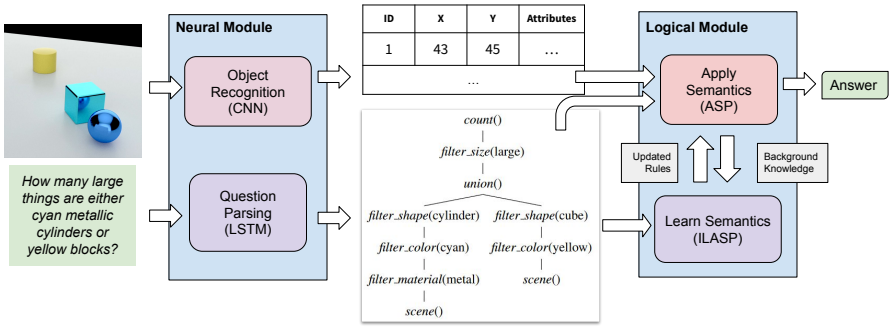


Fig. 1: Neurosymbolic VQA with ILP integration.

Equal and Count Between Groups. For example, the Between Group contains the questions *Between projection*, *Between bbox* and *Between proper*. The first question asks whether the projection of an object is Between the projection of two other objects to one dimension ($x$-axis), the second one expands this with an additional dimension and the last one uses 3D coordinates to test the predicate. We use the question generator provided with CLEVR to generate about 500k new questions using the original images and new templates. We begin by training the MAC on a combination of the original dataset and the new dataset. This first experiment shows that the MAC handles the predicates without major issues, but a following one, where we train the MAC solely on the new question says otherwise, as most of the predicates were not learnt.

Given that the combination shows good results, we hypothesize that learning from the original dataset may be transferable to the new ones, as the original CLEVR also deals with spatial relations, Equality and counting. A zero-shot experiment (i.e., without further learning) shows no indication of such, as the results are similar to the ones of the last experiment. We then consider NS-VQA for solving the task. Its modular architecture allows us to analyze each component separately. As we used the original images to generate the new questions, the CNN module will not change its accuracy. Furthermore, the logical module is handcoded and should make no mistakes whenever its input is correct. For these reasons, we only train the LSTM on the new questions, where the experiments show that it is much faster and more accurate.

**Current Work and Conclusion.** The neurosymbolic approach was able to learn the questions and has the benefit of being explainable, as the logic module is transparent in its execution. To cover the downside of writing semantics for the logical module, we introduce an intended framework shown in Figure 1, where we integrate ILP to learn the semantics as rules, which extends transparency and explainability. Preliminary experiments using ASP and ILASP [11] show that for some predicates this works, but many are still a challenge to learn.

**Acknowledgements.** This work was supported by funding from the Bosch Center for AI at Renningen, Germany.

## References


1. Yeyun Zou and Qiyu Xie. A survey on VQA: Datasets and approaches. In *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*. IEEE, dec 2020.
2. Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding, 2021.
3. Artur S. d'Avila Garcez and L. Lamb. Neurosymbolic ai: The 3rd wave. *ArXiv*, abs/2012.05876, 2020.
4. Zhun Yang, Adam Ishay, and Joohyung Lee. Neurasp: Embracing neural networks into answer set programming. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1755–1762. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.
5. Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in deepproblog, 2019.
6. Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017.
7. Drew Arad Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018.
8. Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1039–1050, Red Hook, NY, USA, 2018. Curran Associates Inc.
9. Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015.
10. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
11. Mark Law, Alessandra Russo, and Krysia Broda. The ILASP system for inductive learning of answer set programs. 2020.