Research Statement

Thu Trang Nguyen
thu.nguyen@ucdconnect.ie

Introduction

The last decade witnesses the enormous integration and impact of machine learning into everyday life. Machine learning algorithms nowadays work stunningly well and grow superficially complex with models of millions parameters [1, 2]. However, we are far behind in explaining why these algorithms often work so well and occasionally fail to perform [3].

This unmatched growth of complexity and explainability of many machine learning algorithms, including those dealing with time series data, undermines application of the technology in critical, human-related areas such as medicine, healthcare, and finance [4]. As sequential, time series data is prevalent in these applications [5–7], Time Series Classification algorithms often call for reliable algorithm explanations [8, 9]. This explanation is usually presented in the form of a *saliency map* [10], highlighting the parts of the time series which are *informative* for the classification decision.

Recent efforts both in designing intrinsically explainable machine learning algorithms, as well as building post-hoc methods explaining black-box algorithms, have gained significant attention [11–15]; yet, these efforts present us a new challenge: *How to assess and objectively compare such methods?* In other words, if two explanation techniques give different explanations, i.e., two different saliency maps, which technique and explanation should we trust?

Assessing and comparing explanations is a non-trivial problem, requiring an effective and scalable method to filter out problematic techniques and to identify trustworthy ones. This assessment helps alleviate the need for, or at least reduce some of the effort required to, conducting user-studies which are very difficult to reproduce [4]. Identification of useful explanation techniques for a specific type of problem not only enables machine learning to make impact in critical application areas, but also pinpoints the potential problems, such as biases, that may exist in the training data [16]. More importantly, this evaluation further empower machine learning to discover new knowledge from the data. Often times, that piece of insights is even more important than the model correct decisions. For example, given a complex, sequential gene database of people with and without auto-immune diseases, it is interesting to know which part of the gene might associate with this specific condition. The knowledge enables identification of high-risk group of people and recommendation of suitable diets and lifestyles that prevent the symptoms to develop and help these people have better life quality.

I specifically seek to understand and finding solutions to the problem of identifying legible explanations (in the form of *saliency maps* [10]) to machine learning algorithms, including time series classification algorithms.

- *Given a time series classification task and a dataset, how can we evaluate if a given explanation method is useful? How can we quantitatively compare different explanations? What metrics should be used to quantify the informativeness of explanation methods?*

- *What are good explanation methods for a specific type of datasets? How robust are these methods when noises are present?*

- *How to design an optimized explanation method based on the proposed evaluation framework?*

My response to these research questions is: 1) to develop approaches to automatically and computationally assess the quality of explanation methods for Time Series Classification (TSC), and 2) to devise metrics to quantify such assessments.

Current Research

My work focuses on quality assessment of explanation methods using explanation-driven perturbation. As a good saliency-based TSC explanation highlights the discriminative parts of the time series, perturbation of these parts will remove critical, class-specific information of the data and make it harder for machine learning models to make correct prediction.

Step-wise Explanation-based Perturbation. A saliency-based time series explanation can either be presented as a heatmap or a non-negative numeric vector (w) that contains the importance weight of each time step. As a result, discriminative parts are identified by a threshold k ($0 \leq k \leq 100$) that represents the top k-percent of w. Varying this threshold k allows step-wise explanation-based perturbation of data, i.e. step-wise replacement of most important parts with noises.

A Model-Agnostic Framework to Comparing Explanations. When multiple explanation methods are available, application users naturally raise a question: *What method should I trust?* My proposal to answer this question based on the idea that good explanations highlights informative, critical parts of the data; thus, making changes to such parts will reduce accuracy of an independent classifier more than those based on other parts. In other words, perturbation of data based on a good explanation will affect the independent classifier most significantly. Our work presents a framework that first builds a time series classifier using the original, non-perturbed training data. This classifier serves as the evaluation classifier for the explanation methods in question, i.e. a *referee classifier*. Each explanation method is evaluated independently in the following step, in which the explanation-in-assessment is used to generate perturbed/noisy datasets. Each dataset corresponds to a value of the threshold k ($0 \leq k \leq 100$). Using the referee classifier, we measure the accuracy in each perturbed datasets and compare the magnitude of accuracy reduction at each step. Best explanations generally experience huge accuracy drops at the first few steps, while others have marginal changes.

The intuition of this process is similar to have an independent referee to judge whether one explanation is making more sense than others. In that fashion, it is possible to invite multiple referees, i.e. using multiple referee classifiers, to assess performance of the explanation methods that are evaluated.

Extension of this framework covers the situation when only one explanation method is available. Two possible options of assessing explanation quality in this case are to compare the method in question with 1) its reverse one, where the most salient region of one is the least salient region of the other; or 2) a random one, in which the salient region is determined randomly.

Informativeness of An Explanation: An Evaluation Metrics. We quantify the informativeness of an explanation using the relationship between the accuracy of a referee classifier on test datasets perturbed at different threshold k. The impact of the explanation methods is estimated using the area under the (explanation) curve described by accuracy at each step k ($0 \leq k \leq 100$), using the trapezoidal rule. We call this metric Explanation-AUC. A better explanation method should have a lower Explanation-AUC than other methods. Ranking of this metric represents ranking of explanation quality [17].

My collaborators and I have demonstrate that our proposed approach, using step-wise explanation-based perturbation to create a model independent framework to quantitatively compare and evaluate explanation methods by calculating Explanation-AUC, can correctly point out a legible explanation which is confirmed by a domain-expert in a human activity recognition dataset [18]. Our work provides a ready-to-use recipe that is compatible with any referee classifiers and saliency-based interpretations, including machine-generated and expert-based explanations, a resulted explanation method that is useful in detecting data and model potential biases, and a tool to automate insights learning from data.

Important questions remain. While we are continuously improve the proposed algorithms (that we already have work in progress) to handling different pertubation types, involving more independent referee classifiers, and testing with different types of time series data, we are still in search for a final design of an optimized explanation method based on our proposed framework, and an recommendation of strengths and limitations of each commonly used explanation methods [11–15]. That results would almost certainly shed new light to many applications, from healthcare, medicine, finance, to any areas that involves sequential data.

Future Directions

Optimized Saliency-based Explanation Method. The notion of informativeness and the proposed framework to evaluate explanation methods help identify the most informative method out of many available explanation techniques. The final, long term aim of my research is to design a new, optimized explanation method that maximizes the informativeness metric (Explanation-AUC) using our evaluation framework. One potential option is to look into an iterative process for tweaking one explanation to optimize informativeness metric.

Learning New Knowledge from Data. Explanations are often thought of as a way to confirm existing knowledge about the data. Yet, there are situations which humans have little knowledge of grouth-truth *explanations*, for example subsequences of gene/DNA/protein chains that are responsible for specific conditions. Our research aims to perform a demonstration of this application, in which we use optimized explanation to find out insights from data that was previously unknown to us. The result would be presented to a domain expert to verify correctness. We believe that this demonstration will show an powerful use-case of Explainable Machine Learning beyond confirming algorithm's inner workings and fairness.

References

1. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* **abs/1810.04805.** arXiv: `1810.04805`. `http://arxiv.org/abs/1810.04805` (2018).
2. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *CoRR* **abs/2005.14165.** arXiv: `2005.14165`. `https://arxiv.org/abs/2005.14165` (2020).
3. Goodfellow, I. J., Shlens, J. & Szegedy, C. *Explaining and Harnessing Adversarial Examples* 2014. `https://arxiv.org/abs/1412.6572`.
4. Doshi-Velez, F. & Kim, B. *Towards A Rigorous Science of Interpretable Machine Learning* 2017. arXiv: `1702.08608 [stat.ML]`.
5. Petitjean, F. *et al. Dynamic Time Warping Averaging of Time Series Allows Faster and More Accurate Classification* in *2014 IEEE International Conference on Data Mining* (2014), 470–479.
6. Ramgopal, S. *et al.* Seizure detection, seizure prediction, and closed-loop warning systems in Epilepsy. *Epilepsy & behavior : E&B* **37C,** 291–307 (Aug. 2014).
7. Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R. & Havinga, P. *Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey* in (Jan. 2010), 167–176.
8. Bostrom, N. & Yudkowsky, E. *The Ethics of Artificial Intelligence*
9. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery* **14,** 1611–1617. ISSN: 1861-6429. `https://doi.org/10.1007/s11548-019-02039-4` (Sept. 2019).
10. Adebayo, J. *et al. Sanity Checks for Saliency Maps* in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Montréal, Canada, 2018), 9525–9536.

11. Zhou, B., Khosla, A., A., L., Oliva, A. & Torralba, A. Learning Deep Features for Discriminative Localization. *CVPR* (2016).

12. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* **abs/1602.04938.** arXiv: `1602.04938`. `http://arxiv.org/abs/1602.04938` (2016).

13. Selvaraju, R. R. *et al.* Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR* **abs/1610.02391.** arXiv: `1610.02391`. `http://arxiv.org/abs/1610.02391` (2016).

14. Smilkov, D., Thorat, N., Kim, B., Viégas, F. B. & Wattenberg, M. SmoothGrad: removing noise by adding noise. *CoRR* **abs/1706.03825.** arXiv: `1706.03825`. `http://arxiv.org/abs/1706.03825` (2017).

15. Lundberg, S. M. & Lee, S. I. *A Unified Approach to Interpreting Model Predictions* (eds Guyon, I. *et al.*) 2017. `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

16. Kim, B. *et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)* 2018. arXiv: `1711.11279 [stat.ML]`.

17. Agarwal, S., Nguyen, T. T., Nguyen, T. L. & Ifrim, G. *Ranking by Aggregating Referees: Evaluating the Informativeness of Explanation Methods for Time Series Classification* in *Advanced Analytics and Learning on Temporal Data* (eds Lemaire, V. *et al.*) (Springer International Publishing, Cham, 2021), 3–20. ISBN: 978-3-030-91445-5.

18. Nguyen, T. T., Le Nguyen, T. & Ifrim, G. *A Model-Agnostic Approach to Quantifying the Informativeness of Explanation Methods for Time Series Classification* in *Proceedings of the 5th Workshop on Advanced Analytics and Learning on Temporal Data at ECML* (2020).