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

M.Sc. Thesis Defense

Atreya Shankar (799227), shankar.atreya@gmail.com
Cognitive Systems: Language, Learning, and Reasoning (M.Sc.)
1st Supervisor: Dr. Sharid Loáiciga, University of Potsdam
2nd Supervisor: Mathias Müller, M.A., University of Zurich

Foundations of Computational Linguistics
Department of Linguistics
University of Potsdam, SoSe 2021

July 8, 2021

Overview

Motivation

- Trend of increasingly complex deep learning models achieving SOTA performance on ML and NLP tasks (Figure 1)
- To address emerging concerns such as inductive biases, several studies make arguments for research into XAI; for example Danilevsky et al. (2020) and Arrieta et al. (2020)
- Schwartz et al. (2018) approach XAI in NLP by proposing an explainable hybridized neural architecture called **Soft Patterns** (SoPa; Figure 2)
- SoPa provides **localized** and **indirect** explainability despite being suited for globalized and direct **explanations by simplification**

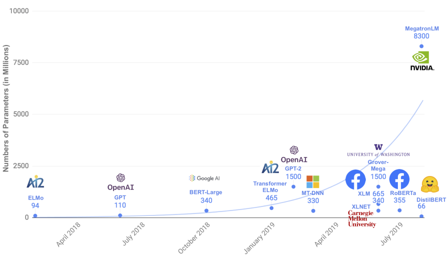

Figure 1: Parameter counts of recently released pre-trained language models; figure taken from Sanh et al. (2019)

SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines

Roy Schwartz*◇♡ Sam Thomson*♣ Noah A. Smith◇
◇Paul G. Allen School of Computer Science & Engineering, University of Washington
♣Language Technologies Institute, Carnegie Mellon University
♡Allen Institute for Artificial Intelligence
{roysch,nasmith}@cs.washington.edu, sthomson@cs.cmu.edu

Figure 2: Excerpt from Schwartz et al. (2018)

Objective and research questions

Objective:

- Address limitations of SoPa by proposing **SoPa++**, which could allow for effective explanations by simplification.

Process:

- We study the performance and explanations by simplification of SoPa++ on the **FMTOD** data set from Schuster et al. (2019); focusing on the English-language intent classification task.

Research questions:

1. Does SoPa++ provide **competitive** performance?
2. To what extent does SoPa++ contribute to **effective** explanations by simplification?
3. What **interesting and relevant** explanations can SoPa++ provide?

Overview

Explainability

- Transparency is a passive feature that a model exhibits
- Explainability is an active feature that involves target audiences (Figure 3)
- Arrieta et al. (2020) explore a taxonomy of post-hoc explainability techniques
- Prominent explainability techniques include local explanations, feature relevance and **explanations by simplification**
- Explainability techniques can provide meaningful insights into decision boundaries within black-box models (Figure 4)

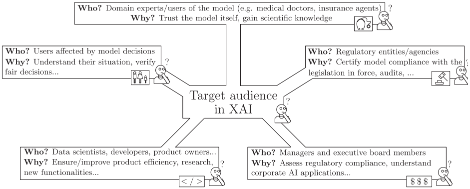

Figure 3: Examples of various target audiences in XAI; figure taken from Arrieta et al. (2020)

(a) Husky classified as wolf (b) Explanation

Figure 4: Local explanation for "Wolf" classification decision, figure taken from Ribeiro et al. (2016)

SoPa: Weighted Finite-State Automaton (WFA)

Definition 1 (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., $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0}$.

- Semirings follow the following generic notation: $\langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$.
- **Max-sum** semiring: $\langle \mathbb{R} \cup \{-\infty\}, \max, +, -\infty, 0 \rangle$
- **Max-product** semiring: $\langle \mathbb{R}_{>0} \cup \{-\infty\}, \max, \times, -\infty, 1 \rangle$

Definition 2 (Weighted finite-state automaton; Peng et al. 2018)

A weighted finite-state automaton over a semiring $\mathbb{K}$ is a 5-tuple $\mathcal{A} = \langle \boldsymbol{\Sigma}, \mathcal{Q}, \boldsymbol{\Gamma}, \boldsymbol{\lambda}, \boldsymbol{\rho} \rangle$, with:

- a finite input alphabet $\boldsymbol{\Sigma}$;
- a finite state set $\mathcal{Q}$;
- transition matrix $\boldsymbol{\Gamma} : \mathcal{Q} \times \mathcal{Q} \times (\boldsymbol{\Sigma} \cup \{\epsilon\}) \rightarrow \mathbb{K}$;
- initial vector $\boldsymbol{\lambda} : \mathcal{Q} \rightarrow \mathbb{K}$;
- and final vector $\boldsymbol{\rho} : \mathcal{Q} \rightarrow \mathbb{K}$.

SoPa: Computational graph

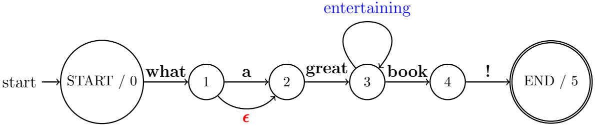

Figure 5: Linear-chain NFA with self-loop (blue), ϵ (red) and main-path (black) transitions; figure adapted from Schwartz et al. (2018)

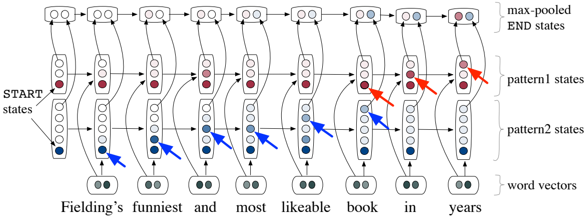

Figure 6: SoPa's partial computational graph; figure taken from Schwartz et al. (2018)

SoPa: Post-hoc explainability techniques

- SoPa provides two post-hoc explainability techniques; namely **local explanations** and **feature relevance**
- Local explanations gather highest scoring phrases across the training data (Figure 7)
- Feature relevance perturbs inputs using an occlusion technique to determine the highest impact phrases for a classification decision (Figure 8)
- Overall, both techniques are **localized** and **indirect**
- WFAs have a rich theoretical background which can be exploited for more direct and globalized explanations

	Highest Scoring Phrases				
Patt. 1	thoughtful	,	reverent	portrait	of
	and	astonishingly	articulate	cast	of
	entertaining	,	thought-provoking	film	with
	gentle	,	mesmerizing	portrait	of
	poignant	and	uplifting	story	in
Patt. 2	's	ϵ	uninspired	story	.
	this	ϵ	bad	on	purpose
	this	ϵ	leaden	comedy	.
	a	ϵ	half-assed	film	.
	is	ϵ	clumsy $,_{SL}$	the	writing

Figure 7: Ranked local explanations from SoPa; table taken from Schwartz et al. (2018)

Analyzed Documents
it 's dumb , **but more importantly** , *it 's just not scary*
though moonlight mile is replete with **acclaimed actors and actresses** and tackles a subject that 's **potentially moving** , the movie is *too predictable* and *too self-conscious to reach a* level of **high drama**
While **its careful pace and** seemingly *opaque story* may not satisfy every moviegoer 's appetite, the film 's final scene is **soaringly , transparently moving**

Figure 8: Feature relevance outputs from SoPa; table taken from Schwartz et al. (2018)

Overview

Overview

Overview

Overview

Overview

Bibliography I

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115.

Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen (Dec. 2020). "A Survey of the State of Explainable AI for Natural Language Processing". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 447–459. URL: https://www.aclweb.org/anthology/2020.aacl-main.46.

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

Peng, Hao, Roy Schwartz, Sam Thomson, and Noah A. Smith (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.

Bibliography II

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, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *NeurIPS EMC^2 Workshop*.

Schuster, Sebastian, Sonal Gupta, Rushin Shah, and Mike Lewis (June 2019). “Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3795–3805. DOI: 10.18653/v1/N19-1380. URL: https://www.aclweb.org/anthology/N19-1380.

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.