On Generating Symbolic Explanations for Recurrent Neural Networks
— A Position Paper —

Manuel de Sousa Ribeiro[0000−0002−5526−1043] and
João Leite[0000−0001−6786−7360]

NOVA LINCS, School of Science and Technology, NOVA University Lisbon, Portugal
mad.ribeiro@campus.fct.unl.pt, jleite@fct.unl.pt

Abstract. In this paper, we address how to generate symbolic explanations for the outputs of recurrent neural networks for real-time data-stream classification. We discuss how a recurrent neural network could be interpreted by mapping its neuronal activations to human-defined concepts from a logic-based formalism. Finally, we elaborate on how stream reasoning could be used in this setting to produce real-time symbolic explanations for a recurrent neural network's output.

Keywords: Neuro-Symbolic AI · Knowledge Representation · Machine Learning · Neural Networks · Stream Reasoning

1 Introduction

Recurrent Neural Networks (RNNs) are a class of neural network known for its effectiveness in domains involving sequential data, with successful applications spanning across multiple areas including text classification, translation, speech recognition, and music generation, to name a few [18]. RNNs are able to achieve impressive performance in this kind of tasks by performing parameter sharing over their input sequences, allowing for generalization across input sequences of different lengths, and by using feedback connections, where the output of a given unit in a model may be fed back to that or a previous unit, allowing prior input elements to influence the way latter ones are processed – simulating memory.

Despite the success of RNNs, they are still considered black boxes. On one hand, the size and complexity of these models typically renders an explanation purely based on their internal parameters, e.g., weights, biases, etc., unfeasible. On the other, these models are subsymbolic in nature, meaning that their internal representations are generally based on a high-dimensional Euclidean space, i.e., real-valued tensors, which do not possess an associated declarative meaning [9], and thus they do not provide any human-interpretable indication regarding why a given output was produced.

In this work, we discuss how to produce human-understandable explanations for the outputs of recurrent neural networks, particularly for models which are fed real-time data from a data stream, e.g, RNNs for real-time intelligent video

surveillance. In Section 2, we briefly introduce and discuss some of the existing state-of-the-art methods on interpreting RNNs. In Section 3, we analyze how to produce symbolical explanations for the outputs of RNNs, and elaborate on how this might motivate new research directions in Stream Reasoning.

2 On Interpreting Recurrent Neural Networks

Many methods have been developed with the goal of increasing the interpretability of neural networks. Current popular approaches include proxy-based methods, where the model being interpreted is substituted for one that is inherently interpretable and with similar behavior, and saliency and attribution methods, which approximate the input features' contributions for a given prediction.

Most work directed at increasing the interpretability of RNNs focuses on the development of saliency and attribution methods, where multiple popular approaches exist. Gradient-based methods, as the ones described in [14] and [17], compute the contribution of each feature based on the gradient of the output with respect to the input. Backpropagation-based methods, such as layer-wise relevance propagation [2], propagate the prediction backwards using a set of propagation rules to compute the input feature's relevancy. Perturbation-based methods, e.g., [11], estimate the features' contributions by measuring how the output changes when different parts of the input are masked.

There has also been work on developing specialized proxy-based methods for RNNs, such as LIMSSE [12], based on LIME [13], but changing the way the model being interpreted is probed, performing an order-sensitive sampling.

Although all of the above described methods do increase the interpretability of RNNs, they do so in terms of the inputs of the model being interpreted, with the explanation consisting only on the input features and their corresponding contribution values. We argue that this approach is too simplistic, and typically only useful in settings where the relationship between the networks' input and output is simple, intuitive, and well-understood, e.g., identifying familiar objects in images, or where the input is already symbolic, such as in the domain of natural language processing. This happens because this kind of explanation is solely based on the input features of a neural network, and thus, does not explicitly present any clarification of the underlying phenomena that lead a particular output to being produced from a given input, leaving the burden of understanding why those contribution values justify the network's output to the end user.

User studies [1] corroborate our argument, often finding that the explanations produced by these methods end up being ignored or unhelpful to end users.

3 Symbolic Explanations for Recurrent Neural Networks

Recently, in [16], we explored the use of ontologies as a means to provide the necessary language and background knowledge, at the appropriate level of abstraction, to adequately convey justifications for a neural network's output. To establish a mapping between the neural network and the concepts existing in an

ontology, we found inspiration in the research conducted in the field of neuroscience, where ensembles of neurons and how they respond to stimuli have been investigated to comprehend what information they encode [8]. We hypothesized that if a human-defined concept is relevant to the task of a trained neural network, then we should be able to relate it with the representations encoded in the model of that network. For instance, if a neural network was trained to identify *mixed trains*, then we should be able to relate the representations encoded in the network's model with concepts like *passenger car* and *freight wagon*, given that they are generally used to define mixed trains.

To test this hypothesis, in [16] we explored the path of establishing mappings from the activations produced in the neurons of a feedforward neural network to concepts from a chosen logic-based ontology. These mappings are established through the so-called *mapping networks*, i.e., small neural networks each built to predict a single human-defined concept from the activations of a given neural network, dubbed *main network*. Through the use of the mapping networks, when input is fed to the main network, it is possible to observe whether their corresponding concepts were identified, and thus acquire additional knowledge about the main network's input. While these mappings would allow the interpretation of a neural network's internal representations in human-understandable symbolical terms, the ontology would provide background knowledge about the domain of task of the neural network.

The outputs of the main network could then be explained by providing justifications, i.e., minimal sets of axioms from the ontology that, together with assertions made for a given input regarding each mapped concept, entail the output of the main network. These justifications can be obtained through the use of an axiom pinpointing algorithm to identify the sets of axioms in a knowledge base responsible for a given entailment, such as the one described in [10]. The justifications so obtained are symbolic and declarative, hence human-understandable, while providing a useful bridge between a neural network's sub-symbolic *behaviour* and the symbolic world of ontologies and reasoning.

The results obtained in [16] were very promising, indicating that it is possible to leverage on the knowledge existing in a neural network's model, and to use that knowledge to establish mappings to human-defined concepts. For the dataset and ontology used, the resulting justifications, were correct in 90% of the cases. Additional experiments carried out in [16] attest to the validity of the proposed method, e.g., by providing evidence that the extracted concepts were instrumental to determine the output of the network, and that the mapping networks were correctly localizing the concepts they were trained to identify.

Towards adapting the method of [16] to deal with sequential data and justify the outputs of RNNs, we performed some preliminary experiments using a video classification task. Based on the Explainable Abstract Trains Dataset [15], we built a dataset containing 60-frame animations of trains transversing the canvas from right to left, with varying speed and angle. Our experiments, performed on three different long short-term memory networks trained to identify anima-

tions containing trains with different visual characteristics indicate that mapping networks can successfully be applied to RNNs.

In this new setting, where RNNs are fed with data streams, i.e., where the input is an unbounded sequence of data elements, the outputs of both the RNNs and the mapping networks can also be seen as data streams. Unlike with feedforward neural networks, here the temporal aspect becomes central. To fully extend our method to this new setting, we need a logic-based formalism that supports the representation of knowledge with a temporal dimension – to represent the background knowledge such as, for example, that a train is of some specific type if a passenger car precedes a cargo wagon – but also reasoning mechanisms that are able to deal not only with such temporal knowledge, but also with the data streams produced by both the main RNN network and the mapping networks, to generate the justifications, which should themselves also be streams.

This is where stream reasoners come into play. Expressive stream reasoners already allow for the representation of knowledge with a temporal dimension, and can certainly deal with the kind os symbolic data streams produced by the RNNs and the mapping networks. One of the most well-known stream reasoners, C-SPARQL [3], provides a language for continuously querying data streams with the ability to incorporate background knowledge expressed in RDF which allows for the representation of ontologies. One of the state-of-the-art stream reasoners, LASER [4], builds upon LARS [5], a rule-based language that extends Answer Set Programming (ASP) [7] with temporal operators. LASER allows one to declaratively encode a query over data streams as a set of LARS logic programming rules, which can also be used to encode background knowledge.

However, for a stream reasoner to be able to deal with the problem at hand, additional work is required. Besides the adoption of optimizations to render the production of justifications in a timely manner feasible, they would have to be extended with the sort of reasoning underlying axiom pinpointing. This would certainly be the case with C-SPARQL. As for LASER, it might be the case that the expressiveness afforded by its ASP-based rules is enough [6].

4 Conclusions

As we witness novel results stemming from the research area of neural-symbolic AI, new opportunities arise for the use of older and newer logic-based reasoners, while posing new challenges and opportunities for their development. Our goal with this paper was to illustrate one such case, namely the need for expressive stream reasoners to help produce symbolic justifications for the output of RNNs, following the methods developed in [16] for feedforward neural networks.

Acknowledgments

The authors were supported by FCT grant UI/BD/151266/2021, projects FORGET PTDC/CCI-INF/32219/2017 and RIVER PTDC/CCI-COM/30952/2017, and strategic project NOVA LINCS UIDB/04516/2020.

References

1. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging Tests for Model Explanations. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS) (2020)
2. Arras, L., Montavon, G., Müller, K., Samek, W.: Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA@EMNLP). Association for Computational Linguistics (2017)
3. Barbieri, D.F., Braga, D., Ceri, S., Valle, E.D., Grossniklaus, M.: C-SPARQL: a Continuous Query Language for RDF Data Streams. International Journal of Semantic Computing (2010)
4. Bazoobandi, H.R., Beck, H., Urbani, J.: Expressive Stream Reasoning with Laser. In: The Semantic Web - 16th International Semantic Web Conference (ISWC), Proceedings. Lecture Notes in Computer Science, Springer (2017)
5. Beck, H., Dao-Tran, M., Eiter, T., Fink, M.: LARS: A Logic-Based Framework for Analyzing Reasoning over Streams. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press (2015)
6. Fandinno, J., Schulz, C.: Answering the "why" in Answer Set Programming - A Survey of Explanation Approaches. Theory Pract. Log. Program. (2019)
7. Gelfond, M., Lifschitz, V.: Classical Negation in Logic Programs and Disjunctive Databases. New Generation Computing (1991)
8. Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P.D., Maguire, E.A.: Decoding Neuronal Ensembles in the Human Hippocampus. Current Biology (2009)
9. Hitzler, P., Bianchi, F., Ebrahimi, M., Sarker, M.K.: Neural-symbolic Integration and the Semantic Web. Semantic Web (2020)
10. Horridge, M.: Justification Based Explanation in Ontologies. Ph.D. thesis, University of Manchester, UK (2011)
11. Li, J., Monroe, W., Jurafsky, D.: Understanding Neural Networks through Representation Erasure. CoRR (2016)
12. Pörner, N., Schütze, H., Roth, B.: Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics (2018)
13. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2016)
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: 2nd International Conference on Learning Representations (ICLR), Workshop Track Proceedings (2014)
15. de Sousa Ribeiro, M., Krippahl, L., Leite, J.: Explainable Abstract Trains Dataset. CoRR (2020)
16. de Sousa Ribeiro, M., Leite, J.: Aligning Artificial Neural Networks and Ontologies towards Explainable AI. In: Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21). AAAI Press (2021)
17. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, PMLR (2017)
18. Yu, Y., Si, X., Hu, C., Zhang, J.: A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Computation (2019)