Wait a MOMENT, What Do You Know? Interrogating Time Series Foundation Models

Michał Wiliński
Auton Lab, Robotics Institute, Carnegie Mellon University
Poznan University of Technology
Poznan, Poland
mwilinsk@andrew.cmu.edu

Mononito Goswami
Auton Lab, Robotics Institute, Carnegie Mellon University
Pittsburgh, USA
mgoswami@andrew.cmu.edu

Nina Żukowska
Auton Lab, Robotics Institute, Carnegie Mellon University
Poznan University of Technology
Poznań, Poland
nzukowsk@andrew.cmu.edu

Chi-En Teh
Auton Lab, Robotics Institute, Carnegie Mellon University
Pittsburgh, USA
cteh@andrew.cmu.edu

Artur Dubrawski
Auton Lab, Robotics Institute, Carnegie Mellon University
Pittsburgh, USA
awd@cs.cmu.edu

***Abstract*—Time series foundation models have significantly impacted the machine learning community, demonstrating notable adaptability across various downstream tasks. These models offer easy-to-use and powerful tools for analyzing and predicting temporal data across diverse applications, from finance to healthcare. However, large scale, in terms of both parameters and training data, poses challenges in evaluating their limitations and determining appropriate use cases. The rapid evolution of this field raises questions about the optimal architecture and training data composition, as well as the mechanisms employed for time series processing. Furthermore, to responsibly deploy these models in real-world scenarios, it is crucial to understand their limitations and potential failure points, enabling stakeholders to make informed decisions.**

To address these knowledge gaps, we introduce Time Series Interrogator, a novel framework for analyzing time series foundation models, leveraging representation analysis and mechanistic interpretability techniques. Our study applies it to multiple publicly available models, offering a comparative analysis of their learned representations and underlying processes. This approach not only provides insights into the workings of time series foundation models, but also paves the way for more informed model development and application. By elucidating these complex systems, our work contributes to the responsible advancement of time series analysis, enabling researchers and practitioners to harness the full potential of foundation models while understanding their inherent strengths and limitations.

***Index Terms*—foundation models, time series, interpretability, model analysis**

I. Introduction

The foundation model paradigm has significantly advanced the field of machine learning, introducing a new class of models that demonstrate adaptability and performance across a wide range of domains and tasks. These models leverage large-scale architectures and pretraining data to learn representations of a wide variety concepts within a given modality. This trend is particularly evident in natural language processing (NLP) [1] and computer vision [2].

One such modality is time series data, which captures temporal dynamics and is prevalent in fields such as finance [3], healthcare [4], and climate modeling [5] among others. Applying foundation models to time series data has enabled a seamless plug-and-play approach, achieving strong predictive performance even for non-specialized stakeholders [6]–[10].

Despite the widespread adoption of foundation models in various domains, their internal workings remain largely opaque. Previous research has focused on interpreting large language [11] and vision [12] models. However, to the best of our knowledge, there has been no comprehensive study on the interpretability of time series foundation models.

In this paper, we propose a suite of three studies to address this gap:

- **Representation Analysis:** Investigating the hierarchy and similarity of internal representations.
- **Performance Comparison:** Evaluating the performance of different components of model representations.
- **LogitLens Study:** Analyzing which parts of the model are responsible for specific operations and assessing whether the model generalizes effectively.

By providing a deeper understanding of the internal mechanisms of time series foundation models, our research aims to enhance their interpretability and inform the development of more effective and reliable models for time series analysis. Additionally, we hope to facilitate the responsible application

Funded by Auton Lab, Carnegie Mellon University

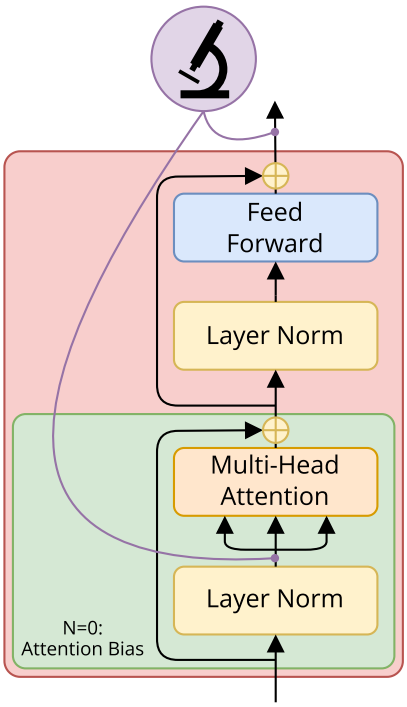

Fig. 1. Activation probing method, we take activations from the outputs of transformer blocks and layer norms of the next transformer blocks.

of these models in critical domains, ensuring that stakeholders can deploy them with greater confidence and understanding.

II. Related Work

A. Time Series Foundation Models

The introduction of transformer-based foundation models for time series data began with TimeGPT-1 [13], which utilized an encoder-decoder architecture pretrained on a collection of publicly available time series datasets, encompassing over 100 billion data points. Although the weights of this model and the specific datasets used in pretraining are not publicly available, the model itself can be accessed through an API interface.

The first fully open time series foundation model was introduced by [6], which was pretrained on time series reconstruction tasks. The authors provided comprehensive resources, including inference and research code, model weights, and the pretraining data. This model was also based on the T5 architecture, employing an encoder-decoder framework.

Following these foundational works, several other studies have emerged, proposing various approaches to the design of foundation models for time series data. These approaches vary both in terms of architectural ideas and the specific pretraining tasks and setups employed. Notable contributions include [7]–[10], each presenting unique methodologies and design choices for creating foundation models in this domain.

Recent advancements in time series foundation models have also explored the integration of large language models (LLMs) with time series data. For instance, [14] provides a comprehensive survey of these approaches, highlighting the potential of LLMs in enhancing time series analysis and forecasting.

B. Analyzing Deep Learning Models

Deep learning often functions as a black-box technique, where the underlying mechanisms and representations within the models are not well understood. One method to gain insights into these models is through the comparison of intermediate representations. This approach uses similarity metrics to determine how similar or dissimilar representations are at different stages of a model, providing insights into the hierarchy, homogeneity, and repeatability of learned features.

Work by [15] introduced novel techniques for analyzing and comparing representations in vision transformers and CNNs, offering valuable insights into the functioning of these models. Similarly, [16] explored how varying the depth and width of different models impacts similarity patterns, providing insights into the effects of model size and training data ratios.

C. Interpretability of Foundation Models

The interpretability of foundation models is an increasingly important area of research, given their growing influence and application across various domains. [17] provides a comprehensive overview of the opportunities and challenges in interpreting and understanding foundation models. In [18], a variety of analysis techniques for transformer-based language models are presented, establishing a well-defined taxonomy and detailing multiple popular techniques.

III. Methodology

To facilitate various experimental setups, we developed **Time Series Interrogator** a library that serves as the foundation for all our experiments. This library includes integration of multiple models: MOMENT, Chronos-T5, and MOIRAI. Users can easily add more models and propose custom methods for inspecting or intervening in different types of models.

Our experiments focused on processing intermediate representations in multiple different ways. The proposed probing setup is illustrated in Fig. 1. We conducted three main studies to analyze these representations and derive meaningful insights.

A. Similarity Analysis of Intermediate Representations

The first study investigates the similarity between different layers within the same model, among models from the same family, and across models from different families. To avoid bias from a single metric, we used three different metrics:

a) Cosine Similarity: Cosine similarity measures the cosine of the angle between two vectors, providing a simple yet effective way to assess similarity. In our case, we work with activation matrices for multiple samples and compute the average cosine similarity. Given two matrices $\mathbf{X}$ and $\mathbf{Y}$, representing the activations from two layers, the cosine similarity is computed as:

$$\text{cosine\_similarity}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \cdot \mathbf{y}_i}{\|\mathbf{x}_i\| \|\mathbf{y}_i\|} \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the i-th columns of matrices $\mathbf{X}$ and $\mathbf{Y}$, and n is the number of samples.

b) Singular Vector Canonical Correlation Analysis (SVCCA) [19]: SVCCA is a method that compares the similarity of representations by aligning subspaces spanned by the top singular vectors. It effectively reduces the dimensionality and then compares the correlations of the principal components. The SVCCA similarity between two activation matrices $\mathbf{X}$ and $\mathbf{Y}$ is computed as follows:

$$\text{SVCCA}(\mathbf{X}, \mathbf{Y}) = \text{CCA}(\mathbf{U}_k, \mathbf{V}_k) \tag{2}$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ are the top k singular vectors obtained from the singular value decomposition (SVD) of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and CCA denotes the canonical correlation analysis.

c) Centered Kernel Alignment (CKA) [20]: CKA measures the similarity of representations by comparing the centered kernel matrices. It has been shown to be effective in capturing similarities between layers in deep networks. The general form of CKA between two sets of representations $\mathbf{X}$ and $\mathbf{Y}$ is defined as:

$$\text{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\text{HSIC}(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{HSIC}(\mathbf{X}, \mathbf{X}) \cdot \text{HSIC}(\mathbf{Y}, \mathbf{Y})}} \tag{3}$$

where HSIC denotes the Hilbert-Schmidt Independence Criterion. In our study, we use CKA with a linear kernel, which simplifies the computation. For linear kernel, CKA is computed as:

$$\text{CKA}_{\text{linear}}(\mathbf{X}, \mathbf{Y}) = \frac{\|\mathbf{X}^T\mathbf{Y}\|_F^2}{\|\mathbf{X}^T\mathbf{X}\|_F \cdot \|\mathbf{Y}^T\mathbf{Y}\|_F} \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

B. Contextualized Time Series Representation

The second study examined the utility of intermediate representations as contextualized time series representations. We trained models on different datasets using these representations and evaluated their predictive performance across various layers. This study aimed to identify which layers provided the most informative representations for downstream tasks and to identify interesting layers for further research.

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i$ is the input time series data and y_i is the corresponding label, we passed the data through a model f with L layers. Let $\mathbf{h}_i^{(l)}$ denote the intermediate representation of the i-th sample at layer l. We then trained a classifier g_l, such as a Support Vector Machine (SVM), on these intermediate representations for each layer. The classifier g_l is trained using the dataset $\mathbf{H}^{(l)} = \{(\mathbf{h}_i^{(l)}, y_i)\}_{i=1}^N$. The performance of each classifier is then evaluated using a suitable metric, such as accuracy or F1 score.

This process can be described as follows:

1. Obtain intermediate representations for each layer:

$$\mathbf{H}^{(l)} = f^{(l)}(\mathbf{X}) \tag{5}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ is the input data matrix, and $f^{(l)}$ denotes the function representing the model up to layer l.

2. Train a classifier g_l on the intermediate representations:

$$g_l = \text{train\_classifier}(\mathbf{H}^{(l)}, \mathbf{y}) \tag{6}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_N]$ is the label vector.

3. Evaluate the performance of the classifier g_l on a validation set to determine the most informative layers for downstream tasks.

C. Feature Attribution and Analysis

Inspired by [21], our third study aimed to assess which parts of the model were responsible for processing different time series features: trend, seasonality, and frequency. Given the absence of an output embedding layer and tokenizer/dictionary, we adopted a novel approach to analyze the model's feature processing capabilities.

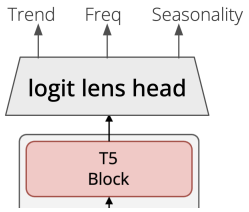

Fig. 2. Illustration of the logit lens head integrated into the MOMENT model. The logit lens head is designed to extract and interpret trend, frequency, and seasonality components from the model's outputs, providing a detailed understanding of the temporal dynamics captured by the T5 blocks.

We trained a special classification head for each feature (trend, seasonality, and frequency) on the whole model. These classification heads were then reused to work with the intermediate representations, enabling a detailed analysis of the contributions of different model components to specific time series features.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ be the input time series data matrix. For each sample $\mathbf{x}_i$, the intermediate representation at layer l is $\mathbf{h}_i^{(l)}$. We denote the true values of the features (trend, seasonality, and frequency) for the i-th sample as t_i, s_i, $and\text{f}_i$, respectively.

Initially, we trained three separate classification heads, one for each feature, on the output of the entire model. The classification heads are denoted as g_t, g_s, and g_f, corresponding to trend, seasonality, and frequency, respectively. These classification heads are trained using the dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where y_i represents the target feature (either t_i, s_i, or f_i).

The classification heads are trained to minimize the following loss function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \ell(g(\mathbf{x}_i), y_i) \tag{7}$$

where $\ell(\cdot, \cdot)$ denotes the loss function, such as cross-entropy loss for classification, and g represents the classification head corresponding to the feature of interest.

After training the classification heads on the whole model, we reused them to evaluate the intermediate representations. For each layer l, we used the classification heads g_t, g_s, and

g_f to predict the values of the features from the intermediate representations $\mathbf{h}_i^{(l)}$. The performance of these predictions was evaluated to determine which layers contribute most significantly to the processing of each feature.

This process can be summarized mathematically as follows: For each layer l and each feature $y_i \in \{t_i, s_i, f_i\}$, we compute the prediction $\hat{y}_i^{(l)}$ using the trained classification head:

$$\hat{y}_i^{(l)} = g(\mathbf{h}_i^{(l)}) \tag{8}$$

We then evaluate the performance of these predictions on a validation set to identify the layers responsible for processing trend, seasonality, and frequency, providing valuable insights into the feature processing capabilities of the model.

These three studies provide a comprehensive analysis of the internal mechanisms of time series foundation models, offering insights into their representations, performance, and feature processing capabilities.

IV. Results

Our experimental studies yielded several insights into the behavior and capabilities of time series foundation models. We summarize the key findings from each of the three main studies: similarity analysis of intermediate representations, contextualized time series representation, and feature attribution and analysis.

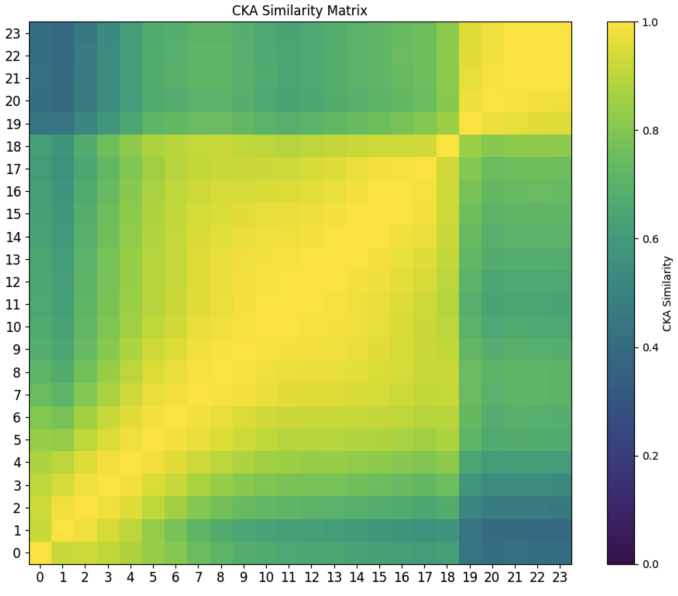

Fig. 3. CKA similarity matrix showing intra-layer similarity patterns across different layers of the MOMENT model. The consistent high-similarity blocks indicate that scaling the model preserves the similarity structure within layers.

A. Similarity Analysis of Intermediate Representations

a) Insight 1: Layer-wise Similarity Patterns: Our similarity analysis revealed distinct patterns in the similarity metrics across layers within the same model, among models from the same family, and across different model families. Specifically, we observed that:

- **Within the same model:** Visible block structure (Fig. 3, scalable with increasing model size as seen in Fig. 5, indicating hierarchy of representation inside of the model.
- **Among models from the same family:** Layers at similar depths show higher similarity, suggesting that models with similar architectures encode information in around the same parts of the network. Decrease in absolute values of the similarity suggest that different model sizes exhibit different feature learning dynamics as visible in Fig. 6.

b) Insight 2: Metric Robustness: The use of multiple metrics (cosine similarity, SVCCA, and CKA) provided a robust assessment of similarity. For instance:

- **Cosine Similarity:** Captured the overall alignment of activation vectors but was sensitive to scaling and because of comparison in feature space, the dimensionality of the models had to be the same.
- **SVCCA:** Effectively reduced dimensionality and provided a nuanced view of subspace alignment.
- **CKA:** Was particularly effective in comparing deep layers, emphasizing the independence of different representations.

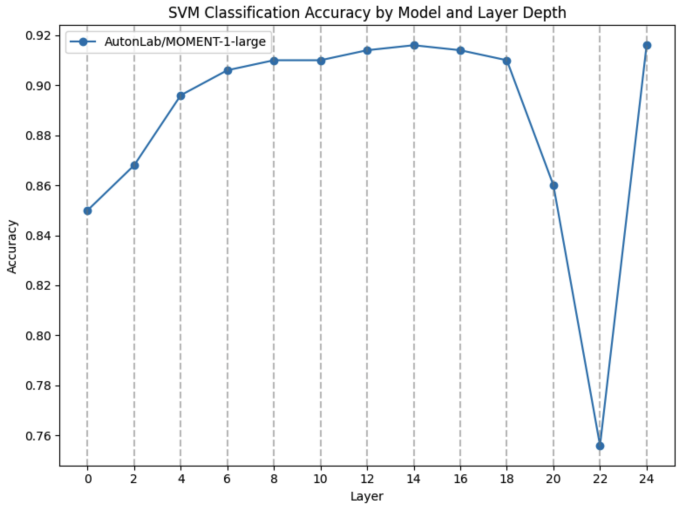

Fig. 4. SVM classification accuracy across different layers of the MOMENT model. The plot shows how the intermediate representations at various layers contribute to downstream task performance, with middle layers often providing the highest accuracy, indicating their importance in feature extraction.

B. Summary of Findings

These results underscore the effectiveness of our methodology in dissecting the internal mechanisms of time series models. The insights gathered from similarity analysis offer a preliminary understanding of how different layers contribute to model performance and feature processing.

Our findings provide a foundation for further research and development, aiming to enhance the design and application of time series models in various domains. Future work will focus on extending these analyses to more diverse datasets and exploring additional metrics and methods for deeper insights.

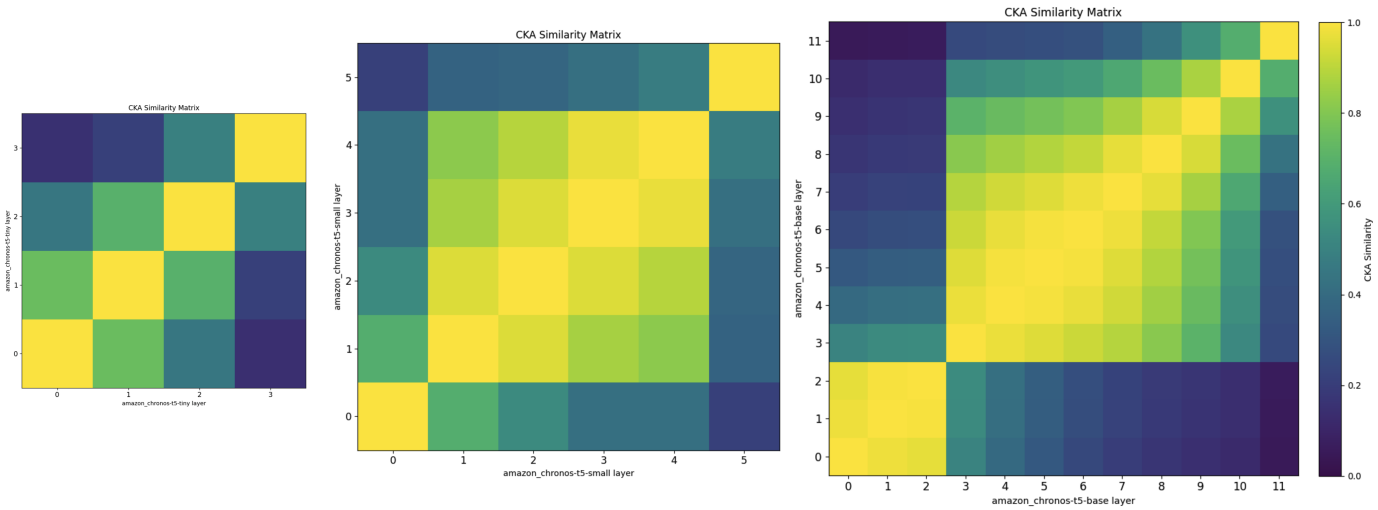

Fig. 5. Intra-similarity patterns across different model scales. The patterns are consistent, demonstrating that scaling the model preserves the similarity structure within layers. Each subplot represents the similarity heatmap for a specific model size.

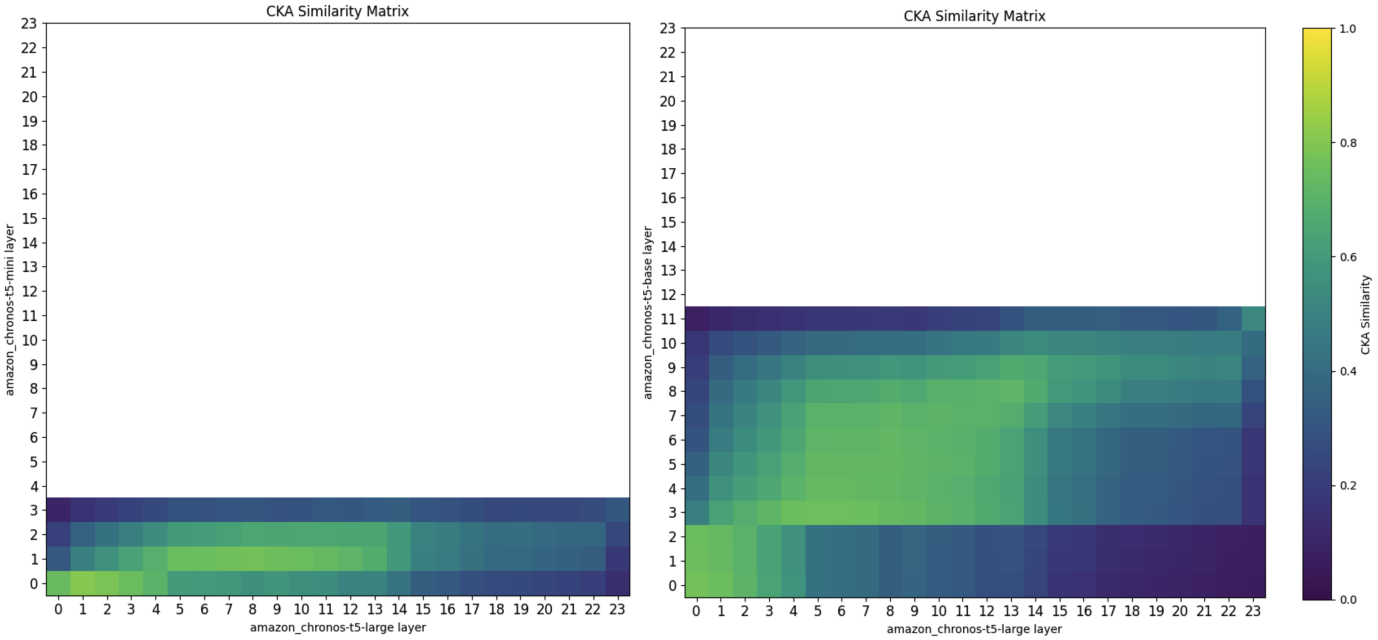

Fig. 6. Inter-similarity patterns across different model scales. The patterns show that model size influences highly the hierarchy of representation and although we have patterns as in the Fig. 5, there is little similarity between representations.

V. Conclusions

In this study, we conducted analysis of time series foundation models, focusing on similarity analysis of intermediate representations, contextualized time series representation, and feature attribution and analysis. Our findings provide several key insights:

- **Layer-wise Similarity Patterns** Intra-similarity patterns are preserved across different model scales, suggesting robustness in model architecture and training.

The integration of multiple similarity metrics (cosine similarity, SVCCA, and CKA) offered a robust framework for assessing representation similarity, while the evaluation of intermediate representations highlighted the importance of layer-specific features for downstream tasks. Our novel approach to feature attribution provided detailed insights into the roles of different model components in processing specific time series features.

These insights lay the groundwork for further research and development in time series modeling, suggesting avenues for optimizing model architectures and improving feature extraction techniques. Future work will involve extending these analyses to more diverse datasets and exploring additional metrics and methods to gain deeper insights into the internal workings of time series models.

A. Future Work

In addition to the current findings, future research will focus on two key areas:

a) Data Quality: Ensuring high-quality data is crucial for the robustness and reliability of time series models. Future work will involve developing methods to assess and improve data quality, handling issues such as missing values, outliers, and noise. Enhanced preprocessing techniques and robust data augmentation methods will be explored to create more reliable datasets for training and evaluation.

b) Inter-Dataset Representation Similarity: Understanding the similarities and differences in representations across different datasets can provide valuable insights into the generalization capabilities of time series models. Future studies will investigate inter-dataset representation similarity, aiming to identify common patterns and unique characteristics across various domains. This analysis will help in transferring knowledge between datasets and improving model performance on diverse time series tasks.

Overall, our study contributes to a better understanding of the internal mechanisms of time series foundation models, offering valuable guidance for their application and improvement in various domains. The continued exploration of data quality and inter-dataset representation similarity will further enhance the effectiveness and generalizability of these models.

ACKNOWLEDGMENT

This research wouldn't be possible without my mentors and coauthors; I couldn't dream of better people to learn from. I am deeply grateful to the Auton Lab for sponsoring my stay and fostering a vibrant space for idea exchange. Special thanks to Dr. John Dolan, Rachel Burcin, and Morgan Grimm for their efforts in making the RISS program a reality. Lastly, thank you to the entire RISS 2024 cohort for their support and camaraderie.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[3] S. J. Taylor, *Modelling financial time series*. world scientific, 2008.

[4] M. Goswami, B. Boecking, and A. Dubrawski, "Weak supervision for affordable modeling of electrocardiogram data," in *AMIA Annual Symposium Proceedings*, vol. 2021. American Medical Informatics Association, 2021, p. 536.

[5] S. H. Schneider and R. E. Dickinson, "Climate modeling," *Reviews of Geophysics*, vol. 12, no. 3, pp. 447–493, 1974.

[6] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "Moment: A family of open time-series foundation models," in *International Conference on Machine Learning*, 2024.

[7] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," in *Forty-first International Conference on Machine Learning*, 2024.

[8] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. Gordon Wilson, M. Bohlke-Schneider, and Y. Wang, "Chronos: Learning the language of time series," *arXiv preprint arXiv:2403.07815*, 2024.

[9] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Unified training of universal time series forecasting transformers," in *Forty-first International Conference on Machine Learning*, 2024.

[10] K. Rasul, A. Ashok, A. R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M. J. D. Bayazi, G. Adamopoulos, R. Riachi, N. Hassen, M. Biloš, S. Garg, A. Schneider, N. Chapados, A. Drouin, V. Zantedeschi, Y. Nevmyvaka, and I. Rish, "Lag-llama: Towards foundation models for probabilistic time series forecasting," 2024.

[11] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2021.

[12] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 782–791.

[13] A. Garza and M. Mergenthaler-Canseco, "Timegpt-1," 2023.

[14] J. Ye, W. Zhang, K. Yi, Y. Yu, Z. Li, J. Li, and F. Tsung, "A survey of time series foundation models: Generalizing time series representation with large language mode," *arXiv preprint arXiv:2405.02358*, 2024.

[15] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in neural information processing systems*, vol. 34, pp. 12 116–12 128, 2021.

[16] T. Nguyen, M. Raghu, and S. Kornblith, "Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth," in *International Conference on Learning Representations*, 2021.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[18] J. Ferrando, G. Sarti, A. Bisazza, and M. R. Costa-jussà, "A primer on the inner workings of transformer-based language models," *arXiv preprint arXiv:2405.00208*, 2024.

[19] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," *Advances in neural information processing systems*, vol. 30, 2017.

[20] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International conference on machine learning*. PMLR, 2019, pp. 3519–3529.

[21] B. Wang, X. Yue, Y. Su, and H. Sun, "Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization," 2024. [Online]. Available: https://arxiv.org/abs/2405.15071