On Generating Symbolic Explanations for Recurrent Neural Networks

— Extended Abstract —

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

Recurrent Neural Networks (RNNs) are a class of artificial neural network known for its effectiveness in domains involving sequential data, with successful applications spanning across multiple areas including text classification [5], translation [4], speech recognition [8], music generation [7], driver intention prediction and trajectory prediction in self-driving cars [13]. 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 inputs to influence the way latter ones are processed – simulating memory.

Despite the success of RNNs, they are still considered black boxes [9]. 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 [11], and thus they do not provide any human-interpretable indication regarding why a given output was produced.

In the last few years, many methods have been developed with the goal of increasing the interpretability of artificial neural networks. Most of the current approaches may be broadly categorized into two groups:

- Proxy-based methods: methods where the model being interpreted is substituted for one that is inherently interpretable and that behaves similarly to the original model;
- Saliency and attribution methods: methods which focus on approximating the contribution of each input feature for a given prediction.

Most of the work directed at interpretability of RNNs focuses on the development of saliency and attribution methods, where multiple different popular approaches exist. Gradient-based methods, such as the ones described in [21] and [24], 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 [3, 2] and DeepLIFT [20], propagate the prediction backwards using a set of propagation rules to compute the relevancy of each

input feature. Perturbation-based methods [25, 14] estimate the features' contributions by measuring how the output changes when different parts of the input are removed or masked. There has also been work on developing proxy-based methods specialized for RNNs, such as LIMSSE [16], this method is inspired by LIME [17] and changes the way that the model being interpreted is probed, by performing a word-order sensitive sampling.

Although such 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, such as 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, 6, 19] corroborate our argument, often finding that such explanations produced by these methods end up being ignored or unhelpful to end users.

In [23], we argued that to justify the output of a neural network, a language containing human-understandable concepts and meaningful relations between those concepts is needed, allowing for a comprehensible description of the reasoning that led the neural network to attain its output. Inspired by 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 [10], 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*, since they are generally used to define such trains.

To test this hypothesis, in [23] we explored the path of establishing mappings from the values of the activations produced by the neurons in the internal layers of a feedforward neural network, dubbed *main 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 main neural 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 the necessary language and background knowledge to adequately convey justifications for the neural network's output. Such an ontology, which may

be obtained by either adopting an existing one or creating one for this specific purpose, should contain human-understandable concepts and meaningful relations between those concepts at an appropriate level of abstraction to allow for a comprehensible symbolical description of the reasoning that led a neural network to attain its output. Using logic-based reasoning methods over the ontology, together with the observations made for a given input regarding each mapped concept, we can create a justification for the main network's output. The justifications would be minimal sets of axioms from the ontology that, together with the observations, entail the output of the main network. The justifications so obtained are symbolic and declarative, hence human-understandable, while providing a useful bridge between the sub-symbolic *behaviour* of the neural network and the symbolic world of reasoning and ontologies. It should be noted, however, that despite the justifications being based on the observations of the mapping networks, they are produced at an ontological level. Therefore, they should not be taken as a representation of how a neural network *really* reached its outputs, but rather as plausible and understandable justifications for how it might have achieved its results, based on human-understandable concepts.

The results obtained in [23] were very promising, indicating that it is possible to leverage on the knowledge existing in the architecture of a neural network and to use that knowledge to establish mappings to concepts from an ontology. For the dataset and ontology used, [23] reports that the resulting justifications, produced by using a reasoner together with the ontology and the output of the mapping networks, were correct in 90% of the cases. The built mapping networks were able to *extract* concepts that were ontologically relevant for the output of the main network with fewer examples and achieving higher accuracy values than non-relevant concepts, providing evidence that these concepts were instrumental to determine the output of the network. Further evidence that neural networks somehow encode internal representations of these (intermediate) concepts was obtained when we observed that using the internal activations of a neural network to extract concepts through the mapping networks was better than if we were to use neural networks built and trained to identify those same ontological concepts directly from the dataset's images, given that the former outperformed the latter, while being considerably smaller. In addition, using an occlusion method [25], it was possible to visually verify that the mapping networks were correctly localizing the concepts that they were trained to identify.

It is worth noting that there has been some work where human-defined concepts outside the scope of the neural networks' input were utilized to increase the interpretability of neural networks. In [12], neural networks are built to compute propositional logic programs, and thus are interpretable by design. However, it seems to miss on the neural networks' capacity to learn from examples. Other works try to integrate both neural networks and logical reasoning, such as [18, 15], either by leveraging on annotated data to learn a concept representing a neural networks' output, or by having neural networks working alongside reasoners within an elaborate architecture aimed at addressing complex cognitive robotics problems in an explainable way, a different problem from the one we

consider, which is to provide human-understandable explanations for previously independently trained neural-networks.

Our current goal is to adopt a similar method to deal with sequential data processed by Recurrent Neural Networks. To this end, we address questions such as, which concepts might be extracted by applying mapping networks on recurrent main networks, whether those concepts match our understanding, what are the costs associated with the development of these mapping networks, and how to integrate the mapping networks' outputs with a logic-based theory in order to justify the main networks' outputs. Furthermore, we investigated whether mapping networks should have access to the internal state vector of each recurrent layer of their main networks, and thus have as input both the main network's neural activations and internal state vectors. Additionally, we also considered whether the mapping networks should themselves be RNNs, due to the temporal/sequential nature of the concepts typically involved in the tasks of RNNs, and the kind of temporal features the ontology should have to enable the adequate generation of justifications.

To address such questions, we conducted preliminary experiments using a video classification task. The dataset built for this problem is based on the Explainable Abstract Trains Dataset [22] and contains 500 000 animations where a representation of a train traverses the frame from right to left, with varying speed and angle. Each animation is composed by 60 frames, where each frame has a size of 30×30 pixels, guaranteeing that the information regarding the train is distributed across different frames.

Our experiments, performed on three different long short-term memory networks, built and trained to identify animations containing trains with different visual characteristics, indicate that mapping networks can successfully be applied to RNNs. The obtained results are to some extent similar to those described in [23]. The mapping networks are typically able to extract concepts that are relevant to the task of their main networks with high accuracy values, with the extracted concepts matching our understanding of those concepts, while still requiring few training samples and using only the activations of a small percentage of the neurons of their main networks. The resulting explanations, despite requiring some filtering of the mapping networks' outputs, were typically able to correctly justify the main networks' results.

In the talk corresponding to this extended abstract, we will discuss how to generate symbolic explanations for the outputs of RNNs. We will address how mapping networks can be applied to RNNs to extract human-defined concepts from a logic-based theory, and present our preliminary results.

Acknowledgments

The authors would like to thank the support provided by FCT through PhD grant (UI/BD/151266/2021), project FORGET (PTDC/CCI-INF/32219/2017), and strategic project NOVA LINCS (UIDB/04516/2020).

References

1. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
2. Arras, L., Montavon, G., Müller, K., Samek, W.: Explaining recurrent neural network predictions in sentiment analysis. In: Balahur, A., Mohammad, S.M., van der Goot, E. (eds.) Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017. pp. 159–168. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/w17-5221
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Mller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), 1–46 (07 2015). https://doi.org/10.1371/journal.pone.0130140
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
5. Bai, X.: Text classification based on LSTM and attention. In: 2018 Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, September 24-26, 2018. pp. 29–32. IEEE (2018). https://doi.org/10.1109/ICDIM.2018.8847061
6. Chu, E., Roy, D., Andreas, J.: Are visual explanations useful? A case study in model-in-the-loop prediction. CoRR **abs/2007.12248** (2020)
7. Garoufis, C., Zlatintsi, A., Maragos, P.: An lstm-based dynamic chord progression generation system for interactive music performance. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. pp. 4502–4506. IEEE (2020). https://doi.org/10.1109/ICASSP40776.2020.9053992
8. Graves, A., Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013. pp. 6645–6649. IEEE (2013)
9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019). https://doi.org/10.1145/3236009
10. Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P.D., Maguire, E.A.: Decoding neuronal ensembles in the human hippocampus. Current biology : CB **19**(7), 546–554 (4 2009)
11. Hitzler, P., Bianchi, F., Ebrahimi, M., Sarker, M.K.: Neural-symbolic integration and the semantic web. Semantic Web **11**(1), 3–11 (2020). https://doi.org/10.3233/SW-190368
12. Hitzler, P., Hölldobler, S., Seda, A.K.: Logic programs and connectionist networks. J. Appl. Log. **2**(3), 245–272 (2004). https://doi.org/10.1016/j.jal.2004.03.002
13. Jili, F.E.: An effective driver intention and trajectory prediction for autonomous vehicle based on LSTM. In: Rocha, A.P., Steels, L., van den Herik, H.J. (eds.) Proceedings of the 13th International Conference on Agents and Artificial Intelligence,

ICAART 2021, Volume 2, Online Streaming, February 4-6, 2021. pp. 1090–1096. SCITEPRESS (2021). https://doi.org/10.5220/0010321710901096
14. Li, J., Monroe, W., Jurafsky, D.: Understanding neural networks through representation erasure. CoRR **abs/1612.08220** (2016)
15. Mota, T., Sridharan, M., Leonardis, A.: Integrated commonsense reasoning and deep learning for transparent decision making in robotics. SN Comput. Sci. **2**(4), 242 (2021). https://doi.org/10.1007/s42979-021-00573-0
16. Pörner, N., Schütze, H., Roth, B.: Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. pp. 340–350. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/P18-1032
17. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144. ACM (2016). https://doi.org/10.1145/2939672.2939778
18. Sarker, M.K., Xie, N., Doran, D., Raymer, M.L., Hitzler, P.: Explaining trained neural networks with semantic web technologies: First steps. In: Besold, T.R., d'Avila Garcez, A.S., Noble, I. (eds.) Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning, NeSy 2017, London, UK, July 17-18, 2017. CEUR Workshop Proceedings, vol. 2003. CEUR-WS.org (2017)
19. Shen, H., Huang, T.K.: How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels. CoRR **abs/2008.11721** (2020)
20. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. PMLR (2017)
21. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings (2014)
22. de Sousa Ribeiro, M., Krippahl, L., Leite, J.: Explainable abstract trains dataset. CoRR **abs/2012.12115** (2020)
23. de Sousa Ribeiro, M., Leite, J.: Aligning artificial neural networks and ontologies towards explainable AI. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 4932–4940. AAAI Press (2021)
24. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (2017)
25. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision -

ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. Lecture Notes in Computer Science, vol. 8689, pp. 818–833. Springer (2014). https://doi.org/10.1007/978-3-319-10590-1_53