Integrating Non-monotonic Logical Reasoning and Inductive Learning with Deep Learning for Explainable Visual Question Answering

Mohan Sridharan[1] , **Heather Riley**[2]
[1]School of Computer Science, University of Birmingham, UK
[2]Department of Electrical and Computer Engineering, The University of Auckland, NZ
m.sridharan@bham.ac.uk, hril230@aucklanduni.ac.nz

Deep neural network architectures and algorithms represent the state of the art for many perception and control problems. Examples include object recognition, gesture recognition, object manipulation, and obstacle avoidance, in domains such as healthcare and surveillance. Common limitations of deep networks are that they are computationally expensive to train, require a large number of labeled training examples, and make it difficult to understand or explain the observed behavior. These challenges are more pronounced in integrated systems that include knowledge-based reasoning and data-driven learning components. Humans interacting with such systems designed for complex domains, with autonomy in some components, are likely to want to know why and how the system arrived at particular conclusions; this "explainability" will help designers improve the underlying algorithms. Understanding the operation of these systems will also help human users build trust in the decisions made by these systems. Despite considerable research in recent years, providing explanatory descriptions of reasoning and learning continues to be an open problem in AI.

We consider Visual Question Answering (VQA) as a motivating example of a complex task that inherently requires explanatory descriptions of reasoning and learning. Given a scene and a natural language question about an image of the scene, the objective of VQA is to provide an accurate answer to the question. These questions can be about the presence or absence of particular objects, the relationships between these objects, or the potential outcome of executing particular actions on objects in the scene. For instance, a system recognizing and responding to traffic signs on a self-driving car may be asked "what is the meaning of this traffic sign?" or "how should the driver respond to this sign?", and a system controlling a robot arm constructing stable arrangements of objects may be asked "why is this structure unstable?" or "what would happen if the structure is pushed?". We assume that such questions are provided as (or transcribed into) text, and that answers are also generated as text (that may be converted to speech). We consider answers to such *explanatory questions* to be relational descriptions of the related decisions and beliefs in terms of the domain objects, domain attributes, and agent actions. Deep networks are the state of the art for VQA, and we seek to address the known limitations of these networks by drawing inspiration from research in cognitive systems, which indicates that reliable, efficient, and explainable reasoning and learning can be achieved by jointly reasoning with commonsense domain knowledge and learning from experience (Laird et al. 2017). Specifically, *our architecture tightly couples knowledge representation, reasoning, and learning, and exploits the complementary strengths of non-monotonic logical reasoning with incomplete commonsense domain knowledge, deep learning, and inductive learning* (Riley and Sridharan 2019). The architecture, as shown in Figure 1, supports the following methodology:

- For any input image of a scene of interest, Convolutional Neural Networks (CNNs) extract concise visual features characterizing the image.
- Non-monotonic logical reasoning with the extracted features and incomplete commonsense domain knowledge is used to classify the input image, and to answer questions about the classification and the scene.
- Feature vectors that the non-monotonic logical reasoning is unable to classify are used to train a decision tree classifier that is also used to answer questions about the classification during testing.
- Feature vectors not classified by non-monotonic logical reasoning, along with the output of the decision tree classifier, train a Recurrent Neural Network (RNN) that is used to answer questions about the scene during testing.
- Feature vectors not classified by non-monotonic logical reasoning are also used to inductively learn and reason with constraints governing domain states; and
- Reasoning with commonsense knowledge is expanded (when needed) to support planning, diagnostics, and the ability to answer related explanatory questions.

Our implementation of this architecture uses CR-Prolog (Balduccini and Gelfond 2003), an extension of Answer Set Prolog, for non-monotonic logical reasoning with incomplete commonsense domain knowledge.

Although VQA is our motivating example, it is not the main focus of our work. State of the art VQA algorithms focus on generalizing to different domains, and are evaluated on benchmark datasets of thousands of images from different domains (Shrestha, Kafle, and Kanan 2019). Our focus, on the other hand, is on reliable, efficient, and transparent reasoning and learning in any given domain in which

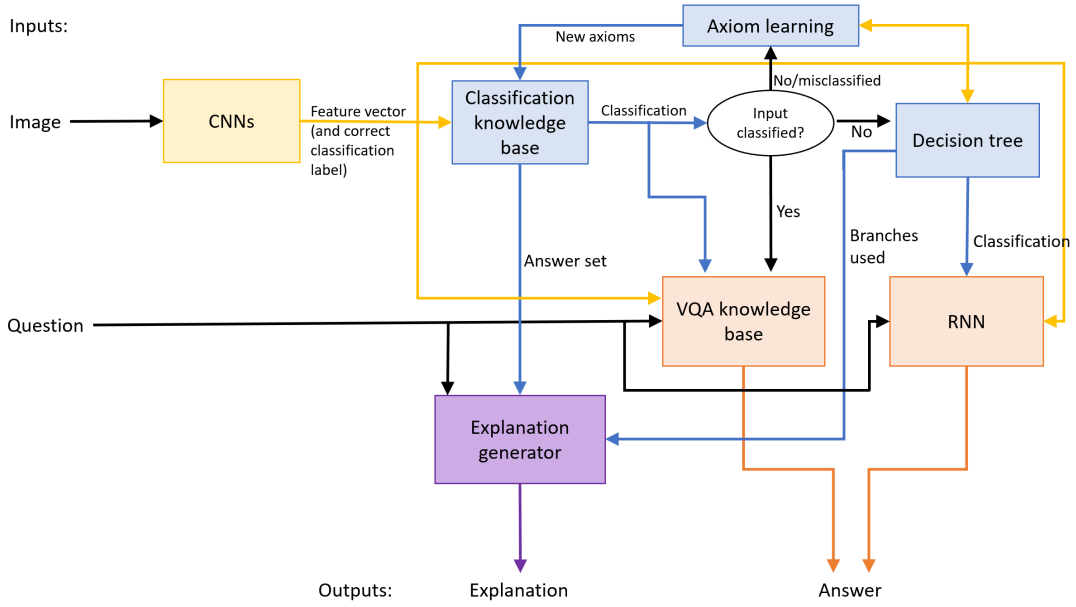

Figure 1: Architecture combines the principles of deep learning, non-monotonic logical reasoning, and decision-tree induction.

a large, labeled dataset is not readily available. We thus neither compare our architecture and algorithms with state of the art algorithms for VQA, nor use large benchmark VQA datasets for evaluation. Instead, we focus on the interplay between reasoning and learning, and evaluate our architecture's capabilities in the context of: (a) estimating the stability of configurations of simulated blocks on a tabletop; (b) recognizing different traffic signs in a benchmark dataset of images; and (c) a simulated robot delivering messages to the intended human recipients at different locations. The characteristics of these tasks and domains match our objective. In the first two domains, we focus on answering explanatory questions about images of scenes and the underlying classification problems (e.g., recognizing traffic signs). We also demonstrate how our architecture can used by a robot assisting humans (in the third domain) to compute and execute plans, and to answer questions about these plans. Experimental results indicate the following benefits in comparison with an architecture based only on deep networks:

1. Better accuracy and reduced computational effort on classification problems when the training dataset is small, and comparable accuracy with larger datasets while still using only a small set of training samples;
2. Ability to provide answers to explanatory questions about the scenes and the underlying decision making problems (e.g., classification, planning);
3. Incremental learning of previously unknown domain constraints, whose use in reasoning improves the ability to answer explanatory questions; and
4. Ability to adapt the complementary strengths of non-monotonic logical reasoning with commonsense domain knowledge, inductive learning, and deep learning, to address decision-making problems, e.g., to compute minimal and correct plans, on a robot.

Future work will explore complex real-world domains and investigate the use of knowledge-based reasoning systems to better understand the observed behavior of deep networks.

Acknowledgments

This work was supported in part by the US Office of Naval Research Science of Autonomy award N00014-17-1-2434, and the Asian Office of Aerospace Research and Development award FA2386-16-1-4071. All claims and conclusions are those of the authors.

References

Balduccini, M., and Gelfond, M. 2003. Logic Programs with Consistency-Restoring Rules. In *AAAI Spring Symposium on Logical Formalization of Commonsense Reasoning*.

Laird, J. E.; Gluck, K.; Anderson, J.; Forbus, K. D.; Jenkins, O. C.; Lebiere, C.; Salvucci, D.; Scheutz, M.; Thomaz, A.; Trafton, G.; Wray, R. E.; Mohan, S.; and Kirk, J. R. 2017. Interactive Task Learning. *IEEE Intelligent Systems* 32(4):6–21.

Riley, H., and Sridharan, M. 2019. Integrating Non-monotonic Logical Reasoning and Inductive Learning With Deep Learning for Explainable Visual Question Answering. *Frontiers in Robotics and AI, special issue on Combining Symbolic Reasoning and Data-Driven Learning for Decision-Making* 6:20. https://www.frontiersin.org/articles/10.3389/frobt.2019.00125/full.

Shrestha, R.; Kafle, K.; and Kanan, C. 2019. Answer Them All! Toward Universal Visual Question Answering Models. In *International Conference on Computer Vision and Pattern Recognition*.