We propose a library and methods to enhance interpretability of time series foundation models

Wait a MOMENT, what do you know?
Interrogating Time Series Foundation Models

Michał Wiliński[1 2], Mononito Goswami[1], Chi-En Teh[1], Artur Dubrawski[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
[2] Poznan University of Technology, Poznań, Poland

Motivation

Foundation models are a leading paradigm in terms of **streamlined model deployment.**

This calls for the **inspection** of those models and how they are actually performing certain tasks.

Methods

The base for this research is interpretability work done in the last decade for **large vision & language models**

We adopted multiple methods to work with time series foundation models including:

- **representation analysis**
- **mechanistic interpretability**
- **machine learning explainability**

Results

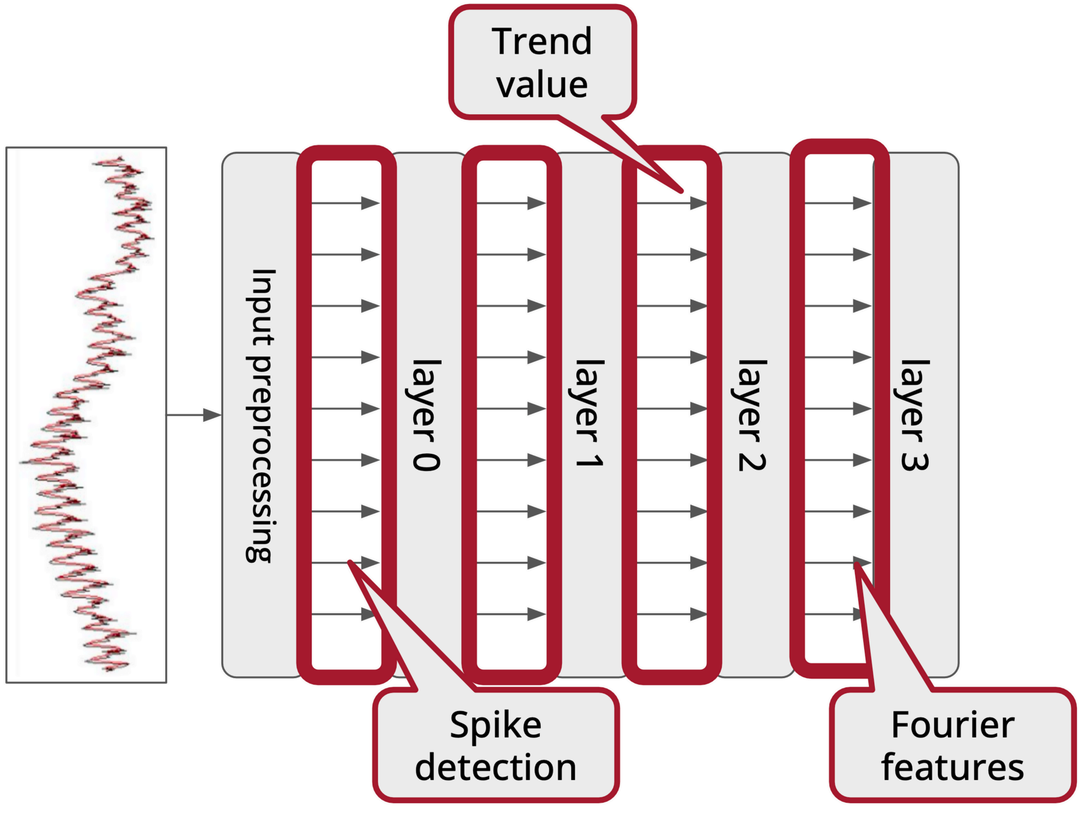
To demonstrate our methods we performed 3 studies:

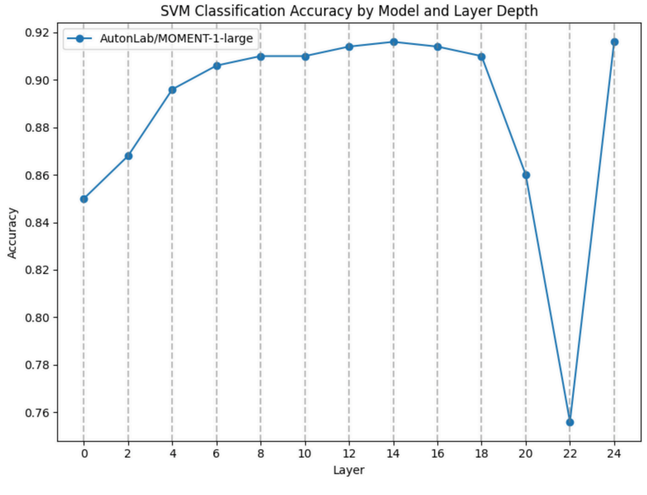
- **representation similarity**
- **task performance**
- **logit lens**

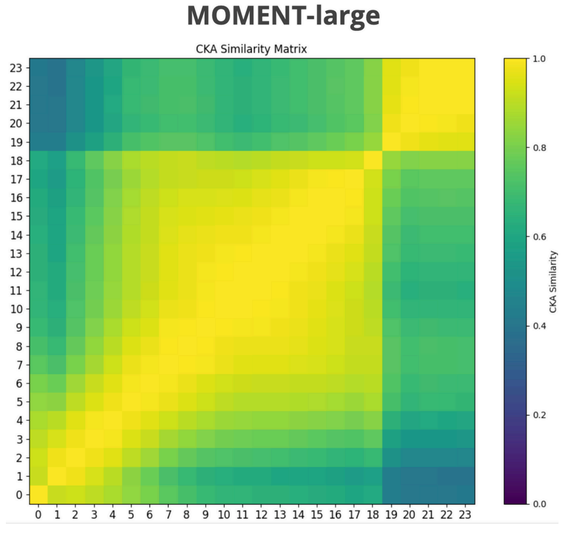
Conclusions

The interpretability of foundation models is still in its early days.

The results of our studies show that there's a lot of promise in this field and this work could help in **the wider adoption of AI models.**







Future work

Currently, the team is working on novel methods based on representation analysis which are suited also for other modalities.

Based on our research, it is possible to suggest changes in how foundation models for time series are designed.

Acknowledgements

This section will be added at the very last minute. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud.

See more:

References

[1] Goswami, M., Szafer, K., Choudhry, A.*, Cai, Y., Li, S., & Dubrawski, A. (2024). MOMENT: A Family of Open Time-series Foundation Models. In International Conference on Machine Learning. PMLR.*

[2] Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, May). Similarity of neural network representations revisited. In International conference on machine learning (pp. 3519-3529). PMLR.

[3] Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842-866.

[4] Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2024). Leace: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems, 36.