Experiments With and Development of Neural Additive Models as an Interpretable Machine Learning Method

Prasil Adhikari
k12049801@students.jku.at

ABSTRACT

Interpretable Machine Learning deals with extracting human understandable insights from any machine learning model. To interpret a model, features in the model and the effect of these features on the predictions are of utmost importance. This report summarises the work of Agarwal et.al in the paper titled "Neural Additive Models: Interpretable Machine Learning with Neural Nets". The Neural Additive model uses neural network architecture to explain with transparent feature graphs, thus empowering the interpretability achieved by using Generalized Additive Models and Explainable Boosting Machines. The architecture and mechanism, experiments and conclusion, and the functionality of Neural Additive Models are analysed in a comprehensive manner.

1 INTRODUCTION

Although Deep Neural Networks have been demonstrated to have excellent performance, their adoption in the real-world applications and mainly high stake domains has been hampered by the fact that they are challenging to interpret. The black box nature of the machine learning model is investigated by the interpretability measures. For many real-world applications, prediction accuracy is only a limited and partial description, and interpretability expands the information by dissecting the underlying transformations that the model has processed. This helps in debugging and detecting bias, improving feature engineering, reliable decision making, and building trust on the system. The challenge lies with the trade-off of accuracy with the increased interpretability.

Neural Additive Models (NAMs) [1], is built on the foundation of Generalized Additive Models (GAMs) [5] and Explainable Boosting Machines (EBMs) [8] and it achieves similar accuracy as standard Deep Neural Networks (DNNs) while simultaneoulsy also maintaining the transparency and interpretability as Logistic Regression. This is therefore a state-of-the-art interpretable model based on the neural network architecture and the paper [1] showcases its usefulness with different experimentation and observations.

2 BACKGROUND

NAMs combine the inherent intelligibility of GAMs while maintaining the expressivity of DNNs. GAM is an extension of the multiple linear model by replacing each linear component with smooth nonlinear function. The additivity allows for the interpretability of each predictor, but also is the main limitation and GAMs might miss the nonlinear interactions among the predictors. Moreover, GAMs are only interpretable if the features they are trained on are interpretable. GAMs have the form

$$g(E_Y(y|x)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_k(x_k)$$

where x is the input with k features, y is the target variable, g is the link function (e.g., logistic regression) and each f (a univariate shape function) is parametrized by a neural network with expectation zero.

Building on the idea of GAMs, EBMs are a tree-based, cyclic gradient boosting generalized additive model with automatic interaction detection. It has the form

$$g(E[y]) = \beta_0 + \sum f_k(x_k)$$

where g is the link function that adapts the GAM to different settings such as regression or classification. Although being often as accurate as the state-of-the-art black box models, EBMs are slower to train and GAMs trained with the boosted trees are not differentiable, thus reducing the flexibility.

NAMs offer the advantages over GAMs and EBMs with its neural network structure, differentiability, composability, flexibility, and multitask learning. They are faster to train, able to train more complex interpretable models, achieve comparable accuracy, and provide intrinsically interpretable models.

3 RELATED WORK

3.1 Architecture

NAMs learn a linear combination of the separate DNN subnets for each input feature, i.e., they are learned in parallel and additively combined. Then the combination is subjected to a nonlinear function (e.g. Sigmoid) to generate the prediction. As based on the neural network architecture, each subnets are differentiable and they can be trained iteratively via backpropagation. These networks can learn arbitrarily complex relationships (shape functions) between their input feature and the output. The impact of a feature on the prediction does not rely on other features and can be understood by visualizing its corresponding shape function, i.e., each learned subnets can be visualised. After training, these subnets can be replaced with feature graphs, which show how the NAM made its prediction.

In order to accurately learn additive models, it is necessary to model functions with sharp jumps, which are common in real-world datasets. Rectified Linear Unit (ReLU) activations were used in the experiments that were carried out in the original paper to model the over-parametrized NNs based on the observations that standard neural networks fail to model highly jumpy 1D functions. However, this tends to overfit, and ReLU's biasedness toward smoothness prevents them from understanding large fluctuations. To overcome this neural network failure, a special activation function Exp-centered (ExU) hidden units is proposed.

$$h(x) = f(e^w * (x - b))$$

i.e., they simply learn the weights in the logarithmic scale with inputs shifted by bias. Because the ExU unit calculates a linear function of the input, the slope of which can be extremely steep with low weights, it is simpler to alter the output during training. The ExU unit makes the model for fitting jumpy functions easier to learn. Any activation function can be used with ExU-units, but ReLU-n works well in practice because it ensures that ExU works well within a small input range without affecting the overall behavior. Strong regularization is necessary to avoid overfitting with ExU, and when fitting NAM with ExU, regularization methods like dropout, weight decay, output penalty, and feature dropout are typically used.

3.2 Characteristics

NAM allows global interpretations by visualizing each sub network as a one-dimensional function. With such visualization, one can easily detect biases w.r.t sensitive features (such as risk and gender) and this representation also allows direct additives on the graph to remove undesirable biases from the learned model. Similarly, NAM also allows faithful explanations to individual predictions (local interpretations), and it achieves all of the above while maintaining near maximum arc accuracies to standard DNNs.

In addition, the comprehension and intelligibility of NAM is also evident from its visualization, where it is self-interpretable by plotting the individual independent shape function which parametrizes each individual feature. These shape function plots provide not just an explanation but an exact and accurate representation of how NAMs compute a prediction. It has been shown by plotting the shape functions of MIMIC II-ICU dataset [9]. It is to note than the NAMs are not causal models, and do not explain how the model came to make the assumptions that it did. However, the shape plots reveal exactly how the model arrived at its predictions.

Also, when comparing the accuracy of NAM with different baselines on different datasets, NAMs achieve comparable performance to EBMs with an competitive edge to EBM with its differentiability and composability.

3.3 Multitask Learning

The composability of NAM also makes it easier to train multiple subnets per feature. In case of the Multitask learning, the model jointly learns a task specific weighted sum over their inputs. The final prediction score is then calculated by adding bias to the outputs associated with each task. For each task, the NAM architecture jointly learn many feature representations while maintaining the interpretability and modularity.. The multitask functionality of NAM has been demonstrated [1](Section 4.2) using synthetic data and COMPAS recidivism data [10].

It is observed from the COMPAS recidivism data that in this case the multitask NAM achieve less error than single task NAM while fitting the model precisely. In addition, it is observed than in some settings, multitask learning can increase accuracy and intelligibility by learning task-specific shape plots that expose task-specific patterns in the data that would not be learned by single task learning. It was observed with experiments on COMPAS recidivism data that the multitask NAM reveals different relationships between race, charge degree, and recidivism risk for men and women while achieving slightly higher overall accuracy. It also concluded that the transparency and modularity of NAMs allows the detection of unanticipated biases in data and simple correction of the bias in the learned model. The extension of NAM to multitask setting is intuitive and provides a competitive edge to the tree-based GAMs and also makes the Multitask NAM architecture powerful in the real-world applications.

3.4 Higher Order NAMs

The differentiability of NAM allows them to train more intricate interpretable models, which was also showcased by training an interpretable parameter generation model for COVID-19 [1](Section 4.1). This demonstrates the value of differential nonlinear additive models, such as NAMs, however because NAM lacks feature interactions, it is difficult to analyze correlations between individual features.

In a recent study by Minkyu et. al., there is the proposition of Higher Order NAM (HONAM) [7], a novel interpretable machine learning method that can model arbitrary orders of feature interactions. The HONAM has been developed by redefining the existing NAM architecture to model high-order feature interactions. A novel feature interaction method for high interpretability has also been proposed. Moreover, there is proposition of the ExpDive unit [7], which overcomes the limitations of the ExU unit, to effectively learn sharp shape functions. The ExpDive Unit is defined as follows

$$h(x) = \sigma(x - b) * (exp(W) - exp(-W))$$

Since the ExpDive unit exponentially diverges in both the negative and positive ranges, it is not only suitable for learning sharp shape functions, but also solves the vanishing gradient problem. In addition, it has been shown that it is possible to create a fair HONAM by eliminating the biases of HONAM, which leads towards the realization of trustworthy AI.

4 CONCLUSION

The report summarized the Neural Additive Models building on the foundations provided by Generalized Additive Models and Explainable Boosting Machines. The architecture and characteristics of the NAM has been described clearly, and the extension of simple NAM architecture to the Multitask setting has been explored. Moreover, it touches the advancement towards HIgher Order NAMs with a novel activation function. The recent development in NAM has seen the utilisation of the NAM architecture in various real-world scenarios, like explainable heart attack prediction [2], the machine learning survival model (SurvNAM) explanation [11], neural additive models for exploring multivariate Nowcasting problems [6] , monotonic NAM for Credit scoring [4], Neural generalized additive models (NODE-GAMs) [3]. As being able to interpret the black box models with such a high precision, there is a wide range of research, experimentation and development future for NAMs with utilisation in image data, sequential data, and integrating with other standard neural network architectures.

REFERENCES

[1] agarwal2021neuralAgarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R. Hinton, GE. 2021. Neural additive models: Interpretable machine learning with neural nets Neural additive models: Interpretable machine learning with neural nets. Advances in Neural Information Processing Systems344699–4711.

[2] balabaeva2022neuralBalabaeva, K. Kovalchuk, S. 2022. Neural Additive Models for Explainable Heart Attack Prediction Neural additive models for explainable heart attack prediction. Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part III Computational science–iccs 2022: 22nd international conference, london, uk, june 21–23, 2022, proceedings, part iii (113–121).

[3] chang2021nodeChang, CH., Caruana, R. Goldenberg, A. 2021. Node-gam: Neural generalized additive model for interpretable deep learning Node-gam: Neural generalized additive model for interpretable deep learning. arXiv preprint arXiv:2106.01613.

[4] chen2022monotonicChen, D. Ye, W. 2022. Monotonic Neural Additive Models: Pursuing Regulated Machine Learning Models for Credit Scoring Monotonic neural additive models: Pursuing regulated machine learning models for credit scoring. Proceedings of the Third ACM International Conference on AI in Finance Proceedings of the third acm international conference on ai in finance (70–78).

[5] hastie2017generalizedHastie, TJ. 2017. Generalized additive models Generalized additive models. Statistical models in S Statistical models in s (249–307). Routledge.

[6] jo2022neuralJo, W. Kim, D. 2022. Neural Additive Models for Nowcasting Neural additive models for nowcasting. arXiv preprint arXiv:2205.10020.

[7] kim2022higherKim, M., Choi, HS. Kim, J. 2022. Higher-order Neural Additive Models: An Interpretable Machine Learning Model with Feature Interactions Higher-order neural additive models: An interpretable machine learning model with feature interactions. arXiv preprint arXiv:2209.15409.

[8] nori2019interpretmlNori, H., Jenkins, S., Koch, P. Caruana, R. 2019. Interpretml: A unified framework for machine learning interpretability Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223.

[9] saeed2011multiparameterSaeed, M., Villarroel, M., Reisner, AT., Clifford, G., Lehman, LW., Moody, G.Mark, RG. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Critical care medicine395952.

[10] ProPublicathejefflarson. 2016. COMPAS Data and analysis for ‘Machine Bias’. Compas data and analysis for ‘machine bias’. https://github.com/propublica/compas-analysis.

[11] utkin2022survnamUtkin, LV., Satyukov, ED. Konstantinov, AV. 2022. SurvNAM: The machine learning survival model explanation Survnam: The machine learning survival model explanation. Neural Networks14781–102.