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

March 2, 2021

Contents

Chapter 1

Introduction

1.1 Motivation

With the recent trend of increasingly large deep learning models achieving State-Of-The-Art (SOTA) performance on a myriad of Machine Learning (ML) tasks (Figure 1), several studies argue for focused research into Explainable Artificial Intelligence (XAI) to address emerging concerns such as security risks and inductive biases associated with such 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 the aforementioned models. Of particular relevance to this study is the *explanations by simplification* post-hoc post-hoc explainability method, which Arrieta et al. (2020) describe 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."*

Through a survey of recent literature on explanations by simplification applied in the Natural Language Processing (NLP) field, we came across several prominent studies employing techniques to simplify black-box neural networks into constituent Finite-State Automata (FSA) and/or Weighted Finite-State Automata (WFSA) (Schwartz, Thomson, and Smith, 2018; Peng et al., 2018; Suresh et al., 2019; Wang and Niepert, 2019; Jiang et al., 2020).

In this thesis, we build upon the work of Schwartz, Thomson, and Smith (2018) by further developing their **So**ft **Pa**tterns (SoPa) model; which represents a hybridized RNN, CNN and Weighted Finite-State Automaton 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.

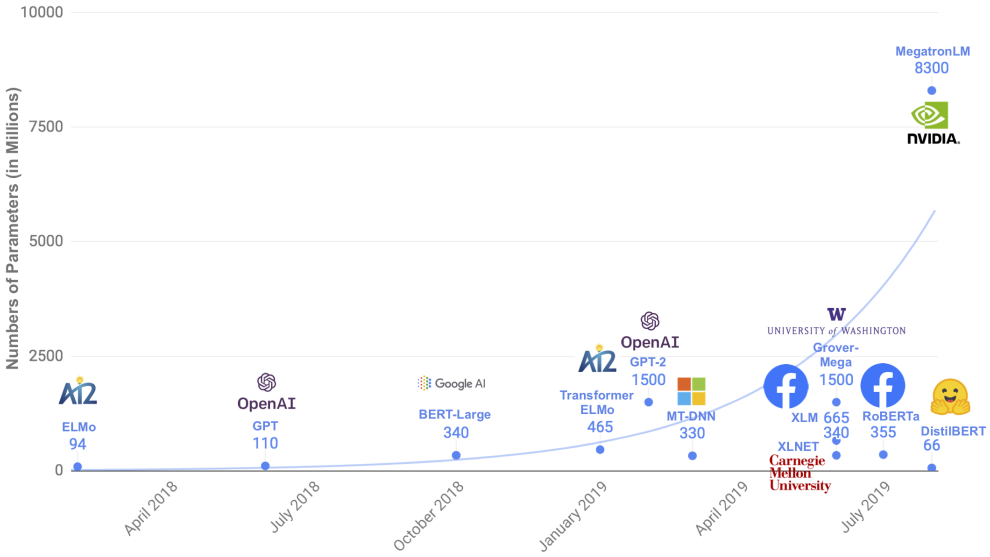

FIGURE 1: Parameter counts of recently released pre-trained language models which showed competitive or SOTA performance when fine-tuned over a range of NLP tasks (Sanh et al., 2019)

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 performance[1] on the FMTOD data set?
2. To what extent does SoPa++ contribute to effective explanations by simplification 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: Introduce this thesis, its contents and our research questions.

Chapter 2: Describe the background concepts utilized in this thesis.

Chapter 3: Describe the methodologies pursued in this thesis.

Chapter 4: Describe the results obtained from our methodologies.

Chapter 5: Discuss the implications of the aforementioned results.

Chapter 6: Conclude this thesis by answering the research questions.

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

[1] We 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

2.1 Explainable artificial intelligence

In this section, we lay out background concepts for Explainable Artificial Intelligence (XAI) which have been largely adopted from Arrieta et al. (2020). The study is particularly helpful for us since it summarizes the findings of approximately 400 XAI contributions and presents these findings in the form of well-defined concepts and taxonomies. In addition, the study discusses future directions of XAI research. We start off by providing definitions from the study, along with accompanying remarks taken either directly from the study or paraphrased for brevity.

2.1.1 Transparency

Definition 1 (Transparency)**.** A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models are divided into three categories: simulatable models, decomposable models and algorithmically transparent models.

Remark 1.1. *Simulatability* denotes the ability of a model of being simulated or thought about strictly by a human, hence complexity takes a dominant place in this class.

Remark 1.2. *Decomposability* stands for the ability to explain each of the parts of a model (input, parameter and calculation).

Remark 1.3. *Algorithmic transparency* deals with the ability of the user to understand the process followed by the model to produce any given output from its input data.

Remark 1.4. A model is considered transparent if it falls into one or more of the aforementioned transparency categories.

Remark 1.5. If a model cannot satisfy the requirements of being transparent, then it is classified as a *black-box* model.

Examples of well-known transparent Machine Learning (ML) models are linear/logistic regressors, decision trees and rules-based learners. Similarly, common examples of non-transparent or black-box ML models are tree ensembles and deep neural networks. Arrieta et al. (2020) provide extensive justifications using the aforementioned three criteria in conducting model classifications into the transparent and black-box categories. We would direct the reader to their study for a full analysis and justification of these classifications.

2.1.2 Explainability and XAI

Definition 2 (Explainability)**.** Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.

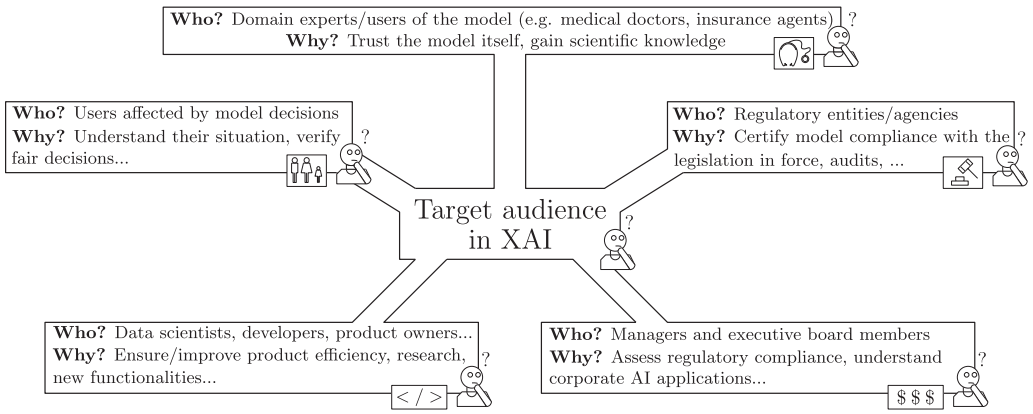

FIGURE 2: Examples of various target audiences in XAI (Arrieta et al., 2020)

Definition 3 (Explainable Artificial Intelligence)**.** Given an audience, an **explainable** Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

Arrieta et al. (2020) observe that black-box ML models are increasingly being employed to make important predictions in critical contexts, citing high-risk areas such as precision medicine and autonomous vehicles. Of particular relevance to the field of Natural Language Processing (NLP), the study notes a myriad of issues related to inductive biases within training data sets and the ethical issues involved with using black-box models trained on such data sets. As a result, they describe the increased demand for transparency in black-box ML models from the various stakeholders in Artificial Intelligence (AI). In addition, Arrieta et al. (2020) concretize the presence of a target audience for XAI; implying that different XAI techniques should be employed for different target audiences. In their study, they provide examples of target audiences such as domain experts, end-users and managers (Figure 2).

2.1.3 Explainability techniques

Based on the aforementioned classification of ML models into transparent and black-box models, Arrieta et al. (2020) expound on explainability techniques for each of these model types. Due to their transparent nature, the study states that transparent ML models are usually explainable in themselves to most target audiences and therefore usually do not require any external technique to extract explanations. The study does however highlight some target audiences, such as non-expert users, who may require external explainability techniques such as model output visualizations in order to explain the inner workings of transparent ML models.

For the case of non-transparent or black-box models, Arrieta et al. (2020) argue that separate or external techniques must be utilized in order to reasonably explain these models. Such explainability techniques are referred to in the study as post-hoc explainability techniques; which is derived from the idea that explanations for such models are usually extracted after the modelling procedure. Notable examples of post-hoc explainability techniques include local explanations, feature relevance and explanations by simplification. Below we provide definitions for these methods, which have been adapted from Arrieta et al. (2020):

Definition 4 (Local explanations)**.** Local explanations tackle explainability by segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole model.

Remark 4.1. Two well-known examples of local explainability techniques are Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro, Singh, and Guestrin 2016) and G-REX (Konig, Johansson, and Niklasson, 2008).

Definition 5 (Feature relevance)**.** Feature relevance explanation methods clarify the inner functioning of a model by computing a relevance score for its managed variables. These scores quantify the affection (sensitivity) a feature has upon the output of the model.

Remark 5.1. A well-known feature relevance explainability technique is known as the Shapley Additive Explanations (SHAP; Lundberg and Lee 2017). Another similar feature relevance explainability technique is known as the occlusion sensitivity method (Zeiler and Fergus, 2014).

Definition 6 (Explanations by simplification)**.** 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.

Remark 6.1. We hereby refer to the original black-box model as an *oracle* model and the simplified version of the model as the *proxy* model. Furthermore, we qualify that all proxy models must be able to globally approximate their respective oracle models. This is in constrast to local explanations which only approximate subsets of oracle models.

Remark 6.2. Bastani, Kim, and Bastani (2017) and Tan et al. (2018) are examples of studies that extract and distill simpler proxy models from complex oracle models.

Through a survey of recent literature on explanations by simplification applied in the Natural Language Processing (NLP) field, we came across several prominent studies employing explanations by simplification to simplify black-box neural networks into constituent Finite-State Automata (FSA) and/or Weighted Finite-State Automata (WFSA) (Schwartz, Thomson, and Smith, 2018; Peng et al., 2018; Suresh et al., 2019; Wang and Niepert, 2019; Jiang et al., 2020). We expound more on WFSA and Schwartz, Thomson, and Smith (2018) in Sections 2.3 and 2.4 respectively.

2.1.4 Key insights

2.2 Straight-through estimator

2.3 Weighted finite-state automata

Definition 7 (Semiring; Kuich and Salomaa 1986)**.** A semiring is a set $\mathbb{K}$ along with two binary associative operations $\oplus$ (addition) and $\otimes$ (multiplication) and two identity elements: $\bar{0}$ for addition and $\bar{1}$ for multiplication. Semirings require that addition is commutative, multiplication distributes over addition, and that multiplication by $\bar{0}$ annihilates, i.e., $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0}$.

Remark 7.1. Semirings follow the following generic notation: $\langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$.

Remark 7.2. A simple and common semiring is the real or sum-product semiring: $\langle \mathbb{R}, +, \times, 0, 1 \rangle$. Two important semirings for this thesis are shown below.

Remark 7.3. **Max-sum** semiring: $\langle \mathbb{R} \cup \{-\infty\}, \max, +, -\infty, 0 \rangle$

Remark 7.4. **Max-product** semiring: $\langle \mathbb{R}_{>0} \cup \{-\infty\}, \max, \times, -\infty, 1 \rangle$

Definition 8 (Weighted finite-state automaton; Peng et al. 2018)**.** A weighted finite-state automaton over a semiring $\mathbb{K}$ is a 5-tuple $\mathcal{A} = \langle \Sigma, \mathcal{Q}, \mathcal{T}, \lambda, \rho \rangle$, with:

- a finite input alphabet Σ;
- a finite state set $\mathcal{Q}$;
- transition weights $\mathcal{T} : \mathcal{Q} \times \mathcal{Q} \times (\Sigma \cup \{\epsilon\}) \to \mathbb{K}$;
- initial weights $\lambda : \mathcal{Q} \to \mathbb{K}$;
- and final weights $\rho : \mathcal{Q} \to \mathbb{K}$.

Remark 8.1. $\epsilon \notin \Sigma$ refers to special ϵ-transitions that may be taken without consuming any input.

Remark 8.2. Self-loop transitions in $\mathcal{A}$ refer to special transitions which consume an input while staying at the same state.

Remark 8.3. Σ^* refers to the (possibly infinite) set of all strings over the alphabet Σ.

Definition 9 (Path score; Peng et al. 2018)**.** Let $\boldsymbol{\pi} = \langle \pi_1, \pi_2, \ldots, \pi_n \rangle$ be a sequence of adjacent transitions in $\mathcal{A}$, with each $\pi_i = \langle q_i, q_{i+1}, z_i \rangle \in \mathcal{Q} \times \mathcal{Q} \times (\Sigma \cup \{\epsilon\})$. The path $\boldsymbol{\pi}$ derives the ϵ-free string $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle \in \Sigma^*$; which is a substring of the ϵ-containing string $\boldsymbol{z} = \langle z_1, z_2, \ldots, z_n \rangle \in (\Sigma \cup \{\epsilon\})^*$. $\boldsymbol{\pi}$'s score in $\mathcal{A}$ is given by:

$$\mathcal{A}[\boldsymbol{\pi}] = \lambda(q_1) \otimes \left(\bigotimes_{i=1}^{n} \mathcal{T}(\pi_i) \right) \otimes \rho(q_{n+1}) \tag{1}$$

Definition 10 (String score; Peng et al. 2018)**.** Let $\Pi(\boldsymbol{x})$ denote the set of all paths in $\mathcal{A}$ that derive $\boldsymbol{x}$. Then the string score assigned by $\mathcal{A}$ to string $\boldsymbol{x}$ is given by:

$$\mathcal{A}[\![\boldsymbol{x}]\!] = \bigoplus_{\boldsymbol{\pi} \in \Pi(\boldsymbol{x})} \mathcal{A}[\boldsymbol{\pi}] \tag{2}$$

Remark 10.1. Since $\mathbb{K}$ is a semiring, $\mathcal{A}[\![\boldsymbol{x}]\!]$ can be efficiently computed using the Forward algorithm (Baum and Petrie, 1966). Its dynamic program is summarized below without ϵ-transitions for simplicity. $\Omega_i(q)$ gives the aggregate score of all paths that derive the substring $\langle x_1, x_2, \ldots, x_i \rangle$ and end in state q:

$$\Omega_0(q) = \lambda(q) \tag{3a}$$

$$\Omega_{i+1}(q) = \bigoplus_{q' \in \mathcal{Q}} \Omega_i(q') \otimes \mathcal{T}(q', q, x_i) \tag{3b}$$

$$\mathcal{A}[\![\boldsymbol{x}]\!] = \bigoplus_{q \in \mathcal{Q}} \Omega_n(q) \otimes \rho(q) \tag{3c}$$

Remark 10.2. The Forward algorithm can be generalized to any semiring (Eisner, 2002) and has a runtime of $O(|Q|^3 + |Q|^2|\boldsymbol{x}|)$ (Schwartz, Thomson, and Smith, 2018); notably with a linear runtime with respect to the length of the input string $\boldsymbol{x}$.

Remark 10.3. A special case of Forward is the Viterbi algorithm, where the addition $\oplus$ operator is contrained to the maximum function (Viterbi, 1967). Viterbi therefore returns the highest scoring path $\boldsymbol{\pi}$ that derives the input string $\boldsymbol{x}$.

2.4 Soft patterns

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.

Bastani, Osbert, Carolyn Kim, and Hamsa Bastani (2017). "Interpretability via model extraction". In: *arXiv preprint arXiv:1706.09773*.

Baum, Leonard E and Ted Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains". In: *The annals of mathematical statistics* 37.6, pp. 1554–1563.

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.

Eisner, Jason (2002). "Parameter estimation for probabilistic finite-state transducers". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 1–8.

Jiang, Chengyue et al. (2020). "Cold-start and Interpretability: Turning Regular Expressions into Trainable Recurrent Neural Networks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3193–3207.

Konig, Rikard, Ulf Johansson, and Lars Niklasson (2008). "G-REX: A versatile framework for evolutionary data mining". In: *2008 IEEE International Conference on Data Mining Workshops*. IEEE, pp. 971–974.

Kuich, Werner and Arto Salomaa (1986). "Linear Algebra". In: *Semirings, automata, languages*. Springer, pp. 5–103.

Lundberg, Scott and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874*.

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 (2016). ""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.

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108*.

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.

Tan, Sarah et al. (2018). "Distill-and-compare: Auditing black-box models using transparent model distillation". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 303–310.

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.

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.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, pp. 818–833.