Validation, Robustness, and Accuracy of Perturbation-based Sensitivity Analysis Methods for Time-Series Deep Learning Models

Zhengguang Wang

University of Virginia
Computer Science Department
zw4re@virginia.edu

Introduction

Deep Learning techniques for analyzing time-series data and developing associated interpretation methods have gained significant importance, given that a substantial portion of the world's information is inherently time-dependent. With state-of-the-art deep learning models and appropriate interpretation methods, we identify temporal patterns from the time-series data. The outcomes can be used for prediction and inference tasks in many fields, thus aiding scientific discoveries and helping governments make decisions.

I want to undertake studies to evaluate **Interpretability Methods for Time-Series Deep Learning**. Sensitivity analysis assesses how input changes affect the output, constituting a key component of interpretation. Among the post-hoc interpretation methods such as back-propagation, perturbation, and approximation, I will investigate perturbation-based sensitivity Analysis methods on modern Transformer model to benchmark their performances. My research focuses on evaluating the validation, robustness, and accuracy of these methods and the social impact of their applications. For example, if these methods yield a similar ranking of the importance of factors, then policymakers will be more effective in making policies based on the result of sensitivity analysis methods.

Background

Sensitivity Analysis is a post-hoc method that has been widely used in the field of deep learning interpretability. It can be categorized into three groups:

- Back-propagation group: Methods like Deep Lift (Shrikumar, Greenside, and Kundaje 2019) and Integrated Gradients (Sundararajan, Taly, and Yan 2017) use gradient-based approaches to attribute model outputs to input features.
- Perturbation group: Methods such as Feature Ablation (Meyes et al. 2019)and Feature Occlusion (Zeiler and Fergus 2014) involve altering or masking individual attributes to observe changes in outputs. Morris Method (Morris 1991) involves varying one parameter at a time while holding others constant, often employing a grid of points in the parameter space.

- Approximation group: Methods such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2016) compute an approximate explanation to a complex model, in a simpler and interpretable way such as fitting an Ordinary Least Sqaures (OLS) /Lasso regression.

In this work, I will study the perturbation group because of its model-agnostic, non-parametric, and interpretable nature.

Prior work done by the Applicant

My research focus is based on two prior works: **1) Interpreting County-Level COVID-19 Infections Using Deep Learning for Time Series** (Islam et al. 2024), in which I reproduced existing work, and **2) Interpreting Time Series Transformer Models and Sensitivity Analysis of Population Age Groups to COVID-19 Infections** (Islam and Fox 2023) project, in which I have conducted sensitivity analysis experiments. We applied the Morris sensitivity analysis method which is defined as:

$$Sensitivity(X, i) = \frac{f(x_1, x_i + \Delta, \ldots, x_k) - f(X)}{\Delta} \quad (1)$$

where Δ is a small change. In short, the contributions are the following:

- 1. Collecting eight different population age group features and COVID-19 cases info for 2years from 3,142 US counties.
- 2. Train a Temporal Fusion Transformer on the dataset to predict COVID-19 cases for next 15 days using past 13 days of input.
- 3. Extend the original Morris method to propose a scaled Morris index. Then calculate the sensitivity of the age group features and rank the age groups by their sensitivity scores.
- 4. Use feature ranks to globally interpret the sensitivity of age groups. Then finally evaluate the ranking with ground truth aggregated from reported cases by those age groups.

Approach

Based on the prior work, I will implement several perturbation-based sensitivity analysis methods and apply them to the U.S county-level COVID-19 cases dataset. The features of interest is the percentage of age groups in the

population.To examine the robustness of these sensitivity analysis methods, I will train other popular time-series deep learning models like DLinear (Li et al. 2019), Autoformer (Wu et al. 2021), PatchTST(Liu et al. 2023), and Times-Net(Xiao et al. 2023) to the same U.S county-level COVID-19 cases dataset and apply the sensitivity analysis methods.

The feature importance produced by the sensitivity analysis methods will be ranked, and we rank the ground truth is aggregated from reported cases by those age groups as well. From here, the evaluation metric, Spearman rank correlation coefficient, can be applied. A value of 1 for the Spearman rank correlation coefficient means perfectly monotonically increasing relationship while a value of -1 implies no monotonic relation.

Evaluation

Components	Implementations
SA Methods	Feature Ablation, Feature Occulsion, and Morris Method
DL Models	TFT, TimesNet, Autoformer, DLinear, PatchTST
Dataset	2-year U.S County-Level COVID cases with 8 age feature groups
Ground Truth	Cases of certain age groups
Evaluation	Spearman correlation coefficient

Table 1: Experiment Components and Implementations. "SA" refers to Sensitivity Analysis, and "DL" refers to Deep Learning

The experimental setup is described in detail in Table 1. I aim to answer three research questions:

- Within the same model, do different sensitivity analysis (SA) methods yield comparable outputs and attribute importance rankings? (Validation)

- Using the same sensitivity analysis method, do different Deep Learning (DL) models impact the output of the sensitivity analysis? (Robustness)

- How well do the results from sensitivity analysis methods align with the ground truth? (Accuracy)

With the experimental data, I will first process the data into ranks within each experiment groups and apply Spearman correlation coefficient to observe the consistencies between different groups of experiments. Furthermore, to evaluate which methods produce the best result, their respective Spearman rank correlation coefficient with the ground truth will also be calculated. These experiments aim to provide benchmark results that answer my research questions on DL models and SA methods.

Discussion

Based on the answers to my three proposed research questions, I can conclude the validation, robustness, and accuracy of sensitivity analysis methods.

- The answer to my first research question will shed light on whether it is necessary to interpret a Deep Learn-

ing model with multiple perturbation-group post-hoc method.

- The answer to my second research question will provide insight into the robustness of sensitivity analysis. If they are robust, an reliable interpretation method that is insensitive to different models is achieved.

- The answer to my third research question will inform us on how confident we can trust the results generated from Sensitivity Analysis.

AI transparency can warrent trust, especially for policymakers who are using the outputs from sensitivity analysis to make decisions of great social impact.

Conclusion

In conclusion, my research focuses on the validation, robustness, and accuracy of Perturbation-based Sensitivity Analysis Methods. My research aims to evaluate if these methods yield comparable outputs, if different models impact sensitivity analysis results, and how well the feature importance from Sensitivity Analysis Methods align with the ground truth. The results will inform best practices for interpretability in time-series data. Consequently, policymakers can adopt policies with greater confidence, relying on insights from deep learning models and sensitivity analysis. Furthermore, my research carries great value in other related fields. For example, in Uncertainty Quantification, reliable Sensitivity Analysis methods will help determine how much uncertainty an individual source contributes to the total uncertainty in a simulated or experimental quantity, analyzing the effect of noisy data and perturbations.

References

Islam, M. K.; Valentine, T.; Sue, T.; Karmacharya, A.; Benham, L.; Wang, Z.; Kim, K.; and Fox, J. 2024. Interpreting Time Series Transformer Models and Sensitivity Analysis of Population Age Groups to COVID-19 Infections. In *2024 The Association for the Advancement of Artificial Intelligence AI4TS Workshop.*

Islam, T. V. R. W. L. D. M. M., Md Khairul; and Fox, J. 2023. Population Age Group Sensitivity for COVID-19 Infections with Deep Learning. ArXiv:2307.00751 [cs.LG], arXiv:2307.00751.

Li, Y.; Liu, T.; Wang, X.; Zhang, J.; Zhang, M.; and Chen, Y. 2019. DLinear: A Simple Way to Improve the Performance and Interpretability of Deep Learning Models. *arXiv preprint arXiv:1906.02799.*

Liu, C.; Cao, Z.; Cao, W.; Liu, Q.; and Zhang, T. 2023. PatchTST: A Transformer-based Model for Time Series Forecasting with Patch Tokenization and Splitting. *arXiv preprint arXiv:2301.02628.*

Lundberg, S. M.; and Lee, S.-I. 2016. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1605.06063.*

Meyes, R.; Lu, M.; de Puiseau, C. W.; and Meisen, T. 2019. Ablation Studies in Artificial Neural Networks. arXiv:1901.08644.

Morris, M. D. 1991. Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics*, 33(2): 161–174.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Model-agnostic interpretability of machine learning models. *arXiv preprint arXiv:1606.05386*.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2019. Learning Important Features Through Propagating Activation Differences. arXiv:1704.02685.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365.

Wu, S.; Wang, S.; Gao, F.; Zhang, X.; Zhu, H.; Gao, J.; Lu, S.; and Zhang, K. 2021. Autoformer: A Transformer-based Model for Time Series Forecasting. *arXiv preprint arXiv:2108.06051*.

Xiao, Z.; Liu, C.; Cao, Z.; Cao, W.; and Liu, Q. 2023. Times-Net: A Temporal 2D-variation Modeling Framework for Time Series Forecasting. *arXiv preprint arXiv:2302.00666*.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 818–833.