Consistency and Robustness of Perturbation-based Sensitivity Analysis Methods

Anonymous Submission

Introduction

Deep Learning for Time-Series data and their associated interpretation methods has become an important topic as much of the world information are stored sequentially. With state-of-the-art deep learning models and appropriate interpretation methods, we could mine information from 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 into the **Time Series Sensitivity Analysis Methods**. 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 see their performances and consistencies. My research focuses on evaluating the reliability and robustness of these methods and the social impact of their applications. For example, if these methods yield 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 such as Deep Lift (Shrikumar, Greenside, and Kundaje 2019) and Integrated Gradients (Sundararajan, Taly, and Yan 2017) employ gradient-based approaches to attribute the model's output to its 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 more interpretable way such as fitting an 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. 2023) and **2) Time Series Sensitivity Analysis of Population Age Groups in Multi-Horizon COVID-19 Forecasting** (Islam and Fox 2023) project. We applied the Morris sensitivity analysis method which is defined as:

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

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 the Temporal Fusion Transformer model on the dataset to predict COVID-19 cases for the next 15 days using the 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 subgroups (e.g. 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, Autoformer, PatchTST, and TimesNet to the same U.S county-level COVID-19 cases dataset and apply the sensitivity analysis methods.

The evaluation metric is the Spearman rank correlation coefficient while the ground truth is aggregated from reported cases by those age groups.

Evaluation

From my experiments setup in Table 1, there are dual-level research questions:

- Within the same model, I will examine whether different sensitivity analysis (SA) methods yield comparable outputs and attribute importance rankings. (Consistency)
- Using the same sensitivity analysis method, I will investigate if different Deep Learning (DL) models impact the output of the sensitivity analysis. (Robustness)

Table 1: Experimental Setup

Components	**Implementation**
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

If I succeed in my experiments, the experimental results would suffice to answer the my research questions.

Discussion

Based on the answers to my two proposed research questions, I can conclude the robustness and consistency of sensitivity analysis methods. If they are consistent with each other and robust, a singular method is sufficient for conducting sensitivity analysis, and policymakers can be more confident to utilize the conclusions from sensitivity analysis. Otherwise, a cautious approach is warranted, 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 Consistency and Robustness of Perturbation-based Sensitivity Analysis Methods. Moreover, my research aims to determine if these methods yield comparable outputs and attribute importance rankings within the same model, and if different models impact sensitivity analysis results. The results will inform best practices for interpretability in time-series data. Consequently, policymakers can approach policy-making 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.; Liu, Y.; Erkelens, A.; Daniello, N.; Marathe, A.; and Fox, J. 2023. Interpreting County-Level COVID-19 Infections using Transformer and Deep Learning Time Series Models. In *2023 IEEE International Conference on Digital Health (ICDH)*, 266–277.

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.

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.

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