Project Proposal

Amirhossein Sohrabbeig, Matin Tavakoli

March 2022

1 Goal

In this project, we try to unwrap the black-box behavior of recurrent neural network (RNN) models. To do so, we will first study the state-of-the-art methods of interpretability for RNNs and implement a few of them. We will then attempt to compare them based on the existing evaluation metrics. Our hope is that this study further enables the analysis of such models, allowing them to be observed more thoroughly and with great care.

2 Problem Statement

RNNs have achieved prominent results in different domains in recent years, ranging from machine translation problems to speech recognition tasks. However, we do not know what the model has actually learned due to its black-box nature. This black-box behavior makes the process of troubleshooting the model's mistakes and validating their functioning a seemingly vague process. Therefore, we need to find ways to understand what exactly the model has learned to overcome these issues.

3 Previous Work

Existing interpretable AI methods can be categorized based on complexity, scope, and model-relatedness.

A more complicated model is more difficult to understand and interpret in terms of complexity. The easiest method of interpretability is to utilize an

intrinsically explainable model, such as a decision tree, which is considered interpretable due to its simple structure. However, these models typically perform worse than more complicated models. Post-hoc interpretability is a different method of interpretability in which the explanation is generated after model training by creating perturbations in the input and then observing the model outputs for the modified inputs. Most interpretable AI studies fall into this category [1–3].

In terms of scope, global methods describe how features affect the prediction on average. In contrast, local methods aim to explain individual predictions. Mokhtari et al. [4] use SHAP [5] to interpret financial time series models, where the contribution scores provided by SHAP allow financial experts to better understand the model's decisions.

Another way to classify interpretable AI methods is whether they are model-specific or model-agnostic. Model agnostic methods are usually favored since they can be applied to any type of machine learning model. On the other hand, model-specific approaches can take advantage of a machine learning model's intrinsic features while still being computationally less expensive. Lundberg and Lee [5] developed different SHAP algorithms for Post-hoc interpretability.

4 Approach

This study will examine and evaluate local explanations provided by model agnostic Post-hoc methods like SHAP and LIME [6] on different models trained on four multivariate time-series datasets. We will also assess insights provided by a global model-specific interpretable LSTM called IMV-LSTM [7] using both quantitative and qualitative measures.

5 Metrics

In order to evaluate and compare the existing interpretability methods, we consider two evaluation metrics, namely, Area Over the Perturbation Curve Regression (AOPCR) and Ablation Percentage Threshold (APT).

AOPCR [8] determines the influence of the top k features introduced by the interpretability method. It is formulated as the average perturbation of the model between the presence and absence of the top features. The area over the perturbation curve for regression at time horizon τ, denoted as AOPCR, is obtained as:

$$AOPCR_\tau = \frac{1}{K} \sum_{k=1}^{K} F_\tau(X_t) - F_\tau(X_{t,\setminus 1:k}) \tag{1}$$

Then, the area over the perturbation for regression is the average of all the time steps $\tau = 1, ..., t_0$, where

$$AOPCR = \frac{1}{t_0} \sum_{\tau=1}^{t_0} AOPCR_\tau \tag{2}$$

On the other hand, APT [8] is based on the switching point of the model. The switching point is defined as a point above and below the original prediction by a predefined threshold distance. APT then calculates the percentage of features that need to be removed before the prediction switches to another class. The lower the score, the lower percentage of features needed to remove and thus, resulting in higher local fidelity. We define APT at time horizon τ with significance factor α as follows:

$$APT_{\tau,\alpha} = \arg \min_{k \in 1,...,J} \frac{k}{J} \tag{3}$$

such that:

$$F_\tau(X_t)(1 + \alpha) > F_\tau(X_{t,\setminus 1:k}) \tag{4}$$

The total APT is a simple average over the time index:

$$APT_\alpha = \frac{1}{t_0} \sum_{\tau=1}^{t_0} APT_{\tau,\alpha} \tag{5}$$

References

[1] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[2] F Mujkanovic. Explaining the predictions of any time series classifier. *Bachelor's Thesis, Universität Potsdam, Potsdam, Germany*, 2019.

[3] Beau Norgeot, Dmytro Lituiev, Benjamin S Glicksberg, and Atul J Butte. Time aggregation and model interpretation for deep multivariate longitudinal patient outcome forecasting systems in chronic ambulatory care. *arXiv preprint arXiv:1811.12589*, 2018.

[4] Karim El Mokhtari, Ben Peachey Higdon, and Ayşe Başar. Interpreting financial time series with shap values. In *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, CASCON '19, page 166–172, USA, 2019. IBM Corp.

[5] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[6] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[7] Tian Guo, Tao Lin, and Nino Antulov-Fantulin. Exploring interpretable LSTM neural networks over multi-variable data. *CoRR*, abs/1905.12034, 2019.

[8] Ozan Ozyegen, Igor Ilic, and Mucahit Cevik. Evaluation of interpretability methods for multivariate time series forecasting. *Applied Intelligence*, 52(5):4727–4743, 2021.