Position Paper: Reasoning upon Learning: A Generic Neural-Symbolic Approach

Marjan Alirezaie, Martin Längkvist, Michael Sioutis, and Amy Loutfi

Center for Applied Autonomous Sensor Systems, Örebro University, Örebro, Sweden
FirstName.LastName@oru.se

Abstract. Machine learning techniques including deep learning methods, despite their increasing success in handling human-level tasks, still seldom perform without error. To be error free, such methods, need huge amount of suitable data, which is not always possible to provide. To deal with this problem, we may enable learning methods to explain themselves and understand their mistakes without the interference of human. To achieve an automated explanation, we propose to integrate the learning methods with ontologies as publicly available symbolic models representing human knowledge. The integration process is twofold: 1) integration of symbolic module (ontologies and reasoning) with the output of the learning method (i.e., error explain and also justification), 2) integration of symbolic module with the hidden layers of the learning module (i.e., feature extraction explanation).

Keywords: Neural-based learning · hidden layer interpretation, · symbolic reasoning

1 Purpose and aims

Achieving human-level intelligence and enabling machines to perform tasks which are intuitive for humans, has been one of the ultimate goals of much research work in Artificial Intelligence (AI). Recently, with the enormous amount of data, advanced machine learning techniques including deep learning methods have flourished in performing human-level tasks such as image recognition, speech recognition, language translation [18] or even more advanced ones including driving cars [6] or playing games [20].

The structure of many deep learning methods which are categorized as subsymbolic systems (vs. symbolic systems), is based on a metaphor of the human brain as a network of computational nodes (i.e., neurons) that trigger each other by transferring signals. The connections between neurons are assigned with weight values that determine the strength and the weakness of the connections in transferring signals. Deep learning algorithms, like other machine learning methods, are trained by optimizing a cost function that measures the training errors during learning, and adapts the weight values in order to minimize these errors. What makes deep learning methods special is their layered structure that takes the raw data (e.g., pixels of a given picture) as input at the first layer and

maps it into a high level concept (e.g., a Cat) at the last layer by passing the data through several hidden layers (i.e., the layers in the middle).

However, it is not clear how this mapping works. Although the math behind the construction of the network is understandable, it is not straightforward to understand which neurons (or group of neurons) at the hidden layers capture which features of data (e.g., color of pixels, shape of objects in figures, their distances, etc.,). In other words, hidden layers are seen as a black box that maps the raw input data into the outputs [19].

The lack of transparency on how a network of neurons has made the eventual decisions is becoming problematic, especially when the learning is used in crucial decision-making processes (e.g., in people's safety or health) [14]. Furthermore, again due to the lack of transparency, the process of understanding why the algorithm has failed is often the task of the human who, using domain knowledge and contextual information, can explain the likely reason(s) behind the errors. Let us illustrate with an example in Figure 1. The figure shows a deep neural network (DNN) [17], as a deep learning method, designed to label all the containers in a given picture. Although the DNN has labeled all the containers in the scene correctly (e.g., bowl and cups in green), there are a number of non-container objects that are wrongly labeled as containers (e.g., the ceiling lights in red).

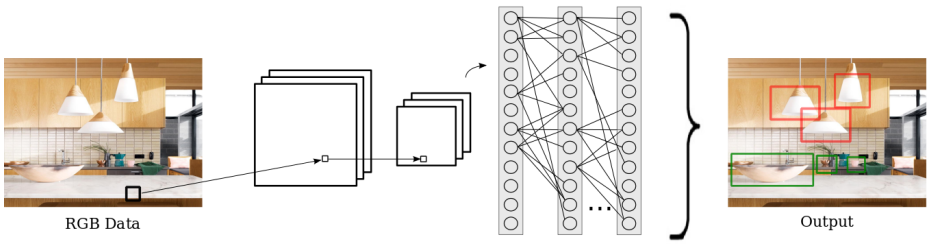

Fig. 1. A Deep Neural Network (DNN) used to classify and label containers in a given input image. The correctly labeled objects are shown in green and the wrongly labeled ones (e.g., ceiling lights) are shown in red.

In the given example, it is a human observer who using his/her knowledge about the context, can discover that the network has taken into account only the shape of objects (as the ceiling lights looks like a container), and not their functionality or their relations with their environment (a container logically cannot be hanged with a wire from the ceiling).

However, even if the human manage to discover and explain the problems, it does not necessarily help to find the main culprits in the structure of the network. Deep learning methods only rely on data. The human generated explanation can

only provide a better insight on how to refine the training dataset with better data whose details help the network to adjust its parameters. **More precisely, in order to be error free, the current deep learning methods, might need even more suitable data, which is not always possible to provide.** For instance, in the case of the container example, we (as human who understood the problem) have to feed the DNN with a large enough number of pictures in which the lamps hanging from the ceiling are labeled as non-container objects.

To deal with this problem we need to enable deep learning methods to explain themselves. In this way, they will also find the possibility to understand their mistakes by themselves and learn from their mistakes without the interference of human. Ultimately, to achieve an automated explanation, **symbolic models that represent human knowledge** should be used. In the area of AI, a common representation model based on formal languages, so called ontologies, was suggested by the Semantic Web community in 2001 [4] to represent the human knowledge available on the traditional Web. This knowledge is represented in the form of concepts that are interlinked together via their relations. Due to their formal languages, these publicly available ontologies are also understandable for machines. In other words, using ontologies, computer systems are able to understand the meaning of concepts based on their relations with the other concepts. For instance, the concept container is defined in an ontology as an object which is deep to hold things such as food or water. it can also be specialized into concepts bowls, cups, etc. The number of available, interlinked and public ontologies representing knowledge about different domains (e.g., medicine, agriculture, sport, geology, social science, etc.) was 12 dataset in 2007 and has reached to 1163 datasets in 2017[1]. Given this wealth of fast growing symbolic knowledge, one might think of integrating it with those systems relying on the human knowledge in different areas.

The main objective of this position paper is to develop a well founded framework which integrates ontological knowledge and reasoning with a sub-symbolic learning system. This integration (called neural-symbolic integration) will be explored at various phases of the learning process. The proposed integration of symbolic module (including ontological knowledge and reasoning) with sub-symbolic approaches is twofold:

1. The ontological reasoning process is applied upon the *output* of a sub-symbolic system (i.e., the output layer). This would be useful when the objective of the system is to have a semantic referee which either justifies the outputs or arbitrate or explains the errors. For example in our scenario above, a reasoning process upon an ontology that represents a kitchen in terms of its objects, their locations and affordances, could find the "container" label assigned to the ceiling lamp as an inconsistency. According to the ontology, a container is not designed to be hanged with a wire from the ceiling.
2. The ontological reasoning process is applied throughout the learning process (i.e., upon the hidden layers). This would be an attempt towards opening

[1] http://lod-cloud.net/

up the black box of the hidden layers by explaining the features extracted. Again using the above scenario, a reasoning process upon a hidden layer could recognize some hidden neurons which are likely dedicated to specific features of the objects such as size, shape, color, or affordances. Finding such neurons, the reasoner can decide on strengthening or weakening their weight values depending on their roles in differentiating objects.

In both aforementioned integration processes, the result of the reasoner will be sent back as a new set of data to the classifier. In other words, the explanations provided by the reasoner can to some extent play the role of good data required for learning and compensate the lack of such data in many domains of work.

2 State-of-the-art

Integration of data-driven learning methods with symbolic reasoning has been identified in the literature as the key factor of developing computationally robust systems [2] and has become one of the key challenges in Artificial Intelligence [10]. Depending on the approaches to represent both low and high level data, such integration has been addressed under different names that include abduction-induction reasoning in learning [15], structural alignment [1], and neural-symbolic methods [5, 3]. With the increasing interest in sub-symbolic systems, and in particular in deep learning methods, research on integrated neural-symbolic systems has recently made considerable progress. Such integrations are routinely referred to as explainable Artificial Intelligence (XAI) [7], and used to provide better insights into the learning process [9]. These insights becomes necessary when the reliability (and not only the precision) of the learning (or classification) methods matters specially in crucial decision-making processes.

As discussed in [22], in neural-symbolic systems where the learning (or the classification process) is based on a sub-symbolic method, one way of interpreting the classification process is to explain its outputs using the concepts related to the classifier's decision. The work presented in [11] introduces a learning method based on a convolutional deep network, LRCN [8], used to classify and label objects in a picture given as input. The learning method also provides textual explanations over the labels assigned to the objects in the picture. These generated texts discriminate features of the objects found in the scene. However, in this work there is no specific symbolic representation used, and the features of the objects are taken from the sentences that were already available for each image in the dataset (CUB dataset [21]). In other words, the sentences were not inferred by any reasoner, and instead were directly taken from the labels assigned to the images.

With focus on symbolic representation, the work presented in [16] proposes a system that uses an ontology to explain the classifier's outputs. The key tool of the system, called DL-Learner, works in parallel with the classifier and accepts the same data as input. Using the Suggested Upper Merged Ontology (SUMO)[2]

[2] http://www.adampease.org/OP/

as the symbolic knowledge model, the DL-Learner is also able to categorize the images by reasoning upon the objects together with the concepts defined in the ontology. The compatibility between the output of the DL-Learner and the classifier can be seen as a reliability support and at the same time as an interpretation of the classification process. However, in this work, there is no interaction between the classifier and the ontological reasoner. They are just used to accept the same input and work in parallel and independent from each other. Due to the gap between the symbolic reasoner and the learning process, the reasoner cannot contribute in resolving misclassifications.

Likewise, the work detailed in [12] relies on an ontological knowledge model called ConceptNet [13]. In this work, the integration of the symbolic model and a sentence-based image retrieval process based on deep learning is again used to explain the classification process. For this, the knowledge about different objects (e.g., their affordances, their relations with other objects) is aligned with concepts derived from the deep learning method. However, except on the output layer of the network, there is again no interaction between the learning process and the reasoning.

To summarize, although the study of the literature shows the interest of the AI community in integrating symbolic reasoning with subsymbolic systems, however we are far from achieving a robust and stand alone computational system as the result of neural symbolic integration. There are a number of shortcomings in the state of the art summarized as follows:

1. Each of the neural-symbolic approach discussed in the literature is bounded and constrained to a specific learning method.
2. Except the few efforts taking the ontological knowledge into account, the structure of the symbolic knowledge and reasoning methods have not been the main focus of the works. However, even the ontology based approaches were focusing on specific ontologies with simplified concepts and relations. There is no research on how upper level ontologies with general knowledge can be automatically found by and integrated with a learning method.
3. The main reason behind the neural-symbolic integration efforts has been mainly for the sake of interpretation as a reliability support on the neural-based classifier. Improving the performance of the neural-based classifier by learning from its mistakes has not been the reason behind using symbolic reasoning. In other words, there is no interaction implemented between the neural classifiers and the symbolic reasoners.
4. Although explainability of a sub-symbolic learning process is important, none of the neural-symbolic approaches has focused on interpreting the hidden layers to open the black box.

3 Significance and scientific novelty

We can implement more robust neural-symbolic integration by addressing the shortcomings of the state of the art mentioned in Section 2.

First, the ideal integration is meant to be independent of the architecture of the sub-symbolic (neural based) system. The integration only considers the common features of the neural-based learning methods, regardless of their technical details considered in specific architectures.
The second novel aspect is the idea of using publicly available ontologies to further automate the process of reasoning. Again in the aforementioned research work, the symbolic representation has been adhocly designed and not necessarily in the form of ontologies. Using ontologies is one essential step towards excluding human from the loop of explaining a learning process or providing good data for it.
Finally, the main goal behind the ideal integration approach is to enable learning methods, instead of only relying on data, to employ symbolic reasoner at different learning phases. It is one further step towards modeling human-like learning. In the proposed approach the reasoning is applied within the learning process to first let it understand its mistakes, and then adjust its parameters to avoid making the same errors without the interference of human. For this, we propose to apply symbolic reasoning also upon the hidden layers which has not been studied in the literature.

4 Preliminary and previous results

The proposed idea was emerged while we were implementing our neural-symbolic system. This work which has been submitted to the Semantic Web Journal and under public review[3], is as a preliminary but practical support of the proposed idea in this manuscript. The idea in this work is to resolve the misclassification made by a deep neural network (CAE) applied on satellite imagery data using ontological reasoning. The publicly available ontology used in this work is called OntoCity[4] and represents general knowledge about the structure of cities in terms of their physical landmarks (e.g., streets, buildings, rivers, roads, railroads, etc) and their spatial relations and constraints (e.g., railroads cannot be directly connected to buildings). The CAE classifier is used to classify and segment the satellite imagery data belonging to the central part of Stockholm. The CAE classifier is able to learn from its mistakes by using a semantic referee. The semantic referee can explain the errors (by reasoning behind the misclassification) based on spatial reasoning and the ontological knowledge about cities. Given the explanation about the errors, the proposed solution is also able to correct the errors by informing the classifier with extra information inferred by the reasoner.
This work, as a preliminary step, shows how by establishing interaction between the classifier and the reasoner we can improve the classification results by understanding the semantics behind the errors. This work, although, only takes the first and the last layers of the neural-based network into account,

[3] http://www.semantic-web-journal.net/content/improving-image-classification-geospatial-data-using-semantic-referee
[4] https://w3id.org/ontocity/ontocity.owl

however, to the best of our knowledge, is the first attempt where symbolic techniques are used to explain the errors made by such learning methods. We envision the design of the semantic referee to be more integrated with the structure of the learner such that the interaction between the two systems is not limited to only the first or last layers but instead involves the hidden layers of the learner as well.

5 Discussion

We are primarily interested in developing the basic theoretical framework for the outlined objective. The neural-symbolic integrated system presented in this paper is expected to be generic and self-standing, independent from a specific deep learning architecture per se and to provide three different types of explanations as follows:

(a) **Feature explanation**: explaining the feature extraction phase saying which features have been selected to be used in the learning process.
(b) **Error explanation**: explaining the errors that the learning process commits.
(c) **Justification**: explaining the process based on other features of data that are not necessarily used or extracted within the learning process.

In order to keep the neural-symbolic integration independent of the domains or uses cases at hand, we divide the the development phase in the form of three different tasks:

Task 1: includes investigation of different deep learning methods in order to formalize a generic domain-independent model. More specifically, the investigation is about finding relations between the parameters of different learning models and the (conceptual) features of the training data. The formalization of these relations are evolved while the networks are trained within a number of training cycles. For the sake of generality, during this phase, each deep learning method is investigated with different types of data including imagery data, environmental sensor data, etc.

Task 2: includes the investigation of existing upper level ontologies which represent upper level concepts and relations used (borrowed) as the basis of more specialized concepts in other ontologies. The complexity of the reasoning methods highly depends on how the knowledge is represented in the ontologies (e.g., in terms of the logical operators, quantifiers, and other logical constraints used in the definition of axioms). Given the results of the first task in the form of formalized relations between the parameters of the generic learning model and the features of data, this investigation is required to adjust or extend the available ontologies to guarantee computationally efficient reasoning process.

Task 3: includes designing algorithms to establish a communication link between the ontological reasoner and the learning methods. The link is required to translate the features or other parameters of the learning model into concepts, relations or in general a query understandable for the reasoner. At the same time, the interface has to also be able to translate the reasoning outputs (the inferred suggestions, explanations, or any extra information) into a representation model used by the learning system. This task can be summarized as the design and development of an ontology based query-answering module for deep learning methods.

To validate the applicability and the generality of the proposed framework, an extensive experimental validation will be conducted on different domains. As the continuation of the preliminary work explained in Section 4, we will start using satellite imagery data for which OntoCity has been developed as an available geo-related representation model. Indoor environments (such as smart homes) are the second domain in which classification/segmentation approaches are highly applied to address different problems such as object recognition used by assisting robots, recognition of activities of daily livings (ADLs) of inhabitants by a sensory system, etc.

References

1. Alirezaie, M., Loutfi, A.: Ontology alignment for classification of low level sensor data. In: International Conference on Knowledge Engineering and Ontology Development (KEOD). pp. 89–97. SciTePress (2012)
2. Bader, S.: Neural-Symbolic Integration. Ph.D. thesis, Chapter 1- TechnischeUniversitt Dresden, Dresden, Germany (2009)
3. Bader, S., Hitzler, P.: Dimensions of neural-symbolic integration - A structured survey. CoRR **abs/cs/0511042** (2005)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5), 34–43 (May 2001)
5. Besold, T., Garcez, A., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kuehnberger, K., Lamb, L., Lowd, D., Lima, P., de Penning, L., Pinkas, G., Poon, H., Zaverucha, G.: Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. CoRR **abs/1711.03902** (2017)
6. Bojarski, M., Testa, D.D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K.: End to end learning for self-driving cars. CoRR **abs/1604.07316** (2016)
7. DARPA: Explainable artificial intelligence (xai). https://www.darpa.mil/program/explainable-artificial-intelligence (2017), [Online; Retrieved 17 July 2017]
8. Donahue, J., Hendricks, L., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017)
9. Doran, D., Schulz, S., Besold, T.: What Does Explainable AI Really Mean? A New Conceptualization of Perspectives (2017)
10. Garcez, A., Besold, T., Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K., Lamb, L., Miikkulainen, R., Silver, D.: Neural-symbolic learning and reasoning:

Contributions and challenges. In: AAAI 2015 Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches. T.R. SS-15-03 (2015)
11. Hendricks, L., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. Lect. Notes Comput. Sci. **9908 LNCS**, 3–19 (2016)
12. Icarte, R., Baier, J., Ruz, C., Soto, A.: How a general-purpose commonsense ontology can improve performance of learning-based image retrieval. IJCAI pp. 1283–1289 (2017)
13. Liu, H., Singh, P.: Conceptnet — a practical commonsense reasoning toolkit. BT Technology Journal **22**(4), 211–226 (Oct 2004)
14. van Otterlo, M.: From algorithmic black boxes to adaptive white boxes: Declarative decision-theoretic ethical programs as codes of ethics. CoRR **abs/1711.06035** (2017), http://arxiv.org/abs/1711.06035
15. Raymond, J.: Integrating abduction and induction in machine learning. In: Abduction and Induction, pp. 181–191. Kluwer Academic Publishers (2000)
16. Sarker, M.K., Xie, N., Doran, D., Raymer, M., Hitzler, P.: Explaining trained neural networks with semantic web technologies: First steps. CoRR **abs/1710.04324** (2017)
17. Schmidhuber, J.: Deep learning in neural networks. Neural Netw. **61**(C), 85–117 (Jan 2015)
18. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61**, 85 – 117 (2015)
19. Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information. CoRR **abs/1703.00810** (2017), http://arxiv.org/abs/1703.00810
20. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of go with deep neural networks and tree search. Nature **529**, 484–503 (2016)
21. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. rep. (2011)
22. Xie, N., Sarker, K., Doran, D., Hitzler, P., Raymer, M.: Relating Input Concepts to Convolutional Neural Network Decisions (2016) (2017)