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

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Motivation

- Trend of increasingly complex deep learning models achieving SOTA performance on ML and NLP tasks, as per Figure 1
- To address emerging concerns such as security risks and inductive biases, several studies make argument for research into XAI (Arrieta et al., 2020; Danilevsky et al., 2020)
- Schwartz, Thomson, and Smith (2018) approach XAI in NLP by proposing an explainable hybridized RNN, CNN and WFA neural architecture called **So**ft **Pa**tterns (SoPa)
- SoPa provides localized and indirect explainability despite being suited for **globalized and direct** explanations by simplification

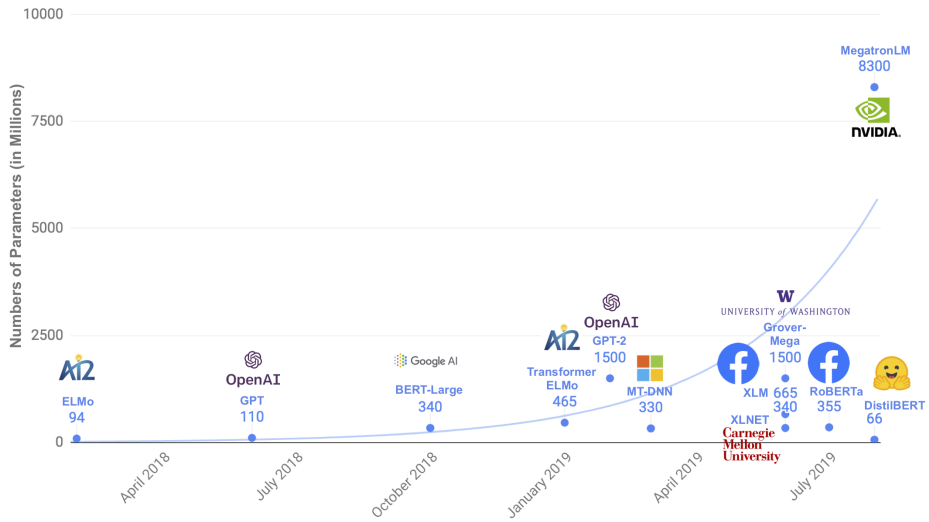

Figure 1: Parameter counts of recently released pre-trained language models; figure taken from Sanh et al. (2019)

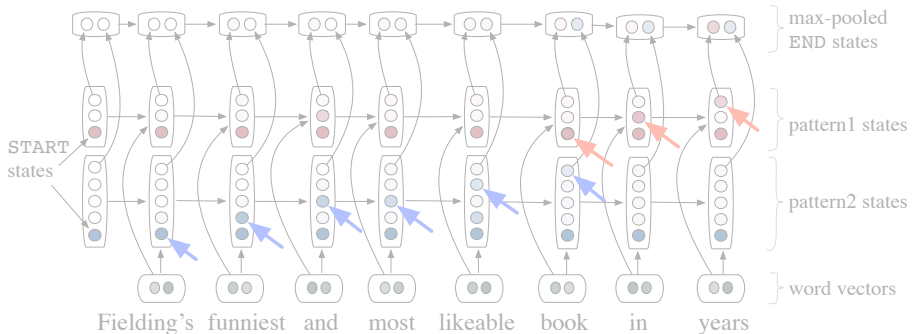

Figure 2: SoPa's partial computational graph; figure taken from Schwartz, Thomson, and Smith (2018)

Motivation

- Trend of increasingly complex deep learning models achieving SOTA performance on ML and NLP tasks, as per Figure 1
- To address emerging concerns such as security risks and inductive biases, several studies make argument for research into XAI (Arrieta et al., 2020; Danilevsky et al., 2020)
- Schwartz, Thomson, and Smith (2018) approach XAI in NLP by proposing an explainable hybridized RNN, CNN and WFA neural architecture called **So**ft **Pa**tterns (SoPa)
- SoPa provides localized and indirect explainability despite being suited for **globalized and direct** explanations by simplification

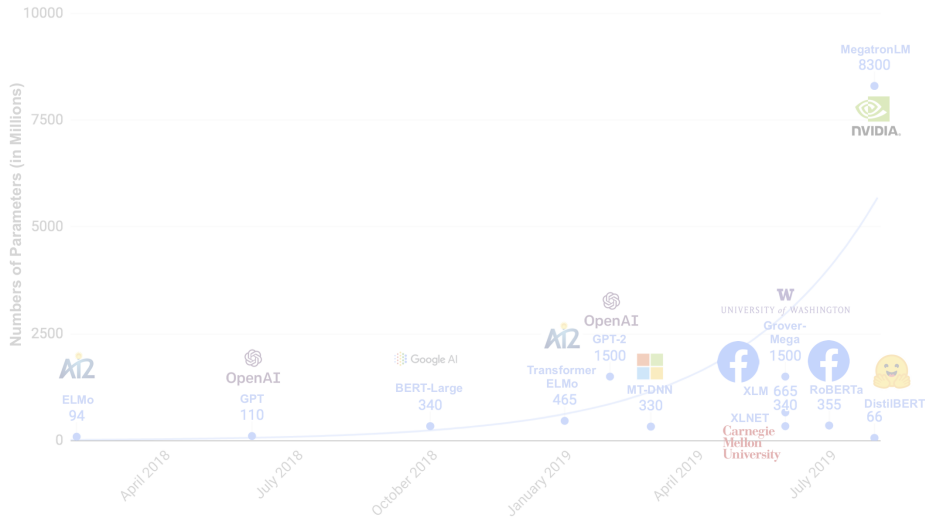

Figure 1: Parameter counts of recently released pre-trained language models; figure taken from Sanh et al. (2019)

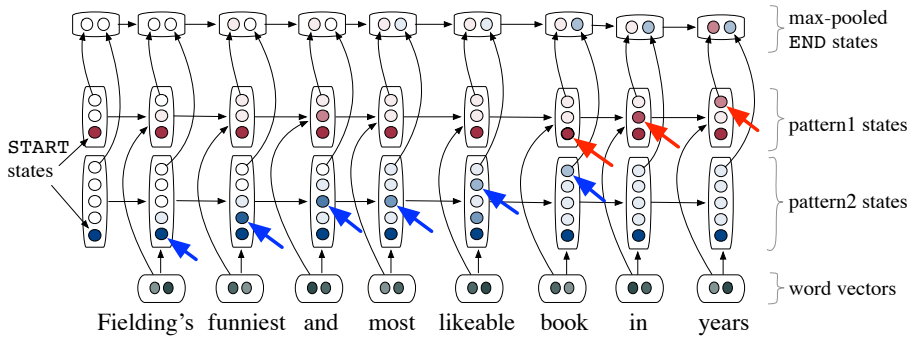

Figure 2: SoPa's partial computational graph; figure taken from Schwartz, Thomson, and Smith (2018)

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Overview

1 Introduction

2 Background concepts

3 Data and methodologies

4 Results

5 Discussion

6 Conclusions

7 Future work

Bibliography I

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.

Danilevsky, Marina et al. (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.

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *NeurIPS EMC^2 Workshop*.

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.