

UNIVERSITY OF POTSDAM

MASTER'S THESIS

SoPa++: Leveraging explainability from hybridized RNN, CNN and weighted finite-state neural architectures

Author:

Atreya SHANKAR

1st Supervisor:

Dr. Sharid LOÁICIGA
University of Potsdam

2nd Supervisor:

Mathias MÜLLER
University of Zurich

*A thesis submitted in fulfillment of the requirements
for the degree of Cognitive Systems: Language,
Learning, and Reasoning (M.Sc.)*

in the

Foundations of Computational Linguistics Research Group
Department of Linguistics

February 24, 2021

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research questions	1
1.3	Thesis structure	2
2	Background concepts	3
	Bibliography	4

Chapter 1

Introduction

1.1 Motivation

With the recent progress of large deep learning models on a myriad of machine-learning and Natural Language Processing (NLP) tasks, several studies argue for focused research into Explainable Artificial Intelligence (XAI) to address emerging concerns such as adversarial security and inductive biases associated with black-box models (Doran, Schulz, and Besold, 2017; Townsend, Chaton, and Monteiro, 2019; Danilevsky et al., 2020; Arrieta et al., 2020). Of these studies, Arrieta et al. (2020) conduct an extensive survey into the spectrum of XAI taxonomies and provide the following definition of XAI:

*“Given an audience, an **explainable** Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.”*

In addition, Arrieta et al. (2020) explore and classify a variety of machine-learning models depending on the degree of their transparencies; as well as document taxonomies of explainability methods associated with varying degrees of model transparencies. Of particular relevance to this study is the *explanations by simplification* explainability method, which Arrieta et al. (2020) document as:

“Explanations by simplification collectively denote those techniques in which a whole new system is rebuilt based on the trained model to be explained. This new, simplified model usually attempts at optimizing its resemblance to its antecedent functioning, while reducing its complexity, and keeping a similar performance score.”

In this thesis, we build upon the work of Schwartz, Thomson, and Smith (2018) by further developing their **Soft Patterns** (SoPa) model; which represents a hybridized RNN, CNN and Weighted Finite-State Automaton (WFSA) neural network architecture. We modify the SoPa model by changing key aspects of its architecture which ultimately allow us to conduct effective explanations by simplification; which was not possible with the previous SoPa architecture. We abbreviate this modified model as **SoPa++**, which signifies an improvement or major modification to the SoPa model.

Finally, we evaluate both the performance and explainability of the SoPa++ model on the Facebook Multilingual Task Oriented Dialog data set (FMTOD; Schuster et al. 2018); focusing on the English-language intent classification task.

1.2 Research questions

With the aforementioned modifications to the SoPa architecture and the introduction of the SoPa++ architecture, we aim to answer the following three research questions:

1. To what extent does SoPa++ contribute to competitive performance¹ on the FMTOD data set?
2. To what extent does SoPa++ contribute to effective explanations by simplification, as exemplified on the FMTOD data set?
3. What interesting and relevant explanations can SoPa++ provide on the FMTOD data set?

1.3 Thesis structure

With the aforementioned research questions, we summarize the structure and contents of this thesis.

Chapter 1: We introduce this thesis, its contents and our research questions.

Chapter 2: We describe the fundamental background concepts utilized in this thesis.

Chapter 3: We describe the methodologies pursued in this thesis.

Chapter 4: We describe the results obtained from our methodologies.

Chapter 5: We discuss the implications of the aforementioned results.

Chapter 6: We conclude this thesis by answering the research questions.

Chapter 7: We document future work to expand on our research questions.

¹We define competitive performance as the scenario where a mean performance metric on a certain data set falls within the range obtained from other recent studies on the same data set

Chapter 2

Background concepts

Bibliography

- Arrieta, Alejandro Barredo et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115.
- Bocklisch, Tom et al. (2017). "Rasa: Open source language understanding and dialogue management". In: *arXiv preprint arXiv:1712.05181*.
- Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.
- Danilevsky, Marina et al. (2020). "A survey of the state of explainable AI for natural language processing". In: *arXiv preprint arXiv:2010.00711*.
- Doran, Derek, Sarah Schulz, and Tarek R. Besold (2017). "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives". In: *CoRR* abs/1710.00794. arXiv: 1710.00794. URL: <http://arxiv.org/abs/1710.00794>.
- Evans, Richard and Edward Grefenstette (2018). "Learning explanatory rules from noisy data". In: *Journal of Artificial Intelligence Research* 61, pp. 1–64.
- Hornik, Kurt, Maxwell Stinchcombe, Halbert White, et al. (1989). "Multilayer feed-forward networks are universal approximators." In: *Neural networks* 2.5, pp. 359–366.
- Hou, Bo-Jian and Zhi-Hua Zhou (2018). "Learning with Interpretable Structure from RNN". In: *CoRR* abs/1810.10708. arXiv: 1810.10708. URL: <http://arxiv.org/abs/1810.10708>.
- Jiang, Chengyue et al. (2020). "Cold-Start and Interpretability: Turning Regular Expressions into Trainable Recurrent Neural Networks". In: .
- Kepner, Jeremy et al. (2018). "Sparse deep neural network exact solutions". In: *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, pp. 1–8.
- Kuich, Werner and Arto Salomaa (1986). "Linear Algebra". In: *Semirings, automata, languages*. Springer, pp. 5–103.
- Law, Mark, Alessandra Russo, and Krysia Broda (2015). *The ILASP system for learning answer set programs*.
- Li, Shen, Hengru Xu, and Zhengdong Lu (2018). "Generalize symbolic knowledge with neural rule engine". In: *arXiv preprint arXiv:1808.10326*.
- Payani, Ali and Faramarz Fekri (2019). "Inductive Logic Programming via Differentiable Deep Neural Logic Networks". In: *CoRR* abs/1906.03523. arXiv: 1906.03523. URL: <http://arxiv.org/abs/1906.03523>.
- Peng, Hao et al. (2018). "Rational Recurrences". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1203–1214. DOI: 10.18653/v1/D18-1152. URL: <https://www.aclweb.org/anthology/D18-1152>.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). "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, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144.

- Ribeiro, Marco Túlio, Sameer Singh, and Carlos Guestrin (2016b). ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *CoRR* abs/1602.04938. arXiv: 1602.04938. URL: <http://arxiv.org/abs/1602.04938>.
- Schuster, Sebastian et al. (2018). “Cross-lingual transfer learning for multilingual task oriented dialog”. In: *arXiv preprint arXiv:1810.13327*.
- Schwartz, Roy, Sam Thomson, and Noah A. Smith (July 2018). “Bridging CNNs, RNNs, and Weighted Finite-State Machines”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 295–305. DOI: 10.18653/v1/P18-1028. URL: <https://www.aclweb.org/anthology/P18-1028>.
- Suresh, Ananda Theertha et al. (2019). “Approximating probabilistic models as weighted finite automata”. In: *CoRR* abs/1905.08701. arXiv: 1905.08701. URL: <http://arxiv.org/abs/1905.08701>.
- Townsend, Joseph, Thomas Chaton, and João M Monteiro (2019). “Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective”. In: *IEEE transactions on neural networks and learning systems* 31.9, pp. 3456–3470.
- Viterbi, Andrew (1967). “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In: *IEEE transactions on Information Theory* 13.2, pp. 260–269.
- Wan, Alvin et al. (2020). “NBDT: Neural-Backed Decision Trees”. In: *arXiv preprint arXiv:2004.00221*.
- Wang, Cheng and Mathias Niepert (2019). “State-Regularized Recurrent Neural Networks”. In: ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. *Proceedings of Machine Learning Research*. Long Beach, California, USA: PMLR, pp. 6596–6606. URL: <http://proceedings.mlr.press/v97/wang19j.html>.
- Yin, Penghang et al. (2019). “Understanding straight-through estimator in training activation quantized neural nets”. In: *arXiv preprint arXiv:1903.05662*.